The evolving role of energy storage systems in AI data center load management

Part one of a two-part series
By Seth Miller, Ben Campbell, Joshua Sorensen
March 6, 2026

The evolving role of energy storage systems in AI data center load management

Key takeaways

- AI data centers combine huge power demands with rapid fluctuations, reshaping how operators, technology providers, and utilities shape infrastructure.
- Data centers consume so much power that operators must consider how to match data center demand to utility supply instead of matching supply to demand.
- AI data center power strategies must address four distinct but interrelated challenges: bridging short-duration outages, smoothing ultrafast load fluctuations, managing peak demand, and accelerating access to grid capacity.

Operators are deploying AI data centers around the world wherever power is available at scale, reliable, and sufficiently cheap. Data centers use large amounts of electricity to power the AI processes that produce revenue. The synchronized behavior of AI workloads concentrates demand changes across thousands of graphics processing units (GPUs) at once. This turns routine fluctuations in process workload into facility-scale power spikes that require fast, high-powered, and tightly coordinated solutions. Data centers also need their power supply to remain stable and consistent. This, coupled with the size of their power demands, slows the rate at which data center build-out can occur. Understanding all these challenges lays the foundation for evaluating which battery solutions matter and why they're becoming increasingly necessary to data center development.

Uncertainty around long-term workload growth and profitability increases risk for utilities. Operators are asking to invest in firm infrastructure for loads whose longevity isn't yet proven. Huge power demands and rigid reliability standards are outpacing the grid's slow interconnection process.

Access to power is now among the most serious constraints on AI data center build-out. These dynamics create a set of recurring electrical challenges that every AI data center must address to remain stable, efficient, and commercially viable.

The massive power needs and strict reliability requirements of data centers are running up against grid connection processes that move far more slowly. This convergence of load, speed, and stability needs fundamentally reshapes facility power architecture. As a result, a wider range of solutions has emerged, including uninterruptable power supply (UPS) systems, battery energy storage systems (BESS), generators, supercapacitors, and software controls, each suited to different operational needs and time scales. For utilities and storage providers, it's important to understand data centers' power challenges, business priorities, and the growing range of solutions. Utilities must plan for loads that are not only large but dynamically volatile. Storage providers must design systems capable of delivering unprecedented combinations of response speed, power density, cycling performance, safety, and cost efficiency. These issues are critical because growing workloads, data center profitability, and AI and storage technologies are all changing quickly.

Stakeholders now must make capital-intensive decisions under shifting technical and economic assumptions. In this environment, rigorous evaluation of power challenges, customer economics, and solution trade-offs are essential for managing investment risk, maintaining system reliability, and developing effective strategies for participating in the rapidly evolving data center market.

How can utilities and data centers form strategic partnerships?

In our exchange Smart loads and smart strategies: Utility programs at the data center frontier, we explored how utilities can engage local data centers as partners for energy efficiency, demand response, and load flexibility opportunities and programs.

The complex power requirements of a data center

A hyperscale data center can expand its internal infrastructure as demand grows, often absorbing all power generation available at its point in the network. The original definition of hyperscale described an operation using at least 5,000 chips and megawatts (MW) of power. Today's hyperscalers operate at 10 to 100 times higher scale than this. Demand for AI training and inference has driven hyperscalers to commit hundreds of billions of dollars in annual capital expenditure to purchase GPUs, IT infrastructure to connect them, physical plants to cool them, and of course power for their operation. These hyperscalers include:

- Amazon
- Microsoft
- Google
- Meta
- Oracle
- IBM
- Bytedance
- Alibaba

- Tencent
- Huawei
- CoreWeave

- Apple
- Nvidia
- Snowflake
- Salesforce
- ServiceNow

From their perspective, AI data centers function as information-generating machines, offering dedicated support for scaled digital services (like Meta's ad engine) or selling use of dedicated computing infrastructure to a variety of clients (CoreWeave's business model). Electrical power is the primary variable cost for data centers, which convert electricity directly into revenue and then recycle much of that revenue into continuous hardware upgrades to keep their infrastructure at the edge of technical capability. Hundreds of thousands of GPUs make up the heart of these machines. Each GPU operates at a peak power of about 1.2 kilowatts (kW). Operating these clusters requires hundreds to thousands of MW of power that is quickly accessible, reliable, and stable.

Most data centers draw power from an existing power grid, which must deliver enormous amounts of power across transmission and distribution lines that are already facing bottlenecks. Delays in securing power, or failing to meet the required 99.98% uptime, can have major financial impacts. At the scale of a gigawatt (GW) -class facility, the site generates billions of dollars in annual revenue, so even a 0.1% reduction in uptime equates to millions of dollars in lost revenue. According to CleanView, the opportunity cost of delayed operations is driving "46 data centers with a combined capacity of 56 GW" to plan to bypass the grid and build their own power behind the meter.

Beyond securing timely access to reliable electrical supply, AI data centers must also maintain exceptionally stable power conditions. The compute function of AI training requires operational synchrony. This means GPU racks exchange data in precise choreography with each other. Switching the operational state of all these GPUs in a hyperscale data center can cause load swings of tens of MWs in milliseconds. These changes can occur inside a single cycle of a 60-hertz grid and require enormous resources to keep the grid frequency undisturbed. To access the grid and ensure its own stable operation, a data center must be able to react to and manage disturbances that are larger, faster, and more frequent than any other electrical infrastructure project in history.

Power costs in context

Including cooling and overhead, each modern GPU requires about 1.6 kW of power. At this density, a 1 GW data center can support between 600,000 and 700,000 GPUs, though this depends on the site's configuration and operating mode. For comparison, Tesla reported in its document Megapack for Data Centers: Powering the Growth of AI Infrastructure (PDF) that it installed 200,000 GPUs at about 1.25 kW per GPU for its 250 MW Colossus data center.

While the costs of electricity for a data center are meaningful, they aren't dominant. At an electricity price of $0.087 per kilowatt-hour (kWh), each modern GPU costs about $0.14 per hour in energy and about $0.35 per hour in total including demand charges and other variable operating expenses. When the investment spreads over the four years of a GPU's depreciation, the ongoing cost for the GPUs and the infrastructure that supports them can be three to five times higher than the initial figure. The cost of GPUs is continually increasing because companies are willing to pay higher and higher prices for more-powerful and better-integrated chips. This has reportedly raised the cost of a data center from $35 billion per GW to over $50 billion per GW. The substantially higher computational throughput of the

state-of-the-art systems easily bears this cost.

GPU cost as a function of capital and operational expenditure

The depreciation of GPUs and their infrastructure account for most of a data center's capital expenditure. The base electricity rate for operation of the GPUs accounts for 29% of a data center's operational expenditure; the remaining operational costs include elements like demand charges to account for the high costs of making power available at scale.

The capital expenditure for data centers is astonishing in scope and inherently short-lived. A high-end GPU may command $3 per GPU per hour, but a slightly older model may only command $2 per GPU per hour. These processors have depreciated at alarming rates over the last two years, with the price of training dropping about 50% over two years as newer, more capable chips deliver greater processing power despite higher prices per chip.

The industry is actively discussing the best way to track processor depreciation, since frequent releases of improved models make high performance less expensive over time. There is precedent for keeping high-performance chips in operation even at seven years old, after they have fully depreciated. Operators will continue to do this if they produce enough processing workload to justify their power consumption. The ultimate profitability of data center investment hinges, in part, on whether these GPUs continue to operate and provide marginal value three, four, or six years after their initial purchase.

The power infrastructure these systems need is a longer-lived asset. Electricity cost represents far less than 10% of the total value (revenue) a new data center produces. However, the future price for these processing operations is the subject of considerable debate. A simple economic model would assume that market prices will scale with the cost of processing workload, which trends rapidly downward. We should then expect today's generation of processors to command less revenue over time. As data center equipment ages, power costs make up a growing portion of total revenue.

Check out our tools for more information

The E Source Battery Forecast Database provides visibility in our battery demand and battery price forecasts for over 40 battery applications, including data on 12 stationary applications.

The E Source Battery Factory Tracker is an analysis of all active and announced Li-ion battery factories. You can use the data to improve your understanding of the manufacturing landscape.

The four power challenges for data centers

These challenges consolidate into four universal problem categories:

- Bridging short-duration outages with backup power
- Smoothing ultrafast transients
- Managing peak loads
- Accelerating or stabilizing interconnections

Bridging short-duration outages and smoothing ultrafast transients both occur on time scales from milliseconds to minutes. While these challenges require similar technical solutions, namely fast-response, high-power energy storage systems (ESS), they arise from fundamentally different sources. Short-duration outages come from external grid disturbances, while ultrafast transients come from the internal behavior of synchronized AI workloads.

The last two problems, peak management and faster interconnection, both play out over minutes to hours and also stem from different sources. Peak load management comes from training-cycle technical requirements, while speeding interconnection comes from reducing the amount of firm grid uptime needed so a site can energize earlier. Understanding these four problems provides the foundation for evaluating which battery solutions matter and why they're becoming increasingly crucial to data center development.

Data center power challenges by time scale

While these challenges can overlap in practice, this simplified framework groups data center power challenges by their typical time scale from milliseconds to hours.

Bridging short-duration outages with backup power

Data centers must protect themselves from power outages. They must also protect the grid from failure if the data center suddenly disconnects. Even when the grid remains online, brief voltage dips, known as low-voltage ride-through (LVRT), pose a major challenge for both data centers and the electric grid. LVRTs happen when a fault somewhere on the grid triggers short-lived voltage dips that can last from milliseconds to a few seconds. Each dip forces a data center's power systems to react instantly.

UPS systems detect the disturbance and, within milliseconds, trigger a switch to backup power to maintain continuous server operation. Backup power extends to the seconds timescale to sustain facility power. This rapid switching from grid to backup lets data centers maintain uptime. But, from a grid perspective, it's dangerous. This switching means a data center can disappear from the grid in an instant. When a data center's load of hundreds of MW suddenly vanishes, the entire grid must react to the imbalance between energy supply and demand. This shock can cause voltage and frequency swings that, in extreme cases, can cascade into generator trips or broader outages. A real example occurred in 2024, when 1.5 GW of Virginia data centers (PDF) simultaneously disconnected due to a fault.

As AI load and the number of data centers grows, so does the risk of LVRT events. Once rare, LVRT events now represent a systemic reliability risk. Data centers must configure their power systems to stay online through short grid disturbances and longer outages. They must also work with utilities to manage power quality and system-level impacts.

Data centers also need long-duration back-up power so they can operate during extended outages. Data centers have relied mostly on diesel generators for long-duration backup power because they can ramp up to full output in seconds. But with data center construction surging and supply tight, new diesel capacity is difficult to obtain, with most availability already committed for the next 12 to 18 months.

Natural gas generators are an alternative solution. These generators often take a minute or longer to ramp and also require battery support for seamless transitions. But even these gas generators are in

short supply. This creates space for previously niche technologies, like natural gas fuel cells from Bloom Energy, to move into the market. While Bloom's systems are substantially more expensive than conventional turbines, at the time of writing its capacity backlog is only four months. Power is a sufficiently low fraction of overall costs that Bloom's higher cost of generation is feasible if it allows a data center to start operations sooner.

According to Tesla's Megapack for Data Centers: Powering the growth of AI Infrastructure (PDF), batteries are generally more cost-effective than diesel for durations under eight hours. This is because of the high fuel and maintenance costs of fossil-fuel solutions. Beyond eight hours, the value of storage varies widely by location and reliability requirements. For data centers already tied to the grid, batteries can earn value by supporting grid operations as well as serving as backup power, helping reduce their overall cost. A BESS system can sit between the data center and the grid, providing value to both.

On the data center side, a UPS is the first—and entirely nonnegotiable—line of defense against disruptions. Operators use high-power, ultrafast technologies to protect data center operations and the grid between the instance of a voltage dip and the ramping of long-duration storage. A conventional data center builds in 5 to 10 minutes of UPS for a backup generator to power on. By rough estimate, a 1 GW data center will need 100 megawatt-hours (MWh) of storage. Older designs required 100% redundancy when data center scale was smaller, while modern designs realistically require the UPS storage to have at least 20% redundancy.

Smoothing ultrafast transients

Earlier generations of data centers had a portfolio of applications so no single process dominated the workload. Today, in contrast, load can swing in less than a millisecond as GPUs switch computational states. This can imbalance the system's power supply, so manufacturers must design systems to accommodate this switch.

Local capacitors at the chip or rack level provide the first line of defense against power spikes. These capacitors can supply energy for about the first 20 milliseconds of a disturbance. After that, integrated supercapacitors, or short-duration batteries in UPS systems, provide several seconds of support to stabilize the facility if a synchronized training job suddenly halts. After this, Li-ion UPS batteries provide around five minutes of ride-through. This is enough time to cover multisecond demand spikes or bridge the gap until backup generators start.

According to the North American Electric Reliability Corporation's (NERC) Incident Review: Considering Simultaneous Voltage-Sensitive Load Reductions (PDF), data centers often reduce load or disconnect from the grid during voltage disturbances because their protection systems automatically shift to on-site backup systems. Computing and cooling equipment are highly sensitive to voltage fluctuations, so data centers can rely on UPS systems that take over instantly when disturbances occur. The specific response varies by architecture:

- **Centralized.** 2–5 MW UPS units switch entire load centers onto battery banks
- **Decentralized.** Rack-level UPS or battery backup unit systems protect individual racks

Some facilities use a dynamic rotary UPS that combines a flywheel with a fast-starting diesel engine. Although these UPS systems provide short-duration support that's just long enough to ride through a disturbance or start generators, their differing designs produce different system-wide behaviors when

grid events occur.

While bridge-to-backup challenges are about overcoming fast, involuntary disturbances on the grid, load smoothing focuses on managing the data center's contribution to grid volatility.

Today, the trend is for a data center to operate solely for inference or training during specific time periods. Because training runs involve synchronized computation, these periods can terminate simultaneously. This causes large, sudden shifts in power use as the processing workload for training shifts to exporting data from memory.

The scale of these swings is dramatic. As seen in Figure 1 of Google's article Balance of power: A full-stack approach to power and thermal fluctuations in ML infrastructure, AI workloads can swing by 15 MW. This is 10 times greater than the 1.5 MW swings of non-AI workloads that SemiAnalysis mentions in its June 2025 newsletter. The Google data is also several years old and therefore may not represent the power fluctuations a new data center might experience today. The simplest solution for data centers to smooth their power loads is to run dummy calculations that prevent demand from dropping catastrophically. While effective, this approach wastes energy—so fast-response energy storage is an attractive alternative.

Managing peak loads

Larger battery systems can smooth renewable variance and shave peak loads. Peak shaving is most likely to be economically valuable if it can reduce peak demand during training cycles since they consume the most power.

Unlike standard daytime peak shaving, which occurs once daily, peak shaving during training requires addressing multiple high-demand events each day. If an electric utility specifies that the highest demand charges apply only during specific hours, the shaving profile would look like the profile of a regular energy user, just at a larger scale. But if a data center limited its energy consumption to a specific cap—for example, to keep use underneath the capacity of an existing interconnection—this cycling profile would be more difficult to meet.

The cost of electricity for a 1 GW data center is around $700 million per year. Demand charges for a data center of that size cost at least the same. These costs create a notable financial sensitivity to peak-period pricing. Reducing a data center's grid power needs would open up more sites and lower operating costs. More traditional uses of BESS could provide meaningful operational savings. Tesla suggests that a two-hour Megapack BESS would represent only about 1% of a data center's capital cost. These numbers are reasonable based on today's market prices: a containerized system priced at $250 per kWh would cost about $500 million to provide two hours of backup to support 1 GW of data center load. Compared to a rough cost of $50 billion per GW to build an AI data center, the cost of batteries is trivial. A $500,000,000 battery could pay for itself in less than the three-to-six-year data center depreciation timeline by reducing demand charges or incorporating cheaper renewables.

In practice, operators must first maximize the use of their expensive GPUs; demand charges are a cost of doing business. BESS adoption for peak shaving will depend primarily on the strength of underlying rate incentives. The growing difficulty of finding stable grid connections will drive BESS adoption. As the most reliable grid connection sites are used up, BESS may become essential for future data center development.

Accelerating or stabilizing interconnections

AI data centers face a fundamental economic and operational bottleneck: the speed at which the grid can deliver power to a new data center. Even when capital is available and construction proceeds rapidly, the data center can't produce revenue until it has enough energy.

In many regions, transmission and distribution upgrades can take up to seven years. For hyperscalers deploying at GW scale, this delay has massive financial implications. A 1 GW training campus houses about 650,000 GPUs, each producing around $0.60–$0.70 per GPU-hour in gross margin. At about 80% utilization, this represents nearly $3 billion in annual gross margins per GW of powered capacity. These margins evaporate with interconnection delays.

Recent policy signals reinforce how the requirement to provide uninterrupted, firm service to large new loads is driving much of today's interconnection delays. The value of flexibility varies by market structure. PJM Interconnection, the largest US grid operator, doesn't treat or value large loads as dispatchable resources like Energy Reliability Council of Texas does. Yet Secretary of Energy Chris Wright's directive to the Federal Energy Regulatory Commission (PDF) explicitly gives faster grid interconnections to customers who can reduce or adjust their power use when needed. This could give flexible data centers a key advantage in the race to power if the Federal Energy Regulatory Commission adopts rules aligned with the Department of Energy's guidance.

This aligns with the Duke University report Rethinking Load Growth: Assessing the Potential of Large Flexible Loads in US Power Systems (PDF, 23).The study found that many parts of the US grid could support large AI loads if data centers can accept brief reductions in reliability—from about 99.98% today to around 99.5%. Together, these insights highlight that, if on-site resources like batteries can moderate and support reliability expectations, far more sites become interconnection-ready. This would let large data centers come online years sooner than under traditional firm-service rules.

This makes accelerating time-to-power one of most important economic attributes of batteries. Time-to-power refers to the ability of storage to bring new load online faster than the grid can upgrade. In 2025, a 31 MW, 2-hour battery helped bring a proposed 108 MW data center online "years earlier than would be possible with traditional utility upgrades." In this case, the data center operator provided both the equipment and the land for siting the BESS installation. This provides a clear example where batteries that can support load-leveling or provide temporary, non-firm services to accelerate revenue realization, even if they require upfront capital or temporarily increase financing costs.

The Duke study also found that such data centers should be able to operate fully with only two hours of storage. This small amount of capacity is still effective because grid stress usually occurs in brief, predictable windows. Telsa has published claims that roughly agree, stating, "Capacity equivalent to 10% of US's peak load can be unlocked with 0.25% curtailment of new load."

In the Duke modeling, this modest reliability adjustment supports the existence of the data center, recovering around 0.5% of the nearly $9 billion of revenue per year from a 1 GW facility renting its capacity at about $2 per GPU per hour. This way, a 2-gigawatt-hour BESS that allows interconnection enables about $450 million per year in revenue while costing roughly $500 million at the $250 per kWh that Tesla reported. Assuming a gross margin of 33%, a data center could pay back its storage capital in just over three years. This is a phenomenal return on investment compared to utility battery projects. Even a four-hour battery could be profitable, as LFP batteries for BESS usually have a 15- to 20-year life

and will depreciate much more slowly than the data center. However, we shouldn't consider the Duke study definitive for any specific site since the required duration would vary based on:

- Grid conditions
- Contracts between the data center developer and its customers
- Other system factors, like market structures

Together, the value of time-to-power, reliability-enablement, load stabilization, and demand-charge mitigation show that batteries can materially improve the economics of large data centers and increase the amount of available processing workload.

Read part two of this series to learn more

For more information about data centers, read part two of this series The evolving role of ESS in AI data center load management. In part two, we focus on the various competing solutions for providing data centers with stable power and their associated market risks.

© 2008 - 2026 E Source Companies LLC. All rights reserved.
Distribution outside subscribing organizations limited by license.