

©SHUTTERSTOCK.COM/WHITEMOCCA

From Load to Asset: Rethinking AI Data Centers for Sustainable Power Grids

Dual Grid Identity of AI Data Centers

By Nabindra Shrestha, Ahmad Ali, Iqbal Husain, Srdjan Lukic, and Hantao Cui

ELECTRICITY DEMAND FROM DATA CENTERS HAS grown rapidly over the past several years. Artificial intelligence (AI) training and inference workloads now account for a large and increasing share. Industry and government forecasts identify data center load growth as one of the dominant drivers of new electricity demand in North America through 2030 (Aljbour et al., 2024). From an energy-use and management viewpoint, AI facilities differ from traditional hyperscale data centers designed

Nabindra Shrestha is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7001 USA. Ahmad Ali is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7001 USA. Iqbal Husain is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7001 USA. Srdjan Lukic is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7001 USA. Hantao Cui is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7001 USA.

Digital Object Identifier 10.1109/ESM.2026.3726869
Date of current version: 28 October 2026

2998-3991 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.

primarily for large-scale cloud computing and storage services. AI data centers often concentrate on large campuses, use multiple stages of power electronic converters to serve the IT load, and rely heavily on chillers and motor-driven pumps for cooling. Their computational workloads synchronize tightly across thousands of accelerators (computing devices) during distributed training and inference. Together, these characteristics place AI data centers in a category that is neither well represented by classical industrial load models nor fully captured by long-term demand growth forecasts.

Two framings of the AI data center coexist in the power systems literature. One views AI facilities as flexible grid-interactive assets capable of curtailment, ramping, and workload migration (Colangelo et al., 2025); (Norris et al., 2025). The other characterizes the same facilities as voltage- and frequency-sensitive disturbance sources, whose protection schemes and power electronic converters have already produced high-impact disconnection events (NERC, 2025); (Jiménez-Ruiz and Milano, 2026). The two framings are not physically contradictory, yet they have evolved within different research communities, draw on different incidents, and inform different regulatory processes.

We argue that asset and threat framing describe the same facility observed at different timescales and that a structural risk arises from the mismatch between the pace at which reliability standards and asset-side frameworks are developing. The contributions of this article are as follows. First, we propose a multi-timescale taxonomy that unifies the asset and threat literature by organizing AI data center behavior into bands from subcycle to seasonal. Second, a coordination matrix is developed that maps four actor groups (vendors, facility operators, utilities, and regulators) onto the timescale bands over which their decisions take effect. Third, a balanced empirical synthesis is presented showing that asset-side and threat-side evidence are broadly comparable in scale and rigor despite a marked asymmetry in institutional response. Finally, we outline a research agenda organized around the timescales at which the technical and regulatory challenges arise.

AI data centers bring together stakeholders ranging from chip and equipment manufacturers to facility designers, AI developers, operators, utilities, and end users. Their sustainability therefore depends not only on the efficiency of the individual technologies but also on how effectively these actors coordinate decisions across the facility and the power grid.

The remainder of this article is organized as follows. The "Internal Architecture of an AI Data Center" section describes the internal architecture of an AI data center as seen from the grid interface. The "A Multi-Timescale Framework for Dual Identity" section introduces the multi-timescale taxonomy. The "The Coordination Gap" section develops the coordination matrix with three case examples and a cybersecurity discussion. The "Empirical Evidence" section synthesizes empirical evidence. The "Institutions in Transition" section surveys institutions in transition. The "AI's Energy Sustainability" section outlines sustainability frameworks, pathways, and challenges. The "A Research Agenda" section proposes a research agenda. The "Conclusion" section concludes the article.

Internal Architecture of an AI Data Center

Server and GPU Power

The compute load of an AI data center is dominated by graphics processing units (GPUs) and

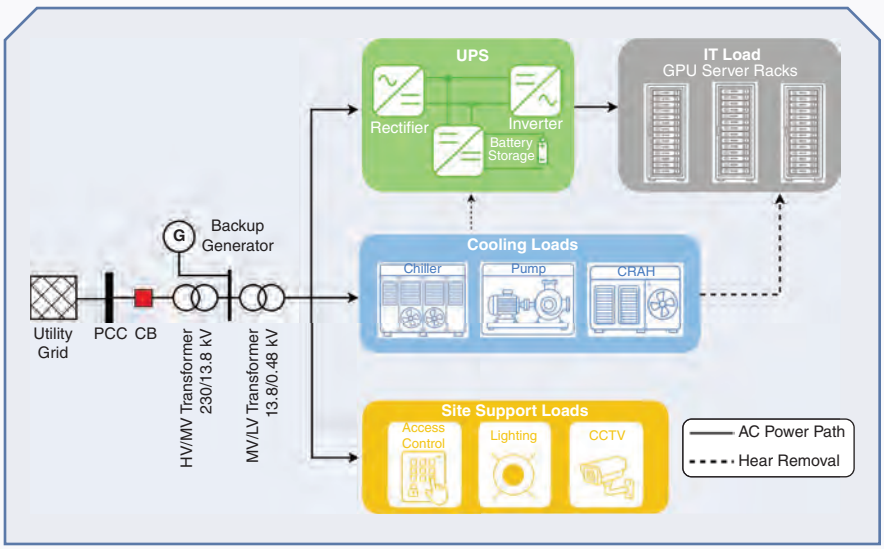


figure 1. A block diagram of an AI data center interfaced with the grid, showing the centralized double-conversion UPS bus, rack-level power distribution to GPU servers, and parallel cooling plant comprising chillers, pumps, and computer room air handlers (CRAHs). The discussion of this section is organized around four functional blocks: compute, cooling, UPS, and protection. CCTV: closed-circuit TV.

dedicated AI accelerators, such as Google's tensor processing units (TPUs), deployed in increasingly dense racks (Figure 1). Modern GPUs range from the NVIDIA H100, with a thermal design power (TDP) of about 700 W, to the B200 at around 1,200 W under liquid cooling. At the rack level, an integrated GB200 NVL72 system can draw approximately 120–140 kW, about an order of magnitude more than the traditional enterprise rack. Similarly, a single 64-chip rack based on Google's Ironwood TPU can draw roughly 90 kW.

Power is delivered to these accelerators through several conversion stages, including a medium-to-low-voltage transformer, a centralized uninterrupted power supply (UPS), rack-level rectifiers or power shelves, and on board point-of-load regulators. Each stage incorporates fast control loops that respond on timescales much faster than electromechanical dynamics. The aggregate compute load therefore behaves as a power electronic-based load, with low effective inertia and well-defined undervoltage thresholds. These thresholds are often determined by the original design manufacturer (ODM) of the server platform rather than by the facility operator. Data center power architectures are also evolving toward higher-voltage DC distribution and solid-state transformer (SST) interfaces, which could significantly change how internal dynamics of the data center load are exposed to the grid.

More recently, agentic AI is shifting the compute-power profile. Unlike conventional training and inference, which are largely GPU-dominated, agentic workloads rely more heavily on CPUs to orchestrate multistep execution loops, tool calls, application programming interface requests, and memory retrieval between GPU inference passes. ARM estimates that agentic AI could require approximately 120 million CPU cores per gigawatt of data center capacity, compared with about 30 million cores per gigawatt in current AI facilities. Such a trend introduces a broader mix of converter and voltage levels, making the overall load profile more heterogeneous.

Cooling Load

The cooling plant is typically the second-largest electric load in an AI data center, accounting for roughly 10–40% of the total site demand, depending on climate, water availability, and cooling architecture. Major loads include induction-motor-driven chillers, condenser and circulation pumps, and computer room air-handling equipment. These motors are directly coupled to the AC bus, drawing substantial reactive current during voltage sags, and may stall during disturbances, causing fault-induced delayed voltage recovery (FIDVR).

As liquid cooling is adopted for higher rack densities, a growing share of the motor load is supplied through variable-frequency drives, shifting part of the cooling load behavior toward the inverter-based class. At the same time, the cooling system has considerable thermal inertia, typically on the order of minutes to tens of minutes. This stored thermal capacity creates opportunities for short-duration demand flexibility through strategies such as precooling and temporary setpoint relaxation (Norris et al., 2025).

UPS

The UPS forms a critical interface between the utility supply and a data center's internal DC and AC distribution. It is also frequently involved in the facility response to grid disturbances. Centralized double-conversion UPS units in current AI facilities are commonly rated at 1–3.5 MW per module. Multiple modules operate in parallel to serve the aggregate IT load. Conventional data centers may also rely on on-site generators to provide longer-duration backup when utility service is lost. The role of the UPS system is also expanding beyond backup power toward active participation in grid-support services.

Voltage-Sensitive Protection and Ride Through

The protection settings of UPS units, server power supplies, and motor drives collectively determine the ride-through behavior of a facility. These settings are typically configured by equipment manufacturers and system integrators rather than coordinated with the interconnecting utility. Consequently, they may not align with the fault-ride-through requirements that apply to inverter-based generation under bulk power-system reliability standards, such as the NERC PRC-029-1 standard.

In some cases, equipment undervoltage thresholds may exceed the voltage at the load bus following a successfully cleared transmission fault, causing the facility to disconnect even when the bulk system remains stable. Such voltage-sensitive disconnections have been observed at the gigawatt scale (NERC, 2025); (Jiménez-Ruiz and Milano, 2026). The asset identity of the facility requires that this protection layer must be characterized, modeled, and coordinated with system-level requirements.

A Multi-Timescale Framework for Dual Identity

Subcycle to One Second

The subcycle- to-1-s band is dominated today by the threat identity (Figure 2). The main mechanisms

include 1) voltage-sensitive tripping of UPS modules, server power supplies, and motor drives following cleared transient faults; 2) FIDVR caused by motor stalling in the cooling plant; and 3) interactions of fast power electronic controls and weak grid conditions at the point of interconnection.

The 2024 Loudoun County, VA, USA, event is a prominent example of data center load loss related to protection and control behavior. A similar event occurred in Northern Virginia in February 2025. In ERCOT, NERC reported 25 crypto-mining load-loss events from 2023 to early 2025, each involving approximately 100–400 MW, while ERCOT later reported 26 large-electronic-load ride-through events since the beginning of 2023.

Importantly, the dominant behavior in this band is set by equipment settings established before the facility participates in any grid-flexibility program. Flexibility in this band is therefore limited. Where available, it comes primarily from co-located battery energy storage systems or the dynamic voltage support capability of the UPS rather than from the computational workload itself. A detailed modeling treatment of these transient phenomena, including the coupling between UPS protection and bulk-system stability, is presented in (Jiménez-Ruiz and Milano, 2026).

One Second to One Minute

The 1-s-to-1-min band is a crossover band in which both identities coexist. On the threat side, sustained low-frequency oscillations have been observed. Dominion Energy, for example, reported a sustained 14.7–14.8-Hz oscillation, which was traced to a UPS-system mode interacting with the surrounding network (NERC, 2025).

Synchronized GPU workloads introduce another concern. Mini-batch synchronization in large training jobs can produce oscillations in the approximately 0.2–3-Hz range, with cluster-level

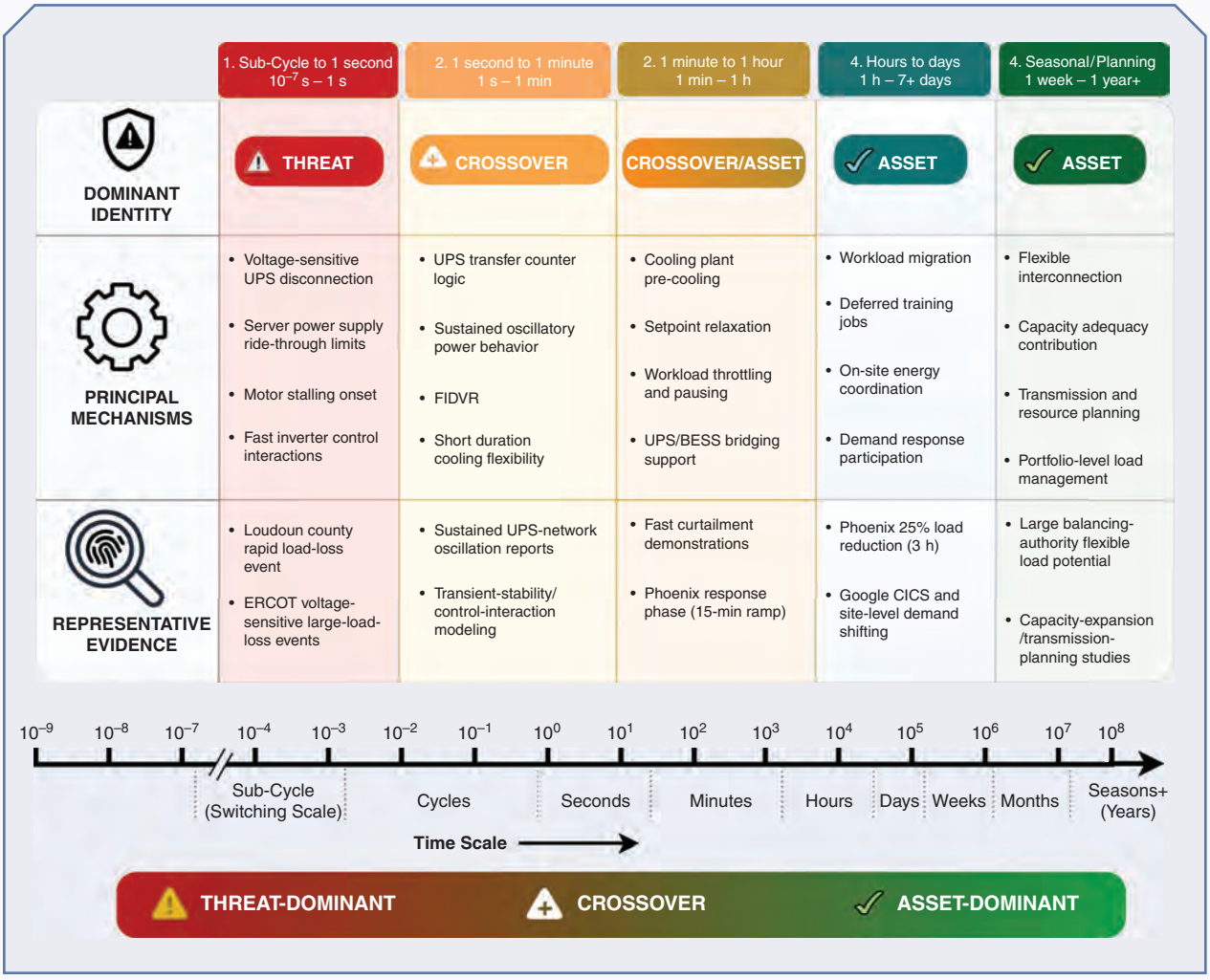


figure 2. The multi-timescale taxonomy of AI data center behavior from subcycle to seasonal timescales.

amplitudes reaching 50–60% of TDP. This range overlaps frequencies associated with electromechanical modes in large power systems. In (Jiménez-Ruiz and Milano, 2026), for example, the WECC-179 benchmark exhibits modes near 0.852, 1.06, and 1.44 Hz, all within the spectral range of synchronized AI workloads.

On the asset side, this band supports fast frequency response through modulation of compute power, dispatch of co-located storage, or both. The NVIDIA GB300 NVL72 platform incorporates power-smoothing capability that can reduce peak grid demand by up to 30% by buffering synchronization-induced oscillations within the rack (Colangelo et al., 2025). This crossover character of this band is particularly important; the same fast power modulation that can provide grid services may, when uncoordinated, interact with electromechanical modes if workload oscillations overlap their characteristic frequencies.

One Minute to One Hour

The 1-min-to 1-h band is dominated by the asset identity and encompasses most operational flexibility products. Key mechanisms include workload throttling at the orchestrator level, batch-job rescheduling, and dispatch of behind-the-meter storage.

The May 2025 Phoenix demonstration is reported in detail in (Colangelo et al., 2025). Using the Emerald AI platform, a 256-GPU NVIDIA A100 cluster sustained a 25% load reduction for 3 h during a Salt River Project peak event. Demand-response participation in this band is technically straightforward for inference and batch-training workloads; training jobs can be rescheduled, while inference queues can be throttled or delayed. On the other hand, interactive and latency-sensitive services, such as interactive chatbots, have tighter response-time constraints that limit their flexibility.

Threat behaviors are less prominent in this band but do not disappear. Cascading effects of upstream events, including repeated UPS transfers to battery, can persist over minutes and reduce the load served at the local substation. At the upper end of this band, the cooling plant also provides a useful source of flexibility because its thermal inertia allows temporary setpoint adjustments with limited impact on service. This band has therefore emerged as a particularly attractive target for early grid integration because it aligns well with established ancillary-service and demand-response programs.

Hours to Days

The hours-to-days band supports some of the largest aggregate flexibility estimates in the literature. Main mechanisms include shifting workloads away from daily peak periods, migrating jobs across geographically dispersed campuses, coordinating demand with co-located generation and storage, and participating in day-ahead energy and reserve markets.

The Google Carbon-Intelligent Computing System (CICS), for example, has demonstrated carbon-aware temporal and spatial workload shifting across more than 20 sites worldwide (Colangelo et al., 2025). Similarly, a Duke Nicholas Institute study estimates that the 22 largest U.S. balancing authorities can accommodate approximately 76 GW of new flexible load with 0.25% annual curtailment, increasing to 126 GW at 1% annual curtailment (Norris et al., 2025).

Threat behavior in this band is largely indirect. Large fleets of data centers that shift training-intensive workloads toward off-peak periods can reshape system net load and, in turn, affect renewable-generation and storage dispatch. Overall, the asset identity dominates because the relevant decisions are made through workload orchestration and scheduling processes that operate on timescales aligned with power-system operational planning.

Seasonal and Locational

The seasonal and locational band concerns siting, interconnection, and long-term planning rather than real-time operation. On the asset side, location-aware capacity planning can align new AI campuses with available transmission capacity and renewable generation, while flexible interconnection agreements can accelerate connection in exchange for curtailment rights (Norris et al., 2025). On the threat side, geographic concentration increases exposure to large-load-loss events and places greater planning demands on the host transmission system.

Flexible interconnection can reduce the transmission expansion required to host new large loads, provided that facilities can deliver reliable curtailment when needed (Norris et al., 2025). Regulatory frameworks are gradually moving in this direction. Texas Senate Bill 6, enacted in 2025, introduced curtailment and disclosure requirements for large new loads above specific thresholds (NERC, 2025). Whether such flexibility can be treated as a dependable long-term planning resource remains an open question.

In summary, the asset and threat identities are best understood as behaviors that merge at different timescales within the same facility, rather than as properties of distinct facility classes. Threat behavior dominates from the subcycle-to-seconds range, where equipment protection and settings are decisive. Asset behavior becomes more prominent from the minutes-to-seasonal timescales, where workload orchestration and planning decisions take effect. The crossover band contains both identities and is therefore where coordination among stakeholders becomes most consequential.

The Coordination Gap

GPU and Platform Vendors

GPU and platform vendors, including NVIDIA, AMD, and ODMs, exercise significant authority over the subcycle-to-1-s band through hardware and firmware defaults (Figure 3). Voltage and frequency ride-through thresholds, transfer-to-battery logic, and the spectral signature of computational workloads are shaped by design choices made before the facility is interconnected. Since key parameters governing grid-interface behavior are often absent from interconnection documentation and standard

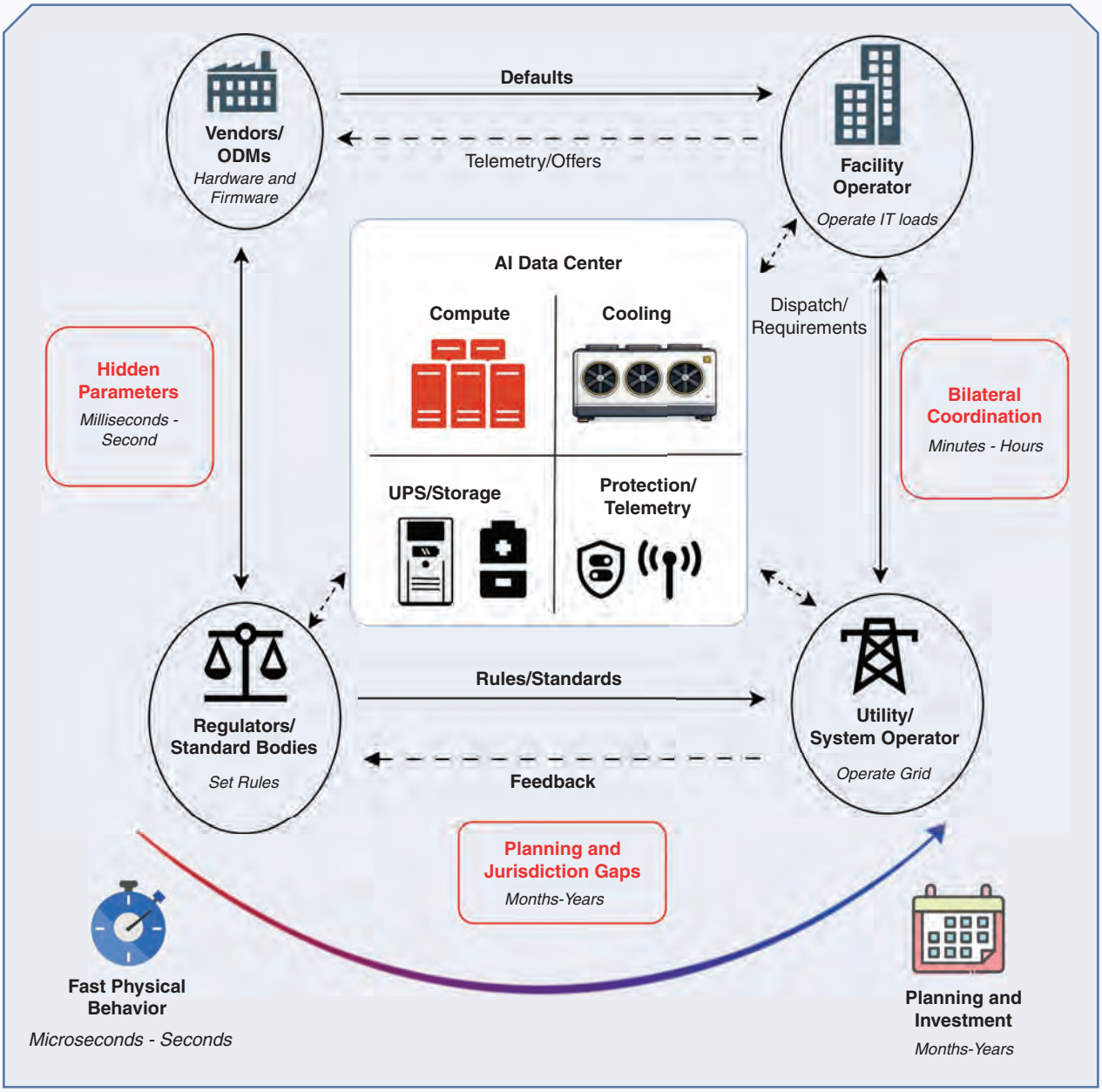


figure 3. A coordination matrix mapping stakeholders onto the timescales over which decisions take effect.

load-model libraries, closing this visibility gap will require more direct coordination between equipment vendors and the power-system community.

Facility Operators

Facility operators have authority from the 1-min to seasonal bands through orchestration software, infrastructure controls, and scheduling policies. The scope of this authority, however, depends on the data center business model. Hyperscale operators typically control both the IT workload and supporting infrastructure, giving them direct visibility into workload-level metrics and, in principle, the ability to expose aggregate load characteristics to the power system. Co-location operators, by contrast, manage the shared infrastructure while hosting workloads controlled by multiple tenants. This creates a split-authority in which the facility operator controls power and cooling systems but not necessarily the computational workloads that drive demand.

To date, participation in flexibility programs has been limited to a small number of hyperscale demonstrations (Colangelo et al., 2025); (Norris et al., 2025). Co-location facilities may have fewer options beyond conventional demand-response programs because workload control ultimately resides with individual tenants. Another issue is economic incentive; the value of maintaining compute availability can exceed the compensation offered for demand response.

Utilities and System Operators

Utilities and system operators, including regional transmission operators and independent system operators, exercise authority from the 1-s to seasonal bands through interconnection studies, ancillary-service market design, and operating procedures. The interconnection process is a key point of leverage because protection settings, ride-through behavior, and other grid-interface requirements can, in principle, be specified before the facility enters service.

Large-load interconnection studies have traditionally focused on steady-state thermal and voltage limits, but some system operators are now extending their requirements to dynamic studies. ERCOT requires dynamic stability analysis for large loads under its 2022 interim interconnection process, while PJM requires positive-sequence transient-stability models and, at later stages, electromagnetic transient models. However, these requirements remain inconsistent across regions.

Regulators and Standards Bodies

Regulators, including FERC, NERC, and state public utility commissions, together with standards organizations such as the IEEE Standards Association, shape the seasonal band and indirectly influence shorter timescales through rulemakings and standard development. Industry research organizations, including EPRI and CIGRE, support these processes through technical studies and best-practice guidance but do not impose enforceable requirements.

The institutional responses to the two identities have followed different paths. Threat-related concerns have increasingly been addressed through enforceable channels, including NERC alerts, FERC actions, and state legislation. Asset-side opportunities, by contrast, have largely advanced through voluntary collaborations, working groups, and demonstration programs. As discussed in the "Institutions in Transition" section, this asymmetry affects how quickly reliability requirements and flexibility frameworks are translated into facility design and operating practice.

Three Case Examples

The coordination matrix can be illustrated through three cases spanning different timescales. The 10 July 2024 Loudoun County event exposed a gap between equipment-level settings and system-level visibility. Approximately 1,500 MW of data center load disconnected following a cleared 230-kV transmission fault and reclosing sequence. Much of the sustained load reduction was linked to customer-side protection and control logic. The event shows how locally reasonable protection settings can become systemically significant when replicated across geographically concentrated facilities.

A different form of coordination is illustrated by Google's demand-response agreements with five U.S. utilities, under which up to 1 GW of data center demand can be curtailed during periods of grid stress. Here, flexibility is provided through workload orchestration rather than emergency disconnection. However, these arrangements remain largely bilateral, with no common framework for defining firm flexibility, response time, baselines, or performance verification across different facilities and workloads.

At the planning timescale, Amazon's data center campus co-located with Talen Energy's Susquehanna nuclear plant exposed a regulatory gap. A proposal to increase the load from 300 MW to 480 MW raised concerns over reliability,

transmission service, and cost allocation, ultimately requiring FERC intervention and a broader framework for large-load co-location in PJM.

Together, the cases illustrate three distinct forms of coordination gap: hidden technical settings, bilateral flexibility without common standards, and new commercial arrangements that exceed existing interconnection rules. The coordination gap is therefore not a single missing interface but a set of coordination failures that emerge differently across timescales.

Cybersecurity as Cross-Cutting Coordination Layer

The coordination failures described previously involve information or control crossing organizational boundaries: vendor defaults influencing system-level behavior, facility telemetry flowing to utilities, and new commercial arrangements interacting with interconnection processes. Cybersecurity governs whether these interfaces can be trusted. Currently, NERC CIP standards apply to bulk electric system cyber assets, while data centers generally fall outside their scope as loads.

This gap spans all four actor categories. Vendors may retain remote access to UPS and cooling control systems for diagnostics and firmware updates, creating potential pathways from equipment-level settings to system-level consequences. Facility operators participating in demand response must exchange telemetry and control signals with utilities, yet these interfaces are typically governed through bilateral agreements rather than a common security standard. Utilities and system operators likewise lack a standardized framework for securely dispatching and verifying large-load flexibility. The January 2026 NERC CIP Roadmap acknowledges large data centers and original equipment manufacturer remote-access pathways as an emerging regulatory concern.

Although baseline cybersecurity requirements for large loads are under development, enforceable standards remain limited. The same telemetry channels that enable data centers to act as grid assets can therefore also create pathways for erroneous or malicious commands to affect large amounts of load. A credible grid-interactive role for AI data centers will require a security framework that spans vendors, operators, utilities, and regulators.

Empirical Evidence

The empirical record has expanded rapidly, but the evidence is not uniform in scale or form (Figure 4). On the asset side, recent work now

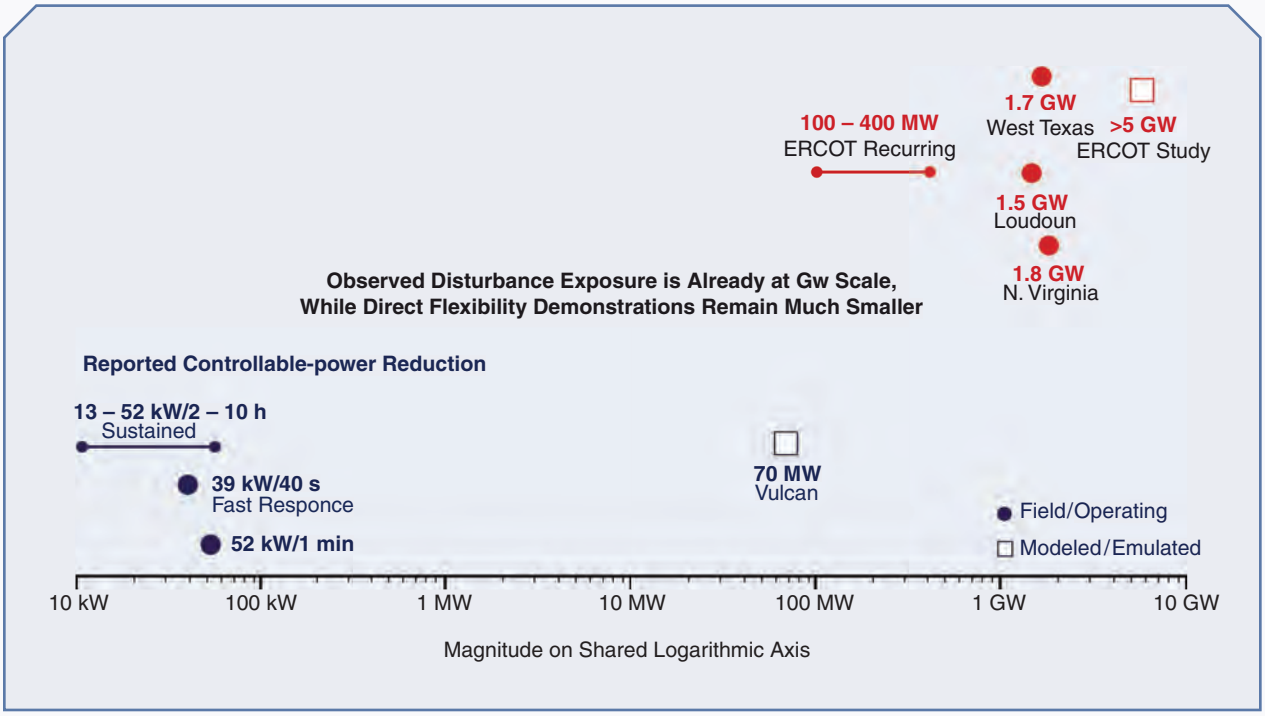


figure 4. Reported disconnected-load events and studies compared with demonstrated or emulated controllable-power reductions. Operating disturbances have reached gigawatt scale, while directly demonstrated flexibility remains substantially smaller. N. Virginia: Northern Virginia.

includes field deployments on operational GPU clusters, hardware tests with measured response times, and system-level simulations using measured AI workload traces. On the threat side, the strongest evidence continues to come from disturbance events and utility studies of large-load ride-through. The relevant question is therefore not whether both identities exist but what has been demonstrated, at what scale, and under what degree of experimental control.

Asset-Side Demonstrations

The asset-side record has moved beyond isolated curtailment demonstrations. A 2026 study reports deployment on a 130-kW NVIDIA Blackwell Ultra cluster that successfully responded to more than 200 National Grid power events. The same study reports the migration of 10% of live inference traffic from an H100 cluster in Virginia to Illinois during a Dominion Energy winter peak event. Although modest in facility scale, these demonstrations provide strong empirical evidence because the control actions were applied to operating GPU infrastructure in response to real grid conditions.

Evidence is also emerging at faster timescales. GridPilot, evaluated on a three-GPU NVIDIA V100 testbed, achieved an end-to-end trigger-to-target response of 97.2 ms. The controller also accounts for instantaneous power usage effectiveness, allowing the delivered response to be defined at the facility meter rather than only at the IT load. While this result does not establish megawatt-scale deliverability, it shows that the software and hardware control path need not be the limiting factor for subsecond response.

Taken together, the evidence shows that AI workloads can be modulated on real hardware without necessarily violating service requirements, with responses ranging from subsecond control to minute-scale curtailment and geographic workload migration. The evidence does not yet justify treating an arbitrary data center as a firm grid resource with a fixed fraction of nameplate capacity available on demand.

Threat-Side Incidents

The threat-side record is stronger at system scale because it is grounded in operating events and utility studies. The Northern Virginia and ERCOT incidents discussed earlier show that geographically concentrated electronic loads can reduce consumption coherently following voltage disturbances. More recent studies suggest that the potential exposure is larger than the events already observed to date.

In a May 2026 assessment of approximately 20 GW of large loads, ERCOT identified four groups for which normally cleared transmission faults produced modeled aggregate tripping above 5 GW, with estimates ranging from about 5.4–6.0 GW. These were not observed in the 5-GW disconnection event; the assessment modeled loads at approved capacity, and none of the four groups was then operating above 3.2 GW. Even so, the results show how correlated ride-through behavior can create an aggregate exposure far larger than any individual facility.

The central empirical threat is therefore not simply that an individual facility may fail to ride through a disturbance but that many facilities may respond in a correlated manner. Similar protection and ride-through behavior, combined with geographic concentration, can transform a local voltage event into a bulk-system power imbalance. The magnitude of that imbalance depends not only on facility size but also on the correlation of protection, transfer, and reconnection behavior.

Empirical Synthesis

Both evidence records are substantial, but they are not symmetric. Threat-side evidence already reaches the gigawatt scale in real operating systems, whereas asset-side evidence still comes largely from demonstrations, contracts, and system-level studies rather than repeated dispatch of gigawatt-scale AI facilities.

This asymmetry does not weaken the asset framing; it defines the present burden of proof. Threat behavior becomes visible to system operators when failures create an event record. Asset behavior, by contrast, must be intentionally instrumented, dispatched, measured, and verified before it can be treated as an operational resource. The asymmetry is therefore both evidentiary and institutional; one side is revealed by disturbance, while the other must be established by coordination.

The next empirical question is no longer whether AI data centers can provide flexibility but whether that flexibility can be characterized before interconnection, maintained across hardware and workload generations, exposed through secure telemetry, and delivered under the same contingencies for which ride-through is required. A facility that can curtail 100 MW during a scheduled peak event but disconnects 500 MW during a normally cleared fault is not yet a

grid asset. The transition from load to asset requires both flexibility and ride-through capabilities to be demonstrated within the same facility.

Institutions in Transition

The institutional response no longer divides cleanly into a threat-side regulatory track and an asset-side voluntary track, although reliability concerns moved first (Figure 5). NERC's Large Loads Task Force issued its white paper in July 2025 (NERC, 2025), followed by industry alerts culminating in a Level 3 Essential Action in May 2026. This shifted the discussion from risk characterization toward mitigation planning. But alerts remain temporary interventions rather than permanent performance standards. Now, the central question is how ride-through, dynamic modeling, protection settings, and operating visibility should be incorporated into routine interconnection practice.

Binding action has also expanded beyond disturbance response. In December 2025, FERC directed PJM to establish rates and conditions for large loads co-located with generation. In June 2026, FERC issued show-cause orders to six regional grid operators addressing large-load interconnection procedures, cost allocation, self-supply arrangements, and related planning practices. Texas Senate Bill 6 provides a state-level counterpart through interconnection, curtailment, and demand-reduction requirements for large new loads. Together, these actions signal that large computational loads can no longer be treated only as passive retail customers.

On the asset side, initiatives such as EPRI's DCFlex program and bilateral utility agreements are testing workload orchestration, storage dispatch, cooling control, and other forms of demand flexibility. Demonstrating flexibility, however, is not the same as defining a firm grid resource. That transition requires common performance metrics for available load reduction, response time, duration, trajectory, measurement location, telemetry, cybersecurity, and behavior during grid disturbances.

The remaining institutional gap is therefore a translational problem between reliability requirements and flexibility programs. Until dispatchable flexibility and ride-through behavior are verified together, planners will continue to treat data center demand as firm when sizing infrastructure and its disconnection as credible when evaluating contingencies. The same load will therefore be modeled conservatively in both directions.

AI's Energy Sustainability

A data center's asset-side identity must also be backed by its long-term commitment to sustainability. This section reviews current sustainability practices at large data centers, some guiding frameworks, and underlying challenges.

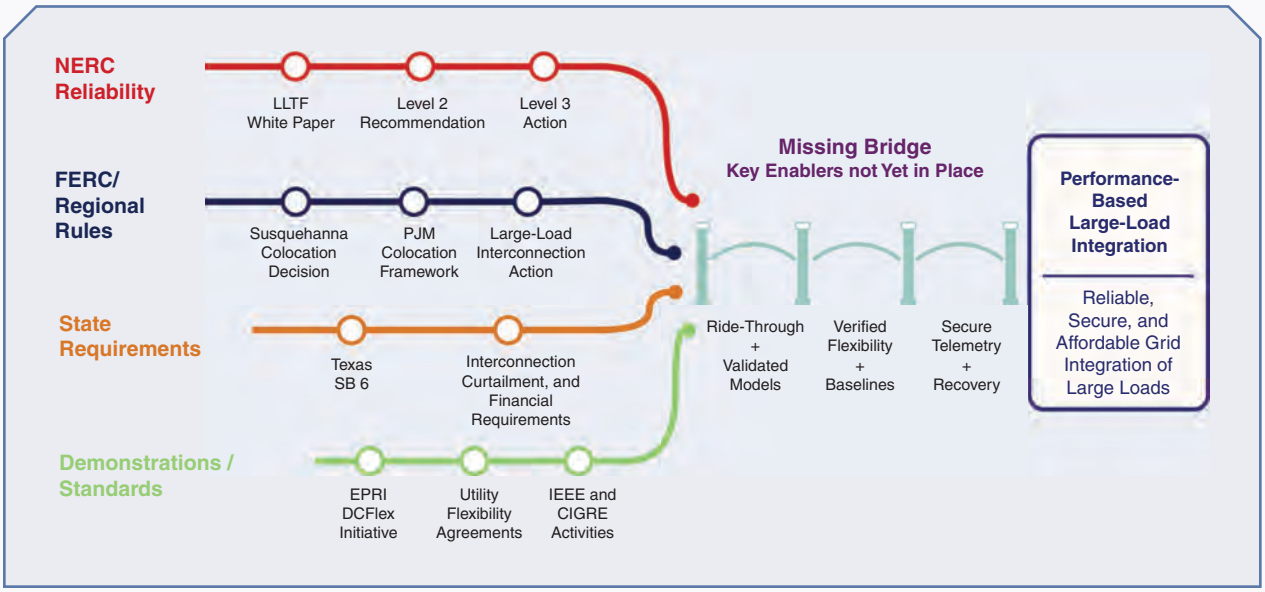


figure 5. Institutional pathways toward performance-based large-load integration. Institutions are moving toward a common framework, but the translation to enforceable grid performance remains incomplete. Reliability, regulatory, state, and voluntary initiatives are advancing, but additional steps remain. LLTF: large loads task force; SB: senate bill.

Frameworks and Accountability

Although some AI hyperscalers voluntarily adopt standards for operational efficiency and energy management, no single global framework governs their sustainability. They instead work within a mix of standards and regional disclosure rules, the main ones being the following:

- *ISO 50001 and ISO/IEC 30134:* These are standards for improving energy management and for efficiency indicators such as power usage effectiveness (PUE) and water usage effectiveness.
- *Carbon accounting*: This refers to measuring and tracking greenhouse gas emissions arising either directly from a data center's operations or indirectly from its broader value chain.
- *EU Energy Efficiency Directive, Article 12 (mandatory):* This requires European Union (EU) member states to make data centers above 500 kW of installed IT power report their energy, water, and renewable-use performance.

None of these standards, however, impose an absolute emission ceiling or indicate whether connecting a particular AI campus improves or degrades the regional grid. Closing this gap requires a framework that links facility metrics such as PUE to grid-level emissions and reliability.

The Dual Reality

The major U.S. hyperscalers, including Amazon, Google, Microsoft, and Meta, are all committed to various sustainability goals and are working to meet them. However, the rapid expansion of data centers is making those goals harder to reach. The evidence points to a dual reality: real commitment and effort on one side, real struggle on the other.

- Google-owned European data centers are certified under ISO 50001:2018. Google reported a 12% cut in data center energy emissions in its 2025 reporting cycle and has also set a tougher target: 24/7 carbon-free energy (CFE) in every grid region by 2030. However, the company's 2024 Environmental Report puts its total emissions at 48% above the 2019 level, reflecting both data center electricity use and AI-driven growth in its supply chain.
- Amazon Web Services (AWS) reported a global PUE of 1.14 for 2025, and Amazon says its company-wide annual electricity consumption was matched with renewable energy for a third consecutive year, alongside a modest emissions decline. That matching is measured annually, using unbundled certificates rather than hourly CFE. An annual measure therefore does not indicate how much of each hour's demand was served by carbon-free power. Its own 2025 report shows that purchased electricity emissions rose as data center consumption expanded.
- Meta states that 100% of the electricity used by its owned operations is matched with clean and renewable energy and that every owned data center is certified LEED Gold or higher. It also reports funding grid infrastructure directly and adding new generation to support its data center expansion. Nevertheless, it identifies data centers as its largest operational source of energy and water use.
- Microsoft pursues LEED Gold certification for newly built data centers, measures PUE and water usage effectiveness, and has goals for carbon-negative electricity operations. Yet its 2024 Environmental Sustainability Report puts overall emissions at 29.1% above the 2020 baseline, "primarily due to the construction of more data centers."

Across all four companies, efficiency and certification keep improving, yet absolute emissions keep climbing with AI capacity. Their largest carbon-free investments, advanced nuclear in particular, are expected to come online between 2029 and 2035, while demand is expanding now. Gas generation and annual certificate matching are likely to cover much of that gap.

Further, long interconnection delays can also increase emissions. If grid service is not available when a data center is ready to operate, the facility may use on-site gas turbines or diesel generators as temporary power. In this case, the delay does not eliminate the electricity demand. It shifts part of the demand to more carbon-intensive on-site generation.

Sustainability Pathways and Challenges

Several measures can cut the carbon footprint of AI data centers and make their operation more sustainable. This section reviews the main ones and the challenges each brings.

1) *Carbon-aware computing:* A data center can reduce its carbon footprint by shifting its flexible AI workloads to times when renewable output is high or to places where the grid currently relies more on renewable

resources. The challenge is that only latency-tolerant workloads can be delayed or relocated. Data centers, usually built close to users, may not have easy access to renewable resources and may draw high power during periods of low renewable generation. Installing battery storage to cover peak demand can help close this gap.

2) *Workload distribution:* Spreading workloads across more data centers also eases the load on any single facility and slows thermal and mechanical wear on components, which extends component life and cuts waste over time. The challenge is that objectives 1 and 2 may not be achievable at once; the case for favoring sites with more renewables can outweigh the case for extending component life.
3) *Sustainable thermal management:* Cooling should be treated as a site-specific design problem rather than preferring one technology. Closed-loop mechanical or dry cooling may be appropriate for regions with limited water resources, while dry climates with adequate water supply can use direct or indirect adiabatic cooling. Moreover, hybrid systems that switch between dry, adiabatic, and mechanical cooling according to ambient conditions can also be developed. Using direct-to-chip water cooling can capture heat close to the processors, reducing dependence on energy-intensive room-air cooling. That heat can then be reused in facility space heating and hot-water production. Further, AI workload scheduling can be coordinated with outdoor temperature, renewable-energy availability, and water conditions.
4) *Improved energy efficiency:* Reducing power losses inside the data center raises electrical efficiency and lowers energy demand. Modern AI server racks are transitioning to 800-V DC architecture or higher. This higher-voltage power distribution reduces conduction losses within the racks, resulting in lower cooling energy requirements.
5) *Transparency and active communication:* Sharing workload schedules with power utilities lets both sides agree on moving flexible workload execution to periods of high renewable energy generation.

The greatest challenge is the sheer pace of growth. Data centers are being built at a rate few anticipated, driven by competition for technological leadership and future economic gains. In contrast, regulators, standards bodies, and the broader energy sector are advancing at their traditional pace. Faster regulatory action and more proactive intervention are needed to enforce sustainability practices rather than relying on voluntary adoption.

A Research Agenda

A research agenda organized by timescale band follows from the framework of the "A Multi-Timescale Framework for Dual Identity" section and the coordination gap of the "The Coordination Gap" section. Five open problems are identified.

1) *Subcycle-to-1-s hardware-software co-design:* Research at the fastest timescales requires progress on both modeling and power-delivery architecture. On the modeling side, public-domain transient-stability models are needed to capture UPS protection logic, server power-supply ride-through, cooling plant motor dynamics, and their interactions with the grid. High-fidelity transient analysis tools such as open source tools can support this effort (Ali et al., 2025). In parallel, hardware and control research should explore coordinated UPS systems, SST front ends, higher-voltage DC distribution, and embedded storage. These architectures could consolidate multiple converter and protection functions behind fewer coordinated grid-facing interfaces, providing defined points for ride-through, current limiting, damping, and fast grid support. The two research directions should inform the design and settings of these interfaces.
2) *One-minute to one-hour operating products:* Credible demand-response and ancillary-service products for AI data center participation require standardized telemetry, baseline methodologies that account for workload heterogeneity, and verification procedures that operate under the cyberphysical constraints discussed earlier. Cooling plant flexibility deserves particular attention because of its substantial untapped potential.
3) *Hours-to-days and seasonal planning tools*: Capacity-expansion and production-cost models that recognize flexible interconnection without suppressing threat behaviors are required for resource-adequacy

and transmission-planning processes. The estimates in (Norris et al., 2025) establish the upper bound on aggregate potential. The modeling task is to translate that bound into operationally credible planning inputs that account for protection and orchestration realities.

4) *Multidisciplinary research and joint field trials:* Lastly, sustainability requires multidisciplinary research and joint field trials. Electrical engineers should pursue more efficient power-delivery architectures; mechanical and civil engineers should design data center buildings with improved thermodynamic properties such as natural cooling alongside low-water or water-free cooling systems; and AI researchers should develop more advanced carbon-aware scheduling algorithms. Their combined impact should be measured through energy use, water use, hourly emissions, and grid performance.

Conclusion

The multi-timescale framework developed in this article supports two conclusions. First, the threat and asset identities are not competing descriptions of different facility classes. They are behaviors of the same facility at different timescales. Subcycle and second-scale behavior is determined primarily by power electronics, UPS controls, motor dynamics, and protection settings. Minute-to-seasonal behavior is determined primarily by workload orchestration, cooling management, storage dispatch, and planning decisions. Second, the institutional treatment of AI data centers should therefore be performance-based rather than category-based. A facility cannot be adequately represented only as a voltage-sensitive load or only as a flexible computational resource.

The practical object of interconnection is the coupled stack of compute hardware, cooling equipment, UPS systems, storage, protection, orchestration software, and telemetry. A credible grid-interactive facility must demonstrate both sides of its behavior. It must remain connected within a defined disturbance envelope, and it must deliver promised flexibility with measurable activation time, magnitude, duration, and recovery. Flexibility should not offset reliability margins merely because it is technically possible. It should receive operational or planning credit only when it is observable, controllable, secure, and contractually deliverable.

The transition from load to asset is therefore a verification problem as much as a technology problem. Physical models must connect the fast and slow timescales. Interconnection studies must expose protection and control settings that are presently hidden from system operators. Secure telemetry must make flexible operation observable. Operating products must translate technical capability into enforceable performance obligations. Planning tools must assign value only after those obligations are credible. An AI data center becomes a grid asset not when it can occasionally reduce demand but when its behavior across the full timescale range can be predicted, measured, and coordinated by the actors responsible for grid reliability.

Acknowledgment

The authors would like to thank Andrew Meyers and George Navarro at M.C. Dean Inc. for helpful suggestions and discussions related to this work.

For Further Reading

"Characteristics and risks of emerging large loads," North Amer. Electric Rel. Corporation, Washington, DC, USA, NERC Large Loads Task Force White Paper, Jul. 2025. [Online]. Available: https://www.nerc.com/globalassets/who-we-are/standing-committees/rstc/whitepaper-characteristics-and-risks-of-emerging-large-loads.pdf

P. Colangelo et al., "AI data centres as grid-interactive assets," *Nature Energy*, vol. 11, no. 2, pp. 254–261, 2025, doi: 10.1038/s41560-025-01927-1.

T. H. Norris, T. Profeta, D. Patiño-Echeverri, and A. Cowie-Haskell, "Rethinking load growth: Assessing the potential for integration of large flexible loads in US power systems," Duke Nicholas Inst. for Energy, Environ. & Sustain., Durham, NC, USA, Feb. 2025. [Online]. Available: https://nicholasinstitute.duke.edu/publications/rethinking-load-growth

J. Aljbour, T. Wilson, and P. Patel, "Powering intelligence: Analyzing artificial intelligence and data center energy consumption," Electric Power Res. Inst., Palo Alto, CA, USA, 2024. [Online]. Available: https://www.epri.com/research/products/3002028905

A. Jiménez-Ruiz and F. Milano, "Data centre model for transient stability analysis of power systems," *IET Gener. Transm. Distrib.*, vol. 20, no. 1, 2026, Art. no. e70381, doi: 10.1049/gtd2.70381.

A. Ali, J. Yin, F. F. Li, and H. Cui, "Pathway toward an open source ecosystem in power systems: A blueprint for collaborative innovation and software sustainability," *IEEE Electrific. Mag.*, vol. 13, no. 2, pp. 58–67, Jun. 2025, doi: 10.1109/MELE.2025.3558711