THE NEW ERA OF THE AI PC

WHAT DOES THE AI PC MEAN FOR USERS, AND HOW DOES IT WORK FOR THEM?

EXECUTIVE SUMMARY

The PC market has been on quite a rollercoaster the last few years, with chip shortages caused by COVID and buying frenzies to accommodate people's ability to work from home. That cycle naturally caused the PC industry to surge, at least initially, but then come back down hard, dropping as much as 16% in 2022 and 28% in the fourth quarter of last year.[1] Since then, the industry has normalized, with the market back to its normal cadence and companies like Intel[2] and AMD[3] reporting stronger earnings and upgrading projections for the coming quarter.

During the depths of the PC downturn, around January 2023, the IT industry began to pivot to AI, predominantly powered by the cloud, as one of the key enablers of new tools and services. This was especially enabled by OpenAI and the public availability of GPT-3, along with the heavy use of large language models (LLMs), which had previously not been performant or accurate enough to be useful, especially for business applications.

A significant multibillion-dollar investment by and partnership with Microsoft, in which OpenAI's capabilities would become infused with Microsoft's Azure cloud services, further accelerated this momentum. Eventually, these AI capabilities would extend into services like Bing Search and Office applications. This showed Microsoft's commitment to AI as an enabler for consumers and the enterprise, accelerating the industry's appetite for generative AI (GenAI). This also resulted in Microsoft's announcement of Copilot[4], its own AI assistant brand that exists across Windows, productivity apps and Bing.

Part of what enabled the industry to pivot toward AI was the availability of hardware to accelerate AI workloads – hardware for developers and cloud providers to enable AI applications at scale. This initially started with GPUs and CPUs but has since extended to specialized accelerators from companies like AMD, Google, Intel, Microsoft and Nvidia.

[1] Gartner, January 11, 2023, Gartner Says Worldwide PC Shipments Declined 28.5% in Fourth Quarter of 2022 and 16.2% for the Year
[2] Intel, October 27, 2023, 3rd Quarter Earnings Presentation
[3] AMD, October 31, 2023, AMD Reports Third Quarter 2023 Financial Results
[4] Microsoft, September 21, 2023, Announcing Microsoft Copilot, your everyday AI companion

While the cloud has initially dominated the AI discussion, there are limits to scale, with models like GPT-4 with 8K context costing as much as $0.03 per 1,000 tokens for prompts and $0.06 for completion according to Microsoft's own price estimates.[5] Scaling this pricing across millions of queries and users simply won't work long term for many applications and can lead to cost overruns and may explain why so many AI startups need ample access to compute to train and run their models. The tech industry is still very much in the early stages of implementing AI, which is why training workloads rather than inference is so much of what we've seen so far. As inference becomes more crucial, with more applications integrating AI capabilities, there will be more challenges with cost models and privacy and security (i.e., keeping data from leaking through the internet). This is where the AI PC becomes beneficial.

The AI PC needs to deliver high-performance AI compute that can run native applications like Microsoft's Copilot and other AI applications that can compete with the cloud in a scalable, performant and secure manner. This means that the AI PC needs to run most of its AI applications locally and utilize the cloud only when necessary, which is a major departure from the way most AI applications run today. As such, the AI PC needs to be as power-efficient as possible in how it runs AI applications that require an AI accelerator, whether that comes in the form of a CPU, GPU or NPU. Using the proper accelerator for the AI workload can make all the difference in user experience and battery life.

An AI PC must effectively run all the different types of AI applications that might need to run on a PC, especially Microsoft Copilot, which can improve user productivity without sacrificing battery life, security or privacy. By offering local performant and efficient AI computing at little to no cost to the developers, the AI PC will solve the challenges cloud-only AI applications have today as well as the overall growth challenges of AI applications. The AI PC will not only usher in the next generation of the PC platform but redefine what it means to work on a PC and its power as a productivity platform. AI acceleration needs to be a component of people's PC-buying decisions.

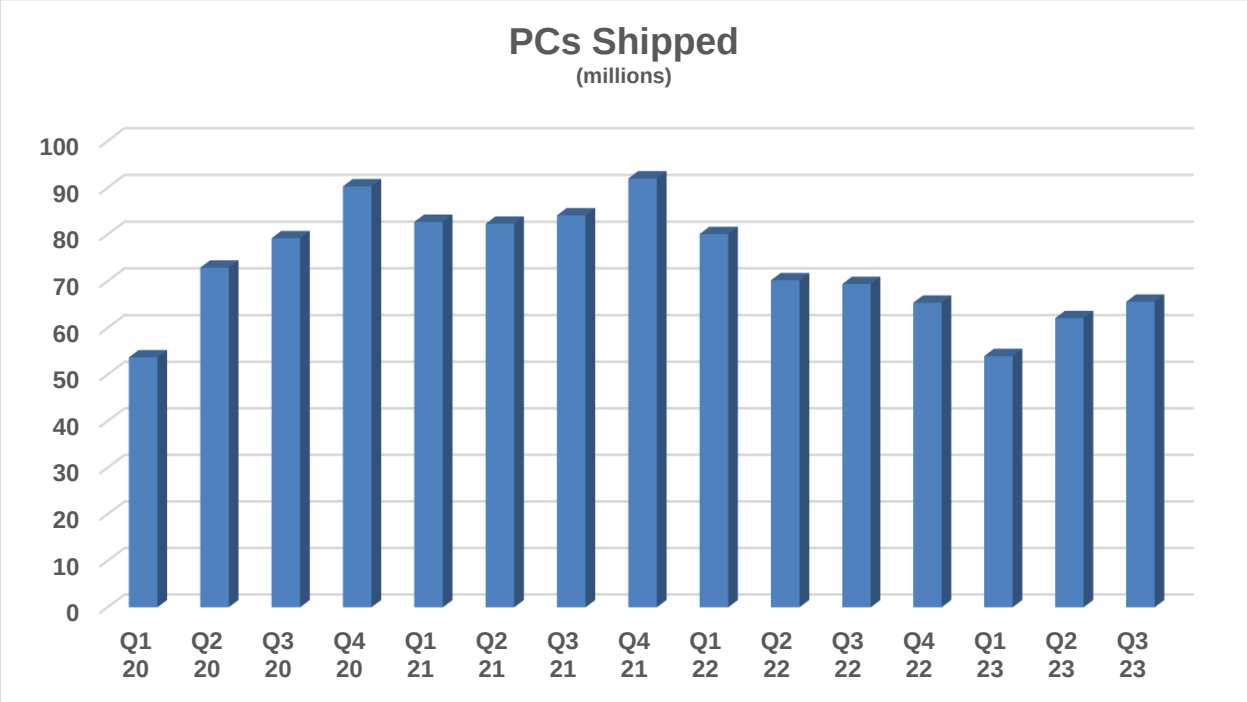
THE PC MARKET AND AI PC

Earlier, we mentioned the turbulence experienced by the PC market over the past several years. It appears to have bottomed out in Q1 2023, however, beginning a

[5] Microsoft, Azure OpenAI Service pricing

normalization of PC shipments through the rest of the year and an expectation of a full return to normal volumes by 2024.[6]

FIGURE 1: GLOBAL PC SHIPMENTS – CANALYS QUARTERLY 2020-2023



Source: Moor Insights & Strategy

This normalization of the PC market in 2024 syncs well with a refresh cycle for systems purchased in 2020 and 2021 that might be starting to feel a little long in the tooth. Opinions on the PC refresh cycle vary widely, but when you consider the power and performance improvements delivered by PC OEMs and their silicon partners, there have been considerable improvements over the last three years alone – ignoring the new AI-accelerated tools, features and applications that have gained momentum over the years. We are coming upon a new post-COVID refresh cycle for PCs that also coincides with the recent momentum in AI applications, which has been driven predominantly by a few factors.

The new AI-accelerated age we live in has mostly been driven by the growth of new GenAI capabilities from technologies like LLMs. These models are possible thanks to

[6] Canalys, July 11, 2023, Global PC Market decline eases as shipments drop 12% in Q2 2023

the huge data sets collected during what was known as the 'big data' era immediately preceding COVID. Without these large datasets, it would be difficult to attain any kind of accuracy when training these models. Additionally, companies had to develop hardware to accelerate the training time of these models as new data came in to help retrain and hone models to improve their efficacy. Much of this training and inference happened on CPUs and GPUs in large clusters, predominantly in HPC clusters or the cloud.

However, as these models start to be quantized[7] to run on smaller, more local devices, there is an increased need for local AI compute that works in conjunction with the larger LLMs to provide quick and accurate results. With Microsoft embracing OpenAI, LLMs like Llama 2 from Meta,[8] and AI as a critical component of the Windows and Office experience with Copilot,[9] there is significant buy-in from the business community and developers to take advantage of AI.

This leads us to define exactly what an AI PC is and how it might work differently from current PCs. First and foremost, an AI PC will be a computer that predominantly runs Windows (although there is a considerable amount of AI model development and research happening on Linux). AI is already so central to the Windows 11 experience that it has a key command – Windows logo key + C – to launch Copilot.

But it isn't enough to run the latest version of Windows to be considered an AI PC. An AI PC also needs to have accelerators with AI-specific enhancements that enable improved AI experiences in the most power-efficient manner possible. This means that an AI PC is a computer that can run most of the AI frameworks designed to work on those AI accelerators, whether they be CPU, GPU or NPU.

There is a need for an AI PC to take advantage of the trends of applications adding AI features or being entirely based on AI-accelerated capabilities based on LLMs. With tools like ChatGPT,[10] workers and consumers create new things without needing to spend time gathering the required components to create them from scratch. They can refine and improve upon existing work rapidly and cleanly with features like AI noise removal, content-aware photography and AI video and audio editing tools. There is also a growing set of collaboration tools that have embraced AI for face tracking, eye contact, background auto-blur and green-screenless background removal.

[7] Hugging Face, Quantization
[8] Meta, Introducing Llama 2
[9] Microsoft, Discover the power of AI with Copilot in Windows
[10] OpenAI, November 30, 2022, Introducing ChatGPT

The AI PC is an evolution of the PC platform – not a radically new platform. This means that while AI capabilities can be new and enable new use cases, there are still a lot of things about the PC that will remain the same. AI is best seen as a force multiplier for many of the things that you are already doing. Apps like Descript aren't necessarily doing something that wasn't possible before, but they are doing it quicker and with less effort. There will also be applications that would have never existed if not for AI, like ChatGPT and the many applications that are likely to be built upon it or use it as a feature.

There is also the balance of how AI work gets done – how much gets processed locally and how much gets done in the cloud. Commonly referred to as hybrid AI, this concept of fusing cloud with local AI does not escape the cost challenges of running AI workloads in the cloud, especially with the cost of GPUs in the cloud not coming down due to demand. Client-based solutions like the AI PC solve not only performance and cost considerations but also address security and privacy concerns around sending data back and forth to the cloud. When running locally, AI PCs can help lower the cost of running an application to mostly free while also significantly lowering latency.

For the foreseeable future, there will be a tradeoff between how long it takes the AI model to run locally and how much it costs to run it in the cloud. There are also environmental considerations, like how much power a PC might consume running the AI model versus a high-performance CPU or GPU in the cloud.

AI workloads might be features that accelerate existing workloads or entirely new tools with more complexity, requiring more memory and AI-acceleration performance. We believe that many application developers will initially explore hybrid AI approaches that balance the cost and performance tradeoffs that exist in AI today but eventually transition to the best experience for the user, which will heavily depend on the application's most important workload.

AI PC Applications And AI PC Future

AI PC applications already exist; in fact, Microsoft has already dedicated an entire section of the Microsoft Store called the AI Hub[11] to AI apps. In addition to Microsoft, PC hardware vendors like Intel are already deploying programs like the AI PC acceleration

[11] Microsoft, May 23, 2023, Welcoming AI to the Microsoft Store on Windows

program[12] to improve developer engagement and understanding of how to develop for AI PCs.

Companies like Blackmagic Design have been working with partners like Intel and Qualcomm to accelerate its wildly successful DaVinci Resolve video editing software. The latest release of DaVinci Resolve includes a long list of AI tools, including AI speech-to-text transcription, AI-based voice isolation, AI audio classification, upscaling, person masking, smart reframe, stylizing with OpenFX, face refinement, dead pixel fixer, object removal and patch replacer.

Blackmagic has already committed to optimizing its software for Intel's Core Ultra processors as well as Qualcomm's Snapdragon X processors for AI PCs. Other developers like Topaz Labs and Skylum already make photo editing tools that help to improve the quality of images, whether it's by removing noise or offering a complete suite of AI-powered features. The gold standards of photography editing tools – Adobe's Lightroom and the more mainstream Adobe Express – both have AI capabilities built in and are also featured in Microsoft's AI Hub as popular apps.

If you're looking to produce or edit audio, Audacity is a popular open-source app that has embraced AI and is partnering with Intel to bring free AI tools to the masses. Audacity isn't alone on the AI audio side, as plug-ins like Krisp (which works with Audacity and many other apps) use AI to help remove background noise, clarify accents and summarize call transcripts. Descript, which we mentioned earlier, is an extremely popular tool in the podcasting community. It is also another AI application that already has a PC client desktop app and uses AI heavily to help edit photos and videos even though many of its AI features are still cloud-based.

Collaboration apps like Zoom and Webex by Cisco have also committed to using AI tools and features in their respective apps on Intel's platform as part of the company's AI PC initiative. Other video-based software developers like XSplit and VideoCom have also spoken about their efforts to target the improved CPU and GPU capabilities of Intel's platform and utilize direct NPU capabilities to further offload the CPU and GPU whenever possible. XSplit specifically quoted a 30% improvement in removing inaccuracies on live video thanks to using the newer larger AI model on the NPU.[13]

[12] Intel, Intel AI PC Acceleration Program
[13] Intel, What Our ISVs Are Saying

Microsoft has also identified other AI applications for the PC that already improve the user experience:

- Spark Mail helps to centralize your email and uses an AI assistant to help you write better emails.
- Yoodli is a private speech coach that helps you practice speeches and gives you grades and feedback in real time with AI.
- Taskade embraces many AI features to help you manage thoughts and tasks, plan and collaborate.
- Moises helps musicians remove their vocals and instruments from any song and create new mixes quickly and easily.
- CapCut is an AI-enhanced video editing tool from ByteDance originally created for mobile video editing that has since become superpowered on the PC.
- Kickresume is a faster resume builder that helps users quickly and easily generate and modify resumes based on their target audience.
- Gamma is presentation software powered by AI helping people to generate entire presentations and make quick stylistic changes on the go.

Microsoft is not limiting its AI engagement to just the AI Hub; it has also effectively integrated AI into the rest of the Windows operating system with Copilot. It is already found[14] in Teams, Outlook, Word, Excel and PowerPoint. The capabilities and features found in Copilot inside of Windows and Microsoft 365 (formerly Office) are merely the beginning of AI on the PC.

As you can see, content creators and coders are the ones who benefit the most from AI apps today through LLMs and other generative AI tools (although marketing and communications professionals do as well). Beyond these individuals, there is an even broader future for far more professionals to benefit from AI tools outside of collaboration applications. However, those applications may take some time. This is why AI applications – those mentioned above and others that don't even exist – are likely to drive buying decisions soon if not already.

Call to Action

The IT industry is undergoing a major shift with the injection of AI into many of the different applications that people use today. Combined with an expected upcoming PC refresh cycle, there is a major opportunity to prepare the industry appropriately with

[14] Microsoft, Copilot for work

well-equipped AI PCs that have the necessary dedicated compute capabilities to deliver a responsive and secure AI-accelerated experience. The AI PC will be pivotal in scaling AI applications beyond the cloud with a considerably better cost profile for developers as well as improved latency, security and privacy for the user or enterprise. People looking to buy PCs within the next year or two should seriously consider the importance of AI as a component of their buying decisions.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR
Anshel Sag, Vice President and Principal Analyst, PC, Mobility & Spatial Computing

PUBLISHER
Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES
Intel commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.