GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

AI Data Center Demand Response and Interconnection Plan

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-10-13 · ai data center demand response · flexible data center load · ai workload curtailment · flexible interconnection · nvidia dsx flex · grid interactive data center · gpu power management · demand response measurement · curtailment acceptance testing

Original article: https://gpusmith.com/articles/en/ai-data-center-demand-response-interconnection-plan

In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. Define the Curtailment Contract Before Sizing the Fleet
5. Implementation Considerations and Process Changes
6. Measurement, Verification, and Control Design
7. AI Data Center Curtailment Testing and Acceptance
8. Economics, Ownership, and the Contract Evidence Pack
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Conclusion

Executive Summary

Artificial intelligence (AI) data center demand response is a commitment to reduce metered electricity demand under defined conditions while keeping protected workloads within their service level objectives (SLOs). The relevant engineering question is how much power a particular workload mix can release, how quickly, for how long, and with what recovery obligations. ASHRAE describes data center flexibility as **partial and situational**, constrained by operating requirements and service agreements. That makes fleet-specific acceptance evidence more useful than a generic percentage of installed graphics processing unit (GPU) capacity. (www.ashrae.org)

The September announcements provide a useful trigger for that evidence plan. In its September 15, 2026 account, NVIDIA reported that Emerald AI Conductor reduced a Santa Clara AI factory from **4 megawatts (MW) to 3 MW**, responded in under a minute, and handled more than **200 demand signals** while priority inference continued. (blogs.nvidia.com) These are vendor-reported observations, not independently obtained utility event records. The September 16 announcement of the **AI Energy Management Alliance (AEMA)** emphasized response speed, duration, predictability, and emergency behavior. (blogs.nvidia.com) Those principles support performance-based contracting; they do not themselves establish a utility tariff, certification, or guaranteed connection date. AEMA's own policy page similarly emphasizes verified performance rather than a particular technology. (www.aema.ai)

A credible commitment starts with protected inference capacity, restartable training, deferrable batch work, and facility constraints. Google's published implementation limits non-urgent compute and reschedules it after grid events. (cloud.google.com) Scheduler priority alone is insufficient: Kubernetes documents that a pod configured never to preempt others can still be preempted by higher-priority pods, while PyTorch warns against assuming checkpoint compatibility across versions. (kubernetes.io) (docs.pytorch.org) Settlement also depends on the program: the **California Independent System Operator (CAISO)** requires a scheduling coordinator and certified revenue-quality metering, with additional direct telemetry for ancillary-service participation. (www.caiso.com) (www.caiso.com)

This report proposes a technology-neutral acceptance method: define the contractual meter boundary; register workload constraints; measure a conservative curtailment envelope; test dispatch, communications loss, degraded capacity, sustained operation, and recovery; and deliver an auditable evidence pack. The quantitative context is substantial but remains distinct from site capability: the **International Energy Agency (IEA)** projects global data center consumption rising from **485 terawatt hours (TWh) in 2025 to 950 TWh in 2030** in its 2026 outlook. (www.iea.org) Operators should sign only for the reduction demonstrated under the agreed workload, thermal, telemetry, and failure conditions.

Introduction and Background

Demand response connects electricity-system needs to controllable consumption. OpenADR describes standardized exchanges of dynamic price and reliability signals between grid participants and energy-management systems. The signal is an input to the facility's policy; delivered curtailment is the measured outcome. (www.openadr.org) An AI workload being flexible in software does not establish that the utility meter will register an eligible reduction.

A **flexible interconnection** addresses a different decision: the conditions under which a site may connect and consume power. ASHRAE describes such agreements as a potential route to faster access to available capacity. (www.ashrae.org) A payment-oriented demand-response program and an import-limited connection can share controls, but their obligations must be compared explicitly. A project may need both an absolute import ceiling and an incremental reduction from an agreed baseline.

This distinction matters to operators, capacity planners, site reliability engineering (SRE) teams, facilities engineers, utilities, and procurement teams. Google's published demand-response design demonstrates one relevant operating pattern: defer non-urgent work while preserving service performance. Its report is evidence of an implemented approach, not a transferable guarantee for another fleet or tenant mix. (cloud.google.com)

GPU Smith's first-party scope describes workload sizing derived from throughput targets. That engineering perspective is relevant here because a power commitment must be reconciled with the useful work the fleet is expected to deliver. It does not make the consultancy a utility program, software platform, or interconnection authority. (gpusmith.com)

The following plan distinguishes **published evidence**, **vendor-reported observations**, and **proposed acceptance criteria**. The criteria are analyst recommendations for a specific project to negotiate and validate, not universal standards. ASHRAE's framework supplies recommended practices and describes operational constraints; the site's applicable tariff, connection agreement, equipment limits, and tenant contracts remain the governing inputs. (www.ashrae.org) An accountable facilities engineer should reconcile the electrical and thermal design, while procurement and appropriate counsel translate the demonstrated operating envelope into the actual agreement.

Key Changes

Production reporting is more concrete than a flexibility claim

NVIDIA's Santa Clara account explicitly says the deployment is **not a DSX Flex installation**. It describes Emerald AI Conductor operating at NVIDIA Eos. Operators should preserve that attribution rather than treating the report as proof that a particular DSX Flex release has passed their site's acceptance tests. (blogs.nvidia.com)

The report's claim of more than **200 signals** is useful evidence of repeated signaling, but it does not establish that there were that many separate settlement events. (blogs.nvidia.com) The reviewed report does not provide event-level durations, counterfactual baselines, complete response traces, or independent failure definitions. The commitment should therefore be sized from project evidence, with the public account serving as a reason to test rather than a performance warranty.

Performance principles need contract-specific thresholds

AEMA calls for ride-through, curtailment, and contingency-response obligations to be defined before connection. **Ride-through** means the ability to continue operating through specified electrical disturbances; it should be treated separately from a dispatch instruction to reduce demand. (blogs.nvidia.com) The operator should ask the utility to translate each principle into an observable condition: a response deadline, a sustained import envelope, allowed exceptions, and an agreed fallback.

The alliance's policy framing favors accelerated interconnection for flexible loads. That is an advocacy position, not an executed agreement granting a particular site additional capacity. (www.aema.ai) Procurement should require the utility's written treatment of the proposed flexibility before allocating any connection benefit to the investment case.

Orchestration interfaces leave important decisions with the site

DSX Flex documentation describes the exchange of power targets, timing, operating state, and exceptions without transferring control of the factory response. Factory-side systems determine how to operate within the target. (docs.nvidia.com) (docs.nvidia.com) A target interface therefore needs an implementation behind it: scheduler decisions, admission controls, equipment power settings, facility coordination, and tenant policy.

The current documented Flex contract does not define control interfaces for onsite generation, renewables, or battery storage. (docs.nvidia.com) Buyers should request written availability, version support, licensing, and integration terms separately; the reviewed product page offers a schema and documentation entry point, without supplying a complete procurement or service commitment. (www.nvidia.com) These details belong in the implementation schedule and acceptance scope.

Define the Curtailment Contract Before Sizing the Fleet

A curtailment commitment should be expressed as a **power-time-service envelope**, with the meter boundary and protected services stated beside the power figure in **kilowatts (kW)** or MW. AEMA's emphasis on verified speed, duration, and predictability identifies relevant dimensions but does not provide universal numeric

thresholds. (www.aema.ai) The project should negotiate the following fields:

- **Quantity:** an incremental reduction, absolute import ceiling, or both, with the settlement location identified.
- **Notification:** dispatch channel, acknowledgment, advance notice, cancellation, and conflicting-instruction handling.
- **Response:** deadline, measurement window, allowed overshoot, and acceptable ramp trajectory.
- **Duration:** maximum continuous obligation, minimum sustainable reduction, and conditions for extension.
- **Frequency:** permitted event clustering, seasonal availability, and the recovery time between events.
- **Recovery:** authorized ramp, backlog clearance, storage recharge, and cooling restoration.
- **Exceptions:** emergency behavior, communications loss, tenant priority changes, and manual override.
- **Consequences:** settlement treatment, under-delivery provisions, evidence disputes, and liability allocation.

Program examples show why these terms cannot be inferred from the phrase demand response. PG&E's **Base Interruptible Program (BIP)** offers notification options of at least **15 or 30 minutes**. (www.pge.com) Its published business-program page describes limits including **one event per day** and **six hours per event**. (www.pge.com) A site validated only for a short power step has not demonstrated this sustained obligation.

PG&E's **Capacity Bidding Program (CBP)** uses different operational arrangements, including remotely readable meters recording **15-minute intervals**. (www.pge.com) This is a settlement-meter example, not a recommended control-loop sampling rate. An event requiring fast response needs higher-resolution internal traces even where billing uses longer intervals.

Eligibility is equally specific. CAISO describes a scheduling coordinator and system-operator-certified revenue-quality meters for participation; its ancillary-services route adds special certification and direct telemetry. (www.caiso.com) (www.caiso.com) PJM describes retail customer participation through member **Curtailment Service Providers (CSPs)**. (www.pjm.com) Neither route should be assumed to apply to a site outside that operator's territory or without its required arrangements.

Storage and generation need explicit treatment. ASHRAE distinguishes outage-only backup storage from resources used continuously in microgrids. (www.ashrae.org) PG&E's BIP tariff limits incentives associated with prohibited resources. (www.pge.com) The contract should specify whether battery discharge, generator operation, and recharge are eligible, how they are metered, and whether emergency reserves constrain availability. Utility eligibility and the site's permitting conditions require local confirmation.

Implementation Considerations and Process Changes

Register flexibility at the workload level

The register covers **information technology (IT)** and facility loads. The **workload flexibility register** should be owned jointly by the platform team and each tenant's SLO owner. PyTorch's training guidance requires saving more than model weights for recovery, including optimizer state. (docs.pytorch.org) Distributed

Checkpoint supports loading across different cluster topologies, but that feature does not establish the application's usable restart time or checkpoint compatibility. (docs.pytorch.org) (docs.pytorch.org)

Table 1 proposes a register for **AI data center workload curtailment**. Every cell marked “measure” is a project input to establish through tests, not a default performance promise. Scheduler priority and application recovery are distinct, as Slurm's requeue rules make clear. (slurm.schedmd.com)

Workload class	Priority and SLO owner	Minimum GPU count	Checkpoint interval	Deferrable minutes	Maximum power step	Restart time
Real-time inference	Protected; service owner	Measure at peak admitted demand with headroom	Application-specific state; do not assume restartable sessions	Only within the latency and availability budget	Measure the step compatible with tail latency	Measure routing, warm-up, cache refill, and backlog
Batch inference	Deadline-based; tenant owner	Measure for the completion deadline	Output and progress durability as applicable	Derive from queue slack	Measure scheduler action plus meter response	Measure reload and resumed throughput
Checkpointable training	Deferrable by agreement; training owner	Validate the supported distributed topology	Record actual durable-save completion	Derive from training deadline and event clustering	Measure checkpoint, drain, and cap options	Demonstrate complete-state restore and stable progress
Non-checkpointable distributed jobs	Protect running jobs or admit only outside reserved windows; tenant owner	Preserve the job's supported configuration	No recovery assumption	Measure admission deferral; avoid assuming running-job flexibility	Count only tested actions with agreed lost-work treatment	Measure full restart, including work repeated
Data preparation	Deadline-based; pipeline owner	Record actual accelerator dependence	Record durable task boundaries	Derive from downstream dependency slack	Measure central processing unit (CPU), storage, and network effects	Measure dependency and queue recovery
Cooling and auxiliary loads	Facility-protected; facilities owner	Not applicable; record minimum service capacity	Not applicable	Derive from validated thermal margins	Measure within equipment and thermal limits	Measure temperatures, flow, and restoration

A registered inference SLO should distinguish **time to first token (TTFT)** from throughput and end-to-end latency. The original two-region experiment explicitly measured TTFT, providing a useful precedent for recording service behavior alongside power. (arxiv.org) The tenant should specify its protected tail and error measures before testing.

The register prevents installed GPU capacity from being mistaken for **available curtailment**. A highly utilized fleet can have little controllable demand if its running jobs are protected. Conversely, admission control may prevent a scheduled peak without evicting any job. Kubernetes documents that higher priority helps scheduling only when resource requirements can be satisfied, reinforcing the need to reserve feasible protected capacity. (kubernetes.io)

Make scheduler actions explicit

The checkpoint path needs its own completion semantics. PyTorch distinguishes the local staging operation from the signal that checkpoint saving has completed. (docs.pytorch.org) The proposed policy should permit resource release only after the application's durable-save condition is met.

The controller should select among distinct actions rather than treating “pause” as a single operation:

- **Admission deferral:** stop launching agreed flexible work while tracking its deadlines.
- **Graceful drain:** finish bounded tasks and reject new discretionary work.
- **Checkpoint and requeue:** confirm durable state, release resources, and restore against a tested environment.
- **Power capping:** apply supported settings and measure the resulting useful throughput.
- **Workload relocation:** verify destination capacity, data access, and service latency before shifting.
- **Controlled recovery:** resume work within the utility ramp and tenant deadline constraints.

Slurm's **SUSPEND** mode leaves memory allocated, and its scheduling configuration affects how suspended work resumes. (slurm.schedmd.com) Kubernetes gives preemption victims a termination period and can kill those that do not finish within it. (kubernetes.io) The acceptance plan should therefore test the configured behavior, not infer a graceful checkpoint from the scheduler action's name.

Google's described approach limits non-urgent compute during events and reschedules it afterward. (cloud.google.com) An enterprise implementation needs its own priority policy, deadline accounting, and tenant authorization, particularly where workload relocation is unavailable because of data locality or disconnected operation.

Measurement, Verification, and Control Design

Separate settlement from engineering telemetry

Measurement and verification (M&V) compares eligible delivery with the program's reference consumption. PG&E's published CBP tariff gives a concrete example: its weekday baseline averages corresponding hours from the preceding **ten similar weekdays**, with specified exclusions. (www.pge.com) A fast-growing AI facility should test how its proposed baseline behaves when utilization, workload mix, maintenance, and tenant occupancy change. A convenient pre-event reading is not automatically the settlement counterfactual.

Workload shifting also changes consumption after the event. Google's controller explicitly reschedules deferred tasks, supporting the inclusion of a recovery window in the engineering record even when settlement focuses on the dispatch period. (cloud.google.com)

Retain two coordinated evidence streams: the utility-approved meter record and higher-resolution internal traces. CAISO's separate requirements for revenue-quality metering and ancillary-service telemetry illustrate that these functions have different purposes. (www.caiso.com) (www.caiso.com) Internal observations should explain the meter response, including protected load, flexible work, storage flows, and cooling effects.

The proposed evidence schema should include:

- **Dispatch record:** event identifier, requested quantity, timing, acknowledgment, cancellation, and revisions.
- **Meter record:** boundary, units, sampling interval, quality flags, missing data, and time reference.
- **Workload record:** admitted jobs, priority changes, deferred work, checkpoints, and restore outcomes.
- **Service record:** throughput, tail latency, queue delay, errors, and availability against owner-approved SLOs.
- **Facility record:** intake temperatures, cooling flow, auxiliary power, storage state, and relevant reserve limits.
- **Audit record:** policy version, operator actions, exceptions, raw-data locations, and evidence hashes.

These are proposed audit fields, not claims that a particular utility accepts this schema. AEMA calls for standardized technical requirements, metrics, and operational data sharing, which supports negotiating the evidence format before commissioning. (blogs.nvidia.com)

Close the loop at the contractual boundary

NVIDIA documents supported GPU power-limit settings and distinguishes a GPU-only limit from wider module scope. (docs.nvidia.com) (docs.nvidia.com) GPU telemetry is consequently an actuator-level observation, not proof of facility-level delivery. The controller should reconcile measured import with the agreed target and retain a margin for untargeted CPUs, networking, storage, and cooling.

Cap persistence must also be tested: NVIDIA documents a module limit returning to its default after driver unload. (docs.nvidia.com) The control design should detect resets, restore approved settings, and report when the current fleet cannot maintain the target. DSX Flex documentation includes a breach-alert mechanism for that condition. (docs.nvidia.com)

Cooling control should follow equipment conditions. LBNL guidance recommends using IT intake temperatures rather than cooling-equipment return temperature as the control reference. (datacenters.lbl.gov) OCP's modular cooling guidance includes meter-based flow validation and telemetry handoffs. (www.opencompute.org) A power reduction should therefore trigger a coordinated assessment, not an automatic proportional reduction in pump or fan capacity.

Define communications and override behavior

OpenADR supplies interoperable signal formats; protocol support does not establish delivered MW, program eligibility, or baseline acceptance. (www.openadr.org) The interface test should verify authentication, timestamps, duplicate and expired commands, acknowledgments, and behavior during loss of connectivity.

NIST's operational technology (OT) guidance addresses performance, reliability, and safety alongside cybersecurity. (csrc.nist.gov) Its detailed guidance includes segmentation, isolation, logging, and monitoring. (nvlpubs.nist.gov) Apply those principles to utility gateways and facility controls, with limited privileges and an auditable path for manual intervention. A protected-service emergency override should produce an explicit exception record and the contractually required notification.

AI Data Center Curtailment Testing and Acceptance

The **curtailment testing plan** should progress from signal-only validation to bounded operational events, then sustained and degraded-condition tests. OCP cooling guidance includes system-validation plans and witnessed acceptance testing; this is a useful commissioning precedent rather than a universal demand-response certification. (www.opencompute.org)

The utility, platform owner, and facilities engineer should agree on the pass criteria before dispatch:

- **Dry run:** verify signal receipt, policy selection, time alignment, and evidence collection without changing load.
- **Controlled step:** demonstrate the requested reduction at the settlement boundary with protected services active.
- **Sustained event:** hold the commitment for the contract duration and observe thermal and workload drift.
- **Communications loss:** exercise missing, stale, duplicated, and canceled commands against the agreed fallback.
- **Degraded capacity:** repeat with a rack, fabric path, or capacity pool unavailable.
- **Recovery ramp:** resume deferred work and recharge storage within the agreed import trajectory.
- **Repeated events:** test the worst permitted clustering, checkpoint overhead, backlog, and recovery debt.
- **Operator override:** exercise authority, notification, and restoration without losing the audit trail.

An original AI-flexibility preprint reports tests lasting **two to ten hours**, demonstrating why sustained behavior deserves separate examination from the initial step. The authors' workload and hardware conditions bound that result. (arxiv.org) Kubernetes termination behavior and Slurm requeue eligibility are separate reasons to inspect jobs killed, completed checkpoints, and useful work restored. (kubernetes.io) (slurm.schedmd.com)

Table 2 gives the proposed **event acceptance record**. Entries are evidence fields, not manufactured test results.

Test	Requested vs delivered kW	Response seconds	Sustained duration	SLO breaches and jobs killed	Recovery ramp	Evidence hash and decision
Controlled step	Request, baseline method, meter trace, minimum delivered reduction	Dispatch-to-delivery under the agreed rule	Actual interval at target	Separate protected-service outcomes from authorized deferrals	Peak import and settled trajectory	Raw meter, scheduler, and SLO bundle; pass or rework

Test	Requested vs delivered kW	Response seconds	Sustained duration	SLO breaches and jobs killed	Recovery ramp	Evidence hash and decision
Sustained event	Delivery distribution and any target excursions	Initial response and subsequent adjustments	Actual continuous delivery	Queue drift, deadline misses, checkpoints, and restores	Backlog clearance and cooling restoration	Full event bundle with quality flags
Communication s-loss test	Actual import under the fallback state	Timeout and fallback-action timing	Period in defined fallback	Service effects and manually stopped jobs	Transition after connectivity returns	Signal, controller, and operator logs
Degraded-fleet test	Revised available reduction with unavailable capacity identified	Response under the degraded configuration	Hold time within the approved envelope	Protected-service headroom and recovery outcomes	Return after capacity restoration	Configuration, topology, and evidence bundle
Repeated-event sequence	Delivery for every event, including low-flexibility periods	Per-event response distribution	Each hold and inter-event interval	Cumulative lost work and unresolved deadlines	Recharge and restart trajectories across the sequence	Linked event bundles and consolidated decision

The table should contain actual measurements with consistent time references. Averages alone can hide a missed delivery interval or a tail-latency excursion. The inference experiment discussed below reports an average time-to-first-token effect; the proposed acceptance record deliberately adds tail and deadline measures because site contracts can protect different outcomes. (arxiv.org)

The proposed thermal record includes installed intake and onboard sensor observations. OCP describes thermal ride-through in terms of maintaining safe equipment temperatures, with duration dependent on heat load, thermal mass, and margin. (www.opencompute.org) OCP's liquid-cooling guidance also recommends thermal-cycling reliability tests. (www.opencompute.org) These sources support testing the installed system; they do not establish a generic cooling curtailment duration.

The initial response deadline should also be tested separately from the hold. The controlled preprint's **40-second** response result is an experimental observation, not a proposed deadline for every grid program. (arxiv.org) Agree the project's timing rule and evaluate it at the meter boundary.

The final release criterion should be the **lowest demonstrated feasible commitment** across the agreed test conditions. An unsuccessful staged test is a reason to revise scope, controls, or commitment before execution of the agreement. It should not be replaced by a larger theoretical figure derived from accelerator nameplate power.

Economics, Ownership, and the Contract Evidence Pack

Demand-response economics should value **useful work delivered**, not just electricity reduced. Research on **high-performance computing (HPC)** power capping notes that performance changes can lead to compensating GPU jobs, illustrating why a lower instantaneous draw does not automatically establish lower

total energy. (arxiv.org) PyTorch also notes that a general training checkpoint can be **two to three times larger** than the model alone, so restart planning must account for storage and transfer work. (docs.pytorch.org)

Use buyer-supplied values in the following proposed event contribution calculation:

Net event value = incentive + allocated avoided interconnection cost - lost-work value - restart energy cost - SLO risk provision - platform and telemetry cost.

The incentive follows the applicable program's settlement rules. PJM's published participation model, for example, routes retail customer payments through member CSPs. (www.pjm.com) Interconnection value should be allocated only from a written, project-specific utility position. ASHRAE describes potential faster capacity access, not a guaranteed schedule saving. (www.ashrae.org)

A revenue model should retain sensitivity to uncertain future utilization. IEA explicitly identifies uncertainty in current and future data center consumption; a sector growth forecast cannot establish a particular fleet's event availability or incentive income. (www.iea.org)

Lost-work value should include repeated computation, deadline consequences, and the opportunity cost of occupied capacity. Restart energy should include checkpointing, loading, warm-up, cooling response, and any storage recharge. The operator should analyze overlapping events and lower-than-expected flexible workload availability rather than valuing every contracted hour as an isolated event.

Table 3 proposes a **responsibility assignment matrix**, using responsible, accountable, consulted, and informed (RACI) roles. The allocations are recommendations to negotiate. Assign the vendor an interface and support responsibility, while keeping an accountable operator for the facility response.

Obligation or failure mode	Accountable	Responsible	Consulted or informed	Evidence and response
Dispatch definition or conflicting instruction	Utility/program authority for instruction; operator for response policy	Utility dispatcher and site gateway owner	Facilities, scheduler team, tenants	Approved command hierarchy, acknowledgment, and exception log
Protected workload exceeds reserved capacity	Tenant/service owner	Platform and scheduler team	Operator, utility as required	Admission policy, capacity trace, SLO record, commitment revision
Checkpoint cannot restore	Training owner	Application and platform teams	Vendor support, operator	Complete-state restore evidence; protect job or reduce commitment
Thermal margin becomes insufficient	Facilities owner	Cooling controls and facility operations	Platform team, utility as required	Intake/flow trace, safe operating action, event exception
Meter or telemetry quality loss	Operator	Metering and telemetry owners	Utility settlement team	Quality flags, missing-data treatment, auditable fallback
Controller loses connectivity or cap state	Operator	Controls and SRE teams	Utility, vendor, tenant owners	Tested local fallback, alert, and state reconciliation

Obligation or failure mode	Accountable	Responsible	Consulted or informed	Evidence and response
Recovery produces excessive import	Operator	Scheduler, storage, and facility teams	Utility and tenant owners	Authorized ramp, backlog trace, recharge schedule

The matrix exposes operational dependencies before they become contract assumptions. PyTorch's version-compatibility warning supports assigning restore validation to the application and platform owners. (docs.pytorch.org) NIST's OT guidance supports retaining facilities ownership where reliability and safety constraints shape control actions. (csrc.nist.gov)

The **contract evidence pack** should contain the signed workload register, meter-boundary diagram, approved baseline, control policy and version, telemetry dictionary, witnessed event records, exception procedures, recovery limits, tenant authorization, and economic assumptions. Procurement should connect each promised capability to its acceptance record and support obligation.

Data Analysis and Evidence

Distinguish scale forecasts from controllable capacity

IEA's 2026 outlook estimates **485 TWh** of global data center electricity consumption in **2025** and projects **950 TWh in 2030**. (www.iea.org) These cover data centers generally, not exclusively AI, and the latter is a forecast rather than measured consumption. The earlier IEA analysis estimated **415 TWh in 2024**, (www.iea.org) and projected approximately **945 TWh in 2030** in its Base Case. (www.iea.org) The update should not be described as a measured increase in demand-response capacity.

IEA identifies near-term server shipment projections as an important modeling input and explicitly acknowledges uncertainty in present and future consumption. (www.iea.org) (www.iea.org) Annual TWh forecasts consequently provide planning context, not the instantaneous MW or event duration a utility can rely on at a specific site.

As of **October 13, 2026**, the **Electric Power Research Institute (EPRI)** DCFlex homepage reports **nine active demonstration sites**, (dcflex.epri.com) while its demonstration page describes a target of **at least ten field projects**. (dcflex.epri.com) Active and planned are different categories; neither is a count of sites that have passed the acceptance plan proposed here.

Compare the public response observations carefully

NVIDIA's reported **4 MW to 3 MW** reduction implies **1 MW**, or **25%**, by arithmetic. (blogs.nvidia.com) The same production account does not disclose the complete event evidence needed for an independent settlement reconstruction. The arithmetic should not be converted into an annual energy saving, a guaranteed fleet capability, or a utility-approved interconnection benefit.

An author-affiliated preprint reports an experiment on a **130 kW GPU cluster**, with an approximately **30% reduction within 40 seconds**. (arxiv.org) It also tests requested reductions of **10% to 40%** for **two to ten hours**. (arxiv.org) Those are controlled-study results under specified workloads, not independent

measurements of the Santa Clara production event.

In its geographically shifted inference experiment, the authors report a sustained increase of approximately **30 ms** in average **time to first token (TTFT)** at the power-restricted Ashburn cluster. (arxiv.org) This is a concrete service trade-off to evaluate against a buyer's SLO; an average does not establish the tail, errors, or behavior under destination failure. Google's own report of scheduled reductions during **5 p.m. to 9 p.m.** winter peaks illustrates another workload-shifting pattern, but does not publish the achieved MW reduction or latency distribution on that page. (cloud.google.com)

Conservative available-curtailment calculation (Hypothetical Example)

The following numbers are **analyst-selected illustrative inputs**, not a measured site result or an external benchmark. Use the equation to replace them with buyer evidence:

Available curtailment = baseline facility kW - protected IT kW - protected cooling/auxiliary kW - uncertainty reserve kW, floored at zero.

Assume a facility baseline of **4,000 kW**, protected IT of **2,300 kW**, protected cooling and auxiliaries of **700 kW**, and an additional uncertainty reserve of **250 kW**. The central illustrative result is **750 kW**. Assume separately that measured baseline uncertainty is plus or minus **100 kW**, and the protected IT and auxiliary estimates each have plus or minus **100 kW** bounds. The resulting conservative arithmetic interval is **450 to 1,050 kW**: lower baseline minus higher protected loads gives the lower endpoint.

These bounds are an **assumed uncertainty interval**, not an empirically established statistical confidence band. The illustrative commitment would start no higher than the lower endpoint, then be reduced further if sustained tests, recovery, or degraded operation show less available flexibility. Establish statistical confidence from actual meter and workload observations, with the chosen baseline and sampling procedure documented. The counterfactual must be an agreed program-specific input, using the tariff methodology discussed earlier.

The equation is a screening limit, not proof that all residual consumption is controllable. The workload register and event tests must show which actions actually remove the demand. NVIDIA's distinction between GPU and wider module power limits reinforces the need to close this calculation at the facility meter. (docs.nvidia.com)

Implications and Future Directions

The useful procurement shift is from buying a flexibility feature to buying **demonstrated operating behavior**. AEMA's call for standardized metrics and operational data sharing creates a basis for comparable evidence, but utility-specific acceptance and settlement still have to be defined. (blogs.nvidia.com) A site should be able to explain a delivery trace without depending on a vendor dashboard alone.

For platforms, **AI workload flexibility for grid demand response** becomes a scheduling policy linked to service ownership. Capacity planners should reserve protected resources, measure flexible deadline slack, and account for the backlog created by event clustering. Slurm's partition-specific gang scheduling behavior illustrates why configured topology and scheduler policy matter to the available actions. (slurm.schedmd.com) PyTorch's requirement for more than weight-only state is a concrete reason to include application-owned restore evidence. (docs.pytorch.org) Neither a priority label nor a checkpoint file is a sufficient acceptance record.

For facilities teams, cooling and power controls become joint constraints. The recommended responsibility matrix gives facilities owners authority over thermal margins and telemetry handoffs. Repeated-event qualification should include storage recharge and thermal recovery as well as useful-work completion.

For project finance, the critical distinction is **modeled opportunity versus contracted capacity**. EPRI's demonstration program is testing multiple approaches; its active-site count indicates activity, not a universal demonstrated availability or return on investment. (dcflex.epri.com) Connection timing benefits should remain conditional until the utility approves the relevant terms.

GPU Smith's published scope includes capacity planning, telemetry, failure-mode analysis, and operating procedures. Its adjacent engineering perspective is therefore to require reproducible workload and facility evidence before sizing a commitment, while leaving eligibility and connection authority with the applicable utility and program. (gpusmith.com) This report makes no claim that the consultancy has validated the cited deployments or secured a specific interconnection outcome.

Conclusion

AI data center demand response is credible when the operator can demonstrate a defined reduction at the contractual meter boundary while meeting protected-service and recovery obligations. ASHRAE's description of flexibility as partial and situational is the right starting assumption. (www.ashrae.org) The commitment should follow measured workload behavior and local program rules.

The proposed **go/no-go rubric** is straightforward:

- **Go:** workload owners approve the policy; metering and eligibility are accepted; staged tests demonstrate the quantity, response, duration, degraded operation, and recovery; evidence and support obligations are complete.
- **Conditional go:** a smaller commitment or restricted availability window is demonstrated, with those limits written into the agreement.
- **No-go:** required delivery depends on untested restore behavior, unapproved resource use, unresolved telemetry, insufficient protected-service capacity, or an unvalidated recovery ramp.

The technology remains useful only within that envelope. PyTorch documents why checkpoint compatibility requires attention, and CAISO demonstrates that metering and participation prerequisites are independent of the workload controller. (docs.pytorch.org) (www.caiso.com) Acceptance should connect those layers rather than approve them as isolated components.

Before signing a flexible interconnection agreement, the buyer should possess the workload register, meter and baseline definition, control and fallback policies, witnessed event results, responsibility matrix, and a buyer-valued economic model. AEMA's emphasis on preconnection curtailment and contingency obligations supports making that evidence a prerequisite. (blogs.nvidia.com) A smaller auditable commitment provides a clearer basis for planning than a larger estimate whose service, thermal, and recovery assumptions have not been tested.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.ashrae.org/technical-resources/topics-and-initiatives/ai-data-center/grid-interactive-design-demand-flexibility
2. https://blogs.nvidia.com/blog/from-megawatts-to-tokens-how-nvidia-maximizes-ai-factory-production/
3. https://blogs.nvidia.com/blog/ai-energy-management-alliance/
4. https://www.aema.ai/solutions
5. https://cloud.google.com/blog/products/infrastructure/using-demand-response-to-reduce-data-center-power-consumption
6. https://kubernetes.io/docs/concepts/scheduling-eviction/pod-priority-preemption/
7. https://docs.pytorch.org/docs/2.14/distributed.checkpoint
8. https://www.caiso.com/generation-transmission/load/demand-response
9. https://www.iea.org/reports/key-questions-on-energy-and-ai/executive-summary
10. https://www.openadr.org/faq
11. https://gpusmith.com/
12. https://docs.nvidia.com/dsx/flex/overview
13. https://www.nvidia.com/en-gb/data-center/products/dsx/flex/
14. https://www.pge.com/en/save-energy-and-money/energy-saving-programs/demand-response-programs/business-programs.html
15. https://www.pjm.com/markets-and-operations/demand-response.aspx
16. https://www.pge.com/tariffs/assets/pdf/tariffbook/ELEC_SCHEDS_E-BIP.pdf
17. https://docs.pytorch.org/tutorials/beginner/saving_loading_models
18. https://slurm.schedmd.com/preempt.html
19. https://arxiv.org/html/2606.25098v1
20. https://www.pge.com/tariffs/assets/pdf/tariffbook/ELEC_SCHEDS_E-CBP.pdf
21. https://docs.nvidia.com/deploy/nvidia-smi/index.html
22. https://datacenters.lbl.gov/sites/default/files/FINAL%20Thermal%20Guidelines%20and%20Temp%20Measurements%209-15-2020.pdf
23. https://www.opencompute.org/documents/ocp-modular-tcs-rev-1-final-2025-pdf
24. https://csrc.nist.gov/pubs/sp/800/82/r3/final
25. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-82r3.pdf
26. https://www.opencompute.org/documents/oai-system-liquid-cooling-guidelines-in-ocp-template-mar-3-2023-update-pdf
27. https://arxiv.org/abs/2402.18593
28. https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai
29. https://dcflex.epri.com/
30. https://dcflex.epri.com/demonstrations

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.