PowerX ENERGY BLADE

Converting AI Infrastructure into a Grid-Native Asset

PowerX Energy Blade Whitepaper v1.0, May 2026

Core thesis

By using high power, high-capacity energy storage, coupled with isolated power converters with layered mechanisms for compute management in data centers, AI infrastructure can respond to grid stress and renewable curtailment without reducing delivered compute.

Power becomes a controllable grid variable; throughput and customer-visible service do not have to move in lockstep.

Table of contents

1. Defining the problem

Data centers use massive amounts of energy that utilities cannot always provide without grid upgrades; renewable curtailment and grid congestion are becoming direct barriers to AI deployment.

AI compute is becoming a large, constant, high-value electrical load, but the grid is planned and operated around scarcity, congestion, and intermittency. The result is a mismatch: AI infrastructure wants firm power now, while utilities often need time, upgrades, and controllable demand.

AI compute represents one of the largest new sources of electricity demand the grid has seen in a generation. A single AI training cluster can draw 50 to 100 MW continuously, while inference fleets may be smaller individually but are proliferating rapidly across regions. The core challenge is not electricity consumption alone. It is that this consumption is largely inflexible: compute loads expect power to be available continuously, regardless of grid congestion, renewable output, peak demand, or local system stress.

Why conventional data centers create a grid bottleneck

- They create large step-changes in demand, often faster than generation, transmission, and distribution infrastructure can be planned or built.
- They expect firm, high-availability capacity even when the local grid is constrained during peak load or under-frequency events.
- They are poor matches for renewable surplus: when solar or wind generation exceeds grid demand or transmission capacity, that energy is curtailed rather than absorbed by useful work.
- They monetize compute but usually do not monetize flexibility, so the grid sees cost and risk while the operator sees only capacity shortage.

Three visible failures

Failure mode	Implication for AI infrastructure
Waiting in line for power	In datacenter-heavy regions, new grid connections can take years. Land and capital wait for utility upgrades.
Paying for being inflexible	Operators are exposed to peak-hour tariffs and local grid constraints they could avoid if they had controllable load.
Wasting clean energy	Renewable surplus can be curtailed when no load can absorb it quickly or reliably enough.

What makes this more urgent for AI

AI clusters differ from many industrial loads because the business value of each watt can be extremely high, while the grid value of flexibility can also be high. A useful architecture should preserve both forms of value: it should let compute continue serving customers while allowing the grid to call on the load when it needs help.

In the current architecture, these two values are often treated as mutually exclusive. The data center either consumes power and delivers compute, or reduces power and compromises throughput or latency. That is the constraint this white paper is designed to challenge.

2. Our thesis

A battery-buffered, isolated power converter with specific layered mechanisms for compute management can adapt to grid incidents and renewable curtailment without losing delivered compute.

Thesis

Our thesis is that AI infrastructure can become a grid-native asset if the compute layer is paired with fast, bidirectional energy storage and a controller that understands both grid state and inference state.

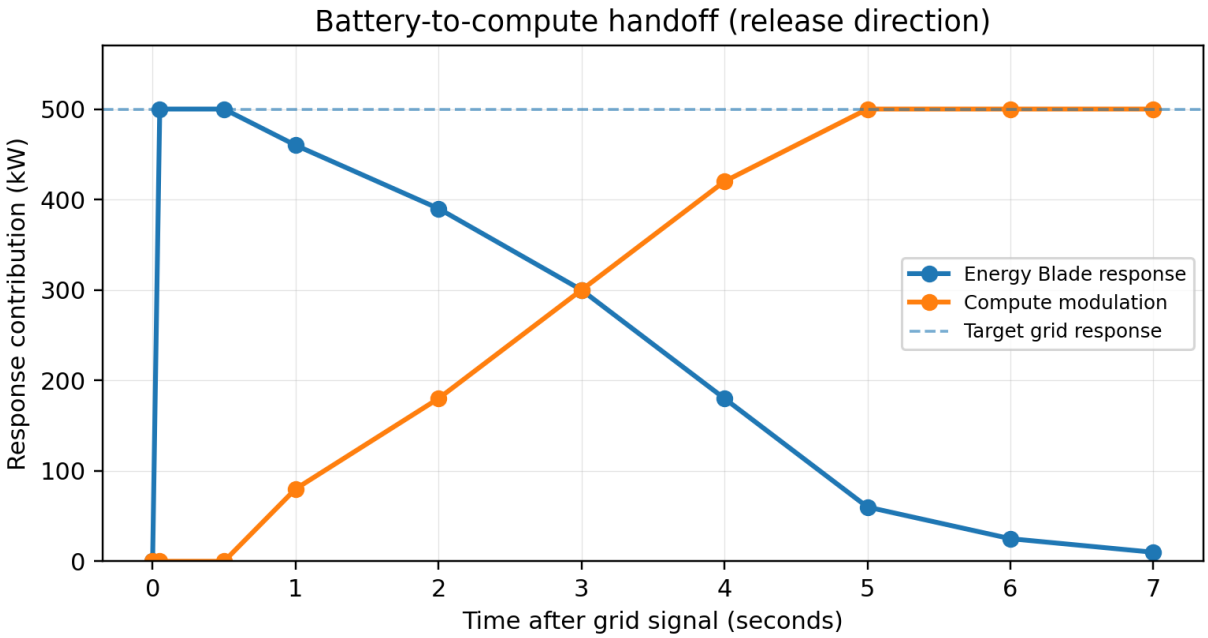
The key claim is simple: power does not have to equal throughput. Conventional controls reduce power by reducing work. PowerX Energy Blade (Energy Blade) proposes two layers that break that link: a compute mechanism toolkit that changes how work is routed and executed, and an Energy Blade layer that supplies or absorbs power while compute adjusts.

Two directions, one architecture

Grid condition	Required response	Energy Blade response	Intended outcome
Grid stress / under-frequency	Release power	Energy Blade discharges in milliseconds while compute consolidates, caps, or reroutes work over seconds.	Grid receives a fast reduction in net draw without an immediate throughput loss.
Renewable surplus / over-frequency	Absorb power	Energy Blade charges in milliseconds while compute can increase useful demand or shift work to the surplus period.	Curtailment becomes useful energy consumption rather than wasted generation.

Energy storage buys time

Grid frequency response is measured in milliseconds. Compute mechanisms are fast by software standards, but still operate over routing, scheduling, and convergence timescales. Our Energy Blade bridges that gap: the battery and inverter respond immediately, then the compute system reshapes work, so the battery does not have to carry the full duration.



Conceptual handoff between storage and compute: the Energy Blade provides the immediate grid response, covering the transient period while compute modulation ramps in to sustain the adjustment. The release direction is shown; the absorb direction follows the same architecture in reverse.

The compute mechanisms that preserve value

The document highlights two measured mechanisms that can reduce power without reducing delivered work: fleet consolidation and mixed-precision routing. Fleet consolidation moves traffic onto fewer busy servers and idles the rest. Mixed-precision routing sends suitable requests to lower-precision model copies that have been validated as quality-equivalent for the relevant task.

Together, these make the data center more than a flexible load. A flexible load can turn down. A grid-native asset can release and absorb power on grid timescales while preserving the economic value of the compute workload.

3. Our experiment

The experiment measured how real AI inference hardware behaves under controlled load and how grid-response mechanisms can be characterized.

Purpose

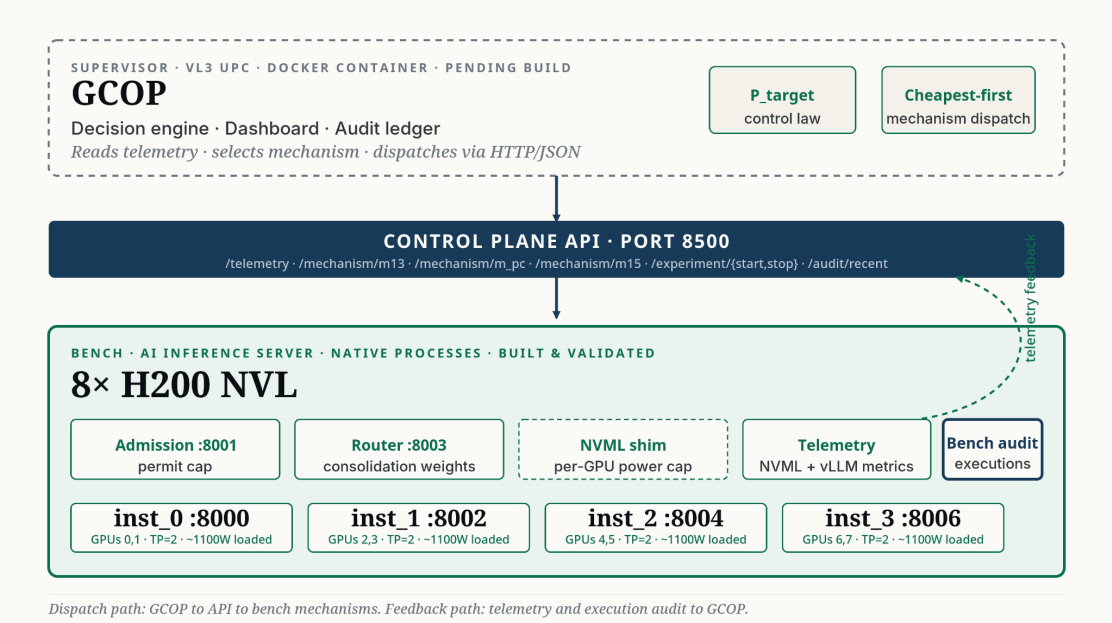
The experiment was designed to measure the relationship between AI workload shape, GPU fleet power, throughput, and controllability. The goal was not to simulate a production data center, but to build a precise measurement rig that could produce controller-ready data.

Bench setup

Component	Description from source
Site	POWERD Lab Tokyo[*1], Phase A
Compute fleet	8 x H200 NVL inference server
Model	Llama-3.1 70B BF16
Serving configuration	Four tensor-parallel instances, TP=2; vLLM metrics used for telemetry
Grid layer	Programmable grid simulator imposes frequency profiles; bidirectional inverters and battery layer release or absorb power
Controller concept	GCOP: Grid Compute Orchestration Platform (internal naming for this system). Decision logic reads telemetry, selects mechanisms, dispatches setpoints, and writes audit records

[*1] POWERD Lab Tokyo： PowerX's R&D facility

Test bench system architecture



Dispatch path: GCOP to API to bench mechanisms. Feedback path: telemetry and execution audit to GCOP.

Protocol

Step	Action
1. Select test case	Choose request shape, concurrency level, grid stress profile, and mechanism under test.
2. Establish baseline	Serve steady-state traffic until power, throughput, and latency converge; record baseline.
3. Inject grid event	Programmable grid simulator drives under-frequency for release or over-frequency for absorb; event is timestamped.
4. Bridge with storage	Energy Blade discharges or absorbs to cover the full target while compute is initially unaffected.
5. Activate compute mechanism	Controller activates consolidation, precision routing, per-GPU cap, admission control, or combinations; battery hands off as compute reshapes.
6. Return and audit	Grid returns to nominal and all dispatches, telemetry, measured power, throughput, latency, and quality scores are saved.

Measured variables

- Fleet AC or server power draw under each request shape and concurrency level.
- Throughput in requests per second.
- Latency and customer-facing service quality, where available.
- Mechanism response time, reversibility, and value cost.
- Battery dispatch timing and compute handoff timing for cross-domain response.
- Audit trail of decisions and execution records for reproducibility.

What was actually run

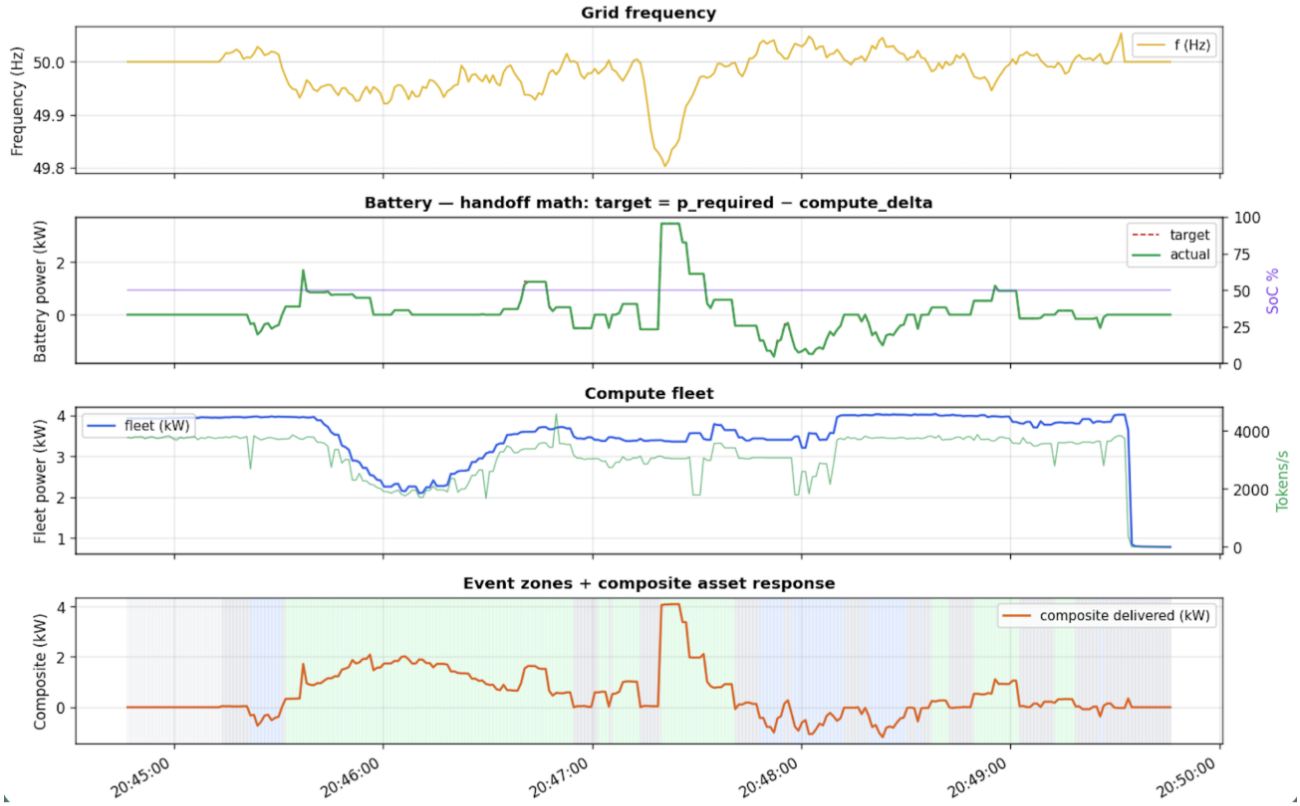
The clearest complete data set in the HTML is the 25-run Llama-3.1 70B fleet map: five request shapes across five concurrency levels, each measured as a 60-second steady-state sample May 2026. The source also includes measured or planned status for individual mechanisms and an empirical M-13 validation at smaller scale.

The shape of the requests used:

requests/sec at →	32 users	64 users	128 users	160 users	256 users
short input / short output	21.2	42.5	63.8	63.9	63.8
medium input / short output	33.0	82.5	125.4	125.8	125.5
long input / short output	15.0	23.4	25.6	28.1	26.1
short input / long output	21.1	42.5	63.6	63.4	63.5
short input / very long output	21.0	42.4	63.4	63.4	63.5

A single 5min long standard Japan grid realtime input looped :

The live measurement data:



Limitations of the workbench:

This demonstration runs on a single-rack Phase-A bench at POWERD Lab Tokyo[*1], consisting of a custom AI server with 8× H200 NVL GPUs, paired with an LFP battery cluster and inverter stack behind a grid emulator. The bench is instrumented using off-the-shelf interfaces, including HTTP-based telemetry, NVML for GPU state, IPMI/RAPL for node power, and Modbus TCP to the PLC. These interfaces establish a working baseline, but they also limit achievable response latency and telemetry resolution.

Production-grade performance will be enabled by the Energy Blade Module / Energy Blade Rack (EBM/EBR), which is purpose-built for in-band power coordination, together with emerging industry standards such as CXL fabric attach and OCP DC-bus rack power. Mechanism performance has been validated for Llama-3 70B and 8B using Mooncake-style trace replay. The frequency profile is OCCTO-derived but replayed in a closed-loop environment. Behavior under multi-rack aggregation, live-grid interconnection, and sustained mixed-tenant traffic remains within the scope of Phase B.

4. Findings

The source data shows that loaded GPU fleet power sits inside a narrow envelope while throughput varies sharply by request shape. With only software that relies on classifiers and SLA, the limit of modulation is capped at 24% per our experimentation.

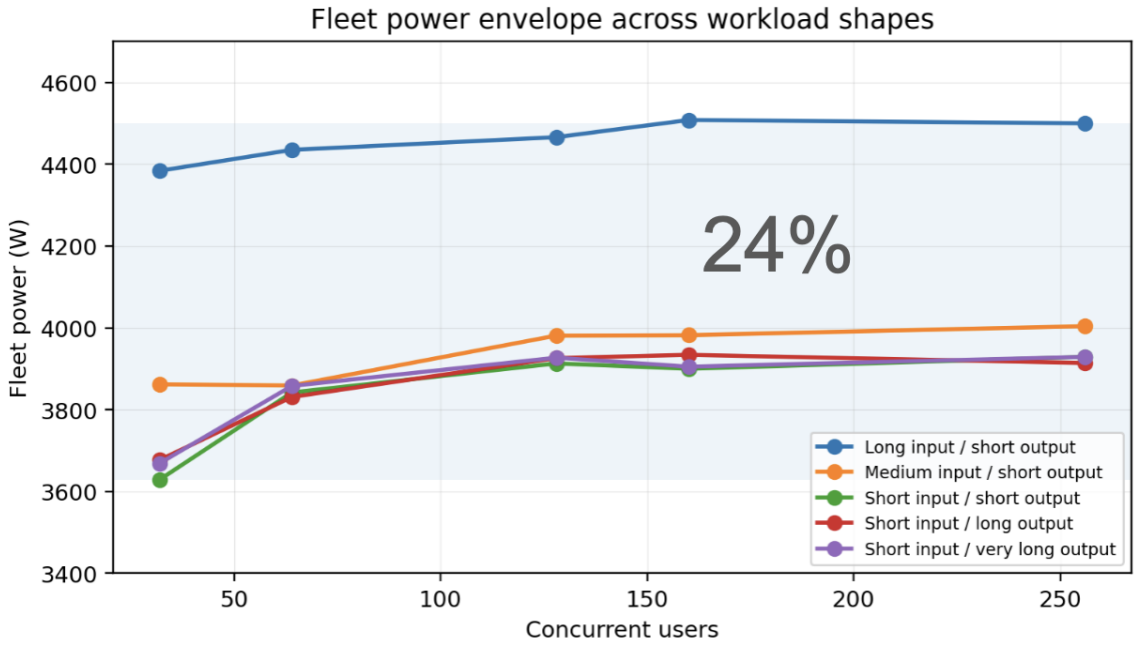
[*1] POWERD Lab Tokyo： PowerX's R&D facility

Data summary

The current data supports four early findings: first, once GPUs are busy, fleet power varies less than expected; second, throughput is highly sensitive to request shape; third, simple admission control has limited depth because busy GPUs maintain high draw; fourth, mechanisms that change which GPUs are active or which precision tier serves a request can move power without necessarily moving delivered work.

Finding 1: loaded power is a narrow band

Across the 25-run 70B data set, loaded fleet power ranges from 3,629 W to 4,500 W. That is roughly a 24% band. In practical terms, it means reducing traffic alone may not reduce power deeply once GPUs are already active.



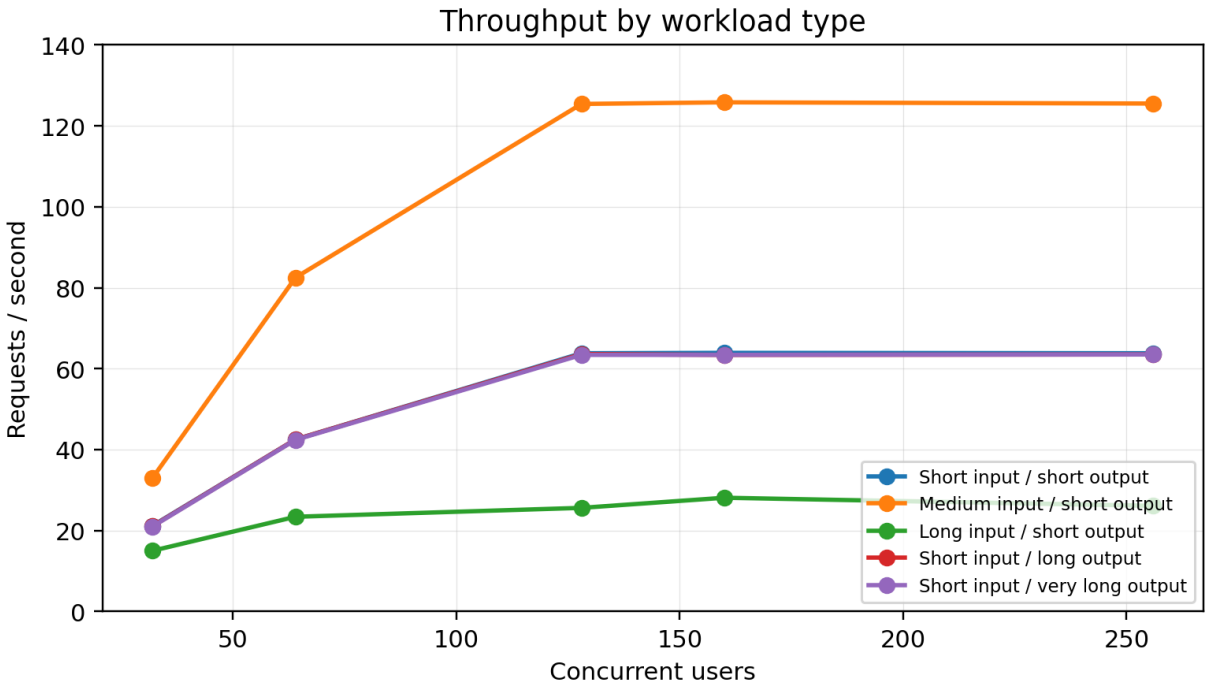
Fleet power envelope across five request shapes and five concurrency levels.

Workload shape	32	64	128	160	256
Long input / short output	4384	4435	4466	4508	4500
Medium input / short output	3862	3859	3981	3982	4004
Short input / short output	3629	3842	3913	3900	3929
Short input / long output	3677	3831	3926	3934	3914
Short input / very long output	3669	3858	3927	3905	3929

Fleet power in watts. Each test was a 60-second steady-state sample on the 8 x H200 NVL Llama-3.1 70B fleet.

Finding 2: request shape decides throughput

The same hardware and concurrency levels produced materially different throughput depending on the mix of prefill and generation work. The source figure reports a fastest workload type above 125 requests per second and a slowest workload type near 26 requests per second at high concurrency.



Throughput by workload type. Data recovered from Figure 04 in the HTML source.

Workload shape	32	64	128	160	256
Short input / short output	21.2	42.5	63.8	63.9	63.8
Medium input / short output	33.0	82.5	125.4	125.8	125.5
Long input / short output	15.0	23.4	25.6	28.1	26.1
Short input / long output	21.1	42.5	63.6	63.4	63.5
Short input / very long output	21.0	42.4	63.4	63.4	63.5

Throughput in requests per second across concurrency levels.

Finding 3: the mechanism catalogue matters

No single control is enough on its own. Reducing power while preserving compute value requires a coordinated set of small adjustments, each with a different speed, depth, and impact on the workload.

Energy Blade treats these adjustments as a control toolkit. Some actions are light-touch and low-cost, such as shifting requests across servers or routing suitable tasks to more efficient model versions. Others are deeper interventions, such as power capping or broader workload reshaping. The controller can start with the least disruptive actions first, then layer in stronger measures only when the grid event requires a larger response.

This matters because flexibility is not just about how much power can be moved. It is also about how quickly, how reversibly, and at what cost to latency, throughput, or model quality. The goal is to create a smooth grid response while using the lowest-cost compute actions available.

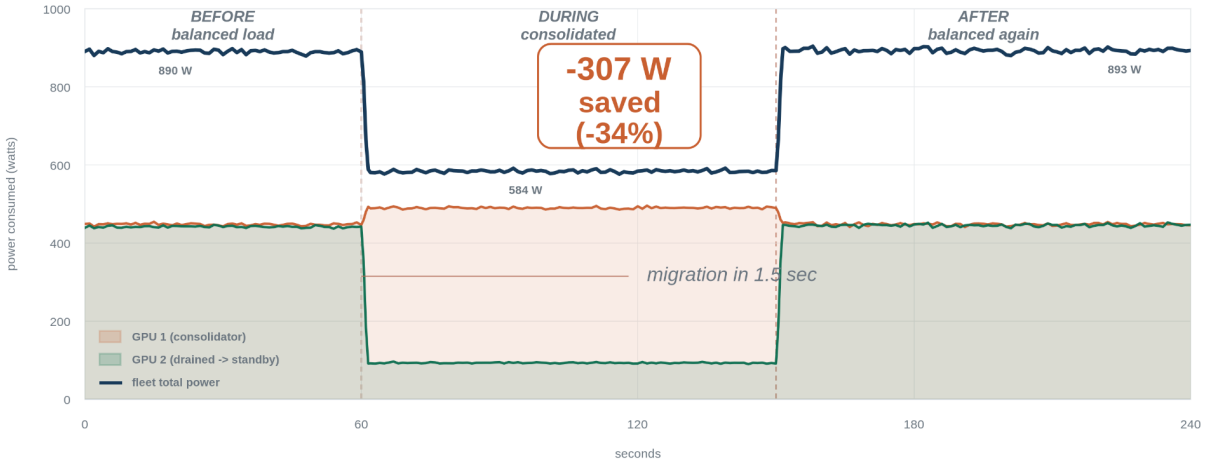
Mechanism	What it does	Speed	Value cost / concern
Fleet Consolidation	Re-weight request routing so traffic runs on fewer servers and empty GPUs drop toward idle.	Under 1 second to redirect; 10-30 seconds for power to settle	Throughput can hold if remaining servers have headroom; latency may rise.
Per-GPU Power Cap	Set firmware-enforced watt ceiling on each GPU through the NVIDIA driver.	Sub-second enforcement	Deep caps reduce throughput; useful as a deeper dial after cheaper mechanisms.
Admission Control	Cap concurrent requests at the front door.	Immediate on next request	No quality impact, but requests may wait or be rejected; source cites 5-15% power savings.
Mixed-Precision Routing	Run model copies at different precisions and route each request to the right precision tier.	Immediate routing decision	Can reduce power per request while preserving quality where lower precision is benchmark-equivalent.
Workload Sharding	Separate urgent interactive work from background/batch work and modulate lower-priority pools first.	Immediate policy change	Background work runs later; urgent workloads protected.
KV Cache Pre-Warm	Pre-load working memory for protected workloads when a grid event is forecast.	Requires a few seconds of lead time	Some speculative work may not be used.
Compute-Battery Bridge	Energy Blade responds in under 100 ms while compute reshapes over seconds.	Sub-100 ms both directions	Battery wear and coordination complexity; essential for grid-timescale response.

Mechanism catalogue

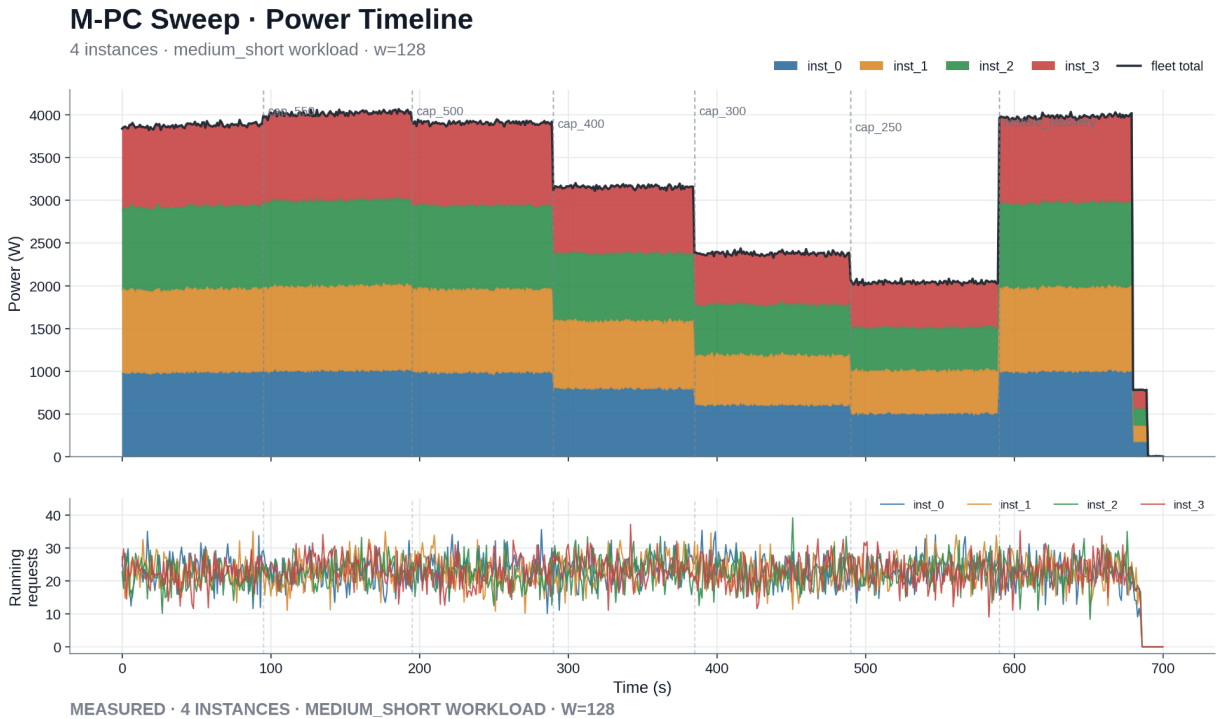
MEASURED · 2× H200 NVL · LLAMA-3 8B · MAY 2026

M-13 Fleet Consolidation

Same workload. Same throughput. One GPU asleep, the other doing both jobs.



Live experiment on 2× NVIDIA H200 GPUs running Llama 3 8B · Synapse bench · May 2026



MEASURED · 4 INSTANCES · MEDIUM_SHORT WORKLOAD · W=128

PER-INSTANCE DETAIL

INST	GPUS	POWER	RUNNING	TOKENS/S	REQ/S
inst_0	[0,1]	833W	27	854.7	29.83
inst_1	[2,3]	895W	15	858.7	31.45
inst_2	[4,5]	895W	29	905.5	24.99
inst_3	[6,7]	992W	26	932.9	26.61

PER-GPU DETAIL

GPU 0 51°C 395W cap 400W · util 100%	GPU 1 53°C 438W cap 450W · util 100%	GPU 2 53°C 447W cap 450W · util 100%	GPU 3 52°C 447W cap 450W · util 100%
GPU 4 53°C 447W cap 450W · util 100%	GPU 5 52°C 448W cap 450W · util 100%	GPU 6 57°C 496W cap 500W · util 100%	GPU 7 54°C 496W cap 500W · util 100%

TELEMETRY SNAPSHOT · 4 INSTANCES · 8 GPUS · MEDIUM_SHORT WORKLOAD · W=128

MEASURED · 2× H200 NVL · LLAMA-3 8B · MAY 2026

M-13 Fleet Consolidation Validation

Empirical validation across 20× workload variation

Six experiments · Llama 3 8B Instruct · 2× H200 NVL · vLLM 0.20.1 BF16 · prefix-cached

Power saved = 281–307 W across all loads

What the data does not yet prove

These results should be treated as a strong early demonstration, not yet as a complete production-scale proof. The current data shows that AI compute power can be measured, shaped, and partially decoupled from delivered work. It also gives us a clear view of the operating envelope of the GPU fleet and the first set of mechanisms that can move power without simply shutting down compute.

In short, the experiment shows that the architecture is promising. The next step is to prove that the same behavior holds when the full system is operating continuously, at larger scale, under real grid and customer constraints.

5. Our thoughts and extrapolations

The early data suggests a path from flexible AI load to grid-native AI asset, but production-scale economics and grid qualification still need to be tested.

What we think the data means

The most important implication is that AI power flexibility should not be treated as a simple load-shedding problem. The loaded GPU envelope is too narrow for traffic reduction alone to provide deep, high-value grid response. The useful design space is a composed system: storage for grid-timescale response, routing for customer-value preservation, and compute mechanisms for sustained power modulation.

Extrapolation 1: storage should sit close to compute

The closer the storage is to the compute load, the more precisely it can control the net grid interface. Storage near the compute layer can answer fast grid signals while the controller makes workload-aware decisions. This is different from a remote battery that only arbitrages energy at the site boundary; it is a battery integrated into the operating model of the compute fleet.

Extrapolation 2: flexibility can unlock more than revenue

A grid-flexible AI site may have a different interconnection profile than a conventional constant load. If the load can reduce draw during constrained periods and absorb surplus when asked, it may support more compute on existing connections, make constrained edge or sovereign sites more viable, and support non-firm or flexible-load discussions with utilities.

Extrapolation 3: the business case has three revenue layers

Revenue layer	Interpretation
AI inference workloads	Primary compute revenue from serving AI customers. The control system must preserve quality, latency, and availability.
Grid services	Potential frequency response, capacity, demand response, and ancillary services revenue where market rules allow fast bidirectional flexibility.
Energy arbitrage	Charge or absorb when energy is cheap/surplus and discharge or reduce draw when expensive/scarce; can also be carbon-aware.

Extrapolation 4: Unlock layers of AI deployment

Deployment	Interpretation
More compute on existing connections	Non-firm uplift on contracted capacity. Existing grid connections can host more compute and new interconnections are not required for scale.
Edge and sovereign AI deployment	Making distributed AI deployments possible, including sovereign sites where data must remain in-country. Flexibility makes local data easier to deploy.
Faster grid connection approvals for AI Infra	A flexible AI load presents an easier planning case for grid operators leading to fast-tracking interconnection approvals in areas where non-firm connections are allowed.

6. Looking to the future

The next step is to move from mechanism characterization to integrated, bidirectional, production-relevant demonstrations.

Near-term technical roadmap

Roadmap item	Purpose
1. Complete 70B mechanism validation	Run per-GPU cap, fleet consolidation, admission control, and mixed-precision routing on the full 70B setup with latency, quality, and power telemetry.
2. Integrate the compute-battery bridge	Demonstrate measured battery response, compute handoff, and audit logging on the same time axis.
3. Prove absorb direction	Run over-frequency or renewable-curtailment-style events where the system absorbs energy without disrupting compute service.
4. Build controller automation	Move from scripted runs to GCOP-driven dispatch with cheapest-first mechanism selection and immutable audit records.
5. Pilot at rack or site scale	Validate hardware integration, interconnection assumptions, thermal behavior, and grid-facing telemetry in a production-like environment.

Future product direction

The long-term opportunity is a data center architecture where compute, storage, inverter control, and workload routing are sold as one grid-native platform. In that model, AI infrastructure is not only a consumer of scarce firm capacity; it is an asset that can participate in grid balancing, absorb curtailed renewables, and accelerate deployment in constrained regions.

Closing

If AI demand continues to grow faster than grid infrastructure, the limiting factor for compute will not be model architecture or GPU supply alone. It will be power availability. Energy Blade reframes the data center as a controllable, bidirectional grid asset: one that can keep serving compute customers while also helping the grid operate with more renewables, less congestion, and faster deployment of AI infrastructure.