AI Inference

How the demands of AI are redefining what energy storage must do—and why zinc meets the moment

By Francis Richey and Justin Vagnozzi

The conversation about AI and energy has focused on volume. The harder problem is speed.

Every time someone asks an AI a question, a power system somewhere has to respond—not in seconds, but in the time it takes to blink. Millions of people, businesses, and governments are using that AI simultaneously, and every request creates a small spike in power demand. Aggregated across a data center, those individual requests produce rapid, unpredictable swings. This is an energy use scenario most power generation systems were not built to handle.

To meet the urgent and enormous energy needs of AI—and bypass multi-year grid interconnection bottlenecks—data center developers are increasingly planning to generate their own power rather than draw from the grid. Building that behind-the-meter generation capacity is a real but solvable engineering and construction problem. The real question is whether those systems will keep up with what AI workloads actually require.

Most traditional power generation—like gensets and turbines—ramps slowly and runs most efficiently at a steady output, while renewables—like solar and wind—produce intermittently. AI inference is the opposite—fast, volatile, and unrelenting. That gap is what Eos energy storage infrastructure was built to close.

Data centers weren't always this complex to power.

As data centers have evolved, the demands they place on power infrastructure have changed in kind. Today's hyperscale facilities often support three distinct types of workloads, frequently within a single site. Each one asks something different of the power system.

Cloud services: steady and predictable.

Traditional cloud services—hosting websites, running enterprise applications, storing data—follow a predictable daily pattern. The volume of energy is large, but most power generation sources handle it well. Energy storage plays a supporting role: absorbing minor demand fluctuations or smoothing short periods of renewable intermittency throughout the day.

AI training: scheduled but not smooth.

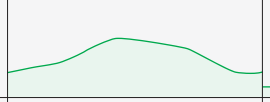
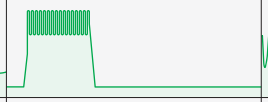
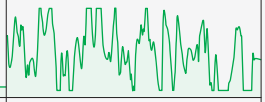
Training AI models—building the large language models and other systems that power today's AI applications—is schedulable work, typically run as batch jobs during off-peak hours. It's predictable, but not smooth. Thousands of processors work in sync—all drawing heavy power at the same moment, then briefly pausing together while they exchange data. That coordinated rhythm creates a pulsing pattern in the facility's total power demand: a high sustained baseline with step-like oscillations on top. Whether the site runs on generators or renewables, energy storage becomes a critical element of the system—handling startup ramps, smoothing the synchronized swings and short-term intermittencies. And on solar-only campuses, they must now serve as the primary overnight power source. Training is easier to plan for than inference, but it demands substantially more from the power system than traditional cloud.

AI inference: volatile and unpredictable.

Inference—the work of running trained models in real time—is a new paradigm: it cannot be planned for. Consumers, businesses, and governments access these systems around the clock—asking questions, generating content, running analyses, and performing tasks, all at varying levels of complexity, and all with urgent expectations of delivery. When aggregated, the resulting sharp and unpredictable swings can double demand in milliseconds and drop just as fast. To put specific numbers to it: inference workloads typically operate at 30–70% of a facility's capacity, with load changes of 10–50% per minute. Power response is expected in under a second—often in double-digit milliseconds. There is no schedule, no gradual ramp, and no predictable pattern.

AI inference introduces an entirely new type of workload that most conventional power generation systems and many battery technologies struggle to handle. Eos engineered its long-duration, zinc-based chemistry to solve for this type of complexity—a capability that repositions its energy storage for AI applications from backup equipment to critical infrastructure.

Three workloads. Three power profiles.
How cloud services, AI training, and AI inference differ in what they demand from a power system.

	Cloud services	AI training	AI inference.
			
Load variability	Low	High, synchronized	Very high, random
Ramp rate	Minutes-hours	Seconds within cycle	10-50% per minute
Response time	Seconds	Subseconds during cycles	As fast as 3-5 millseconds
Predictability	High	Scheduled, not smooth	None

What happens when power can't keep up.

The emerging strategy for AI-powered data centers is to place energy storage between the power generation source and the facility. The storage acts as a buffer—absorbing the rapid swings in demand from the data center so that generators can operate at a steady, efficient output.

In principle, batteries respond faster than generators. But AI inference doesn't just ask storage systems to respond quickly once. It asks for continuous, rapid charging and discharging—deep swings in power, sustained over hours, day after day, night after night. That sustained stress is where many battery technologies begin to struggle.

While conventional battery technologies are well-suited to many applications, AI inference can expose a common, important constraint. Under sustained, aggressive cycling for long periods of time, repeated continuously, they can generate more heat, and lose more capacity over time, forcing a choice no operator wants to make: deliver the required response speed and accept accelerated degradation, or throttle performance to protect longevity and fall behind the load fluctuations. There is often no way to configure or engineer around that tradeoff.

On the generation side, the constraints are equally real. Generators—whether piston-driven or turbine-based—are designed to run at or near full output. Piston-driven generators lose efficiency significantly below about 50% of their rated output; turbines begin to fall off closer to 70%. When AI inference pulls demand down sharply, the generator either runs in that inefficient range or shuts off entirely—requiring a restart when the next spike arrives. Without storage that can absorb those swings reliably and keep the generator in its optimal band, the entire power system suffers.

The consequences are not just technical. An energy storage system that runs hot to keep up with power demands requires its cooling system to work even harder—pushing up operating costs. A GPU that sits idle because power could not ramp fast enough is wasted capital. Every millisecond of latency is a cost—measured in processing capacity, in user experience, and in revenue.

Not all storage is built for this.

How zinc works.
Eos's battery chemistry is inspired by zinc plating baths—the same industrial process used to coat metals with a protective zinc layer. Inside each Z3 module, energy is stored by depositing zinc from a water-based solution (electrolyte) onto an electrode. When the battery discharges, the zinc dissolves back into the electrolyte, releasing energy.

Energy storage is often discussed as though any battery can fill any role. For AI inference, that assumption does not hold. The requirements—response times measured in single-digit milliseconds, sustained rapid cycling over large states of charge, tolerance for unpredictable load patterns across thousands of cycles—go beyond what most conventional battery chemistries can sustain without accelerated degradation.

The Eos long-duration technology is based on a proprietary zinc-halide chemistry that behaves differently under these conditions. It can cycle from fully empty to fully charged and back—its complete energy range—without the kind of degradation that constrains how aggressively other chemistries can be used. It is water-based and non-flammable, which reduces the thermal safety constraints that limit how hard many other systems can be pushed. And its behavior under rapid, dynamic loading—sharp transitions at millisecond intervals—has been tested and validated specifically for AI inference workloads.

Tested at every level.

To validate its technology for AI data center applications, Eos conducted a series of tests designed to answer progressively harder questions. Could a single battery module handle the large power fluctuations at the millisecond time scale without generating significant heat? Could it sustain that performance over multiple hours of continuous cycling? Could a full, commercial-scale system—with all of its real-world components—deliver the same results?

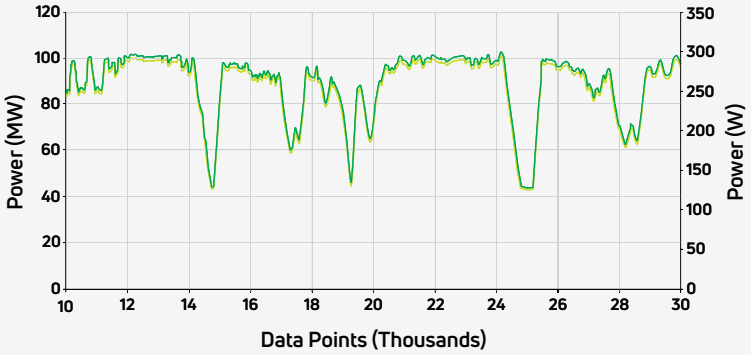
Can a single module keep up?

Eos took a real AI data center load profile—originally scaled for a 100 MW facility—and applied specific intervals directly to individual Eos Z3™ battery modules. In the initial testing, the signal delivered a new power command every 3 to 5 milliseconds, simulating the fastest fluctuations seen within the AI workload. The modules tracked the signal precisely. At the 3-millisecond measurement interval, there was negligible lag and no deviation between the command and the battery's response.

Figure 1
A real AI data center load profile applied to a single module—new command every 3 to 5 milliseconds.

— GPU load profile (MW)
— Module power (W)

Note: The chart shows ~4 seconds of a 15-minute continuous test.



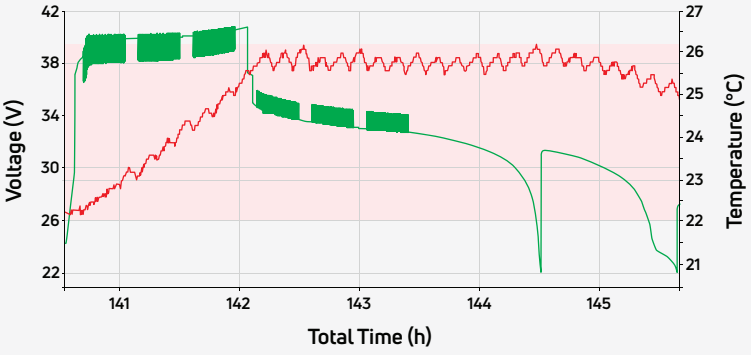
The module tracked the signal precisely. Negligibile lag and deviation.

Can it sustain that over time?

Next, Eos extended the test to larger intervals, longer durations, and deeper cycling. The dynamic workload signal was applied across multiple full charge-and-discharge cycles, each lasting nearly an hour. The module maintained consistent efficiency—approximately 78% round-trip—across every test configuration. Temperature barely changed: less than 4°C of increase. There was no measurable performance degradation from the fastest transitions in the dynamic workload.

Figure 2
Voltage and temperature during a full charge-and-discharge cycle with three consecutive dynamic AI workload runs.

— Voltage (V)
— Temperature (°C)



Temperature rose less than 4°C.

Approximately 78% round-trip efficiency.

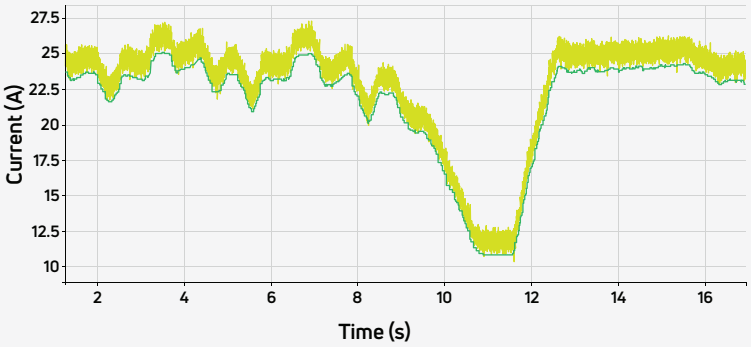
No active cooling needed.

Can a full commercial system do it?

The next stage moved from individual Z3 modules to a complete Eos Indensity Core™—a single unit containing 112 modules along with every component of a commercial, grid-connected energy storage system: inverter, DC/DC converter, printed circuit boards with microcontrollers and MOSFETs, and Eos DawnOS™ software and controls. This is the building block of an Eos Indensity™ solution. When the fastest fluctuations from the dynamic workload signal were again applied, the full system responded in 3 milliseconds—from a standing start to full power output. None of the additional commercial components—the inverter, the DC/DC converter, the control systems—introduced any meaningful delay.

Figure 3
How closely the full system's output tracked the incoming power signal.

— Signal
— DC-DC setpoint (A)



Full system responded in 3 milliseconds from a standing start.

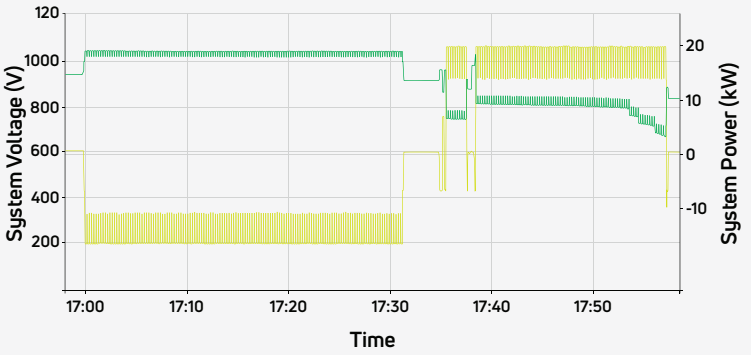
No meaningful delay from commercial components.

Can it handle real GPU demand patterns?

Finally, Eos applied a second data center load profile with a signal designed to more closely mimic the normal power behavior of a GPU array—the sharp, repeating idle-to-peak transitions that characterize AI inference in the real world. The maximum expected variation in a typical data center was exceeded deliberately, pushing the system harder than it would face in practice, but not as fast as the initial tests. This final test was run at 1000V, the full system voltage, with fluctuations of approximately 40% of the system's rated power at 10 to 20 milliseconds intervals for 30 minutes consecutively on charge and discharge. The Indensity Core matched every transition.

Figure 4
Voltage and power as the system responds to rapid, repeating GPU load spikes.

— System voltage
— System power



30 minutes of sustained 10 to 20 millisecond transitions at full system voltage. Power swung ~40% every cycle. Every transition matched.

Millisecond response, sustained without degradation.

The core finding from the test program: the Eos Indensity system can respond to AI inference workloads at both the fastest expected intervals of 3 to 5 millisecond and more normal transitions of 10 to 20 milliseconds, sustained over hours of continuous dynamic cycling, without degradation in performance or accelerated wear or significant temperature changes. Speed alone is not enough—most storage systems can respond quickly for a short burst. The combination of speed, endurance, and longevity is what this application demands, and it is what Eos's proprietary zinc-halide chemistry delivers.

Eos Indensity: an energy storage architecture ideal for AI.

Indensity doesn't just solve the AI inference millisecond-response problem. It meets the exacting operational demands of AI data centers: energy density that constrained sites require; safe co-location with computing infrastructure; a viable route to carbon-free energy, and a return on investment beyond load support.

Four times the energy density.

Data centers are space-constrained environments. Server halls, cooling infrastructure, fire suppression, and on-site generation all compete for acreage. Energy storage is often treated as an afterthought in site planning, forced to fit in whatever space remains. Where conventional storage provides approximately 250 MWh per acre, Eos Indensity can reach roughly four times the energy in the same footprint by stacking Indensity Cores vertically—targeting 1 GWh per acre.

Safe enough to co-locate.

For most battery chemistries, safety codes require setback distances, dedicated enclosures, and supplemental fire suppression. The Eos zinc-halide battery chemistry at the heart of Indensity is water-based and non-flammable, so Indensity systems can be placed immediately adjacent to computing infrastructure or even inside the facility itself. The result is a shorter cable run between storage and the data center, with the ability to feed DC power directly to the native bus. That proximity reduces resistive energy losses, response latency, and capital costs.

Duration that delivers for solar.

Off-grid solar farms represent an opportunity for data centers to lower carbon emissions and fuel costs—if the paired energy storage system can act as the primary power source for overnight runs. Unlike conventional battery technologies that are optimized for short duration applications, the Eos zinc-halide chemistry is a long-duration solution. With a discharge range of 4 to 16+ hours, Eos Indensity provides the firm reservoir that makes 24/7 carbon-free energy viable for AI.

Built for dual use.

Most data centers are being built today with on-site generation, but many will eventually connect to the grid. When that connection exists, the Indensity architecture, through Eos DawnOS controls, can be partitioned at the individual Indensity Core level, allowing an operator to dispatch one portion to buffer the AI load and the other to trade excess stored energy with the grid. A capital investment sized to deliver reliability takes on a second role as a revenue-generating asset when workload drops below full capacity—no additional capacity required.

How Indensity works in a real data center.

A data center running cloud services, AI training, and AI inference moves through different operating states over the course of a day. The power demands of each state are distinct, and the Eos Indensity system adapts to each one.

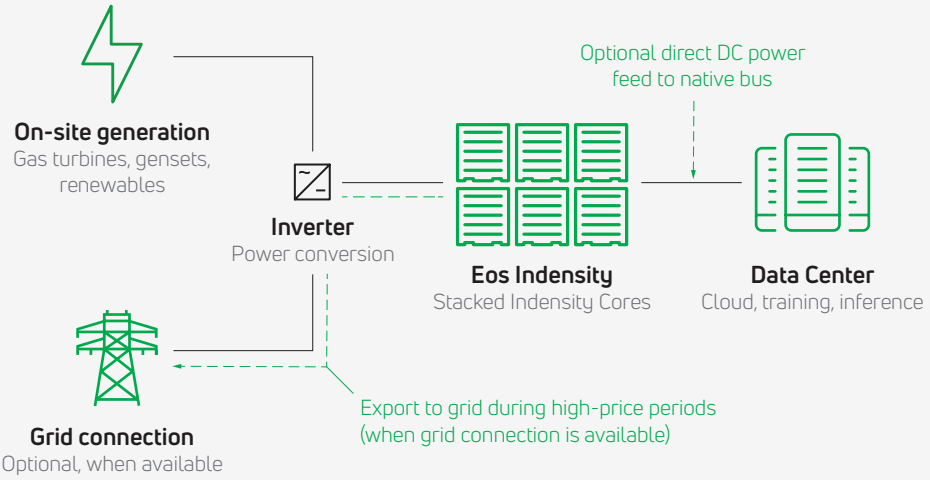
During normal business hours, when cloud services dominate, the generator—or solar farm—runs at a steady, efficient output. The Indensity system manages small load fluctuations or intermittencies in the background—absorbing minor peaks and filling minor dips so the power source doesn't have to chase them.

When inference activity picks up—as it often does unpredictably throughout the day—the Indensity system absorbs the rapid, volatile swings at millisecond speed. The generator stays in its optimal range. It doesn't see the volatility. Indensity handles it. And when co-located, with a direct DC power feed to the native bus, resistive losses, response latency, and capital costs can also be reduced.

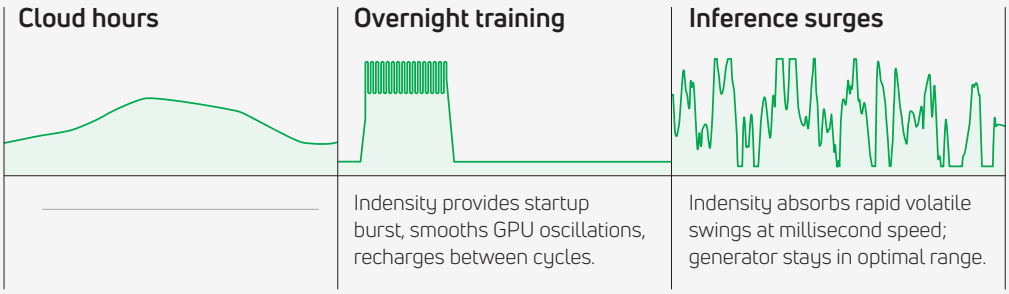
During off-peak hours, when large-scale training jobs run overnight, Indensity provides the initial startup burst while the generator ramps up. Once the run is underway, Indensity smooths the steplike oscillations from synchronized GPU cycles, keeping the generator in its efficient range. Or, if the campus is solar-powered, Indensity simply steps in as the primary power source. Throughout, the system uses excess generation capacity between cycles to replenish reserves for the next day's inference demands.

Across all of these states, DawnOS controls manage the energy level and health of each individual Core—and each Z3 module—keeping the Indensity system within its optimal operating range for maximum efficiency and longevity. DawnOS also feeds real-time system status back to the plant's energy management system, informing critical decisions such as when to ramp generators up or hold them at steady state.

Three operating stages. One storage system.
Power flow architecture: on-site generation, optional grid, Indensity buffer, and facility load



How power flows in each operating state



This is what zinc makes possible.

AI data centers face a power challenge that goes beyond how much energy they need. It is the speed and volatility of that demand, driven by millions of spontaneous, real-time inference requests, that exposes the limits of conventional power infrastructure.

Eos Indensity, built on the company's zinc-halide chemistry and managed by DawnOS at the Z3 module level, delivers the millisecond response speed that inference demands and sustains it across the kind of relentless, unpredictable cycling that would shorten the life of conventional storage. The capability has been tested at the module level, validated at the system level, and designed into the architecture from the ground up.

This is not energy storage adapted for AI. It is storage that was engineered with deep operational flexibility, right from the start, to handle the kind of complexities that AI inference—and increasingly, other modern workloads—now bring. And it was all made possible by zinc.

About the authors

Mr. Richey is Chief Technology Officer at Eos. He is an electrochemical engineering expert with more than a decade of experience in developing, scaling, and commercializing zinc-based battery technology for grid-scale energy storage. He holds a Ph.D. in Chemical and Electrochemical Engineering from Drexel University, 9 patents, and has published in multiple peer-reviewed journals.

Mr. Vagnozzi is Senior VP of Technical Sales and Commercial Operations at Eos, leading the company's commercial operations and talent. He holds a B.S. in Mechanical Engineering from Temple University and more than 20 years' of executive experience in technical sales. He most recently led Sales & Demand Generation for the Sustainable Solutions business of Duke Energy, one of the country's largest energy utilities.