From design intent to measured performance

The Vertiv™ ATLAS (Advanced Test Lab for Architecture and Systems)

AI factories represent a class of infrastructure investment without a direct or modern precedent. Billions in capital committed against a single operational premise: that electrical capacity can be converted into productive compute quickly, reliably, and at a cost structure that justifies the commitment.

The economics can be unforgiving: Every megawatt that reaches a facility but fails to produce useful AI output represents stranded capital due to thermal constraints, power delivery limitations, controls misalignment, or commissioning delays, among other factors. At a multi-megawatt scale, the financial exposure of even modest delays or inefficiencies compounds rapidly, which could amount to millions of dollars in deferred productive capacity per week.

This is emerging and turning from a procurement challenge to a systems engineering challenge. The physical infrastructure supporting AI compute must be designed, validated, and operated as an interdependent system instead of a collection of individually specified components assembled in the field and hoped to behave as intended.

The question facing AI factory operators is straightforward: how to turn more of every available megawatt into productive AI capacity, sooner and with greater confidence?

Five forces reshaping AI factory infrastructure

Five intensifying forces are converging, rendering traditional approaches to infrastructure design and deployment insufficient for AI-scale operations. None of these forces operate in isolation, and the compounding effect is what makes a systems-level response necessary.

Density

From 50 kW, 100 kW, to 150 kW per rack and beyond, rack power is climbing rapidly and forcing tighter coupling between compute, power delivery, liquid cooling, heat rejection, and serviceability. At these densities, the margin between stable operations and thermal exceedance narrows considerably. Every interface between power and cooling becomes a potential failure point if not validated as a coupled system. The physics of heat removal at 100+ kW per rack are fundamentally different from those at 10–20 kW, and the infrastructure must be validated accordingly.

Speed

The competitive landscape for AI capacity means success is measured not by how much infrastructure is built, but by how quickly that infrastructure begins producing useful AI output. Time to productive capacity[1] has become the defining metric. Weeks of commissioning delays translate directly into lost competitive position and deferred revenue. In a market where AI capability compounds, late infrastructure is a strategic disadvantage.

[1] *Time to productive capacity is the interval between capital commitment and first productive workload.*

Scale

AI factories are expanding from room-scale and pod-scale deployments to multi-megawatt and gigawatt-scale programs. At this scale, infrastructure interactions that were manageable in smaller deployments become dominant engineering challenges. A thermal imbalance affecting one rack at the pod scale can cascade across an entire hall at the factory scale. Power distribution architectures that performed adequately at 5 MW face fundamentally different challenges at 50 MW or 500 MW. The number of interdependent variables grows non-linearly, and the consequences of unvalidated assumptions grow with them.

Complexity

Power, thermal, controls, software, construction sequencing, commissioning, and maintenance services now interact in ways that traditional field-integrated models were never designed to absorb. The number of system interfaces grows non-linearly with scale, and each interface represents a potential point of divergence between design intent and physical reality. Controls logic must coordinate across subsystems that were historically specified, procured, and commissioned independently. The integration challenge has moved from the equipment level to the system level.

Dynamic workload profiles

AI infrastructure does not behave like traditional steady-state IT loads:

- Training workloads produce sustained high-power draws with periodic spikes during checkpoint operations.
- Inference workloads create rapid, unpredictable load transients as request volumes fluctuate. Failover events redistribute thermal and electrical load across the facility in milliseconds.
- Mixed workloads (i.e., training and inference running simultaneously across different zones) create compound dynamic profiles that no single steady-state specification can characterize. Load profiles, thermal response characteristics, failover behavior, and control loop timing must be understood together.

These five forces are compounding. A facility operating at high density and scale, with complex system interactions and dynamic workloads, must reach productive capacity faster than ever. The infrastructure validation approach must account for all five simultaneously, because the physical system will experience them all.

Beyond component selection: A systems-level imperative

The conversation between infrastructure providers and AI factory operators has shifted. Operators have refrained from asking which uninterruptible power supply (UPS) to include, which coolant distribution unit (CDU) to select, or which busway to route. They are asking how the entire physical system behaves when all components operate simultaneously under realistic AI workload conditions.

Individual component performance[2] remains necessary, but is no longer sufficient. A UPS that meets its specification in isolation may interact with a busway and power distribution unit (PDU) configuration in ways that create voltage transients under dynamic AI load profiles. A CDU operating at rated capacity may encounter flow distribution challenges when integrated with a specific rack-level liquid cooling topology at density. Controls logic that functions correctly in simulation may exhibit timing conflicts when orchestrating real power and thermal systems simultaneously.

The gaps that matter most in AI factory infrastructure exist at the interfaces and seams between systems. Addressing this reality requires a disciplined, systems-level approach built on five principles:

1. **Repeatable building blocks:** Configurable infrastructure modules that reduce reinvention and enable validated configurations to be deployed consistently across multiple sites and phases.
2. **Defined interfaces:** Mechanical, electrical, controls, software, and service boundaries engineered intentionally.
3. **System orchestration:** Power, thermal, and controls operating as a coordinated system rather than as isolated subsystems managed independently.
4. **Digital continuity:** Design intent preserved from engineering through deployment to operation; what was modeled, what was validated, and what is operating remain aligned throughout the lifecycle.
5. **Lifecycle assurance:** Day One treated as the beginning of a monitored, governed, and optimized operational lifecycle instead of at the end of a delivery project.

[2] Individual component performance is verified through factory acceptance testing, type testing, and certification.

The closed-loop validation cycle

Reducing the variance between design intent and physical reality is not a one-time event. It is a closed loop that connects virtual simulation, physical validation, field deployment, and operational learning into a single, iterative process.

Each stage feeds the next. Each iteration starts from a higher baseline of confidence. The loop never fully closes; it spirals forward, with each cycle reducing uncertainty and improving system performance.

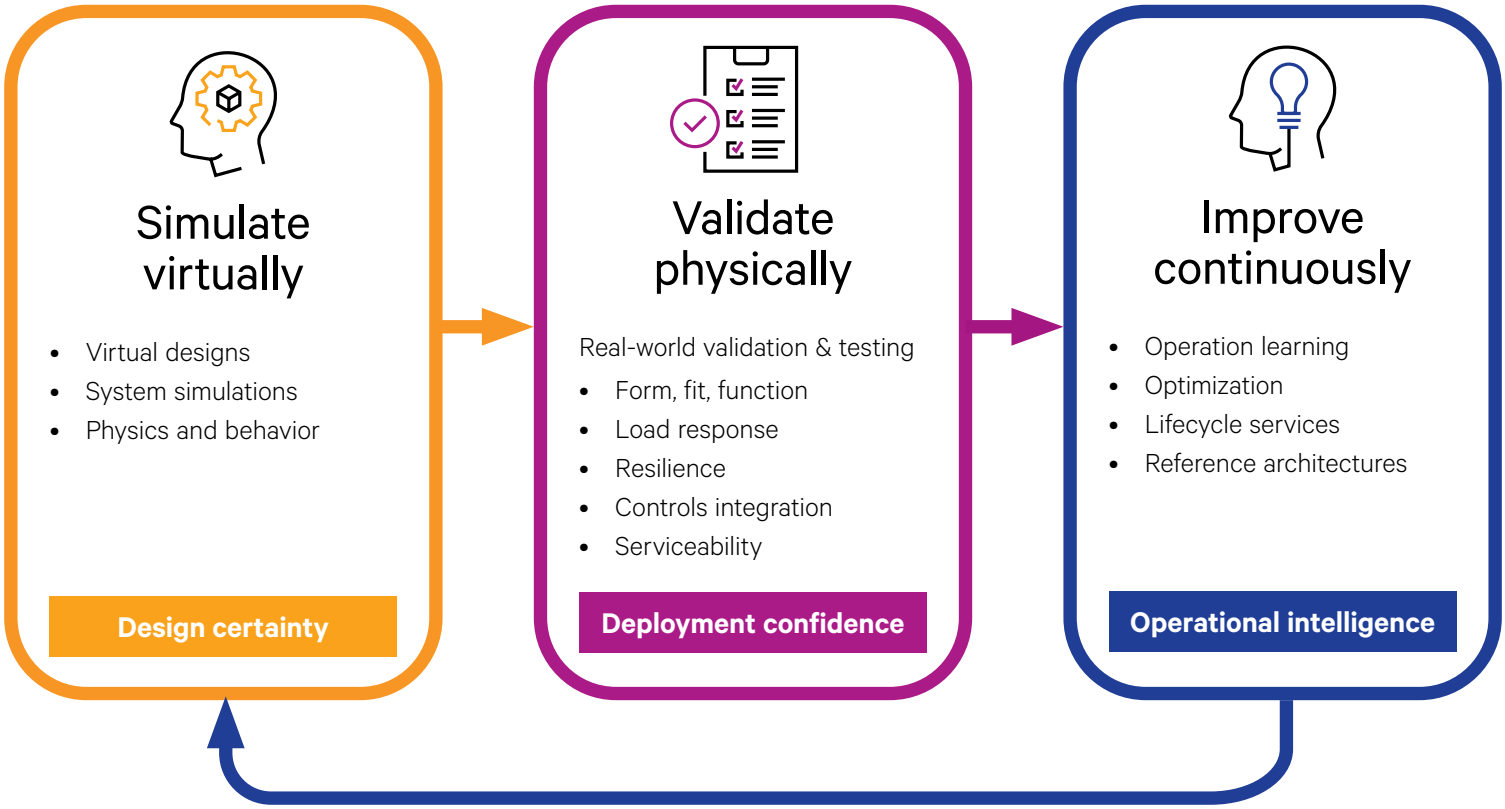


Figure 1. In designing and implementing the modern AI factory at scale, the goal is to continuously reduce the variance between design intent and physical reality through continued simulation and physical validation. Source: Vertiv.

Stage 1: Simulate virtually

Outcome: Design certainty

The cycle begins with model-based systems engineering, virtual representations of complete AI factory infrastructure that incorporate geometry, load profiles, thermal boundaries, fluid dynamics, and control logic. Physics simulation layers real-world behavior onto the model: how heat moves through a liquid cooling loop under transient load, how voltage propagates through a power distribution architecture during failover, and how control loops respond to simultaneous thermal and electrical events.

The result is a working virtual system that behaves like the physical infrastructure it represents. Within the fidelity of the model, design decisions can be tested, alternatives compared, and failure modes explored before any physical hardware is committed. The economic value is clear: errors caught in simulation cost less than errors discovered during commissioning.

Stage 2: Validate and calibrate physically

Outcome: Deployment confidence

Virtual models carry assumptions regardless of sophistication. Material properties, manufacturing tolerances, installation variables, and real-world environmental conditions introduce variance that simulation alone cannot fully capture. Stage 2 is where design intent meets physical reality.

Full systems-level testing stress-tests the complete infrastructure configuration.[3] Power trains and thermal chains operate simultaneously. Controls logic executes against real electrical and thermal dynamics. Interfaces between systems are exercised under realistic conditions: steady-state operation, dynamic load transients, failover scenarios, and recovery sequences.

This is the critical validation node at which the inherited risk is identified and resolved before capital is committed at scale and before the operator assumes ownership of a system with behavior that has only been predicted, not proven.

What physical validation reveals:

- Form, fit, and function confirmation across integrated systems.
- Load response characteristics under dynamic AI workload profiles.
- Resilience and recovery behavior during fault conditions.
- Controls integration timing and coordination across power and thermal subsystems.
- Serviceability, access, and maintenance sequencing in dense configurations.

Physical validation generates:

1. Deployment intelligence: Real performance data that informs commissioning sequences, control loop tuning, and operational baselines.
2. Model calibration data: Measured physical behavior that feeds back into simulation tools, improving their accuracy for subsequent designs.
3. Interface validation: Confirmation that the seams between systems perform as designed, not just validating that the equipment on either side functions.

The distinction is important: factory acceptance testing validates individual equipment against its specification. Systems-level physical validation tests what happens between the equipment; at the interfaces and seams where design assumptions are most likely to diverge from physical reality.

Stage 3: Improve continuously

Outcome: Operational intelligence

Deployment is not the end of the cycle. Operational data from fielded systems — including thermal performance trends, power utilization patterns, control loop behavior under real workloads, and maintenance event histories — feeds back into the design and simulation environment. Reference architectures are refined. Simulation models are recalibrated against field reality. The next design iteration begins from a foundation of operational truth rather than a theoretical assumption.

This closed loop, from the model to the field and back, is the mechanism that continuously reduces the distance between what was designed and what is delivered. It transforms infrastructure deployment from a series of independent projects into a compounding body of validated knowledge.

[3] Full systems-level testing uses load emulation to replicate AI server behavior without the cost or degradation of actual compute hardware.

Consolidating AI factory validation at system scale

Vertiv has tested and validated critical infrastructure globally for decades across power systems, thermal management, controls architectures, and integrated solutions. Engineering centers, test facilities, and service organizations worldwide have continuously developed and refined validation methodologies for mission-critical environments.

The Vertiv™ Advanced Test Lab for Architecture and Systems (ATLAS) consolidates and connects a much broader set of those capabilities under one roof, designed specifically for the scale, complexity, and dynamic behavior of AI factory infrastructure. It validates how complete power trains, thermal chains, controls, software, and compute-aligned infrastructure behave together as an interdependent system at greater scale, speed, and fidelity. This is a consolidation and maturation in the industry with decades of distributed validation capability focused and scaled to meet the specific demands of AI factory infrastructure.

What the lab unlocks

The outcomes are measured in operational terms:

- **Faster time to productive capacity:** Validated systems commission faster because interface issues are resolved before they reach the field.
- **More productive use of available MW:** System-level optimization identifies and eliminates sources of stranded capacity.
- **Reduced stranded power and cooling:** Thermal and power architectures validated together prevent oversizing driven by uncertainty.
- **Lower integration and commissioning risk:** Interfaces proven in the lab do not become surprises during deployment.
- **Greater confidence across compute generations:** Validated reference architectures help accelerate adaptation to new compute hardware without full re-validation.
- **Better AI factory economics:** The cumulative effect of reduced uncertainty across every phase of the lifecycle.

What is new

The distinction between the ATLAS Lab and traditional testing approaches lies in focus. Factory acceptance testing validates individual equipment against its specification. The ATLAS Lab validates what happens at the interfaces and seams where systems interact, specifically:

- End-to-end system visibility across power, thermal, controls, and software simultaneously.
- Validation of interfaces and seams between subsystems, the points where design assumptions are most likely to diverge from physical reality.
- AI-relevant dynamic workload emulation at density and scale.
- Combined power, thermal, and control behavior under simultaneous operation.
- Model-to-physical calibration measured data feeding back into simulation environments.
- Failure injection and recovery scenario testing.
- Serviceability and commissioning sequence validation in integrated configurations.
- Closed-loop learning that feeds validation results into future reference architectures.

The question worth asking any infrastructure provider: do you have a validation environment specifically designed to test AI factory systems at this level of integration?

The Vertiv™ ATLAS: Physical evidence of measured results

It is the capability within Vertiv where a complete integrated system is tested and validated end-to-end as an interdependent whole. Power train, thermal chain, controls, software, and compute-aligned infrastructure are designed, simulated, and tested from the component level to the system in real world scenarios.

Its purpose is precise: to improve design and validate AI factory systems, reduce the uncertainty between design intent and physical performance, and contribute to better economics and higher productive capacity for operators.

The lab uses row-level bulk load testing in 2026, adding the full dynamic AI server workload emulation to replicate real workload behavior in early 2027. These emulation capabilities allow systems-level tuning and stress testing under dynamic workload conditions at a fidelity that static testing cannot achieve; from training profiles and inference transients, to mixed workloads and failover events in environments testing compute generations that are still in development.

Vertiv™ ATLAS roadmap:

Planning & construction
2026

Row-level bulk load capability
2026

Dynamic AI workload emulation
Early 2027

What distinguishes the lab from conventional test environments is the questions it is designed to answer:

- Does this system behave as designed when every subsystem operates simultaneously under realistic AI factory conditions?
- Does the power train respond correctly when the thermal system shifts operating mode?
- Do controls coordinate across subsystems within acceptable timing windows during a failover event?
- Does the interface between the CDU and rack-level liquid cooling perform as modeled at full density?

These are questions that can only be answered physically at system scale, under dynamic conditions. They are also the questions whose answers determine whether a deployment commissions in weeks or months, and whether the operator inherits a validated system or an unproven assumption.

The validation data generated in this environment feeds forward into deployment intelligence, informing commissioning sequences, control loop parameters, and operational baselines for fielded systems. This feeds back into simulation environments to calibrate models against measured physical reality so that the next design iteration starts from a higher floor of accuracy.

This is the mechanism by which the closed loop operates in practice: the lab is not an endpoint, but a capacity in a continuous cycle of design, validation, deployment, and improvement.

AI factory economics: The metrics that matter

The relationship between infrastructure validation and AI factory economics is indirect but consequential. Rigorous system-level validation improves the ability to design and deploy infrastructure that performs as intended. That reduces uncertainty and eliminates sources of stranded capacity. That contributes to greater productive capacity and better capital economics.

Metric	What it measures
Time to productive capacity	How quickly infrastructure begins supporting productive AI workloads.
Productive MW	How much contracted or available power actually reaches productive compute.
AI output per MW	The relationship between infrastructure performance and AI factory output.
Capital productivity	Reduce stranded infrastructure, rework, commissioning delays, and unnecessary capacity margins.

Table 1. A summary of the metrics for measuring key operational indicators (KPIs) at scale. Source: Vertiv.

Four metrics frame this relationship:

- **Time to productive capacity:**
  The interval between facility energization and first productive AI workload. Every source of commissioning delay extends this interval, such as interface conflicts, control timing issues, or thermal imbalances discovered in the field. Validation that resolves these issues before deployment compresses the timeline directly. At scale, compressing this interval by even weeks represents significant economic value.
- **Productive MW:**
  The proportion of contracted or available electrical capacity that actually reaches productive compute. Stranded power[4] reduces this proportion. System-level validation identifies and eliminates sources of waste before they become embedded in the deployed system.
- **AI output per MW:**
  The relationship between infrastructure performance and useful AI factory output. When power and thermal systems operate as a validated, optimized system, more of every megawatt supports productive computation rather than compensating for infrastructure uncertainty or absorbing inefficiency from unvalidated interfaces.
- **Capital productivity:**
  The aggregate economic effect of reducing stranded infrastructure, eliminating rework, compressing commissioning timelines, and removing unnecessary capacity margins. Infrastructure validated at the system level before deployment carries less embedded waste than infrastructure assembled and debugged in the field. Over multi-megawatt programs spanning multiple phases, the cumulative economic impact is substantial.

These metrics are interconnected. Faster time to productive capacity improves capital productivity. Higher productive MW improves AI output per MW. Reduced stranded capacity improves all four simultaneously. The common thread: system-level validation reduces the uncertainty that degrades each metric.

[4] Stranded power is the capacity consumed by infrastructure inefficiency, thermal overhead from unvalidated configurations, or margins added to compensate for uncertainty.

Decision framework: Assessing AI factory designs with measured performance

For operators and decision-makers evaluating their infrastructure's overall posture, the following questions map to each phase of the closed-loop cycle:

Design and simulate

- Has the complete infrastructure system been modeled as an interdependent whole or just at the equipment/component-level?
- Do simulation models account for dynamic AI workload profiles or only steady-state conditions?
- Are interfaces between power, thermal, and controls architectures explicitly defined and simulated?
- Is there a clear methodology for translating simulation results into physical validation requirements?

Validate and calibrate

- Has the integrated system been physically validated under realistic operating conditions before deployment commitment?
- Have interfaces and seams between subsystems been tested?
- Has the system been exercised under failure and recovery scenarios?
- Is there calibration data flowing from physical testing back into simulation models?
- Are dynamic workload profiles part of the validation protocol?

Deploy and operate

- Are commissioning sequences informed by validation data or based on generic procedures?
- Is there visibility across the complete system during operation: power, thermal, controls, and software simultaneously?
- Are operational baselines established from validated performance data rather than theoretical projections?
- Is the deployment leveraging reference architectures refined through prior validation cycles?

Learn and improve

- Does operational data feed back into design and simulation environments?
- Are reference architectures updated based on field performance?
- Is the next deployment starting from a higher confidence baseline than the last?
- Is the organization capturing and applying lessons across sites, phases, and compute generations?

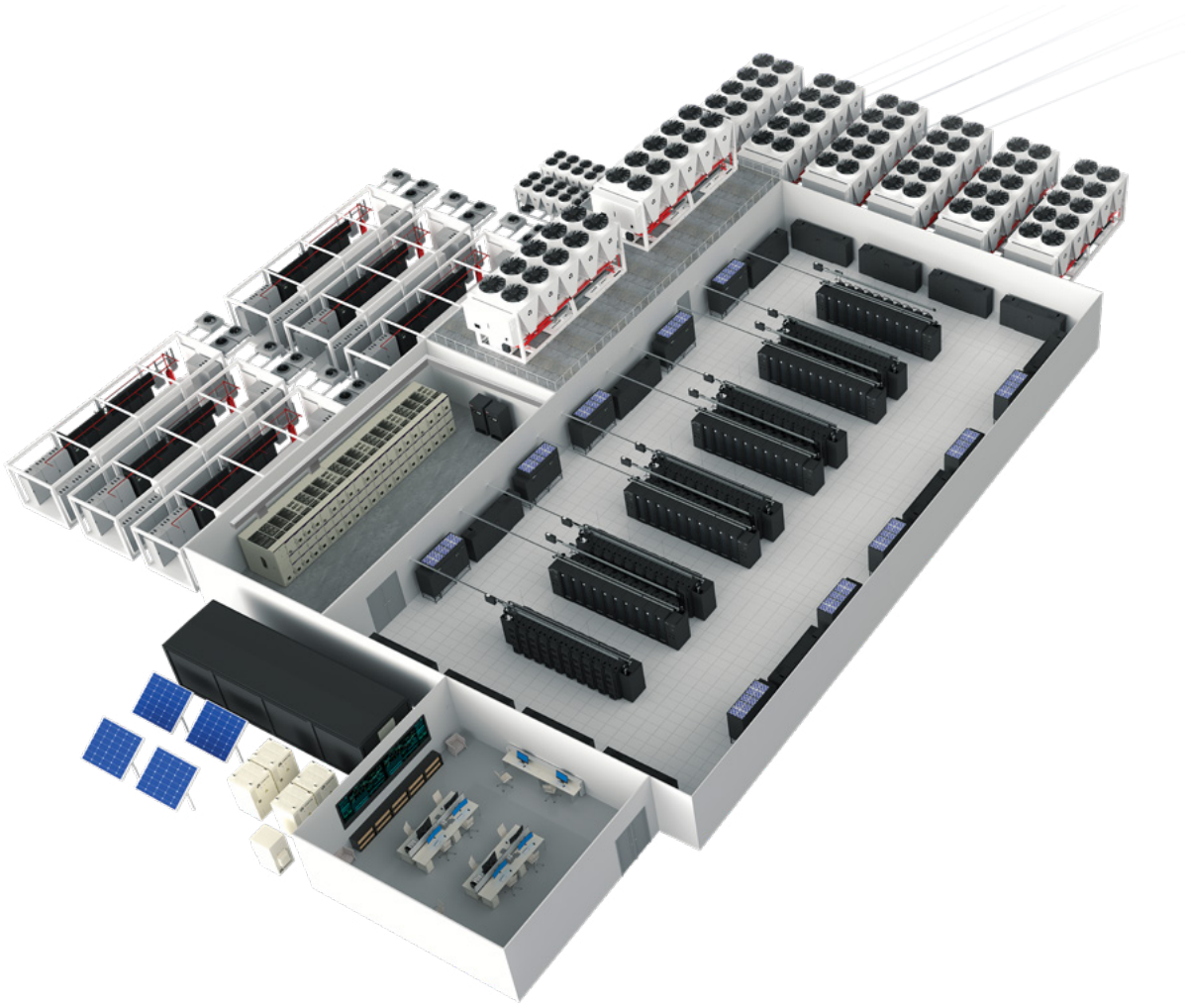
The answers to these questions can vary and do not necessarily point to right or wrong answers. These simply define the infrastructure's distance between design intent and operational confidence, and enable operators and organizations to measure and evaluate their own position. Closing that distance systematically, continuously, and at system scale is the engineering challenge that determines whether AI factory capital produces productive compute or stranded infrastructure.

Closing the loop

AI factory infrastructure operates as a system. Power trains, thermal chains, control architectures, software layers, and the interfaces between them either perform as a validated, convergent whole or perform as a collection of assumptions awaiting field confirmation.

Vertiv has tested and validated critical infrastructure globally for decades. The Vertiv™ ATLAS consolidates and connects a much broader set of those capabilities to validate complete AI factory infrastructure as an interdependent system; at a scale, speed, and fidelity designed to meet the demands of the evolving AI factory infrastructure than any single facility or test program has previously addressed.

The closed loop from model to field is the mechanism that continuously reduces the gap between theoretical design and operational reality. Each cycle compresses the distance between what was designed and what is delivered, each deployment informs the next, and each validation sharpens the one that follows. The infrastructure either performs as a validated system, or it performs as an assumption. AI factory economics leave increasingly little room for the latter.



SL-80657 (R09/26)