加速实现深度学习

专用于支持 AI 和 ML 工作负载的思科 UCS C480 ML M5 机架式服务器

在数据中心支持大规模人工智能工作负载

思科® 机器学习计算解决方案可以有效应对 IT 组织和数据研究人员面临的挑战：在满足机器学习 (ML) 工作负载需求的同时，将这些工作负载整合到企业数据中心。借助思科解决方案，您不仅能有力地支持大规模人工智能 (AI) 工作负载，而且能帮助相关人员以更智慧的方式提取数据，做出更好的决策。

真正适合深度学习工作负载的成熟方法

除了支持机器学习的思科 UCS® C480 ML M5 平台外，我们目前还提供与 AI 生命周期各个阶段紧密对应的一整套计算服务，包括：在接近网络边缘的位置执行数据收集和分析；在数据中心核心准备数据并进行训练；对 AI 核心加以实时干预。不仅如此，我们基于云的管理方法还便于您将高速计算扩展至分布式 IT 环境的适当位置。

- **按需配置性能和容量：**思科 UCS C480 ML M5 不仅提供灵活的 CPU、内存、网络和存储选项，还具备出色的 GPU 加速性能
- **从容实现机器学习**：软件生态系统与验证解决方案相结合，让机器学习更简单
- **简化操作**：利用 Cisco Intersight™ 平台轻松将高速计算扩展至所需位置，简化您的操作

优势

- **获得敏锐洞察力，加快决策速度：**借助高性能系统支持不断变化的数据密集型工作负载。
- **轻松实现机器学习模型融合：**经过验证的解决方案确保您更迅速地实现更可靠的部署。
- **降低成本和复杂性：**基于云的管理确保您的整个计算环境中的操作一致而统一。

AI 技术在各行各业的应用

- **零售业：**预测购物模式；优化供应链；预防损失；消除结算长龙，打造“流畅的”购物体验。
- **医疗业：**快速筛选或分类 X 光片、CT 影像和皮肤影像；预测患者治疗效果；为治疗提供指导。
- **智慧城市：**更快地在视频中识别人脸、车牌号和可疑物品；观察汽车、自行车和行人的交通模式；监视是否有人进入安防区域。

我们为您提供贯穿整个 AI 和 ML 生命周期的支持

当 IDC 发现某个行业有 30% 的增长率时，很可能意味着，您的 IT 组织很快将迎来变革。目前，已有一些企业和政府机构开始借助人工智能、机器学习和深度学习应用来加快决策速度，提高决策质量。

但是，传统数据中心技术并非设计用于应对生产级 AI 和 ML 工作负载庞大的数据量、超快的速度和固有的易变性。这就需要一种与传统业务应用截然不同的应用形式。由于所需使用的数据量极为庞大，数据获取速度无比快速，应用的行为方式正在发生根本性变化。换而言之，应用现在是围绕数据而设计的，所以必须具备高性能的系统来适应这些新型工作负载。但是这对您的 IT 团队来说，会负担很重。面对不断变换数据源和软件工具的数据研究人员，IT 团队必须竭力紧跟他们的步伐，满足不断变化的基础设施要求。另一方面，数据研究人员则苦于难以将机器学习成果转化为富有竞争力的商业工具。

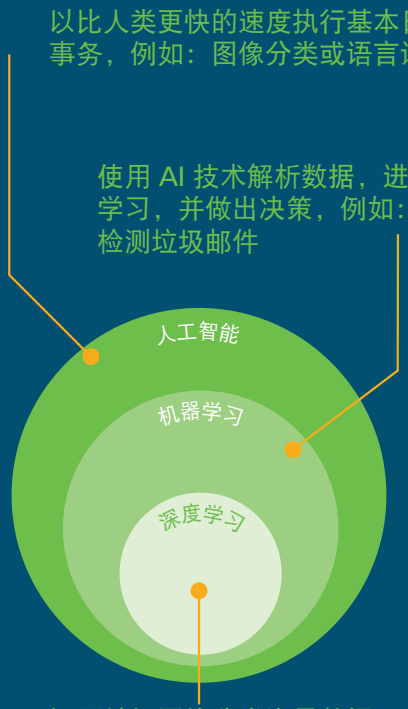
人工智能项目的生命周期始于获取和准备数据，然后是开发和测试机器学习软件。软件成型后，就要使用海量数据对其进行训练，并通过人工干预引导其做出正确决策（见图 1）。

支持深度学习工作负载

思科 UCS C480 ML M5 机架式服务器专门面向 AI 和 ML 生命周期中计算密集程度最高的阶段而设计，这个阶段称为“深度学习”。这款服务器搭载多个支持高速互联技术的 GPU，具备大容量存储，网络连接速度最高可达 100 Gbps。

图 1 思科专为 AI 和 ML 设计的全套服务器产品可为您的数据研究人员提供有效支持。

进化至深度学习



思科 UCS C480 ML M5 的产品特性（详见图 2）：

- **GPU 加速：**8 个 NVIDIA V100 SMX2 32 GB 模块通过 NVIDIA NVLink 技术互联在一起，可实现 GPU 间的快速通信，从而加快计算速度。NVIDIA 提供的 TensorFlow 性能为每模块最高 125 万亿次浮点运算 (teraFLOP)，所以每个服务器的总体 TensorFlow 性能为最高 1 千万亿次浮点运算 (petaFLOP)。
- **最新的 Intel Xeon 可扩展 CPU：**2 个 CPU（每个 CPU 最多具有 28 个核心），用于管理机器学习进程，并将计算任务分配至 GPU。
- **存储容量和性能：**对于深度学习应用，数据本地性至关重要。UCS C480 ML M5 最多可配备 24 个硬盘驱动器或 SSD，用于将数据存储在使用位置附近，并允许通过安装在中板上的 RAID 控制器进行访问。多至 6 个磁盘驱动器插槽可用于安装 NVMe 驱动器，从而提供业界一流的性能。
- **最高 3 TB 主内存：**基于高速 2666 MHz DDR4 DIMM。
- **高速网络连接：**配备 2 个内置万兆以太网接口，可加快数据传入和传出服务器的速度。
- **PCIe 扩展支持：**4 个 x16 PCIe 插槽可容纳 4 个 PCIe 交换芯片，用于提供高性能网络连接。可选择使用思科 UCS 虚拟接口卡 (VIC) 或第三方网络接口卡 (NIC)，提供高达 100 Gbps 的连接速度。
- **统一管理：**通过在思科 UCS 产品组合中添加这款新的思科 UCS C480 ML M5 服务器，我们可以在不增加管理复杂性的情况下更好地支持各种工作负载。

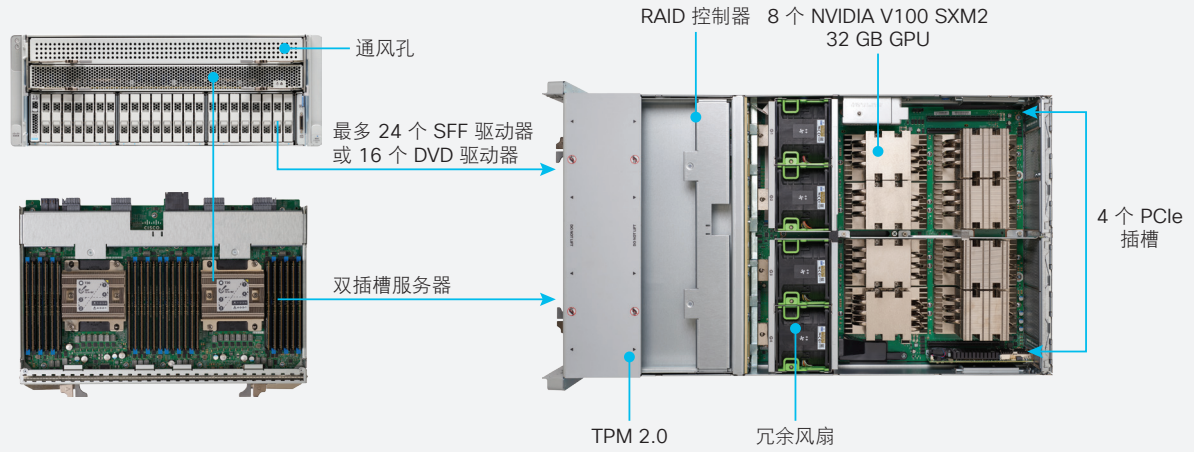


图 2 思科 UCS C480 ML M5 机架式服务器外观设计

相关详细信息

- cisco.com/go/ai-compute
- 思科 UCS C480 ML M5 产品手册

为什么选择思科

支持整个 AI 和 ML 数据生命周期

在帮助客户将不断变换的数据源整合至动态数据管道方面，思科拥有丰富的经验。通过将 GPU 集成到思科统一计算系统™（思科 UCS）和 HyperFlex™ 系统，我们可以帮助您将现有的大数据环境扩展为支持 AI 和 ML 的环境，让您能够利用思科 UCS 系统的适应能力和编程能力支持大规模 AI 工作负载。

消除孤岛

借助 Cisco Intersight，您可以轻松地在任何位置部署新技术，在所有位置彻底消除独立服务器孤岛（无论是在数据中心、远程位置和分支机构，还是在网络边缘）。

轻松实现机器学习模型融合

为了确保您信心十足地部署思科服务器产品，我们针对各种 AI 和 ML 解决方案进行了测试和验证。思科验证设计可为您提供涵盖解决方案部署各个环节的实践指南。有思科工程人员提供的部署验证保障，您不仅能提高解决方案的部署速度，还能降低部署带来的风险。

利用 AI 技术，您可以从数据中进行学习，以便更快地做出更好的决策。我们的计算解决方案组合可以满足 AI 生命周期所有阶段的要求。思科服务部门将与认证合作伙伴一起，帮助您妥善组合分析、深度学习和自动化功能，更快实现数据中心转型。

LE-66901-00 08/18