Lenovo to Deliver Enterprise AI Compute for NetApp AIPod Through Collaboration with NetApp and NVIDIA

Article

Enabling the AI infrastructure

Artificial intelligence (AI) has transcended beyond a technological buzzword into a key driver of turning valuable data into actionable insights, delivering real business value and competitive advantage for customers. With the advent of ChatGPT in late 2022 and the ensuing development of large language models (LLMs), AI has gained traction within enterprises, enabling new use cases, applications, and workloads across many industries. AI has moved beyond just training LLMs in the cloud to become more hybrid in nature, driving the need for running private AI use cases leveraging on-prem AI infrastructure for enterprise customers. These private AI use cases require new converged infrastructure solutions targeted at enterprise customers to simplify and accelerate AI infrastructure implementations at scale for on-premises solutions. However, enterprise customers face the following types of challenges, which can limit on-prem AI deployments:

- **Lack of data scientists** who can build, deploy, and manage LLMs.
- **Datacenter power and cooling limitations** for enterprise customers typically are limited to 8 to 10KW per physical rack space that houses the IT. Typical LLMs/Generative AI (GenAI) utilizing GPUs consume in a single server over 10KW of power. This limits the utilization of datacenter real estate.
- **Challenges with configuring the right AI solution deployments** for customers who don't have the skilled resources to determine the amount of GPU accelerators, storage, and networking needed to support the training, retraining, and inference of their data for their AI deployment.

Lenovo collaboration with NetApp and NVIDIA

Lenovo is focused on simplifying our customers' AI journey and providing comprehensive solutions to unleash the power of AI to drive intelligent transformation in every aspect of our lives and every industry – delivering AI for All. We do this by delivering customer-valued innovation and partnering with industry leaders – such as NetApp and NVIDIA – to bring you the right set of solutions for your enterprise AI deployments.

Lenovo is further expanding our strategic partnership with NetApp by delivering the world's most reliable and secure enterprise AI compute in the NetApp AIPod™ solution. NetApp AIPod is an integrated solution designed to simplify the planning, sizing, deployment, and management of tuning and inferencing enterprise AI models.

As announced on May 14, 2024, the Lenovo ThinkSystem SR675 V3 servers are foundational to this solution and deliver the powerful computing resources needed for accelerating the complex GenAI computations used in today's enterprise AI models.

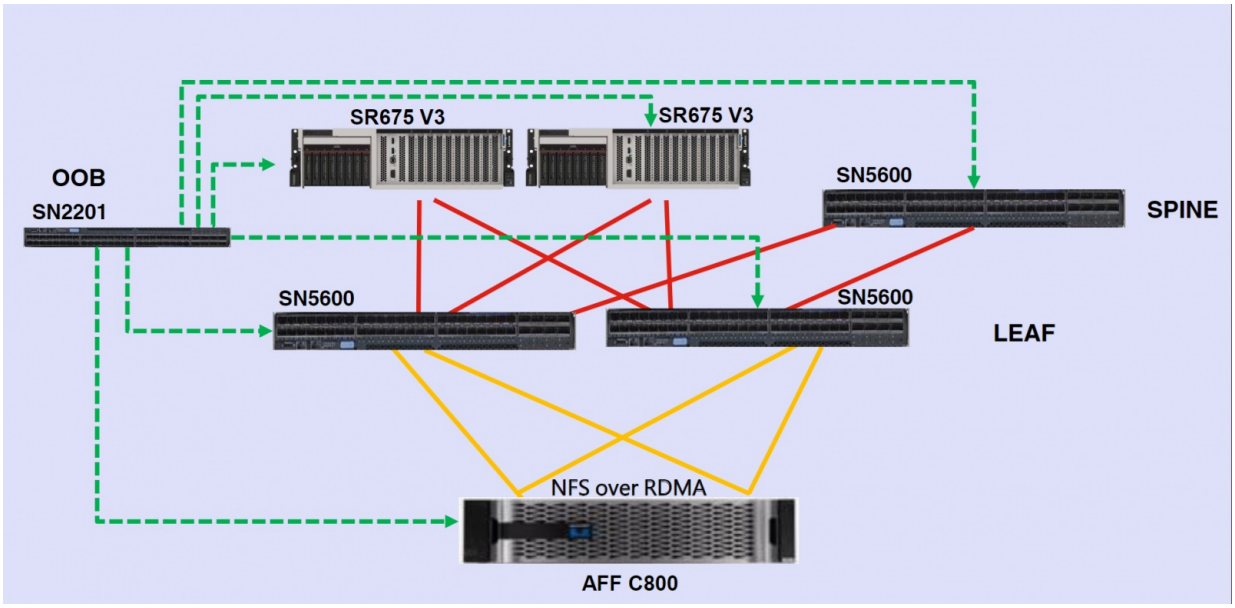

Figure 1. NetApp AIPod solution featuring ThinkSystem SR675 V3 servers

AI has emerged as one of the most transformative technologies in the industry, with the potential to unlock faster and better data insights to help customers transform their business operations. In the past few years, the industry has been focused on developing and honing AI models in the cloud, leveraging massive investments in technology infrastructure. As these AI models mature in the cloud, the logical evolution is to bring them into private AI deployments so enterprise customers can turn their own business data into valuable business insights and actions.

Given the continued growth in data and the large datasets common in most enterprise deployments, data gravity and data privacy trends have driven a higher demand for an on-premises and hybrid approach to AI. This hybrid AI approach enables customers to use their own data when running AI models to deliver the optimal, unique, and tailored results for their business operations. The same Lenovo infrastructure and technology that helped mature and hone these AI models in the cloud is the foundation for our private AI enterprise infrastructure deployments. With our comprehensive AI portfolio and services, such as AI Discover Center of Excellence and AI Innovators, Lenovo has the proven AI expertise, solutions, services, and support to turn your organization's data into actionable insights.

Customer data is the most valuable part of an organization's AI strategy, and our partner NetApp is the storage platform for most enterprise customers' unstructured data. The Lenovo strategy to bring AI to customers' data makes NetApp the ideal partner for AI solutions. We are expanding on our five-year partnership to deliver new innovation and efficiency for enterprise customers' AI deployment.

Lenovo integrated AIPod solution

The NetApp AIPod solution is a new validated, integrated solution composed of Lenovo ThinkSystem SR675 V3 servers, NetApp C800 high performance storage, NVIDIA L40S GPUs, the NVIDIA Spectrum-X networking platform, and the NVIDIA AI Enterprise software platform. This is an ideal solution to host AI workloads, including the targeted workloads of training and inferencing for AIPod. Lenovo works closely with NVIDIA to deploy NVIDIA GPUs and DPUs and NVIDIA AI Enterprise, which includes NVIDIA NIM and other microservices, throughout the Lenovo AI infrastructure portfolio. Our collaboration with NetApp and NVIDIA enables us to deliver the first integrated AI solution in the market with retrieval augmented generation (RAG) to support chatbot, knowledge management, and object recognition use cases with critical benefits, including:

- **Data Management Simplicity**: The NetApp AIPod with Lenovo addresses the complexities of data

management by providing tools and features that simplify infrastructure management, enhance automation, and ensure scalability and data protection.

- **High Performance**: Combining the raw power of Lenovo ThinkSystem servers with advanced NVIDIA L40S GPUs and NVIDIA Spectrum-X networking, the AIPod with Lenovo for NVIDIA OVX delivers the computational strength required for the most intensive AI tasks, supported by the unparalleled speed and efficiency of NetApp storage systems.
- **Integrated Solution**: The AIPod with Lenovo for NVIDIA OVX integrates NVIDIA AI Enterprise to streamline the deployment and scaling of AI workloads, while NVIDIA NeMo allows organizations to customize, build, and deploy AI models with ease, leveraging pre-trained models for quicker deployment and NVIDIA TensorRT software for accelerating and optimizing inference performance.
- **Trusted Secure Data**: Security is a foundational element of the AIPod with Lenovo, ensuring that data is not only rapid and accessible but also rigorously protected.

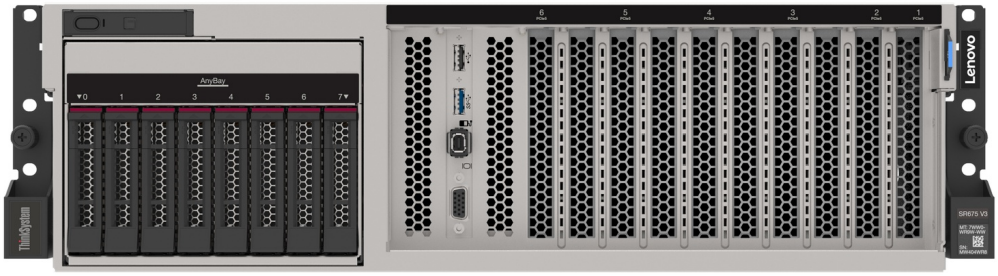

Figure 2. Lenovo ThinkSystem SR675 V3 server

Lenovo is proud to collaborate with NetApp and NVIDIA to deliver a simplified, efficient, and robust solution that helps customers operationalize their enterprise AI workloads to transform their business. The reference architecture with a detailed bill of material (BOM) for this solution will be published later in June 2024 and will be available from the network of channel partners for Lenovo, NetApp, and NVIDIA.

We look forward to continuing to deliver Smarter AI for All for our customers!

Learn More

To learn more about the Lenovo, NetApp and NVIDIA collaboration on NetApp AIPod, please read the Lenovo press release.

You can also contact your Lenovo sales representative or authorized channel partner to learn about this and other Lenovo Data Management offerings.

About the author

Kamran Amini is the Vice President and General Manager of Server, Storage & Software Defined Infrastructure at Lenovo ISG.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- ThinkSystem SR675 V3 Server

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, LP1962, was created or updated on May 14, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1962
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1962.

Trademarks