nature energy

Article

https://doi.org/10.1038/s41560-025-01927-1

# AI data centres as grid-interactive assets

Received: 3 July 2025

Accepted: 28 October 2025

Published online: 05 December 2025

Check for updates

Philip Colangelo[1], Ayse K. Coskun[1] ✉, Jack Megrue[1], Ciaran Roberts[1], Shayan Sengupta[1], Varun Sivaram[1], Ethan Tiao[1], Aroon Vijaykar[1], Chris Williams[1], Daniel C. Wilson[1], Brandon Records[2], Zack MacFarland[3], Daniel Dreiling[4], Nathan Morey[4], Anuja Ratnayake[5] & Baskar Vairamohan[5]

The exponential growth in electricity demand driven by artificial intelligence (AI) is threatening grid reliability, increasing energy costs for communities funding new infrastructure and slowing AI innovation as data centres await interconnection to constrained grids. Here we present a field demonstration of a software-based method that enables AI data centres to operate as flexible grid resources. Tested on a 256-Graphics Processing Unit (GPU) cluster running representative AI workloads in a hyperscale cloud facility in Phoenix, Arizona, the system reduced power usage by 25% for 3 hours during peak demand while maintaining AI quality of service guarantees. By coordinating workloads in response to real-time grid signals, without hardware modifications or energy storage, this approach demonstrates the potential for data centres to contribute to grid stability and affordability while sustaining computational performance within existing power-system constraints.

The global proliferation of artificial intelligence (AI) technologies has led to surging demand for high-performance data centres, increasingly powered by Graphics Processing Unit (GPU) clusters[1]. As these AI clusters scale, their energy consumption poses a growing strain on power grids[2]—particularly during periods of high demand or low renewable output. In the USA alone, projections estimate that AI-related data centre demand could reach tens of gigawatts by 2030, exacerbating grid congestion and delaying project deployments[3,4]. This growing demand, when paired with today's conservative power-system planning processes that assume data centres are continuously high-power customers, lead to large and expensive grid-infrastructure upgrades and long delays to connect new data centres.

Historically, demand response of computational load in data centres has been explored in academic settings, mostly using Central Processing Unit (CPU)-based clusters running high-performance computing (HPC) applications[5] or demonstrating the potential of demand response via simulation and analytical models[6,7]. These studies provided valuable insights but did not account for the rigid performance demands and distinct energy profiles of AI training and inference workloads on GPUs. Other work in industry has demonstrated that data centres can reduce operational carbon emissions by allocating fewer compute resources to jobs broadly classified as having flexible performance needs or via load shedding[8–10].

Scalable, software-based solutions that respect AI service-level agreements (SLAs) while offering real-time power modulation are urgently needed. Our central hypothesis is that GPU-driven AI workloads contain enough operational flexibility—when smartly orchestrated—to participate in demand response and grid stabilization programmes[11]. Although utilities offer financial incentives for power flexibility, other adoption costs, such as impacts on workload performance and delays in deploying new data centres, can limit participation[12], especially amid rapid AI growth. Utilities and system operators can further prioritize flexible AI data centres for accelerated interconnection and offer them lower tariffs and flexibility payments, recognizing their benefits to system reliability and ability to utilize existing system headroom, estimated as nearly 100 GW in the USA when large loads such as data centres provide 0.5% annualized curtailment[13].

In this Article, we present results from a real-world demonstration of a software-based power-orchestration platform that transforms a production AI cluster into a grid-responsive asset. Conducted in Phoenix, Arizona, on a 256-GPU, production-grade cluster of a data centre in collaboration with partners representing industries in cloud infrastructure, compute hardware and grid energy, the demonstration validated the ability to deliver sustained, accurate power reductions using only workload orchestration—without requiring any hardware retrofits or energy-storage systems.

[1]Emerald AI, Washington, DC, USA. [2]Oracle Corporation, Austin, TX, USA. [3]NVIDIA Corporation, Washington, DC, USA. [4]Salt River Project (SRP), Tempe, AZ, USA. [5]Electric Power Research Institute (EPRI), Palo Alto, CA, USA. ✉e-mail: ayse.coskun@emeraldai.co

https://doi.org/10.1038/s41560-025-01927-1

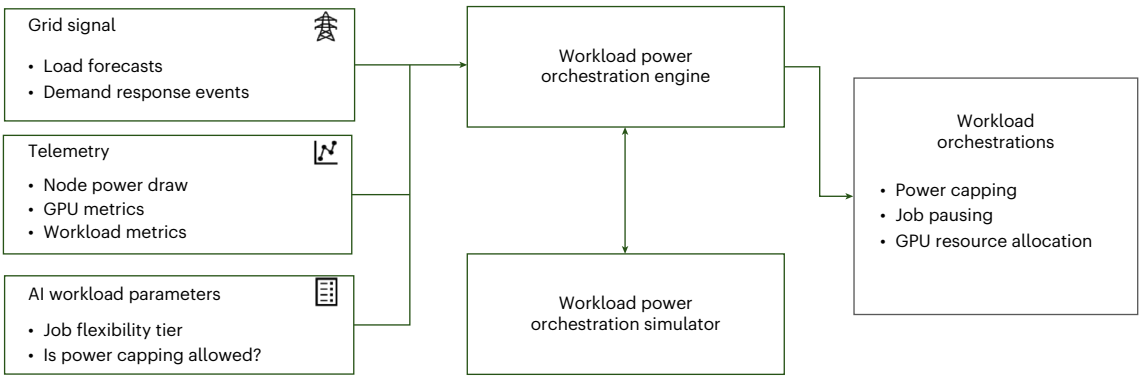


**Fig. 1 | An overview of the software-based power-orchestration architecture.** The AI Workload Orchestration Engine takes three inputs: a grid power signal, application and power telemetry, and job parameters. It uses power and load-management algorithms and an internal simulator to explore and select an optimal job schedule and power allocations. The final output is the optimal set of per-job power orchestrations chosen to meet the grid's power target while maintaining performance based on job priorities.

## System architecture and flexibility framework

At the core of the demonstration is a software-based power-orchestration platform that interfaces with AI workload managers and grid signal sources. The power-orchestration platform dynamically schedules jobs, modifies resource allocations for each job and applies power-limiting techniques such as GPU frequency scaling. To guide its decisions, the platform uses a system-level simulator trained to predict the power-performance behaviour of AI jobs. The simulator evaluates the trade-offs of various orchestration strategies under operational constraints and grid needs, recommending an orchestration strategy to assure AI workload Quality-of-Service (QoS) guarantees while also meeting power grid response commitments. Figure 1 demonstrates the overall architecture of the software-based platform.

Our workload-tagging schema classifies jobs into flexibility tiers based on user tolerance for runtime or throughput deviations (Table 1). The tags allow applying control policies that differentiate jobs and ensure each job meets the compute user's desired QoS threshold, which could pave the way for an SLA that upholds user-specified QoS while enabling flexible power management. For example, using feedback and guidance from our industry partners, we identified the following flexible SLAs for representative AI workloads: (1) Flex 0: no performance reduction (strict SLA); (2) Flex 1: up to 10% performance (average throughput) reduction allowed over a 3–6 hour period; (3) Flex 2: up to 25% allowed; (4) Flex 3: up to 50% allowed. These tiers enable intelligent, non-disruptive throttling that preserves workload commitments while unlocking power flexibility.

## Phoenix field trial

The demonstration was conducted at an Oracle Phoenix Region Cloud data centre on a 256-GPU cluster built on NVIDIA A100 Tensor Core GPUs, orchestrated through Databricks MosaicML (https://www.databricks.com/research/mosaic), instrumented via Weights & Biases (https://wandb.ai/) for telemetry and integrating Amperon's grid-demand forecasting tools. Four representative workload ensembles were selected—each combining varying proportions of jobs representing different types of AI workload. AI systems go through training (where models learn patterns from large datasets), inference (where trained models make predictions) and fine tuning (where models are adapted to a new task with smaller amounts of data); inference can happen in batch mode (processing many inputs at once, such as overnight analysis) or in real time (responding instantly to user requests, such as a chatbot). Methods provide more details about how we combine those types of workload.

In consultation with the regional utilities Arizona Public Service (APS) and Salt River Project (SRP), we set stringent targets for grid-responsive demand to prove that AI compute power load could provide meaningful relief during periods of system-coincident peak stress, for example, during a hot Phoenix day with high-air-conditioning load. We executed two events to demonstrate these capabilities: (1) 1 May 2025 (addressing APS system peak) and (2)

**Table 1 | Workload ensembles and their flexibility SLAs used in the experiments**

| Ensemble | Workload | Number of nodes | Flex level |
|---|---|---|---|
| Ensemble 1 80% training, 20% user inference | MPT 13B - training | 8 | Flex 3 |
| | MPT 7B - training | 6 | Flex 3 |
| | LLaMA 8B - fine tune | 6 | Flex 2 |
| | LLaMA 8B - fine tune | 6 | Flex 3 |
| | LLaMA 8B - inference | 4 | Flex 0 |
| | LLaMA 8B - inference | 2 | Flex 0 |
| Ensemble 2 50% training, 50% user inference | MPT 13B - training | 8 | Flex 3 |
| | MPT 7B - training | 6 | Flex 3 |
| | LLaMA 8B - fine tune | 4 | Flex 2 |
| | LLaMA 8B - inference | 4 | Flex 0 |
| | LLaMA 8B - inference | 6 | Flex 0 |
| | LLaMA 8B - inference | 4 | Flex 0 |
| Ensemble 3 50% training, 50% latency-tolerant inference | MPT 13B - training | 6 | Flex 3 |
| | MPT 7B - training | 6 | Flex 3 |
| | LLaMA 8B - fine tune | 4 | Flex 2 |
| | LLaMA 8B - inference | 4 | Flex 3 |
| | LLaMA 8B - inference | 6 | Flex 2 |
| | LLaMA 8B - inference | 6 | Flex 1 |
| Ensemble 4 90% training, 10% user inference | MPT 13B - training | 6 | Flex 3 |
| | MPT 7B - training | 6 | Flex 3 |
| | LLaMA 8B - fine tune | 4 | Flex 2 |
| | LLaMA 8B - fine tune | 4 | Flex 3 |
| | LLaMA 8B - inference | 2 | Flex 0 |
| | LLaMA 8B - inference | 2 | Flex 0 |
| | LLaMA 8B - fine tune | 4 | Flex 3 |
| | LLaMA 8B - fine tune | 4 | Flex 2 |

In the cluster under test, each GPU node consisted of eight A100 GPUs, and each workload ensemble ran throughout the experiment duration, with control actions determined by the power-orchestration platform.

https://doi.org/10.1038/s41560-025-01927-1

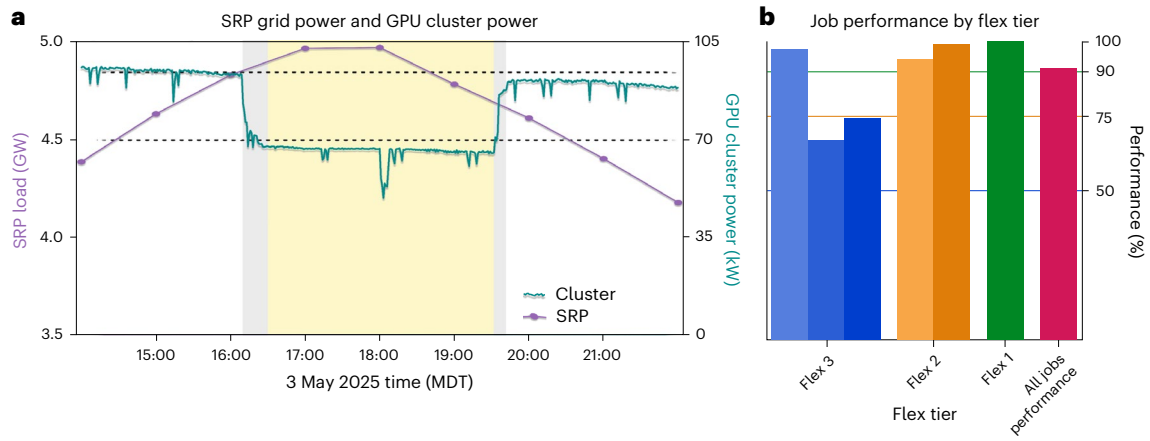


**Fig. 2 | SRP utility experiment summary. a**, The reduced AI cluster power demand from 16:30 to 19:30 achieved by running the power-orchestration platform during the time of peak SRP load. The purple line shows energy demand (GW) in SRP over time. The green line shows the total power draw (kW) of the GPU cluster over time. The yellow shaded region represents the event window, during which the cluster held its 25% power reduction. The grey shaded region represents the time during which the cluster power ramped down and up around the event. The power reduction was achieved by pausing jobs, reducing the number of GPUs allocated to jobs and reducing the power of the GPUs themselves. Horizontal lines demonstrate the average base power and the power curtailment target during the power event. MDT, Mountain Daylight Time. **b**, How the performance of each flex tier was impacted by applying power control. Flex 1 jobs (green) had a minimum performance threshold of 90%, Flex 2 jobs (orange) had a minimum performance threshold of 75% and Flex 3 jobs (blue) had a minimum performance threshold of 50%. Note that all jobs exceeded their minimum performance threshold.

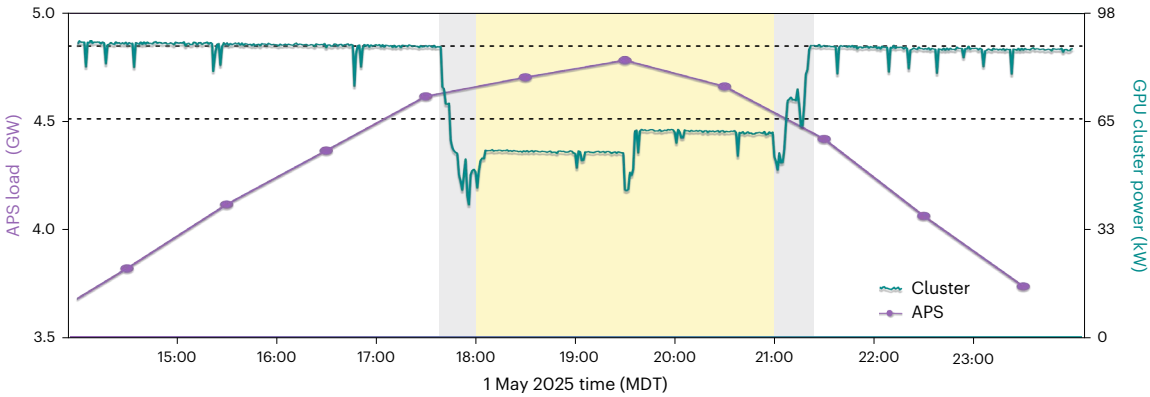


**Fig. 3 | APS utility experiment summary.** The cluster's power was reduced from 18:00 to 21:00, achieved by running the power-orchestration platform during peak APS load. The purple line shows energy demand (GW) in APS over time. The green line shows the total power draw (kW) of the GPU cluster over time. The yellow shaded region represents the event window, during which the cluster held its 25% power reduction. The grey shaded region represents the time during which the cluster power ramped down and up around the event. The power reduction was achieved using a mixture of control knobs, including job pausing and GPU reallocation. Horizontal lines demonstrate the average base power and the power curtailment target during the power event.

3 May 2025 (addressing SRP system peak), where we tracked the grid load and identified the upcoming peak demand periods. Each event required the cluster to reduce power by 25% with respect to the average base load during the peak demand period, sustain the reduction for 3 hours and ramp down and up gracefully over 15 minutes, avoiding so-called 'snap back' at the conclusion of the event by staying around the same pre-event baseline (that is, without a substantial increase in power consumption). These technical requirements of the test were set by our utility partners. On both occasions, the power-orchestration platform met the utility-set requirements precisely (Figs. 2 and 3).

In both APS and SRP events, all jobs completed within allowable SLA envelopes. We confirmed via power measurements that our software solution achieved sustained and accurate reductions.

We also modelled a synthetic emergency event as part of the field trial, based on California Independent Service Operator (CAISO)'s August 2020 emergency load shed event, where a power plant failure during an extreme weather event triggered CAISO to request reduction of power via available reserves in the grid[14]. In this re-enactment, the power-orchestration system responded to an initial 15% curtailment followed by an emergency 10% further step down. The system delivered both reductions smoothly, matching the desired power profile (Fig. 4), while continuing to assure the AI QoS thresholds.

In addition to demonstrating capabilities for APS, SRP and CAISO sample events, we ran a total of 33 experiments (each 3–6 hours long, Table 2) including 212 total individual jobs during the field trial, where we demonstrated the impact of applying several different power management policies on our select workload ensembles (Methods). In every single experiment, the system performed as expected, guided by simulated predictions to achieve the required power reduction and job-specific SLA requirements.

## Simulator performance accuracy

The simulator predictions closely matched real-world cluster behaviour. Across control intervals in our experiments, the model achieved 4.52% root mean square error (RMSE) in power predictions, relative to average experiment power (Fig. 5). Individual job behaviours, including fine tuning and batch inference workloads, were predicted

https://doi.org/10.1038/s41560-025-01927-1

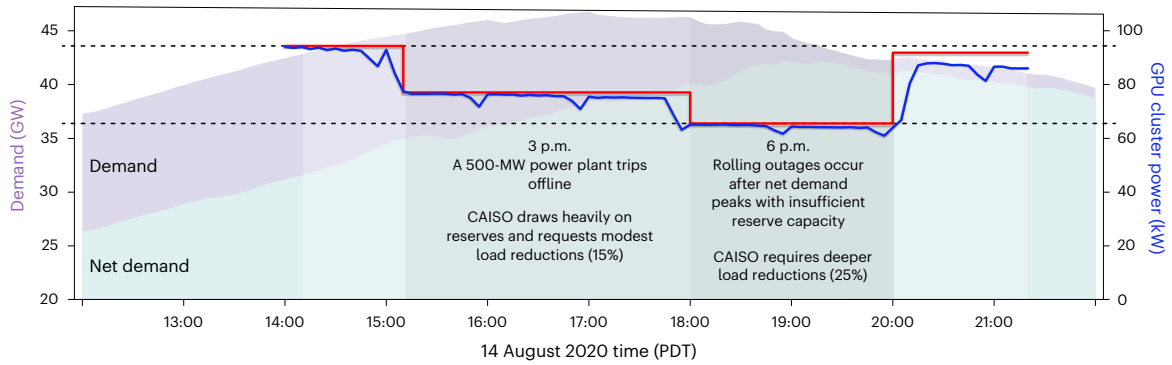


**Fig. 4 | Re-enacted historical CAISO power event from 2020.** The power-orchestration platform responded to the curtailment requirement, which involved a 3-hour curtailment from 15:00 to 18:00 and a 2-hour curtailment from 18:00 to 20:00. The red line shows the target power reductions needed to follow the curtailment requirements, and the blue line shows the AI cluster power draw over time, closely following the targets. Power consumption shown in the plot is measured and averaged at continuous 5-minute windows. The shaded purple and green regions represent CAISO energy demand and net demand, respectively. Horizontal lines demonstrate the average base power and the power curtailment target during the power event. PDT, Pacific Daylight Time.

## Table 2 | Summary of experimental runs

| Ensemble | Control knob | Policy | Power target (%) | Event length (min) |
|---|---|---|---|---|
| 1 | Resource allocation | Fair | 75 | 5 |
| 1 | Resource allocation | Fair | 75 | 60 |
| 1 | Resource allocation | Greedy | 75 | 60 |
| 1 | Job pausing + resource allocation | Greedy | 75 | 60 |
| 1 | Job pausing + resource allocation | Fair | 75 | 60 |
| 1 | Resource allocation | Fair | 75 | 60 |
| 1 | Resource allocation | Fair | 90 | 60 |
| 1 | Job pausing + resource allocation | Greedy | 60 | 60 |
| 1 | DVFS | Fair | 75 | 60 |
| 1 | DVFS | Fair | 75 | 60 |
| 1 | DVFS + job pausing | Fair | 75 | 60 |
| 2 | Resource allocation | Fair | 75 | 60 |
| 2 | Job pausing + resource allocation | Greedy | 75 | 60 |
| 2 | Resource allocation | Fair | 90 | 60 |
| 2 | DVFS | Fair | 75 | 30 |
| 2 | DVFS | Fair | 75 | 122 |
| 3 | Job pausing | Greedy | 75 | 60 |
| 3 | Resource allocation | Fair | 75 | 60 |
| 3 | Job pausing + resource allocation | Fair | 75 | 180 |
| 3 | Job pausing + resource allocation | Greedy | 75 | 60 |
| 3 | Job pausing + resource allocation | Fair | 75 | 60 |
| 3 | Job pausing + resource allocation | Greedy | 75 | 60 |
| 3 | DVFS + job pausing | Fair | 75 | 180 |
| 3 | DVFS | Fair | 75 | 5 |
| 3 | DVFS | Greedy | 75 | 60 |
| 3 | DVFS + job pausing | Fair | 60 | 60 |
| 3 | DVFS | Fair | 90 | 60 |
| 3 | Resource allocation | Greedy | 75 | 60 |
| 4 | Job pausing | Greedy | 75 | 60 |
| 4 | Resource allocation | Fair | 75 | 60 |
| 4 | Job pausing + resource allocation | Greedy | 75 | 60 |
| 4 | Job pausing + resource allocation | Fair | 75 | 60 |
| 4 | DVFS | Fair | 90 | 60 |

All experiments met both utility power targets and workload performance constraints.

https://doi.org/10.1038/s41560-025-01927-1

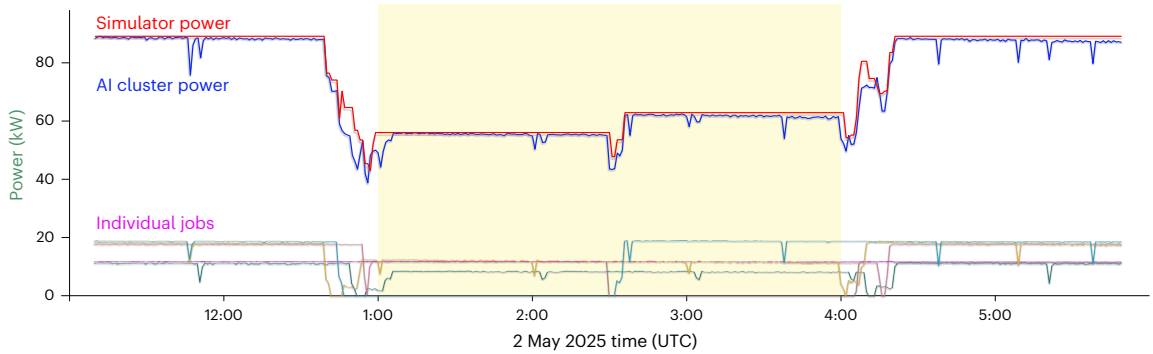


**Fig. 5 | Power prediction of the platform's simulator vs measured power during a power-reduction event.** The red line shows the power draw predicted by the simulator after it was given the workload ensemble, power targets, and power and load-management algorithms. The blue line shows the actual power draw of the AI cluster that occurred when our system responded to the power-reduction event. The bottom part of the plot shows the individual job power traces that, in aggregate, sum to the blue AI cluster power-draw line. Each trace is different, reflecting a diversity of controls applied to each job in the workload ensemble running during this event. UTC, Universal Time Coordinated.

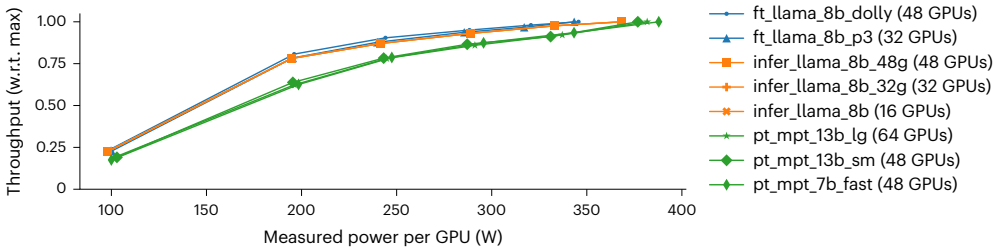


**Fig. 6 | Throughput of power-capped AI workloads on the 256-GPU cluster.** Each colour-symbol pair represents a different workload configuration. This plot shows the relationship between different power cap levels and workload throughput. w.r.t., with respect to.

with sufficient fidelity to inform real-time orchestration without breaching SLAs.

## Discussion

This demonstration represents an important advance in AI data centre management, moving facilities from static, high-load consumers to active, controllable grid participants. By integrating software intelligence with workload management, AI clusters can shape their power consumption profiles dynamically in response to grid needs.

Recent studies demonstrate that load flexibility for AI data centres could unlock 100 GW of new data centre capacity in the USA without requiring extensive new generation or transmission infrastructure[13,15]—enough to meet projected AI growth for the next decade. The new power capacity is unlocked by reducing new load with 0.5% annualized curtailment over periods lasting on average up to 2.1 hours.

A central advantage of the power-orchestration platform is deployability on standard hardware. Unlike hardware retrofits or battery deployments, which incur substantial capital costs and operational complexity, this software platform runs atop existing cloud and HPC stack environments. This means the approach is cloud native, scalable and already compatible with emerging AI cluster standards, including NVIDIA's modern infrastructure and containerized machine learning platforms. An opportunity exists to deploy this method in other major AI data centre regions, particularly those with constrained grid conditions such as Northern Virginia, Silicon Valley and Texas[4] and globally in countries with strong data centre growth such as the UK, Ireland, Germany, Singapore and others[3].

Our demonstration explored several control knobs to regulate power consumption: (1) power capping via the NVIDIA-System Management Interface (SMI), which primarily uses dynamic voltage frequency scaling (DVFS) to effectively reduce power with minor throughput impacts for many workloads[16] (Fig. 6); (2) job pausing (de-prioritization), which temporarily pauses running jobs for steep reductions in power and (3) changing the allocated resources for jobs, which reduces the number of allocated GPUs to reduce power while allowing job progress.

We evaluated performance sensitivity to power capping across multiple GPU allocations and workload types, including: fine tuning (ft) LLaMA 8B on two datasets, LLaMA 8B inference (infer) across three different GPU allocation sizes and three configurations of Mosaic Pretrained Transformer (MPT) pre-training (pt) workloads (Workloads section provides details). Across all experiments, job performance exhibited sensitivity to power limits, with modest degradation observed as power caps approached 400 W—the typical thermal design power of the target GPU. As power caps decreased further below this threshold, performance degradation became more pronounced. The number of GPUs allocated had minimal influence on power sensitivity; however, power-performance trade-offs varied by workload. In particular, MPT pre-training jobs were notably more sensitive in the mid-range of power caps, showing greater performance drops compared to fine-tuning and inference tasks.

The power capping control (via NVIDIA-SMI) overhead is negligible, making real-time response feasible even in busy clusters. Both pausing and re-allocating resources for running jobs require check pointing for training jobs to ensure forward progress and minimize the overhead associated with these knobs. Although prior work offers methods to optimize for long-term objectives where changes to control state incur a cost[17], we are able to treat check pointing overhead as negligible for our relatively infrequent demand response events since training jobs often run for days (or even weeks) at a time. An essential ingredient in the power-orchestration platform is to apply these knobs in a carefully calculated manner to avoid SLA violations—therefore, our software was able to meet SLAs even with performance estimates that are amortized over a more conservative 6-hour span.

In experiments, our 'DVFS + Job Pausing, Fair' policy, which spreads performance slowdowns across all jobs proportionally with their flexible SLAs (Methods), provided a good balance between achieving higher average job throughput and reducing the number of jobs impacted by power management.

Software-defined flexibility in AI data centres could transform utility interconnection processes. Flexibility may accelerate data centre interconnection and siting approvals, particularly in regions where permitting delays have become a bottleneck[18,19]. Data centres that are able to respond to grid signals can also avoid expensive infrastructure upgrades and receive demand response credits

Several limitations remain in the current approach to grid-interactive AI data centres. First, not all AI workloads are temporally flexible; AI customers with jobs with strict latency or reliability requirements may not be willing to accept workload throttling at a single site. In this study, batch-style training, fine-tuning and inference tasks could be slowed or paused, whereas other 'Flex 0' jobs such as real-time inference, streaming and model serving were not modified. To unlock broader flexibility, future work will explore geographically shifting AI workloads[9,20], harnessing spatio-temporal flexibility to preserve performance with minimal latency penalty across the broadest range of AI workloads. In addition, the industry's incentives and SLAs will need to evolve, encouraging users to opt into flexibility tiers in exchange for cost or compute availability benefits that are enabled by the massive new capacities of AI compute that can be brought online to power grids thanks to flexibility. Future work could explore trade-offs between workload flexibility mixes and data centre power flexibility to guide facility power-planning decisions.

Our field demonstration focused on a single cluster's ability to curtail load to respond to system peaks and reduce the system-coincident peak demand of an AI data centre, enabling faster interconnection and avoiding infrastructure buildout to serve higher system peaks. To fully understand system-level impacts, larger-scale deployments involving full data centre telemetry and multiple data centre zones are needed. These would allow for the validation of broader control strategies and coordination mechanisms. Our future work will also explore participation, in addition to emergency demand response, in a wider range of grid programmes such as day-ahead demand response, frequency regulation and other ancillary services[21,22] to assess economic viability and technical readiness in real-world grid markets.

Our software-only demonstration does not require on-site generation or energy-storage equipment to achieve flexibility, but such energy resources can provide complementary flexibility. Prior works have demonstrated opportunities for improved grid integration when adding storage and generator awareness to power-aware schedulers[17,23–25]. Future work could explore trade-offs between life-cycle costs and flexible power capacity by harnessing varied degrees of distributed energy resources.

This Phoenix-based field demonstration marks a transformative moment in energy–AI integration. With solely a software-based solution, existing AI clusters can become reliable, precise assets for grid support. As demand from AI accelerates, flexible data centres offer a scalable, sustainable way to meet electricity system needs—without delaying innovation or sacrificing reliability.

The path forward will require deeper engagement with grid operators, AI developers and regulators to standardize these capabilities and unlock a new era of grid-interactive computing.

## Methods

### Workloads

For our field trial, we worked with Databricks to design four representative workload ensembles consisting of training, inference and fine-tuning jobs. We downloaded MPT and LLaMA 3.1 Family models from Hugging Face and used the C4 dataset for pre-training and Dolly and P3 datasets for fine tuning. These workloads and their flexible SLA tiers are shown in Table 1. Each job was tagged into one of four flexibility tiers (discussed earlier in Results), based on tolerance to runtime or throughput degradation. Training workloads are generally the most flexible (Flex 3) in our ensembles because they are expected to run for many days. Inference loads incorporate a range of flexibility, such as Flex 0 to represent real-time user inference or Flex 3 for inference as part of a batch data pipeline. We determined these ensembles and tags in consultation with our industry expert partners and based on typical SLA expectations for different types of AI job.

### Orchestration algorithms

We designed a suite of power and load-management algorithms for the power-orchestration platform. These algorithms utilize the three control knobs described earlier—both individually and in various combinations (for example, DVFS with job pausing, job pausing with resource reallocation or others).

We implemented two main algorithmic strategies across different knob combinations: (1) Greedy: this algorithm prioritizes applying control knobs to the most flexible jobs first, aiming to maximize power reduction while minimizing the number of jobs affected. (2) Fair: this algorithm distributes the projected performance overhead more evenly across all jobs, ensuring a more balanced impact on workload performance.

All policies we implemented solved for power-reduction objectives, subject to constraints from job SLAs. The following 'control knob, algorithmic strategy' combinations provided the most desirable results while meeting power reduction and job performance constraints: 'DVFS + Job Pausing, Fair' provided the best trade-off between average job throughput and the number of jobs impacted; 'DVFS, Fair' provided the best average job throughput overall as the knob can be applied at a finer granularity than other knobs; 'Job Pausing, Greedy' affected the fewest number of jobs to achieve the desired power reduction; 'Job Pausing + Resource Allocation, Fair' achieved the best average throughput among policies without DVFS.

Our experiments also covered the following control strategies, however they did not perform as well as the other strategies: 'DVFS, Greedy'; 'Resource Allocation, Fair'; 'Resource Allocation, Greedy'; 'Job Pausing + Resource Allocation, Greedy'.

### Simulation and real-time execution

In this field trial, we determined the policy decisions in the platform's simulator, which then drove the platform's decisions at runtime. We received the grid load signals via Application Programming Interface (API) from the Amperon platform, which provides forecast grid load and actual historical grid data that we used to verify the timing of our cluster's demand response performance. We timestamp aligned all power and job completion telemetry.

The simulator estimates power-performance relationships based on profiles we measured for each job in advance of the power-reduction event experiments. Although prior characterization may not be feasible for all workloads, the data centre operator may be able to characterize its non-urgent internal workloads[9]. Future work may extend system-wide job characterization efforts[26–28] to infer power-performance properties of non-internal workloads.

### Evaluation metrics

Three key metrics were used to assess performance in our experiments (1): power-reduction compliance, which compares percentage power reduction achieved vs utility target; (2) QoS preservation, which checks whether individual job SLAs are met and (3) simulator accuracy, where we calculate the root mean square error (RMSE) of power prediction vs measured power. We measured power consumption of GPUs via NVDIA-SMI and throughput (for example, steps per second or tokens per second) of applications via our custom scripts.

In all power-reduction events (APS, SRP, CAISO) and our 33 total power-reduction experiments (Table 2), we fully maintained compliance thresholds while maintaining zero SLA violations. Across our experiments, we achieved 4.52% simulator accuracy RMSE relative to the average experiment power.

Energy efficiency is not a direct metric in our flexible power-demand policies because the consumed energy is a function of the utility's power target and event duration. Energy impact beyond the demand response event may come as a side effect from following a reduced power target under performance constraints because job performance during the event will impact how much work remains in a job after the event is over. As an example from our SRP flagship experiment (Fig. 2), we scaled each job's progress during the experiment relative to the job's total progress when it completes as a proxy for slowdown, and we scaled each job's average measured power relative to the job's average power when executed without the power-orchestration platform. The product of these two values represents each job's relative energy savings through the platform's policy. We applied a weighted average of those energy savings, weighted by each job's baseline power demand as an expected energy reduction, considering the system's schedule would continue to execute those jobs after the power-reduction event. In this example case, we observed a 5% reduction in energy.

### Implementation of the software-based power-orchestration platform

We implemented the software-based orchestration platform as a centralized Python application that issues commands to compute-node management processes hosted alongside executing jobs. The platform starts, stops and queries the scheduling state of jobs through the MosaicML mcli API. GPU power and application throughput metrics are queried from the Weights & Biases API.

The power-orchestration platform issues new commands through a distributed memory object cache that is monitored by the management process on each node. Whenever the cache is updated with a new power cap, the cap is enforced by launching *nvidia-smi -pl* to set a power limit on the compute node's GPUs. A MosaicML Composer callback handles early stop requests received in the cache to allow jobs to checkpoint immediately before suspension/rescaling.

## Data availability

The data that support the findings of this study are available within the article and via GitHub at https://github.com/ai-emerald/emerald-ai-demo-may-2025. The data directory in our GitHub artefacts archive contains our DVFS control sweep (source data for Fig. 6) and time-series power data from our experiments (source data for Figs. 2–5). The README.md file explains the purpose of each data file.

## Code availability

The README.md file in our GitHub artefacts archive (https://github.com/ai-emerald/emerald-ai-demo-may-2025) contains Python code and a Docker container image to apply power control (DVFS power caps, job start/stop commands and forced checkpoints) to LLM Foundry jobs, such as the ones used in our experiments. The README.md file also contains pseudocode describing how to apply power controls for each of the orchestration algorithms described in the 'Orchestration Algorithms' section.

## References


1. National Academies of Sciences, Engineering, and Medicine. *Implications of Artificial Intelligence–Related Data Center Electricity Use and Emissions: Proceedings of a Workshop* (National Academies Press, 2025); https://doi.org/10.17226/29101
2. Bianchini, R., Belady, C. & Sivasubramaniam, A. Data center power and energy management: past, present, and future. *IEEE Micro* **44**, 30–36 (2024).
3. Çam, E., Casanovas, M. & Moloney, J. *Electricity 2025: Analysis and Forecast to 2027* (IEA, 2025).
4. Aljbour, J., Wilson, T. & Patel, P. *Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption* (EPRI, 2024).
5. Zhang, Y., Wilson, D. C., Paschalidis, I. C. & Coskun, A. K. HPC data center participation in demand response: an adaptive policy with QoS assurance. *IEEE Trans. Sustain. Comput.* **7**, 157–171 (2022).
6. Zhang, Y., Wilson, D. C. Paschalidis, I. C. & Coskun, A. K. A data center demand response policy for real-world workload scenarios in HPC. In *2021 Design, Automation & Test in Europe Conference & Exhibition* 282–287 (IEEE, 2021); https://doi.org/10.23919/DATE51398.2021.9474075
7. Xing, J. et al. Carbon responder: coordinating demand response for the datacenter fleet. Preprint at https://arxiv.org/abs/2311.08589 (2023).
8. Radovanović, A. et al. Carbon-aware computing for datacenters. *IEEE Trans. Power Syst.* **38**, 1270–1280 (2023).
9. Mehra, V. & Hasegawa, R. Supporting power grids with demand response at Google data centers. *Google Cloud Blog* https://cloud.google.com/blog/products/infrastructure/using-demand-response-to-reduce-data-center-power-consumption (2023).
10. Terrell, M. How we're making data centers more flexible to benefit power grids. *Google Data Centers and Infrastructure Blog* https://blog.google/inside-google/infrastructure/how-were-making-data-centers-more-flexible-to-benefit-power-grids/ (2025).
11. Sivaram, V. *Taming the Sun: Innovations to Harness Solar Energy and Power the Planet* (MIT Press, 2018).
12. Tan-Soo, J.-S., Qin, P., Quan, Y., Li, J. & Wang, X. Using costbenefit analyses to identify key opportunities in demand-side mitigation. *Nat. Clim. Change* **14**, 1158–1164 (2024).
13. Norris, T., Profeta, T., Patino-Echeverri, D. & Cowie-Haskell, A. *Rethinking Load Growth: Assessing the Potential for Integration of Large Flexible Loads in US Power Systems* (Nicholas Institute for Energy, Environment & Sustainability, 2025).
14. Mainzer, E., Batjer, M. & Hochschild, D. *Final Root Cause Analysis: Mid-August 2020 Extreme Heat Wave* (CAISO, 2021).
15. Lin, L. et al. Exploding AI power use: an opportunity to rethink grid planning and management. In *Proc. 15th ACM International Conference on Future and Sustainable Energy Systems* 434–441 (ACM, 2024); https://doi.org/10.1145/3632775.3661959
16. Zhao, D. et al. Sustainable supercomputing for AI: GPU power capping at HPC scale. In *Proc. 2023 ACM Symposium on Cloud Computing* 588–596 (ACM, 2023); https://doi.org/10.1145/3620678.3624793
17. Lechowicz, A. et al. The online pause and resume problem: optimal algorithms and an application to carbon-aware load shifting. *Proc. ACM Meas. Anal. Comput. Syst.* **7**, 45 (2023).
18. Ratnayake, A., Lopez, I. D., Vairamohan, B. & Lannoye, E. *Grid Flexibility Needs and Data Center Characteristics* (EPRI, 2025).
19. Satchwell, A. et al. *Electricity Rate Designs for Large Loads: Evolving Practices and Opportunities* (Lawrence Berkeley National Laboratory, 2025).
20. Zheng, J., Chien, A. A. & Suh, S. Mitigating curtailment and carbon emissions through load migration between data centers. *Joule* **4**, 2208–2222 (2020).
21. Zhou, Z., Levin, T. & Conzelmann, G. *Survey of U.S. Ancillary Services Markets* (Argonne National Laboratory, 2016); https://doi.org/10.2172/1236451
22. *ERCOT Ancillary Services Study* (ERCOT, 2024); https://www.ercot.com/files/docs/2024/10/07/ERCOT-Ancillary-Services-Study-Final-White-Paper.pdf
23. Liu, Z., Wierman, A., Chen, Y., Razon, B. & Chen, N. Data center demand response: avoiding the coincident peak via workload shifting and local generation. *Perform. Eval.* **70**, 770–791 (2013).

24. Goiri, Í, Katsak, W., Le, K., Nguyen, T. D. & Bianchini, R. Parasol and GreenSwitch: managing datacenters powered by renewable energy. In *Proc. Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems* 51–64 (ACM, 2013); https://doi.org/10.1145/2451116.2451123
25. Pahlevan, A., Zapater, M., Coskun, A. K. & Atienza, D. ECOGreen: electricity cost optimization for green datacenters in emerging power markets. *IEEE Trans. Sustain. Comput.* **6**, 289–305 (2021).
26. Patel, T. et al. What does power consumption behavior of HPC jobs reveal?: demystifying, quantifying, and predicting power consumption characteristics. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 799–809 (IEEE, 2020); https://doi.org/10.1109/IPDPS47924.2020.00087
27. Hu, Q. et al. Characterization of large language model development in the datacenter. In *21st USENIX Symposium on Networked Systems Design and Implementation* 709–729 (USENIX, 2024); https://www.usenix.org/conference/nsdi24/presentation/hu
28. Tang, B. J. et al. The MIT supercloud workload classification challenge. In *2022 IEEE International Parallel and DistributedProcessing Symposium Workshops (IPDPSW)* 708–714 (IEEE, 2022); https://doi.org/10.1109/IPDPSW55747.2022.00122

## Acknowledgements

We would like to thank the following individuals for their insights and support throughout this work: P. Vincent, J. Jackson, M. Sweeney, A. Springborn, S. Campbell, R. Caputo, S. Chen and M. Zablocki at Oracle; J. Frankle, Y.-L. Yang, L. Ladrech and S. Sherpa at Databricks; S. Trivedi, V. Troy, M. Spieler at NVIDIA; R. Toews, J. Jacobs, D. Katz and D. Mulet at Radical Ventures; M. Specks at Aventurine Partners; D. Rousseau at SRP; G. Bernosky, C. Lynn and B. Brazis at Arizona Public Service; A. S. Frank at Luminary; T. Norris at Duke University; I. Brown at 38 North; S. Kelly at Amperon; A. Patterson at Ceramic AI; A. Atkinson at Camus Energy; G. Desai; S. Goodman; M. Boomer; R. Stuebi at Boston University; and D. Porter, D. Larson and T. Wilson at EPRI.

## Author contributions

Emerald AI authors are listed in alphabetical order. The Emerald AI team conceived the idea and initiated the research and development. The Emerald AI tech team (S.S., C.W., P.C., D.C.W., E.T., C.R. and A.K.C.) implemented and executed the field test. B.R. (Oracle) gave feedback on integration with cluster management infrastructure. D.D. and N.M. (SRP) and A.R. and B.V. (EPRI) advised on power grid and flexibility aspects, and Z.M. (NVIDIA) provided feedback on the field test goals and results. A.K.C., D.C.W. and V.S. drafted the paper, and E.T., C.W., S.S., C.R., E.T. and P.C. contributed to its editing.

## Competing interests

This work was conducted by Emerald AI in collaboration with SRP, NVIDIA, Oracle and EPRI. Authors affiliated with Emerald AI, SRP, NVIDIA, Oracle and EPRI are employees of their respective organizations. Authors from Emerald AI, NVIDIA and Oracle hold stock or stock options in their companies. Authors from Emerald AI also have patents pending related to the technologies discussed in this paper. The authors declare no other competing interests.

## Additional information

**Correspondence and requests for materials** should be addressed to Ayse K. Coskun.

**Peer review information** *Nature Energy* thanks Benjamin Lee and the other, anonymous, reviewer for their contribution to the peer review of this work.