White Paper

Providing Users the Best Generative AI Experience on Workstations

Sponsored by: Lenovo and NVIDIA

Linn Huang
May 2025

SITUATION OVERVIEW

Historically speaking, technology curves have been less linear and more punctuated equilibrium. Every decade or so, a new technology comes along that disrupts the landscape. Innovation for, investments in, and adoption of this technology rapidly accelerate in a short period of time.

This results in sweeping transformations of workers, businesses, and industries at large. Eventually, the technology matures, and the honeymoon of disruption gives way to prolonged cost cutting and optimizations. While the concept of artificial intelligence (AI) has been around since long before the advent of computers, the growth of generative AI (GenAI) promises to be the inflection point that launches AI into its era of disruption.

Generative AI uses deep neural networks, such as large language models (LLMs), to generate text, code, images, and audio/video from simple prompts. As it improves, scientists hypothesize that we could reach artificial superintelligence, one that surpasses even humans. Consequently, generative AI promises to be a boom to end-user productivity. Today, generative AI tools can help users summarize meetings, draft reports, and create graphics, allowing them to automate simple tasks and ideate faster.

As such, we expect enterprise investments in all AI to soar in the coming decade. IDC projects AI spending to grow at a 29% CAGR from $177 billion in 2023 to $632 billion in 2028. Generative AI spending is projected to grow faster than all other AI technologies at a 70% CAGR from $3 billion in 2023 to $40 billion in 2028.

With so much industry focus on AI technologies, we need to differentiate between activities that fall under consumption of AI and those that fall under the development

of AI. Consuming AI is about using AI technologies and services to accelerate outcomes. Copilot+ is starting to bring more AI-enabled features and use cases to the OS.

Up until now, most AI experiences have existed in the cloud or datacenter. We're talking about closed models with trillions of parameters. The benefit of such a large number of parameters is the ability to answer broad inquiries with robust outcomes. The downsides can come from accrued cloud costs, user latency, increased security risks, and potential loss of privacy.

To help remedy some of these issues, device makers are making their devices increasingly AI enabled. Much attention has turned toward neural processors (NPUs) as a means of AI enablement. However, the NPU is very much a coprocessor, designed not for AI acceleration but for lightweight AI inference tasks at superior power efficiency. Today, NPUs can achieve up to 50TOPS.

Over the better part of the prior decade, power users have already been running AI workloads on dGPUs, the latest and greatest of which was announced with AI performance reaching up to 4,000TOPS. Consequently, GPUs are better suited for data preparation, small model training, and inference. Generative AI is inherently very compute intensive, requiring immense acceleration. This is why the GPU is so valuable in the broader AI stack.

Developing AI is vastly different from consuming it through inference. AI development is the domain of AI and data specialists and scientists. It involves foundational model development and tuning to one's own data before deployment. Typically, this requires significantly more compute than that for AI consumption.

Workstations pack the requisite level of performance that makes them best for AI tuning, inferencing, and consumption. Workstations can be configured with professional GPUs, which offer up to triple the frame buffer compared with consumer cards, and they can also be configured with multiple such GPUs. In terms of CPUs, workstations can be configured with up to server-grade CPUs offering superior clock speed as well as thread and core count. Multi-socketed workstations can additionally take more than one CPU.

Organizations that choose to deploy workstations to their AI developers could enjoy several benefits. Workstations save your AI developers from constantly sending your data off on round-trip voyages to the cloud. This allows your developers to work with more agility in a more controlled sandbox environment. Developing on workstations will also help cut down on associated cloud costs. Deploying workstations can also increase both the scalability of GPUs and overall workforce agility. Finally, they can also help provide on-ramps for enterprise AI initiatives.

FUTURE OUTLOOK

While we have been speaking singularly about AI development, it's important to note that AI development is a multi-disciplinarian process. Data management is how data is organized and governed. Model development involves model training, fine-tuning, and inferencing. AI deployment is about maintenance and updates.

For most organizations, AI development will be a critical capability in the coming decade just as web design and ecommerce became a critical capability for most businesses in the early 2000s. Most AI activity right now is occurring in open source spaces. AI developers in search of solutions can find something that's already been done for free by looking through open source foundational models that closely match their use case.

The open ecosystem of software, libraries, models, and tools can be liberating for AI developers, but it can also pose some challenges. These challenges can include slowness and inefficiencies when deploying and scaling, poor collaboration, painful ecosystem dependencies, security/privacy concerns, and costs of scaling.

Today, organizations can speed agentic AI development and deployment with NVIDIA AI Enterprise software platform and optimize locally using NVIDIA AI Workbench on Lenovo Workstations. From a workstation perspective, reducing large language models (>20 billion parameters) to a 10 billion–parameter model or less speeds up the inference response and decreases power consumption with only a slight decrease in model accuracy.

These small language models (SLMs) can retain much of the original functionality of the parent LLM, but with far less complexity and lower resource needs. This makes SLMs quite powerful in their ability to solve specific issues. And with such an abundance of open source LLMs to tune from, the impact of SLMs on organizations will be as broad reaching as it is profound, if they have the right AI development capabilities.

As part of this study, IDC conducted in-depth interviews with enterprises across various industries and at varying stages of their AI journey to see what they're developing AI for and how that development is going.

One of the world's leading automakers is using AI across multiple operations. It is fine-tuning models with its own internal data for things like RFPs, legal review, document generation, and HR recruitment. Currently, it's testing different LLMs in building advanced customer use cases, something that has become significantly more difficult for the automaker with new EU regulations in place.

This company has identified AI development as the next crucial skill it must build. "We've got a really strong data science capability, internally. However, the GenAI-specific developments are still reasonably new to us and also very new to our stakeholders," states the company's AI product manager. Allowing data scientists to work faster in secured and sandboxed environments is key to upskilling their internal AI development capabilities.

A Fortune 500 retailer is also developing AI for diverse use across the organization. At the store level, the company has already deployed AI for inventory management and merchandising assortment. At the corporate level, it has deployed AI for HR and business analysis. These have been turnkey AI solutions the company has gotten through its partners.

Internally, the company is developing vision models that can analyze foot traffic, manage queues, speed up checkout, and identify theft. The company initially shopped around with partners for such solutions but ultimately decided to go down a different route due to cost and security. "Instead of going to buy it, we can probably build it in-house and also potentially stand this up ourselves," says the company's director of IT.

A company that provides a global digital investment platform for consumers is building out its own models for two reasons: because it holds so much personal consumer data, governance is a key issue, and because it is a digital platform company, cloud costs can run high.

Workstations are helping the company tackle both issues concurrently. As its senior director illustrates, "Because of the kind of data we work with, open source isn't really an option unless there's a way to sandbox it. It would be really difficult to sit and explain to someone 'Hey? Your personal information is now floating out somewhere in the ether.'" As mentioned previously, the ability to fine-tune LLMs into proprietary SLMs will be a crucial skill for all companies in the next decade.

There are several techniques to achieve this objective, one of which is quantization in which the precision of the numbers that represent a model's parameters is reduced from the commonly used 32-bit precision down to 8 bits. Additional techniques are related to how the inference code is run using parallelization and other methods. While a billion-parameter model may not be as accurate as one with 70x the size, it's likely more than enough for most organizations to address specific use cases and issues. At billions of parameters instead of trillions, models can be run locally in a sandbox environment on device as opposed to in the cloud or datacenter.

In five years, organizations will have dozens of AI-enabled apps using their own IP, accessing their own data lakes, behind their own firewalls. Here, we will see the industry shift from models to agents. AI agents refer to code that can make decisions

and act upon them. Agents will be designed for specific tasks but may fail to generalize across different domains.

What separates agents from models is the agent's ability to work autonomously with minimal human intervention to perform tasks and achieve set goals. Agents will become more powerful and pack more feature functionality than domain-specific models and will orchestrate multiple models. Agentic AI will be able to assist in coding, handle customer service, and manage patient care.

Given the transformative nature of AI, organizations are racing to get ahead of the technology. In doing so, they must not forget to keep security and ethics at the core of their AI strategy. AI consumes significant amounts of data in exchange for accelerated action or insight. This exposes the organization to data breaches and leaks. AI will also have human and IP impacts that organizations must responsibly navigate. Every company's strategy regarding AI security and AI responsibility will differ, and finding the right technology partner can ease the journey. Lenovo can serve as your partner along this journey.

Enter Lenovo Workstations with NVIDIA AI Workbench (for more information, go to **www.nvidia.com/en-us/deep-learning-ai/solutions/data-science/workbench/**) for local AI development and NVIDIA AI Enterprise (for more information, go to **www.nvidia.com/en-us/data-center/products/ai-enterprise/**) for secure, scalable deployment. Lenovo has powerful desktop ThinkStations that can be configured with multiple NVIDIA RTX professional GPUs. Lenovo also has sleek and powerful mobile workstations in the ThinkPad P Series. Lenovo's workstation portfolio has over 3 million possible specification configurations, and the company can help you pair the right device with the right developer.

NVIDIA AI Workbench streamlines local AI development by simplifying GPU setup, container management, and version control, making it easy for developers to fine-tune models on Lenovo Workstations. NVIDIA AI Enterprise extends this workflow by providing a secure, optimized software stack for deploying AI models at scale, ensuring stability, performance, and enterprise-grade support.

CHALLENGES/OPPORTUNITIES

Challenges

- Workstations can be more expensive than PCs owing to high-performance components.
- The workstation market is highly consolidated compared with the PC market offering fewer choices of vendor.

- Open source environments can be daunting for those that don't know how to navigate.

Opportunities

- Workstations drive the best on-device AI inferencing and can be purpose built for AI model development.
- Workstations are critical tools in developing AI models and software of tomorrow.
- AI-enabled workforce promises to be a highly productive one in the future.

CONCLUSION

If AI or any other deep learning technologies are critical to your company's strategy over the next decade, you must consider putting workstations into the hands of your developers. Workstations will provide your end users the best AI experience on device. These devices will also help cut down development time as well as cloud costs. And if data security and privacy in the coming era of AI disruption is of concern to you, workstations will provide the safe sandbox environment for your AI development.

MESSAGE FROM THE SPONSORS

Lenovo is a US$57 billion revenue global technology powerhouse, ranked #248 in the Fortune Global 500, and serving millions of customers every day in 180 markets. Focused on a bold vision to deliver Smarter Technology for All, Lenovo has built on its success as the world's largest PC company with a full-stack portfolio of AI-enabled, AI-ready, and AI-optimized devices (PCs, workstations, smartphones, tablets), infrastructure (server, storage, edge, high performance computing and software defined infrastructure), software, solutions, and services. Lenovo's continued investment in world-changing innovation is building a more equitable, trustworthy, and smarter future for everyone, everywhere. Learn more at https://www.lenovo.com.

NVIDIA pioneered accelerated computing to tackle challenges no one else can solve. Our work in AI and digital twins is transforming the world's largest industries and profoundly impacting society. Founded in 1993, NVIDIA is the world leader in accelerated computing. Our invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, revolutionized accelerated computing. Our invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, revolutionized accelerated computing, ignited the era of modern AI, and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing infrastructure company with datacenter-scale offerings that are reshaping industry. Learn more at www.nvidia.com/en-us/about.

ABOUT IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com