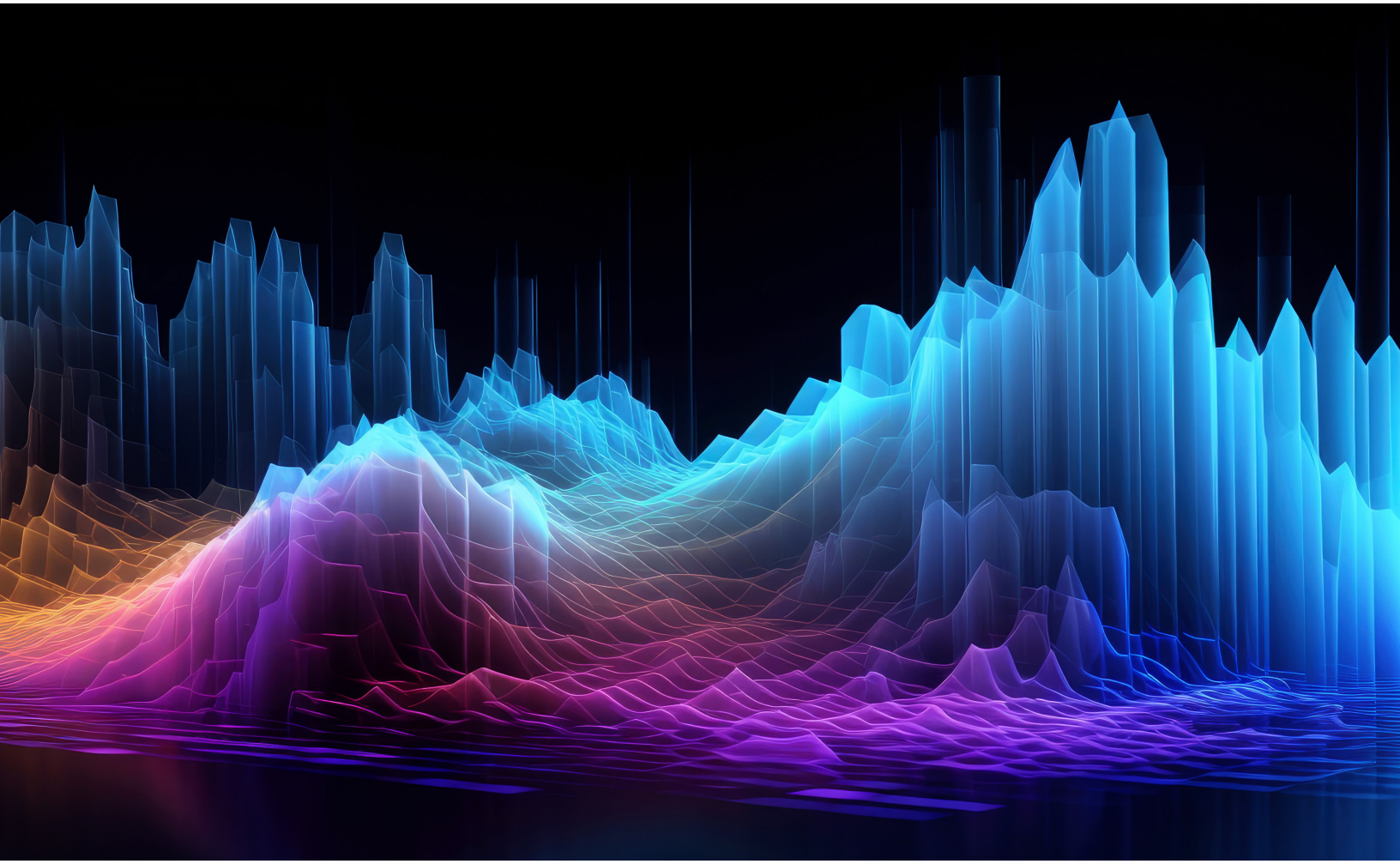
AMD
RYZEN AI
MAX PRO Series



AI 如何推动移动紧凑型工作站设计的发展演变

过去，绝大多数移动工作站都采用性能较高的独立显卡 (dGPU) 来处理 3D 渲染、视频编辑和 CAD/CAM 项目。虽然这种策略对于一些用户行之有效，但独立显卡的功耗更高，从而迫使系统制造商放弃最终用户重视的一些其他功能特性，包括机身厚度、更低的运行温度和更持久的电池续航能力。

工作站市场的长期结构性调整无疑会使这种权衡取舍所付出的代价变得更加昂贵。AI 模型及依赖这些模型的新兴应用对工作站制造商提出了全新的特殊要求。4K 和 8K 视频在编辑工作室中广泛采用，建筑信息建模 (BIM) 技术在工程领域越来越多地应用，照片级真实感渲染技术稳步崛起，这些都对 GPU 提出了更高的要求。但是，与此背道而驰的是，消费者普遍青睐于更轻、更快、更节能的系统。

不断拓展的工作站市场

移动工作站销量预计将在未来五年内持续增长，IDC 预测，到 2030 年，“随着越来越多的组织认识到工作站对于任务关键型工作负载的价值”，工作站在商用系统中的占比将达到十分之一。负责 IDC 全球客户端设备跟踪报告的调研经理 Jay Chou 表示：“目前围绕 AI 的探索工作应该会使 AI 在各个行业中得到更广泛的使用，包括用于进行模型开发[1]。”

虽然工作站的应用场景可能在不断扩大，但系统占用空间不一定如此。疫情期间，随着工作环境发生改变，移动工作站的采用率出现激增。虽然塔式工作站依然是整个行业的中流砥柱，但笔记本电脑和小型化 (SFF) 台式工作站变得日益重要。

企业级工作站客户希望设计的系统做到两点：物理占用空间尽可能小，数据流尽可能大。外形规格和应用场景同时发生转变，这既带来了设计上的严峻挑战，也带来了重新定义企业对于移动紧凑型工作站期望的宝贵机会。

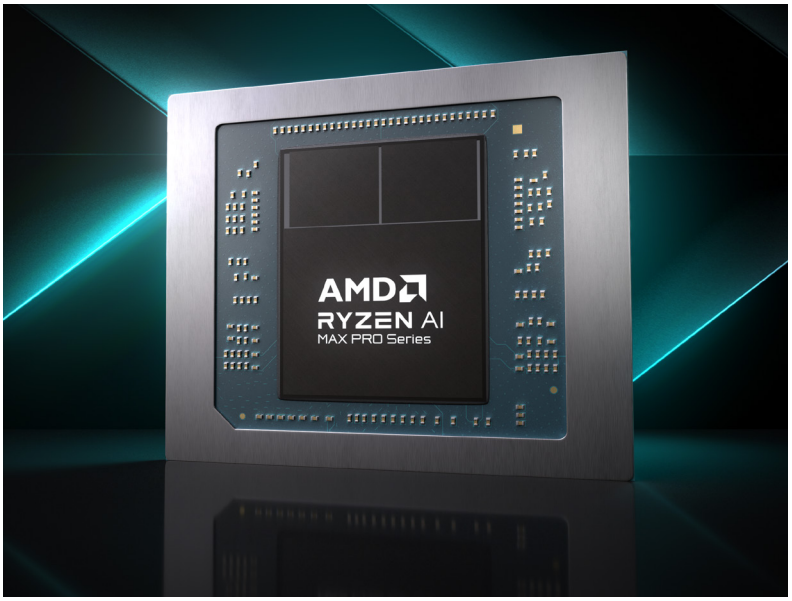
挑战常规，让工作站客户无需权衡取舍

传统设计通常要求客户在移动紧凑型工作站与台式工作站这两种产品或预期结果之间做出二选一的抉择。一方面，台式工作站由于较大的内部空间和强大的冷却系统，性能和功能更不受限制。

另一方面，移动紧凑型工作站用户不得不在功能和性能方面进行妥协。这种二择一的观点与当前的大趋势并不同步，因为当前系统规格在不断缩小且 AI 的计算需求在不断增大。如今，可以通过一种全新方式以及更好的解决方案来满足计算需求。

AMD 锐龙 AI MAX PRO 重磅来袭：

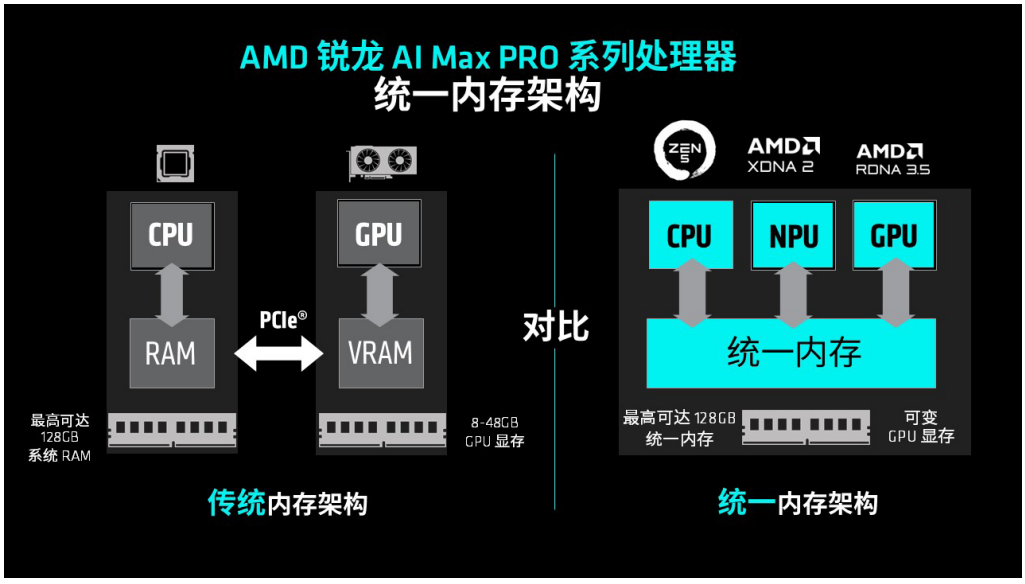
AMD 锐龙 AI Max PRO 系列处理器是 x86 系统以及更广泛的 Windows 工作站市场的一个重要里程碑。该系列处理器开创 x86 系统先河，率先将性能堪比独立显卡的集成 GPU、台式机级 CPU 核心以及神经处理单元 (NPU) 整合到一个芯片之中。如果需要获得出色性能以及专业应用优化和认证，并且需要能够运行当前常见商用笔记本电脑和 SFF 台式机中采用的大多数 dGPU 所无法承受的 AI 工作负载，该系列处理器是合适之选。AMD 锐龙 AI Max PRO 系列处理器经过精心设计，能够处理需要并行运行多个应用的复杂 3D 项目，还能够使用本地大语言模型探索创新想法。



消除 GPU 显存瓶颈

现代移动独立显卡通常可提供 8GB 到 16GB 的专用显存 (VRAM)。虽然这足以满足许多工作站应用的需求，但是要在本地运行带大型数据集的工作负载和 AI 模型，显然会给显存带来更大挑战。例如，Stable Diffusion 3.5 Large 的需求甚至可能超出高端移动 GPU 通常可达到的 16GB VRAM 容量。为此，客户必须使用复杂度较低的模型版本以便*能够*适应有限的可用显存，或者购买云服务。

为解决此问题，AMD 锐龙 AI Max PRO 系列处理器在 CPU、GPU 和 NPU 之间共享一个通用内存池，如下图所示：



此类共享称为统一内存架构 (UMA)。这种架构不同于传统的分布式内存架构（如左图所示），其中 CPU 和 GPU 各自拥有自己的专用内存池。

传统的内存架构为 GPU 提供了超大带宽，但总内存远远少于 CPU。在传统架构下，CPU 和 GPU 之间的通信速度更慢。如右图所示，通过在片上整合这两个组件，即可让它们共享一个通用内存池。如果系统能够提供足够的内存带宽，那么这种架构将具有显著优势。AMD 锐龙 AI Max PRO 系列处理器能够满足这一点。

大多数 x86 移动工作站处理器使用两个内存通道，而 AMD 锐龙 AI Max PRO 系列处理器使用四个通道。因此，该系列处理器可带来更高的总系统带宽，足以同时支持 GPU、CPU 和 NPU。AMD 锐龙 AI Max PRO 系列处理器还可提供最高可达 32MB 的 MALL (Memory Attached Last Level) 高速缓存，能够增加显卡带宽并支持堪比独立显卡的内核性能。

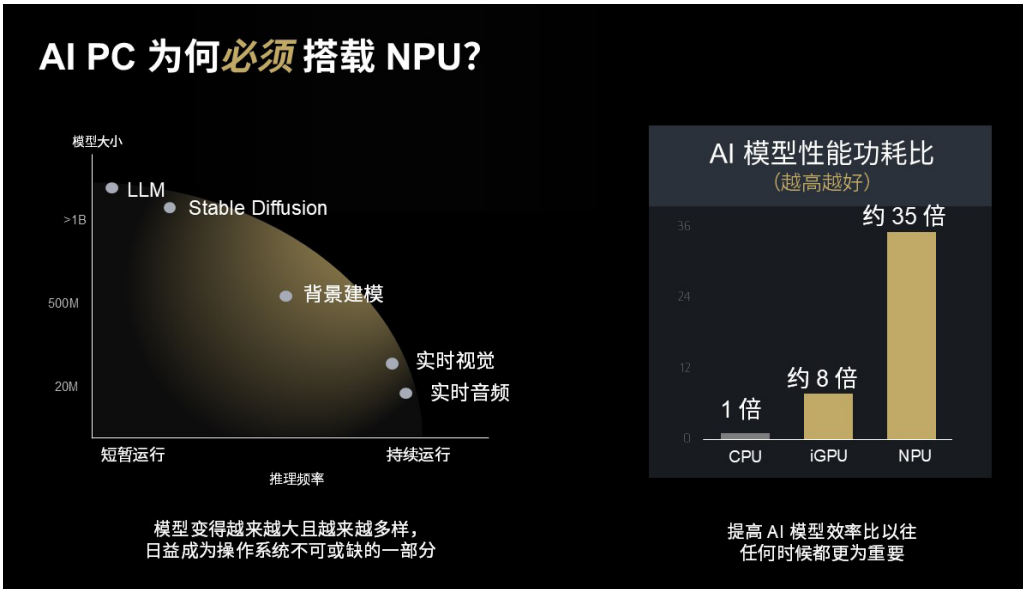
AMD 以这种方式打造了锐龙 AI Max PRO 系列，旨在充分发挥芯片级集成的功耗和效率优势，同时依然提供媲美独立显卡的强大 GPU。此外，在最高可达 128GB 的总可用系统内存中，多达 96GB 内存可专用于图形处理，远远超出当今普遍 dGPU 上的可用显存。

这一巨大的存储缓冲区对 AI 的影响尤其值得关注。如前所述，许多消费类和商用显卡只能加载版本较低或较简单的模型，这些模型经过精简处理以适应移动和迷你台式机系统中常见的 8GB 到 16GB 帧缓冲区。相较之下，AMD 锐龙 AI Max PRO 系列处理器可以提供充足的 GPU 显存，支持在本地运行 Stable Diffusion 3.5 Large 或 Llama 3.1 70B-Q4 等模型。其他移动或台式机系统无法承受的推理工作负载，都可以在 AMD 锐龙 AI Max PRO 系统上运行。

专为新兴 AI 工作负载打造的 NPU

迄今为止，大多数 AI 工作负载要么由 CPU 处理，要么由 GPU 处理。神经处理单元 (NPU) 是一种新兴的处理器，由 AMD 于 2023 年创新引入 x86 PC。NPU 性能飞速提升，从第一代 AMD 锐龙 PRO 7040 系列处理器的 10 TOPS 提升至现今锐龙 AI Max PRO 处理器的 50 TOPS。

随着 AI PC 数量激增以及 ISV 对其优势和能力的进一步了解，NPU 软件支持范围预计将持续扩大。与 CPU 或传统集成 GPU 相比，NPU 在能效方面的巨大潜力使其成为富有吸引力的优化目标，如下图所示：



将 NPU 引入 AMD 锐龙 AI Max PRO 系列处理器的决定反映了 AI 工作负载将来的潜在运行位置，而媲美独立显卡的 GPU 和桌面级 CPU 核心是专为企业当前使用的传统应用和以 AI 为中心的应用而设计的。将 AI 应用迁移到 NPU 后，不仅能节省更多能效，还能释放 CPU 和 GPU 资源以专注于处理其他任务。

结语

传统的移动/塔式二元设计方法限制了移动紧凑型工作站可能处理的工作负载和场景，企业需要超越这种传统设计的新型工作站。随着企业向 Windows 11 过渡，加上 ISV 将人工智能技术集成到现有的照片和视频编辑工具、内容管理平台、知识库和办公套件中，工作站用户可能会在操作系统和应用级别接触到 AI，无论是开发人员还是创意人士都是如此。

如果企业和最终用户需要处理规模更大的项目集，应对更为复杂的 AI 加速项目，以及在本地开发基于 LLM 的新应用，那么 AMD 锐龙 AI Max PRO 系列处理器绝对是合适之选。AMD 锐龙 AI MAX Pro 系列处理器配备 ISV 认证的强大显卡，支持 AMD Pro 技术带来的安全性和可靠性功能，并利用内在集成优势推动移动紧凑型工作站实现突破。

脚注

1.请参阅 IDC 报告“IDC 调查显示，2023 年全球 PC 工作站出货量下降近 9%，但在多个市场驱动因素的联合推动下，2024 年预计将迎来复苏”，2024 年 3 月 13 日