

On-Prem AI Infrastructure: Comparing Dell, HPE, & More

By Adrien Laurent, CEO at IntuitionLabs • 10/22/2025 • 45 min read

- on-prem ai
- ai infrastructure
- nvidia blackwell
- gb200 nvl72
- enterprise ai
- private ai
- liquid cooling data center
- dell ai factory
- hpe private cloud ai



[Revised February 28, 2026]

Executive Summary

The leading IT vendors have each introduced advanced [on-premises AI infrastructure solutions](#), centered on [NVIDIA GPUs](#), to meet the exploding demand for enterprise-scale Generative AI. These offerings vary in architecture, cooling, and software integration, but all aim to deliver massive GPU compute density, energy efficiency, and turnkey deployment. Dell's **AI Factory** platform (e.g. PowerEdge XE97xx/XE9712) provides high-density rack-scale clusters (72 GPUs per rack with NVLink, ~30× LLM inference speed-up and up to 25× energy efficiency advantage over prior-gen systems ([cloud.watch.impress.co.jp](#))) with both liquid- and air-cooled options. HPE's **Private Cloud AI with NVIDIA** (a co-developed turnkey on-prem generative AI stack) combines GPUs (H100s, H200s, GH200s, etc.), NVIDIA's AI Enterprise software (including NIM microservices), and its GreenLake cloud management for full lifecycle support (^[1] [www.hpe.com](#)) (^[2] [nvidianews.nvidia.com](#)). Lenovo's **Hybrid AI Advantage** solutions emphasize efficiency – e.g. designing GB200 (Grace+Blackwell) GPU racks with Lenovo Neptune liquid cooling that cut data center PUE to ~1.1 and reduce power by ~40%, yielding claims of up to 45× faster inference and 40× lower energy/TCO than comparable configurations (^[3] [news.lenovo.com](#)) (^[4] [ir.supernmicro.com](#)). Supermicro's **AI SuperCluster** approach offers plug-and-play racks built from its "Building Block" servers: complete 48U systems with 72 GB300 NVL72 GPUs + 36 Grace CPUs (20 TB HBM3e in a 72-GPU NVLink fabric) or 8-GPU HGX B300 nodes, in both liquid- and air-cooled form factors (^[5] [ir.supernmicro.com](#)) (^[6] [www.supernmicro.com](#)). Cisco's **Secure AI Factory with NVIDIA** focuses on integrating AI networking and data management: for example, validated AI PODs now incorporate VAST Data's *InsightEngine* to accelerate [RAG pipelines](#) (cutting query latency from minutes to seconds) while maintaining security via Cisco AI Defense; additionally Cisco's Nexus HyperFabric clusters fuse Cisco 6000-series switches (400–800 Gbps fabric) with NVIDIA H100/H200 GPUs and BlueField DPUs for AI-optimized networks (^[7] [newsroom.cisco.com](#)) (^[8] [newsroom.cisco.com](#)). VMware's **Private AI Foundation with NVIDIA** is a vSphere/Cloud Foundation add-on that can run LLM inference workloads virtually on any OEM hardware (Dell/HPE/Lenovo, etc.), offering features like live-migratable GPU/bare-metal-like performance and enterprise-grade security (^[9] [blogs.vmware.com](#)) (^[10] [blogs.vmware.com](#)). Finally, Nutanix's **GPT-in-a-Box 2.0 (Enterprise AI)** is a validated hyperconverged solution routing standard servers (with NVIDIA L40S, H100, etc.) through Nutanix's HCI platform; it provides a simplified on-prem AI stack (including NVIDIA NIM microservices, Hugging Face integration, vector DBs, and unified data services) from edge to core (^[11] [www.nutanix.com](#)) (^[12] [www.nutanix.com](#)).

This report provides an in-depth, side-by-side comparison of these offerings. We analyze [hardware specs](#) (GPU counts, networking, cooling, and memory), software integration, performance claims (throughput, efficiency), security/governance features, case examples, and future directions. Detailed tables compare architectures and key metrics. We draw on vendor announcements, press reports, independent analysis, and expert commentary to assess each solution's strengths and use cases. The comparison reveals that while all platforms exploit NVIDIA GPUs, they differentiate on *scale versus ease-of-use versus data integration*. For instance, Dell and Supermicro push extreme GPU density (72-GPU racks) for training "superclusters," whereas Cisco and VMware emphasize secure enterprise deployments and data pipelines. HPE and Nutanix target AI adoption with managed cloud-like experiences and turnkey configuration. Lenovo highlights energy-efficient hardware. All vendors cite orders-of-magnitude improvements (e.g. 30×–45× inference speed-ups, 40% power savings) over previous generations ([cloud.watch.impress.co.jp](#)) (^[3] [news.lenovo.com](#)).

We conclude by discussing implications for enterprises: the trade-offs of such on-prem solutions (cost, complexity, data control), their role alongside public cloud AI services, and emerging trends (e.g. new NVIDIA architectures like Blackwell Ultra, Rubin—which all vendors pledge to support (^[13] [nvidianews.nvidia.com](#)) (^[14] [www.dell.com](#))). This comprehensive analysis equips IT decision-makers with the nuanced understanding needed to choose among these next-generation AI infrastructure platforms.

Introduction and Background

Modern generative AI demands unprecedented compute power while preserving data privacy and security. Enterprises often prefer on-premise or hybrid solutions for sensitive AI workloads rather than relying solely on cloud services (^[15] [blogs.vmware.com](#)) (^[9] [blogs.vmware.com](#)). To meet this need, hardware and software vendors have partnered with NVIDIA (the market leader in AI accelerators) to create **turnkey private AI platforms**. These solutions integrate NVIDIA's latest GPUs (the Blackwell architecture, memory-rich HBM, and NVSwitch interconnect) with OEM servers and racks, management software, and services. Key trends include:

- **GPU Superclusters:** Rack-scale systems combining dozens of NVIDIA GPUs (often in a fully NVLink-connected domain). Examples include single racks with 36 Grace CPUs + 72 NVIDIA GPUs (forming 1.8 TB/s NVLink fabrics) ([cloud.watch.impress.co.jp](#)) (^[5] [ir.supernmicro.com](#)), effectively turning a rack into an exascale "supercomputer". These enable training and inference of trillion-parameter LLMs with extremely high throughput.
- **Cooling and Efficiency:** High-density GPU clusters generate megawatts of heat. Vendors now emphasize liquid cooling and thermal design. For instance, Cisco and Supermicro show racks operating with 35°C–40°C coolant while capturing ~98% of heat and cutting power use by up to ~40% (^[4] [ir.supernmicro.com](#)) (^[6] [www.supernmicro.com](#)). Lenovo's Neptune system touts PUE=1.1 and 40× lower energy consumption for certain AI tasks (^[3] [news.lenovo.com](#)) (^[4] [ir.supernmicro.com](#)).
- **Integrated Software Stacks:** Beyond hardware, vendors integrate AI software. This includes NVIDIA's **AI Enterprise** suite (with Triton, RAG frameworks, NIM inference microservices, etc.), as well as vendor-specific tooling (e.g. HPE's OpsRamp AIOps, Nutanix's unified data services). The aim is a cloud-like user experience on-prem, with single-pane management, secure multi-tenancy, and automation.

- **Ecosystem Partnerships:** Each offering often bundles services or partners. HPE and VMware stress go-to-market with large SIs (Deloitte, Infosys, etc.) ⁽¹⁶⁾ www.hpe.com). Nutanix and Cisco have partner programs (e.g. Cisco + VAST, Nutanix + Hugging Face). This syndication accelerates enterprise adoption and solution validation.
- **Data and Security Focus:** With RAG and agentic AI rising, solutions now incorporate data fabrics and security. Cisco's Secure AI Factory explicitly integrates data pipelines (VAST Data's InsightEngine) to feed AI agents securely ⁽¹⁷⁾ newsroom.cisco.com) ⁽¹⁸⁾ newsroom.cisco.com). VMware's platform leverages vSphere security features (vTPM, encryption) to keep AI data on-prem ⁽¹⁹⁾ blogs.vmware.com). These measures address enterprise compliance and governance.

Key Terminology: Across vendors we see terms like NVL72 (an NVIDIA reference architecture for 72 GPUs in one cluster) and NVL2/NVL16 (multi-GPU server interconnect), as well as names like **Grace CPU**, **Blackwell/GD100 GPUs**, **InfiniBand Quantum-X800**, and **Spectrum-X Ethernet**. Vendors often cite performance in exaFLOPS (10¹⁸ operations) and memory (tens of TB) for their solutions. We detail these terms and metrics in context.

Vendor Solutions Overview

Dell Technologies – PowerEdge XE97xx Series (Part of Dell AI Factory)

Dell's strategy, branded "**Dell AI Factory with NVIDIA**", offers turnkey AI solutions spanning servers, racks, and storage. A centerpiece is the PowerEdge XE97xx line:

- **Dell PowerEdge XE9712/ XE9780:** These rack-scale servers implement NVIDIA's latest **GB200 NVL72** (Grace CPU + 72 Blackwell GPU) or **GB300 NVL72** (Blackwell Ultra) designs. In the XE9712, a single rack connects 36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell GPUs in one NVLink domain (cloud.watch.impress.co.jp). Because all 72 GPUs act as one massively parallel unit, Dell claims up to *30x faster LLM inference* compared to equivalent systems (cloud.watch.impress.co.jp). This design is part of Dell's broader *Integrated Rack 7000 (IR7000)* architecture, a 48U Open Rack with up to 480 kW cooling capacity (cloud.watch.impress.co.jp). The XE97xx can be configured in both liquid-cooled and air-cooled variants – e.g. Xeon-based nodes or AMD nodes – to suit customer needs.
- **Dell AI Factory:** This initiative brings together these hardware components with management and services. Dell provides pre-integrated racks ("plug-and-play") that Dell itself assembles and tests. Notably, Dell collaborated with CoreWeave (a cloud provider) to install the *first* Dell GB300 NVL72 racks in a production environment (particle.news) ⁽²⁰⁾ www.pcgamer.com). These racks (72 GPUs + 36 Grace + 36 BlueField DPUs each) came fully assembled from Dell and required minimal setup on-site ⁽²¹⁾ www.pcgamer.com). Such turnkey deployment illustrates Dell's solution approach.
- **Performance and Efficiency:** Dell highlights efficiency gains. For example, liquid-cooled XE97 racks at CoreWeave operate at rack-level power ~2 MW with ~1.75 MW of heat rejection via cooling (news.rambler.ru). In that deployment, Dell cites roughly 21 TB of GPU memory per rack and over 1x10¹⁸ mixed-precision operations (exaFLOPS) capability (news.rambler.ru). Moreover, Dell claims that its liquid-cooled GB200 NVL72 achieves up to 25x the energy efficiency of an equivalent air-cooled H100-based system (cloud.watch.impress.co.jp). The GTC 2025 announcement noted Dell racks can contain up to 144 Blackwell GPUs per Dell IR7000 integrated rack ⁽²²⁾ www.dell.com), underscoring their density.
- **Networking and Interconnect:** Dell's PowerEdge servers support ultra-fast networking. They added support for NVIDIA ConnectX-8 800 Gb Ethernet or InfiniBand ⁽²³⁾ www.dell.com), and use NVIDIA's Quantum-X switches for GPU-to-GPU communication. In the CoreWeave racks, ConnectX-8 and Quantum InfiniBand provided 800 Gbps fabrics (news.rambler.ru) (particle.news). Such fabrics are essential to turn 72 GPUs into a coherent single domain.
- **Use Cases:** Dell positions XE97xx for "massive LLM training and real-time inference" in enterprises, HPC centers, and research institutions (cloud.watch.impress.co.jp). The self-service, rack-level design (IR7000) is intended for easy scalability at scale. End customers include cloud providers (CoreWeave) and large organizations running giant AI models. Dell's AI Factory also includes solutions with Intel Gaudi accelerators and more, but the NVIDIA-based offerings are at the high end of performance.

👉 **Data Point:** A Reuters report confirms Dell's strategy: in May 2025 Dell unveiled servers (air/liquid variants) with up to **192 Blackwell GPUs** (with options to expand to 256) on a single platform, offering *4x faster training* than prior systems ⁽²⁴⁾ www.reuters.com). This aligns with Dell's aggressive GPU-dense designs.

👉 **2025–2026 Update:** Dell has continued to expand the AI Factory portfolio significantly. At Dell Technologies World in May 2025, Dell unveiled "**AI Factory 2.0**" with over 200 updates, including **Project Lightning**, a new parallel file system claimed to be the world's fastest for AI training workloads. In November 2025, Dell introduced the **PowerEdge XE8712**, a liquid-cooled system supporting up to 144 NVIDIA Blackwell GPUs per IR7000 rack (generally available December 2025), and the **PowerEdge XE9785/XE9785L** with AMD EPYC CPUs and eight AMD Instinct MI355X GPUs—marking Dell's first major non-NVIDIA AI server option. Dell also introduced **PowerCool** liquid cooling technology and claims 100x faster token generation per second for distributed AI inferencing and 80%+ latency reduction vs. previous generation systems (dell.com). Additionally, Dell partnered with IREN to deploy GB300 NVL72 systems in Canada, expanding the platform's geographic reach.

HPE – Private Cloud AI with NVIDIA (“NVIDIA AI Computing by HPE”)

HPE and NVIDIA launched a jointly developed portfolio dubbed “NVIDIA AI Computing by HPE”. The flagship is **HPE Private Cloud AI**, a turnkey on-prem GenAI solution:

- **Solution Composition:** HPE Private Cloud AI is essentially a pre-integrated private cloud stack. It combines NVIDIA's computing platforms (GPUs/GPUs), NVIDIA AI Enterprise software (including Triton inference server and NIM microservices), networking, and HPE's own infrastructure. Specifically, it integrates: HPE compute nodes (ProLiant servers supporting NVIDIA L40S, H100 NVL, and GH200 NVL2 GPUs (^[26] [nvidianews.nvidia.com](#))), HPE storage (GreenLake for File Storage), and NVIDIA Spectrum-X Ethernet (or InfiniBand) for networking (^[2] [nvidianews.nvidia.com](#)). The entire stack is managed by HPE GreenLake cloud-like interface, offering self-service and AIOps (via the OpsRamp AI copilot) for full lifecycle management (^[27] [www.hpe.com](#)) (^[2] [nvidianews.nvidia.com](#)).
- **Turnkey and Lifecycle:** A key selling point is “turnkey” simplicity for enterprise customers. HPE & NVIDIA claim the stack can be delivered in **right-sized configurations** (four pre-defined hardware scales) to match different organizational needs (^[28] [www.hpe.com](#)). Because it is delivered as a service-like offering, HPE provides installation, configuration, training, and ongoing monitoring. The integration with OpsRamp brings automated IT workflow and a conversational AI assistant to help manage AI workloads (^[27] [www.hpe.com](#)). In effect, enterprises get a private cloud for AI with built-in GPU virtualization, orchestration, and monitoring.
- **Support and Ecosystem:** HPE emphasizes enterprise features: data privacy/governance (NVIDIA's microservices enable isolated inference), and broad partner integration. All solutions are sold jointly by HPE/NVIDIA and through top SI partners (Deloitte, Infosys, etc.) (^[29] [nvidianews.nvidia.com](#)) (^[27] [www.hpe.com](#)). HPE also announced collaboration with major hardware vendors: for example, HPE ProLiant NX servers supporting upcoming NVIDIA Rubin/Vera chips, and HPE storage certified for NVIDIA DGX BasePOD (^[30] [nvidianews.nvidia.com](#)). At launch, HPE also highlighted global deals (like support from major consultants) and timetables (General Availability expected in fall 2024 (^[31] [nvidianews.nvidia.com](#))).
- **Hardware Details:** While partly virtualized, HPE lists reference hardware: e.g. HPE Cray XD670 can house 8 NVIDIA H200 NVL GPUs (for LLM “builders”), HPE DL384 Gen12 handles 2x GH200 NVL2 for large models, and DL380a Gen12 supports 8x H200 NVL (flexible scaling) (^[30] [nvidianews.nvidia.com](#)). HPE explicitly states it will support NVIDIA's latest GPU/CUD architectures (GB200 NVL72, GH200 NVL2) and upcoming ones (Blackwell Ultra, Rubin, Vera) as they arrive (^[13] [nvidianews.nvidia.com](#)). Networking is typically NVIDIA Quantum InfiniBand or SN switches behind the scenes to link nodes.
- **Use Cases:** HPE targets enterprise customers developing proprietary LLMs on-prem – especially ones needing private data use (finance, healthcare, gov't). Private Cloud AI supports fine-tuning, inference, and RAG pipelines on local data (^[32] [nvidianews.nvidia.com](#)). The integration with GreenLake means customers have consumption-based billing even for on-prem hardware. HPE also points out use in HPC/AI convergence (via its Cray supercomputer line) and co-marketing across industries.
- **Performance:** HPE's announcement was light on raw performance numbers, focusing instead on TCO and synergy. However, it mentions support for NVLink fabrics and NVIDIA's accelerated AI software, implying enterprise-grade throughput. One independent report noted that Dell's new H200-based servers (likely similar spec) achieved ~4x training speedup over prior-gen (^[24] [www.reuters.com](#)), suggesting HPE systems would be in a comparable range.

👉 **Data Point:** In bench tests, VMware-NVIDIA collaboration (which HPE's stack builds on) achieved performance comparable to bare-metal. VMware reported that running AI workloads on the Virtual Private AI Foundation yields “performance similar to, and sometimes better than bare metal” due to efficient GPU passthrough and NVLink support (^[9] [blogs.vmware.com](#)). HPE's nascent platform aims for this level of virtualization efficiency while adding enterprise management.

👉 **2025–2026 Update:** HPE has significantly upgraded its portfolio. The **HPE ProLiant Compute XD685** (5U, direct-liquid-cooled, 8x NVIDIA B300 Blackwell Ultra GPUs) became generally available in January 2026, while the **HPE Compute XD690** (8x Blackwell Ultra GPUs) shipped from October 2025. HPE also added NVIDIA RTX PRO 6000 Blackwell GPU support and STIG-hardened, FIPS-enabled AI software for government and defense use cases (^[33] [hpe.com](#)). In December 2025, HPE rebranded and expanded GreenLake as “**GreenLake Intelligence**”, an agentic AI-powered hybrid cloud platform that embeds autonomous AI agents across networking, storage, compute, observability, and FinOps functions. The new **HPE Alletra Storage MP X10000** (available January 2026) adds Model Context Protocol (MCP) server support for AI-native storage workloads (^[34] [hpe.com](#)).

Lenovo – Hybrid AI Advantage (Collaboration with NVIDIA)

Lenovo's **Hybrid AI Advantage** initiative extends previous NVIDIA collaborations (e.g. Tech World announcements) and focuses on delivering *energy-efficient, hybrid cloud-friendly* AI infrastructure:

- **Hardware Platforms:** At NVIDIA GTC 2024, Lenovo unveiled new **ThinkSystem AI servers**. Highlights include an 8-GPU node (like ThinkSystem SR780a V3) with liquid cooling (Lenovo Neptune bath-water technology) achieving PUE as low as 1.1 (^[35] [news.lenovo.com](#)) to sustain high density. It runs 8x NVIDIA B200 GPUs (Grace Blackwell), delivering supposedly *45x faster large-model inference with 40x reduced energy use* vs older systems (^[3] [news.lenovo.com](#)). Lenovo also introduced an air-cooled SR680a V3 for mixed CPU/GPU use, and a compact 1U PG8A0N node with open-loop liquid cooling for a single GB200 Superchip support (^[3] [news.lenovo.com](#)). The result is that Lenovo will offer “GB200 rack systems” (composed of GB200 Superchips) and high-density GPU servers to “supercharge AI training” (^[3] [news.lenovo.com](#)).

- **Neptune Liquid Cooling:** Lenovo's proprietary cooling (Neptune) uses direct water loops. For example, their new SR780a achieves up to 40% reduction in power draw and 3.5× higher thermal efficiency than traditional air cooling (^[35] [news.lenovo.com](#)). This lets Lenovo pack 8×1100W GPUs in 5U or 8U racks without throttling. They claim reusing waste heat "enables more AI performance without thermal limits". Lenovo touts this as a key sustainability edge, and notes the company's rank on the Green500 list for energy-efficient designs (^[36] [news.lenovo.com](#)).
- **AI Ecosystem:** Lenovo's solutions are certified for NVIDIA AI Enterprise software. They support NVIDIA DLSS and inference microservices, and provide Lenovo XClarity and LiCO orchestration stacks for managing clusters. The portfolio also embraces NVIDIA MGX modular design: e.g. new MGX 1U/2U/4U systems (HG630N, HG650N, HG660X V3) designed for H200/Hopper GPUs and future Grace/MGX architectures (^[37] [news.lenovo.com](#)). Importantly, Lenovo collaborates with CSPs via its MGX reference designs (for NVIDIA OVX and Omniverse).
- **Use Cases:** Lenovo emphasizes enabling "AI anywhere" (cloud, enterprise, edge). The Hybrid AI solutions target enterprises building AI at the data source – from private cloud to edge devices. For example, it cites use cases in retail personalization, smart factories, and urban IoT analytics (^[38] [news.lenovo.com](#)). The partnership with NVIDIA ensures support for cutting-edge models (via Hugging Face plug-ins, NVIDIA NIM). Lenovo often ties in PC offerings too (e.g. new RTX-powered workstations for developers).
- **Performance:** Lenovo's disclosure highlights inference acceleration: up to 25× lower-latency LLM inference on Blackwell B200 GPUs (^[39] [news.lenovo.com](#)). Combined with better cooling, they claim "hosting massive AI workloads without compromise". No single exaFLOPS claim is given for the racks, but the GB200 (Grace) design in particular is said to support "trillion-parameter models with 40 TB of fast memory per rack" ([chubutdigital.com.ar](#)), in partnership with CoreWeave. (Lenovo's role in that deployment is not detailed, but it indicates support for NVIDIA's largest memory GPUs).

👉 **Data Point:** Lenovo notes that its direct-water-cooled nodes allow a 100% increase in GPU density per rack vs legacy systems. Citing independent benchmarks, they state its new 8U servers with 8×H200 GPUs achieve 2.3 TB HBM3e and 800 Gb/s networking, matching up to 288 GB per GPU in aggregate memory (^[39] [news.lenovo.com](#)) (^[5] [ir.supermicro.com](#)).

👉 **2025–2026 Update:** At CES 2026 (January 2026), Lenovo launched three purpose-built AI inferencing servers: the **ThinkSystem SR675i** (for full LLM deployments in manufacturing, healthcare, and financial services), the **ThinkSystem SR650i** (high-density GPU compute for existing data centers), and the **ThinkEdge SE455i** (ultra-compact for edge deployments in retail, telco, and industrial environments). All include Neptune air- and liquid-cooling options and are available through Lenovo's TruScale pay-as-you-go pricing model (^[40] [news.lenovo.com](#)). Lenovo also announced a **gigawatt-scale AI factory program with NVIDIA**, combining GB300 NVL72 rack-scale systems with Lenovo liquid cooling and global manufacturing capabilities to reduce deployment timelines from months to weeks (^[41] [news.lenovo.com](#)).

Supermicro – AI Building Block Solutions and SuperClusters

Supermicro's approach is to offer **modular building-block** systems and pre-integrated *SuperClusters*. Key elements:

- **NVIDIA Blackwell Platforms:** Supermicro provides both **8U and 4U servers** optimized for NVIDIA's newest GPUs. Its 8U systems can house 8× NVIDIA HGX B300 (Blackwell Ultra) GPUs (with 2.3 TB HBM3e total) or 8× HGX B300 Tensor Core GPUs (^[42] [ir.supermicro.com](#)) (^[6] [www.supermicro.com](#)). Similarly, smaller 4U variants hold 4× GPUs for lighter workloads. For the largest scale, Supermicro's **GB300 NVL72 SuperCluster** integrates **72 Blackwell Ultra GPUs plus 36 Grace CPUs in one rack** (^[5] [ir.supermicro.com](#)). This rack has ~20 TB of GPU memory and 1.8 TB/s NVLink across all GPUs (^[5] [ir.supermicro.com](#)). In other words, a single GB300 NVL72 furnace is exascale-class.
- **Cooling and Plumbing:** Supermicro is at the forefront of data center liquid-cooling. Their *AI SuperClusters* use direct-to-chip water blocks with 98% heat capture (DLC-2 tech). PR material states their 4U liquid-cooled systems achieve **40% energy savings** from recycling heat (^[43] [www.supermicro.com](#)). In a 16-node 48U configuration, racks operate at 35°C-40°C inlet water, cutting power ~40% (^[44] [ir.supermicro.com](#)). They even offer 250 kW in-rack heat exchangers and in-row chillers. This "free cooling" is often cited as making liquid cooling essentially "costless" for customers (^[45] [ir.supermicro.com](#)).
- **Networking and Turnkey Fit:** These systems include integrated NVIDIA InfiniBand (Quantum-X800) or Spectrum-X E800 for cluster fabrics. Notably, the product pages show each rack includes 9 NVLink switches to fuse GPUs (^[46] [www.supermicro.com](#)). Supermicro emphasizes "plug-and-play" delivery: each rack arrives fully cabling and networking pre-validated (termed L11/L12 validation) (^[44] [ir.supermicro.com](#)). This speed-to-market is critical in GTC demonstrations: for example, vendors often say "we can deploy a SuperCluster in weeks not months." The Supermicro PR mentions that its building blocks power "the world's largest liquid-cooled AI data center deployment" (^[47] [www.supermicro.com](#)).
- **Cutting-edge Support:** Supermicro explicitly announced support for NVIDIA's newest architectures: besides GB300 NVL72, it provides all-new HGX B300 NVL16 servers and will flow support for upcoming Rubin/Vera. They note "first-to-market" support for all NVIDIA Blackwell generations (B100, B200, GB200, GB300) (^[48] [ir.supermicro.com](#)). The IR announcements pledge joint GH200/H200 platforms already shipping, and imminent support for GB300 NVL72-tier.
- **Use Cases:** Supermicro targets both hyperscale AI clouds and enterprise HPC. They talk about "AI SuperClusters" for training and inference at scale (^[49] [ir.supermicro.com](#)). Clients include service providers (e.g. CoreWeave), research institutions, and enterprises needing maximum performance per rack. Supermicro also partners with the NVIDIA AI Enterprise stack (Triton, NIM) and expects heavy use in DGX BasePOD and OVX solutions.

👉 **2025–2026 Update:** Supermicro has expanded its Blackwell Ultra portfolio with new form factors: a **4U front-I/O liquid-cooled HGX B300** server (standard 19-inch EIA rack, 8× B300 GPUs per node, DLC-2 technology capturing 98% of heat) and a **2-OU OCP liquid-cooled HGX B300** (Open Rack V3 spec, up to 144 GPUs per rack at 18 nodes) targeting hyperscale and cloud deployments. A full SuperCluster configuration now supports 8

HGX B300 compute racks + 3 NVIDIA Quantum-X800 InfiniBand racks + 2 in-row CDUs, totaling 1,152 GPUs per scalable unit. Volume shipments of Blackwell Ultra systems have begun ⁽⁵⁰⁾ [supermicro.com](#)). On the corporate side, Supermicro filed its delayed SEC reports in late February 2025, regaining Nasdaq compliance. However, replacement auditor BDO issued an adverse opinion on internal controls over financial reporting, and DOJ/SEC investigations remain ongoing as of early 2026 ⁽⁵¹⁾ [cnbc.com](#)).

Cisco – Secure AI Factory with NVIDIA (and Nexus HyperFabric)

Cisco addresses an emerging niche: integrating enterprise data management and security with AI computing. Two major announcements illustrate this:

- Secure AI Factory with NVIDIA:** This is Cisco's validated reference architecture for on-prem AI, focusing on security at the data-level. It consists of Cisco UCS servers with GPUs, Nexus networking fabric, and Cisco AI software (AI Defense, Splunk integration for monitoring) ⁽⁵²⁾ [newsroom.cisco.com](#)). The **AI POD** building blocks can be ordered with NVIDIA RTX PRO (Blackwell) GPUs for inference, plus connected storage (initially Cisco's own or partner SAN) for large data. In September 2025, Cisco extended Secure AI Factory to include **VAST Data's InsightEngine** for RAG (Retrieval Augmented Generation) use cases ⁽¹⁷⁾ [newsroom.cisco.com](#)) ⁽¹⁸⁾ [newsroom.cisco.com](#)). This upgrade means Cisco AI nodes now deliver ultra-fast data I/O (NVIDIA Quantum InfiniBand under the hood) to reduce RAG latency from "minutes to seconds" ⁽¹⁸⁾ [newsroom.cisco.com](#)), enabling near-real-time AI agents. Cisco's emphasis is on end-to-end security: their AI Defense tools provide token-level audit/logging and compliance. A press quote summarizes: "Securing every token in the AI pipeline" to enable enterprise-grade agentic AI ⁽⁵³⁾ [newsroom.cisco.com](#)).
- Nexus HyperFabric (AI Clusters):** Earlier in 2024, Cisco unveiled *Nexus HyperFabric*, an all-in-one AI fabric solution for data centers ⁽⁵⁴⁾ [newsroom.cisco.com](#)). It combines Cisco's high-performance switches (Silicon One-based 6000-series with 400G/800G ports) with NVIDIA accelerated computing and VAST's storage to simplify multi-node AI clusters ⁽⁵⁴⁾ [newsroom.cisco.com](#)) ⁽⁵⁵⁾ [blogs.nvidia.com](#)). The system provides single-pane management (cloud control) across compute and network, essentially acting like a "hypervisor for networking". At launch, it was demonstrated with H100 NVL GPUs, BlueField DPUs (for hardware offload/security), and the NVIDIA AI Enterprise software stack ⁽⁶⁾ [newsroom.cisco.com](#)) ⁽⁵⁶⁾ [blogs.nvidia.com](#)). This solution is pitched to organizations that want enterprise-grade scale-out AI without building it from scratch. Cisco claims its customers "will be able to deploy AI infrastructure without deep networking expertise" thanks to Nexus HyperFabric's automation ⁽⁵⁷⁾ [newsroom.cisco.com](#)).
- Performance and Scale:** Cisco's public statements on performance focus on data throughput rather than raw FLOPS. For example, their VAST-enabled AI POD can now reach over exaflop-class workload capacity per rack and provide ~40 TB of fast memory to AI tasks [chubutdigital.com.ar](#)). Cisco also highlights multi-agent capabilities: enabling *agentic AI* at enterprise scale, meaning "AI agents operate continuously, dynamically learning, using all relevant data." The infrastructure is built to support many simultaneous RAG tasks with high security ⁽⁵⁸⁾ [newsroom.cisco.com](#)) ⁽¹⁸⁾ [newsroom.cisco.com](#)).

👉 **2025–2026 Update:** Cisco has expanded its Secure AI Factory hardware lineup. The **Cisco UCS C845A M8** (with NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs) is now orderable, with the **UCS C240 M8** and **UCS X580p PCIe Node** (supporting up to 4x RTX PRO 6000 GPUs) expected in early 2026. The infrastructure stack now spans Cisco UCS + NVIDIA AI Enterprise + NVIDIA BlueField DPUs + Cisco Nexus networking + Cisco Secure Workload + Cisco Intersight + Splunk observability. At VAST Forward 2026, Cisco demonstrated a full reference implementation combining NVIDIA compute, VAST data services, Cisco infrastructure, and the Isovalent Enterprise Platform (Cilium/Tetragon) for cloud-native AI security.

VMware – Private AI Foundation with NVIDIA

VMware's offering is primarily software, enabling a standardized platform for on-prem GenAI:

- Platform:** VMware's **Private AI Foundation with NVIDIA** is an add-on to VMware Cloud Foundation (vSphere/vSAN/NSX stack). It includes NVIDIA's NIM inference microservices and GPU virtualization, and provides pre-configured DL VM images and RAG workflows. In practice, enterprises deploy their own servers (Dell, HPE, Lenovo, or VMware's own EVO:RAIL etc) running vSphere/VCF; this add-on enables them to very easily spin up GPU-accelerated workloads.
- Virtualization Efficiency:** A key point is that VMware vSphere now supports NVIDIA GPUs in a near-bare-metal manner. According to VMware, the overhead is minimal: in internal benchmarks, AI workloads on this platform achieved *similar or even better performance than on comparable bare-metal clusters* ⁽⁹⁾ [blogs.vmware.com](#)). This is attributed to technologies like GPU physical passthrough, NVIDIA NVLink bridging, and optimized I/O. Thus, companies can enjoy virtualization benefits (mobility, snapshots, security) without sacrificing GPU speed.
- Security & Compliance:** By building on vSphere, the solution inherits strong security (Secure Boot, VM encryption, role-based access) and regulatory compliance controls. VMware emphasizes that *privacy* is architected into the platform: data never leaves the customer's data center, and VRAM is virtualized securely ⁽¹⁹⁾ [blogs.vmware.com](#)). Administrators gain unified management of AI models, vector DBs, and infrastructure, with governance policies enforced through vCenter.
- OEM Adoption:** VMware notes that major OEM hardware vendors have committed to support this platform ⁽⁵⁹⁾ [blogs.vmware.com](#)) ⁽⁶⁰⁾ [blogs.vmware.com](#)). For instance, Dell, HPE, and Lenovo provided reference architectures for Private AI Foundation (i.e. validated node configurations with NVIDIA GPUs, drivers, etc.). This means customers can choose their server hardware and be assured it will work "out of the box" with VMware's AI stack.

- **Usage:** VMware envisions enterprises using this for private LLM inference, RAG, and fine-tuning tasks. The platform provides automation wizards (Catalog Setup Wizard) to simplify provisioning GPU cluster templates for DevOps teams (^[61] [blogs.vmware.com](#)). Because it runs on the same control plane as an organization's existing vSphere environment, IT staff can apply familiar processes.

👉 **2025–2026 Update (VMware):** A major shift occurred with **VMware Cloud Foundation (VCF) 9.0** (generally available August 2025 at VMware Explore 2025): **VMware Private AI Services is now bundled as a standard component of VCF 9.0**, no longer a separate add-on. VCF 9.0 is positioned as an "AI-native platform" and includes GPU Monitoring, Model Store, Model Runtime (multi-accelerator support for AMD and NVIDIA GPUs), Agent Builder, Vector Database, and Data Indexing/Retrieval services. New hardware support includes NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, NVIDIA B200 GPUs, and ConnectX-7/BlueField-3 400G DPUs with Enhanced DirectPath I/O (^[62] [broadcom.com](#)). VCF 9.0 also introduced **VCF Intelligent Assist** (tech preview), an AI-driven support assistant. Note that Broadcom's licensing changes for VCF (moving to subscription bundles) have caused significant disruption for some existing VMware customers.

Nutanix – GPT-in-a-Box 2.0 and Enterprise AI

Nutanix addresses AI with its **GPT-in-a-Box** line, evolved into an **Enterprise AI Foundation**:

- **GPT-in-a-Box 2.0:** Announced in mid-2024, this is a pre-configured software/hardware solution for GenAI. It is built on Nutanix's Acropolis HCI platform, with added AI features. Hardware-wise, it can run on standard x86 servers (Dell, HPE, Lenovo, Nutanix-branded NX) equipped with NVIDIA data-center GPUs (L40S, L40, H100) (^[12] [www.nutanix.com](#)) (^[11] [www.nutanix.com](#)). It supports high-density GPU setups (e.g. racks with NVL72 or smaller clusters) but crucially does *not* require any special proprietary hardware. This allows customers to leverage existing infrastructure easily.
- **Software Stack:** GPT-in-a-Box integrates NVIDIA's AI Enterprise (Triton, NIM), Hugging Face model catalogs, and specialized data services. For example, it includes vector databases and connectors for RAG, unified management of models and endpoints, and a point-and-click UI for deploying NVIDIA NIM microservices (^[63] [www.nutanix.com](#)) (^[64] [www.nutanix.com](#)). The idea is to make it as easy as possible for enterprise developers to spin up private Generative AI services (chatbots, code generation, content creation) without building infrastructure pipelines manually.
- **Data Services:** Being on the Nutanix Cloud Platform, the solution relies on Nutanix's distributed storage (Objects/Files) to serve model artifacts and data. Notably, Nutanix engineered very high-density storage now (multi-PBs in small footprint) and will add support for NVIDIA GPUDirect Storage to boost data-to-GPU throughput (^[65] [www.nutanix.com](#)). These enhancements mean feeding GPUs with data can scale with the GPU compute power.
- **Ease of Deployment:** Nutanix touts that GPT-in-a-Box 2.0 requires "no special architecture" – IT can simply attach high-end NVIDIA GPUs to any Nutanix cluster. The deployment is streamlined through Nutanix Prism/Ansible automation and pre-validated designs. Customers can run LLM inference workloads at the edge (IoT devices or branch offices) or core, with consistent management. (^[66] [www.nutanix.com](#)) (^[67] [www.nutanix.com](#)) The solution also emphasizes governance: RBAC, audit trails, and hybrid-cloud consistency (via Nutanix Cloud Infrastructure) allow maintaining privacy and auditability.
- **Partnerships and Compatibility:** The 2024 press release emphasizes collaboration with NVIDIA, Hugging Face, and infrastructure OEMs (^[68] [www.nutanix.com](#)). For example, customers can deploy okay-run Hugging Face's Transformers with full Nutanix support. Nutanix validates hardware such as the new **NX-9151** (a modular NVIDIA MGX rack with Rubin Superchips) in their system roadmap (^[12] [www.nutanix.com](#)), indicating support for future NVIDIA architectures. It also explicitly mentions support for Dell/HPE/Lenovo GPU servers, enabling high GPU densities to reduce TCO (by packing more compute in fewer racks) (^[12] [www.nutanix.com](#)).

👉 **2025–2026 Update (Nutanix):** GPT-in-a-Box has been effectively superseded by **Nutanix Enterprise AI (NAI)** as the primary branding. At .NEXT 2025 (May 2025, Washington, D.C.), Nutanix announced "Nutanix Enables Agentic AI Anywhere," featuring deep integration with NVIDIA AI Enterprise (NIM microservices + NeMo framework), a new **Shared Model Service** methodology (allowing multiple applications to share common embedding, reranking, and guardrail models), generative AI safety/guardrails features (query filtering, topic control, jailbreak detection), and function calling capabilities for agentic applications. NAI supports deployment from edge to on-premises private cloud to public cloud (^[69] [nutanix.com](#)).

Comparative Analysis

To systematically compare the solutions, we construct a summary table of key attributes:

Feature/Metric	Dell AI Factory (PowerEdge XE97xx)	HPE Private Cloud AI	Lenovo Hybrid AI Solutions	Supermicro AI Clusters	Cisco Secure AI Factory	VMware Private AI Foundation	Nutanix GPT-in-a-Box 2.0
GPU Architecture	NVIDIA GB200/GB300 NVL72 (Grace+Blackwell)	NVIDIA H100/H200/GH200,	NVIDIA B200 (Grace+Blackwell),	NVIDIA B300 NVL16 (8xGPUs) &	NVIDIA RTX PRO 6000 Blackwell (inference) & H100/H200 in clusters	Any NVIDIA GPU (L40S, H100, etc) on vSphere	Any NVIDIA GPU (L40S, H100, etc) on Nutanix nodes
	(up to 72 GPUs/ rack)	L40S on ProLiant, planning Rubin/Vera support (^[13] nvidianews.nvidia.com)	Stated upcoming H200/Others via MGX (^[37] news.lenovo.com)	NVIDIA GB300 NVL72 (72 GPUs/rack) (^[5] ir.supermicro.com)	(with Nexus HyperFabric includes H100 NVL, BlueField) (^[3] newsroom.cisco.com)	(virtualized GPU via NVLink) (^[9] blogs.vmware.com)	(standard servers with GPUs)

Feature/Metric	Dell AI Factory (PowerEdge XE97xx)	HPE Private Cloud AI	Lenovo Hybrid AI Solutions	Supermicro AI Clusters	Cisco Secure AI Factory	VMware Private AI Foundation	Nutanix GPT-in-a-Box 2.0
GPUs/Server-Rack	Up to 8 GPUs per node, 72 GPUs per rack (NVL72) (cloud.watch.impress.co.jp)	Varies by config; 8 GPUs per node, clusters up to ~64 GPUs+	Lenovo PGBA0N: 1 GPU (1U); SR780a V3: 8 GPUs (5U) (35) news.lenovo.com)	8 GPUs/node (HGX B300) (70) ir.supermicro.com); 72 GPUs/rack (GB300 cluster) (5) ir.supermicro.com)	Cisco AI POD: e.g. RTX PRO Servers with 8 GPUs; NX HyperFabric clusters: 8-Node racks with H100/H200 NVLs (8) newsroom.cisco.com)	Node-level GPUs via vSphere (no fixed count)	Cluster GPUs per node depends on hardware; supports racks of NVL72 by reference
On-Chip Mem (HBM)	~21–40 TB per 72GPU rack (news.rambler.ru) (chutudigital.com.ar)	Up to 1–2 TB per node; depends on GPUs (e.g. GH200 500GB, H100 80GB)	Up to 45x inference speed (GB200) implies multi-TB per rack (3) news.lenovo.com)	2.3 TB per 8-GPU (HGX B300); ~20 TB per 72-GPU rack (5) ir.supermicro.com)	Data Platform spans racks; Cisco doesn't publish per-server HBM (uses RTX6000 Ada: 48GB each)	Virtual – leverages hardware GPUs as available	Depends on installed GPUs; currently H100 80GB, L40 48GB;
CPU	NVIDIA Grace (36 per rack, integrated) (cloud.watch.impress.co.jp); AMD/Intel for host	Intel Xeon/AMD EPYC (e.g. Cray XD670 double-socket) (71) nvidianews.nvidia.com)	Intel Xeon (SR680a) or ARM Grace (in GB200 SoC)	Dual Xeon or AMD per node; 36 Grace (NVL72) at rack-level (5) ir.supermicro.com)	Cisco UCS servers (Intel Xeon/APL, or Grace Hopper in future OVX nodes)	Any (runs on existing VMware clusters)	AMD EPYC or Intel Xeon; focus on heterogeneous compute
Cooling	Liquid cooling (CDU racks); air-cooled nodes also offered (cloud.watch.impress.co.jp) (news.rambler.ru)	Primarily air-cooled designs (GreenLake colocation); some liquid in HPC lines	Direct liquid (Neptune) cooling standard in high density; also air-cooled options (35) news.lenovo.com)	Both: DLC-2 liquid-cooled (98% capture) and air-cooled versions (43) www.supermicro.com)	Air/water cooling for servers; Cisco Nexus fabric uses air 400–800G switches	Virtual, uses existing data center cooling	Standard cooling of X86 servers; liquid optional with partner
Power/Heat	Up to ~1.8–2.0 MW per rack (72 GPUs) (news.rambler.ru); 25%–40% power savings with liquid vs air (cloud.watch.impress.co.jp)	HPE promises "sustainable" design; no public power numbers. GreenLake monitoring	Claims up to 40% facility power reduction with liquid vs comparable air (35) news.lenovo.com)	8-node rack: 40°C inlet, 16-node: 35°C (up to 40% savings) (4) ir.supermicro.com)	Cisco AIM: not specified per rack; focus on efficiency of network-managed clusters	Depends on cluster design; VMware supports DV switching for power efficiency	Depends on chosen hardware; Nutanix platform itself is HW-agnostic
Network Fabric	NVIDIA Quantum-X800 InfiniBand and ConnectX-8 800GbE; Dell's IR7000 bus	NVIDIA Spectrum-X (Ethernet) and InfiniBand; HPE GreenLake					
	NVIDIA InfiniBand (Quantum-X800) / Spectrum-X800 (100/400GbE) (72) news.lenovo.com)						
	Cisco Silicon switches (Cisco 6000 series) 400G/800G fabrics (8) newsroom.cisco.com) (56) blogs.nvidia.com)	VMware: uses NVIDIA's ConnectXBlueField fabrics; vDS network virtualization	Nutanix: HCI fabric (vSwitch/KBs networking), optional Infiniband via ConnectX				

Scale & Density | Rack-scale (up to 36 nodes/rack on IR7000; 144 GPUs/rack) (22) www.dell.com) | Configurable: from small Private Clouds to full DGX BasePOD scale | Racks of 4–8 nodes; example 5U/8U configurations, or MGX racks (open standards) | Rack-level "Superclusters" (5–9 racks per cluster); turnkey installed systems | Multi-rack solutions for large enterprise sites; secure multi-cluster architecture | Scales on existing vSphere clusters; multi-site federation via SDDC | HCI clusters of any size; "multi-cloud" federation; portal management |

Software Stack | Dell OpenManage & QuickSync plus NVIDIA AI Enterprise (Triton, NIM); Dell HPC & APEX Cloud services | HPE GreenLake management; HPE AI software (OpsRamp AI Ops); NVIDIA AI Enterprise (Triton, NIM, GPUs drivers) (2) nvidianews.nvidia.com) | Lenovo XClarity & LiCo management; NVIDIA AI Enterprise (Triton); becomes part of Lenovo AI Ready stack | Supermicro SuperDoctor/OneClick; NVIDIA AI Enterprise (Triton/TensorRT, NIM); plus optional Slurm/HPC toolchain | Cisco Intersight/AI Ops for infra; NVIDIA AI Enterprise/Triton; Cisco AI Defense suite; Splunk monitoring (73) newsroom.cisco.com) | VMware Cloud Foundation UI; NVIDIA NIM, Triton; marketplace for LLMs; vCenter/Grafana dashboards (GPU monitoring) (9) blogs.vmware.com) | Nutanix Prism/CALM; NVIDIA AI Enterprise (Triton/NIM); Hugging Face integration; pgvector for RAG; Nutanix Objects/Files (NUS) data platform |

Deployment Model | Pre-integrated racks (Dell-managed assembly); air- or liquid-cooled; enterprise installation services | Turnkey rack or cluster delivered and managed by HPE; offers consumption-based GreenLake billing | On-prem servers or racks sold through OEM channels; includes Adobe-style "co-engineered" reference architectures | Factory-integrated racks, global delivery; "ready out of box" with full stacking, cabling; optional in-house integration services | Pre-validated architecture (Cisco internal or partners install clusters); focus on enterprise data center environments | Delivered as software license; runs on customer's existing servers/clusters with vSphere & Network | Delivered on validated hardware (customer choice); uses standard Nutanix software subscriptions |

Primary Use Cases | **LLM training/inference** at scale (enterprises, research), data center acceleration (HPC/AI convergence) (cloud.watch.impress.co.jp) (news.rambler.ru) | **Private generative AI clouds** for corporate; fine-tuning LLMs on proprietary data; RAG workflows with compliance | **Hybrid AI at the edge and core** – running inference and smaller models anywhere (from IoT/edge to cloud) efficiently (36) news.lenovo.com) (67) www.nutanix.com) | **AI data centers and clouds** – building "AI factories" for model training and simulation in hyperscale settings (74) ir.supermicro.com) (5) ir.supermicro.com) | **Enterprise AI with secure data access** – deploying AI agents and RAG-based assistants while preserving data security (finance, telecom, government) (73) newsroom.cisco.com) (18) newsroom.cisco.com) | **Private-genAI deployments** – serving LLM inference (NLP copilots, RAG apps) on-prem with cloud-like agility and security (9) blogs.vmware.com) (10) blogs.vmware.com) | **Enterprise GenAI**

acceleration – pilot to production of GPT use cases (chatbots, code generation, document AI) on on-prem infrastructure (^[75] www.nutanix.com) (^[76] www.nutanix.com) |

Each cell above is supported by cited sources. For example, Dell's XE97xx is documented by Dell and news sources as employing the NVIDIA GB200 NVL72 design (72 GPUs per rack) (cloud.watch.impress.co.jp). HPE's platform details come from the official NVIDIA/HPE announcement (^[1] www.hpe.com) (^[2] nvidianews.nvidia.com). Lenovo's entries reflect the Gizmodo/Lenovo press release describing Neptune cooling and GPU counts (^[35] news.lenovo.com) (^[3] news.lenovo.com). Supermicro's specs are from its PRNewswire releases and product pages (^[5] ir.supermicro.com) (^[6] www.supermicro.com). Cisco information is drawn from Cisco press releases and blogs (^[52] newsroom.cisco.com) (^[8] newsroom.cisco.com). VMware's capabilities are sourced from VMware Cloud Foundation blog posts (^[9] blogs.vmware.com). Nutanix information comes from its own press and blog (^[76] www.nutanix.com) (^[11] www.nutanix.com).

Performance and Efficiency Data

Because each system prioritizes different metrics, performance comparisons must factor in architecture:

- **Compute Throughput:** Dell and Supermicro's NVL72 solutions deliver *exaFLOPS* of AI throughput. Supermicro explicitly quotes **2.3 TB/s NVLink** domains and 20+ TB of GPU RAM per GB300 NVL72 rack (^[5] ir.supermicro.com), while Dell's XE9712 promises similar ~30× LLM acceleration (cloud.watch.impress.co.jp). Cisco's VAST-enabled racks claim **1.1 exaFLOPS** dense inference (FP4) performance per rack (particle.news). In contrast, HPE and VMware do not emphasize raw FLOPS; they stress adequate acceleration (e.g. HPE's 192× NVL GPUs for 4× speed-up (^[24] www.reuters.com)).
- **Inference vs Training:** Lenovo, Cisco, and Nutanix largely highlight inference. Lenovo's "45× faster inference" stat (^[3] news.lenovo.com) assumes use of B200's Tensor Core improvements. Cisco's agentic AI focus implies multi-step inference chains with RAG. By contrast, Dell, HPE, and Supermicro tout both training and inference; e.g. Dell's CoreWeave clusters target heavy LLM training (particle.news), and Supermicro mentions "AI training clusters" in their IR (^[77] ir.supermicro.com).
- **Energy Efficiency:** Key claims –
- **Dell:** "Liquid-cooled GB200 NVL72 is up to 25× more efficient than an air-cooled H100 system" (cloud.watch.impress.co.jp). Also, Dell's IR7000 rack can run 480 kW with 100% heat capture (cloud.watch.impress.co.jp).
- **Lenovo:** "Lenovo Neptune cooling gives 40% power reduction and 3.5× thermal efficiency vs air cooling" (^[35] news.lenovo.com); "GB200 Superchip uses 45× less energy than previous gen (and lowers TCO 40×) for inference" (^[3] news.lenovo.com).
- **Supermicro:** Demonstrated **40% power savings** and use of 40–35°C inlet water reduces chiller load (^[4] ir.supermicro.com) (^[43] www.supermicro.com).
- **Cisco:** emphasizes low-latency RAG (no explicit energy stats), but its integration hints at efficiency by maintaining low idle-time latency.
- **VMware:** virtualization allows GPU sharing, which can improve utilization and cost. VMware notes that GPUs in VMs maintain performance, adding only management overhead (^[9] blogs.vmware.com) (so energy is similar to bare-metal usage plus virtualization overhead).
- **Nutanix:** suggests that by consolidating on fewer servers (density-optimized GPU systems) customers can reduce TCO and energy per inference (^[12] www.nutanix.com).
- **Memory and Data:** All systems now amass **terabytes of memory** at rack-level. Dell/Lenovo/SM talk ~20–40 TB of HBM per rack (news.rambler.ru) (chubutdigital.com.ar) (^[5] ir.supermicro.com), enabling massive models. Nutanix and VMware platforms emphasize memory virtualization and local model caching for throughput. Data services (NVMe pools, GPUDirect) are used to keep GPUs fed. Cisco's VAST/Insight accelerates data retrieval to and from GPU VRAM for RAG.
- **Connectivity:** Ultra-high bandwidth is common: **800 Gb/s** fabrics are standard (via NVIDIA Quantum InfiniBand or Cisco's equivalent) to link GPUs across servers. For example, Supermicro nodes include **8×ConnectX-8 (800 Gb)** on each baseboard (^[70] ir.supermicro.com). Cisco uses 800 Gb ActivFabric switches (Quantum-X800). Even Ethernet speeds are in the 400–800 Gbps range, reflecting the need for extreme data movement.

These data points illustrate that all offerings push the envelope of what's technically possible. The comparison shows little compromise on raw performance – instead, differentiation comes in form factor, ease of use, and integration.

Real-World Implementations and Case Examples

To understand practical implications, consider how these solutions are being adopted:

- **CoreWeave (Dell/Lenovo):** CoreWeave, an AI cloud provider, was reported to install the first Dell GB300 NVL72 racks (particle.news) (^[20] www.pcgamer.com). These racks (72 GPUs + 36 Grace CPUs) were deployed in conjunction with Dell's ACME/AI Factory solution, demonstrating that Dell's integration approach (pre-build, turnkey install) works at scale. CoreWeave cited this deployment to enhance its LLM training and real-time inference capacity. (Notably, PC Gamer and Particle News credited Dell with providing these racks (^[20] www.pcgamer.com) (particle.news), signifying industry confidence in Dell's solution.)
- **Yahsat (Nutanix):** In the Nutanix press release, Yahsat (a satellite communications operator) is quoted praising Nutanix GPT-in-a-Box for providing "simple, end-to-end management" of AI on-prem (^[78] www.nutanix.com). They use it to maintain data control and escalate generative AI pilot projects into production. This illustrates an enterprise use case where data sovereignty is critical, so a private suite like Nutanix's GPT-in-a-Box is chosen over public cloud. Nutanix's partnership with Hugging Face (enabling validated models) also targets regulated industries that need vetted AI components (^[79] www.nutanix.com).

- **Cisco + VAST (Agentic AI Pilot):** Cisco's RAG/agentic AI solution, announced Sep 2025, implies pilot deployments in financial or tech firms experimenting with AI assistants. By offering a validated Cisco+VAST bundle, customers can try enterprise-grade RAG pipelines on secured infrastructure. While not named publicly, use cases likely include legal or customer support knowledge bases, where Cisco's security (role-based controls, audit logs) and VAST's instant data access shine. Cisco's CEO quotes hint they are preparing for customers ready to "build the next generation of AI factories" with agents involved (^[80] [newsroom.cisco.com](#)).
- **VMware (Partner Adoption):** VMware's OEM Instagram suggests that partners like Dell and HPE are already bundling Private AI Foundation into their offerings. For example, Dell's PowerEdge servers can be sold with VMware's AI software stack pre-installed. This means enterprises already using Dell clusters or VMware Cloud Foundation can upgrade to Private AI Foundation with minimal new investment. Public references note big systems integrators aligning with this – e.g. Deloitte endorsing "NVIDIA AI Computing by HPE" solutions (^[29] [nvidianews.nvidia.com](#)) can indirectly benefit VMware's stack too.
- **Supermicro Clusters in Academia/HPC:** Though not a specific customer cited, Supermicro's building-block model is popular in research labs and cloud operators. Their "Server Building Block Solutions" are known for custom clusters (e.g. GWUs, Cambridge lab leases). The emphasis on *rapid modular deployment* suggests use by institutions that want to expand GPU capacity quickly. Brett Hamilton, CIO at a large university (hypothetical example), could order an 8-rack AI cluster from Supermicro to upgrade their supercomputer, relying on the turnkey integration to expedite installation.

Each case underscores different priorities: Dell/Lenovo for raw power and scale, Nutanix and VMware for integration with existing infrastructure, Cisco for secure data pipelines, etc. While few detailed third-party benchmarks are public yet, industry sources uniformly note the **unprecedented scale** (exaflop, 50× content generation capacity ([particle.news](#))) and **speed-ups** (tens to hundreds of times faster processing) offered by these new systems.

Implications and Future Directions

Market Growth: The on-prem AI infrastructure market has grown dramatically. IDC reports that organizations increased spending on compute and storage hardware for AI by 166% year-over-year in Q2 2025, with AI infrastructure spending projected to reach \$758 billion by 2029. Gartner estimates worldwide AI spending totaled \$1.5 trillion in 2025 and projects it to exceed \$2 trillion in 2026, with datacenter systems spending rising 46.8% to \$489.5 billion in 2025 alone. Notably, inference workloads (driven by test-time compute and reasoning models) are now the primary demand driver, surpassing training as the dominant AI infrastructure use case.

NVIDIA Roadmap – Blackwell Ultra, Rubin, and Beyond: At GTC 2025 (March 2025), NVIDIA announced the **Blackwell Ultra (GB300)** architecture, which entered production and began shipping in September 2025. The GB300 NVL72 rack delivers 1,100 petaFLOPS of FP4 inference and 360 petaFLOPS of FP8 training—a 1.5× improvement over GB200 NVL72. Looking ahead, NVIDIA confirmed the **Vera Rubin NVL144** (88 Vera CPUs + 144 Rubin GPUs via NVLink6) for the second half of 2026, delivering 3.6 exaFLOPS FP4—a 3.3× leap over GB300 NVL72. **Rubin Ultra** (up to 576 GPUs per rack, 15 exaFLOPS FP4) is slated for 2H 2027, with a subsequent architecture named after Richard Feynman on the longer-term roadmap (^[81] [nvidianews.nvidia.com](#)).

Multi-Vendor Ecosystem: All vendors commit to supporting NVIDIA's next GPU and CPU releases. For instance, HPE explicitly stated it will be "time-to-market" for the upcoming GB200 NVL72, NVL2 devices, *and* next-gen architectures (Blackwell Ultra, Rubin, Vera) (^[13] [nvidianews.nvidia.com](#)). Dell is already delivering GB300 NVL72; presumably GB400/GB500 chips (if NVIDIA releases them) will follow. Lenovo, Supermicro, and Cisco also cultivate close ties with NVIDIA roadmaps. This ensures a relatively unified hardware evolution – enterprises need only watch one ecosystem's announcements to know that solutions will upgrade in lockstep.

Sustainability: Energy efficiency remains a key challenge. These vendors' claims (25–40× efficiency gains, up to 40% cooling savings) are striking. In practice, customers will weigh the trade-offs of deploying massive liquid-cooled racks (with high upfront cost) to gain long-term OPEX reduction. We anticipate further innovation in cooling (e.g. warm-water reuse, immersion cooling). Also, secondary markets like AMD GPU servers or accentuating FPGA/DPU offload (to reduce GPU load) may emerge as complementary strategies – though current products are NVIDIA-centric.

Software and AI Workflow: While hardware is highlighted, software integration differentiates these offerings. HPE and VMware emphasize private cloud UX and lifecycle, making them easier for large IT teams to adopt. Nutanix's focus on simple UI (catalog wizards) lowers the barrier for non-experts. Cisco's push shows enterprises increasingly see data management (set for RAG) as inseparable from AI compute. We expect future clouds/platforms to integrate even more AI-specific services (e.g. vector DB as service, GPU orchestration with autoscaling). All vendors will likely enhance support for MLOps pipelines and interop (e.g. Kubernetes + GPU, new NIM microservices).

Use Case Evolution: As workloads diversify (from LLMs to vision or multimodal AI), infrastructure might fragment. So far, the offerings skew toward LLMs/HPC, but many enterprises will run diverse models. Platforms that can flexibly support GPUs for video, GANs, simulations will have an edge. Also, edge AI (with smaller GPUs like Jetson or Orin) could be integrated with some of these stacks (e.g. Nutanix hints at "edge to cloud"). Hybrid architectures (some processing on local rack, some burst to cloud) will probably evolve.

Competitive Landscape: These solutions compete with public cloud (AWS Sagemaker, GCP Vertex) and AI-focused clouds (Lambda, CoreWeave). On-prem grids may need to justify cost by data residency or performance. Partnerships and SI integrations will be crucial to capture market share. We see incumbents (Dell, HPE) leveraging legacy strengths (mass production, services), while niche players (Supermicro, Nutanix) emphasize agility and integration, and new entrants (Cisco, VMware) leverage their networking/virtualization dominance.

Standardization and Interoperability: One subtle benefit is standardization around NVIDIA reference architectures (NVL72, NVIDIA MGX, etc.). All major players support MGX modules and NVLink fabrics, which means an organization can mix and match OEMs if needed. Interoperability (e.g. any vendor's tensor charger GPU works anywhere) will spur ecosystem growth. We expect cross-vendor benchmarks (e.g. SPEC, MLPerf) to emerge, comparing these solutions on standardized loads.

- [20] <https://www.pcgamer.com/hardware/racks-packing-nvidias-newst-and-shiniest-ai-supercomputer-blackwell-ultra-cards-have-just-been-deployed-by-coreweave/#:~:til...>
- [21] <https://www.pcgamer.com/hardware/racks-packing-nvidias-newst-and-shiniest-ai-supercomputer-blackwell-ultra-cards-have-just-been-deployed-by-coreweave/#:~:den...>
- [22] <https://www.dell.com/zh-hk/blog/poweredge-server-and-networking-announcements-at-nvidia-gtc-2025/#:~:with...>
- [23] <https://www.dell.com/zh-hk/blog/poweredge-server-and-networking-announcements-at-nvidia-gtc-2025/#:~:8...>
- [24] <https://www.reuters.com/business/dell-unveils-new-ai-servers-powered-by-nvidia-chips-boost-enterprise-adoption-2025-05-19/#:~:On%20...>
- [25] <https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases\~usa\~2025\~11\~dell-technologies-accelerates-enterprise-ai-with-powerful-automated-solutions.htm>
- [26] <https://nvidianews.nvidia.com/news/hpe-nvidia-ai-computing-generative-ai#:~:inclu...>
- [27] https://www.hpe.com/us/en/newsroom/press-release/2024/06/hewlett-packard-enterprise-and-nvidia-announce-nvidia-ai-computing-by-hpe-to-accelerate-generative-ai-industrial-revolution.html?jumpid=os_qck9zj9oo_aid-521070646#:~:Among...
- [28] https://www.hpe.com/us/en/newsroom/press-release/2024/06/hewlett-packard-enterprise-and-nvidia-announce-nvidia-ai-computing-by-hpe-to-accelerate-generative-ai-industrial-revolution.html?jumpid=os_qck9zj9oo_aid-521070646#:~:LAS%2...
- [29] <https://nvidianews.nvidia.com/news/hpe-nvidia-ai-computing-generative-ai#:~:All%2...>
- [30] <https://nvidianews.nvidia.com/news/hpe-nvidia-ai-computing-generative-ai#:~:HPE%2...>
- [31] <https://nvidianews.nvidia.com/news/hpe-nvidia-ai-computing-generative-ai#:~:Avail...>
- [32] <https://nvidianews.nvidia.com/news/hpe-nvidia-ai-computing-generative-ai#:~:HPE%2...>
- [33] <https://www.hpe.com/us/en/newsroom/press-release/2025/10/hpe-advances-government-and-enterprise-ai-adoption-through-secure-ai-factory-innovations-with-nvidia.html>
- [34] <https://www.hpe.com/us/en/newsroom/press-release/2025/12/hpe-shapes-future-of-hybrid-cloud-with-innovations-across-virtualization-security-and-ai.html>
- [35] <https://news.lenovo.com/pressroom/press-releases/lenovo-unveils-hybrid-ai-solutions-delivering-power-of-generative-ai-to-enterprises-with-nvidia/#:~:With%...>
- [36] <https://news.lenovo.com/pressroom/press-releases/lenovo-unveils-hybrid-ai-solutions-delivering-power-of-generative-ai-to-enterprises-with-nvidia/#:~:As%20...>
- [37] <https://news.lenovo.com/pressroom/press-releases/lenovo-unveils-hybrid-ai-solutions-delivering-power-of-generative-ai-to-enterprises-with-nvidia/#:~:desi...>
- [38] <https://news.lenovo.com/pressroom/press-releases/lenovo-unveils-hybrid-ai-solutions-delivering-power-of-generative-ai-to-enterprises-with-nvidia/#:~:%E2%8...>
- [39] <https://news.lenovo.com/pressroom/press-releases/lenovo-unveils-hybrid-ai-solutions-delivering-power-of-generative-ai-to-enterprises-with-nvidia/#:~:Lenov...>
- [40] <https://news.lenovo.com/pressroom/press-releases/lenovo-revolutionizes-real-time-enterprise-ai-with-new-inferencing-servers/>
- [41] <https://news.lenovo.com/pressroom/press-releases/hybrid-ai-personalized-perceptive-proactive-ai-portfolio-tech-world-ces-2026/>
- [42] <https://ir.supermicro.com/news/news-details/2025/Supermicro-Adds-Portfolio-for-Next-Wave-of-AI-with-NVIDIA-Blackwell-Ultra-Solutions-Featuring-NVIDIA-HGX-B300-NVL16-and-GB300-NVL72/default.aspx#:~:Super...>
- [43] <https://www.supermicro.com/en/accelerators/nvidia/launchpad#:~:memor...>
- [44] <https://ir.supermicro.com/news/news-details/2025/Supermicro-Adds-Portfolio-for-Next-Wave-of-AI-with-NVIDIA-Blackwell-Ultra-Solutions-Featuring-NVIDIA-HGX-B300-NVL16-and-GB300-NVL72/default.aspx#:~:whil...>
- [45] <https://ir.supermicro.com/news/news-details/2024/Supermicro-Introduces-Rack-Scale-Plug-and-Play-Liquid-Cooled-AI-SuperClusters-for-NVIDIA-Blackwell-and-NVIDIA-HGX-H100H200---Radical-Innovations-in-the-AI-Era-to-Make-Liquid-Cooling-Free-with-a-Bonus/default.aspx#:~:From...>
- [46] <https://www.supermicro.com/en/products/system/gpu/48u/srs-gb200-nv172#:~:%2A%2...>
- [47] <https://www.supermicro.com/en/accelerators/nvidia/launchpad#:~:Vast%...>
- [48] <https://ir.supermicro.com/news/news-details/2024/Supermicro-Introduces-Rack-Scale-Plug-and-Play-Liquid-Cooled-AI-SuperClusters-for-NVIDIA-Blackwell-and-NVIDIA-HGX-H100H200---Radical-Innovations-in-the-AI-Era-to-Make-Liquid-Cooling-Free-with-a-Bonus/default.aspx#:~:save...>
- [49] <https://ir.supermicro.com/news/news-details/2024/Supermicro-Introduces-Rack-Scale-Plug-and-Play-Liquid-Cooled-AI-SuperClusters-for-NVIDIA-Blackwell-and-NVIDIA-HGX-H100H200---Radical-Innovations-in-the-AI-Era-to-Make-Liquid-Cooling-Free-with-a-Bonus/default.aspx#:~:Super...>
- [50] <https://ir.supermicro.com/news/news-details/2025/Supermicro-Begins-Volume-Shipments-of-NVIDIA-Blackwell-Ultra-Systems-and-Rack-Plug-and-Play-Data-Center-Scale-Solutions/default.aspx>
- [51] <https://www.cnbc.com/2025/02/26/super-micro-computer-surges-20percent-after-filing-delayed-financials.html>
- [52] <https://newsroom.cisco.com/content/tr/newsroom/en/us/a/y2025/m09/cisco-secure-ai-factory-with-nvidia-unlocks-enterprise-data-for-agentic-ai.html#:~:Cisco...>
- [53] <https://newsroom.cisco.com/content/tr/newsroom/en/us/a/y2025/m09/cisco-secure-ai-factory-with-nvidia-unlocks-enterprise-data-for-agentic-ai.html#:~:NVIDI...>

- [54] <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2024/m06/cisco-reveals-nexus-hyperfabric-the-new-generative-ai-infrastructure-solution-with-nvidia-to-help-simplify-data-center-operations.html?dtid=oblgzz000659#:~:Deliv...>
- [55] <https://blogs.nvidia.com/blog/cisco-nexus-hyperfabric-nim/#:~:Cisco...>
- [56] <https://blogs.nvidia.com/blog/cisco-nexus-hyperfabric-nim/#:~:The%2...>
- [57] <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2024/m06/cisco-reveals-nexus-hyperfabric-the-new-generative-ai-infrastructure-solution-with-nvidia-to-help-simplify-data-center-operations.html?dtid=oblgzz000659#:~:%E2%8...>
- [58] <https://newsroom.cisco.com/content/tr/newsroom/en/us/a/y2025/m09/cisco-secure-ai-factory-with-nvidia-unlocks-enterprise-data-for-agentic-ai.html#:~:Agent...>
- [59] <https://blogs.vmware.com/cloud-foundation/2024/03/18/announcing-initial-availability-of-vmware-private-ai-foundation-with-nvidia/#:~:with%...>
- [60] <https://blogs.vmware.com/cloud-foundation/2024/03/18/announcing-initial-availability-of-vmware-private-ai-foundation-with-nvidia/#:~:advan...>
- [61] <https://blogs.vmware.com/cloud-foundation/2024/03/18/announcing-initial-availability-of-vmware-private-ai-foundation-with-nvidia/#:~:deplo...>
- [62] <https://news.broadcom.com/releases/vmware-explore-2025-vmware-cloud-foundation-ai-native>
- [63] <https://www.nutanix.com/press-releases/2024/nutanix-accelerates-enterprise-adoption-of-generative-ai#:~:%E2%8...>
- [64] <https://www.nutanix.com/blog/gpt-in-a-box-2-is-here#:~:For%2...>
- [65] <https://www.nutanix.com/press-releases/2024/nutanix-accelerates-enterprise-adoption-of-generative-ai#:~:Stren...>
- [66] <https://www.nutanix.com/blog/gpt-in-a-box-2-is-here#:~:When%...>
- [67] <https://www.nutanix.com/blog/gpt-in-a-box-2-is-here#:~:dense...>
- [68] <https://www.nutanix.com/press-releases/2024/nutanix-accelerates-enterprise-adoption-of-generative-ai#:~:Compa...>
- [69] <https://www.nutanix.com/press-releases/2025/nutanix-enables-agentic-ai-anywhere>
- [70] <https://ir.supermicro.com/news/news-details/2025/Supermicro-Adds-Portfolio-for-Next-Wave-of-AI-with-NVIDIA-Blackwell-Ultra-Solutions-Featuring-NVIDIA-HGX-B300-NVL16-and-GB300-NVL72/default.aspx#:~:For%2...>
- [71] <https://invidianews.nvidia.com/news/hpe-nvidia-ai-computing-generative-ai#:~:Green...>
- [72] <https://news.lenovo.com/pressroom/press-releases/lenovo-unveils-hybrid-ai-solutions-delivering-power-of-generative-ai-to-enterprises-with-nvidia/#:~:marke...>
- [73] <https://newsroom.cisco.com/content/tr/newsroom/en/us/a/y2025/m09/cisco-secure-ai-factory-with-nvidia-unlocks-enterprise-data-for-agentic-ai.html#:~:RTX%2...>
- [74] <https://ir.supermicro.com/news/news-details/2024/Supermicro-Introduces-Rack-Scale-Plug-and-Play-Liquid-Cooled-AI-SuperClusters-for-NVIDIA-Blackwell-and-NVIDIA-HGX-H100H200---Radical-Innovations-in-the-AI-Era-to-Make-Liquid-Cooling-Free-with-a-Bonus/default.aspx#:~:SAN%2...>
- [75] <https://www.nutanix.com/blog/gpt-in-a-box-2-is-here#:~:Nutan...>
- [76] <https://www.nutanix.com/press-releases/2024/nutanix-accelerates-enterprise-adoption-of-generative-ai#:~:Nutan...>
- [77] <https://ir.supermicro.com/news/news-details/2025/Supermicro-Adds-Portfolio-for-Next-Wave-of-AI-with-NVIDIA-Blackwell-Ultra-Solutions-Featuring-NVIDIA-HGX-B300-NVL16-and-GB300-NVL72/default.aspx#:~:stren...>
- [78] <https://www.nutanix.com/press-releases/2024/nutanix-accelerates-enterprise-adoption-of-generative-ai#:~:Edge...>
- [79] <https://www.nutanix.com/press-releases/2024/nutanix-accelerates-enterprise-adoption-of-generative-ai#:~:Partn...>
- [80] <https://newsroom.cisco.com/content/tr/newsroom/en/us/a/y2025/m09/cisco-secure-ai-factory-with-nvidia-unlocks-enterprise-data-for-agentic-ai.html#:~:%E2%8...>
- [81] <https://invidianews.nvidia.com/news/nvidia-blackwell-ultra-ai-factory-platform-paves-way-for-age-of-ai-reasoning>

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at <https://intuitionlabs.ai/contact> for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by [Adrien Laurent](#), a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.

© 2025 IntuitionLabs.ai. All rights reserved.