rehlko

AI Readiness Starts with *Power*

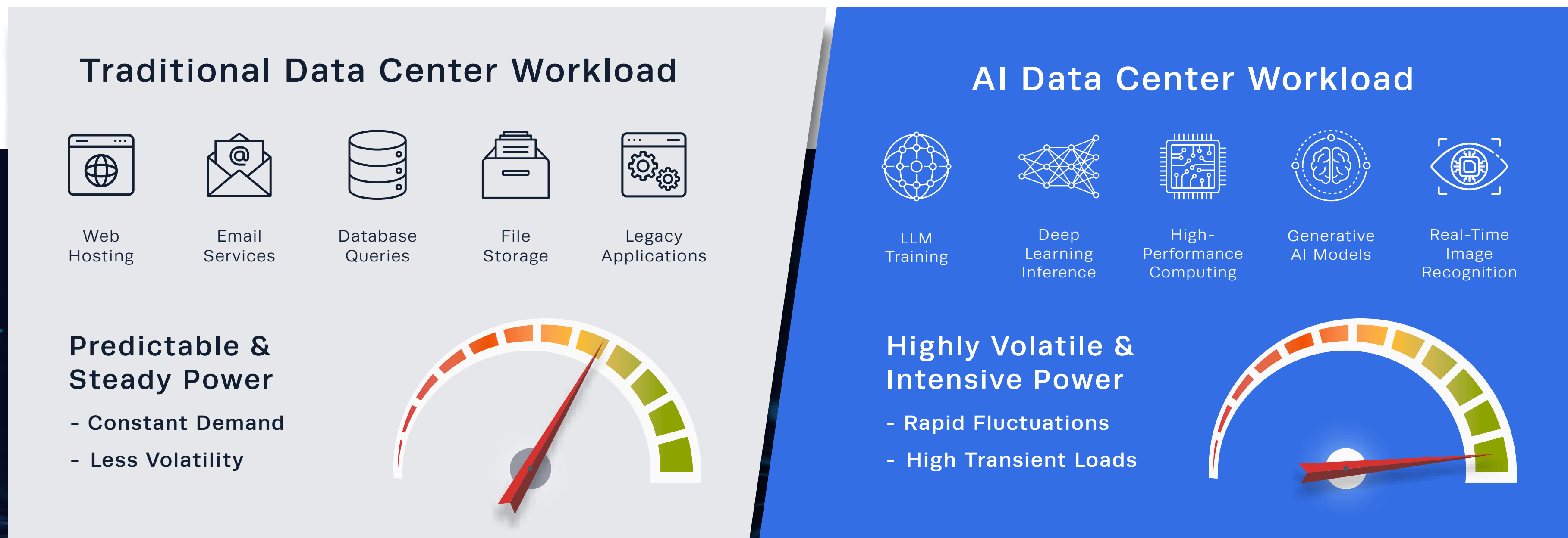
Designing for Real-World Load Conditions

Introduction

Artificial intelligence is rapidly reshaping the infrastructure requirements of the data center industry. While much of the current discussion focuses on the scale of Artificial Intelligent power demand, AI workloads are also introducing operating conditions that differ from the assumptions that have historically guided critical power system design, testing, and validation.

Traditional power architectures were developed around relatively stable operating conditions, where major fluctuations were infrequent and temporary. As AI infrastructure scales, operators and infrastructure providers are increasingly evaluating whether traditional approaches to power architecture, resilience, and performance validation remain sufficient for highly dynamic AI environments.

These changes have important implications for how infrastructure should be designed, validated, and operated. The industry is increasingly shifting from a focus on installed capacity and backup availability toward a greater emphasis on system-level performance under real operating conditions.



The Challenge: Legacy Power Systems Were Not Designed for AI Workloads

Traditional data center infrastructure was designed around relatively stable operating conditions, where major power fluctuations occurred during isolated events such as startup sequences, equipment failures, or load transfers. Power systems were engineered to absorb these transient events, stabilize, and return to steady-state operation.

Large-scale AI training environments rely on thousands of GPUs operating in tightly synchronized compute and communication cycles. Rather than producing smooth and predictable electrical demand, these workloads generate rapid, repetitive, and highly variable power fluctuations across racks, rows, and entire facilities.

In conventional environments, transient behavior was temporary. Under AI workloads, it becomes continuous. This matters because most generator systems, alternators, and supporting controls were optimized for infrequent, high-consequence events rather than sustained dynamic operation. Standards such as ISO 8528-5 reflect this assumption by evaluating generator performance during discrete load acceptance and rejection events with recovery periods between disturbances.

As GPU clusters scale, recovery periods shrink or disappear entirely. Control systems remain continuously active, while mechanical and thermal stresses accumulate rather than dissipate. What was once considered exceptional operating behavior increasingly becomes normal operation.

The challenge is maintaining stable and predictable performance across the power chain under sustained and highly variable AI load conditions.

Why Traditional Generator Assumptions Break Down

Conventional generator systems were designed for a different electrical environment. Mechanical inertia within the engine helps stabilize rotational speed during sudden load events, while alternators and voltage regulation systems respond to transient disturbances within predefined tolerances. These systems perform effectively when disturbances are isolated and recovery periods exist between events.

AI workloads compress recovery windows. Highly synchronized GPU activity creates repeated ramping behavior and rapid shifts in power demand. While UPS and battery systems absorb the fastest transients, generator systems still need to accommodate sustained dynamic loading. Understanding those capabilities can help optimize the size and cost of supporting UPS and battery infrastructure. . Instead of responding to occasional disturbances, systems must continuously regulate against persistent variability.

Under traditional load conditions, the objective was surviving a transient event without unacceptable voltage or frequency deviation. Under AI-driven load behavior, the challenge becomes maintaining consistent performance during ongoing dynamic operation.

That shift has major implications for:

- Voltage and frequency stability
- Equipment stress and thermal loading
- Control system behavior
- Long-term reliability
- and the effectiveness of traditional mitigation strategies

AI infrastructure scales, performance under load increasingly becomes the defining measure of infrastructure readiness.

2.1 The Reality of AI Site Development

AI infrastructure is being deployed faster than traditional power infrastructure can expand to support it. In many markets, utility interconnection timelines now extend several years due to grid constraints, transmission limitations, permitting delays, and rising demand from large-scale data center development. As a result, access to utility power can no longer be assumed as a prerequisite for bringing AI capacity online. This is changing how AI facilities are designed and energized.

Some sites will operate as conventional grid-connected facilities where the utility remains the primary source of power. Others will be partially constrained, requiring on-site generation to supplement limited grid capacity. In some cases, AI campuses may operate as fully islanded environments where on-site generation effectively forms the grid itself.

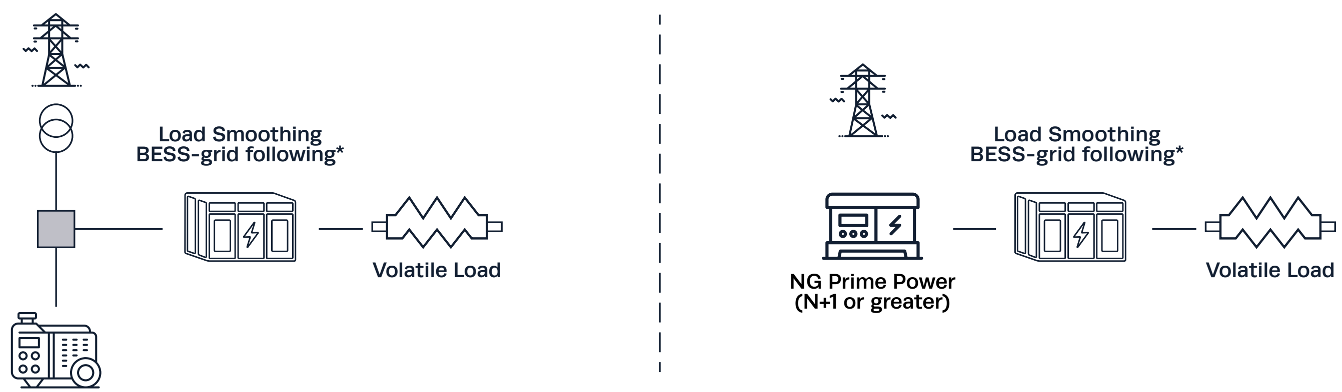
These architectures differ operationally, but they are being driven by the same underlying reality: AI workloads introduce sustained, highly dynamic load behavior that power systems must continuously manage.

Generator systems designed primarily for backup operation behave differently from systems operating as continuous prime power assets. Duty cycles, control strategies, and exposure to dynamic loading changes. Power infrastructure must increasingly sustain stable performance as part of normal AI operations, whether operating behind the grid, alongside it, or independently from it entirely.

Grid-Tied vs. Islanded Generator and Load Smoothing Solutions

Many customer sites do not have a grid tie right now.

Rehlko can furnish diesel and gaseous prime or backup generator solutions for Grid Tied or Islanded DCs.



*Grid forming mode is optimal for fast response load smoothing

2.2 Grid-Tied AI Facilities: Backup-Dominant Power Systems

In grid-tied AI facilities, the utility grid remains the primary source of power while on-site generation serves a backup and resiliency role. These architectures typically rely on diesel generators, automatic transfer switches (ATS), UPS systems, and grid-following battery storage to maintain continuity during outages or disturbances.

Historically, these systems were designed around short-duration, high-consequence events. Generators were expected to start, stabilize, accept load, and operate until utility power returned. Outside of testing and emergency operation, systems spent most of their lifecycle in standby mode.

Even in grid-connected environments, GPU-driven load variability can create operating conditions that differ significantly from traditional enterprise or cloud facilities. During utility disturbances, transfers to generator power may occur under highly dynamic load conditions rather than relatively stable demand profiles.

This places greater stress on generators, alternators, and control systems during the moments when stability matters most.

UPS and battery systems can help absorb fast transients and smooth portions of the load profile. However, these layers reduce exposure to variability rather than eliminate it entirely. System performance still depends on how effectively the underlying generation architecture responds to sustained and rapidly changing demand.

For grid-tied AI facilities, can backup systems maintain stable and predictable performance under AI-driven load behavior when called upon?

The implications of these operating conditions extend beyond generator sizing and capacity planning, influencing how reliability, mitigation, modeling, and validation must be approached in grid-tied AI environments.

3. The Risk: Performance Under Load as a New Reliability Criterion

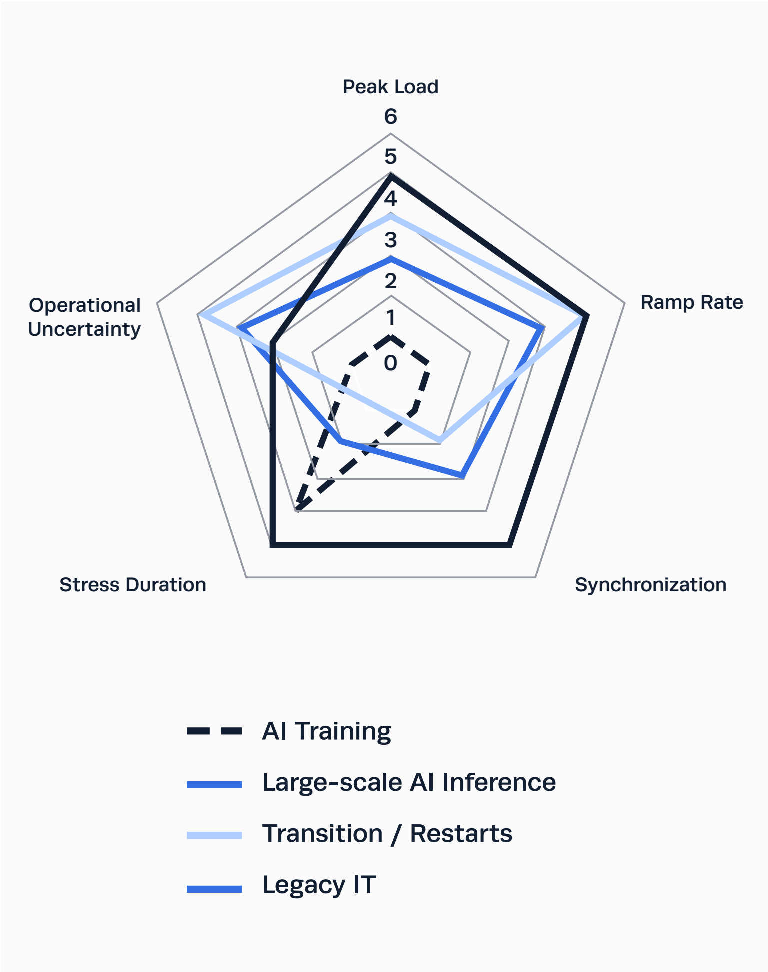
Historically, reliability in critical power environments has been treated as a binary outcome: power is either available or unavailable. Backup systems were evaluated based on uptime, transfer success, and the ability to recover from isolated events.

Under sustained GPU-driven load variability, a power system may remain online while continuously operating near voltage, frequency, or thermal limits. Even without a visible failure, prolonged operation under these conditions can reduce dynamic margin, increase equipment stress, and introduce greater operational uncertainty over time.

Reliability therefore becomes more than system availability. It becomes a question of how consistently power systems can maintain stable performance under continuous dynamic load.

Publicly available research from leading AI infrastructure companies has shown that large-scale AI training environments can cause thousands of GPUs to ramp power consumption simultaneously within extremely short time windows, creating steep and repetitive power transients throughout the electrical infrastructure.

Traditional Reliability Model	Performance-Based Reliability Model AI-Ready
Primary question: Is power available?	**Primary question:** How does power perform under load?
Event-based thinking: - Grid outage - Generator start	**Continuous operating condition:** - Sustained dynamic AI load - Repetitive fast fluctuations
Metrics: - Uptime - Start success rate	**Metrics:** - Voltage stability - Frequency control - Dynamic response time - Margin under stress
Result: - Power present - Behavior assumed acceptable	**Result:** - Power delivered with controlled and validated performances



4. Limitations of Traditional Mitigation Approaches

To manage dynamic load behavior, most critical power architectures rely on mitigation layers designed to smooth or buffer fluctuations before they reach upstream generation systems.

These approaches typically include:

- **Oversizing**
- **UPS systems**
- **Battery energy storage systems (BESS)**
- **Layered control architectures**

UPS systems and battery storage can absorb fast transients, dampen ramp rates, and temporarily decouple GPU-driven load behavior from generators. In doing so, they help recreate operating conditions closer to those assumed by traditional generator standards and validation methods.

However, these systems are designed to attenuate dynamic behavior, not eliminate it entirely.

As GPU clusters scale and variability becomes more continuous, portions of the load dynamics are still transmitted through the power chain. Under sustained operation, system performance increasingly depends on the intrinsic response of generators, alternators, controls, and supporting infrastructure rather than mitigation layers alone.

At the same time, additional attenuation layers introduce greater architectural complexity. Control interactions multiply, system-level validation becomes more difficult, and long-term stability becomes increasingly dependent on the coordinated performance of multiple intermediate systems.

Industry research from Microsoft and NVIDIA has similarly concluded that load-smoothing approaches can reduce the severity of power fluctuations, but do not remove the underlying dynamic behavior associated with large-scale AI workloads.

AI readiness cannot be achieved solely through stacked mitigation technologies. Reliability increasingly depends on how integrated power architectures perform under sustained and variable operating conditions.

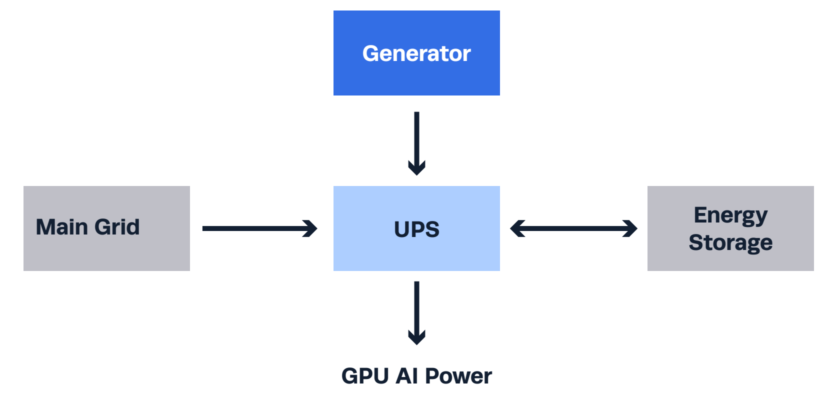
4. Limitations of Traditional Mitigation Approaches

As GPU clusters scale and variability becomes more continuous, portions of the load dynamics are still transmitted through the power chain. Under sustained operation, system performance increasingly depends on the intrinsic response of generators, alternators, controls, and supporting infrastructure rather than mitigation layers alone.

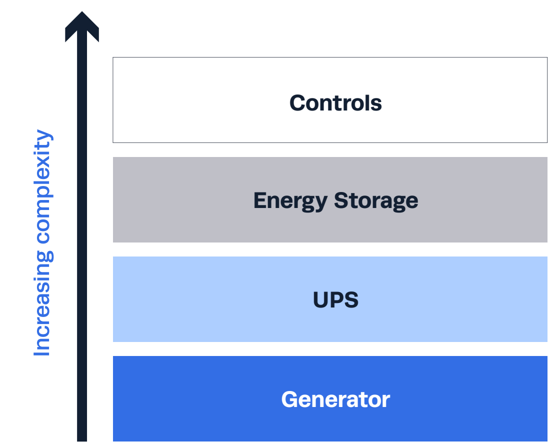
At the same time, additional attenuation layers introduce greater architectural complexity. Control interactions multiply, system-level validation becomes more difficult, and long-term stability becomes increasingly dependent on the coordinated performance of multiple intermediate systems.

Research into AI power consumption has shown that load-smoothing approaches can reduce the severity of power fluctuations, but they do not eliminate the underlying dynamic behavior associated with large-scale AI workloads.

AI readiness cannot be achieved solely through stacked mitigation technologies. Reliability increasingly depends on how integrated power architectures perform under sustained and variable operating conditions.



Power Architecture With Ups As An Attenuation Layer



Limitations of Layered Attenuation

5. Modeling for Reality: Understanding Power System Behavior Before Deployment

As AI workloads introduce new demands on power infrastructure, modeling is becoming an increasingly important tool for understanding how systems will perform before they are deployed.

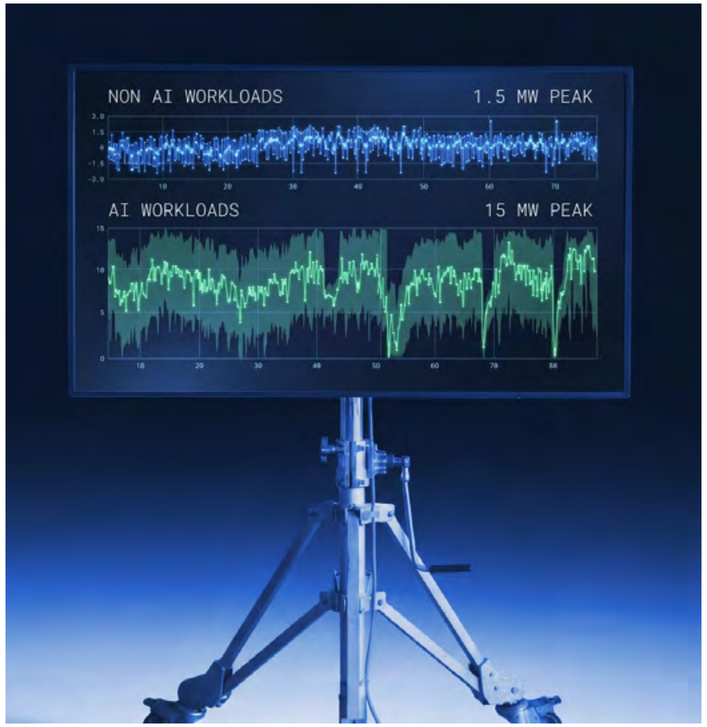
Traditional modeling approaches assume systems recover to a stable operating condition after a transient event. AI workloads challenge that assumption by introducing rapid, repeated load changes that can keep power systems operating in a near-continuous state of recovery. Modeling must therefore capture not only the final operating point, but the system behavior throughout the transient response.

Modeling helps infrastructure teams evaluate how power systems respond to changing operating conditions, including rapid load ramps, transient events, utility disturbances, and generator transitions. By simulating these scenarios before deployment, potential issues can be identified earlier and system performance can be assessed under a wider range of conditions than would be practical through physical testing alone.

Rehlko's approach begins with physical generator testing, using measured performance data to develop validated models that accurately represent both steady-state and transient system behavior. These test-based models are designed to provide a representation of generator behavior based on measured performance data under dynamic operating conditions.

Rather than serving as standalone engineering tools, these validated generator models support broader system-level analysis by providing UPS manufacturers and infrastructure teams with more accurate inputs for digital twin environments and power system optimization. This enables designers to evaluate interactions across generators, energy storage systems, controls, and facility loads before commissioning, supporting more informed decisions around generator sizing, energy storage requirements, and overall system architecture.

By reducing uncertainty early in the design process, Rehlko helps infrastructure teams move beyond conservative oversizing toward validated, performance-driven power system design.



5.1 Advancing Infrastructure Design: Digital Twins and System-Level Thinking

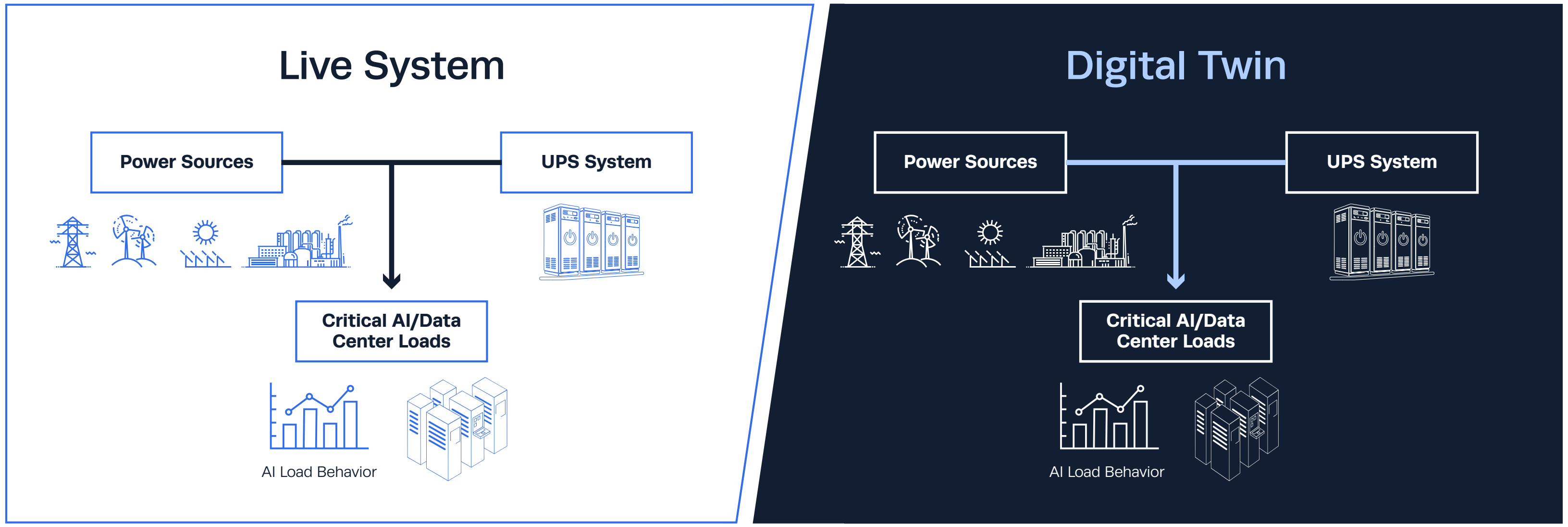
As AI power systems become more interconnected, infrastructure teams are increasingly using digital twins to evaluate system behavior before equipment is deployed. These models bring together validated representations of generators, UPS systems, battery storage, controls, and facility loads to simulate how the complete power architecture will respond under a wide range of operating conditions.

By combining simulation, testing, and operational data, these tools can provide a more complete view of how power architectures behave under changing operating conditions. Instead of evaluating assets independently, infrastructure teams can assess how generators, energy storage, controls, and AI workloads interact across the full system.

Rehlko contributes to this process by testing physical generator systems, using measured performance data to create validated generator models, and sharing those models with technology partners. Those models become part of a larger system-level digital twin, allowing designers to evaluate equipment interactions, optimize system architecture, and explore operating scenarios before construction begins.

This approach can help infrastructure teams evaluate and optimize generator sizing, control strategies, and energy storage requirements before deployment. By evaluating system behavior in a virtual environment, designers can reduce reliance on conservative assumptions and unnecessary oversizing while gaining greater confidence in overall system performance.

Looking ahead, digital twin technologies may enable continuous monitoring, optimization, and performance assessment throughout the operational life of a facility, helping operators better understand system behavior as AI workloads evolve.



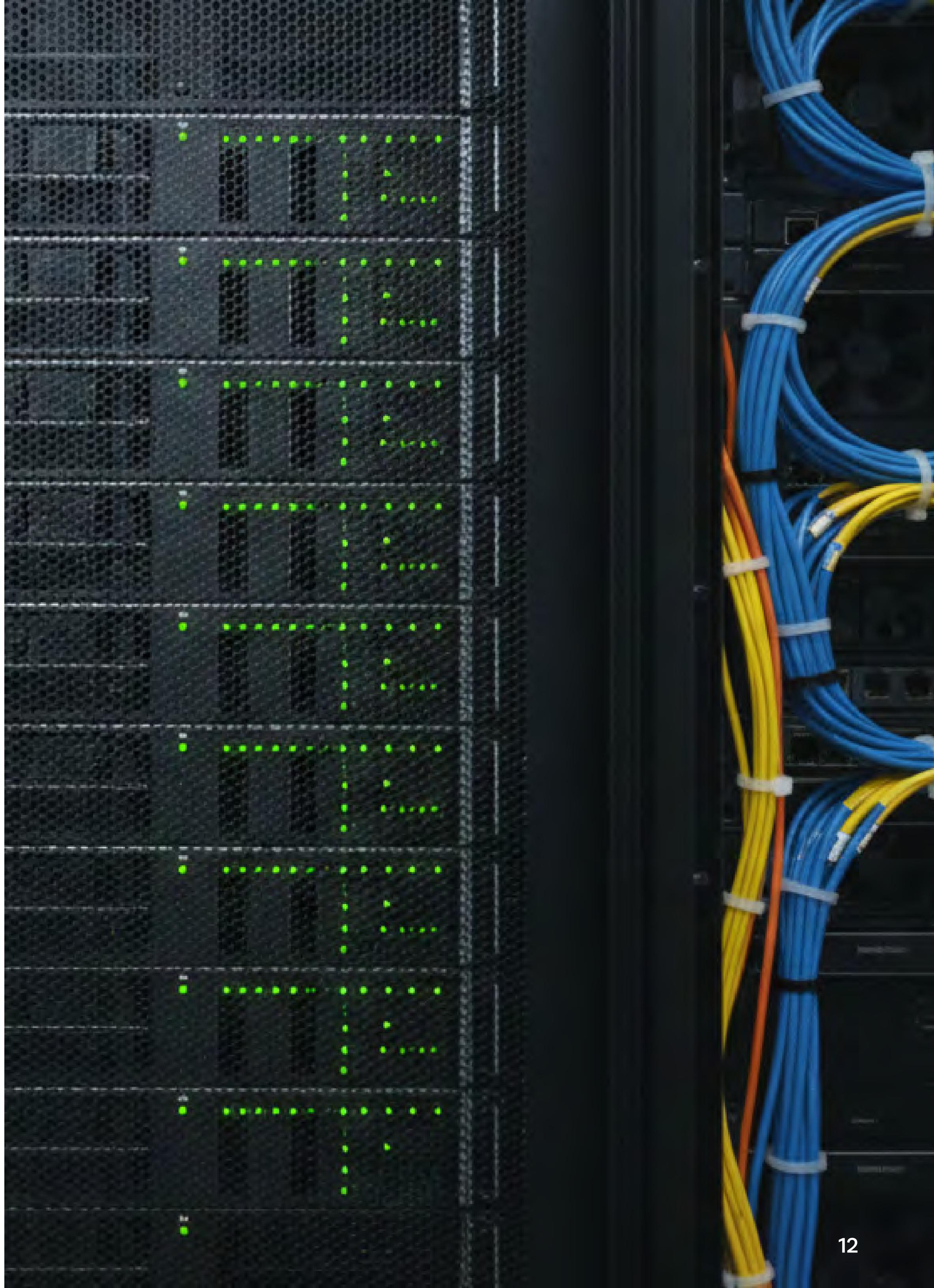
6. Validating Model Predictions Through Testing

Modeling provides valuable insight into how power systems are expected to behave under dynamic AI load conditions. However, simulation alone is not sufficient. Models must be validated against measured system performance to ensure predictions accurately reflect real-world operating behavior.

Testing remains a critical part of the validation process. Traditional methodologies, including load acceptance and rejection testing, continue to provide important performance benchmarks for generators and supporting infrastructure. These tests help verify system response, stability, and recovery characteristics under controlled conditions.

As AI workloads introduce more dynamic operating profiles, validation increasingly requires testing approaches that extend beyond traditional steady-state assumptions. Rapid load changes, fluctuating demand profiles, and longer periods of variable operation can help provide a more complete understanding of system behavior and performance margins.

By comparing modeled predictions against measured results, infrastructure teams can refine assumptions, improve model accuracy, and build greater confidence in system performance before deployment. This process also helps identify potential control, stability, or operating issues earlier in the design cycle, reducing risk and improving overall system readiness.



7. Islanded AI Facilities: Prime Power Systems

In islanded AI facilities, on-site generation does not function as backup power. It functions as the primary source of power.

These environments emerge when utility capacity is unavailable, delayed, or insufficient to support large-scale AI deployment. Rather than waiting for full grid access, operators increasingly rely on natural gas generation, grid-forming battery systems, and integrated controls to create self-sustaining power architectures capable of supporting continuous operation.

Unlike standby generators that operate intermittently, prime power systems remain exposed to AI-driven load variability at all times. As a result, system stability depends not only on generation capacity, but on the coordinated behavior of generators, energy storage, power conversion systems, and load management across the entire architecture.

Battery systems play a critical role by absorbing fast transients, supporting frequency and voltage stability, providing active and reactive power response, and reducing stress on generation assets. However, long-term reliability still depends on whether the integrated system can maintain stable performance under sustained dynamic load conditions.

In this context, the power plant is not simply serving a predictable facility load. It is supporting a variable compute environment where load steps, load drops, and reactive power behavior may occur repeatedly as part of normal operation. As AI infrastructure scales, readiness is no longer defined by power availability alone. It is defined by how effectively the integrated system performs under real operating conditions.

7.1 Prime Power Changes the Test Requirement (Islanded)

In many AI facilities, particularly grid-constrained or islanded environments, natural gas generation may operate as the primary source of power rather than a standby asset. This fundamentally changes the operating profile of critical power infrastructure.

Traditional backup systems are designed around short-duration emergency events. Prime power systems operate continuously under variable load conditions as part of normal operation. As a result, testing methodologies designed around isolated load acceptance events are no longer sufficient on their own.

A conventional resistive load-bank test may demonstrate that a generator or power block can accept load, but it does not fully validate how a coordinated prime power system will respond to fast AI-driven transients, voltage and reactive power (VAR) requirements, load rejection events, or repeated variability over time.

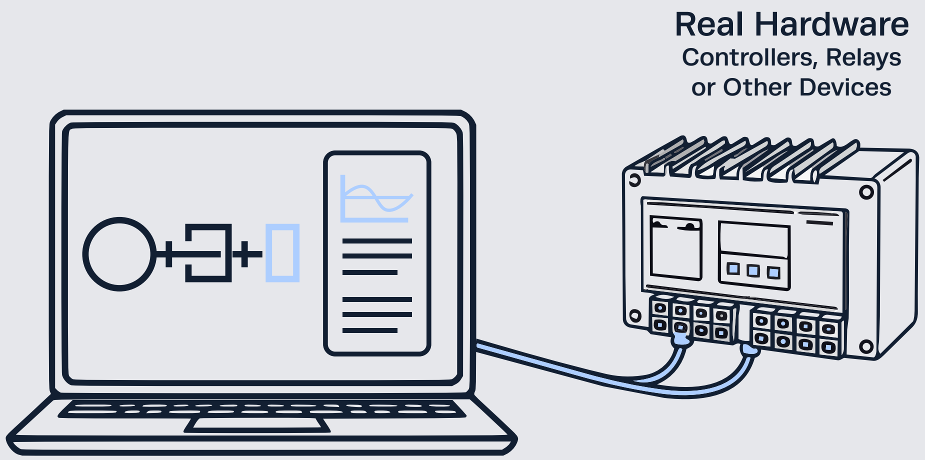
For AI prime power systems, the key question is no longer whether enough megawatts are installed. The key question is whether the complete power architecture can deliver stable, repeatable performance under dynamic operating conditions.

7.2 Building Confidence Before Commissioning with HIL Testing

Hardware-in-the-loop, or HIL, testing plays an important role before equipment is installed on site. It can be used during project feasibility, controls development, and integration planning to confirm that the intended architecture, control logic, communication pathways, and major equipment selections are capable of interoperating as designed. In many cases, a digital twin — a validated software model of the microgrid or power system — is used within the HIL environment to simulate plant behavior while real controllers, relays, or other hardware are tested against it.

HIL testing is valuable because it helps identify control conflicts, communication issues, equipment compatibility concerns, and sequence-of-operation gaps early in the project lifecycle. However, HIL is a precursor to site validation. It does not replace commissioning once physical assets are installed.

In simple terms, HIL testing helps answer whether the system architecture should work. On-site validation answers whether the installed system actually does work under real electrical conditions.



7.3 Asset-to-Fleet Validation

For islanded AI facilities, commissioning should follow an **Asset-to-Fleet Validatio**n approach.

Asset-to-Fleet Validation is a staged testing process that first verifies each controllable asset independently, then progressively validates the coordinated response of the full power plant.

This approach recognizes that a prime power system serving AI load is not a collection of independent assets operating in parallel. It is a coordinated fleet of generation, storage, conversion, and control systems that must respond together in real time.

The first step is to test and characterize each major asset block independently. This includes MBMS controllers, BMS/DC blocks, inverters, reciprocating engine controls, and the reciprocating master control unit. Each asset should be validated for local control behavior, communications, protection settings, response time, ramp capability, voltage regulation, and active/reactive power performance.

For BESS and inverter assets, testing should include load banks capable of both real and reactive power behavior rather than purely resistive load. AI facilities do not only create active power challenges. They also create voltage, power factor, and reactive power coordination requirements that may not be visible during purely resistive testing.

Once each individual asset block has been validated, testing should progress to subsystem-level operation. Each MBMS and associated BMS/DC block should be tested as a functional BESS block. Each inverter should be tested with its associated battery system and local controls. Engine units should be tested individually and then as a coordinated block.

The final step is fleet-level validation under the PMS/PPC. At this stage, the objective is no longer simply to prove that individual equipment can accept load. The objective is to prove that the full power architecture can respond to AI-style load steps, load drops, reactive power swings, and sustained variability while maintaining stable voltage, frequency, and coordinated dispatch across BESS and engine assets.

This is especially important where a plant-level power management system is dispatching a distributed fleet. The PMS/PPC must relay voltage, VAR, real power, and operating setpoints across multiple asset types. If one asset responds too slowly, overreacts, reaches a protection limit, or behaves differently than modeled, the fleet-level response can become unstable even if each individual asset appears acceptable in isolation.

For this reason, commissioning should not jump directly from individual equipment checks to full plant operation. Each layer of the dispatch chain must be proven before the next layer is tested.

7.4 Implications for AI Power Plant Design

As AI workloads reshape power behavior, testing and validation increasingly become inputs to infrastructure design rather than final-stage compliance exercises.

Results from dynamic load testing should directly inform decisions around generator sizing, BESS capacity, inverter capability, reactive power requirements, control tuning, operating envelopes, and overall system architecture.

In prime-power AI environments, generators, BESS, UPS systems, inverters, controls, and load behavior continuously interact as part of a single operating system. Long-term reliability therefore depends not only on individual asset performance, but on how effectively the integrated power architecture performs under sustained operating conditions.

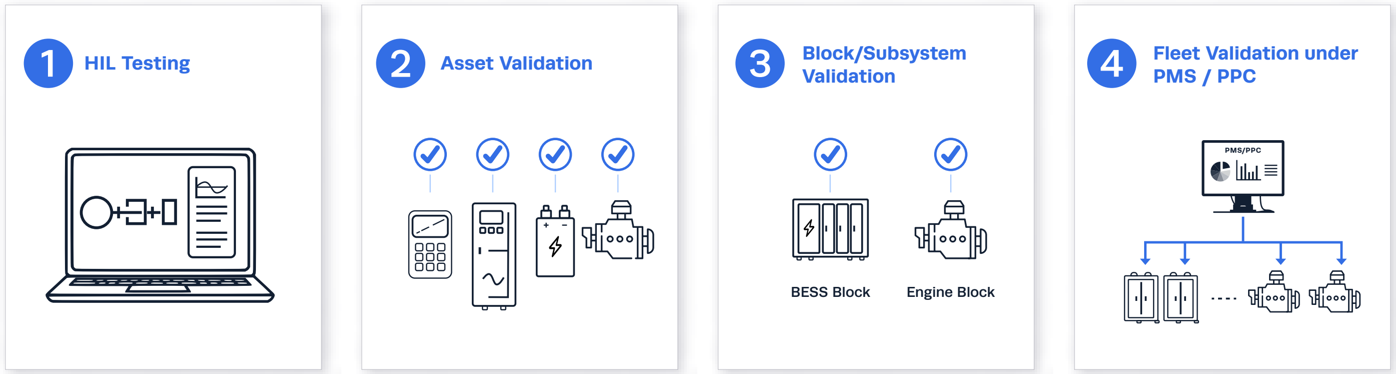
For AI infrastructure, readiness is no longer defined by installed megawatts or backup availability alone. It is defined by whether the power system can demonstrate stable, repeatable performance under real-world load conditions.

HIL testing helps establish confidence before equipment is installed. Asset-to-Fleet Validation proves the system after installation. Together, they create a practical validation pathway for islanded AI facilities where power quality, dispatch coordination, and sustained dynamic performance are critical to long-term reliability.

* Asset validation and block/subsystem validation are typically accomplished through Factory Acceptance Testing (FAT) and project commissioning, respectively, and are discussed here as an integrated process flow from control validation to fleet-wide coordinated commissioning and dispatch.

From HIL to Asset-to-Fleet Validation

Simple Commissioning Pathway for Islanded AI Power Systems



8. Engineering AI Infrastructure with Confidence

Historically, power systems were designed around relatively predictable operating conditions, established redundancy models, and individual equipment performance. AI changes that equation. As workloads become increasingly dynamic, infrastructure teams must evaluate how generators, UPS systems, energy storage, controls, and cooling systems perform together under sustained operating conditions rather than as isolated components.

This shift places greater emphasis on testing, modeling, validation, and digital engineering earlier in the design process. By combining validated component models into system-level digital twins, operators can better understand power system behavior, optimize design decisions, and reduce uncertainty before infrastructure is deployed. Rather than relying on conservative oversizing, future AI facilities will increasingly depend on accurately modeling and validating complete power architectures.

As AI deployments continue to scale, infrastructure readiness will no longer be defined by installed capacity alone. It will be measured by the ability to deliver stable, predictable performance under real AI load conditions. Rehlko is helping advance this next generation of infrastructure design by combining generator testing, high-fidelity modeling, and system-level collaboration to support more resilient, efficient, and adaptable digital infrastructure.