生成式和代理式 AI 就绪基础设施战略

路线图: 从概念验证到规模化生产

Mary Johnston Turner
IDC未来数字基础设施议程研究副总裁

目录

单击下方各个标题，跳转到相应章节。

本简报内容

人工智能，尤其是生成式人工智能（GenAI），正处于加速开发和部署阶段。企业的投资激增，测试了数百个场景，并确定了加速计算、云基础设施以及自动化开发和数据管理工具链在推动这类强大技术应用于大规模生产过程中起到的关键作用。

AI可实现日常工作自动化，提高效率，**因此可能会彻底改变从客服到各种内部流程的运营方式。**AI技术的快速普及必然会产生深远的经济影响，重塑行业，开辟新市场，改变竞争格局。

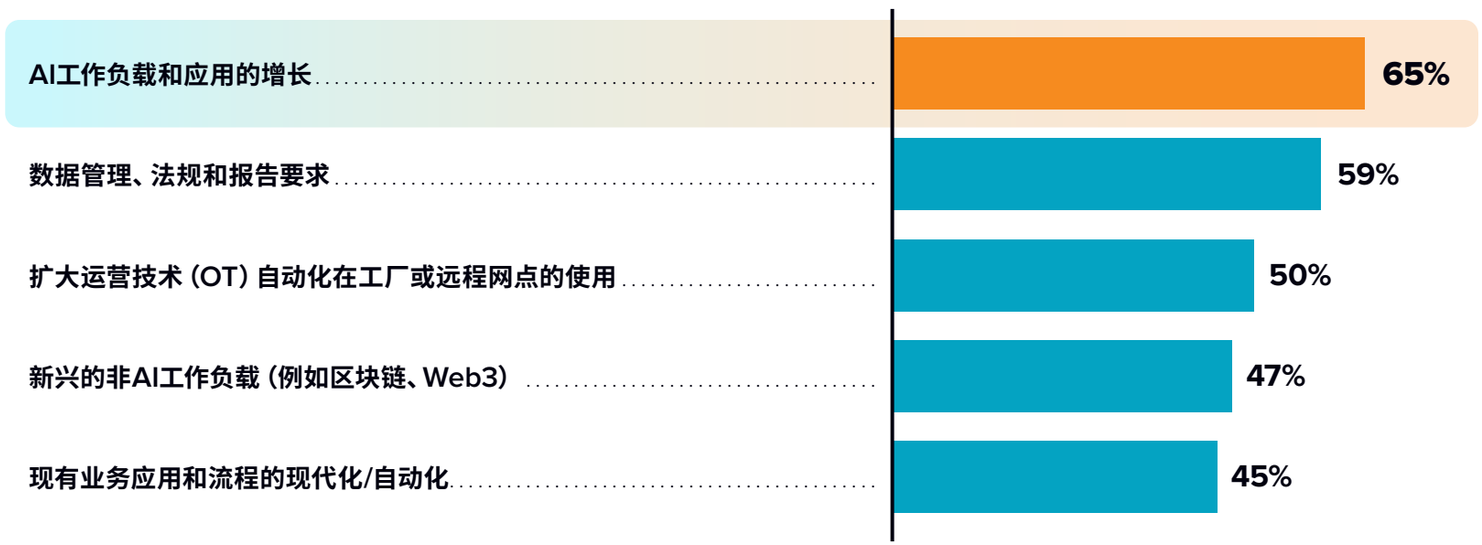
本简报将介绍IDC对AI全球使用情况、影响和价值创造的研究，聚焦企业在基础设施现代化、治理和运营模式等方面面临的机遇和挑战。它还为技术买家提供了一些建议，确保他们成功部署和规模化运营AI。

预计到2030年，AI采用者在AI解决方案和服务上每花一美元，就会为经济带来4.60美元的间接和诱导效应。

来源: IDC *Macroeconomic Center of Excellence*, 2024

未来两年，AI工作负载将成为企业基础设施投资和变革的头号驱动力

未来两年，以下哪些趋势将对贵企业的计算和存储资源利用产生最大的影响？



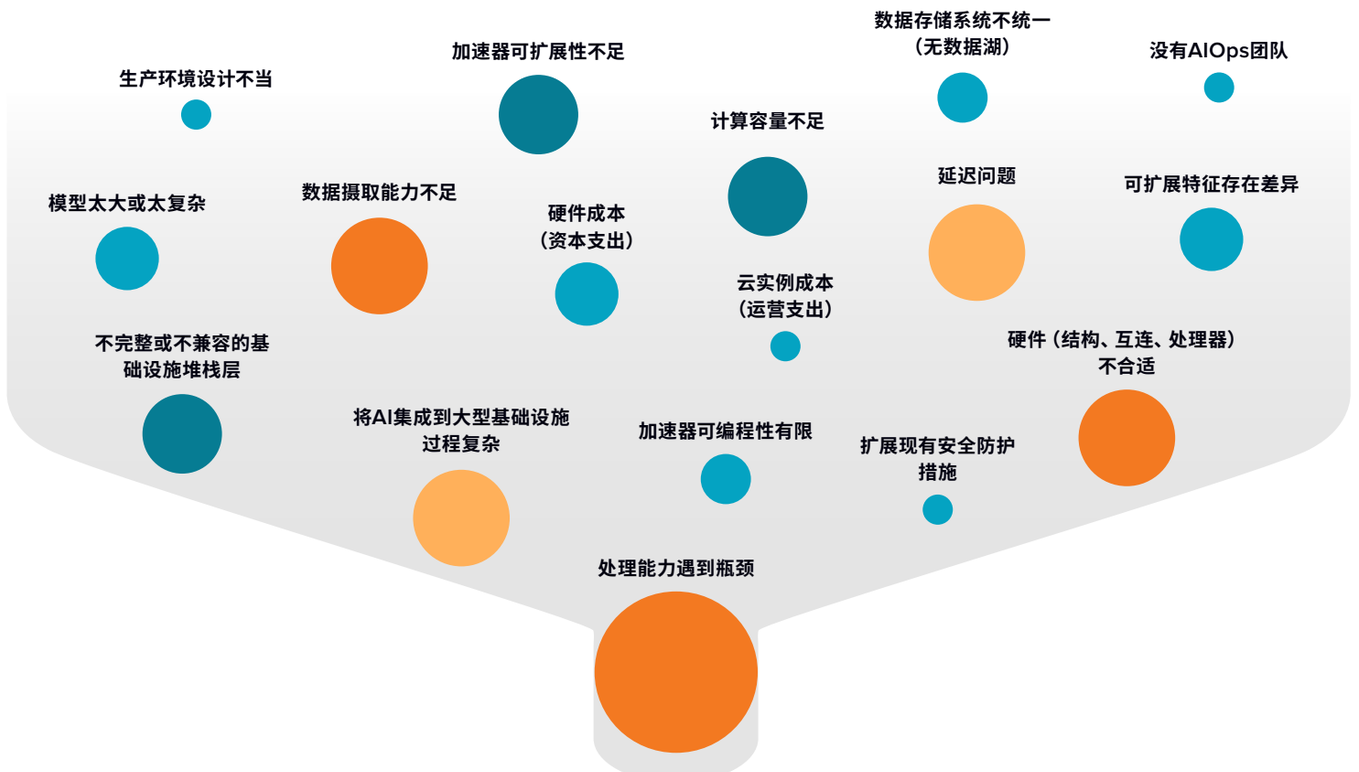
注意：按国家/地区IT支出加权。n = 1,129；来源：IDC *Worldwide Digital Infrastructure Sentiment Survey*，2024年6月

AI能否从小规模概念验证 (POC) 成功过渡到应用于大规模生产取决于企业能否从性能、成本、安全和合规等关键角度优化基础设施战略。

数据量、连接和监管要求等对许多有关部署位置和计算平台的决策起决定作用。

许多AI项目之所以失败，是因为对AI应用层和数据层的基础设施要求不够了解

- 计算、存储和网络基础设施的方法和投资水平将取决于企业投资的AI。
- 不了解这些要求会增加成本并带来许多运营挑战。
- 了解场景、业务目标和性能要求至关重要。
  - AI的计算要求可能因AI项目的生命周期而异。
  - 并非所有AI计划都必须有立竿见影的效果；许多计划可能需要分批进行，需要数天甚至数周才能交付结果。



来源: IDC *Harnessing Hybrid Infrastructure to Fuel AI Business at Scale: A C-Suite Playbook*, *#US52101325*, 2024年8月

创建AI就绪型基础设施需要了解场景的性能、安全、合规、成本和可持续发展要求

性能

许多(但不是全部)AI工作负载需要高性能计算和数据基础设施,这样才能为实时分析和决策提供所需的处理能力和吞吐量。许多需要分批进行。

安全与合规

企业基础设施必须具备强大的安全控制和合规功能,以保护敏感数据,符合AI监管要求。

成本管理

AI工作负载生命周期各个阶段对基础设施的消耗情况差异很大。按量计费的订阅模式可能成本高昂,但时间越长灵活性越大。专用资产成本是可预测的,但难以扩展。

可持续发展

高性能基础设施对电源和冷却的要求可能很高。有关AI模型大小和规模的决策将直接影响能源使用和可持续发展。

的全球企业一致认为,GenAI是一项重要的新型企业工作负载,与ERP和电子商务一样,未来几年需要增加技术支出。

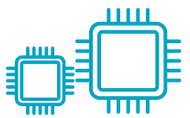
n = 889; 来源: IDC *Future Enterprise Resiliency & Spending Survey Wave 4*, 2024年4月

AI模型选择关系到成本、成果和基础设施要求

利用现有模型	创建私有模型
对特殊技能的需求更低	成本/复杂性更高
访问高质量模型	数据隐私与合规
降低成本	模型集成
快速部署	模型准确性

影响模型选择的因素

- 紧迫性与准确性
- 内外部数据的混合情况
- 更新频率
- 数据加权和参数调整
- 数据质量和数量
- 法律与合规
- 可解释性和集成
- 检索增强生成 (RAG) 的使用



模型大小相差很大，
少则100万个参数，
多则超过2万亿个。

数据量可以从几
GB到数PB不等。

大多数现成的商业或开
源模型都需要一定程度的调优，
并且都需要频繁更新。

不同的AI模型和场景对基础设施的要求也不同

处理器和存储之间的权衡

所需的数据规模和算力水平

数据隔离和主权控制的程度

部署位置和集成

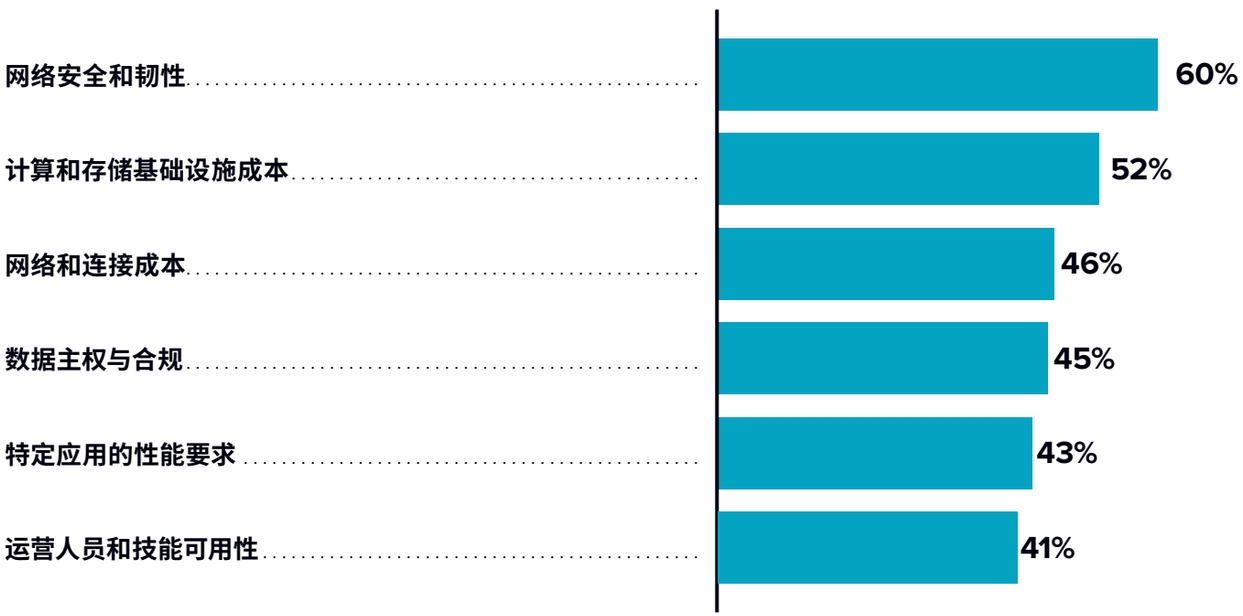
部署位置取决于延迟和性能

模型定制化级别

- 模型要求差异很大——一刀切的基础设施战略可能效率低且成本高昂。
- 基础设施与AI场景必须相匹配，才能实现业务目标。
  - 模型规模和性能
  - 数据量和速度
  - 所需的价值实现时间
  - 所需的输出精度
  - 数据安全、合规和主权
  - 使用量和容量
  - 模型自定义和更新的级别和频率
  - 数据和工作流互操作水平
  - 延迟和性能
- 确保应用开发和测试标准能预测生产需求。
- 制定动态持续的模型更新和迁移计划，以满足不断变化的数据科学和业务需求。

在制定部署决策时必须考虑安全、成本、连接性和数据主权/合规

在判断应用及其数据集部署位置和部署方式（跨本地数据中心、托管和数据中心托管站点、边缘或公有云）时，哪些标准最重要？



注意：按国家/地区IT支出加权。n = 1,129；来源：IDC *Worldwide Digital Infrastructure Sentiment Survey*，2024年6月

制定跨专用和共享基础设施部署数据政策框架

运营模式优先项会影响部署方法

● 客户责任
◐ 共同责任
● 提供商责任




	传统的本地部署资本支出	数据中心托管/托管	专用基础设施即服务	共享公有云服务
设施、电力和热力	●	●	●	●
硬件配置和容量	●	●	●	●
基础设施软件配置和容量	●	●	●	●
生命周期管理和更新	●	●	●	●
网络体系结构	●	◐	●	●
数据管理和安全	●	●	●	●
基础设施运营工具和人员	●	●	◐	◐
避免技术债务	●	●	●	●

来源: IDC *Build Versus Buy Decision-Making:Optimizing AI-Ready Infrastructure ROI*, *#US51930224*, 2024年3月

治理很重要——AI卓越中心可确保整个企业齐头并进，加速做好AI就绪工作

AI卓越中心是一个由专家组成的核心小组，旨在为AI投资、政策和战略提供指导，应牵头制定决策框架和部署指南。

- 围绕以业务为中心的KPI、政策和成果协调IT、业务部门 (LOB) 、开发和数据团队。
- 联合内外部最好的基础设施。
- 指导选择最适合的专用平台和基础设施技术。
- 定义并实施数据合规、道德和安全护栏。
- 在整个企业范围内共享经验教训。
- 建立融资模式，促进跨AI场景的数据和工作流集成。

来源: IDC *Harnessing Hybrid Infrastructure to Fuel AI Business at Scale: A C-Suite Playbook*, *#US52101325*, 2024年8月

根据生产规模要求设计概念验证

AI与任务相匹配

了解不同类型的AI如何使不同的场景受益

- POC期望和时间限制
- 最低可接受能力
- 生产途径

价值

潜在效益评估

- 成本
- 效率
- 用户体验

可行性

复杂性和实施障碍评估

- 工具可用性
- 技能和文化就绪情况
- 安全与合规

治理

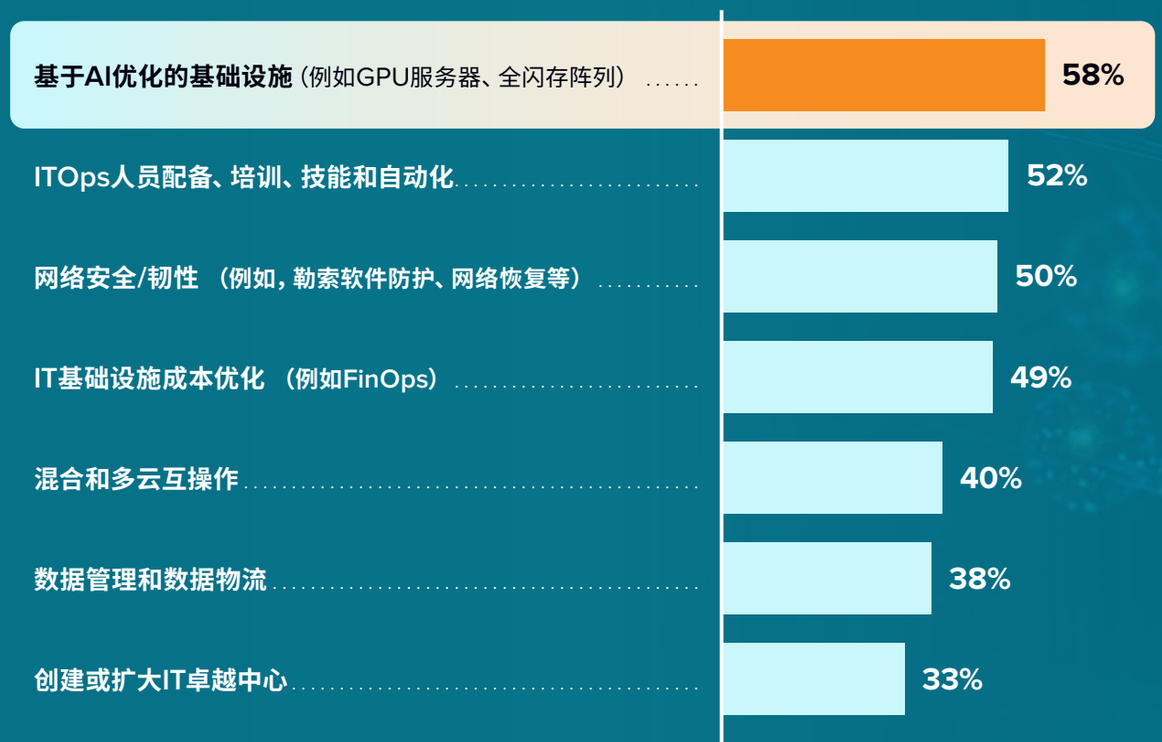
投产路线图

- 与主要利益相关者的互动
- KPI和ROI跟踪系统和流程
- 人员、流程和工具转型

来源: IDC *Executive Playbook:Optimizing Use of AI for IT Operations*, *#US52579724* 2024年9月

基本指南

未来两年，贵企业最具战略意义的IT基础设施投资是什么?



AI就绪型基础设施成功清单

- 调整现有治理和运营模式，以适应AI赋能型业务。
- 贯彻工作负载和数据驱动的决策框架。
- 根据使用场景，将应用和数据匹配到最合适的平台和部署模式。
- 打造互操作性、可移植性，实现动态持续的模型更新和迁移。
- 利用开源社区和供应商群体，获取最新的创新成果和最佳实践。
- 投资人员、流程和技能，实现自动化和智能运营。

关于IDC分析师

Mary Johnston Turner
IDC未来数字基础设施议程研究副总裁

Mary Johnston Turner是IDC未来企业研究团队成员，担任未来数字基础设施研究副总裁。她分析企业IT和业务战略如何利用部署在专用数据中心和共享公共服务环境中的无处不在的自主云基础设施解决方案。她的研究工作以企业改革数字基础设施解决方案采购、保护和优化的最佳实践为基础开展调查和深入分析，强调企业客户的声音。她的研究着重探讨即用即付订阅、跨云控制面和协作式企业基础设施治理模式如何帮助企业更好地将基础设施投资与关键业务成果和创新优先点保持一致。

有关Mary Johnston Turner的详细信息

赞助商寄语

释放 AI 创新潜力: NVIDIA 全栈解决方案在 AWS 云平台的强大实力

NVIDIA 全栈解决方案是 AI 革命的核心，它提供了一套端到端平台，其价值远超高性能芯片本身，重新定义了企业加速创新、部署解决方案以及在数据驱动型世界中实现蓬勃发展的方式。通过将 NVIDIA 行业领先的端到端硬件、软件和工具与 AWS 全面的云服务、全球化规模和安全能力相结合，各类组织能够获得无可比拟的灵活性、可扩展性和运营卓越。

什么是 AWS 上的 NVIDIA 全栈解决方案?

NVIDIA 全栈解决方案整合了高性能硬件 (GPU、AI 超级计算机、边缘计算系统) 、行业优化软件 (CUDA、TensorRT、Dynamo、RAPIDS、Omniverse、Earth-2、NVIDIA AI Enterprise) 以及一套丰富的开发者工具，所有组件协同工作，以加速 AI 生命周期的每个阶段。从数据采集、模型训练，到推理与决策，该解决方案旨在实现最高效率: 将部署时间从数周缩短至数分钟，并以前所未有的速度将数据转化为可执行的洞察。

为各类应用场景加速创新

AWS 的托管服务、预训练模型，以及与 NVIDIA 的深度集成，消除了基础设施瓶颈，为医疗健康、金融科技、汽车、科研等各行业的 AI 驱动型创新缩短了价值实现周期。NVIDIA 硬件与软件的紧密集成，确保性能不被浪费，投入到基础设施的每一分电力和资金都能产生最大价值。

便捷快速获取先进 AI 技术

与 AWS 托管服务的直接集成，使客户能够大规模部署 NVIDIA AI 模型、工具和框架，同时享受 AWS 额外的数据服务和开发者服务。

- NVIDIA NIM 微服务可轻松部署在 Amazon Elastic Kubernetes Service 上，并在 Amazon Bedrock 和 SageMaker 市场中提供。NVIDIA NIM 是一套易于使用的微服务，专为高性能 AI 模型推理的安全可靠部署而设计。依托 NVIDIA 及社区领先的推理技术 (包括 NVIDIA Dynamo Triton、TensorRT-LLM、vLLM 等) ，NIM 能够实现大规模无缝 AI 推理，确保 AI 应用可在任何场景下放心部署。这些预制容器支持广泛的 AI 模型，涵盖开源社区模型、NVIDIA 基础模型以及自定义模型。例如，NVIDIA Nemotron Super 49B 和 Nano 8B 推理模型，已在 Amazon Bedrock 市场和 Amazon SageMaker JumpStart 中上线。这些模型属于 NVIDIA Nemotron 多模态模型系列，可提供顶尖的代理推理能力，适用于研究生水平的科学推理、高级数学计算、代码编写、指令遵循、工具调用以及视觉推理等场景。不同模型针对不同用户需求优化: Nano 模型侧重成本效益，Super 模型则在单 GPU 上兼顾精度和计算效率。

赞助商寄语（续）

- NVIDIA Dynamo 是一款开源推理框架，提供创新解决方案以优化性能和可扩展性。它支持多项 AWS 服务，如 Amazon Simple Storage Service (Amazon S3)、Elastic Fabric Adapter (EFA) 和 Amazon EKS，且可部署在 NVIDIA GPU 加速的 Amazon EC2 实例上，包括由 NVIDIA Blackwell 加速的 P6 实例。
- 借助 AWS 上的 NVIDIA Run:ai 实现高级工作负载编排，并与 Amazon EKS 集成，可简化集群管理、最大化 GPU 利用率，并实现部署流程自动化。

AWS 托管的 NVIDIA DGX Cloud 是一套开箱即用的全栈 AI 平台，由 AWS 与 NVIDIA 联合打造，用于模型构建和微调 — 包含加速基础设施、GPU 编排软件、云原生工具，以及 AI 专家支持，每一层都经过优化。DGX Cloud 整合了 AWS 上 NVIDIA 的优势资源，兼具灵活性和专业能力，可加速 AI 项目推进。

可扩展、灵活的 AI 基础设施

AWS 通过优化的 Amazon EC2 实例与 DGX Cloud，提供对 NVIDIA 最新 GPU 架构（包括 Blackwell）的即时访问能力。这使企业能够灵活扩展 AI 工作负载 — 从概念验证到大规模全球部署 — 无需资本投入或受硬件限制。无论您是开展快速实验、大规模训练任务，还是实时推理，由 NVIDIA 加速的 AWS 基础设施都具备低延迟、高吞吐量和无缝资源分配的特点。

安全、可靠与合规

AWS 将 NVIDIA “设计即安全” 的全栈解决方案与加密网络、身份认证及严格的跨区域合规（GDPR、HIPAA、SOC 2）相结合。AWS Nitro 系统进一步在 AI 全生命周期中保护数据、模型权重和知识产权。客户可受益于 AWS 强大的全球云基础设施，该基础设施能为关键 AI 应用提供 99.99% 的可用性和业务连续性保障。

成本效益与运营简化

AWS 基于使用量的计费模式，让您只为实际使用的资源付费 — 无论企业规模和预算如何，都能利用 NVIDIA 的先进加速能力，无需过度配置资源或造成资源闲置。通过消除硬件采购、数据中心管理和冗余瓶颈，AWS 最大限度降低了运营复杂性并节省成本，同时确保计算资源始终保持最新状态。

丰富的生态系统与全球合作网络

AWS 与 NVIDIA 的合作生态系统提供一流支持、丰富工具和蓝图，让您能够借助全球数千家组织的创新成果。通过 AWS 的数百万客户网络与 14 万家全球合作伙伴，结合 NVIDIA 超过 400 万开发者的社区资源，可加速您的 AI 计划推进。

AWS 与 NVIDIA 全栈解决方案的结合，使各类规模的组织都能在 AI 革命中占据领先地位 — 在真正的全球范围内，将潜力转化为价值，将想法转化为竞争优势。

有关AWS和英伟达的更多信息，请访问aws.amazon.com/nvidia

IDC Custom Solutions

IDC Research, Inc.
140Kendrick Street, Building B, Needham, MA 02494, USA
T +1 508 872 8200

idc.com @idc @idc

国际数据公司 (IDC) 是全球著名的信息技术、电信和消费科技咨询、顾问和会展服务专业提供商。IDC在全球拥有超过1,300名分析师，他们针对110多个国家/地区的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。IDC的分析和洞察帮助IT专业人士、业务主管和投资机构做出以事实为基础的技术决策，实现他们的关键业务目标。