Production-Ready AI Platforms for Real-Time Enterprise Inferencing
Article

Lenovo’s NVIDIA-accelerated AI servers and hybrid AI platforms

Lenovo's NVIDIA-accelerated AI servers and hybrid AI platforms are redefining enterprise AI economics and now delivering ROI in less than six months and up to 8x lower cost per token compared to comparable cloud IaaS, helping enterprises bring AI workloads on-premises with greater efficiency and control.

To learn more about the benefits of on-prem and hybrid AI implementations, see the Lenovo paper, On-Premise vs Cloud: Generative AI Total Cost of Ownership (2026 Edition).

The expanded portfolio features pre-validated system designs integrated with NVIDIA AI Enterprise software, including:

- Two Lenovo Hybrid AI agentic inferencing & tunning platform optimizations:
    - With NVIDIA RTX PRO 6000 GPUs for scale-out AI and multi-modal inferencing. For details, see the Lenovo Hybrid AI 285 Platform Guide
    - With NVIDIA B300 - among the first NVIDIA design review board-certified systems for enterprise AI training. For details, see the Lenovo Hybrid AI 289 Platform Guide
- Lenovo Hybrid AI inferencing starter platform with NVIDIA RTX PRO 4500 Blackwell, delivering up to 3X performance gains for video and data processing and 4X better performance for content generation compared to NVIDIA L4 for single-node deployments. For details, see the Lenovo Hybrid AI 221 Platform Guide
- Lenovo ThinkAgile HX650a with Nutanix Enterprise AI and Nutanix Kubernetes Platform, providing a validated foundation for protected inferencing and agentic workloads. For details, see the Lenovo Solution Brief, ThinkAgile HX Solution for AI with Nutanix Enterprise AI.
- Lenovo Hybrid AI platforms with Cloudian deliver scalable, sovereign data pipelines, while Veeam Kasten provides Kubernetes-native protection to safeguard AI models and services. For details, see the Lenovo Validated Design: On-Prem High Performance AI Data Platform.
- Enterprises are scaling generative AI and RAG into production — and protecting models, vector databases, and Kubernetes state is now mission-critical. Downtime can exceed $300K per hour. Data breaches average $4.88M. Retraining large AI models can cost millions more. The Lenovo Validated Design for Lenovo Hybrid AI Platform with Veeam Kasten provides a validated blueprint to reduce that exposure.

Lenovo NVIDIA GB300 NVL72

The Lenovo NVIDIA GB300 NVL72 is a rack scale solution leading the new era of AI with optimized compute and increased memory. Based on the NVIDIA GB300 NVL72 platform, the Lenovo NVIDIA GB300 NVL72 server combines the high-performance NVIDIA B300 GPUs, NVIDIA Grace processors, and a hybrid cooling solution, which results in extreme HPC/AI performance in dense packaging.

For more information on this new rack-scale offering see the following:

- GB300 NVL72 datasheet
- GB300 NVL72 Product Guide
- GB300 NVL72 3D Tour

Read more about an implementation of the Lenovo offering in the Storyhub press release, Lenovo Teams with NVIDIA on Gigawatt AI Factories Program to Accelerate Enterprise AI.

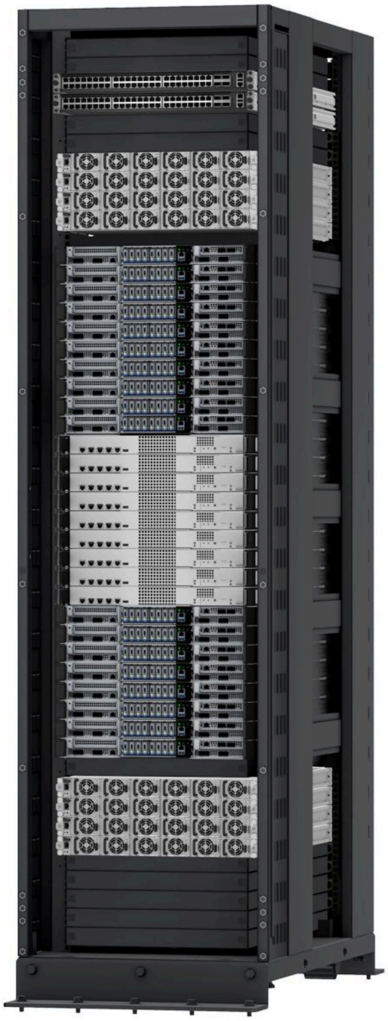

Figure 1. Lenovo NVIDIA GB300 NVL72 Rack Scale AI

Lenovo Inferencing Servers

As introduced at CES in January, Lenovo announced new inferencing-optimized ThinkSystem and ThinkEdge servers, combined with enhanced Hybrid AI platforms and integrated partner solutions, enable real-time AI Inferencing across retail, manufacturing, healthcare, sports, and smart city environments.

- ThinkSystem SR650i V4
- ThinkSystem SR675i V3
- ThinkEdge SE455i V3

Read the datasheets and product guides:

- SR650i V4: Datasheet and Product Guide
- SR675i V3: Datasheet and Product Guide
- SE455i V3: Datasheet and Product Guide

Learn more about these offerings with the following materials:

- Lenovo Storyhub press release: Lenovo Revolutionizes Real-Time Enterprise AI with New Inferencing Servers
- Thought leadership paper by Futurum: AI Inference: Enterprise Infrastructure and Strategic Imperatives and Infographic
- Lenovo Storyhub article: The Agentic AI Era: Why Hybrid AI Advantage with NVIDIA Will Define Enterprise Leadership
- AI Innovators Solution Brief: Smart cities AI for public safety with Vaidio
- AI Innovators Solution Brief: Spatial intelligence for Retail with AiFi
- AI Innovators Solution Brief: Adaptive loss prevention from RocketBoots
- Inferencing Infographic: AI Inferencing at Scale - Lenovo Hybrid AI Advantage with NVIDIA
- Hybrid AI Infographic: Build Your Enterprise AI Factory Faster - Lenovo Hybrid AI Advantage with NVIDIA

IBM Technology Lifecycle Services

These solutions are backed by an expanded global collaboration with IBM Technology Lifecycle Services to accelerate hybrid AI adoption worldwide.

IBM Technology Lifecycle Services complements the Lenovo Hybrid AI foundation with end-to-end delivery and managed services—from Day 0 design and deployment through ongoing operations and lifecycle optimization. Together, Lenovo and IBM bring joint validation, proven blueprints, and real-world operational experience to customers worldwide.

The result is a pragmatic path to production grade hybrid AI—where infrastructure is treated as a strategic asset, operations are designed for scale, and enterprises can move from AI ambition to sustained business impact. Because in enterprise AI, infrastructure isn't an afterthought. It's what makes everything else possible.

Read more in the article by Lenovo's Robert Daigle and IBM's Omkar Nimbalkar, Inference Economics Are Here: Why Enterprise AI Lives or Dies on Infrastructure.

Explore the Lenovo 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 and explore the full portfolio of NVIDIA-accelerated solutions and services:
https://www.lenovo.com/us/en/servers-storage/solutions/ai/

Author

Maria Belén Rotelli is the ISG Worldwide Enterprise AI and Sports Solutions Marketing Manager at Lenovo. She specializes in executive messaging and enterprise AI positioning across the company's AI infrastructure and Hybrid AI portfolio, including its application within the sports vertical. She brings more than 15 years of experience at Lenovo, IBM and Microsoft, with deep expertise in corporate communications, global campaign leadership and enterprise technology marketing within large-scale organizations.

Related product families

Product families related to this document are the following:

- AI Servers
- Artificial Intelligence
- Rack Scale AI

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.

This document, LP2393, was created or updated on March 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/LP2393
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2393.

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
ThinkAgile®
ThinkEdge®
ThinkSystem®

The following terms are trademarks of other companies:

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

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