

The Megawatt You Don't Build For

Large-load flexibility is becoming an infrastructure alternative. But a promised curtailment is not capacity: utilities can defer firm network investment only when flexible megawatts are contractible, observable, dispatchable and auditable.

Scope: U.S. electric utilities, with a primary focus on transmission and grid-planning decisions associated with data centers and other very large loads.

A. At a glance

Five signals for the executive committee

9–17%

EPRI's scenario range for data centers' share of US electricity demand in 2030, up from 4–5% in 2024.[\[1\]](#)

Six RTOs/ISOs

FERC ordered every regional grid operator under its jurisdiction to justify or reform large-load connection rules in June 2026.[\[5\]](#)

25% for 3 hours

A Phoenix field demonstration reduced power consumption in a 256-GPU AI cluster by 25% for three hours while maintaining the trial's stated quality-of-service guarantees. It demonstrates compute-level flexibility, not facility-wide availability or planning credit.[\[14\]](#)

Nearly 50% in Texas forecasting

NERC reports that new Texas data-center demand projections are reduced by almost 50% from initially requested levels to reflect observed early-operation consumption; this is an ERCOT forecasting practice, not a universal planning factor.[\[3\]](#)

Measurement matters

Virginia's commission explicitly placed measurement and verification of flexible load reductions on its regulatory agenda.[\[9\]](#)

Rules remain regional

Federal, state and RTO actions point in the same direction, but service definitions, dispatch authority and retail cost allocation are not uniform.[\[5\]](#)[\[7\]](#)

B. What the evidence says

The grid does not need another unqualified peak forecast

The large-load problem is commonly framed as a race to build. That is only half the diagnosis. Electric utilities and grid planners face a compound uncertainty: the volume of proposed demand is unusually large, the probability and pace of project realization are uncertain, and the eventual electrical behavior of AI-intensive facilities may differ materially from a flat 24/7 block. EPRI estimates US data centers could consume 383–793 TWh annually by 2030, equivalent to 9–17% of national electricity demand under its scenarios.^[1] Berkeley Lab's June 2026 update, using a bottom-up model based on planned IT-equipment shipments, device energy use and cooling performance, independently estimates a 9.5–15.3% share by 2030, with an 11.8% reference case.^[2] The ranges are not interchangeable forecasts, but their convergence on a high-demand outcome strengthens the evidence on scale while their spread reinforces the uncertainty around pace and realization.

That uncertainty already enters planning practice. NERC explains that system forecasts rely on interconnection information and customer agreements covering peak demand, flexibility and operating characteristics; projects must advance beyond speculative stages before they drive formal studies.^[3] In Texas, NERC reports that new data-center demand projections are reduced by almost half from their originally requested load levels to reflect observed consumption during the initial operation of existing sites. ERCOT also removes most new large loads that have not yet entered a transmission-service contract from its long-term demand forecast.^[3] These adjustments illustrate how project maturity and realized operating behavior can affect forecasting; they should not be read as a standardized derating factor applicable outside ERCOT. PJM's 2026 load-forecast supplement likewise separates firm from non-firm data-center additions in AEP territory.^[20]

The implication is uncomfortable but useful: a request for 500 MW is not yet a 500 MW planning fact. It is a planning input whose treatment depends on evidence of project maturity, expected energization and ramp, contractual commitment and operating characteristics. Those adjustments vary by utility, planning authority and jurisdiction.^[3] If electric utilities plan network investment around every nameplate request, they risk stranded or prematurely timed assets and cost transfer to existing customers. If they discount too aggressively, they risk reliability and foregone economic development. DOE's rate-design brief therefore treats cost allocation, stranded-asset risk, resource adequacy and on-site capacity as a connected design problem.^[16]

Regulation is moving from “how much?” to “under what service?”



The clearest federal signal now comes from FERC, but turning flexible load into a dependable service remains a multi-jurisdictional problem. FERC's large-load docket sought comment on whether customers willing to operate flexibly and curtail usage should qualify for accelerated interconnection studies, including whether such studies could be completed within 60 days.^[4] On 18 June 2026, FERC issued show-cause orders to all six RTOs/ISOs under its jurisdiction. The five reform categories included new transmission services for flexible large loads, alongside study reform, cost protection, co-location and joint study of load with proximate generation.^[5] A commissioner's accompanying statement made the economic logic explicit: non-firm service can reduce network upgrades and generation capacity, accelerate connection and lower costs, but requires enhanced data reporting and telemetry.^[6]

PJM is developing that logic through its Connect and Manage Senior Task Force. The stakeholder process is exploring flexibility arrangements that could allow certain large loads to connect before necessary transmission upgrades are completed. The PJM Board has directed development of a framework in which qualifying new large loads without associated new generation could be curtailed before pre-emergency Demand Response, but the operational design remains subject to stakeholder development and applicable governing-document approvals.^[7] The associated Large Load Customer Flexibility problem statement highlights unresolved implementation needs, including identification of non-firm retail load, tools to determine the location and magnitude of curtailment, and enforceable retail service arrangements.^[8]

State and regional evidence makes the governance gap even clearer. Virginia's State Corporation Commission asked how load reductions should be measured and verified, how large-load flexibility could enter demand-side programs, and whether an inflexible customer could obtain flexibility from another location.^[9] Texas Senate Bill 6 established several large-load provisions in PURA, including authority for specified curtailment arrangements and a requirement for the PUC to establish a large-load demand-management service.^[10] ERCOT is now developing the operational implementation. Its September 2026 notice describes an October workshop on proposed registration requirements, curtailment blocks and sequencing, deployment mechanics and a proposed Large Load Curtailment Manager, with stakeholder feedback preceding formal revision requests.^[10]

The emerging model is therefore layered rather than uniform:

- FERC shapes interstate transmission rules,
- RTOs/ISOs define regional planning and operating mechanisms,
- state commissions govern important retail and cost-allocation questions,
- and utilities must translate those rules into enforceable customer service arrangements.

Convergence of regulatory attention, 2025–2026

The policy questions are increasingly similar; service design and implementation remain jurisdiction-specific.



Sources: FERC, Virginia SCC and ERCOT.^{[4][5][9][10][11]}

The technical pathways exist, but flexibility is not homogeneous

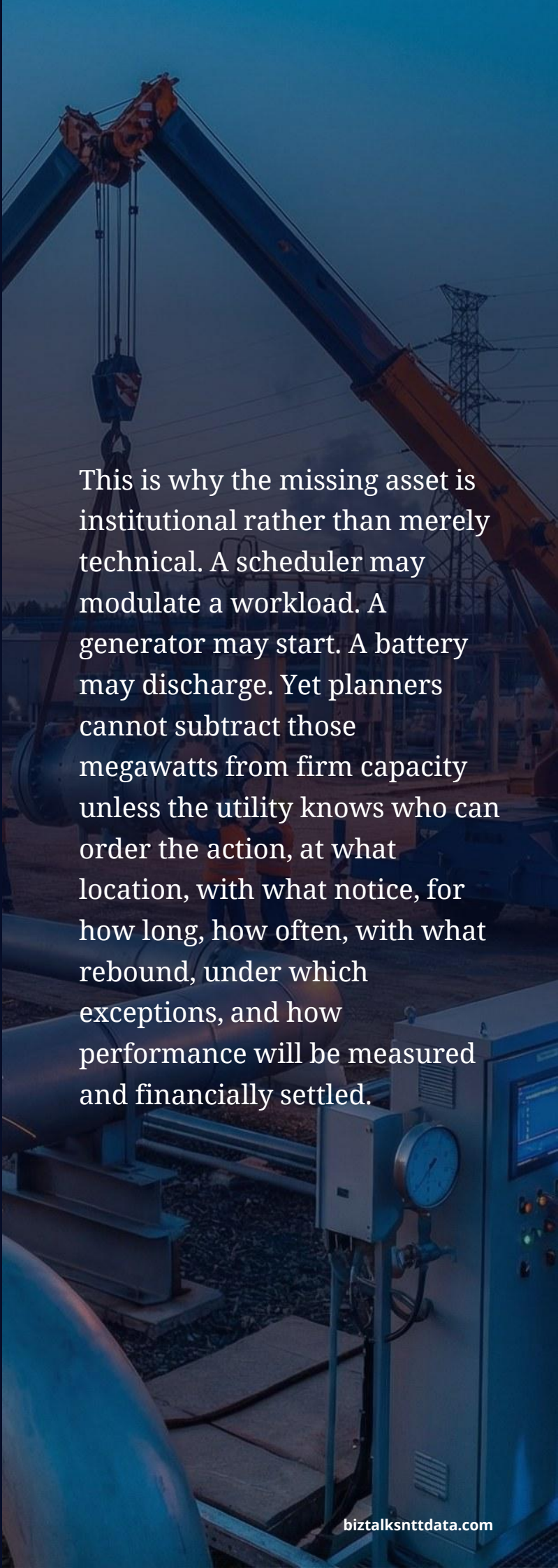


EPRI's DCFlex framework identifies compute assets, balance-of-plant systems and power assets as distinct sources of flexibility. It characterizes each by potential modulation, notice time, duration and availability, and warns that ownership models, workloads and market rules affect what can actually be delivered.^[12] Its demonstrations span workload modulation, demand response, backup-power integration, power quality and carbon-aware operation.^[13]

The Phoenix demonstration provides a concrete public proof point for compute-level flexibility. In a commercial hyperscale data-center environment, a software-controlled 256-GPU cluster reduced cluster power by 25% for three hours during peak grid events while maintaining the trial's stated quality-of-service guarantees.^[14] The result demonstrates that selected AI workloads can provide measurable demand reduction through software orchestration without dedicated storage or new generation. It does not establish an equivalent flexible share for the full facility, nor does it demonstrate year-round availability, repeatability across workload types or planning-grade firmness. EPRI's broader program also tests backup integration and other pathways precisely because no single lever fits every facility.^[13]



There is also a reliability counterweight. NERC reported a 2024 event in which approximately 1,500 MW of data-center load disconnected simultaneously and unexpectedly from the Eastern Interconnection following a transmission fault, causing system voltage and frequency to rise rapidly. NERC also reports several similar events in ERCOT at a smaller scale, in the range of 100–400 MW.^[3] Separate NERC incident analysis shows that some data-center protection and control schemes can transfer load from the grid to backup systems following repeated voltage disturbances, reinforcing the need to model facility-specific protection behavior.^[21] DOE separately reports that synchronized AI workloads can create electrical oscillations that exceed the useful frequency range of conventional phasor measurements, leading it to recommend hybrid monitoring with higher-resolution waveform data.^[15] Flexible load is therefore not simply negative generation. Its ramps, rebounds, transfer behavior, protection logic and power quality must be modeled.

A large blue and orange crane is lifting a heavy metal component at an industrial site. The crane's arm extends diagonally across the upper right portion of the image. In the foreground, there is a large blue cylindrical object, possibly a tank or part of a machine. The background shows industrial structures and power lines under a clear sky.

This is why the missing asset is institutional rather than merely technical. A scheduler may modulate a workload. A generator may start. A battery may discharge. Yet planners cannot subtract those megawatts from firm capacity unless the utility knows who can order the action, at what location, with what notice, for how long, how often, with what rebound, under which exceptions, and how performance will be measured and financially settled.

Illustrative mechanism: from technical flexibility to planning credit

Inputs

Nameplate commitment:	100 MW	Demonstrated technical flexibility:	25 MW*
Performance / availability factor:	A	Planning or accreditation adjustment:	B



Outputs

Technical flexible MW 25 MW	Creditable flexible MW $25 \times A \times B$	Residual firm MW $100 - (25 \times A \times B)$
--------------------------------	--	--

Illustrative analytical structure, not a tariff, accreditation methodology or engineering standard. The 25% value is used solely to connect the example to the Phoenix cluster demonstration.^[14] Performance, availability and planning adjustments must be determined from site-specific evidence and the applicable planning or regulatory framework; no default values are implied.



C. Implications and scenarios

Three analytical service models

NTT DATA analytical taxonomy synthesizing approaches observed across current large-load planning and flexibility discussions; the categories are not standardized regulatory service classes.

Model	Executive logic	Primary risk	Use when
Firm-first	Build and procure for the full qualified load.	Time, cost and stranded capacity if demand ramps slowly.	Curtailment is infeasible or failure consequences are unacceptable.
Conditional connection	Connect earlier with enforceable non-firm rights until upgrades arrive.	Dispatch failure, unclear authority or customer operational loss.	Constraints are identifiable and the load can respond at the required location and duration.
Portfolio-backed flexibility	Combine workload modulation, storage, on-site power or contracted remote flexibility to meet a defined reduction obligation.	Complex baselines, double counting and jurisdictional fragmentation.	Multiple resources can be orchestrated and settled as one obligation.

Analytical synthesis informed by FERC, PJM, EPRI and state/regional proceedings.[5][7][9][12]

Decision hypothesis: when conditional connection can create value: Conditional connection should be treated as a distinct service option where three conditions can be demonstrated: the relevant network constraint is identifiable, the customer's reduction can be contractually controlled at that constraint, and non-performance can be managed without compromising reliability or shifting undue risk to other customers. Where those conditions are met, planning credit should be earned progressively through tested and settled performance rather than assumed from technical capability alone. Technical capability may establish eligibility; contractual control and demonstrated performance determine whether any planning credit is warranted.[5][7][8][9]

The Firmness Ledger: the proof stack for flexible capacity

Each layer converts a customer promise into a more credible planning input.

01

Commitment

Nameplate MW, energization schedule, site control, commercial readiness and ramp.

02

Physical capability

Flexible assets, workload classes, location, response time, duration, frequency and rebound.

03

Contractual control

Dispatch authority, curtailment priority, exceptions, notice, customer compensation and liability.

04

Operational proof

Telemetry, baseline, event signal, delivered MW, power quality, failure reason and recovery.

05

Planning and settlement

Creditable MW by constraint and season, audit trail, penalties, payment and periodic derating.

Proposed NTT DATA analytical artifact; qualitative assessment to be validated. Informed by FERC, PJM, Virginia SCC, ERCOT and EPRI requirements.[\[6\]](#)[\[8\]](#)[\[9\]](#)[\[10\]](#)[\[12\]](#)



D. NTT DATA point of view

The consequence of the evidence

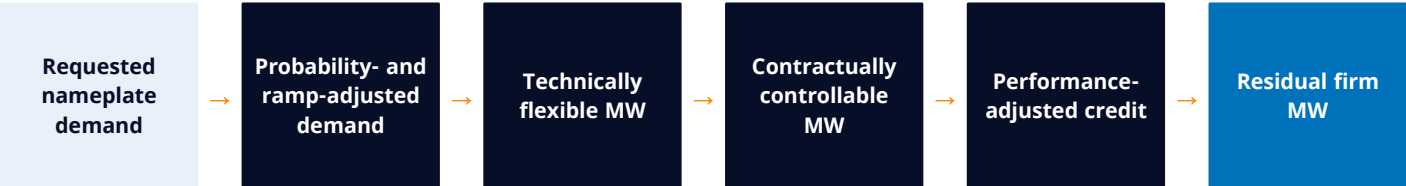
Our reading is that the industry should stop asking whether large loads are flexible in the abstract. The operative question is whether a particular portfolio of load, compute, storage and on-site power can produce a verifiable reduction at the constraint that matters, under a contract the operator can enforce.

This is hard because the evidence is owned by different parties and lives on different clocks. Developers own energization commitments. Data-center operators understand workload criticality. Utilities see network constraints and customer meters. RTOs/ISOs hold system conditions and dispatch logic. Regulators define cost recovery and consumer protection. No single dataset can establish firm need; the answer must be reconciled across them.



Proposed NTT DATA point of view, subject to SME validation.
“Flexibility-to-Capacity” and “Firmness Ledger” are proposed analytical constructs for this paper; they are not presented as existing NTT DATA products or commercial offerings.

We therefore propose a governed flexibility-to-capacity capability centered on the Firmness Ledger. It would not replace SCADA, distribution management, market systems, customer contracts or planning tools. It would connect their authoritative evidence into one decision record:





For an electric utility facing concentrated large-load growth, the strategic advantage is not merely “more flexibility.” Where the required tariff, operational authority and evidence framework are in place, a large-load customer may gain a clearer or earlier connection pathway; the operator may gain an enforceable and observable service; and the regulator may gain a stronger evidentiary basis for assessing whether existing customers are protected from speculative or unperformed commitments.

Existing public NTT DATA capabilities are adjacent to several components required by this model, but they do not constitute evidence of an existing end-to-end large-load flexibility offering. “Digital Twin Connectivity for Real-Time Utility Operations” describes the synchronization of telemetry, topology, asset status and operational events using common identifiers, APIs, event streams, timestamping and data-quality controls.^[18] The public Utilities portfolio also describes real-time utility platforms and edge-to-cloud OT data orchestration that connect operational and enterprise information across utility environments.^[17] Separately, “How to Decide Where to Invest in the Grid? AI and Optimization Make the Difference” describes data integration, predictive analytics and scenario-based optimization for network-investment decisions.^[19] These capabilities illustrate relevant architectural components; applying them to flexibility accreditation, contractual curtailment and planning credit remains a proposed extension that would require utility-specific design and validation.

Economically, the potential value mechanisms are identifiable even where net value is not. Better proof may reduce unnecessary early commitment of network capital, improve utilization of existing assets and give some large loads a priced route to earlier service. Against those benefits sit new costs and risks: telemetry, integration, testing, cyber controls, customer compensation, operating reserves, non-performance exposure and potentially deferred rather than avoided investment. The business case should therefore be evaluated as a portfolio of incremental effects—not as a headline CAPEX saving—including upgrade timing, avoided or deferred capacity, connection revenue, flexibility payments, operational cost, reliability exposure and customer bill impact.^[16]

E. What leaders should do

2026–2030 decision agenda

1	Separate demand truth from queue volume.	Adopt common readiness, ramp and duplication criteria before a request enters capital planning.
2	Define the service before selecting technology.	Specify constraint, notice, duration, frequency, rebound, priority and failure consequence.
3	Build the evidence contract.	Align telemetry, baseline, event records, audit access and settlement with the curtailment right.
4	Pilot by constraint pocket.	Choose locations where conditional service can materially change an upgrade decision and where dispatch authority is clear.
5	Make firmness earned, not assumed.	Award planning credit only through the applicable accreditation methodology and demonstrated performance under relevant seasonal and operating conditions.



Boardroom questions:

- Which planned upgrades are driven by a small number of large-load peaks?
- What flexible-MW claim could we defend today before a regulator?
- Who has the legal and operational authority to curtail?
- What happens to customers and reliability when the flexibility does not arrive?

Bibliography

All sources accessed 21 September 2026

- 1. Powering Intelligence 2026: Executive Summary** — Electric Power Research Institute — 2026 — scenarios to 2030 — accessed 21 Sep 2026.
- 2. United States Data Center Energy Usage Report: 2025 Update** — Lawrence Berkeley National Laboratory — June 2026 — scenarios to 2030 — accessed 21 Sep 2026.
- 3. 2025 Long-Term Reliability Assessment** — North American Electric Reliability Corporation — Jan 2026 — ten-year assessment — accessed 21 Sep 2026.
- 4. Interconnection of Large Loads to the Interstate Transmission System, Docket No. RM26-4-000** — Federal Energy Regulatory Commission — ANOPR proceeding opened Oct 2025 — accessed 21 Sep 2026.
- 5. FERC Launches Aggressive Targeted Action to Speed Large Load Integration** — FERC — 18 Jun 2026 — accessed 21 Sep 2026.
- 6. Commissioner Rosner's Remarks on Large Load Show Cause Orders** — FERC — 18 Jun 2026 — accessed 21 Sep 2026.
- 7. Connect and Manage Senior Task Force** — PJM Interconnection — 2026 — accessed 21 Sep 2026.
- 8. Large Load Customer Flexibility Problem Statement** — PJM stakeholder process — 25 Mar 2026 — accessed 21 Sep 2026.
- 9. SCC Schedules Technical Conference on Data Center Load Flexibility** — Virginia State Corporation Commission — 24 Oct 2025 — accessed 21 Sep 2026.
- 10. Workshop to Share Proposed Operational Framework for Implementing SB6-related Large Load Curtailments, Including Registration Requirements** — ERCOT Market Notice M-A090226-01 — 2 Sep 2026 — accessed 21 Sep 2026.
- 11. 2026 Summer Weather and Operations Outlook** — ERCOT — May 2026 — accessed 21 Sep 2026.
- 12. Grid Flexibility Needs and Data Center Characteristics** — EPRI DCFlex — Jun 2025 — accessed 21 Sep 2026.
- 13. DCFlex Demonstrations** — EPRI — updated 2026 — demonstration program and 2025 update — accessed 21 Sep 2026.
- 14. Turning AI Data Centers into Grid-Interactive Assets: Results from a Field Demonstration in Phoenix, Arizona** — Colangelo et al. — 1 Jul 2025 — preprint / field demonstration conducted with Emerald AI, NVIDIA, Salt River Project and EPRI — accessed 21 Sep 2026.
- 15. Monitoring Oscillations from Large Data Centers** — US Department of Energy, Office of Electricity — 28 May 2026 — accessed 21 Sep 2026.
- 16. Electricity Rate Designs for Large Loads** — US Department of Energy, Office of Policy — 17 Jan 2025 — accessed 21 Sep 2026.
- 17. Utilities: How NTT DATA helps clients** — NTT DATA BizTalks — public Utilities portfolio covering Real-Time Utility Platforms, Edge-to-Cloud OT Data Orchestration and Digital Twin Connectivity — accessed 21 Sep 2026.
- 18. Digital Twin Connectivity for Real-Time Utility Operations** — NTT DATA BizTalks — Sep 2026 — accessed 21 Sep 2026.
- 19. How to Decide Where to Invest in the Grid? AI and Optimization Make the Difference** — NTT DATA BizTalks — 16 Jul 2026 — accessed 21 Sep 2026.
- 20. PJM Load Forecast Report Supplement 2026** — PJM Interconnection — 2026 — forecast horizon 2026–2046 — accessed 21 Sep 2026.
- 21. Incident Review: Considering Simultaneous Voltage-Sensitive Load Reductions** — North American Electric Reliability Corporation — 2024/updated Jan 2025 — accessed 21 Sep 2026.

