Reference Architecture for Generative AI Based on Large Language Models (LLMs)

Reference Architecture

Executive summary

Lenovo is driven by two core principles: enabling universal access to advanced technology and establishing unparalleled trust as a partner in the era of intelligent transformation. These principles underscore our unwavering dedication to fostering innovation, empowering global partners and customers to harness, develop, and deploy AI on a large scale across diverse industries, with a steadfast commitment to safety and efficiency. Our current initiatives are centered on fortifying the infrastructure that underpins Generative AI Models (GenAI) and Large Language Models (LLMs), aiming to mitigate risks and elevate the support framework for these technologies.

In the landscape of commonly employed regression models, the unique characteristics of LLMs come to the fore. Unlike conventional regression models that consist of a modest number of parameters, LLMs challenge conventions with billions of parameters, thereby pushing the boundaries of hardware, software, networking, storage, input/output operations, and computational resources. The intricate process of training LLM models from scratch mandates purpose-built systems designed for optimal performance, often integrating potent accelerators such as GPUs. While deploying LLMs for inferencing might not demand the same level of resource intensity, it necessitates meticulous attention to code optimization. This optimization encompasses the fine-tuning of accelerator utilization, adept parallelization, and judicious core utilization.

In this document, we discuss the interaction between hardware and software stacks that are pivotal for the successful implementation of LLMs and GenAI. Furthermore, we will provide a comprehensive Bill of Materials (BoM) that can serve as the basis for building robust and high-performance AI infrastructure. By delving into these critical aspects, we aim to equip professionals like CIOs, CTOs, IT architects, system administrators, and those with an AI background with the knowledge and insights needed to navigate the complex landscape of AI-powered technologies effectively.

Introduction

Generative AI (GenAI) and large language models (LLM) involve algorithms that can generate new content based on patterns learned from existing data. The following definition of Generative AI is actually an example in itself - this text was created by Bing AI, a search engine integration of ChatGPT.

Generative AI is a form of AI that can create new and original content, such as text, images, audio, video, music, art, and code. GenAI is based on various technical foundations, such as: model architecture, self-supervised pre-training, and generative modeling methods.

- Model architecture: The structure and design of the neural networks that generate the content. Examples of model architectures include transformers, convolutional neural networks, recurrent neural networks, and attention mechanisms.
- Self-supervised pretraining: The process of training a model on a large amount of unlabeled data is to learn general features and patterns that can be transferred to specific tasks. Examples of self-supervised pretraining methods include masked language modeling, contrastive learning, and denoising autoencoders.
- Generative modeling methods: The techniques and algorithms that enable a model to learn the probability distribution of the data and sample new data from it.

Common models for Generative AI are generative adversarial networks (GANs) to generate multi-media and realistic speech, variational autoencoders (VAEs) for signal processing, autoregressive models, and diffusion models for waveform signal processing, multi-modal modeling, molecular graph generation, time series modeling, and adversarial purification.

Large language models are trained on a broad set of unlabeled data that can be used for different tasks and fine-tuned for purposes across many verticals. These models have billions of parameters, 65 billion to 540 billion for latest models, that require large numbers of accelerators, like GPUs, and extended time to train (e.g., BloombergGPT took 1.3 million hours of GPU time to train).

A recent key finding of language model research has been that using additional data and computational power to train models with more parameters consistently results in better performance. Consequently, the number of parameters is increasing at exponential rates and as such puts enormous strain on training resources, which makes pre-trained models attractive. Pre-trained models, either open-source or commercial, only require fine tuning for the specific use case and can be deployed relatively quickly for inferencing.

Lenovo chatbot

Lenovo has started using a pre-trained LLM, specifically Llama 2 from Meta, to drive a chatbot that helps our sales community quickly find technical and esoteric details regarding our hardware. The process by which this was created is outlined in the figure below.

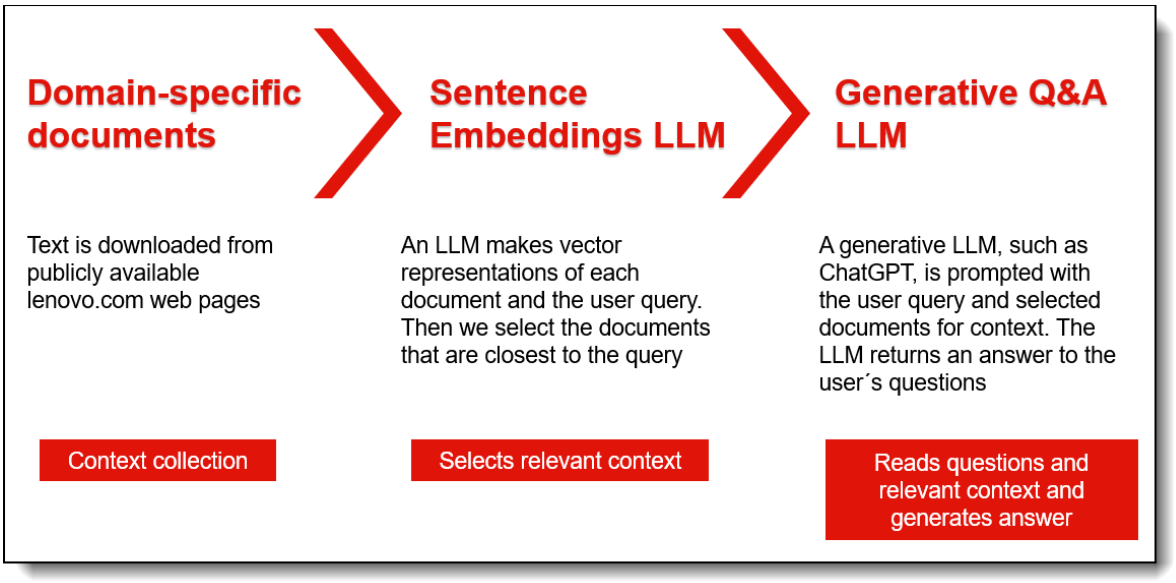

Figure 1. Process to create the Lenovo LLM chatbot

The results of the process are illustrated in the 30-second video embedded below. This bot references content from Lenovo.com, which provides some context for the responses. For more control over the responses, completely air-gapped data, stored in secure data center storage, offers additional protection.

Chatbots, like the Lenovo LLM Chatbot, have potential end uses that include question answering over warranty and service manuals for a comprehensive understanding, an online shopper asking for simple explanations of tech specs on a webpage, or a customer seeking quick online product support. With the right documents for context, a custom Q&A chatbot can provide users an easy and informative experience.

The current state of Generative AI and LLMs

The content below was written by Bing AI using the phrase “what is Generative AI” and separately using the phrase “current state of Generative AI, LLM”. The purpose of showing this is to illustrate how comprehensive and effective a response can be provided using LLMs.

Generative AI is a rapidly evolving and expanding field that offers many opportunities and challenges for businesses and society. Some of the benefits of generative AI include:

- Enhancing creativity and innovation: Generative AI can augment human creativity by providing new ideas, insights, and perspectives that can inspire novel solutions and products.
- Improving efficiency and productivity: Generative AI can automate tedious and repetitive tasks that are time intensive and require significant resources, such as data collection, data labeling, data analysis, and content creation.
- Increasing accessibility and diversity: Generative AI can make content and services more accessible and inclusive by generating personalized and customized content that meets the needs and preferences of different users and audiences.

Some of the challenges of generative AI include:

- Ensuring quality and reliability: Generative AI can produce inaccurate or misleading content that can harm the reputation or performance of a business or an Examples of quality and reliability issues include factual errors, logical inconsistencies, grammatical mistakes, and plagiarism.
- Maintaining security and privacy: Generative AI can pose risks to the security and privacy of data and systems by enabling malicious actors to generate fake or spoofed content that can deceive or manipulate users or bypass authentication mechanisms. Examples of security and privacy issues include identity theft, fraud, phishing, deepfakes, cyberattacks, and data breaches.
- Regulating ethics and responsibility: Generative AI can raise ethical and social questions about the ownership, accountability, transparency, explainable, fairness, trustworthiness, and impact of the generated content on individuals and society.

The output from Bing's Chatbot always include references for the content provided. The references for these results are below.

- Papers with Code - A Complete Survey on Generative AI (AIGC): Is https://paperswithcode.com/paper/a-complete-survey-on-generative-ai-aigc-is
- What is ChatGPT, DALL-E, and generative AI? | McKinsey https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai
- The state of generative AI in the enterprise | survey results https://writer.com/guides/generative-ai-survey/
- KPMG Generative AI https://info.kpmg.us/news-perspectives/technology-innovation/kpmg-generative-ai-2023.html
- Salesforce survey shows IT interest in generative AI tempered with technical, ethical concern https://siliconangle.com/2023/03/06/salesforce-survey-shows-interest-generative-ai-tempered-technical-ethical-concerns/

There are also challenges of concern when it comes to Generative AI. These concerns include:

- **High energy usage**: An analysis has shown that training an LLM model with 200 billion parameters produces approximately 75,000 kg of CO_2 emissions, compared to only 900 kg CO_2 emissions for a flight from NYC to San Francisco. For details of the analysis, see https://huggingface.co/learn/nlp-course/chapter1/4?fw=pt.
- **Model bias**: Presence of systematic and unfair inaccuracies or prejudices in the predictions or decisions made by a machine learning model.
- **Toxic comments**: A toxic comment is a text-based input, usually in the form of a comment, message, or text snippet, that contains harmful, offensive, or inappropriate content. Specialized training processes, such as reinforcement learning from human feedback (RLHF), have been developed to prevent toxic comments from LLMs.
- **Hallucination**: Hallucination refers to Generative AI responses that are produced when the search context changes, or a response is not supported by the underlying data. Hallucination results in query

responses that are illogical and deceptive.

- **Conversational AI leakage**: Conversational AI leakage occurs when sensitive data is input into a LLM and is unintentionally exposed. The inadvertent exposure of sensitive data in LLMs, raises significant concerns due to privacy violations, potential data breaches, erosion of trust, legal repercussions, competitive disadvantages, and ethical considerations. Protecting against such leakage is essential to maintain user trust, uphold legal compliance, prevent misuse of information, and safeguard reputations. Addressing this concern involves technical improvements in model behavior, robust data handling processes, and adherence to data privacy and security best practices.

These concerns are an active area of research and *Responsible Generative AI* guidelines are only now emerging. At Lenovo, we create rules engines and other AI models for preventing these concerns.

Lenovo is deeply concerned that AI be developed and used consistently with our values. We have developed the Lenovo Responsible AI Committee to ensure our solutions and those of our AI innovator partners meet requirements that protect end users and ensure that AI is used fairly, ethically, and responsibly.

The difference between Generative AI and LLMs

Generative AI, large language models and foundation models are similar, but different and are commonly used interchangeably. There is not a clear demarcation between terms, and this becomes challenging when a needed delineation is required. Where distinction in terms is required, what the intent is and is not serves as a guide.

The following definitions clarify the difference and are excerpts from the following article from Georgetown University Center for Security and Emerging Technology:

What Are Generative AI, Large Language Models, and Foundation Models?
https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

"*Generative AI* is a broad term that can be used for any AI system whose primary function is to generate content. This is in contrast to AI systems that perform other functions, such as classifying data (e.g., assigning labels to images), grouping data (e.g., identifying customer segments with similar purchasing behavior), or choosing actions (e.g., steering an autonomous vehicle).

"Typical examples of generative AI systems include image generators (such as Mid-journey or Stable Diffusion), large language models, code generation tools (such as Copilot), or audio generation tools (such as VALL-E or resemble.ai)."

"Large language models (LLMs) are a type of AI system that works with language. The LLM aims to model language, i.e., to create a simplified—but useful—digital representation. The "large" part of the term describes the trend towards training language models with enormous number of parameters. Typical examples of LLMs include OpenAI's GPT-4, Google's PaLM, and Meta's LLaMA."

Large language models perform well for content and code generation, translations, content summarization, and chatbots. Foundation models, which are not constrained to language, are trained models on enormous data that is adapted to many applications. GPT-4, PaLM, LLaMA are examples of foundational LLMs. Foundation models can be used as the starting point for retraining to a specific use case, which saves on compute, infrastructure needs, resources, and training time.

Both Generative AI and LLM models extract value from enormous data sets and provide straightforward learning in an accessible manner. Details on the more commonly used pre-trained LLMs (foundation models) are provided below.

- ChatGPT, by OpenAI, can generate an answer to almost any question it's asked and is free to use. ChatGPT using the GPT-4 LLM, released in 2023, is the latest version as of this writing and is unique in that it is multimodal, which can process a combination of image and text inputs to produce text outputs.
- DALL-E2, developed by OpenAI, creates AI-generated images that are realistic and art from

descriptive natural language. DALL-E2 can merge styles and concepts in new creative ways.

- LLAMA 2, by Meta, is open-source LLM series based on up to 70 billion parameters and trained on 2 trillion tokens. Smaller versions exist that can be fine-tuning for a variety of tasks. This model is open source and available for research and commercial purposes.
- BLOOM, by BigScience, is an open-source multilingual large language model, which can generate text in 46 languages and 13 programming languages. Soon, an inference API will be released.
- LaMDA 2, language model for dialogue applications by Google, is an advanced AI-driven chatbot with conversational skills trained on dialogue and is built on a transformer neural network architecture.
- MPT-7B-Instruct is built on a modified decoder-only transformer architecture and has ~ 7 billion parameters trained on 1 trillion tokens from a variety of text datasets. It is open-source, which can be used commercially and is a model for short-form instruction following.

Some of these models are subjected to common performance benchmarks (like MLPerf Training 3.0 and MMLU benchmark) and are limited by the freshness of the data upon which they were trained.

Use cases

The landscape of Generative AI and LLM use cases are illustrated in the following figure. This figure shows that starting with various inputs: audio, video, text, code, etc. enhanced outputs can be produced and converted to other forms like chatbots, translations, code, avatars, etc.

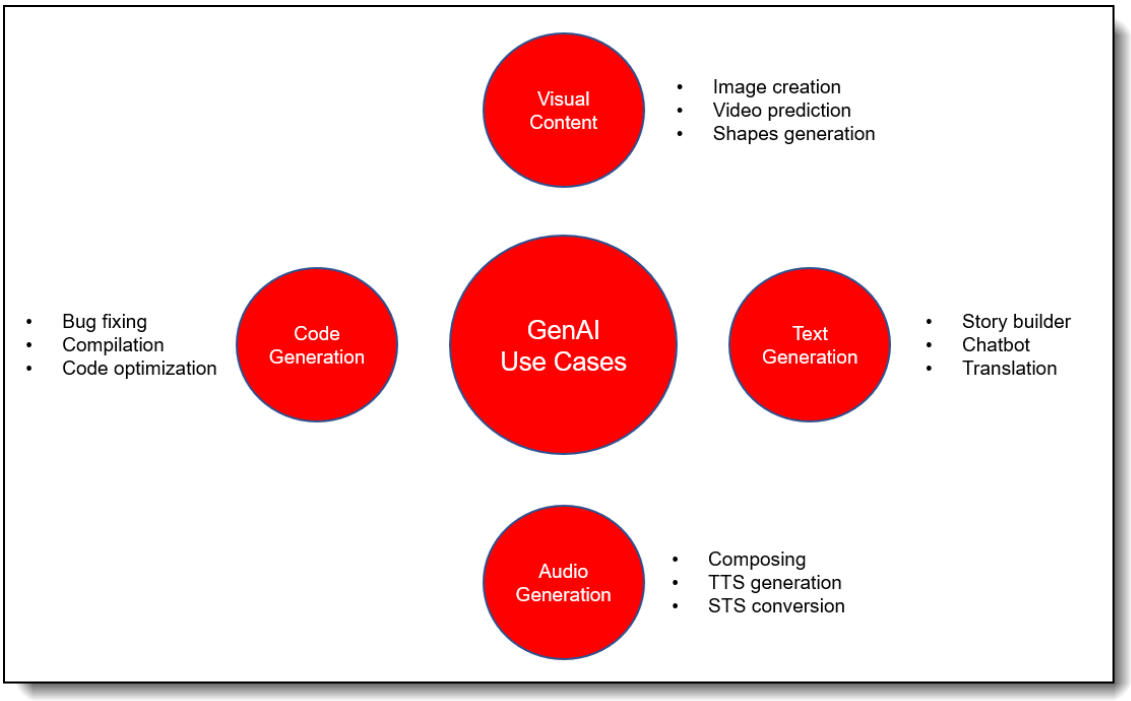

Figure 2. Generative AI and current use cases

These use cases cross most verticals: healthcare, manufacturing, finance, retails, telecommunications, energy, government, technology, and drive significant business impact.

Large language models can translate text into various languages, classify and organize customer feedback for an improved experience, summarize legal documents and earning calls, and create innovative marketing content. Models like, ChatGPT, can create sales pitches, create ad copy, find computer code bugs, create blog posts, and draft customer support emails. Most of the value that Generative AI provides falls into customer operations, marketing and sales, software engineering and R&D, Life Sciences, Finance, and High Tech are the industries projected to see the biggest impact as function of revenue by using Generative AI.

Use cases:

- Life Science use cases
- Finance use cases
- Manufacturing use cases
- Operations use cases
- Other use cases

Life Science use cases

Life Science use of Generative AI and LLM is expanding and offering great promise for foundational research. These models are being used to generate images of biological structures and processes facilitating enhanced understanding. Gene expression profiles of individual cells are now being understood using GenAI models. Denoising raw gene expression data is accomplished using these models. Data imputation and synthetic data generation, commonly used in life science and other verticals, can be provided by GenAI and LLM.

Creating this computer-generated data is important for more accurate models and helps with the study of rare diseases where real-world large datasets do not exist. GenAI models can be used to create novel protein sequences with specific properties and functionalities. These models can predict protein structures, which facilitates new gene therapies, and are helpful for protein engineering, development of novel therapeutics and enzyme design.

Generative adversarial networks (GAN) and variational autoencoders (VAE) are types of generative models that are being used to develop new drug like molecules that target binding affinity of target proteins. Other life science use cases that utilize generative AI are drug discovery, medical imaging, and personalized medicine.

Finance use cases

The Financial services industry is already adopting Generative AI models for certain financial tasks. These institutions are using these models because of their ability to enhance efficiency, improve customer experiences and reduce operational costs. These methods are especially adept at automating services, other financial processes, and decision-making. An example of which is Morgan Stanley's use of OpenAI-powered chatbots to support financial advisors by drawing upon the firm's internal collection of research and data as a knowledge repository. BloombergGPT is a LLM specifically trained on finance data and is capable of sentiment analysis, news classification and other financial tasks.

Much like other industry verticals, synthetic data generation and risk modelling using Generative AI are commonly employed in finance. The use of natural language processing, language understanding, and language generation are widely employed to create conversational responses to customer queries and to personalize their experience. Additionally, finance is using Generative AI's ability to create code to address the challenge they face with legacy systems that are based on outdated languages. These no longer supported languages can be replaced with contemporary supported code that can run current applications and support modern deployment methods.

Forecasting, document analysis and financial analysis is a task well performed by Generative AI models, which analyze, summarize, and extract new insights from historical data. This approach can understand complex relationships and patterns to make predictions about future trends, asset prices and economic indicators. By using properly fine-tuned Generative AI models, reports can be generated, fraud detected, and nefarious patterns identified. These tuned models can generate various scenarios by simulating market conditions, macroeconomic factors, and other variables, providing valuable insights into potential risks and opportunities.

Manufacturing use cases

Manufacturing generative AI use cases have some similarities to those of Life Science and Finance as it pertains to denoising raw data and producing synthetic data for improved model performance. Generative AI enables industries to design new parts that are optimized to meet specific goals and constraints like performance and manufacturing. Using these model types, engineers can analyze large data sets to help improve safety, create simulation datasets, explore how a part might be manufactured or machined faster, and bring products to market more quickly. For factory operations, it is early days for generative AI use cases; however, these methods can help to optimize overall equipment effectiveness and serve to provide an effective method to “read” repair manuals, service bulletins and warranty claims for new insights and quicker problem resolution.

In manufacturing, significant bottlenecks exist where legacy systems and traditional management operations are in place. Generative AI can transform data insights to drive operations, whether they are organizational or on the factory floor. These methods help companies overcome data-quality barriers and unleash the full potential of AI in manufacturing while structuring, cleaning up, and enriching existing data.

Operations use cases

For operations, Generative AI models can help optimize supply chains, improve demand forecasting, provide better supplier risk assessments, and improve inventory management. Generative AI can analyze large amounts of historical sales data, incorporating factors such as seasonality, promotions, and economic conditions. By training an AI model with this data, it can generate more accurate demand forecasts. This helps businesses better manage their inventory, allocate resources, and anticipate market changes.

For supply chain optimization, Generative AI models can perform data analysis on various sources, such as traffic conditions, fuel prices, and weather forecasts, to identify the most efficient routes and schedules for transportation. These models can generate multiple possible scenarios, and based on the desired optimization criteria, they can suggest the best options for cost savings, reduced lead times, and improved operational efficiency across the supply chain.

For supplier risk assessment, generative AI models can identify patterns and trends related to supplier risks by processing large volumes of data, including historical supplier performance, financial reports, and news articles. This helps businesses evaluate the reliability of suppliers, anticipate potential disruptions, and take proactive steps to mitigate risk, such as diversifying their supplier base or implementing contingency plans.

For inventory management, generative AI models can analyze demand patterns, lead times, and other factors to determine the optimal inventory levels at various points in the supply chain. By generating suggestions for reorder points and safety stock levels, AI can help warehouse management by minimizing stockouts, reducing excess inventory, and lowering carrying costs.

Other use cases

Generative AI has many applications and use cases across various domains and industries, other examples are as follows:

- **Content generation and editing**: These models can generate or modify multimedia content that is realistic, diverse, and creative. Examples of content generation and editing applications include art, music, 3D models, audio, video, and synthetic data.
- **Application development**: Generative AI can support code development and deployment. Example uses in application development include code generation, code analysis, code completion, and code testing.
- **Cybersecurity**: The task of protecting data and systems from malicious attacks or unauthorized access can be accomplished using Generative AI. Examples of cybersecurity applications include anomaly detection, malware detection, intrusion detection, and encryption.

These use cases can solve previously unsolved mathematical theorems leading to new physics, engineering, and statistical methods.

Generative AI models can generate new candidate chemicals, molecules, and materials. These models along with computational chemistry are used to develop new materials by analyzing unstructured data as well as structured data that exist in virtual chemical databases, which contain billions of identified and characterized compounds. The vast size of these repositories, even when constrained to molecular data, has been intractable to fully research. Recent developments in AI technology based on pre-trained language models and Generative Adversarial Networks (GANs), have been applied to materials discovery.

Performance considerations

Generative AI and large language models are extremely computationally intensive. Significant improvements in workload performance and usage cost for compute resources can be gained by using optimized software, libraries, and frameworks that leverage accelerators, parallelized operators and maximize core usage. There are many approaches that are used to address these challenges on the transformer architecture side, modeling side, and on the code deployment side.

The transformer-based architecture uses a self-attention mechanism, which enables a LLM to understand and represent complex language patterns more effectively. This mechanism increases the parallelizable computations, reduces the computational complexity within a layer, and decreases the path length in long range dependencies of the transformer architecture.

For efficient resource utilization during training and deployment, modeling should have appropriate data parallelism and model parallelism. Algorithmic considerations to speed up applications with parallel processing is involved with data parallelization. This approach increases performance and accuracy especially when using PyTorch Distributed Data Parallel to spread workloads across GPUs. Model parallelism is accomplished by using techniques of activation checkpointing and gradient accumulation to overcome the challenges associated with large model memory footprints.

On the code deployment side, the limiting factor is bandwidth when moving weights and data between compute units and memory. A compute unit (CU) is a collection of execution units on a graphics processing unit (GPU) that can perform mathematical operations in parallel. Optimizing the use of compute units and memory is required to run these models efficiently, quickly and to maximize performance.

Complex network architectures, such as the ones discussed in the architectural overview section, challenge efficient real-time deployment, and require significant computation resources and energy costs. These challenges can be overcome through optimizations such as neural network compression. There are two types of network compression: pruning and quantization. There are various methods to accomplish pruning, whose goal is to remove redundant computations. Quantization reduces precision of the data types to achieve reductions in computations. For these model architectures, 8-bit integers are typically used for weights, biases, and activations.

Mixed precision increases speed with the goal of intelligently managing precision while maintaining accuracy and gaining performance from smaller faster numerical formats.

Hardware stack

As previously highlighted, the demanding nature of Generative AI and Large Language Models (LLMs) necessitates a meticulously crafted solution. Each facet of this solution was scrutinized to optimize performance. Attention was devoted to tackling areas that typically introduce latencies, and the architecture was conceived with a holistic view, considering the collaborative impact of all components. This has led to the creation of an architecture featuring finely tuned components that precisely cater to these rigorous requirements.

The following illustrates the cornerstone of this architecture—the primary building block built on the foundation of the Lenovo ThinkSystem SR675 V3 AI ready server, equipped with 8 NVIDIA H100 NVLink (NVL) GPUs.

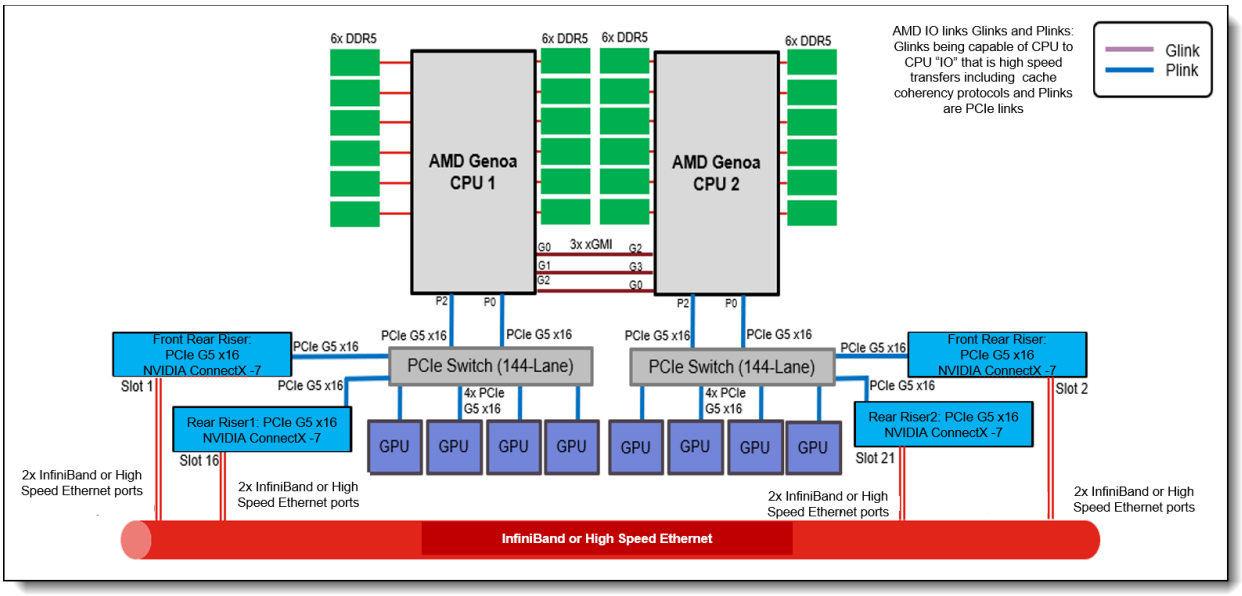

Figure 3. Primary hardware building blocks

To ensure optimal performance, all elements are seamlessly connected through an NDR InfiniBand fabric or High-Speed Ethernet. Notably, this configuration remains applicable when deploying latest NVIDIA L40S GPUs, NVIDIA BlueField-3.

This reference architecture follows the philosophy of our approach for Lenovo EveryScale solutions where customer can start from simple and scale depending on their needs.

Lenovo EveryScale provides Best Recipe guides to warrant interoperability of hardware, software, and firmware among a variety of Lenovo and third-party components. Addressing specific needs in the data center, while also optimizing the solution design for application performance requires a significant level of effort and expertise. Customers need to choose the right hardware and software components, solve interoperability challenges across multiple vendors, and determine optimal firmware levels across the entire solution to ensure operational excellence, maximize performance, and drive best total cost of ownership.

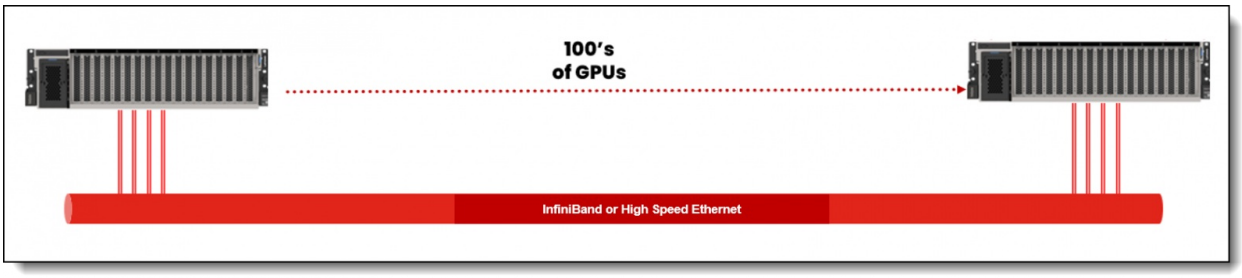

Figure 4. Reference Architecture's scalability

A pivotal inclusion in this solution is NVIDIA AI Enterprise a high-performance, secure, cloud-native AI software platform built to maintain a consistent, secure, and stable software used in creating and deploying AI models.

Topics in this section:

- GPUs
- Server configuration
- Networking connectivity
- Management Ethernet switches

GPUs

Generative AI, characterized by its ability to create new data instances that resemble existing ones, has revolutionized various fields such as image synthesis, text generation, and even drug discovery. The significance of Graphics Processing Units (GPUs) in this context is multifaceted, owing to their unique capabilities that align perfectly with the requirements of generative models.

Here's why GPUs are pivotal for driving the advancement of generative AI:

- **Parallel Processing Power**: One of the key features of GPUs is their exceptional parallel processing capability. Generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), involve complex mathematical operations and iterative processes. GPUs excel at performing these operations in parallel, significantly accelerating training times and enabling the exploration of larger, more complex models.
- **High Computational Throughput**: The massive number of cores within a GPU allows it to handle a high volume of computations simultaneously. This is particularly beneficial for generative AI, where model architectures can consist of numerous layers and parameters. The computational throughput of GPUs accelerates model training and inference, resulting in faster iterations and quicker results.
- **Complex Neural Architectures**: Modern generative models often comprise intricate neural network architectures with numerous layers and connections. GPUs excel at efficiently handling these complex structures, enabling researchers and engineers to experiment with diverse model architectures to achieve better performance.
- **Large-Scale Data Processing**: Generative AI models often require extensive datasets for training. GPUs facilitate the processing of large-scale datasets by efficiently distributing the computational load across their many cores. This enables faster data preprocessing, augmentation, and model training.
- **Real-Time Inference**: In applications like image synthesis or style transfer, real-time or near-real-time performance is crucial. GPUs are optimized for high-performance parallel computation, allowing them to process incoming data and generate responses rapidly, making them well-suited for real-time generative applications.
- **Transfer Learning and Fine-Tuning**: Many generative AI tasks benefit from transfer learning, where pre-trained models are fine-tuned for specific tasks. GPUs enable rapid fine-tuning by efficiently updating model weights and parameters based on new data.
- **Hardware Acceleration for Optimized Models**: With the increasing complexity of generative models, efficient hardware acceleration becomes essential. GPUs offer hardware-level optimizations for neural network operations, enhancing both training and inference efficiency.
- **Scalability**: The parallel nature of GPUs lends itself well to scalability. Researchers and organizations can leverage multiple GPUs or even GPU clusters to further accelerate training and inference, allowing them to tackle larger and more ambitious generative AI projects.

In summary, GPUs play a pivotal role in the advancement of generative AI by providing the computational power, parallel processing capability, and hardware acceleration required to train and deploy complex models efficiently. Their ability to handle large-scale data and complex neural architectures makes them a crucial tool for researchers and practitioners aiming to push the boundaries of generative AI technology.

In this document, we will use to GPU options from NVIDIA, NVIDIA H100 and NVIDIA L40S. The H100 is recommended for traditional and highly Intensive AI workloads and the L40S for those workloads that requires both graphics and mainstream AI workloads

NVIDIA H100

The NVIDIA H100 Tensor Core GPU is the next-generation high-performing data center GPU and is based on the NVIDIA Hopper GPU architecture. A primary driver for this GPU is to accelerate AI training and inference, especially for Generative AI and LLMs.

The H100 securely accelerates workloads from small enterprise scale, to exascale HPC, to trillion parameter AI model training. This GPU has 80 billion transistors customized for NVIDIA. The is available in two H100 GPU form factors, PCIe and SXM; the SXM variant is shown in the following figure.

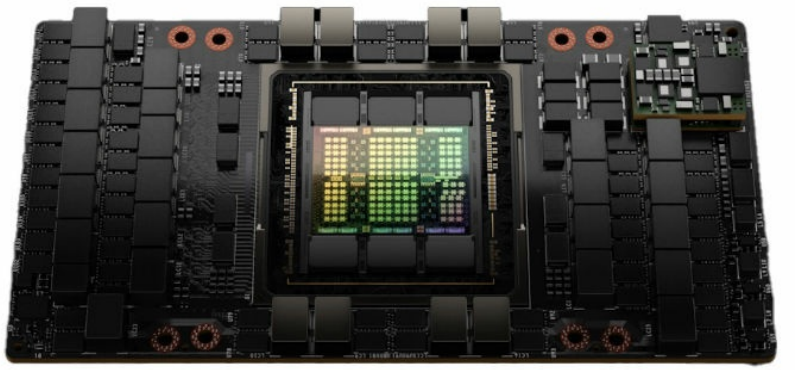

Figure 5. NVIDIA H100 SXM GPU

There are substantial performance enhancements when comparing NVIDIA H100 GPU with previous generation A100 GPU as shown in the following figure.

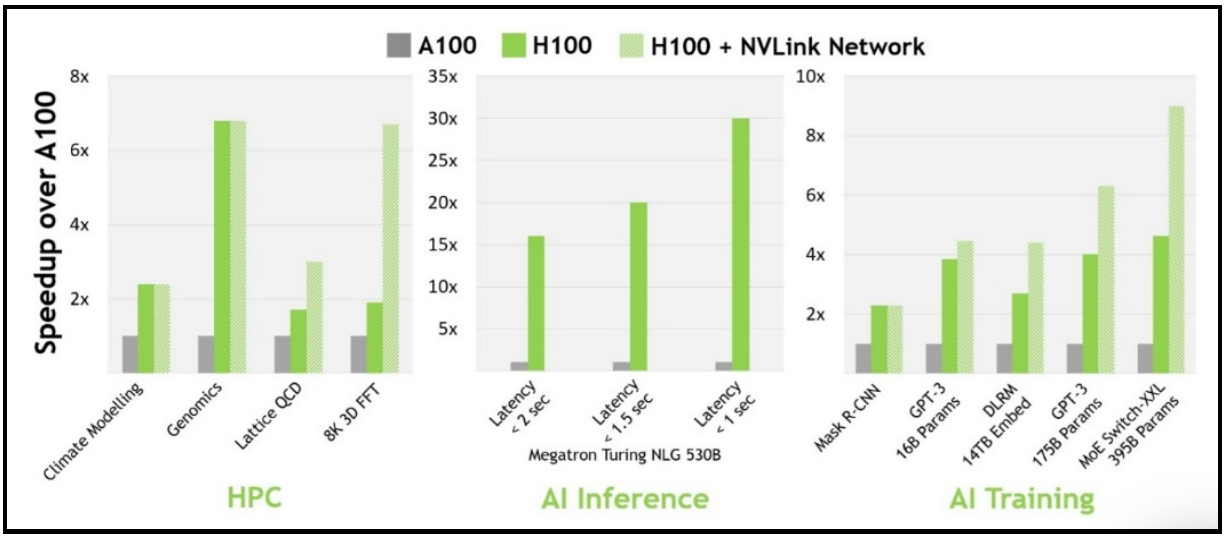

Figure 6. NVIDIA H100 performance comparison

NVIDIA L40S

The NVIDIA L40S is a powerful universal GPU for the data center, delivering end-to-end acceleration for the next generation of AI-enabled applications, from GenAI model training and inference to 3D graphics to media acceleration. The L40S's powerful inferencing capabilities combined with NVIDIA RTX accelerated ray tracing and dedicated encode and decode engines, accelerate AI-enabled audio, speech, 2D, video and 3D Generative AI applications.

Figure 7. NVIDIA L40S GPU

When compared to the A100, the L40S has 18,176 NVIDIA Lovelace GPU CUDA cores that produce a 5x improved single precision floating point (FP32) performance. In addition, the L40S enables two FP32 data paths doubling the peak FP32 operations. For mixed precision workloads, enhanced 16-bit math capabilities are available.

The L40S has 568 4th generation NVIDIA Tensor Cores that include NVIDIA's Transformer Engine and new FP8 data format. A 2x improvement over previous generations Tensor Cores is achieved for tensor matrix operations. This is due to the L40S Tensor Cores being able to accelerate more data types while still supporting fine-grained structured sparsity feature. The L40S Transformer Engine dramatically accelerates AI performance and improves memory utilization for both training and inference. It also intelligently scans the layers of transformer architecture neural networks and automatically re-casts between FP8 and FP16 precisions to deliver faster AI performance and accelerate training and inference.

For MLPerf models, eight L40S in a mainstream server allow for a 0.8x increase in the training performance when compared to an A100 8-GPU system.

In the figure below, L40S performance is compared to the NVIDIA A100 for two types of common GenAI models, GPT-408 LoRA and Stable Diffusion.

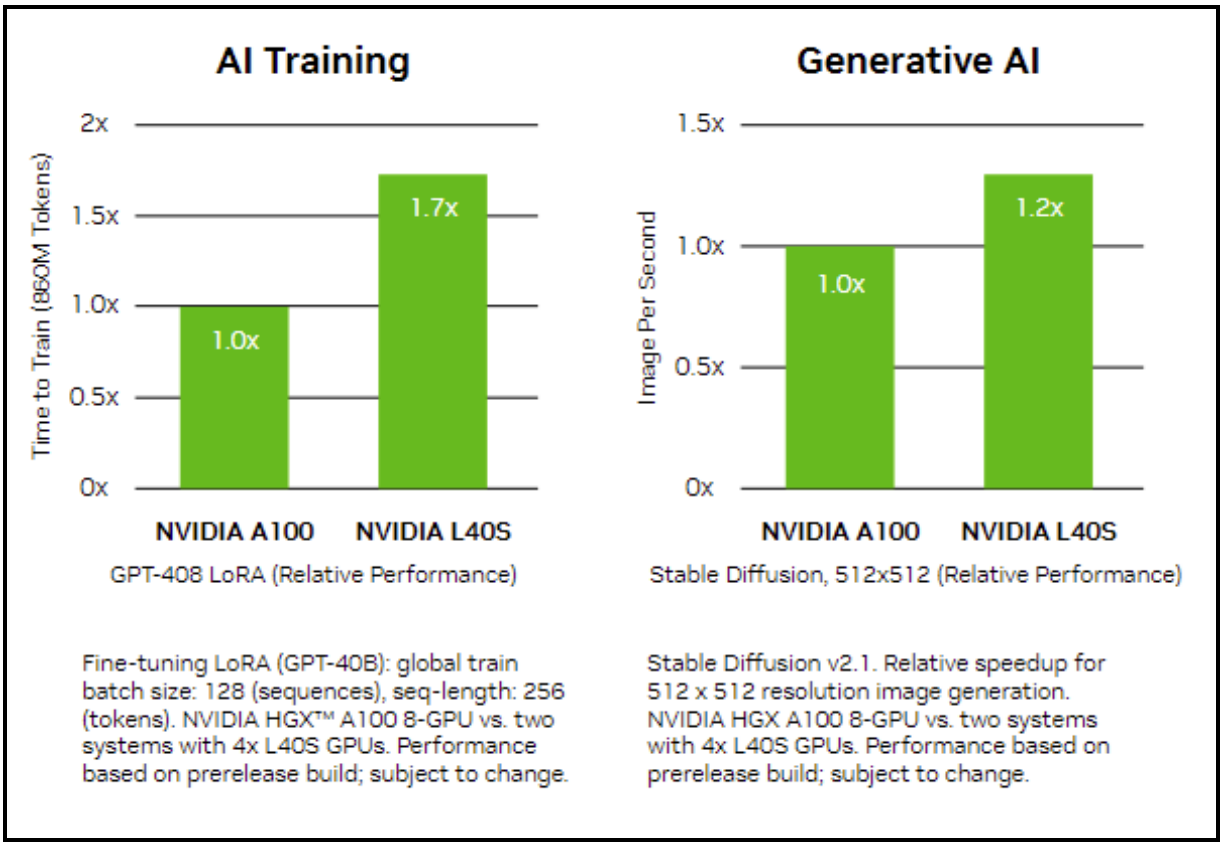

Figure 8. NVIDIA L40S performance comparison

Server configuration

This Lenovo reference architecture is built on the best-in-class x86 server line in the market, our servers are suitable for a broad range of customers, from small businesses and line-of-business applications, technical/finance customers with HPC workloads, through to multi-national corporations with mission-critical workload needs.

For Generative AI it is important to find a balance between performance, power consumption and scalability. Among our ThinkSystem server line we can meet that balance on our ThinkSystem SR675 V3 AI ready node.

SR675 V3 with H100 or L40S GPUs

The Lenovo ThinkSystem SR675 V3 is a versatile GPU-rich 3U rack server that supports eight double-wide GPUs including the new NVIDIA H100 and L40S Tensor Core GPUs, or the NVIDIA HGX H100 4-GPU offering with NVLink and Lenovo Neptune hybrid liquid-to-air cooling. The server is based on the new AMD EPYC 9004 Series processors (formerly codenamed "Genoa", “Genoa-X” and “Bergamo”).

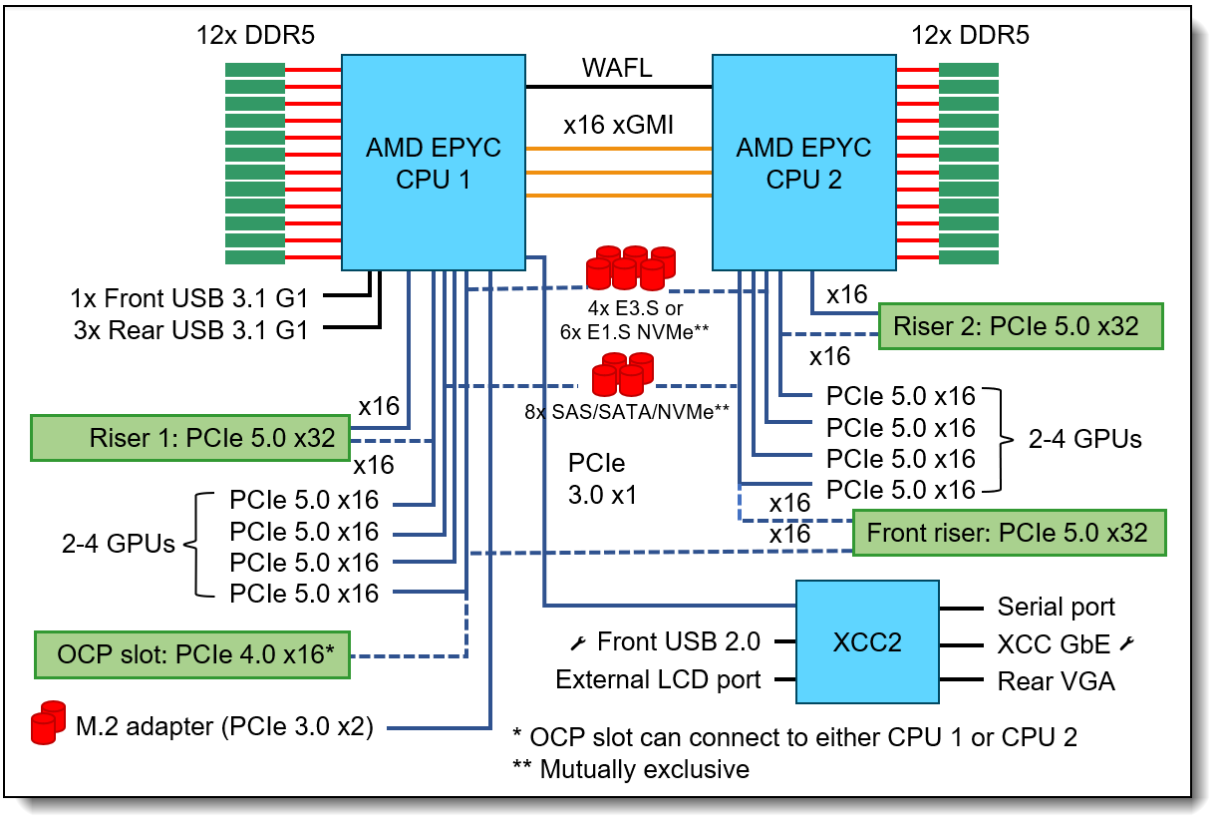

Figure 9. SR675 V3 system architectural block diagram

The SR675 V3 features a modular design for ultimate flexibility. Multiple configurations are supported, including:

- One or two 4th Generation AMD EPYC™ processors
- Up to eight double-wide GPUs with NVLink bridges
- NVIDIA HGX H100 4-GPU with NVLink and Lenovo Neptune hybrid liquid cooling
- Choice of front or rear high-speed networking
- Choice of local high speed NVMe storage

There are three different base configurations of the SR675 V3 as shown in the following figure. The configurations determine the type and quantity of GPUs supported as well as the supported drive bays.

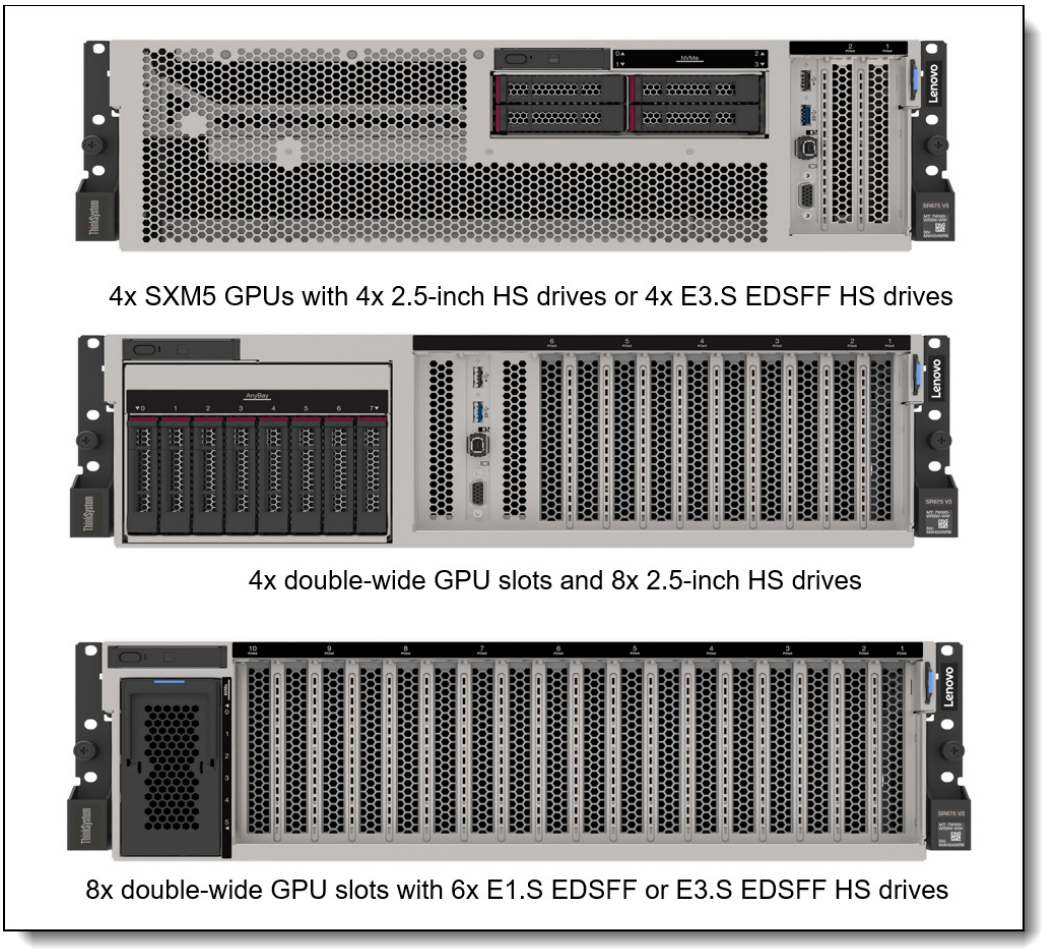

Figure 10. ThinkSystem SR675 V3 available configurations

It is important that you refer to the Deployment options section to have deep understanding on the server configuration that we choose for this reference architecture. However, you can use the following recommendation for maximum benefits. This configuration needs to be replicated on each node.

- 2x AMD EPYC 9634 processors, 84C, 2.25GHz
- 24x 128GB TruDDR5 RDIMMs
- 2x 1.92TB NVMe SSDs
- 6x 7.68TB NVMe SSDs, this storage configuration is dependent on your storage needs and can be modified.
- 1x 1G 4-port OCP Adapter – Management Connector
- 8x NVIDIA H100 NVL or NVIDIA L40S
- 4x NVIDIA ConnectX-7 200 GbE or NDR dual port
- 1x NVIDIA ConnectX-7 200 GbE or NDR dual port (optional network storage)
- 1x NVIDIA ConnectX-6 Lx 10/25GbE, this card is only needed when deploying a Virtual environment. Refer to Deployment to a virtual environment for details.
- 4x 2400W power supplies (230V)

SR675 V3 OVX L40S Node

The evolution and broader adoption of the Generative AI will require not only a high-performance GPU but also graphic capabilities. Our Lenovo OVX L40S option is facilitating businesses with the capability to refine foundational generative AI models. This server config empowers enterprises to seamlessly customize and implement generative AI applications, encompassing cutting-edge functionalities such as intuitive chatbots, advanced search systems, and efficient summarization tools.

This server config is different to the SR675 V3 described in the previous section, integrating the NVIDIA BlueField-3 DPU card and the L40S as the standard GPU selection. With this config we can now empower Generative Ai workloads but also the inclusion of Metaverse applications.

Furthermore, our servers are equipped with the potent prowess of NVIDIA-accelerated infrastructure and software, elevating the potential of Virtual Private AI Foundation in synergy with NVIDIA's exceptional technsology.

Below you have the recommend server configuration for an SR675 V3-based OVX L40S node.

- 2x AMD EPYC 9634 processors, 84C, 2.25GHz
- 12x 128GB TruDDR5 3DS RDIMMs
- 2x 1.92TB NVMe SSDs
- 6x 7.68TB NVMe SSDs, this storage configuration is dependent on your storage needs and can be modified.
- 1x 1GbE 4-port OCP Adapter – Management Connector
- 4x NVIDIA L40S
- 2x NVIDIA ConnectX-7 200 GbE dual port
- 1x NVIDIA BlueField-3 200 GbE dual port DPU (North to South connectivity - Storage integration)
- 4x 2400W power supplies (230V)

This configuration needs to be replicated on each node.

We provide an example of our Reference Architecture using this SR675 V3 OVX L40S node in Appendix 1: NVIDIA OVX L40S Nodes Integration.

Networking connectivity

Ethernet and InfiniBand are both popular networking technologies used in high-performance computing (HPC) environments, including generative AI clusters.

The choice between Ethernet and InfiniBand depends on various factors, and there are cases where Ethernet might be a better option:

- Cost-Effectiveness: Ethernet is generally more cost-effective compared to InfiniBand. If budget constraints are a significant concern, Ethernet can be a more practical choice for deploying a generative AI cluster.
- Compatibility: Ethernet is more commonly found in standard data center infrastructure, making it easier to integrate with existing networking equipment and systems. If your organization's infrastructure is already based on Ethernet, transitioning to InfiniBand might require significant changes.
- Simplicity of Setup: Ethernet setups are typically easier to configure and manage compared to InfiniBand, which might require specialized knowledge and additional effort to set up and maintain.
- Moderate Workloads: If your generative AI workloads are not extremely demanding in terms of low-latency communication and high-speed data transfer, Ethernet can still provide adequate performance without the complexities of InfiniBand.
- Long-Distance Communication: Ethernet is more suitable for longer-distance communication due to

its widespread availability and support for various topologies. InfiniBand's benefits become more pronounced when dealing with short-range, high-speed communication within a data center.

- Standard Software Stack: If your generative AI applications rely on standard software libraries and frameworks that are optimized for Ethernet, making the switch to InfiniBand might not yield significant performance improvements.
- Interoperability: Ethernet offers better interoperability with a wide range of devices and platforms. InfiniBand might require additional effort to ensure compatibility with various hardware components.
- Scalability: If you're planning to scale your generative AI cluster gradually and want to keep the option open for future expansion, Ethernet might be a more flexible choice.

InfiniBand typically excels in scenarios where ultra-low latency and extremely high data transfer rates are critical, such as large-scale simulations, weather forecasting, and certain scientific research applications. However, for many generative AI workloads, Ethernet can provide sufficient performance while being more straightforward to implement and manage.

Ultimately, the decision should be based on your specific requirements, existing infrastructure, budget, and the expertise available within your organization to manage and optimize the chosen networking technology.

Adapter selection

Depending on your selection for InfiniBand or Ethernet you may select the appropriate HCA options for you server.

For this reference architecture we are using 2 options:

- ConnectX-7 for both InfiniBand or Ethernet implementation
- BlueField 3 + ConnectX-7 for the OVX L40S config

A summary of each adapter is presented below.

NVIDIA ConnectX-7 supports both 200 GbE/s and NDR400. These InfiniBand adapters provide the highest networking performance available and are well suited for the extreme demands of Generative AI and LLM. These adapters provide ultra-low latencies and higher throughputs. The acceleration engines have collective operations, MPI All-to-All, MPI tag matching, and programmable data path accelerators.

NVIDIA BlueField-3 data processing unit (DPU) is the 3rd-generation infrastructure compute platform that enables organizations to build software-defined, hardware accelerated IT infrastructures from cloud to core data center to edge. With 400Gb/s Ethernet or NDR 400Gb/s InfiniBand network connectivity, BlueField-3 DPU offloads, accelerates, and isolates software-defined networking, storage, security, and management functions in ways that profoundly improve data center performance, efficiency, and security.

Providing powerful computing, and a broad range of programmable acceleration engines in the I/O path, BlueField-3 is perfectly positioned to address the infrastructure needs of the most demanding applications, while delivering full software backward compatibility through the NVIDIA DOCA™ software framework.

InfiniBand switches

The InfiniBand switches used in the reference architecture are the NVIDIA QM9700 and QM9790 Quantum-2-based switch systems that provide 64 ports of NDR 400Gb/s InfiniBand per port in a 1U standard chassis. A single switch carries an aggregated bidirectional throughput of 51.2 terabits per second (Tb/s) with more than a 66 billion packets per second (BPPS) capacity.

These support the latest NDR technology providing a high-speed, ultra-low latency, and scalable solution that incorporates Remote Direct Memory Access (RDMA), adaptive routing, and NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP).

NVIDIA InfiniBand provides self-healing network capabilities and quality of service, enhanced virtual lane mapping and congestion control to provide the highest overall application throughput. These switches enable a variety of topologies including Fat Tree and more.

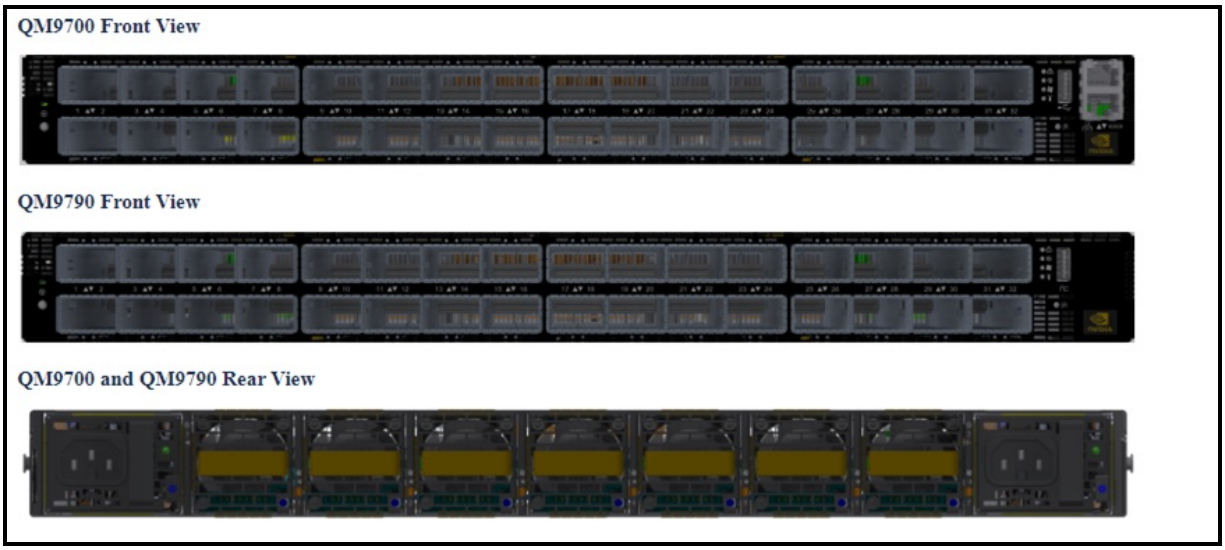

Figure 11. NVIDIA QM9700 and QM9790 InfiniBand switches

Ethernet switches

The NVIDIA Spectrum-3 SN4600 Ethernet switch offers 64 ports of 200GbE in a 2U form factor doubling the networking capacity over the SN3700. The SN4600V can be used as a high-density leaf, fully splittable to up to 128x 10/25/50GbE ports when used with splitter cables. SN4600 allows for maximum flexibility, with ports spanning from 1 to 200GbE and port density that enables full rack connectivity to any server at any speed, and a variety of blocking ratios.

Figure 12. NVIDIA Spectrum-3 SN4600

Management Ethernet switches

These switches are used for monitoring activities through Lenovo XClarity software and virtual operations if choosing a virtual environment. We use 1GbE or 10GbE for management with XClarity, and 25GbE for virtual operations.

- 1GbE XClarity management network: NVIDIA Spectrum SN2201 1GbE switch
  The NVIDIA Spectrum SN2201 switch has two key use cases:
  - A top-of-rack switch, connecting up to 48x 1G/100M/10M Base-T host-ports with non-blocking 100 GbE spine uplinks
  - Out-of-band (OOB) management switch

  Featuring highly advanced hardware and software, along with ASIC-level telemetry and a 16MB fully shared buffer, the SN2201 delivers unique and innovative features to 1G switching.
- 10GbE XClarity management network: NVIDIA AS4610 10GbE switch
  The managed Ethernet Switch chosen was the AS4610 series Gigabit Ethernet Layer 2/3 switch. This switch family has 48x 10/10/1000BASE-T ports and 4x 10G SFP+ uplink ports. Specifically, the AS4610-54T is well suited for building a managed network and was employed in the Reference Architecture. The 4x SFP+ uplink ports support either 1 GbE or 10GbE depending on the cabling type.
- Virtualization operations: NVIDIA SN2410 or SN2010 25GbE switches

For managing virtual operations, we use a 25GbE Ethernet switch. The NVIDIA Networking Spectrum Ethernet portfolio is an ideal top-of-rack solution for HPC, hyperconverged, and storage fabric deployments.

Lenovo EveryScale offers two 25GbE switches from this family:

- SN2410, with 48x QSFP28 ports running at 25Gb/s and 8x QSFP28 ports running at 100Gb/s
- SN2010, with 18x SFP28 ports running at 25Gb/s and 4x QSFP28 ports running at 100Gb/s.

The SN2410 is ideal for 25GbE fabrics and larger aggregation scenarios, while the SN2010 is more suited for storage and hyperconverged use cases, as well as small aggregation for HPC deployments.

Figure 13. NVIDIA SN2000 switches

Software stack

A robust end-to-end software platform is critical to ensure success of building generative AI and LLMs. In this reference architecture we carefully selected the most comprehensive and performance optimized software stack for inference and training generative AI.

In the following figure, we show the full stack of our Lenovo Reference Architecture for Generative AI and Large Languages Models.

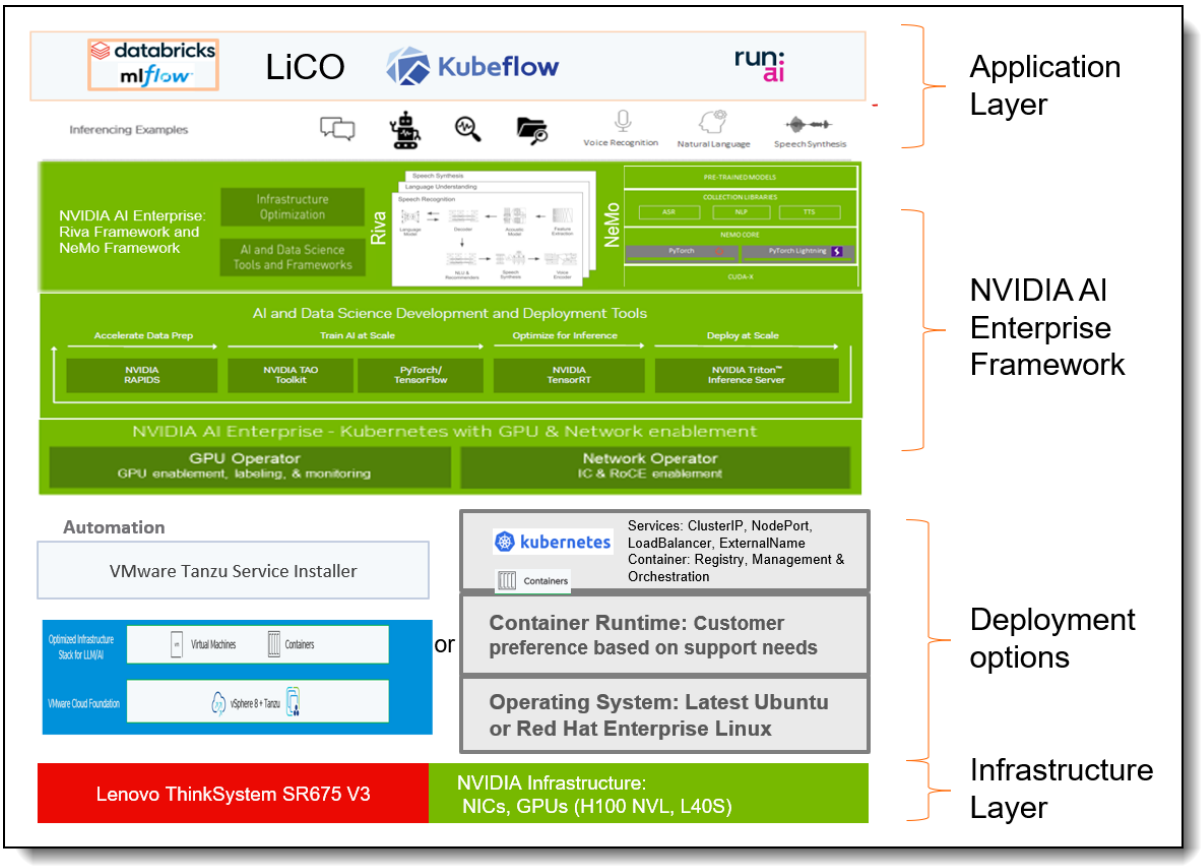

Figure 14. Reference Architecture full stack components

In the next section we will cover the application layer. NVIDIA AI Enterprise and Deployment option are covered in the NVIDIA AI Enterprise and Deployment options sections, respectively.

Application layer

In this section, we cover the major components of the application layer, as introduced in the Software stack section.

Topics in this section:

- DataBricks MLflow
- Kubeflow
- Run.ai
- Lenovo Intelligent Computing Orchestration (LiCO)
- Lenovo XClarity Administrator

DataBricks MLflow

MLflow is an open-source project created by DataBricks that provides a comprehensive framework for managing the entire machine learning lifecycle. It offers a centralized platform for tracking experiments, parameters, and results, as well as for packaging and deploying machine learning models.

There are four main components of MLflow:

- MLflow Tracking: This component provides a UI and API for logging parameters, metrics, artifacts, and code versions. It works in any environment, including notebooks, and allows users to track experiments through the development process.
- MLflow Projects: This component provides a format for packaging reusable code for sharing and moving to production.
- MLflow Models: This component standardizes the packaging of machine learning models and contains tools for deploying them. It allows for deployment on multiple serving platforms.
- MLflow Registry: This component provides a centralized model store with a UI and API. It allows users to manage the entire model lifecycle, including versioning, annotations, and step transitions collaboratively.

Machine learning applications can benefit from another resource from Databricks: the medallion architecture. The medallion architecture, also known as the multi-hop architecture, is a data design pattern that logically organizes data in a lake house.

The architecture consists of three data layers, each with a specific purpose and function:

- Bronze layer is the raw data layer, where data from external sources is initially stored.
- Silver layer is the cleansed and conformed data layer, where data is standardized to create a unified view of the data. This layer is designed to be a source for data scientists and data engineers: it enables self-service analytics and ad-hoc reporting, advanced analytics, and machine learning. Speed and agility are prioritized in this layer.
- Gold layer is the curated business-level tables layer, where data is transformed into a format that is optimized for visualization and reporting.

The medallion architecture is compatible with the data mesh design, in which bronze and silver layers are joined in a one-to-many fashion. The medallion architecture is designed to incrementally improve structure and quality of data as it flows through each layer. It is useful in MLOps because it provides a clear lineage of data and a silver layer that prioritizes data quality and to the level required and efficient data processing.

Kubeflow

Organizations seeking comprehensive model life cycle management can optionally deploy MLOps platforms, like Kubeflow and MLflow. These platforms streamline the deployment, monitoring, and maintenance of AI models, ensuring efficient management and optimization throughout their life cycle.

Kubeflow is an open-source set of tools designed to simplify the end-to-end development of machine learning applications on Kubernetes. Kubeflow uses Kubernetes to handle all code execution and resource management. It is highly scalable and efficient, and other required Kubernetes applications can share the cluster.

The main goal of Kubeflow is to create a standard for machine learning applications that considers each phase of the machine learning lifecycle, from experimentation to prediction. To achieve this, Kubeflow provides a simple UI for controlling ML projects, known as Kubeflow Pipelines, which allows developers to easily create, manage, and run ML workflows. Each step in a Kubeflow pipeline is isolated in its own container, improving the developer experience by reducing the risk of contamination between steps.

Highlights of Kubeflow are its responsive UI, with status changes in real time, and high level of customizability, as developers can customize Dockerfiles, containers, and use of nodes and memory to fit their specific needs. While this adds complexity, it also provides a high degree of control over the workflow.

Overall, Kubeflow is a powerful tool for building and managing machine learning applications on Kubernetes.

Run.ai

Run.ai is a cluster management platform designed to speed up the development, scaling, and cost optimization of AI infrastructure. It provides a unified dashboard for managing the entire cluster, including compute, jobs, user permissions, and an audit log of job history.

Key features:

- Automatically splits and joins GPU resources between users and jobs, allowing for optimal use and cost minimization.
- Enables the creation of Kubernetes node pools tailored to the needs of specific jobs.
- Provides team-level quotas to prevent interference between teams.
- Offers integrations with popular data science tools, such as MLflow, Jupyter, TensorBoard, and VSCode. Data scientists can also setup environments or project templates in just a few clicks.

Overall, Run.ai is designed to provide a scalable, cost-effective, and user-friendly solution for managing an AI cluster.

Lenovo Intelligent Computing Orchestration (LiCO)

Lenovo Intelligent Computing Orchestration (LiCO) is a software solution that simplifies the use of clustered computing resources for Artificial Intelligence (AI) model development and training, and HPC workloads. LiCO interfaces with an open-source software orchestration stack, enabling the convergence of AI onto an HPC or Kubernetes-based cluster.

The unified platform simplifies interaction with the underlying compute resources, enabling customers to take advantage of popular open-source cluster tools while reducing the effort and complexity of using it for HPC and AI.

LiCO enables a single cluster to be used for multiple AI workloads simultaneously, with multiple users accessing the available cluster resources at the same time. Running more workloads can increase utilization of cluster resources, driving more user productivity and value from the environment.

There are two distinct versions of LiCO, LiCO HPC/AI (Host) and LiCO K8S/AI, to allow clients a choice for the which underlying orchestration stack is used, particularly when converging AI workloads onto an existing cluster. The user functionality is common across both versions, with minor environmental differences associated with the underlying orchestration being used.

A summary of the differences for user access is as follows:

- LiCO K8S/AI version: AI framework containers are docker-based and managed outside LiCO in the customer’s docker repository. Custom job submission templates are defined with YAML, which do not include HPC standard job submission templates.
- LiCO HPC/AI version: AI framework containers are Singularity-based and managed inside the LiCO interface Custom job submission templates are defined as batch scripts (for SLURM, LSF, PBS) Includes HPC standard job submission templates.

Lenovo offers LiCO as an outstanding alternative tool for orchestrating your AI workloads on bare metal environments.

Lenovo XClarity Administrator

Lenovo XClarity Administrator is a centralized resource management solution that simplifies the management of Lenovo ThinkSystem infrastructure and ThinkAgile solutions. It provides a unified platform for managing various aspects of Lenovo's datacenter infrastructure, including servers, storage, networking, and software. Its key value is in reducing complexity, speeding response, and enhancing the availability of Lenovo server systems and solutions.

XClarity Administrator runs as a virtual appliance and supports managing a maximum of 1,000 devices. It uses no CPU cycles or memory on agent execution, saving up to 1GB of RAM and 1-2% CPU usage compared to a typical managed system. It provides a HTML-based UI that provides real time updates and alerts. XClarity Administrator is a key tool in device management and security, and it can be integrated with external, higher-level management from several providers, including VMware.

NVIDIA AI Enterprise

NVIDIA AI Enterprise is a high-performance, secure, cloud-native AI software platform built with enterprise-grade security, stability, manageability, and support for creating and deploying AI models. It addresses the complexities of organizations trying to build and maintain a complex AI software stack that builds on over 4,500 unique software packages: 64 NVIDIA CUDA® libraries and more than 4,471 third-party and open-source software (OSS) packages. The platform maintains API stability and the 9,000+ dependencies between these unique software packages.

As a full AI software stack, NVIDIA AI Enterprise accelerates AI pipelines and streamlines development and deployment of production AI covering the range of use cases from computer vision to Generative AI, to LLMs.

The Enterpriser-grade software includes:

- NVIDIA NeMo, an end-to-end framework for organizations to easily customize pretrained foundation models from NVIDIA and select community models for domain-specific use cases based on business data
- 100+ frameworks, pre-trained models, and development tools optimized for building and running AI on NVIDIA GPUs
- Continuous monitoring and regular releases of security patches for critical and common vulnerabilities and exposures (CVEs)
- Production releases that ensure API stability
- End-to-end management software including cluster management across cloud and data center environments, automated model deployment, and cloud nativeorchestration.
- Enterprise support with service-level agreements (SLAs) and access to NVIDIA AI experts
- NVIDIA AI Enterprise runs only on the most common versions of enterprise grade Linux, virtualization stacks, and multiple versions of Kubernetes.

The following figure shows the NVIDIA AI Enterprise software platform.

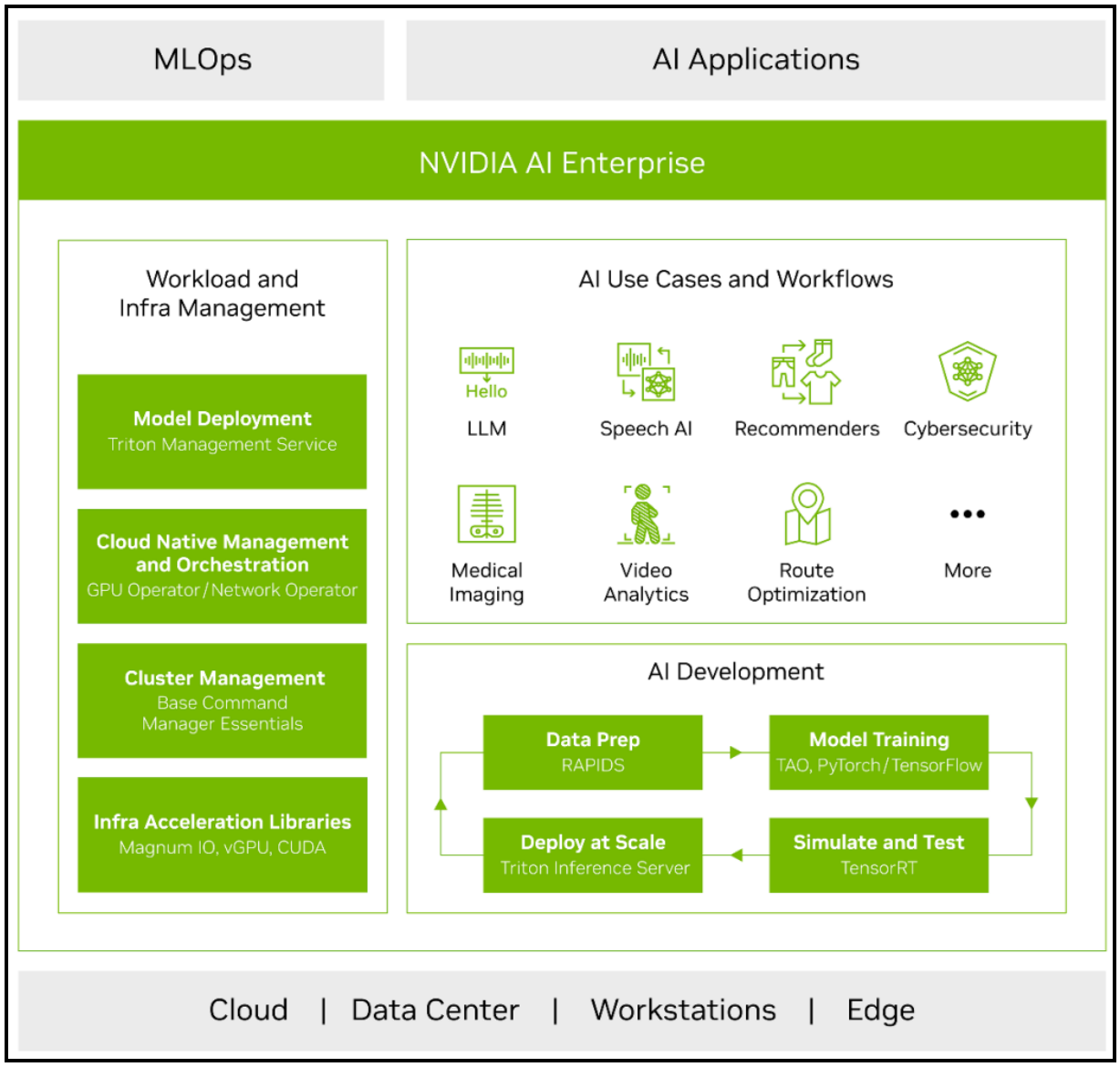

Figure 15. NVIDIA AI Enterprise software platform

In this section, a discussion will be provided on the NeMo and Riva Frameworks, the pre-trained models that are part of the solution, and the Triton Inference Server.

Topics in this section:

- NVIDIA NeMo
- NVIDIA Riva
- Triton Inference Server

NVIDIA NeMo

NVIDIA NeMo, included with NVIDIA AI Enterprise, is an end-to-end, cloud-native framework to build, customize, and deploy generative AI models anywhere. It includes training and inferencing frameworks, guard railing toolkits, data curation tools, and pretrained models. It provides tooling for distributed training for LLMs that enable advanced scale, speed, and efficiency. NeMo enables integration of real-time, domain-specific data via Inform.

NeMo Guardrails assist defining operational boundaries so that models stay within the intended domain and avoid inappropriate outputs. This framework supports Reinforcement Learning from Human Feedback (RLHF) technique, allowing enterprise models to get smarter over time, aligned with human intentions.

Pre-packaged scripts, reference examples, and documentation across the entire pipeline are provided by NeMo. An auto-configuration tool that is part of NeMo aides building of foundation models by automatically searching for the best hyperparameter configurations to optimize training and inference for any given multi-GPU configuration, training, or deployment constraints.

NeMo offers pretrained models that address various categories, including Automatic Speech Recognition (ASR), NLP, and Text-to-Speech (TTS). In addition, pretrained models from both NVIDIA's NGC Catalog and Hugging Face are offered that are tested and optimized for best performance.

For completeness, NVIDIA offers other AI foundries for language, biology, visual design, and interactive avatars. There is a tool that is part of the NeMo Framework that facilitates the frameworks usage. A user only needs to use this tool and complete the appropriate configuration files and the launcher associated with this tool will auto-generate the needed scripts. The is also a service (early release) associated with this framework that allows for the fine tuning of existing models that can be hosted and then deployed.

NVIDIA Riva

NVIDIA Riva is a GPU-accelerated SDK for building Speech AI applications that are customized for your use case and deliver real-time performance. It offers pretrained speech models in NVIDIA NGC™ that can be fine-tuned with NVIDIA NeMo on a custom data set, accelerating the development of domain-specific models. These models can be easily exported, optimized, and deployed as a speech service on premises or in the cloud with a single command using Helm charts.

Riva's high-performance inference is powered by NVIDIA TensorRT™ optimizations and served using the NVIDIA Triton™ Inference Server, which are both part of the NVIDIA AI platform. Riva, like NeMo, is fully containerized and can easily scale to hundreds and thousands of parallel streams.

Riva can be used to access highly optimized Automatic Speech Recognition (ASR) and speech synthesis services for use cases like real-time transcription and virtual assistants. The ASR skill is available in multiple languages. It is trained and evaluated on a wide variety of real-world, domain-specific datasets. With telecommunications, podcasting, and healthcare vocabulary, it delivers world-class production accuracy. Riva's text-to-speech (TTS) or speech synthesis skills can be used to generate human-like speech.

Triton Inference Server

At the heart of the AI inference system is the Triton Inference Server, part of NVIDIA AI Enterprise, that handles AI models and processes inference requests. Triton is a powerful inference server software that efficiently serves AI models with low latency and high throughput. Its integration with the compute infrastructure, GPU accelerators, and networking ensures smooth and optimized inferencing operations.

Triton enables the deployment of any AI model from multiple deep learning and machine learning frameworks, including TensorRT, TensorFlow, PyTorch, ONNX, OpenVINO, Python, RAPIDS FIL, and more. Triton supports inference across cloud, data center, edge, and embedded devices on NVIDIA GPUs, x86 and ARM CPU, or AWS Inferentia. Triton delivers optimized performance for many query types, including real time, batched, ensembles and audio/video streaming.

Deployment options

The primary objective of this reference architecture is to provide enterprises with unparalleled flexibility in deploying a Generative AI solution. This reference architecture has been meticulously designed with a keen understanding of diverse workload requirements, operational considerations, and application nuances.

To empower our customers, we've provided them with the autonomy to tailor their deployment approach, offering a choice between a robust bare metal setup or a versatile virtual deployment. In either scenario, an assurance of optimal performance and unmatched flexibility accompanies their decision, fostering an environment conducive to realizing the full potential of their Generative AI Solution.

Deployment options discussed here are:

- Deployment to bare metal
- Deployment to a virtual environment

Deployment to bare metal

Topics in this section:

- Benefits
- Recommended configuration for a bare metal deployment
- Red Hat and Kubernetes

Benefits

Deploying a bare metal environment for a generative AI cluster using InfiniBand can offer several benefits:

- **Performance**: InfiniBand is known for its low latency and high bandwidth capabilities. In a bare metal setup, where there is minimal virtualization overhead, you can fully harness the performance potential of InfiniBand, resulting in faster communication between nodes and reduced data transfer times. This is crucial for real-time or time-sensitive generative AI workloads.
- **Dedicated Resources**: Bare metal servers provide dedicated hardware resources to your generative AI workload. This exclusivity translates to consistent performance and reduced contention for resources that might occur in virtualized environments.
- **Customization**: With a bare metal setup, you have greater control over hardware configuration and optimization. You can fine-tune hardware components to match the specific requirements of your generative AI workloads, resulting in improved performance and efficiency.
- **Isolation**: Bare metal environments provide better isolation between different workloads or applications. This separation prevents interference from other workloads, ensuring that your generative AI cluster operates without disruptions.
- **Lower Overhead**: Virtualization introduces overhead due to the virtualization layer. By deploying on bare metal, you eliminate this overhead, allowing your generative AI applications to make full use of the available resources.
- **GPU Utilization**: Many generative AI workloads rely heavily on GPUs for parallel processing. In a bare metal setup, you can directly allocate GPUs to your workload without any virtualization layers, maximizing GPU utilization and performance.
- **Predictable Performance**: In virtualized environments, performance can be impacted by the variability introduced by other virtual machines sharing the same host. Bare metal deployments offer more predictable and consistent performance levels.
- **Scalability**: While virtualized environments offer flexibility in scaling resources, a bare metal environment can also be scaled by adding more physical servers to the cluster. This can be advantageous for larger-scale generative AI applications.
- **Simplicity**: Virtualization introduces additional complexity in management and maintenance. A bare metal environment can be simpler to manage, reducing the potential for configuration errors or

compatibility issues.

- **HPC Workloads**: If your generative AI cluster involves complex simulations or modeling tasks typical of high-performance computing (HPC), a bare metal setup with InfiniBand is better suited to handle these demanding workloads efficiently.

In summary, deploying a bare metal environment for a generative AI cluster using InfiniBand can lead to superior performance, customization, and resource utilization, making it an ideal choice for high-performance and time-sensitive applications. However, it's important to carefully assess your workload's requirements, available resources, and management capabilities before deciding on the deployment approach.

Recommended configuration for a bare metal deployment

Our recommendation for hardware and software for bare metal deployment is as follows. Note that this deployment is designed with an InfiniBand network.

- Server - SR675 V3
  - 2x AMD EPYC 9554 processors, 64 cores, 3.1GHz
  - 24x 64GB TruDDR5 RDIMMs
  - 2x 960GB NVMe SSDs
  - 1x 1G 4-port OCP adapter for management connector
  - 8x NVIDIA H100 NVL GPUs
  - 4x NVIDIA ConnectX-7 NDR200 dual-port
  - 1x NVIDIA ConnectX-7 NDR200 dual port (optional network storage)
  - 4x 2400W power supplies (230V)
  - Premier Essential support - 3 years, 24x7, with 4-hour response time + Your Drives Your Data (YDYD)
- Operating system: Latest Ubuntu or Red Hat OpenShift
- Software - NVIDIA AI Enterprise: 2 options
  - NVIDIA AI Enterprise Essentials; 1,3 year or 5 year subscription)
  - NVIDIA Riva (includes NVIDIA AI Enterprise Essentials plus support for Riva SDK for Automatic Speech Recognition (ASR), text-to-speech (US), and neural machine translation (NMT) applications)
- NeMo framework for inferencing and training: toolkit for building GenAI and LLMs models
- Optional Lenovo storage
- Networking
  - 2x NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch, Compute
  - 2x NVIDIA QM9790 HDR IB Managed Switch (NDR200) for parallel file system, optional storage
  - 1X NVIDIA 2201 1GbE Managed Switch

It is important to mention that this deployment option takes advantage of Red Hat and Kubernetes, on top the NVIDIA AI Enterprise is deployed to complete the full Stack of this reference Architecture.

Red Hat and Kubernetes

Red Hat and Kubernetes are container orchestration platforms that can be effectively used in deploying and managing generative AI solutions. They provide a robust framework for automating the deployment, scaling, and management of containerized applications, which is particularly advantageous in the context of complex and resource-intensive generative AI workloads.

Red Hat and Kubernetes can be used for the following workloads:

- **Containerization of Generative AI Workloads**: Generative AI applications can be packaged into containers using technologies like Docker. Containers encapsulate the application, its dependencies, and runtime environment. Kubernetes and OpenShift excel at managing and orchestrating these containers, ensuring consistent deployment across various environments.

- **Resource Allocation and Scaling**: Kubernetes and Red hat enable automatic scaling of generative AI workloads based on resource utilization. As the demand for computational resources increases, these platforms can dynamically allocate additional resources to ensure optimal performance. Conversely, when demand decreases, excess resources can be reclaimed.
- **High Availability and Fault Tolerance**: Both platforms facilitate the deployment of generative AI workloads across multiple nodes, providing high availability and fault tolerance. If a node fails, the orchestration system automatically reallocates workloads to healthy nodes, minimizing downtime.
- **Persistent Storage**: Generative AI solutions often require storing and managing large amounts of data, such as training datasets and model checkpoints. Kubernetes and OpenShift integrate with various storage solutions to provide persistent storage for containers.
- **Environment Consistency**: Kubernetes and Red hat ensure consistent deployment environments, regardless of the underlying infrastructure. This consistency is vital for generative AI applications that rely on specific software libraries, configurations, and dependencies.
- **Efficient Resource Utilization**: These platforms optimize resource utilization by packing multiple containers onto the same physical node. This can help in utilizing hardware resources effectively while maintaining performance isolation between containers.
- **Application Updates and Rollbacks**: Kubernetes and Red Hat streamline the process of deploying updates to generative AI applications. They support rolling updates, which ensure that new versions are gradually deployed while maintaining the application's availability. If issues arise, rollbacks can be easily executed.
- **Configuration Management**: Both platforms allow you to define and manage the configuration of generative AI applications using declarative manifests. This makes it easier to maintain consistency and automate the deployment process.
- **Monitoring and Logging**: Kubernetes and Red hat provide tools for monitoring the health and performance of generative AI workloads. They offer integration with various monitoring and logging solutions, helping you gain insights into the behavior of your applications.
- **Integration with CI/CD Pipelines**: Generative AI solutions often involve continuous integration and continuous deployment (CI/CD) pipelines. Kubernetes and Red hat can integrate seamlessly with CI/CD tools, enabling automated testing, deployment, and validation of new AI models.
- **Multi-Cloud and Hybrid Deployments**: Both platforms support multi-cloud and hybrid deployment scenarios, allowing you to deploy generative AI solutions across different cloud providers or on-premises infrastructure.

In summary, Red Hat and Kubernetes provide a powerful foundation for deploying, scaling, and managing generative AI solutions by leveraging containerization and orchestration capabilities. They help streamline deployment workflows, improve resource utilization, enhance availability, and simplify the management of complex AI workloads.

Deployment to a virtual environment

As we describe previously it is our intention to offer the flexibility on finding the right deployment option that satisfies your needs. Virtualization technologies are an essential component of today’s enterprises.

Topics in this section:

Benefits

Deploying a generative AI cluster on a virtual environment also offers several benefits, which can contrast with a bare metal deployment:

- **Resource Sharing**: Virtual environments allow for efficient resource sharing among multiple virtual machines (VMs). This can be advantageous when you have a mix of workloads with varying resource needs, as you can dynamically allocate resources based on demand.
- **Isolation and Security**: Virtualization provides a higher degree of isolation between VMs. This can enhance security by preventing one workload from affecting others. It also allows for better testing and development environments without impacting the production setup.
- **Resource Utilization**: In virtual environments, you can achieve higher overall resource utilization by consolidating multiple workloads on a single physical server. This is especially useful when individual workloads don't require the full capacity of a dedicated server.
- **Flexibility and Scalability**: Virtual environments offer greater flexibility in scaling resources up or down as needed. You can easily provision new VMs or adjust resource allocations without the need for additional physical hardware.
- **Hardware Independence**: Virtualization abstracts the underlying hardware, allowing you to migrate VMs between different physical servers without compatibility concerns. This can simplify hardware maintenance and upgrades.
- **Cost-Efficiency**: Virtual environments can be more cost-effective, as you can utilize hardware resources more efficiently by sharing them among multiple VMs. This can lead to better ROI, especially if your workloads don't require dedicated physical servers.
- **Snapshot and Recovery**: VM snapshots enable you to capture the state of a VM at a specific point in time. This makes it easier to recover from failures or to test changes without risking the production environment.
- **Rapid Deployment**: Setting up new VMs is generally faster than procuring and configuring physical hardware. This agility can be beneficial for quickly deploying and experimenting with new generative AI models.
- **Workload Diversity**: Virtual environments are well-suited for environments where multiple workloads with varying operating systems, software requirements, and configurations need to coexist.
- **Reduced Environmental Footprint**: By consolidating workloads onto fewer physical servers, virtualization can lead to a reduced physical footprint in data centers, resulting in lower energy consumption and cooling costs.
- **Ease of Management**: Virtualization platforms often provide management tools for monitoring, provisioning, and scaling VMs, which can simplify administrative tasks.

The decision between bare metal and virtual deployment should be based on a careful analysis of your generative AI workload's requirements, available resources, budget constraints, and the trade-offs associated with each deployment approach.

For the virtual environment we are basing our reference architecture design on the latest VMware Private AI Foundation with NVIDIA.

VMware Private AI Foundation with NVIDIA

This platform enables enterprises to fine-tune LLM models and run inference workloads in their data centers, addressing privacy, choice, cost, performance, and compliance concerns. The platform includes the NVIDIA NeMo framework, NVIDIA LLMs, and other community models (such as Hugging face models) running on VMware Cloud Foundation. Lenovo's world-class servers will provide hardware layer for this platform.

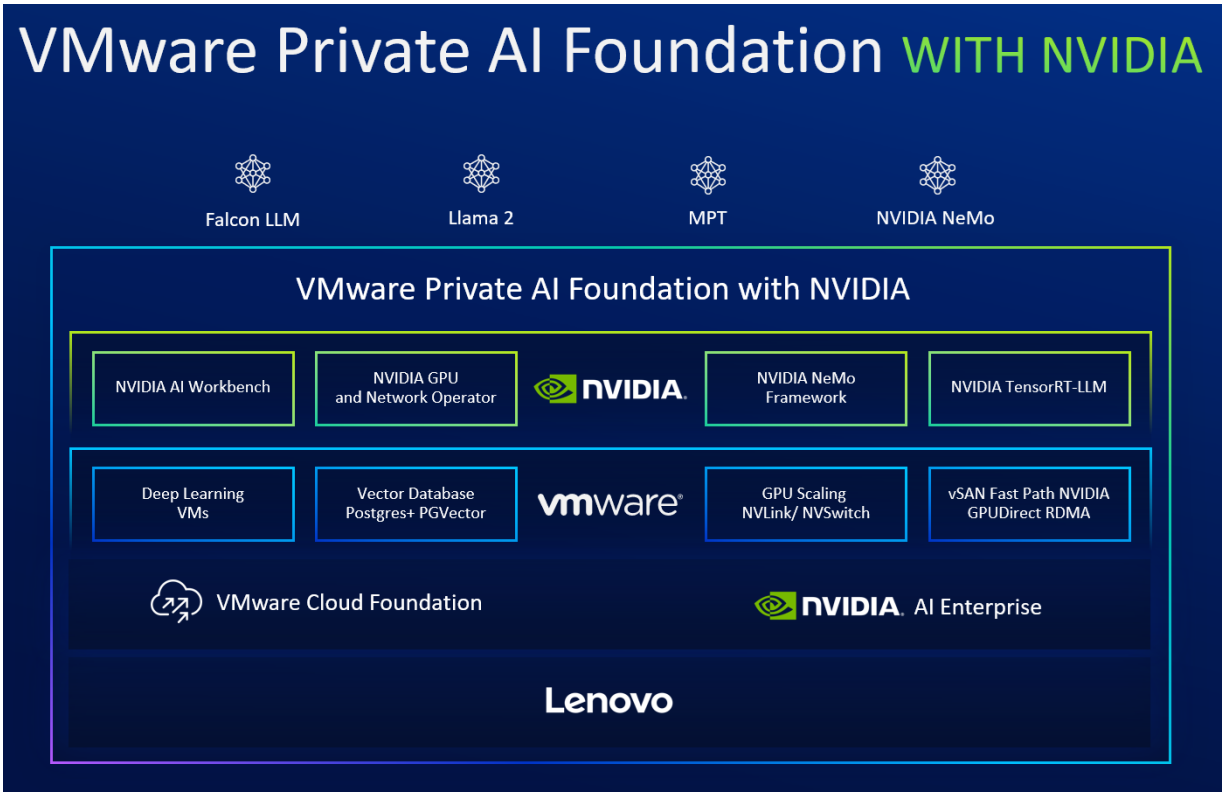

Figure 16. VMware Private AI Foundation with NVIDIA building blocks

Features of the VMware Private AI foundation include the following:

- **Deep Learning VMs**

  Data scientists often spend a significant amount of time on version and dependency management, especially when dealing with the complexities of machine learning (ML) applications and GPU utilization. The layers of libraries, frameworks, toolkits, and drivers between the ML application and GPU can be delicate and interdependent. To help expedite the process for our customers and enable them to get started quickly, this foundation offers customized Deep Learning VM images, specifically optimized for running deep learning workloads and come pre-configured with popular frameworks like TensorFlow and PyTorch, as well as various NVIDIA libraries and integrated development environments (IDEs), Libraries while being aligned to the infrastructure. This will save valuable time and effort for data scientists, allowing them to concentrate on developing their machine learning models.

- **Vector database (Postgres + PGVector)**

  GenAI systems, especially applications using LLMs use a Vector database (Vector DB) for context retrieval or similarity search. In general, a Vector DB is cheaper and faster than retraining an LLM, allowing real-time updates of the available data to the LLM. A Vector DB can supply the LLM with more information during a prompt generation and allows the LLM to have more up-to-date, contextually relevant information.

  We believe Vector DBs can be a critical differentiator for organizations with rapidly changing knowledge bases or any other source of information. VectorDB provides data access segmentation because of its role-based authentication, and it restricts particular users to specific parts of the database index and allows an organization to segment the data accordingly. Offering a Vector DB functionality in the Private AI Foundation aligns with our focus on delivering a platform that offers data security and economical use of computational power.

- **Model repository and converter**

  LLMs consume lots of GPU memory space. The containers that LLMs run on also consume a lot of disk space, on average 10-30GB. So big models equal heavy lifting for the Kubernetes schedulers. This is where a localized container repository could offer a range of substantial benefits for efficient software management and deployment. By establishing a repository closer to the worker nodes, the process of pulling container images is significantly accelerated, thus reducing latency, and enhances overall performance. The VMware Private AI Foundation should enable pulling from a localized repository, shielding the organization from configuration errors. In addition, we want to make the experience of pulling from the localized model repository as seamless as possible, by creating appliances which can be easily deployed via the submenu of the vCenter UI.

- **GPU Scaling (NVLink/NVSwitch)**

  As the ML models get bigger and bigger, they don't fit into the graphics memory of a single GPU, so you need to use multiple GPUs. With NVLink and NVSwitch, customers have the flexibility where they can combine multiple GPUs on the same host, and support bigger models, without a significant communication overhead.

- **NVIDIA GPUDirect RDMA**

  Whereas NVIDIA NVLink and NVSwitch optimize inter-GPU communication within a single ESXi host, NVIDIA GPUDirect RDMA (Remote Direct Memory Access) optimizes the complete path between GPUs in separate ESXi hosts. It provides a direct peer-to-peer data path between the GPU memory to and from the high-performance NIC. It decreases latency and reduces overall ESXi host overhead for distributed training processes. vSphere Device Groups allow a GPU device and a NIC that share the same PCIe Switch for communicating to be presented to a VM as one unit. Automatic discovery of hardware topology assists the VI-admin in provisioning optimized VMs, while DRS uses vSphere device groups for VM placement decisions.

- **NVIDIA GPUDirect Storage (vSAN Fast Path)**

  Like GPUDirect RDMA, GPUDirect Storage avoids system overhead by creating a direct path between local or remote storage systems and GPU memory. GPUDirect RDMA helps accelerate machine-learning single-host and multi-host machine workloads. Note: vSAN is currently not part of this reference architecture; we plan to include it in future release.

Recommended configuration for a virtualized deployment

Our recommendation for hardware and software for a virtual deployment is as follows. Note that all virtual components are included as part of VMware Cloud Foundation.

- Servers
  - 4, 8, or 16x ThinkSystem SR675 V3
- Each server:
  - 2x AMD EPYC 9634, 84C 2.25GHz
  - 24x 128GB TruDDR5 3DS RDIMMs
  - 2x 1.92TB NVMe SSDs
  - 6x 7.68TB NVMe SSD s
  - 1x 1Gb 4-port OCP Adapter for management connector
  - 8x NVIDIA H100 NVL
  - 4x NVIDIA ConnectX-7 200GbE dual-port
  - 1x NVIDIA ConnectX- 7 200GbE dual-port (optional network storage)
  - 1X NVIDIA ConnectX 6 Lx 10/25GbE
  - 4x 2400W power supplies (230V)
  - Premier Essential service & support - 3 years, 24x7, 4-hour response time, with Your Drive Your Data (YDYD)
  - Guest OS Latest Ubuntu or Red Hat Enterprise Linux
- VMware software
  - VMware vSphere 8 Enterprise Plus with Tanzu Standard

    - VMware NSX Advanced load Balancer Enterprise
    - VMware vCenter Server 8 Standard for vSphere 8
    - VMware Tanzu Service installer
- NVIDIA AI Enterprise - 2 options:
    - NVIDIA AI Enterprise Essentials (3-year or 5-year subscription)
    - NVIDIA Riva (includes NVIDIA AI Enterprise Essentials plus support for Riva SDK for Automatic Speech Recognition (ASR) text-to-speech (TTS) and neural machine translation (NMT) applications)
- NeMo framework for inferencing and training toolkit for building GenAI and LLMs models
- Networking:
    - 1x or 2x NVIDIA SN4600 200GB Ethernet High Speed Spectrum Switch
    - 1x SN2410 25GbE Virtual Services Networking
    - 1x SN2201 1GbE Hardware Management Networking
- NVAIE operators:
    - GPU: v22.9.1
    - Network: v23.5.0
- Optional Lenovo DM or DE series, storage example:
    - Storage, 830TB useable
    - 8x 7.68TB NVMe SSDs
    - Premier Essential service & support - 3 years, 24x7, 4-hour response time, with Your Drive Your Data (YDYD)

VMware Cloud Foundation

On the Bare Metal deployment, we described Red Hat and Kubernetes as the core elements to build our environment. For the virtual environment we are selecting VMware Cloud Foundation as robust alternative for those customers seeking a Generative AI with the benefits of the virtualization technology.

VMware Cloud Foundation provides a ubiquitous hybrid cloud platform for both traditional enterprise and modern applications. Based on a proven and comprehensive software-defined stack that includes VMware vSphere with Tanzu, VMware vSAN™, VMware NSX®, and VMware vRealize® Suite. The result is an agile, reliable, efficient cloud infrastructure that offers consistent operations across private and public clouds.

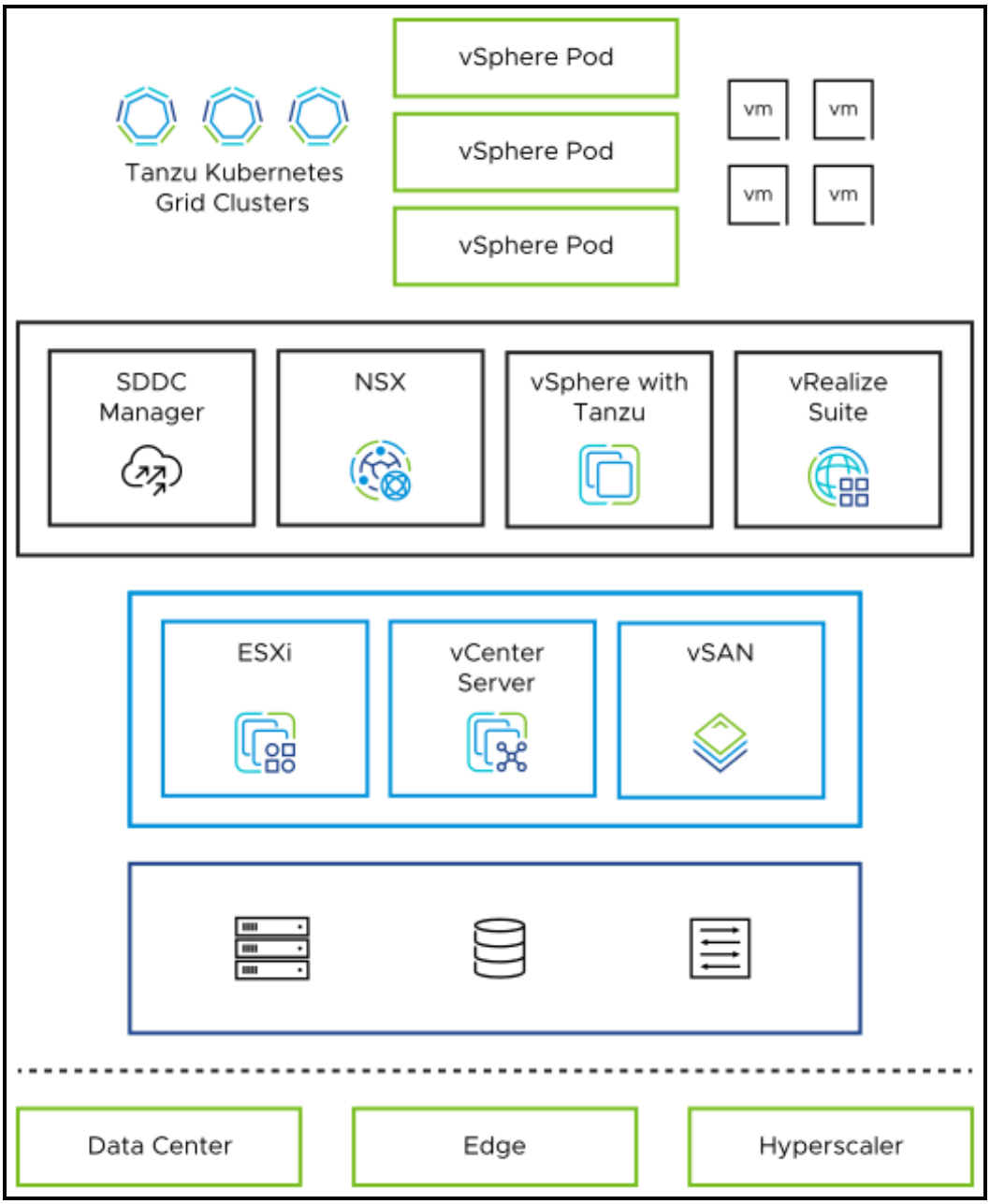

Figure 17. VMware Cloud Foundation

By using VMware Cloud Foundation (VCF), cloud infrastructure administrators can provision application environments in a rapid, repeatable, and automated way versus the traditional manual processes.

As teams look to adopt and operationalize AI/ML and LLMs the need for each individual team to have an environment that meets their needs is critical. Admins can look to VCF to help solve this paradigm shift by leveraging Workload Domains that fit the persona needs. For example, one team will need to support the models and datasets in production, but another team will be responsible for working on the bleeding edge models in a development context. These teams all require their own infrastructure components, and they should not be influencing each other's performance or touching each other's datasets. That isolation is provided by Workload Domains (WLD) in VCF.

The following figure shows some of the constructs for the Workload Domains in VCF.

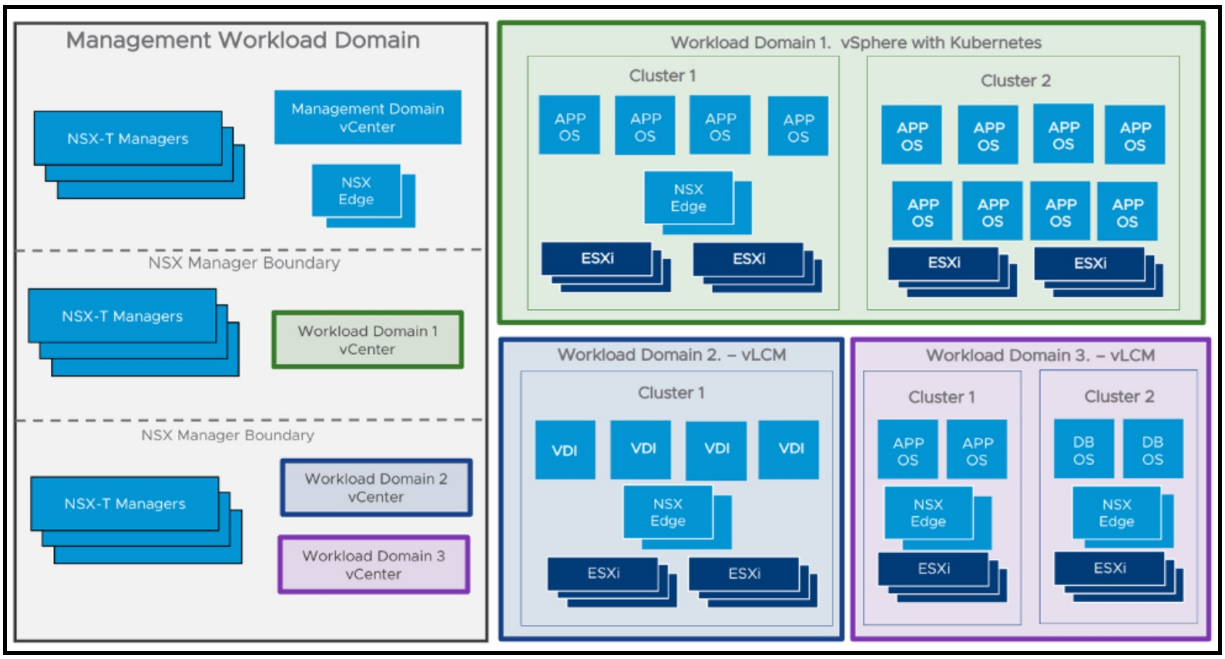

Figure 18. Workload Domains in VMware Cloud Foundations

VMware vSphere with Tanzu

VMware vSphere with Tanzu, transforms traditional virtualization infrastructure into a robust platform for containerized workloads. Of course, it also enables customers to support their traditional virtual machines and run them alongside their Kubernetes clusters. This provides a more heterogeneous approach to maintaining existing resources that have access to GPUs.

VMware Tanzu Kubernetes Grid™ facilitates the creation and management of Tanzu Kubernetes clusters natively within vSphere, seamlessly integrating Kubernetes capabilities with the reliable features of vSphere. With vSphere's networking, storage, security, and high availability features, organizations achieve better visibility and operational simplicity for their hybrid application environments.

vSphere with Tanzu also enables organizations to run application components in containers or virtual machines (VMs) on a unified platform, streamlining workload management and deployment. By converging containers and VMs, IT teams can leverage the benefits of both paradigms while maintaining a cohesive and scalable infrastructure. It also enables a fast path to GPU enabled VMs for development work during the creation of algorithms and models and a dynamic path to deploying these elements on Kubernetes clusters once models are slipstreamed into DevOps/MLOps pipelines so they can be promoted into production.

Deployment considerations

Prior your deployment it is important that you configure the following features on your environment.

Multi Node (Distributed) Learning

Many data scientists are looking to scale compute resources to reduce the time it takes to complete the training of a neural network and produce results in real-time. Taking a Multi-GPU approach brings scientists closer to achieving a breakthrough as they can more rapidly experiment with different neural networks and algorithms. To maximize the throughput, multi-node learning takes the approach by using two or more worker nodes of Tanzu Kubernetes Grid Cluster on different VMware ESXi hosts with GPU installed, with each worker node assigned a virtual GPU. These worker nodes can transfer data across a network with TCP or Remote Direct Memory Access (RDMA) protocol.

For data scientist that have access to systems with multiple GPUs the same outcome can be achieved without having to leverage distributed frameworks as they start their initial work. This can simplify the initial development phase and requires less advanced networking designs.

GPUDirect

GPUDirect on VMware vSphere works by leveraging NVIDIA vGPU technology and RDMA-capable network adapters like the NVIDIA ConnectX-6 and ConnectX-7. For Virtual Machines ACS relax VMX settings as well as NUMA affinity or Device Groups must be configured, to enable Peer-to-Peer communication between the PCIe devices on the same Root Complex.

Benefits of GPUDirect RDMA include:

- Lower latency: By bypassing the CPU and enabling direct GPU-to-device communication, GPUDirect RDMA significantly reduces data transfer latency, enhancing the responsiveness of GPU-accelerated applications.
- Improved bandwidth: GPUDirect RDMA optimizes data transfers between GPUs and other devices, increasing the available bandwidth for data-intensive workloads.
- Reduced CPU overhead: The technology offloads data movement tasks from the CPU, freeing up valuable CPU resources for other computational tasks.
- Enhanced scalability: GPUDirect RDMA allows multiple GPUs to access remote data sources simultaneously, enabling better scalability in large-scale GPU clusters.

After the RDMA network device Virtual Functions (VFs) are configured in vSphere and vGPU software is installed on the ESXi hosts and the NVIDIA GPUs are configured for vGPU mode, GPUDirect RDMA can be used. In this setup, the guest VMs must have the compatible NVIDIA vGPU driver and network driver installed. On Tanzu Kubernetes Grid, NVIDIA GPU Operator and NVIDIA Network Operator can be leveraged to install the proper vGPU and network drivers, applying the proper configuration and settings to the devices.

When a VM requests data transfer, GPUDirect RDMA facilitates direct communication between the vGPU and the remote device without involving the hypervisor or the CPU avoiding an extra buffer copy of the data. The vGPU's memory buffers are registered with the RDMA-capable device, and data can be transferred directly between these memory buffers, bypassing the hypervisor's intervention.

By enabling direct communication between VMs with vGPUs and other devices, GPUDirect RDMA improves data transfer efficiency, reduces data path latency, and enhances the overall performance of GPU-accelerated applications such as AI/ML in VMware virtualized environments.

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 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.

The GPU Operator allows administrators of Kubernetes clusters to manage GPU nodes just like CPU nodes in the cluster. Instead of provisioning a special OS image for GPU nodes, administrators can rely on a standard OS image for both CPU and GPU nodes and then rely on the GPU Operator to provision the required software components for GPUs.

The GPU Operator also enables GPUDirect RDMA; a technology in NVIDIA GPUs that enables direct data exchange between GPUs and a third-party peer device using PCI Express. The third-party devices could be network interfaces such as NVIDIA ConnectX SmartNICs or BlueField DPUs among others.

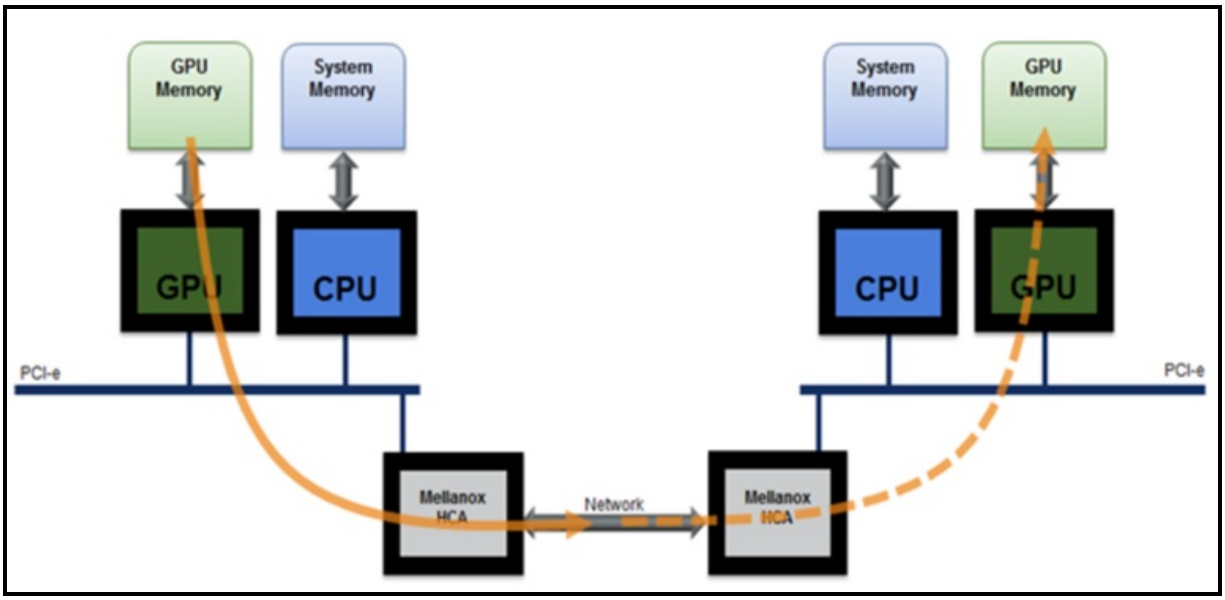

Figure 19. GPU Direct, Direct communication between GPUs

NVIDIA Network Operator

NVIDIA Network Operator leverages Kubernetes CRDs and Operator SDK to manage networking related components, to enable fast networking, RDMA and GPUDirect for workloads in a Kubernetes cluster. The Network Operator works in conjunction with the GPU-Operator to enable GPU-Direct RDMA on compatible systems and high-speed networking like InfiniBand or RoCE. The goal of the Network Operator is to manage the networking related components, while enabling execution of RDMA and GPUDirect RDMA workloads in a Kubernetes cluster.

NVIDIA Network Opertaor includes:

- NVIDIA Networking drivers to enable advanced features.
- Kubernetes device plugins to provide hardware resources required for a fast network.
- Kubernetes secondary network components for network intensive workloads.

The NVIDIA network operator can be deployed in different modes. For this reference architecture, we deploy a network operator with Host Device Network. This deployment includes:

- SR-IOV device plugin, single SR-IOV resource pool
- Secondary network
- Mutlus CNI
- Container networking-plugins CNI plugins
- Whereabouts IPAM CNI plugin

The Network Operator can be deployed on virtualized deployments as well. It supports both Ethernet and InfiniBand modes. From the Network Operator perspective, there is no difference between the deployment procedures. To work on a VM (Virtual Machine), the PCI passthrough must be configured for SR-IOV devices. The Network Operator works both with VF (Virtual Function) and PF (Physical Function) inside the VMs or Tanzu Worker nodes.

Like the Bare Metal Environment, NVIDIA AI Enterprise will complement the full stack of this deployment option.

Configuration examples

In the Hardware stack section, we explained how this reference architecture adopts our philosophy of EveryScale. Following the recommendations and guidance of this document you will be able to build your own solution at the scale that you need.

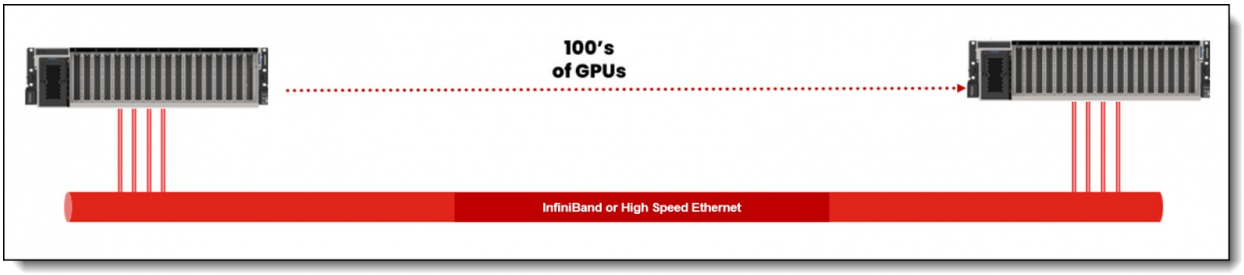

Figure 20. Solution Scalability

To aid in the deployment of this reference architecture, we have tailored three configuration examples ("T-Shirt sizes") starting from:

- Small (4 nodes config)
- Medium (8 node config)
- Large (16 nodes config)

We provide the detailed bill of materials (BOM) for each of these configurations in the Bill of materials section.

Following this document, you can build a larger solution.

The three T-shirt size sized examples have an InfiniBand network. This solution is based on the NVIDIA H100 NVL and NVIDIA L40S GPUs. We also provide complementary design for OVX L40S configuration powered by BlueField-3 and NVIDIA L40S GPUs. Included in this architecture are the required VMware components and NVIDIA AI Enterprise.

The following figures show the three example configurations based on 8x double-wide GPU slots. The ThinkSystem SR675 V3 can have either a 4DW PCIe GPU base or an 8DW PCIe GPU Base.

The following figures show the three configuration examples.

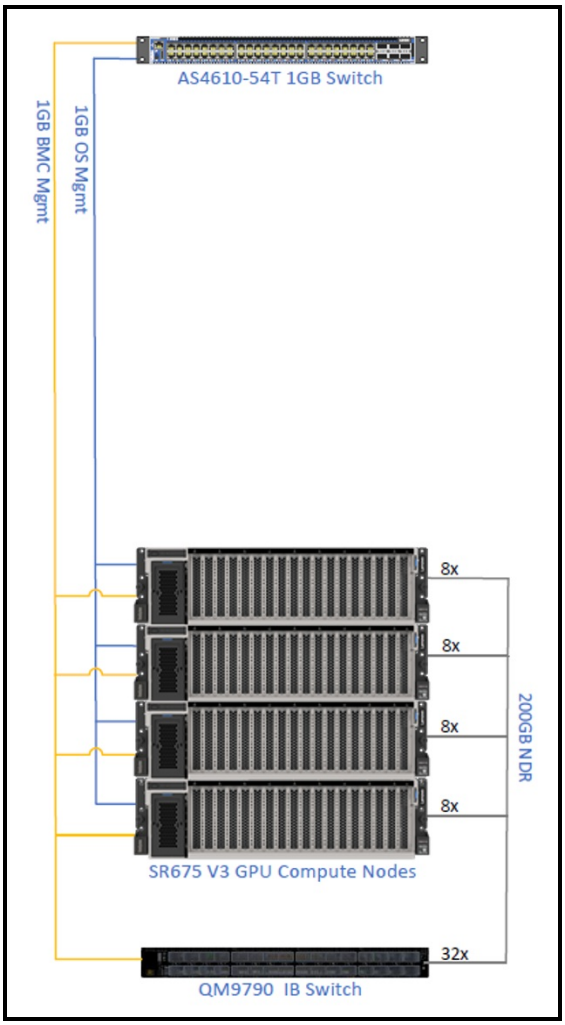

Figure 21. Small configuration (4 nodes)

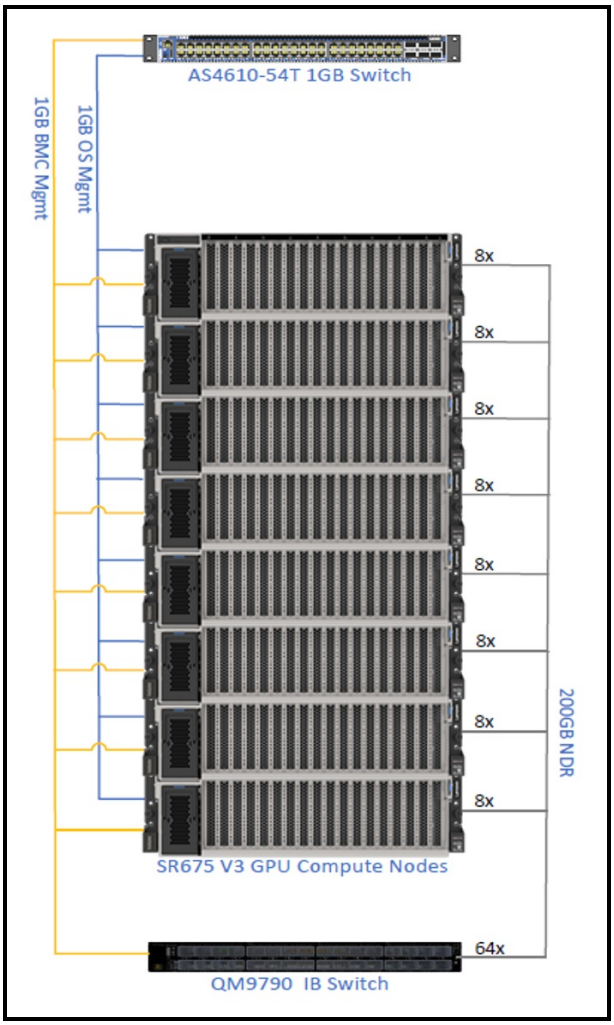

Figure 22. Medium configuration (8 nodes)

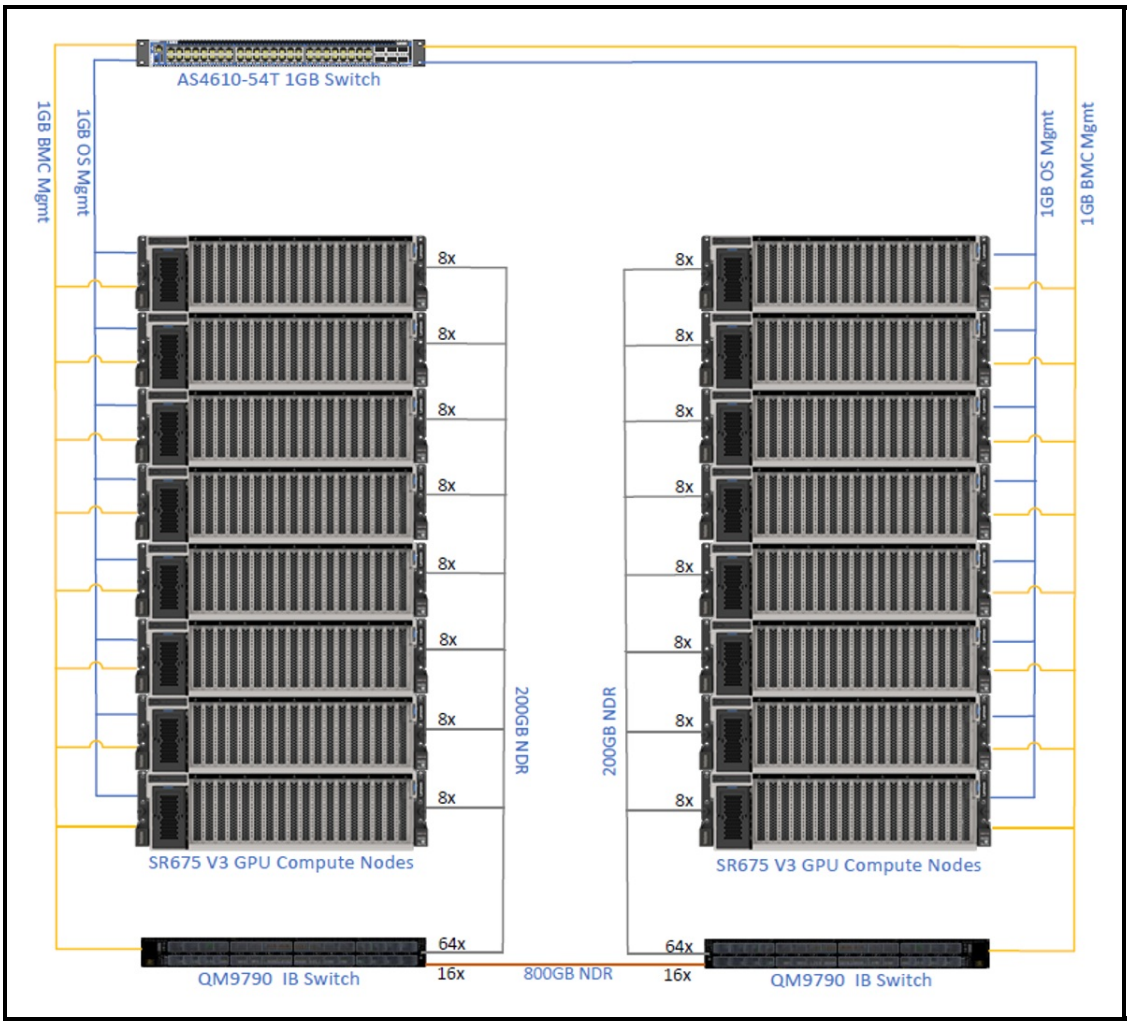

Figure 23. Large configuration (16 nodes)

For detailed server specs on this configuration please refer to the Build of materials section.

Design recommendations

Building a generative AI cluster that leverages GPUs and InfiniBand networking involves careful planning, hardware selection, software configuration, and optimization. Here are some design recommendations to consider.

Server configuration

- **GPUs**: Select servers with multiple GPUs, ensuring sufficient GPU memory and compute power to handle the complexity of generative AI models.
- **CPU**: Choose CPUs that can keep up with the GPU processing demands, especially in multi-GPU setups.
- **Memory**: The memory installed in the server should be 2x the size of the GPU memory.

Networking

Building a network for a generative AI use case using a non-blocking topology involves designing a network infrastructure that minimizes data transfer bottlenecks, maximizes communication bandwidth, and ensures low latency.

Here are some best practices to consider:

- **Understand Workload Communication Patterns**: Analyze how your generative AI workloads communicate with each other. Identify patterns such as all-to-all communication, one-to-all communication, or point-to-point communication.
- **Choose a Non-Blocking Topology**: A non-blocking topology ensures that every node can communicate simultaneously with all other nodes without contention or blocking. Fat Tree, Clos, and Hypercube are examples of non-blocking topologies.
- **Topology Selection**: Evaluate the communication patterns of your workload to determine the best non-blocking topology. Consider the number of nodes, switch ports, and potential future scalability.
- **High-Speed Interconnect**: Select a high-speed networking technology like InfiniBand or high-speed Ethernet (e.g., 200 Gbps or higher) to ensure sufficient bandwidth for data-intensive AI workloads. It is important to provide 200Gb bandwidth to each GPU
- **Switch Selection**: Choose switches with low latency and high throughput to minimize communication delays. Consider managed switches that support features like Quality of Service (QoS) and traffic prioritization.
- **Network Design**: Aim for a balanced design where each node has multiple paths to other nodes. This reduces the likelihood of bottlenecks and ensures redundancy.
- **Minimize Network Congestion**: Implement techniques like flow control, traffic shaping, and adaptive routing to minimize network congestion and prevent hotspots.
- **Low Latency and Jitter**: Prioritize low-latency communication for real-time generative AI applications. Ensure that your network design minimizes packet jitter (variation in delay) to maintain consistent performance.
- **Optimized Routing**: Utilize adaptive routing algorithms that dynamically choose the best path based on network conditions to avoid congested links.
- **Future Scalability**: Design the network with scalability in mind. Choose a topology and infrastructure that can accommodate the addition of more nodes and resources as your generative AI workload grows.

Management software

- **Containerization**: Utilize containerization platforms like Docker and Kubernetes to manage and deploy generative AI workloads consistently across the cluster.
- **Orchestration**: Leverage Kubernetes or a platform like VMware Tanzu to automate workload deployment, scaling, and management.

AI software stack

- **Deep Learning Frameworks**: Choose frameworks like TensorFlow, PyTorch, or Apache MXNet for building and training generative AI models. These frameworks have GPU acceleration support.
- **GPU Drivers**: Install the latest GPU drivers and CUDA toolkit for optimal GPU utilization and performance.
- InfiniBand Drivers: Ensure the correct InfiniBand drivers are installed and configured to make the most of the high-speed networking.

Parallelism and optimization

- **Model Parallelism**: If your generative AI model is too large for a single GPU, consider model parallelism techniques to distribute the model's layers across multiple GPUs.
- **Data Parallelism**: Utilize data parallelism to distribute training data across GPUs, speeding up training by processing batches concurrently.

Deployment flexibility

- **Scalability**: Design the cluster architecture to accommodate future scalability needs by adding more GPUs, servers, and networking resources as required.

By following these practices, you can build a robust and high-performance generative AI cluster that takes full advantage of GPU acceleration and InfiniBand networking for faster training and inference of complex AI models.

Bill of materials

In this section, we provide a detailed and repeatable configurations for the example configurations covered in the Configuration examples section, using InfiniBand networking. We also provide the configuration bill of materials for the server node with the high-speed Ethernet option.

In this section, we provide a detailed and repeatable configuration. The configurations the Configuration examples section use InfiniBand networking. We also provide the configuration bill of materials for the server node with the high-speed Ethernet option. It should be noted, that the ThinkSystem SR675 V3 can be either a 4 DW PCIe GPU base or an 8 DW PCIe GPU Base. The part numbers provided below are for the 8 DW option.

You can build your solution using the Lenovo Data Center Solution Configurator, DCSC: https://dcsc.lenovo.com

Tables in this section:

Small config

Table 1. Small config

Part number	Product Description	Qty
7D9RCTOLWW	Server: ThinkSystem SR675 V3 - 3yr Warranty - HPC&AI	4
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	4
BFYA	Operating mode selection for: "Maximum Efficiency Mode"	4
BPVJ	ThinkSystem AMD EPYC 9554 64C 360W 3.1GHz Processor	8
BQ3D	ThinkSystem 64GB TruDDR5 4800MHz (2Rx4) 10x4 RDIMM-A	96
5977	Select Storage devices - no configured RAID required	4
BFTQ	ThinkSystem 1x6 E1.S EDSFF Backplane Option Kit	4
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	4
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	8
BG3F	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	48

Part number	Product Description	Qty
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/HDR QSFP112 2-Port PCIe Gen5 x16 InfiniBand Adapter	20
B93E	ThinkSystem Intel I350 1GbE RJ45 4-port OCP Ethernet Adapter	4
BR9U	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	32
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	8
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	4
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	4
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	8
B962	ThinkSystem 2400W 230V Platinum Hot-Swap Gen2 Power Supply	16
B4L2	2.0m, 16A/100-250V, C19 to IEC 320-C20 Rack Power Cable	16
BFTL	ThinkSystem SR670 V2/ SR675 V3 Toolless Slide Rail	4
3803	3m Blue Cat5e Cable	4
3793	3m Yellow Cat5e Cable	4
B7XZ	Disable IPMI-over-LAN	4
BR7V	ThinkSystem SR675 V3 System Board	4
BK15	High voltage (200V+)	4
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	4
BE0D	N+1 Redundancy with Over-Subscription	4
BR82	ThinkSystem SR670 V2/ SR675 V3 WW Packaging	4
BR80	ThinkSystem SR675 V3 Agency Labels	4
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	4
B993	ThinkSystem V2 EDSFF Filler	24
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	4
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	4
BRUC	ThinkSystem SR675 V3 CPU Heatsink	8
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	4
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	4
BR7U	ThinkSystem SR675 V3 Root of Trust Module	4
BS03	ThinkSystem SR675 V3 2400W Power Supply Caution Label	4
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	32
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	4
BR8G	ThinkSystem SR675 V3 Rear PCIe Riser Cable 4	4
BU22	ThinkSystem SR675 V3 Rear PCIe Riser Cable 6	4
BU23	ThinkSystem SR675 V3 Front OCP Cable 2	4
BR8Q	ThinkSystem SR675 V3 Front PCIe Riser Cable 6	4
BR8V	ThinkSystem SR675 V3 Front PCIe Riser Cable 2	4
BFTM	ThinkSystem SR670 V2/ SR675 V3 EDSFF Cage	4
BRUL	ThinkSystem SR675 V3 EDSFF Drive Sequence Label	4
BFGZ	ThinkSystem SR670 V2/ SR675 V3 Backplane Power Cable 4	4
BRUQ	ThinkSystem SR675 V3 EDSFF to Riser Cables	4
BF94	AI & HPC - ThinkSystem Hardware	4

Part number	Product Description	Qty
5PS7B09635	Premier Essential - 3Yr 24x7 4Hr Resp + YDYD SR675 V3	4
5AS7A82992	Hardware Installation (Business Hours) for SR67x	4
5641PX3	XClarity Pro, Per Endpoint w/3 Yrs. SW S&S	4
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yrs. SW S&S	4
3444	Registration only	4
0724HEC	Switch: NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BP63	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
3791	0.6m Yellow Cat5e Cable	1
BQK3	Lenovo 3m NVIDIA NDRx2 OSFP800 to 4x NDR200 QSFP112 Passive Copper Splitter Cable	8
BRQ6	2.8m, 10A/100-250V, C15 to C14 Jumper Cord	2
BRET	NVIDIA QM97xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1
5WS7B14266	Premier Essential - 3Yr 24x7 4Hr Resp NVID QM9700 PSE	1
7D5FCTO1WW-HPC	Switch: Mellanox AS4610-54T 1GbE Managed Switch with Cumulus (PSE)	1
BE2J	Mellanox AS4610-54T 1GbE Managed Switch with Cumulus (PSE)	1
3792	1.5m Yellow Cat5e Cable	1
6311	2.8m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
BEGG	Mellanox AS46xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1

Medium config

Table 2. Medium config

Part number	Product Description	Qty
7D9RCTOLWW	Server: ThinkSystem SR675 V3 - 3yr Warranty - HPC&AI	8
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	8
BFYA	Operating mode selection for: "Maximum Efficiency Mode"	8
BPVJ	ThinkSystem AMD EPYC 9554 64C 360W 3.1GHz Processor	16
BQ3D	ThinkSystem 64GB TruDDR5 4800MHz (2Rx4) 10x4 RDIMM-A	192
5977	Select Storage devices - no configured RAID required	8
BFTQ	ThinkSystem 1x6 E1.S EDSFF Backplane Option Kit	8
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	8
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	16
BG3F	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	96
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/HDR QSFP112 2-Port PCIe Gen5 x16 InfiniBand Adapter	40
B93E	ThinkSystem Intel I350 1GbE RJ45 4-port OCP Ethernet Adapter	8
BR9U	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	64
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	16
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	8

Part number	Product Description	Qty
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	8
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	16
B962	ThinkSystem 2400W 230V Platinum Hot-Swap Gen2 Power Supply	32
B4L2	2.0m, 16A/100-250V, C19 to IEC 320-C20 Rack Power Cable	32
BFTL	ThinkSystem SR670 V2/ SR675 V3 Toolless Slide Rail	8
3803	3m Blue Cat5e Cable	8
3793	3m Yellow Cat5e Cable	8
B7XZ	Disable IPMI-over-LAN	8
BR7V	ThinkSystem SR675 V3 System Board	8
BK15	High voltage (200V+)	8
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	8
BE0D	N+1 Redundancy with Over-Subscription	8
BR82	ThinkSystem SR670 V2/ SR675 V3 WW Packaging	8
BR80	ThinkSystem SR675 V3 Agency Labels	8
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	8
B993	ThinkSystem V2 EDSFF Filler	48
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	8
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	8
BRUC	ThinkSystem SR675 V3 CPU Heatsink	16
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	8
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	8
BR7U	ThinkSystem SR675 V3 Root of Trust Module	8
BS03	ThinkSystem SR675 V3 2400W Power Supply Caution Label	8
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	64
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	8
BR8G	ThinkSystem SR675 V3 Rear PCIe Riser Cable 4	8
BU22	ThinkSystem SR675 V3 Rear PCIe Riser Cable 6	8
BU23	ThinkSystem SR675 V3 Front OCP Cable 2	8
BR8Q	ThinkSystem SR675 V3 Front PCIe Riser Cable 6	8
BR8V	ThinkSystem SR675 V3 Front PCIe Riser Cable 2	8
BFTM	ThinkSystem SR670 V2/ SR675 V3 EDSFF Cage	8
BRUL	ThinkSystem SR675 V3 EDSFF Drive Sequence Label	8
BFGZ	ThinkSystem SR670 V2/ SR675 V3 Backplane Power Cable 4	8
BRUQ	ThinkSystem SR675 V3 EDSFF to Riser Cables	8
BF94	AI & HPC - ThinkSystem Hardware	8
5PS7B09635	Premier Essential - 3Yr 24x7 4Hr Resp + YDYD SR675 V3	8
5AS7A82992	Hardware Installation (Business Hours) for SR67x	8
5641PX3	XClarity Pro, Per Endpoint w/3 Yrs. SW S&S	8
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yrs. SW S&S	8
3444	Registration only	8
0724HEC	Switch: NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1

Part number	Product Description	Qty
BP63	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
3791	0.6m Yellow Cat5e Cable	1
BQK3	Lenovo 3m NVIDIA NDRx2 OSFP800 to 4x NDR200 QSFP112 Passive Copper Splitter Cable	8
BRQ6	2.8m, 10A/100-250V, C15 to C14 Jumper Cord	2
BRET	NVIDIA QM97xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1
5WS7B14266	Premier Essential - 3Yr 24x7 4Hr Resp NVID QM9700 PSE	1
7D5FCTO1WW-HPC	Switch: Mellanox AS4610-54T 1GbE Managed Switch with Cumulus (PSE)	1
BE2J	Mellanox AS4610-54T 1GbE Managed Switch with Cumulus (PSE)	1
3792	1.5m Yellow Cat5e Cable	1
6311	2.8m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
BEGG	Mellanox AS46xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1

Large config

Table 3. Large config

Part number	Product Description	Qty
7D9RCTOLWW	Rack 1 - Compute: ThinkSystem SR675 V3 - 3yr Warranty - HPC&AI	8
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	8
BFYA	Operating mode selection for: "Maximum Efficiency Mode"	8
BPVJ	ThinkSystem AMD EPYC 9554 64C 360W 3.1GHz Processor	16
BQ3D	ThinkSystem 64GB TruDDR5 4800MHz (2Rx4) 10x4 RDIMM-A	192
5977	Select Storage devices - no configured RAID required	8
BFTQ	ThinkSystem 1x6 E1.S EDSFF Backplane Option Kit	8
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	8
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	16
BG3F	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	96
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/HDR QSFP112 2-Port PCIe Gen5 x16 InfiniBand Adapter	40
B93E	ThinkSystem Intel I350 1GbE RJ45 4-port OCP Ethernet Adapter	8
BR9U	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	64
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	16
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	8
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	8
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	16
B962	ThinkSystem 2400W 230V Platinum Hot-Swap Gen2 Power Supply	32
B4L2	2.0m, 16A/100-250V, C19 to IEC 320-C20 Rack Power Cable	32
BFTL	ThinkSystem SR670 V2/ SR675 V3 Toolless Slide Rail	8
3803	3m Blue Cat5e Cable	8

Part number	Product Description	Qty
3793	3m Yellow Cat5e Cable	8
B7XZ	Disable IPMI-over-LAN	8
BR7V	ThinkSystem SR675 V3 System Board	8
BK15	High voltage (200V+)	8
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	8
BE0D	N+1 Redundancy with Over-Subscription	8
BR82	ThinkSystem SR670 V2/ SR675 V3 WW Packaging	8
BR80	ThinkSystem SR675 V3 Agency Labels	8
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	8
B993	ThinkSystem V2 EDSFF Filler	48
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	8
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	8
BRUC	ThinkSystem SR675 V3 CPU Heatsink	16
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	8
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	8
BR7U	ThinkSystem SR675 V3 Root of Trust Module	8
BS03	ThinkSystem SR675 V3 2400W Power Supply Caution Label	8
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	64
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	8
BR8G	ThinkSystem SR675 V3 Rear PCIe Riser Cable 4	8
BU22	ThinkSystem SR675 V3 Rear PCIe Riser Cable 6	8
BU23	ThinkSystem SR675 V3 Front OCP Cable 2	8
BR8Q	ThinkSystem SR675 V3 Front PCIe Riser Cable 6	8
BR8V	ThinkSystem SR675 V3 Front PCIe Riser Cable 2	8
BFTM	ThinkSystem SR670 V2/ SR675 V3 EDSFF Cage	8
BRUL	ThinkSystem SR675 V3 EDSFF Drive Sequence Label	8
BFGZ	ThinkSystem SR670 V2/ SR675 V3 Backplane Power Cable 4	8
BRUQ	ThinkSystem SR675 V3 EDSFF to Riser Cables	8
BF94	AI & HPC - ThinkSystem Hardware	8
5PS7B09635	Premier Essential - 3Yr 24x7 4Hr Resp + YDYD SR675 V3	8
5AS7A82992	Hardware Installation (Business Hours) for SR67x	8
5641PX3	XClarity Pro, Per Endpoint w/3 Yrs. SW S&S	8
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yrs. SW S&S	8
3444	Registration only	8
0724HEC	Rack 1 - Compute Switch: NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BP63	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
3791	0.6m Yellow Cat5e Cable	1
BQK3	Lenovo 3m NVIDIA NDRx2 OSFP800 to 4x NDR200 QSFP112 Passive Copper Splitter Cable	16
BQJK	Lenovo 3m NVIDIA NDRx2 OSFP800 to NDRx2 OSFP800 Active Copper Cable	8

Part number	Product Description	Qty
BRQ6	2.8m, 10A/100-250V, C15 to C14 Jumper Cord	2
BRET	NVIDIA QM97xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1
5WS7B14266	Premier Essential - 3Yr 24x7 4Hr Resp NVID QM9700 PSE	1
7D5FCTO1WW-HPC	Rack 1 - Mgmt. Switch: Mellanox AS4610-54T 1GbE Managed Switch with Cumulus (PSE)	1
BE2J	Mellanox AS4610-54T 1GbE Managed Switch with Cumulus (PSE)	1
3792	1.5m Yellow Cat5e Cable	1
6311	2.8m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
BEGG	Mellanox AS46xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1
7D9RCTOLWW	Rack 2 - Compute: ThinkSystem SR675 V3 - 3yr Warranty - HPC&AI	8
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	8
BFYA	Operating mode selection for: "Maximum Efficiency Mode"	8
BPVJ	ThinkSystem AMD EPYC 9554 64C 360W 3.1GHz Processor	16
BQ3D	ThinkSystem 64GB TruDDR5 4800MHz (2Rx4) 10x4 RDIMM-A	192
5977	Select Storage devices - no configured RAID required	8
BFTQ	ThinkSystem 1x6 E1.S EDSFF Backplane Option Kit	8
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	8
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	16
BG3F	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	96
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/HDR QSFP112 2-Port PCIe Gen5 x16 InfiniBand Adapter	40
B93E	ThinkSystem Intel I350 1GbE RJ45 4-port OCP Ethernet Adapter	8
BR9U	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	64
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	16
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	8
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	8
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	16
B962	ThinkSystem 2400W 230V Platinum Hot-Swap Gen2 Power Supply	32
B4L2	2.0m, 16A/100-250V, C19 to IEC 320-C20 Rack Power Cable	32
BFTL	ThinkSystem SR670 V2/ SR675 V3 Toolless Slide Rail	8
3803	3m Blue Cat5e Cable	8
3793	3m Yellow Cat5e Cable	8
B7XZ	Disable IPMI-over-LAN	8
BR7V	ThinkSystem SR675 V3 System Board	8
BK15	High voltage (200V+)	8
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	8
BE0D	N+1 Redundancy with Over-Subscription	8
BR82	ThinkSystem SR670 V2/ SR675 V3 WW Packaging	8
BR80	ThinkSystem SR675 V3 Agency Labels	8
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	8

Part number	Product Description	Qty
B993	ThinkSystem V2 EDSFF Filler	48
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	8
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	8
BRUC	ThinkSystem SR675 V3 CPU Heatsink	16
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	8
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	8
BR7U	ThinkSystem SR675 V3 Root of Trust Module	8
BS03	ThinkSystem SR675 V3 2400W Power Supply Caution Label	8
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	64
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	8
BR8G	ThinkSystem SR675 V3 Rear PCIe Riser Cable 4	8
BU22	ThinkSystem SR675 V3 Rear PCIe Riser Cable 6	8
BU23	ThinkSystem SR675 V3 Front OCP Cable 2	8
BR8Q	ThinkSystem SR675 V3 Front PCIe Riser Cable 6	8
BR8V	ThinkSystem SR675 V3 Front PCIe Riser Cable 2	8
BFTM	ThinkSystem SR670 V2/ SR675 V3 EDSFF Cage	8
BRUL	ThinkSystem SR675 V3 EDSFF Drive Sequence Label	8
BFGZ	ThinkSystem SR670 V2/ SR675 V3 Backplane Power Cable 4	8
BRUQ	ThinkSystem SR675 V3 EDSFF to Riser Cables	8
BF94	AI & HPC - ThinkSystem Hardware	8
5PS7B09635	Premier Essential - 3Yr 24x7 4Hr Resp + YDYD SR675 V3	8
5AS7A82992	Hardware Installation (Business Hours) for SR67x	8
5641PX3	XClarity Pro, Per Endpoint w/3 Yrs. SW S&S	8
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yrs. SW S&S	8
3444	Registration only	8
0724HEC	Rack 2 - Compute Switch: NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BP63	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
3793	3m Yellow Cat5e Cable	1
BQK3	Lenovo 3m NVIDIA NDRx2 OSFP800 to 4x NDR200 QSFP112 Passive Copper Splitter Cable	16
BQJK	Lenovo 3m NVIDIA NDRx2 OSFP800 to NDRx2 OSFP800 Active Copper Cable	8
BRQ6	2.8m, 10A/100-250V, C15 to C14 Jumper Cord	2
BRET	NVIDIA QM97xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1
5WS7B14266	Premier Essential - 3Yr 24x7 4Hr Resp NVID QM9700 PSE	1

Repeatable node with high-speed Ethernet

Table 4. Repeatable node with high-speed Ethernet

Part number	Product Description	Qty
7D9RCTO1WW	Server: ThinkSystem SR675 V3 - 3yr Warranty	1
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	1
BFYB	Operating mode selection for: "Maximum Performance Mode"	1
BR2Z	ThinkSystem AMD EPYC 9634 84C 290W 2.25GHz Processor	2
BQ3A	ThinkSystem 128GB TruDDR5 4800MHz (4Rx4) 3DS RDIMM-A	24
5977	Select Storage devices - no configured RAID required	1
BPKW	ThinkSystem E1.S 5.9mm 7450 PRO 7.68TB Read Intensive NVMe PCIe 4.0 x4 HS SSD	6
BFTQ	ThinkSystem 1x6 E1.S EDSFF Backplane Option Kit	1
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	1
BKSS	ThinkSystem M.2 7450 PRO 1.92TB Read Intensive Entry NVMe PCIe 4.0 x4 NHS SSD	2
BG3F	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	12
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/HDR QSFP112 2-Port PCIe Gen5 x16 InfiniBand Adapter	5
BE4T	ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1
BR9U	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	8
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	1
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	2
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	1
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	2
BKTJ	ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supply	4
6252	2.5m, 16A/100-250V, C19 to IEC 320-C20 Rack Power Cable	4
BFTL	ThinkSystem SR670 V2/ SR675 V3 Toolless Slide Rail	1
B7XZ	Disable IPMI-over-LAN	1
BR7V	ThinkSystem SR675 V3 System Board	1
BK15	High voltage (200V+)	1
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	1
BE0D	N+1 Redundancy with Over-Subscription	1
BR82	ThinkSystem SR670 V2/ SR675 V3 WW Packaging	1
BR80	ThinkSystem SR675 V3 Agency Labels	1
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	1
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	1
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	1
BRUC	ThinkSystem SR675 V3 CPU Heatsink	2
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	1
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	1
BR7U	ThinkSystem SR675 V3 Root of Trust Module	1
BRNM	ThinkSystem SR670 V2/SR675 V3 2600W Power Supply Caution Label	1
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	8
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	1

Part number	Product Description	Qty
BR8G	ThinkSystem SR675 V3 Rear PCIe Riser Cable 4	1
BU22	ThinkSystem SR675 V3 Rear PCIe Riser Cable 6	1
BU23	ThinkSystem SR675 V3 Front OCP Cable 2	1
BR8Q	ThinkSystem SR675 V3 Front PCIe Riser Cable 6	1
BR8V	ThinkSystem SR675 V3 Front PCIe Riser Cable 2	1
BFTM	ThinkSystem SR670 V2/ SR675 V3 EDSFF Cage	1
BRUL	ThinkSystem SR675 V3 EDSFF Drive Sequence Label	1
BFGZ	ThinkSystem SR670 V2/ SR675 V3 Backplane Power Cable 4	1
BRUQ	ThinkSystem SR675 V3 EDSFF to Riser Cables	1
7S02CTO1WW	NVIDIA Software	1
S6Z3	NVIDIA AI Enterprise Subscription License and Support per GPU Socket, 3 Years	8

NVIDIA AI Enterprise, Lenovo Intelligent Computing Orchestration and VMware

Table 5. NVIDIA AI Enterprise, Lenovo Intelligent Computing Orchestration and VMware

Part Number	Description	QTY
7S06077MWW	VMware Cloud Foundation 4 Advanced (Per CPU) w/Lenovo 3Yr S&S	Depends on the number of CPU sockets in your Design
7S02001GWW	NVIDIA AI Enterprise Subscription License and Support per GPU Socket, 3 Years	Depends on the number of GPU Sockets in your Design
7S090008WW	Lenovo K8S AI LiCO Software 4GPU w/3Yr S&S	Depends on the number of GPU Sockets in your Design

NVIDIA RIVA needs to be integrated through the Lenovo VLS process.

Appendix 1: NVIDIA OVX L40S Nodes Integration

We offer the flexibility to integrate NVIDIA OVX L40S nodes to this reference architecture. These nodes are based on the ThinkSystem SR675 V3.

The following figure show the configuration. Each SR675 V3 server has an BlueField-3 DPU installed for north-to-south bandwidth.

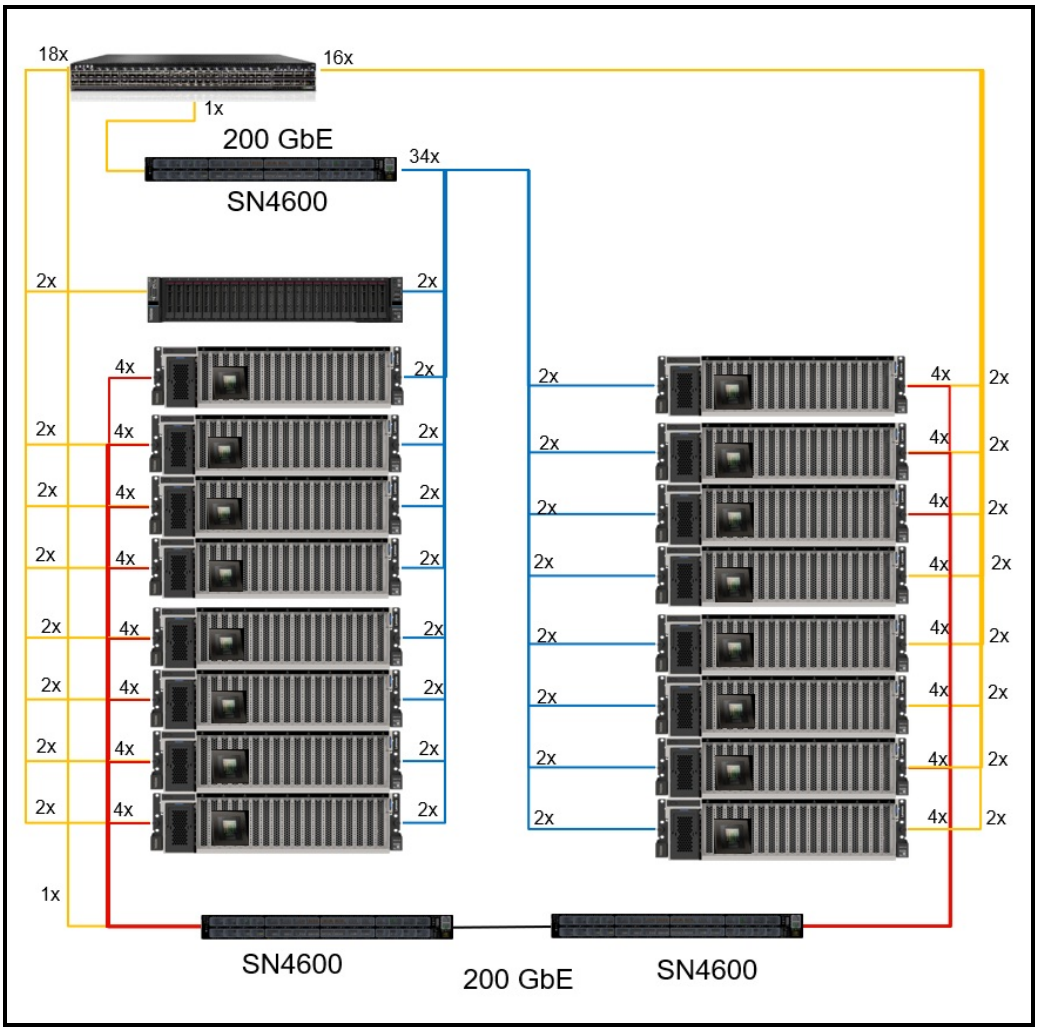

Figure 24. NVIDIA OVX L40S Nodes Integration

The configuration is as follows:

- 1 Master Node
- 16 Compute Nodes – memory could be reduced to 1.5TB
  - 4x NVIDIA L40S GPUs
  - 2x ConnectX-7 dual-port 200GbE
  - 1x BlueField-3 dual-port 200GbE Ethernet for virtualization
  - 1 GbE adapter for management network
- VMware Cloud Foundation
- VMware Tanzu + Service installer
- 64 NVIDIA AI Enterprise licenses
- 3 Spectrum SN4600 switches

References

For more information, see these resources:

- Papers with Code: A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need?
  https://paperswithcode.com/paper/a-complete-survey-on-generative-ai-aigc-is
- McKinsey: What is ChatGPT, DALL-E, and generative AI?
  https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai
- Writer: The State of Generative AI in the Enterprise
  https://writer.com/guides/generative-ai-survey/
- KPMG U.S. survey: Executives expect generative AI to have enormous impact on business, but unprepared for immediate adoption
  https://info.kpmg.us/news-perspectives/technology-innovation/kpmg-generative-ai-2023.html
- SilionANGLE: Salesforce survey shows IT interest in generative AI tempered with technical, ethical concerns
  https://siliconangle.com/2023/03/06/salesforce-survey-shows-interest-generative-ai-tempered-technical-ethical-concerns/
- NVIDIA H100 Tensor Core GPU Architecture Overview
  https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper
- NVIDIA L40S datasheet
  https://resources.nvidia.com/en-us-l40s/l40s-datasheet-28413
- Lenovo ThinkSystem SR675 V3 Server
  https://lenovopress.lenovo.com/lp1611-thinksystem-sr675-v3-server
- ThinkSystem NVIDIA ConnectX-7 NDR InfiniBand OSFP400 Adapters
  https://lenovopress.lenovo.com/lp1692
- NVIDIA BlueField-3 DPU datasheet
  https://resources.nvidia.com/en-us-accelerated-networking-resource-library/datasheet-nvidia-bluefield?lx=LbHvpR&topic=networking-cloud
- NVIDIA Spectrum SN4000 Series Switches
  https://www.nvidia.com/content/dam/en-zz/Solutions/networking/br-sn4000-series.pdf
- Lenovo EveryScale (formerly Lenovo Scalable Infrastructure or LeSI)
  https://lenovopress.lenovo.com/lp0900-lenovo-everyscale-lesi#ethernet-switches
- MLflow vs KubeFlow: Architecture and Key Differences (run.ai)
  https://www.run.ai/guides/machine-learning-operations/mlflow-vs-kubeflow
- Databricks: What is a medallion architecture?
  https://www.databricks.com/glossary/medallion-architecture
- Medium: Kubeflow Pros and Cons: Kubeflow/Vertex AI vs Airflow vs SageMaker
  https://medium.com/datasparq-technology/kubeflow-pros-and-cons-kubeflow-vs-airflow-vs-sagemaker-4942d7e7910a
- Atlas AI Cloud Platform - Run:ai
  https://www.run.ai/platform
- NVIDIA AI Enterprise Release Notes
  https://docs.nvidia.com/ai-enterprise/latest/release-notes/index.html
- Lenovo XClarity Administrator
  https://lenovopress.lenovo.com/tips1200-lenovo-xclarity-administrator

Authors

Robert Daigle is the Director for Marketing Strategy & Planning at Lenovo. He has held multiple leadership roles within the tech industry, from software startups to Fortune 500 companies. Robert was instrumental in launching one of the first AI recruiting platforms before his current post in Lenovo, where he leads the strategy & business development for Lenovo's AI business.

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.

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, LP1798, was created or updated on December 15, 2023.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1798
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1798.

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 Neptune®
ThinkAgile®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

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

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

Bing® is a trademark of Microsoft Corporation in the United States, other countries, or both.

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