AI PC卷出“芯”趋势

本报记者 张心怡

虽然苹果公司的AI PC预计于今年晚些时候才能与消费者见面，但在6月11日凌晨进行的苹果WWDC 2024上，苹果团队详细描绘了Mac系列将具备的AI能力。现场公布的信息显示，苹果将为全产品线配置个人化智能系统Apple Intelligence，将生成式模型置于iPhone、iPad、Mac的核心，根据“个人情景”为用户提供智能协助。对于Mac系列，Apple Intelligence的测试版本将在今年秋季随MacOS Sequoia推出，支持配备M1的Mac以及后续机型。在苹果展示的AI PC图景背后，是苹果的软硬件整合能力。苹果软件工程高级副总裁Craig Fedrighi表示，A17 Pro和M系列处理器，为驱动Apple Intelligence提供了坚实的算力基础。

从微软的Copilot+PC到苹果由Apple Intelligence加持的Mac系列，AI正在重塑个人计算体验，驱动PC向着个性化、情景化的方向发展。在这个过程中，AI PC处理器也在全维度进化。

NPU算力战火升级 既要好用也要用好

如何使NPU算力真正发挥效能，以满足AI大模型的部署要求，仍需要厂商进一步细化架构与系统设计。

截至2024年6月，主要PC处理器厂商最新产品和架构的NPU（神经网络处理器）算力已经来到40TOPS级别，最高来到50TOPS。“卷”NPU算力也成为第三方处理器厂商争夺AI PC市场份额的重要手段。

而在去年年底之前，NPU甚至还不是PC处理器的标配。NPU作为能够以更低功耗加速AI负载的处理单元，率先应用于手机。2017年9月，华为推出全球首款内置独立NPU的手机处理器麒麟970。同年，苹果、高通分别将NPU搭载于A11 Bionic处理器和Hexagon DSP（数字信号处理器）。这或许也解释了为什么高通、苹果这两家兼具手机和PC业务的Fabless，会先于英特尔、AMD等传统PC处理器厂商，将NPU应用于PC处理器。2020年，苹果在M1芯片搭载了NPU，这枚NPU与苹果手机处理器A14搭载的NPU类似，算力达到11TOPS。高通在2020年9月面向PC发布的Snapdragon 8cx Gen2 5G处理器，搭载了9TOPS算力的Hexagon NPU。

时间来到2023年，面对生成式AI向设备侧蔓延的凶猛势头，老牌PC处理器厂商们以AI PC为目标市场，终于将NPU武装进自家产品线。AMD为2023年发布的Ryzen Mobile 7040系列处理器的部分型号配置了NPU，能提供最高10TOPS的算力。英特尔在2023年12月发布的Meteor Lake处理器首次搭载NPU，算力为11TOPS。

而同一年，苹果最新PC处理器M2 Ultra的NPU算力已经来到31.6TOPS，高通的骁龙X Elite平台NPU算力达到45TOPS。这让作为第三方处理器供应商的高通占据了先机。2024年5月，微软发布了“有史以来速度最快、最智能的Windows PC”Copilot+PC这一全新AI PC品类，45TOPS的高通骁龙X系列处理器成为首批Copilot+PC的处理器。微软Windows与设备副总裁帕万·达武鲁里（Pavan Davuluri）认为，每台Copilot +PC都需要一个至少能够处理40TOPS的NPU。

有了微软划定的“起跑线”，有着“Wintel”基因的英特尔和它在PC市场的老对手AMD也不甘落后，在6月的台北国际电脑展上纷纷亮出大招。英特尔宣布下一代AI PC旗舰处理器架构Lunar Lake所搭载的NPU能提供48TOPS的AI性能，AMD更是将Ryzen AI 300系列的NPU算力拉到了50TOPS，以倍速提升的算力实现了与Copilot+PC的适配。

但算力的提升只是驱动设备侧AI算力的第一步，如何使NPU算力真正发挥效能，以满足AI大模型的部署要求，仍需要厂商进一步细化架构与系统设计。

比如，在AI推理任务中，FP16（半精度浮点数）和INT8是常见的数据类型。NPU当前以INT8数据类型为主，运行模型所需的计算和内存较少，但牺牲了一定的精度。而FP16有着更高的精度，却不具备INT8的计算和内存特性。这让开发者在设计计算方案时，需要对精度和性能做出平衡。针对不同数据类型的特点，AMD采用了块16位浮点（Block FP16）这一新的数据格式，将FP16的精度与INT8的性能结合在一起。XDNA2也成为首款支持Block FP16的NPU。

此外，由于CPU和GPU也能够处理特定的AI任务，如何协调这两类芯片与NPU的计算负载，也影响着系统性能的发挥。例如英特尔的Lunar Lake架构，就在AI任务的协调上做了工作，CPU、GPU与NPU分别负责轻型AI负载、创作AI和AI助手类任务。

全天候AI需要更高能效比 WoA阵营迎来机遇

无论是Arm架构还是x86架构，都在提升架构的集成度，通过将内存整合进封装来优化功耗表现。

AI PC强调始终在线，也就意味着设备端的AI工具全天候在后台运行。这需要处理器具备更高的能效比，以保证AI PC的续航能力。除了采用NPU这一能够用更低功耗处理AI负载的处理单元，处理器厂商还通过制程、IPC、减少内存访问距离等方式，进一步提升处理器的能效比。

芯片的制程尺寸越小，则电流传输距离越短，功耗也就越小。在最新一代AI PC处理器中，苹果M4采用台积电第二代3nm技术，高通骁龙X Elite和AMD Ryzen AI 300都采用了4nm制程，与当前的旗舰型手机同步。

同时，IPC（每时钟周期执行的指令数）越高，意味着CPU在相同频率下的性能越高。英特尔Lunar Lake性能核的IPC较上一代提升了14%，在相同功耗下能实现10%~18%的性能提升。AMD Ryzen AI 300系列处理器采用的Zen5架构也实现了16%的IPC提升。

同样值得注意的是，无论是Arm架构还是x86架构，都在提升架构的集成度，通过将内存整合进封装来优化功耗表现。Arm在5月底推出的终端计算子系统（CSS）中，采用了系统级高速缓存（SLC），以减少DRAM带宽和访问次数，提升系统能效。苹果M系列处理器，也一直采用封装级内存，将SoC和DRAM芯片安装在一起。而此前的大多数x86处理器，都将主内存外置。在Lunar Lake架构中，英特尔首次把内存集成到封装内。这样的封装方式，让计算核心以更短距离、更低延迟访问内存，将PHY功耗降低了40%。

随着能效比的重要性日益凸显，“WoA”（Windows on Arm）也受到了更多OEM厂商的关注。从架构来看，Arm在功耗和边缘侧AI推理具有优势。从生态来看，Windows正在深化与原生Arm的适配。据统计，在运行Windows10和Windows11的iGPU（集成GPU）笔记本电脑中，用户在87%的应用程序使用时间中，使用的是原生支持Arm的版本。加上Arm在最新的计算平台中，又面向AI设备侧的发展趋势，进一步提升了能效比。比如Arm于5月底推出的终端计算子系统（CSS）所采用的GPU Arm Immortalis-G925，在提供与上一代产品相当的游戏性能时，功耗降低了30%。

当前，WoA主要的芯片玩家是高通，但英伟达和联发科也对这一市场虎视眈眈。英伟达在5月宣布，搭载RTX GPU的Windows11 AI PC将在未来几个月推出，且发布了帮助开发者在Windows PC优化及部署生成式AI模型的开发工具NVIDIA RTX AI Toolkit。而联发科也在近日传出正在开发基于Arm架构的AI PC芯片。摩根士丹利分析师Charlie Chan预计，WoA AI PC芯片出货量将在2024年达到200万台左右，2025年将增至1500万台。

混合AI时代的隐私挑战 苹果芯片进入服务器

AI PC这一产品类型的核心价值，是围绕个人情境的计算体验，这就需要结合用户的个人数据。

AI PC这一产品类型的核心价值，是围绕个人情境的计算体验，这就需要结合用户的个人数据。无论微软Copilot+PC主打的Recall功能，还是Apple Intelligence加持的Siri提供的跨APP协作能力，底层逻辑都是对用户信息进行语义索引。这也是为什么逐鹿AI PC的处理器厂商，都强调终端侧AI的概念，如Pavan Davuluri所言，设备侧AI意味着更快的响应时间、更好的隐私和更低的成本。

但是，如果为了保护个人数据，就限制用户获得云端大模型的支持，未免“因噎废食”。2023年，高通提出了混合AI架构，也就是根据模型和查询需求的复杂度等因素，选择不同方式在云端和终端侧之间分配处理负载。当用户发起请求时，终端侧神经网络或基于规则而运行的判决器（arbiter）将决定是否需要使用云端。

而苹果在WWDC 2024上，展现了混合式AI的具体图景。当用户向Siri提出专业问题时，Siri会提示用户是否询问ChatGPT，并询问能否把照片、文档等信息分享给ChatGPT。同时，苹果各产品线的写作工具和图像生成工具，也可以运用ChatGPT进行创作。

但是，一旦在设备侧通过网络获取ChatGPT等部署在服务器端的大模型服务，就有可能面临个人数据被服务器存储的风险。Craig Fedrighi表示，在传统方式中，服务器会存储个人数据，甚至未经同意就使用这些数据，且用户难以验证个人数据是否被滥用。

面向个人数据在混合式AI时代的安全挑战，苹果推出了私有云计算技术，为服务器大模型提供芯片级安全保护。当苹果设备判断用户的请求无法用设备侧AI解决，会引入基于服务器的模型来处理更复杂的请求，而服务器端的模型会在以苹果芯片（Apple Silicon）打造的服务器上运行，可提供与iPhone相同的芯片级别的隐私安全保护。Apple Intelligence会将仅与任务相关的数据发给苹果芯片服务器，同时，独立专家能够检查服务器运行代码，以验证用户的隐私安全能否得到保障。“‘私有云计算’通过加密形式，确保iPhone、iPad、Mac可以拒绝与服务器对话，除非这个服务器的软件已经得到公开的安全标准认证。”Craig Fedrighi说道。

车规级存储持续创新 “存力”解决“算力”焦虑

本报记者 王信豪 许子皓

汽车形态和功能的持续进化，为车规级芯片带来了巨大的市场增量，也对车规级芯片的集成度、算力、功耗等指标，以及芯片供应商的研发和响应能力提出了更高要求。在汽车智能化、网联化的浪潮中，SoC处理能力的提升，传感器数据量的扩张，都推动了存储技术的不断提升。

“传统汽车的功能对数据存储需求不高，而随着智能化程度的加深，汽车存储芯片的需求量将持续增长。”得一微电子市场总监罗挺向《中国电子报》记者表示，“一方面，数字仪表盘等显示组件要求画面越来越高清，智能座舱也将搭载多样化的功能并记录用户数据；另一方面，ADAS和自动驾驶不论是预载高精度地图，还是使用大模型推理，都需要存储容量的提升。”

当前一辆车的闪存容量范围在64~256GB，未来在端侧大模型、娱乐系统、传感器精度等方面的升级将使存储容量提升至TB级别。Yole数据显示，2027年车载存储市场规模将达到125亿美元。

存储芯片不仅需要大容量，还要为数据提供端到端的保护。此外，由于汽车行驶环境较为复杂，存储芯片还需具备对高温差环境的适应性。

据悉，得一微电子车规级eMMC5.1存储芯片已被广泛应用于数字仪表、T-Box、ADAS、智能座舱、车载信息娱乐系统等多个场景，且在-40℃至105℃的极端环境中稳定运行，既保证了对车规芯片高可靠性的严苛要求，同时满足了供应链创新的需求。此外，得一微电子基于其创新性车规存储主控芯片，即将推出车规级BGA SSD及UFS产品，以满足汽车行业日益增长且多样化的存储需求。

“面向未来的车规级存储，在存储控制之外，还要在存算互联和存算一体两个方向进行创新。”罗挺表示，传统汽车中的存储和计算彼此分离，二者的数据交互需要大量“沟通成本”，成为系统的瓶颈。存算互联、存算一体（或近存计算）的技术创新将有效解决数据传输的内存墙、功耗墙等问题。

在存算互联方面，过去的CAN和LIN总线已经难以满足当前汽车的数据传输需求。“目前整车计算的瓶颈不仅在于SoC的处理能力，也在于DRAM和Flash如何与多个处理器（如CPU、GPU和NPU）进行联动。不论是PCIe（高速串行通信协议），还是Cache（高速缓存），高速率、低延迟的传输技术在未来都将以存储为中心再次升级。”罗挺表示。

在存算一体方面，车用存算一体不再是将SoC和存储芯片进行简单连接，而是依托车规级存储自身具备的计算能力，在未来智能汽车的运行中高效、精准地完成数据的矩阵运算、数据库加速、加解密、数据销毁等操作，此举可以减少这些数据在SoC和存储器之间的搬运，明显降低SoC的算力负荷、成本以及系统的复杂度。

“想实现存算一体，既需要芯片的创新，也需要系统的创新。”罗挺告诉记者，“这要求当前的市场具备更加完善和成熟的生态体系，以及更多合作伙伴围绕车载存储系统形成紧密联系。”

Rapidus与IBM达成2nm芯片封装技术合作伙伴关系

本报讯 近日，日本晶圆代工企业Rapidus与IBM共同宣布，双方以联合开发2nm节点技术的现有协议为基础，确立了2nm世代半导体芯片封装量产技术合作伙伴关系。作为本合作伙伴关系的一部分，Rapidus的技术人员将在IBM北美的封装制造研发基地进行合作。同时，Rapidus还上调了其尖端半导体的销售目标。

2022年12月，Rapidus就与IBM达成战略性伙伴关系，双方将携手推动基于IBM突破性的2nm制程技术的研发。而早在2021年，IBM就正式公布了全球首款2nm芯片原型。关于此次合作，Rapidus董事长小池淳义表示：“继2nm半导体的共同开发之后，关于芯片封装的技术确立也与IBM达成了伙伴关系。我们将最大限度地运用这次国际合作，推进日本在半导体封装的供应链上发挥比现在更重要的作用。”

据《日经亚洲》最新的报道显示，由于人工智能市场对于芯片需求量的强劲增长，Rapidus近日对外透露，Rapidus上调了其尖端半导体的销售目标，即到2030年实现超过1万亿日元（约64亿美元）的收入，相比之前2040年实现1万亿日元收入的目标提前了10年。

资料显示，Rapidus于2022年8月成立，由丰田、索尼、NTT、NEC、软银（Softbank）、电装（Denso）、NAND Flash大厂铠侠（Kioxia）、三菱UFJ等8家日企共同出资设立，出资额各为73亿日元，且日本政府也将提供补助金作为其研发预算。

Rapidus目标在2027年量产2nm以下最先进逻辑芯片，其位于北海道千岁市的第一座工厂已在2023年9月动工，试产产线计划在2025年4月启用，2027年开始进行量产。

此前，日本经产省已向Rapidus提供了3300亿日元的补助，如果加上媒体报道的2024年度将追加的约5900亿日元补助款，Rapidus预计可获得总共近1万亿日元的补助款。（艾 文）

机构上调今年全球半导体产值预测至6110亿美元

本报讯 近日，世界半导体贸易统计组织（WSTS）宣布上调最新的半导体市场预测。WSTS修正的预测数据显示，2024年全球半导体市场将实现16%的增长，达到6110亿美元。这反映了过去两个季度的强劲表现，特别是在运算终端市场。

2024年预计主要有两个集成电路类别将推动今年的增长，增幅达到两位数：逻辑增长10.7%，存储增长76.8%。相反，分立元件、光电子元件、感测器和类比半导体等其他类别预计出现个位数下降。

按区域来看，美洲和亚太地区预计出现显著增长，增幅分别为25.1%和17.5%。相比之下，欧洲预计出现0.5%的年增长，而日本则预计小幅年减1.1%。

展望2025年，WSTS预测，全球半导体市场将增长12.5%，产值估达6870亿美元，增长预计主要由存储和逻辑产业推动，这两个产业的产值将分别飙升至2000亿美元以上，与上一年相比，存储增长超过25%，逻辑增长超过10%。预计所有其他细分市场都将有个位数增长率。（文 编）