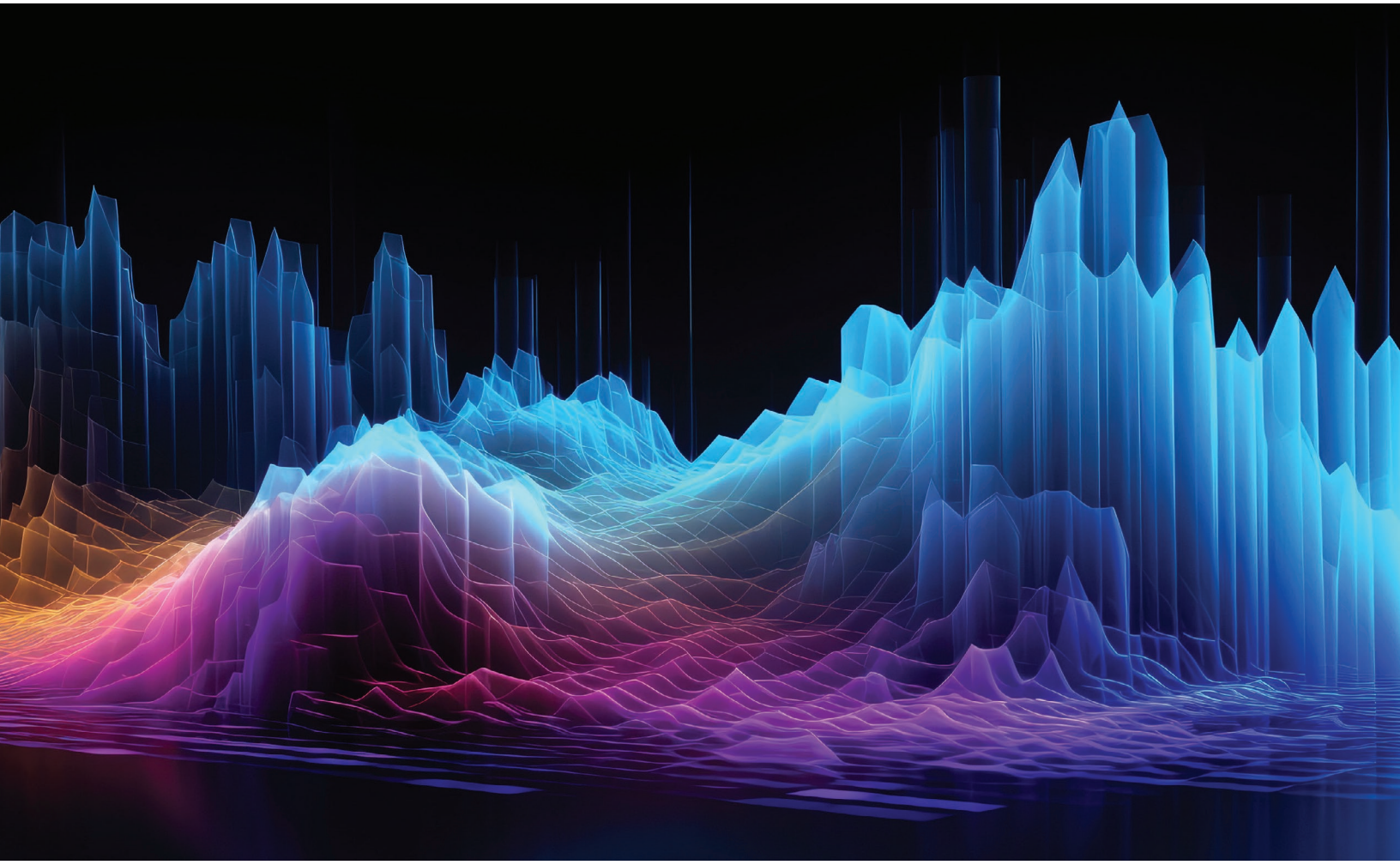
AMD
RYZEN AI
MAX PRO Series



HOW AI IS DRIVING EVOLUTION OF MOBILE, COMPACT WORKSTATION DESIGN

Historically, the vast majority of mobile workstations have relied on a discrete graphics card (dGPU) to provide sufficient performance for 3D rendering, video editing, and CAD/CAM projects. While this has been an effective strategy for some users, the additional power these cards consume has forced system manufacturers to trade away other features end-users value, including system thickness, lower operating temperatures, and increased battery life.

Long-term structural realignments in the workstation market are poised to make such tradeoffs even more expensive. AI models and the emerging applications that rely on them are placing new and unique demands on workstation manufacturers. The widespread adoption of 4K and 8K video in editing booths, the growing use of building information modeling (BIM) in engineering, and the steady shift towards photorealistic rendering have put upward pressure on GPU requirements as well. These pressures run counter to the general consumer preference for lighter, faster, and more power-efficient systems.

THE EXPANDING WORKSTATION MARKET

Mobile workstation sales are expected to grow over the next five years, with IDC predicting that one in 10 commercial systems will be a workstation by 2030 "as more organizations recognize their value for mission-critical workloads." According to Jay Chou, research manager for IDC's Worldwide Client Device Trackers, "Current exploration around AI for work should translate into an expansion of use cases, including model development, across numerous industries[1]."

While workstation use cases may be expanding, system footprints aren't necessarily headed in the same direction. Mobile workstation adoption surged during the pandemic as working conditions changed, and while tower workstations remain essential to the industry, laptops and small form factor (SFF) desktop workstations are increasingly important.

Enterprise workstation customers are looking for systems designed with minimal physical footprints and maximum dataflow. This simultaneous shift in form factor and use case is both a design challenge and an opportunity to redefine what enterprises expect from a mobile, compact workstation.

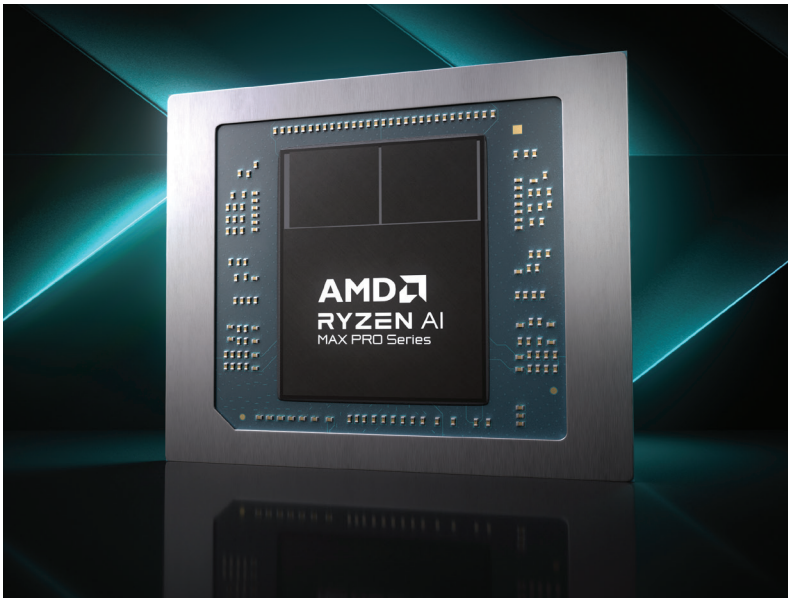
CHALLENGING THE TRADE-OFFS WORKSTATION CUSTOMERS ARE FORCED TO ACCEPT

Too often, mobile and compact workstations are designed in a way that forces customers into binary choices between two desirable product options or outcomes. Desktop workstations are less constrained, thanks to their large internal volumes and aggressive cooling systems.

This dichotomy forces mobile and compact workstation users to compromise on features and performance. It's a view that's out of step with both larger trends around system form factors and AI's increased computational requirements. There's a new way to meet the computing demands of today – and a better solution to meet them with.

INTRODUCING AMD RYZEN™ AI MAX PRO:

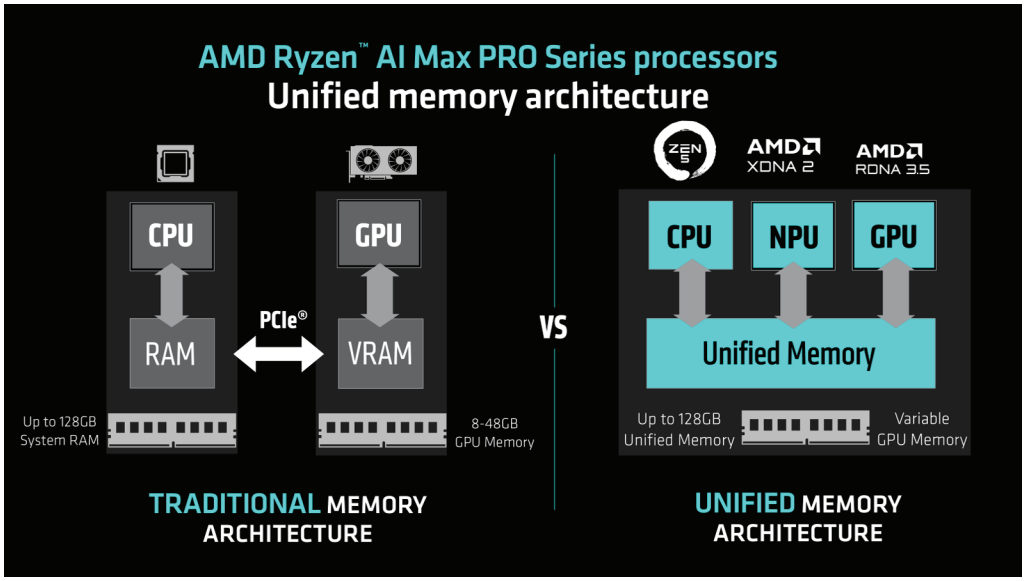
The AMD Ryzen™ AI Max PRO Series of processors represent a significant milestone for x86 systems and the larger Windows workstation market. As the first x86 processors to combine an integrated GPU with discrete-level performance, desktop-class CPU cores, and a neural processing unit (NPU) in a single chip, they are tailored for customers who want excellent performance, professional application optimizations and certifications, and the ability to run AI workloads that would otherwise be too heavy for most of the dGPUs found in commercial notebooks and SFF desktops commonly shipping today. AMD Ryzen AI Max PRO Series processors are designed to tackle complex 3D projects with multiple applications running in parallel or to explore new ideas using local large language models.



ELIMINATING GPU MEMORY BOTTLENECKS

Modern mobile discrete graphics cards typically offer between 8-16GB of dedicated video RAM (VRAM). While this is sufficient for many workstation applications, running workloads with large datasets and AI models locally presents a greater challenge. For example, Stable Diffusion 3.5 Large can exceed even the 16GB VRAM capacity typically available on higher-end mobile GPUs. Customers must either use less complex versions of the model that *can* fit within the limited memory available or buy time on cloud services.

AMD Ryzen AI Max PRO Series processors address this problem by sharing a common pool of memory between the CPU, GPU, and NPU, as shown below:



This type of sharing is known as a unified memory architecture (UMA), as opposed to a traditional distributed memory architecture (shown on the left), where the CPU and GPU each have their own dedicated pools of memory.

A traditional memory architecture provides the GPU with a great deal of bandwidth, but much less total memory compared to the CPU. It also makes communication between the CPU and GPU slower. Connecting these two components together on-die, as shown on the right, allows them to share a common pool of memory. This is advantageous, provided the system offers enough memory bandwidth. AMD Ryzen AI Max PRO Series processors do.

Where most other x86 mobile workstation processors rely on two memory channels, the AMD Ryzen AI Max PRO Series uses four. The resulting boost to total system bandwidth is enough to feed the GPU, CPU, and NPU simultaneously. AMD Ryzen AI Max PRO Series processors additionally include up to 32MB of MALL (Memory Attached Last Level) cache to amplify graphics bandwidth and keep the discrete-class graphics core fed.

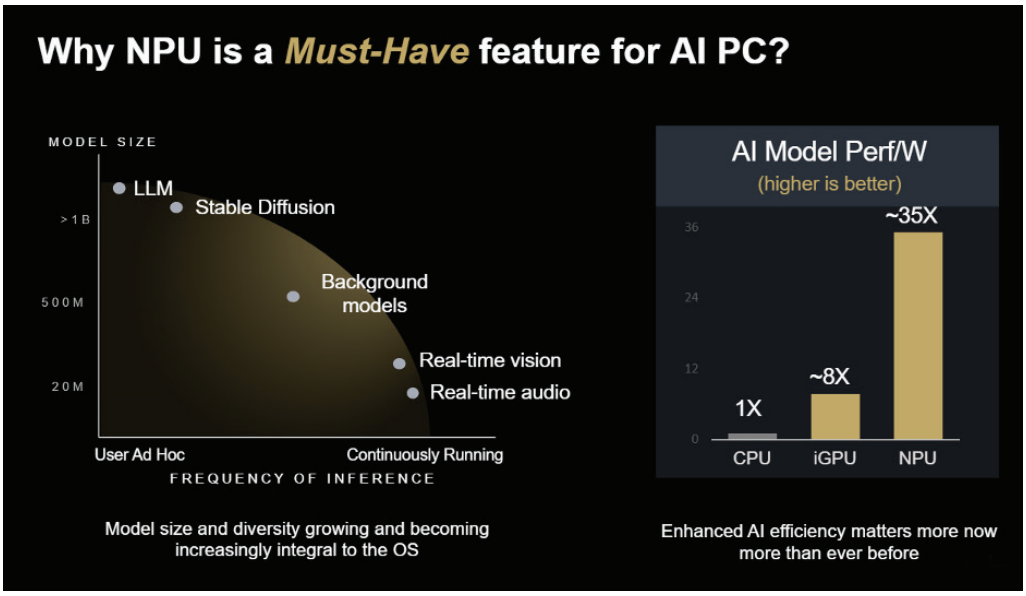
AMD built the Ryzen AI Max PRO Series this way to take advantage of the power consumption and efficiency advantages of die-level integration while still offering a powerful discrete-equivalent GPU. Additionally, up to 96GB of the 128GB maximum total available system memory can be dedicated to graphics processing – far more than is available on any dGPU today.

This enormous memory buffer has particularly interesting implications for AI. As previously discussed, many consumer and commercial graphics cards can only load older or simplified models that have been trimmed to fit within the 8-16GB frame buffers commonly available in mobile and mini desktop systems. AMD Ryzen AI Max PRO Series processors, in contrast, can offer enough GPU memory to run models like Stable Diffusion 3.5 Large or Llama 3.1 70B-Q4 locally. Inference workloads that would otherwise be out of reach on any other system, mobile or desktop, can run on an AMD Ryzen AI Max PRO system.

AN NPU FOR EMERGING AI WORKLOADS

To date, most AI workloads have targeted either the CPU or GPU. Neural processing units, or NPUs, are an emerging processor type first introduced to the x86 PC by AMD in 2023. NPU performance has increased rapidly, from 10 TOPS on the first AMD Ryzen PRO 7040 Series processors to 50 TOPS with the Ryzen AI Max PRO processors available today.

NPU software support is expected to grow as AI PCs proliferate and ISVs become more familiar with their strengths and capabilities. The efficiency potential they offer compared to CPUs or traditional integrated GPUs make them an attractive optimization target, as shown below:



The decision to include an NPU in the AMD Ryzen AI Max PRO Series reflect where AI workloads are likely to run in the future, while the discrete-level GPU and desktop-equivalent CPU cores are designed for the conventional and AI-centric applications enterprises are using today. As applications move to the NPU, they'll both unlock additional power efficiency and free the CPU and GPU to focus on other tasks.

CONCLUSION

Enterprises need workstations that move beyond the traditional mobile/tower dichotomy that limits the workloads and scenarios that mobile and compact workstations could plausibly address. Workstation users, from developers to creatives, are likely to encounter AI at both the OS and the application level as businesses transition to Windows 11 and ISVs integrate artificial intelligence into existing photo and video editing tools, content management platforms, knowledge bases, and office suites.

AMD Ryzen AI Max PRO Series processors are ideal for companies and end users who want to work with larger project assemblies, tackle more complex, AI-accelerated projects, and develop new LLM-based applications locally. AMD Ryzen AI MAX Pro Series processors include powerful ISV-certified graphics, support AMD PRO Technologies security and reliability features, and leverage the intrinsic advantages of integration to push the boundaries of what's possible in a mobile, compact workstation.

FOOTNOTES

1. See IDC Global Shipments of PC Workstations Shrank Nearly 9% in 2023, but Recovery Expected as Several Market Drivers Coalesce in 2024, According to IDC, 13 March, 2024

DISCLAIMERS

COPYRIGHT NOTICE

C2954271-0225