IDC

分析师简报

由戴尔赞助

企业界对 AI 处理基础架构的投资速度正在攀升。好消息是，在 2024 年，超过 50% 的 AI 处理基础架构无需配备专门的加速硬件，完全可以在标准服务器和以太网网络上运行。

企业界开启应用生成式 AI 技术的浪潮，生成式 AI 技术呈星火燎原之势

2024 年 2 月

作者：Brandon Hoff（网络和通信支持技术研究总监）和 Vijay Bhagavath（云和数据中心网络研究副总裁）

简介

IDC 研究对 2024 年以后的 IT 投资趋势做出了预测，并说明了我们认为可能对此趋势产生影响的一些驱动因素。技术部门领导及其业务部门 (LOB) 同仁可利用这份文档，作为制定战略时的指导。

运营团队一直在积极采集数据、创建数据湖，并利用云服务来存储数据。随着 ChatGPT 日益普及，生成式 AI 技术来到关键拐点，运营团队如今明白数据集可以发挥出巨大潜力。由于生成式 AI 技术的优势如今已是人尽皆知，投资者、高管和市场纷纷要求运营团队实施有效的生成式 AI 战略。有多种技术和众多选项可被用于改善企业的业务运营并提高员工的工作效率，其中包括生成式 AI、ML、数字孪生等。能否成功实施适当的技术，将成为决定运营团队乃至整个企业工作成败的重要 KPI。

概览

要点

开始规划生成式 AI 基础架构：

- 要想在应用 AI 技术方面先行一步，企业首先需要将数据从云端迁回本地，在本地保留至少一份数据。
- 企业应当投入时间和资源来了解生成式 AI、AI、ML 和数字孪生技术底层算法的商业价值，并根据价值大小来确定优先级。
- 分作三步：部署标准服务器和以太网网络，尝试应用生成式 AI 技术并进行评估；根据需要，利用标准以太网来扩大生成式 AI 技术的应用规模，以便将其用于企业级工作负载；通过在本地和异地基础架构之间重新分配工作负载，在 3 到 5 年内优化企业的资本支出和运营支出。

了解生成式 AI 技术的发展现状

随着 AI 技术越来越受到追捧，云服务提供商 (SP) 和企业界开始加快扩展基础架构的步伐。云 SP 购买了市场上出售的大部分 AI 加速器，并利用它们搭建了自己的 AI 基础架构。然而，这些加速器价格昂贵，导致云 SP 的 AI 服务成本随之上涨。AI 加速器包括 GPU、TPU、FPGA、ASSP 和 ASIC。这些云 SP 正在搭建用于处理不同公司密集型工作负载的“AI 工厂”。这些工厂大多部署在全球排名前列的大型 IT 环境，大约涉及九家公司。

与此同时，面向企业级工作负载的生成式 AI 基础架构可以使用标准系统构建，无需在这些系统上配备专门的加速硬件。据 IDC 预测，在 2024 年，超过 50% 的生成式 AI 系统无需配备专门的加速硬件。因此，人人都可以着手使用标准服务器和网络部署生成式 AI 基础架构。人们还可以视需要使用 GPU。部署 AI 基础架构有多种选择，同时有多种类型的生成式 AI、AI、ML 和数字孪生技术将以不同方式使不同公司受益。生成式 AI 软件堆栈通常可在标准服务器上顺利运行，因此使用标准服务器运行生成式 AI 可带来许多优势。投资购置本地标准基础架构的公司可比同行先行一步，更快速推进生成式 AI 计划。IT 团队需要评估各种生成式 AI、AI、ML 和数字孪生算法，从中找出更符合业务需求的算法。

优势

生成式 AI 技术和其他基础模型开启一代变革浪潮，推动辅助技术不断进步，并为非技术用户带来了许多强大的功能。生成式 AI 技术可提高效率和生产力、创造增长机会、降低成本，还可为采用此项技术的公司带来竞争优势。

通过搭建自己的生成式 AI 基础架构，企业可以快速将此项颠覆性技术应用于业务运营，并获得有关生成式 AI 技术栈的现场专业知识。通过优先投资构建基于标准企业服务器和以太网网络的本地初始生成式 AI 基础架构，采用此项变革性技术的企业将可在产品上市时间方面获得竞争优势。

考虑事项

利用生成式 AI，时不我待

考虑到 ChatGPT 和其他模型取得的出色成果，生成式 AI 技术在企业界引发的兴奋情绪可以想见。生成式 AI 技术可为企业创造价值，价值大小视企业采用的专有数据源和算法而有所不同。董事会、投资者和高管们开始提出各种各样的问题，希望了解如何利用生成式 AI 技术推动业务发展。

对于那些已经积累了大量非结构化专有数据的企业，生成式 AI 技术可利用现有专有数据创建原创内容，而这将有助于企业实现持续创新。企业应当采用逐步深入的方法，更好地理解生成式 AI 技术对业务的影响，以及如何在业务中深化应用。

在以太网上构建初始生成式 AI 基础架构

对于企业级工作负载，使用标准系统可以满足企业使用生成式 AI 技术的性能需求。此外，使用标准服务器和以太网网络构建生成式 AI 基础架构，还可方便企业同时使用企业操作系统、企业管理工具和企业网络管理工具。企业了解适合业务需求的 LLM 的计算要求后，就可以通过选择合适的生成式 AI 和 AI 加速器来提高计算性能。关键在于要为 AI 基础架构设计一个良好的体系结构。一个良好的体系结构可以支持少则数十、多则数千个 AI 计算节点。

业界针对生成式 AI 工作负载提供了多种不同的网络方案，其中以太网网络以其开放性获得了普遍接受并得到了多家供应商支持。企业可以使用目前已有的标准以太网网络，完成生成式 AI 群集的初始部署。

利用超以太网构建生成式 AI 基础架构

随着企业步入生成式 AI 技术开发的起步或高速发展阶段，企业或许应当考虑构建自己的可横向扩展型生成式 AI 基础架构。构建可横向扩展型生成式 AI 基础架构需要增加两个关键组件：数据中心 AI 加速器和 AI 网络。在典型的可横向扩展型生成式 AI 基础架构中，每台服务器上部署八个数据中心 GPU，并且对每个 GPU 都部署了一个高速 NIC 或 DPU 以实现高性能网络。

IDC 的 Vijay Bhagavath 表示：“预计企业领域的生成式 AI 数据中心以太网交换市场，将以 158.2% 的复合年增长率增长，从 2023 年的 4,190 万美元增长到 2027 年的 10 亿美元。”

高性能网络对于可横向扩展型 AI 基础架构至关重要。生成式 AI LLM 的一大处理瓶颈在于数据在网络中的传输时间。对于某些工作负载，网络中的传输时间可能占到 LLM 处理时间的 60%。这导致在数据在不同计算群集之间移动的过程中，计算基础架构长时间处于闲置状态。至于 AI 网络，目前已经有了超以太网联盟提供的改进后网络。该网络提供的互连性能可媲美超级计算机网络，可扩展至云数据中心，并且在成本效益和普及程度上与以太网同样出色。AI 网络对于满足生成式 AI 和 HPC 大规模网络需求增长至关重要。好消息是，超以太网联盟得到了多数以太网交换机供应商的支持。

为了实现优异性能，有三项关键技术必不可少：高速 SerDes、PHY 和光学器件。这三项技术被应用于以太网和其他网络技术，因此没有哪种网络技术从本质上就具有性能优势。为了充分发挥以太网的性能，InfiniBand 行业协会推出了基于融合以太网的 RDMA (RoCE) 计划，并制定了 RoCE 协议。RoCE 技术受标准数据中心交换机支持，同时还将在市场上推出多项性能改进措施，例如高基数以太网交换、直通交换、负载均衡以及带宽高达 800 GbE (4 x 200 GbE) 的链路。

对生成式 AI LLM 进行初步测试，可以帮助企业初步了解生成式 AI 技术可以带来的好处，从而帮助他们制定生成式 AI LLM 战略并明确所需的基础架构类型。本质上，软件堆栈驱动了企业在生成式 AI 技术革新中进一步发展半导体的需求。了解软件堆栈将有助于企业部署优化的硬件基础架构。

半导体成本已趋于稳定，需重新分配本地与异地基础架构之间的工作负载

随着数据中心 GPU 供应量的增加，将有更多供应商提供数据中心 GPU，更多 AI 加速器将陆续上市，并且更多的生成式 AI 处理能力将可用于本地部署。与此同时，云服务提供商的处理瓶颈也将消失，成本有望趋于稳定。如果这成为现实，在接下来的 3 到 5 年内，企业可通过在本地基础架构和云基础架构之间重新分配生成式 AI 工作负载，来优化资本支出和运营支出。

总结

生成式 AI 是 AI 领域的突破性技术。企业需要立即着手制定一项适用于企业级工作负载的生成式 AI 战略/计划，以便将这项颠覆性技术应用于企业运营。

旺盛的需求，推动了组件价格和云 SP 服务定价的上涨。另一方面，据 IDC 预测，到 2024 年，将有超过 50% 的生成式 AI 系统无需配备专门的加速硬件。因此，人人都可以开始使用标准服务器和网络，部署生成式 AI 基础架构。人们还可以视需要使用 GPU。部署 AI 基础架构有多种选择，同时有多种类型的生成式 AI、AI、ML 和数字孪生技术将以不同方式使不同公司受益。

IDC 的 Brandon Hoff 表示：“市场一直低估了生成式 AI 的增长速度。IDC 预计，生成式 AI 基础架构和半导体将实现强劲增长。”

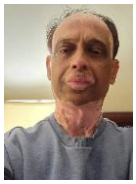
IDC 预测，企业将把数据从云端迁回本地进行生成式 AI 处理，以此来降低运营支出成本。企业将开始在标准计算和以太网网络硬件上，使用生成式 AI 进行开发和测试，并在了解如何利用 LLM 推动业务发展以及如何从专有数据中挖掘价值以后，逐步增加投资。

通过使用市面上现成的服务器和企业以太网网络，构建用于生成式 AI LLM 测试的基础架构，企业将可尽享生成式 AI 带来的优势。

分析师介绍

Brandon Hoff，网络和通信支持技术研究总监

Brandon Hoff 是 IDC 支持技术团队的一员，负责领导 IDC 的网络和通信基础架构相关工作。企业 IT 和 Web、云以及电信服务提供商数据中心相关技术趋势、工作负载、产品、供应商、供应链以及终端用户采用策略等等，都属于 Hoff 先生的工作范围。



Vijay Bhagavath，云和数据中心网络研究副总裁

Vijay Bhagavath 针对云和数据中心网络市场与技术，提供切实可行的前瞻思想与求真务实的深刻见解。Vijay 对整体网络市场、技术、产品路线图、竞争优势和部署策略都有深刻见解，可为供应商、云提供商、企业 IT 采购者与使用者提供富有见地的评论和指导。

赞助方致辞

将 AI 应用于数据

Dell Technologies 不仅拥有众多创新技术，还拥有全面的专业服务和庞大的合作伙伴网络，可帮助企业将理论上的可能性转化为现实中的成果。

- 化繁为简。通过提供战略指导、路线图以及经过验证的解决方案，帮助企业更快取得成果。
- 量身定制。通过提供根据企业业务需求量身定制的基础架构，帮助企业从数据中挖掘更大价值。
- 值得信赖。通过提供用于保护企业数据与知识产权的安全基础，帮助企业未来更好地利用 AI 技术。

Dell Technologies 不仅提供技术与创新，还提供其特有的种种优势，帮助企业悦享卓越的 AI 性能，并轻松采购、部署和管理面向生成式 AI 时代的 AI 基础架构，从而提升企业的工作效率与智能化程度。

欲了解更多信息，请访问 www.dell.com/AI。

IDC Custom Solutions

本文内容根据 www.idc.com 上发布的现有 IDC 研究改编而成。

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
电话：508.872.8200
传真：508.935.4015
Twitter：@IDC
idc-insights-community.com
www.idc.com