PROWESS

Technical Research Study

How AI-Ready Solutions Can Help Your Business Hit the Ground Running in Today's AI Race

Prowess Consulting examined how AI-optimized solutions using Intel® and Lenovo® hardware can help accelerate your business toward an increasingly AI-powered future.

*The analysis in this document was done by Prowess Consulting and commissioned by Intel.

Executive Summary

AI technologies are rapidly evolving and transforming various industries and domains. However, deploying AI successfully can be challenging and costly if you lack a trusted AI partner. In this report, sponsored by Intel, Prowess Consulting examines how Intel® and Lenovo® solutions can help your business keep moving forward into an increasingly AI-powered future by accelerating innovation, modernizing your data center, spanning cloud, edge, and core environments, and facilitating DevOps tasks. The Intel and Lenovo portfolio of AI solutions includes Lenovo servers and clients, Intel processors, the Lenovo alliance of ISV partners, and Lenovo resources and support.

To exemplify how AI can benefit your organization, we present five industry scenarios illustrating how AI can help optimize customer experiences, operational efficiency, company profitability, and risk management. After reading this report, you should have a good idea of how you can proceed with your AI initiatives while minimizing cost and risk.

Exploring Today's AI Landscape

AI is a multifaceted and dynamic collection of technologies that can provide innovative solutions for a wide variety of business challenges. Generative AI tools like ChatGPT® and DALL-E have captured the public's attention, as evidenced by a Forrester Consulting study finding that three-quarters of surveyed organizations used generative AI to solve specific business problems.[1] And work continues in artificial general intelligence (AGI) and other future applications.

It is important to remember that not all AI is generative AI. In fact, there are many AI technologies you can apply today to foster innovation, increase customer engagement, enhance productivity, and optimize efficiency. Among these numerous technologies, predictive machine learning (ML) algorithms and deep learning (DL) neural networks are the reliable AI workhorses many businesses rely on to deliver business value. This diversity of AI technologies indicates that AI is not a uniform, one-size-fits-all solution; rather, it's a collection of distinct technologies suited for different tasks (see Figure 1).

The Business Case for Adopting AI

This report examines how you can effectively deploy a best-practice AI solution for your organization. We suggest you start by selecting solutions from trusted AI partners who rely on open standards and industry standards. These standards are essential for creating an AI platform that gives you plenty of hardware options and flexible code deployment. Whenever possible, use prebuilt AI frameworks and libraries that can enable your hardware to perform as optimally as possible.

In short, we believe the right AI solution should allow you to:

- Accelerate innovation across your organization
- Cost-effectively modernize your data center
- Seamlessly span cloud, core, and edge environments
- Support deployments using the latest technical expertise

Highlights

The right AI solution should deliver the following business value:

Accelerate innovation

Modernize your data center

Span cloud, core, and edge environments

Use the latest technical expertise

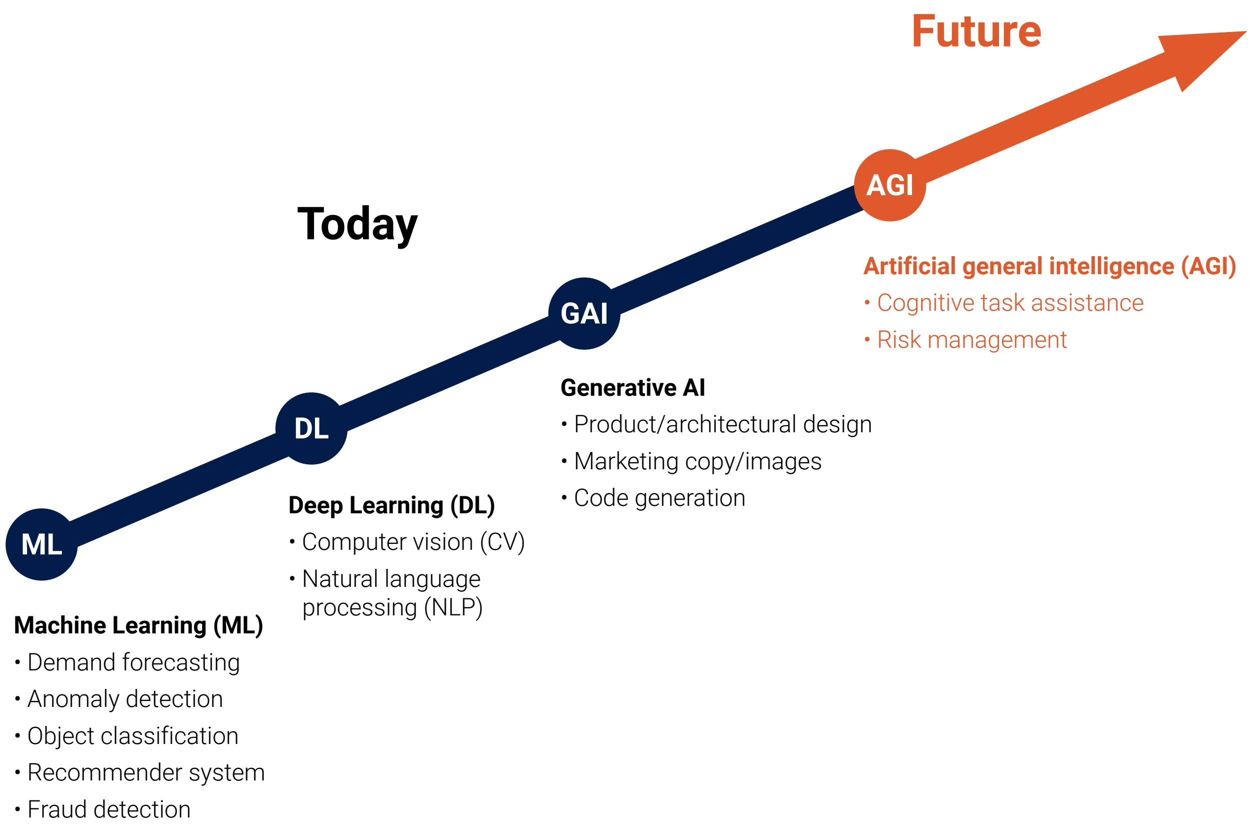

Figure 1 | The ongoing evolution of AI technologies

Paving the Way to AI Success

AI can be difficult to deploy successfully for various reasons. As AI technologies evolve, they require upgraded infrastructures to support their expansive capabilities. That upgraded software and hardware might require more IT expertise to administer. In addition, when AI is implemented and accessed across more of your organization, you might see licensing and administration costs go up. Maintaining global security is another concern as you extend AI into the cloud and out to the edge.

To help ensure a positive outcome when using AI, we strongly recommend choosing a trusted partner who can give you the best value now and in the future. For example, it would be extremely unusual, not to mention time-consuming, to build AI models from scratch. A much more cost-effective approach would be to choose optimized frameworks and pre-trained models that have been error-checked and verified to run on your servers.

We cannot overemphasize the value of an AI partner who lets you build using open and industry standards. Open standards–based infrastructures can support heterogeneous computing, which can help maximize the return on your AI investments. With such an approach, you can mix and match hardware into a combination that best addresses your compute needs without having to completely rewrite code every time you switch hardware brands. And open-/industry-standard infrastructure also lets you efficiently deploy write-once, run-everywhere code using open-source tools such as oneAPI. This is a critical capability for executing your AI solutions across the entire enterprise and helping eliminate application silos.

Planning a new AI strategy offers an excellent opportunity to assess your current security technologies and policies. We recommend including security as an essential criterion when deciding on an AI solution. Make sure your partner offers technologies that help protect your assets and expertise that addresses your needs.

Modeling Your First Steps in AI

To help you find AI's business value for your organization, this section describes deployments built on Lenovo systems, Intel processors, and popular AI apps from ISVs. Table 1 summarizes the benefits of using AI in five industry scenarios. Click an industry scenario to jump directly to the specifics of that AI deployment.

Table 1 | Summary of AI deployments

Industry Scenario	Business Need	Deployment Type	Lenovo® System	Intel® Processors	Lenovo Partner ISV
Using generative AI to be agile and competitive	Meeting customer demands, streamlining operations, and boosting employee productivity.	• Data center closet • Core server	• Lenovo® ThinkSystem™ SR650 V3	• Intel® Xeon® Scalable processors	• Open-source AI frameworks, libraries, and toolkits
Improving restaurant food quality and services	Supporting the latest applications while keeping operating costs low.	• Base station • Edge server	• Lenovo® ThinkEdge SE50 and SE450 • ThinkSystem SE350	• Intel® Core™ i5 processors • Intel Core i7 processors • Intel® Xeon® D processors • Intel Xeon Scalable processors • Intel® Data Center GPU Flex series	• byteLAKE • Hyperconverged infrastructure (HCI)
Optimizing the customer experience in retail spaces	Quickly analyzing massive volumes of real-time and historical data.	• Data center closet • Edge server	• ThinkEdge SE450 • ThinkSystem SR630	• Intel Core processors	• WaitTime server • WaitTime dashboard
Increasing profitability and reducing risk in oil and gas markets	Using real-time data collection and analysis to remotely manage oil and gas wells.	• Base station • Edge server • Cloud gateway • Edge client	• ThinkEdge SE450, SE30, and SE50 • ThinkSystem SE50 and SE350	• Intel Core i5 processors • Intel Xeon D processors • Intel Xeon Scalable processors	• Pathr.ai®
Making the most of limited resources in manufacturing and smart cities	Optimizing operational efficiencies within constrained budgets and resources.	• Base station • Edge server • Containerized apps • Edge client	• ThinkEdge SE450, SE30, and SE50 • ThinkSystem SE350	• Intel Xeon Scalable processors	• Guise AI containers • Intel® Distribution of OpenVINO™ toolkit

Using Generative AI to Stay Ahead of the Competition

To compete successfully in today's rapidly evolving business landscape, companies need innovative solutions to stay competitive and meet the growing demands of their customers. They must tackle ongoing challenges such as enhancing customer experiences, streamlining operations, and boosting worker productivity. Generative AI has rapidly emerged as a transformative approach to these challenges, offering capabilities such as generating fast, accurate search results, providing natural customer interactions, and automating complex, interconnected processes.

Prebuilt AI frameworks and libraries allow organizations to efficiently deploy and manage generative AI foundational models and large language models (LLMs). Many of these tools—such as PyTorch®, Torch, Intel® Extension for PyTorch (IPEX), Deepspeed, and Transformers—are built using open-source code. This lets IT staff develop applications on open-source ecosystems, enabling organizations to benefit from AI's potential without overhauling their existing data center infrastructures.

The Lenovo® ThinkSystem™ SR650 V3 server, powered by 4th Gen Intel® Xeon® Scalable processors, is a data-center core server solution that delivers scalable and cost-effective AI deployments for enterprise organizations. The ThinkSystem SR650 V3 server provides the application performance, data storage, and memory capacity needed to support demanding AI workloads. Intel Xeon Scalable processors use Intel® Advanced Matrix Extensions (Intel® AMX) to deliver peak performance for large AI models. The ThinkSystem SR650 V3 server uses direct water-cooling systems and high-efficiency power supplies, which helps data centers reduce operational costs as they embrace the cutting-edge capabilities of generative AI.

Improving Restaurant Food Quality and Services

Fast-service restaurants are faced with some particularly tough challenges. Their customers want the option to customize their food orders, have their food prepared quickly, and of course, enjoy a high-quality meal. These restaurants could cost-effectively improve food quality and services with a base-station or edge AI deployment that uses a hyperconverged infrastructure (HCI) built on Lenovo servers and Intel processors.

The byteLAKE app for HCI can help streamline food-service operations and enforce quality controls by recognizing meal orders and speeding up the checkout process. Their compact form factor makes the Lenovo® ThinkEdge SE50, ThinkEdge SE450, and ThinkSystem SE350 servers ideal for space-constrained environments. These edge servers are easily set up and managed—a necessity for restaurants that typically do not have on-site IT administrators. The Intel processors within these servers—Intel Xeon Scalable, Intel Xeon D, Intel® Core™ i7, and Intel Core i5 processors—cater to a range of budgetary and compute needs. If you need a more interactive customer experience, we recommend powering your Lenovo server with an Intel Xeon Scalable processor and adding the Intel® Data Center GPU Flex series to enhance model training.

Optimizing the Customer Experience in Retail Spaces

Shared retail spaces are complex and dynamic environments. Customers want an enriching and engaging shopping experience, which means that businesses need to process collected data in near-real time. One way of harnessing AI to accelerate your data analysis is by running WaitTime server and dashboard apps on a Lenovo ThinkEdge or ThinkSystem server powered by Intel Core processors.

The WaitTime software uses real-time and historical data to monitor crowd behavior, and it then intelligently applies the results to help lower customers' wait times and facilitate crowd management. For example, it can help customers move more efficiently through shared spaces by providing them with interactive digital displays and mobile app integrations. The Lenovo ThinkEdge SE450 and ThinkSystem SR630 edge servers are designed for data-center closets and other high-performance edge deployments. Intel Core processors deliver high-performance, low-latency processing for the massive datasets produced by video cameras, guest kiosks, mobile apps, and other edge devices.

Increasing Profitability and Reducing Risk in Oil and Gas Markets

Minor delays can result in major losses for energy companies doing business in the highly volatile oil and gas markets. They need fast data collection and communications with oil and gas wells scattered across multiple locations. Additionally, they need fast and accurate answers so they can make management decisions for increasing productivity, reducing risk, and improving capital efficiencies at these remote locations. In other words, they need AI. One way they can get it is by deploying AI edge infrastructures built on Lenovo servers and Intel processors and using Pathr.ai® software to provide remote monitoring and management, predictive analytics, and operational optimizations.

The Pathr.ai software delivers accurate data analysis in real time so that companies can proactively detect and prevent system failures, find ways to increase oil and gas production, and maintain constant communications with all their remotely located wells. For your most demanding AI workloads, Lenovo ThinkEdge SE450 and ThinkSystem SE350 data-center servers offer high performance and scalability. The Intel Xeon Scalable processors powering these servers use built-in Intel® AI Engines to accelerate inference and training. If you need rugged systems that can deliver high network availability in harsh environments, we recommend Lenovo ThinkEdge SE50 and SE30 edge clients powered by Intel Xeon D or Intel Core processors.

Making the Most of Limited Resources in Manufacturing and Smart Cities

Manufacturers and smart cities are two examples of organizations that might need to operate within limited budgets and resources. To deliver AI to these resource-constrained organizations, we suggest a cloud gateway deployment using a Lenovo ThinkSystem or ThinkEdge platform powered by Intel Xeon Scalable processors and running the Guise AI containerized app.

You can use Lenovo ThinkSystem SE350 or ThinkEdge SE450 edge servers to move the computing infrastructure closer to the data, which can save on data transmission time and costs. For lighter computing tasks, we recommend the Lenovo ThinkEdge SE50 or ThinkEdge SE30 edge clients.

One way that organizations can drive innovation while conserving resources is by using hardware they already own. Intel Xeon Scalable processors are among the most popular CPUs for running inference workloads. If you are running AI on Intel processors, you can optimize workloads running on Guise AI by using an open standards–based toolkit such as the Intel® Distribution of OpenVINO™ toolkit. This performance tuning can deliver improvements in power consumption, waste reduction, and other day-to-day operations across your organization, without having to purchase new processors.

Intel and Lenovo AI Solutions

Figure 2 illustrates the Intel and Lenovo AI solutions portfolio, which includes servers and clients for core, edge, and cloud deployments. The wide selection lets you bring the AI infrastructure to where your data resides, cutting down on data transmission delays and costs. At the high end, ThinkSystem data center and core servers deliver outstanding performance, inherent scalability, and high availability for your most demanding AI workloads. Lenovo ThinkSystem and ThinkEdge servers deliver high-performance AI at the edge for base stations, data center closets, and edge servers. For cloud gateways, containerized applications, and end users, we suggest looking at the Lenovo ThinkEdge and Lenovo® ThinkCentre® clients.

Infrastructure Deployment	Data Center Core Server	Base Station Data Center Closet Edge Server	Cloud Gateway Containerized Apps Edge Client
Lenovo Platform	 **Lenovo® ThinkSystem™** • SR650 V3 • SR630 V3 • SD650-I V3 • SR680a V3 • SD650 V3	 **Lenovo ThinkSystem** • SE350 V2 • SE350 **Lenovo ThinkEdge** • SE450 • SE360 V2	 **Lenovo ThinkEdge** • SE50 • SE30 • SE10-I • SE10 **Lenovo ThinkCentre®** • M90n-1
Intel Processor	• Intel® Xeon® Scalable Processor • Intel® GPU Max Series	• Intel Xeon Scalable Processor • Intel® GPU Flex Series	• Intel® Core™ Ultra Processor • Intel Core Processor

Figure 2 | Intel and Lenovo AI infrastructure solutions

Intel Xeon, Intel Core Ultra, and Intel Core processors use built-in AI accelerators to boost inference and training. Intel Xeon processors are purpose-built to handle the full AI pipeline, from data ingestion to deployment.[2] In other words, you can take a direct path to AI without having to add a discrete AI card. 4th Gen Intel Xeon Scalable processors use built-in Intel AMX to accelerate ML training and inference performance. Intel AMX enables 4th Gen Intel Xeon Scalable processors to improve inference performance by up to 10x, compared to 3rd Gen Intel Xeon Scalable processors.[3]

For dedicated at-scale model training that won't break the bank, we recommend Intel® Data Center GPU Max series, Intel® Data Center GPU Flex series, or Intel® Gaudi® processors. MLPerf® Training 3.1 results show that an ML server powered by Intel Xeon Platinum 8380 processors and Intel Gaudi2 processors was able to train a demanding GPT3 LLM in 153.58 minutes.[4] If a couple of hours is an acceptable wait time for training your LLMs, the Intel Gaudi2 processor -based system offers significant value for your investment.

If you need to deploy the highest-performing ML/DL inference workloads at the edge, we suggest installing Intel Xeon Scalable processors. Intel Core Ultra and Intel Core processors provide multiple options for AI at the edge, depending on your requirements for performance, form factor, or power consumption.

Lenovo AI Innovators

The Lenovo AI Innovators program partners with Intel and ISVs to help ensure your software ecosystem is optimized to deliver the best-possible AI performance from your hardware. The alliance brings together trusted AI partners with the knowledge and experience to fast-track your AI journey, from start to finish and at any stage in between. This close collaboration between hardware and software vendors improves the odds of you enjoying success in your AI journey.

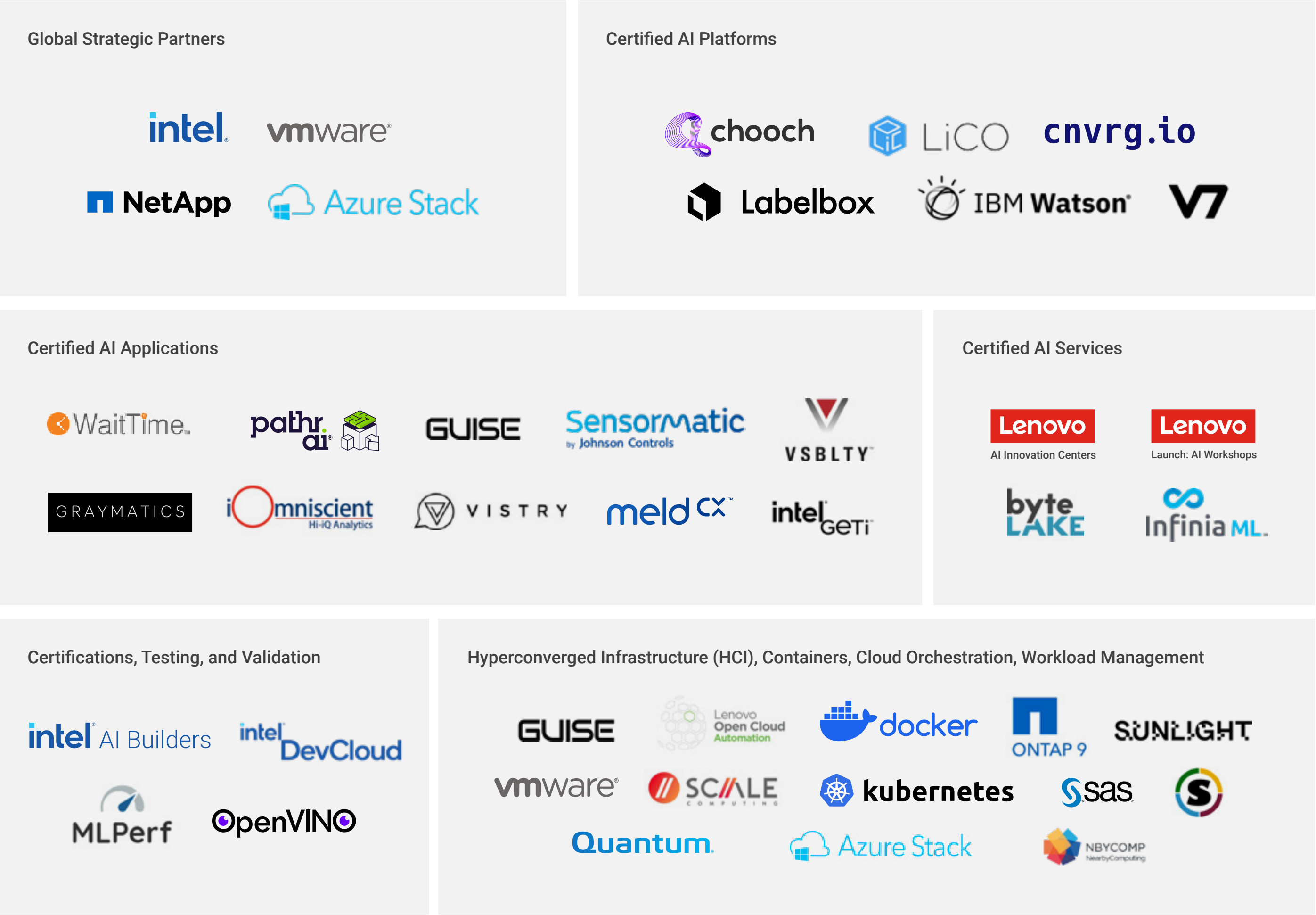

Figure 3 | Lenovo® AI Innovators partner ecosystem

Lenovo Resources and Support

A trusted AI partner offers comprehensive post-sales product and technical support that can help minimize system downtime, empower users, and ensure smooth operations. Lenovo assures that its customers have access to technical experts who are trained to keep up with the evolving capabilities of their Lenovo solutions. Up-to-date expertise is particularly beneficial for maximizing the lifecycle of your infrastructure investments. Lenovo also promises that its tech support covers partner technologies, providing you with a single source of assistance for multiple hardware and software components.[5]

Learn More

- Discover Lenovo AI-ready infrastructure solutions for the data center, edge, and cloud.
- Learn more about the Lenovo AI Innovators alliance of ISV partners.
- Visit Lenovo Support for drivers, updates, how-to guides, and technical help.
- Find customized service plans at Lenovo Premium Care and Premium Care Plus.
- Read about Intel's "AI everywhere" initiative to enable AI on every platform.

[1] Forrester Consulting. "Maximizing Business Potential With Generative AI: The Path to Transformation." Commissioned by Grammarly. July 2023.

[2] Intel. "5th Gen Intel® Xeon® Processors: Workload-Optimized Performance and Power Efficiency Gains."

[3] Intel. See claim [A17] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[4] Comparison based on MLPerf® v3.1 training results. **Intel® Gaudi2 processor–based system:** 96 × Intel® Xeon® Platinum 8380 processors, 384 × Intel Gaudi2 processors, 153.58 minutes. Source: MLCommons. "MLPerf Results: Training." Data as of: 2/12/2024, round v3.1.

[5] Lenovo. "Lenovo Premier Support Services for the Data Center."

PROWESS

The analysis in this document was done by Prowess Consulting and commissioned by Intel.