White Paper

为什么在工作站上开发部署 AI 技术行得通?

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

IDC 观点

AI 已成为一项可为所有行业带来独特优势的重要能力，运行 AI 所需的硬件也在快速发展。技术行业往往非常关注先进 AI 模型正在经历的指数级规模增长。人们讨论着 AI 训练和推理方面的许多需求，包括数百亿计的参数、降低精度、扩大内存、高性能计算 (HPC)，以及大量的加速服务器。但在现实生活中，这种超大规模的 AI 计算并不多见，在企业领域尤其如此。

目前，许多企业都在大力推行不需要超级计算机的 AI（包括生成式 AI）计划。事实上，许多 AI 开发工作（以及越来越多的 AI 部署工作，尤其是在边缘位置）都是在功能强大的工作站上进行的。工作站在 AI 开发和部署领域具有许多优势。它们让 AI 科学家或开发人员无需再耗时费力地协商服务器使用问题；它们提供了 GPU 加速功能，这一点在数据中心依然不易获得基于服务器的 GPU 的当下尤为难得；它们相较于服务器十分经济实惠；只需一次性支付小金额费用，不会像持续购买云实例那样快速累积高额费用；此外，使用工作站时敏感数据安全地保存在本地，因而可以让用户安心。这样一来，科学家和开发人员也不用那么焦虑：他们不会因为只开展一些 AI 模型试验便产生高昂成本。

IDC 发现，在 AI 部署方面，边缘的增长要快于本地和云环境。同样，在边缘位置上，工作站作为 AI 推理平台在发挥着越来越重要的作用，往往甚至不需要 GPU，而是在软件优化的 CPU 上执行推理。在边缘位置基于工作站进行 AI 推理的应用场景正在快速增加，包括 AIOps、灾难响应、放射学、石油和天然气勘探、土地管理、远程医疗、交通管理、制造工厂监控和无人机。

本白皮书探讨了工作站在 AI 开发和部署中发挥的日益重要的作用，并简要讨论了戴尔 AI 工作站产品组合。

情况概览

AI 爆炸和基础架构影响

全球组织参与的 AI 项目数量在迅速增长。在所有行业中，已经有许多任务由部分或完全依靠 AI 模型驱动的软件执行。IDC 在多个层面跟踪 AI 的发展现状，其中一个重要指标是企业和云服务提供商为了开发和运行 AI 而在服务器方面预计支出的金额。到 2026 年，这一数字将达到 346 亿美元，占全球服务器总支出额的近 22%。

但除了服务器，还有其他支出。大量的 AI 准备、开发、原型设计工作 — 以及越来越多的*部署*工作 — 是在工作站上进行的。随着大大小小的组织发现可以通过为其应用程序注入或多或少的 AI 功能来实现新的商机，AI 模型试验的数量开始飞速增长。而性能强劲的工作站因其即时可用且更靠近数据来源，非常适宜被用来进行这类试验。

我们部署 AI 算法已经有了几十年的时间，缘何 AI 现在才突然火爆起来？主要原因是，让神经网络这种大火特火的 AI 算法类型取得重大进展的两个基本条件，在过去几年间均已具备：其一是可轻松获得数量庞大、价格低廉且类型多样的数据，如非结构化数据和半结构化数据；其二是通过并行处理模式增强了线性计算能力，从而能够在可接受的时间范围内处理这些神经网络。由于这两个基本条件均已具备，数据科学家在开发神经网络方面取得了巨大进展 — 这些神经网络会自动学习如何执行越来越庞大的任务。传统机器学习 (ML) 仍然适用于处理文本和数值型数据，但对于视频、音频、语言等类型的数据，深度学习 (DL) 更为有效。

传统机器学习模型通常可以利用工作站 CPU 进行开发，这些工作站最多只有几十个核心，而神经网络需要使用协处理器在数千个核心上执行并行处理。其主要原因是：在 ML 中，功能提取和分类是手动过程，而在 DL 中这一过程是自动化的，要求使用大数据集通过不断重复对模型进行训练。目前通常使用的协处理器是 GPU，但由初创企业开发的新型 AI 专用处理器也开始上市。这种类型的加速使用独立协处理器进行并行处理，彻底改变了服务器和工作站市场，从而使 IDC 所说的大规模并行计算应运而生。

2022 年，加速服务器的全球市场规模达到 218 亿美元，到 2026 年将增长到 434 亿美元，运行 AI 的加速服务器占到了其中的 57%。与此同时，用于工作站的独立 GPU 的销量在 2022 年达到了 640 万；IDC 估计，随着 AI 开发需求的增长，科研或软件工程用工作站的市场规模增长趋势也将日益凸显，到 2026 年将增加到近 20 亿美元。

AI 开发的各个阶段

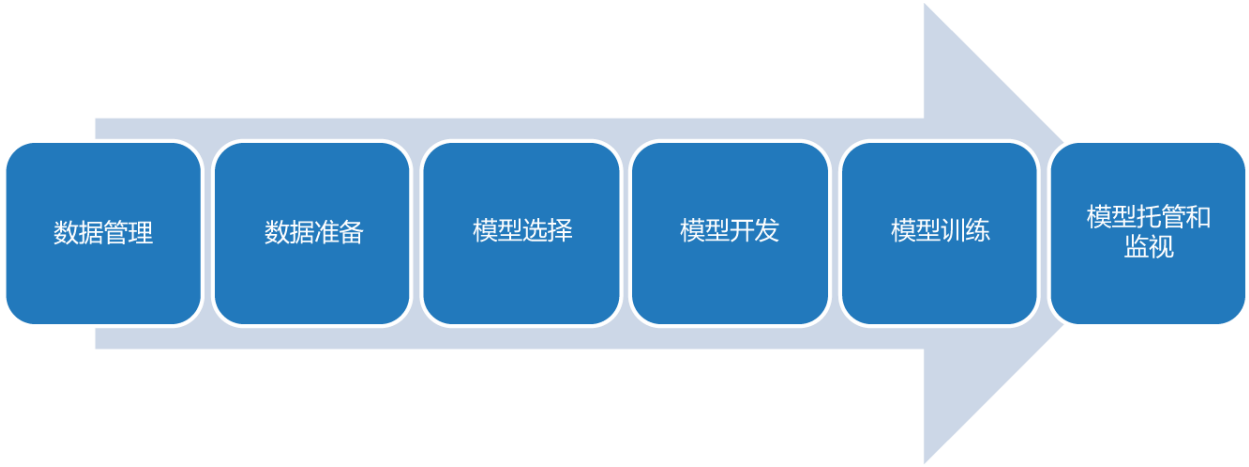
如前所述，由于不断扩大的数据类型和数据量以及新计算思路的出现，神经网络变得切实可行。此等式中的第一部分，亦即数据量和数据类型，并非一个无足轻重的部分 — 据某些统计，一项深度学习 AI 计划中多达 80% 的工作量都在数据的管理和准备方面。需要先接收、管理和准备数据，然后才能开始模型设计和训练。按照 IDC 的说法，以下是 AI 开发的各个阶段（参见图 1）：

- **数据管理：**从组织在数据中心、边缘和云环境等不同位置接收、生成和/或获取的海量数据中，识别和管理 AI 模型相关数据（此数据可以是任何类型的，可以是事件驱动型或流式传输数据，而且其中有许多可能要求某种类型的治理。）
- **数据准备：**将数据（文件、数据块或对象）存储在数据仓库或数据湖中，对其进行净化处理，确保数据完整且质量高，然后将其转换为一种使其可在 AI 模型中使用的形式 — 例如，使用 Spark 或者像 Pandas 这样的工具
- **模型选择：**根据错误率和/或性能，确定哪个模型能够更好地执行其编程所针对的 AI 任务
- **模型开发：**使用诸如 XGBoost、LightGBM、GLM、Keras、TensorFlow、PyTorch、Caffe、RuleFit、FTRL、Snap ML、scikit-learn 或 H2O 这样一些框架设计 AI 模型
- **模型训练：**使用足够多的处理器和/或协处理器核心在计算基础架构上训练模型以实现并行处理（越来越多地，还包括解释、验证和记录模型的决策以确保公平性、可问责性和透明性的这一能力）（这其中包括原型设计 — 通过在其上运行推理来测试经过训练的模型。）
- **模型托管和监视：**将模型部署在生产环境中以执行其设计的目标任务，通常称为“AI 推理”，并监视其表现

通过与数据中心、云或边缘基础架构结合，工作站可在这六个阶段中的任意一个阶段发挥重要作用。

图 1

AI 开发的各个阶段



资料来源：IDC，2023 年

在工作站上开发 AI 模型

工作站与个人电脑

众所周知，个人计算机 (PC) 的性能不足以支持 AI 开发。数据科学家和 AI 开发人员通常会参与其组织的具有战略重要性的项目，而无障碍工作至关重要。与 PC 相比，工作站往往有更可靠的性能，因为它们通常使用更高性能的组件构建，并针对在其上运行的软件进行了优化。

这些组件包括：

- **高级处理器：**例如，英特尔至强可扩展处理器。
- **强劲 GPU：**例如，包括 NVIDIA RTX 6000 Ada 在内的 NVIDIA RTX 专业级 GPU。
- **更多存储：**某些工作站可提供高达 60 TB 的容量，而 I/O 速度也往往明显高于 PC。
- **更多内存：**工作站现在提供高达 6 TB 的内存。
- **散热：**高性能组件将产生大量热量。数据科学家需要工作站具备强大的散热能力，以防止过热并保持良好性能。
- **网络接口卡 (NIC)：**数据科学家需要处理存储在远程服务器上的大型数据集，因而能够快速、高效传输数据的高速网络接口卡将至关重要。
- **显示器**：高质量的显示器对于数据可视化任务非常重要。数据科学家需要具有高分辨率、色彩精度和大屏幕尺寸的显示器。
- **纠错码 (ECC) 内存：**ECC 检测并纠正常见的内部数据损坏类型，防止在长时间的 AI 训练运行过程中因硬错误（坏位）或软错误（翻转位，导致错误的值）而导致蓝屏；ECC 还确保结果的准确性，这在医疗保健等生命攸关工作中是一项至关重要的要求。
- **专用硅片：**例如，英特尔 Movidius 视觉处理器 (VPU)，它们是用于在零售、安保和工业自动化等环境中使用的计算机视觉和边缘 AI 应用程序的并行处理协处理器。工作站中还使用了 FPGA，例如用于运行财务应用程序。

- **优化软件**：例如，OneAPI 是英特尔的标准化编程模型，旨在方便使用 CPU、GPU、FPGA 和其他加速器开发和部署以数据为中心的工作负载；CUDA 是 NVIDIA 的并行计算平台和应用程序编程接口，旨在利用 GPU 运行通用工作负载。

AI 开发：CPU 与 GPU

工作站可在 AI 开发的各个阶段中使用，并且通常具备多种功能。尽管强调使用 GPU 进行并行处理，但在工作站上开发 AI 模型时，CPU 起着关键作用。CPU 和 GPU 一样，可以用于数据操作，当然也可以用于开发传统的 ML 模型。CPU 还可用于数据探索，亦即使用数据集的直观表示来了解数据特征。

在 DL 训练中，在实际训练期间，随着 GPU 接管处理任务，主机 CPU 的角色会略有削弱，但即便在此时，CPU 仍将继续用作关键软件（如操作系统或 CUDA）的处理层，并用于在 GPU 之间或与其他硅片之间协调进程。此外，在使用工作站在生产环境中运行 AI 模型的情形中，CPU 越来越多地充任 AI 推理引擎这一新角色。IDC 预计，到 2024 年，在 AI 推理基础架构方面的开支将超过在 AI 训练基础架构方面的开支，而且很大一部分 (39%) 推理任务将在主机 CPU 上进行。

工作站与服务器：一种共生关系

对于大多数组织而言，在考虑为 AI 开发部署工作站、本地服务器、云实例或这三者的任意组合时，实用性是他们的经验法则。在 AI 项目的不同开发阶段，工作站、服务器和云实例之间存在共生关系。

与数据中心服务器相比，工作站的优势在于数据科学家可以随时随地工作 — 在当前新冠疫情肆虐的情况下这是一个重要考量因素，但在正常形势下也是如此。他们还可以在工作站上随心所欲地进行 AI 模型试验，并视需要进行多次迭代。现代工作站搭载性能强劲的 GPU，通常可以让迭代过程更具互动性，数据科学家可以获得即时反馈和结果，无需请求访问服务器或面临其他数据中心限制。工作站让他们能够灵活地将计算移到更接近数据的位置，而不是相反，从而节省带宽、减少网络拥塞和增加吞吐量。此外，工作站可以针对不同的需求进行配置：例如传统的 ML 任务，或者 DL 成分更多的工作。

此外，尽管加速服务器市场有了显著增长，但加速服务器在企业数据中心中仍不能普遍提供。在本白皮书撰写之时，企业数据中心中平均有 4% 的服务器是加速的，这意味着许多组织无法在现成的本地 GPU 上开发或运行 AI。为此原因，加速工作站是 AI 开发的一个有用的替代方案。

高度加速的工作站现在足够强大，只要 AI 模型不是太大，它们就可以执行 DL 训练，从而无需在服务器上进行训练。在配备 GPU 的工作站上完成训练的模型可以部署在没有 GPU 的工作站或服务器上，从而利用 CPU 中的推理功能。软件技术 — 如英特尔的 DL Boost 和 OneAPI — 可以运行在 CPU 上进行的 AI 推理，从而使数据中心中已经部署的非加速服务器能够支持 AI 应用程序。

工作站与云

云计算彻底改变了组织对基础架构、数据和应用程序的看法。凭借近乎没有限制的可扩展性，云使开发人员能够即需调配资源，从而有可能通过减少束缚和限制而加快创新步伐。从表面上看，云似乎是 AI 开发的理想模式。

然而，实际情况并非总是如此。实际上，据 IDC 研究显示，组织越来越多地将某些工作负载从公有云迁回本地基础架构。这样做有以下几个驱动因素：

- **云可用性**：云服务用户都经历过服务中断的情形 — 可能是由于云提供商自身的问题，或者是由于超大规模数据中心与终端用户之间的网络连接失败。在这些情况下，用户需要完全依靠服务提供商来解决问题，而在此期间正常工作会陷于停顿。
- **安全性和合规性**：在许多行业中，公司治理策略决定了数据可以传输和存放到哪里，而这限制了云服务的使用。政府法规，如欧洲的 GDPR 和加利福尼亚州消费者隐私法案等，也强制规定了有关数据主权的规则。

- **成本**：组织往往会低估云服务费用的增长速度，尤其是对于那些需要高性能计算能力和大量存储空间的工作负载。云经济性需要计量所有类型的资源消费，包括将数据传回现场基础架构。
- **试错压力**：多数 AI 计划一开始都要进行大量的试验，开发过程中出现模型失败在所难免；在此过程中，随着云资源使用费不断累积而 AI 科学家和开发人员又迟迟拿不出可执行的结果，他们会有一种心理压力。

工作站可消除这些限制，同时仍可利用云原生技术，例如基于微服务的体系结构和 API 驱动的自动化。工作站因此提供媲美数据中心服务器的优势：

- **随时随地工作**：通过消除对公有云的依赖，断连工作模式现在成为可能。许多高安全性环境需要与公用网络进行物理隔离，而 AI 工作站正好可以满足这一需求。本地资源还减少了对昂贵的网络连接的需求。
- **数据本地化**：物联网设备和其他互联设备的激增正在推动边缘位置的数据呈指数级增长。在许多情况下，使用专用工作站让计算资源与数据同在一处是有意义的。这也通过限制数据移动而满足了许多合规性要求。
- **自由试验**：AI 模型的训练和优化是一个迭代过程，往往会有一些试错元素。开发人员需要能够自由地进行试验，不会因为可能会产生额外的服务费用而缩手缩脚。工作站还为自定义配置提供了更多灵活性。

在自定义配置方面，可以相对轻松地比较工作站的价格与云部署的成本，因为大多数云服务提供商都可以立即给出终端用户所需部署配置的大致成本。例如，单个配有一个 NVIDIA T4 和一个 375GiB SSD 存储实例的常规虚拟机 (VM)，每天使用 8 个小时，每周 5 天，一家主流云提供商给出的成本估算将是 140 美元。将虚拟机、T4 和 SSD 翻倍，每月成本将增加到 365 美元。保持使用两个虚拟机，但将 T4 翻倍到 4 个，存储容量增加到 4 个 375GiB，并在此环境上进行一次全时训练，每月成本将高达 2700 美元。所以，公平地说，利用云进行 AI 开发的成本可以轻松达到每年数万美元，大大超过高端工作站的年度折旧值。

在工作站上进行 AI 原型设计

与本地服务器和云相比，工作站在 AI 模型的原型设计方面具有明显的优势。数据中心中的服务器可能处于满载状态，或者可能非常需要用于关键任务，因此不适合用于进行 AI 原型设计和测试；而且，如前所述，云实例在真的用作测试环境时，可能会迅速导致成本超支。工作站使 AI 科学家和开发人员不必再协商服务器使用情况，而且再也不用在整个原型设计阶段期间时刻担忧云账单不断增加这一问题。它们以较低的一次性成本带来了完全的自由：可以随时随地进行原型设计，而又不会产生额外成本。

在工作站上部署 AI 模型

虽然多年以来在工作站上开发 AI 模型是一项常见的战略，但 IDC 发现，在工作站（通常在边缘）上*部署* AI 模型的应用场景越来越多 — 换言之，就是通过让工作站运行 AI 模型推理，而将生产环境中的 AI 模型放在工作站上。作为一个 AI 部署位置，边缘环境在迅速增长 — 从年度硬件开支上看，服务器部署量从 2020 年到 2024 年增加三倍；而且随着终端用户发现工作站在边缘位置的优势，工作站的部署量也没有落后多远。

IDC 将边缘定义为一种分布式计算模式，其中包括将基础架构和应用程序部署在集中式云环境和本地数据中心之外，根据需要尽可能靠近数据生成和消费的位置。这包括远程和分支办公室，以及一些特定于行业的位置，如工厂、仓库、医院和零售店。

数据密集型和计算密集型工作负载越来越多地部署在本地或边缘位置。这样做是为了减少公有云存在的固有限制，例如上传大型数据集所需的时间以及执行 AI 训练时无法预料的成本 — 尤其是在需要大量数据科学试验的情形中。

据 IDC 研究显示，边缘是 AI 部署增长飞快的领域之一，预计组织在 2023 年将投资 29 亿美元用于在边缘位置采用 AI 计算，并且这个数字在 2026 年将增长到 69 亿美元（参见*《Worldwide AI Hardware Forecast, 2022–2026: Strong Market Growth for AI Compute and Storage》*，IDC #US49671722，2022 年 9 月）。此外，边缘日益成为企业部署工程和技术等 HPC 工作负载时的选择。目前，企业投资了近 10 亿美元用于在边缘位置部署此类工作负载，这个数字在 2027 年将达到 24 亿美元（参见*《Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs》*，IDC #US50525123，2023 年 4 月）。这些是部署 AI 工作站较为有意义的领域。

当在边缘位置在工作站上部署 AI 模型时，并不是像 AI 开发时那样总需要高端 GPU。配置较低的 GPU 就能够执行 AI 推理，而且在很多情况下完全不需要 GPU。在这些情况下，CPU 就足以胜任推理任务，特别是在搭配使用像英特尔 DL Boost 这样的优化软件时；该软件是英特尔微处理器上的一组指令集功能，旨在加速 AI 工作负载 — 包括 AI 推理。据英特尔表示，与上一代产品相比，使用支持英特尔 DL Boost 的第 4 代英特尔至强可扩展处理器可将 INT8 实时推理吞吐量提高 1.45 倍（以 BERT-Large SQuAD 为例）。这还有助于使工作站更适合部署在边缘，因为在此类环境中，电力、移动性和散热管理等考虑因素要求更小的处理器功率。英特尔 Movidius Myriad (M2) 正好在此功率范围之内，而这得益于其 12 瓦的低能耗。

在工作站上部署 AI 的应用场景

在若干情形下，非常适合在本地部署的工作站上部署 AI。这类情形的常见特征包括有大量机器生成的时间序列数据和非结构化数据，如视频流和图像。这类情形还包括：相关领域专家需要通过人工解释来增强 AI 模型。

示例包括：

- **AIOps：**随着 IT 系统规模和复杂性的增加，人们越来越需要从被动事件管理转变为主动监控。随着基础架构和应用程序分布到边缘位置，人们尤其需要这一转变，因为在这些位置很少甚至没有技术人员。通过建模一个正常性能基准，就能够识别性能异常并自动执行修正步骤。
- **灾难响应：**在紧急情况下，急救人员必须快速评估情况、跟踪关键设备，并部署资源来帮助那些急需帮助的人。这通常必须在没有网络连接的环境中进行，这就要求有一个能够聚合数据馈送、根据 AI 模型进行推理并自动与关键人员进行通信的本地工作站。
- **放射学：**成像技术的进步导致单次扫描生成的数据量增加了，这些数据需要保留在现场，以便及时进行分析。使用数百万个以前的示例训练过的 AI 模型可以比人眼更准确地识别模式，从而提高准确率。
- **石油和天然气勘探：**上游石油和天然气公司使用遥测、地震和图像数据的组合来定位自然资源矿藏，选择钻井位置，并在生产过程中优化设备的性能。这往往要求在只能使用昂贵的卫星通信的区域能够执行信息分析。
- **癌症研究和药物开发：**科研型医院和学术界的研究人员通过综合运用 AI 技术和自然语言处理技术，帮助肿瘤科医生确定对患者更有效的个体化癌症治疗方案。他们还将机器学习技术与计算机视觉技术相结合，帮助放射科医生更好地了解患者的肿瘤进展情况。他们使用算法来帮助理解癌症的发展进程，并确定哪种治疗方案能更有效地抗击癌症。
- **保险索赔评估：**人工理赔需要耗费大量人力，并且容易出现人为错误。能够评估索赔有效性的 AI 使保险理算员能够专注于需要更多调查的案例，从而降低成本。这样可以提高运营的总体吞吐量，而又不会牺牲准确性。

- **远程医疗**：通过基于来自可穿戴设备的实时生命体征数据定制个人治疗计划，AI 正在提高患者康复率。此信息将与患者病历和类似病案知识库相结合。这在更依赖远程医疗的农村地区尤为重要。
- **零售安保（防盗）**：使用应用于视频流的实时分析功能来预测可能发展成犯罪活动的人类行为。这通常需要将多个视频馈送拼接在一起，以跟踪一个人在商店内的动作。考虑到识别重大事件的时效性，这是个过程最好在本地运行。
- **交通管理**：负责交通运输的政府机构在逐渐增加对 AI 的使用，以协调红绿灯和数字路牌，从而改善车流并保障公民人身安全。这要求结合使用多项数据输入 — 包括视频摄像头和来自道路传感器的遥测数据 — 以优化车流模式。
- **制造车间监控**：对于车间经理而言，确保关键流程的正常运行和不误工期是至关重要的。这就要求要对关键设备进行预测式维护，自动检测缺陷，并进行现场和场外的供应链优化。这是一个 AI 可以帮助人类操作员提高性能并同时维护安全标准的领域。
- **无人机**：自动分析无人机捕获的图像，这使人们能够在以前无法达到的规模上监视一系列状况。这将对天然气和电力设施检查、保险调查、搜索和救援工作、精准农业以及渔业和野生动物保护区的维护产生重大影响。
- **日常办公环境**：日常办公环境越来越多地受益于 Microsoft Copilot 等 AI 生产力工具。
- **可再生能源**：可再生能源站点（如风力发电场、水力发电站和太阳能电站）需要进行实时监测、维护和数据收集工作。这些工作需要在站点本地完成数据的生成和分析。

适用于 AI 的戴尔工作站

戴尔针对不同级别的 AI 部署和/或实施工作提供一系列专门工作站。这类工作站被全部归入到“数据科学工作站 (DSW)”产品品牌下。本节将简要介绍相关规格，探讨不同 AI 角色/应用程序（比如数据科学家）以及戴尔 DSW 技术产品的优势。这些支持 AI 的数据科学工作站专为数据科学家而设计。新款 Precision 数据科学工作站利用 AI 功能微调设备，让数据科学家在使用常用应用程序时获得更佳性能。数据科学家将可以更快速完成重要工作。此外，戴尔 Precision 工作站经独立 ISV 测试与认证，可顺利运行戴尔客户完成日常工作所需的高性能应用程序。

戴尔工作站的优秀之处

戴尔 Precision 工作站搭载 NVIDIA RTX GPU，旨在提供出色的可扩展性和性能，助力组织顺利推进分析和 AI 计划。Dell Technologies 提供全面的硬件解决方案。这些解决方案经过优化，可运行新款 AI 软件。

- **强大的硬件配置**：戴尔 Precision 工作站提供多种强大的硬件配置，包括多核处理器、大容量 RAM 和多种 GPU 选项。这些组件为 AI 任务提供了必要的计算资源，助力训练和推理过程高效进行。
- **可扩展性和可定制性**：戴尔 Precision 工作站具有可扩展性和可定制性，允许用户根据特定 AI 需求自定义硬件配置。凭借这种灵活性，工作站可针对 AI 工作负载的特定需求进行优化。
- **认证和优化**：戴尔携手 NVIDIA 对 Precision 工作站进行认证，确保工作站与 NVIDIA RTX GPU（包括 NVIDIA RTX 6000 Ada Generation 显卡）相兼容并可提供出众性能。此项认证可确保用户将配备 NVIDIA RTX GPU 的戴尔 Precision 工作站用于执行 AI 任务时，可获得无缝集成与优化性能。
- **强大的处理能力**：搭载英特尔处理器的戴尔 Precision 工作站可提供执行 AI 任务所需的计算能力。这些工作站配备多核处理器并具备较高的时钟速率，可提供 AI 工作流中训练和推理过程所需的性能。
- **软件和工具支持**：戴尔 Precision 工作站预装了支持 AI 开发和部署的软件和工具。这包括可充分利用 NVIDIA RTX GPU 强大性能的优化软件堆栈、AI 框架和库，帮助用户轻松启动 AI 项目。

此外，以下技术也让戴尔工作站在众多同类产品中脱颖而出：

可靠内存技术

戴尔在 ECC 之上提供了一种称为 Reliable Memory Technology Pro (RMT Pro) 的技术，旨在帮助更大限度地增加正常运行时间。它与 ECC 内存配合工作，以实时检测和纠正内存错误。据戴尔称，RTM Pro 技术可以在 DIMM 仍在全面工作的情况下防止系统再次访问坏内存，从而几乎完全消除了内存错误。系统重新启动后，RTM Pro 将隔离故障内存区域并对操作系统隐藏该区域。因此，AI 数据科学家和开发人员将不会遇到因为坏内存仍可以寻址而导致的持续崩溃，以此大幅提升工作效率。

Dell Optimizer for Precision

戴尔还在其多数工作站产品中包括了 Dell Optimizer for Precision，后者将自动调整系统设置，这样工作站将能够以尽可能快的速度运行各种常见的商业应用程序。这提高了数据科学家和开发人员的工作效率。该工具还可以为 IT 部门创建关于处理器、存储、内存和显卡利用率的实时性能报告。DOP 尚不能在 Linux 上运行，因此主要用于部署 AI，因为 AI 开发工作往往是使用基于 Linux 的开放源代码软件完成的。Dell Optimizer for Precision 还提供了 ExpressSign-in 智慧感知登录、ExpressCharge 快速充电（在移动设备上）、Intelligent Audio 智能音频以及报告和分析工具，以帮助微调工作站。

挑战/机遇

对于企业

IDC 观察到，AI 市场正在出现两极分化。一方面，一些企业正在部署数据策略来保持竞争力，包括大规模采用 AI 技术。比如，企业可能看到某些同行利用排名前 100 的超级计算机等企业级 AI 基础架构取得了非凡的成果。另一方面，一些企业的日常现实却是不得不依靠有限的预算和性能欠佳的硬件，利用数据中心或云端服务器尝试小型 AI 计划。

对于许多企业说，第一种情况与他们无关，第二种情况才是他们的现实。对于他们来说，挑战在于为他们的 AI 数据科学家和/或开发人员提供适当的工具，让他们能够及时执行 AI 训练，而又无需在云实例或 GPU 加速数据中心服务器方面花费巨资。IDC 认为，通过为其科学家和开发人员提供功能强大的 GPU 加速工作站，这些企业的需求可以得到满足。

对于戴尔

市场上存在一种误解，那就是 AI 开发和部署需要昂贵的加速服务器硬件 — 甚至是服务器群集。对于包括数十亿个参数的最大的 AI 算法来说，可能确实如此，但大多数企业没有在开发如此庞大的算法。他们在通过其 AI 计划执行一些有用、有影响且可管理的工作，而且许多企业没有意识到这种常见规模的 AI 模型可以在工作站上开发和部署。戴尔面临的挑战是破除这一误解，向市场宣传使用其工作站产品组合完成这些工作的可能性。

同时，戴尔必须确保其工作站能够提供预期的性能，并且不会随着时间的推移成为技术瓶颈。这意味着需要进行快速、持续的创新，以便不让以适当方式使用工作站的终端用户（也就是，那些没有尝试运行包含几十亿个参数的庞大算法的终端用户）失望。这也意味着，对于那些突然开始非常快地扩展其算法或者其算法确实变得非常大的客户，他们将可以从工作站无缝过渡到戴尔的 AI 服务器系列。当然，戴尔也将有机会为每家客户提供合适的解决方案 — 无论他们正在实施多大规模的 AI 计划。

总结

IDC 认为工作站当前被低估了，实际上在许多应用场景中，它们都可以作为用于 AI 开发和部署的主力机型。它们为 AI 科学家和开发人员提供了一个强大的 GPU 加速平台，与服务器相比资本支出更低，而且运营支出明显低于云实例，并且允许更自由地对 AI 模型进行试验。企业在制定 AI 计划时，如果 AI 计划不需要包含数十亿个参数的庞大算法（多数情况下不需要），应考虑为其 AI 团队配备工作站，以实现不受限制的 AI 开发和基于边缘的轻松部署。

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com