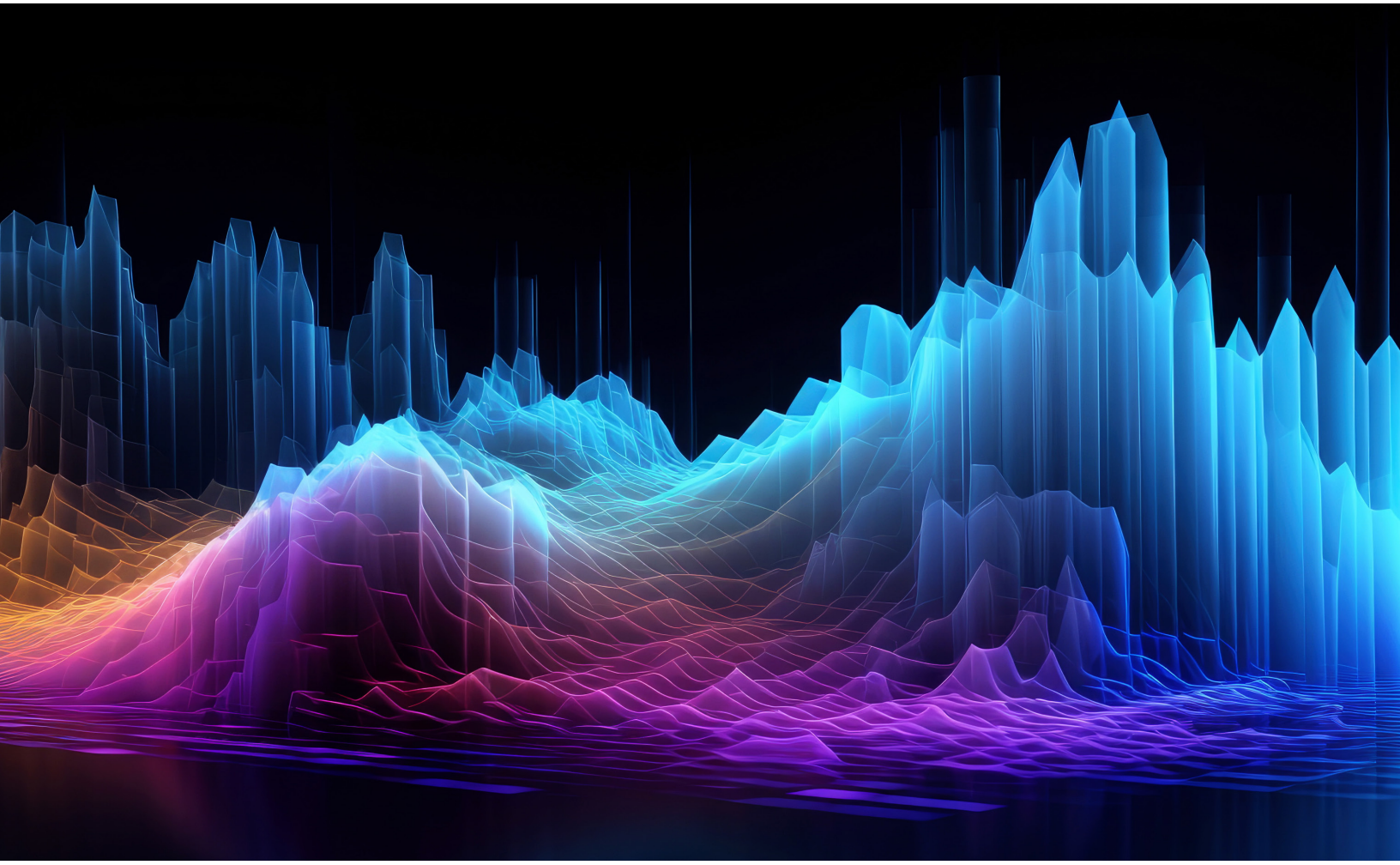
AMD
RYZEN AI
MAX PRO Series



WIE KI DIE ENTWICKLUNG MOBILER, KOMPAKTER WORKSTATIONS VORANTREIBT

In der Vergangenheit beruhte der Großteil mobiler Workstations auf einer diskreten Grafikkarte (dGPU), um ausreichend Performance für 3D-Rendering, Videobearbeitung und CAD-/CAM-Projekte bereitzustellen. Obwohl dies eine effektive Strategie für einige Benutzer war, hat der zusätzliche Stromverbrauch dieser Karten die Systemhersteller dazu gezwungen, bei anderen von den Benutzern geschätzten Funktionen, wie z. B. Systemstärke, niedrigere Betriebstemperaturen und längere Akkulaufzeit, Kompromisse einzugehen.

Langfristige strukturelle Neuausrichtungen auf dem Workstation-Markt werden solche Kompromisse noch teurer gestalten. KI-Modelle und neue Anwendungen, die auf diesen beruhen, stellen neue und spezifische Anforderungen an Workstation-Hersteller. Die verbreitete Einführung von 4K- und 8K-Video in Schneideräumen, die zunehmende Nutzung von Building Information Modeling (BIM) im Engineering und die konstante Verlagerung in Richtung des fotorealistischen Renderings üben bei den GPU-Anforderungen zusätzlich mehr Druck aus. Diese Belastungen stehen im Widerspruch dazu, dass sich Verbraucher leichtere, schnellere und energieeffizientere Systeme wünschen.

DER EXPANDIERENDE WORKSTATION-MARKT

Es wird erwartet, dass der Verkauf von mobilen Workstations in den nächsten fünf Jahren steigen wird. IDC prognostiziert, dass bis 2030 eines von 10 kommerziellen Systemen eine Workstation sein wird, „da mehr Organisationen deren Wert für unternehmenskritische Auslastungen erkennen werden". Jay Chou, Research Manager für Worldwide Client Device Trackers von IDC, sagt hierzu: „Die aktuelle Erkundung rund um KI für die Arbeit sollte zu einer Erweiterung der Anwendungsfälle, einschließlich Modellentwicklung, über zahlreiche Branchen hinweg führen."[1]

Auch wenn es immer mehr Anwendungsfälle für Workstations gibt, gilt dies nicht notwendigerweise für den Platzbedarf von Systemen. Die Einführung mobiler Workstations stieg während der Coronapandemie stark an, da sich die Arbeitsbedingungen änderten. Und während Tower-Workstations für die Branche unverzichtbar bleiben, gewinnen Notebooks und Desktop-Workstations mit kleiner Bauform zunehmend an Bedeutung.

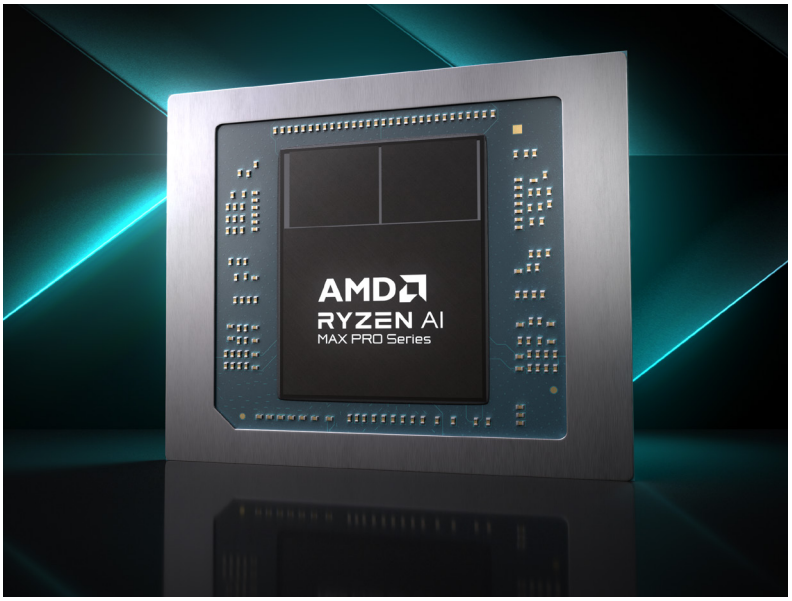
Kunden von Unternehmens-Workstations sind auf der Suche nach Systemen mit minimaler physischer Stellfläche und maximalem Datenfluss. Diese gleichzeitige Verschiebung bei Bauform und Anwendungsfall ist sowohl eine Herausforderung für das Konzept als auch eine Chance, neu zu definieren, was Unternehmen von einer mobilen, kompakten Workstation erwarten.

HINTERFRAGEN DER KOMPROMISSE, DIE WORKSTATION-KUNDEN AKZEPTIEREN MÜSSEN

Viel zu oft sind mobile und kompakte Workstations so konzipiert, dass sie Kunden zu binären Entscheidungen zwischen zwei wünschenswerten Produktoptionen oder -ergebnissen zwingen. Bei Desktop-Workstations gibt es weniger Einschränkungen, da sie große interne Kapazitäten und leistungsstarke Kühlsysteme haben.

Diese Gegensätzlichkeit zwingt Benutzer mobiler und kompakter Workstations dazu, in puncto Funktionen und Performance Kompromisse einzugehen. Dies ist weder mit größeren Trends rund um Systembauformen noch mit den steigenden Rechenanforderungen von KI in Einklang. Es gibt eine neue Möglichkeit, die Rechenanforderungen von heute zu erfüllen – und eine bessere Lösung, um diese zu meistern.

WIR STELLEN VOR: AMD RYZEN™ AI MAX PRO:

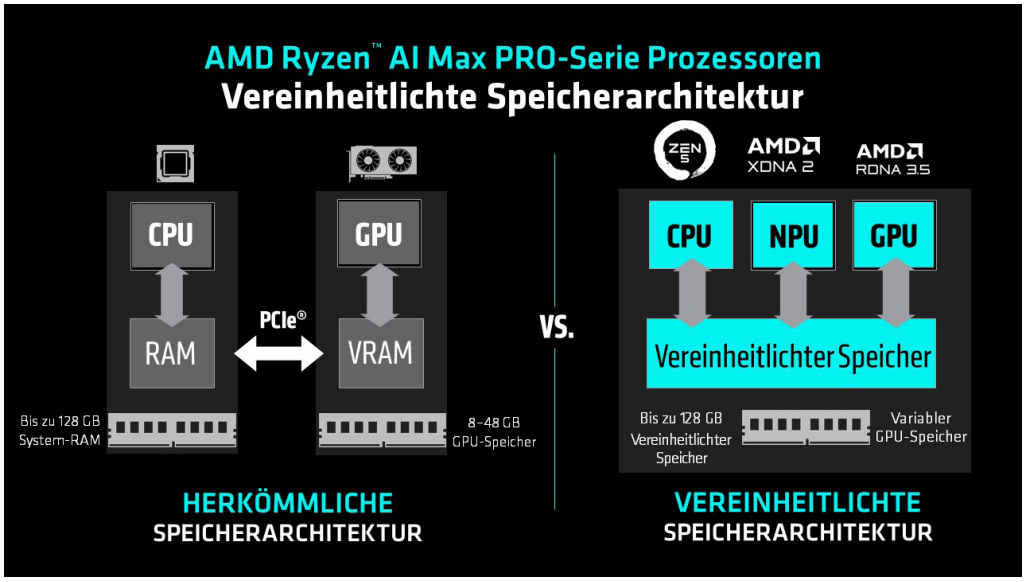


AMD Ryzen™ AI Max PRO-Serie Prozessoren repräsentieren einen wichtigen Meilenstein für x86-Systeme und den größeren Windows Workstation-Markt. Sie sind die ersten x86-Prozessoren, die eine integrierte GPU mit der Performance einer diskreten GPU, CPU-Kernen der Desktop-Klasse und einer Neural Processing Unit (NPU) in einem einzigen Chip kombinieren. Als solche sind sie für Kunden konzipiert, die sich herausragende Performance, professionelle Anwendungsoptimierungen und -zertifizierungen sowie die Möglichkeit zur Ausführung von KI-Auslastungen wünschen, die für die meisten dGPUs in derzeit bereitgestellten kommerziellen Notebooks und Desktops mit kleiner Bauform zu hoch sind. AMD Ryzen AI Max PRO-Serie Prozessoren sind darauf ausgelegt, komplexe 3D-Projekte, bei denen mehrere Anwendungen parallel ausgeführt werden, zu meistern bzw. mit lokalen Large Language Models neue Ideen zu erkunden.

ENGPÄSSE IM GPU-SPEICHER BEHEBEN

Bei modernen mobilen diskreten Grafikkarten werden normalerweise 8 bis 16 GB dedizierter Video-RAM (VRAM) bereitgestellt. Für viele Workstation-Anwendungen ist dies meist ausreichend. Bei der Ausführung von Auslastungen mit großen Datensätzen und KI-Modellen auf lokaler Ebene stellt dies allerdings eine größere Herausforderung dar. Stable Diffusion 3.5 Large z. B. kann selbst die VRAM-Kapazität von 16 GB, die normalerweise auf leistungsfähigeren mobilen GPUs verfügbar ist, überschreiten. Kunden müssen entweder weniger komplexe Versionen des Modells verwenden, die in den begrenzten verfügbaren Speicher *passen*, oder Zeit für Cloud-Services kaufen.

AMD Ryzen AI Max PRO-Serie Prozessoren gehen auf dieses Problem ein, indem sie einen Speicherpool für CPU, GPU und NPU gemeinsam nutzen. Dies ist in der folgenden Abbildung dargestellt:



Diese Art der gemeinsamen Nutzung wird als vereinheitlichte Speicherarchitektur bezeichnet. Sie steht im Gegensatz zu einer herkömmlichen verteilten Speicherarchitektur (links dargestellt), bei der CPU und GPU jeweils einen eigenen dedizierten Speicherpool haben.

In einer herkömmlichen Speicherarchitektur wird der GPU eine hohe Bandbreite zur Verfügung gestellt, allerdings wesentlich weniger Gesamtspeicher im Vergleich zur CPU. Außerdem ist hier die Kommunikation zwischen CPU und GPU langsamer. Wenn diese beiden Komponenten wie auf der rechten Seite dargestellt mit einem On-Die-Ansatz miteinander verbunden werden, können sie einen Speicherpool gemeinsam nutzen. Dies ist vorteilhaft, sofern das System ausreichend Speicherbandbreite zur Verfügung stellt. Bei AMD Ryzen AI Max PRO-Serie Prozessoren ist genau das der Fall.

Während die meisten anderen mobilen x86-Workstation-Prozessoren auf zwei Speicherkanälen basieren, nutzt die AMD Ryzen AI Max PRO-Serie vier. Die daraus resultierende Gesamtbandbreite des Systems reicht aus, um GPU, CPU und NPU gleichzeitig zu versorgen. AMD Ryzen AI Max PRO-Serie Prozessoren haben zusätzlich bis zu 32 MB MALL-Cache (Memory Attached Last Level), um die Grafikbandbreite zu erhöhen und den Kern der Grafikeinheit auf dem Niveau diskreter Modelle zu versorgen.

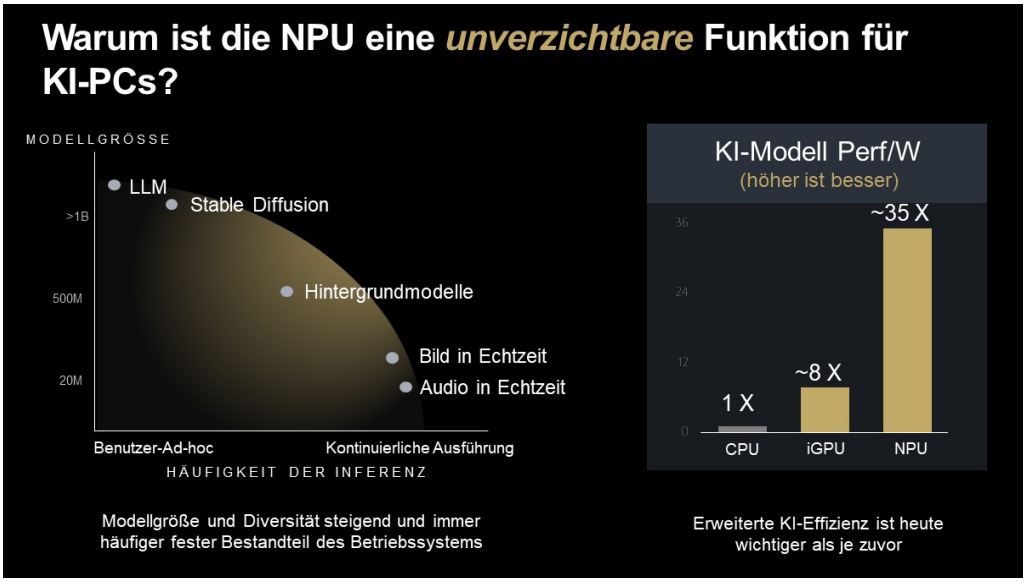
AMD hat die Ryzen AI Max PRO-Serie so konzipiert, um die Vorteile des Stromverbrauchs und der Effizienz der Integration auf Die-Ebene zu nutzen und gleichzeitig eine leistungsstarke GPU bereitzustellen, die äquivalent zu einem diskreten Modell ist. Außerdem können bis zu 96 GB des maximal verfügbaren Gesamtsystemspeichers von 128 GB für die Grafikverarbeitung verwendet werden – weit mehr als bei allen anderen momentan verfügbaren dGPUs.

Dieser riesige Speicherpuffer hat besonders interessante Auswirkungen für KI. Wie bereits erwähnt, können viele Verbraucher- und kommerzielle Grafikkarten nur ältere oder vereinfachte Modelle laden, die so modifiziert wurden, dass sie in die 8 bis 16 GB großen Frame-Puffer passen, die häufig in mobilen und Mini-Desktop-Systemen verfügbar sind. AMD Ryzen AI Max PRO-Serie Prozessoren können im Gegensatz dazu ausreichend GPU-Speicher bieten, um Modelle wie Stable Diffusion 3.5 Large oder Llama 3.1 70B-Q4 lokal auszuführen. Inferenzauslastungen, die sonst auf jedem anderen System, egal, ob mobil oder Desktop, nicht möglich wären, können auf einem AMD Ryzen AI Max PRO System ausgeführt werden.

EINE NPU FÜR NEUE KI-AUSLASTUNGEN

Bislang sind die meisten KI-Auslastungen entweder auf die CPU oder die GPU ausgerichtet. Neural Processing Units, kurz NPUs, sind ein neuer Prozessortyp, der 2023 erstmals von AMD auf einem x86-PC eingeführt wurde. Die NPU-Performance hat sich schnell erhöht: von 10 TOPS bei den ersten AMD Ryzen PRO 7040-Serie Prozessoren auf 50 TOPS bei den derzeit verfügbaren Ryzen AI Max PRO Prozessoren.

Es wird erwartet, dass die NPU-Softwareunterstützung weiter wächst, wenn KI-PCs häufiger eingesetzt werden und die ISVs deren Stärken und Möglichkeiten besser kennenlernen. Das Effizienzpotenzial, das sie im Vergleich zu CPUs oder herkömmlichen integrierten GPUs bieten, macht sie wie unten dargestellt zu einem attraktiven Optimierungsziel:



Die Entscheidung, eine NPU in die AMD Ryzen AI Max PRO-Serie einzubinden, spiegelt wider, wo KI-Auslastungen in Zukunft wahrscheinlich ausgeführt werden. Die GPU auf dem Niveau diskreter Modelle und die Desktop-äquivalenten CPU-Kerne sind wiederum für konventionelle und KI-orientierte Anwendungen konzipiert, die Unternehmen heute nutzen. Wenn Anwendungen auf die NPU verlagert werden, setzen sie zusätzliche Energieeffizienz frei und geben die CPU und GPU für andere Aufgaben frei.

FAZIT

Unternehmen benötigen Workstations, die über die herkömmliche Zweiteilung in Mobil- und Tower-Lösungen hinausgehen, die die Auslastungen und Szenarien, die für mobile und kompakte Workstations infrage kommen, einschränkt. Workstation-Benutzer, von Entwicklern bis zu kreativen Profis, werden wahrscheinlich sowohl auf Betriebssystem- als auch auf Anwendungsebene auf KI stoßen, wenn Unternehmen auf Windows 11 umsteigen und ISVs künstliche Intelligenz in bestehende Foto- und Videobearbeitungs-Tools, Content-Management-Plattformen, Wissensdatenbanken und Office Suites integrieren.

AMD Ryzen AI Max PRO-Serie Prozessoren sind ideal für Unternehmen und Endbenutzer, die mit größeren Projektbaugruppen arbeiten, komplexere, KI-beschleunigte Projekte in Angriff nehmen und neue LLM-basierte Anwendungen lokal entwickeln möchten. AMD Ryzen AI Max PRO-Serie Prozessoren beinhalten leistungsstarke ISV-zertifizierte Grafikeinheiten, unterstützen Sicherheits- und Zuverlässigkeitsfunktionen der AMD PRO Technologien und nutzen die spezifischen Vorteile der Integration, um die Grenzen dessen, was mit einer mobilen, kompakten Workstation möglich ist, neu zu definieren.

FUSSNOTEN

1. Siehe „Global Shipments of PC Workstations Shrank Nearly 9% in 2023, but Recovery Expected as Several Market Drivers Coalesce in 2024, According to IDC", vom 13. März 2024