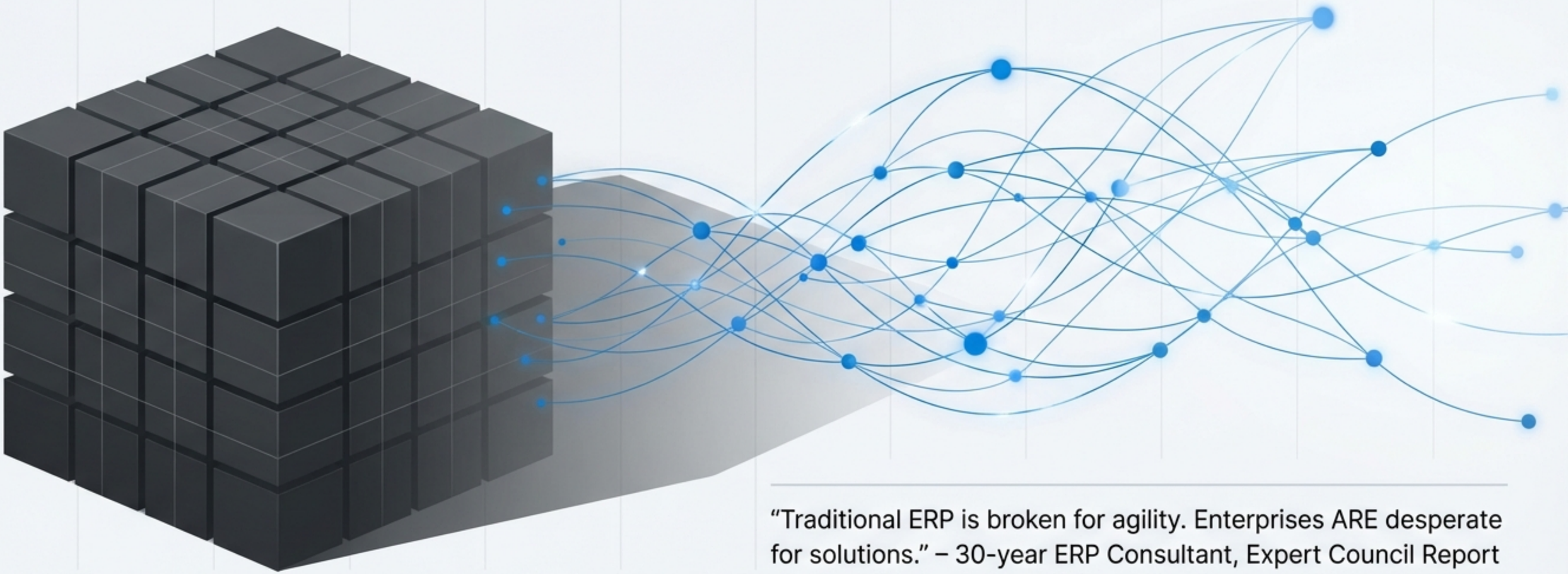
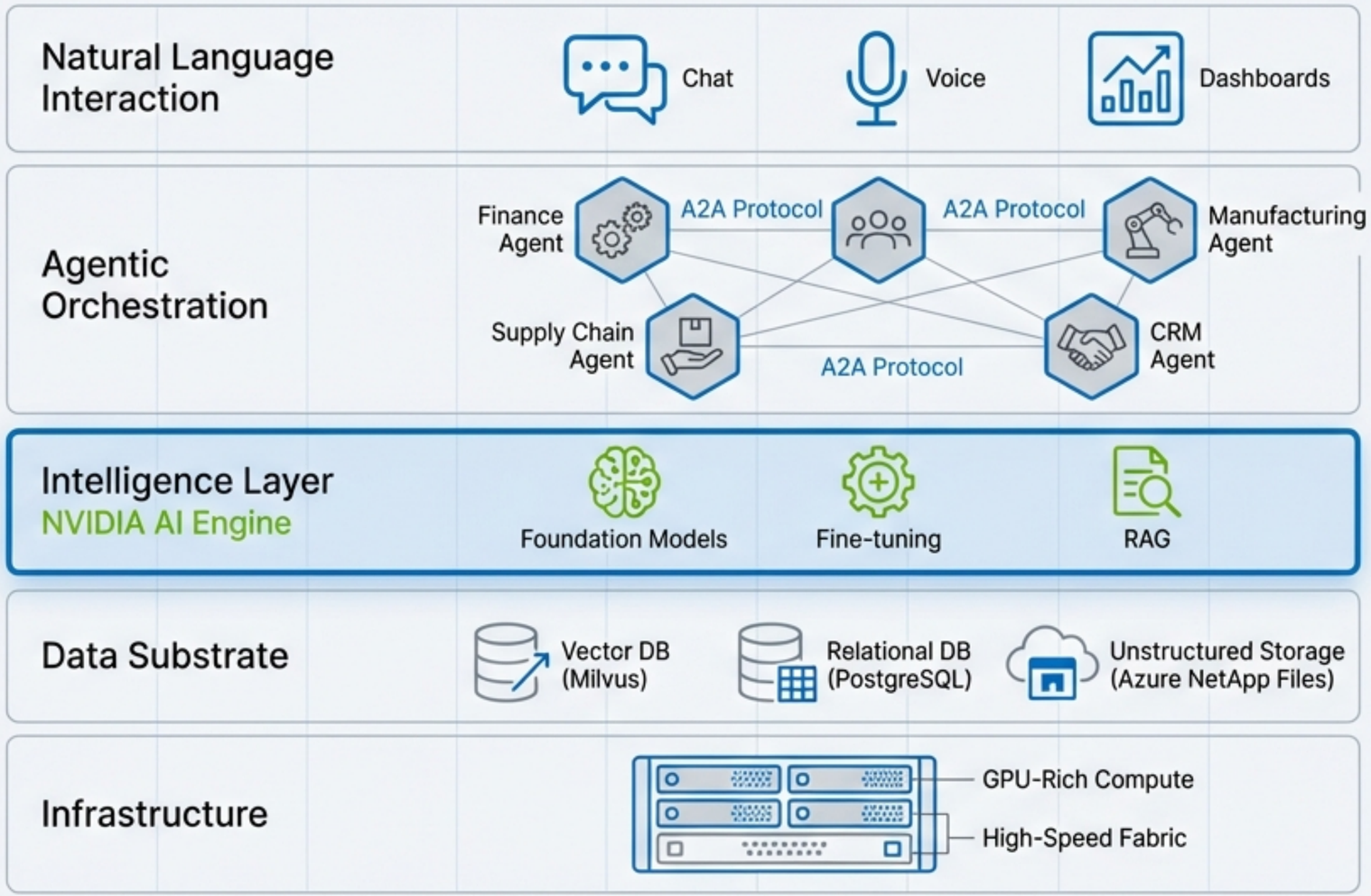
The Paradigm Shift in Enterprise Operations is Here.

Traditional, monolithic enterprise systems **are failing**. **They are slow**, **rigid**, and **incapable** of meeting the demands of a dynamic business environment. The future **is not an iteration; it is a replacement**—a move from static applications to a **responsive, intelligent organism of AI agents** that operate, adapt, and learn in real time.



"Traditional ERP is broken for agility. Enterprises ARE desperate for solutions." – 30-year ERP Consultant, Expert Council Report

The Agent-Native Enterprise: A Conceptual Blueprint



This is not a future dream; it is an architectural pattern. Autonomous agents form the new application layer, interacting with a unified intelligence and data substrate to execute complex business workflows.

The Vision is Validated: Market Demand Meets Technical Feasibility

1. The Market Problem is Acute

"Enterprises ARE desperate for solutions."

	Pilot Interest
Finance: **Critical Pain**	Pilot in 6 months, $500K-$1M/year.
Retail Supply Chain:	**Critical Pain**, Pilot in 6 months, ROI: $2-3M/year.
Healthcare Revenue Cycle:	**Critical Pain**, Pilot in 9 months, ROI: $2-3M/year.

2. The Technology is Feasible Today

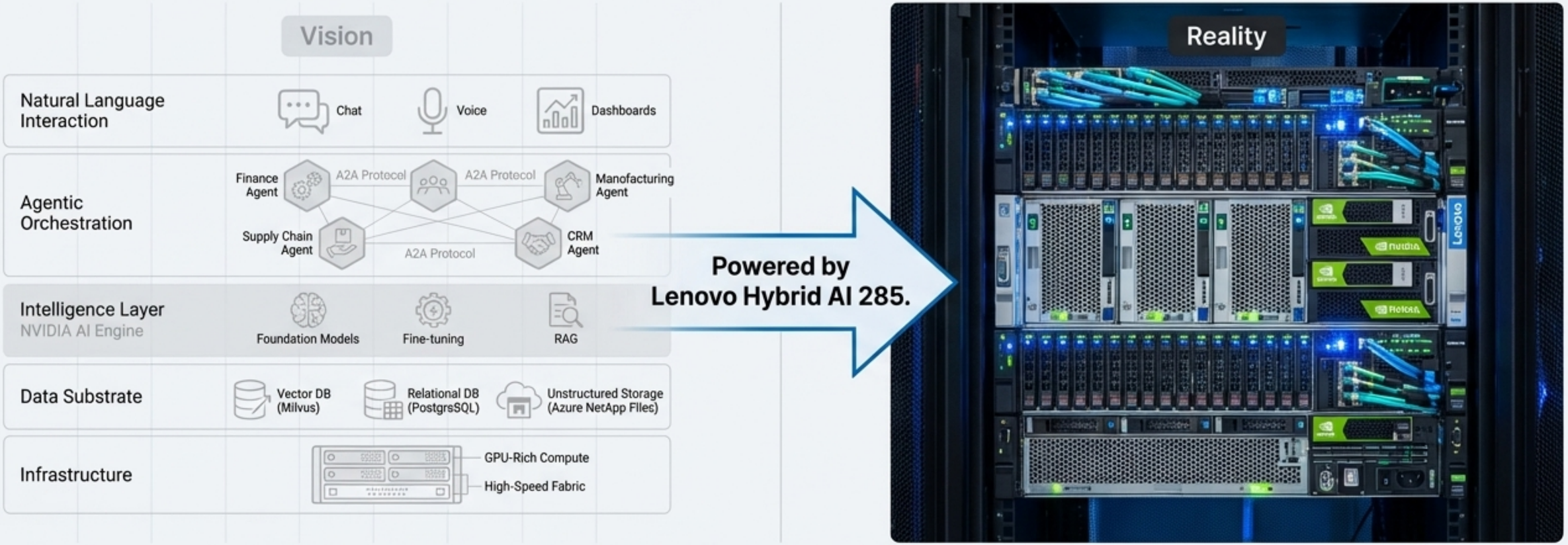
"From a pure technical perspective: **Yes, this is feasible. Today. In 2025."** – Dr. Sarah Chen, Former NVIDIA VP, AI Infrastructure

✓ Llama-3.1-70B (Production-ready)
✓ NVIDIA NIM (Enterprise-ready)
✓ RAG (Solved)

3. Customers Are Eager to Pilot

"I'd pilot AP automation. Investment: $500K. Expected ROI: $500K-$1M/year." – Fortune 500 CIO

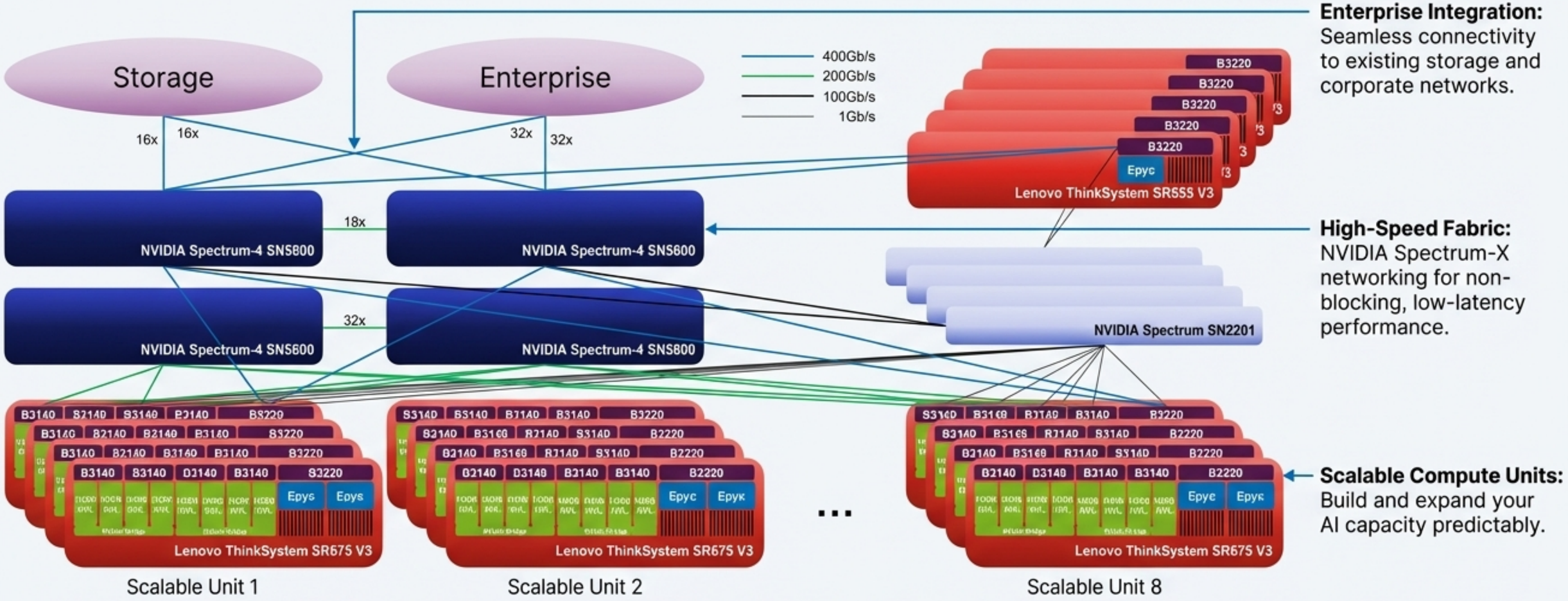
From Vision to Reality: The Foundation for the Agent-Native Enterprise.



A visionary architecture requires an engineered foundation. The challenges of deploying enterprise-grade AI——integrating GPU compute, high-performance networking, and a sophisticated software stack—are significant.

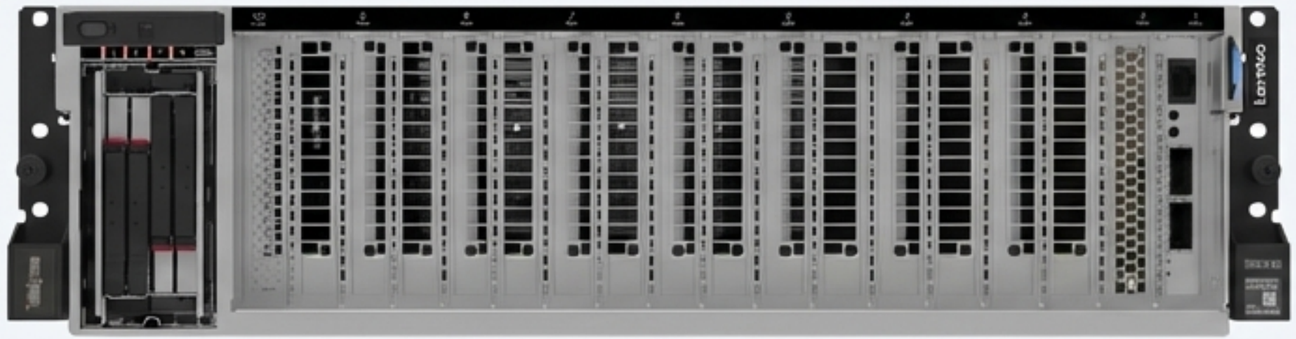
The **Lenovo Hybrid AI 285 platform** is the pre-validated, performance-tuned infrastructure that makes the agent-native vision achievable. It is the exact match for the NVIDIA AI Enterprise Reference Architecture, designed to run the most demanding AI workloads at scale.

The Anatomy of an Enterprise AI Factory.



The Lenovo Hybrid AI 285 platform is an integrated solution, combining GPU-rich compute, advanced networking, and management nodes into a cohesive, scalable AI factory.

The Building Block: The ‘2-8-5’ AI Compute Node.



The ‘2-8-5’ DNA

2x CPUs

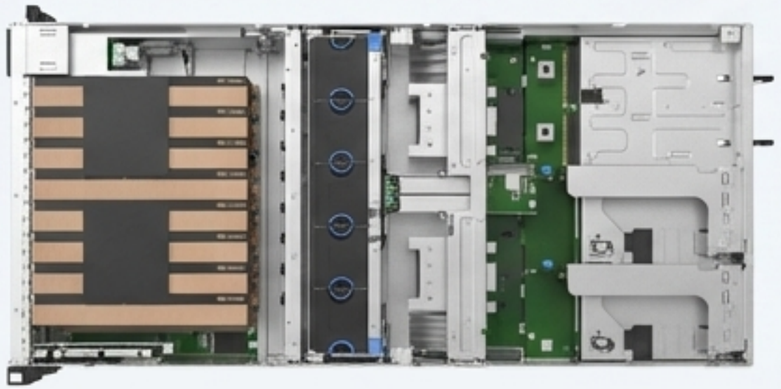
5th Gen AMD EPYC 9005 processors.

8x GPUs

Support for 8 double-width PCIe GPUs, ideal for inference and training.

5x Network Adapters

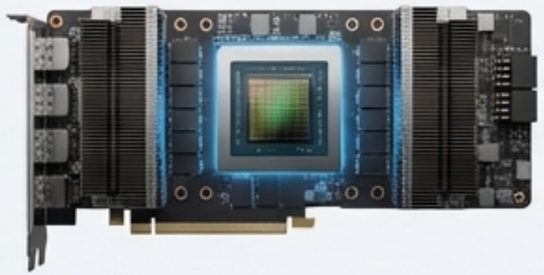
High-bandwidth connectivity for both internal (East-West) and external (North-South) traffic.



The heart of the platform is the **Lenovo ThinkSystem SR675 V3**, a 3U GPU-rich server.

Its density and PCIe-optimized design make it the ideal choice for NVIDIA’s demanding 2-8-5 configuration requirement, delivering maximum AI performance in a compact footprint.

Pillar 1: Unprecedented Compute for Generative AI & HPC



NVIDIA H200 NVL

Accelerates the largest Generative AI and HPC workloads.

141GB
of ultra-fast HBM3e memory per GPU.

4.8 TB/s
of memory bandwidth.

Includes a 5-year license for **NVIDIA AI Enterprise** at no extra charge.

NVIDIA RTX PRO 6000 Blackwell Server Edition

A powerful combination of AI and visual computing capabilities for the data center.

96GB
of ultra-fast GDDR7 memory.

Agentic AI, physical AI, scientific computing, rendering, and 3D graphics.

Pillar 2: The High-Performance Network Fabric.

Converged (North-South) Fabric

Purpose: Handles storage access, in-band management, and connects the AI platform to the enterprise IT environment.

Technology: Built on Ethernet with RDMA over Converged Ethernet (RoCE).

Performance: Guarantees a minimum of **25Gb/s** bandwidth per GPU to the enterprise network.

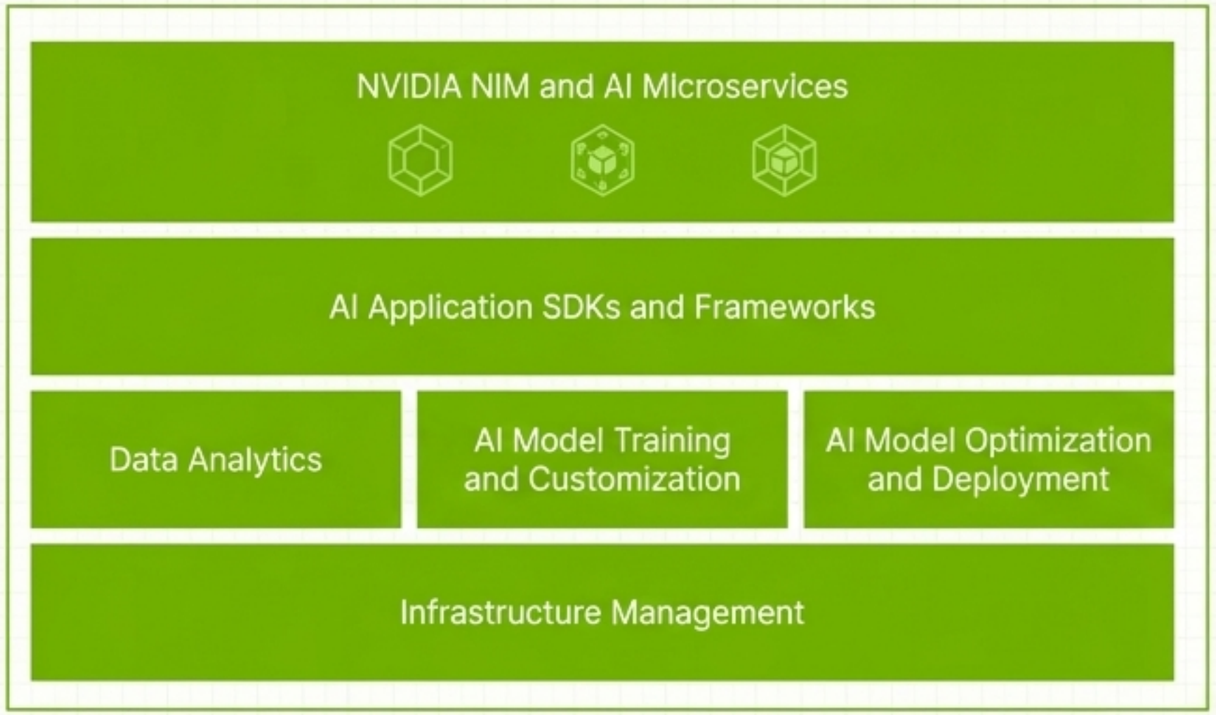
Compute (East-West) Fabric

Purpose: Crucial for training and fine-tuning, enabling direct, low-latency communication between GPUs across nodes.

Technology: Utilizes NVIDIA BlueField-3 DPUs and Spectrum-4 switches to create the **NVIDIA Spectrum-X** platform.

Performance: Delivers up to **800Gb/s** of non-blocking performance for AI workloads.

Pillar 3: The Software Stack for Enterprise AI Deployment



NVIDIA AI Enterprise

Infrastructure Management

NVIDIA BaseCommand Manager provisions and manages the entire AI environment, from the OS to Kubernetes and GPU operators.

AI Frameworks & SDKs

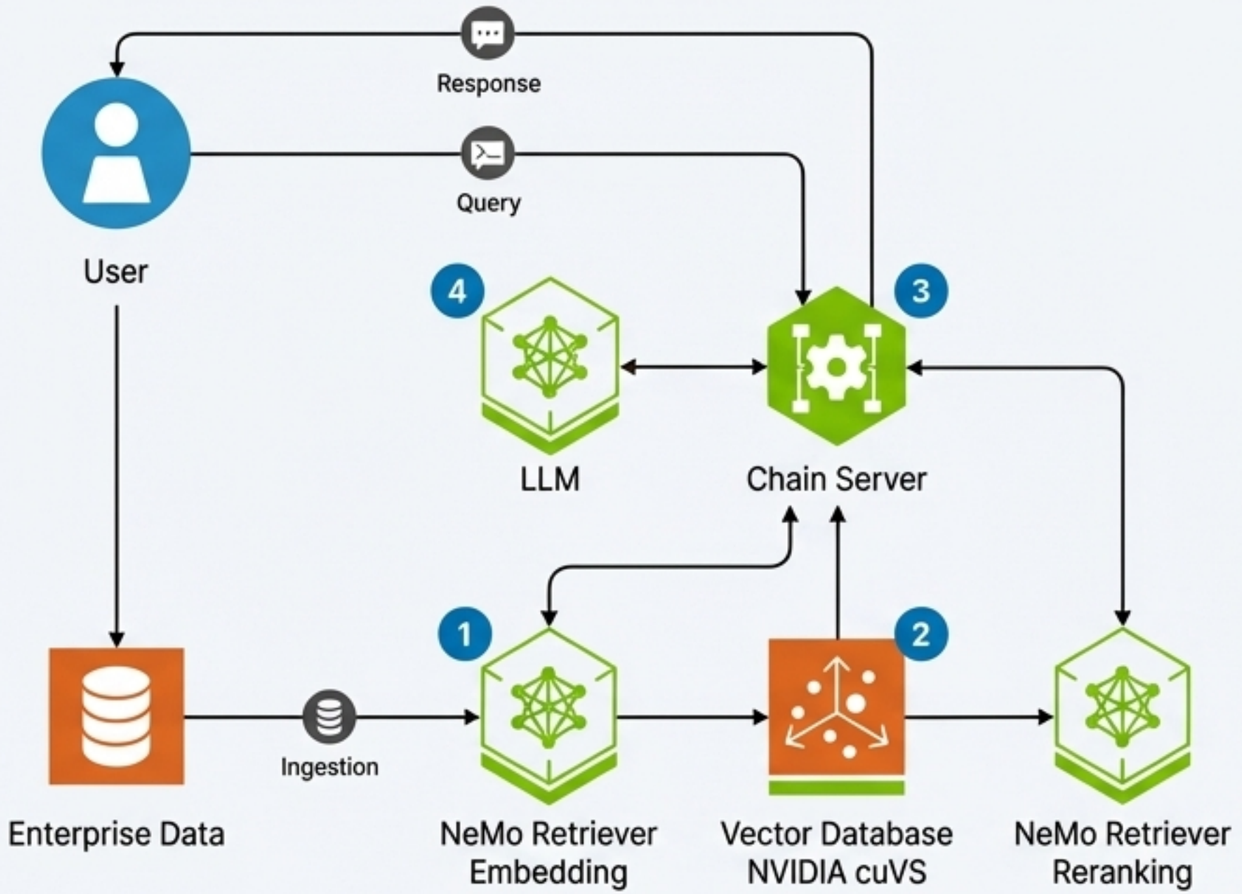
Provides access to optimized frameworks like NvIDIA NeMo (for building LLMs), RAPIDS (for data science), and TensorRT (for inference).

NVIDIA NIM & AI Microservices

The crucial top layer. A catalog of pre-built, containerized, and performance-optimized microservices that dramatically simplify the deployment of AI models for use cases like RAG, computer vision, and speech.

This is a comprehensive software suite that moves you from bare metal to a fully functional AI development and deployment environment.

A Practical Application: Powering the Enterprise RAG RAG Pipeline.



1. **Ingestion:** Enterprise data is processed by **NeMo Retriever Embedding** to create vector embeddings.
2. **Storage:** These embeddings are stored in a **Vector Database** (e.g., Milvus).
3. **Retrieval:** When a user makes a query, **NeMo Retriever** finds the most relevant data from the vector database.
4. **Generation:** The LLM, augmented with this retrieved context, generates an accurate, grounded, and relevant response.

In its simplest form, a RAG pipeline requires a minimum of three NVIDIA Inference Microservices (Retriever, Reranker, LLM). This platform is built to run them efficiently.

Your Path to Adoption Starts Here: Entry Sizing.

The ideal starting point for development, application trials, or small-scale use cases. This approach reduces initial hardware cost, control plane overhead, and networking complexity.

Option 1: Single Node

Compute
1x Lenovo ThinkSystem SR675 V3

Capacity
4-8 GPUs

Networking
Connects to your existing data center network.

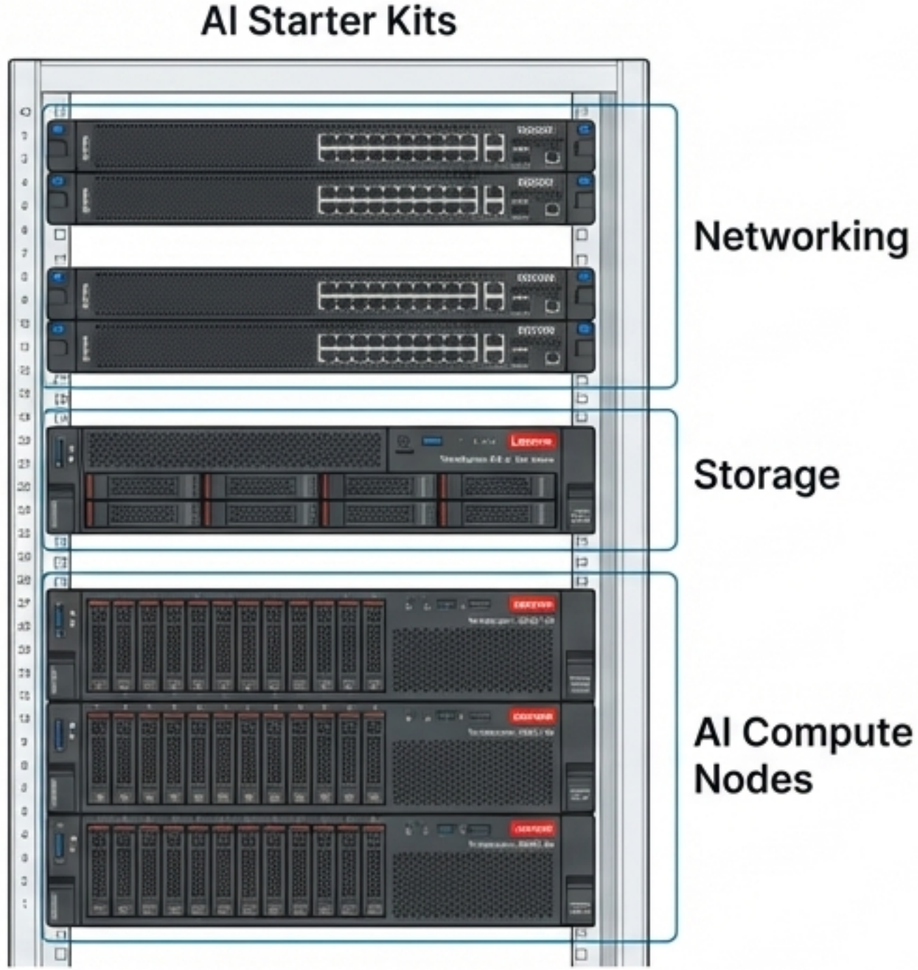
Option 2: Two Nodes

Compute
2x Lenovo ThinkSystem SR675 V3

Capacity
8-16 GPUs

Networking
Can be directly connected without external switches for maximum simplicity.

Scaling Up: The AI Starter Kit.



Use Case

For customers who require integrated, high-performance storage and networking but do not plan to scale beyond 24 GPUs in the near future. This is an end-to-end solution for a dedicated AI environment.

Components of the Kit

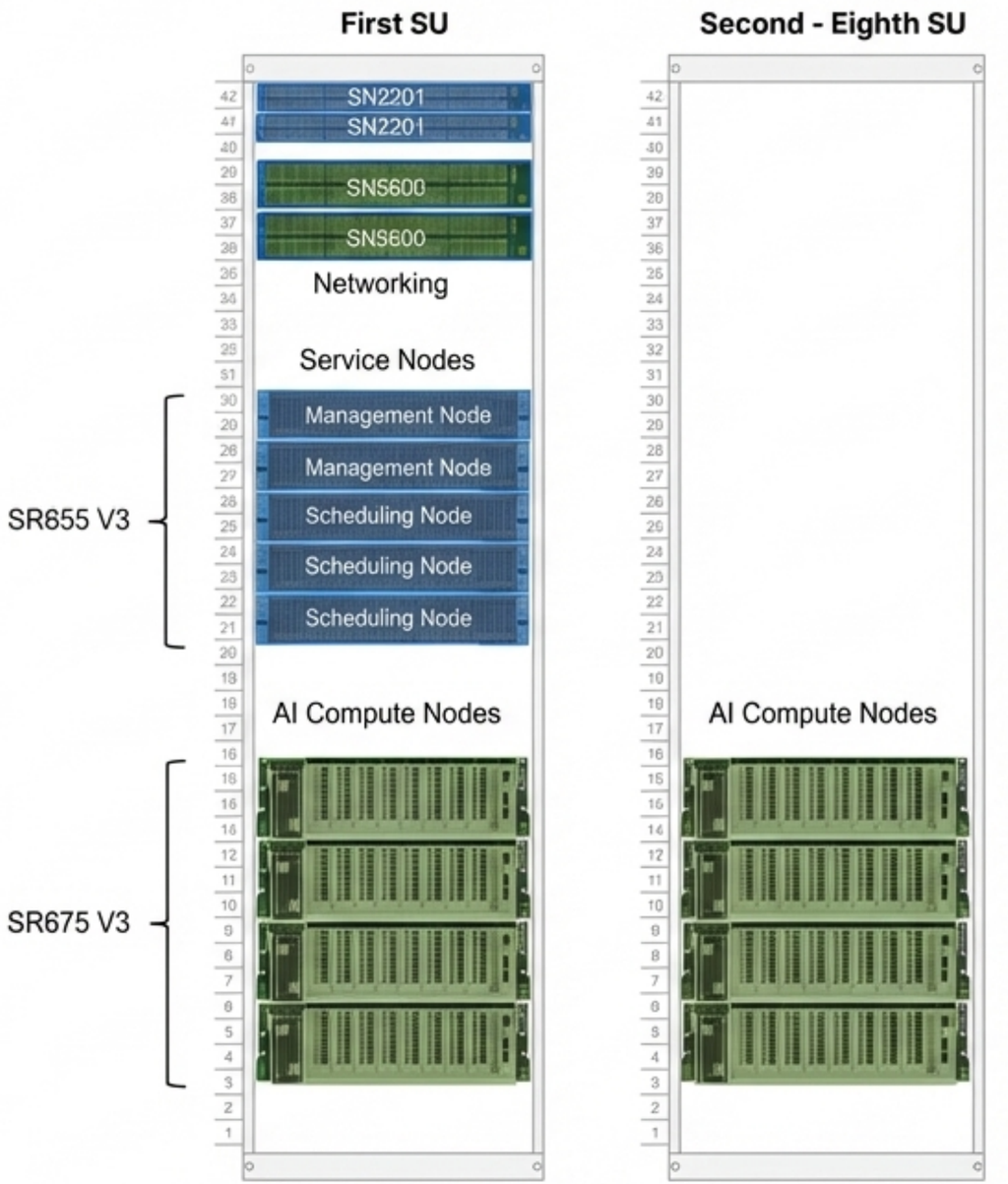
- **AI Compute Nodes:** 1-3x Lenovo ThinkSystem SR675 V3 servers (up to 24 GPUs).
- **Storage:** Integrated Lenovo ThinkSystem DM or DG Series All-Flash Storage arrays.
- **Networking:**
  - NVIDIA SN3700 200GbE switches for compute fabric.
  - NVIDIA SN2201 switches for management.
- Available in High-Availability (HA) and non-HA options.

The Blueprint for Growth: The Scalable Unit Deployment.

For deployments beyond the starter kit, the Scalable Unit provides a robust foundation for unlimited growth. New units can be added seamlessly and without downtime to expand capacity as use cases grow.

Composition of a Scalable Unit (SU):

- **AI Compute Nodes**
  Up to 4x ThinkSystem SR675 V3 servers.
- **Service Nodes**
  The first SU includes 5x ThinkSystem SR655 V3 servers to manage the cluster (2 Management, 3 Scheduling for Kubernetes).
- **Networking**
  NVIDIA SN5600 Spectrum-4 switches, enabling the full power of the Spectrum-X platform for multi-node training and inference.



De-Risking Your AI Journey with the Lenovo EveryScale Solution

Lenovo leverages its world-leading expertise in High Performance Computing (from Exascale to EveryScale™) to deliver **enterprise-class AI factories** that are **ready for immediate use.**

Best Recipe Guides

Warrants interoperability of all hardware, software, and firmware, eliminating guesswork and integration failures.

Pre-Integrated Delivery

Systems are fully pre-built, pre-cabled, and pre-loaded with the “Best Recipe” stack in Lenovo manufacturing.

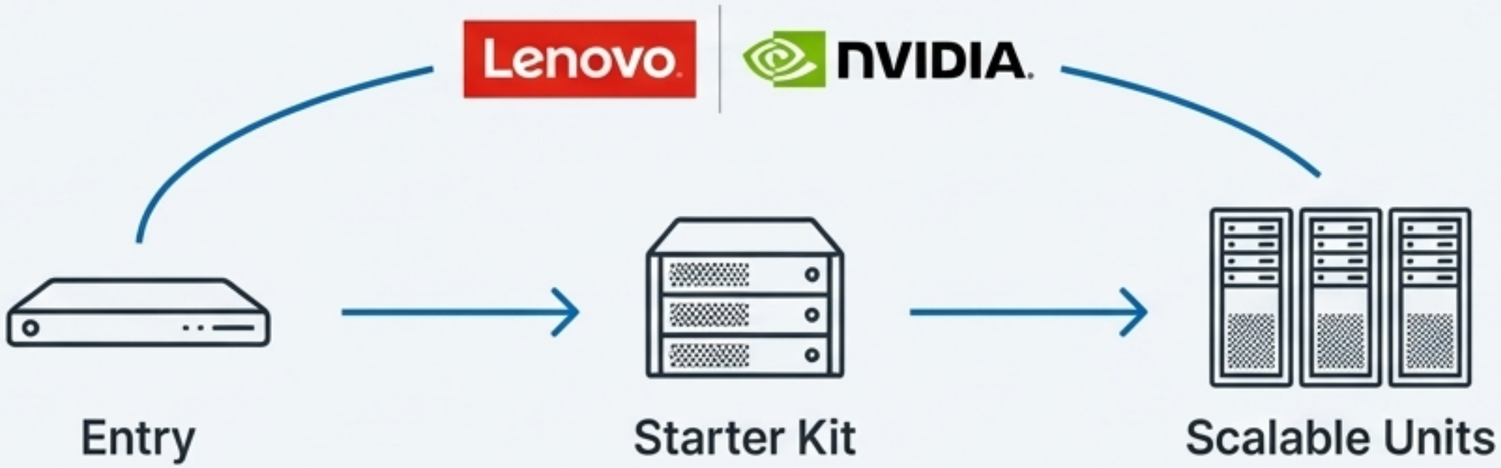
Rack-Level Testing

Ensures a reliable, turnkey delivery that minimizes installation time in your data center.

“Following their excellent experience with Lenovo on Omniverse, NVIDIA has once again chosen Lenovo technology as the foundation for the development and test of their NVIDIA AI Enterprise Reference Architecture (ERA).”

The Future of the Enterprise is Not a Vision. It’s a Blueprint.

The promise of an intelligent, agent-driven enterprise is now within reach. The journey begins not with abstract strategies, but with a concrete, engineered foundation.



A Clear Path

Start small with entry sizing, prove value with a starter kit, and expand predictably with scalable units.

An Engineered Foundation

A purpose-built platform with best-in-class Compute, Networking, and Software, perfectly aligned with NVIDIA’s reference architecture.

A De-Risked Deployment

Delivered as a fully integrated and validated solution through the Lenovo EveryScale framework.

You have the vision. We provide the blueprint to build it.