



# Boost Efficiencies and Optimize Performance with Turnkey AI Enterprise Solutions

## AUTHOR

**Steven Dickens**

Chief Technology Advisor | The Futurum Group

**Ron Westfall**

Research Director | The Futurum Group

**AUGUST 2024**

IN PARTNERSHIP WITH





## Executive Summary

AI has become a top priority for enterprises and organizations, but its rapid adoption faces multifaceted challenges. Despite the popularity of public generative AI applications, IT organizations struggle with issues such as lack of domain-specific knowledge, limited enterprise readiness, security concerns, and insufficient expertise. These problems are further compounded by inadequate data management solutions, compliance regulations, and high costs related to cloud fees and energy consumption. The fragmented independent software vendor (ISV) market for use-case-driven AI applications adds complexity, potentially slowing progress.

Hybrid AI emerges as a strategic solution to the challenges faced in AI adoption. This approach combines machine learning, deep learning, and neural networks with human expertise to create highly accurate, domain-specific AI models. By using a mix of private and public AI resources, companies can get the benefits of both - the control and customization of private infrastructure alongside the scalability and capabilities of cloud-based AI. This hybrid model allows organizations to optimize their AI strategy based on their specific needs, budget, and technology requirements.

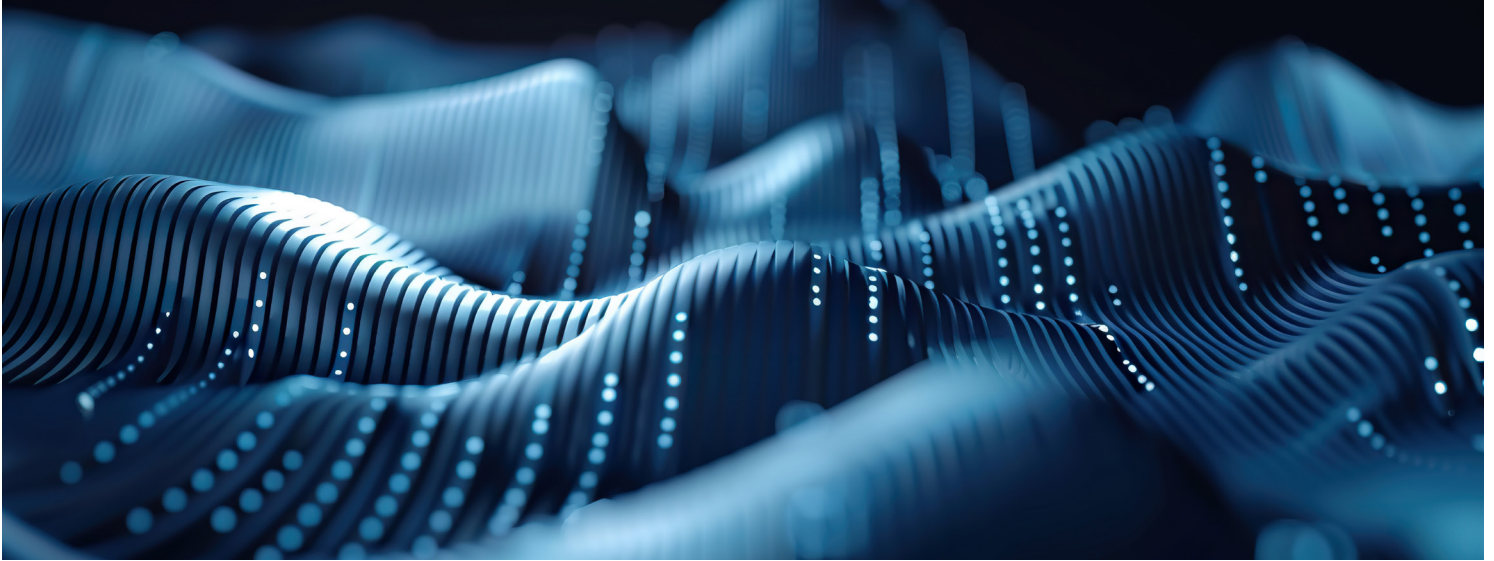
For customers who have concerns about data privacy, security, or controlling their own AI systems, private AI infrastructure rises as the effective solution for them. By hosting AI models and processing on their own secure infrastructure, companies can maintain full control and visibility over how their data and AI systems are being used. This can be especially important for industries with strict data regulations, sensitive intellectual property, or mission-critical applications. Private AI infrastructure also allows for more customization and integration with a company's existing IT systems and workflows.

In this scenario, IT business decision-makers (ITBDMs) such as CIOs, CTOs, and Heads of Architecture find that private AI enables tailoring AI models to their organization's specific needs with greater control over their entire AI stack, including models, applications, infrastructure, and operating frameworks. This enhanced control enables businesses to meet specific requirements for customization, flexibility, privacy, real-time inferencing, compliance, and security. As a result, hybrid AI with private models becomes a powerful tool for organizations seeking to harness AI's potential while addressing the complexities of implementation and data management.

This paper covers how organizations can plan and build private AI solutions and why private AI addresses the growing demand for greater control over AI solutions, including protection and transparency. Lenovo Smarter AI for All strategy prioritizes enterprise AI, harnessing its power to drive intelligent transformation across industries and enable organizations to innovate rapidly and maintain a technological edge over competitors.



Discover why building successful private AI solutions requires an ecosystem of partners, AI expertise, a carefully curated group of ISVs and their AI apps, right-sized infrastructure for targeted workloads and developer support to ensure organization-wide AI adoption and innovation.



## Section 1: The Private AI Landscape

As AI adoption accelerates, protecting sensitive data has become paramount. Private AI techniques, such as differential privacy and encrypted computation, offer a powerful solution, enabling organizations to ensure successful AI model training while maintaining robust security.

As private AI gains momentum, we identify industries that are ready to benefit the most from private AI today:

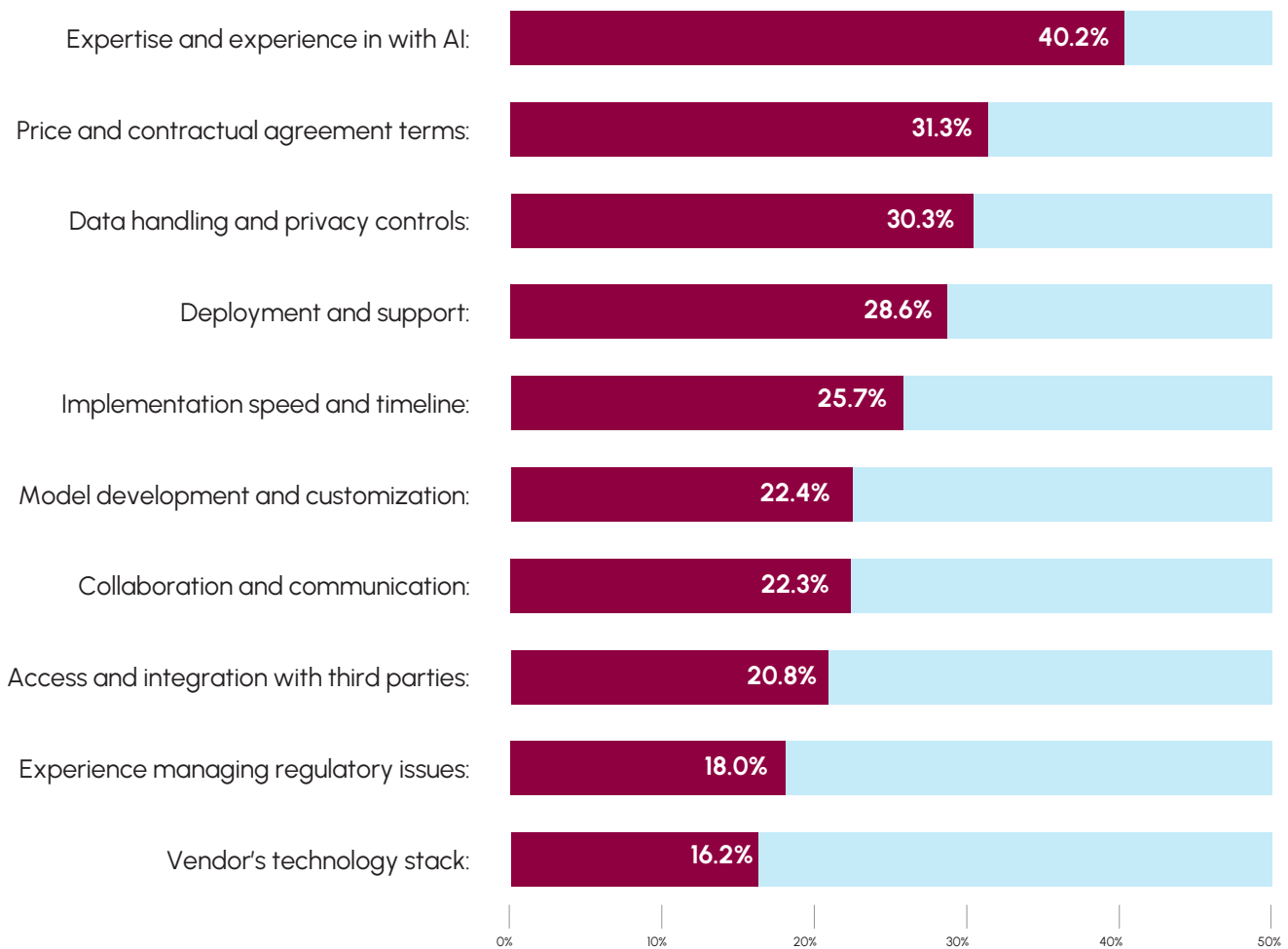
- Highly regulated industries such as financial services, healthcare, government, utilities, and oil & gas
- Distributed industries with heavy edge play, distributed locations such as retail, hospitals, airports, and stadiums
- Organizations that already have significant on prem data and data centers supporting big private environments
- Industries that depend on highly trained/tuned, domain specific AI assistants such as legal, engineering, and medical organizations

In today's dynamic enterprise landscape, businesses (and developers) are adopting a sophisticated approach to AI implementation. They're strategically combining various elements - from small and large models to diverse data sets, and from CPUs to GPUs - to create optimized private AI solutions. This customization allows them to balance crucial factors such as cost, security, compliance, performance, and latency, making private AI a practical reality for their specific needs.

Furthermore, organizations exploring Generative AI (GenAI) are embracing a hybrid model. This approach integrates both public and private GenAI models, leveraging the strengths of each. Public models offer accessibility and off-the-shelf capabilities for text, image, video, and code generation. In contrast, private models, whether self-hosted or cloud-based, provide enhanced customization. These private models are often built upon public foundation models, incorporating tailored applications that offer specialized outputs, stronger data controls, and heightened security measures. This hybrid strategy enables businesses to harness the power of GenAI while maintaining the flexibility to address their unique requirements and constraints.

## Choosing the Right Vendor

In the current AI landscape, organizations are evolving their vendor selection criteria. When choosing AI partners, companies now place equal importance on private AI capabilities and general AI expertise. Specifically, they prioritize vendors who excel in secure data handling and implement robust privacy controls, while also demonstrating extensive experience and technical proficiency in AI development and deployment. This shift reflects a growing awareness that effective AI solutions must not only be innovative but also trustworthy and compliant with data protection standards:



Source: The Futurum Group

Organizations are rapidly embracing private AI solutions, prioritizing vendors who meet their evolving criteria. This shift is driven by key advantages inherent to private AI:

1. Enhanced security measures.
2. Cost-effective long-term AI inferencing.
3. Consistent compliance with regulations.
4. Customization capabilities.
5. Improved governance frameworks.

Private AI is particularly valuable for organizations operating under stringent data privacy laws, such as Europe's GDPR or California's CCPA. By ensuring data is managed securely and in compliance with these regulations, private AI helps businesses navigate complex legal landscapes while leveraging AI's potential.

AI-specific regulations are coming on-board as the EU AI Act establishes a governance structure at both the European and national levels that requires a conformity assessment before deploying high-risk AI systems and puts enforcement mechanisms in place. Additionally, the US, UK, and China are developing similar regulations. Critically, an organization will require a private model if:

- **Data privacy is imperative:** A healthcare provider could use a GenAI model to develop a patient diagnosis system trained on patient data. This would allow the provider to deliver patients more accurate and personalized private diagnoses.
- **Up-to-date transactional information is required:** A retailer could use a GenAI model to develop a dynamic pricing system that is updated with the latest sales data. This would allow the retailer to optimize prices and maximize profits.
- **There is an opportunity to leverage proprietary data and intellectual property:** A financial services provider could use a GenAI model to develop a proprietary investment trading strategy trained on its data and research.

## Customization Capabilities: Private GenAI

Private AI has become crucial in securing businesses' top priority: the successful adoption of GenAI technology. This shift highlights the growing importance of leveraging Private Generative Pretrained Transformer (GPT) models effectively. By doing so, businesses can create their own secure GenAI systems and private AI-driven decision engines. In our view, organizations aiming to build a private cloud for AI should focus on these key considerations:

- Enterprise grade integrated platforms that can accommodate governance, data management, and ML management built typically for Fortune 500 businesses.
- AI and data stacks that are purposely designed to offer choice of AI apps, agents and assistants, expertise and complemented through validated ISV and partner ecosystem.

- Selecting and exploring the blending of open-source and some closed platforms, such as the NVIDIA AI Enterprise platform, alongside other data platform stack owners.
- Prioritize the examination of open-source platforms that can be customized for AI applications.
- Edge needs fulfillment in meeting the data and latency requirements vital to building edge solutions at the edge with ISVs.
- In choosing Retrieval-Augmented Generation (RAG) capabilities, having a well-defined knowledge base in place enhances defining the specific domain or context to improve language model responses.

## The AI Ecosystem: Opportunities and Challenges

The AI landscape is rapidly evolving, with new independent software vendors (ISVs) and AI applications emerging as open-source models become more powerful and affordable. Platforms like Hugging Face Hub, hosting over 350,000 models, 75,000 datasets, and 150,000 demo apps, exemplify this trend. This expanding ecosystem enables businesses to tailor private AI solutions to their specific needs.

However, enterprises face significant challenges in navigating this complex landscape and staying current with new developments. Partnering with trusted experts can help organizations:

1. Validate and integrate diverse ecosystem elements.
2. Access critical AI and data skills.
3. Select optimal AI infrastructure balancing price, performance, and security.

## Responsible AI Development

Ethical AI development hinges on responsible data management and protection. This ensures AI models are trained on representative, unbiased, and accurate data. Businesses prioritize flexibility in choosing private AI solutions that best safeguard and manage their AI workloads.

Key success factors for enterprise-ready AI include:

1. Adhering to Responsible AI principles.
2. Implementing robust MLOps practices.
3. Developing comprehensive data management strategies across all organizational locations.

These elements are crucial for building trust in AI applications and models, making them truly enterprise ready.

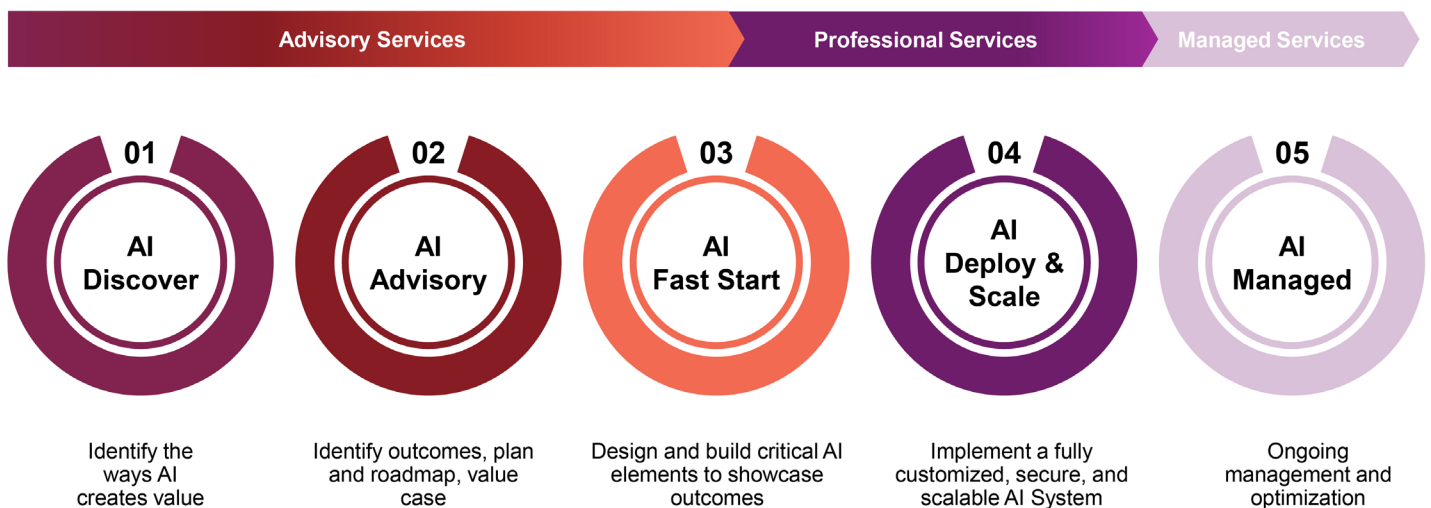
## Section 2: Enterprise AI Solutions from Lenovo

Lenovo Smarter AI for All vision offers a wide array of AI solutions, from personal AI devices to enterprise-level AI infrastructure. Lenovo enterprise AI solutions are designed to be secure, efficient, and scalable, aiding businesses in using AI responsibly and successfully.

We discern that Lenovo offers a comprehensive suite of enterprise AI solutions that prioritize data protection and management. These solutions strike a balance between flexibility, security, and efficiency, enabling organizations to harness AI's potential while safeguarding sensitive information. Lenovo's approach includes on-device and on-premises AI capabilities, ensuring data remains isolated from public networks. This strategy allows businesses to implement AI securely and responsibly, whether within individual devices or across on-site server infrastructure. By focusing on these critical aspects, Lenovo empowers enterprises to leverage advanced AI technologies while maintaining strict data security and regulatory compliance.

Organizations planning to build a private AI must start with understanding business goals and data capabilities. Although this important process is multi-faceted and depending on the organization's capabilities and the focus, the following six steps can lead to project success:

1. Identify the business problem and desired outcome.
2. Source and analyze data.
3. Build the business case.
4. Plan for the operating model, infrastructure, model design, and deployment.
5. Build/ train or tune the model.
6. Deploy, monitor performance, and refine as needed.

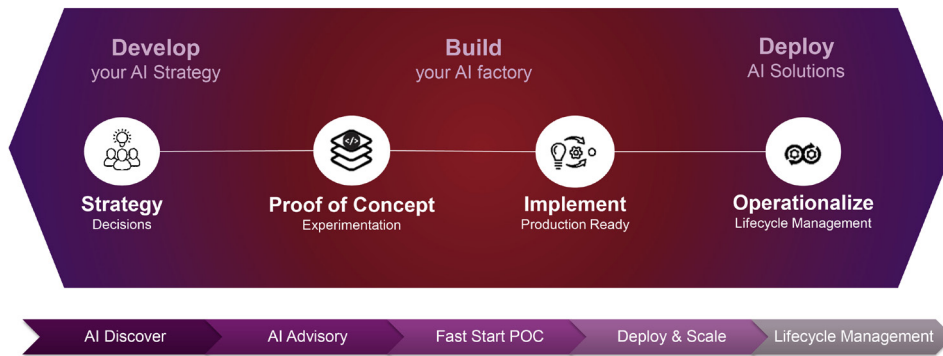


Source: Lenovo

Lenovo teams accelerate the process through the AI CoE and AI Innovators program with partners and ISVs. Key solution areas include:

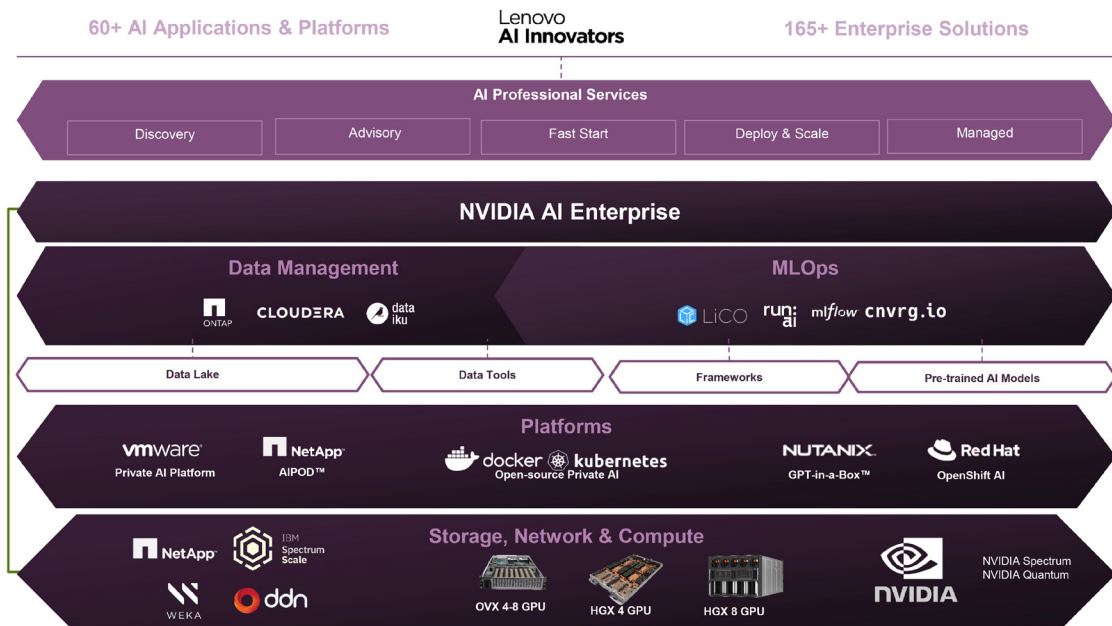
- AI PCs and workstations powering employees, data scientists and developers
- Data management and protection solutions
- AI services and expertise that can accelerate the planning, strategy and execution process
- AI Fast Start solutions that deliver use case and industry pilots with organizations own data
- Energy efficient AI Infrastructure solutions for Edge and datacenter needs offering right size compute (CPU, GPU), storage, liquid cooling innovations and choice

### Start your AI journey anywhere and grow with Lenovo and NVIDIA



Source: Lenovo

Lenovo is building a modular architecture that provides flexibility that can adapt to customers unique needs while creating standardization with common building blocks. This starts from use cases and business needs and ends at the server and storage level and this concept is carried through Lenovo's solution architecture for customers that want to build their AI factory or center of competency.



Source: Lenovo



From our perspective, the Lenovo enterprise AI portfolio and NVIDIA alliance stand out. Using the NVIDIA AI Enterprise software platform, which includes the NVIDIA NeMo framework, Lenovo's newest Reference Design for GenAI based on LLMs shows businesses how to deploy and commercialize GenAI tools and foundation models by leveraging a pre-validated, fully integrated, and performance-optimized solution. With the NVIDIA partnership and the co-development of solutions, Lenovo can aid businesses in improving their LLM workflows as well as maximize performance, security, and right-sizing.

The AI infrastructure landscape is complex, with considerable uncertainty surrounding the components needed for RAG architectures and large language models. Lenovo addresses this challenge by offering modular, scalable infrastructure solutions. These range from standard configurations to high-performance options, allowing customers to make informed decisions based on their specific needs.

Key offerings include:

1. NVIDIA OVX Server with 4 or 8-GPU configurations for inferencing and RAG architectures. Utilizes NVIDIA L40S GPUs, along with NVIDIA Connect-X and BlueField networking technologies upgradable to NVIDIA H100NVL or HGX 8-GPU for enhanced performance.
2. High-performance AI networking with NVIDIA Spectrum Ethernet or NVIDIA Quantum InfiniBand platforms.
3. Flexible storage solutions that accommodate Software-Defined Infrastructure (SDI). Lenovo DM / NetApp options. High-performance storage from Weka or DDN.

This approach enables organizations to tailor their AI infrastructure, balancing performance requirements with budget constraints and future scalability needs.

Lenovo takes a similar approach to network topologies and architectures and up through the software stack ranging from open-source offerings to fully supported enterprise offerings (e.g. NVIDIA AI Enterprise, NVIDIA NIMs, OpenShift AI, etc.) and where applicable the company leverages Lenovo AI Innovator partners to provide the application layer that are use-case determined. Additionally, Lenovo has developed professional services that can assist customers along their adoption journey from discovery to deploy and management.

In the era where data is the new 'oil' and AI the refinery, Lenovo has developed a comprehensive strategy to maximize value for its customers. This approach spans from infrastructure to software tools, ensuring organizations' data is AI-ready, properly governed, and efficiently managed.

Lenovo's data management foundation is built on a diverse storage portfolio, including:

- Traditional storage solutions
- Software-defined storage (SDS)
- High-performance offerings from partners like Weka and DDN

This flexibility extends to data processing, supporting both open-source data lakes and enterprise-grade solutions like Cloudera. For hybrid environments, Lenovo's professional services facilitate seamless integration of on-premises and cloud-based offerings such as Databricks.

Building on this robust data foundation, Lenovo has made significant strides in developing enterprise AI solutions. These offerings are designed to be right-sized, energy-efficient, and reliable, catering to various AI applications including training, modeling, and inferencing.

Key features of Lenovo's AI infrastructure include:

- TruScale "as-a-service" (aaS) options for: Data Management, High-Performance Compute, GPUs
- GPU as a Service (GPUaaS) utilizing ThinkSystem servers (SR680a V3, SR685a, and SR780a V3) that support up to eight high-performance GPUs, Suitable for enterprise AI, high-performance computing (HPC), and graphical workloads

This integrated approach ensures that Lenovo customers can efficiently manage their data and seamlessly deploy powerful AI solutions, all within flexible and scalable infrastructure solutions.

Lenovo integrates AI across its offerings to guide customers through their AI adoption journey. The company has expanded its AI Services Practice with new solutions like AI Discover and AI Fast Start. These can be implemented as a service via Lenovo TruScale, helping businesses deploy and scale AI and GenAI solutions efficiently.

For infrastructure management, Lenovo offers XClarity One, a hybrid-cloud platform with a user-friendly interface for managing servers and edge devices. This solution incorporates AI-Powered Smarter Support, leveraging machine learning for predictive maintenance and automation of routine tasks. Additionally, XClarity One features a Secure Management Hub that uses generative AI to continuously adapt and enhance cybersecurity defenses.





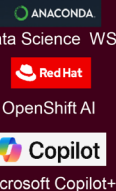
To address environmental concerns, Lenovo introduced LISSA (Lenovo Intelligent Sustainability Solutions Advisor), an AI-powered tool that assists companies in reducing their IT environmental impact. LISSA represents a new approach to IT asset management throughout their lifecycle providing actionable insights that help businesses understand and manage their carbon footprints more effectively.

From our view, Lenovo's 6th Generation Neptune® Liquid Cooling solution is a proven and market-leading choice for meeting unique AI workload requirements. This advanced liquid cooling method significantly outperforms traditional air cooling, particularly in heat dissipation, ensuring improved performance while substantially improving energy efficiency. And Lenovo's power and cooling services are designed to help data centers manage their energy consumption more efficiently by proactively managing power and cooling systems to ensure reliable operations and reducing the risk of downtime.

Lenovo's AI-driven XClarity Energy Manager boosts energy efficiency optimization by identifying low-usage servers, including turnoff of unused components and evaluating how servers can accommodate new workloads based on available resources. In addition, the platform is well suited for harnessing AI predictive failure analytics by leveraging AI-Powered Smarter Support capabilities included in Lenovo XClarity One.

Lenovo Smarter AI for All underpins its enterprise AI proposition driven by a \$1 billion investment over three years, targeted at accelerating AI deployment across multiple industry verticals. This investment supports the creation of over 150 AI solutions in partnership with ISVs. Lenovo offers a diverse suite of over 90 hybrid AI platforms designed to meet a wide range of needs across public, private, and personal sectors.

Moreover, Lenovo AI solutions have amassed numerous awards including the CRN 2024 AI 100 and CRN 2023 Product of the Year.

The right People	<b>400+</b> AI staff globally	<b>4</b> Global AI Centers	<b>182</b> markets served	<b>10+</b> Policy consortiums guiding the future of AI	 <p>Our team members participate in OECD AI, NAIAC, AAAI, AAAS, ACM, MLCommons, Linux Foundation + Many more</p>	<p><b>\$1.2B+</b> AI Investment commitment</p> <p><b>\$2B+</b> In AI Revenue</p> <p>Award winning AI software, hardware &amp; solutions since 2018</p>  
The right Partners & Solutions	<b>60+</b> AI Solution Providers	<b>165+</b> Enterprise AI Solutions			<b>30K</b> Channel & delivery partners	
The right infrastructure & Platforms	<b>80+</b> AI-Ready Platforms	<b>#1</b> in infrastructure Reliability	<b>#1</b> supercomputer provider globally	<b>Top 3</b> AI Infrastructure provider globally	<b>#1</b> PC provider delivering a diverse portfolio of AI PCs	

**The most diverse portfolio from one brand.  
The Leader in sustainable computing with Lenovo Neptune®**

Source: Lenovo

Lenovo's ODM+ strategy further bolsters its enterprise AI offering. This approach entails Lenovo owning its development and manufacturing processes, leveraging its \$20 billion supply chain to decrease costs and deliver end-to-end services. The Lenovo AI Innovators program is an ecosystem of software partners working with Lenovo to provide customers with ready-to-deploy AI solutions. The program focuses on decreasing AI software developers' costs and provides access to global AI labs for testing and benchmarking. The program supports ISVs and their customers in deploying AI on a scalable and secure basis.







## Section 3: Conclusions and Recommendations

We believe that Lenovo Smarter AI for All proposition ensures the implementation of market ready enterprise AI solutions. Lenovo offers the portfolio and services vital to shepherding the collaboration of ecosystem partners, AI expertise, ISVs, essential to organization-wide enterprise AI success.

Lenovo's approach ensures that customers can retain and exercise control over their AI systems, data infrastructure, and deployment environments. Lenovo's approach ensures data security and privacy, regardless of geographical location. With these considerations, we make the following recommendations to organizations and ITBDMs in evaluating Lenovo Smarter AI for All to address their enterprise AI requirements.

**Comprehensive Portfolio Capabilities.** Prioritize consideration of Lenovo Smarter AI for All vision offers a wide array of AI solutions, from personal AI devices to enterprise-level AI infrastructure. Lenovo's enterprise AI solutions are designed to be secure, efficient, and scalable, aiding businesses in using AI responsibly and successfully.

**Accelerate Enterprise AI Benefits.** Lenovo teams accelerate the process through the AI Innovators program with ISVs to deliver the business outcomes with AI applications consisting of data management and protection solutions working with partners, AI services and expertise that can accelerate the planning, strategy and execution process, AI Fast Start solutions that deliver use case and industry pilots with organizations own data, as well as Energy efficient AI Infrastructure solutions offering right size and choice.

**NVIDIA Alliance Advantages.** Explore how using the NVIDIA AI Enterprise software platform, which includes the NVIDIA NeMo framework, alongside Lenovo's newest Reference Design for GenAI based on LLMs can deploy and commercialize GenAI tools and foundation models by leveraging a pre-validated, fully integrated, and performance-optimized solution. With the NVIDIA partnership in co-developing solutions, Lenovo can aid businesses in improving their LLM workflows as well as maximize performance, security, and right-sizing.



# Important Information About this Report

## CONTRIBUTORS

### Steven Dickens

Chief Technology Advisor | The Futurum Group

### Ron Westfall

Research Director | The Futurum Group

## PUBLISHER

### Daniel Newman

CEO | The Futurum Group

## INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

## CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations

## LICENSING

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

## DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.



## ABOUT LENOVO AND NVIDIA

Lenovo brings the new era of AI-powered innovation to everyone. Our full-stack portfolio delivers powerful, flexible, and responsible AI solutions to transform industries and empower individuals. We create a future of Smarter AI for all. At Lenovo, we believe the future of AI involves the co-existence of public and enterprise AI. Lenovo brings AI to you and your data.

In partnership with NVIDIA, hybrid AI solutions are purpose built through engineering collaboration to efficiently bring AI to customer data, where and when users need it the most, advancing Lenovo's vision to enable AI for all and delivering time to market support of breakthrough technologies and architecture for the next generation of generative AI. Lenovo hybrid solutions, already optimized to run NVIDIA AI Enterprise software for secure, supported and stable production AI, also provide developers access to NVIDIA microservices, including NVIDIA NIMs and NeMo Retriever.



## ABOUT THE FUTURUM GROUP

[TheFuturum Group](#) is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.



## CONTACT INFORMATION

The Futurum Group LLC | [futurumgroup.com](http://futurumgroup.com) | (833) 722-5337 |

© 2024 The Futurum Group. All rights reserved.

