

智算革命：重塑未来计算

在人类文明的长河中，计算始终扮演着至关重要的角色。从远古的结绳记事到现代计算机的普及，无论是提及算力、智算还是 AI，其底层的核心往往都离不开“计算”，计算方式的不断演进，推动了人类社会的进步。

文 | 方磊 九章云极 DataCanvas 董事长

智底层计算模式变化带来的一场产业变革

计算技术正处于革命性发展时期，最引人瞩目的推动力无疑是大模型的崛起。从传统的机器学习、深度学习以及预训练模型等 AI 领域的众多进步，到当前这种规模庞大的万亿参数及以上的大模型，都为计算领域带来了前所未有的变化。

底层计算模式的演进，一直悄然地为产业带来深刻变革。以 80386 为例，当时缺乏数字协处理器，浮点数计算能

力有限，主要处理整型数。随着浮点数计算需求的增长，80486DX 版本应运而生，配备了数字协处理器。这一创新迅速成为标准配置，到了奔腾芯片时代，更是融入了 CPU 的核心功能。

如今，计算需求发生了显著变化，从以 CPU 为主导的条件判断和顺序执行，转向以 GPU 为核心的高密度并行计算，GPU 的并行计算能力远超 CPU。

大模型的崛起更是标志着人类迈入了计算密集的新纪元。与移动互联网时代相比，过去手机 APP 可能更多时间用



于数据传输，而非计算。但如今，计算模式发生了翻天覆地的变化，从简单的浏览图片、视频、文章，转变为生产视频、文章和内容，计算的主要模式已从消费转向生产，整个计算生态正在经历前所未有的变革。

而现在所面临的挑战是从加密货币到 AI 的快速发展。在计算密集型时代，大量的数据需要实时计算，而非预先计算好的简单获取。

因此，我们见证了几个关键的推动力：大模型的发展正式驶入计算密集型时代的蓝海；处理设备从 CPU 时代跨越到 GPU 时代；计算规模也在急剧扩大。自从人类掌握计算技术以来，算力实现了惊人的飞跃，从最初的简单计算到如今的庞大规模，已经增长了 10 的 7 次方，甚至 10 的 8 次方。可以预见，如果再实现一次类似的增长，或许能够触及通用人工智能的门槛。

然而，仅仅通过增加处理芯片的数量来实现这一增长并不现实，因为这将超越全球现有的 GPT 能力。相反，我们通过研发更高效、更先进的芯片、算法和系统来达成目标。

计算的持续演变带来软件新形态

随着计算规模的不断扩张和计算方式的转变，当前已进入一个算力密集型的时代。处理器也从 CPU 转向 GPU，这预示着我们z需要探索新的技术和解决方案。因为需求在变化，历史总是在不断地演进中上升。

从大型机到 PC，再到移动设备，计

算需求一直在不断演变。大型机更侧重于集中式的强大算力，而 PC 则实现了计算力的分散化，让每个人都能拥有自己的算力装置。

随着技术的迭代，PC 从单一的 CPU 发展到多核 CPU，甚至孕育出了 GPU 这样的新型处理芯片。到了移动设备时代，计算变得更加移动和便捷，包括地理位置信息的处理等功能也变得日益重要。

因此，我们可以预见，随着计算的持续演变，必将催生硬件的新形态，进而带动软件的新形态。计算领域有一个恒久不变的发展逻辑。每当计算需求发生转变，无论是从简单到复杂，从单机到并行，还是从云端到边缘计算，硬件形态都在持续演进。而在硬件革新的基础上，软件形态也紧随其后，尽管它们始终以“操作系统”这一标签呈现，但其核心作用在于有效整合新硬件形态，满足时代所需的计算能力。

为 AI 而生的“新电脑”

如今，我们正站在一个全新的历史节点。构建智算中心或集群时，是否应该延续传统的 PC 时代思维？像在 CPU 主板上插入 GPU 卡，再将其置于铁皮盒子中通过网络连接，这种做法无疑带有浓厚的历史痕迹，正如铁轨宽度沿袭自古罗马马车，这些传统做法虽有其历史渊源，但未必代表最优方案。

在计算领域，从 PC 到现代的服务器，都承载着过去的技术痕迹。当计算重心从 CPU 转向 GPU 时，我们是否应该重新审视这些传统架构？例如，将 GPU 置于

一个大型池中，通过 DPU 进行高速连接，或在 CPU 与 GPU 之间设置大容量内存作为缓存，这些都是值得探讨的问题。

回归第一性原理，如果计算的核心目标是为大模型提供训练、调优和推理支持，那么我们需要一个全新的计算平台。未来的计算机可能不再是我们熟悉的 PC 式服务器集群，而是集成了 GPU、CPU、DPU 以及各类协处理器的混合体。这种新型计算机将为大模型驱动的 AI 应用提供强大的计算支持，其芯片间的组织方式也将从网络连接到内存缓冲等方面发生深刻变革。

以 GPU 为核心的芯片发展仅是起始，接下来，我们要根据 AI 的需求，在系统上高效地组织这些计算能力。正如发动机的发明至关重要，但将其应用于不同类型的车辆，才开启了更广阔且充满活力的市场。同样，将各类芯片整合起来，满足不同场景和行业的需求，这将孕育出无数的产业机遇。

AIDC OS 迎接“软件定义算力”的时代

英伟达 H100 的 SuperPOD 产品，以及 IBM、英伟达、戴尔、HP 等推出的类似大型集成设备——我们称之为 POD，集成了众多 GPU、CPU 及内存存储的组合，正预示着硬件发展的新篇章。而新的计算平台诞生后，其操作系统成为我们关注的焦点，九章云极 DataCanvas 将其命名为“AIDC OS”，预示着软件定义算力的新时代的到来。

为何如此说？当拥有如此丰富的芯

片和计算能力，并将其集成于大型设备中，如何定义和组织这些算力，便成为关键。而这在很大程度上取决于软件与硬件连接方式的协同配合，以确保外部用户能够充分利用这些算力。正如 PC 和 Service 的形态是经过历史发展而成，新的计算平台也将经历一系列变革，最终确定其形态。

在这一过程中，软件的驱动力和定义作用至关重要。正如鼠标的交互方式定义了 PC 的形态，手机的触摸方式影响了手机的设计，AIDC OS 也将在新的计算机时代中，定义如何组织和管理算力，以满足用户的需求，我们称之为“软件定义算力”的时代。

然而，目前仍处于硬件堆砌的阶段。尽管将各种芯片和网络连接设备放在一起，在某种程度上提供了可用的算力，但离“好用”和“经济”仍有很远的距离。缺乏专业技能和团队，很难实现垂类大模型的落地、训练和推理部署。因此，我们需要在软件层面进行更多的创新和优化，以实现更为高效、便捷和经济的算力应用。

总之，随着硬件和软件的协同发展，我们将迎来一个全新的计算时代。在这个时代里，软件将发挥更为核心的作用，定义和管理算力，以满足不断增长的 AI 需求。而我们，作为这个时代的参与者和见证者，有责任也有能力推动这一变革的进程。

AIFS 基础软件是核心能力

谈及现有的技术进展，大家可能会联

想到众多 GPU 与新型设备的涌现。我们将这些设备整合成 AIDC，即“智算中心”，进而为用户提供服务。其与云计算之间存在显著的差异，这不仅仅是一个自然的延伸，更是一次跨越式的变革。

过往，我们提及的 CPU 云中的算力，通常被称为“通用算力”。它是从下至上的资源驱动视角，即以拥有的众多 CPU 机器为基础，通过虚拟化技术，将这些机器切割成多个虚拟机，进而将硬件资源分时段、分片地提供给用户。然而，虚拟机内的工作负载类型却是五花八门，如网站、APP 后台、流媒体直播等，这使得管理变得极为复杂。简而言之，我们主要关注资源管理，而至于资源如何被使用，则完全取决于用户。

在 GPU Cloud 中，情况发生了根本性的变化。这里采用的是从上至下的视角，因为 GPU Cloud 的工作负载相对固定，主要集中在大模型的训练、优化和推理上。这使得资源管理和驱动方式需要更加高效，以满足上层服务的需求。

这种根本性的变化导致了软硬件协调逻辑的显著差异。在 CPU 中，资源驱动主要依赖于虚拟化技术，而在 GPU Cloud 中，我们更加注重 AI 使用能力，涉及高性能分布式计算、存储和网络等一系列技术。

计算需求决定了计算形态，而计算形态又影响着硬件的设计。硬件进而影响了软件的发展，软件则成为整合这些元素的综合驱动力。智算中心的 GPU Cloud 与 CPU Cloud 是两个截然不同的领域，它们各自拥有独特的挑战和机遇。

站在历史节点，迈入计算的新世界

大学时期，我翻阅了比尔·盖茨的《未来之路》，其中他使用的 28 英寸液晶显示器让我印象深刻。那时，我买的是 14 英寸的飞利浦显示器，心中不禁感叹：比尔·盖茨真是领先时代。这本书让我们见证了个人计算能力随着 PC 的普及而大幅提升，数字化水平日新月异，进而开启了互联网连接的新纪元。

如今，我们再次站在了这样的历史交汇点。智算操作系统的出现，标志着我们正步入一个可以重新定义算力基础设施的新时代。它将塑造我们未来的电脑形态，释放并优化计算能力的组合，引领人类进入一个全新的计算世界。

这个新世界与当年的 PC 时代有着诸多相似之处：首先是硬件设备的革新，其次是这些设备之间的互联性，不仅限于单个智算中心内部，更能实现跨中心的算力调度。我们同样面临着单一计算单元能力的巨大提升，多种处理器和连接技术的组合涌现，以及新型操作系统的诞生。

这些新操作系统为计算带来了更多的可连接单元，正如 PC 时代的 Windows 和 IE 网络版，它们共同推动了互联网的蓬勃发展。如今，从计算的角度看，我们又站在了新的历史起点。人类即将迈入计算的新纪元，而智算操作系统正是引领我们走向未来的关键。

✎ 责任编辑：杜玢翰 dubh@staff.ccidnet.com