March 2026

Power-Flexible AI Factories

A UK-First Demonstration of Grid-Responsive AI Infrastructure

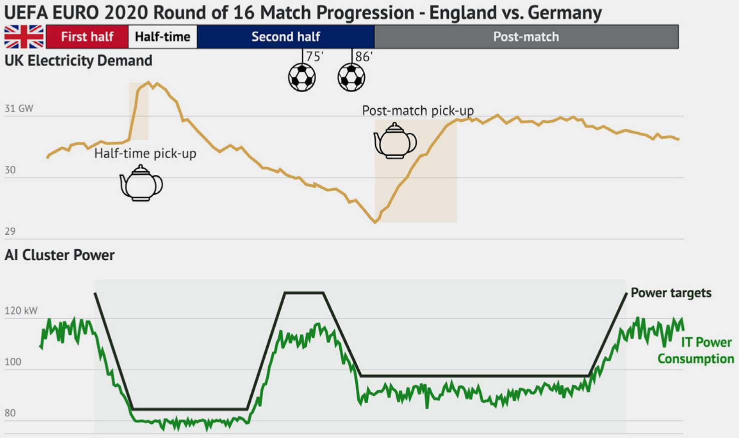
emeraldai   EPRI ELECTRIC POWER RESEARCH INSTITUTE   nationalgrid   NEBIUS

01 Executive Summary

As artificial intelligence (AI) adoption accelerates, the power required to support large-scale computing is growing fast.

Grid planners are statutorily required to assume large industrial loads will draw peak capacity even during extreme system scenarios, such as winter evening peaks. Under this “firm load” modeling, the grid currently lacks the headroom to accommodate gigawatt-scale AI factories during the highest-demand hours of the year without further infrastructure upgrades.

Emerald AI, EPRI, National Grid, Nebius, and NVIDIA recently conducted a live, UK-first demonstration at Nebius’s new AI Factory in London. The objective was to demonstrate that high-performance AI infrastructure could operate as a **power-flexible, grid-responsive asset** without disrupting mission-critical workloads.

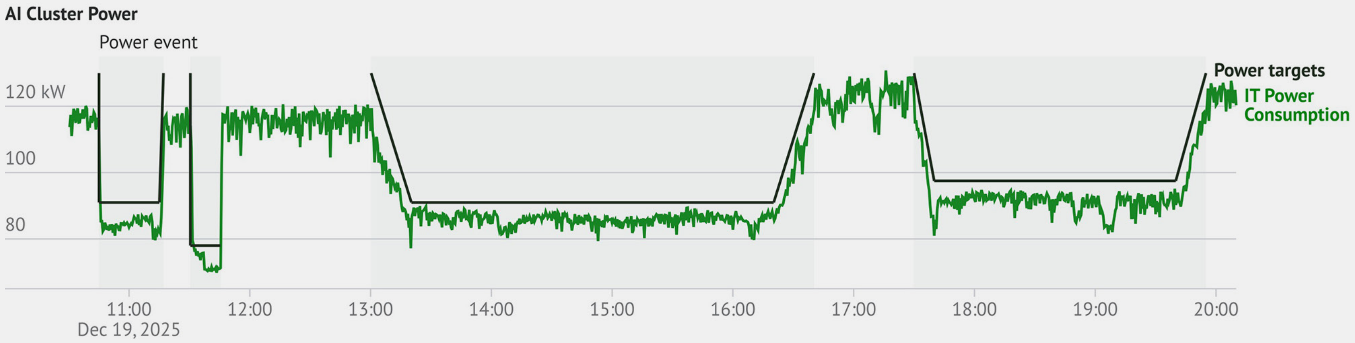


Tea Kettle Balance: System orchestrates AI cluster power to offset a reenactment of the UK “TV pickup” phenomenon that occurred during England vs. Germany Euro 2020 Round of 16 match

Over a five-day period in December 2025, a cluster of 96 NVIDIA Blackwell Ultra GPUs at the AI Factory was subjected to 22 live dispatch events from the power grid, including “surprise” emergency signals with zero advance notice. The results support a “Power-Flexible” operational paradigm:

- **The “Lightning Strike” Response:** The cluster autonomously slashed power by **30% in under 40 seconds** following an emergency scenario, a reflex speed comparable to industrial battery storage.
- **The “Tea Kettle” Balance:** In a re-enactment of the UK’s famous “TV pickup” phenomenon, the AI Factory successfully modulated its load in inverse correlation to a residential demand spike, effectively neutralising grid stress.
- **Zero-Compromise Reliability:** Emerald AI’s Conductor platform achieved **100% compliance** across 200+ power targets while preserving Service Level Agreements (SLAs) for critical workloads.

By converting data centres from static consumers into responsive virtual power plants, the UK can unlock the twin imperatives of its industrial strategy: affordability and speed to power. Power-flexible AI Factories can help keep rates low for billpayers by avoiding further grid infrastructure upgrades, while simultaneously turbocharging UK AI innovation and competitiveness by enabling larger and faster power connections for the data centres that drive the future economy.



Lightning Strike Response: System achieves sub-minute response times and reduces power by up to 40% in response to system operator dispatch commands, while orchestrating dozens of realistic AI workloads

02 Introduction: Enabling AI Factories to Operate as Responsive Grid Assets

The United Kingdom stands at a pivotal juncture in its energy transition. The digitalisation of the global economy has entered a new phase of energy intensity, driven by generative AI. Unlike previous generations of data centre growth, which were driven by storage and general-purpose compute, AI is defined by massive, synchronous computational workloads that require gigawatts of power.

The UK is seeing a surge of large demand connection applications, including data centres, at a time when the grid cannot expand on software timescales. The Office of Gas and Electricity Markets has highlighted rapid growth in the demand connections queue (from 41 GW in November 2024 to 125 GW by June 2025), explicitly calling out data centres as a significant share of that increase [OFGEM].

This shift presents a challenge of timescales. Building high-voltage transmission lines, substations, and generation capacity needed to support exponentially growing AI power demand typically involves long and complex planning approvals, with specialist supply chains, and community and environmental assessments and engagement. Conversely, the AI sector moves at the speed of software, with capital and hardware ready to deploy now.

These pressures reflect the fact that grid operators are statutorily required to assume large industrial loads will draw peak capacity even during extreme system scenarios, such as winter evening peaks. Under this “firm load” modeling, the grid currently lacks the headroom to accommodate gigawatt-scale AI factories during the highest-demand hours of the year without further infrastructure upgrades.

Because these major upgrades require time to plan and build, data centres can face longer lead times for connections in areas of high demand. This can influence where investment lands: innovation capital is fluid; if power cannot be secured in London, it will move to markets with faster connection timelines.

The solution is not only to build more wires and make better use of existing capacity, but also to make the demand side smarter. By enabling AI Factories to operate not as firm, inflexible loads, but rather as responsive grid assets that can adjust power consumption in real time based on grid conditions, this demonstration helps show that more data centres could fit within existing gaps of grid capacity, and overall less additional grid capacity will be needed to meet their needs. This “non-firm” approach allows for faster connections (using existing headroom) and lower system costs (avoiding unnecessary build-outs).

Box 2.1: Firm vs. Flexible Load - Definitions that matter for network planning in Great Britain

Term	Definition
Firm load (traditional assumption)	A customer whose connection is designed to supply them with full capacity at any time, including during system stress and standard contingencies.
Non-Firm	A Customer allowed to connect on the condition that their access can be curtailed when the system is constrained. This typically allows faster connection to the network as not all network reinforcements are required. Curtailment up to 100% of demand is permitted (typically unpaid) during faults, outages, or high system loading.
Flexible loads (planning-relevant definition)	A firm customer or non-firm customer (as long as they aren’t curtailed for another reason) that is paid to deliver measurable reductions under defined conditions - characterised by: • response time, • ramp rate, • power reduction, • duration, and • number of events per year
Flexibility markets	Typically operated by System Operators, flexibility markets provide structured arrangements to procure flexibility services from generation (such as batteries) and demand (such as large loads) to help grid operators manage peak periods and balance supply and demand effectively, in exchange for compensation.

Why this matters: Network planning and connections outcomes depend on whether the system can rely on flexible load behaviour *as if it were infrastructure*—repeatable, enforceable, and verifiable.

03 Defining the Power-Flexible AI Factory

A Power-Flexible AI Factory can be operated so that:

1. **Power is software-controllable** at fine time granularity (seconds to minutes);
2. **Flexibility is delivered autonomously,** without human operators manually intervening;
3. **Workload priorities are explicit,** so critical workloads keep their service levels while flexible workloads absorb constraints; and
4. **Performance and compliance are measurable,** enabling a regulator, system operator, or utility to trust the capability.

Where flexibility comes from in modern AI infrastructure

AI factories run mixed workloads. Some are latency-sensitive (e.g., serving inference), while others are throughput-oriented (e.g., training, fine-tuning, and batch inference). Within training and fine-tuning, many jobs include natural “flex points”: checkpoint intervals, gradient accumulation windows, or parallelism strategies that can tolerate short slowdowns.

Modern GPU clusters also have direct controllability:

- GPU power caps can reduce instantaneous consumption at the device level.
- Schedulers can pause or deprioritise jobs.
- Work can be shifted temporally—e.g., moving lower-priority fine-tuning to off-peak hours.
- AI workloads can be shifted geographically—so even serving inference queries can be redirected to regions where the power grid is less constrained while meeting the latency expectations of customers.

Why software-defined flexibility is different from traditional curtailment

Traditional large-load curtailment often implies a blunt instrument: Stop a process line, drop a feeder, or disconnect. That is operationally disruptive and often incompatible with high-performance, high-availability computing.

Power-flexible AI Factories instead aim to achieve demand reduction through controllable, smooth ramps to partial load reductions with workload-aware orchestration while preserving service levels. This is what makes power-flexible AI Factories powerful for both day-to-day grid optimisation and contingency events.

04 The UK Demonstration: Methodology and Architecture

4.1 Cluster set-up and operation

To demonstrate the flexibility of AI Factories under realistic operating conditions, the partnership executed a live trial at **Nebius's new London AI Factory.** This facility is among the first deployments in the United Kingdom to bring **NVIDIA Blackwell Ultra GPUs** online, connected through NVIDIA Quantum-X800 InfiniBand Platform. The demonstration was executed as the site came online and was conducted as part of **EPRI's DCFlex initiative**, which aims to demonstrate how data centres can support and stabilise the grid while improving the interconnection process [EPRI DCFlex].

To approximate production-grade conditions, Emerald AI, in conjunction with Nebius, selected a set of commercially representative AI training workloads, including:

- **GPT OSS 20B:** an OpenAI LLM optimised for powerful reasoning and agentic tasks with low latency, run in full-parameter supervised fine-tuning mode.

- **Llama 3 8B/70B:** a Meta LLM for general dialogue use cases with strong code generation capabilities (8-billion and 70-billion parameter sizes), run in supervised fine-tuning and direct preference optimisation (DPO) modes.

- **Llama 4 Scout 17B:** a Meta multimodal model optimised for text and image understanding, run in full-parameter supervised fine-tuning mode.

- **Qwen 2.5-7B VL:** a visual language model from Alibaba Cloud proficient at analysing texts, charts, images, videos, and structured outputs like JSON, run in full-parameter supervised fine-tuning mode.

The experiments kept the cluster continuously utilised by running workload ensembles under a scheduler that randomly sampled from the jobs listed above. The scheduler launched new runs whenever GPU capacity became available. In most experiments, some jobs started and finished multiple times per hour, creating a noisy power profile typical of a real-world production environment. The Emerald Conductor system had no foreknowledge of any of this randomized behaviour, so that the trial resembled realistic production conditions.

4.2 Grid signals and power experiments

The power experiments underpinning this demonstration were designed to be rigorous and representative of realistic grid operating conditions under grid stress events.

- **Scale:** A 130 kW AI cluster, consuming roughly the equivalent energy of **400 UK households.**

- **Duration:** Five days of continuous operation.

- **Dispatch signals:** National Grid Electricity Transmission (NGET) and EPRI issued **22 distinct real-time dispatch events.**

NGET and EPRI submitted dispatch signals through an event submission portal developed by Emerald AI. This portal enabled the submitter to specify the notice period, power reduction percentage, ramp-down duration, ramp-up duration, and overall event duration. These events were submitted remotely with no prior knowledge from Emerald AI.

Critically, some events were intentionally issued as "surprise" signals, including:

- no advance notice,
- no ramp time, requiring immediate response, and
- no knowledge of the workload ensembles or schedules.

Box 4.1: What does a grid power signal look like?

A grid-relevant flexible load dispatch signal specifies:

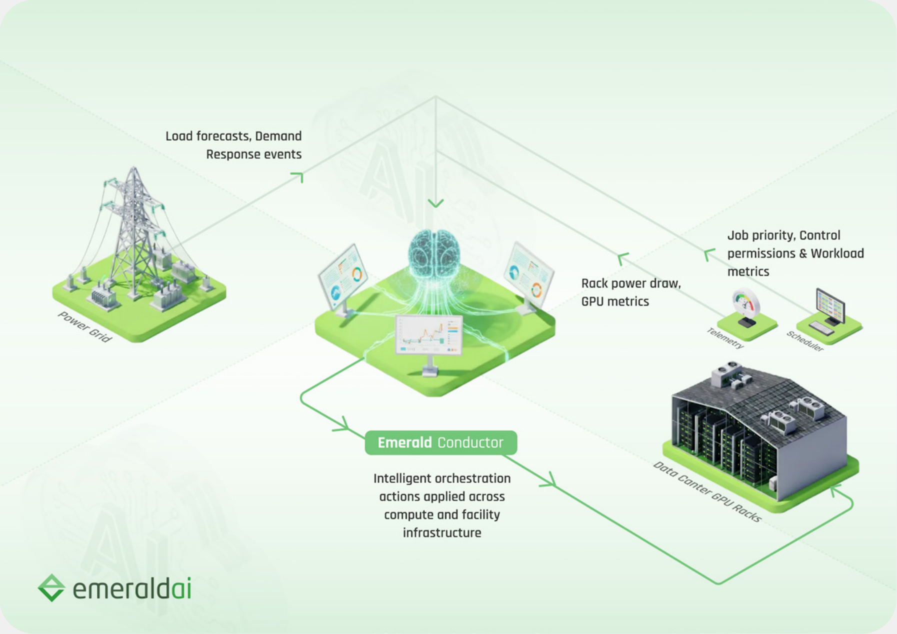
- Lead time (minutes until event goes into effect)
- Target level (absolute kW or % reduction)
- Hold/Event duration (e.g., 15 minutes, 2 hours)
- Ramp down constraint (e.g., reach target in 1 minute)
- Ramp up constraint (e.g., ramp back to baseline in 5 minutes)

During these dispatch events, the AI cluster continued running dozens of concurrent, production-grade workloads. Jobs continuously started, stopped, and competed for resources.

Figure 4.1: Emerald AI Conductor software architecture diagram

Emerald AI managed this power flexibility through its **Conductor** platform, a software layer that sits between the grid operator and the IT infrastructure. Conductor autonomously interprets grid signals and manages IT-level power control. This approach ensures that the highest-priority AI workloads, such as inference or urgent training, maintain high performance, while lower-priority work, such as non-urgent fine-tuning, absorbs flexibility while still meeting service level agreements (SLAs).

Result
Conductor achieved 100% compliance with 200+ requested power targets and ramp rates, including power reductions of up to 40% and sub-minute response times.


Load forecasts, Demand Response events
Power Grid
Job priority, Control permissions & Workload metrics
Rack power draw, GPU metrics
Telemetry
Scheduler
Emerald Conductor
Intelligent orchestration actions applied across compute and facility infrastructure
Data Center GPU Racks
emeraldai

Emerald AI Conductor software architecture diagram

4.3 Power Telemetry and Methodology

Emerald AI collected granular power telemetry at both the GPU level and the rack infrastructure level to assess cluster compliance with grid power targets. In the Nebius AI Factory, a rack consists of **32 NVIDIA Blackwell Ultra GPUs.**

- The **NVIDIA System Management Interface (smi)** was used to retrieve seconds-level GPU power telemetry.
- Nebius provided **minute-level and 20-second level rack power meter data** to help validate the NVIDIA smi readings and train cluster power models.

Emerald AI developed a high-granularity power model to bridge the gap between GPU-level telemetry and total rack power consumption–accounting for CPU, network, and storage overhead. Integrated into the **Conductor platform,** this model enabled real-time control of rack power during grid stress event scenarios. These predictions were subsequently validated against high-fidelity **Nebius rack meter data.**

During the UK demonstration, Conductor utilised an adaptive job power model to refine its predictive accuracy. As the platform orchestrated workloads, it dynamically characterised how specific job archetypes responded to power capping and pausing. By learning these “power signatures” in real time, the system improved provisioning accuracy for novel workloads and shifted the strategy from reactive closed-loop control to proactive, model-driven management.

05 Results: What the Grid Tested and What the AI Factory Delivered

The 130 kW AI cluster achieved **100% compliance** with all 200+ requested power targets and ramp rates. The trial demonstrated flexibility across distinct use cases that correspond to the most pressing challenges facing the UK grid.

5.1 Peak Load Relief ("The TV Pickup Effect")

The UK has a distinctive demand phenomenon: the "TV Pickup". During major televised events—most notably football matches—millions of people do the same thing at the same time: turn on kettles, open refrigerators, and use appliances during breaks.

The impact on the grid is sudden. For instance, following the penalty shootout of the Euro 2020 Round of 16 match between England and Germany, National Grid Electricity System Operator recorded a demand spike of nearly 1 gigawatt (GW) in a matter of minutes [NESO report]: roughly the output of a standard nuclear reactor.

NESO must carefully manage these demand peaks to keep the system stable. Their engineers normally draw on pumped storage hydroelectric power stations to deliver extra power quickly – highlighting the benefits of flexible response.

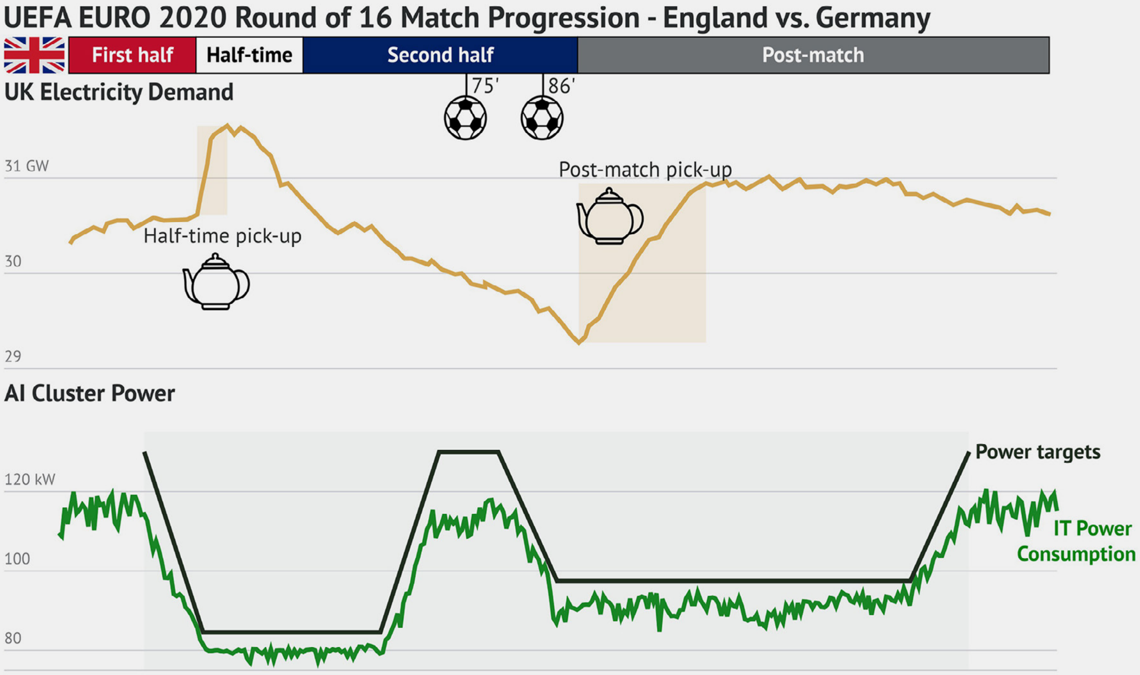
The Test
To mimic this exact stressor, a dispatch profile was issued to the Nebius AI Factory that mirrored the historical demand curves of the Euro 2020 "TV pickup." The goal was to see if the AI Factory could act as a shock absorber.

The Result
As the emulated "kettles" turned on across the country, the AI cluster autonomously ramped down its power consumption with seconds-level precision *(see Figure 5.1)*.

- **Inverse Correlation:** The AI power draw created a perfect inverse mirror of the demand spike. As residential load rocketed up, the AI load dropped, effectively neutralising the net load increase on the local substation.
- **SLA Preservation:** Despite the rapid de-powering, the Emerald AI Conductor ensured that critical high-priority jobs continued without latency spikes, while some lower-priority jobs were momentarily throttled.

This result is profound. It suggests that AI Factories, rather than being a liability during peak events, can actively dampen system volatility. By shedding load during a "TV pickup," an AI Factory "generates" virtual power (Negawatts) instantly.

Figure 5.1
AI cluster power response timed to offset a "tea kettle" spike, overlaid with grid demand curves.



5.2 Emergency Load Reduction (Generation Loss / Lightning Strike)

Grid operators must maintain stability under N-1 contingencies, including the sudden loss of a major generator or a fault on high-voltage transmission infrastructure. Lightning strikes are particularly challenging because they can trigger rapid, unexpected losses in generation, causing frequency to fall on seconds-to-minute timescales—often faster than conventional generation can respond.

A clear example occurred on August 9, 2019, on the Great Britain power system, as documented by National Grid Electricity System Operator (NESO). Prior to the event, the system was operating normally, with typical summer demand and a well-diversified generation mix [Lightning Strike Event].

At 4:52 pm, there was an unexpected reduction in generation from two large plants, Hornsea offshore wind farm and Little Barford gas station. Together, these units shed 1.38 GW of generation within seconds, far exceeding the grid's assumed single-largest contingency.

As a result, system frequency fell rapidly outside its normal operating range despite the deployment of all available automatic reserves, including battery storage. Frequency ultimately dropped to 48.8 Hz, triggering the grid's last-resort protection mechanism. Approximately 1 GW of demand (~5% of national load) was automatically disconnected to stabilise the system, temporarily interrupting power to about 1.1 million customers.

The Test
To evaluate whether AI infrastructure can respond effectively during such extreme, time-critical events, the system was subjected by NGET and EPRI to a surprise emergency curtailment signal, explicitly modeled on the August 9, 2019 lightning-strike contingency. The test assumed no advance notice and required an immediate, material reduction in load—mirroring the sub-minute response window observed in the UK event [Lightning Strike Event].

The Result
The AI cluster autonomously reduced its power consumption by 30% within 40 seconds *(see Figure 5.2)*. Responding on a sub-minute timescale places these facilities among the most agile demand-side resources available today. Demonstrating that a computing facility can shed load on this timescale shows that AI infrastructure can act as a grid reliability asset, helping arrest frequency collapse during rare but high-impact physical faults like the 2019 lightning-strike event. In a separate test, NGET and EPRI dispatched a surprise 40% reduction, which the cluster achieved in roughly one minute, all while preserving the performance of the highest-priority workloads.

Why this is critical for accelerated grid connections

Emergency response capability is one of the strongest arguments for allowing flexible loads to connect earlier. If a large new load can be treated as dispatchable during contingencies, planners can consider connection agreements that rely on curtailment during these rare events instead of overbuilding permanent backup infrastructure.

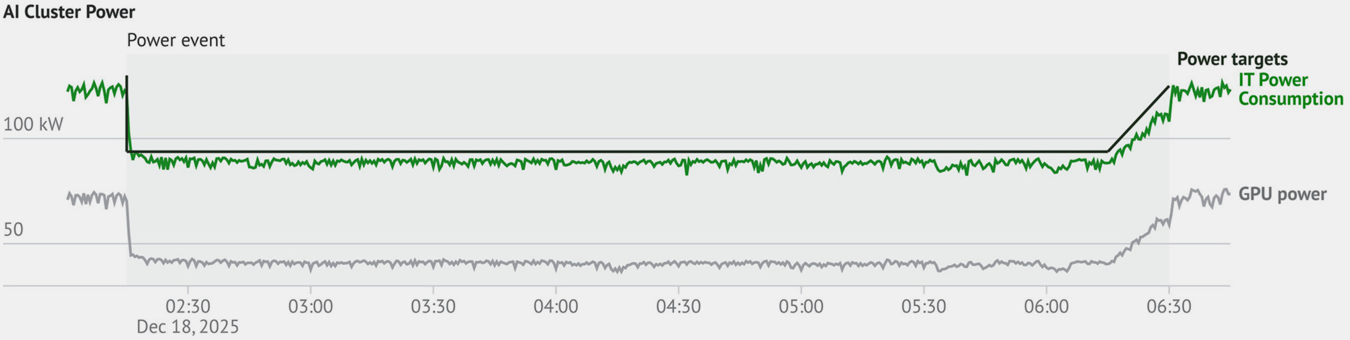
5.3 Sustained Load Reductions

Electricity systems are increasingly challenged not only by sudden disturbances but by extended periods of system stress lasting hours rather than seconds. These conditions can arise from multiple drivers: prolonged periods of low renewable output, extreme weather that elevates demand, or high coincident load from large electricity consumers, including data centres. In such scenarios, the core reliability challenge is maintaining adequacy over time, not rapid frequency stabilisation.

Conventional flexibility resources are poorly suited to this problem. Battery storage, while effective for short-duration balancing, typically provides 2-4 hours of discharge and becomes costly or operationally constrained when required to sustain multi-hour or repeated reductions. As demand grows and the grid relies more heavily on intermittent supply, operators require controllable resources capable of maintaining reduced load for extended durations without exhausting finite energy reserves.

The Test
Partners issued requests for 10-40% load reductions

Figure 5.2
AI cluster responds to a historical replication of the 2019 lightning strike that reduced grid frequency levels in the UK.



sustained over periods ranging from 2 to 10 hours, reflecting realistic stress conditions such as high-demand windows or regional capacity constraints.

The Result
The system accurately followed the reduced load targets for the full duration *(see Figure 5.3)*. By shifting lower-priority compute tasks to off-peak hours (e.g., late night when wind may be stronger), the facility sustained critical operations while relieving grid strain for durations some storage technologies cannot match. This capability is essential for integrating higher percentages of intermittent renewables into the UK energy mix.

The extended power reduction had minimal impact on job performance, achieving **98.8% performance on highest priority jobs** *(see Figure 5.4).*

5.4 Live Rapid Response to National Grid Events

Grid operators rely on assets that can accept live instructions, respond repeatedly, and remain stable under evolving conditions. Dispatch signals may arrive back-to-back, without warning, and with constraints that test the physical and operational limits of the resource. For flexible AI Factories, this means maintaining precise, reliable power control, even as underlying compute workloads start and stop in real time.

Power-flexible AI Factories need live interfaces through which grid partners can issue real dispatch events, at times of their choosing, with no coordination or advance notice.

The Test
Using the live submission portal, National Grid and EPRI remotely submitted four distinct grid events within a 10-hour window. Emerald AI had no prior knowledge of the timing, magnitude, duration, or ramp constraints of any event.

The Result
The system successfully met all requested power targets across all four events. The two immediate ramp-down events were executed in under 30 seconds and 70 seconds, respectively, while the multi-hour events were held for their full duration without deviation *(see Figure 5.5)*.

Taken together, the results demonstrate a key capability: The system can function as a live, dispatchable resource, not a one-off or pre-scheduled demand response asset. It can accept real-time commands from operators, handle dense and heterogeneous event sequences, and respond even when events push the limits of lead time and ramp speed.

This experiment also showcased Emerald Conductor's ability to seamlessly handle jobs stopping and starting

Figure 5.3
AI cluster reduces power demand by 10% for over 10 hours to help ease grid strain during the “doldrums”.

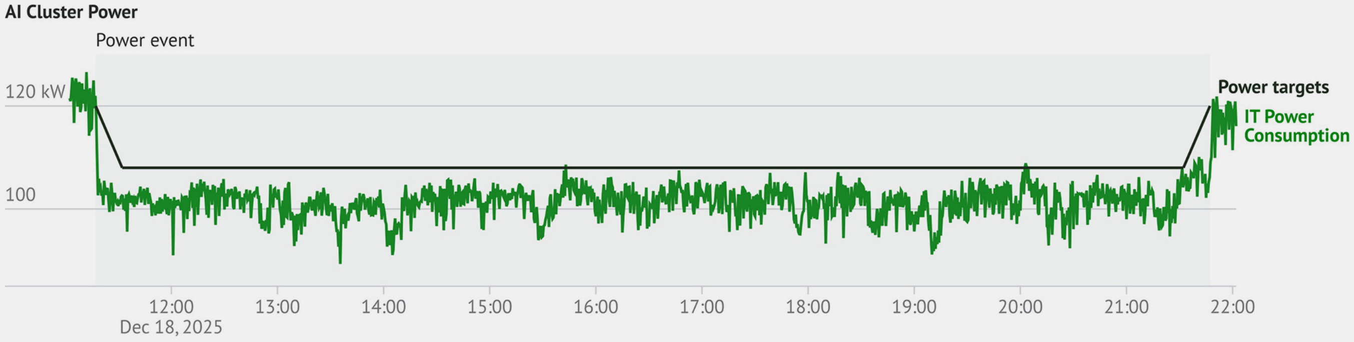


Figure 5.4
Workload performance for 10-hour experiment, showing preservation of performance (measured as normalised throughput) for highest-priority workloads.

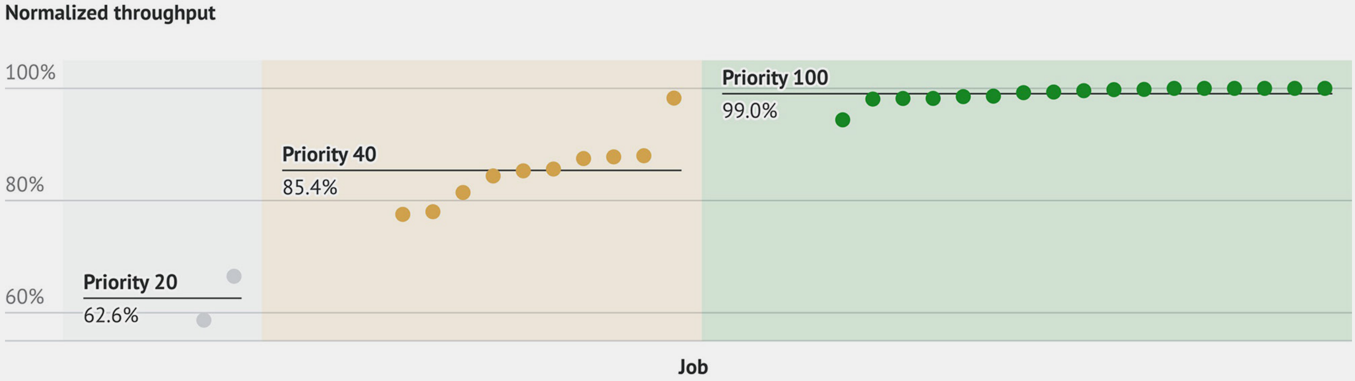


Figure 5.5
AI cluster power response to live grid events submitted by National Grid and EPRI, demonstrating ~30-second and ~70-second response times to zero-notice, immediate-ramp-down events

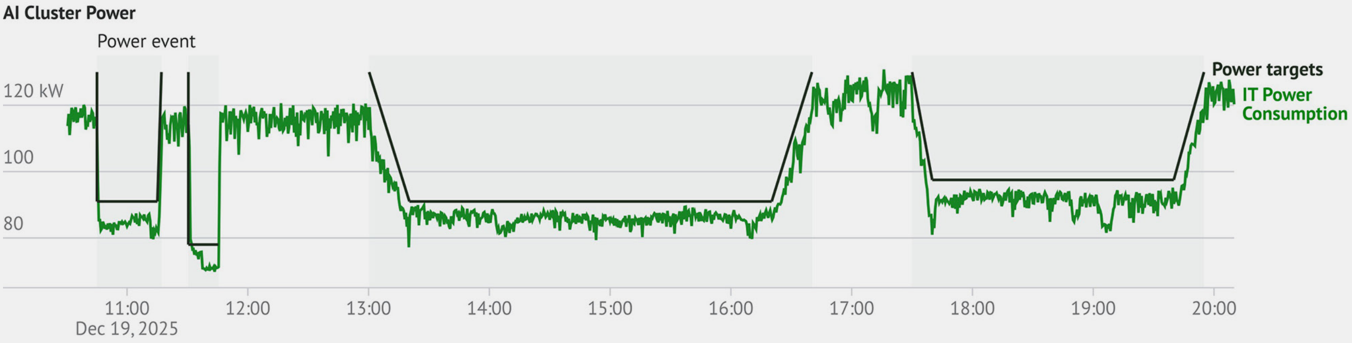
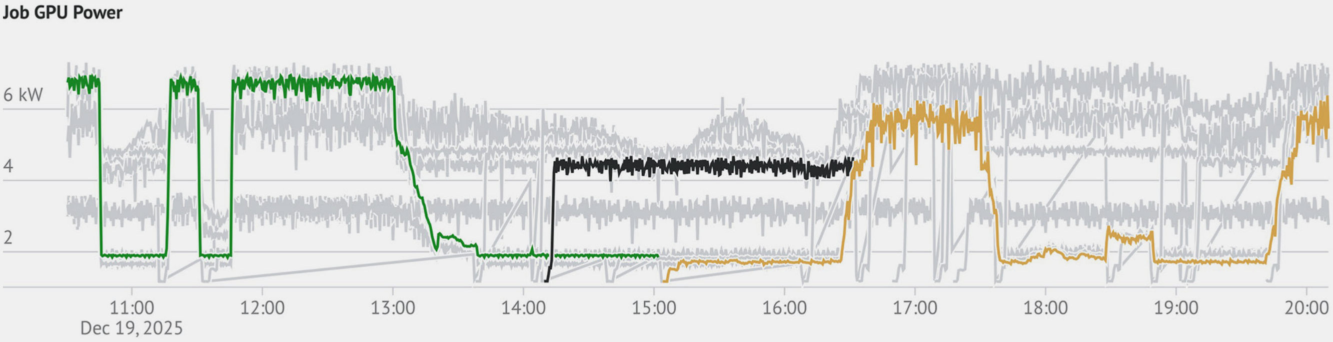


Figure 5.6
Per-job power traces for a period of National Grid and EPRI live event submissions. Jobs can be seen stopping and starting throughout the testing window. Three example jobs representing jobs starting and stopping are highlighted in green, black, and yellow.



frequently throughout the demonstration without violating power targets in a production environment *(see Figure 5.6).*

5.5 Five-Minute Carbon Signal Tracking

As renewable generation increases on the grid in the UK, system carbon intensity varies throughout the day; this is in response to changes in wind output, solar availability, imports, and the thermal generation setting dispatch. These variations occur at intra-hour timescales, reflecting real operational shifts in the generation mix rather than long-term averages.

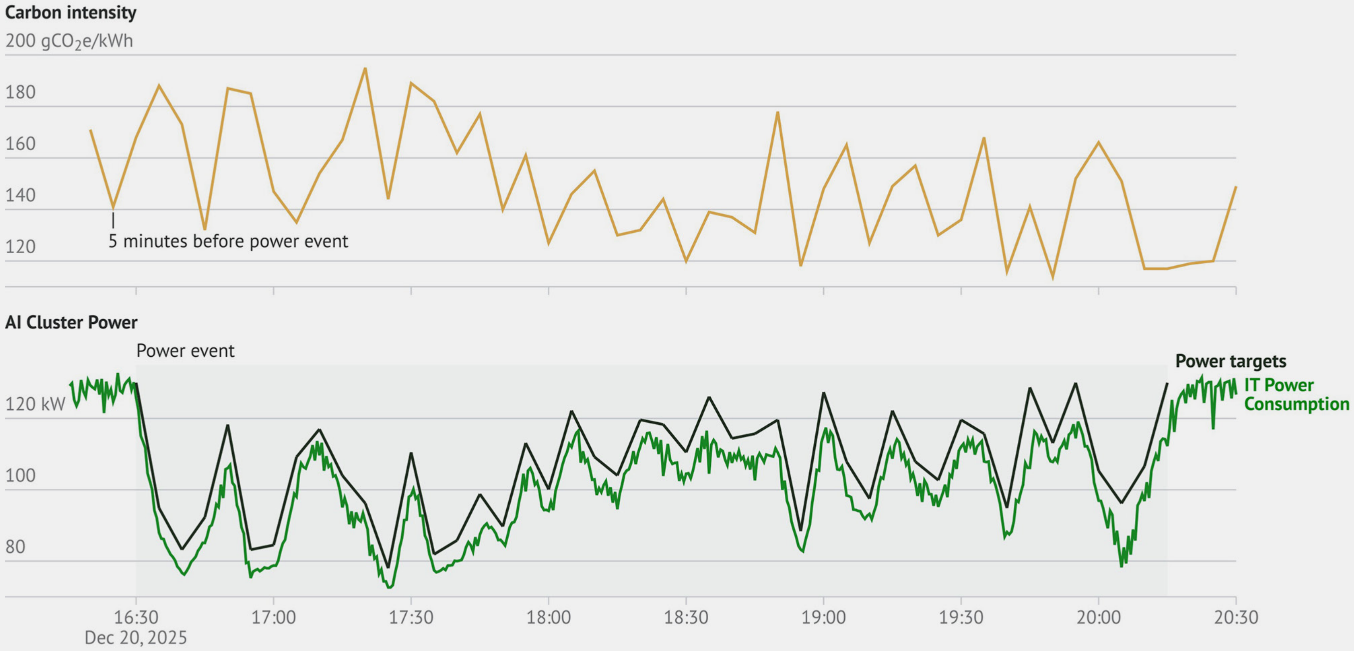
For demand-side flexibility to contribute meaningfully to emissions reduction, it must be able to respond to time-varying carbon intensity signals at comparable temporal resolution. Coarse schedules or static time-of-use assumptions do not capture these dynamics and can misalign demand with actual system conditions. Following high-granularity signals requires stable control, frequent adjustment, and the ability to modulate load repeatedly without introducing instability or operational disruption.

The Test
The cluster was instructed to follow a carbon intensity signal at 5-minute granularity derived from Great Britain's active generation mix over time; this reflected changes in system emissions intensity driven by short-term variations in the generation mix.

The system was required to continuously adjust power consumption in response to each signal update, rather than responding to discrete events or thresholds.

The Result
The cluster successfully tracked the carbon intensity signal, reducing power consumption during higher-intensity periods and increasing utilisation when system emissions were lower. Power adjustments were executed consistently at the required 5-minute cadence, demonstrating the ability to follow frequent, fine-grained signals over time *(see Figure 5.7)*.

Figure 5.7
Carbon-aware load following and power tracking across grid conditions.



06 Enabling Faster Connections and Smarter Standards

Power-flexible AI has now arrived, as evidenced by the results above. However, regulation must also evolve to operationalise and leverage it.

We propose three specific mechanisms to fast-track larger and faster connections for flexible AI data centres:

1. **Alternative Connection Agreements (Non-Firm Access):** Currently, most data centres apply for “firm” connections, which guarantee 100% power availability 24/7. This requires the grid to be built for the “worst-case scenario”. By utilising an alternative connection agreement model, an AI Factory accepts a connection where the grid operator has the right to curtail power to the AI Factory during specific constraint windows.

2. **A Market for “Compute-as-Flexibility”:** Flexibility market mechanisms need to be attractive to encourage participation from a broad range of large loads willing to reduce or shift their consumption. This revenue stream improves the economics of the data centre, ultimately lowering the cost of compute for UK innovators, lowers costs of the connection for customers and the cost of network and generation build for consumers.

3. **Dynamic Tariffs for Gigascale Loads:** Data Centres already connected to the grid could help grid operators through greater flexibility. For this, energy system players should introduce dynamic transmission tariffs and more attractive contractual compensation arrangements that incentivise load shifting, and the ability for large loads to be available and dispatch the required flexibility. If an AI Factory can shift its training runs to follow wind generation patterns (as demonstrated in the “Doldrums” test), it should pay lower energy costs. This aligns the economic incentives of the AI operator with the physical reality of the grid.

07 From Proof Point to Gigascale: The Aurora Project

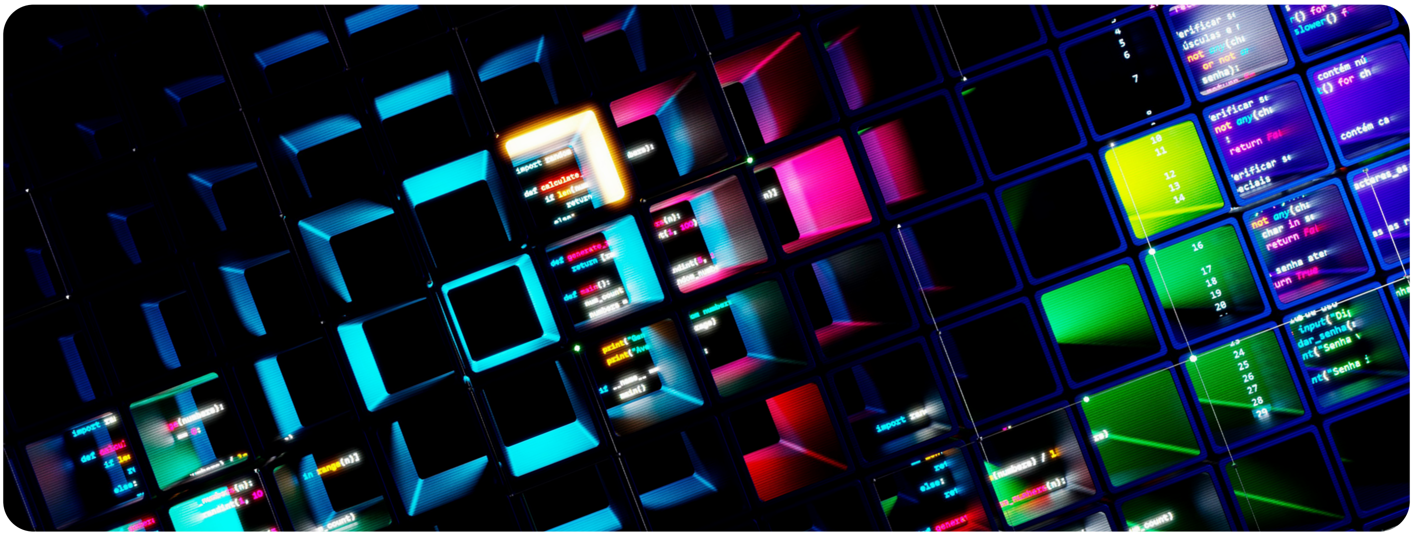
The London demonstration provides operational evidence that AI power flexibility is reliable, fast, and precise. However, 130 kW is a pilot; the market requires hundreds of megawatts.

Building on this foundation and an earlier demonstration at an Arizona data centre in the US, Emerald AI, EPRI, NVIDIA, Digital Realty, and PJM have announced plans for Aurora, a nearly 100 MW power-flexible AI Factory in Virginia, planned for 2026.

Moving from a single cluster to a 100 MW facility introduces additional complexity, particularly in facility-level thermal management, where power, airflow, and cooling capacity must be coordinated across thousands of GPUs and multiple electrical and mechanical subsystems. However, the core logic demonstrated in London, namely software-defined, autonomous power control, remains the governing principle and can scale from the kilowatt level to megawatt-scale deployments and beyond. At larger scales, Emerald AI's Conductor platform can support a more holistic form of facility management by incorporating thermal and cooling constraints directly into its power orchestration decisions.

Emerald AI is supporting NVIDIA's DSX Omniverse reference design for gigascale AI Factories. This initiative aims to make power flexibility a built-in operational capability for NVIDIA Cloud Partners, enabling providers like Nebius to potentially adopt these capabilities as a standard feature.

08 Conclusion



The data gathered from the London trial offers a constructive path toward incorporating demand side flexibility into AI facilities.

Power-flexible AI Factories prove that performance and power responsibility are not mutually exclusive; instead, they can be designed together. By replacing the rigid “firm load” models of the past with measurement-based flexibility, grid operators and policymakers can create new options for delivering capacity efficiently and enabling timely connections for the data centres that will power growth. This approach supports the UK’s strategic goals by making best use of existing network capability, while major infrastructure investments continue as needed.

As stated by Steve Smith of National Grid, “As the UK’s digital economy accelerates, there’s concern that data centres could add pressure to an already constrained system. This trial proves the opposite can be true. High-performance data centres don’t have to place additional strain on the grid. With our partners, we’ve shown they can be connected and managed without major new network capacity, flexing their power up or down in real time to support the whole system. This approach will enable us to connect significant new demand more quickly and help to lower network charges for customers over time.”

Authors & Acknowledgements

Authors

Chris Williams, Philip Colangelo, Ayse Coskun, Ethan Levine, Andy Neale, Ciaran Roberts, Shayan Sengupta, Nikhil Shirolkar, Varun Sivaram, Sarah Soares, Ethan Tiao, Scott Underwood, Daniel Wilson at **Emerald AI;** Frank Sharp at **EPRI;** Ben Davis, Neha Moturi, Luke Wainwright at **National Grid;** Aaron Miller, Igor Znamenskii at **Nebius**

Acknowledgements

This report benefited greatly from input and review by collaborators across the National Grid, Nebius, NVIDIA, EPRI, and Emerald AI teams, including: Tony Bedard, David Porter, Anuja Ratnayake at **EPRI;** Steve Smith, Pradeep Tagare, Sonia Brown, David Adkins, Jason Avramidis, Abbie Badcock-Broe, Miranda Cochrane, Gerard Kelly, Niki Kesharaju, Lydia Ogilvie, Caroline Perry, Alex Roache, Peter Thorp, Luke Wainwright, Richard Woodward at **National Grid;** Marc Boroditsky, Meghan Horvath, Peter Morley, Daria Mukhortova, Mike Welsh at **Nebius;** Nico Caprez, Josh Parker, Vladimir Troy, Marc Spieler, Ahsan Yousufzai, Zack MacFarland, and Kira Sarkisian at **Nvidia;** and Mia Beams, Swasti Jain, Jack Megrue, and Aroon Vijaykar at **Emerald AI.**

Contacts

press@emeraldai.co
media.uk@nationalgrid.com
media@nebius.com
news@epri.com

Suggested Citation

Williams et al. Power-Flexible AI Factories: A UK-First Demonstration of Grid-Responsive AI Infrastructure. Emerald AI, EPRI, National Grid, Nebius, March 2026.