市场分析

AI PC 发展机遇

赞助方 intel

概述

人们对 AI（尤其是生成式 AI）的浓厚兴趣激发了技术行业的高涨情绪。然而，在兴奋之余，大家也逐渐意识到并非所有 AI 支持的新功能都能在云端计算。因此，在 PC 以及其他客户端设备上运行 AI 工作负载成为新的关注焦点。针对人们对 AI 的这种关注，英特尔等 PC 半导体供应商推出了为运行这类应用而专门优化的新型芯片和软件。从在英特尔® 酷睿™ Ultra 等新一代片上系统（SoC）中内置新型 AI 加速器架构（如 NPU），到进一步利用 CPU 和增强型 GPU 的强大性能，英特尔公司正集中力量，抓住在 PC 上运行生成式 AI 及其他类型 AI 应用的发展机遇。从中，我们可以明确的一点是，仅关注 NPU 的 TOPS 性能并不能准确地评估在 PC 上使用 AI 所能达到的效果。各方面的软件支持，包括开发工具、嵌入在操作系统中的 API、运行时 AI 框架、部署工具以及系统级驱动程序，这些都是在客户端设备上充分利用 AI 潜能的必要条件。

> “AI 最令人振奋的影响是，它开阔了人们的思维，让人们敢于大胆挖掘计算设备的潜能。”——首席分析师 Bob O'Donnell

简介

基于人工智能（AI）的应用激起了人们的兴奋之情，其中一个最令人振奋的影响就是，它开阔了人们的思维，让人们敢于大胆挖掘计算设备的潜能。关于释放 AI 的潜力，人们已经经历了数十年的殷切期盼，在这份期盼似乎要以失望告终时，支持生成式 AI 应用（如 OpenAI 的 ChatGPT）的基础模型的广泛推出和近乎立竿见影的影响开启了一个令人振奋的全新计算时代。

生成式 AI（GenAI）变革了人们对计算、创作、生产、通信等问题的思考，激励着全球利用这项技术来创造影响。这种影响也正在从 GenAI 扩展至许多现有的“传统”AI 应用。从图像和视频分析与编辑，到办公效率、会议记录和总结、3D 建模和纹理渲染、图像/视频中的对象擦除等，各种 AI 应用都开始进入全新时代。除此之外，由于以 AI 为中心的现有计算资源可以开始被加以利用，人们也开始以新的视角去看待更多“传统”AI 应用，如背景模糊和音频降噪等。

在此之前，对于基于 AI 的计算，人们主要关注的是在云中运行的应用和服务。但实际上，在 PC 及其他客户端设备上直接运行这些应用可带来许多有趣的新机遇。这种做法不仅在快速成为现实，而且在某些情况下，本地运行 AI 的性能和输出也会更高。此外，当您可以在自己的设备上使用数据，而不是将其发送至公有云环境时，隐私性和安全性相关的优势也非常显著。

过去几个月里，端侧 AI 和 GenAI 解决方案取得了巨大的进步，使得一些可能性正在成为现实。由于开源基础模型的快速发展和缩小，以及模型量化等技术的进步，许多行业观察者认为，未来几年客户端设备上都无法实现的技术可能会突然在未来几个月内成为现实。事实上，在过去几个月里，设备端的创新速度甚至比 GenAI 的整体进步速度还要快，这恰恰说明了一些问题！

除了惊人的技术进步，端侧 AI 发展的显著加速很大一部分原因是一些非常实际的问题。值得一提的是，由于 GenAI 工具采用率的飞速增长以及大量新产品的上线，人们普遍认为，现有公有云数据中心基础设施根本无法满足所有的预期需求。此外，人们对于这些基于云的资源在电力方面的需求也颇有顾虑。最后，围绕成本、安全性和效率的问题都表明，在云中运行所有或者大部分 AI 工作负载并不是长期可持续的选择。因此，AI 应用要保持持续的发展势头，端侧 AI 解决方案变得至关重要。更多 AI 工作负载必须转移至 PC。

均衡片上系统的重要性

在这种环境下，人们将大量关注放在了可在 PC 等客户端设备上运行的 AI 应用和工作负载。当然，这与现代 PC 提供的计算资源类型直接相关。（我们也将在后文中进一步讨论人们对充分利用计算硬件所需软件越来越多的关注。）

所幸，今年新推出了几款 PC 片上系统（SoC）架构，在运行 AI 工作负载时比之前的硬件更加强大和高效。特别是英特尔® 酷睿™ Ultra SoC（之前的代号为“Meteor Lake”）等芯片现已包含更灵活的 CPU、更强大的 GPU 以及一种叫做神经处理单元（NPU）的新型组件，专为多种 AI 工作负载进行了优化。

尤其是 NPU 的加入，使人们对这些全新的现代 SoC 架构及其用于支持端侧 AI 的前景产生了极大的兴趣。NPU 旨在提高计算矩阵乘法及其他数学公式时的性能，这些数学公式常用于大型应用中的 AI 应用或功能。对于算法以及其他在后台持续运行的软件组件（如数字助手应用和其他“智能代理”），NPU 能够显著提升它们的性能。

尽管对于某些 AI 应用的作用非常强大，但 NPU 也并非是所有 AI 应用的“灵丹妙药”。事实上，许多 PC 上的 AI 推理负载仍是由 CPU 完成的，而 GPU 则能更高效地完成其他与 AI 工作负载相关的计算。值得注意的是，在 PC 上运行的几乎所有 AI 进程实际上都可以由 SoC 的任意架构组件（CPU、GPU 或 NPU）进行计算，不同的只是效率。此外，在英特尔® 酷睿™ Ultra 这种既有性能更高但功耗更大的 P-Core（性能核）、又有性能较低但更节能的 E-Core（能效核）的芯片上，有时候一种 CPU 会更适合某一类 AI 工作负载。一般来说，CPU 会用于要求低时延的单个轻量型推理 AI，GPU 会用于 AI 密集型工作负载，而 NPU 则用于持续型 AI 和 AI 卸载。

这就是很典型的“办事要选对工具”。正如我们在现实生活中学到的经验：锤子除了其最初设计的用途，还可以用来做很多其他事，但有些工具却能让我们更轻松（更快速）地完成特定任务！

说到这，除了新功能外，我们还有必要讨论一下在 PC 上运行 AI 应用的性能和效率基准问题。许多公司一开始往往会先关注 TOPS，即每秒的万亿次运算数，这是一种最初为衡量数学计算而设计的衡量机制。人们尤其关注系统 NPU 的 TOPS。

事实证明，要说真实环境中的实际体验，TOPS 并不是最优的衡量标准，原因如下：首先，许多人认为 TOPS 并不能很好地反映真实的性能，它更像是一个简单的综合指标，在真实环境中很难遇到这样的情况。这是因为 TOPS 衡量的是执行的计算次数，没有考虑计算类型，而计算类型通常会对实际性能产生更大影响。另一个受到关注的指标就是单位能耗的 TOPS，该指标根据执行特定计算时产生的功耗来衡量整体效率。尽管人们普遍认为这个指标更合理，但它与真实环境仍然不具备可比性。

另一个巨大挑战就是，这些指标忽略了之前讲到的一个基本事实，即 AI 工作负载会（而且经常会）在完整 PC 系统的几个不同组件上运行。因此，人们开始更多地去讨论系统整体的 TOPS，这样做综合考虑了 CPU、GPU 和 NPU 的潜在 TOPS，可以获得一个能够更好地反映运行多个不同类型 AI 应用的系统指标。

然而，即使是系统 TOPS 也不是理想指标，因为它不一定融合了使用不同应用的经验。衡量 PC 上 AI 性能的另一大难点是，完全基于速度的传统指标在涉及 AI 时就变得不再那么合理。比如说，有些人可能会关心在本地运行基于大语言模型 (LLM) 的聊天机器人时，它对指令的响应速度有多快，但这并非大多数人关心的点。在比较 AI PC 的性能时，响应的质量以及对电池续航的影响等因素可能要重要得多。

在对 AI PC 进行基准测试时还有一项挑战就是：事实证明，与任意指定组件的 TOPS（或系统的整体 TOPS）相比，其他因素会对性能表现产生更显著的影响。对于许多 LLM 来说，鉴于其使用的数据集大小，系统内存以及访问内存的速度对实际性能产生的影响在方方面面都要比 TOPS 更大。举一个例子，与具有更高 TOPS 性能指标的其他系统相比，系统内存更大、内存访问速度更高，即使 TOPS 低一些，在真实场景中的表现可能也会比 TOPS 更高的系统要好。

AI 软件工具

拥有先进的硬件功能固然重要，但就像计算领域的大多数情况一样，如果没有合适的软件工具，任何硬件功能都将变得毫无意义。基于 AI 的软件也是一个发展日新月异的领域，因此 AI/机器学习 (ML)/深度学习 (DL) 模型、算法和开发框架变得尤为重要。

过去，基于云的超大型基础模型为 GenAI 应用和服务提供了支持。但正如前文所述，如今我们在缩小这些模型规模方面取得了巨大进展，使其已经可以在 PC 上原生运行。

大规模模型的缩减版（如 Meta 的 Llama 2、Google 新推出的 Gemini 以及许多其它公司的模型）无需云连接就可以在 PC 上直接运行少于 100 亿参数的基础模型，这为用户提供了更多选项。丰富的开源模型以及 Hugging Face 等市场的增长也为开发人员创造了许多机会，去构建专门在客户端设备上运行的模型。除此之外，近期我们在缩小大型模型的量化规模上取得了大量振奋人心的成果，使其能够适应 PC 的有限资源。总的来说，这些进展以及随之而来的更多发展已迅速将端侧 AI 从短期内的“科幻小说”变成了实时的“科技现实”。

就在 PC 上运行的 AI 工作负载而言，不同 PC 的各种系统和应用级软件组件也对端侧 AI 的体验质量至关重要。例如，在基于 Windows 的 PC 上，操作系统的某些元素在将基于 AI 的工作负载和功能分配给系统硬件的各个组件方面发挥着重要作用。特别是，

DirectML 在 Windows 系统中起到了 AI 应用的“交通警察”作用，负责将给定应用中的各种软件元素或子例程分流至 PC SoC 上合适的硬件元件。DirectML 的新版本集成了英特尔专门为微软创建的一些软件优化功能，以增强 PC 上 AI 应用的整体软件生态系统（这些优化也涵盖了运行其他供应商 SoC 的系统）。像 DirectML 这样的工具对于提高真实环境中的性能至关重要，因为在真实环境中可能会同时运行多个 AI 应用或代理。当发生这种情况时，就必须平衡在不同 SoC 硬件元件上运行的不同软件组件组合，以实现不同的功耗和性能效果。

除了这些系统级增强外，要充分发挥给定应用的性能潜能通常需要与软件开发人员直接合作，以确保他们的代码针对特定架构进行了针对性优化。这就是英特尔的规模及其所具备的大量内部软件开发人员的优势所在，因为他们能够联系并接触到许多开发 AI PC 应用的独立软件供应商 (ISV) 并与其合作。为此，英特尔新推出了 AI 软件计划，通过此计划，英特尔将与前 100 大 AI 应用开发商合作，来提升他们的应用在英特尔® 芯片上的运行效率。

最后一点，却也是常常被忽略的一点就是，在软件方面，还需要考虑开发工具。软件开发人员在创建应用时经常会使用 CPU 供应商提供的工具。诸如英特尔® OpenVINO™ 这样的开发环境有助于加快和改进他们的工作。OpenVINO™ 自带一个包含 200 多个预训练 AI 模型的“模型库”，这些模型专为在 PC 上运行而构建，并且已经经过测试（在英特尔® SoC 上运行可获得更高效率）。此外，OpenVINO™ 还包含一个模型转换 API，使开发人员能将新的公有或开源模型纳入 OpenVINO™，从而更灵活地构建 AI 应用。OpenVINO™ 还支持在 Pytorch 和 TensorFlow 中训练的模型，并可用作 Hugging Face Optimum 和 Pytorch torch.compile 的集成后端，为应用开发人员提供大量选择。

PC 上的 AI 应用

说到这，目前已经出现了许多利用 AI 功能的 PC 应用和系统级功能。微软的 Windows Studio Effects 已经面向 NPU 经过专门优化，可在具有 NPU 的 PC 上用该组件来运行，在实时消息传递功能中实现增强的视频背景模糊和音频降噪效果。喜人的是，与在 PC 的 CPU 或 GPU 上运行相同功能相比，NPU 的效率明显更高。

诸如 Rewind.ai 这类工具所实现的各种可能性更令人感到兴奋，英特尔在其近期的创新活动中也进行了展示。顾名思义，Rewind.ai 可记录您在 PC 上的一言一行，从电子邮件到文档，从聊天到在线会议等，并可提供基于 GenAI 的概括总结，也让用户可以访问并获取所有这些信息。它标志着许多人从 Cortana、Siri、Alexa 等语音助手面市早期就梦寐

以求的真正的数字助手的到来。但与前者相比，Rewind.ai 的功能要强大得多，也实用得多。

在另一条赛道上，新版 Adobe Lightroom 和 Vegas Magix 纳入了 GenAI 图像和视频增强技术，并可使用本地 PC NPU 来加速工作。此外，市场上开始推出其他无需基于云连接就能运行的 GenAI 图像生成工具。与所有本地运行的应用一样，这大大提高了使用这些应用时的隐私性和安全性，因为云端再也无法采集到您的任何信息。

近期 LLM 模型规模的减小对提升 PC 应用通用性的潜在影响也不容小觑。尽管微软的新版 M365 与谷歌的新版 Workspace 生产力工具套件目前都还在利用云实现大多数 GenAI 功能，但如果可以通过这些规模更小的 LLM 直接在 PC 上执行其中一部分功能，其发展前景将十分诱人。更激动人心的是，您可以使用自己（或公司内部）的数据对这些 LLM 进行定制。与所有要在云中访问的工具相比，这种基于本地存储或公司内网存储数据进一步定制的能力能够打造出更强大、更优化的工具。此外，如果使用的所有数据都存储在本地设备上，就可以更快速地执行这些操作。事实上，这也是体现端侧 AI 应用价值的一个很有说服力的重要原因。

许多公司也在开始探索的另一个有趣的方向就是利用混合式 AI，即一项工作的某些部分在云中完成，其他部分则在 PC 上完成。试想一下，图像编辑程序在 PC 上创建了一个屏幕友好的低分辨率图像版本，但随后又通过基于云的模型单独创建了一个更高分辨率的版本。您可以在 PC 上快速编辑低分辨率的版本，但最终保存的是基于云的版本。在监管严格的医疗行业等业务环境中，之前也有一些公司做过类似的事情，比如通过多个模型生成与医疗程序有关的定制化邮件。在这种情况下，个人身份信息等私人信息在 PC 上的本地模型中处理，而邮件中更通用的格式信件部分则通过基于云的大型 LLM 生成，最终再把这两部分元素合并到生成的电子邮件中。这些以及 2024 年可能实现的许多其他用例都表明，未来生成式 AI 的运行将更加顺畅。这也说明 PC 将在生成式 AI 领域发挥比人们最初想到的更重要的作用。

需要明确的是，有一些基于 PC 的应用需要功能强大的 NPU 才能高效运行（有些甚至要完全依赖 NPU 运行），但绝大多数专为在 PC 上运行而构建的应用仅仅是将 NPU 用作加速器（如果设备上有 NPU 的话）。理论上讲，它的作用更类似于 GPU。与搭载集成显卡解决方案的系统相比，在搭载更强大的独立 GPU 的系统上，某些功能的运行速度可能更快，某些游戏可以在更高分辨率模式下或者以更高的帧率运行，但在大多数情况下，没有独立 GPU 时，它们也仍然可以运行。随着时间的推移，将会有越来越多的装机 PC 会搭载更强大的 NPU，软件开发人员也将更充分地利用这些功能。但与大多数技术进步一样，这些演进需要时间才能实现。

结论

毫无疑问，GenAI 以及其他广泛的 AI 开启了一个可以通过计算设备实现的全新发展前景。尽管大多数人已经不得不接受使用云连接来获取这些工具的功能，但他们很快就会发现，端侧 AI 不仅可行，而且是必然。在不久的将来，其体验甚至会优于目前基于云的体验。

基于这些种种原因，我们就可以理解为什么那么多人对计算机领域正在发生的变化感到如此兴奋。正如许多人说的那样，这就像是一个千载难逢的机遇，我们可以用一种令人兴奋的新方式开展工作。

尽管早期有人质疑，但现在我们已经可以清楚地看到，PC 将在未来发展中发挥极其重要的作用。从芯片架构振奋人心的进步到基于 PC 的软件应用和工具方面的重要发展，可以说，PC 正在以一种全新且让人倍受鼓舞的方式焕发新生。诚然，关于如何以更合理的方式去衡量即将实现的性能增益仍然存在许多问题，但公允地说，现在，也许是时候要采取一种全新的方式来思考基准测试和其他衡量标准了。

无论这些问题最终将如何解决，这绝对是 PC 行业激动人心且值得珍视的时刻。