AI Training Load Fluctuations at Gigawatt-scale - Risk of Power Grid Blackout?

108GW Large Load Queue, Tesla Megapacks, Supercapacitors, Gigawatt-scale Batteries, PyTorch No Power Plant Blow Up

JEREMIE ELIAHOU ONTIVEROS, DYLAN PATEL , ANDAJEY PANDEY

JUN 26, 2025 · PAID

The largest AI labs are racing to build multi-gigawatt-scale datacenters, and stressing our century-old power grid to an unprecedented extent. Not only is the scale massive, but **AI training workloads have a very unique load profile,** unexpectedly rising and falling from full load to nearly idle in fractions of a second. Our power grids were never designed to handle this pattern. At Gigawatt-scale, the worst-case scenario is **a blackout for millions of Americans.**

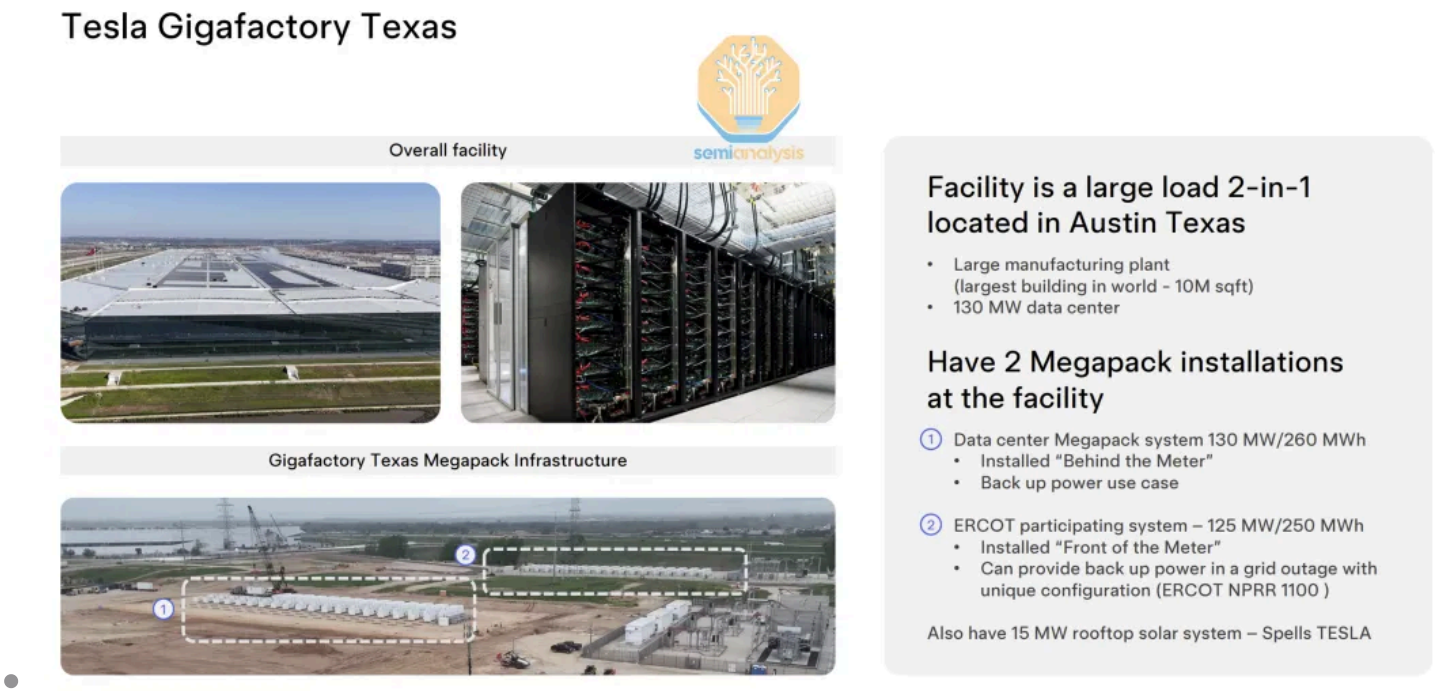
The issue caught leading AI labs by surprise. Meta's LLaMa 3 paper mentions challenges with power fluctuations, and that is "only" a 24,000 H100 Cluster (30MW of IT capacity).

> During training, tens of thousands of GPUs may increase or decrease power consumption at the same time, for example, due to all GPUs waiting for checkpointing or collective communications to finish, or the startup or shutdown of the entire training job. When this happens, it can result in instant fluctuations of power consumption across the datacenter on the order of tens of megawatts,

> stretching the limits of the power grid. This is an ongoing challenge for us as we scale training for future, even larger Llama models.

Out of desperation, engineers built the command "**pytorch_no_powerplant_blowup=1**" to generate dummy workloads, smoothing out the power draw. But at gigawatt-scale, the annual energy expense caused by such workloads sums up to tens of millions! Hardware vendors have since lined up to propose serious solutions.

In Memphis, xAI's "Colossus" opted for Tesla's Megapack system. Musk's carmaker leads the Battery Energy Storage System (BESS) market and is now actively engaging with utilities and datacenter operators to make its solution the standard. Is Tesla set to take over the market, or are there credible alternatives to BESS to handle AI Training load fluctuations?



Source: Tesla

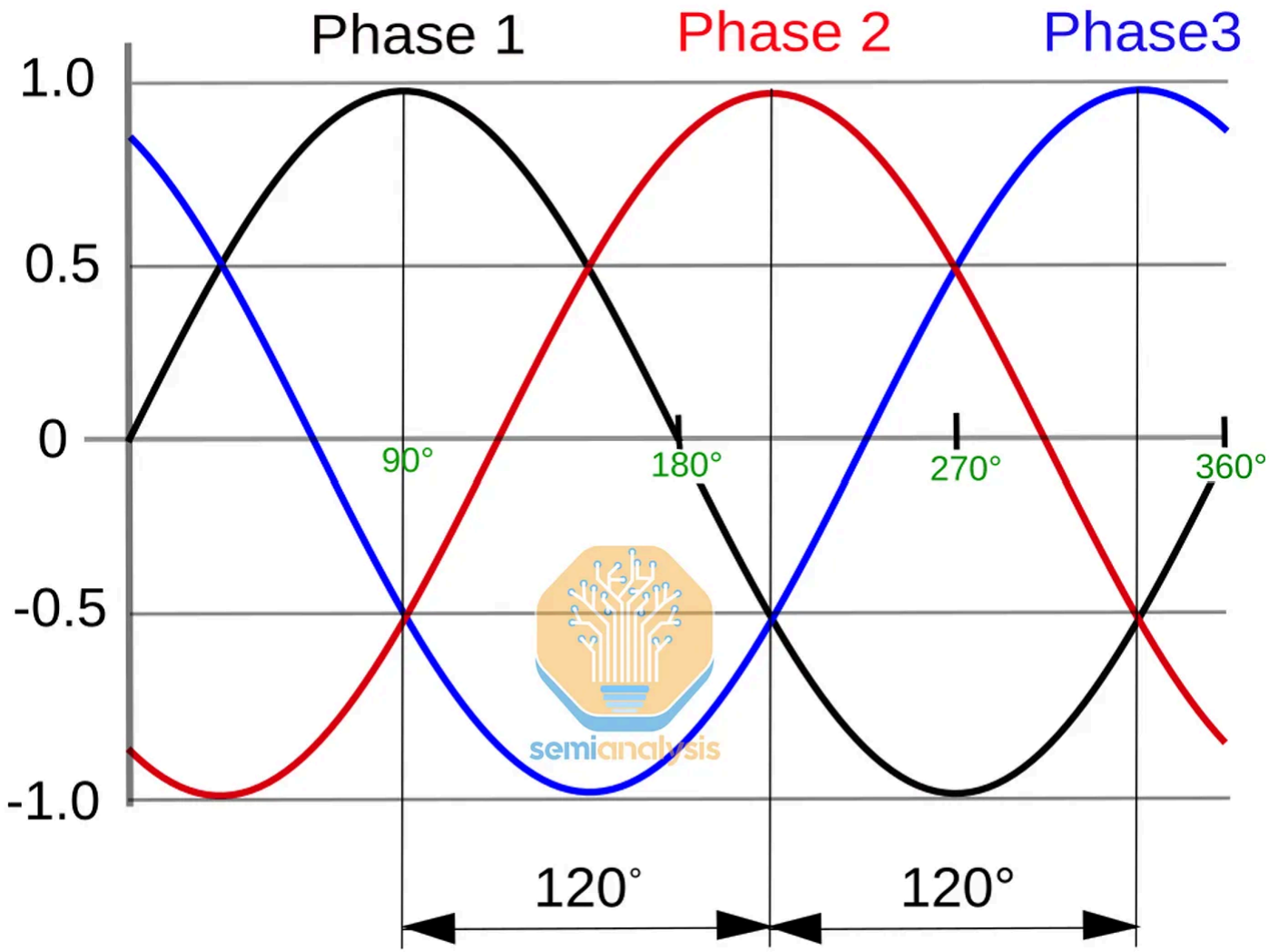
To understand market implications, we start from first principles and explain why power quality matters and some basic power grid design considerations. We then explain the load profile of AI training and inference and compare it to traditional workloads, and lay out how a gigawatt-class AI training datacenter could trigger a blackout. We then examine solutions, from supercapacitors to UPS and Battery Energy Storage Systems (BESS), and identify the likely winners. Our datacenter project-by-project forecast enables us to get an early understanding of what's coming ahead, and we believe a few firms are set to benefit disproportionally.

SemiAnalysis will be posting exclusive contents on Instagram Reels and TikTok starting next week. Follow our socials to get the latest insights on the AI and GPU industry.

Power Quality, Briefly

It is a testament to the competence of utility engineers that **power quality** has not entered public vocabulary. Most readers simply flip a switch and trust that nothing will flash, fry, or trip. But that confidence rests on **balancing electric generation and electric load on a fractions-of-a-second basis**.

Nearly every part of the grid, fossil-fuel and nuclear plants, transformers, high-voltage lines, runs on **alternating current** (**AC**). Within AC electric systems, **voltage** and **current** oscillate at a *very* tightly managed region-specific frequency: 60 Hz (60 cycles per second) in North America and 50 Hz in Europe and Asia. Residential loads typically operate with one oscillating line, but industrial loads like datacenters typically receive **three-phase** power, in which each power line is in fact three wires with three oscillation cycles running offset from each other.



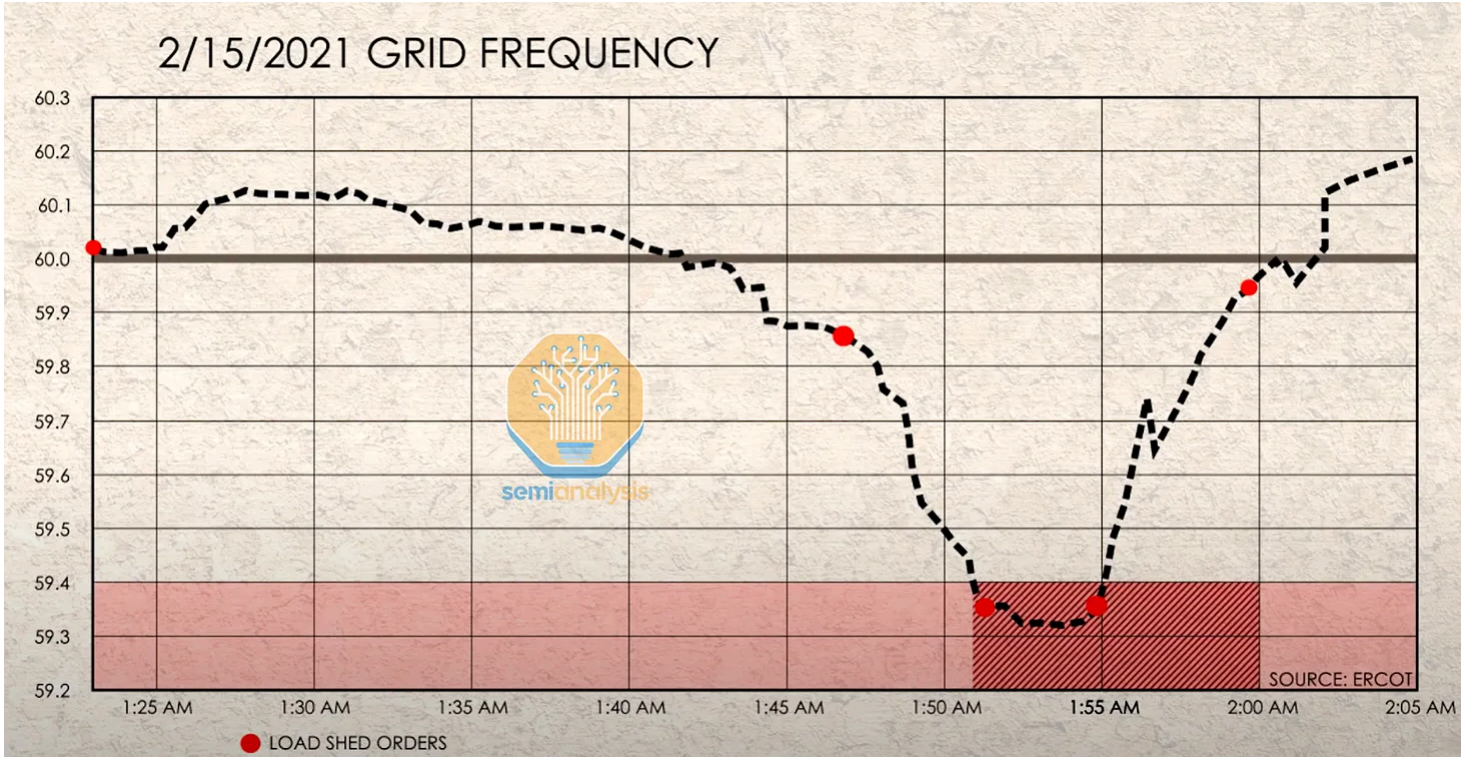
Source: Wikipedia - Three-phase Electric Power

However, voltage and frequency are fragile properties of electricity. If the supply and demand of electricity do not closely match, both voltage and frequency drift away from set points. If supply exceeds demand, voltage and frequency increase from baseline. If supply falls *below* demand, voltage and frequency fall below that baseline. A mere 10 %

swing can fry motors, trip breakers, and crash electronics, and a grid operator's job is to maintain a threshold of **power quality**.

The winter-2021 freeze in Texas proved the point. Extreme cold sent heating demand soaring and knocked several large gas plants offline. Supply lagged, and system frequency sank below 59.4 Hz. In ERCOT (Texas's grid), staying under 59.4 Hz for nine minutes triggers protective breakers and plunges the state into a **multi-day blackout with lasting damage**.

To keep the lights on, ERCOT cut power to homes and businesses, slashing demand until it matched the crippled supply.



Source: Practical Engineering

This highlights how grid stability depends on a stable balance of supply and demand, and the risk of an imbalance. Fortunately, household demand is quite predictable, and large loads like electrical steel manufacturing, chip fabs and Cloud datacenters usually draw a stable load. The rise of GenAI changes the playbook.

AI Load profile deep dive

AI computing systems are typically synchronous. A large GPU training run can involve hundreds of thousands of GPUs working together, in sync. We explain the basics here. This pattern is at odds with traditional computing profiles:

- Cloud Computing is the business of selling multiple Virtual Machines to a large number of users - each with very different use cases. Some large customers can rent thousands of VMs, but even then, they generally have a heterogeneous load profile. Keep in mind that a 100MW datacenter can host millions of CPU cores (and VMs).
- Conventional inference such as Meta's DLMRs (AdRec, feed ranking, etc) typically involves using multiple small models where each one has a small compute requirement. The end result is a non-synchronous pattern.

The chart below published by Google Cloud suggests a ~15x difference in load fluctuations between a Cloud datacenter and an AI datacenter, from 1.5 MW to 15MW.

Source: Google at OCP EMEA Summit 2025, SemiAnalysis

Large-scale training clusters

This is easiest to understand in the context of large AI training datacenters, where up to hundreds of thousands of GPUs are networked and act as a single supercomputer. Read our deep dive on the networking architecture of a 100k H100 cluster for more details. There are many reasons causing AI training loads to fluctuate so much, such as:

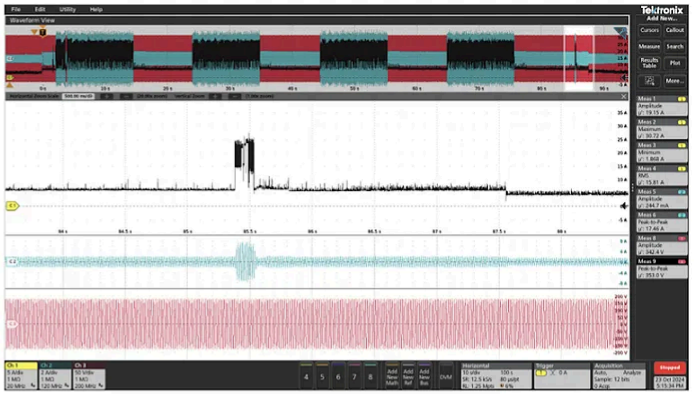
- Intra-batch spikes and dips (milliseconds): as a batch is processed, power spikes during matrix computations, and dips during lighter operations such as data transfers and synchronization.
- Checkpointing / restoring (milliseconds): loads can drop to near zero during a checkpoint, which typically lasts a few milliseconds.
- Synchronization (up to a few seconds): as cluster sizes rise to hundreds of thousands, AllReduce operations are plagued with network issues, sometimes leading to up to a few seconds of idle GPU compute activity.

- End of a training run: after a very large run, if there is no immediate workload to use the GPUs at max power, it can lead to a huge load drop.

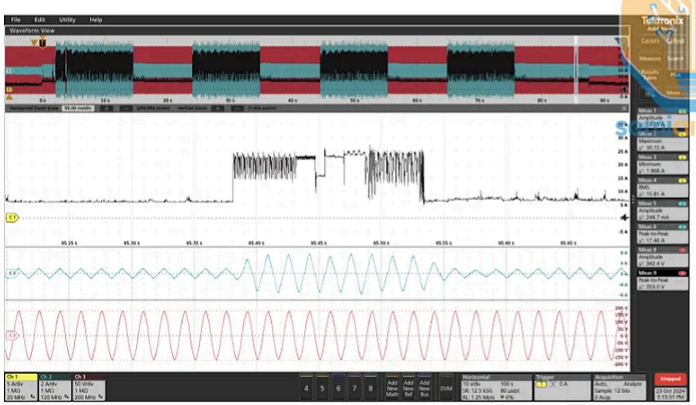
This is a non-exhaustive list and, to be clear, many of these can be partially solved by software modifications and workload & cluster management optimizations. But the problem remains, and AI training workloads are very unique in that regard. A **hardware-based solution** is needed.

The below paper shows some empirical results of a training run.

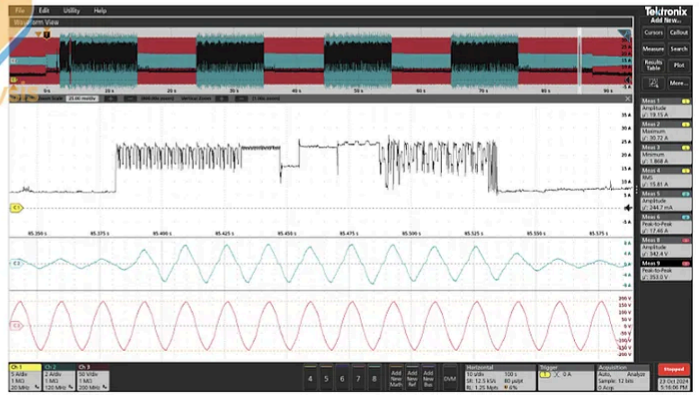
(a) Initial capture showing the first load drop around 14.5 s. Checkpoints induce a sudden drop from peak current to near-idle, underscoring the abrupt nature of AI training workload transitions.



(b) Closer view (10 s scale) of the last checkpoint event, where training is intentionally interrupted, causing rapid up/down current surges.



(c) Detailed capture (500 ms scale) focusing on the steep transitions as the GPU goes from full compute load to near idle. Several fundamental AC cycles are visible in the PSU current waveform.



(d) Millisecond-level zoom of the final interrupt event, pinpointing the instantaneous swings in GPU motherboard current across just a few AC cycles.

Fig. 4: Progressive zoom-in of GPU current transients during GPT-2 (124 M) training checkpoints. (a) illustrates the first load drop, while (b)–(d) progressively zoom into the last interrupted training phase, where rapid up/down surges occur within a span of several AC cycles. Waveform from top to bottom: current of the GPU mother board, PSU current, PSU voltage.

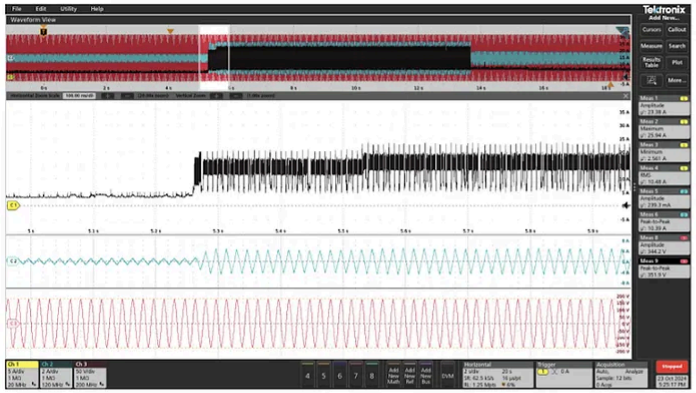
Source: AI Load Dynamics–A Power Electronics Perspective

Inference workloads

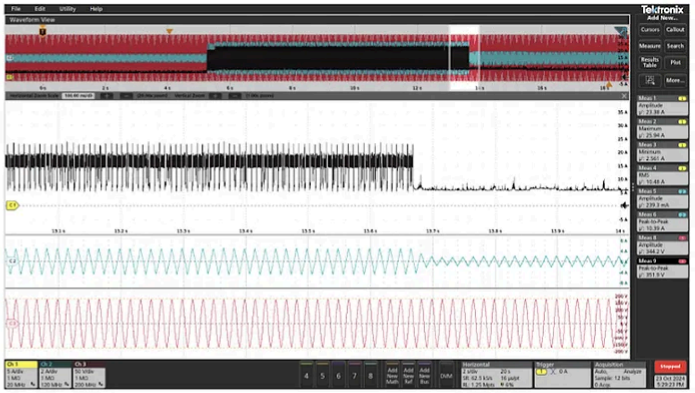
Empirical evidence of large-scale inference deployments (DLRMs) from the likes of Google, Meta, TikTok suggests the issue is much less pronounced with inference. But GenAI once again brings very new dynamics:

- Prefill and decode: each LLM query has two distinct phases, prefill and decode. The former is generally much more FLOPS-heavy than the latter, meaning that GPUs run at max power during Prefill, but often less than 50% during decode. This is mitigated by modern disaggregated prefill & decode techniques.
- Inter-node communication stalls: high batching is crucial to profitably serve millions of users, and in the context of SOTA reasoning models, many nodes are often required. Inference workloads then resemble more training.

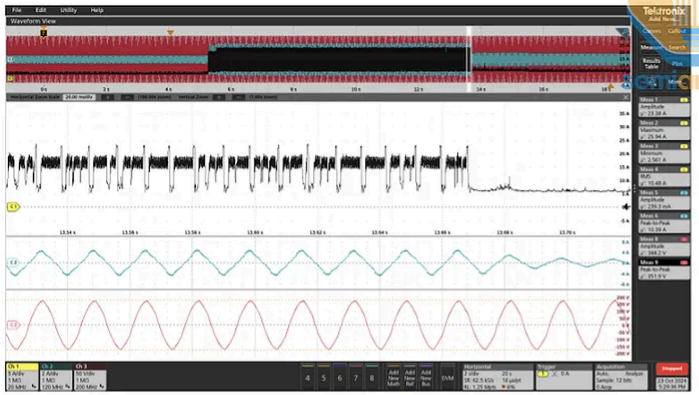
The second point is best exemplified by DeepSeek's very unique inference deployment to efficiently serve millions of users with a small GPU footprint - which we explained in depth to our Core Research clients.



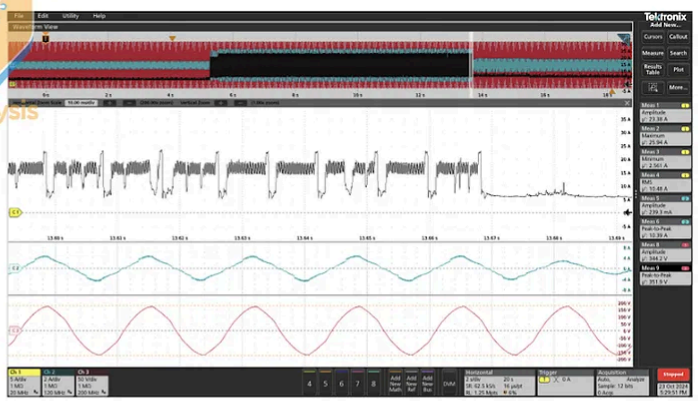
(a) Initial capture showing the inference load-up transition for an RTX-4090 running a LLaMA-3.1 8B model. The GPU current (black trace) rapidly increases, indicating the onset of inference operations.



(b) Subsequent capture illustrating the initial portion of the load-down event (GPU returning to a lower-power state). Note the abrupt decline in GPU current and the near-sinusoidal PSU input current.



(c) More detailed zoom of the negative transient, showing repeated downward spikes in GPU current as the inference request load diminishes.



(d) Millisecond-scale view capturing the final ramp to idle power. The high-frequency ripples on the GPU current trace illustrate rapid control-loop adjustments before settling.

Fig. 5: Waveforms recorded during inference operations on an RTX-4090 running a LLaMA-3.1 8B model. (a) The GPU load-up event draws substantial current as the inference request begins, while (b)–(d) document the load-down stages with progressively closer zoom, highlighting the abrupt negative transients and the PSU's response in stabilizing the supply voltage and current. Waveform from top to bottom: current of the GPU mother board, PSU current, PSU voltage.

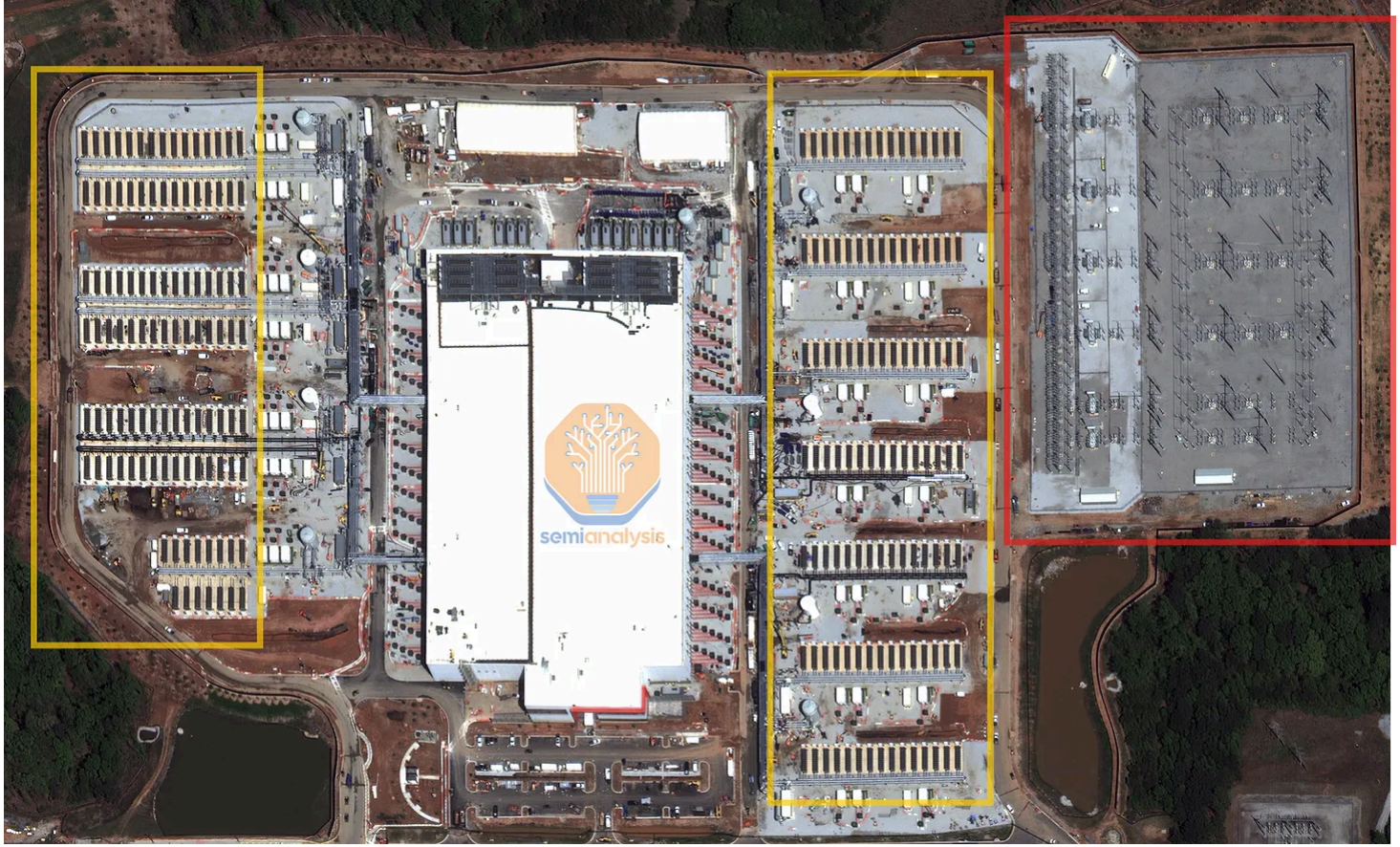
Source: AI Load Dynamics–A Power Electronics Perspective

Both inference and training are subject to load fluctuation issues, but training workloads are much more challenging, as they involve up to Gigawatt-scale systems working synchronously. However, given trends in Scaling Laws and Reinforcement Learning, we see inference workloads as likely to increasingly rely on large scale-out clusters - making them problematic as well, but not to the same extent.

Power grid impact - AI Datacenters are flooding the grid

To understand the magnitude of the problem and potential risk, we take a step back and look at the scale of today's largest AI Datacenters. We show below one of OpenAI's key training clusters. **This is simply the world's largest single building** (alongside a "sister" site in Wisconsin), at ~300MW IT capacity and ~400MW nameplate, by a wide margin. The scale is obvious to readers of our Datacenter Anatomy reports (Cooling and Electrical systems), by looking at the 210 air-cooled chillers or the massive on-site substation.

A second identical building is under construction since January 2025, taking the campus to Gigawatt-scale by mid-2026.

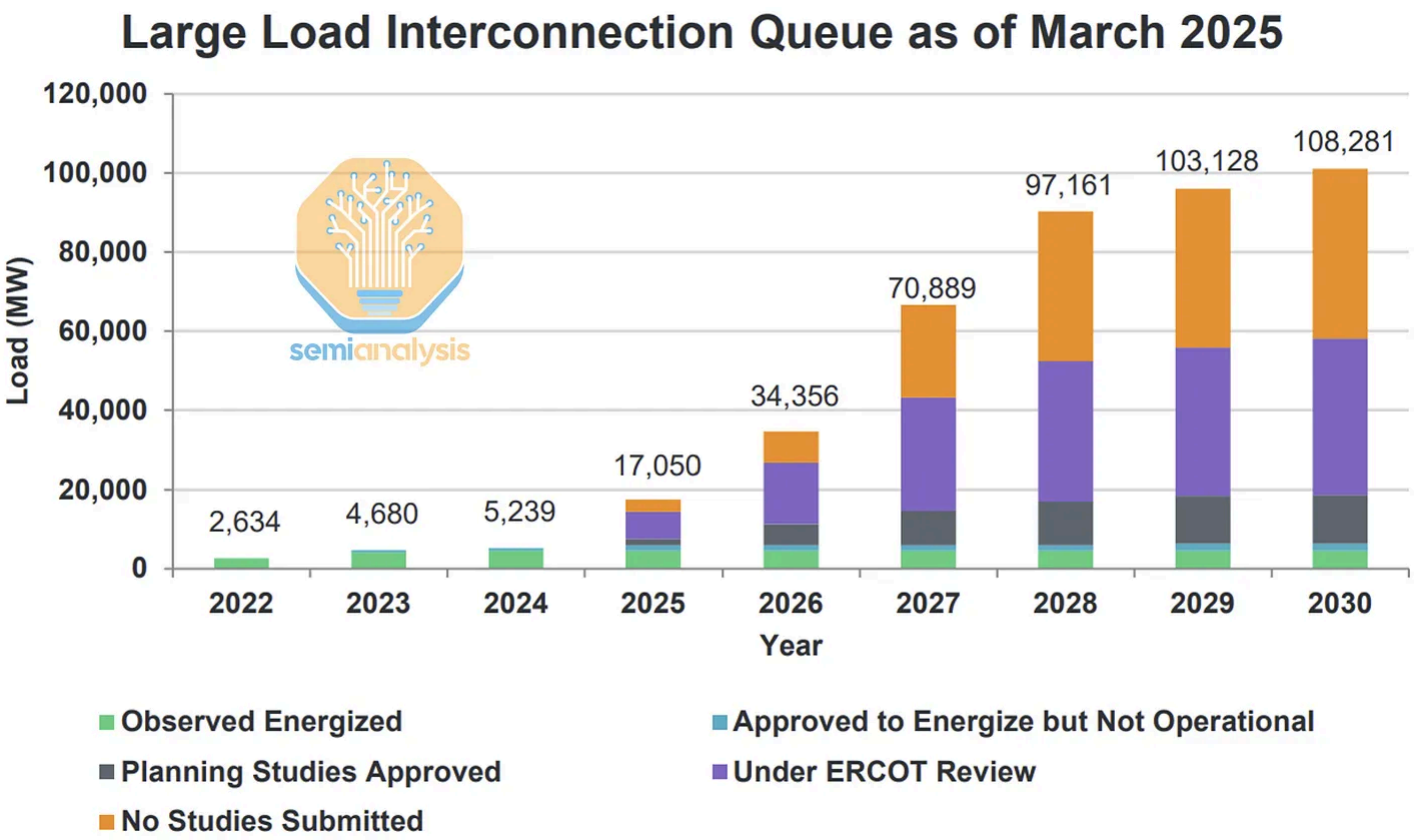


Source: SemiAnalysis Datacenter Model

This caught ERCOT's (Electric Reliability Council of Texas) attention - the organization that oversees the Texan power grid. The chart below makes it easy to understand: **more than 108GW of "large loads" are looking to connect to ERCOT**, of which the majority are datacenters. To put this in perspective, the US' peak load is 745GW!

To be clear, datacenter load queues all around the world are filled with duplicates, and ERCOT is no exception. The 108GW figure is not realistic (and neither is its

datacenter load forecast). A future SemiAnalysis report will dig deeper into this topic, but data is already available in our Datacenter model.



Source: ERCOT

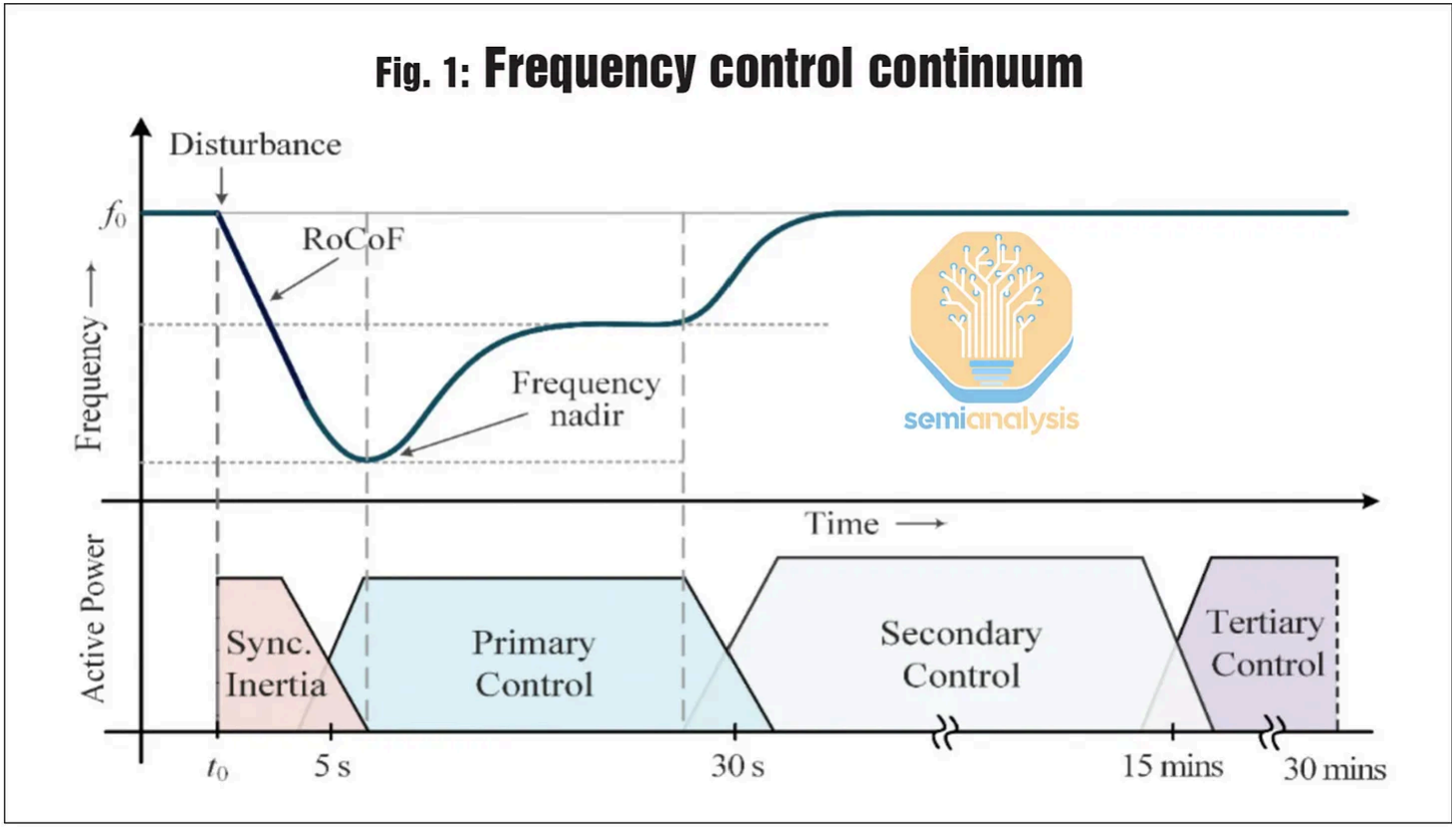
The NERC (North American Electric Reliability Corporation), a regulator that oversees all of North America, is concerned as well, and asking all major transmission utilities how they model datacenter loads when conducting interconnection studies. We dug into these studies, filings, ERCOT meeting documents, and more, to better understand the magnitude of the problem. We explain everything below.

Problem 1: Managing Fast Power Fluctuations

Electric demand changing over time as loads turn on or off is nothing new, and has been managed for decades by electric supply change in tandem on a split-second basis. But managing hundreds of megawatts in a fraction of a second is an unprecedented challenge for operators. And this is precisely the threat posed by Gigawatt-scale AI Datacenters.

Supply change typically involves activating or deactivating electric generators, or directing generators to **ramp** output up or down. The **ramp rate** for generators is measured in **MW per minute** (MW/min), such that a generator with a ramp rate of 10 MW/min can increase or decrease its output by 100MW within 10 minutes. Fossil fuel generators have ramp rates between 5-50 MW/minute, and nuclear power plants have ramp rates too slow to react to any grid conditions.

Typically, sub-second voltage and frequency balancing was managed by system **inertia.** Because conventional electric generators are very large spinning magnets, the inherent momentum of those rotating masses could soak up small imbalances in electric supply and demand, at the cost of excess heat and less efficient operation.



Source: Balancing the Grid: POSOCO report on assessment of inertia in the power system

This is increasingly challenged by a changing generation mix. An increasing quantity of electricity is generated by **intermittent renewable** resources, particularly wind and solar. These systems do not generate alternating current electricity at a frequency synchronized to the rest of the grid. Instead, they generate **direct current** (DC) electricity, which is converted into AC electricity through an **inverter**.

Because these inverters are not built around a large rotating mass, they do not have the inertia necessary to *passively* compensate for the imbalances in supply and demand that lead to drift in voltage and frequency. And because these intermittent renewable resources are dependent on weather conditions to generate electricity, they cannot be

dispatched at a MW/min ramp rate like fossil fuel generators, unless they are paired with batteries. Newer tools exist to manage power quality, including dedicated power quality devices like capacitor banks, synchronous condensers, static VAR compensators, and static synchronous compensators (STATCOMS).

Problem 2: Risk of Cascading Blackouts

Although ERCOT discussed power quality concerns at length, their notes suggested they had a bigger concern: cascading blackouts.

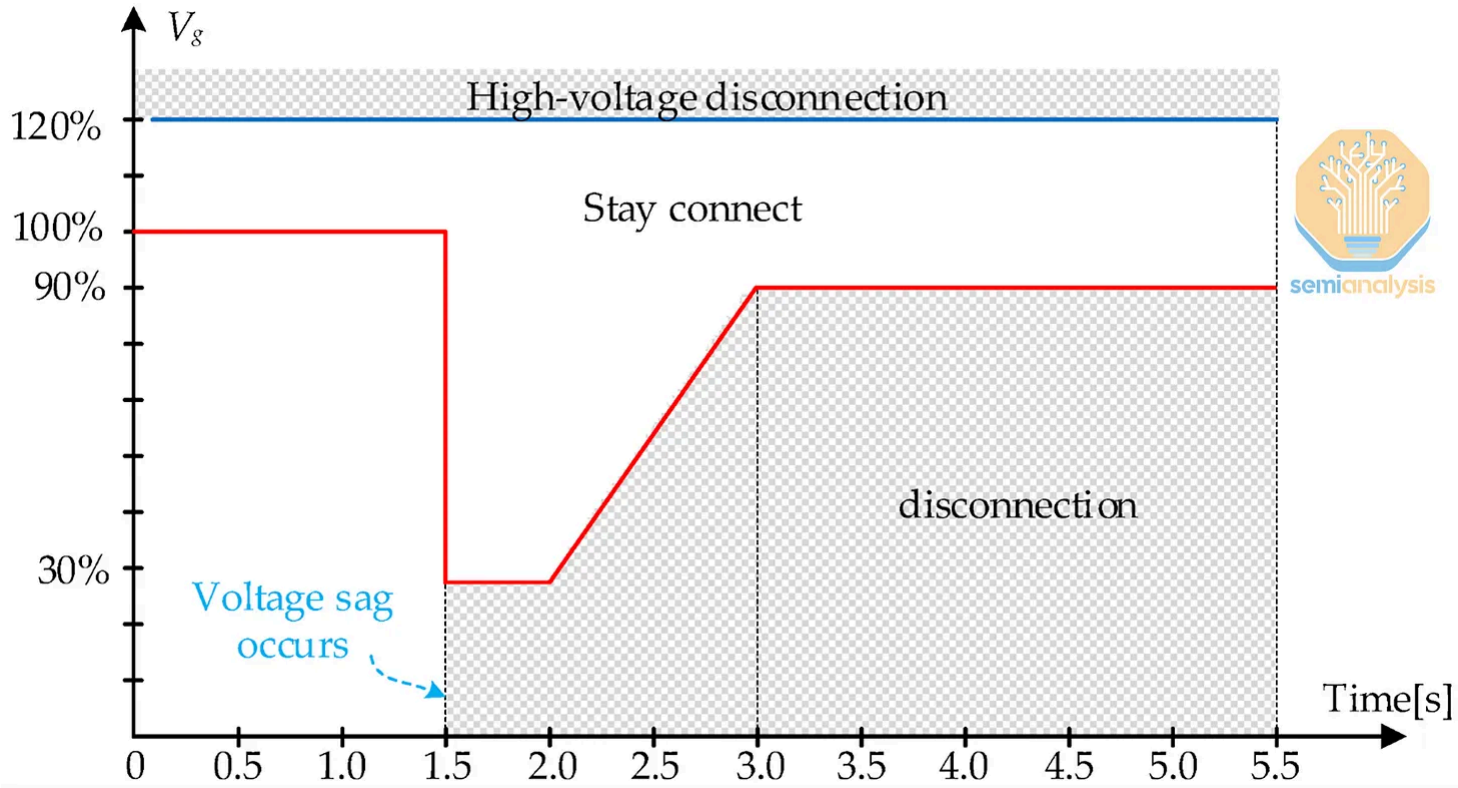
Low Voltage Ride-Throughs (LVRTs), Briefly

ERCOT considered a particular fault response relevant to datacenters: the **low voltage ride-through** (**LVRT**). A low-voltage ride-through is not a response to a total power outage so much as a **transient** blip in which input voltage may sag, for example, 30% below baseline for a length of time between 30 milliseconds and 5 seconds. This type of outage would reflect standard operation of a distant **recloser** clearing a **fault**. Reclosers are, in a sense, breakers that can “re-close” automatically. If a recloser senses a problem, it will trip, wait for a set amount of time, and reconnect.

Source: Tavrida Electric

Often, the recloser will run this cycle of trip-wait-reconnect-trip two or three times before permanently tripping. This repetition is particularly important for clearing wildlife. Some of the most common causes of electric faults are birds, squirrels, and trees. The wildlife typically touch power lines in the wrong way and cause short circuits. Resetting a recloser can literally zap an object off a power line, allowing the recloser to clear a fault without requiring a line team to drive over and fix the problem. The animals do have problems after this sequence of events.

If the fault was on the circuit directly feeding a datacenter, the datacenter would simply see an outage for a short period of time. However, because the grid is a deeply interconnected system, a fault on a different circuit would send shockwaves through the grid in the form of abrupt drops in voltage. In an LVRT, the datacenter would see voltage drop because of that distant fault, and then the voltage would return once that recloser trips. If that recloser resets without issue, then the datacenter doesn't see any other dips in voltage. But if that recloser cycles a few times before clearing the fault or giving up, then the datacenter may see a few voltage sags in succession. The challenge in an LVRT is to stay online, "riding through" the low voltage blip, without disconnecting from the grid.



Source: Low-Voltage Ride-Through Operation of Grid-Connected Microgrid Using Consensus-Based Distributed Control

Datacenters typically use **uninterruptible power supplies** (**UPSs**) and backup power generation to manage LVRTs. Our Datacenter Anatomy - Electrical Systems report explains how power flows in a datacenter and relevant equipment. If grid-supplied voltage dips, the UPS can react near-instantaneously, switching the datacenter from grid power to battery energy storage (typically good for five minutes of operation). This handoff is seamless enough that it does not force electronics to shut down. If grid voltage recovers, the UPS can reconnect the datacenter to the grid. However, if the UPS senses multiple voltage sags in a row, like what would happen when a recloser cycles to clear a fault, then the UPS may permanently disconnect from the grid and switch the datacenter to backup power generation (typically diesel generators).

This switch to backup power is fine for a datacenter, the diesel backup fuel is expensive, but this double handoff from the grid to the UPS to the backup generators *does not interrupt operation.* However, this switching operation would **cause serious problems for the broader electric grid**, because it takes hundreds of megawatts if not gigawatts of electric demand off the grid in an instant. This in turn causes voltage and frequency fluctuations from the sudden imbalance between electric supply and demand, which then may cause *other* generators or large loads to trip offline in a **cascading grid failure**.

Note that this isn't a new issue. In July 2024, a faulty transmission line caused 1.5GW of datacenters in Virginia to unexpectedly disconnect from the grid and turn on their backup power. Dominion Energy successfully managed the issue without a major outage, but had to take drastic action. But given the load growth coming to the US, and the aforementioned AI training load profile, the Virginia issue could become much more common.

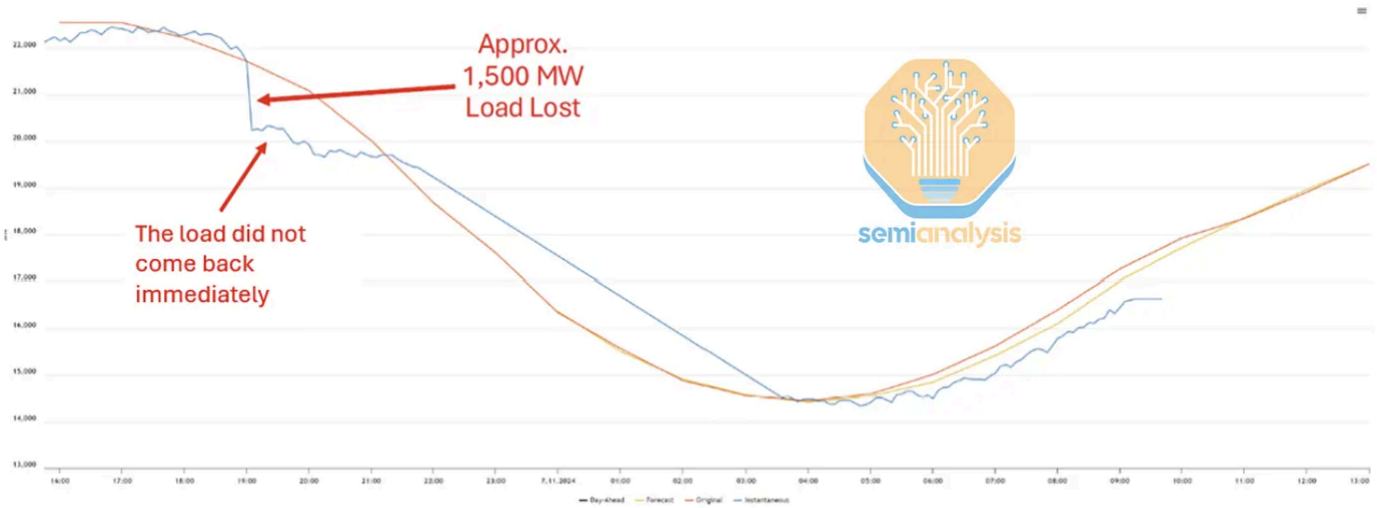


Figure 2: System Load Chart

Source: North American Electric Reliability Corporation (NERC)

The Nightmare Scenario Pt. 1: Datacenter Disconnection Risk

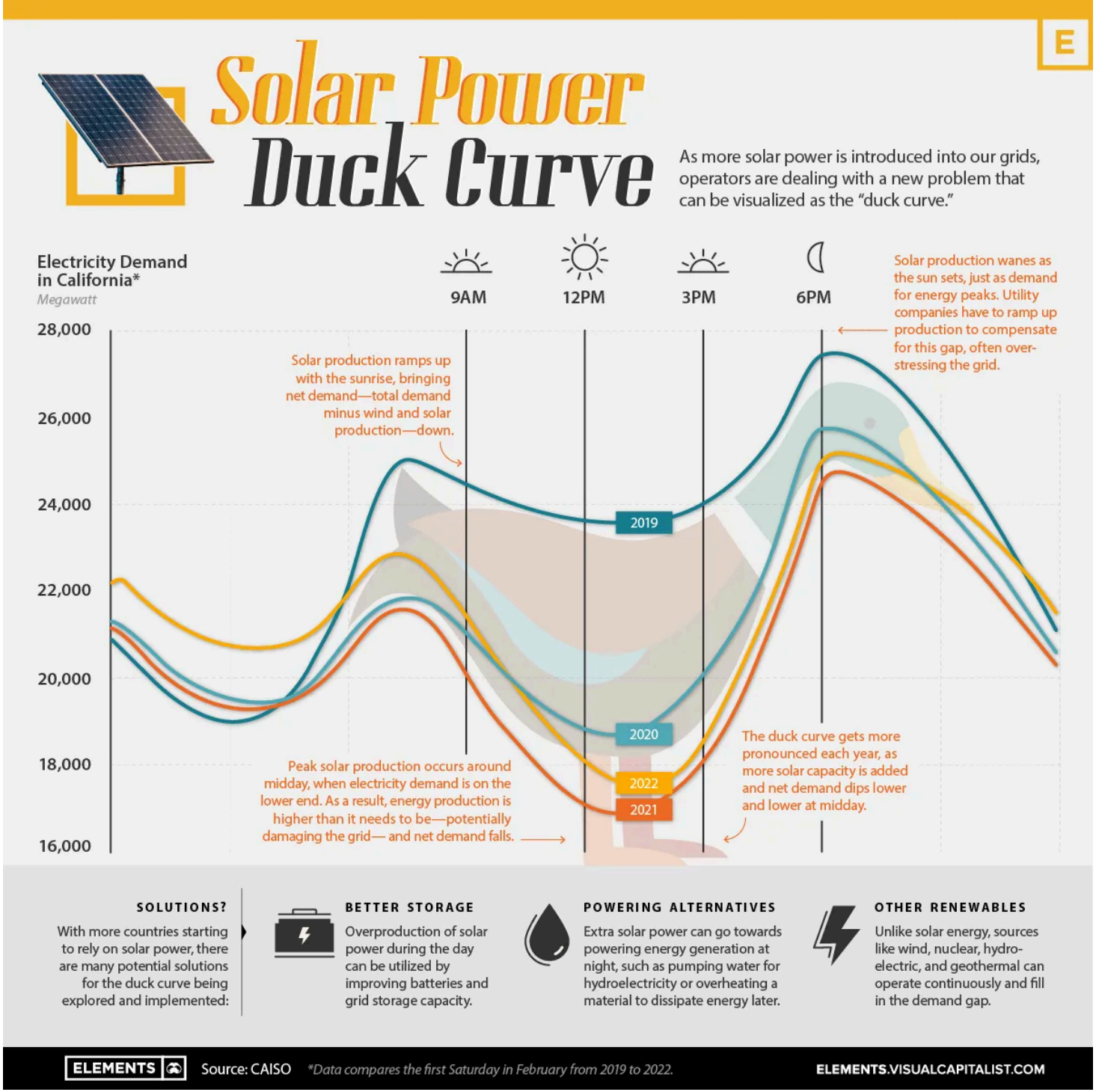
A pair of ERCOT presentations at a May 2025 meeting described a potential nightmare scenario.

The first presentation by Yunzhi Cheng presented a model of what it would take to knock out datacenters with a low voltage ride-through failure. The model looked at two weather scenarios overlapping with two fault response scenarios.

The weather scenarios were:

- Summer Peak (SP): maximum electric load across Texas; a late afternoon three days into a heat wave.

- High Renewable Minimum Load (HRML): "**duck curve**" electric load across Texas; midday on a sunny spring or fall day, the intersection of minimum electric load versus maximum behind-the-meter solar production.



Source: Visual Capitalist

The fault response scenarios were:

1. Datacenters immediately trip if voltage drops below 75% of baseline
2. Datacenters can manage an LVRT of 70% voltage for 20 milliseconds, but no lower voltage, and no longer than 20 milliseconds.

Cheng modeled a fault on a 345kV transmission line (about 1/6 of what is necessary to supply Austin, TX) in a West Texas substation. Combining the two sets of scenarios, he modeled fault outcomes based on four potential assumption sets:

- Fault at Summer Peak, trip if voltage drops below 75% baseline
- Fault at Summer Peak, manage LVRT of 70% voltage for 20 ms
- Fault at High Renewable Minimum Load, trip if voltage drops below 75% baseline
- Fault at High Renewable Minimum Load, manage LVRT of 70% voltage for 20 ms

Cheng found that in all four assumption sets the ERCOT grid system would see at least 1.5 GW of datacenter load disconnect from the grid almost immediately. If that fault happened during a duck curve day with datacenters that are not equipped for an LVRT, then the grid could see 2.5 GW of load, every datacenter currently in West Texas, disconnect from the grid at approximately the same time. Note this load of West Texas Datacenters will soar past 10GW rapidly.

Base Case Datacenter Disconnection Risk

Base Case Data Center Disconnection Risk		
Study Cases	Immediate Trip	LVRT 20 ms
Summer Peak (SP)	~1,500 MW	~1,500 MW
High Renewable Minimum Load (HRML)	~2,500 MW	~1,500 MW

Source: ERCOT

Installing a synchronous condenser (basically, a giant electromagnetic flywheel) at the point of interconnection for *every West Texas datacenter* helped, by adding system inertia right next to each load. But even this countermeasure left 1.3-1.9 GW of load at risk of disconnecting.

Datacenter Disconnection Risk + SynCon

Data Center Disconnection Risk + SynCon		
Study Cases	Immediate Trip	LVRT 20 ms
Summer Peak (SP)	~1,500 MW	~1,500 MW
High Renewable Minimum Load (HRML)	~2,500 MW	~1,500 MW

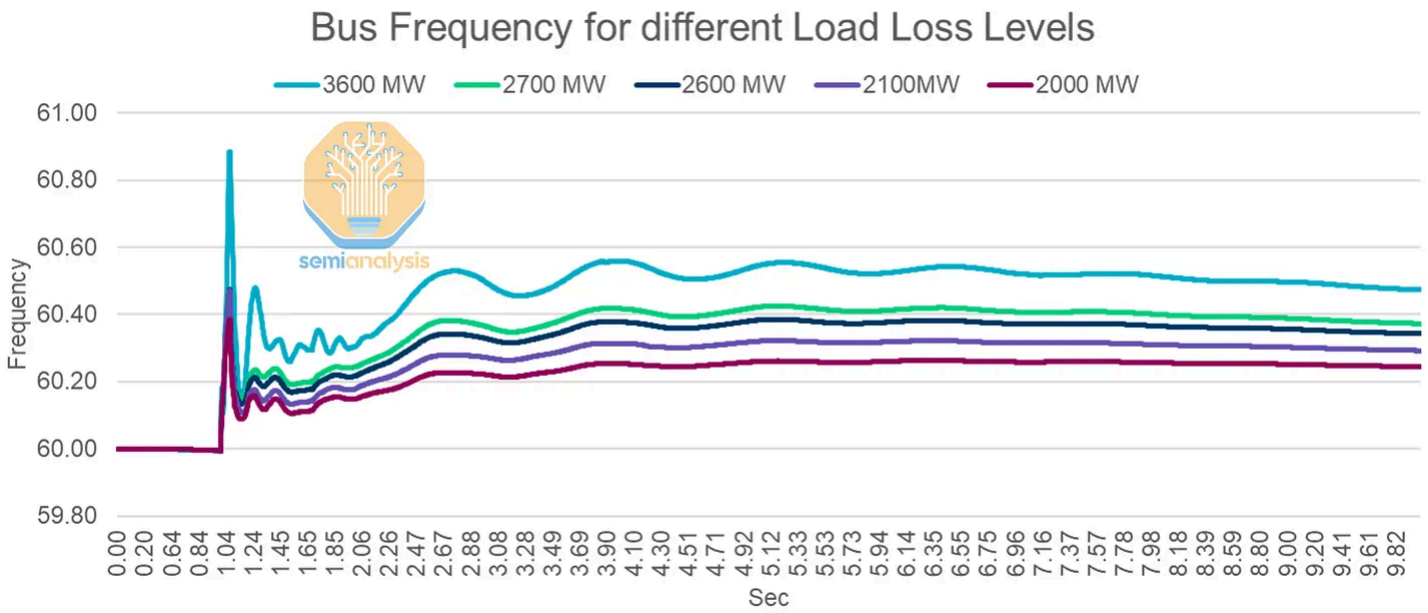
Source: ERCOT

Moreover, synchronous condensers are expensive systems. The capital cost for these systems comes out to $30k-60k per MVA Reactive. At the installation spec used in Cheng's model, this would cost $10M-20M to install for a 1 GW datacenter.

Source: Wikipedia - Synchronous Condenser

The Nightmare Scenario, Pt 2: Cascade Risk

The second presentation by Luis Hinojosa carried forward the knock-on consequences of so many datacenters disconnecting from the grid in response to a transient fault. He found that if more than about 2.6 GW of electric load disconnected from the grid at once, grid frequency across the ERCOT system would rise beyond the 60.4 Hz “danger zone” set by the ERCOT Dynamics Working Group.

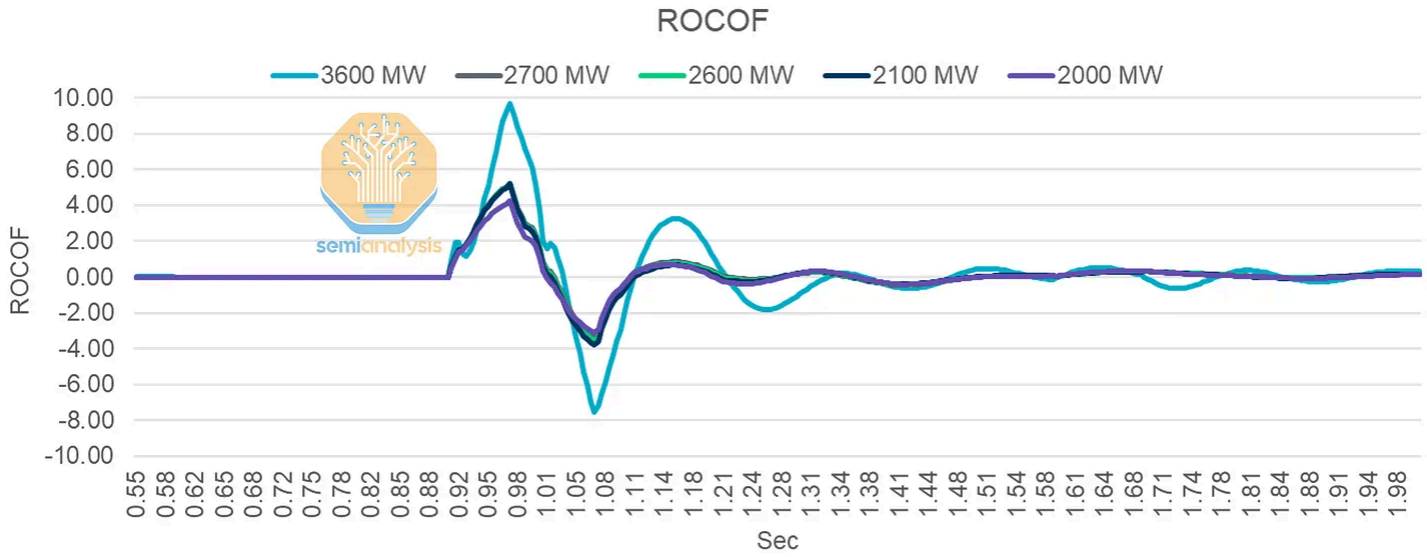


Source: ERCOT

MW Load Loss	Steady State Freq
~3,600	**60.58**
~2,700	**60.42**
~2,600	**60.37**
~2,100	60.27
~2,000	60.25

Source: ERCOT

Even a smaller 2 GW disconnection would also cause **rate of change of frequency** (**ROCOF**) instabilities beyond what ERCOT considers safe.



Source: ERCOT

MW Level	Max Absolute ROCOF
~3600	9.70
~2700	5.23
~2600	5.21
~2100	5.18
~2000	4.25

Source: ERCOT

That scale of disconnection would also cause problems for voltage quality, if more than 2.5 GW of load disconnected at once, large swaths of the Texas grid would see damaging voltage issues.

Scenario (Load Loss Level - MW)	Status
~3,600	Insecure - Wide Area Voltage Issues
~2,600	Insecure - Wide Area Voltage Issues
~2,450	Insecure - Wide Area Voltage Issues
~2,100	Insecure - Local Voltage Issues
~2,000	Insecure - Local Voltage Issues
~1,950	Secure

Source: ERCOT

Hinojosa aggregated his findings into two **operating limits** of load loss: if the entire ERCOT system lost 2.6 GW of load in rapid succession, or if the West Texas load zone lost 2.0 GW, then the Texas grid would be dangerously unstable and at risk of cascade blackouts.

Nightmare Scenario, Pt. 3: This has Already Happened in the Iberian Peninsula

The grid stability issues outlined by Cheng's and Hinojosa's analyses reveal a pathway to grid instability remarkably similar to the 28 April 2025 blackout in the Iberian Peninsula. In that case, 2.2 GW of generation tripped offline, reportedly because of miscalculated dispatch decisions by the local grid operator. This lead to cascading voltage and frequency fluctuations that tripped breakers all over Spain and Portugal. Because the Iberian grid is relatively isolated from the rest of Europe, external connections could not stabilize the grid, leading to a total collapse within 27 seconds.

The same scenario is possible in Texas: if 2-2.5 GW of datacenter load tripped off the grid in short succession, then similar voltage and frequency fluctuations could cause cascade failures through Texas. And because the Texas Interconnection has *only four* connections to other grid networks, those external connections can do little to stabilize the grid. And once those failures begin to echo through Texas, it would be too late to catch. All this, potentially started by a squirrel stepping on the wrong power line at the wrong time near a West Texas substation.

How to avoid the nightmare scenario - solutions

Note that **the responsibility of degrading system-level power quality largely falls on the datacenter side**, if a datacenter causes harmonics issues, he has to pay the bill. Of course, this has led the industry to actively look for solutions.

Below we start by discussing Battery Energy Storage Systems (BESS), and for subscribers we will discuss other hardware-based solutions, their associated supplier, and it all fits in Nvidia's new 800V DC power architecture.

The Promise of Battery Energy Storage Systems (BESS)

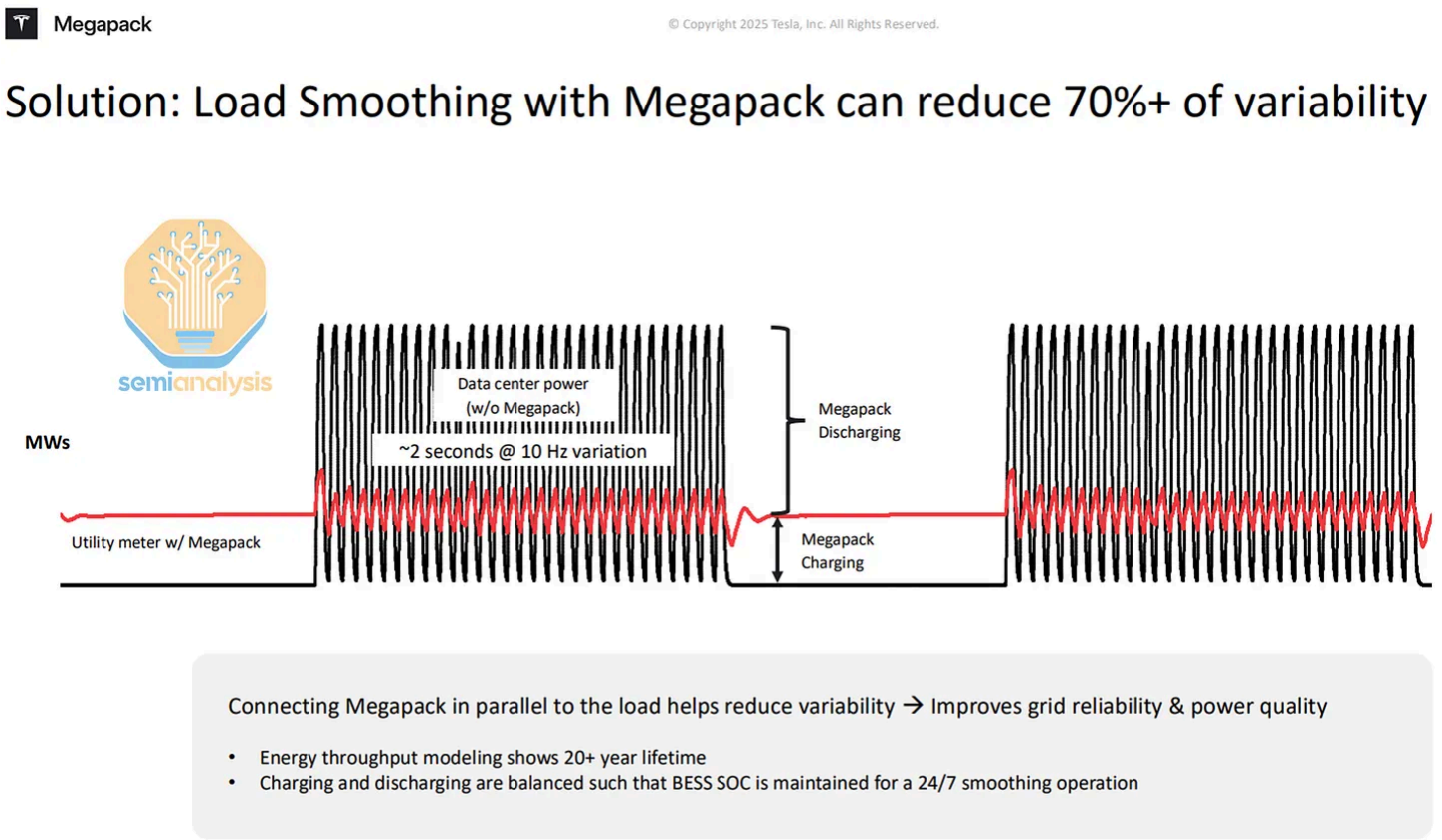
Tesla believes that the best solution to the power quality problems that datacenters face are large-scale batteries on the scale of 100s of megawatts or gigawatts. At a May 2025 ERCOT meeting, Tesla presented a slide deck basically unchanged from the deck presented at a larger April 2025 workshop on large loads run by the North American Electric Reliability Council (NERC). The slide deck focused on their Megapack 2 XL battery pack, shown below.

Source: Tesla

It demonstrates the promise of **battery energy storage systems** (**BESS**) for datacenters, regardless of manufacturer. The killer feature of BESS within datacenters is that these systems can charge and discharge hundreds of megawatts within seconds, allowing these batteries to react to datacenter load fluctuations at the appropriate reaction speed *and* power output.

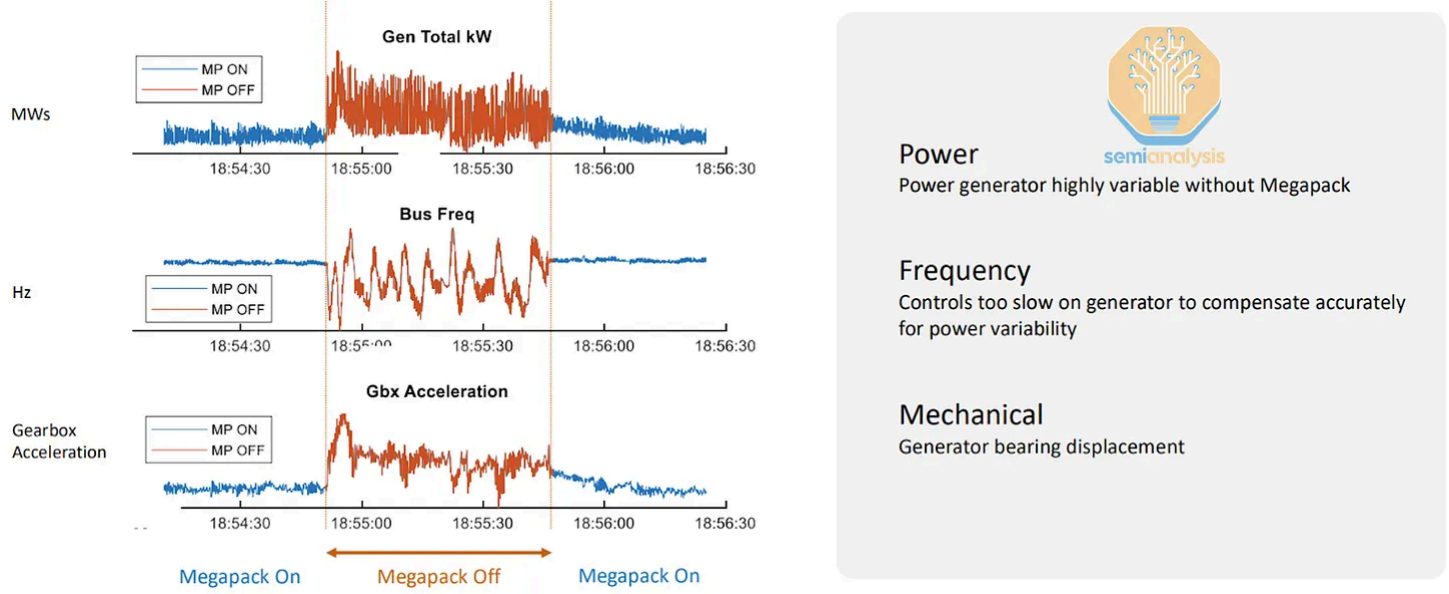
BESS for Power Quality and Grid Stability

A megawatt-scale battery connected to an inverter can manage power quality issues through rapid charging and discharging, this is called **fast frequency response.**



Source: Tesla

Tesla describes BESS as a more viable option for managing demand fluctuations than alternative solutions like diesel generators or capacitor banks. We explain how datacenter capacitors work and whether we think Tesla's claim is true behind paywall. Tesla's deck assumes a Megapack 2 XL would be installed *in tandem* with existing measures like generators and UPSs. One slide suggests that installing a BESS in series with a generator allows for smoother operation (and by extension improved lifespan) for that generator.



Source: Tesla

Tesla mentions that capacitor banks are also an option, but they note correctly that those capacitor banks cannot manage load smoothing at the second scale. By contrast, BESS can manage load fluctuations at the MW/millisecond, MW/second, *and* MW/minute basis, which offers more flexibility than capacitor banks, diesel generators, or grid-scale resources can manage.

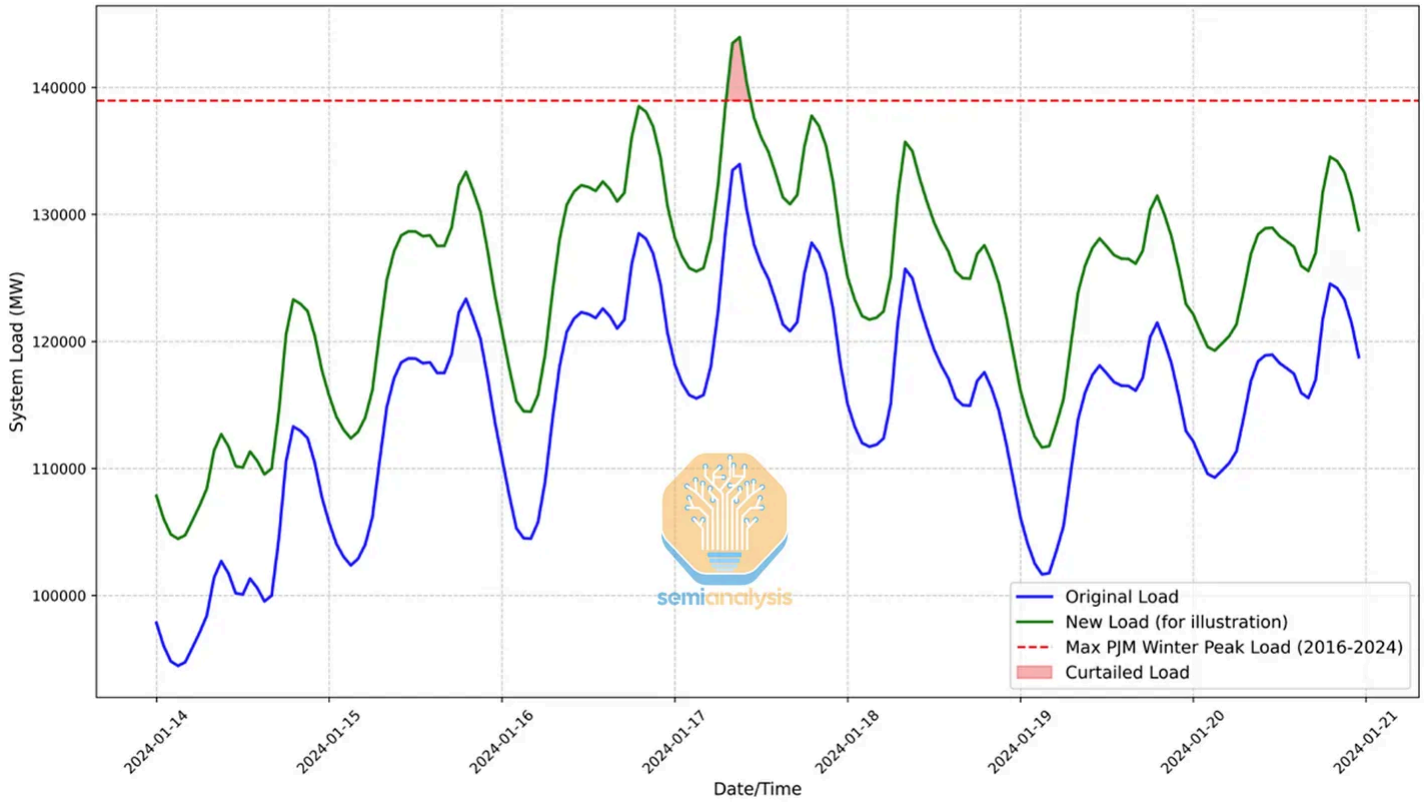
BESS can also improve responses to low voltage ride-throughs (LVRTs) as described above. Notably, Tesla describes the functionality of Megapack *in tandem with an existing UPS*, instead of describing their BESS solution as a *replacement* for that UPS. Specifically, Tesla describes their BESS as a means of compensating for the baseline UPS behavior of tripping offline if voltage sags multiple times in a row, if the UPS clicks to off-grid operation, then the BESS would charge from grid load, so that the grid sees “mimicked” load while the UPS resets manually.

BESS for Demand Response

Tesla also suggests an ancillary benefit to datacenters, **Demand Response**. This practice has multiple names, grid-edge response, flexible load management, load curtailment, load adjustment, but as of today, adoption has not been particularly high (besides cryptominers in Texas), due to a lack of incentives. The concept of Demand Response is simple, if you participate to such a program, the grid can force you to shut down your load, but will compensate you for it.

In today's power-constrained environment, the incentives shift. Demand response enables transmission systems to unlock more capacity, and faster time-to-power. Per one study performed by Duke University, if new load could achieve demand response for 20-90 hours per year, the ERCOT system alone could enable 6.5-14.7 GW of new load (not exclusive to datacenters) without additional system upgrades.

This is due to fundamental grid design principle. Many potential locations have limitations on how much electricity can be generated or transmitted to that given location. However, those constraints are only relevant for 20-90 hours per year, or 0.25-1% of the year. Those **peak times** when the grid sees maximum electric load for the year are the *specific* design specification for much physical infrastructure the grid needs. Notably, because those peak times are driven primarily by air conditioning and behind-the-meter solar, they're reasonably predictable: late afternoons, deep into summer heat waves, as behind-the-meter solar generation falls away for the day.



Source: Rethinking Load Growth - Assessing the Potential for Integration of Large Flexible Loads in US Power Systems

xAI's participation to a demand-response program in Memphis, TN was key to accessing grid power faster than typical timelines. While onsite natural gas turbines enabled the cluster to be set up in four months, xAI has also built a substation and is drawing 150MW from the grid - less than a year after requesting the load, which is remarkably fast.

Source: Tesla

However, implementing demand response has challenges on the customer and the utility side. On the customer side, no one likes *doing* demand response, in many cases, it requires cutting lights, air conditioning, and "nonessential" process loads. Backup power becomes a necessity, and Tesla suggests that BESS are a good fit: instead of cutting load, the datacenter can discharge energy from the BESS to reduce electric demand at the meter.

Source: Tesla

Notably, this requires charging the battery before a **peak event** is called and discharging the battery after it is called. Charging the battery can be a challenge, because utilities typically only identify the possibility of a peak event with 24-hour notice and identify the likely 3-6 hour window of the peak event with 3 hour notice. Even if utilities are reliably prompt about notifying customers about peak events (there is reason to doubt that), there is *very little time to react* unless the BESS was fully charged. Without careful **state-of-charge** (**SOC**) management across multiple large loads, an advanced demand response program might see peak loads shift to 1 PM or 2 PM, driven by large loads charging on-premise batteries in preparation for a *projected* peak event at 5 PM. Additionally, any SOC spent on demand response is SOC *not* held ready in the event of an LVRT event or larger power outage. Every BESS must be programmed to trade off the twin mandates of demand response and backup power, prioritizing one use case deprioritizes the other in equal proportion.

However, even installing a BESS does not address the utility-side challenges with demand response. First and foremost, utilities are typically very bad at demand response. Utilities are often 10-20 years behind on IT infrastructure, and Demand Response Management Software (DRMS) remains an immature market. Utilities have broadly struggled at technical building blocks of demand response, like:

- Collecting and managing the data necessary for good peak forecasts
- Writing and operating good peak forecasting tools
- Notifying customers about peak events
- Integrating demand response measures into patchwork building management systems (BMS) in commercial and industrial buildings
- Accurately measuring customer demand response
- Converting demand response into bill credits

Beyond implementation, utilities struggle to offer incentive payments to make demand response worth the effort. As an energy-only market, ERCOT has no strict cost for electric capacity, which would put a hard price on peak demand. The organization has approved a market reform called a **Performance Credit Mechanism (PCM)** that will likely be instituted in 2026 or 2027. However, even if that PCM cost reflects high peak costs like MISO's and PJM's controversially high $8-15 per kW-month ($270-500 per MW-day) (inclusive of capacity and transmission), that may come out to $160,000-300,000 per month for 20 MW of demand reduction, inclusive of utility labor, DRMS SaaS fees, and bill credits to customers. That might look like five-figure bill credits on a datacenter electric bill. That is at best a rounding error and at worse an insult for all the effort and capital that went into *implementing* demand response.

The Cost of BESS

Tesla's slide deck is cagey about net cost for the Megapack system, because the likely cost is substantial. Per the Lazard June 2024 LCOE report, a 100 MW BESS would cost **$38-80M** for a two-hour battery (as described in Tesla's slide deck) and **$76-157M** for a four-hour battery (as would be necessary for functional demand response or backup power). At those installation prices, a BESS suitable for a GW-scale datacenter would cost **close to a billion dollars**, and for that price, Tesla would *not* consider BESS a replacement for a UPS or a diesel generator. This is simply an *additional* cost in construction time, CAPEX, land use, supply chain vulnerability, and headache.

Therefore, is BESS the best solution to manage datacenter load fluctuations? Today we focus on behind-the-meter BESS, but a future SemiAnalysis report will explore the broader role of BESS and renewables in the power grid, in the context of a significantly higher load growth than the last 20 years.

Below, we explore hardware-based alternatives, explain how they compare and discuss associated suppliers.

Hardware-based solutions to manage AI Training Load Fluctuations

Previous

Next

A guest post by

Ajey Pandey

Analyst, SemiAnalysis