Getting started with AI

How Lenovo and Intel are powering practical applications for AI today

intel. XEON PLATINUM

Lenovo ThinkSystem SR650 V3 servers built on 5th Gen Intel® Xeon® Scalable processors designed for AI.

Smarter technology for all

Lenovo

Table of Contents

The rapid rise of AI

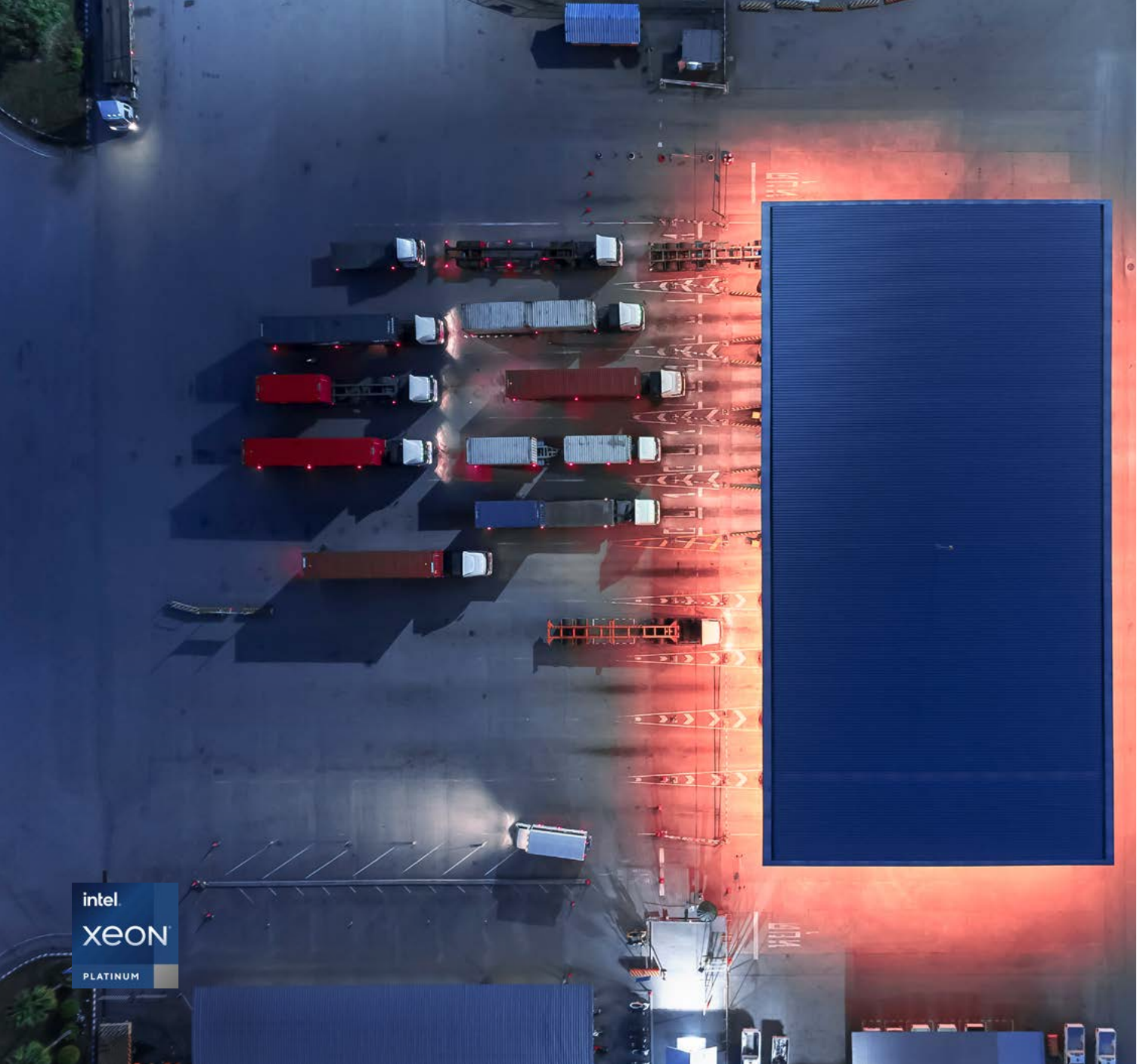
Artificial intelligence (AI) has made tremendous strides since its pioneering days in the 1950s. The predefined static algorithms designed for statistical analysis and prediction running on the earliest computers gave way to early instances of machine learning in the 1980s, when algorithms were taught to recognize relationships and build models of complex systems.

The advent of large neural networks in the 2000s paved the way for massive expansions of computational capability and the introduction of generative AI and large language models capable of working with complex and abstract patterns.

From a business perspective, the potential to derive insights, reduce workloads, and accelerate productivity looks almost limitless — and businesses are actively investigating ways to put AI to work.

80%

of CIOs today are tasked with researching and evaluating possible AI additions to their technology stacks.[1]



Enabling AI everywhere

Tackling AI initiatives can be daunting. Historically, AI has only been in the realm of search engines, financial institutions, and scientific research. Beyond the cost of acquiring the computing hardware itself, in many cases existing data centers can't support the additional power and cooling requirements, which necessitates additional time and capital expenditures.

The good news is that the introduction of broad-based AI models trained on public data has lowered the barriers for organizations to implement advanced AI solutions.

Lenovo and Intel are putting their long-standing partnership to work, delivering solutions that enable businesses to leverage all the great work that's been done so far and apply AI in practical ways that deliver measurable results.

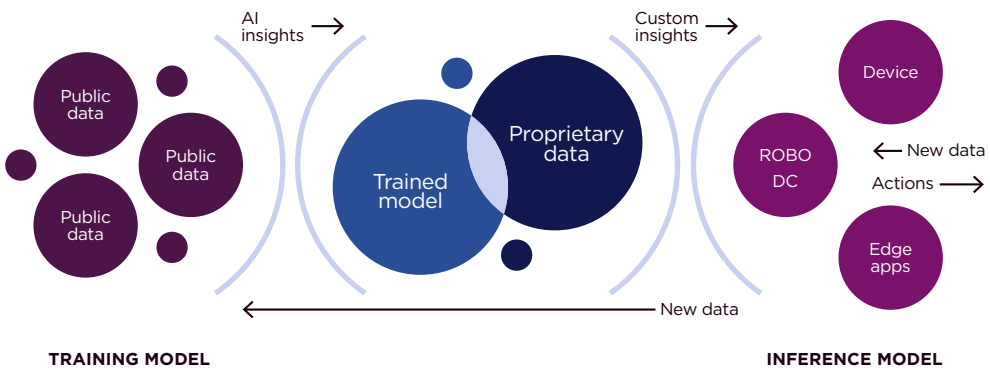
Let's start with some basics

In the simplest terms, AI is broadly defined as any automation system that simulates human intelligence by learning on the job. AI is implemented in two phases:

1. **Training or model development**
This is the process where data scientists develop and optimize foundational models with a curated data set.

2. **Inferencing**
Applying new data to a trained model to derive new insights and accelerate automation.

Systems of Action



Training model development

Training is achieved through a process called machine learning (ML), in which a model is trained based on specific parameters that define the task (for example, color, shapes, and edges) and uses techniques like clustering, regression, and neural networking that process enormous amounts of data to develop predictions. From there, the model continues to consume and analyze data while improving its understanding of those features.

The data sets used for foundational model training have grown to a scale that requires large amounts of specialized computing hardware that relies on thousands of processors running in parallel to deliver the capabilities needed. That's why training foundational AI models has traditionally been the exclusive domain of academic research, financial, and government organizations.

The amount of computing power needed to train the largest AI models is doubling every

3 to 10 months.[2]

Introducing AI inferencing

AI inferencing involves taking existing trained models and applying them to new proprietary data sets for application-specific tasks. The outcomes and insights are then adapted for new applications that are tailored to deliver more precise and relevant experiences.

Inferencing builds off the learning already accomplished, so the processing demands of generating predictions and insights are significantly lower than those required during the initial training.

With the reduced processing requirements, AI inferencing is opening the doors for more businesses and organizations of all sizes to leverage the power of AI for a wide range of applications.

See how Lenovo and Intel are accelerating Industry 4.0 with AI-assisted visual inspections and faster data analytics. **Learn more.**

The number of businesses using AI grew by **300%** in **5 years.**[3]

Since organizations don't need to develop the foundational training models, this dramatically accelerates development and helps them move toward real applications more quickly. Additionally, since the approach needs only new information applied to the model, the data and processing demands can extend beyond the data center. That means inferencing can happen where the data is collected, including at the edge.

This is important because it allows for real-time control of critical functions that won't incur latency penalties going back and forth to the cloud, such as the autonomous systems found in self-driving cars or automated factories.

The trained AI inferencing model works only with the data necessary to make the decisions, which speeds the decision process and reduces the need to move large amounts of data across networks.

For example, in a manufacturing environment, edge servers on the line running AI inference models can use computer vision (based on existing trained models) to identify defects, make decisions, and take the appropriate action (using proprietary local data) to address the defect while maintaining line productivity.

Unlock insights in your data faster

As AI applications evolve, the technology supporting them is evolving to adapt to these new expectations while accelerating and enabling AI deployment at every step from edge to cloud. Lenovo and Intel have teamed up to deliver purpose-built solutions designed specifically for AI inferencing applications.

The latest generation of ThinkSystem servers, like the **ThinkSystem SR650 V3**, are built on 5th Gen Intel® Xeon® Scalable processors designed for AI. The built-in acceleration delivers increased performance for AI inferencing tasks and reduces the power and cooling requirements, which means ThinkSystem SR650 V3 servers can be deployed in existing data centers as opposed to building out new centers.

Up to
2.7x the AI performance
of any other CPU[4] with 5th Gen Intel® Xeon®
Scalable processors featuring Intel® AI Engines.

Up to
14x higher real-time
object detection inference performance vs.
3rd Gen Intel® Xeon® processors.[5]

In addition, Lenovo offers an industry-leading portfolio of edge AI solutions like the **ThinkEdge SE350 V2** and **ThinkEdge SE360 V2** using Intel® Xeon® D processors to deliver real-time insights. The enhanced computer power and flexible deployment capabilities support multiple types of AI workloads with advanced performance and efficient designs. With AI on the edge, organizations can capitalize on dynamic real-time information and deliver greater automation, remediation, and insight where it's most actionable — on the front lines.

Leverage AI transformation expertise

Designing and implementing AI inferencing models that deliver reliable, actionable insights takes a very specific set of skills and extreme attention to detail.

The **Lenovo AI Discover Center of Excellence** brings together Lenovo and Intel AI experts to help your developers create and accelerate the delivery of AI applications and AI inferencing models.

Our AI experts conduct a wide range of workshops to deliver comprehensive business assessments, IT evaluations, and documented design blueprints.

Technical engineers, partners, and data scientists optimize your AI codes using open-source frameworks to run on ThinkSystem servers with Intel hardware and software.

We can help you leverage Intel's comprehensive suite of resources like the OpenVINO™ toolkit and oneAPI Deep Neural Network Library (oneDNN) to simplify deep learning inference deployment for hundreds of pretrained models.

Lenovo also offers a wide range of Lenovo Professional Services workshops to accelerate your AI transformation journey.

intel. XEON PLATINUM

Lenovo

The OpenVINO™ toolkit

The barriers to AI adoption typically include the need for large, optimized, diverse models, a wide range of xPU architectures (often deployed together), and an expansive ecosystem of API software frameworks to choose from. Deploying AI can be a difficult and time-consuming process involving many vendor ecosystem choices.

With all these complexities, proofs of concept often never make it to production, creating a "POC graveyard."

These barriers need to be broken down to create opportunity, and that's what **OpenVINO™** does by offering an open-source toolkit that supports a wide array of xPU architectures and AI software frameworks.

intel. XEON PLATINUM

OpenVINO™

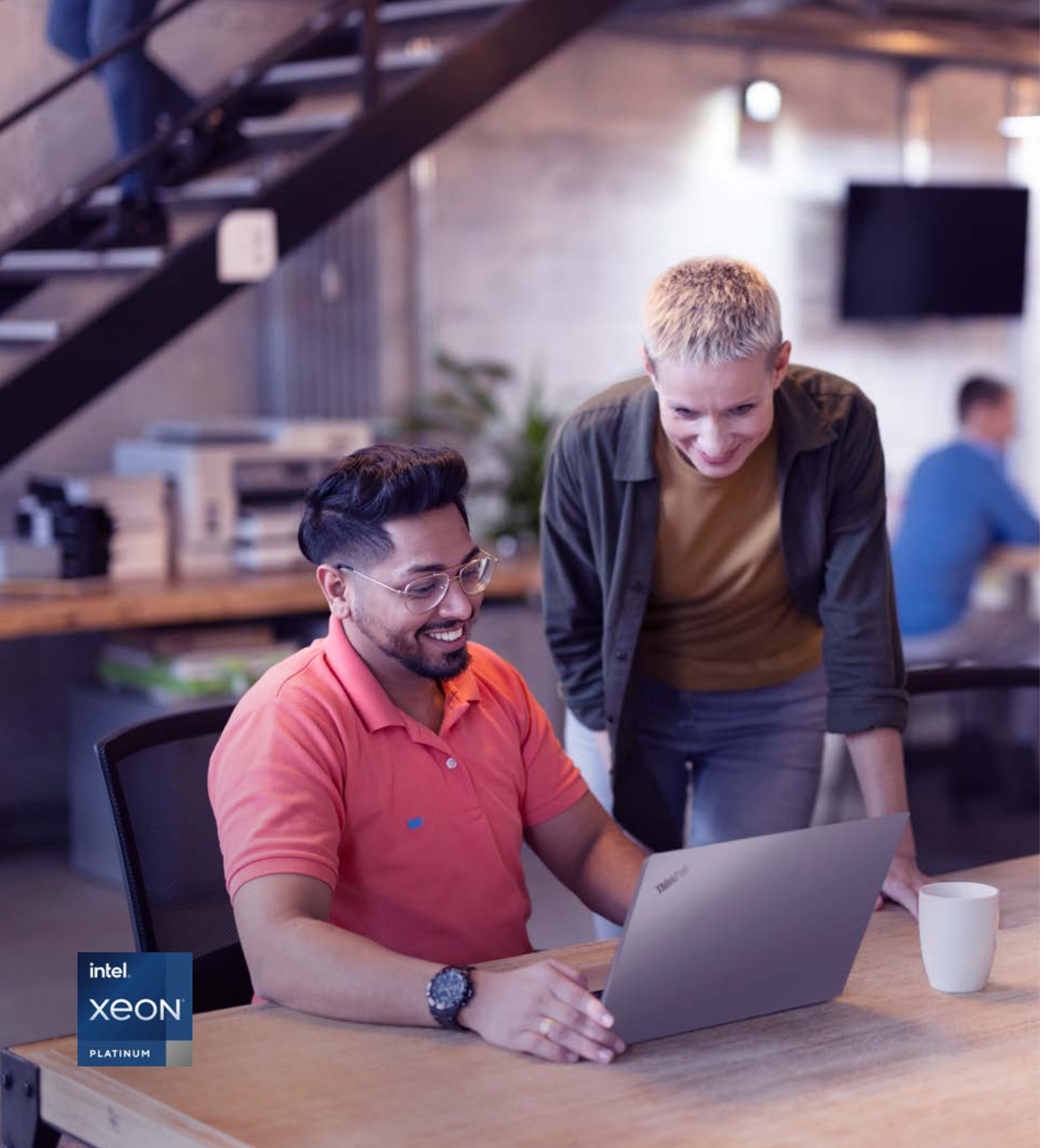
Benefits:

1. Broad accessibility for multiple xPU architectures through an open-source model.
2. An affordable and efficient AI inference solution that reduces the costs of adopting and applying AI technology from the edge to the cloud to local PCs.
3. Open architecture that enables collaboration across the ecosystem — from data scientists creating models for deep-learning frameworks to application developers in a variety of vertical markets leveraging multi-modal AI functions of vision, natural language processing, recommender systems, and generative AI.

Paired with **Intel's Edge Platform**, complete edge-native solutions can be built to accelerate edge AI initiatives with AI model training, optimization, and application development resources.

Companies can also securely onboard and manage a fleet of edge nodes, leveraging the most suitable and cost-effective brownfield or greenfield components in partnership with our unmatched ecosystem for lower total cost of ownership.

See how Lenovo and Intel are streamlining AI adoption with OpenVINO™ **Learn more.**



Accelerate your journey with proven deployment solutions

When the time comes to deploy your AI inferencing solution, Lenovo's AI Innovators program streamlines the process with proven solutions using best-in-class ISV software on Lenovo and Intel's AI-optimized infrastructure.

Lenovo and Intel build, test, and validate AI inferencing solutions with a partner ecosystem of proven AI Innovators to ensure smooth and optimal implementations that keep you on schedule and budget.

- **Nybl's** remote management solution
- **byteLAKE's** AI-assisted visual inspection solution
- **Guise AI's** computer vision, predictive maintenance, and anomaly detection solutions
- **WaitTime's** Queue and Crowd Analytics solution
- **Sunlight.io's** solution that accelerates the digital transformation of restaurants and drive-throughs
- **Smartia's** industrial intelligence solution that connects and transforms data into actionable insights

We continue to monitor, evaluate, and build relationships with ISV partners as their solutions evolve.

Case study:
AI is transforming spectator experiences

Lenovo and **WaitTime** unveiled a groundbreaking venue solution for transforming the Formula 1® spectator experience using cutting-edge technology. By combining 18 cameras strategically installed throughout Circuit of The Americas (COTA) racetrack with WaitTime's patented AI technology on Lenovo ThinkEdge servers powered by Intel® Xeon® Scalable processors, COTA operators can meticulously monitor crowds of people in queues.

"This real-time data analytics platform provides invaluable insights, enabling operators to dynamically understand how crowds are growing, moving, and changing," said Zachary Klima, founder and CEO of WaitTime.

"Such instantaneous information empowers them to make on-the-fly adjustments to operations and revenue strategies, ensuring an optimal and seamless experience for the spectators while maximizing efficiency and revenue for the event."

intel. XEON PLATINUM

You can read more about the solution **here.**

Lenovo

Gain the flexibility to scale seamlessly

Implementing AI inferencing requires far less in initial expenditures compared to building and training foundational models from scratch, but there are still costs to be considered for hardware, software, and services.

Lenovo TruScale offers the flexibility of a scalable pay-as-you-go model for your AI inferencing initiatives — providing you with access to expertise that accelerates your initiatives.

The OpEx model reduces upfront investment and scales with your changing business needs, seamlessly allowing you to take projects from proof of concept through deployment and beyond.

Faster implementation

By replacing CapEx approval requirements and moving to an OpEx model, TruScale can increase flexibility and accelerate procurement and deployment times.

Scalable options

Choose from a fixed contract or metered consumption to match your organization's needs.

Built-in AI expertise and services

Lean into Lenovo's expert services to close skill and resource gaps and help ensure implementation success. In addition, dedicated Lenovo customer success managers can help facilitate and coordinate with Lenovo resources.

This flexibility not only makes it easier for a broader range of organizations to leverage AI inferencing, it also future-proofs the technology and eliminates the risk of obsolescence as technology evolves.

intel. XEON PLATINUM

AI with an eye on sustainability

The increased computational power required for training and operating AI models means more electricity consumed and more heat generated, which continues to be a source of concern across the globe.

And, as AI becomes integrated into more aspects of everyday life, the resulting increase in required computing power will only accelerate.

For example, a typical Google search uses **less than 0.3 watt-hours (Wh)** per request. Adding a large language model interaction to the request raises that power requirement to somewhere between **7Wh and 9Wh** per request. Given their current search volume, if every Google search request included an AI component, Google's AI alone could consume around **30 terawatt-hours (TWh)** per year, or roughly as much as the country of Ireland.[7]

Training a single AI model can produce

626,000

pounds of CO_2 equivalent.[6]

Lenovo and Intel are committed to sustainable, energy-efficient, environmentally responsible solutions for AI inferencing.

5th Gen Intel® Xeon® Scalable processors are Intel's most sustainable data center processors ever, delivering up to 10x higher performance per watt using built-in accelerators on targeted workloads.[8] And they can be deployed into existing data centers with no additional power or cooling requirements.

In the data center, TruScale metering technology can help you monitor power consumption, utilization, and temperature to manage usage and costs more efficiently. In addition, our Energy Aware Runtime (EAR) software and xClarity Energy Manager help deliver optimal performance at a low level of energy consumption by optimizing power states, turning off unused components, and routing workloads to the most efficient resources.

Optimizing your data center with Lenovo TruScale helps reduce CO_2 emissions and power consumption by up to 20%.[9]

Google search <0.3Wh

AI-powered Google search 7–9Wh

All AI Google searches 30TWh per year

A smarter approach to AI inferencing everywhere

AI inferencing holds tremendous promise to accelerate business growth, reduce workloads, and optimize efficiency for businesses across every industry.

No matter where you are on the journey to implementing AI-based solutions in your organization, Lenovo and Intel are ready to help with purpose-built solutions, industry-leading expertise, and best-in-class partners.

Visit the **Intel AI Alliance page** to learn more.

Sources

1 Foundry, "State of the CIO Survey 2024"
2 Accenture, "Technology Vision 2023," March 2023
3 Tidio, "10+ Essential AI Statistics You Need to Know for 2023," October 2023
4 Based on performance gains of 1.19x to 2.69x with Intel® Advanced Matrix Extensions (Intel® AMX) for inference on GPT-J, LLaMA-2 13B, DLRM, DistilBERT, BERT-Large, and ResNet50v1.5 compared to AMD EYPC 9654 and 9754. See A201, A202, A208-A211 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.
5 Up to 1.4x (BF16) and 1.34x (INT8) vs. 4th Gen and up to 14x (BF16) and 6.7x (INT8) vs. 3rd Gen Intel® Xeon® processors. See A20 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.
6 University of Massachusetts, "Energy and Policy Considerations for Deep Learning in NLP," June 2019
7 De Vries, "The growing energy footprint of artificial intelligence," October 2023
8 Based on performance per watt gains of 1.46x to 10.6x with built-in accelerators on a range of AI, database, and networking workloads. See A19-A25, D1, D2, D5, N16 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary. Intel®, "Intel® Performance Index," 2024
9 TruScale IaaS accurately reports on power and CO_2 emissions, allowing managed infrastructures to be designed, implemented, and tuned not only for performance and capacity but also for CO_2 emissions. Ongoing monitoring of the system using Lenovo XClarity Power Monitor and systems performance figures is used to optimize power consumption by the infrastructure. CO_2 emissions are calculated based on the localized carbon footprint of the power source used.

intel. XEON PLATINUM

Lenovo ThinkSystem SR650 V3 servers built on 5th Gen Intel® Xeon® Scalable processors designed for AI.