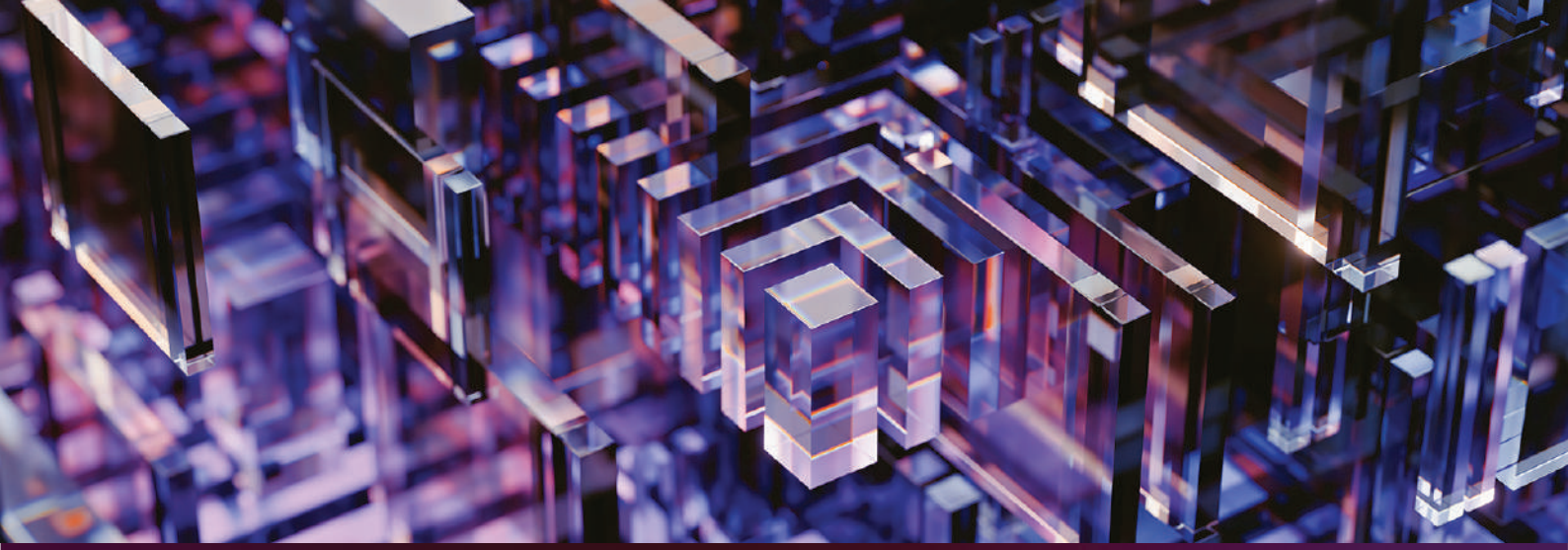


Delivering High-Performance Infrastructure for Generative and Enterprise AI

Lenovo and NVIDIA® Solutions to Enhance Productivity, Innovation, and Time-to-Market

Executive Summary

Across many sectors, Artificial Intelligence (AI) and Generative AI (GenAI) can accelerate innovation and improve a company's competitive position, the quality of products/services, operations, and customer engagement. However, numerous implementation challenges exist while deploying real-life use cases.

Lenovo helps enterprises overcome these obstacles by providing a comprehensive set of best-practices services and an AI-optimized infrastructure of servers, storage, workstations, mobile devices, and software from the edge to the data center to the cloud. For example, the Lenovo ThinkSystem and ThinkEdge servers, powered by NVIDIA graphics processing units (GPUs) and software, can accelerate a customer's AI and GenAI journey with:

- Faster time-to-results for training and inference workloads across text, video, image, and other data modalities
- More flexibility to customize and optimize various AI, GenAI, and other related enterprise workloads from the edge to the data center to the cloud
- Better energy efficiency and lower total cost of ownership (TCO)
- Several complementary immersive offerings and services to facilitate a company's digital transformation with AI and GenAI, including access to the Lenovo AI Discover Center of Excellence (AIDiscover@lenovo.com)

Introduction

AI and GenAI solutions are rapidly growing in the enterprise, providing many benefits across many industries. Companies are implementing AI and GenAI to accelerate innovation and improve their competitive position, the quality of products/services, operations, and customer engagement.

While AI's promise and economic value are immense, so are the challenges of implementing AI, given the large datasets and the need to effectively store, analyze, and protect all your valuable data throughout its lifecycle. With GenAI, these issues get even more acute.

This whitepaper discusses how Lenovo and NVIDIA partner with their respective unique technologies to provide the optimal architecture to deliver AI and GenAI for enterprises. Based on Lenovo and NVIDIA's engagements with customers and partners, this paper offers valuable guidance to select performance-optimized configurations for several AI and GenAI use cases across many industries to gain a competitive edge. Lenovo continues investing[1] in AI partnerships, including NVIDIA, to accelerate AI deployment for enterprises worldwide to help customers start their AI and GenAI journey today!

AI and GenAI Drive Enterprise Value Across Many Industries

AI, which includes Machine Learning (ML) and Deep Learning (DL), is rapidly growing and transforming a wide range of industries and applications. The global AI market is expected to reach US $241.80 billion in 2023 and is forecasted to grow at a CAGR of 17.30% from 2023 to 2030, resulting in a market volume of US $738.80 billion by 2030.[2]

GenAI, including Large Language Models (LLMs), is a powerful, new type of DL that can create new content, such as text, images, audio, and video. It does this by learning patterns from existing data and then using this knowledge to generate new and unique outputs. GenAI can produce highly realistic and complex content that mimics human creativity. It is growing even faster than AI[3] (over 58%) and becoming a valuable tool for many industries, such as financial services, healthcare, manufacturing, retail, telecommunications/media, etc.

Figure 1 depicts several prominent GenAI use cases summarized from a recent Deloitte study[4] that add significant business value across these industries by extracting deep, actionable insights across many data modalities (Text, Audio, Image, Video, Code, and 3D/Specialized artifacts). Solid, rich colors in the pie chart represent the data modalities typically prominent for each use case. The white color in the pie chart is for data modalities that are usually not significant.

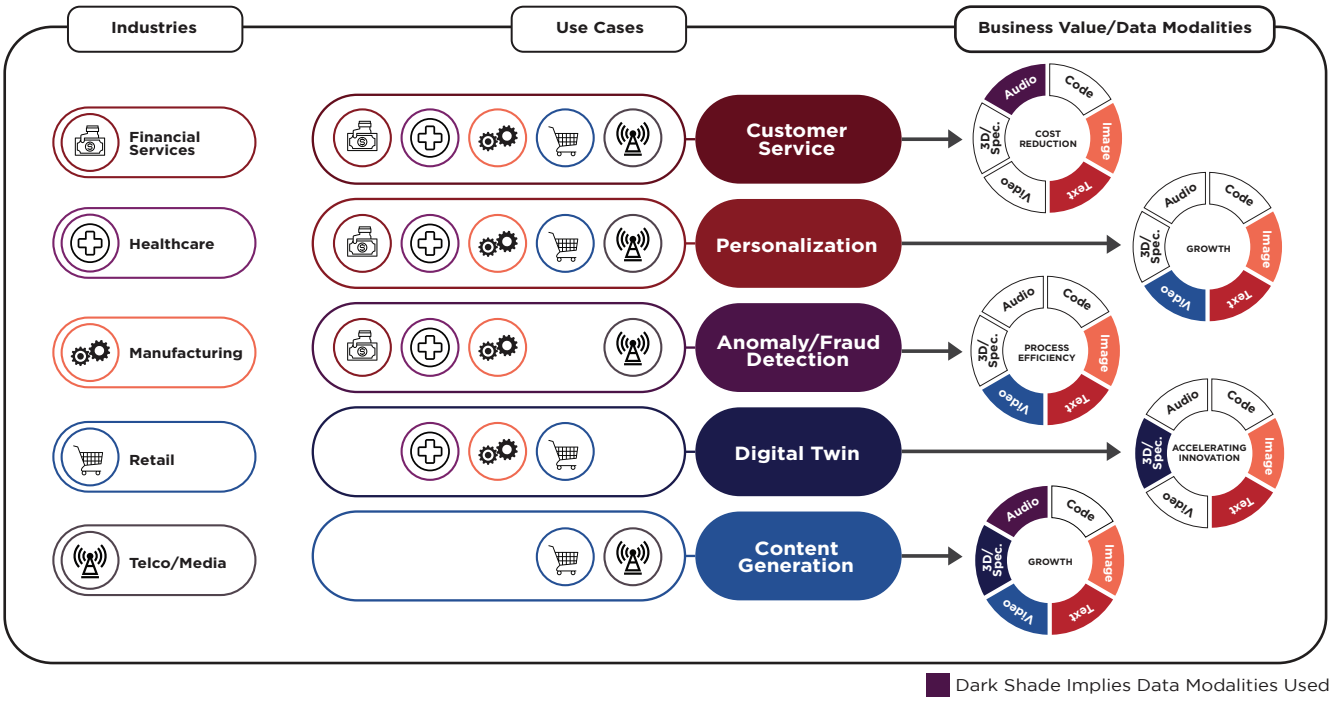


Figure 1: High-Value GenAI Use Cases Across Many Industries and Data Modalities[4]

Financial Services: Banks and insurance companies are adding GenAI to their data-intensive processes to improve:

- **Customer Service:** GenAI-powered digital avatar interface options with text, audio, and imagery enhance 24/7 customer support, answer queries, and assist with financial tasks to improve process efficiency and customer engagement.
- **Personalization:** Deliver regulatory-compliant marketing materials, product promotions, and sales engagement with customized text, images, and videos across different geographies to drive growth and acquire new customers.
- **Fraud Detection:** Identify in real-time fraudulent transactions by analyzing patterns and anomalies in real and synthetic data across several modalities, helping to improve processes and prevent financial losses.

Healthcare: Payers, providers, pharmaceutical, and biotech organizations are adding GenAI for:

- **Customer Service:** Accelerate prior authorization for patients and generate responses to questions about the claims process, insurance coverage, and other plan details to improve the process.

  Enable 24/7, continuous proactive patient monitoring and care with IoT devices and AI-powered analytics on data to monitor patient vital signs, alert healthcare providers to deviations from typical values, and take remedial action.
- **Personalization:** Discover and tailor treatments and drugs to individual patients based on their genetic makeup and medical history to enhance the effectiveness of care and grow the business and competitive advantage.
- **Fraud/Anomaly Detection:** Identify fraudulent claims in real-time by analyzing patterns and anomalies in data across several modalities, helping improve processes and prevent financial losses. Analyze medical images such as X-rays, MRIs, and CT scans to assist in the early detection of diseases and abnormalities.
- **Digital Twin:** Build end-to-end patient-centric digital replicas to analyze patient data, including medical records/images and symptoms, to improve diagnosing diseases and recommending treatment plans. Identify potential drug candidates, predict their effectiveness, and optimize molecular structures.

Predict disease outbreaks, patient readmissions, and healthcare resource utilization, helping hospitals and clinics allocate resources efficiently. All these drive more innovation across the healthcare ecosystem.

Manufacturing: Automotive, Aerospace, and Semiconductor manufacturers are adding GenAI for:

- **Maintenance:** Analyze machine sensor data across several modalities to predict when they could fail. It helps perform preventive services and avoid costly downtime and disruptions.
- **Personalization:** Enable mass customization by analyzing data and efficiently adapting manufacturing processes to produce personalized products. It drives greater customer appeal and business growth.
- **Anomaly Detection:** AI-powered computer vision systems can rapidly and accurately inspect products for defects, reducing the number of faulty items in the production line to drive process efficiency.
- **Digital Twin:** Build an end-to-end digital replica of the entire product lifecycle, from development to manufacturing to service. It helps generate and evaluate new product designs, optimizing them for performance, cost, and manufacturability. Optimize manufacturing processes by analyzing data from sensors and production lines to improve efficiency and product quality. Provides real-time insights into the supply chain and customer operation, helping manufacturers track raw materials, monitor production progress, respond to disruptions, track actual customer use of their products, and ensure timely maintenance. All this dramatically enhances product and process innovation and quality.

Retail: Consumer-facing companies such as large, small, and specialized retailers are using GenAI for:

- **Customer Service:** GenAI-powered digital avatar interface options with text, audio, and imagery enhance 24/7 customer support, answer queries, and assist with product recommendations to improve process efficiency and empathetic customer engagement to build brand loyalty and equity. It also frees up expensive human resources to tackle more complex customer issues.
- **Personalization**: Deliver marketing materials, product promotions, and sales engagement with customized text, images, and videos across different geographies to drive growth and acquire new customers. Generate more specific and targeted recommendations across many modalities than search engines to make buying more personalized and convenient.
- **Digital Twin:** Create virtual showrooms, product demonstrations, and planograms, personalize the customer experience, forecast demand, simulate the layout and operations of a store, and identify areas for improvement. All this can help retailers innovate more.
- **Content Generation:** Create product descriptions, imagery, video, and more much faster and consistently than traditional tools and processes. To grow the business, personalize this content by geography, language, cultural nuances, and local regulations.

Telco/Media: With GenAI, these companies are accelerating digital transformation with better:

- **Customer Service:** GenAI-powered digital avatar and virtual assistance interface options with text, audio, and imagery enhance 24/7 customer support, answer queries, and assist with service recommendations to improve process efficiency and empathetic customer engagement to build loyalty and contain switching costs. It also frees up expensive human resources to tackle more complex customer issues. Optimize network performance and reduce congestion, improving customer experience.
- **Personalization:** Organize and manage complex file types, analyze content before translation to optimize localization, and integrate other language tools into the workflow to increase conversations and engagement to build loyalty. Speech recognition helps transcribe video and audio content into text and translate spoken content into other languages to grow the business.
- **Fraud detection:** Use real and synthetic data to improve process efficiency and detect fraudulent activity on telecommunications networks, such as SIM swapping and unauthorized access.
- **Content Generation**: Mimic the style of company marketing materials and generate new and high-quality multiple versions of content rapidly and on-demand tailored to different audiences. Improve the language quality of marketing materials with phrasing, grammar, company style, and adherence to company values. Rapidly create numerous versions of content in various styles to identify the best option to grow the business.

While the ROI from GenAI can be substantial for the enterprise, deploying DL and the associated high-performance information technology (IT) infrastructure can be complex and expensive. There are numerous implementation challenges.

To discuss your specific use case contact the Lenovo AI Discover Lab by emailing AIDiscover@lenovo.com.

AI and GenAI Implementation Challenges

Deploying DL and GenAI workflows in production typically has four stages (Figure 2)[5]:

1. **Data Management** to prepare data needed to build the DL and GenAI model
2. **Model Learning (Training)** to define, select, and train the DL and GenAI model
3. **Model Verification (Training)** to ensure the model meets specific functional and performance requirements
4. **Model Deployment (Inference**) to integrate the trained model into the IT infrastructure and run, maintain, and update the model as needed.

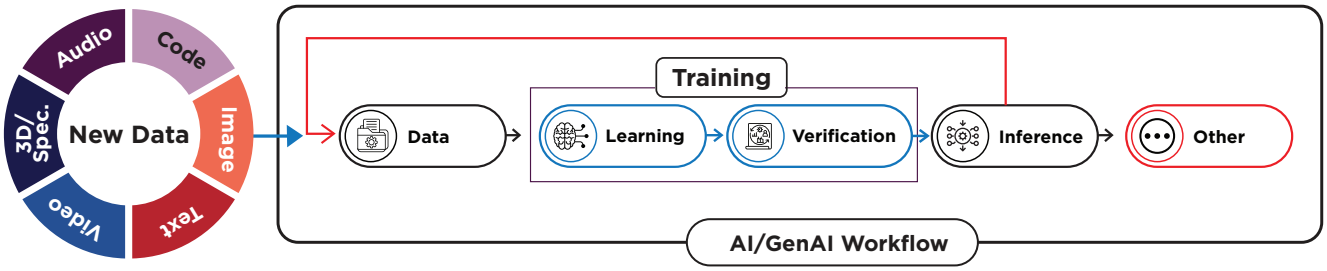


Figure 2: Key Phases in the DL and GenAI Workflow

These stages have smaller steps (Figure 2) that can run in parallel and with feedback. In addition, there are other ethical, legal, trust, and security considerations. All of this makes GenAI implementation very challenging. Figure 3 depicts these Data, Process, Business, Infrastructure, and Other challenges:

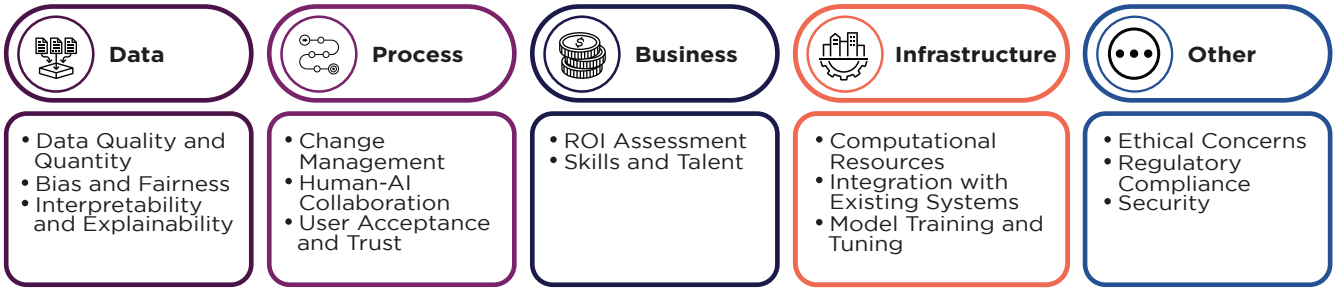


Figure 3: GenAI Implementation Challenges

Data

Data Quality and Quantity

- Data Availability: GenAI models often require vast amounts of high-quality data, which can be challenging to collect and curate.
- Data Diversity: Ensuring the training data represents a wide range of scenarios and demographics can be complex.

Bias and Fairness

- Data Bias: GenAI models can inherit biases in the training data, leading to biased or unfair outputs.
- Fairness: Ensuring fairness in the generated content, especially in sensitive domains like finance and healthcare, is a significant challenge.

Interpretability and Explainability

- Black Box Models: Many GenAI models are like "black boxes," making it difficult to understand their decision-making processes. It can be problematic for applications where transparency is crucial.

Process

Change Management

- Organizational Culture: Implementing GenAI may require significant changes to an organization's culture, processes, and workflows, which can require overcoming organizational resistance.

Human-AI Collaboration

- Training and Monitoring: Enterprises need to establish processes for collaboration between human operators and GenAI systems, including ongoing monitoring and maintenance.

User Acceptance and Trust:

- User Skepticism: Users may be skeptical of AI-generated content, affecting adoption rates.
- Building Trust: Building trust in AI-generated content is crucial for user acceptance.

Business

ROI Assessment

- Measuring Impact: Assessing the return on investment (ROI) of GenAI implementation can be challenging, especially in quantifying the value generated by AI solutions.

Skills and Talent

- Talent Shortage: There may be a shortage of AI experts and data scientists with the necessary skills to implement and maintain GenAI systems effectively.

Infrastructure

Computational Resources

- High-Performance Infrastructure: Training and deploying large-scale GenAI models require substantial computational resources, leading to high infrastructure costs.
- Scalability: Ensuring the infrastructure can scale to handle increased computational demands as GenAI models evolve is a continuous challenge.
- Energy Efficiency: The infrastructure must be energy-efficient. Analysis has shown that training a LLM, a GenAI model with 200 billion parameters, produces approximately 75,000 kg of CO_2 emissions, compared to only 900 kilograms of CO_2 emissions for a flight from New York to San Francisco.[6]

Integration with Existing Systems

- Legacy Systems: Integrating GenAI with existing IT infrastructure and legacy systems can be complex and require substantial effort.

Model Training and Tuning

- Training Time: Training complex GenAI models can be time-consuming, delaying the deployment of AI solutions.
- Hyperparameter Tuning: Fine-tuning models for specific tasks and optimizing their performance can require significant effort.

Other

Ethical Concerns

- Malicious Use: There are concerns about the misuse of GenAI for generating fake content, deepfakes, or other malicious purposes.
- Privacy: Generating highly personalized content can raise privacy concerns, necessitating robust data protection measures.
- Hallucinations: They are model outputs that are either nonsensical or outright false.

Regulatory Compliance

- Data Privacy: Compliance with data privacy regulations, such as GDPR or HIPAA, can be complex when handling user-generated data.
- Content Regulations: Some industries, like pharmaceuticals, banking, and finance, have strict regulations governing the content they produce and share.

Security

- Vulnerabilities: GenAI models can be vulnerable to adversarial attacks, potentially compromising their reliability and security.
- Intellectual Property: GenAI models and the processes used to build them are often an organization's "crown jewels" and must be protected.

Lenovo expects AI to be developed and used consistently with its core values. The Lenovo Responsible AI Committee ensures all solutions and those of AI Innovator partners meet requirements that protect end users and ensure that AI use is fair, ethical, and responsible, focusing on:

- Diversity & Inclusion
- Privacy & Security
- Accountability & Reliability
- Explainability
- Transparency
- Environmental & Social Impact

Lenovo and NVIDIA are working with a large and growing ecosystem of partners and customers to develop best practices and solutions to help enterprises overcome these

implementation challenges. Lenovo has also created a reference architecture[7] for GenAI based on NVIDIA GPUs and software.

High-Level Architecture of Lenovo Solutions Powered by NVIDIA

Lenovo simplifies AI and GenAI implementation with optimized, ready-to-deploy infrastructure (hardware, software, and services), proven expertise, and pre-validated ISV and partner solutions designed for any size or scale. At the foundation of this high-level architecture (Figure 4) is leadership expertly engineered high-performance systems and storage for AI and GenAI ranging from workstations to the edge to the data center to the cloud.

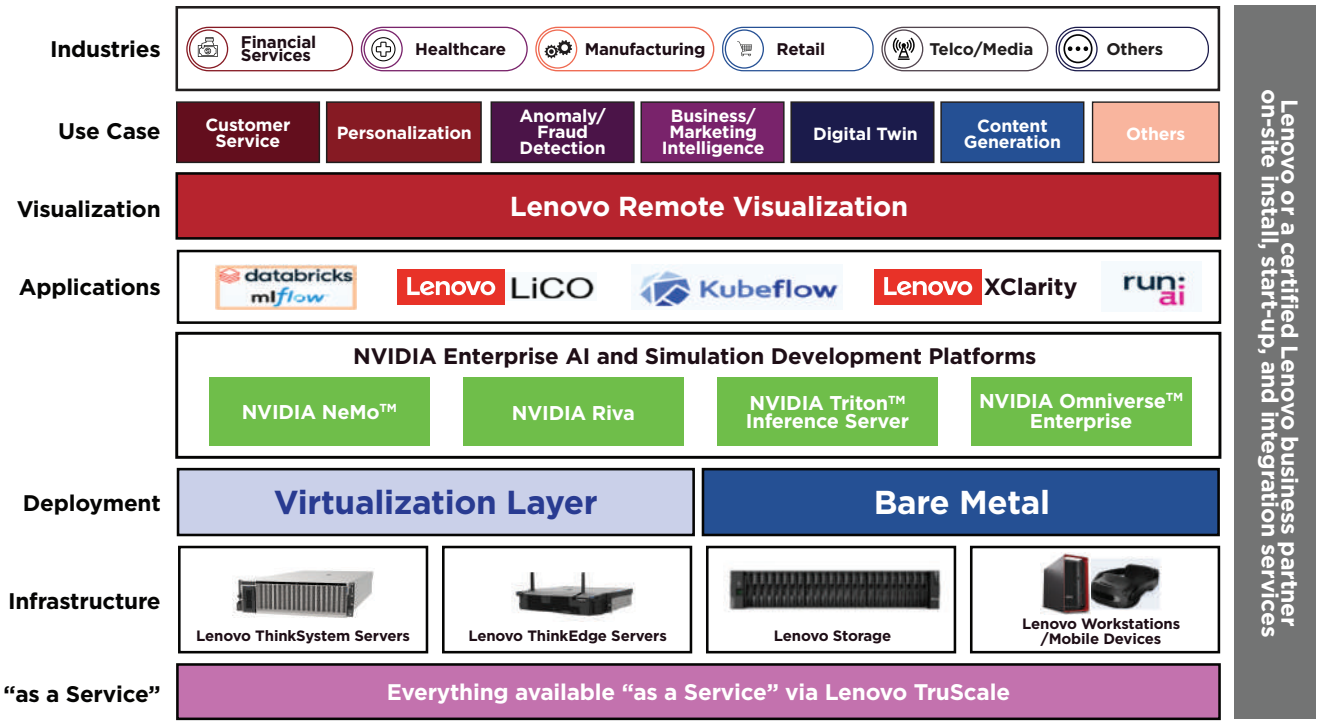


Figure 4: High-Level AI and GenAI Architecture

Some components (not covered earlier) of this high-level architecture, starting at the infrastructure layer, include:

- **Lenovo Performance-optimized ThinkSystem Servers:** Highly reliable, scalable, high-performance servers for significantly accelerating AI and GenAI. This Lenovo portfolio of servers includes the GPU-rich Lenovo ThinkSystem SR675 V3. Leveraging Lenovo's Neptune™ liquid cooling technologies, some systems range from direct water cooling for CPUs and GPUs to liquid-enhanced systems where liquid augments standard air cooling.
- **Lenovo ThinkEdge Servers:** Deliver purpose-built and secure platforms suitable for compute-intensive and latency-sensitive applications like the Lenovo ThinkEdge SE455 V3 deployed outside traditional data centers.
- **Lenovo Storage:** Direct-Attached Storage JBODs and expansion units provide flexible, cost-effective, high-capacity storage and are ideal for space-constrained environments and cost-sensitive customers. Lenovo ThinkSystem DE Series All-Flash Arrays are designed for extreme performance with up to 2.0M IOPS and sub-100 microsecond latency and include industry-leading, enterprise-proven availability and security features.
- **Lenovo Workstations:** ThinkStation P Series workstations with NVIDIA GPUs deliver powerful performance and are ISV-certified, energy-efficient, and highly versatile.
- **Deployment Options:** Provides the autonomy to tailor the deployment approach, choosing between a robust bare metal setup or a versatile virtual deployment.
- **NVIDIA AI Enterprise:** As a full AI software stack, NVIDIA AI Enterprise (Key components include NVIDIA NeMo™, NVIDIA Riva, and NVIDIA Triton™) accelerates AI pipelines and streamlines the development and deployment of production AI for a wide range of use cases from computer vision to GenAI, including LLMs.
- **NVIDIA Omniverse™ Enterprise** is a native OpenUSD software platform for connecting complex 3D pipelines and developing applications for industrial digitalization. Easily unify your 3D tools and data to break down information siloes, minimize tedious data preparation, and supercharge collaboration across enterprise teams. Take advantage of easy-to-use developer tools to build advanced, real time 3D applications that enable you to visualize and simulate your products, assets,

and facilities in full design fidelity. Deploy the platform in your preferred environment, whether on NVIDIA RTX™ professional mobile workstations, NVIDIA-Certified Workstations and Servers, or NVIDIA OVX™.

- **Application:** Major components in this layer include:
  - **Databricks MLflow™** provides a unified platform for managing the machine learning lifecycle, from experiment tracking and model registry to model deployment and monitoring.
  - **Lenovo XClarity** is a family of software that simplifies and automates the deployment and management of Lenovo infrastructure so customers can focus on their high-value projects.
  - **Lenovo Intelligent Computing Orchestration (LiCO)** reduces the complexity of using a massive HPC cluster and simplifies application deployment, operation, and acceleration.
  - **Run:ai** is a scheduler that manages tasks in batches using multiple queues on top of Kubernetes®, allowing system administrators to define different rules, policies, and requirements for each queue based on business priorities.
- **Lenovo Remote Visualization:** Provides reliable and secure access to graphics-intensive applications anytime, anywhere. Instead of issuing new expensive workstations to all design staff, IT can deploy less expensive enterprise or consumer-class personal computers. In addition, IT departments can maintain security and keep costs down by using remote virtualization hosted in an internal data center or from the cloud. Remote visualization performs intensive graphics operations on a high-end graphics server and generates a 2D pixel version that users can receive quickly. In addition, server-side rendering considerably speeds up the process of using graphics in remote sessions.
- **Lenovo or Certified-partner Services:** Lenovo and its global ecosystem of highly specialized AI software and services partners can deliver the entire or parts of the integrated Lenovo stack depicted in Figure 4. They also can provide onsite installation and start-up services to integrate this into a customer's work environment, including installing AI and GenAI applications across various industries.
- **"As a Service":** Subscribe to innovation that scales with you with Lenovo TruScale, which provides end-to-end delivery, management, and refresh services, meaning your IT teams don't have to lift a finger when they deploy new devices and scale their IT infrastructure.

At the core of this high-level architecture are Lenovo servers with NVIDIA software and GPUs that deliver excellent performance for AI and GenAI.

High-Value NVIDIA Software and GPUs for AI and GenAI

High-value NVIDIA software depicted in the high-level architecture include:

- **NVIDIA AI Enterprise** is a high-performance, secure, cloud-native AI software platform with enterprise-grade security, stability, manageability, and support for creating and deploying AI models. It accelerates AI pipelines and streamlines the development and deployment of production AI, covering the range of use cases from computer vision to AI and GenAI. NVIDIA AI Enterprise includes:
  - **NVIDIA NeMo™** (Neural Models) is an end-to-end, cloud-native framework for building, customizing, and deploying AI and GenAI models. It comes with a comprehensive set of tools and resources, including:
    - A library of pre-trained models for a variety of tasks, including text generation, translation, speech recognition, and image generation
    - A set of tools for customizing and training models
    - A cloud-based platform for deploying and managing models at scale
    - NeMo Guardrails helps enterprises keep applications built on LLM aligned with their safety and security requirements.
  - **NVIDIA Riva** is a GPU-accelerated speech and translation AI SDK for building and deploying fully customizable, real-time conversational AI pipelines for:
    - Automatic speech recognition (ASR)
    - Conversational AI digital avatars
    - Interactive voice response (IVR) systems
    - Neural machine translation (NMT)
    - Text-to-speech (TTS)
    - Voice assistants.
  - **NVIDIA Triton™ Inference Server** is open-source software standardizing AI model deployment and execution across every workload. Triton accelerates and optimizes the deployment and execution of AI models across cloud, data center, and edge devices.
- **NVIDIA Omniverse™** is a powerful platform on NVIDIA GPUs that facilitates integrating augmented reality (AR) and virtual reality (VR) technologies in enterprises. It provides a collaborative environment where teams can create, simulate, and visualize virtual worlds and enhance various aspects of their workflows

NVIDIA offers several high-performance GPUs to help customers deploy AI, GenAI, and other workloads.

Here are some affordable GPUs based on the NVIDIA Ada Lovelace architecture that are a good fit for these workloads:

- **NVIDIA L40S** (2U) delivers 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 GenAI.
- **NVIDIA L40** (2U) delivers revolutionary neural graphics, virtualization, compute, and AI capabilities for GPU-accelerated data center workloads.
- **NVIDIA L4** (1U) is a universal, cost-effective, energy-efficient accelerator designed to meet AI needs across video, visual computing, graphics, virtualization, and numerous applications, including cloud gaming, simulation, and data science. It delivers high throughput and low latency in every server, from the edge to the data center to the cloud.

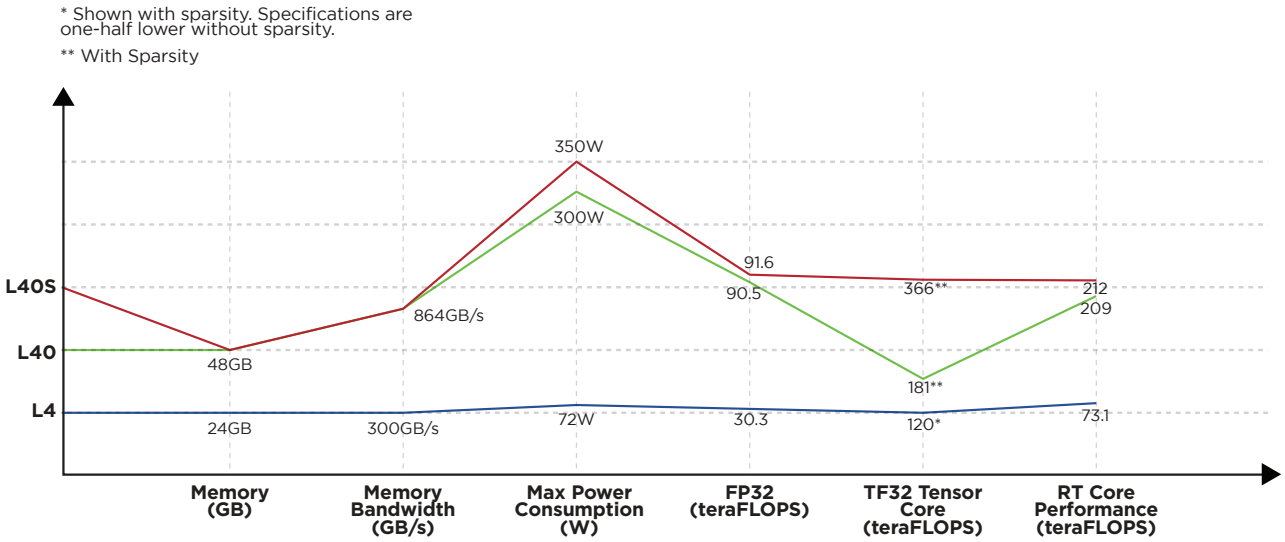


Figure 5: Comparative GPU Features

Figure 5 depicts the key features of these three GPUs. Table 1 shows a suggested map of workload affinity (Best, Better, and Good) by GPU, although it would ultimately depend on specific customer requirements.

The blank cells in Table 1 mean the corresponding GPU is excessive or insufficient for that particular workload. For example, for virtual desktop (VDI) and far-edge acceleration, the L4 is the only suggested GPU. The L40 and the L40S are excessive, more expensive, and take up more slots since both are 2U. For DL Training and high-performance computing (HPC), the L40S is the only suggested GPU because of its significantly better TF32 Tensor Core performance. The L40 is the best for rendering with its excellent RT Core performance and better affordability than the L40S.

NVIDIA GPU Portfolio and Workload Affinity

GPU	DL Training	DL Inference	HPC/AI	Rendering	Virtual Wkstn.	Virtual Desktop (VDI)	AI Video	Far Edge Accln.
L40S	◎	◎	◎	◎	◎		◎	
L40				●	●			
L4		○		○	●	●	●	●

● Best ◎ Better ○ Good

Table 1: NVIDIA GPU Portfolio for Workloads Affinity

Tables 2 and 3 depict the suggested GPUs for AI and GenAI training (only L40S) and inference workloads, including LLMs.

NVIDIA GPU Training Portfolio						
GPU	NLP/LLM				Image/Video Gen AI	Recsys
	Up to 5B	6B to 65B	66B to 175B	>175B		
L40S	◎	◎	●		◎	●

● Best ◎ Better ○ Good

Table 2: NVIDIA GPU Training Portfolio

NVIDIA GPU Inference Portfolio								
GPU	NLP/LLM				Image/Video Gen AI	Recsys	Computer Vision	AI Video
	Up to 5B	6B to 65B	66B to 175B	>175B				
L40S	◎	◎	●		●		◎	◎
L4	●				●		●	●

● Best ◎ Better ○ Good

Table 3: NVIDIA GPU Inference Portfolio

Based on these NVIDIA GPUs and software, Lenovo provides customers across many industries with validated and performance-optimized solutions with the choice and flexibility to customize based on specific use cases, workloads, budgets, and other requirements.

Lenovo Delivers the Optimal Architecture with NVIDIA for AI and GenAI

Figures 6 and 7 show a high-level map of Lenovo ThinkSystem servers with specific NVIDIA GPUs. These systems are designed and engineered from the ground up to meet and exceed the rigorous performance requirements of demanding industry AI and GenAI applications and workflows.

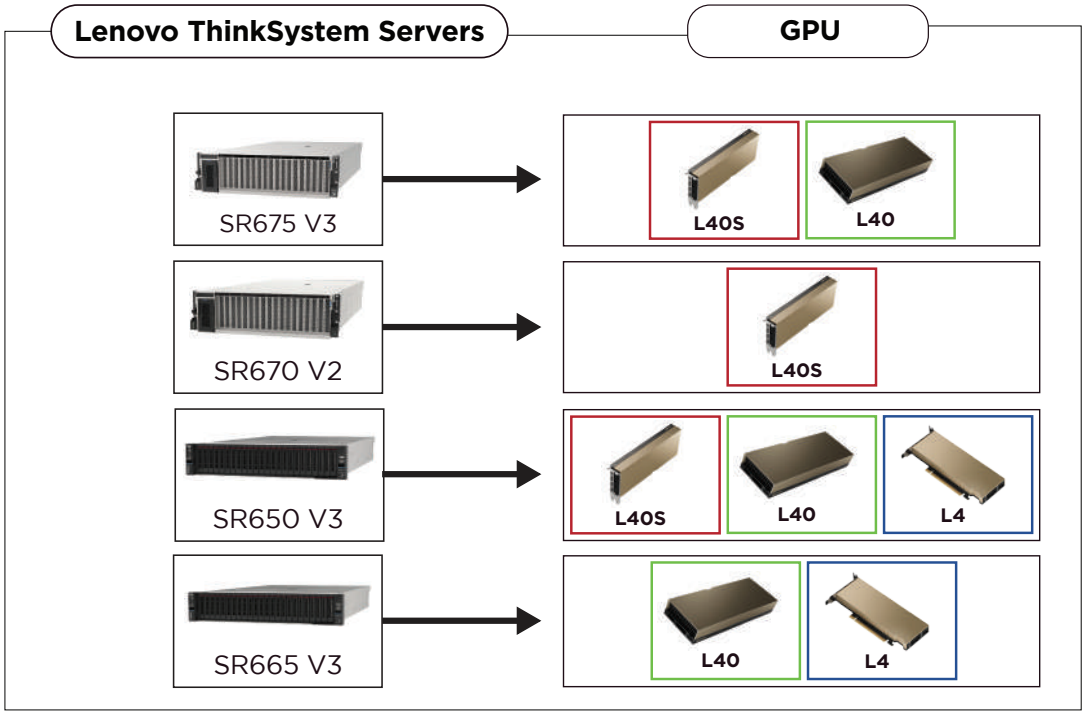


Figure 6: Lenovo ThinkSystem Servers with Relevant Supported GPUs for AI and GenAI

- The Lenovo ThinkSystem SR675 V3 Server and the Lenovo ThinkSystem SR670 V2 Server are versatile GPU-rich 3U rack servers that support eight double-wide GPUs, including the L40S Tensor Core GPUs, with NVLink and Lenovo Neptune hybrid liquid-to-air cooling. These servers deliver optimal performance across many industries for AI, GenAI, HPC, and graphics workloads.
- The Lenovo ThinkSystem SR665 V3 Server offers the ultimate two-socket server performance in a 2U form factor. It is ideal for dense workloads utilizing GPU processing and high-performance NVMe drives.
- The Lenovo ThinkSystem SR650 V3 Server is an ideal 2-socket 2U rack server for industry-leading reliability, management, and security, maximizing performance and flexibility for future growth. It can handle various enterprise workloads, such as databases, virtualization, cloud computing, streaming media, etc.

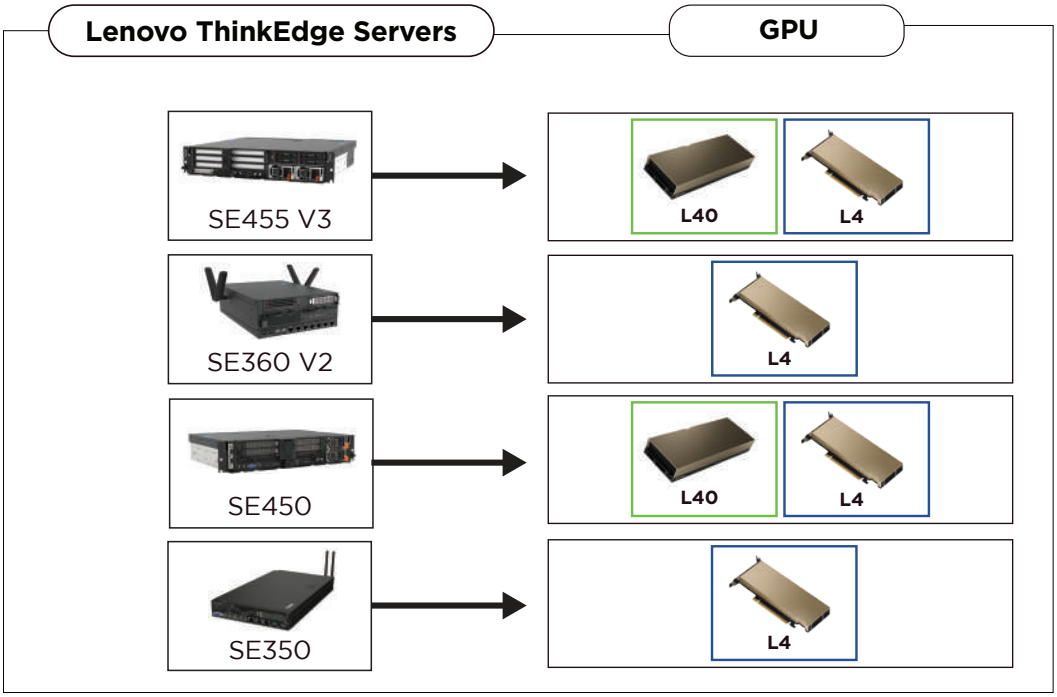


Figure 7: Lenovo ThinkEdge Servers with Relevant Supported GPUs for AI and GenAI

- The ThinkEdge SE455 V3 Edge Server is for AI and Telco-specific solutions and supports the emerging workload consolidation strategies at the edge, with a large core count in a smaller footprint.
- The Lenovo ThinkEdge SE450 Edge Server is a single-socket server with a 2U height and short depth case that can go almost anywhere, operate quietly across a wide range of temperatures, and tolerate dust and vibration.
- The Lenovo ThinkEdge SE360 V2 Edge Server and the Lenovo ThinkEdge SE350 V2 Edge Server are half the width and significantly shorter than a traditional server, ideal for deployment in tight spaces. They provide increased processing power, storage, and network closer to the source of data generation for real-time workloads such as AR/VR, surveillance, AI, etc.

Lenovo also provides additional value and several complementary services and solutions to help customers with their AI and GenAI journey with:

- The Lenovo Reference Architecture for GenAI based on LLMs: Lenovo recently created this reference architecture to help customers in their AI and GenAI journey.
- **Lenovo Innovative Energy Efficient Cooling:** As processor frequencies and the number of cores increase and GPUs become more powerful to deliver the best performance, it is critical to cool these systems efficiently to avoid system overheating issues that cause shutdowns, slower performance and potential data loss. For over a decade, Lenovo has been leading in data center power and cooling technology and has several innovative and unique solutions with Specialized or Liquid to Air (L2A) heatsinks and high-speed fans with low impedance. If air-cooling is not feasible, customers can use other liquid-cooling technologies in the Lenovo Neptune portfolio.

- **AI Discovery and Adoption Acceleration:** Many companies face implementation challenges due to resource limitations and infrastructure complexities, stalling the rollout of AI and GenAI initiatives. The Lenovo AI Innovators program includes an ecosystem of best-in-class software partners collaborating with Lenovo to provide customers with tailored, proven, ready-to-deploy AI and GenAI solutions for their use cases.
- **AI Discover Lab:** Work with Lenovo and NVIDIA's AI experts to get the most value while lowering project risks. Lenovo has been pushing boundaries at the forefront of AI for almost a decade. Benefit from the Lenovo AI Discover Lab, AI assessment workshops, and an AI committee driving AI adoption for customers on every continent.

AI Discover Lab can offer the following services:

  - Access to Data Scientists, Solutions Architects, and GPU Performance Engineers.
  - Help in identifying and delivering AI solutions that meet or exceed the KPIs set out by your business. Will identify and deliver AI solutions that generate ROI, not just AI projects, for the sake of AI projects.
  - Focus on use cases in manufacturing, retail, healthcare, and finance but have also done many projects in various other industries.
  - Help determine AI strategy and adapt to new GenAI technologies.
  - Deliver computer vision deployment use cases as we have done in many projects ranging from NASCAR to Island Conservation to manufacturing defect detection.
  - Deliver GenAI solutions for on-prem deployments that preserve privacy and security, as we have done with many open-source LLMs.

- **Lenovo AI Professional Services Practice:** Offering a breadth of services, solutions, and platforms, the Lenovo AI Professional Services Practice helps businesses of all sizes navigate the AI landscape, find the right solutions, and put AI to work for their organizations quickly, cost-effectively, and at scale. It helps bring AI from concept to reality — from designing AI roadmaps to deploying platforms and providing transparency into technology utilization with the Lenovo TruScale Hub.
- **Innovative Complementary Solutions:** Lenovo is delivering many leading-edge technologies in workstations, laptops, tablets, mobile devices, AR/VR (ThinkReality), and cloud computing (TruScale), which address integration, flexibility, and immersive experience needs of customers across several industries.
- **Edge to the Data Center to a Cloud Platform:** The AI and GenAI computing model is hybrid, with training done in the data center with ThinkSystem servers and inference done at the edge with ThinkEdge servers (Figure 8).

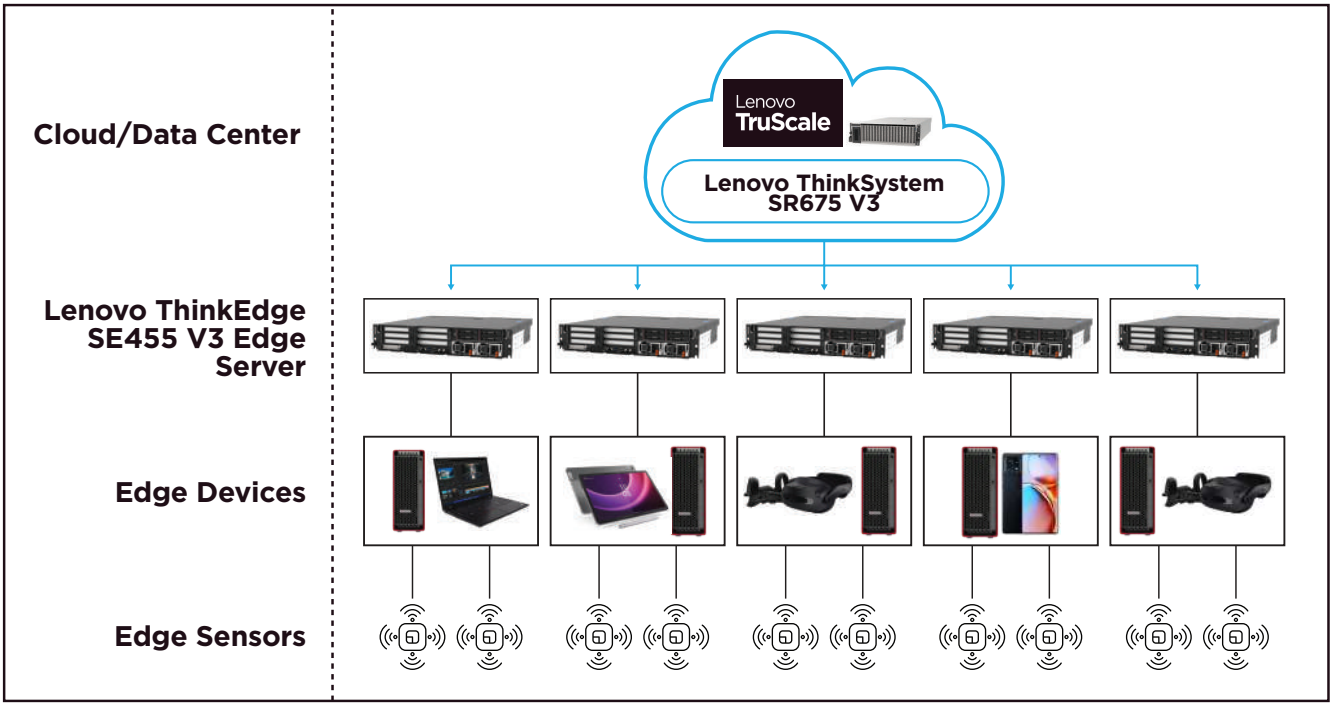


Figure 8: Lenovo Edge to Data Center to Cloud Solutions for AI and GenAI

Industry-Specific Edge to the Data Center Use Cases

Here are some industry-specific examples:

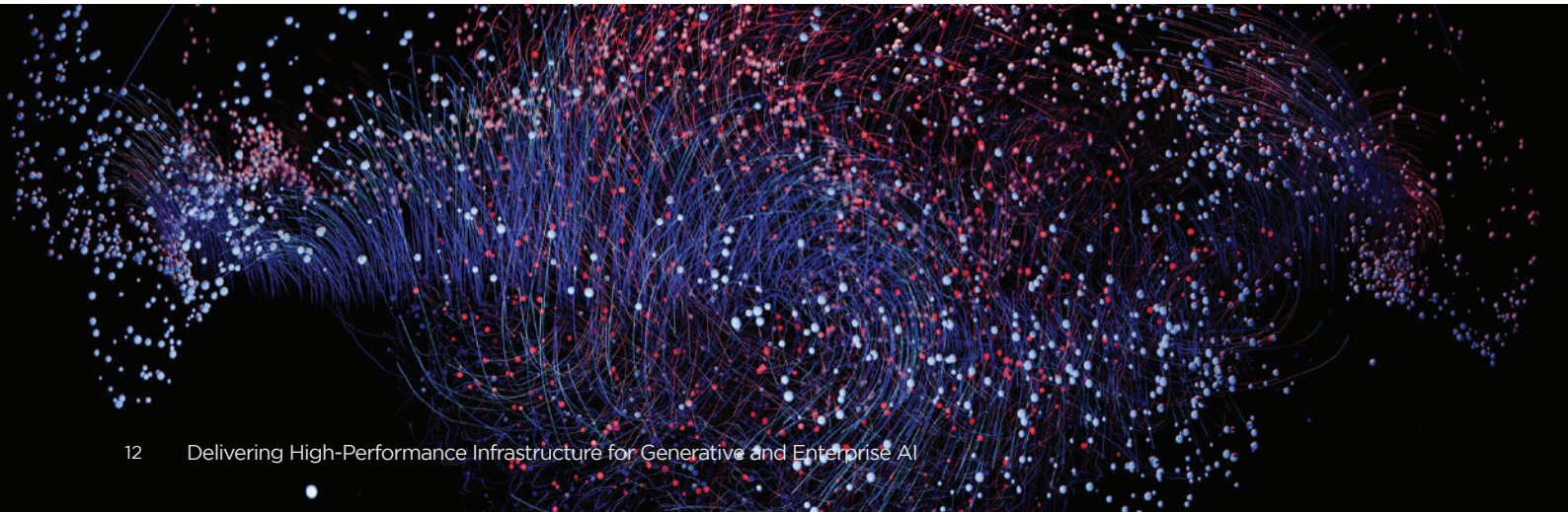
- **Financial Services** firms can use Lenovo ThinkEdge servers for real-time fraud detection. These servers now have the power to run biometrics authentication (inference) using models trained in the data center that are periodically updated with actual and synthetic data to improve accuracy.
- **Healthcare** providers can use Lenovo ThinkEdge servers to monitor patients' vital signs and other real-time health data from their wearable devices. By analyzing this data (inference) using ThinkEdge servers in their offices, providers can identify potential health problems early on and provide patients with personalized health recommendations. Providers can share this data with affiliated large healthcare organizations who can use this anonymized patient data and build better AI training models to make more accurate predictions in the future.
- **Retailers** can use GenAI to create personalized recommendations for customers in their stores. The retailer trains a GenAI model on sales data to learn what products customers will likely buy together. A ThinkEdge server at a specific store can run this model to generate (inference) personalized recommendations when a customer enters the store.
- **Manufacturers** can inspect products with image analysis (inference) on Lenovo ThinkEdge servers for defects on the assembly line to reduce waste and improve product quality, design, and manufacturability. These insights influence new product designs analyzed on high-performance Lenovo ThinkSystem servers in the data center.
- **Telco/Media** companies can personalize the TV experience for their customers. By training an AI model on customer data on a Lenovo ThinkSystem server, a company can learn what types of shows and movies customers will most likely enjoy. This model could then be deployed on edge devices like set-top boxes to generate personalized recommendations.

Get Started with Lenovo and NVIDIA on Your AI and GenAI Journey

As enterprises include AI and GenAI as part of their core business processes, they cannot afford performance problems, delays, or downtime. Therefore, support must be proactive, carried out by technical specialists who work closely with the customer and deeply understand their environment.

As part of their Lenovo contract, companies can receive a dedicated technical account manager or system admin as their single contact point. Whether onsite, working remotely, or a mixture of both, support professionals can quickly pinpoint and resolve any issues and ensure the AI environment runs optimally 24/7.

However, Lenovo goes way beyond specialized technical support. Lenovo's end-to-end service for AI and GenAI includes initial consultation, workshops, analysis, and configuring the right environment through ongoing cooling assessment and monitoring/maintenance services to billing and administration. These comprehensive services can help customers maximize the ROI from their AI investments.



The Lenovo NVIDIA Advantage

As AI, particularly GenAI, becomes an integral part of an enterprise's core business processes, it must overcome several implementation challenges. Lenovo and NVIDIA help companies maximize the ROI from their AI investments, accelerate time to value, and drive innovation and productivity by delivering:

- **Performance-Optimized Systems:** ThinkSystem and ThinkEdge Servers powered by NVIDIA GPUs and software deliver excellent performance for demanding training and inference workloads across text, video, and image data modalities for use cases across several industries.
- **High-Value Services and Software:** The Lenovo AI Innovators program includes a best-in-class ecosystem of software and services partners to provide customers with tailored, proven, ready-to-deploy AI and GenAI solutions for their use cases from initial consultation, workshops, analysis, and configuring the right environment.
- **Leadership in Energy Efficiency:** Lenovo has leadership in data center power and cooling technology and several innovative and unique air- and liquid-cooling solutions, including Neptune™ liquid cooling technologies.
- **Enterprise-level Support:** Systems are tested, validated, and optimized for performance, manageability, security, and scalability. Lenovo, or a certified business partner, provides onsite installation, start-up, integration, and proactive monitoring and remediation of any operational issues.
- **A Complete Portfolio of Solutions:** With Lenovo, customers can implement end-to-end AI solutions using a vast portfolio of intelligent mobile devices, workstations to ThinkEdge servers, and the most scalable ThinkSystem Servers. These systems come with a full range of storage, software, and comprehensive services that provide excellent performance, reliability, and security for a customer's IT environment from the edge to the data center to the cloud.
- **Solid Roadmap with Continuing Innovation:** NVIDIA continues to lead the GPU market by consistently providing a high-performance portfolio of GPUs and software to accelerate the most demanding GenAI inference and training workloads while lowering the TCO. Likewise, Lenovo delivers data center and edge servers that quickly integrate these NVIDIA GPUs with other leading-edge technologies in cloud computing (TruScale) and AR/VR (ThinkReality), which address future performance, affordability, energy efficiency, and immersive experience needs for enterprises and their customers.

Maximize the ROI from Your AI Investment

Please get in touch with your Lenovo representative or email AIDiscover@lenovo.com to schedule an initial consultation with a Lenovo AI Expert or request a customized AI workshop.

[1]Lenovo Grows AI Infrastructure Revenue to Over US$2 Billion and Brings AI to the Data with Industry's Most Comprehensive Portfolio - Lenovo StoryHub
[2]"Artificial Intelligence - Worldwide," Statista
[3]Generative AI software market to exceed $36bn in aggregate revenues by 2028, with 58% CAGR between 2023 and 2028 | S&P Global Market Intelligence (spglobal.com)
[4]Deloitte AI Institute, "Generative AI Dossier: A selection of high-impact use cases across six major industries," 2023.
[5]Challenges in Deploying Machine Learning: A Survey of Case Studies, 2011.09926v2.pdf (arxiv.org), Jan 2021
[6]How do Transformers work? - Hugging Face NLP Course
[7]https://lenovopress.lenovo.com/lp1798-reference-architecture-for-generative-ai-based-on-large-language-models#authors.