

Large-Scale Data Center Integration in Power Systems: Uncertainty and Multi-Scale Variation

Le Xie, *Fellow, IEEE*, Dongjoo Kim, *Student Member, IEEE*, Qian Zhang, *Student Member, IEEE*, Grant Ruan, *Member, IEEE*, and Paul L. Joskow

Abstract—Hyperscale artificial intelligence (AI) data centers are emerging as a structurally new class of electricity demand. Unlike conventional loads, they materialize as spatially concentrated, transmission-level point loads with rapid deployment timelines and converter-dominated dynamics. Their integration is governed by two defining properties: forecast uncertainty in load realization and multi-scale temporal variation spanning sub-seconds to planning horizons.

We develop a unified framework that links reliability and affordability to these dimensions through a four-quadrant structure: (reliability, affordability) $\times$ (uncertainty, variation), which integrates probabilistic resource adequacy, generation and transmission expansion, operational security, and market design.

Measured traces from a public generative-AI workload dataset show strongly *tonal* GPU power spectra rather than smooth broadband behavior: harmonic combs at the training-step cadence, and a distinct sub-second periodicity in inference. Across these traces the dominant variance band shifts with workload type—seconds for training and offline inference, tens of seconds and longer for online inference. We report this as a characterization of the measured workloads rather than a universal signature, and take it to indicate that such loads warrant a spectral description rather than a peak-demand one. These signatures directly affect reserve requirements, ramping products, voltage stability, ride-through performance, and harmonic interactions in converter-rich grids.

Taken together, the analytical framework and empirical results establish a system-level basis for assessing the reliability and cost implications of large AI data center integration across planning and operational horizons.

Index Terms—AI data centers, large flexible loads, power system reliability, resource adequacy, operational security, power quality, market design, uncertainty.

I. INTRODUCTION

Electricity demand from computing is projected to grow several-fold over the next decade, driven by the rapid expansion of artificial intelligence (AI) training and inference workloads and periodic surges and geographic shifts in cryptocurrency mining. However, these forecasts remain uncertain

and reflect variability in AI adoption trajectories, improvements in hardware efficiency, and the cyclical nature of cryptocurrency markets. This growth represents a fundamental shift in electricity consumption, but it is not uniform across the data-center fleet. In practice, there are thousands of data centers, and the majority are small-to-medium facilities that typically interconnect at distribution grids and exhibit load profiles similar to other commercial customers. The bulk power system planning and operation challenges, however, are driven disproportionately by hyperscale data centers: spatially lumpy, megawatt-to-gigawatt-class campuses that often pursue transmission-level (or sub-transmission) interconnection and develop substantial “inside-the-fence” infrastructure, including on-site distribution, short-duration storage/uninterruptible power supply (UPS) systems, and backup generation. The magnitude and pace of this transition are already evident: major forecasting organizations project U.S. data-center electricity consumption to reach hundreds of terawatt-hours by 2030, while individual hyperscale AI facilities can approach gigawatt-scale demand, comparable to the capacity of a large nuclear unit or the peak load of a mid-sized city.

This transition from cryptocurrency mining to AI-driven computing has amplified both the scale and the complexity of grid integration challenges. While crypto mining facilities demonstrated that computing workloads could materialize rapidly as large point loads with minimal geographic diversification, AI data centers introduce additional dimensions of uncertainty and variation that differ in terms of their economic drivers, operational flexibility, and long-horizon predictability. Compared with the rule-based predictability of crypto mining, AI demand growth is driven by factors largely exogenous to traditional load forecasting models [1].

The rapid construction of large-scale AI data centers is transforming electricity demand in ways that challenge conventional power system planning and operation. These facilities can arrive as concentrated, high-power point loads on short timelines (often within 12–24 months), while generation and transmission additions typically require longer lead times of 3–7 years. This temporal mismatch creates new stress for both grid reliability and customer affordability. Moreover, AI data centers exhibit pronounced spatial concentration, which amplifies the localized impact of forecast errors and operational variability. These spatial features, combined with the temporal uncertainty of project duration and the multi-scale variation of operational profiles, create a fundamentally different risk landscape from that of traditional industrial/commercial loads.

In this paper, we interpret reliability broadly as encompassing three interrelated dimensions: i) resource adequacy:

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Le Xie, Qian Zhang are with the Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Allston, MA, USA.

Dongjoo Kim is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA.

Grant Ruan is with the Laboratory for Information & Decision Systems (LIDS), Massachusetts Institute of Technology, Cambridge, MA, USA.

Paul L. Joskow is with the Department of Economics, Massachusetts Institute of Technology, Cambridge, MA, USA.

Corresponding author: Le Xie, (e-mail: xie@seas.harvard.edu).

planning reserve and energy adequacy, ii) operational security: deliverability and contingency-secure operation, and iii) stability: voltage and frequency stability, and oscillations. We argue that two properties dominate the landscape of AI data center integration: uncertainty (forecast ambiguity in size, timing, location, and firmness) and variation (multi-scale temporal dynamics from sub-second to minutes and beyond). These properties manifest across different time horizons and interact with both reliability and affordability concerns, which motivate a structured analytical framework that can systematically address the challenges and opportunities of large-scale data center integration.

Recent studies have explored the grid integration issues of data centers, specifically focusing on demand response potential and ancillary service provision [2], power quality and harmonic distortion [3], and planning methods for accommodating large loads [4]. However, existing efforts address only one of the uncertainty and variation aspects, without explicitly connecting how these properties interact across time scales and propagate into both reliability and affordability outcomes. Moreover, most works treat data centers as homogeneous entities, without distinguishing how AI training workloads (synchronized, step-like dynamics) differ fundamentally from inference workloads driven by bursty user requests, or how these operational signatures translate into practical metrics and planning requirements.

The separation between engineering and economic studies is another notable gap. Engineering studies of dynamic performance, power quality, and stability often do not address cost allocation or tariff design questions, while economic and policy analyses of cost recovery and rate design seldom capture the load variability that determines infrastructure upgrades. This disciplinary division can lead to suboptimal outcomes: tariffs that fail to align incentives with operational needs, or infrastructure investments that deviate from the least-cost portfolio. There are clear concerns regarding the treatment of uncertainty: While stochastic optimization is popular in generation expansion planning, its application to demand-side uncertainty remains largely limited and cannot capture the discrete, spatially concentrated demand with uncertain timing and scale.

This paper bridges these gaps through three main contributions. First, we develop a unified analytical framework that explicitly connects uncertainty (forecasting challenges in size, timing, location, and firmness) and variation (multi-scale temporal dynamics from sub-seconds to years) to both reliability and affordability outcomes. This framework is structured as a four-quadrant matrix (see Table I): (reliability, affordability) $\times$ (uncertainty, variation), which systematically organizes the existing challenges, evidence, and mitigation strategies across these dimensions.

Second, we provide an empirical characterization of AI data center variability across multiple time scales, including spectral analysis of measured workload traces that reveals strongly *tonal* signatures—harmonic combs at the training-step cadence and a distinct inference periodicity—and a dominant variance band that shifts with workload type. We present these as properties of the measured traces rather than as universal

characteristics of AI workloads. This finding benefits both forecasting methodologies and reserve sizing. We connect these temporal signatures to practical metrics and operational requirements, establishing quantitative links between workload characteristics (training versus inference) and power-system performance.

Third, we synthesize evidence from multiple domains (planning models, operational experiences, tariff design practices, and dynamic performance requirements) to identify actionable techno-economic policy guidance. Rather than isolating the engineering, economic, and policy discussions, we explicitly examine how tariff and contracting choices affect which infrastructure portfolios are built, how uncertainty in load forecasts propagates into adequacy risks and cost allocation decisions, and how operational flexibility mechanisms can be designed to align incentives with system needs.

The contribution of this paper is integrative rather than technology-specific. Detailed dynamic simulation is indispensable for evaluating a particular data-center topology, control strategy, and grid condition, and a growing body of work—including several of our own prior studies—already addresses those questions in depth: load characterization and grid-impact reviews [5], electromagnetic-transient models for ride-through assessment [6], large-signal data-driven impedance models [7], Lyapunov-certified adaptive dynamics models [8], and ISO-scale flexibility-aware siting studies [9]. A single additional case study would represent one system under one operating condition and could not validate a framework that spans microsecond dynamics to multi-decade planning; we therefore ground the framework in this quantitative literature and in measured workload data.

The rest of this paper is organized as follows:

Section II establishes the foundational concepts of uncertainty and multi-scale variation by characterizing how AI data centers differ from conventional loads, introducing spatial-concentration metrics that contextualize why these issues matter for grid planning, and validating the multi-scale structure against measured generative-AI workload traces. Section III addresses reliability under uncertainty by examining how forecasting challenges propagate into resource adequacy risks, with separate treatment of systems with and without capacity markets. Section IV examines affordability under uncertainty, reviewing how tariff design and cost allocation mechanisms can discipline speculative load and limit cross-subsidies. Section V addresses operational security and stability under variation, covering both sub-second dynamic performance and seconds-to-minutes ramping requirements. Section VI examines the incremental infrastructure and operational capabilities needed to accommodate fast variation and connects these engineering requirements to cost assessment and allocation. Section VII outlines a techno-economic policy framework that integrates these dimensions and proposes directions for aligned incentives, commitments, and physical constraints. Finally, Section VIII concludes this paper with a concise summary.

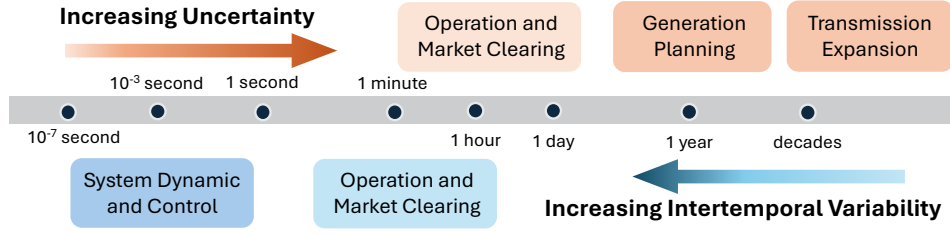


Fig. 1. Timescale map linking the two defining properties of AI data-center demand to power-system functions. The upper orange progression denotes increasing forecast uncertainty as the horizon extends from operations and market clearing to generation planning and transmission expansion. The lower blue progression denotes increasing inter-temporal variation toward the shorter horizons governed by system dynamics and control and by operations and market clearing. Together, the two progressions motivate the reliability–affordability framework summarized in Table I.

TABLE I
THE PROPOSED FOUR-QUADRANT LOGIC: (RELIABILITY, AFFORDABILITY) × (UNCERTAINTY, VARIATION)

	Uncertainty	Variation
Reliability (Resource Adequacy, Operational Security, Dynamic Stability)	Resource adequacy risk: planning reserve errors, procurement mismatch; Deliverability/siting risk: clustered load vs. network lead times; Mitigation: milestone-based interconnection, probabilistic scenarios. <i>See Sec. III</i>	Operational security: ramps, reserves, congestion; Stability: ride-through, oscillations, protection interactions; Mitigation: demand response, UPS/storage participation, dynamic modeling. <i>See Sec. V</i>
Affordability (New Facility Requirements, Cost Allocation)	Cost allocation under load risk: stranded upgrades, bill volatility; Tariff design: minimum bills, deposits, large-load classes; Equity: limit unjustified cost shifts to other customers. <i>See Sec. IV</i>	Incremental infrastructure needs: storage, reactive power (VAR) support, filtering, monitoring; System vs. behind-the-meter choice; New ramp products market. <i>See Sec. VI</i>

II. AI DATA CENTERS AS A NEW CLASS OF DEMAND: UNCERTAINTY AND MULTI-SCALE VARIATION

A. A New Class of Demand

Conventional power system planning and rate design largely evolved under gradual growth of industrial customers whose operating schedules and energy use are comparatively stable. AI data centers depart from this paradigm in three interrelated aspects: i) they materialize rapidly as large step changes at a small number of substations; ii) the magnitude, timing, and even the realization of announced projects are highly uncertain; and iii) once in operation, their power consumption exhibits distinctive temporal dynamics driven by workload characteristics. This section formalizes two core concepts used throughout the paper: *uncertainty* (Section II-B) and *multi-timescale variation* (Section II-C). A spatial concentration metric is introduced to contextualize why these issues matter for grid planning. We also use cryptocurrency mining as a comparative baseline: crypto and AI share electrical infrastructure and compute hardware, but differ in economic drivers, operational flexibility, and long-horizon predictability.

1) *Spatial Concentration:* Emerging compute loads exhibit pronounced *uneven spatial distribution*: large capacity expansions often concentrate at a small number of substations rather than spread across multiple feeders. To quantify this concentration, we define an area load intensity metric:

$$I_{\text{area}}(\mathcal{A}) := \frac{\sum_{j \in \mathcal{D}(\mathcal{A})} P_j^{\text{conn}}}{|\mathcal{A}|} \quad [\text{MW}/\text{mi}^2], \quad (1)$$

where $\mathcal{A}$ is a geographic region (e.g., a county, utility planning zone, or buffer around a substation), $|\mathcal{A}|$ is its land area, and $\mathcal{D}(\mathcal{A})$ is the set of large-load facilities within $\mathcal{A}$, each with interconnection capacity P_j^{conn} .¹

Table II places compute-driven load intensities in perspective. The U.S. nationwide average is approximately 0.125 MW/mi², derived from 2023 retail sales of 3,861 billion kWh [13] spread over 3,532,316 mi² of land area [14]. Typical suburban residential areas fall in the 5–20 MW/mi² range, and New York City as a whole averages roughly 36 MW/mi² (2024 forecast summer peak $\approx 11,000$ MW [15] across ≈ 300 mi² of land area [16]).

The latest generation of AI-focused data centers exhibits substantially higher densities than traditional hyperscale facilities. Vantage Data Centers’ Frontier campus in Texas is designed for 1.4 GW of IT load on 1,200 acres, implying a campus-average intensity near 750 MW/mi² [17]; smaller-footprint AI facilities with liquid-cooled, ultra-high-density racks (≥ 250 kW per rack [17]) can exceed 2,000 MW/mi² (authors’ extrapolation; industry surveys report rack densities above 100–200 kW becoming increasingly common [18]). Cryptocurrency mining facilities occupy a similar range:

¹In addition to AI data centers and cryptocurrency mining, several other emerging load categories exhibit similarly concentrated spatial footprints and share many of the planning challenges discussed here. These loads include Megawatt Charging Systems for freight trucks (1–3.75 MW per charger; station-level loads of 20 MW or more) [10], industrial-scale green-hydrogen electrolyzers (100–700 MW projects operational, under construction, or at final investment decision) [11], and semiconductor fabs (roughly 200 MW for a single leading-edge fab, exceeding 1 GW for full multi-fab campus build-out) [12].

TABLE II
REPRESENTATIVE AREAL LOAD INTENSITIES

Context	Intensity (MW/mi ²)
U.S. nationwide average	≈ 0.125
Suburban residential	5–20
New York City (citywide avg.)	≈ 36
Mid-sized city downtown core	50–200
Large-scale AI/DC campus	500–1,000
e.g., Stargate, TX (1.2 GW)	≈ 700
e.g., Meta Hyperion (5 GW planned)	≈ 1,000
Crypto mining campus	1,000–2,000
Manhattan (peak)	up to 2,000
Ultra-dense AI data center	2,000–3,000

Note: the U.S. row is an average power density; other rows are on a peak-demand or connected-capacity basis. Sources and derivations are given in the text; ranges without sources cited in the text are authors' estimates.

Riot Platforms' Corsicana campus delivers 400 MW (Phase 1) on approximately 265 acres (≈ 1,000 MW/mi²), with a planned build-out to 1 GW [19]. Two flagship projects under construction illustrate these campus-scale intensities and the contrasting supply arrangements they induce. Meta's Hyperion campus in Richland Parish, Louisiana—announced in December 2024 at 2 GW on 2,250 acres and expanded in 2026 to 5 GW of compute capacity on roughly 3,200 acres (≈ 1,000 MW/mi²)—will be served by the vertically integrated utility Entergy Louisiana through dedicated new resources approved by the Louisiana Public Service Commission: three combined-cycle units totaling approximately 2.26 GW and up to 1.5 GW of solar, with Meta contractually covering their revenue requirement [20], [21]. By contrast, in ERCOT's energy-only market, the OpenAI/Crusoe Stargate campus in Abilene, Texas (1.2 GW on approximately 1,100 acres; ≈ 700 MW/mi²) pairs its grid interconnection with developer-procured behind-the-meter generation, including roughly 360 MW of permitted on-site gas turbines [22], [23]. For comparison, Con Edison reports that portions of Manhattan reach load densities up to 2,000 MW/mi², but that urban load is distributed across many end users and feeders, whereas a data-center or mining campus of comparable intensity concentrates demand at one or a few interconnection points, amplifying both planning uncertainty and operational variability at specific grid locations. National Laboratory of the Rockies (NLR, formerly NREL)'s *Speed-to-Power* data viewer provided a geospatial basis to evaluate (1) by aggregating mapped data-center demand with infrastructure attributes [25], [26].

2) *Uncertainty and Variation*: Spatial concentration sets the stage for the two core themes developed below. Because large loads cluster geographically, forecast errors in siting and energization timing do not average out. They translate directly into localized infrastructure stress (substations, feeders, import interfaces) and, in some cases, zonal adequacy and deliverability risks. Section II-B examines the sources and

implications of this *uncertainty*. Once a facility is operational, its workload-driven power profile can differ markedly from traditional industrial loads; Section II-C characterizes this inter-temporal variability across timescales from sub-seconds to years, linking workload signatures (training vs. inference) to practical metrics.

Fig. 1 summarizes how these two properties manifest across time. Uncertainty is most consequential at longer horizons (typically months to decades) because it determines whether planners procure sufficient firm capacity and whether transmission and substation reinforcements are initiated early enough to meet lead times. In this regime, the dominant risks are commitment errors (over-build versus under-build) and stranded-cost exposure when costs cannot be shifted to other customers, which can lead to project delays or downsizing. These risks motivate the planning and adequacy focus of this paper.

By contrast, variation is an operational timescale phenomenon layered on top of concentrated spatial footprints. Power-system operations and market designs already manage routine variability (load following, regulation, contingency reserves), but AI data centers can introduce unfamiliar patterns and amplitudes of change. The most acute gaps tend to arise in the *sub-second to minutes* regime, where power-electronic interactions, protection/ride-through behavior, and fast ramps from coordinated workload scheduling can stress stability and operating security beyond what existing assumptions and products were designed to cover. On-site UPS and battery storage can partially mitigate ramps and provide ride-through support (Sections V–VI), but deployment is not universal, energy and duration are limited, and phenomena such as harmonic interactions and protection coordination often remain system-level concerns that behind-the-meter assets alone do not fully address. Accordingly, our variation-focused discussion emphasizes sub-second dynamics relevant to stability, and seconds-to-minutes ramps relevant to reserves, congestion management, and flexibility procurement, while also noting that longer-term variability (hours to seasons) feeds back into planning through peak coincidence and capacity accreditation.

B. Uncertainty: Forecasting Under Rapid Technical and Policy Change

AI data center demand is unusually difficult to forecast because it is shaped by rapidly moving and interacting factors that are largely *exogenous* to traditional load forecasting models: model and hardware innovation, competitive “AI race” dynamics, supply-chain bottlenecks, and evolving policy and geopolitical conditions. Efficiency breakthroughs can *lower* energy per unit of service (e.g., less compute per token or better utilization), while simultaneously *accelerating* adoption and shifting where workloads run (cloud, colocation, on-prem, edge). The net effect is often an *expanded scenario envelope*, not a narrowing one, because both demand growth and intensity can change direction quickly depending on technical and business trajectories.

This uncertainty differs from many conventional industrial loads, where demand is more tightly coupled to local economic activity, permitting timelines, and relatively stable process

²See [24]: EIA Energy Conference presentation (Energy Storage Activities), slide reporting Manhattan load density.

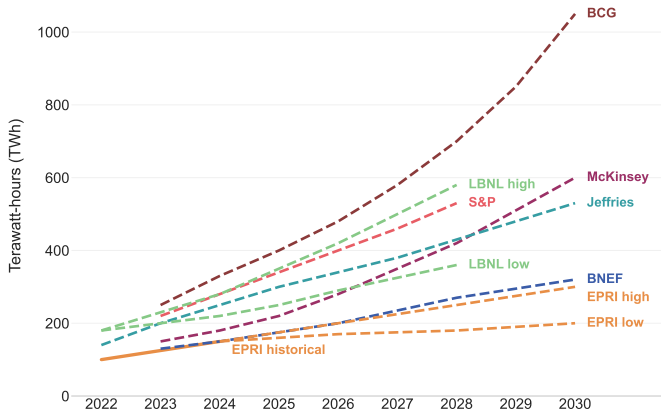


Fig. 2. Published projections of U.S. data-center electricity consumption through 2030 from major forecasting organizations [27]. The trajectories and reported high–low ranges reflect different assumptions about load growth and infrastructure deployment; their spread at 2030 illustrates forecast uncertainty rather than a statistical confidence interval.

technologies. It also differs from cryptocurrency mining in an important long-horizon sense: while crypto mining can be highly price-responsive in the short run, protocol-level issuance rules and supply caps are set far in advance, providing a structurally different long-term context for baseline adoption trajectories. For power systems, the forecasting challenge is therefore not only the magnitude of AI load, but also its *conditionality*: which projects are real, when they will energize, how quickly they will ramp, and how firm or interruptible demand will be under different grid and price conditions.

A practical consequence is that credible organizations can arrive at materially different views of 2030 demand even when they agree that growth will be rapid. This divergence reflects different assumptions about AI adoption and model efficiency, the share of workloads that migrate to hyperscale facilities, siting constraints and interconnection timelines, and the availability of generation and delivery infrastructure. Fig. 2 illustrates this dispersion: scenario ranges published by major analysts and research organizations span multiple hundreds of TWh by 2030, implying that the standard deviation of 3+ year forecasts is substantially larger than what planners typically face for conventional industrial loads.

Moreover, forecasting is itself a moving target: estimates are repeatedly revised as new information arrives about interconnection queues, procurement strategies, hardware supply, and policy constraints. In some cases, published outlooks have shifted significantly within months as assumptions about build rates, load realization, and infrastructure bottlenecks are updated. Fig. 3 provides an example of this “forecast drift” over time. For system planners and regulators, the implication is methodological: forecasts of data center demand should be updated iteratively as new information becomes available, with explicit milestones used to distinguish exploratory inquiries from committed projects and to revise scenarios as projects advance toward construction and energization.

Another related and often underappreciated source of forecast instability is the risk of *double counting* large-load projects across jurisdictions and interconnection queues. Data

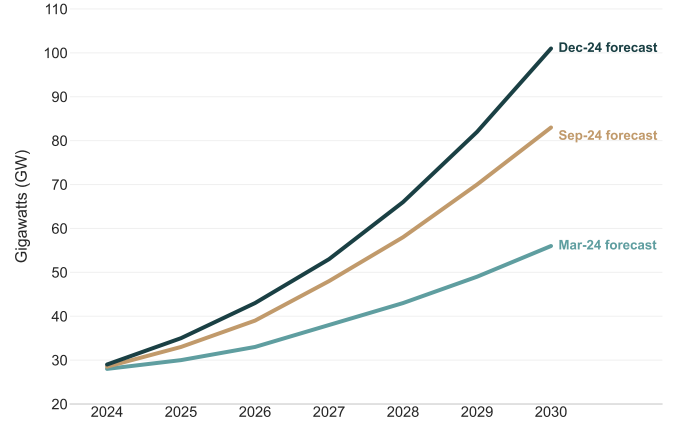


Fig. 3. Successive McKinsey forecast vintages for U.S. data-center capacity through 2030 [28]. The March, September, and December 2024 curves are alternative publication vintages over the same forecast horizon; their upward shifts show revisions as new information became available, not realized capacity growth between the publication dates.

center developers have strong incentives to explore multiple siting and procurement options in parallel: initiating service inquiries, interconnection requests, or preliminary regulatory reviews with more than one utility or transmission provider before committing to a single location. From the developer’s perspective, such redundancy is rational risk management in the face of uncertain power availability, pricing, and permitting timelines. From the system planner’s perspective, however, these parallel inquiries can inflate aggregate load forecasts if each request is treated as an independent, additive project. This phenomenon closely mirrors the well-documented experience on the supply side, where wind, solar, and storage developers routinely enter multiple interconnection queues, and historically less than a quarter of queued projects ultimately reach commercial operation. Without explicit mechanisms to identify mutually exclusive alternatives, load forecasts can systematically overstate realized demand and bias resource adequacy assessments toward over-procurement.

C. Inter-temporal Variability and Multi-Scale Load Dynamics

1) Workload-driven variability: training vs. inference.:

Even after a data center’s nameplate capacity is known, its inter-temporal variation can be atypical for large loads. *Training* workloads synchronize iterative computation across many accelerators; that synchronization produces coherent, step-like aggregate swings (cause) that propagate through the facility power-delivery chain (mechanism), and appear as fast ramps at the grid interface (impact). *Inference* workloads are constrained by latency and driven by bursty user demand; those demand bursts (cause) concentrate power changes into diurnal and event-driven patterns (mechanism), elevating congestion exposure and coincident-peak contribution even when annual energy is unchanged (impact). These behaviors span milliseconds to years: sub-second excursions interact with voltage control, protection, and power-quality compliance, while minute-scale ramps stress reserves, congestion management, and dispatch feasibility. “Average MW” therefore fails

to describe the operational and planning risk profile in the AI era.

2) *A timescale crosswalk from workload signatures to grid observables.*: We adopt a two-level timescale hierarchy that links i) representative phenomena, ii) primary facility- and grid-side components (“actors”), and iii) quantitative metrics used later in Sections III–VI.³ The fast-dynamics layer (μs –60 s) captures converter-rich behavior and control interactions that determine *survivability* and compliance (power quality, ride-through, and sub-minute regulation behavior) [29], [30]. The operational layer (1 min–24 h) captures dispatch- and market-relevant variability (net-load ramps, reserve needs, congestion exposure, and diurnal/event-driven structure) [2], [31].

Fast-timescale phenomena impose *hard* feasibility constraints that define an operating envelope; within that envelope, operational coordination allocates scarce flexibility (or reveals its absence), shaping near-term congestion and curtailment risk [32].

a) *Fast dynamics (μs –60 s): feasibility and compliance (overview)*: At this layer, the central questions are “*can the system ride through and remain compliant?*” rather than “*what is the marginal price?*” [33]. Converter-dominated power delivery (server power supply units (PSUs) and voltage regulator modules (VRMs), rectifier front ends, harmonic filtering, and UPS and battery energy storage system (BESS) interfaces) makes data centers salient, and synchronized *training* can compress many server-level transitions into coherent steps at the point of interconnection [5]. Representative observables include harmonic/interharmonic distortion and flicker (e.g., IEC 61000-4-7/4-15), voltage unbalance, sag/swell tolerance (e.g., ITI/CBEMA), RoCoF and frequency nadir, and sub-minute ramp rates relevant to regulation behavior [3], [34]–[37]. Appropriate tools include EMT simulation or impedance-based screening for interaction risks, and validated dynamic models of UPS and IT power controls [6].

b) *Operations (1 min–24 h): dispatch and market-relevant variability (overview)*: Minute-to-day variability is where feasibility constraints translate into economic outcomes: commitment feasibility, reserve sufficiency, congestion, scarcity pricing, and curtailment [38]. Representative metrics include MW/min ramp rates (and rebound), reserve sufficiency margins, binding-constraint indicators, scarcity adders, and locational marginal price (LMP) statistics (mean/variance/spikes), along with response-time and performance measures for demand response participation [39]. Mechanisms that matter include workload orchestration (shifting/deferral), use of UPS/BESS as short-duration buffers, and thermal/cooling coupling that shapes sustained loading during hot conditions [40]–[42].

3) *Spectral characteristics and multi-scale validation.*: We validate this multi-scale structure directly using the NLR Dataset of Generative AI Workload Power Profiles [43], [44]:

³This hierarchy is a *modeling crosswalk*, not a claim that any single timescale alone determines system risk. The intent is to clarify which physics and controls are dominant, what the observables are, and which tools (electromagnetic transient (EMT), phasor-domain dynamics, production-cost models, or planning models) are appropriate.

area-normalized Welch power spectral densities (PSDs) for the GPU and CPU-package channels for five representative training and inference traces (Fig. 4). Training workloads show clean harmonic combs at a ~ 10 –13 s step cadence, recovering the training-step period directly from the power spectrum; offline and online-rate inference share a distinct ~ 0.75 s peak suggestive of inference-engine-level periodicity, and GPU spectra are markedly more tonal than the broadband CPU-package spectra throughout.

Decomposing variance by time scale (Fig. 5) shows that the dominant band depends on the workload: training and offline inference concentrate variance in the 1–10 s band, whereas the online inference variants shift toward 10–100 s and above. Sub-second content accounts for under 1% of GPU variance in every trace, although the CPU-package channel carries substantially more sub-second content for the online variants. These are properties of the five measured traces rather than universal characteristics of AI workloads, and the spectral structure should be expected to differ with model, serving configuration, and hardware. The point we draw from them is methodological: the relevant structure of these loads is workload-dependent and lives at time scales that peak-demand descriptors do not express, which is also what identifies where demand-flexibility and dynamic voltage and frequency scaling (DVFS) interventions are best matched [45].

III. RELIABILITY UNDER UNCERTAINTY: SYSTEM-LEVEL RESOURCE ADEQUACY

In this section, we use *reliability* in the planning sense, focusing primarily on system-level *resource adequacy*: whether the system has sufficient deliverable resources to serve demand across credible realizations of uncertainty. As illustrated in Fig. 6, there are two broad institutional pathways by which jurisdictions seek to ensure resource adequacy. The first is centralized planning, where a utility or system planner selects cost-effective investments and adequacy is enforced through planning criteria and regulatory processes [46]. The second is market-based planning, where investment incentives are mediated through capacity markets with reliability requirements translated into market demand curves, performance rules, and deliverability requirements [47]. Importantly, these two pathways both rely on probabilistic resource adequacy assessment (RAA) to map uncertainties and constraints into quantitative risk metrics and compliance determinations [48].

An important resource option in this context is behind-the-meter (BTM) generation at the data-center site. It is useful to distinguish BTM generation that is intended to serve the data-center load on a routine basis from on-site backup generation. BTM generation may supply all or part of the facility’s electricity demand as a substitute for grid-supplied generation, thereby reducing the net load for which the bulk power system must plan and maintain resource adequacy. By contrast, backup generation is primarily intended to maintain facility operation during grid outages or other emergency conditions and should not automatically be treated as a resource that offsets system-supplied demand for resource-adequacy purposes. This distinction is important because BTM supply

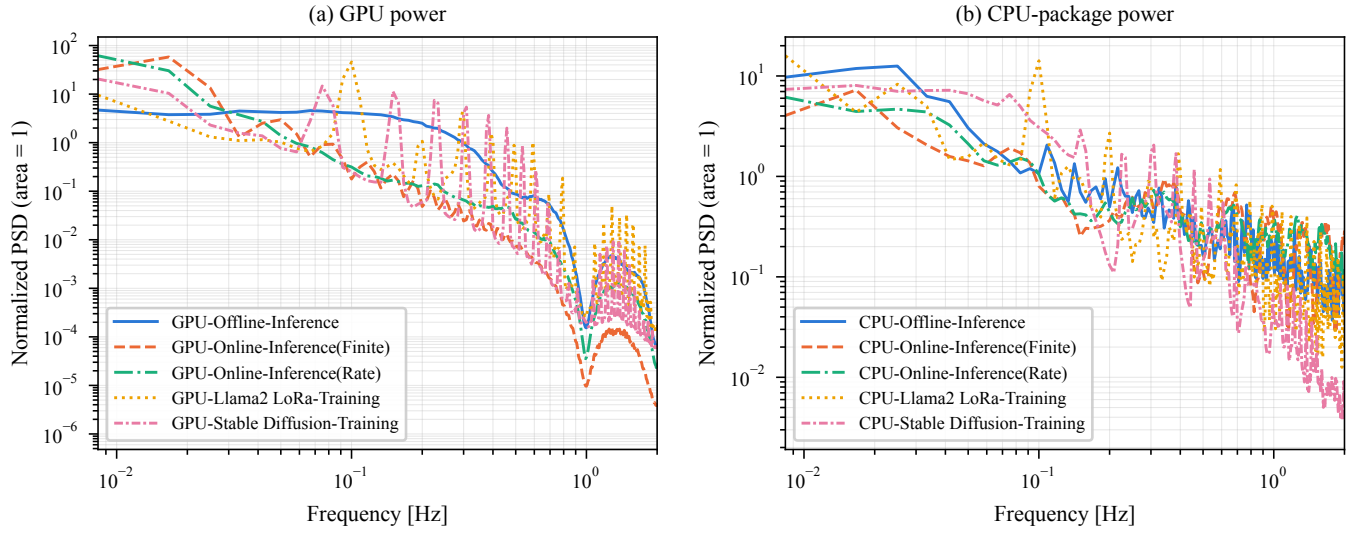


Fig. 4. Area-normalized Welch power spectral densities (120 s Hann window, 50% overlap) for five representative single-node AI workload traces from the NLR Dataset of Generative AI Workload Power Profiles [43], [44], shown for (a) GPU power and (b) CPU-package power.

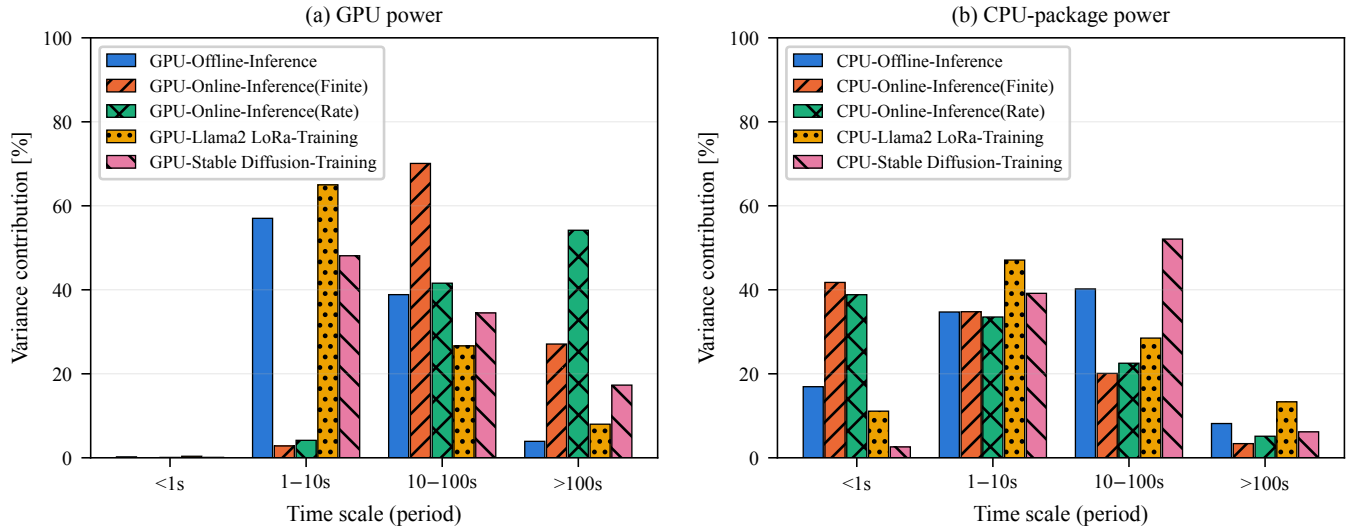


Fig. 5. Variance decomposition by time scale (<1 s, 1–10 s, 10–100 s, >100 s, the last bounded by the 120 s Welch-window resolution) for the same five traces, shown for (a) GPU power and (b) CPU-package power.

and backup generation are sometimes conflated, even though their operating roles and implications for system planning can be substantially different.

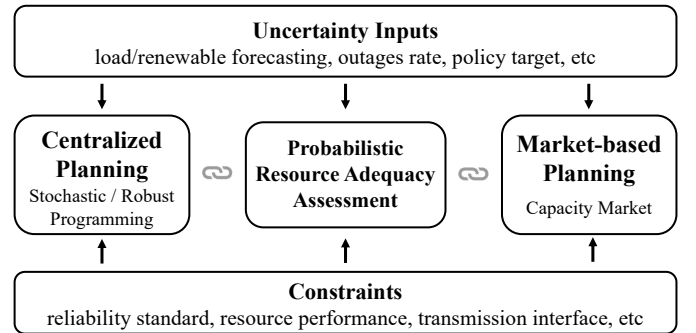


Fig. 6. Two institutional pathways to resource adequacy and their common analytical core. Uncertainty inputs such as load and renewable forecasts, outage rates, and policy targets enter probabilistic resource adequacy assessment under reliability, resource-performance, and transmission constraints. The resulting adequacy information supports either centralized stochastic/robust planning or market-based procurement through capacity markets.

In the following, we first briefly introduce the probabilistic RAA model and show how AI data-center growth expands the uncertainty affecting load modeling, complicating the interpretation of standard adequacy metrics. We then discuss how these uncertainties challenge centralized planning by increasing the value of staged commitments and tightening the coupling between generation expansion and transmission deliverability. Finally, we turn to market-based planning and explain how data-center uncertainty stresses capacity market design, including demand-curve calibration, deliverability requirements, and performance incentives.

A. Probabilistic Resource Adequacy Assessment

1) *System Model in RAA*: Regardless of whether adequacy is pursued through centralized planning or market-based mechanisms, compliance is often evaluated by Monte Carlo simulation tools rather than being explicitly embedded in the optimization program or market-clearing process. Conventional RAA models were designed around the forced-outage behavior of thermal generators, simulating unit-level outages using technology-specific outage rates and repair-time distributions [49]. Modern RAA simulators, such as GE-MARS [50] and PRAS [51], extend this framework to include variable renewable resources and storage: renewable availability is typically represented using historical (or weather-simulated) production profiles, while storage is modeled through an assumed dispatch policy that maps net-load conditions into charging and discharging decisions.

Let $\mathcal{T}$ denote the set of all hours in a year (or the study periods of interest), $\mathcal{D}$ the set of all days, and Ω the set of all possible year-long scenarios for the system under study, assumed to be finite. Following the notation in [52], we construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ represents the σ -algebra of events on Ω and $\mathbb{P}$ is a probability measure defined on $\mathcal{F}$ satisfying standard measure-theoretic properties. Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable representing our resource adequacy (RA) metric for any given scenario. For each generator $g \in \mathcal{G}$, let $x_g \in \mathbb{R}_+$ denote its installed capacity. In the absence of storage, and assuming that generation resources have no intertemporal operating constraints, we define $A_{g\tau\omega} : \mathcal{T} \times \Omega \rightarrow [0, 1]$ as a stochastic process representing generator g 's availability, where $A_{g\tau\omega}$ denotes its realization in hour τ of scenario ω . Then, the total available capacity is the stochastic process on $\mathcal{T} \times \Omega \rightarrow \mathbb{R}_+$ defined as

$$\hat{x}_{\tau\omega} = \sum_{g \in \mathcal{G}} x_g A_{g\tau\omega}. \quad (2)$$

The system load profile is typically sampled from a set of representative scenarios, with $L_{\tau\omega}$ denoting load in hour τ of scenario ω . For notational compactness, we define the vectors of hourly values within scenario ω as

$$\hat{\mathbf{x}}_{\omega} := (\hat{x}_{\tau\omega})_{\tau \in \mathcal{T}}, \quad \mathbf{L}_{\omega} := (L_{\tau\omega})_{\tau \in \mathcal{T}}, \quad (3)$$

representing total available capacity and system load, respectively.

2) *Resource Adequacy Metrics*: For semantic clarity, we distinguish three related but distinct concepts: the *RA metric*, the *risk measure*, and the *reliability standard*. The RA metric specifies how shortage events are quantified and aggregated within a single assessment period, such as Unserved Energy (UE), Loss of Load Hours (LOLH), and Loss of Load Days (LOLD). The risk measure defines how these per-scenario outcomes are aggregated across multiple stochastic realizations (e.g., expectation or percentile). Finally, the reliability standard specifies the target or benchmark value of the risk measure that the system is expected to satisfy, such as the ‘one day in ten years’ LOLD criterion commonly used in North America since the 1960s [53]. We define the RA metric for each scenario ω as a functional

$$M : \mathbb{R}_+^{|\mathcal{T}|} \times \mathbb{R}_+^{|\mathcal{T}|} \rightarrow \mathbb{R}_+, \quad M(\hat{\mathbf{x}}_{\omega}, \mathbf{L}_{\omega}), \quad (4)$$

which aggregates unserved load over time according to the chosen reliability measure. For example, LOLH can be represented as:

$$M_{\text{LOLH}}(\hat{\mathbf{x}}_{\omega}, \mathbf{L}_{\omega}) = \sum_{\tau \in \mathcal{T}} \mathbb{1}\{(L_{\tau\omega} - \hat{x}_{\tau\omega}) > 0\}. \quad (5)$$

The system-level RA measure is defined by aggregating the per-scenario metric across the probability space:

$$\mathcal{M}(\hat{\mathbf{x}}, \mathbf{L}) := \Phi_{\omega}[M(\hat{\mathbf{x}}_{\omega}, \mathbf{L}_{\omega})], \quad (6)$$

where $\Phi_{\omega}[\cdot]$ is a risk operator, typically expectation.

3) *Effective Load Carrying Capability*: Effective Load Carrying Capability (ELCC) provides a reliability-based measure of the capacity contribution of a resource. In words, the (marginal) ELCC of a candidate unit is the largest constant load increase that can be served after adding the unit while maintaining the same system-level resource-adequacy risk.

Consider a candidate resource g with nameplate capacity $\Delta x > 0$ and scenario-dependent availability $\mathbf{A}_g := (A_{g\tau\omega})_{\tau \in \mathcal{T}, \omega \in \Omega}$. Using the system-level RA measure $\mathcal{M}$ defined above, its marginal ELCC is defined as

$$L_c^{\text{ELCC}} := \max_{L_c \geq 0} \left\{ L_c : \mathcal{M}(\hat{\mathbf{x}} + \Delta x \mathbf{A}_g, \mathbf{L} + L_c \mathbf{1}) \leq \mathcal{M}(\hat{\mathbf{x}}, \mathbf{L}) \right\}, \quad (7)$$

where $\mathbf{1}$ is the all-ones vector, so that $L_c \mathbf{1}$ is a constant load increase across all study periods and scenarios. Equivalently, at the limiting value, the additional resource and additional load leave system reliability unchanged. The corresponding ELCC capacity accreditation is

$$\alpha_g^{\text{ELCC}} := \frac{L_c^{\text{ELCC}}}{\Delta x}, \quad \alpha_g^{\text{ELCC}} \in [0, 1]. \quad (8)$$

ELCC therefore provides an important link between probabilistic RAA and planning or market decisions. In centralized planning, ELCC-based capacity accreditation converts heterogeneous resources into comparable firm-capacity contributions and thereby informs whether a portfolio satisfies a planning reserve-margin (PRM) requirement. In capacity markets, the same principle determines the quantity of accredited capacity that a resource may offer or receive compensation for, effectively converting resources with different availability

characteristics into a common reliability product. Because ELCC depends on the existing resource portfolio, load profile, and joint availability conditions, it is inherently system- and portfolio-dependent rather than a fixed rating.

4) *Impact of Data Center and Increasing Uncertainty in Resource Adequacy Study*: Rapid expansion of data centers affects probabilistic RAA *primarily* by introducing more uncertainty in the load profile L_ω and by strengthening common-mode stress conditions under which high demand coincides with degraded resource availability [54]. In the notation above, the key modeling implications are: i) enlarging the scenario set Ω to represent uncertainty in data-center timing and scale, and ii) representing an uncertain operational characteristic that reduces *effective* load during scarcity, which decomposes $L_{\tau\omega}$ into baseline load plus an incremental data-center component [55]. Taken together, these changes tend to increase the variance and tail risk of shortage outcomes, and also impact the interpretation of standard adequacy metrics.

a) *Data Center Demand and Flexibility*: Recent work emphasizes that modern resource adequacy is no longer a purely supply-side problem with fixed demand; flexible demand can materially affect both the frequency and severity of load shedding, and large loads such as data centers introduce additional uncertainty because their flexibility is often imperfectly known *ex ante* [54], [56]. In our notation, a compact way to represent data-center-driven uncertainty is to decompose hourly load into a baseline component and a data-center component:

$$L_{\tau\omega} = L_{\tau\omega}^{\text{base}} + L_{\tau\omega}^{\text{DC}}, \quad L_{\tau\omega}^{\text{DC}} = K_\omega s_{\tau\omega}, \quad (9)$$

where K_ω captures the realized scale (MW) of data-center additions in scenario ω , and $s_{\tau\omega} \in [0, 1]$ summarizes the energization and utilization ramp. This representation makes explicit why expanding the scenario set Ω is often necessary: uncertainty in $(K_\omega, s_{\tau\omega})$ can dominate the upper tail of $L_{\tau\omega}$ even when $L_{\tau\omega}^{\text{base}}$ is sampled from representative weather years.

To reflect the modern distinction between *static* demand and *flexible* demand in adequacy studies, one model is to treat part of the data center load as potentially controllable during system stress [57]. Let $r_{\tau\omega} \geq 0$ denote the realized voluntary reduction (e.g., contracted demand response, managed curtailment, or behind-the-meter supply netting) available in hour τ and scenario ω . Then the system-served (effective) load becomes

$$L_{\tau\omega}^{\text{eff}} = L_{\tau\omega} - r_{\tau\omega}, \quad 0 \leq r_{\tau\omega} \leq \bar{r}_{\tau\omega}, \quad (10)$$

where $\bar{r}_{\tau\omega}$ is the flexibility envelope. In probabilistic RAA, the value of flexibility is driven mainly by its *coincidence* with scarcity hours rather than total curtailed energy. Consequently, uncertainty in the availability of $r_{\tau\omega}$ should be reflected in the scenario set Ω whenever data center flexibility is material.

b) *Metric Sensitivity*: Because data-center growth often introduces “lumpy” additions and steeper peak conditions, it can shift the shortage profile from many small deficits to fewer but larger events (or vice versa), which affects the relative

informativeness of common per-scenario metrics [53]. For instance, severity-sensitive metrics such as unserved energy,

$$M_{\text{UE}}(\omega) = \sum_{\tau \in \mathcal{T}} (L_{\tau\omega}^{\text{eff}} - \hat{x}_{\tau\omega})_+, \quad (11)$$

can distinguish plans that have similar event counts but very different deficit magnitudes, whereas frequency-based metrics such as LOLH (5) primarily reflect how often scarcity occurs. Under increased uncertainty, tail-aware risk aggregation $\Phi_\omega[\cdot]$ (e.g., conditional value at risk, CVaR) can become more relevant when adequacy concerns are driven by adverse combinations of high load growth and stressed system conditions.

c) *Co-Location*: The term *co-location* admits multiple interpretations across jurisdictions and regulatory contexts. For example, under PJM’s proposed definition, a co-located load refers to end-use customer load that is physically connected to the facilities of an existing or planned generator on the generator’s side of the point of interconnection (POI) to the transmission system [58]. Data Center co-location with on-site generation, storage, or firm imports can reduce the contribution of data centers to system risk by decreasing the load seen by the bulk system during stressed hours [59]. In the RAA abstraction, this effect is naturally captured through the same $r_{\tau\omega}$ term (netting behind-the-meter supply from load) and, crucially, by allowing $(r_{\tau\omega}, L_{\tau\omega})$ and the system availability process $(A_{g\tau\omega})$ to be correlated under common drivers (e.g., weather and fuel constraints). This correlation structure poses a new challenge when data-center growth is rapid and geographically concentrated.

B. Centralized Planning

In vertically integrated systems, resource adequacy is primarily ensured through integrated resource planning, utility procurement, and regulatory approval of long-lived investments. Under uncertain data center growth, the core challenge is not simply selecting an optimal portfolio for a given forecast, but sequencing irreversible investment decisions in the presence of evolving information. Over-commitment risks stranded assets and rate impacts, while under-commitment increases loss-of-load risk and reliance on emergency measures.

1) *Formulation of Generation Expansion Planning Under Uncertainty*: Generation expansion planning (GEP) determines long-lived capacity investments while ensuring that the resulting system can be operated feasibly and economically. A standard modeling backbone is a two-level (investment–operation) structure: a *first stage* commits capacity expansion, and a *second stage* performs system operation (e.g., economic dispatch) subject to operational and network constraints. Deterministic and stochastic variants have a long history in the GEP literature [60]–[65]. This framework also provides a natural entry point for discussing how AI data-center uncertainty shifts attention from static planning toward more adaptive (but computationally harder) multi-stage formulations [66]–[69].

Let x_c denote the installed (or incremental) capacity decision for technology c (MW), with unit investment cost Inv_c and upper bound $\bar{x}_c$. Let $\omega \in \Omega$ index uncertainty scenarios with probability p^ω . Let y^ω denote operating decisions in

scenario ω (dispatch, flows, reserves, etc.). Let d^ω denote scenario-dependent net load (demand net of variable supply if modeled that way). Let $X(d^\omega)$ denote the operational feasibility set, such as power balance, generator limits, ramping, transmission constraints, etc.

a) *Deterministic Two-stage Formulation*: The deterministic formulation replaces uncertainty with a single load forecast $d = (d_t)_{t \in \mathcal{T}}$ and chooses capacity investments x to minimize investment plus operating cost:

$$\begin{aligned} \min_x \quad & \sum_c \text{Inv}_c x_c + Q(x) \\ \text{s.t.} \quad & 0 \leq x_c \leq \bar{x}_c, \quad \forall c, \end{aligned} \quad (12)$$

where the second-stage operating cost is given by

$$\begin{aligned} Q(x) = \min_y \quad & \sum_{t \in \mathcal{T}} \sum_g C_{gt}^E P_{gt}^E \\ \text{s.t.} \quad & (y, x) \in X(d). \end{aligned} \quad (13)$$

Here y collects operational decisions, P_{gt}^E denotes generation output, and $X(d)$ represents standard operational and network constraints under the deterministic load forecast d .

b) *Stochastic Two-stage Formulation*: To represent uncertainty explicitly, the deterministic operating term is replaced by an expectation over scenarios:

$$\begin{aligned} \min_x \quad & \sum_c \text{Inv}_c x_c + \sum_{\omega \in \Omega} p^\omega Q(x, \omega) \\ \text{s.t.} \quad & 0 \leq x_c \leq \bar{x}_c, \quad \forall c, \end{aligned} \quad (14)$$

where the second-stage operating cost in each scenario ω is given by

$$\begin{aligned} Q(x, \omega) = \min_{y^\omega} \quad & \sum_{t \in \mathcal{T}} \sum_g C_{gt}^E P_{gt}^{E\omega} \\ \text{s.t.} \quad & (y^\omega, x) \in X(d^\omega). \end{aligned} \quad (15)$$

In this “here-and-now” structure, capacity decisions x are committed before uncertainty is realized, while operational decisions y^ω adapt to each scenario. Depending on the planner’s risk preferences, the expected-cost objective can be replaced by alternative risk-aware formulations, such as robust optimization or chance-constrained planning, to hedge against adverse load realizations [70], [71].

c) *Multi-stage Stochastic Optimization*: Two-stage models can under-value flexibility when uncertainty resolves gradually over multiple planning cycles, as is common with large load additions. Multi-stage stochastic GEP generalizes (14)–(15) by allowing capacity decisions to adapt over multiple periods as uncertainty unfolds on a scenario tree, subject to non-anticipativity [66], [67]. The corresponding challenge is computational: scenario trees grow rapidly with horizon length and branching, and realistic multi-stage models (often with discrete investment decisions) can become large-scale stochastic programs that are difficult to solve and update frequently [68].

2) *Impact of Data Center Uncertainty in Centralized Planning*: Rapid growth in data center demand changes centralized planning primarily by increasing i) the value of *adaptivity* in investment decisions, ii) the coupling between

generation and transmission expansion, and iii) the difficulty of enforcing reliability requirements *endogenously* within an optimization model. Unlike ex post Monte Carlo RAA, a planning model must commit long-lived decisions before key uncertainty drivers are revealed, so the formulation must trade computational tractability against fidelity to adequacy risk.

a) *Increasing Need for Multi-stage and Adaptive Planning*: Large data center additions are often uneven and only partially revealed over time (e.g., as interconnection milestones are met, construction progresses, and tenant/utilization becomes firm). This information structure is not well captured by a “here-and-now” two-stage model, which commits all investments before uncertainty is realized. Multi-stage formulations instead allow investment decisions to adapt across planning stages as new information arrives.

The computational cost, however, rises quickly with the number of stages. Each additional stage typically requires a richer scenario representation of how uncertainty unfolds (i.e., more branching in the scenario tree) and introduces a new set of stage-dependent decision variables replicated across scenario nodes. As a result, model size can grow roughly in proportion to the number of nodes in the scenario tree, making realistic multi-stage stochastic GEP substantially harder to solve.

b) *Increasing Coupling with Transmission Planning*: Geographically concentrated data center growth makes deliverability constraints bind more frequently, so planning must increasingly co-optimize generation and transmission. In compact form, this expands the feasibility set from $X(d^\omega)$ to a network-aware set that includes transmission build decisions $(y^\omega, x) \in X(d^\omega)$ to $(y^\omega, x, z) \in X(d^\omega; z)$. A planner-initiated siting framework that pre-screens interconnection-ready locations under standardized operational-flexibility envelopes (firm, pause, shift) operationalizes this coupling directly: on a synthetic 2000-bus Texas system it expands the siting frontier by 9–21% and cuts first-energization timelines from 5–8 years to 12–18 months relative to firm operation [9].

Transmission expansion often introduces binary investment variables and lumpy project choices, so even if the generation expansion model is linear/convex, joint planning typically becomes a mixed-integer program [72]. In a multi-stage setting, integrality interacts with non-anticipativity and scenario growth, substantially increasing solve time and limiting how frequently the plan can be re-optimized as new data center information arrives.

c) *Embedding Reliability Requirements into GEP*: Centralized planning models often enforce adequacy through simplified proxies (reserve margins, capacity constraints, or deterministic peak requirements), while probabilistic adequacy is evaluated ex post via Monte Carlo simulation and metrics such as loss of load expectation (LOLE) and expected unserved energy (EUE). Embedding the same reliability requirement directly into the investment optimization is difficult because reliability metrics are defined over a large scenario space and depend on detailed scarcity dynamics. A reduced-form approach is to introduce an explicit shortage variable in the dispatch layer and penalize it with the value of lost load (VoLL) or to impose an explicit risk constraint on unserved

TABLE III
SIMPLIFIED U.S. RA CONSTRUCTS BY ISO/RTOS

ISO/RTO	LSE RA Obligation	Capacity Market
CAISO	Yes	No
ERCOT	No	No
SPP	Yes	No
MISO	Yes	Yes
PJM	No	Yes
NYISO	Yes	Yes
ISO-NE	No	Yes

energy, consistent with the RAA notation (6). These embeddings make reliability endogenous, but they also highlight the core tension: higher-fidelity reliability constraints push the model toward large-scale stochastic optimization, motivating approximations, decomposition, or hybrid simulation-optimization frameworks [73].

C. Systems With a Capacity Market

Capacity markets ensure long-term resource adequacy in deregulated electricity markets by providing financial incentives for resources to commit their capacity in advance [47]. In these markets, the system operator procures *accredited* capacity in advance of the delivery year to satisfy an administratively chosen reliability criterion, most commonly expressed as a target LOLE. Procurement is implemented through a forward auction that clears against a demand curve, which translates the reliability target and the underlying probabilistic RAA into a willingness-to-pay for capacity.

Capacity markets act as a *risk-management layer* between probabilistic adequacy modeling and decentralized investment, using forward price signals to encourage entry when scarcity risk rises and to deter overinvestment when the system is long. In practice, they coexist with heterogeneous state policies, bilateral contracting, and complementary reliability tools such as local deliverability requirements and out-of-market procurement. Table III summarizes key differences in resource adequacy institutions across U.S. ISO/RTOs: some regions rely primarily on load-serving-entity (LSE) resource adequacy obligations and bilateral/self-supply arrangements, while others rely more heavily on ISO-run centralized capacity auctions. NYISO represents a hybrid case, where LSEs retain RA obligations but may satisfy them through bilateral contracts, self-supply, and NYISO-administered capacity auctions. The remainder of this subsection focuses on how data-center-driven load uncertainty challenges the design of centralized capacity markets, particularly through demand-curve calibration, deliverability constraints, and performance incentives.

Remark: ERCOT relies on its operating reserve demand curve (ORDC) and contracted demand response. PJM has an ISO-level centralized capacity market and also offers the Fixed Resource Requirement (FRR) alternative, under which LSEs may self-supply or contract for capacity. NYISO is classified as having both an LSE RA obligation and a capacity market: LSEs certify bilaterally contracted or self-supplied capacity before the NYISO-administered auction process, and any

residual shortfall can be met through NYISO's monthly/spot capacity market.

1) *Demand Curve Under Load Uncertainty:* Modern capacity markets have largely moved away from clearing a fixed (vertical) capacity target and toward *downward-sloping* capacity demand curves [74]. The key motivation is to explicitly represent the economic trade-off between higher procurement cost and higher reliability, and to reduce extreme price volatility that can occur when a vertical demand curve intersects a relatively inelastic supply stack near the target quantity. In practice, these curves are implemented as administratively defined demand schedules in forward auctions, anchored to a system reliability requirement and a benchmark for investment economics such as Cost of New Entry (CONE), which is used to indicate the amortized capital cost of constructing a power plant.

While the high-level principle is shared, implementations differ materially across ISO/RTOs. For example, PJM's curve is parameterized around a region-wide reliability requirement in UCAP terms and Net CONE, and includes an explicit price-cap segment before transitioning to downward-sloping segments [75]. ISO-NE historically cleared a fixed quantity before adopting a sloped curve and subsequently evolving toward a probabilistic *Marginal Reliability Impact* (MRI) construct, where the slope is tied to the incremental reduction in EUE [76]. NYISO uses a multi-segment zonal/system curve with a cap and a zero-crossing quantity, translating between installed capacity (ICAP) and unforced capacity (UCAP) using derate factors [77]. This heterogeneity is not merely cosmetic: it reflects differing views on how well current curves map to consumer preferences and how closely demand-curve parameters should be benchmarked to the incremental reliability value of capacity.

2) *The Impact of Data Center Uncertainty:* AI data-center growth stresses capacity market mechanisms because incremental demand is often concentrated in a small number of large, milestone-dependent projects with uncertain energization dates and realized utilization. As a result, the capacity market demand curve becomes unusually sensitive to load forecast error. High-biased forecasts shift the demand curve outward, raising near-term procurement and clearing prices, while low-biased forecasts suppress prices but increase the likelihood of future scarcity and corrective actions. Geographic concentration further tightens local delivery constraints even when system-wide capacity appears sufficient. Under these conditions, the demand curve functions as the primary *risk-allocation interface*: its slope, cap-and-collar features, and update frequency determine how forecast errors propagate into price volatility, investment timing, and exposure to market power when the system clears near scarcity.

a) *Uncertainty in Demand Curve Estimation and Potential Market Power:* Data-center growth increases not only the level of forecasted peak demand, but also the variance and skewness of forecast error because many projects are discrete, milestone-driven, and geographically clustered. This can make the capacity requirement "jump" across auction years, increasing volatility in the clearing price even if the underlying supply stack changes slowly. In tight conditions,

the price impact of forecast error is amplified because the market clears near the steep portion of the demand curve, where small quantity shifts translate to large price movements.

When reserve margins tighten, capacity markets can become more susceptible to structural and strategic market power [78]. Under uncertain data center growth, a demand curve calibrated on an aggressive load forecast can push the market toward these tight conditions even before load is realized, increasing the probability that prices are set by scarcity rents or mitigation rules rather than by competitive fundamentals. This is one reason that capacity market design and market power mitigation become more central when load growth is dominated by a small number of uncertain mega-load projects.

b) *Mitigation Levers*: Common design responses aim to reduce the sensitivity of auction outcomes to forecast noise while maintaining investable signals. These include: i) price caps and collars in the demand curve to limit extreme outcomes when forecasts are wrong or supply is temporarily constrained; ii) prompt or closer-to-delivery procurement to correct forecast errors with better information and shorter supply response requirements; and iii) more granular/seasonal products where reliability risk is increasingly seasonal. Operationally, these instruments also reduce the value of strategic behavior by increasing procurement frequency and by narrowing the time window over which a single forecast error can dominate procurement [79].

IV. UTILITY TARIFF DESIGN AND COST ALLOCATION

Large-scale AI data centers introduce an affordability risk that differs from incremental, diversified load growth: they induce *discrete* and often front-loaded transmission and distribution (T&D) capacity expansion, while their realized demand trajectory (MW level, ramp profile, and site persistence) remains uncertain. When realized utilization falls below the planning assumption, the associated fixed-cost recovery can be shifted to legacy customers through subsequent rate cases. This section summarizes how U.S. jurisdictions are adapting tariff design and cost allocation to i) reduce speculative build risk, ii) assign incremental delivery costs to cost-causing entities, and iii) limit cross-subsidies to residential and small commercial customers.

A. Why “cost shifting” arises under cost-of-service regulation

Under cost-of-service regulation (COSR), retail rates are set by i) establishing an allowed revenue requirement, ii) functionalizing and allocating that revenue requirement across customer classes and functional cost “buckets,” and iii) mapping each class responsibility to specific rate elements. A compact representation is

$$RR_y = O_y + \delta_y + r_y RB_y + T_y, \quad (16)$$

where RR_y is the allowed revenue requirement in year y (\$/year), O_y denotes approved operating expenses (including O&M and pass-through items treated as operating costs), δ_y is the year- y book depreciation that increases accumulated depreciation (and therefore reduces rate base), RB_y is the (depreciated original cost) rate base, r_y is the allowed rate of

return, and T_y aggregates income/property taxes and other regulatory adjustments (e.g., riders, true-ups). Rate base evolves according to

$$RB_{y+1} = RB_y + I_y - \delta_y - \rho_y, \quad (17)$$

where I_y denotes new plant additions placed in service and included in rate base, and ρ_y captures net plant retirements/write-offs removed from rate base. Eqs. (16)–(17) emphasize the *burnout* property: if new investment slows ($I_y \approx 0$), then RB_y declines mechanically as depreciation and retirements continue, making sustained utilization important when large assets are built in anticipation of future load [80].

Functionalization is important because customers connected at different voltage levels do not make use of the same set of utility facilities. Residential and small commercial customers are generally allocated costs associated with secondary and primary distribution, as well as appropriate shares of sub-transmission, high-voltage transmission, and generation. Medium commercial and industrial customers typically avoid secondary-distribution costs, while large customers connected at sub-transmission voltage generally avoid both primary and secondary distribution costs. Very large loads connected directly at transmission voltage therefore primarily bear costs associated with high-voltage transmission and, in vertically integrated systems, generation. For hyperscale data centers, the principal cost-pooling concern consequently arises not from lower-voltage distribution facilities that they do not use, but from transmission and generation costs that are traditionally pooled across customers and allocated using load-based determinants such as coincident-peak demand. Creating a distinct hyperscale rate class or special tariff can alter this pooling arrangement by ring-fencing clearly incremental transmission and generation costs and assigning them more directly to the large loads that cause those investments.

Cost shifting arises primarily through the pooling and allocation of these functional cost categories. In practice, costs are organized into buckets $k \in \mathcal{K}$ (e.g., customer-related, demand-related, energy-related, transmission, distribution, and public-purpose costs). Class $c \in \mathcal{C}$ revenue responsibility is then computed via bucket-specific allocators:

$$RR_{c,y} = \sum_{k \in \mathcal{K}} \alpha_{c,k,y} C_{k,y}, \quad \sum_{c \in \mathcal{C}} \alpha_{c,k,y} = 1 \quad \forall k, \quad (18)$$

where $C_{k,y}$ is total cost in bucket k during year y , and $\alpha_{c,k,y}$ is the corresponding class allocator constructed from cost drivers such as coincident peak demand, non-coincident peak demand, energy usage, number of customers, and voltage level. Traditional COSR typically does not maintain a systematic “new” versus “existing” separation within a class. As a result, incremental upgrades that enter RB_y can propagate into delivery-related buckets via the return and depreciation components embedded in $C_{k,y}$ and then flow into class responsibilities through Eq. (18). When hyperscale-driven upgrades are large relative to the embedded legacy system, these embedded-cost channels can raise delivery charges for customers who did not trigger the upgrades.

This mechanism is not universal: if incremental investments are modest and load growth increases utilization, embedded

costs can be diluted and average rates can fall. The hyperscale context elevates upward-pressure risk when i) capacity is built ahead of realized load (front-loading), ii) the probability of partial build-out, delayed ramp, or relocation is nontrivial, increasing stranding tail risk, and iii) the nominal cost of new T&D and generation equipment has risen well above the embedded cost basis of existing facilities, so marginal additions are recovered at materially higher unit costs.

B. Jurisdictional layers: interconnection, transmission, and retail obligations

Hyperscale data centers often interconnect at transmission voltages (at or above 69 kV). In ISO/RTO regions, engineering assessment and categorization of network upgrades is frequently performed under federal-jurisdictional transmission processes. However, the end-use customer remains a *retail load*: state commissions determine retail cost allocation, tariff applicability, and the class/rate-element mapping through which transmission and distribution costs are recovered.

Transmission cost recovery applies analogous embedded-cost logic at a regional scale. Annualized transmission revenue requirements are allocated using local, zonal, or regional methods intended to reflect cost-causation principles (costs roughly commensurate with burdens caused or benefits received), yet the incidence varies substantially across regions and methodologies. These allocated transmission charges are then flowed through to retail customers via state-jurisdictional class allocation (Eq. (18)) and rate design [81].

A related rate-design issue arises when hyperscale data centers provide measurable services or benefits to the bulk power system. Such services should, in principle, be compensated through direct payments or bill credits when their value can be identified and verified. Historically, industrial retail tariffs have more commonly imposed technical performance requirements—for example, acceptable power-factor ranges—than provided explicit compensation for grid-support capabilities. For hyperscale loads, however, their scale and controllability may make explicit valuation of such services increasingly important. A practical approach is to separate wholesale valuation from retail compensation: the ISO/RTO could quantify the value of qualifying grid services and reflect those values in the wholesale grid costs or credits allocated to LSEs; the portion attributable to a specific hyperscale data center could then be unbundled and passed through by the utility, subject to state commission approval, as a corresponding bill credit or other compensation. Such an arrangement preserves the respective wholesale and retail regulatory roles while allowing data centers to receive the economic value of verifiable services they provide to the grid. Compensation should therefore be based on measured service delivery rather than installed capability. Infrastructure installed principally to serve the facility itself should remain the customer's cost responsibility; socialized recovery or long-term contracting is defensible only when an asset demonstrably avoids or substitutes for grid-side investment that otherwise would have been planned.

C. Generation procurement and when “generation cost shifting” matters

The role of generation in cost shifting depends on retail market structure. In retail-competitive jurisdictions (and some large-customer access regimes), hyperscale customers often secure energy via bilateral contracts such as power purchase agreements (PPAs), making generation costs primarily contractual rather than embedded in a regulated utility revenue requirement. In vertically integrated jurisdictions, utility-backed generation procurement undertaken to serve a hyperscale load can enter embedded cost recovery; if the load ramps below forecast or exits early, a stranded-cost dispute can emerge over whether unrecovered balances remain in RB_y and how they are allocated across classes in subsequent proceedings.

Even in competitive procurement settings, other customers are not fully insulated from system-wide wholesale market effects of large demand growth. Accordingly, many regulatory proceedings focus on long-lived network investments (T&D and network upgrades) that are more directly tied to embedded-cost recovery and class allocation.

D. Regulatory tools to limit cross-subsidies and stranding risk

Because network upgrades are long-lived assets recovered over multi-decade horizons, commissions have increasingly emphasized mechanisms that reduce the probability that unrecovered balances are reassigned to other classes if a data center ramps slowly, reduces load, or exits within a decade.

a) *Up-front cost responsibility (customer-funded upgrades)*: A first lever is to treat certain facilities as customer-specific rather than system-shared. When service requests exceed standard tariff obligations (e.g., dedicated feeders, non-standard facilities, or site-specific reinforcements), utilities may apply contributions in aid of construction (CIAC), higher deposits, and customer-funded line extensions. In engineering terms, this approach assigns specific capital additions to the triggering customer rather than injecting the full I_y into the pooled rate base used for class allocation.

b) *Long-term commitments (minimum bills, contract demand, take-or-pay)*: Commissions have supported multi-year commitments by very large loads that bind cost recovery to a minimum billing level over a horizon aligned with asset recovery. A generic take-or-pay formulation can be expressed at the billing-period level as

$$\text{Bill}_t \geq \underline{\text{Bill}}, \quad \forall t \in \{1, \dots, H\}, \quad (19)$$

where Bill_t is the billed amount in billing period t and $\underline{\text{Bill}}$ is a contractual minimum over commitment horizon H (often 10–20 years). This structure reduces exposure to underutilization and lowers the probability that unrecovered fixed costs are reallocated to other classes in future rate cases.

c) *Cost segregation and nonlinear rate design (separate “buckets” for hyperscale loads)*: A structural approach isolates incremental costs through special tariffs or contracts that effectively create a parallel COSR track (a separate “bucket”) for hyperscale loads rather than pooling upgrades into legacy embedded costs. Nonlinear rate design provides a technical reconciliation between marginal-efficiency objectives

and distributional constraints. A standard tariff formulation is shown below:

$$\text{Bill}_t = F + p_E E_t + p_D D_t, \quad (20)$$

where F is a fixed charge (\$/billing period), E_t is energy (kWh) with unit price p_E (\$/kWh), and D_t is a billing demand determinant (kW, e.g., monthly maximum kW or coincident-peak kW) with unit price p_D (\$/kW). In this framing, p_D (and time-varying demand determinants, when used) can be tuned to reflect capacity-driven marginal costs, while F and/or minimum-bill provisions (Eq. (19)) recover infra-marginal costs and support class equity without distorting marginal operational incentives.

E. “Behind-the-fence” strategies and unresolved allocation questions

Developers increasingly consider behind-the-fence generation (often paired with storage) to reduce interconnection exposure and improve reliability. However, commissions may retain authority over certain overhead-cost allocations and policy obligations (e.g., renewable portfolio standards). Moreover, complete cost isolation is technically and institutionally difficult: network expansions can produce shared benefits (e.g., reduced congestion or improved contingency performance), and stakeholders often dispute what capacity is truly incremental versus what would have been built under baseline planning trajectories. These debates sit at the intersection of cost-causation norms, embedded-cost recovery, and the practical limits of disentangling joint network benefits [81]. These questions become sharper if the interconnection itself evolves from a single metered point into a multi-port interface serving generation, storage, and load simultaneously. Cost responsibility is presently assigned at the level of a customer and a service connection, whereas a multi-port architecture would require attribution at the level of individual ports and bidirectional flows—a granularity that the CIAC and take-or-pay constructs above are not designed to express. A regulatory precedent nonetheless exists in distributed energy resources: a rooftop-solar customer is already simultaneously a generator and a load, and rate design accommodates this not through a new customer class but through metering arrangements, i.e., net metering, or separate production and consumption meters. A multi-port interface generalizes this to a multi-meter system in which each port is metered, and potentially charged or credited, according to the direction and nature of its flows; the same architecture therefore governs both what a port is charged for the service it draws and what it may be credited for a service it verifiably delivers. How such a credit could be arranged across the wholesale and retail layers is set out earlier in this section.

The analogy holds at the level of metering and settlement design rather than compensation rates: the net-metering cost-shift debate is itself a caution that such design choices carry significant allocation consequences, and at transmission scale the closer precedents are co-located generation and station-power rules rather than retail distributed energy resource (DER) tariffs. Designing such multi-meter, port-differentiated

tariffs is future work; we return to this prospect, and to its dependence on the maturity of the enabling technology, in Section VII.

F. Broader rate design transition and implications

These issues arise alongside an ongoing rate design transition. Many residential tariffs remain dominated by volumetric charges, whereas large C&I tariffs more commonly include demand charges and voltage differentiation; performance-based and incentive-oriented mechanisms can further change the timing and incidence of cost recovery [80]. Overall, the prevailing direction is not to prohibit growth, but to require large new loads to i) internalize a larger share of clearly incremental delivery costs, ii) commit to billing determinants over horizons aligned with asset lives, and iii) face tariffs with capacity-reflective marginal components, thereby reducing speculative build risk and limiting bill impacts on residential and small commercial customers while preserving efficient marginal price signals.

Sections III–IV emphasized uncertainty at planning and regulatory horizons; we now shift to *variation* at operational horizons, where sub-second dynamics and minute-scale ramps create new security and stability requirements. Section V examines these reliability implications and the mechanisms through which data centers can either exacerbate or mitigate them.

V. OPERATIONAL SECURITY AND STABILITY UNDER VARIATION

Building on the timescale crosswalk in Section II-C, we focus here on operational reliability under variation: i) sub-second dynamic performance (power quality, ride-through, oscillations, and stability) and ii) seconds-to-minutes ramps that drive reserve deployment, congestion management, and operational flexibility requirements. The high power density of AI compute clusters and the prevalence of power-electronic interfaces (UPS systems, rectifiers/inverters, on-site storage and generation) can introduce grid-visible dynamics that are unfamiliar to operators who historically treated load as largely passive.

A. Sub-Second Dynamics: Transients, Power Quality, Ride-Through, and Oscillatory Stability

The rapid emergence of large data centers has altered demand profiles and influenced the operations of modern power systems. At sub-second time scales, data centers exhibit complex transient behaviors driven by power-electronic circuits and high-speed control loops. It is therefore essential to investigate these transients in detail and examine their role in shaping power quality, fault ride-through capability, and oscillatory stability.

In Table IV, we compare data centers (power-electronics loads) with classical loads and highlight how data centers differ in governing mechanisms, harmonics, nonlinearity, fluctuation, inertia, and modeling characteristics. More details are provided in the subsequent discussions. It is clear that prior models and intuition are no longer adequate to address the challenges of data center integration.

TABLE IV
COMPARISON OF CLASSICAL LOADS AND DATA CENTERS

	Conventional Loads	Data Centers
Mechanism	mostly physics-driven (passive, loose regulation)	control-driven (active, tight regulation)
Harmonics	limited	rich (covering low to high frequencies)
Nonlinearity	limited	highly nonlinear (waveform distortion, saturation)
Fluctuation	slight to moderate	huge and rapid
Inertia	low to high	almost zero
Characteristics	constant impedance ($P \propto V^2$, dominant), constant current ($P \propto V$), constant power (P fixed)	constant power (P fixed, negative impedance)

1) *Transient Dynamics*: Transient dynamics of data centers refer to the milli- to sub-second variations of voltage/current waveforms or power swings in response to external disturbances [82]. As a result, these dynamics tend to trigger sophisticated interactions between data centers and the utility grid. They can damage sensitive electronic facilities (e.g., compute servers, communication networks), trigger protective devices (e.g., fuses, protection relays), or even cause cascading system failures. Transient dynamics are uniquely able to capture real-time operational details of data centers, which are unobservable in other long timescales (steady state). In particular, these details can provide a full picture of nonlinear characteristics and fast-acting responses (e.g., harmonics, current saturation) during faulty conditions. It turns out that a deep understanding of data center transients is the key to mitigating risks and ensure reliable operation.

Specialized models have been developed to capture the transient dynamics. Two common modeling options are root-mean-square (RMS) models and EMT models, and a detailed comparison is provided in Table V. RMS models apply electric phasors, which are vectors representing a sinusoidal component of the fundamental frequency (60/50 Hz), to analyze sinusoidal dynamics. Dq-frame formulations [83] are widely used to transform the sinusoids into algebraic constants. This formulation is effective at capturing slowly varying waveforms under balanced or weakly unbalanced conditions. While computationally efficient, it cannot represent dynamic modes during asymmetrical faults or rapid voltage variations due to the neglect of high-frequency harmonics. On the other hand, EMT models capture the full process of electromagnetic transients, including saturation, harmonics, converter switching, and fast control loops. For instance, in [6], an EMT model was developed for large cryptocurrency mining loads. Another aggregate model combining resistive loads, switch-mode power supply loads, and lighting loads was formulated in [84]. EMT models are essential for tracking data center transients, but industry-grade EMT models for data centers are presently lacking. Other modeling options include frequency-domain models (which rely on local linearization) and averaging models (which focus on low-frequency dynamics) [85].

Modern data centers utilize multiple stages of power con-

TABLE V
COMPARISON OF RMS AND EMT MODELS

	RMS Models	EMT Models
Variables	voltage/current phasors	instantaneous voltages/currents
Modeling	electromechanical oscillations, voltage/frequency variations	electromagnetic transients, switching, harmonics
Computation	efficient, highly scalable	intensive, poorly scalable
Accuracy	moderate precision	highest precision
Use Cases	sinusoidal (quasi) steady-state, three-phase balanced systems	broad, including nonlinear, non-steady-state, unbalanced systems

version from utility feeders (AC) to terminal chips (DC) [86]. In this regard, their transient behaviors are largely driven by hardware known as power-electronic converters (AC-to-DC rectifiers and DC-to-DC choppers). Formally, data center dynamics are governed by two essential aspects of power electronics: topology and control [87]. There is a clear industry trend toward reducing the number of power conversion stages to improve overall efficiency. Novel topologies along with wide bandgap semiconductors allow for higher switching frequencies and near-peak efficiency [88]. Despite rapid advances in hardware innovation, the dynamical performance of these new converters has not been thoroughly explored, and many of them may exhibit different (or troublesome) behaviors in sub-second timescale. Additionally, control designs play an essential role in data center transient dynamics. From a feedback control perspective, the tight regulation of power electronic converters will shape the feedback loops and govern the convergence/evolution of interconnected systems. Many converters in data centers implement “constant power” mode, which isolates compute chips from input voltage fluctuations but introduces negative impedance to the upstream grid (thus degrading stability performance) [85].

2) *Power Quality*: Power quality describes how well the electric power supplies (e.g., voltage/current magnitudes, frequencies, waveforms) stay within a desirable range [89]. It can be regarded as the approximate equilibrium of transient dynamics measured at the point of common coupling. A wide range of standardized metrics is used in current practice, including voltage/frequency deviation (sag/swell depth), total harmonic distortion, and voltage unbalance factor.

Most power-quality issues are attributed to the internal power electronics equipment, iron-saturating devices, and arcing devices. In data centers, the involved nonlinear power conversions and nonlinear loads [90] greatly distort the sinusoidal waveforms and inject rich harmonics into the upstream grid. Well-known symptoms include equipment failure and misoperation, increased energy losses and overheating [91], often resulting in unplanned downtime and reduced system reliability. Power factor correction (PFC) converters, a type of AC-to-DC circuits, become a popular option to enforce the input current waveform sinusoidal and in phase with the input voltage (close to unity power factor) [92]. Another

option is hybrid transformers, which combine a traditional low frequency transformer with a power electronic module for harmonic isolation [93]. The current practices show that, by employing proper mitigation strategies, data centers can achieve energy savings while maintaining the stability, purity, and cost-efficiency of the power supply.

Power quality modeling must account for the above intricate dynamics while allowing rapid validation of performance across diverse topological and control designs. RMS models are effective for analyzing voltage magnitude and unbalance status but they lack the fidelity to capture harmonics and fast dynamics. There are harmonic-based extensions that replace a single-frequency representation with a weighted sum of harmonic components. EMT models provide the highest level of modeling fidelity and track the full details of switching events and harmonics. Furthermore, the current infrastructure has introduced a new power-quality issue of supraharmonics in the 2–150 kHz range [94], typically generated by high-frequency control signals.

3) *Fault Ride-Through*: Fault ride-through describes the ability to remain connected and operational for a short period (e.g., 150 ms) during voltage or frequency disturbances [95]. These disturbances may come from a nearby fault or an uncontrolled fast transient or surge. As defined, fault ride-through is a resilience measure that mandates operational support during external disturbances [96]. Fault ride-through has historically been a requirement for interconnected generators, but recent practices have begun to extend these rules to large loads, including data centers [97].

Data centers face unique physical and operational hurdles during external disturbances because of their intense deployment of power electronics. Specifically, most power-electronic equipment acts as a constant-power load. It increases current draw in response to voltage drops (a positive feedback), which can dangerously strain a weak grid and hinder disturbance rejection. If massive loads trip simultaneously, the abrupt power imbalance can induce sharp voltage and frequency spikes, which greatly stress the grid and increase the risk of cascading failures [98]. Physically, power-electronic converters are inherently sensitive to voltage variations, and without specific control, they may disconnect compute loads during voltage sags or swells. Many power electronic circuits do not equip enough energy capacity reserve to enable reactive power injection and prevent the rapid voltage collapse [99]. Backup power and storage are found useful in extending the fault ride-through performance [89]. There are challenges around the unpredictable transient behaviors for power electronics as well, and rigorous site-specific tests are often needed to manage the recovery process and grid reconnection dynamics.

The power industry is shifting toward active management of data centers for system supports. It becomes increasingly common for grid codes to require system participants to inject reactive power during faults [100]. Grid operators are also advocating for better site-specific load modeling and transparency [101]. There is also a growing trend for data centers to evolve into large controllable loads instead of passive consumers.

4) *Oscillatory Stability*: An electric grid is oscillatory stable if it can maintain synchronization and acceptable operating conditions. To achieve this, it must adequately damp all oscillations in the system triggered by arbitrary disturbances. The integration of data centers brings new stability risks to the grid, primarily due to the wide timescale dynamics and frequency-coupling effects of converters [102]. Oscillations can arise from interactions between multiple data centers or between data centers and the electric grid. The negative impedance and weak grid conditions are two major contributors to instability. In recent years, the power swings associated with AI training workloads have raised new concerns regarding high-magnitude, periodic power fluctuations and their impacts on system stability and power quality [103], [104].

Data centers and other converter-based resources have motivated a revisit of the existing power system stability categories. The task force [105] has identified two new stability categories that exhibit a close connection to power-electronics loads and generators. The first category, converter-driven stability, covers instabilities driven by the cross-coupling between converters and the grid electromagnetic transients. The periodic, structured power fluctuations driven by repetitive training patterns may serve as a forcing signal to trigger such stability issues [106]. The second category is resonance stability (harmonic stability), and it focuses on the waveform distortions both above and below the fundamental frequency. In data centers, the harmonic components might be generated and poorly damped at certain frequencies.

Integrated analytical methods are required to evaluate oscillatory stability in the presence of data centers. The common tools include small-signal and eigenvalue analysis, impedance-based methods, and EMT simulations. Small-signal analysis can provide rich information on oscillation modes and participation factors of system parameters. In [107], [108], single-phase power supply units were rigorously modeled to capture DC bus dynamics and the coupled current responses. The new model then supported a more accurate assessment of system stability. A well-known drawback of small-signal analysis is its reliance on linearized dynamics, and this method is often limited to certain operating conditions. Impedance-based methods model different components by their frequency-dependent impedance and investigate the interaction between source and load impedance instead of the full state spaces. While this method is flexible and allows convenient validation through measurements, it still depends on typical operating points and becomes inaccurate during large disturbances. Reference [7] used a neural network to model impedance under a large-signal setting. The data-driven model provided an effective solution to study the adverse interactions between data centers and the grid. EMT simulation is the most precise and time-consuming of all the options. It is particularly suited for power-electronics-rich systems and weak-grid conditions when other methods fail to capture the true dynamics.

A further difficulty is scale. The screening methods above are typically applied to a small number of interacting terminals, whereas a converter-dense facility with many independently controlled ports—some regulating a DC bus, some tracking maximum power, some exchanging reactive power

with the grid—constitutes a substantially higher-dimensional coupled system. Whether impedance-based screening, including its data-driven variants, remains both tractable and appropriately conservative at that scale is an open question; hierarchical screening, aggregated port-level impedance models, and model-order reduction are plausible directions, but validated practice does not yet exist. The same limitation applies across the interface as well as within it: neither the facility nor the upstream system can be screened in isolation, and validated interaction models spanning the two do not yet exist. A practical decomposition nonetheless follows from the interface itself: for grid-interaction screening, what must be characterized is the facility’s aggregate impedance seen at the point of interconnection, with internal stability certified separately by the facility. This does not dissolve the internal problem, but it assigns each half to the party able to observe it and makes the interconnection-facing quantity something that can be specified and verified.

B. Seconds-to-Minutes Variation: Ramping Products, Reserves, and Operating Flexibility

At operational timescales, a dominant concern is the net-load ramp (and rebound) imposed at the point of interconnection by coordinated workload changes, inference bursts, and potentially correlated behavior across multiple facilities (including geographic failover). These ramps translate directly into the deployment of regulation and ramping reserves, and they can exacerbate congestion in regions where large loads cluster [109]. In practice, operators manage this through look-ahead dispatch, explicit ramp procurement, or both.

Advanced power-electronic interfaces are sometimes expected to remove this requirement altogether by absorbing fluctuations before they reach the grid. That expectation should be bounded. The buffering capability of a solid-state interface is limited by its current headroom—on the order of $1.5\times$ rated, against roughly $25\times$ for two seconds in a line-frequency transformer—and by millisecond-scale device thermal constants, so it can reshape the spectral signature seen upstream but cannot absorb the energy-bearing ramps discussed here [110]. Any masking that does occur additionally creates a measurement problem, because the ramp observed at the meter no longer reveals the stress borne by the interface itself; this argues for qualification and telemetry requirements that observe internal state rather than terminal flow alone. The deeper point is that the two domains are not separable. Dynamics inside the facility and dynamics on the upstream system are coupled through the interface, and the coupling runs in both directions: the interface reshapes what the grid observes, while conditions at the point of interconnection determine how much the interface can absorb at any instant. Characterizing either side alone is therefore insufficient, and transient behavior at the boundary has to be coordinated by design rather than assumed away. This is also why the intrinsic characterization in Section II-C3 remains grid-relevant rather than internal: it is the source description that any interaction model across the interface requires.

a) Reserve-relevant ramps and congestion exposure (seconds–60 min): Coordinated job starts/stops, checkpointing phases, ramp-bounded controls, and post-curtailement rebound can produce ramps that must be balanced by committed ramping capability and operating reserves [111]. Data centers can worsen this requirement when ramps are uncoordinated, or partially mitigate it when controllability is explicit through workload orchestration and on-site buffers (UPS/BESS energy shifting, temporary throttling, or limited deferral) [112].

To illustrate the operational relevance, consider CAISO’s Flexible Ramping Product (FRP), which procures upward and downward ramping capability to cover the forecast net-load change *plus* a probabilistic uncertainty margin over rolling 5- and 15-minute intervals [113], [114]. During steep solar ramp periods, CAISO’s duck-curve analysis identifies a system requirement to supply an additional 13,000 MW within approximately three hours as solar output declines [115], corresponding to an average sustained ramp of roughly 70 MW/min across that window. Similarly, ERCOT’s Contingency Reserve Service (ECRS) and Responsive Reserve Service (RRS) are sized based on the largest credible contingency [116]. Production telemetry from hyperscale training clusters indicates that, in extreme cases, aggregate power consumption can oscillate by tens of megawatts within a single data center, with the amplitude growing in proportion to the number of synchronized GPUs and their peak per-GPU power [117]. Swings of this magnitude are not yet comparable to the largest credible contingency, but they are large enough that the associated reserve and ramping burden ceases to be negligible as facility scale increases. Operational descriptors therefore include MW/min ramp rates, reserve sufficiency margins, scarcity price adders, binding-constraint indicators, and response-time requirements for demand response participation [118]. A key question is whether the procured “ramp capability” can actually be delivered as a feasible multi-interval trajectory when the system is tight.

b) Diurnal/event-driven structure and sustained loading (1–24 h): Across hours, inference load can track human activity and exogenous events, imprinting diurnal and event-driven structure that aligns with system stress periods and amplifies congestion exposure even when annual energy is unchanged [119]. Training is more schedulable and can be shifted toward low-price/low-carbon intervals, but rebound after curtailment or deferred execution can re-concentrate demand and reshape net-load and congestion patterns [120]–[122]. Sustained high-utilization loading also couples to thermal limits and cooling-system inertia. Unlike traditional industrial loads with diurnal cool-down periods, AI training runs can persist at near-peak power for days or weeks, denying transformers the thermal relief typically assumed in IEEE/IEC loading guides and accelerating insulation degradation [123], [124]. The coupling to cooling-system inertia, where thermal storage or chiller scheduling can buffer short-term thermal transients, offers a partial mitigation pathway [125].

c) Mechanisms for demand response and ancillary services.: Practical pathways for data centers to provide demand response and ancillary services include workload shifting/deferral, controlled throttling (e.g., GPU frequency scal-

ing), and leveraging UPS systems (and co-located BESS where present) as short-duration buffers to shape ramps and reduce reserve deployment [45], [126]. Recent demonstrations show that coordinated UPS discharge and workload modulation can deliver frequency regulation response within seconds, potentially qualifying data centers for fast-response reserve products [112]. Because flexibility can be time- and state-dependent (e.g., limited by job deadlines, thermal headroom, or UPS state of charge), operational programs must explicitly represent response speed, duration, and rebound characteristics rather than treating flexibility as a static MW quantity. These demand-response mechanisms are moving from concept to field validation: DOE’s Data Center Load Flexibility workshop convened operators, utilities, and national laboratories to scope near-term demand-response opportunities [31], and EPRI’s DCFlex initiative is now running multi-year demonstrations of demand response and workload/load shifting at hyperscale facilities across a planned portfolio of roughly ten sites in the U.S. and Europe [127].

d) *Market design and qualification: pricing response characteristics with verifiable performance.*: Whether existing reserve products correctly price the response characteristics that matter for data-center flexibility (response time, ramp rate, sustainable duration, and rebound effects) remains an open question. FRP-style products compensate resources for maintaining headroom rather than for energy throughput, aligning well with data center capabilities where the goal is to avoid rather than *deliver* rapid load swings [113]. Recent comparisons of look-ahead dispatch and ramp-product implementations highlight two practical gaps: rolling-window operation can drift into different end states, and ramp awards may not guarantee intertemporal deliverability unless additional constraints are enforced; proposed remedies include enforcement terms (e.g., ramp-increment/rolling-difference constraints) and, in some systems, a mid-duration ramp product to cover extended ramps under stress [128]. However, qualification requirements (telemetry granularity, baseline methodologies, performance scoring) must ensure that offered flexibility is measurable and verifiable while remaining compatible with data center service-level agreements [129]. In practice, qualification rules should align incentives with system needs (e.g., fast response during scarcity hours, congestion-aware delivery) and reduce the risk that nominal flexibility fails to materialize under tight operating conditions [130]. As data center loads grow to represent a significant fraction of regional demand, market designs may need to evolve toward dynamic, location- and time-differentiated reserve requirements that explicitly account for the load-side contribution to system ramping needs [131].

VI. AFFORDABILITY UNDER VARIATION: GRID REINFORCEMENT AND NEW FLEXIBILITY REQUIREMENTS

Variation has an affordability dimension because it changes the type and quantity of assets needed to meet reliability standards. A grid built to serve a high-load-factor but slowly varying demand profile may be inadequate or expensive to retrofit when faced with large, fast ramps, voltage-sensitive behavior, and concentrated power-electronic loads. As discussed

in Sec. V-B, these operational choices shape what “counts” as usable flexibility; that, in turn, changes which investments are cost-effective (e.g., storage duration, monitoring, and power-conditioning upgrades). This section identifies the incremental infrastructure and operational capabilities that may be required, from substation and transmission reinforcements to flexibility resources such as energy storage and dynamic reactive power. We then connect these engineering requirements to cost assessment and allocation, emphasizing how tariff and contracting choices determine whether the least-cost portfolio is ultimately built.

A. Network and Interconnection Upgrades for High-Density, Fast-Ramping Loads

AI data centers often arrive as large step changes at specific substations, driving upgrades in transformers, breakers, protection, feeders, and upstream transmission to manage thermal limits and contingency flows. The integration of hyperscale AI facilities introduces a fundamental discontinuity in grid planning due to extreme spatial load density, often exceeding 50 kW per rack for GPU-intensive clusters and reaching approximately 120 kW for a fully populated NVIDIA GB200 NVL72 rack [132], [133], and the rapid temporal volatility of computational workloads.

These characteristics necessitate significant upgrades to physical infrastructure. As noted in Section V-B, sustained training can deny transformers the thermal relief that loading guides assume, accelerating insulation ageing [134]. Inference workloads additionally create “bursty” demand profiles that can fluctuate by tens of megawatts per minute at individual facilities, inducing rapid thermal cycling and mechanical stress on substation equipment. Consequently, utilities must deploy advanced thermal monitoring and dynamic rating systems to maximize the utilization of existing transformer capacity while avoiding accelerated aging. On the facility side, emerging wide-bandgap power electronics (GaN, SiC) are enabling more efficient rectification and power delivery, potentially reducing upstream infrastructure stress [135].

Because these upgrades are capital-intensive and can have multi-year lead times (currently up to four years to secure large power transformers, roughly double 2021 levels [136]), they raise option-value questions: should infrastructure be built ahead of demonstrated need, or incrementally as projects materialize? The traditional “just-in-time” investment model risks creating speculative queues and delaying energization, forcing developers toward carbon-intensive bridging solutions such as backup diesel generation. Conversely, anticipatory investment reduces lead times but increases the risk of stranded assets if projected loads fail to materialize. Recent studies suggest that coordinating transmission planning with data center siting can reduce total system investment costs by over 12% compared to uncoordinated expansion [137].

To mitigate these risks, modular interconnection designs and “Connect-and-Manage” strategies are emerging as viable alternatives. Drawing on approaches pioneered in the UK’s regulatory framework [138], similar concepts are being explored in U.S. markets: data centers may connect prior to the

completion of deep network reinforcements by agreeing to “non-firm” access or flexible operational limits. For instance, a facility might be granted early energization with a contractual obligation to curtail load during specific grid contingencies. This approach trades capital investment (new lines) for operational complexity (active curtailment management), requiring robust software orchestration to ensure reliability margins are maintained without compromising the stability of the sub-transmission network [139]. It also increases the risk of cost shifting to other customers or of stranded assets when cost shifting is constrained by regulators.

B. Flexibility and Power-Quality Investments: Storage, Grid-Forming Controls, and Filtering

Mitigating the variability introduced by AI data centers is best framed as a portfolio problem spanning three complementary resource classes: energy storage, demand-side flexibility (including demand response and workload shifting), and grid-forming control. This framing is increasingly reflected in guidance from international agencies and industry consortia. The IEA’s *Energy and AI* report identifies flexible data-center operation and co-located storage as central levers for reconciling data-center growth with grid reliability, cost, and decarbonization goals [140]. In parallel, EPRI’s DCFlex initiative, launched in 2024 and now spanning grid operators (including ERCOT and PJM), utilities, and hyperscalers (Google, Meta, Microsoft, NVIDIA), is running multi-year field demonstrations of storage, demand response, and load-shifting strategies across a planned portfolio of roughly ten flexibility hubs in the U.S. and Europe, explicitly aimed at transitioning data-center backup power and on-site flexibility into dependable grid assets [31], [127]. The remainder of this subsection details the engineering basis for each mitigation pathway.

1) *Storage and Grid-Forming Control*: Energy storage, and batteries in particular, are key resources for enhancing data center flexibility and power quality [141]. They can reshape facility demand profiles and mitigate rapid fluctuations; the appropriate design depends on what the system operator compensates and can verify. Fast, short ramps call primarily for power capability, while longer ramps and rebound management require sufficient energy duration and appropriate controls (Sec. V-B). Advanced control of on-site batteries can provide inertia-like behavior that supports grid frequency stability and helps maintain voltage stability during disturbances. Storage systems can also mitigate voltage sags, harmonic distortion, and flicker caused by rapidly fluctuating AI training traces. Storage capacity should be selected through a careful trade-off between economic efficiency and the required level of system reliability. Scenario analysis and uncertainty-aware planning can evaluate the benefits of additional storage deployment and compliance with grid requirements for mandatory transient-performance criteria.

Hybrid energy storage is a powerful strategy for combining technologies to exploit complementary performance and cost benefits [142]. Leveraging diverse storage devices (typically supercapacitors and batteries) can provide high power density

and high energy density simultaneously. Hierarchical scheduling of these resources can improve performance, extend operational lifetime, and reduce premature failures [143].

Grid-forming control is a novel mechanism that transforms energy storage from a passive backup into an active, grid-stabilizing asset [144]. Under grid-forming control, batteries act as independent voltage sources and regulate voltage and frequency. They can provide inertia-like response and fault ride-through support to the broader system. For instance, [145] found that parallel grid-forming batteries can help mitigate voltage sags and flicker, alleviating power-quality concerns. On-site UPS batteries have also been modified to implement grid-forming control that supports fault isolation and protection [146]. Grid-forming batteries are increasingly being deployed and are valued for providing voltage/frequency control while enhancing power quality and grid resilience [147]. To ensure that grid-forming storage from different vendors can be qualified and interconnected reliably at scale, the U.S. Department of Energy-supported UNIFI Consortium has published vendor-agnostic technical specifications defining system- and inverter-level requirements for grid-forming resources, providing a standards basis for treating grid-forming storage as a dependable mitigation asset for large, fast-ramping loads such as AI data centers [148].

2) *Power Quality and Filtering*: Fast variation can necessitate incremental investments in flexibility and power quality: short-duration storage for ramps and frequency response; dynamic reactive power devices (e.g., static synchronous compensators (STATCOMs) or synchronous condensers) for voltage support; and harmonic filtering and monitoring to manage interactions between power-electronic equipment and the grid.

The interaction between the grid and the nonlinear switch-mode power supplies (SMPS) of thousands of GPU servers creates a complex power-quality (PQ) environment. These loads inject harmonic currents, particularly triplen harmonics (3rd, 9th) and characteristic 5th/7th orders. These injections can cause neutral conductor overheating and voltage distortion at the Point of Common Coupling (PCC) [149]. Unlike passive loads, AI clusters require active mitigation strategies.

- **Active Harmonic Filtering (AHF)**: Due to the dynamic nature of AI workloads, static passive filters are often insufficient and may create resonance. Active filters, while more capital intensive, dynamically inject counter-currents to cancel harmonic distortion across a wide spectrum, helping ensure compliance with the IEEE 519 standard.
- **Dynamic Reactive Power Support**: The rapid ramp rates of AI clusters can cause severe voltage sags. Traditional capacitor banks are too slow to respond (> 1 second), and their output drops with the square of voltage ($Q \propto V^2$). Advanced STATCOMs can provide millisecond-level voltage support and, when paired with grid-forming controls, deliver virtual inertia to stabilize frequency during large load steps [150].

A key boundary decision is whether mitigation is provided behind the meter (on-site storage and power conditioning) or on the system side (utility-installed devices). Distributed,

behind-the-meter (BTM) mitigation adheres strictly to cost-causation principles but may result in capital inefficiency due to over-provisioning of redundancy at each site. Conversely, centralized utility-side mitigation benefits from economies of scale but raises regulatory challenges regarding cost socialization.

The comparison above shows how cost allocation and tariff structures can either encourage efficient mitigation or unintentionally socialize avoidable costs. Modern “Large Load Tariffs” are increasingly incorporating strict power factor penalties, minimum demand charges (e.g., 85–90% take-or-pay clauses), and CIAC fees to ring-fence these costs [151]. Furthermore, innovative tariffs that incentivize flexibility can transform these loads from liabilities into grid assets, reducing the need for expensive peaking generation. However, discrete engineering upgrades and tariff adjustments alone are insufficient; aligning these physical and financial layers necessitates the integrated techno-economic approach explored in Section VII.

VII. TOWARD A TECHNO-ECONOMIC POLICY FRAMEWORK

A. A Broader Structural Change is Coming

Although this paper is focused on engineering and economic aspects, it is important to emphasize that the future trajectory of the AI–power grid nexus remains highly uncertain. This uncertainty reflects not only technical factors, such as rapid evolution in AI architectures and efficiency; but also broader geopolitical dynamics, industrial policy, and the pace of progress in off-grid or hybrid generation resources. We hypothesize that regions with stronger, more resilient, and more adaptable power grid capabilities are more likely to attract sustained, long-term investment in AI data centers. Relatedly, regions that are able to co-design digital infrastructure with inter-temporal flexibility aligned to local grid characteristics are likely to be better positioned in the emerging global competition for AI capacity.

From an education and workforce development perspective, these dynamics point to a broader structural shift. We hypothesize that AI infrastructure will emerge as a major domain of convergence for electrical engineering, computer science, and economics. This convergence creates an opportunity to develop high-impact curricula that go beyond a narrow focus on AI algorithms or data analysis; instead, we must train students to reason across the full infrastructure stack, spanning power systems, computing architecture, markets, and policy. This underpins the development of advanced AI and, potentially, artificial general intelligence.

B. Potential Solution: PowerAgent-X

Large AI data centers provide a timely and consequential use case for illustrating the value of an agentic, systems-level perspective on the AI–power grid nexus. They are among the fastest-growing sources of electricity demand, with significant implications for grid capacity, reliability, emissions, and governance. At the same time, they possess atypical operational flexibility that makes them natural candidates for coordinated,

grid-aware operation [152]. In regions such as PJM, where large-load interconnection has become a focal reliability and policy challenge, the binding constraint is often not the absence of engineering options, but the lack of trusted, scenario-based analysis capable of evaluating complex, human-in-the-loop pathways under deep uncertainty, including contested forecasts of future data center demand and the system impacts of gigawatt-scale load additions.

Building on recent advances in agentic AI and the PowerAgent community [153], we propose *PowerAgent-X*: an agentic simulation framework designed to study grid planning and interconnection decision-making under data-center-driven uncertainty. The key departure from conventional planning models is that coordination failures arise endogenously. Instead of assuming a single planner with perfect foresight, *PowerAgent-X* represents utilities, system operators, regulators, communities, and data-center developers as heterogeneous agents who learn over time, operate under institutional procedures, and may delay, disagree, or act strategically. This matters for data-center growth because uncertainty is *milestone-driven* and geographically clustered: small delays in queue processing, transmission upgrades, or demand coordination can redirect interconnection outcomes, shift congestion patterns, and trigger costly stopgap measures.

As shown in Fig. 7, *PowerAgent-X* couples three layers through explicit interfaces so that physical constraints shape incentives, incentives shape behavior, and governance determines whether coordination occurs in time. The bottom layer, *PowerAgent-Grid*, provides an engineering-faithful feasibility backbone that translates candidate interconnection locations, timelines, and operating profiles into congestion, deliverability constraints, reliability risk, and emissions outcomes using trusted grid tools in an agentic workflow. The middle layer, *PowerAgent-Market*, models how participants respond to rules and incentives through queue entry strategies, contracting, and flexibility commitments; it also represents AI-aware data-center control (e.g., workload shifting and temporal load modulation) that reacts to prices, reliability signals, and carbon intensity. The top layer, *PowerAgent-Policy*, simulates regulatory and stakeholder processes that frequently dominate timelines, including negotiations over cost allocation, siting, reliability requirements, and community acceptance under asymmetric information.

A practical advantage of this architecture is that it supports mechanism verification for planning under uncertainty. For example, planners can evaluate how alternative interconnection rules, flexibility products, or cost-allocation policies change i) queue outcomes and upgrade timing, ii) the credibility and timing of data-center flexibility, and iii) downstream reliability and emissions impacts. The agentic market layer is particularly useful because it makes high-dimensional preferences and constraints elicitable and simulatable at scale: by converting concise natural-language descriptions of risk tolerance, operational constraints, and contractual priorities into structured experimental economics model [154], the framework can generate auditable distributions of bids, commitments, and strategic responses that are difficult to obtain from purely optimization-based models.

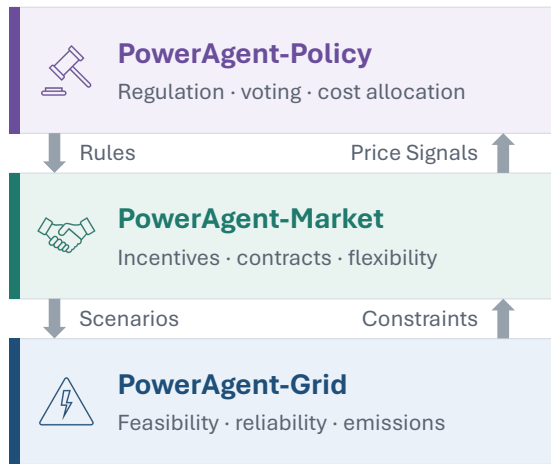


Fig. 7. The PowerAgent-X framework and the roles of its three coupled layers. PowerAgent-Grid evaluates physical feasibility, reliability, and emissions; PowerAgent-Market represents participant behavior, contracting, and flexibility decisions; and PowerAgent-Policy represents governance, cost allocation, and siting processes. Bidirectional interfaces allow physical constraints to shape incentives and policy while market and policy decisions alter the scenarios evaluated by the grid layer.

PowerAgent-X therefore provides a unified environment for diagnosing and mitigating the planning challenges posed by data centers: it can quantify the reliability and cost consequences of delayed transmission and interconnection, identify incentive designs that elicit verifiable flexibility, and surface policy bottlenecks that determine whether technically feasible solutions are implemented on time. This enables transparent comparisons between uncoordinated growth and coordinated, AI-aware pathways in which regulatory enablement and performance-based commitments reduce congestion, accelerate deployment, and lower long-run system costs.

C. Potential Solution: Advanced Grid Integration

New developments are emerging in the grid integration technologies for future AI data centers. An exciting trend is that the interface between the grid and data centers will evolve from a single connection point into a sub-network.

Fig. 8 shows an illustrative layout for the new kind of interfaces. In the future grid integration, there will be an interface expansion to form local microgrids [155] or virtual power plants (VPPs) [156] with internal networks and a number of distributed energy resources. Bidirectional and fluctuating energy flows will alter the coupling between the grid and data centers, and ultimately change whole-system performance.

Fig. 8 also highlights the enabling technology, namely solid-state transformers (SSTs) [157]. By design, SSTs integrate power electronics with high-frequency transformers, which can offer greater operational flexibility, multi-port integration of heterogeneous resources, and—in applications where they also displace rectification stages, as in DC-distribution data centers—improved end-to-end conversion efficiency and a more compact footprint. Ongoing advances in SSTs are therefore closely tied to the feasibility of the microgrid/VPP interface described above.

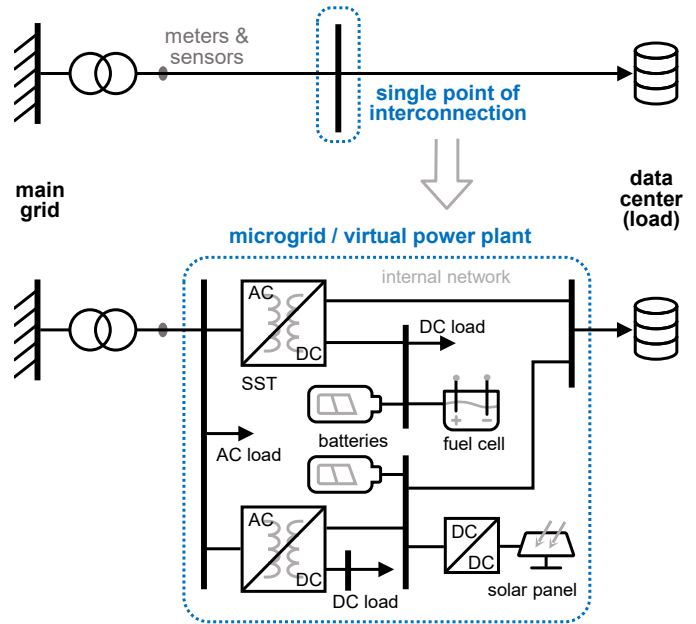


Fig. 8. Conventional and advanced grid-data-center interfaces. The upper diagram shows a data center connected to the grid through a single metered point of interconnection. The lower diagram expands that boundary into an internal microgrid/virtual-power-plant sub-network in which solid-state-transformer-based multi-port converters can coordinate AC and DC loads, batteries, a fuel cell, and solar generation. The architecture permits bidirectional flows among the internal ports while retaining meters and sensors at the external grid boundary.

These prospects must, however, be weighed against the present cost and maturity of the technology, which bound how quickly such interfaces can be expected at scale. On a per-kVA basis SSTs remain considerably more expensive than line-frequency transformers (LFTs): matched 7.2 kV medium-voltage prototypes place the SST at 86.2 \$/kVA against 25 \$/kVA for an LFT, a factor of 3.4 [158], while assessments of AC-AC distribution duty in the 1000 kVA class report a premium of at least $5\times$ together with roughly $3\times$ higher losses [159]. A component-level comparison is nonetheless incomplete, because an SST also displaces rectification, filtering, and reactive-power compensation equipment; on a functional-equivalence basis, and particularly in DC-dominant data centers, the effective premium is materially smaller than the headline multiple suggests. The non-cost barriers are arguably more binding. An LFT can deliver on the order of $25\times$ rated current for two seconds, whereas an SST is typically limited to about $1.5\times$ with millisecond-scale thermal constants at the die, so overcurrent-based protection coordination does not carry over unchanged [110]. Existing test codes cannot presently assign a basic insulation level to a medium-voltage SST, which by itself precludes drop-in substitution within an insulation-coordinated distribution system [160], and standards for medium-frequency insulation and partial-discharge measurement remain under development [161]. Deployment reflects these constraints: SSTs are not in routine service on utility distribution feeders, with field experience concentrated in railway traction, a small number of medium-voltage demonstrations, and pilot fast-charging installations [162], [163]. We therefore treat SST-

based interfaces as a *conditional long-term option* rather than a near-term substitute for conventional interconnection: Fig. 8 indicates a direction of travel, while the timing depends on cost trajectories, protection practice, and standardization. This framing matters for the regulatory questions raised below, since port-level cost attribution and premium products for grid-forming service become pressing only as, and to the extent that, the underlying hardware is actually deployed.

In addition, the new interfaces require an upgrade of the energy management process and toolboxes. There should be a new layer of microgrid/VPP control between the existing grid-side and data center management systems. Note that this extra layer is occasionally viewed as an energy community in the literature. The new energy management process and toolboxes will continue to emerge and represent a direction for future development. This trend is driven by the higher penetration of distributed energy resources and, conditionally, by the anticipated maturation of SST technology. There will be significant challenges and opportunities in the future, including complex operating status, flexibility resources, market design, policy making, and practical implementation.

1) *Complex Operating Status and Flexibility*: The increased number of power flow paths and system components has created a strong need for coordinated operation. As introduced above, a new coordination layer (centralized or distributed) must integrate advanced controllers/schedulers to enhance flexibility resources and mitigate operational risks of power fluctuations and harmonics. Through these coordination layers, many data centers can shift their role from passive loads to proactive grid assets.

Flexibility potential can be unlocked through SSTs [157] because they allow full control of power flow directions and multi-port operations, and they are able to smooth, and partially whiten, the fluctuations of data center demand seen upstream—though only within the current and thermal limits noted in Section V-B, so that the underlying variability is reshaped rather than eliminated. Flexibility can also come from other technologies, e.g., grid-forming batteries [147]. These facilities provide frequency and voltage regulation services for enhancing the whole-system stability margins.

2) *Market and Policy Making*: Unlike the single-point interconnection, the internal sub-network is expected to operate under complex status, which increases the burden for measurement and allocation of energy costs. This issue becomes even more pronounced within a distributed decision-making framework. There is a clear desire to redesign the settlement mechanisms to efficiently coordinate between the external grid operators and the internal energy consumers.

A two-layer settlement structure organizes this. In the upper layer the facility settles with the utility as a single entity, and prevailing practice retains a single metered point of interconnection even when the internal network is extensive, precisely because measurement and billing are tractable there; the cost-responsibility instruments of Section IV continue to apply unchanged at that boundary. What the multi-port architecture changes is the lower layer, where energy-flow tracking, loss allocation, and port-level attribution arise, and for which bilateral contracting and network-usage-based allocation are

the established candidates.

A notable trend is the emergence of new market operators and business models. The new operators may serve as a local balancing agent or a load aggregator. They have strong incentives to innovate their business models to serve different energy stakeholders, e.g., community-based markets and peer-to-peer trading [164].

Meanwhile, policy makers need new analysis and design tools to regulate these new markets and business practices. Typical issues include how to incentivize demand response participation, how to ensure fair revenue allocation among stakeholders, and how to leverage subsidies to encourage investment in new flexibility and grid-supporting technologies.

3) *Towards Practical Implementation*: Sufficient support for practical implementation is of great importance. For instance, a major concern associated with the new grid integration pathway is around the potential risks and uncertain benefits. In this respect, PowerAgent-X may provide a forward-looking assessment of these new issues by adapting its grid-market-policy modeling capacity to local microgrids/VPPs. In addition, industrial standardization and workforce training need to align with future trends of technological development and engineering practices.

Ultimately, the energy community must maintain an open mindset and actively explore diverse technologies and business models to navigate a future of both challenges and opportunities. This mindset will help us identify the optimal pathway that well serves AI and meets energy sustainability goals. There is little doubt that technological breakthroughs alone will not determine future success. Instead, a holistic approach that integrates new technologies, new business models, and new policy interventions will become essential for shaping the pathway forward.

VIII. CONCLUSION

Large data centers are emerging as a major and qualitatively distinct source of electricity demand. Their integration into power systems challenges conventional planning, operational, and regulatory frameworks because large facilities arrive as spatially concentrated point loads with uncertain realization and exhibit workload-driven variability across multiple time scales. This paper organized these challenges around two defining load properties—*forecast uncertainty* and *inter-temporal variation*—and examined how they propagate into reliability and affordability outcomes.

Using a four-quadrant framework linking reliability and cost to uncertainty and variation, we showed that AI data center growth affects power systems across planning and operational horizons. At longer time scales, uncertainty in size, timing, location, and firmness complicates resource adequacy assessment, generation and transmission expansion, and cost allocation under cost-of-service regulation and capacity market designs. At shorter time scales, coordinated and power-electronic-driven variability introduces new requirements for operational security, stability, and ramping reserves. These effects are amplified by the spatial concentration of data centers, which limits diversification and increases the local consequences of forecast and operational errors.

A key insight from this analysis is that engineering, economic, and institutional dimensions are tightly coupled. Reliability risks are shaped not only by physical constraints but also by tariff structures, interconnection rules, and contracting mechanisms that determine how uncertainty is revealed and how commitments are enforced. Similarly, affordability outcomes depend on technical choices regarding mitigation location, flexibility deployment, and the coordination of generation, transmission, and load-side investments. Treating these dimensions in isolation can lead to inefficient infrastructure build-out, misaligned incentives, or unintended cost shifts.

The paper also highlighted that AI data centers are not inherently inflexible loads. Empirical characterization of workload-driven power variability suggests that, particularly for training-dominated facilities, a substantial share of variation occurs at time scales compatible with demand response, reserve provision, and congestion management. Realizing this potential, however, requires explicit modeling, measurement, and verification of response characteristics—including speed, duration, rebound behavior, and deliverability—rather than relying on simplified notions of interruptibility.

Several research directions follow from this work. First, resource adequacy and planning models would benefit from improved representations of load-side uncertainty, including milestone-dependent realization, correlated siting decisions, and conditional flexibility availability. Second, further development and validation of dynamic models for large power-electronic loads are needed to assess stability and interaction risks as data center penetration increases. Third, market and tariff designs should be evaluated with explicit attention to how forecast error and spatial concentration affect procurement outcomes, investment incentives, and cost allocation. Finally, integrated simulation frameworks that combine physical constraints with institutional behavior offer a promising approach for studying coordination challenges under deep uncertainty.

More broadly, the growth of AI data centers illustrates a shift toward electricity demand that is increasingly shaped by digital systems, software-driven control, and rapid technological change. Addressing this shift will require analytical approaches that integrate power system engineering, economics, and policy, and that can adapt as both computational workloads and power systems continue to evolve.

DISCLAIMER

The views expressed in this paper are those of the authors and do not reflect the views of PJM Interconnection, L.L.C. or its Board of Managers, of which Le Xie is a member.

REFERENCES

- [1] Q. Zhang, S. Chung, and L. Xie, "From crypto mining to artificial intelligence training: Economic drivers and grid impacts of the data center transition," *IEEE Energy Sustainability Magazine*, vol. 2, no. 3, pp. 56–64, 2026.
- [2] W. Wang, A. Abdolrashidi, N. Yu, and D. Wong, "Frequency regulation service provision in data center with computational flexibility," *Applied Energy*, vol. 251, p. 113304, 2019.
- [3] *IEEE Standard for Harmonic Control in Electric Power Systems*, IEEE Std. 519-2022, 2022. [Online]. Available: <https://standards.ieee.org/ieee/519/10677/>
- [4] J. Aljbour and T. Wilson, "Powering intelligence: Analyzing artificial intelligence and data center energy consumption," Electric Power Research Institute, Palo Alto, CA, Tech. Rep., 2024. [Online]. Available: <https://www.epri.com/research/products/3002028905>
- [5] X. Chen, X. Wang, A. Colacelli, M. Lee, and L. Xie, "Electricity demand and grid impacts of AI data centers: Challenges and prospects," *arXiv*, no. arXiv:2509.07218, 2025. [Online]. Available: <https://arxiv.org/abs/2509.07218>
- [6] A. Samanta, S. Majumder, H. Ibrahim, P. Enjeti, and L. Xie, "Electromagnetic transient model of cryptocurrency mining loads for low-voltage ride through assessment in transmission grids," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*, 2024, pp. 1–5.
- [7] G. Ruan, M. D. Ilić, and L. Xie, "Data center control against sub-synchronous resonance: A data-driven approach," in *CIGRE Grid of the Future Symposium*, 2025, preprint: arXiv:2511.14141.
- [8] A. Jena, N. Li, and L. Xie, "LILAD: Learning in-context Lyapunov-stable adaptive dynamics models," in *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI-26)*. AAAI Press, 2026. [Online]. Available: <https://github.com/amitjena1992/LILAD>
- [9] D. Kim, L. Dong, and L. Xie, "Flexibility-aware framework for efficient planner-initiated siting of data center," *Nature Communications*, vol. 17, no. 1, p. 6512, 2026.
- [10] National Laboratory of the Rockies, "The dawn of electric trucking calls for high-power charging," NLR News Feature, Jan. 2025. [Online]. Available: <https://www.nlr.gov/news/feature/2025/the-dawn-of-electric-trucking-calls-for-high-power-charging>
- [11] International Energy Agency, "Global hydrogen review 2024," IEA, Paris, Oct. 2024. [Online]. Available: <https://www.iea.org/reports/global-hydrogen-review-2024>
- [12] C. McGeedy, H. Post, and J. Nakano, "Energy considerations at the dawn of strategic manufacturing," Center for Strategic and International Studies (CSIS) Commentary, Apr. 2024. [Online]. Available: <https://www.csis.org/analysis/energy-considerations-dawn-strategic-manufacturing>
- [13] U.S. Energy Information Administration, "Electric power annual 2023, table 2.1: Retail sales of electricity to ultimate customers," EIA Electric Power Annual, Oct. 2024, 2023 total retail sales: 3,861 billion kWh. [Online]. Available: <https://www.eia.gov/electricity/annual/>
- [14] U.S. Census Bureau, "QuickFacts: United states—land area in square miles, 2020," Census QuickFacts, 2021, 2020 Census land area: 3,532,316.24 mi². [Online]. Available: <https://www.census.gov/quickfacts/fact/table/US/LND110220>
- [15] "2024 load & capacity data (gold book)," New York Independent System Operator, Tech. Rep., 2024. [Online]. Available: <https://www.nyiso.com/documents/20142/2226333/2024-Gold-Book-Public.pdf>
- [16] U.S. Census Bureau, "QuickFacts: New york city, new york—land area in square miles, 2020," Census QuickFacts, 2021, 2020 Census land area: 300.5 mi². [Online]. Available: <https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/LND110220>
- [17] Vantage Data Centers, "Vantage data centers unveils plans for frontier, a \$25b mega-campus in texas to meet unprecedented ai demand," Press release, Aug. 2025. [Online]. Available: <https://vantage-dc.com/news/vantage-data-centers-unveils-plans-for-frontier-a-25b-mega-campus-in-texas-to-meet-unprecedented-ai-demand/>
- [18] Black & White Engineering, "Higher loads, faster builds: Six trends defining data centers in 2026," Data Center Knowledge, Dec. 2025. [Online]. Available: <https://www.datacenterknowledge.com/data-center-construction/higher-loads-faster-builds-six-trends-defining-data-centers-in-2026>
- [19] Riot Platforms, "Riot energizes new corsicana facility in navarro county, texas," 2024. [Online]. Available: <https://www.riotplatforms.com/riot-energizes-new-corsicana-facility-in-navarro-county-texas/>
- [20] Meta Platforms, Inc., "Deepening our investment in Richland Parish, Louisiana," Meta Data Centers blog, Jul. 2026. [Online]. Available: <https://datacenters.atmeta.com/2026/07/deepening-our-investment-in-richland-parish-louisiana/>
- [21] Entergy Louisiana, LLC, "Entergy Louisiana receives LPSC approval for major infrastructure investments to support Meta's data center and improve reliability," Press release, Aug. 2025, IPSC Docket No. U-37425. [Online]. Available: <https://www.entergy.com/news/entergy-louisiana-receives-lpsc-approval-for-major-infrastructure-investments-to-support-metas-data-center-and-improve-reliability>
- [22] Crusoe Energy Systems LLC, "Crusoe expands AI data center campus in Abilene to 1.2 gigawatts," Press release, Mar. 2025. [Online]. Available: <https://www.crusoe.ai/resources/newsroom/crusoe-expands-ai-data-center-campus-in-abilene-to-1-2-gigawatts>

- [23] GE Vernova Inc., “GE Vernova and Crusoe announce major 29-unit aeroderivative gas turbine deal to deliver power to AI data centers,” Press release, Jul. 2025. [Online]. Available: <https://www.gevernova.com/news/press-releases/ge-vernova-crusoe-announce-major-29-unit-aeroderivative-gas-turbine-deliver-ai-data-centers>
- [24] Con Edison, “Energy storage activities,” EIA Energy Conference presentation, Jun. 2015. [Online]. Available: <https://www.eia.gov/conference/2015/pdf/presentations/leff.pdf>
- [25] National Laboratory of the Rockies, “Accelerating Speed to Power: Data Viewer,” Interactive web map, 2025, online; retired 2026. [Online]. Available: <https://maps.nlr.gov/speed-to-power/data-viewer>
- [26] U.S. Department of Energy, “Speed to Power,” Webpage, Sep. 2025. [Online]. Available: <https://www.energy.gov/speed-to-power>
- [27] World Resources Institute, “Powering the us data center boom: Why forecasting can be so tricky,” <https://www.wri.org/insights/us-data-centers-electricity-demand>, Sep. 2025.
- [28] McKinsey & Company, “Ai power: Expanding data center capacity to meet growing demand,” <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand>, 2024.
- [29] North American Electric Reliability Corporation (NERC), “Characteristics and risks of emerging large loads,” 2025. [Online]. Available: https://www.nerc.com/comm/RSTC_Reliability_Guidelines/Whitepaper%20Characteristics%20and%20Risks%20of%20Emerging%20Large%20Loads.pdf
- [30] Western Electricity Coordinating Council (WECC), “An assessment of large load interconnection risks in the western interconnection,” 2025. [Online]. Available: https://www.wecc.org/sites/default/files/documents/products/2025/Report_WECC%20Large%20Loads%20Risk%20Assessment%204.pdf
- [31] “DOE Data Center Load Flexibility Workshop Summary,” Lawrence Berkeley National Laboratory, Tech. Rep., January 2025. [Online]. Available: https://eta-publications.lbl.gov/sites/default/files/2025-03/final_doe_data_center_load_flexibility_workshop_summary.v0307.pdf
- [32] North American Electric Reliability Corporation (NERC), “Industry recommendation: Large load interconnection, study, commissioning, and operations,” 2025. [Online]. Available: <https://www.nerc.com/pa/rrm/bpsa/Alerts%20DL/NERC%20Alert%20Level%202%20%20Large%20Loads.pdf>
- [33] Electric Reliability Council of Texas (ERCOT), “Large load stability studies whitepaper,” August 2025. [Online]. Available: https://www.ercot.com/files/docs/2025/08/25/Large_Load_Stability_Whitepaper.pdf
- [34] *Electromagnetic compatibility (EMC) – Part 4-7: Testing and measurement techniques – General guide on harmonics and interharmonics measurements and instrumentation, for power supply systems and equipment connected thereto*, International Electrotechnical Commission Std. IEC 61000-4-7, 2009. [Online]. Available: <https://webstore.iec.ch/en/publication/4227>
- [35] *Electromagnetic compatibility (EMC) – Part 4-15: Testing and measurement techniques – Flickermeter – Functional and design specifications*, International Electrotechnical Commission Std. IEC 61000-4-15, 2010. [Online]. Available: <https://webstore.iec.ch/en/publication/4173>
- [36] Information Technology Industry Council, “ITI (CBEMA) curve: Application and interpretation,” 2000. [Online]. Available: <https://www.itic.org/dotAsset/b7e622fd-7b12-4641-bb0b-00af8c9e5c37.doc>
- [37] North American Electric Reliability Corporation (NERC), “Incident review — considering simultaneous voltage-sensitive load reductions,” January 2025. [Online]. Available: https://www.nerc.com/pa/rrm/ea/DOuments/Incident_Review_Large_Load_Loss.pdf
- [38] Lawrence Berkeley National Laboratory, “Demand response and open automated demand response opportunities for data centers,” 2010. [Online]. Available: https://eta-publications.lbl.gov/sites/default/files/demand_response_and_open_automated_demand_response_opportunities_for_data_centers_lbnl-3047e.pdf
- [39] Electric Reliability Council of Texas, (ERCOT), “Notice of interim voluntary curtailment program for large flexible loads,” 2022. [Online]. Available: https://www.ercot.com/services/comm/mkt_notices/M-A120622-01
- [40] J. Paananen, “Grid-interactive data centers enabling energy transition: Data center’s hidden potential to provide essential grid services of a future power system,” *IEEE Electrification Magazine*, vol. 11, no. 3, pp. 26–34, 2023.
- [41] L. Liu, Q. Zhang, Z. J. Zhai, C. Yue, and X. Ma, “State-of-the-art on thermal energy storage technologies in data center,” *Energy and Buildings*, vol. 226, p. 110345, 2020.
- [42] J. A. Winick, G. Zhu, K. Kitz, and A. Adams, “Reducing data center peak cooling demand and energy costs with underground thermal energy storage (utes),” in *50th Workshop on Geothermal Reservoir Engineering*, 02 2025.
- [43] National Laboratory of the Rockies, “Dataset of generative AI workload power profiles,” NLR Data Catalog, 2026. [Online]. Available: <https://data.nlr.gov/submissions/312>
- [44] R. Vercellino, J. Willard, G. Campos, W. da Silva Pereira, O. Hull, M. Selensky, and J. Mueller, “Measurement of generative AI workload power profiles for whole-facility data center infrastructure planning,” *arXiv preprint arXiv:2604.07345*, 2026.
- [45] D. Al Kez, A. M. Foley, F. W. Ahmed, M. O’Malley, and S. Muyeen, “Potential of data centers for fast frequency response services in synchronously isolated power systems,” *Renewable and Sustainable Energy Reviews*, vol. 151, p. 111547, 2021.
- [46] N. E. Koltsaklis and A. S. Dagoumas, “State-of-the-art generation expansion planning: A review,” *Applied energy*, vol. 230, pp. 563–589, 2018.
- [47] P. L. Joskow, “Competitive electricity markets and investment in new generating capacity,” *AEI-Brookings Joint Center Working Paper*, no. 06-14, 2006.
- [48] M. Tillmanns, J. Schöttler, and A. Praktiknjo, “A review of probabilistic resource adequacy assessments in power systems: Methods, applications, and future challenges,” *Energy Policy*, vol. 209, p. 114924, 2026.
- [49] R. Billinton and R. N. Allan, *Reliability Evaluation of Power Systems*, 2nd ed. Springer US, 1996.
- [50] GE Energy Consulting, *Multi-Area Reliability Simulation Program (MARS) User’s Manual*, March 2021.
- [51] G. Stephen, “Probabilistic resource adequacy suite (pras),” National Laboratory of the Rockies, Tech. Rep., 2021. [Online]. Available: <https://docs.nlr.gov/docs/fy21osti/79698.pdf>
- [52] Q. Zhang, F. Zhao, G. Stephen, C. Singh, and L. Xie, “A gradient-based capacity accreditation framework in resource adequacy: Formulation, computation, and practical implications,” *arXiv preprint*, 2026. [Online]. Available: <https://arxiv.org/abs/2601.22087>
- [53] G. Stephen, S. H. Tindemans, J. Fazio, C. Dent, A. F. Acevedo, B. Bagen, A. Crawford, A. Klaube, D. Logan, and D. Burke, “Clarifying the interpretation and use of the lone resource adequacy metric,” in *2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2022, pp. 1–4.
- [54] G. Mantegna, Z. Huang, G. Van Caelenberg, B. Frew, M. Lynch, and M. O’Malley, “Electric grid reliability in an era of unprecedented uncertainty: a review of advances in electric sector resource adequacy assessment and planning,” *IEEE Transactions on Energy Markets, Policy and Regulation*, vol. 4, pp. 1–9, 2025.
- [55] P. Colangelo, A. K. Coskun, J. Megrue, C. Roberts, S. Sengupta, V. Sivaram, E. Tiao, A. Vijaykar, C. Williams, D. C. Wilson, B. Records, Z. MacFarland, D. Dreiling, N. Morey, A. Ratnayake, and B. Vairamohan, “Ai data centres as grid-interactive assets,” *Nature Energy*, vol. 11, pp. 254–261, 2025.
- [56] Y. Sun, B. Frew, S. Dalvi, and S. C. Dhulipala, “Insights into methodologies and operational details of resource adequacy assessment: A case study with application to a broader flexibility framework,” *Applied Energy*, vol. 328, p. 120191, 2022.
- [57] J. Kuna, G. Stephen, and T. Mai, “Beyond capacity credits: Adaptive stress period planning for evolving power systems,” National Laboratory of the Rockies, Tech. Rep., 2024. [Online]. Available: <https://docs.nlr.gov/docs/fy24osti/89386.pdf>
- [58] M. C. Faconti, K. B. Frank, N. C. Howe, S. B. Robbins, and C. N. Thompson, “Ferc orders pjim to reform tariff for co-located generation and load,” K&L Gates LLP, Tech. Rep., 2026. [Online]. Available: <https://www.klgates.com/FERC-Orders-PJM-to-Reform-Tariff-for-Co-Located-Generation-and-Load-1-15-2026>
- [59] S. Li, L. Tong, and T. D. Mount, “Risk-based capacity accreditation of resource-located large loads in capacity markets,” *arXiv preprint arXiv:2511.17715*, 2025.
- [60] A. S. Manne, “Capacity expansion and probabilistic growth,” *Econometrica*, vol. 29, no. 4, pp. 632–649, 1961.
- [61] J. Freidenfelds, “Capacity expansion when demand is a birth-death random process,” *Operations Research*, vol. 28, no. 3-part-ii, pp. 712–721, 1980.
- [62] J. C. Bean, J. L. Hagle, and R. L. Smith, “Capacity expansion under stochastic demands,” *Operations Research*, vol. 40, no. 3-Supplement-2, pp. S210–S216, 1992.
- [63] D. Bienstock and J. F. Shapiro, “Optimizing resource acquisition decisions by stochastic programming,” *Management Science*, vol. 34, no. 2, pp. 215–229, 1988.

- [64] S. Jin, S. M. Ryan, J.-P. Watson, and D. L. Woodruff, "Modeling and solving a large-scale generation expansion planning problem under uncertainty," *Energy Systems*, vol. 2, no. 3–4, pp. 209–242, 2011.
- [65] X. A. Sun and A. J. Conejo, *Robust optimization in electric energy systems*. Springer, 2021, vol. 313.
- [66] O. Berman, Z. Ganz, and J. M. Wagner, "A stochastic optimization model for planning capacity expansion in a service industry under uncertain demand," *Naval Research Logistics*, vol. 41, no. 4, pp. 545–564, 1994.
- [67] S. Ahmed, A. J. King, and G. Parija, "A multi-stage stochastic integer programming approach for capacity expansion under uncertainty," *Journal of Global Optimization*, vol. 26, no. 1, pp. 3–24, 2003.
- [68] S. Ahmed and N. V. Sahinidis, "An approximation scheme for stochastic integer programs arising in capacity expansion," *Operations Research*, vol. 51, no. 3, pp. 461–471, 2003.
- [69] J. Zou, S. Ahmed, and X. Sun, "Partially adaptive stochastic optimization for electric power generation expansion planning," *INFORMS Journal on Computing*, vol. 30, pp. 388–401, 05 2018.
- [70] F. Pourahmadi and J. Kazempour, "Distributionally robust generation expansion planning with unimodality and risk constraints," *IEEE Transactions on Power Systems*, vol. 36, no. 5, pp. 4281–4295, 2021.
- [71] C. Yan, X. Geng, L. Xie, and Z. Bie, "A scenario-based storage planning framework with probabilistic guarantees," in *2019 IEEE Sustainable Power and Energy Conference (iSPEC)*. IEEE, 2019, pp. 2341–2346.
- [72] I.-C. Gonzalez-Romero, S. Wogrin, and T. Gómez, "Review on generation and transmission expansion co-planning models under a market environment," *IET Generation, Transmission & Distribution*, vol. 14, no. 6, pp. 931–944, 2020.
- [73] M. C. Fu, "Optimization for simulation: Theory vs. practice," *INFORMS Journal on Computing*, vol. 14, no. 3, pp. 192–215, 2002.
- [74] C. Byers, T. Levin, and A. Botterud, "Capacity market design and renewable energy: Performance incentives, qualifying capacity, and demand curves," *The Electricity Journal*, vol. 31, no. 1, pp. 65–74, 2018.
- [75] K. Spees, S. A. Newell, A. W. Thompson, E. Snyder, and X. Bartone, "Sixth review of pjm's variable resource requirement curve," The Brattle Group, Tech. Rep., 2025. [Online]. Available: <https://www.brattle.com/wp-content/uploads/2025/04/Sixth-Review-of-PJMs-Variable-Resource-Requirement-Curve.pdf>
- [76] F. Zhao, T. Zheng, and E. Litvinov, "Constructing demand curves in forward capacity market," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 525–535, 2017.
- [77] New York Independent System Operator (NYISO), "Icap demand curves," 2020. [Online]. Available: <https://www.nyiso.com/documents/20142/3036383/ICAP-Demand-Curves.pdf/a16634ed-20c8-0f50-5912-4dd92793cef8>
- [78] "2025 quarterly state of the market report for pjm: January through june," Monitoring Analytics, LLC, Tech. Rep., 2025. [Online]. Available: https://www.monitoringanalytics.com/reports/PJM_State_of_the_Market/2025/q2-som-pjm-sec5.pdf
- [79] ISO New England, "Capacity Auction Reforms Key Project," <https://www.iso-ne.com/committees/key-projects>, 2025.
- [80] P. L. Joskow, "The expansion of incentive (performance-based) regulation of electricity distribution and transmission in the united states," *Review of Industrial Organization*, vol. 65, pp. 455–503, 2024.
- [81] J.-M. Glachant, P. L. Joskow, and M. G. Pollitt, *Handbook on Electricity Regulation*. Edward Elgar Publishing, 2025.
- [82] E. Taylor, M. Korytowski, and G. Reed, "Voltage transient propagation in ac and dc datacenter distribution architectures," in *2012 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 2012, pp. 1998–2004.
- [83] Q. Lin, B. Wen, R. Burgos, K. Shan, Y. Tang, X. Li, and Q. Wang, "Input impedance characterization of a single-phase pfc in dq frame," in *2021 IEEE 22nd Workshop on Control and Modelling of Power Electronics (COMPEL)*. IEEE, 2021, pp. 1–7.
- [84] M. T. Milani, B. Khodabakhchian, and J. Mahseredjian, "Detailed EMT-type load modeling for power system dynamic and harmonic studies," *IEEE Transactions on Power Delivery*, vol. 38, no. 1, pp. 703–711, 2023.
- [85] A. Emadi, "Modeling of power electronic loads in AC distribution systems using the generalized state-space averaging method," *IEEE transactions on Industrial Electronics*, vol. 51, no. 5, pp. 992–1000, 2004.
- [86] Y. Chen, K. Shi, M. Chen, and D. Xu, "Data center power supply systems: From grid edge to point-of-load," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 11, no. 3, pp. 2441–2456, 2022.
- [87] H. Zhang, H. Li, J. Mao, C. Pan, and Z. Luan, "Model-free control of single-phase boost AC/DC converters," *IEEE Transactions on Power Electronics*, vol. 37, no. 10, pp. 11 828–11 838, 2022.
- [88] P. Sandri, "Increasing hyperscale data center efficiency: A better way to manage 54-v\48-v-to-point-of-load direct conversion," *IEEE Power Electronics Magazine*, vol. 4, no. 4, pp. 58–64, 2017.
- [89] D. J. Carnovale, A. Chiste, and T. M. Blooming, "Price and performance considerations for backup power and ride-through solutions," in *Proc. Power Quality Conference (PowerSystems World 2004)*, Chicago, IL, USA, Nov. 2004.
- [90] A. Hoevenaars and P. Hoevenaars, "Data center cooling harmonics-how to get the 'good' without the 'bad'," in *2020 IEEE/IAS 56th Industrial and Commercial Power Systems Technical Conference (I&CPS)*. IEEE, 2020, pp. 1–7.
- [91] M. Miletic, K. R. Ragz, V. Zelenicic, I. Erceg, and D. Sumina, "Development of single-phase shunt active power filter for reduction of current harmonics in data center power system," in *2023 11th International Conference on Smart Grid*. IEEE, 2023, pp. 1–7.
- [92] T. Zhu, F. Zhao, X. Wang, and G. V. Torrico-Bascope, "Adaptive harmonic conductance control for boost pfc converters at light loads," *IEEE Transactions on Power Electronics*, vol. 39, no. 3, pp. 3175–3185, 2023.
- [93] Y. Prabowo, V. M. Iyer, S. Bhattacharya, and E. Aeloiza, "A control method of hybrid transformer enabled harmonic isolator for sensitive clustered harmonic loads," in *2020 IEEE Energy Conversion Congress and Exposition*. IEEE, 2020, pp. 4872–4878.
- [94] J. Sutaria and S. K. Ronnberg, "Propagation of supraharmonics in a data center with different operating modes of ups," *IEEE Access*, vol. 11, pp. 36 823–36 833, 2023.
- [95] P. Piya, M. Ebrahimi, M. Karimi-Ghartemani, and S. A. Khajehoddin, "Fault ride-through capability of voltage-controlled inverters," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 10, pp. 7933–7943, 2018.
- [96] X. Liu, X. Chen, M. Shahidepour, C. Li, Q. Wu, Y. Wu, and J. Wen, "Active fault current limitation for low-voltage ride-through of networked microgrids," *IEEE Transactions on Power Delivery*, vol. 37, no. 2, pp. 980–992, 2021.
- [97] P. Gravois, "ERCOT large electronic load voltage ride-through performance requirements," 2025. [Online]. Available: https://www.ercot.com/files/docs/2025/07/11/ERCOT-LEL-Ride-Through-Criteria_L_LWG-final.pptx
- [98] J. Conto, Y. Cheng, J. Rose, and J. Schmall, "Texas loads ride toward grid stability: Voltage ride through of large power electronic loads," *IEEE Power and Energy Magazine*, vol. 23, no. 5, pp. 56–67, 2025.
- [99] R. Arghandeh, M. Pipattanasomporn, and S. Rahman, "Flywheel energy storage systems for ride-through applications in a facility microgrid," *IEEE Transactions on smart grid*, vol. 3, no. 4, pp. 1955–1962, 2012.
- [100] S. Yan, M.-H. Wang, J. Chen, and S. Y. Hui, "Smart loads for improving the fault-ride-through capability of fixed-speed wind generators in microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 661–669, 2019.
- [101] A. Samanta, Q. Zhang, and L. Xie, "Prediction of cryptocurrency mining load tripping through learning-based fault classification," in *2024 IEEE Texas Power and Energy Conference (TPEC)*. IEEE, 2024, pp. 1–6.
- [102] X. Wang and F. Blaabjerg, "Harmonic stability in power electronic-based power systems: Concept, modeling, and analysis," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2858–2870, 2018.
- [103] M.-S. Ko and H. Zhu, "Wide-area power system oscillations from large-scale AI workloads," *IEEE Transactions on Power Systems*, 2026, early access.
- [104] G. Valverde, G. Pierrou, J. Krajacic, and G. Hug, "Forced oscillations in power systems induced by ai workloads in data centers," *Authora Preprints*, 2025.
- [105] N. Hatziaegyriou, J. Milanovic, C. Rahmann, V. Ajarapu, C. Canizares, I. Erlich, D. Hill, I. Hiskens, I. Kamwa, B. Pal, P. Pourbeik, J. Sanchez-Gasca, A. Stankovic, T. Van Cutsem, V. Vittal, and C. Vournas, "Definition and classification of power system stability-revisited & extended," *IEEE Transactions on Power Systems*, vol. 36, no. 4, pp. 3271–3281, 2021.
- [106] A. Jiménez-Ruiz and F. Milano, "Data centre model for transient stability analysis of power systems," *IET Generation, Transmission & Distribution*, vol. 20, no. 1, 2026.

- [107] J. Sun, M. Xu, M. Cespedes, and M. Kauffman, "Data center power system stability—part i: Power supply impedance modeling," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 2, pp. 403–419, 2022.
- [108] J. Sun, M. Mihret, M. Cespedes, D. Wong, and M. Kauffman, "Data center power system stability—part ii: System modeling and analysis," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 2, pp. 420–438, 2022.
- [109] S. Sheehan and A. Rakow, "Evolving a data center into a microgrid: Industry perspectives and lessons learned," *IEEE Electrification Magazine*, vol. 11, no. 3, pp. 16–25, 2023.
- [110] T. Guillod, F. Krismer, and J. W. Kolar, "Protection of MV converters in the grid: The case of MV/LV solid-state transformers," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 5, no. 1, pp. 393–408, Mar. 2017.
- [111] Rocky Mountain Institute, "Powering the data-center boom with low-carbon solutions," 2024. [Online]. Available: <https://rmi.org/insight/powering-the-data-center-boom-with-low-carbon-solutions/>
- [112] I. Alaperä, J. Paananen, K. Dalen, and S. Honkapuro, "Fast frequency response from a ups system of a data center, background, and pilot results," in *2019 16th International Conference on the European Energy Market (EEM)*, 2019, pp. 1–5.
- [113] "Flexible ramping product refinements: Final proposal," California Independent System Operator, Tech. Rep., Aug. 2020. [Online]. Available: <https://stakeholdercenter.caiso.com/InitiativeDocuments/FinalProposal-FlexibleRampingProductRefinements.pdf>
- [114] Q. Wang and B.-M. Hodge, "Enhancing power system operational flexibility with flexible ramping products: A review," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1652–1664, 2017.
- [115] "What the duck curve tells us about managing a green grid," CAISO Fast Facts, California Independent System Operator, Tech. Rep., 2016. [Online]. Available: https://www.caiso.com/documents/flexibleresourcehelprenewables_fastfacts.pdf
- [116] Electric Reliability Council of Texas, "Implementation details for ERCOT contingency reserve service," May 2023. [Online]. Available: https://www.ercot.com/services/comm/mkt_notices/M-D050523-01
- [117] E. Choukse, B. Warriar, S. Heath, L. Belmont, A. Zhao, H. A. Khan *et al.*, "Power stabilization for AI training datacenters," arXiv:2508.14318, Aug. 2025, 57 authors (Microsoft and NVIDIA). [Online]. Available: <https://arxiv.org/abs/2508.14318>
- [118] Electric Reliability Council of Texas, "2023 large flexible load task force (lftf): Scheduled meetings and key documents," 2023. [Online]. Available: <https://www.ercot.com/committees/inactive/lftf/2023>
- [119] I. Riu, D. Smiley, S. Bessasparis, and K. Patel, "Load growth is here to stay, but are data centers?" Energy and Environmental Economics, Inc, Tech. Rep., Jul. 2024. [Online]. Available: <https://www.ethree.com/wp-content/uploads/2024/07/E3-White-Paper-2024-Load-Growth-Is-Here-to-Stay-but-Are-Data-Centers-2.pdf>
- [120] Google, "Our data centers now work harder when the sun shines and wind blows," Apr. 2020. [Online]. Available: <https://blog.google/innovation-and-ai/infrastructure-and-cloud/global-network/data-centers-work-harder-sun-shines-wind-blows/>
- [121] V. Mehra and R. Hasegawa, "Supporting power grids with demand response at google data centers," Oct. 2023. [Online]. Available: <https://cloud.google.com/blog/products/infrastructure/using-demand-response-to-reduce-data-center-power-consumption>
- [122] Google LLC, "24/7 by 2030: Realizing a carbon-free future," Sep. 2020. [Online]. Available: <https://sustainability.google/reports/24-7-carbon-free-energy/>
- [123] *IEEE Guide for Loading Mineral-Oil-Immersed Transformers and Step-Voltage Regulators*, Std., 2012.
- [124] T. Gabler, K. Backhaus, S. Großmann, and R. Fritsche, "Dielectric modeling of oil-paper insulation systems at high dc voltage stress using a charge-carrier-based approach," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 26, no. 5, pp. 1549–1557, 2019.
- [125] Nostromo Energy, "Ice-based cooling in the data center: Less carbon, more profit," 2023. [Online]. Available: <https://nostromo.energy/blogs/ice-based-cooling-in-the-data-center-less-carbon-more-profit>
- [126] K. S. Ratnam, K. Palanisamy, and G. Yang, "Future low-inertia power systems: Requirements, issues, and solutions - a review," *Renewable and Sustainable Energy Reviews*, vol. 124, p. 109773, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1364032120300691>
- [127] Electric Power Research Institute, "EPRI launches initiative to enhance data center flexibility and grid reliability," Press release, Oct. 2024. [Online]. Available: <https://www.prnewswire.com/news-releases/epri-launches-initiative-to-enhance-data-center-flexibility-and-grid-reliability-302289874.html>
- [128] Q. Zhang, L. Xie, L. Zhao, and C. Wang, "Comparative assessment of look-ahead economic dispatch and ramp products for grid flexibility," 2026. [Online]. Available: <https://arxiv.org/abs/2601.22120>
- [129] M. C. Offutt, L. Zhu, and A. J. Lawson, "Data centers and their energy consumption: Frequently asked questions," Congressional Research Service, Tech. Rep. R48646, 2026. [Online]. Available: <https://crsreports.congress.gov/product/pdf/R/R48646>
- [130] North American Electric Reliability Corporation (NERC), "2024 summer reliability assessment," May 2024. [Online]. Available: https://www.nerc.com/pa/RAPA/ra/Reliability%20Assessments%20DL/NERC_SRA_2024.pdf
- [131] "2024 long-term reliability assessment," North American Electric Reliability Corporation (NERC), Tech. Rep., Dec. 2024. [Online]. Available: https://www.nerc.com/globalassets/our-work/assessments/2024-ltra_corrected_july_2025.pdf
- [132] W. Torell and M. A. Torres Arango, "How 6 AI attributes change data center design," Schneider Electric, White Paper 110, Version 3, 2026. [Online]. Available: https://download.schneider-electric.com/files?p_Doc_Ref=SPD_WP110_EN
- [133] NVIDIA Corporation, "Hardware—NVIDIA DGX GB rack scale systems user guide," NVIDIA Documentation, 2026, rack power consumption approximately 120 kW. [Online]. Available: <https://docs.nvidia.com/dgx/dgxb200-user-guide/hardware.html>
- [134] Lawrence Berkeley National Laboratory, "2024 united states data center energy usage report," 2024. [Online]. Available: <https://eta.lbl.gov/publications/2024-lbnl-data-center-energy-usage-report>
- [135] Navitas Semiconductor, "Redefining data center power: GaN and SiC technologies for next-gen 800 VDC infrastructure," White Paper, Oct. 2025. [Online]. Available: <https://navitassemi.com/resources/white-papers/>
- [136] International Energy Agency, "Building the future transmission grid: Strategies to navigate supply chain challenges," International Energy Agency, Paris, France, Tech. Rep., Feb. 2025. [Online]. Available: <https://www.iea.org/reports/building-the-future-transmission-grid>
- [137] Z. Shao, N. Yu, and D. Wong, "Stochastic long-term joint decarbonization planning for power systems and data centers: A case study in pjm," *arXiv preprint*, 2025. [Online]. Available: <https://arxiv.org/abs/2510.25118>
- [138] "Monitoring the 'connect and manage' electricity grid access regime," Office of Gas and Electricity Markets, Tech. Rep., dec 2015. [Online]. Available: https://ofgem.gov.uk/sites/default/files/docs/monitoring_the_connect_and_manage_electricity_grid_access_regime_sixth_report_from_ofgem_0.pdf
- [139] C. Horan, M. Zimmerman, C. Jermyn, and D. Godo-Solo, "Let's get flexible: Considerations for unlocking grid capacity using flexible interconnection," Environmental Defense Fund, Tech. Rep., 2024. [Online]. Available: <https://library.edf.org/AssetLink/q812pd5af3hb0i61cm503fprla5ge0p.pdf>
- [140] International Energy Agency, "Energy and AI," International Energy Agency, Paris, Tech. Rep., Apr. 2025. [Online]. Available: <https://www.iea.org/reports/energy-and-ai>
- [141] Y. He, G. Ruan, and H. Zhong, "Performance-aware control of modular batteries for fast frequency response," *IEEE Transactions on Sustainable Energy*, vol. 16, pp. 2547–2559, 2025.
- [142] W. Zheng, K. Ma, and X. Wang, "Hybrid energy storage with supercapacitor for cost-efficient data center power shaving and capping," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 1105–1118, 2016.
- [143] K. Wang, L. Ye, S. Yang, Z. Deng, J. Song, Z. Li, and Y. Zhao, "A hierarchical dispatch strategy of hybrid energy storage system in internet data center with model predictive control," *Applied Energy*, vol. 331, p. 120414, 2023.
- [144] M. Shamseldin, "From liability to asset: A three-mode grid-forming control framework for centralized data center ups systems," *arXiv preprint arXiv:2512.16497*, 2025.
- [145] A. Azizi, A. Aluko, S. Maleki, R. Salehi, and H. Bayat, "Grid-forming inverter applications for improved reliability and power quality in ai data centers," in *2025 IEEE Power & Energy Society General Meeting*, IEEE, 2025, pp. 1–5.
- [146] L. L. Qi, H. Suryanarayana, Y. Zhang, T. Jiang, and S. Colombi, "Modeling, simulation, and protection of grid forming inverter-based ring configuration datacenter," in *2022 IEEE Industry Applications Society Annual Meeting (IAS)*, IEEE, 2022, pp. 1–7.
- [147] B. A. Ross, X. Lyu, U. C. Nwaneto, S. M. Mohiuddin, and A. B. Nassif, "Using grid-forming energy storage systems to provide dynamic active

- power support for hyperscale data center,” *IEEE Access*, vol. 14, pp. 29 250–29 259, 2026.
- [148] UNIFI Consortium, “UNIFI specifications for grid-forming inverter-based resources: Version 2,” National Renewable Energy Laboratory, Tech. Rep. NREL/TP-5D00-89269, Mar. 2024. [Online]. Available: <https://docs.nrel.gov/docs/fy24osti/89269.pdf>
- [149] P. Matwijec and E. Spears, *Mitigating data center harmonics*, Eaton, Jan. 2014. [Online]. Available: <https://www.eaton.com/content/dam/eaton/markets/data-center/Mitigating-DC-harmonics.pdf>
- [150] Hitachi Energy, “Static compensator (STATCOM) for dynamic grid stability,” Product Documentation, 2025. [Online]. Available: <https://www.hitachienergy.com/products-and-solutions/facts/statcom>
- [151] A. Satchwell, N. M. Frick, P. Cappers, S. Sergici, R. Hledik, G. Kavlak, and G. Oskar, “Electricity rate designs for large loads: Evolving practices and opportunities,” Lawrence Berkeley National Laboratory, Tech. Rep., Jan. 2025. [Online]. Available: <https://eta-publications.lbl.gov/publications/electricity-rate-designs-large-loads>
- [152] L. Xie, Q. Zhang, M. Yu, P. L. Joskow, and C. Singh, “Crucial role of foundation models in enhancing the interaction of ai and power systems: Achieving integrated frameworks,” *IEEE Energy Sustainability Magazine*, vol. 2, no. 1, pp. 22–28, 2026.
- [153] Q. Zhang and L. Xie, “Poweragent: A road map toward agentic intelligence in power systems: Foundation model, model context protocol, and workflow,” *IEEE Power and Energy Magazine*, vol. 23, no. 5, pp. 93–101, 2025.
- [154] S. Wang, Z. Yao, S. Zhang, J. Gai, T. X. Liu, and S. Zhong, “When experimental economics meets large language models: Tactics with evidence,” *arXiv preprint arXiv:2505.21371*, 2025.
- [155] J. Li and W. Qi, “Toward optimal operation of internet data center microgrid,” *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 971–979, 2018.
- [156] G. Ruan, D. Qiu, S. Sivaranjani, A. S. Awad, and G. Strbac, “Data-driven energy management of virtual power plants: A review,” *Advances in Applied Energy*, vol. 14, p. 100170, 2024.
- [157] J. Huber and J. W. Kolar, “Next-generation solid-state transformers for powering artificial intelligence data centers: A survey of emerging topologies,” *IEEE Industrial Electronics Magazine*, vol. 20, pp. 42–54, 2026.
- [158] L. Zheng, A. Marellapudi, V. R. Chowdhury, N. Bilakanti, R. P. Kandula, M. Saeedifard, S. Grijalva, and D. Divan, “Solid-state transformer and hybrid transformer with integrated energy storage in active distribution grids: Technical and economic comparison, dispatch, and control,” *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 10, no. 4, pp. 3771–3787, Aug. 2022.
- [159] J. E. Huber and J. W. Kolar, “Applicability of solid-state transformers in today’s and future distribution grids,” *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 317–326, Jan. 2019.
- [160] C. Xu, J. Wei, L. Zheng, X. Han, M. Saeedifard, R. P. Kandula, K. Kandasamy, D. Divan, and L. Graber, “Insulation coordination design for grid-connected solid-state transformers,” *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 10, no. 4, pp. 3746–3758, 2022.
- [161] S. Ozdemir, N. Altin, A. Nasiri, and R. Cuzner, “Review of standards on insulation coordination for medium voltage power converters,” *IEEE Open Journal of Power Electronics*, vol. 2, pp. 236–249, 2021.
- [162] S. A. Saleh, E. Ozkop, B. Alsayid, C. Richard, X. F. St. Onge, K. M. McDonald, and L. Chang, “Solid-state transformers for distribution systems—part II: Deployment challenges,” *IEEE Transactions on Industry Applications*, vol. 55, no. 6, pp. 5708–5716, 2019.
- [163] R. H. Byrne, A. Bera, and T. A. Nguyen, “Reliability of the future smart grid and the role of energy storage,” Presented at SPEEDAM 2024, Ischia, Italy; Sandia National Laboratories SAND2024-07495C, Jun. 2024. [Online]. Available: <https://www.osti.gov/servlets/purl/2586303>
- [164] T. Sousa, T. Soares, P. Pinson, F. Moret, T. Baroche, and E. Sorin, “Peer-to-peer and community-based markets: A comprehensive review,” *Renewable and Sustainable Energy Reviews*, vol. 104, pp. 367–378, 2019.