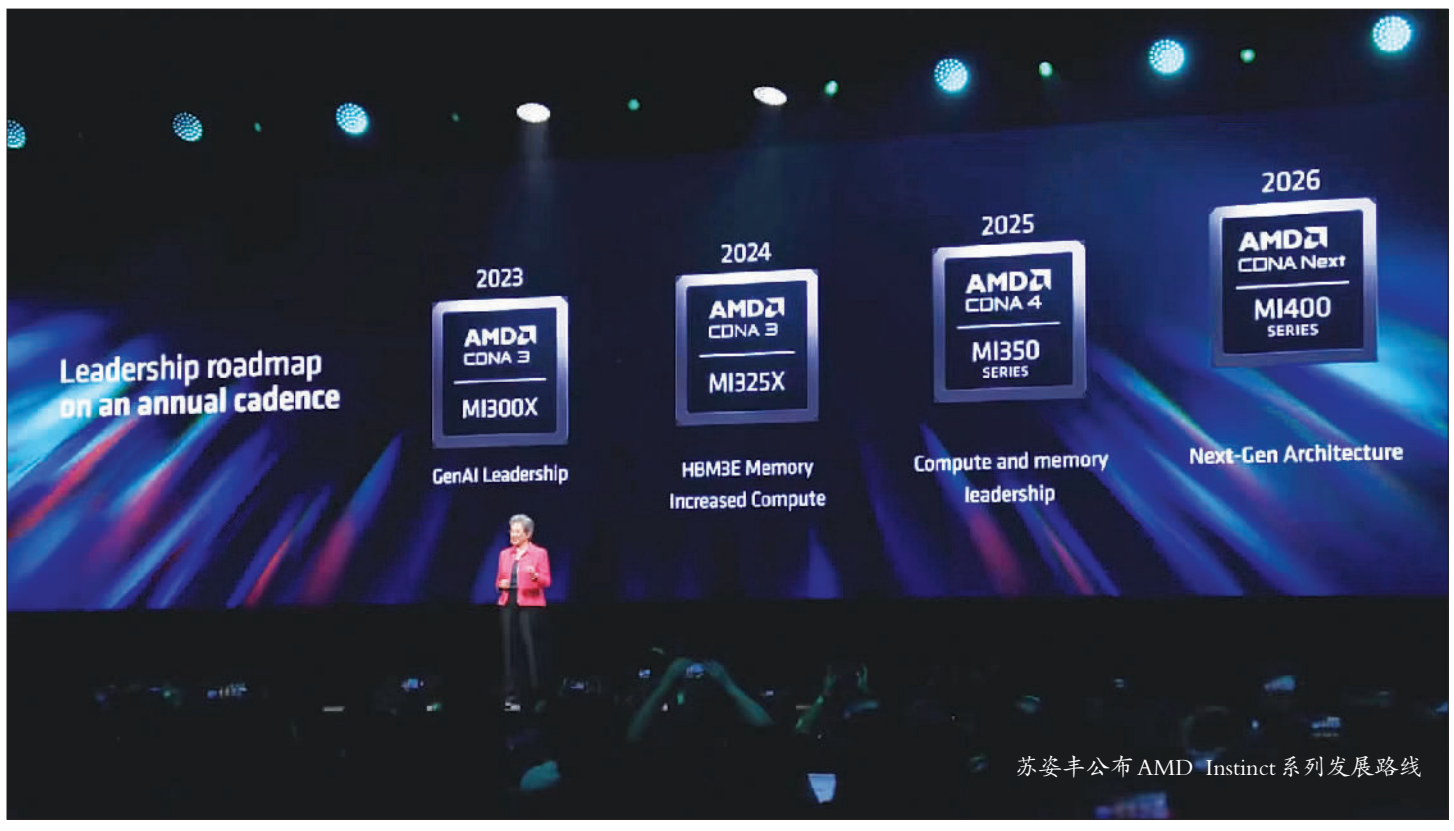
芯片厂商瞄准AI持续发力

6月4—7日，2024年台北国际电脑展（COMPUTEX 2024）在台北南港展览馆举办。英伟达、AMD、英特尔、高通、联发科、Arm等半导体企业的首席执行官齐聚一堂，面向AI PC等人工智能驱动下的新业态、新趋势分享最新洞察，并发布了最新的处理器产品、计算解决方案与技术创新成果。始于1982年的台北电脑展是全球第二、亚洲最大的国际电脑展，一路伴随PC产业和市场发展。此次COMPUTEX 2024，无论是所展示的产品、所输出的观点、所凝聚的人气，都给人一种感觉：PC时代的COMPUTEX又回来了。

本报记者 张心怡 夏冬阳（实习）



苏姿丰公布AMD Instinct系列发展路线

英伟达探索加速计算与AI的未来

作为近年人工智能浪潮的“卖铲人”，英伟达在AI芯片市场“风头正劲”。在COMPUTEX 2024上，英伟达首席执行官黄仁勋在演讲中强调，新的计算时代正在启动，未来计算机产业发展的关键在于加速运算与人工智能。

随着数据量呈指数级增长，传统的CPU性能扩展已大大放缓，难以满足日益增长的计算需求。为此，英伟达发明了一种新架构：通过为CPU添加作为专用辅助处理器的GPU来实现对密集型应用程序的加速。由于这两个处理器可以并行工作，可以让原本需要100个时间单位才能完成的任务，仅需约1个时间单位即可完成。这种架构可以实现100倍的加速计算，而功率仅增加约3倍，每瓦性能比单独使用CPU提高25倍，成本仅上升约50%。

在广受关注的芯片产品上，黄仁勋表示，英伟达Blackwell架构的GPU产品已经开始投产，与2016年Pascal的19 TFLOPS相比，它的AI算力增长了1000倍，几乎“超越了摩尔定律在最佳时期的增速”。此外，英伟达计划于2025年推出增强版Blackwell Ultra GPU，2026年推出下一代GPU架构Rubin。Rubin将配备全新GPU、基于Arm架构的下一代CPU平台Vera CPU，以及采用NV-Link 6、CX9 SuperNIC和X1600，融合InfiniBand/以太网交换机的网络平台。

在今年台北国际电脑展开幕前夕，英伟达宣布其RTX系列显卡将支持微软的Co-pilot+PC。而在本次展会上，英伟达推出了GeForce RTX技术，支持在全新GeForce RTX AI笔记本电脑上运行G-Assist AI助手及NVIDIA ACE数字人生成式AI等内容。黄仁勋认为，未来的笔记本电脑和PC将成为人工智能的载体。这些PC将运行由人工智能增强的应用程序，无论是进行照片编辑、写作，还是使用其他工具，都可享受到人工智能带来的便利和增强效果。

AMD强化AI性能和能效的双重革新

即便英伟达目前以约80%的市场份额主导着人工智能半导体市场，但AMD依旧不甘示弱。在本次台北电脑展上，AMD董事长兼首席执行官苏姿丰表示，AI是AMD的首要任务。“从大规模云服务器和企业集群，到下一代支持AI的智能嵌入式设备和个人电脑，AMD可以为定义人工智能时代的端到端基础设施提供支持。”

为“硬刚”英伟达，AMD在本次展会上公布了第5代架构Zen5的CPU细分产品线，覆盖从台式机到AI PC、从游戏到服务器等应用领域，并推出了NPU和GPU解决方案，包括全新的AMD Instinct系列路线图、AMD Ryzen系列处理器和EPYC处理器。苏姿丰表示，Zen5架构是AMD有史以来性能最好、能效最高的CPU架构。与Zen4相比，Zen5的指令带宽增加了一倍，缓存和浮点单元之间的数据带宽增加了一倍，AI性能也增加了一倍，同时具有完整的AVX 512吞吐量。

而搭载了全新Zen5架构CPU、升级版RDNA3.5架构GPU和全新XDNA2架构NPU的AMD Ryzen AI 300系列处理器，是全球首款浮点NPU，可在不牺牲准确性的情况下将16位应用程序的性能提高一倍，将应用于微软Copilot+PC等下一代AI PC。

此外，在英伟达宣布了“芯片年更”计划后，AMD也提出了要保持“一年一迭代”的速度（行业的惯例通常为“两年一迭代”），并公布了未来几年的AMD Instinct加速器路线图，包括MI325X、MI350和MI400系列等新产品。作为MI300的后继产品，MI325X在保持高性能的同时，优化了能效比。据苏姿丰介绍，该加速器配备了高达288GB的HBM3E内存和6TB/秒的内存带宽，计算性能是英伟达H200的1.3倍，预计于今年第四季度上市。另外，AMD将在未来两年内相继推出MI350X和MI400X，以提供更高性能。

英特尔在AI PC处理器上实现封装级内存

作为老牌IDM，英特尔公司在台北电脑展上公布了包含架构设计创新、先进封装在内的处理器技术进展。首席执行官帕特·基辛格（Pat Gelsinger）在主题演讲中表示，AI的影响力可与当年的互联网媲美，每一台设备都将融入AI，每一家公司都将转型为AI公司。从数据中心和云端，到边缘和PC，英特尔致力于让AI无处不在。

围绕AI PC这一台北电脑展的焦点，英特尔团队公布了下一代AI PC旗舰处理器Lunar Lake的架构细节。该处理器整体算力将达到120TOPS。其中，下一代Xe2 GPU提供67TOPS算力，主要用于游戏和AI创作；第四代英特尔NPU提供48TOPS的AI性能，用于AI助手和创作；CPU提供5TOPS算力，用于轻型AI工作负载。

从整体架构来看，英特尔通过Foveros封装技术，将计算模块、平台控制模块、填料模块（用来拼凑正方形）封装在基础模块上，再与内存一起封装，这也是英特尔首度实现封装级内存。这样的封装方式，节省了250mm²的芯片面积，也将PHY功耗降低了40%。该处理器预计将于2024年第三季度出货。

基辛格表示，开放标准、安全和可持续性是“让AI无处不在”的核心。在数据中心层面，英特尔通过至强6等产品，以高性能、高密度、低功耗的方式，为AI创新项目提供算力和基础设施支持；在网络和边缘领域，英特尔正在向客户提供基于开放标准的安全技术；在客户端，英特尔正在与ISV和OEM生态系统合作伙伴，推动AI PC的发展，以充分发挥Lunar Lake等产品的潜力。

高通强调AI PC处理器要兼顾速度、功耗与散热

AI PC有一个特殊的定位——移动的生产力工具，即AI PC要支持AI在后台全天候运行，这一趋势也对PC处理器提出了新的技术需求。

高通总裁兼首席执行官安蒙在台北电脑展主旨演讲中指出，AI大幅改变了PC的交互和工作流程，并将内容创建的速度提升了一个数量级。AI PC的使用体验更加趋向情境化和个性化。当PC终端的AI与云端AI整体协作时，能够实现混合AI。当移动技术与PC加速融合，开发者的思维方式也在发生革命性的变革。用户要真正受益于AI PC基于意图的主动交互式操作系统，就需要AI持续运行，因此需要考量速度、功耗和散热三个维度。

骁龙X系列（骁龙X Elite和骁龙X Plus）是高通面向AI PC发布的处理器，安蒙在演讲中公布了该系列更多的技术细节。在速度方面，骁龙X Elite搭载了算力为45TOPS的NPU，能够对PC端的AI功能进行加速。比如针对骁龙X Elite NPU优化的视频调色及编辑软件达芬奇，其Magic Mask（魔法遮罩）在基于X Elite运行时，速度是集成GPU的顶级14核Windows处理器的4.7倍。X Elite相比竞品还实现了更快的网页浏览器响应速度。在功耗方面，X Elite采用的定制Oryon CPU在达到相同峰值性能时，功耗相比竞品降低65%，NPU每瓦特性能是竞品的2.6倍以上。在散热方面，骁龙X Elite在NPU持续运行1小时的情况下，能够提供比竞品更高的性能，并保持更低的温度。

在上个月，微软发布了Copilot+ PC，首批产品将搭载高通X系列处理器。微软Windows+终端公司副总裁Pavan Davuluri在台北电脑展现场表示，为了充分利用骁龙X系列处理器的能力，微软重构了Windows 11。对于微软来说，骁龙X系列令人兴奋的是，它是Windows平台上第一个将面向NPU打造的小语言模型Phi Silica内置于操作系统的计算平台。据悉，Phi Silica在NPU运行，能够提供本地推理并优化首词元延迟（模型在收到提示后生成回答的第一个词元所需的时间）。

联发科展示针对Chromebook和4K的SoC

在本次展会上，MediaTek联发科副董事长暨执行长蔡力行介绍了两款AI芯片新品——用于Chromebook的Kompanio 838 SoC和用于4K高级智能电视和显示器的Pentonic 800 SoC，并首度透露联发科正在朝2nm制程与2.5D/3D小芯片（Chiplet）的方向推动AI芯片迭代，以实现无所不在的混合式AI运算。

据悉，Kompanio 838搭载了高能效的八核CPU，集成了AI处理器NPU 650，以支持Chromebook所需的AI增强功能。与前代产品相比，Kompanio 838的内存带宽翻倍，提供了更高的数据吞吐量。同时支持DDR4和LPDDR4X内存，以满足OEM更广泛的产品设计需求。而Pentonic 800则支持多种AI画质增强技术，其中包括AI超分辨率（AI-SR 3.0）、AI-Contrast 2.0、AI场景识别画质监测和AI物体识别画质增强。与上一代产品相比，它的AI性能提升了50%，内存带宽占用降低了60%，支持VRR可变刷新率技术，进一步优化了显示器的画质和游戏体验。

Arm追求提高能效和AI芯片交付能力

在2024台北电脑展的演讲和展示中，Arm公司首席执行官Rene Haas强调了当前计算产业面临的两项挑战，一是AI算力能否获得足够的能源，二是如何通过软硬件服务帮助芯片厂商及开发者提升AI产品交付能力。

数据中心是云服务厂商部署大模型的基础设施。生成式AI和大模型的迅速扩充，为数据中心的能耗带来了严峻的挑战。数据显示，至2023年，仅美国数据中心的用能需求就将达到800kWh。

Arm诞生于以电池供电设备为目标客户的市场，长期为物联网等追求低功耗的边缘设备，以及智能手机等兼顾高性能和长续航能力的移动智能设备提供计算解决方案。如今，Arm希望将能效表现的优势延伸到数据中心领域。Rene Haas现场披露的数据显示，云服务厂商亚马逊、微软、谷歌基于Arm架构自研的Graviton、Cobalt、Axion处理器，在能效表现上分别较上一代产品提升60%、40%和60%，英伟达基于Arm架构研发的Grace Blackwell平台训练大语言模型的能耗较上一代架构降低到二十五分之一。

而Arm关注到的另一个挑战，是在竞争越发激烈、迭代越发快速的AI市场中，芯片设计厂商及广大开发者缩短AI产品上市时间的需求。在硬件层面，Arm展示了Arm终端CSS，该方案是Arm首次交付基于3nm这一最新工艺节点的CPU和GPU物理实现，呈现出设计中的晶体管和线路，以加速芯片设计厂商的流片和产品上市速度。在软件层面，Arm推出了KleidiAI计算库，使开发者能够充分调用Arm架构和硬件性能，以加速基于Arm计算平台的应用层创新。

基于Arm CSS和KleidiAI等软硬件技术，Rene Haas预计到2025年年底，将有超过1000亿台基于Arm架构的设备用于AI。

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号