How AI Factories Can Help Relieve Grid Stress

Emerald AI, an NVIDIA Inception startup, is developing software to control power use during times of peak grid demand while meeting the performance requirements of data center AI workloads.

July 1, 2025

Marc Spieler, Senior Managing Director of Energy, NVIDIA

In many parts of the world, including major technology hubs in the U.S., there's a yearslong wait for AI factories to come online, pending the buildout of new energy infrastructure to power them.

Emerald AI is developing an AI solution that could enable the next generation of data centers to come online sooner by tapping existing energy resources in a more flexible and strategic way.

"Traditionally, the power grid has treated data centers as inflexible — energy system operators assume that a 500-megawatt AI factory will always require access to that full amount of power," said Varun Sivaram, founder and CEO of Emerald AI. "But in moments of need, when demands on the grid peak and supply is short, the workloads that drive AI factory energy use can now be flexible." That flexibility is enabled by the startup's Emerald Conductor platform, an AI-powered system that acts as a smart mediator between the grid and a data center. In a recent field test in Phoenix,

Arizona, the company and its partners demonstrated that its software can reduce the power consumption of AI workloads running on a cluster of 256 NVIDIA GPUs by 25% over three hours during a grid stress event while preserving compute service quality.

Emerald AI achieved this by orchestrating the host of different workloads that AI factories run. Some jobs can be paused or slowed, like the training or fine-tuning of a large language model for academic research. Others, like inference queries for an AI service used by thousands or even millions of people, can't be rescheduled, but could be redirected to another data center where the local power grid is less stressed.

Emerald Conductor coordinates these AI workloads across a network of data centers to meet power grid demands, ensuring full performance of time-sensitive workloads while dynamically reducing the throughput of flexible workloads within acceptable limits.

Beyond helping AI factories come online using existing power systems, this ability to modulate power usage could help cities avoid rolling blackouts, protect communities from rising utility rates and make it easier for the grid to integrate clean energy.

"Renewable energy, which is intermittent and variable, is easier to add to a grid if that grid has lots of shock absorbers that can shift with changes in power supply," said Ayse Coskun, Emerald AI's chief scientist and a professor at Boston University. "Data centers can become some of those shock absorbers."

A member of the NVIDIA Inception program for startups and an NVentures portfolio company, Emerald AI today announced more than $24 million in seed funding. Its Phoenix demonstration, part of EPRI's DCFlex data center flexibility initiative, was executed in collaboration with NVIDIA, Oracle Cloud Infrastructure (OCI) and the regional power utility Salt River Project (SRP).

"The Phoenix technology trial validates the vast potential of an essential element in data center flexibility," said Anuja Ratnayake, who leads EPRI's DCFlex Consortium.

EPRI is also leading the Open Power AI Consortium, a group of energy companies, researchers and technology companies — including NVIDIA — working on AI applications for the energy sector.

Using the Grid to Its Full Potential

Electric grid capacity is typically underused except during peak events like hot summer days or cold winter storms, when there's a high power demand for cooling and heating. That means, in many cases, there's room on the existing grid for new data centers, as long as they can temporarily dial down energy usage during periods of peak demand.

A recent Duke University study estimates that if new AI data centers could flex their electricity consumption by just 25% for two hours at a time, less than 200 hours a year, they could unlock 100 gigawatts of new capacity to connect data centers — equivalent to over $2 trillion in data center investment.

Putting AI Factory Flexibility to the Test

Emerald AI's recent trial was conducted in the Oracle Cloud Phoenix Region on NVIDIA GPUs spread across a multi-rack cluster managed through Databricks MosaicML.

"Rapid delivery of high-performance compute to AI customers is critical but is constrained by grid power availability," said Pradeep Vincent, chief technical architect and senior vice president of Oracle Cloud Infrastructure, which supplied cluster power telemetry for the trial. "Compute infrastructure that is responsive to real-time grid conditions while meeting the performance demands unlocks a new model for scaling AI — faster, greener and more grid-aware."

Jonathan Frankle, chief AI scientist at Databricks, guided the trial's selection of AI workloads and their flexibility thresholds.

"There's a certain level of latent flexibility in how AI workloads are typically run," Frankle said. "Often, a small percentage of jobs are truly non-preemptible, whereas many jobs such as training, batch inference or fine-tuning have different priority levels depending on the user."

Because Arizona is among the top states for data center growth, SRP set challenging flexibility targets for the AI compute cluster — a 25% power consumption reduction compared with baseline load — in an effort to demonstrate how new data centers can provide meaningful relief to Phoenix's power grid constraints.

"This test was an opportunity to completely reimagine AI data centers as helpful resources to help us operate the power grid more effectively and reliably," said David Rousseau, president of SRP.

On May 3, a hot day in Phoenix with high air-conditioning demand, SRP's system experienced peak demand at 6 p.m. During the test, the data center cluster reduced consumption gradually with a 15-minute ramp down, maintained the 25% power reduction over three hours, then ramped back up without exceeding its original baseline consumption (figure 1).

AI factory users can label their workloads to guide Emerald's software on which jobs can be slowed, paused or rescheduled — or, Emerald's AI agents can make these predictions automatically.

Orchestration decisions were guided by the Emerald Simulator, which accurately models system behavior to optimize trade-offs between energy usage and AI performance (figure 2). Historical grid demand from data provider Amperon confirmed that the AI cluster performed correctly during the grid's peak period.

Figure 1: (Left panel): AI GPU cluster power consumption during SRP grid peak demand on May 3, 2025; (Right panel): Performance of AI jobs by flexibility tier. Flex 1 allows up to 10% average throughput reduction, Flex 2 up to 25% and Flex 3 up to 50% over a six-hour period.

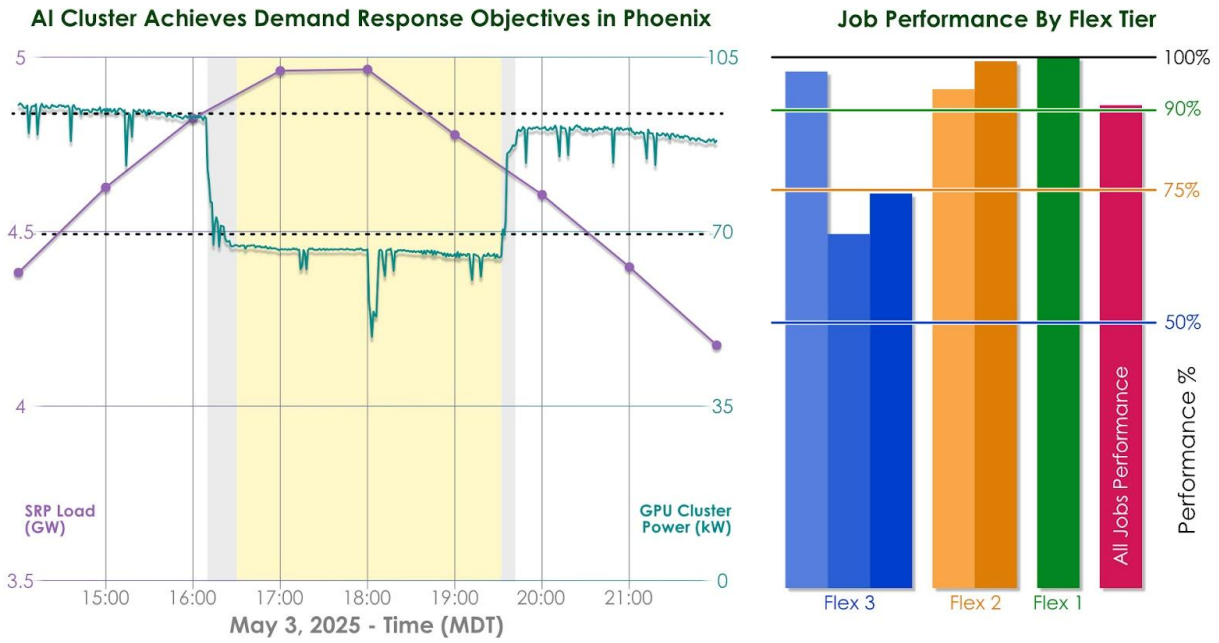
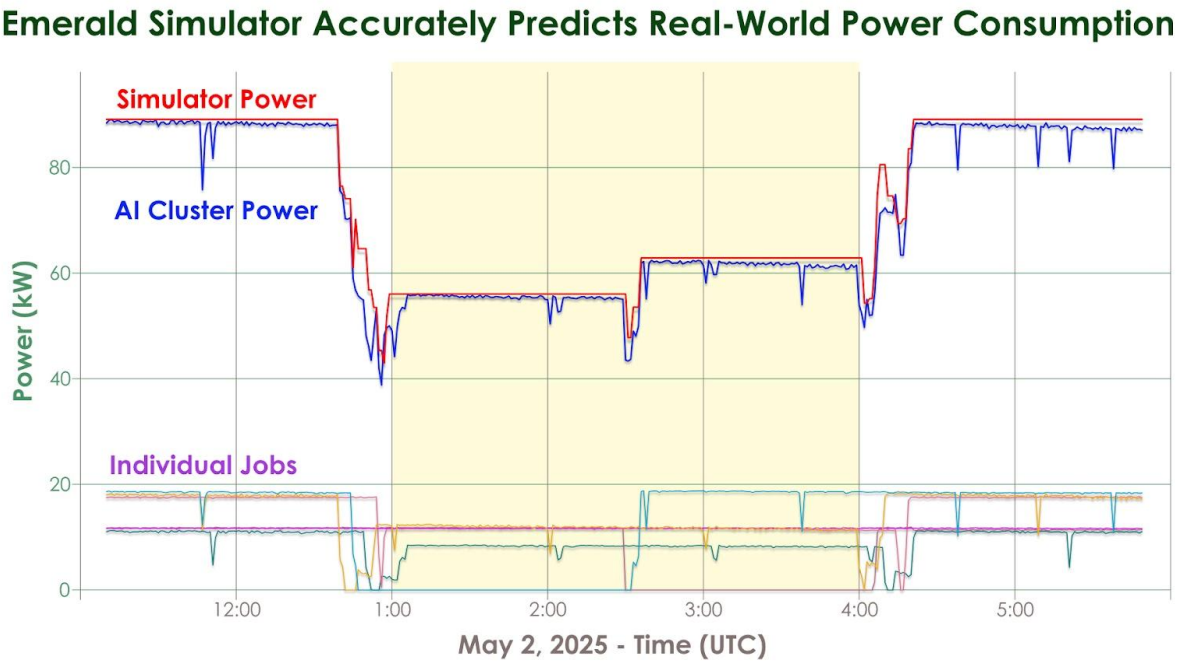


Figure 2: Comparison of Emerald Simulator prediction of AI GPU cluster power with real-world measured power consumption.



https://blogs.nvidia.com/blog/ai-factories-flexible-power-use/

Forging an Energy-Resilient Future

The International Energy Agency projects that electricity demand from data centers globally could more than double by 2030. In light of the anticipated demand on the grid, the state of Texas passed a law that requires data centers to ramp down consumption or disconnect from the grid at utilities' requests during load shed events.

"In such situations, if data centers are able to dynamically reduce their energy consumption, they might be able to avoid getting kicked off the power supply entirely," Sivaram said.

Looking ahead, Emerald AI is expanding its technology trials in Arizona and beyond — and it plans to continue working with NVIDIA to test its technology on AI factories.

"We can make data centers controllable while assuring acceptable AI performance," Sivaram said. "AI factories can flex when the grid is tight — and sprint when users need them to."

Learn more about NVIDIA Inception and explore AI platforms designed for power and utilities.
