AMD CIO

SPONSORED CONTENT

Integrating Generative AI into PC Silicon: The PC Experience at an Inflection

WITH RESEARCH AND ANALYSIS BY IDC

Integrating Generative AI into PC Silicon: The PC Experience at an Inflection

While today's generative AI services reside almost exclusively in the cloud, PC OEMs are working on embedding them natively into the OS and onto the silicon. Moving AI from the datacenter to the edge and the endpoint is a necessity for the industry as it helps ameliorate the overall resource crunch facing AI services today.

Introduction

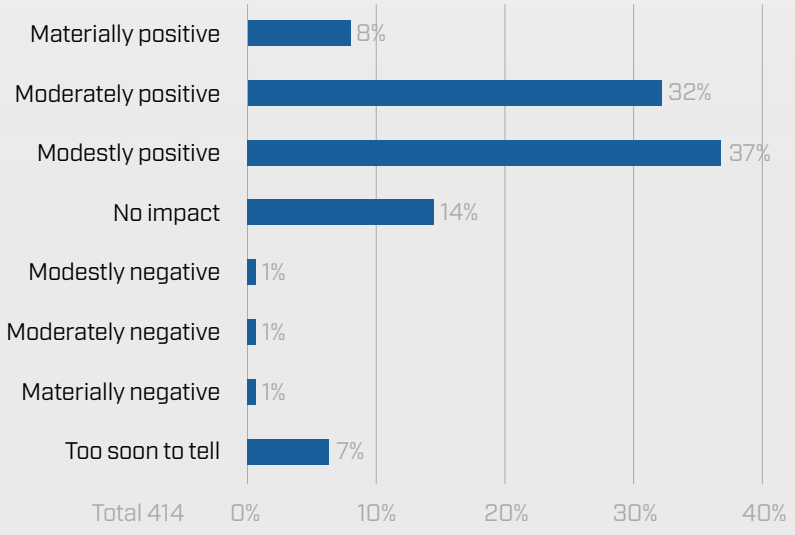
AI Making Commercial Breakthroughs

Since the advent of modern computing, artificial intelligence (AI) has been lauded as the future. Humans have long understood the vast potential of using computers to augment and even supersede human decision making. Computers can think faster serially, process more in parallel, and do so with a higher degree of fidelity and reliability than people. Yet, the leap to get from quick thinker to full intelligence in computers is wide and requires machines to be able to intuit from context and experience.

Over the years, there have been small steps toward this fully realized intelligence. Many brands have deployed chatbots to help customers reduce support time.

Figure 1 | **AI PC Buyer Perceptions**

What kind of impact do you anticipate AI PCs having on your organization?



Source: IDC's 2023 U.S. Commercial PCD Survey

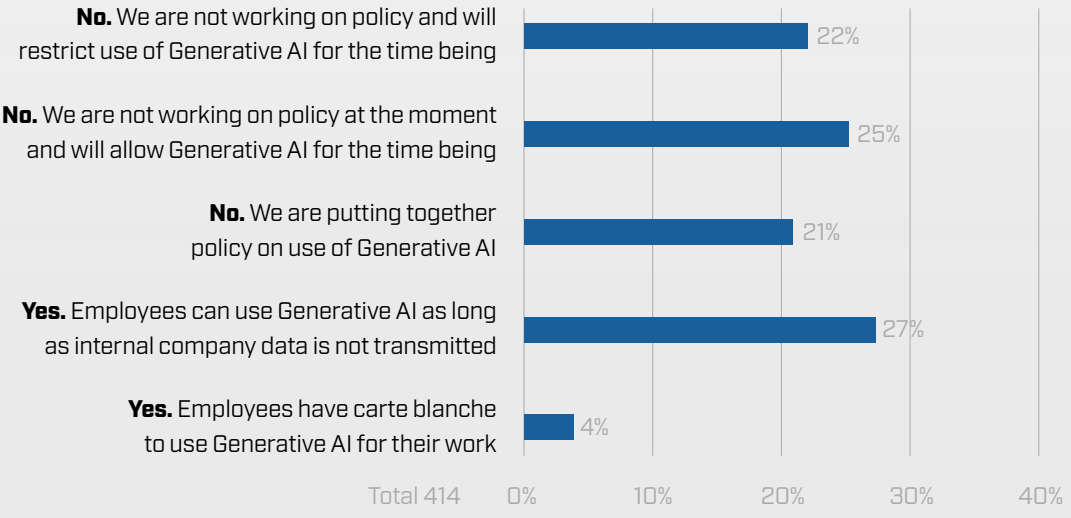
Every smartphone nowadays comes with a personal digital assistant. Companies often use AI for computer vision and language translation. In all these instances, the technology is impressive, but the use case is narrow.

Enter generative AI and the likes of conversational AI and image generation AI applications. Whereas the earlier AI use cases require significant human intervention, generative AI can provide complex and robust outcomes with simple instruction. It can iterate and refine with short follow-ups. Generative AI is a major leap toward realizing full intelligence, and companies worldwide are taking quick notice.

According to IDC's 2023 U.S. Commercial PCD Survey, most organizations are either exploring initial use cases for generative AI (45%) or are already directly investing in the technology (13%), as shown in Figure 2. Likewise, the majority have either already put together an official policy on the use of generative AI (31%) or are currently writing the policy (21%). The era of commercial AI is nearing, and forward-thinking businesses are preparing posthaste.

Figure 2 | **Generative AI Policies**

Does your organization have an official policy in place regarding employees utilizing generative AI for business purposes?



Source: IDC's 2023 U.S. Commercial PCD Survey

Bringing AI to the PC

Over the years, PCs have become more intelligent, leveraging predictive models to help enhance video, collaboration, and speech-to-text. The next generation of AI on PCs will integrate generative AI capabilities, bringing about a revolutionary new experience. Today's generative AI services reside almost exclusively in the cloud. Soon, PCs will be able to access those capabilities locally and privately. In fact, some CPUs are already shipping with dedicated AI chips.

AI deployments from the datacenter to the edge and the endpoint is a necessity for the industry as it helps ameliorate the overall resource crunch now facing AI services. But the advantages to buyers are obvious as well: optimized costs, low latency, and privacy/sovereignty of internal data. More than a third of respondents to the aforementioned IDC survey said their greatest concern regarding generative AI is "privacy of internal data."

Allowing internal company data to leak outside the perimeter walls is clearly detrimental to an organization. It's important to consider the careful handling of data to ensure its privacy. Additionally, it's necessary for you to align with external providers' data retention policies. Putting the generative AI engine on the client and running its workloads locally helps ensure that AI benefits the company and its customers.

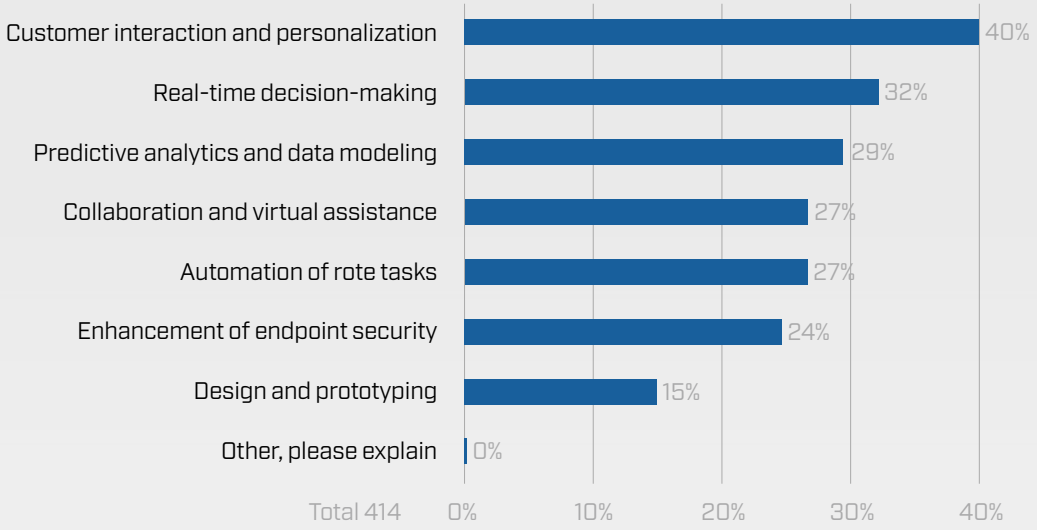
When IDC survey respondents were asked about these benefits, the majority chose either "increase user productivity" (30%) or "improve customer engagement" (28%). Both benefits are driven by AI's transformation of the user experience. The majority of respondents believe that the greatest benefit for their users will come from either "improved automation and efficiency" (38%) or "accelerated data analytics and insight" (23%).

Given the breadth and depth of impact AI PCs could potentially have on an organization, it's unsurprising to see varied initial use cases. When asked about initial use cases, no single option reached majority (see Figure 3). Responses were led by "customer interaction and personalization" (40%), "real-time decision making" (32%), and "predictive analytics and data modeling" (29%).

And so, although AI PCs are just in their infancy stage, IT managers are already dreaming up use cases for the technology.

Figure 3 | **Generative AI Use Cases**

What will be the primary use cases for your initial deployment of AI PCs?



Source: IDC's 2023 U.S. Commercial PCD Survey

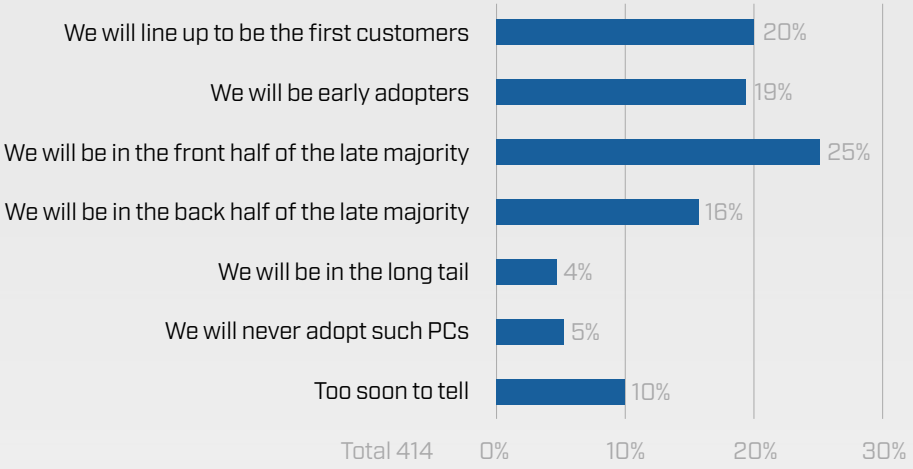
AI PCs to Power Next Hardware Revolution

The dreams of IT decision makers (ITDMs) are lofty ones. In the same survey, two out of three ITDMs indicated that AI PCs would have a positive impact on their organizations. This trend is even more pronounced for large enterprises (94%). Indeed, the expectations are so high that the majority (51%) rate onboard AI capabilities as being somewhere between "very important" and the "top factor" when they go to refresh their fleet.

When asked how far up the adoption curve their organizations will be for AI PCs, the sample skewed earlier. In total, more than a third said they would either be "first in line" for AI PCs (20%) or an "early adopter" of them (19%) (see Figure 4). Again, this attitude is even more pronounced for large enterprises where the combined majority indicate they will either be early (25%) if not first (28%).

Figure 4 | **Adoption Intentions**

How will your company approach AI PCs?



Source: IDC's 2023 U.S. Commercial PCD Survey

ITDMs across all sectors are dreaming up big impacts across a wide variety of use cases, helping their end users improve automation and accelerate data analytics. Subsequently, organizations expect to see material improvement in end-user productivity and customer engagement.

Benefits

In the future, organizations that deploy AI PCs can expect their users to realize the following benefits:

- Improved automation and efficiency
- Accelerated data analysis and insight
- Strengthened security and privacy
- Support for remote collaboration
- Enhanced real-time decision making

This is turn will provide the broader organization the following tangible benefits:

- Increased user productivity
- Improved customer engagement
- Accelerated product innovation
- Reduced operational costs
- Strengthened security and compliance

Considerations

While AI has the potential to transform organizations, several of the following risks should be considered:

- The AI technology market remains fast-moving. Generative AI is still in its relative infancy. As such, devices, services, and offerings are still rapidly evolving, and vendor battles have yet to be won.

- Questions regarding data ethics and governance will also arise. As AI increasingly moves into the role of content creator, companies must sort out who gets credit, and more importantly, ownership of what. Who is responsible for what is said or created at the end of the day?

- AI also has a massive human impact. Innumerable people everywhere see the rapid rise of AI and fear what that means for their long-term livelihoods. Companies need to assess how to integrate AI without displacing too much of the human intelligence.

- Finally, a relationship exists between AI location and the need for data privacy. Companies that are more restrictive of their own internal data might elect to invest more of their AI capabilities in AI PCs, where the work can be done locally without being sent back to massive public models. Consider where your company is in this dynamic.

Trends

Given the substantial potential impact of AI on both the organization and the individual worker, we see the market for this technology as one of the fastest-growing sectors, with a 27% CAGR between 2021 and 2026. In that time span, IDC expects worldwide spending on AI to more than triple from just under $100 billion to just over $300 billion.

While IDC has yet to forecast the AI PC market ahead of launch, we expect it to steadily grow in share and become the most important technology and market driver for PCs moving forward.

Conclusion

AI has always been lauded for its massive potential on the human workforce. With the boom of generative AI, the massive potential energy of AI is starting to convert kinetically into interest, investments, and new real-world use cases. From 2021 to 2026, IDC expects AI investments to grow from less than $100 billion to more than $300 billion.

As AI use cases advance, expect to see AI engines proliferate from the cloud all the way down to the local device. Commercial PCs built with dedicated neural processors will allow their users to access real-time AI engines on the fly to complete their workflows. Human intelligence with AI at the fingertips – the future is now. ■

About the Analyst

Linn Huang, Research Vice President, Devices & Displays

Linn Huang tracks market trends and industry developments that impact the worldwide and U.S. markets for PCs, thin clients, and monitors. He participates in cross-research streams that cover all device categories

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

Get Ready for the Future of Work with Ryzen™ AI

AI solutions are changing the way work gets done. A dedicated AI engine, built right into the PC, empowers end users with fast performance, increased privacy, and immersive experiences.

Many individuals are now trialing artificial intelligence (AI) tools such as conversational AI and image generation AI applications. They are using these solutions not just to create songs, collate news, and ask questions, but also to perform work tasks like writing emails, generating programming code, summarizing notes, and analyzing images.

The excitement is understandable; AI chatbots for example can eliminate the drudgery of mundane daily work and instill the joy of creation. In addition, AI functionality provides immersive video conferencing experiences to improve collaboration and connection.

Businesses should tap into this enthusiasm for the innovative possibilities, as well as to increase productivity in today's world of hybrid work. But to do so, employees need PCs with AI technology baked in and that are built for power with long battery life, processing speed, and near-silent device operations. Anything less can slow down tasks and distract workers.

Read on to discover:

- Why AI technologies require a dedicated processing engine on the PC
- The capabilities that empower localized AI processing
- The gateway to unlock the power of AI in PCs
- How Ryzen™ AI prepares businesses for the new era of work

The need for a dedicated AI processing engine on the PC

Fifty-two percent of IT leaders are using or refining AI-related technologies such as machine learning (ML), natural language processing (NLP), and deep learning, according to the CIO Tech Poll: Tech Priorities 2023, and 32% are actively researching them.

Most companies recognize the unique processing requirements for AI workloads and turn to the cloud for power and resources. Yet, enterprises can gain value in a hybrid model of AI which deploys these solutions at the edge, on-premises, and in the cloud – and particularly on local devices.

The benefits of this approach include performance and cost efficiencies. For example, when AI workloads are processed on users' PCs, individuals can get access to fast data insights, which in turn can accelerate decision making.

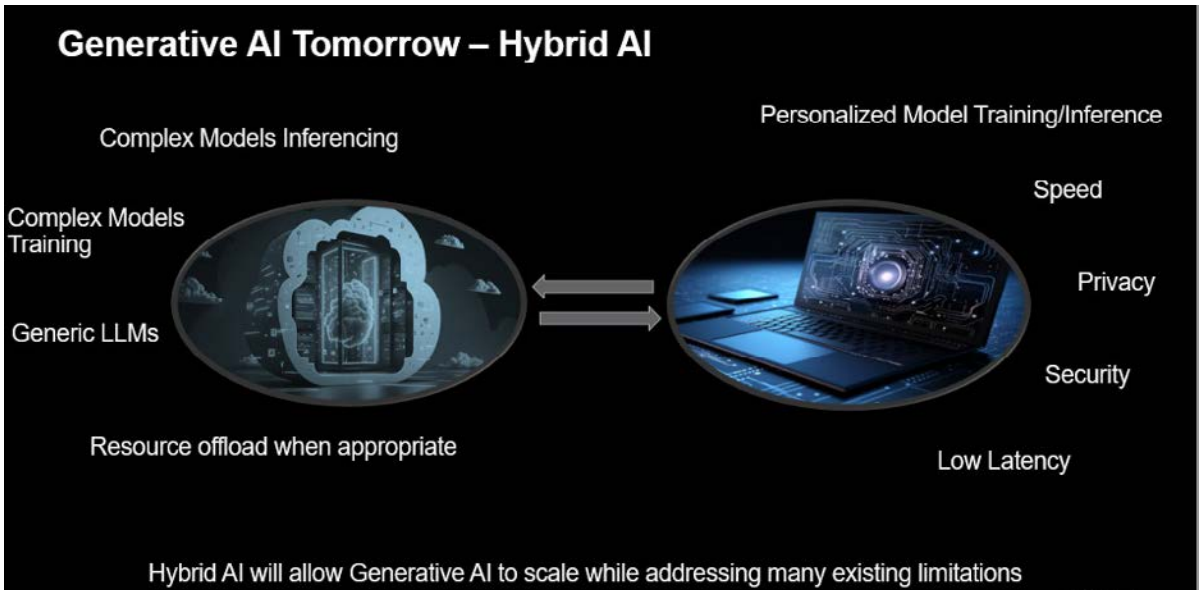
There's another significant benefit to a hybrid AI model: enhanced data privacy. Ensuring the privacy of data is a critical consideration for companies when it comes to responsible data management. As an illustration, a recent news report revealed that employees at a company unintentionally uploaded sensitive source code when leveraging an AI application. A dedicated AI engine on the PC mitigates the risks associated with data transfer by providing local processing of data.

Local AI processing offers other significant benefits:

- **Costs optimization.** Training and operating AI/ML models require massive compute power. Yet, cloud resources can become expensive, especially as AI workloads and projects scale. As any organization that has scaled cloud capacity knows, the costs quickly add up. By using local, dedicated AI processing power, enterprises can help optimize subscription costs – if not eliminate them entirely.

- **Personalization.** The consumerization of technology has demonstrated that individuals want to use devices in ways that best enhance their work. Enterprises can capitalize on AI enthusiasm by giving employees the opportunity to personalize these solutions on their PCs. Imagine the impact to productivity and innovation when workers can, for example, use an AI solution on their devices to create a presentation from their existing documents and their notes, including their tone of voice or personal style – without latency or PC performance issues.

In addition, the hybrid AI model allows for these solutions to scale where it most makes speed and cost sense (see *Generative AI Tomorrow*).



Required functionality: Capabilities that enhance local AI processing

So, what will it take for AI workload processing to efficiently and effectively occur on a user's PC? The answer: an optimized processing engine on the CPU, according to experts from AMD.

"Today AI processing typically demands GPU chips, specialized processors with multiple cores that accelerate performance. But that is changing," says Rakesh Anigundi, Director of Product Management at AMD.

In the same way the GPU has produced efficiencies for certain applications like graphics rendering and data analysis, an AI hardware accelerator integrated on the CPU can enhance local processing, Anigundi says. This AI-focused engine offsets the potential of any performance latency that users might experience when running AI/ML applications on their devices.

The AMD Ryzen™ AI brand means that the AMD Ryzen™ processor in your PC is uniquely capable of performing AI tasks. It means it includes a dedicated AI engine designed for the ultimate in AI processing efficiency; an AMD Radeon™ graphics engine optimized for AI workloads; and Ryzen™ processor cores that also have powerful AI capabilities. All three of these separate AI accelerators are part of an AMD Ryzen™ processor with AMD Ryzen™ AI.

In the future, an integrated AI engine right on the PC empowers new business opportunities for experiences that will necessitate high-powered processors, including:

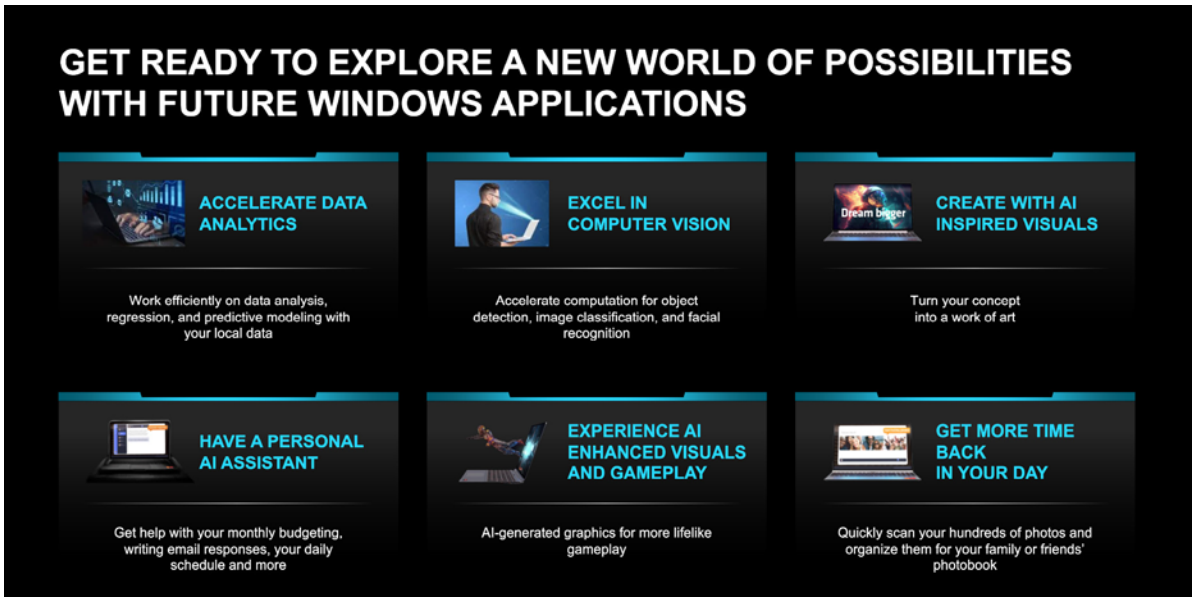
- **Intelligent assistants:** Individuals can customize a digital assistant to help them build presentations, write emails, manage

tasks and calendars, and summarize conversations or notes –helping to increase productivity and enabling employees to work creatively.

- **Content creation:** Fast, private AI imaging models on the PC can generate rich, visual content including video and avatars – for example, improving business presentations and marketing efforts, and providing immersive user experiences.

- **Advanced data analytics:** Rather than potentially expose confidential or sensitive data, employees can run data analysis and predictive modeling directly on their PCs. Not only does this functionality improve data privacy, it also can speed individual decision making.

- **AI threat detection and self-healing:** An integrated AI hardware accelerator on the PC can allow algorithms for anomaly detection to run independent of the CPU for greater performance. In addition, the dedicated AI engine is designed to isolate malware threats from the CPU for improved security.



The gateway to enhanced PC experiences

AMD is powering the future of AI with select Ryzen™ PRO 7040 series processors, which feature the world's first dedicated AI engine built right into the chip.[1] This integrated hardware accelerator is engineered specifically to run AI workloads for high performance, optimal efficiency, and low resource consumption.

"As AI applications reshape the work environment, AI PCs empowered by

AMD Ryzen™ AI are positioned to transform the productivity of business users," says Matt Unangst, Sr. Director of Product Management at AMD, "Ryzen™ AI aims to provide access to new productivity boosting AI apps and a great PC experience with incredible battery life, speed, and quiet operation."

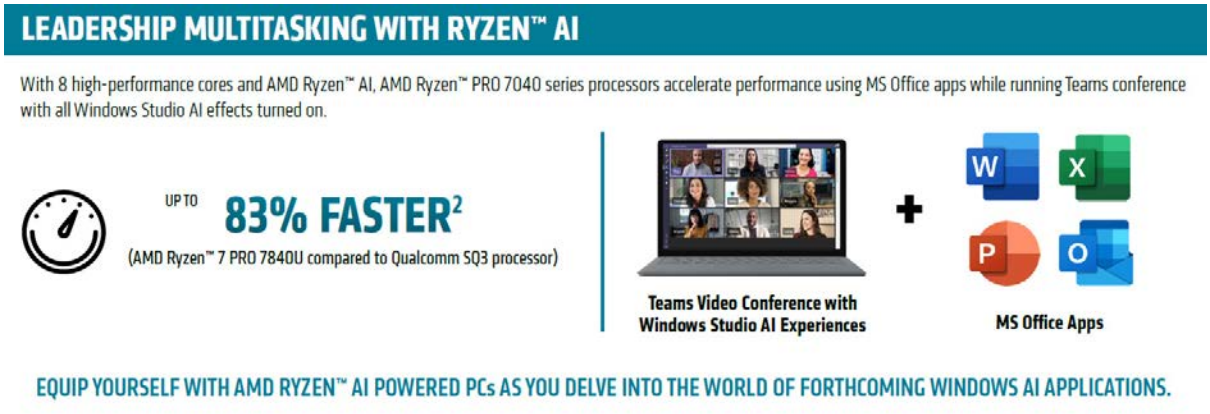
Today, a Ryzen™ AI engine provides advanced collaboration experiences, great power/battery life with near-silent operations ™ to power the future of work.

For example, select AMD Ryzen PRO 7040 series processors can accelerate productivity with Microsoft Office apps while running Teams conference calls, even with all the Windows Studio AI Effects turned on – such as eye-gaze correction, standard or portrait blur, and auto framing. Ryzen AI unburdens the CPU/GPU from running these applications.[2]

Ryzen AI is designed not only to help the user increase productivity and collaboration; the AI engine also improves device efficiencies such as overall performance when leveraging AI experiences. Here's evidence that Ryzen™ AI helps accelerate multi-tasking:

- Up to 83% faster compared with a Qualcomm SQ3 NPU processor based on testing in AMD Performance Labs in June 2023[2]
- Significant power savings vs NVIDIA Broadcast: "If you're doing a Zoom call for work and you want to use eye contact, you're looking at well less than half the power enabled by AMD Ryzen™ AI," according to PCWorld.

In addition, when AI-related tasks can be completed on a local device, data is exposed to fewer attack vectors.



Ryzen™ AI prepares businesses for the new era of work

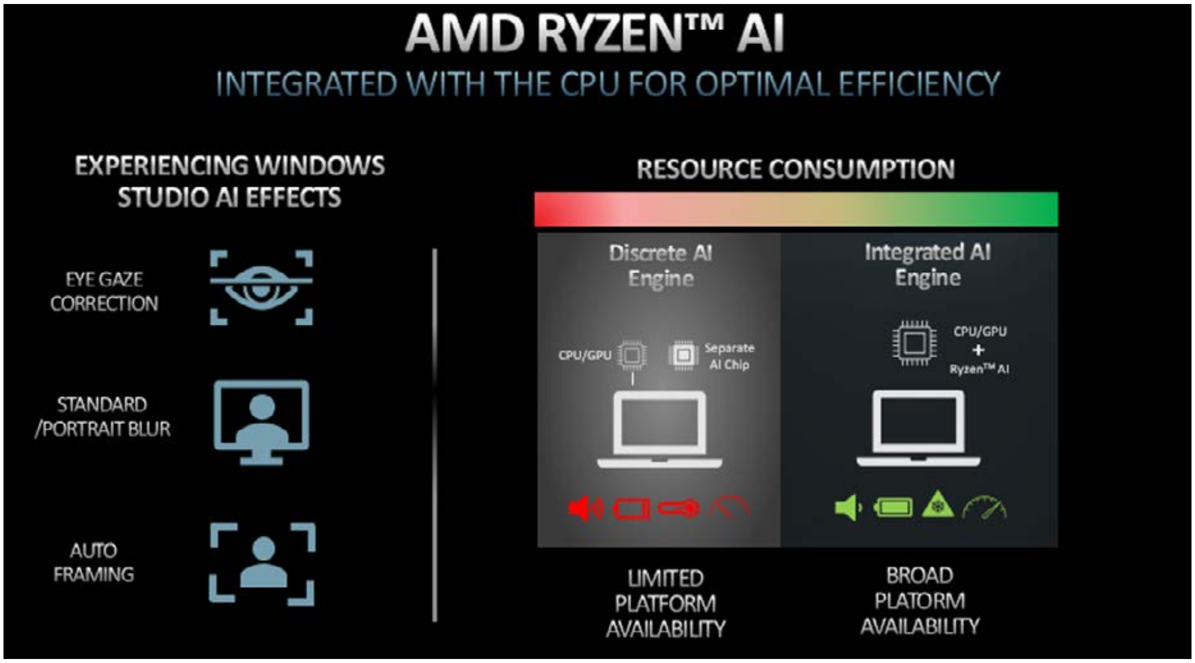
AMD is leading the way in AI chip innovation. Select Ryzen PRO 7040 series processors feature the first dedicated AI engine available in x86 based Windows OEM systems.[1]

Its innovative chip design offers significant advantages over discrete engines, including fast processing and leadership power efficiency. Because Ryzen AI is integrated with the CPU and shares the same power rail, it works in a complementary fashion to unburden the CPU – thus helping to extend battery life with optimal resource consumption versus platforms with a discrete AI chip.

"That translates to an improved user experience," Anigundi said. "For example, powering up the PC can happen faster because everything is synchronized. Ryzen AI is also very power efficient as it does not rely on an external chip for functionality, which can cause latency."

In addition, AI data processing occurs locally on the machine. This minimizes the attack surface, which translates to better device and data privacy.

AMD is committed to helping its partners, including ISVs and OEMs, accelerate and ease the roll out of AI functionality. Ryzen AI can handle the expanding world of AI workloads – even those not yet conceived – making now the right time to invest.



The bottom line: Step up to the biggest opportunity in AI

AI solutions are becoming fundamental to the way we work – as computers help us think, plan, and act. Yet, without the right power and capabilities, devices cannot take advantage of the transformative benefits of AI.

AMD Ryzen AI makes it possible for individuals to use their everyday laptops to delve into the future of work. And in the process, organizations can increase business value, including productivity, collaboration, and efficiencies.

[1] As of May 2023, AMD Ryzen AI was the first available dedicated AI engine on an x86 Windows processor, where 'dedicated AI engine' is defined as an AI engine that has no function other than to process AI inference models and is part of the x86 processor die. **PHX-3**.

[2] Testing as of 6/23/23, by AMD Performance Labs using the following benchmark tests: Procyon Overall, Procyon Word, Procyon Excel, Procyon PowerPoint, each while running a simulated 9-person (3:3) Microsoft Teams video conference call with utilizing system configuration for AMD Ryzen 7 7840U @15W TDP: MAYAN FP7-101DRC3INT-230331 (CRB), 16GB RAM, 1TB NVMe SSD, Integrated Radeon Graphics, Windows 11 Pro running in "high-performance mode," with Advanced Background Blur, eye gaze detection (using a mannequin to simulate the feature) and auto framing enabled via Ryzen AI. System configuration for Qualcomm SQ3 processor: Microsoft Surface Pro 9, 16GB RAM, 512GB NVMe SSD, Qualcomm integrated graphics, Windows 11 Pro running in "best performance mode," and Advanced Background Blur, eye gaze detection (using a mannequin to simulate the feature) and auto framing enabled via Qualcomm integrated NPU (Neural Processing Unit). System configurations may vary yielding different results. **PHXP-38**.