VMware Private AI Foundation with NVIDIA on Lenovo Architecture

Solution Brief

As digital transformation of businesses continues to accelerate, organizations want to harness the power of Artificial Intelligence (AI) to transform existing operations and create new opportunities. Adoption of AI has been hindered by privacy, complexity and cost associated with deploying and managing AI workloads at enterprise scale. Advancement in Generative AI (Gen AI) technology spurred organizations in many industries to start adopting for various use cases. Enterprise-scale Gen AI solutions emerged to simplify the deployment, management, and scaling of AI workloads.

Lenovo, NVIDIA, and VMware By Broadcom are partnering to deliver a private, secure, scalable, and flexible AI infrastructure solution that helps enterprise customers build and deploy AI workloads within their own private cloud infrastructure, ensure the control of sensitive data and compliance with regulatory requirements, ultimately driving faster time to value and achieving their AI objectives.

By combining VMware's industry-leading virtualization and cloud computing capabilities with Lenovo's expertise in AI-optimized infrastructure and best in-class hardware, and leveraging NVIDIA's superior accelerator portfolio, AI ecosystem, and support, this partnership will provide a robust and scalable foundation to accelerate AI adoption, allowing businesses to focus on driving business value rather than managing complex infrastructure.

This solution brief introduces VMware Private AI Foundation with NVIDIA, a joint Gen AI platform by Broadcom and NVIDIA on Lenovo servers, which helps enterprises deploy AI workloads faster and unleash productivity. With this platform, enterprises can fine-tune and customize Large Language Models (LLM), deploy retrieval augmented generation (RAG) workflows, and run inference workloads in their data centers, addressing privacy, choice, cost, performance, and compliance concerns.

Generative AI Adoption

Generative AI (Gen AI) is a general-purpose AI technology that allows every industry to use AI to transform existing business operations and create new opportunities for innovation. A major advantage of Gen AI is the ability to interact with Large Language Models (LLM) via natural language, making AI technology consumable by a much broader set of use cases and personas. Many enterprises use Gen AI for knowledge base Q&A, document summarization, and creating chatbots or smart assistants to improve productivity and enhance service experience.

In Financial Services, Gen AI is used for fraud detection, offering personalized banking, and providing investment insights. In Healthcare, Gen AI is used for molecule simulation to accelerate drug discovery, and for improving clinical trial data analysis. In Media and Entertainment, Gen AI is contributing to character development, enabling sophisticated video editing & image creation, and facilitating style augmentation. Gen AI is enhancing Retail shopping experiences and automating catalog descriptions. Gen AI is enhancing Manufacturing simulation, product design, and predictive maintenance.

To implement these and many other use cases of Generative AI, enterprises need a flexible and scalable AI infrastructure that can customize and fine-tune next-gen LLM with their domain-specific data and up-to-date context, deploy the models rapidly into operational environments, while always ensuring data privacy and security.

VMware Private AI Foundation with NVIDIA

VMware Private AI Foundation with NVIDIA is a joint Gen AI platform between NVIDIA and Broadcom. Now Generally Available, the platform is built on top of VMware Cloud Foundation (VCF) and is an add-on offering to VCF. The platform comprises VCF, VMware supported AI tools, and NVIDIA AI Enterprise. With this platform, enterprise customers can fine-tune and customize LLM, develop and deploy Gen AI applications faster, while addressing privacy, choice, cost, performance, and compliance, all in a secure private cloud environment.

VMware Private AI Foundation with NVIDIA platform simplifies the implementation, operations, and life-cycle management of an AI platform through infrastructure automation, which helps the customers build and deploy Gen AI applications in these four areas:

- **Code Generation:** Generates code for software development and testing accelerates developer velocity, gives customers full control of prompts and ensures the protection of data privacy and IP.
- **Contact Center Experience**: Improves relevance and accuracy of responses from their contact centers and significantly enhances customer experience.
- **IT Operations Automation**: Enables customers to improve their IT operations by automating processes of incident management, reporting, ticketing, and monitoring.
- **Advanced Information Retrieval:** Provides advanced capabilities for document search, policy, and procedure research, significantly improving worker productivity.

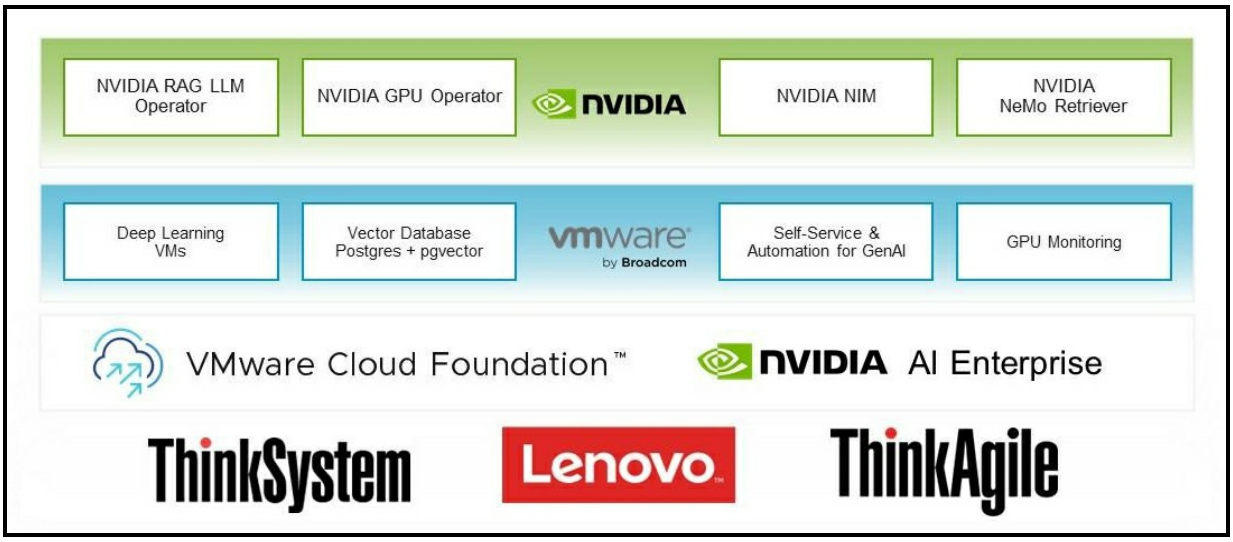

Figure 1. VMware Private AI Foundation with NVIDIA

VMware Private AI Foundation with NVIDIA platform also helps the customers access infrastructure resource through a self-service experience, improve efficiency of iterative development and test of AI workloads, thus accelerating the path to production for these critical workloads. To achieve this, VMware and NVIDIA have been working together to integrate NVAIE with VCF to ensure easy consumption of the key elements such as a retrieval-augmented generation architecture.

Solution Components

Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of Generative AI models with facts fetched from external sources. VMware Private AI Foundation with NVIDIA includes several key components to optimize RAG workflows and speed up Gen AI deployment across enterprises.

The key components are as follows:

- **Vector Database** is a key component developed using pgvector on PostgreSQL and delivered by Data Services Manager (DSM) with VMware enterprise-level support, allowing customers to add domain-specific, up-to-date contexts to LLM, enabling fast querying of real-time data and ensuring data privacy.
- **NVIDIA Nemo Retriever**, part of the NeMo platform, is a collection of generative AI microservices enabling organizations to seamlessly connect custom models to diverse business data and deliver highly accurate responses. NeMo Retriever provides world-class information retrieval with the lowest latency, highest throughput, and maximum data privacy, enabling organizations to make better use of their data and generate business insights in real-time. NeMo Retriever enhances generative AI applications with enterprise-grade retrieval-augmented generation (RAG) capabilities, which can be connected to business data wherever it resides.
- **NVIDIA NIM** is a set of easy-to-use microservices designed to speed up the deployment of Gen AI across enterprises. This versatile microservice supports NVIDIA AI Foundation Models — a broad spectrum of models, from leading community models to NVIDIA-built models to bespoke custom AI models optimized for the NVIDIA accelerated stack. Built on the foundations of NVIDIA Triton Inference Server, NVIDIA TensorRT, TensorRT-LLM, and PyTorch, NVIDIA NIM is engineered to facilitate seamless AI inferencing at scale, helping developers deploy AI in production with agility and assurance.
- **NVIDIA RAG LLM Operator** makes it easy to deploy your RAG application into production. It streamlines the deployment of RAG pipelines developed using AI workflows into production, without rewriting code.

In addition to optimizing RAG workflows and speeding up Gen AI deployment, VMware Private AI Foundation with NVIDIA provides the following toolsets that increase data scientists' productivity and enable self-service:

- **Deep Learning VM** provides pre-configured VM templates with Deep Learning frameworks and software libraries that are validated for version compatibility, reducing errors, and saving time for users.
- **Catalog Setup Wizard** expedites AI deployment, enables rapid creation, customization, and availability of Gen AI catalog items, allowing data scientists to access resources on-demand in a self-service manner. This capability is accomplished through VCF's self-service portal (via Aria Automation) and comes preloaded with example templates to get customers up and running quickly.

Beyond these benefits for end users, VMware Private AI Foundation with NVIDIA also provides toolsets and automation that make it easy for administrators to manage AI infrastructure at scale:

- **VMware Cloud Foundation (VCF)** offers a full-stack software-defined architecture designed to deliver a self-service unified platform and leverage an automated IT environment that simplifies the deployment and management of all workloads utilizing VMs, containers, and AI technologies. The versatility offered through this architecture enables cloud admins to utilize different workload domains which can each be customized to support specific workload types, optimizing for workload performance and resource utilization, specifically GPUs. Leveraging the field-proven integration between VMware Cloud Foundation and the NVIDIA AI Enterprise Suite, the ability to virtualize GPU (vGPUs) to maximize performance and utilization of physical GPUs resources across multiple users. This VCF- NVAIE integration is core to the value that is provided by VCF and when coupled with the VMware Private AI Foundation add-on, customers have a complete solution to support Gen AI workloads across the spectrum of use cases and applications supported by the solution.

- **NVIDIA GPU Operator**: Kubernetes provides access to special hardware resources such as NVIDIA GPUs, NICs, Infiniband adapters and other devices through the device plugin framework. However, configuring and managing nodes with these hardware resources requires the configuration of multiple software components such as drivers, container runtimes or other libraries, which are difficult and prone to errors. The NVIDIA GPU Operator uses the operator framework within Kubernetes to automate the management of all NVIDIA software components needed to provision GPUs. These components include the NVIDIA drivers (to enable CUDA), Kubernetes device plugin for GPUs, the NVIDIA Container Toolkit, automatic node labeling using GFD, DCGM based monitoring and others.
- **GPU Monitoring** provides visibility of total GPU compute and memory usage across hosts and clusters. This monitoring will allow customers to start forecasting GPU usage and with integrations into Aria Operations customers can build out-of-the-box or custom dashboards to help them track the usage of GPU enabled infrastructure and the applications that are leveraging them.

Lenovo Platforms for Private AI Foundation

Dense GPU deployments are crucial for AI due to their unparalleled ability to handle extensive computational workloads efficiently. GPUs, with their thousands of cores, can perform parallel processing at high speeds, making them ideal for training complex neural networks and executing large-scale AI models. This capability significantly reduces training times and enhances real-time inference performance, essential for applications like autonomous driving, natural language processing, and large-scale recommendation systems. Additionally, dense GPU setups optimize resource utilization, lower operational costs, and facilitate the scalability required for modern AI workloads, ensuring robust and responsive AI-driven solutions in various industries.

Lenovo ThinkAgile and Lenovo ThinkSystem are the industry leading solutions for dense GPU deployments, offering unparalleled flexibility for AI workloads. Our servers cater to a wide range of workload needs, providing configurations from 2-GPU setups for smaller projects to 8-GPU configurations for intensive applications. This granularity ensures that enterprises can scale their AI infrastructure precisely to their requirements, optimizing both performance and cost. Lenovo ThinkAgile and Lenovo ThinkSystem are designed to support diverse AI deployments, from research and development to production environments, making them the ideal choice for businesses seeking adaptable, high-performance solutions for their AI initiatives.

- Lenovo systems supporting 3x GPUs:
  - Lenovo ThinkAgile VX665 V3
  - Lenovo ThinkAgile VX655 V3
  - Lenovo ThinkAgile VX650 V3
  - Lenovo ThinkSystem SR665 V3
  - Lenovo ThinkSystem SR655 V3
  - Lenovo ThinkSystem SR650 V3
- Lenovo systems supporting 8x GPUs:
  - Lenovo ThinkSystem SR675 V3
  - Lenovo ThinkSystem SR670 V2

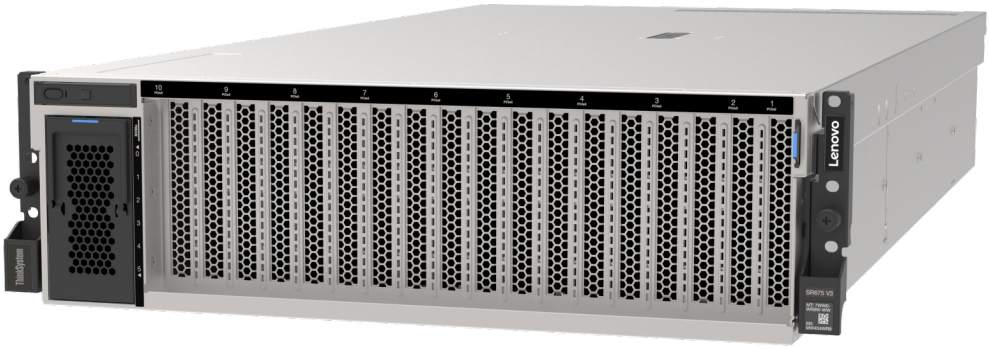

Figure 2. Lenovo ThinkSystem SR675 V3 supports eight double-wide GPUs

Running VMware Private AI Foundation with NVIDIA on Lenovo architecture significantly enhances the value of Lenovo ThinkAgile and ThinkSystem GPU deployments. VMware's robust platform provides a secure and scalable infrastructure for managing AI workloads, ensuring seamless integration, efficient resource allocation, and high availability. Combined with our flexible GPU configurations, customers can effortlessly deploy, manage, and scale their AI models while maintaining control over their data and compliance requirements. The synergy between VMware's advanced virtualization technology and our powerful servers empowers businesses to accelerate their AI initiatives, driving innovation and achieving superior performance in their AI-driven applications.

Architecture for VMware Private AI Foundation with NVIDIA

Architecture for VMware Private AI Foundation with NVIDIA is a solution blueprint for enterprise customers to design a secure, scalable, and flexible Gen AI infrastructure solution that helps them accelerate AI development, increase user productivity, and provide enterprise-level support.

Key benefits of this architecture include:

- **Privacy and Security**: Enables organizations to build and deploy AI models on their own private cloud infrastructure, ensuring the control of sensitive data and compliance with regulatory requirements.
- **Scalability and Performance**: Capable of handling large-scale AI workloads in enterprise environments and supporting AI applications to efficiently process vast amounts of data while maintaining responsiveness and reliability.
- **Accelerated AI Development**: Provides pre-configured AI frameworks and libraries optimized for NVIDIA GPUs, which enables data scientists and developers to quickly prototype, train, and deploy AI models, reducing time-to-market for AI-driven solutions.
- **Collaboration and Integration**: Facilitates collaboration and integration across different teams and departments within an organization by providing tools and frameworks for data sharing, model collaboration, and workflow automation.
- **Support and Ecosystem**: Provides enterprise-level support by Lenovo, NVIDA and VMware, and access to an ecosystem of partners and developers for latest technologies, best practices, and expertise needed to succeed in AI-driven initiatives.

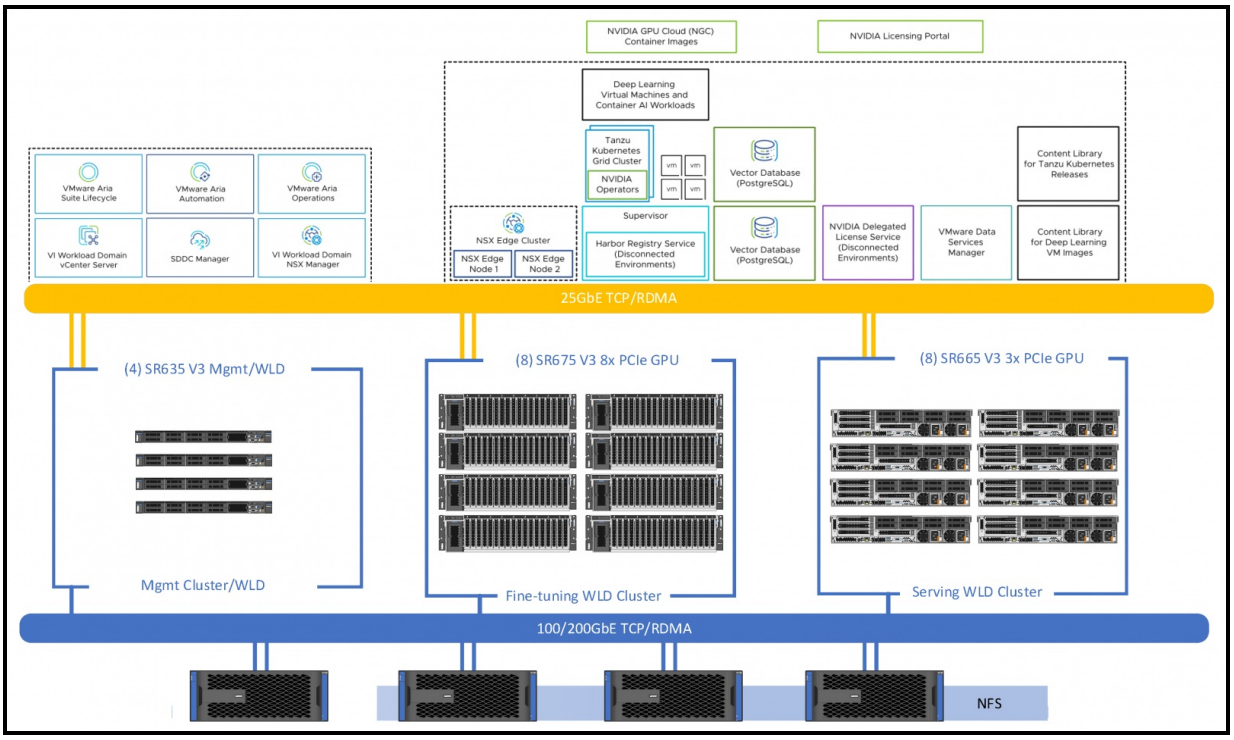

Figure 3. Architecture for VMware Private AI Foundation with NVIDIA

Retrieval-Augmented Generation Architecture

Retrieval-Augmented Generation (RAG) architecture is indeed revolutionary in the realm of Generative AI, particularly for chatbot applications. RAG combines the strengths of both generative and retrieval-based models. Generative models create responses from scratch, while retrieval-based models retrieve responses from a pre-existing database. RAG seamlessly integrates these two approaches, allowing for a more real-time and contextually relevant generation of responses.

1. **Improved Context Understanding**: RAG leverages retrieval mechanisms to obtain and incorporate relevant information from a large knowledge database, enabling it to generate more coherent and contextually accurate responses.
2. **Enhanced Response Quality**: By accessing a vast amount of information during the generation process, RAG can produce responses that facilitate higher-quality interactions with users, enhancing the overall user experience.
3. **Faster AI Objective Achievement**: Users can quickly obtain the information or assistance they need without the need for extensive training or fine-tuning of the model.
4. **Scalability and Adaptability**: RAG allows for easy scalability, and it can be augmented on domain-specific data to enhance its performance and accuracy.
5. **User Engagement and Satisfaction**: With its ability to generate more relevant and engaging responses, RAG helps in keeping users engaged in conversations for longer durations. This increased engagement leads to higher user satisfaction and retention rates.

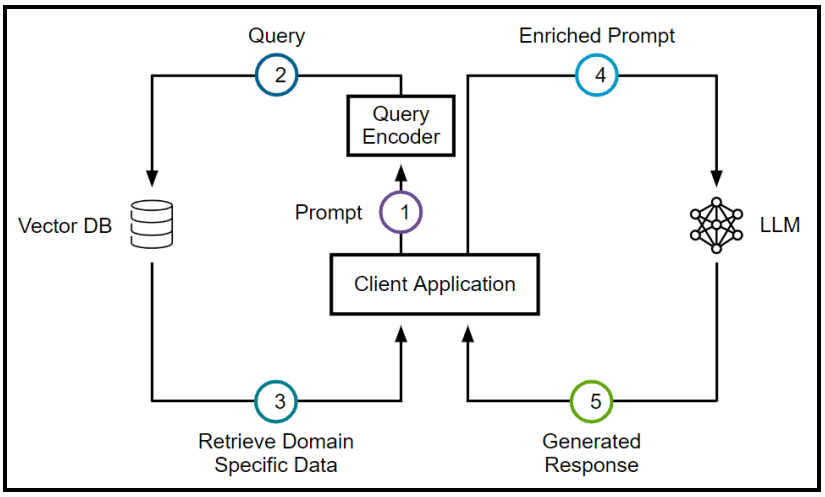

Figure 4. Retrieval-Augmented Generation Workflow

In the figure, the steps are:

1. Prompt: The user generates the prompt "What's new in vSphere 8 update 2"
2. Query: The prompt is forwarded to the Vector DB which looks for documents, articles, any type of information related to vSphere 8.0 U2
3. Retrieve Domain Specific Data: Information related to vSphere 8.0 U2 is sent back to the client application
4. Merging Retrieval with Generation: The retrieved information is combined with the user prompt and sent as a new, enriched prompt to the Large Language Model (LLM):
5. Generated Response: The LLM uses the enriched prompt (original prompt + retrieved information) and its creative writing skills to create a unique response

In summary, the RAG architecture revolutionizes Generative AI use cases, especially in chatbot applications, by combining the strengths of generative and retrieval-based approaches. It enables faster achievement of AI objectives by providing contextually relevant responses, leading to improved user engagement and satisfaction.

Design and Sizing Considerations

When designing a Gen AI infrastructure solution, it is essential to consider factors such as model size, the domains of previously trained data, and the context window size (maximum prompt size) and licensing terms. Optimal resource sizing for LLMs within the constraints of available hardware can significantly enhance application performance.

When selecting an LLM, it is important to consider the model's parameter size, such as 13 billion (13B) parameters, 70 billion (70B) parameters, or even larger models. For inferencing tasks, each parameter can be loaded in 16 bits (2 bytes) of GPU memory; for optimal performance, it is recommended to double this amount. For training or fine-tuning tasks, it is advisable to consider a GPU memory capacity four times larger. A solution based on Lenovo Private AI Foundation architecture can efficiently load a 20 billion parameter model using NVIDIA L40S GPUs.

It is also important to consider the bandwidth between GPUs. According to NVIDIA, each GPU should achieve a bandwidth of 200 Gbps, with newer GPU generations capable of reaching 400 Gbps. A full non-blocking topology is recommended for optimal performance.

During the design and sizing, it is crucial to refer to benchmarks, adhere to guidelines, and follow best practices, ensuring that the solution exhibits strong capabilities for handling enterprise LLM RAG applications. For example, it is necessary to quantitatively assess both the retrieval and generation components using an appropriate framework such as Ragas (Ragas is an open-source toolkit that utilizes LLMs to effectively score the accuracy of both components) so that the RAG application can achieve superior results in responding to user queries and improving overall system efficiency. Stay tuned for the next article, where we will explore the creation of a dataset for validating and evaluating the RAG application.

Solution Configurations

Above we provided an overview of VMware Private AI Foundation with NVIDIA, Lenovo platforms for Private AI Foundation, and the combined architecture as a solution blueprint for enterprise customers. The following table illustrates two configurations that customers can use as starting points then further optimize based on their requirements of Gen AI infrastructure.

The hardware and software hardware bills of materials (BOMs) shown here are a subset of available options. Please engage Lenovo, NVIDIA, and VMware technical teams for design and sizing assistance.

Table 1. Solution Configurations

Solution Configuration	AI Inferencing	AI Inferencing & LLM Fine Tuning
Lenovo Platforms for Private AI Foundation		
Model Size	<13B Parameters	13B to 175B Parameters
Server Model	ThinkAgile VX650 V3 (Intel CPU) ThinkAgile VX665 V3 (AMD CPU)	ThinkSystem SR670 V2 (Intel CPU) ThinkSystem SR675 V3 (AMD CPU)
Compute Cores per Server	VX650 V3: 64C (2x Intel Gold 6548N 32C 2.8GHz) VX665 V3: 64C (2x AMD EPYC 9334 32C 2.7GHz)	SR670 V2: 96C (2x Intel Gold 6448H 48C 2.4GHz) SR675 V3: 96C (2x AMD EPYC 9454 48C 2.75GHz)
GPU per Server	3 x NVIDIA L40 or L40S 48GB	SR670 V2: 4x NVIDIA H100 80GB SR675 V3: 8x NVIDIA H100 80GB or 4x NVIDIA H100 NVL 94GB
System Memory per Server	256GB minimum, 512GB recommended	512GB minimum, 1TB recommended
Storage per Server	Boot drive: 2x 960GB M.2 RI NVMe Storage Tier: 6x 7.68TB RI NVMe	Boot Drive: 2x 960GB M.2 RI NVMe Storage Tier: 6x 7.68TB RI NVMe
Network Adapter per Server	1x Broadcom 57504 10/25GbE 4-Port 1x ConnectX-6 HDR100/100GbE 2-port	1x Broadcom 57504 10/25GbE 4-Port 2x ConnectX-7 NDR200/200GbE 2-port
# of Servers Required for VI Workload Domain	3 servers minimum	3 servers minimum
Private AI Foundation Software License		
Infrastructure Software Subscription & Support	VMware Cloud Foundation (VCF) version 5.1.1 or later 1, 3, or 5-year term subscription, per Core	
NVIDIA AI Software Subscription & Support	NVIDIA AI Enterprise (NVAIE) version x.x or later 1, 3, or 5-year term subscription, per GPU	
VMware AI Software Subscription & Support	VCF Add-on: Private AI Foundation with NVIDIA 1, 3, or 5-year term subscription, per Core	
Additional Solution Offering		
Professional Service	Lenovo ThinkAgile VX Deployment Services VMware VCF Implementation Services	
Solution Support	Lenovo Premier Support Lenovo Enterprise Server Software Support	

Conclusion

Lenovo, NVIDIA, and VMware By Broadcom have a productive and proven history of technical collaboration, this partnership consistently delivers innovative data center solutions and lower TCO for our joint customers.

With Lenovo's leadership in AI-optimized hardware infrastructure, combined with NVIDIA's superior AI hardware and software portfolio, and VMware's industry-leading virtualization and cloud computing software, we are delivering a robust and comprehensive AI solution tailored for enterprise customers, addressing key requirements of security, scalability, performance, and flexibility. It empowers organizations to harness the power of AI to drive innovation, improve operational efficiency, build AI applications, and gain a competitive advantage today and into the future.

We look forward to engaging with your business strategists, AI practitioners, and solution architects to understand your industry use cases and requirements of your AI workloads, and help you design and implement an optimized Gen AI infrastructure solution leveraging Lenovo VMware Private AI Foundation solution.

For More Information

For more information, see these resources:

- VMware Private AI Foundation:
  - VMware Private AI Foundation with NVIDIA
  - General Availability Blog
  - Solutions Brief
- Lenovo AI solutions, and Lenovo platforms for VMware Private AI Foundation:
  - Reference Architecture for Generative AI Based on Large Language Models (LLMs)
  - Lenovo ThinkAgile VX650 V3 Integrated System and Certified Node
  - Lenovo ThinkAgile VX665 V3 Integrated System and Certified Node
  - Lenovo ThinkSystem SR675 V3 GPU Rack Server
  - Lenovo ThinkSystem SR670 V2 Rack Server
- Additional references:
  - Lenovo ThinkAgile VX Series for VMware

Authors

Carlos Huescas is the Worldwide Product Manager for NVIDIA software at Lenovo. He specializes in High Performance Computing and AI solutions. He has more than 15 years of experience as an IT architect and in product management positions across several high-tech companies.

Markesha Parker is the WW Technical Leader for ThinkgAgile VX Hyper-Converged solutions at Lenovo. In this role, she works to deliver quality solutions integrating Lenovo and third party hardware and software. She defines the technical requirements to enable manufacturing, support, and technical sales. Markesha has worked in the IT industry for over 17 years and is currently based in Morrisville, NC.

Chris Gully is a Master Solutions Architect on the Advanced Services team in the VMware Cloud Foundation Org at VMware by Broadcom. Chris has over 20 years of professional experience ranging from Systems Engineering to Cloud Services Product Management to leading technical teams. He has been focused on optimizing virtualization in the world of Advanced Computing and using VMware software solutions to enable the integration and proliferation of AI/ML workloads like GenAI, with an emphasis on business and customer outcomes. Previous roles at Dell, Oracle, Sun Microsystems, and several start-ups have prepared him for the speed at which technology and innovation must move to keep customers and businesses relevant. When not devising new ways of applying emerging technology to solve the problems of today and tomorrow, Chris likes “Keepin’ It Weird” with the live music scene in Austin, Texas and spending time with his family and friends.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- ThinkAgile VX Series for VMware
- ThinkSystem SR670 V2 Server
- ThinkSystem SR675 V3 Server
- VMware 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.

This document, LP2004, was created or updated on August 26, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2004
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2004.

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®
ThinkAgile®
ThinkSystem®

The following terms are trademarks of other companies:

AMD and AMD EPYC™ are trademarks of Advanced Micro Devices, Inc.

Intel® is a trademark of Intel Corporation or its subsidiaries.

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