# Speed to Power

What Flexibility Is Worth to an AI Campus, from Seconds to Months

Abhay Chakra Sadineni
University of California, Berkeley
abhaychakra@berkeley.edu

November 2026
DOI: 10.5281/zenodo.24210931

**Abstract**

Grid operators now ask large AI campuses for three kinds of flexibility. On a scale of seconds, they want synchronised training jobs to stop swinging tens of megawatts. On a scale of hours, they want the right to curtail the campus during the few hundred hours a year when the system is stressed. And on a scale of months to years, campuses that accept curtailment may be connected sooner. This paper prices each layer for a campus built on current Blackwell hardware, using public specifications, published grid studies and current battery costs. The power-supply energy storage NVIDIA now ships, 65 joules per GPU, holds about 1 kWh across 100 MW of racks, enough to ride through a 50 MW swing for about 70 milliseconds. Swings with periods of seconds need tens of kilowatt-hours, which is little energy but a large amount of fast power. For curtailment, a widely cited study finds that 76 to 126 GW of new US load could be added if it accepts curtailment of 0.25 to 1 percent of its maximum annual energy. On GPU rents and capital costs in late 2026, the compute lost at those rates is worth about $18,000 to $167,000 per MW a year. A battery that carries half the load for two hours costs about $67,000 per MW a year. Even on the optimistic assumption that it avoids all curtailment, it pays only above curtailment of roughly 0.4 to 0.9 percent. The largest number is the third. One month of earlier operation is worth 730 full-load hours, as many as 33 years of curtailment at 0.25 percent or 8 years at 1 percent, before discounting and whatever the price of compute. If flexibility shortens the wait for a grid connection, which is plausible but not yet measured, it is mainly a way to buy time.

**Keywords:** data-centre flexibility; demand response; interconnection; battery storage; power quality; AI infrastructure.

## 1 Introduction

On 10 July 2024 a fault on a 230 kV line in northern Virginia caused about 1,500 MW of data-centre load to switch to backup power at once. The North American Electric Reliability Corporation found that none of the load was disconnected by utility equipment. The facilities' own protection systems moved it [1]. On 22 July 2026 a similar fault led to about 3,800 MW of data-centre load tripping offline in the same region, which PJM called the largest such event in its history [2]. On 20 June 2025 Texas enacted Senate Bill 6. It applies to loads of 75 MW or more, lets the grid operator direct new large loads with on-site backup generation able to serve at least half their demand to run it or curtail during emergencies, and requires curtailment protocols for large loads connected after 31 December 2025 during firm load shed [3]. In August 2025 Google announced its first agreements to curtail machine-learning workloads at the request of two utilities, presenting them as a way to connect new capacity faster [4].

A large AI campus is now expected to be flexible in three different ways, on three different time scales. This paper asks what each is worth to the campus, and what it costs. It uses public specifications for NVIDIA's Blackwell rack systems, published grid studies, the 2025 battery cost benchmarks of the National Renewable Energy Laboratory, and GPU rents and prices from earlier work by the author [5].

## 2 Power swings over seconds

Large training jobs run thousands of GPUs in lockstep. Each step alternates between computation, when the GPUs draw close to full power, and communication between GPUs, when they draw much less. Researchers from NVIDIA, Microsoft and OpenAI showed that these swings add up across a campus to tens or hundreds of megawatts. If the swings fall near the resonant frequencies of generators or long lines, they could damage grid equipment [6]. Their proposed remedy combines software and firmware controls on the GPUs with energy storage at the rack.

NVIDIA's GB300 NVL72 puts that storage into the rack's power supplies, about 65 joules per GPU in electrolytic capacitors, and pairs it with power caps and controlled power burning at the start and end of jobs [7]. NVIDIA reports that this can cut peak demand on the grid by up to 30 percent. How far does that storage go?

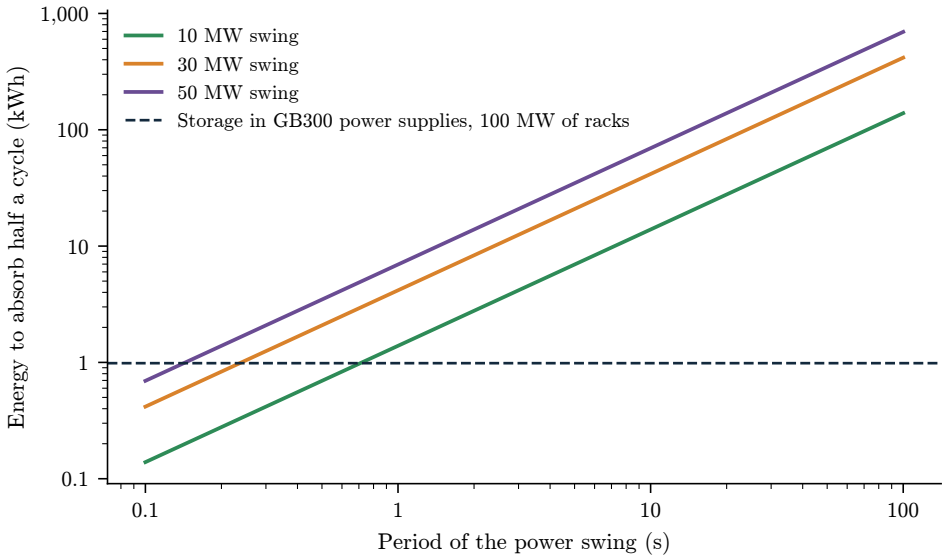



Figure 1. Energy needed to absorb half a cycle of a square-wave power swing, by period and amplitude, against the energy stored in GB300 power supplies across 100 MW of NVL72-class racks (about 54,500 GPUs at 132 kW per 72 GPUs).

A 100 MW campus of NVL72-class racks, at 132 kW per 72 GPUs, holds about 54,500 GPUs. At 65 joules each that is 3.5 MJ, or about 1 kWh. It can absorb a 10 MW swing for 355 milliseconds or a 50 MW swing for 71 milliseconds. Figure 1 compares this with the energy needed to absorb half a cycle of a square swing. For periods of a few seconds and amplitudes of tens of megawatts, the need is tens of kilowatt-hours.

That is very little energy by the standards of grid batteries, and a great deal of power. A 50 MW swing needs 50 MW of storage power regardless of how few kilowatt-hours sit behind it. It also cycles the storage thousands of times an hour. That cycling suits capacitors and flywheels better than lithium-ion cells, and the cost is set by power rating rather than energy. The capacitors in the power supplies deal with sub-second transients. Swings of a second or more have to be damped by the GPUs themselves or by fast facility-scale storage.

# 3 Curtailment over hours

## 3.1 How much curtailment

The most cited estimate of how much flexible load the US grid can absorb comes from Norris et al. [8]. Across the 22 largest balancing authorities, they found that 76 GW of new load could be integrated if it accepted curtailment of 0.25 percent of its maximum annual energy, 98 GW at 0.5 percent, and 126 GW at 1 percent. Curtailment would fall in about 85, 177 and 366 hours a year respectively. Average events would last 1.7, 2.1 and 2.5 hours. For 88 percent of curtailment hours, at least half the new load could keep running [9].

Measured in full-load hours, those rates are modest. A 0.25 percent rate takes away the equivalent of 21.9 hours of full operation a year. A 1 percent rate takes away 87.6.

## 3.2 What curtailed compute is worth

The value of a curtailed kilowatt-hour depends on what the GPUs were doing. Two bounds bracket it. If the campus is training its own model, a pause delays the job, and the cost is the idle capital. A DGX B200 listed at $515,410 for 14.3 kW [10], recovered over five years at 10 percent, carries a capital charge of $1.09 per kWh of IT load. If the GPUs are rented out, the cost is lost rent. At the Silicon Data index of $6.63 an hour for a B200 on 26 October 2026 [11], rent is $3.71 per kWh of IT load. Utilisation is assumed at 70 percent. Dividing by a facility overhead of 1.3 converts each figure to the grid meter. That overhead covers cooling, power conversion and other site loads together, so it is larger than the heat-rejection energy alone, which the companion paper on rack cooling puts at 0.035 to 0.15 kWh per kWh [12]. Finally, and the rent case is net of the electricity saved, at 8.62 cents per kWh [13]. The value of a curtailed kilowatt-hour at the meter is then $0.83 to $1.91.

| Curtailment rate | Hours a year | Events a year | Full-load hours lost | Value lost ($ thousand per MW-year) |
|---|---|---|---|---|
| 0.25% | 85 | 50 | 21.9 | 18 to 42 |
| 0.5% | 177 | 84 | 43.8 | 37 to 84 |
| 1.0% | 366 | 146 | 87.6 | 73 to 167 |

Table 1. Curtailment at the rates studied by Norris et al. [8], and the value of the compute lost, using the bounds in the text. Events are hours divided by average duration.

## 3.3 Battery or pause

A campus can meet a curtailment request by pausing work, by moving it to another site, by running on-site generation, or by drawing on a battery. A battery that carries half the load for two hours would cover the typical event, given that most curtailment hours leave at least half the load running. NREL's 2025 benchmark puts a 2024 utility-scale lithium-ion system at $241 per kWh of energy plus $372 per kW of power, with a 15-year life and 85 percent round-trip efficiency [14]. Recovered at 8 percent over 15 years, with fixed operating costs of 4 percent of capital a year following NREL's convention, such a battery costs about $67,000 per MW of campus load a year. A battery covering the whole load for four hours costs about $210,000.

Figure 2 sets the two side by side, assuming, optimistically, that the battery avoids all curtailment and ignoring the 15 percent of energy lost in each charge and discharge. In practice a half-load, two-hour battery cannot cover the deeper or longer events, so the comparison flatters it. At 0.25 percent curtailment, pausing is cheaper than the smaller battery by a factor of 1.6 to 3.7.

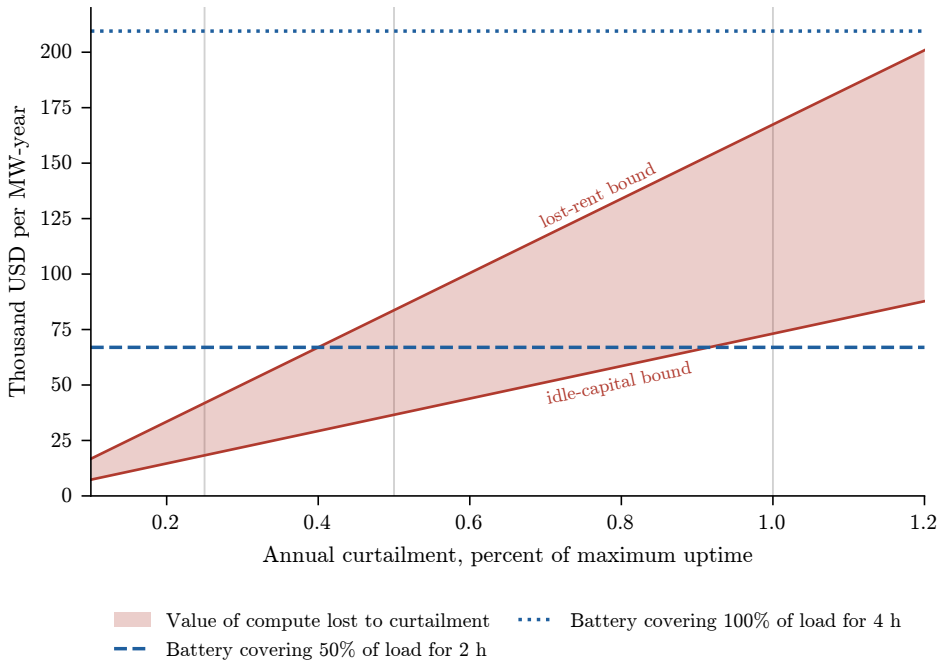



Figure 2. Annual value of compute lost to curtailment, between the idle-capital and lost-rent bounds, against the annual cost of two battery sizes. Vertical lines mark the three rates studied by Norris et al.

The battery pays only when curtailment exceeds about 0.40 percent, if every lost hour would have been rented, or about 0.92 percent, if the work could simply be delayed. A full four-hour battery does not pay within the studied range on avoided curtailment alone. Batteries also earn revenue from energy arbitrage and grid services, which this comparison leaves out, and they can serve the second-scale layer described above. Those revenues could change the answer at a particular site. They do not change the general result that, at the curtailment rates the grid studies contemplate, a battery bought only to avoid pausing GPUs is hard to justify.

## 4 Connecting months sooner

The third layer dwarfs the other two. New large loads in many US markets wait years for grid connections, and both Norris et al. and Google present flexibility as a way to shorten that wait [4, 8]. How much sooner a flexible load is connected depends on the utility and the site, and no public data yet measure it. The value of each month of earlier operation is easy to state.

One month is 730 hours. At the bounds used above, a month of full operation per MW of campus load is worth $0.61 million in capital that would otherwise sit idle, or $1.40 million in rent net of electricity. The idle-capital figure assumes the hardware has been bought and is waiting for power. Set against curtailment, the ratio does not depend on the price of compute at all. It is simply 730 hours divided by the full-load hours lost each year. Before discounting, one month of earlier operation equals 33 years of curtailment at 0.25 percent, 17 years at 0.5 percent and 8 years at 1 percent. Discounting narrows the gap without closing it. At 10 percent, 33 years of curtailment worth $18,000 a year have a present value of about $173,000, against $0.61 million for the month.

For a 300 MW campus, a connection three months earlier at the idle-capital bound is worth about $550 million. The same campus, curtailed at 0.5 percent, would give up compute worth about $11 million a year at the same bound.

## 5 Discussion

Three conclusions follow for anyone planning, financing or regulating an AI campus.

The layers need different equipment. Sub-second transients belong to power-supply capacitors, which now ship with the racks. Swings of seconds belong to GPU firmware and fast facility storage sized by power. Curtailment of hours belongs mostly to scheduling, deferring training or moving inference elsewhere, with batteries justified only at higher curtailment rates or with other revenue.

Flexibility should be sold for time. A campus that offers a utility curtailment of 0.5 percent is giving up compute worth tens of thousands of dollars per MW a year. If the return is a connection even a few weeks earlier, the trade is strongly in its favour, though how much earlier is something each developer will have to establish with its utility. Developers should negotiate for it explicitly, and utilities and regulators should make the trade explicit in interconnection rules, as Texas has begun to do [3].

Ride-through matters as much as curtailment. The 2024 and 2026 Virginia events were unplanned load losses caused by the campuses' own protection systems [1, 2]. A cluster of campuses that drops 1,500 MW in a voltage dip imposes a cost on the grid that no amount of scheduled flexibility offsets. Ride-through settings, coordinated with the utility, are the cheapest reliability measure available.

These results link to two earlier papers. In a power-limited site, the rent each kilowatt earns decides when a GPU should be replaced [5]. Every kilowatt saved on cooling by running hot water is a kilowatt available for computing [12]. Flexibility is the third lever on the same constraint. All three treat grid power as the scarce input of an AI campus.

## 6 Limits

The curtailment rates and durations are averages from one study of 22 balancing authorities, built on historical load and system data. Individual sites and future systems will differ, and events will cluster in extreme weather. The battery comparison assumes the battery avoids all curtailment and ignores charging losses, both of which favour the battery. The value of curtailed compute uses two bounds and an assumed 70 percent utilisation, and treats every lost hour as equally valuable. In practice, inference demand may be highest when the grid is stressed. Battery costs are 2024 national benchmarks, before later tariff changes. On-site generation is not costed, because air permits limit the hours backup engines may run and those limits vary by jurisdiction. The swing analysis uses square waves and ignores the frequency content that matters for resonance. The value of earlier connection assumes that flexibility actually shortens the wait, which is plausible but not yet measured.

## Data and code

The script that reproduces every number in this paper is available from the author's research page.

## Funding and interests

No external funding was received. The author has no financial interest in any company named.

# References


[1] North American Electric Reliability Corporation (2025). Incident review: Considering simultaneous voltage-sensitive load reductions. 8 January 2025. https://www.nerc.com/pa/rrm/ea/Documents/Incident_Review_Large_Load_Loss.pdf

[2] Howland, E. (2026). PJM eyes data center, crypto reliability requirements after 3.8 GW of load trips offline. Utility Dive, 12 August 2026. https://www.utilitydive.com/news/pjm-nerc-data-center-crypto-reliability-standards/827653/

[3] Texas Legislature (2025). Senate Bill 6, 89th Legislature, Regular Session. Signed 20 June 2025. https://capitol.texas.gov/BillLookup/History.aspx?LegSess=89R&Bill=SB6

[4] Terrell, M. (2025). How we're making data centers more flexible to benefit power grids. Google, 4 August 2025. https://blog.google/innovation-and-ai/infrastructure-and-cloud/global-network/how-were-making-data-centers-more-flexible-to-benefit-power-grids/

[5] Sadineni, A. C. (2026). The economic life of an accelerator: Rental prices, power and the six-year depreciation of data-centre GPUs. Preprint, November 2026. https://doi.org/10.5281/zenodo.24210935

[6] Choukse, E., Warrier, B., Heath, S., et al. (2025). Power stabilization for AI training datacenters. arXiv:2508.14318.

[7] Dimitrov, R., Petty, H., Srivastava, N., and Blake, M. (2025). How new GB300 NVL72 features provide steady power for AI. NVIDIA Technical Blog, 28 July 2025. https://developer.nvidia.com/blog/how-new-gb300-nvl72-features-provide-steady-power-for-ai

[8] Norris, T. H., Profeta, T., Patiño-Echeverri, D., and Cowie-Haskell, A. (2025). Rethinking load growth: Assessing the potential for integration of large flexible loads in US power systems. NI R 25-01. Durham, NC: Nicholas Institute for Energy, Environment and Sustainability, Duke University.

[9] Howland, E. (2025). Existing US grid can handle 'significant' new flexible load: report. Facilities Dive, 11 February 2025. https://www.facilitiesdive.com/news/us-grid-headroom-flexible-load-data-center-ai-ev-duke-report/739938/

[10] Zuhair, M. (2024). NVIDIA Blackwell DGX B200 listed, prices start at a half a million dollars. Wccftech. https://wccftech.com/nvidia-blackwell-dgx-b200-price-half-a-million-dollars-top-of-the-line-ai-hardware/

[11] Silicon Data (2026). B200 Rental Price Index, SDB200RT. Value dated 26 October 2026, accessed 13 November 2026. https://www.silicondata.com/products/silicon-index/b200

[12] Sadineni, A. C. (2026). Hot water for hot chips: Thermodynamics, climate and the cooling cost of 130 kW AI racks. Preprint, November 2026. https://doi.org/10.5281/zenodo.24210933

[13] US Energy Information Administration (2026). Electricity explained: Prices and factors affecting prices. https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php

[14] Cole, W., Ramasamy, V., and Turan, M. (2025). Cost projections for utility-scale battery storage: 2025 update. NREL/TP-6A40-93281. Golden, CO: National Renewable Energy Laboratory. https://docs.nrel.gov/docs/fy25osti/93281.pdf