Enterprise Strategy Group™
TechTarget 出品

白皮书

为 GenAI 启用以太网驱动的解决方案

开放式网络的重要性

作者：Enterprise Strategy Group
首席分析师 Bob Laliberte

2024 年 1 月

目录

AI 基础架构正在快速增长

在全球范围内，生成式 AI (GenAI) 已极受关注，促成广泛行动。事实上，2023 年，TechTarget 网站上与 GenAI 相关的搜索活动增长了 900% 以上。需要注意的是，这种兴趣正引发实质性的变革。服务提供商是这项技术的早期采用者，许多提供商扩展了其服务产品组合，包括 GPU 即服务选项，大型企业正在为内部应用场景（如消费者分析以及供应链和库存管理）构建私有 GenAI 基础架构。事实上，许多公司董事会和首席级高管已经制定了将 GenAI 应用于其业务流程的计划。此外，在最近的 Microsoft Ignite 大会上，GenAI 领域的领头羊 Nvidia 的首席执行官黄仁勋预测，GenAI 将产生重大影响，他说：“GenAI 的影响将超过 PC，超过手机，甚至会超过互联网。”[1]

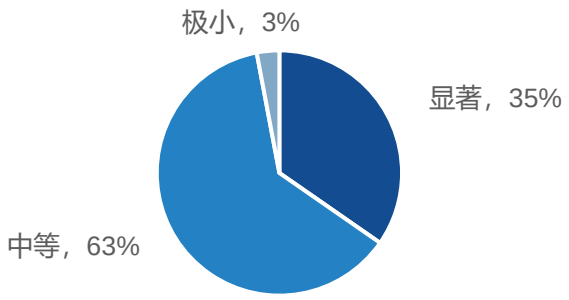
在 TechTarget 的 Enterprise Strategy Group (ESG) 看来，很容易理解为什么组织如此渴望部署 GenAI 解决方案。ESG 研究表明，AI 预计可带来多种好处，包括增强洞察力、提高收入和盈利能力、加快决策速度、改善客户体验和提高运营效率。[2]

还有一点也很明显，这些 GenAI 计划要求组织采用新的基础架构、软件和服务来支持。但是，正如 Dell Technologies 副董事长兼首席运营官 Jeff Clarke 所指出的，这些环境在不同用户那里可能差异巨大。“GenAI 的模式绝非千人一面。它需要端到端解决方案、合适的基础架构、数据计划、无缝协同的软件和服务，这样才能跨云、本地和边缘支持工作负载。”

ESG 研究表明，超过九成 (97%) 组织认为，由于 GenAI，AI 基础架构将出现显著或中等幅度的增长（见图 1）。[3] 这将是支持前端（用户）和后端 (GPU) 环境所必需的，从而确保实现可靠稳固的 GenAI 环境。

图 1.GenAI 带动下 AI 基础架构市场的预期增长

在您看来，就市场增长而言，生成式 AI 将对 AI 基础架构市场产生什么影响（即需要购买更多 AI 基础架构来满足训练和维护大型语言模型的需求）？



来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

[1] 来源：CRN，《Microsoft Ignite 2023: Nvidia CEO Huang Says Microsoft Is Now ‘More Collaborative And Partner-Oriented'》，2023 年 11 月。
[2] 来源：Enterprise Strategy Group 完整调查结果，《*Navigating the Evolving AI Infrastructure Landscape*》，2023 年 12 月。
[3] 同上。

组织不仅对 GenAI 进行了研究，还制定了部署 GenAI 环境的计划，研究表明，绝大多数受访者 (92%) 计划在未来 12 个月内部署 GenAI 环境，这进一步证明组织对采用 GenAI 抱有强烈的期望。[4]

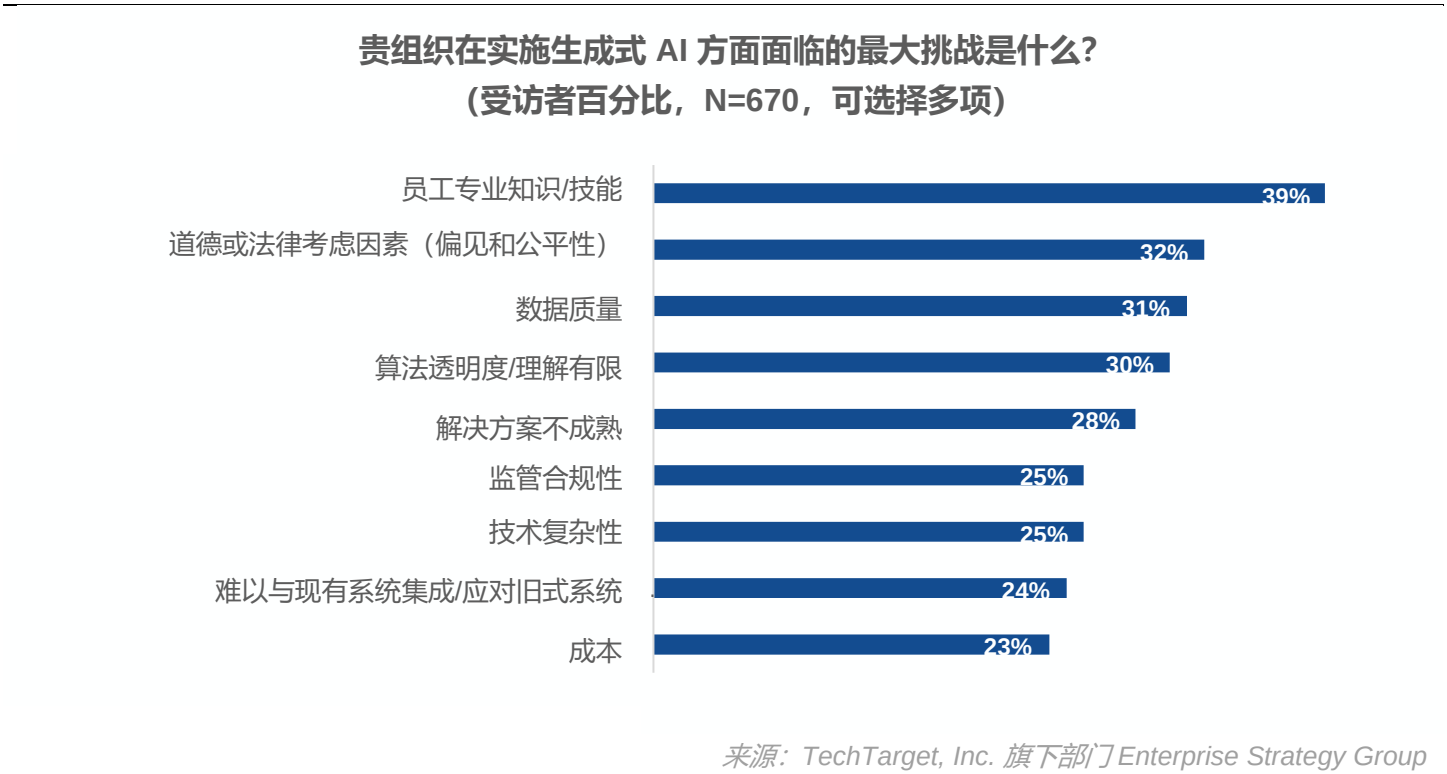
为此，组织需要专门设计的基础架构来满足 GenAI 的特定要求，尤其是对于后端 GPU 环境。然而，部署全新技术可能会在许多不同的层面上带来挑战。

迁移到新技术时面临的挑战

对 IT 部门来说，部署任何新技术都可能具有挑战性，即使只是简单地替换掉现有技术。全新的技术和/或体系结构可能更难部署。GenAI 恰恰就需要新的体系结构，进而需要新的计算、存储和网络基础架构，尤其是对于后端 GPU 环境。这不仅需要更多的基础架构，更重要的是，还需要精心搭建系统，以满足 GPU 群集间的大量连接需求。典型的 50 Gb 以太网 (GbE) 或 100 GbE 架顶式 (ToR) 连接使用 400 GbE 上行链路，会给大型语言模型造成严重的拥塞和延迟，使整个计划面临风险。

当被问及组织在实施生成式 AI 解决方案时面临的最大挑战时，受访者强调了几个问题，包括员工专业知识和技能、技术复杂性、无法与现有或旧式系统集成以及成本，除此之外，还有与数据质量、道德考虑和透明度相关的许多其他挑战（见图 2）。[5]

图 2.GenAI 带来的主要挑战



来源：TechTarget, Inc. 旗下部门 Enterprise Strategy Group

[4] 同上。

[5] 来源：Enterprise Strategy Group 完整调查结果，《*Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*》，2023 年 8 月。

预料之中的是，最大的挑战是缺乏技能和专业知识，尤其是对于生成式 AI 这样的新兴技术而言。大多数组织没有具备必要技能的人才来评估、设计和实施大规模 GenAI 基础架构，尤其是性能密集型后端环境。

技术复杂性也会影响 GenAI 部署，因为某些解决方案利用了通常为高性能计算 (HPC) 环境所专用的专利技术，例如 InfiniBand 网络。因此，掌握相应配套技能的人才数量有限。对于在以太网网络上实现标准化的企业和超大规模业者来说尤其如此。同时，专有解决方案可能更难集成到任何现有的监控或编排平台中，这需要额外的技能、硬件和软件。利用专有解决方案时的另一个考虑因素是交付时间。鉴于过去几年供应链的复杂性，组织可能不愿意采用仅有一家提供商才能提供的解决方案。

由于这些挑战，组织还面临着实施全新 GenAI 解决方案所带来的高成本问题，尤其是专有解决方案，会在扩展时使客户面临除此一家、别无选择的状况。如果缺乏参考设计和体系结构，评估和设计解决方案所需的时间可能会相当长。

组织需要开放且稳健的 GenAI 基础架构

考虑到这些因素，组织需要寻求开放式解决方案来帮助加快 GenAI 基础架构的部署。组织需要创建新的前端环境，使用户能够通过基于 Web 的界面进行交互，并且该界面应非常易于使用和访问。后端基础架构与传统环境，甚或 HPC 环境，都大不相同，需要支持由能够处理海量数据的 GPU 群集驱动的大型语言模型 (LMM)。这些后端基础架构环境对于 GenAI 项目的成功至关重要。

理想情况下，这些解决方案应：

- **全面。**希望部署 GenAI 解决方案的组织需要适用于前端和后端环境的完整解决方案，以加快采用速度。这些解决方案将包括适用于这两种环境的相应计算（包括 GPU 群集）、存储和网络。除了基础架构之外，这些解决方案还需要全面的自动化和监视工具，不仅用于初始配置和日常管理，还需要协助进行结构优化和性能微调。
- **高性能。**对于网络而言，这意味着部署具有可靠交付能力、高带宽、低延迟的无阻塞结构。为此，Linux 基金会在其下属的 Joint Development Foundation 另外创建了 Ultra Ethernet Consortium (UEC)，它汇集了全行业的公司，在开发以太网规范和软件 API 方面进行合作，使 AI 环境具有更高水平的性能、可扩展性、可靠性（例如通过 RoCE v2 协议）和互操作性。[6]
- **经过预先测试和验证。**为了加快这些新 GenAI 环境的采用，就要能部署经过测试和证明可以有效工作的全面解决方案，这将有助于避免常见的部署陷阱。这些解决方案大大节省了研究、分析和设计时间，使组织能够更快地实现其目标并从其 GenAI 环境中获得真正的价值。
- **开放且可扩展。**这将包括利用商品芯片和以太网结构，而不是专有网络技术。GenAI 环境需要尽可能高的网络性能，但应当遵循开放式（而非专有）标准。为了实现这一目标，UEC 将确保以太网能够在 GenAI 环境中发挥重

[6] Ultra Ethernet Consortium。

要作用。此外，组织还可以利用商用开放源代码网络操作系统，例如 SONiC (Software for Open Networking in Cloud)。值得注意的是，SONiC 和 UEC 项目均由 Linux 基金会管理，这简化了行业协作和创新。

Enterprise Strategy Group 的研究强调，希望实现本地数据中心现代化的组织将利用本地超大规模解决方案列为首要优先事项。[7]

- 专业服务带来强化。在能够提供相关专业知识和经验的合作伙伴的帮助下，GenAI 解决方案的价值得以加快实现。这将包括进行适当的评估、构建设计和快速实施解决方案的能力。这还可能包括完全托管服务和技术蓝图或经验证的设计。
- 可扩展。由于大多数组织刚刚开始使用 GenAI，初始部署的规模可能有限，但需要纵向扩展以适应更高的要求。因此，GenAI 基础架构，更具体地说，网络环境必须能够扩展以支持这些需求。
- 高能效。基于 GPU 的解决方案需要大量的电力。因此，组织需要采取一切可能的措施来减少功耗。要实现这一点，应使用优化了吞吐量功耗比的新一代芯片技术。高速交换机所用的机架空间、电力和布线都更少，是更经济高效、更环保的解决方案。除了降低能耗之外，提供可持续性报告的能力也会为运营和管理团队带来帮助。
- 软件驱动。专注于软件可以加快创新的步伐，特别是对于在开放环境中开发的软件，因为它不仅限于单个供应商，而可能有几十个组织为其创新做出贡献。

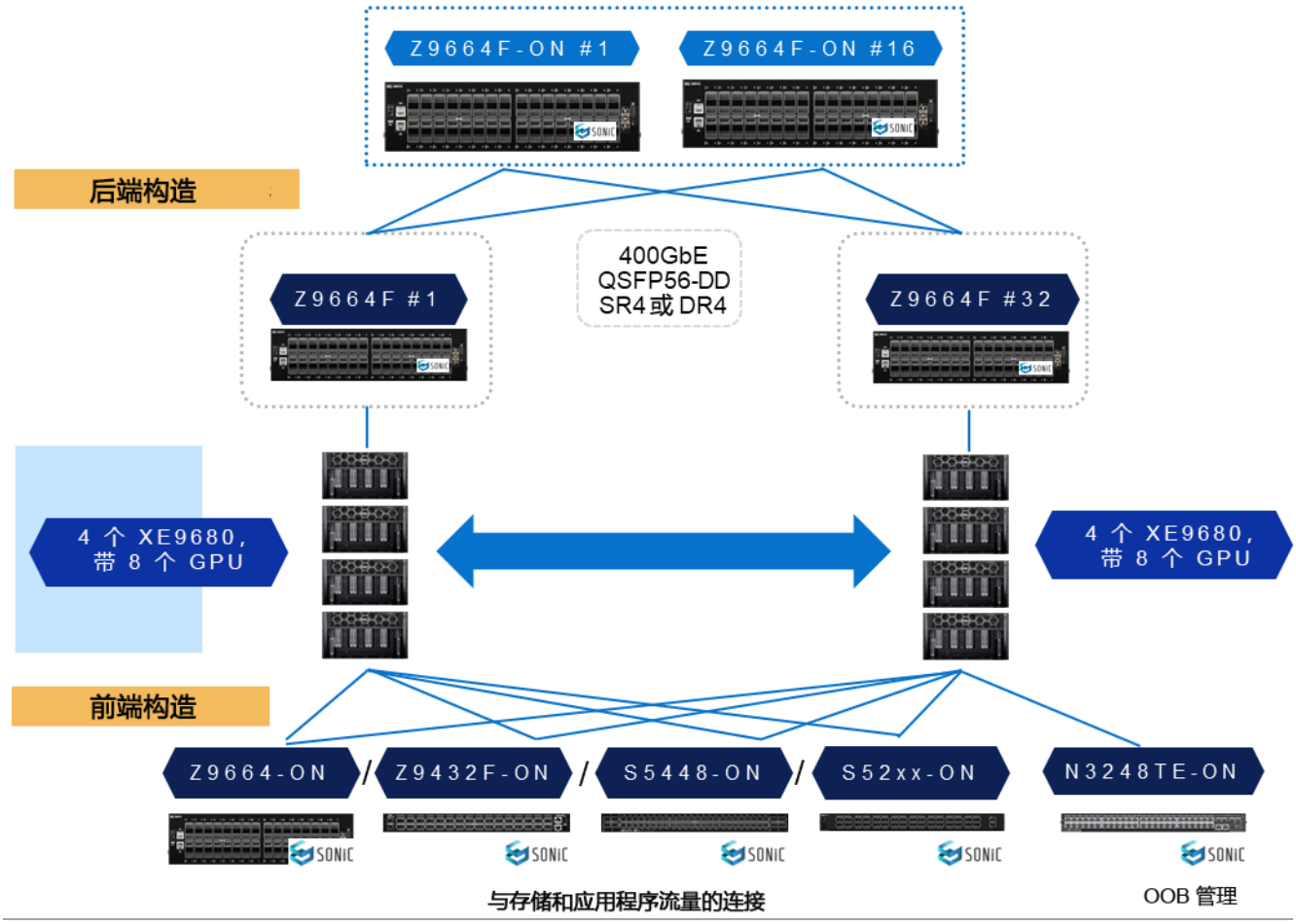
Dell Technologies 为 GenAI 提供基于开放式以太网的解决方案

多年来，Dell Technologies 一直为 AI、建模和 HPC 环境提供全面、开放的基础架构解决方案。该公司正在利用丰富的经验，为前端（应用程序流量、存储访问、常规网络）和包括计算、存储和网络的后端（GPU 结构）环境启用 GenAI 基础架构解决方案。

实现高性能 GenAI 解决方案的关键之一是经过验证的开放式 AI 网络结构，如图 3 所示。

[7] 来源：Enterprise Strategy Group 研究报告，《2023 Technology Spending Intentions Survey》，2022 年 11 月。

图 3.全面的 AI 网络结构解决方案



来源：Dell Technologies。

Dell Technologies GenAI 解决方案包括：

- **模块化的计算系统。**基于 Dell PowerEdge XE 服务器，依托公司为 AI、建模和 HPC 市场提供服务的经验，这些服务器针对此类环境进行了加速优化。戴尔提供风冷或液冷选项以及多种 GPU 数量，并专注于 LLM 的推理或训练，可提供合适的外形规格和高性能解决方案，以满足您的 GenAI 计算需求。计算环境是面向 GenAI 的经验证设计和体系结构解决方案的一部分。
- **以 AI 为中心的存储。**戴尔根据工作负载要求提供一系列存储选项，包括 PowerScale、Elastic Cloud Storage 和 ObjectScale 解决方案。基于以太网的 PowerScale OneFS 存储支持流式读写，可支持 AI 工作负载快速访问数据，并提高 AI 建模能力。戴尔 PowerScale 已经过实地测试，有超过 1,000 家客户在其上运行 GPU 工作负载。因此，戴尔正是依托如此丰富的经验打造了大量“经验证的设计”解决方案。此外，该种类丰富的系列中的所有产品都获得了能源之星认证。

- **新一代以太网结构。**此开放式网络硬件以戴尔 PowerSwitch 为中心，使用下一代芯片（如 Broadcom Tomahawk 4），可提供高达 51.2 Tbps 的共享数据包缓冲。Z9664F-ON 64 端口交换机和 Z9432F-ON 32 端口交换机作为 PowerSwitch Z 系列商用产品，可扩展以支持数千个节点。此外，Dell Technologies 还是 UEC 的成员，将致力于扩展以太网的适用范围，为 GenAI 环境提供支持。
- **软件驱动的体系结构。**Dell Technologies 始终致力于为 GenAI 环境中的网络操作系统、编排和监视提供开放式网络解决方案。对于网络操作系统，Dell Technologies 采用并强化了 SONiC，可提供大型企业所需的全球支持、规模和功能。最新的 Enterprise SONiC Distribution by Dell Technologies（版本 4.2）为 AI 环境提供高级支持，其中包括 RDMA over Converged Ethernet version 2 (RoCE v2)、增强型哈希和直通式交换。即将发布的版本 4.3 为负载均衡和映射提供了增强功能。所有 SONiC 版本均在整个 Z 系列产品组合中进行了测试和验证。这些版本还针对戴尔的第三方应用程序合作伙伴生态系统进行了测试。
- **提供服务以加快采用和优化速度。**除了 24/7 全天候全球支持之外，Dell Technologies 还拥有具有成熟经验的专业服务专家，可帮助组织正确评估、设计和实施全面的 GenAI 解决方案。他们不仅能够了解网络，还能够了解计算和存储域，从而加快了设计过程并降低了出现兼容性问题的风险。这些经验证的设计涵盖推理和模型定制，还有一些服务涵盖 GenAI 管道的数据准备和提取。戴尔还提供托管服务来运营这些 AI 环境。
- **关注可持续发展。**大规模部署 GenAI 环境需要大量的电力资源。戴尔高速交换机拆分模式需要的机架空间、电力和布线更少。利用全新的芯片技术，服务器、网络和存储解决方案能够尽可能节能。专注于能源效率使组织能够降低成本和能源消耗。

通过这些集成，Dell Technologies 完全有能力为后端和前端环境提供完整的 GenAI 基础架构解决方案。

结论

对 GenAI 的兴趣趋热，相应的举措也层出不穷，促使组织评估适合其环境的解决方案。然而，由于它最近很受欢迎，大多数 IT 团队缺乏快速实施解决方案的专业知识或经验。此外，平心而论，这些需要新体系结构和技术的 GenAI 基础架构非常复杂。它们必须经过精心构架，并提供一个平衡的系统，因此分头取得各个组件然后将其胡拼乱凑在一起可能会非常危险。因此，组织需要战略合作伙伴，以获得所需技能以及紧密集成的解决方案，这样才能确保 GenAI 环境取得成功。

但是，组织需要小心那些让客户受制于专利技术的全面解决方案，尤其是在这些环境不断扩展时。开放式解决方案可为大规模 GenAI 环境提供创新、灵活性和成本效益。但是，为了确保环境稳健，还必须确保这些开放式解决方案经过全面测试、验证且受到可靠支持。

Dell Technologies 提供完整的 GenAI 解决方案，其中包含所有基础架构和软件，包括前端和后端环境的编排和管理。它们还整合了开放式计算、存储和网络。此外，组织可以利用托管服务、专业服务以及经过充分验证的设计和体系结构，其中还包含戴尔合作伙伴生态系统的各类产品和服务。这些全面而模块化的解决方案使组织能够加快 GenAI 解决方案的部署，使其更快发挥价值，同时降低风险并确保提高运营效率。

Enterprise Strategy Group 简介

TechTarget 的 Enterprise Strategy Group 提供有针对性且切实可行的市场情报、需求方研究、分析师咨询服务、GTM 战略指导、解决方案验证以及量身定制的内容，为企业采购和销售技术提供支持。

contact@esg-global.com
www.esg-global.com