INTELLIGENCE UPDATE

Handling AI power fluctuations at the root cause

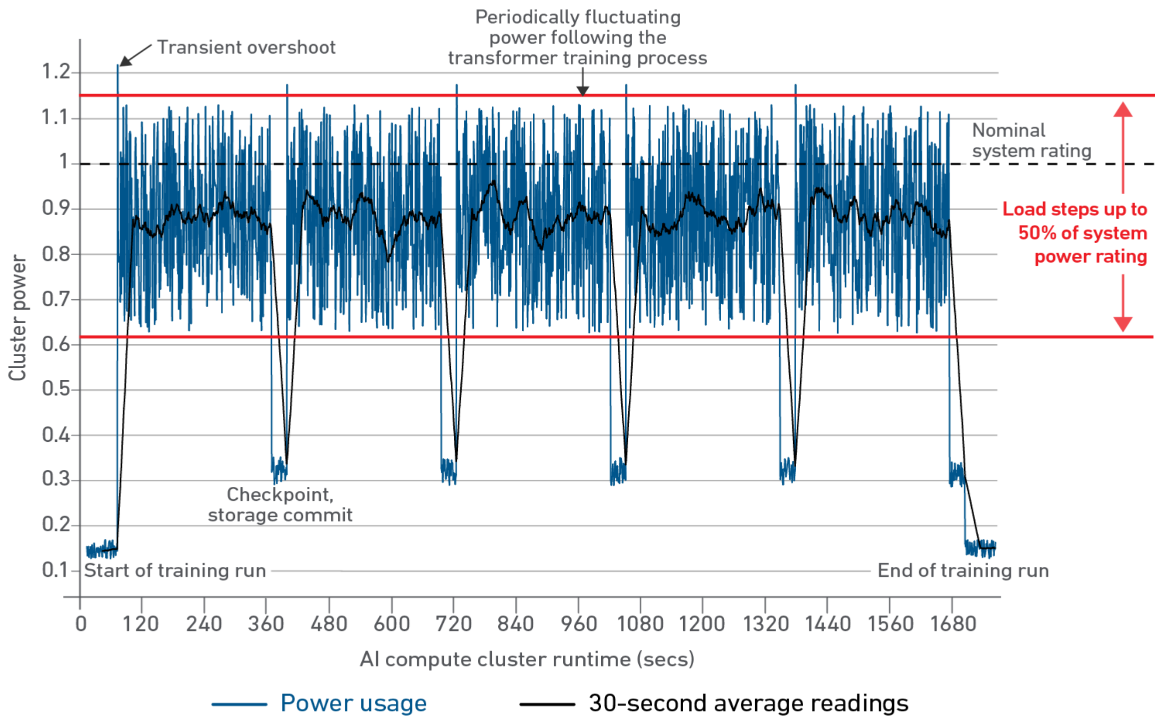
Daniel Bizo 30 Jul 2026

The data center industry first took notice of the erratic power profile of AI training systems in 2023, at the time when the first truly large compute clusters came online to develop AI models even larger and better than OpenAI's ChatGPT-3. Now, the phenomenon is widely documented and heavily researched. In short, training generative pre-trained transformers (GPTs, the dominant crop of popular AI models today) on massively parallel compute clusters will create various types of repeated power fluctuations on a facility scale. AI inference does not create similar electrical patterns.

There are several contributing factors to these power fluctuations, but the dominant one is the training pipeline. Computationally intense stages of the training run across thousands of processing units are followed by memory operations moving large volumes of data. Every few seconds, this creates large-amplitude step changes in the load. Large clusters also rely on regular storage checkpoints to minimize any loss of work, during which power levels drop steeply before shooting up tens of seconds later.

When a training cluster uses several megawatts at full load, such swings in power can become an issue. To make matters worse, modern silicon exacerbates the issue by producing power excursion events above nominal design power ratings — creating stronger power swings and potentially overload conditions if capacity sizing of equipment is not appropriate, also considering inrush currents and power factor. Repeated overloads may cause breakers to trip, make UPS systems to use energy storage to match the load, or even force UPS systems to run in bypass mode (leaving the IT load unprotected against grid disturbances). **Figure 1** shows the power profile issues of AI training (see *Electrical considerations with large AI compute* for a more detailed discussion).

Figure 1 Power profile of simulated GPU-based training cluster



UPTIME INSTITUTE 2026

Areas of concern

Discussions with industry participants highlighted four major areas of concern:

- **UPS system performance**. Large load steps challenge the UPS systems' ability to maintain stable power conditions that meet strict industry standards for sensitive critical loads, such as IT hardware. These variations can also reduce UPS system efficiency by several percentage points. If the system is not correctly set up and sized, power swings may also result in repeated small discharges from energy storage as the UPS controls respond to stabilize output, particularly during overload conditions. Frequent battery discharging and charging can lead to premature aging.
- **Engine generators.** Volatile step loads can prevent the generator set from delivering power within expected voltage and frequency tolerances. This can also result in engine instability, and continuous large variations in the load may shorten engine lifespan due to increased mechanical stress. In the worst-case scenario, engine generators may shut down or fail because they are unable to support the AI workload.
- **Transformers.** Large load steps cause both localized thermal and mechanical stress in transformers arising from harmonics and changes in magnetic fields. These may contribute to accelerated insulation aging, as well as coil deformations. Transformers vary greatly in how they cope with large load steps depending on their type and the engineering choices made by the manufacturer.
- **Grid imbalance.** Large AI data centers have the potential to negatively affect grid stability. AI training workloads are a source of potential imbalances that increase power grid costs

due to the increased use of containment reserves. In the worst-case scenario, sudden load steps may result in grid frequency variations, overloading of transmission lines and resonance in turbine generator shafts of power plants (causing mechanical fatigue).

Mitigation of IT root causes

The challenge of managing these AI training related power swings is still relatively new, and no single solution has yet emerged to address the issues described above for all scenarios. The problem is also becoming more prevalent, as AI training systems grow bigger (surpassing 10 MW) and become increasingly prevalent worldwide — establishing a new class of computer system. For comparison, as of July 2026, a mere handful of research supercomputers worldwide require more than 10 MW of power, and only a few dozen need more than 5 MW.

For operators with pre-existing large loads and spare capacity to host an AI training cluster, load diversity is their first best option to absorb most of the load swings associated with AI training. Load diversity simply means that the electrical effects of these sudden, synchronized changes in power demand are diluted by other workloads, making them less substantial for the data center power system. Some power delivery equipment, such as power distribution units and breakers directly serving the AI training system, will remain exposed to these load swings, but the additional costs (oversizing) are limited and the risks are more contained overall.

Addressing the root of the issue — the IT hardware — can be highly effective, both technically and economically, for new builds. This is especially true in existing data center capacity, where retrofitting electrical systems (such as changing UPS systems and batteries or adding flywheels) can introduce operational risks and may not be financially viable. Below is a summary of major IT options, including a combination of them:

- **Workload scheduling**. Operating AI clusters in a manner similar to research supercomputers can help alleviate many of the problems. Running several jobs simultaneously on the same hardware, or splitting up the system to resource partitions for different tasks depending on their capacity needs, will prevent the formation of large power swings. However, the trade-off of this approach would inevitably be lower overall performance, and the development of cutting-edge AI models may require the entire system to be viable.
- **IT system performance and power controls**. Performance and power management are underused features in IT hardware, despite the long-standing availability and maturity of the control software for all major platforms, whether Nvidia, AMD or Intel. The three most common methods to minimize power swings (available on all hardware platforms) are: limiting or turning off turbo clocks to slow down and limit the amplitude of power swings; outright capping of peak power levels; and turning off power-management features, such as sleep states, which then lifts power consumption during periods of low activity and idling. However, the first two options will somewhat reduce training performance of a given compute cluster (typically by a single digit percentage), while the last option may be deemed controversial because its principle conflicts with energy efficiency and sustainability mandates.

  Nvidia offers additional special controls for systems built around its GPUs through

Nvidia's system management interface (SMI). Since the introduction of the Blackwell (B-series family) generation, SMI "power smoothing" commands allow power to be maintained within a narrow band, as well as to limit the rate of change at the start, during checkpoints and at the end of the training job. This also relies on hardware-integrated features for being able to burn excess power to slow down load drops. Although it does not affect performance, a limitation of the technique is that the integrated hardware has a finite lifespan due to wear. However, it remains unclear what this means in practice given an AI training system's expected useful life is 4-6 years at the current pace of obsolescence.

- **Added energy storage capacity to IT racks**. Even if the above techniques were at the disposal of the infrastructure teams, they will not fully dampen rapid (sub-second and multi-second) power oscillations during training runs. Without workload diversity in the facility and on campus, persistent power oscillations may still be a cause for concern because they accelerate electrical component aging and may create power quality issues in the form of harmonics, which may also be fed back onto the grid.
  Adding larger capacitor banks to the IT power supply is an effective way of dampening these power oscillations. Alternatively, battery-based energy storage systems with suitable chemistries (lithium-ion, sodium-ion, nickel-zinc) for rapid micro-cycling (repeated shallow depth of discharge and recharge) can be considered, although at more cost and electrical complexity. These options, however, all require changes to the IT specifications, which may limit their applicability in some circumstances.
  In the future, a key feature of centralized 800V DC power supply racks proposed by Nvidia for its next-generation of AI training systems (Rubin and Rubin Ultra) is to offer more room for capacitors and batteries to provide dampening against electrical fluctuations (see *Vendors gearing up for 800V DC adoption*). Such row-level power supply architectures will be specific to certain Nvidia rack-scale systems and are unlikely to move into high volume production before 2027, with mass deployments likely to begin in 2028. However, row-level energy storage to protect AI hardware and dampen load fluctuations may more broadly gain currency in the near future.

Outlook

In the foreseeable future, AI training clusters will become denser and larger in power demand, surpassing most research supercomputers. Next-generation AI training systems developing cutting-edge models could exceed 40 MW, with compute rack densities expected to exceed 400 kW before the end of the decade (although using wider, deeper cabinets). Current AI hardware technology roadmaps at Nvidia and AMD call for close coupling of increasing amounts of compute and memory resource to enable faster synchronization. This trend will exacerbate the intrinsic step load behavior, making for mitigation measures an important consideration in the infrastructure planning stage.

Cooperation between IT infrastructure teams (internal tenants or external customers) can avoid or reduce the need for more expensive mitigation layers in the facility power infrastructure. Manufacturers of UPS systems are honing their products to be able to almost fully dampen the sub-second and second-level oscillations of large AI training systems, even without load diversity. Addressing larger and longer-duration load steps to shield engine generators and the grid requires energy storage systems that are both sufficiently large and responsive. These systems can absorb high-energy events, such as checkpoints or starts and ends to training jobs. Additional options include battery energy storage systems and, alternatively, flywheel energy storage systems (where rotating mass stores kinetic energy and converts it into electrical energy when needed and vice versa).

The importance of large step loads extends beyond data center electrical systems. Although owners and operators already face scrutiny over power and water use, a relatively recent concern is how a high concentration of data centers is affecting — or will affect — the grid in some regions. This includes the need for network upgrades and the potential impact on grid stability. Several regulators and power grid operators worldwide are considering revising the rules of grid connection to address data center-specific load issues, which will include fast load changes (see *Draft grid rules position data centers as active grid participants*).

> ## The Uptime Intelligence View
>
> To date, much of the industry's attention has been duly directed toward solving the rack power density problem as training leading-edge generative AI models increasingly resembles supercomputing. However, large and rapid power swings place significant strain on electrical equipment and may affect grid stability. Addressing the issues at their root cause — the IT system software and hardware — can be both highly effective and less costly and complex than mounting layers of defence in the facility infrastructure.

ABOUT THE AUTHOR

Daniel Bizo
30 Jul 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. With over 4,000 awards issued in over 122 countries around the globe, and over 1,100 currently active projects in 80+ countries, Uptime has helped tens of thousands of companies optimize critical IT assets while managing costs, resources, and efficiency. For over 30 years, the company has established industry-leading benchmarks for data center performance, resilience, sustainability, and efficiency, which provide customers assurance that their digital infrastructure can perform across a wide array of operating conditions at a level consistent with their individual business needs. Uptime's Tier Standard is the IT industry's most trusted and adopted global standard for the design, construction, and operation of data centers.

Offerings include the organization's Tier Standard and Certifications, Management & Operations reviews and assessments including SCIRA-FSI financial sector risk assessment, the Sustainability Assessment, and a broad range of additional risk management, performance, availability, and related offerings. Uptime Education training programs have been successfully completed by over 100,000 data center professionals, such as the much-valued ATD (Accredited Tier Designer) and AOS (Accredited Operations Specialist). The Uptime Education curriculum has been expanded by the acquisition of CNet Training Ltd. In 2023.

Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, and Singapore, and full-time Uptime professionals based in over thirty-four countries around the world.

For more information, visit www.uptimeinstitute.com