微软AI服务器陷入硬件短缺的窘境

由于微软内部的人工智能（AI）项目需求日益增长，但是难以满足这个增长所需的硬件资源，因此微软不得不实施了一个配额制。这意味着，每个部门都被赋予了一个固定数量的硬件资源，以确保他们能够执行其AI项目。这种硬件短缺现象不仅仅是在微软内部，也是整个硬件市场都面临的问题，但这些困难已经导致了微软极大的不便，尤其是在快速发展的AI领域内。微软表示，他们正在与硬件制造商密切合作，以解决这个问题，在未来加强硬件供应，从而为AI项目提供更充分的支持。

这迫使微软不得不对硬件资源采取配额供应的方式，限制其他内部 AI 工具开发团队的使用量，从而确保基于 GPT-4
的新版必应聊天机器人和即将发布的新款 Office 工具获得足够的资源。

这种硬件短缺的现状还对微软的客户产生了影响：至少有一个客户表示，需要等待很长时间，才能使用已经通过微软 Azure 云服务提供的 OpenAI 软件。

微软表示正在加紧添置更多硬件，以便增加 AI 计算能力，但如果速度不够快，就有可能限制其 Azure OpenAI 服务对新客户的吸引力。这些客户会借助此项服务将 AI 功能添加到自己的应用中。微软原本在这种新的 AI 工具的商业化方面已经领先于谷歌，但如果迟迟不能解决硬件短缺问题，这种优势也将难以发挥出来。

谷歌抢在微软之前于本周发布了自己的 AI 辅助写作工具，包括能够根据简短的提示，在谷歌 Docs 和 Sheets 中自动编写文本的软件。

服务器短缺正值 OpenAI 和微软试图快速扩大销售范围之际，客户可以使用自己的数据对这种 AI 软件进行微调，从而开发图片生成、文档总结、搜索引擎和聊天机器人等定制化 AI 工具。

微软已经向 OpenAI 投资了数十亿美元，有权通过 Azure OpenAI 服务向创业公司销售软件。OpenAI 也将软件直接授权给客户使用 ——而作为与微软合作的一部分，所有 OpenAI 机器学习模型都必须运行在 Azure 服务器上。这两项服务的定价相同，每次请求收费都不到一美分。

这两家公司在过去 3 个月推出的服务吸引了企业用户的广泛兴趣，他们都希望将 OpenAI 的 ChatGPT 聊天机器人的底层技术整合到自家产品中。但是微软 Azure OpenAI 服务的新客户却要面临很长的等待时间。

硬件资源短缺的核心问题在于，GPT-4 等新的 AI 大模型需要在 GPU 上运行，这种服务器芯片可以一次性处理大量数据。为了处理这个模型的庞大计算需求，微软已将数万个 GPU

整合成一个个群集，分散到各个数据中心里。

在明确用户需求量之前，微软目前不得不将现有的大量 GPU 资源预留给新的必应 GPT-4 聊天机器人和即将推出的 GPT-4 Office 工具。

知情人士表示，微软 2 月刚刚推出这项必应智能聊天功能，目前还在加紧测算平均每天有多少人愿意使用这项服务。知情人士表示，这就导致工程师很难预测该功能所需的计算资源。

而即将推出的 Office GPT-4 工具也可能导致微软客户对 OpenAI 聊天机器人技术的需求大幅激增。这些新功能包括 AI 文档总结、个性化写作建议以及编辑建议等。

其他从事 AI 开发的微软团队不得不让位于必应和 Office，包括为各种机器学习模型的开发团队，其中就包括微软的图灵自然语言生成模型，它可以理解文本，之前曾为电子邮件和 Office 应用中的搜索工具提供底层技术。

知情人士表示，当这些团队想要使用 GPU 开发新的 AI 工具或测试现有 AI 软件时，必须向一位公司副总裁提交特别申请，获批后才能使用硬件。

知情人士称，有些申请需要等待几天甚至几周才能获批。自 2022 年末，微软就一直对内部的 GPU 资源实施配额供给，但 1 月以来的等待时间越来越长。

微软发言人在声明中补充道，该公司正在为 Azure 等服务增加更多 AI 资源，并且遵循“ 优先满足客户需求并根据需求进行调整的流程 ”。他们还补充道，微软并不担心 AI 资源。

至少有一个微软客户透露，他们也经历了延迟访问的问题。市场营销软件创业公司 TouchCast 创始人兼 CEO Edo Segal 说：“ 几乎不可能立刻使用程序 ”。Edo

的公司正在为多家车企开发基于 GPT
的互动式用户手册，而这项技术正是通过微软 Azure 授权的。

现有的 AzureAI 客户几乎未受影响。无人驾驶汽车开发商 Cruise 和 AI
搜索创业公司 Perplexity 的发言人都表示，他们在使用 Azure 的 GPU
资源时并未遇到困难。

去年 11 月，微软和 GPU
生产商英伟达共同宣布，微软将新增数万个处理器来扩大 AI
处理能力，但目前还不清楚采购进度。

这两家公司还在周一共同宣布，Azure
的客户很快就能抢先使用英伟达的新款 H100 GPU—— 这款产品尚未广
泛发布。而知情人士称，已经有少部分微软内部团队开始使用
H100，但多数部门尚未获得这项权限。

本文链接：https://dqcm.net/zixun/167894334212364.html