



Own Your Own Intelligence: The Lenovo Sovereign AI Inference Solution with IBM and NVIDIA

Solution Brief

Enterprises are moving AI assistants, research, code generation, and multimodal analysis from experimentation into production. Cloud services can suit proofs of concept and variable demand, but sustained use can increase requirements for data sovereignty, predictable latency, model control, capacity, and cost visibility.

The Lenovo Sovereign AI Inference architecture addresses these requirements by keeping inference within the enterprise boundary, controlling model and data placement, and scaling owned infrastructure as demand grows. This Solution Brief describes the use cases, components, benefits, and deployment options.

The Sovereign AI Inference Use Case

AI sovereignty has four dimensions:

- **Data sovereignty** — Prompts, context, embeddings, and outputs remain within the organization's boundary and jurisdiction, including deployments without external inference endpoints.
- **Model sovereignty** — Enterprises select, fine-tune, version-pin, and update open-weight models on their own schedule.
- **Operational sovereignty** — The enterprise manages availability, performance, capacity, and workload priority.
- **Economic sovereignty** — Owned infrastructure provides greater cost visibility and control as token volumes grow.

The table below summarizes common sovereign AI inference use cases by industry, workload, and primary sovereignty driver.

Table 1. Common sovereign AI inference use cases

Industry	Typical Inference Workload	Primary Sovereignty Driver
Financial services	Research summarization, client document processing, internal policy assistants	Regulatory jurisdiction and data residency
Healthcare & life sciences	Clinical documentation, literature, and trial-data research agents	Patient data protection and auditability
Government & defense	Classified document search, analyst support, disconnected field systems	Classification and air-gap requirements
Manufacturing	Visual defect analysis, maintenance knowledge assistants, digital twins	Latency and plant-floor autonomy
Legal & professional	Contract review, discovery, matter research at volume	Client confidentiality and privilege

Lenovo Sovereign AI Inference Architecture

The architecture keeps prompts, retrieved context, KV cache, and outputs on enterprise-managed infrastructure, with no required external endpoint in the inference data path. Enterprises can deploy, fine-tune, and version-pin open-weight models; independently scale Kubernetes-based prefill and decode resources; and tier KV cache to IBM shared storage to reduce memory pressure and improve GPU utilization.

Deployment runbooks for IBM and NVIDIA RAG and agentic blueprints, together with the LangFlow visual application builder, help keep retrieval pipelines, embeddings, and agent workflows on premises while accelerating the first production workload.

The following table outlines the software components that support workloads, orchestration, inference serving, models, data services, operations, and observability.

Table 2. Software components of the architecture

Component	Description
Workloads	NVIDIA AI-Q and Video Search and Summarization (VSS) Blueprints, IBM OpenRAG
AI orchestration	LangFlow
Inference serving	NVIDIA Dynamo, vLLM, LMCache, NIXL/UCX transport, LiteLLM
Models	Any open-weight, open-source model that fits the deployed GPU memory profile
Data services	IBM Storage Active File Management
Operating Platform	Kubernetes on Ubuntu or Red Hat
Observability	Customer choice of AIOps and AI SecOps tools

The following table summarizes the compute, storage, and networking components that form the hardware foundation of the architecture.

Table 3. Hardware components of the architecture

Component	Description
Compute	Lenovo Hybrid AI Factory architectures, including GPU server platforms supporting up to eight GPUs per node, following NVIDIA reference designs, plus x86 CPU nodes for client, control-plane, and gateway roles
Storage	Lenovo storage-rich servers running IBM Storage Scale ECE, or IBM Storage Scale System 6000
Networking	NVIDIA Spectrum-X Ethernet switching, with RoCE-capable adapters and BlueField DPUs

Inside the Architecture

The architecture, shown in the figure below, ranges from a single-node proof of concept to a Kubernetes-orchestrated, multi-node cluster. Specialized compute, workload, storage, and network resources can scale independently with token volume and user demand.

IBM Storage Scale provides a unified namespace for inference data, model weights, KV cache, and outputs such as transcripts, agent memory, logs, traces, and telemetry. Common encryption, access, retention, and audit policies simplify governance across the pipeline. The architecture uses scale units rather than a fixed configuration, enabling enterprises to expand compute, workload, storage, and network capacity independently as requirements evolve.

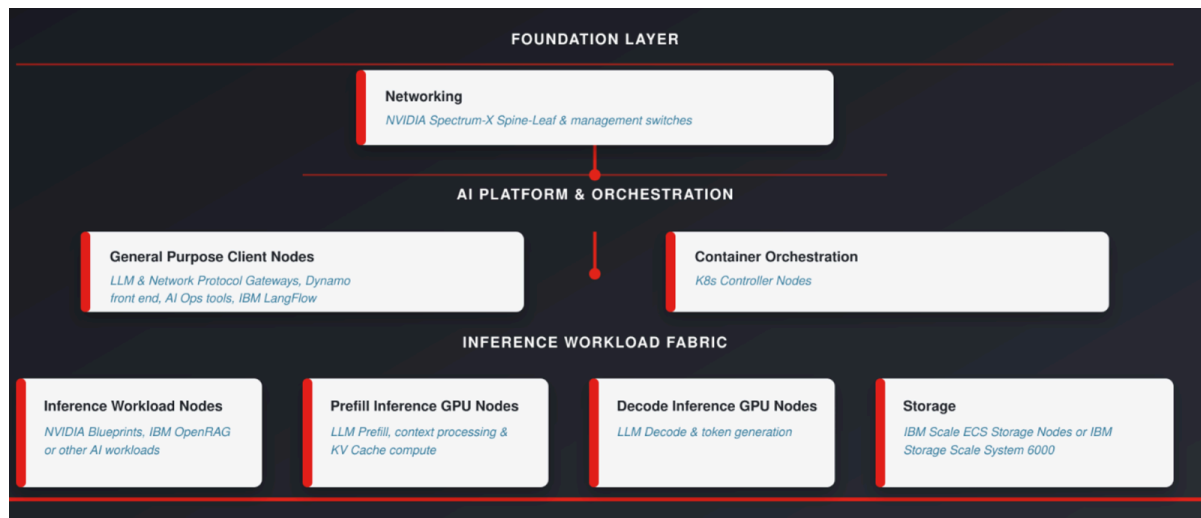


Figure 1. Lenovo Sovereign AI Inference Architecture on a Multi-Node Cluster

The architecture's key platform capabilities mapped to their practical customer outcomes, including stronger sovereignty, flexible scaling, improved GPU utilization, deployment choice, and faster time to production are shown in the following table.

Table 4. Key Features, Customer Benefits and Specifications

Key Platform Feature	Customer Benefits
Sovereign and turnkey by design	Keeps the complete inference path on premises; supports approved open-weight and fine-tuned models; integrates NVIDIA Dynamo, vLLM, and LMCache; and applies encryption, access control, tenant isolation, and audit trails.
Scale inference serving and workloads with ease	Independently scales control-plane, application, prefill, decode, and storage resources. Kubernetes manages container scheduling and recovery, while LiteLLM routes requests across models and endpoints for resilience and efficiency.
Serve more users without more GPUs	IBM Storage Scale extends KV cache beyond GPU HBM, reducing recomputation and helping GPUs serve more concurrent users with improved capacity economics.
Flexible deployment options	Supports single-node and multi-node deployments, with IBM Storage Scale ECE on Lenovo servers or IBM Storage Scale System 6000.
Working AI use cases day one	Includes deployment recipes for NVIDIA AI-Q, Video Search and Summarization, IBM OpenRAG, and the LangFlow visual workflow builder.

Single Node POC and Demo System

The following table presents a compact single-node starting point for proofs of concept and demonstrations using the included AI workloads and orchestration tools.

Table 5. Single-Node POC and Demo System Starting Point

Node Role	Starting Point	Content
POC/Demo	1	NVIDIA AI-Q Blueprint, IBM OpenRAG, LangFlow

For reference and details, view the document [Lenovo Validated Design Hybrid AI 281 with IBM](#).

Multi-Node Cluster

The following table provides recommended starting points and primary scaling factors for each role in a

production multi-node cluster.

Table 6. Multi-Node Cluster Starting Points and Scaling Factors

Node Role	Starting Point	Scales With
Prefill GPU nodes	1 or more	Context length and time-to-first-token targets
Decode GPU nodes	1 or more	Output length and concurrent conversations
Workload nodes	1 or more	Number and type of AI applications in production
General purpose client nodes	1 or more	Request rate and gateway throughput
Kubernetes control plane	3 nodes	High availability requirement
Storage	• 7 or more (IBM Storage Scale ECE), or • 1 or more IBM Storage Scale System 6000	Application datasets, RAG corpus size, KV cache tier capacity, and HA requirements
Spectrum-X fabric	3+ switches, spine-leaf	Node count and network traffic bandwidth

Services and Support

Lenovo AI Services can support workload and data assessment, cluster sizing, model and blueprint selection, system integration, on-site deployment, validation, and performance tuning. After go-live, Lenovo can assist with operations, re-tuning, and expansion as token volume and workload mix evolve, including connected, disconnected, and air-gapped environments.

Conclusion

The Lenovo Sovereign AI Inference architecture gives enterprises greater control over production of AI data, models, infrastructure, and costs. Independent scaling of compute, storage, and inference resources aligns capacity with demand, while support for connected, disconnected, and air-gapped environments provides a practical foundation for owning and operating enterprise intelligence.

Explore the Hybrid AI Advantage with NVIDIA

Discover how Lenovo and NVIDIA are helping enterprises build secure, scalable, and cost-efficient AI factories - bringing generative AI, RAG, and real-time inferencing into production with confidence.

Learn more about the Lenovo Hybrid AI Advantage, the [Lenovo Validated Design Hybrid AI 281 with IBM](#) and explore the full portfolio of NVIDIA-accelerated solutions and services by visiting the [Lenovo Hybrid AI Solutions](#) page.

Authors

Pierce Beary is the AI Solutions Product Manager focused on datacenter solutions, he previously worked on the WW HPC team as a product manager for Lenovo EveryScale (LeSI). Prior to joining Lenovo in 2021, Pierce held positions at Exxon Mobil, where he worked as a Mechanical Engineer and Engineering Project Manager. Pierce has a Bachelor's degree in Mechanical Engineering from North Carolina State University.

Traci Parker is the Worldwide Solutions Marketing Manager for Enterprise IT and AI at Lenovo. She specializes in hybrid cloud, infrastructure modernization and AI solutions. She has more than 15 years of experience as a Marketing Manager and Product Marketing Manager across high-tech, fin-tech and healthcare industries.

Related product families

Product families related to this document are the following:

- [Artificial Intelligence](#)
- [Hybrid AI Factory](#)
- [IBM Alliance](#)

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

© Copyright Lenovo 2026. All rights reserved.

This document, LP2558, was created or updated on October 16, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
<https://lenovopress.lenovo.com/LP2558>
- Send your comments in an e-mail to:
comments@lenovopress.com

This document is available online at <https://lenovopress.lenovo.com/LP2558>.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at <https://www.lenovo.com/us/en/legal/copytrade/>.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®

Lenovo Hybrid AI Advantage™

The following terms are trademarks of other companies:

IBM® is a trademark of IBM in the United States, other countries, or both.

NVIDIA®, NVIDIA Spectrum®, and NVIDIA Spectrum-X are trademarks of NVIDIA Corporation.

Other company, product, or service names may be trademarks or service marks of others.