DELL Technologies

In association with Deloitte

DELL AI CLIENT SOLUTIONS

The Transformative Value of AI PCs

What is driving the value of AI PCs?

Generative AI, a field dominated by cloud and data center computing up to this point, is now seeing evidence of an inflection point in how enterprises and consumers will interact with AI. Rapid advancements in chip technology are being reflected in the latest PC offerings, and complex models such as Large Language Models (LLMs) are becoming increasingly efficient and performant under restricted hardware conditions. These trends are driving the emergence of AI-enabled devices capable of running AI workloads locally, or “on device”.

Behind these rapid advancements is the release of new AI chipsets. Dell’s lineup of AI PCs[1] includes the latest Latitude[2], XPS, Alienware, Inspiron, and Precision models powered by the new Intel® Core™ Ultra Series and select devices with AMD Ryzen™ 7000 and 8000 Series. Further, in June Dell launched the first Copilot+ PCs[3]: the XPS 13, Inspiron 14 Plus, Inspiron 14, Latitude 7455 and Latitude 5455 powered by Qualcomm’s Snapdragon® X Series processors.[4] These Copilot+ PCs are AI PCs with even more AI capability, thanks to the powerful Neural Processing Unit (NPU) integrated across the entire Snapdragon® portfolio.

Simultaneously, advancements in LLM optimization, such as quantization, compression, and pruning, have resulted in smaller, more efficient models suitable for running on device. As both compute capability across devices and model performance relative to footprint continue to increase, the number of workloads capable of running on device will only continue to grow. In fact, interest is already growing around Small Language Models (SLMs) designed specifically for local, on-device inference.

What is an AI PC?

The AI PC embodies a significant departure from traditional PC architecture, introducing a fundamental transformation in its design and operation. While today’s PCs utilize a central processing unit (CPU) and a graphics processing unit (GPU), the AI PC introduces a new type of processing unit dedicated to AI: the neural processing unit (NPU). This NPU handles AI tasks independently, minimizing impact to the CPU and GPU. On Copilot+ PCs, these NPUs are capable of 40 trillion operations per second (TOPS) at minimum.

These NPUs, combined with next-gen CPUs and GPUs, enable energy-efficient AI inferencing alongside all other computer functions. These modern chips can even support the intensive computational demands of LLMs, making on-device processing a practical, efficient, and powerful alternative to server-hosted models. Internal testing at Deloitte on the Dell XPS 13 AI PC with a Snapdragon® X Elite processor generated 80 lines of code in 1 minute using a Llama 2 model (7 billion parameters) on NPU after INT4 quantization, well above traditional PC performance.

The AI PC represents Dell’s latest expansion in its comprehensive AI portfolio, integrating solutions from desktop to data center to cloud in a cohesive offering. This addition broadens the spectrum of options for executing AI tasks, enabling enterprises to strategically align them with the most suitable environment. Whether using public cloud, private cloud, on-premises data centers, or the AI PC, organizations can now tailor their deployment strategies to optimize performance and meet specific operational requirements.

1. AI PCs Laptop Computers | Dell USA
2. Dell Blog, Dell Technologies Annouces New Latitude AI PCs, February 2024
3. Dell Press Release Details, Dell Introduces Comprehensive Portfolio of Copilot+ AI PCs, May 2024
4. Copilot+ PCs | Dell USA

WHAT ARE THE BENEFITS OF AN AI PC?

By harnessing the capabilities of AI PCs, enterprises can future-proof their organizations by unlocking productivity and empowering their teams to accomplish more.

AI PCs have four key advantages. First, NPUs are designed for AI workloads, enabling responsive, personalized intelligence in everyday applications. Second, AI PCs offer reduced security risks by replacing cloud services with on-device data processing, so your data stays on your machine. Additionally, security software can use AI to fortify defenses with advanced encryption and real-time threat detection. Third, AI PCs offer energy efficiency. Because the NPU is specialized for neural arithmetic, AI workloads consume less power during inferencing, generate less heat, and extend battery life compared to traditional CPUs or GPUs. Fourth, AI PCs offer cost effectiveness. When an NPU is available, applications can use local processing over cloud processing, reducing the expenses incurred by cloud computing costs. AI PCs combine the power of a CPU, GPU, and NPU to handle the right workload with the right engine at the right time.

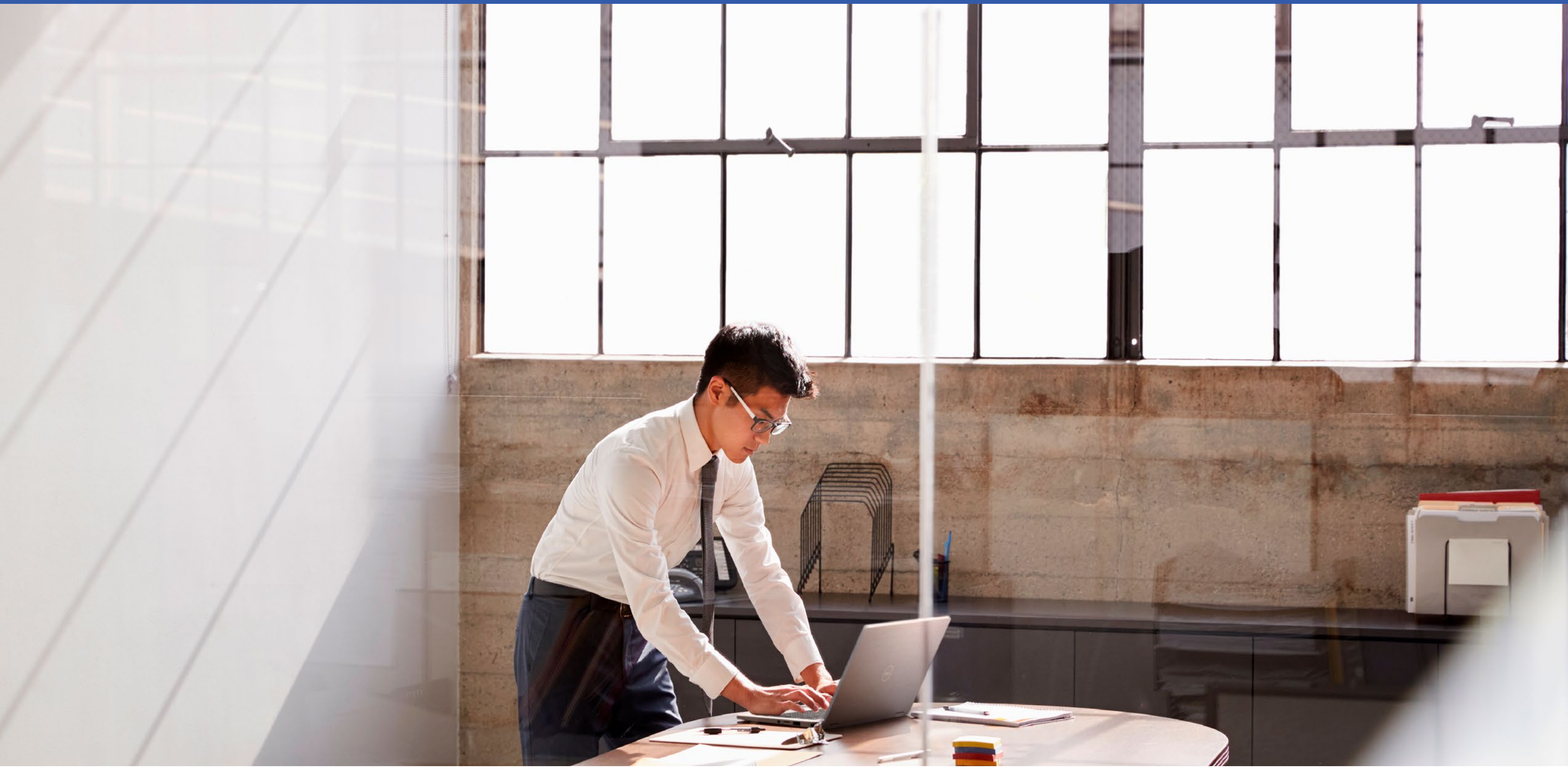
How do AI PCs address challenges faced by enterprises today?

According to Figure 1 of the Gartner® research,[5] the **top 2 barriers** to implement AI Techniques are

“Estimating and demonstrating AI value” at 49% and “Lack of talent/skills” at 42%.

Enterprises are struggling to find ways to immediately recognize the impact of AI for their organizations, due to concerns around their data privacy, legacy tech stacks, and ballooning costs of AI Compute in the cloud.

On-device AI applications can provide benefits to enterprises, sidestepping the traditional challenges faced with integrating cloud options. Creating and implementing these applications for an enterprise is achievable, with Dell and Deloitte leveraging experience and relationships across the AI on- device ecosystem to create Neuron.



[5] Gartner, “Gartner Survey Finds Generative AI Is Now the Most Frequently Deployed AI Solution in Organizations”, 7 May 2024. https://www.gartner.com/en/newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

Neuron Case Study

Eager to take advantage of AI PCs, Deloitte worked with Dell to create Neuron, an AI-powered coding assistant. Neuron runs locally on an AI PC and can be used across diverse industries while remaining user friendly and impactful. Building an AI application capable of running on-device requires expertise in device hardware, model selection and optimization, and software development to package such intelligence within a user facing application.

With the coding assistant use case in mind, Deloitte and Dell aligned on two target platforms: the Intel® Core™ Ultra 7 155H with CPU, Arc™ GPU and AI Boost NPU and the Snapdragon® X Elite with Oryon™ CPU, Adreno™ GPU, and Hexagon™ NPU. Both System-on-a-Chip (SoC) offerings deliver advanced computing capabilities and can handle AI workloads. Specifically, the Latitude 7450 and XPS 13 (9340 and 9345) were selected for development and testing.

Shop AI PCs

After identifying the development systems, it was necessary to choose an appropriate LLM for the coding assistant. With Dell's expertise on LLMs best suited for the target platforms, Deloitte ran internal benchmarking tests of various LLMs and compared performance with the necessary requirements of the code generation application. The Llama 2 7B Chat and Code Llama 7B Instruct models were chosen for their balance of code quality, inference speed, and hardware support.

Every platform has special considerations when it comes to deploying AI

Unique hardware stacks necessitate unique firmware and software stacks. For example, to prepare and run a model, Intel provides OpenVINO™, and Qualcomm provides AI Engine Direct (QNN). Dell used these tools to perform hardware-tailored model optimizations, including quantization of the LLMs to reduce their size and conversion to hardware-accelerated runtime formats. For the Intel platforms, an open-source OpenVINO™ recipe was leveraged, with verification & performance enhancement recommendations from Intel. For the Snapdragon X Series platforms, Dell worked directly with Qualcomm for early enablement of a custom build of Llama to take full advantage of the Hexagon™ NPU. Dell's partnerships with Intel and Qualcomm ensured these libraries were fully utilized and seamlessly integrated into the application for optimized LLM inference on device.

With the optimized backend compute stack set up, Deloitte designed a UI application integrated with Microsoft Visual Studio Code around key features developers wanted in the AI application. It offered inline code completion and a chat assistant for developers to use in code generation, code review and bug fixing, and text to code ideation.

Deloitte distributed AI PCs to developers across various industries.

Each developer used a Dell AI PC (Latitude 7450 or XPS 13) and an AI Coding Assistant Application running 100% on device. They used the AI PCs and Neuron in their everyday work. Since, Neuron runs fully on-device, developers were able to pass confidential information through the assistant and even use it offline during a flight.

Over 4 weeks of testing in the US, developers reported...

41%

increase in efficiency, driven by decreases in time spent coding

83%

developers reported time saved using Neuron

34%

increase in output in everyday development

25%

improvement in code quality driven by decreases in bugs in code

Comparatively, independent research on the impact of cloud-based coding assistants found a potential 55% productivity increase for developers (arxiv.org).

By using Neuron on AI PCs, developers have already achieved 75% of the gains offered by cloud-based coding assistants. Rapid advancements in chipmaking will continue to help drive the closure of this gap.

What strategies can you build around AI PCs?

The commercial sector is likely to be the early adopter of AI PCs, driven by their positioning as productivity tools and the initial high price points exceeding $1,000. The commercial PC market, which is due for a refresh ahead of the Windows 10 end-of-life in October 2025, will see a significant boost from AI PCs. Currently, 75% of CIOs in the US and EU are evaluating or planning to evaluate AI PCs, with an average willingness to pay 6% more for these devices.[6]

Already, the AI PC is a fundamental transformation that is reshaping the industry. According to IDC, AI PCs will grow from nearly 50 million units in 2024 to more than 167 million in 2027, representing nearly 60% of all PC shipments worldwide.[7]

With AI PCs, organizations can increase their pace of innovation in a competitive AI space. Driven by zero cost inferencing and increased data security, AI accelerates the integration and scaling of software solutions. Using AI PCs, Deloitte implemented tailored changes for users based on feedback and tested out various LLMs to deliver Neuron.

Companies that adopt AI PCs can also see productivity increases by utilizing these platforms for features that uniquely help users with their day-to-day work. These features depend on robust and reliable output. Neuron's case study demonstrates that AI PCs can produce quality output for functions that traditionally use cloud computing, such as code generation. More TOPS on devices, more memory, and more efficient model architectures will only continue to develop over time to drive these compute capabilities.

[6] The Dawn of AI PCs: Transforming the PC Market Landscape, May 2024

[7] Based on IDC, Worldwide Artificial Intelligence PC Forecast, 2023–2027, doc #US51747324, January 2024

With 83% of enterprise CIOs planning to bring workloads back from the public cloud to on-premises environments this year,[8] it is a great time for enterprises to adopt a Hybrid AI approach. This strategy can offer scalable, efficient, and enhanced inference capabilities at the edge. Once these solutions scale, inferencing on-premises can be 75% more cost-effective than public cloud solutions over 3 years.[9]

A Hybrid AI Approach

Hybrid AI enhances intelligent inferencing by integrating capabilities across public clouds, co-location centers, on-premises data centers, and edge devices, leveraging their growing computational power to optimize performance in diverse technological settings. This allows enterprises to realize the cost savings and privacy of on-device inferencing alongside the robustness of inferencing in the cloud. For example, an LLM router created by Deloitte conducts performance analysis on prompt responses and routes requests to various cloud or local LLMs, creating cost savings while upholding robust performance for the user. By strategically incorporating on-device inferencing capabilities, organizations can effectively position themselves to adapt to the growing significance of devices in inferencing processes. This proactive approach helps ensure an effective transition as these devices become increasingly integral to inferencing operations.

Organizations poised for intelligent inferencing can gain a competitive edge by delivering exceptional customer experiences that align with company needs and worker preferences, ultimately helping boost satisfaction.

A Hybrid AI approach will be necessary for ISVs (Independent Software Vendors) as AI PC users are driving demand for features that utilize inherent privacy, security, and improved latency.

- ✓ ISVs who utilize on-device inferencing will operate unique and differentiated features at a fraction of the cost of doing so in the cloud.
- ✓ Increasingly capable inferencing platforms will further enable the features ISVs can offer locally, driving more users and demand.
- ✓ Simplified access to local compute, such as through a browser extension, will drive new market entrants and increased competition.

Through their relationship, Dell Technologies and Deloitte can make AI in the workplace a tangible reality. Leveraging their collective experience, these industry leaders have paved the way for businesses to harness the transformative power of AI. Project Neuron highlights the success of this collaboration, emphasizing the impact of the AI PC in revolutionizing organizational operations.

Fast-track your AI Journey with APEX PC-as-a-Service

Explore advanced processing power and seamless scalability of Dell AI PCs and Precision AI workstations. Reach out to your Dell Sales Representative to equip your enterprise with devices that redefine operational efficiency.

Learn More

[8] The Dawn of AI PCs: Transforming the PC Market Landscape, May 2024

[9] ESG Economic Summary Dell Technologies for LLM Inferencing, April 2024

About the Authors

Jake Leland, Software Principal Engineer

Dell Technologies

Jake is a software architect in Dell's CSG CTO organization. With years of experience in cloud computing, web technologies, native applications, and artificial intelligence, he drives state-of-the-art solutions through rapid prototyping and cross-functional relationships. His passion lies in designing, delivering, and demonstrating proof-of-concept projects that push the boundaries of innovation.

Matt Kalman, Managing Director

Deloitte

Matt's 20+ year career has focused on leading cross-functional engineering teams to support market facing outcomes for consumers and enterprises. He is a full-lifecycle consultant who engages at all levels: from upfront strategy and architecture, delivering on that vision from early-stage proof of concepts to complex product launch; to advising clients on how to operate mature products at scale. He has broad and extensive experience working with engineering leaders and CTOs to mobilize large transformational programs that drive strategic outcomes across embedded software, front end development, AI & data and cloud capabilities.

Alex Liebetrau, AI & Data Consultant

Deloitte

Alex drove the creation and testing of the Neuron coding application. He has experience creating early-stage proof of concepts of AI solutions and managing the dev ops required to scale these solutions.

About Dell Technologies

A renowned global technology company dedicated to transforming businesses and societies through the power of technology. With a strong focus on innovation and customer-centric solutions, Dell Technologies provides a wide range of products and services, including cutting-edge AI-enabled PCs. These AI PCs are designed to enhance productivity, streamline workflows, and empower organizations to unlock the full potential of artificial intelligence.

About Deloitte

Deloitte provides industry-leading audit, consulting, tax and advisory services to many of the world's most admired brands, including nearly 90% of the Fortune 500® and more than 8,500 U.S.-based private companies. At Deloitte, we strive to live our purpose of making an impact that matters by creating trust and confidence in a more equitable society. We leverage our unique blend of business acumen, command of technology, and strategic technology alliances to advise our clients across industries as they build their future. Deloitte is proud to be part of the largest global professional services network serving our clients in the markets that are most important to them. Bringing more than 175 years of service, our network of member firms spans more than 150 countries and territories. Learn how Deloitte's approximately 457,000 people worldwide connect for impact at www.deloitte.com. As used in this document, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

DELLTechnologies
In association with Deloitte