White Paper

AI PCs and Workstations: Future Proofing for Tomorrow

Sponsored by: Dell Technologies

Linn Huang Peter Rutten
February 2025

SITUATION OVERVIEW

From an IT perspective, the past decade has centered around digital transformation for most organizations. As consumers increasingly moved online, businesses had to pivot and digitize their operations, offerings, and business models. The next decade of IT investments will be dominated by artificial intelligence (AI).

According to IDC's June 2024 *Future Enterprise Resiliency and Spending Survey, Wave 6* (n = 887 [or appropriate n value]; data weighted by IT spend [500+ employee size]), 93% of enterprises are looking at generative AI technologies in one way or another. That includes 63% that are already investing significantly in AI infrastructure, software, and services.

AI promises to be a game changer, from how the org operates at the broadest levels down to the individual workflows of each user. Software vendors are building exciting new features that leverage the power of local AI to help users iterate faster, collaborate better, and stay more organized. Powering local AI will require newer and more performant hardware. Enter AI PCs and workstations with powerful CPUs, GPUs, and NPUs (neural processing units).

Developing AI on Workstations

Workstations are powerful enough to perform AI model training provided the model is not excessively large. Models trained on workstations can be deployed onto either workstations or servers, with or without discrete GPUs (dGPUs). What makes workstations ideal for developing AI models?

It starts with silicon. Workstations are typically configured with high-grade CPUs, with some running up to server-grade processors. These devices are also typically configured with NVIDIA professional graphics. Mobile workstations can be configured with one GPU, while some tower workstations can be configured with up to four such

GPUs. Workstation OEMs will then spend time optimizing their systems and getting them certified by leading technical and engineering software vendors. ISV (independent software vendor) certification means a company can count on that particular workstation build to run its software optimally and reliably.

Workstations can also be configured with substantial memory, with some going up to 48TB of storage and exceeding 4TB of memory. Workstations also utilize error-correcting code (ECC) memory. ECC memory detects and corrects the most common kinds of internal data corruption in real time. Workstations can also be configured with specialized silicon such as vision processing units. At the optimization level, these devices leverage critical software toolkits such as oneAPI and CUDA.

All of this coalesces into significant benefits for developing AI on workstations. It removes dependence on cloud and datacenter availability. It reduces latency for developers, in turn lessening the pressure to iterate faster. For those who need to collocate a workstation with a data source in the field, mobile workstations provide the necessary power on the go.

Workstations also offer greater security during AI development. Workstations can be used to build, validate, and deploy AI models. They can be used to run complex simulations, optimize algorithms, and perform large-scale experiments. Workstations can act as sandboxes for AI scientists, allowing them to keep sensitive data entirely local and to reduce cloud and server costs.

Workstations not only serve a critical role in developing AI models but also are great at running the models themselves. We're already seeing increasing deployments of AI models on workstations in many different use cases. Workstations are being used to power generative AI technologies used for code generation, digital assistants and twins, and computer vision.

For code generation, companies from defense contractors to financial firms to automotive companies are using generative AI on workstations to automate coding tasks, provide live support, and update legacy systems today. They're also being used to power digital assistants with greater data privacy, controls and customization, and cost savings and lower latency.

Workstations are also used to power digital twins, delivering faster and more accurate simulations for better decision-making. Finally, these workstations are used frequently for computer vision. Security companies use workstations to develop AI models for facial recognition. Social media companies have used computer vision on workstations for content moderation. Manufacturers use workstations to automate quality assurance processes with AI.

Generative AI on PCs

Increasingly, we're seeing AI technologies, experiences, and use cases trickle down from powerful workstations toward PCs. In May 2024, Microsoft announced Copilot+ PCs, a new category of AI-capable PCs that leverage high-TOPS neural processing units, which are specialized processors for AI acceleration, to drive a wholly new user experience. These new experiences include live translations and studio effects for better collaboration as well as Cocreate for rapid ideation, with many more features promised in the coming years.

The AI ecosystem is also booming outside of the OS. Intel is working with over 100 ISVs to create new AI software features that can run on its powerful new Core Ultra NPUs. Before NPUs came along to PCs, AI and machine learning had exclusively been the domain of GPUs. Given the legacy, NVIDIA is working with over 600 software vendors and game developers on AI features today.

So what differentiates the NPU from the GPU? NPU architecture was specifically created to run AI locally and efficiently. Today's top NPUs for PCs can power models with billions of parameters. The newest Intel Core Ultra CPUs will feature NPUs that go up to 48 TOPS (trillions of operations per second). NPUs are optimized to run specific AI operations highly efficiently, trading lower precision for lower power consumption.

GPUs, on the other hand, can exceed 1,000+ TOPS at the high end of the market and can execute a wider range of tasks concurrently. Simply put, today, nothing on the market beats a beefy discrete GPU when it comes to raw number crunching. However, there is a stark trade-off in power consumption utilizing a discrete GPU at full blast, which can tank battery life for notebooks and spike energy costs for desktops. Unlike embedded NPUs, dGPUs can be scaled up to meet rising AI demands and requirements.

Consequently, we see the optimal AI PC of the future as having both powerful NPUs and discrete GPUs. OS features and other persistent agents (such as security agents) that run pervasively would do so better on an NPU. Queries such as data analysis and modeling, as well as content creation, could be done faster or with wider parameters on a GPU. Over the next two years, expect a deluge of new and exciting software features and use cases that will take advantage of AI-enabled hardware.

Today, AI PCs are already doing the following for businesses everywhere. CrowdStrike can now offload multipoint attack workloads to the NPU, providing faster early threat detection for fileless malware on endpoints. For content creation, Adobe continues to introduce new AI features to Lightroom and Premiere Pro, while CyberLink has brought more AI capabilities to PhotoDirector and PowerDirector. Luminar has introduced new

AI features into Neo, while DaVinci has done so for Resolve for faster photo and video editing, respectively.

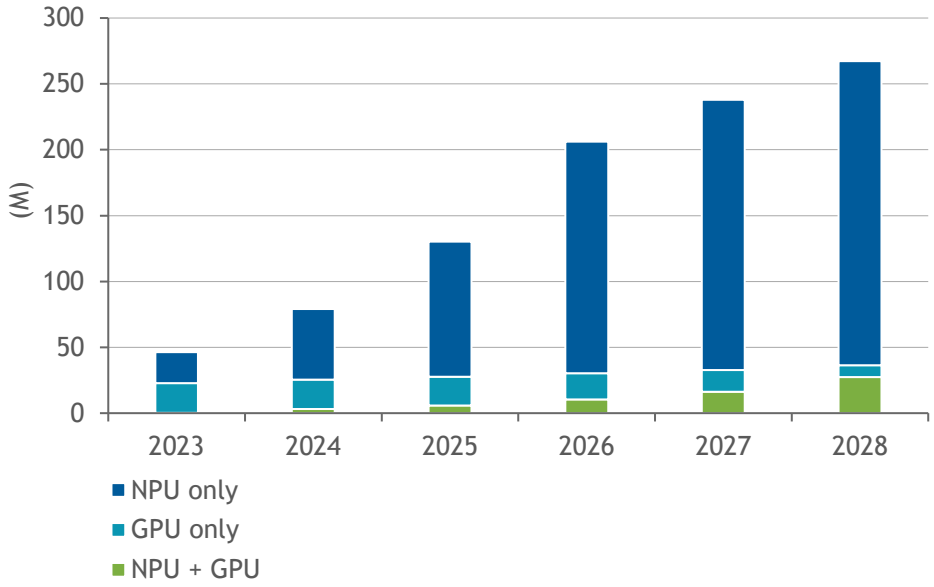
AI is not only changing operations and workflows; it's also changing lives. Medivis SurgicalAR leverages computer vision to provide hands-free 3D medical imaging overlays during procedures and runs on Dell hardware powered by NVIDIA professional GPUs. Cephable uses AI for head and gaze tracking as well as facial mapping to provide more robust adaptive controls to the differently abled.

FUTURE OUTLOOK

AI PCs are projected to grow with a CAGR of 42.1% from 2023 to 2028, from 46.2 million to 267.4 million. Those with high-TOPS NPUs are projected to grow to 60% of the PC market in 2028. Those with discrete GPUs will grow to 13% of the PC market in 2028. Figure 1 provides our forecast of AI-enabled PCs.

FIGURE 1

Worldwide AI-Enabled PC Shipments, 2023–2028

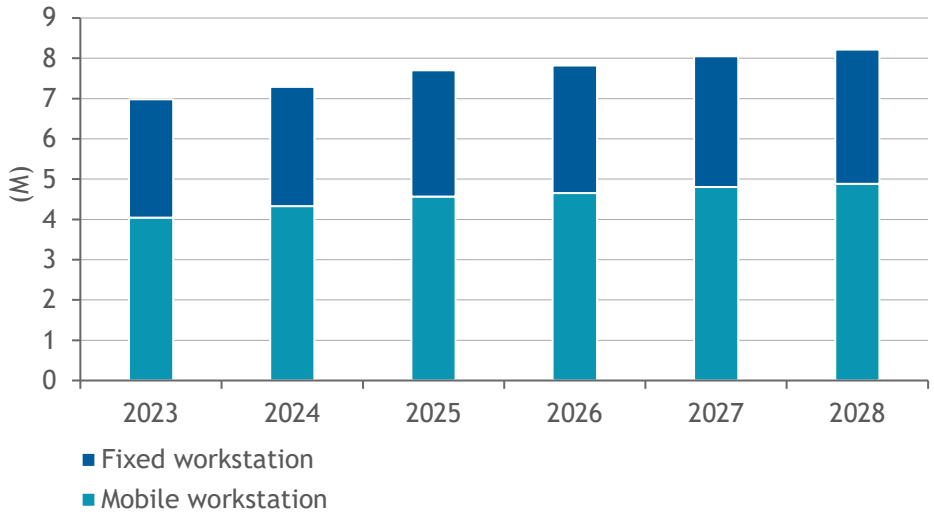


Source: IDC, September 2024

Workstations are projected to grow 3.3% from 2023 to 2028. Figure 2 provides our total workstation forecast.

FIGURE 2

Worldwide Workstation Shipments, 2023-2028



Source: IDC, September 2024

Considering Dell for AI Hardware

Dell is a leading enterprise IT vendor. Since 2022, Dell has led the global commercial PC market in value. Its workstation portfolio has ranked first in year-end market share every year since 2017. In monitors, it has been number 1 since 2014. Dell also boasts robust life-cycle services and software capabilities.

Dell operates the Dell AI Factory with NVIDIA. The Dell AI Factory combines both companies' technologies to provide simple infrastructure solutions and services to quickly scale AI.

CHALLENGES/OPPORTUNITIES

Challenges

- AI PCs and workstations can carry high premiums.
- In terms of the overall AI experience, software is still playing catch-up to the enabling hardware.
- Use cases have yet to be fully elucidated at the moment.

Opportunities

- AI can drive significant end-user productivity gains.

- AI PCs and workstations help future proof your business or organization so you are prepared as more AI technology comes online.
- AI PCs transform workloads for technical users.
- AI PCs augment cybersecurity at the endpoint level.

CONCLUSION

For many companies, AI is the future, and that future has already started today. When investing in AI, IT managers should take care not to overlook endpoints while focusing on infrastructure. AI PCs and workstations will provide wholly new user experiences. And as software catches up, AI will transform how work is done. With future productivity gains on the line, consider AI PCs and workstations today.

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