IDC Quick Take Observations on the Next-Generation AI PC

Welcome to the Next-Generation AI PC Era

June, 2024

Written by: Tom Mainelli, Group VP, IDC Devices and Consumer Research

Microsoft has announced plans for the next generation of AI computers, called Copilot+ PCs. These PCs will include new, fast processors and powerful but efficient neural processing units (NPUs) offering 40+ trillion operations per second (TOPS) of AI performance. PCs with 40+ TOPS NPUs will enhance performance across the growing list of apps and Windows system components designed to leverage local AI capabilities on a purpose-built piece of silicon. But perhaps more importantly, these Copilot+ PCs will bring to market an AI-first OS designed to enable new client-side PC features and functionality that will elevate every user's experience, allowing them to work smarter and not harder. Additionally, Microsoft says that each next-generation AI PC will evolve and improve the more you use the system.

Understanding TOPS.

Most AI workloads today happen in the cloud, but as more companies embrace AI, generative AI (GenAI), and large language models (LLMs), more of these workloads will need to move to the edge to scale. The introduction of the NPU and, more specifically, 40+ TOPS NPUs makes it possible to bring pervasive and efficient AI capabilities to the PC. In IDC's taxonomy, a PC with a 40+ TOPS NPU is called a Next-Generation AI PC.

TOPS as a measure of AI performance didn't appear with the introduction of the NPU. The PC sitting on your desk today likely offers some AI TOPS performance tied to the central processing unit (CPU) and the graphics processing unit (GPU).

What makes the NPU special is its remarkable efficiency in handling AI tasks on the client itself. While both GPUs and NPUs offer parallel processing, the GPU is geared toward rendering graphics, whereas the NPU is optimized for complex computations and can handle a wide variety of data types, including images, speech, and temporal data. This makes it incredibly efficient for running AI models. This is particularly crucial for a notebook PC running on battery. This efficiency can translate to notably less power draw across a fleet of PCs, resulting in a greener solution.

Thanks to the NPU's power-efficient design, the operating system can seamlessly run AI, constantly monitoring system activities and user input. Achieving this performance on the CPU and/or GPU would demand significantly more power, leading to potentially sluggish system performance and shorter battery life. While NPUs are new to the market, forward-thinking IT decision makers (ITDMs) are already keen to adopt next-gen AI PCs, according to a recent Microsoft-sponsored IDC survey, with 30% suggesting that they'll buy as soon as the systems become available and another 45% reporting that they'll buy within the first 12 months of availability.

It's important to note that a system with an NPU isn't going to ignore the AI capabilities of its CPU and GPU. Which silicon an app uses will depend upon how the app was written, and software vendors are evolving their products today to leverage the CPU, the GPU, and the NPU. However, to carry the Copilot+ branding, a PC must meet Microsoft's minimum

system requirements, which include having a greater-than-40 TOPS NPU, a minimum of 16GB of RAM, and 256GB of local storage. The first systems to meet these specs will leverage Qualcomm silicon; compatible silicon from Intel and AMD should appear later this year.

Most people are just beginning to understand the capabilities of GenAI and the power of LLMs and small language models (SLMs). Today, most of these capabilities are almost entirely cloud based. Now imagine running multiple LLMs and SLMs locally, leveraging the NPU, resulting in more flexibility.

Elevating Both AI and Non-AI Apps

Independent software vendors (ISVs)are at work today updating their existing apps and creating new ones to leverage the always-on AI capabilities of Copilot+ PCs. Combined with Microsoft's first-party apps, you can expect to see a growing list of use cases and workloads made easier, faster, and more efficient thanks to these apps' abilities to use the NPU. Examples include image-generation apps that leverage the NPU to allow artists to quickly iterate on designs, security apps that keep a PC safe while eliminating friction for the user, and customer service apps that make it easier for an employee to provide a more customized experience to the customer.

While app-specific AI features will please many users, what gets especially interesting is what happens when the entire operating system gets additional AI integration on top of what already exists. Microsoft's updated version of Windows 11 will be an AI-first OS. This means that AI will be operating persistently and pervasively across the OS, driving a range of new features and functionality, as well as enhanced endpoint security.

Early on, we'll see features such as the ability to find any data easily and quickly that's passed across your screen, live captioning with real-time translation, and new ways to move between apps to pull together new content seamlessly. New research shows that ITDMs find these features compelling. Respondents in the IDC survey said that the top expected benefits of local AI are improved productivity, accelerated business results, and faster innovation. AI apps will be hugely important, but an AI OS lights up functionality across the depth and breadth of a user's workday, and many new experiences may not require an app at all.

A New Beginning

We expect that these initial features and functionalities will be just the start. Over time, an AI-first OS will adapt and evolve to better suit you, the end user. Today, we adapt ourselves to fit the machine, but in the future, the machine will adapt to suit us. This will likely take the form of new interaction modes, more personalized results, new ways of collecting and displaying information, and evolutionary methods of collaborating with other people and AI agents.

As developers get their hands on these Copilot+ PCs, we can expect them to deliver new and unexpected features and functionality. First, this will mean changes to the apps we use today, but over time, it will lead to brand-new types of apps covering workloads we've not yet imagined.

Admittedly, the tech industry is prone to hyperbole when it comes to new product introductions. The rollout of AI broadly, and of next-generation AI PCs specifically, is not one of those moments. IDC's *Future Enterprise Resiliency and Spending Survey*, February 2024, found that 17% of enterprises worldwide have already introduced generative AI apps into their environment, while another 36% are investing significantly, with near-term plans to deploy. The introduction of next-gen AI PCs will help to accelerate this movement toward an AI-infused future. We expect this to fundamentally

change the PC market and set a new course for the industry going forward, leading to happier and more productive end users.

Sources

- IDC *AI Computer Survey Sponsored by Microsoft*, May 2024
- IDC *Future Enterprise Resiliency and Spending Survey*, February 2024

Tom Mainelli, *Group VP, IDC Devices and Consumer Research*

Tom Mainelli manages the Device & Consumer Research Group, which covers a broad range of hardware categories, inclusive of both home and enterprise markets, as well as IDC's growing consumer research practice. The device research includes PCs, tablets, smartphones, wearables, smart home products, thin clients, displays, and virtual and augmented reality headsets. He also manages IDC's supply-side research team that tracks display and ODM production across a wide range of products.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
blogs.idc.com
www.idc.com