Special Analysis

# Department of Energy Announces Nine New Supercomputers

Jaclyn Ludema and Mark Nossokoff
December 2025

## HYPERION RESEARCH OPINION

Recent announcements from the U.S. Department of Energy (DOE) and several of its national laboratories point to a structural shift in how leadership-class computing is delivered. These changes align with the DOE's new Genesis Mission, which aims to create an integrated platform linking the world's most powerful supercomputers, experimental facilities, AI systems, and domain-specific datasets, intended to double the productivity and impact of US government research within a decade. In this context, the systems now planned at Oak Ridge National Laboratory (ORNL), Argonne National Laboratory (ANL), and Los Alamos National Laboratory (LANL) can be seen as early steps toward the Genesis vision.

The announced plans for nine new supercomputers include both traditional leadership-class scientific and engineering systems and a new class of AI-first "factories". The latter are designed not as stand-alone supercomputers, but as tightly integrated, AI-centric environments. Two features of these new AI-first factories stand out. First, DOE is moving beyond its traditional “own, operate, and fully control” model, transferring the majority of system operation responsibility to a commercial vendor. For example, US-based Oracle will operate several of the most advanced systems on government property, using Oracle’s cloud software stack, with DOE receiving only a defined share of capacity rather than exclusive use. Second, these environments are being designed from the ground up around AI and agentic workflows, with conventional FP64-oriented HPC positioned as only one component of a broader AI-centric ecosystem.

From a market standpoint, these supercomputer announcements suggest that the concept of HPC-centric continuum computing (computing seamlessly distributed across edge, cloud, and on-premises infrastructures), which has been discussed for years, is now reaching a national-infrastructure phase. The attributes of the nine supercomputers made public thus far begin to outline the shape of national-scale AI infrastructure over the next decade: a mixed ecosystem that preserves some traditional supercomputing capabilities while introducing public-private AI factories that blur the line between hyperscale cloud and sovereign HPC, accompanied by governance, economic, and technical models that diverge from historical supercomputing standards.

## SITUATION ANALYSIS

ORNL, LANL, ANL, and DOE partners are building a portfolio of AI-centric systems spanning leadership-class complexes and smaller workforce-development platforms. While project timelines and use cases vary or have yet to be released in detail, they share a common orientation toward large-scale AI training, advanced reasoning models, and tightly coupled data and experimental facilities.

**ORNL, LANL, ANL, and DOE partners are building a portfolio of AI-centric systems… seen together as potential building blocks to meet the Genesis Mission objectives.**

### ORNL: Lux and Discovery

At Oak Ridge, Lux is positioned as an AI factory to expand near-term capability for fusion and fission research, materials discovery, quantum, manufacturing, grid science, and national security applications. It is structured as a public-private partnership with co-investment and will be operated by Oracle Cloud Infrastructure (OCI) using Oracle's software stack on government property. Discovery is framed as Frontier's successor and a leadership-class supercomputer intended for converged HPC, AI, and quantum workflows. It follows a more traditional DOE procurement path on the hardware side while still being expected to run under an OCI-operated software environment, emphasizing large scientific AI models and shortened discovery cycles.

While Discovery is not strictly positioned as an AI factory in the same way as Lux, it is an important counterpoint in DOE's evolving portfolio. Architecturally, Discovery follows a more traditional leadership-class HPC design with strong FP64 performance but layers in high-end GPUs and a "bandwidth everywhere" approach to support converged HPC/AI workloads.

#### *ORNL architecture details:*

**Lux**

- HPE ProLiant Compute XD685
- AMD Instinct MI355X GPUs
- AMD EPYC CPUs
- AMD Pensando networking

**Discovery**

- HPE Cray Supercomputing GX5000
- Next-generation AMD EPYC "Venice" CPUs
- AMD Instinct MI430X GPUs
- "Bandwidth Everywhere" design

### LANL: Mission and Vision

At Los Alamos, Mission and Vision are being deployed under the National Nuclear Security Administration (NNSA) to support both classified and unclassified national security workloads. Mission will serve as a stockpile stewardship platform for high-fidelity simulations central to nuclear weapons physics, while Vision targets broader national security science in an unclassified environment.

### *LANL architecture details:*

**Mission**

- HPE Cray Supercomputing GX5000
- NVIDIA Vera Rubin platform
- QuantumX-800 IB

**Vision**

- HPE Cray Supercomputing GX5000
- NVIDIA Vera Rubin platform
- QuantumX-800 IB

## ANL: Solstice, Equinox, Tara, Minerva, and Janus

Argonne's roadmap centers on Solstice and Equinox, positioned as DOE's largest AI supercomputing complex and as a key part of Argonne's AI factory strategy. Solstice and Equinox are being jointly developed with NVIDIA and Oracle: Solstice designed to host large-scale training and inference of frontier AI and reasoning models for open science, and Equinox is intended to operate closely with Argonne's experimental and data facilities, focusing on large-model training, scaled inference, and agentic AI workflows that can drive experiment design, control, and analysis. Supporting these flagship systems, Tara, Minerva, and Janus will serve as smaller NVIDIA-based resources aimed at broader AI access and workforce development.

### *ANL architecture details*

**Solstice**

- Joint DOE/Argonne–NVIDIA–Oracle deployment
- 100,000 NVIDIA Blackwell GPUs
- Part of a complex delivering ~2,200 exaflops of AI-precision compute (combined with Equinox)

**Equinox**

- DOE–NVIDIA–Oracle partnership
- 10,000 NVIDIA Blackwell GPUs
- Integrated with NVIDIA Megatron-Core (training) and NVIDIA TensorRT (inference)

**Tara, Minerva, Janus**

- Smaller NVIDIA GPU-based systems for AI prediction, modeling, and training
- Detailed node-level configurations not yet public

## Strategic Alignment to the Genesis Mission

The architectural designs of these nine systems suggest a deliberate alignment with the DOE's stated Genesis Mission, which seeks to establish a "national discovery platform" integrating supercomputing, AI, and quantum technologies. Rather than viewing these announcements as isolated upgrades, they can be seen together as potential building blocks to meet the Genesis Mission objectives of creating " an intelligent network capable of sensing, simulating, and understanding nature at every scale".

- Experimental Integration: The placement of inference-optimized resources in close proximity to experimental facilities appears targeted at the "sensing" objective

- Simulation and Fidelity: A subset of the planned infrastructure retains the strong double-precision (FP64) performance characteristics essential for traditional modeling. These resources will likely target the "simulating" component of the mission, generating high-fidelity data necessary for model validation and training
- AI-Centric Compute: The emergence of systems designated as "AI factories" suggests a prioritization of the "understanding" component

Additionally, the adoption of the commercial cloud operating models across multiple sites may represent a strategic attempt to address the interoperability challenges that have historically siloed HPC centers.

## A New Operational Paradigm

These technical efforts are unfolding within a broader policy context that emphasizes accelerating AI infrastructure while navigating regulatory and security constraints. Recent executive actions have focused on easing bottlenecks around data center deployment, while legislation such as the One Big Beautiful Bill Act (P.L. 119-21) allocates dedicated $115 million in funds for NNSA to accelerate its AI-enabled national security missions. DOE highlights that it already operates three of the world's ten fastest supercomputers and positions these new AI-centric systems as essential for maintaining leadership in both scientific computing and AI.

The most significant governance shift lies in the operational model. Several of the new AI-oriented systems at Oak Ridge and Argonne will be located on government property with partial DOE funding but operated by Oracle, running largely on Oracle's proprietary OCI's software stack. Oracle is expected to control the majority of the compute cycles, with DOE receiving a reserved portion for its own work. Functionally, this moves these deployments closer to specialized or sovereign cloud regions anchored at national labs, rather than fully lab-owned and lab-operated supercomputers.

Access to physical infrastructure is likely another important motivator for these public–private partnerships. Frontier-scale AI systems demand substantial power capacity, cooling, and floor space that may be difficult to provision rapidly under traditional government facility planning and budgeting cycles. By co-developing on-premises AI factories with commercial partners, DOE can tap into vendor expertise and capital for power distribution, cooling technologies, and data center build-out, while still situating the systems on lab grounds. In practice, this model may allow DOE to move faster on large-scale deployments than would be possible relying solely on internal construction, permitting, and infrastructure upgrades.

## Current Unknowns

Despite the volume of these recent announcements, significant gaps remain in public information about these systems. The unknowns fall into five main categories: architecture, performance, cost, power requirements, and tenancy.

On the architecture side, most announcements provide vendor and platform names but stop short of full configurations. Node counts, memory per GPU and per node, storage capacity, and detailed interconnect designs are not disclosed in a consistent way. Even where a GPU count is publicly known (for example, the 100,000 NVIDIA Blackwell GPUs in Solstice), there is no information as to how these GPUs are packaged (e.g. per node), how much system memory is paired with them, or how the internal and external fabrics are structured. Without these details, any evaluation of the balance between compute, memory, and I/O, or comparisons with other leadership-class installations will be unsatisfying.

On performance, the available numbers emphasize aggregate AI-precision throughput for only part of the portfolio. Solstice and Equinox have a combined AI exaflop figure, but other systems do not have comparable metrics, and there is little publicly available information on double-precision performance. Without standard benchmarks for most of the new systems, it remains difficult to assess their suitability for simulation-heavy workloads relative to AI-optimized tasks. There is also no clear data on real-world performance for representative mixed workloads that combine traditional HPC with large-scale training or inference, even though this is a core justification for the AI factory model.

Regarding power requirements, the announcements are largely silent. There are no public figures for projected or peak power draw at the system level, nor for power usage effectiveness (PUE) targets, cooling strategies, or power density per rack. Given the scale of these deployments, especially as complex in the 2,000-plus AI exaflop range at Argonne, the lack of power data is non-trivial. It is unclear how much incremental electrical capacity is being added at each site, how aggressively power capping or power-aware scheduling will be required, or how these systems will interact with local power grid constraints and resilience planning.

On cost, the headline figures are aggregated and incomplete. For ORNL, Lux and Discovery are described collectively as exceeding $1 billion in public–private investment, but there is no breakdown by system or by funding source (DOE vs. vendor). For LANL and ANL, any references to projected spending are high-level and omit details such as capital versus operating cost, shared infrastructure investments, or vendor co-funding beyond vague “partnership” language. Without more granular information, it is not possible to estimate an effective cost per GPU, per node, or per unit of delivered performance, which limits the ability to benchmark these projects against traditional DOE procurements or commercial cloud consumption at scale.

Tenancy and capacity allocation models are only broadly sketched. Oracle is expected to control the majority of cycles on several systems, with DOE receiving a defined share, but no quantitative split has been disclosed. DOE’s share could be carved out as dedicated partitions, time-based reservations, or flexible quotas that can be traded off against commercial demand. It is also unclear how different categories of DOE work, such as open science, mission-critical national security workloads, and collaborative industry projects, will be prioritized within DOE’s allocation.

Until these gaps are addressed, external observers can only partially evaluate how the AI factory model balances performance, cost, power, and predictable access for DOE and its partners.

## ANALYST INSIGHT

Historically, national labs purchased bespoke systems that they owned and operated, while commercial clouds offered separate, pay-as-you-go AI and HPC services. The new DOE AI factories sit somewhere in between, with national labs acting as anchor tenants in cloud-operated environments on lab grounds.

Oracle’s role may be the most strategically significant shift. By operating some of the most demanding and sensitive AI/HPC systems in the public sector, OCI gains a reference point that extends beyond traditional commercial success stories. If these deployments meet expectations for performance, reliability, and governance, Oracle will be able to position OCI as a platform not just for generic enterprise workloads but for sovereign-grade AI, modelling and simulation, R&D, targeting defense, scientific, and highly regulated markets. This is likely to drive competitive responses from other

hyperscalers, as they pursue similar partnerships with governments and research organizations worldwide.

The DOE AI factories are intended to normalize AI-native and agentic workflows in both science and national security. This includes training large domain-specific foundation and reasoning models, embedding AI into experiment design and operation, and enabling agents that can steer simulations or recommend interventions autonomously, subject to human oversight. Patterns that prove effective and safe in the DOE context may to filter into industrial R&D, drug discovery, energy optimization, and other data-intensive, simulation-heavy domains.

## FUTURE OUTLOOK

DOE's shift away from a pure "own, operate, and fully control" model toward cloud-operated, co-funded AI factories is unlikely to be a one-off experiment. If these deployments perform as advertised, they may provide a template for how governments and other large research stakeholders can access frontier-class infrastructure without bearing the full capital and operational burden. For vendors, this model locks in long-term anchor tenants and highly visible reference sites; for DOE, it offers a way to scale capacity more rapidly and flexibly than traditional procurements.

**DOE's shift away from a pure "own, operate, and fully control" model toward cloud-operated, co-funded AI factories may provide a template for accessing frontier-class infrastructure without bearing the full capital and operational burden.**

At the same time, the design emphasis for these systems has clearly moved from FP64-centric simulation to AI-first, agentic workflows, with traditional HPC becoming one (albeit still important) workload among many. This reframes what "leadership" infrastructure means: less about peak double-precision FLOPs in isolation and more about how effectively an environment can train, host, and orchestrate large models and agents that sit in the loop of scientific discovery or mission workflows.

Combined, these cloud-style operations for national facilities and AI-native design assumptions point toward a new standard for high-end computing. DOE and its partners are effectively piloting what a next-generation research and security computing stack looks like. Organizations watching from the outside will likely take the lessons learned from this shift and will influence how they think about where control really matters, what must remain "in house," and how far they can lean into AI-centric architectures without abandoning the FP64-heavy workloads that still underpin much of science and engineering.

## About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

## Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com