解决方案简介

NetApp ONTAP AI

借助 NetApp 和 NVIDIA 技术，简化、加速和集成深度学习数据管道 (data pipeline)

NetApp | NVIDIA

人工智能基础架构挑战

人工智能 (AI) 和深度学习 (DL) 可以帮助企业检测欺诈行为、加强与客户的关系、优化供应链、交付创新产品和服务，从而在竞争日益激烈的市场中占据一席之地。您的企业可能和其他许多企业一样，正在寻求新的深度学习方法来推动数字化转型，提升竞争优势。如果想要从深度学习中获得最大收益，您必须首先解决几项关键挑战。

自己动手 (Do-It-Yourself, DIY) 集成非常复杂。组装和集成现成的深度学习计算、存储、网络和软件组件会增加复杂性和部署时间，从而导致宝贵的数据科学资源耗费在系统集成工作上。

实现可预测和可扩展性能并非易事。深度学习最佳实践建议企业应该从小规模入手，然后随业务增长逐步扩展。传统做法中，通常使用计算和直连存储为人工智能工作流馈送数据。但是，扩展传统存储可能会导致业务运营出现中断和停机情况。

中断会影响数据科学家的工作效率。DL 基础架构涉及大量硬件和软件互依赖关系，要使 DL 基础架构正常运行，需要具备深度的全堆栈 AI 专业知识。停机或者速度缓慢的人工智能性能可能会引发连锁反应，从而影响开发人员的工作效率，导致运营费用失控。

解决方案

由 NVIDIA DGX 系统和 NetApp 云互联全闪存存储提供动力支持的业已验证的 NetApp® ONTAP® AI 架构可简化、加速和集成数据管道，帮助您充分实现人工智能和深度学习的优势。利用横跨边缘到核心再到云的 Data Fabric，可以可靠地简化数据流，加速分析、训练和推理。

主要优势

借助经过验证的灵活解决方案降低风险

- 通过消除复杂的设计和避免盲目猜测来加快发展步伐
- 利用可用的预配置解决方案简化配置和部署。

提供合适的性能和可扩展性

- 从小规模入手，然后逐步无中断扩展。
- 利用高性能解决方案加速生成结果

构建集成数据管道

- 从边缘到核心再到云，使用集成管道智能管理数据
- 部署以人工智能专业知识和简单支持选项为后盾的解决方案。

统一人工智能工作负载管理

- 消除基础架构孤岛
- 灵活应对不断变化的业务需求

作为最先推出的融合基础架构堆栈之一，NetApp ONTAP AI 将全球首个每秒达 5000 万亿次计算的人工智能系统 NVIDIA DGX A100 和 NVIDIA Mellanox® 高性能以太网交换机集于一体。您可以统一管理 AI 工作负载、简化部署流程并加速获得投资回报。

“深度学习是一场革命，几乎所有市场都深陷其中。我们正在将深度学习应用于不同的市场，从而推动这门艺术发挥无限潜力。由 NVIDIA DGX 系统和 NetApp 全闪存存储提供动力支持的 NetApp ONTAP AI 可以简化并加速深度学习的数据管道。”

Cambridge Consultants 人工智能总监 Tim Ensor

图 1 ）采用 DGX A100 的 ONTAP AI 架构；双节点，四节点和八节点配置。

借助经过验证的灵活解决方案降低风险

人工智能创新的快速发展为设计一款高效的人工智能基础架构带来了挑战。借助 ONTAP AI ，您可以使用经过现场验证的参考架构，消除盲目猜测，加快入门速度。或者，通过选择易于采购和部署的预配置集成解决方案，您可以消除设计和管理复杂性。

ONTAP AI 集成解决方案提供三种预配置选项，其中包括容量扩展和可选高级软件。这种集成的解决方案可以通过单个号码（从意外事件报告到解决）提供现场安装和全面支持，从而进一步降低复杂性。

提供合适的性能和可扩展性

深度学习日常训练需要大量的计算能力。更快速的映像训练可以降低整体计算成本，并加速人工智能创新和工作效率。

DGX A100 系统采用全新 NVIDIA Ampere 架构构建，可提供比上一代产品高出 6 倍的训练性能。您可以获得相当于一个数据计算基础架构中心的分析，培训和推理功能，现在整合到一个系统中。与 CPU 系统相比，DGX A100 占用空间仅为 1/25，功耗仅为 1/20，而且成本也仅为 1/10。

一流的计算搭配一流的存储，才可以每秒处理成千上万的训练映像。您需要高性能数据服务解决方案，才能满足要求最苛刻的深度学习训练工作负载的需求。

借助 NetApp 全闪存存储，您可以预期获得超过 2 Gbps 的持续吞吐量（峰值为 5 Gbps ），而延迟远低于 1 毫秒，而 GPU 的运行利用率则超过 95%。对于 NAS 工作负载，单个 NetApp AFF A800 系统即可支持吞吐量为 25 GB/秒的顺序读取和 100 万次 IOPS 的小型随机读取，同时延迟保持在 500 微秒以下。

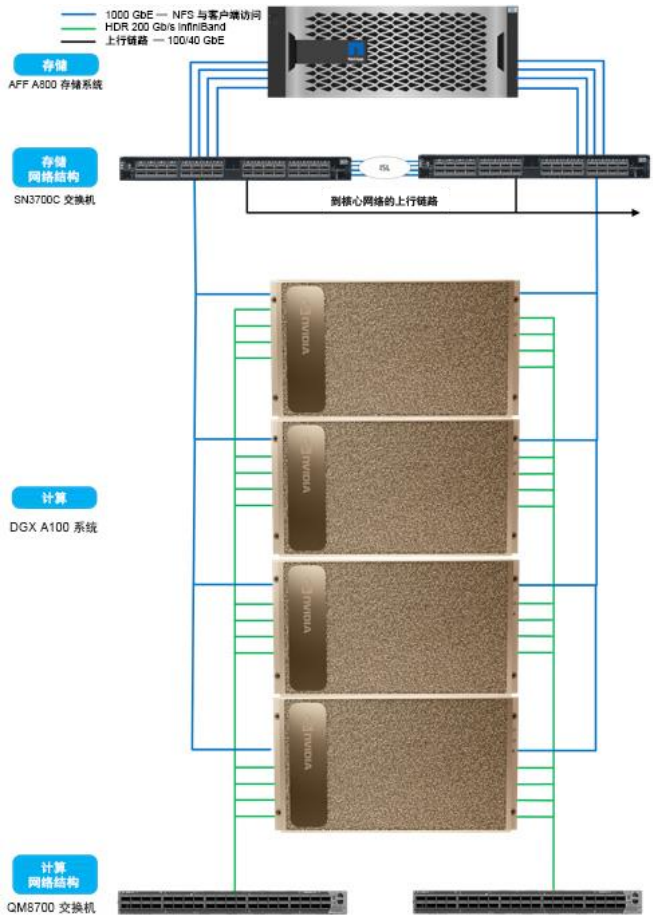

图 2) 采用 Mellanox Spectrum 100GbE 交换机的 ONTAP AI 4 节点配置。

借助 NetApp 机架级架构，您的企业可以利用全闪存存储从数十 TB 扩展到数十 PB 。同时，NetApp ONTAP FlexGroup 中的一个命名空间即高达 20 PB，可处理 4000 亿个文件。

构建从边缘到核心再到云的集成数据管道

ONTAP AI 可利用 Data Fabric 通过一个平台统一管理整个管道的数据。使用相同的工具有助于确保安全地控制和管理数据（无论数据是在传输中、使用中还是空闲状态），从容地满足合规性要求。如果深度学习环境出现问题，您可以依靠我们经验证的可靠支持模式，获得问题解决方案和指导。

统一人工智能工作负载管理

现在，您的组织可以消除基础架构孤岛，这些孤岛要么未充分利用，要么已成为人工智能工作负载的基础。借助 ONTAP AI ，您可以获得基于 DGX A100 系统构建的通用 AI 基础架构解决方案，将分析，培训和推理整合到一个平台上，该平台可以灵活响应您的业务需求。与传统架构相比，您还可以获得更好的 TCO 。

NetApp 与 NVIDIA：携手推动创新

DGX A100 在 ONTAP AI 之中居于核心地位，它是数据中心人工智能的通用组件，支持通过一个平台管理深度学习训练、推理、数据科学及其他对性能要求较高的工作负载。每个 DGX A100 系统都由八个 NVIDIA A100 Tensor 核心 GPU 和两个第二代 AMD EPYC ® 处理器提供支持，并集成了最新的高速 NVIDIA Mellanox 100/200 GB 以太网和支持 InfiniBand 的 ConnectX-6 适配器互连。

现在，可通过分区方法将每个 DGX A100 系统分割为多达 56 个实例，从而采用全新多实例 GPU 技术加速处理多个较小规模的工作负载。通过这种加速，您的组织可以在 ONTAP AI 中极其高效地分配 GPU 性能。整个企业的数据科学团队可以更快地进行迭代，实现可重现性自动化，并将 AI 项目交付时间缩短至 3 个月，同时提高质量。

NetApp AFF 系统配备业内速度最快、最灵活的全闪存存储以及全球首创端到端 NVMe 技术，确保数据流向深度学习流程。AFF A800 系统能够以最快达同类竞争解决方案 4 倍的速度向 NVIDIA DGX 系统传送数据。*

该解决方案集成 Mellanox Spectrum 以太网交换机，具备低延迟、高密度、高性能优势，可满足人工智能环境的功耗要求。

1. 每个全闪存集群的读取吞吐量高达 300 Gbps，领先的竞争友商则为 75 Gbps。

由 NetApp 提供支持的 Data Fabric 采用同类最佳的数据管理和云集成解决方案，可以帮助您加速深度学习，同时管理和保护您的关键数据。ONTAP 则提供无可比拟的 22:1 的整体数据精简率，以及低于直连存储 54% 的 TCO。

DGX A100 系统由 NVIDIA DGX 软件堆栈提供支持，该堆栈包含针对 AI 和数据科学工作负载的优化软件。您可以最大限度地提高性能，使企业更快地获得人工智能基础架构投资回报。

NetApp AI 控制平台可将 Kubernetes 和 Kubeflow 与 NetApp 支持的数据网络结构相集成，从而帮助简化 AI 数据管理，从而为您提供从边缘到核心再到云的最佳数据可用性和可移植性。NetApp 数据科学工具包是一个 Python 库，可帮助您的数据科学家和数据工程师轻松执行大量数据管理任务，从而增强 AI 控制平台。例如，他们可以配置新的数据卷，即时克隆数据卷以及为数据卷创建 NetApp Snapshot ® 副本，以实现可追溯性和基线化。

正确的工具对于成功至关重要。这就是 ONTAP AI 通过领先的机器学习操作（ MLOps ）软件（包括 Domino 数据实验室， Iguazio 等）进行验证的原因。您的团队可以使用熟悉的工具最大限度地发挥 AI 环境的价值，并加快获得洞察力的速度。

解决方案组件

- NVIDIA DGX A100 系统
- 采用 ONTAP 9 的 NetApp AFF A 系列存储系统
- NVIDIA Mellanox Spectrum SN3700C ，NVIDIA Mellanox Quantum QM8700 和 / 或 NVIDIA Mellanox Spectrum SN3700-V
- NVIDIA DGX 软件堆栈
- NetApp AI 控制平台
- NetApp 数据科学工具包

参考架构

针对特定行业的使用情形，NetApp 发布了以下基于 ONTAP AI 参考架构：

- 适用于医疗保健的 ONTAP AI 参考架构：诊断成像
- 适用于无人驾驶工作负载的 ONTAP AI 参考架构：解决方案设计
- 适用于无人驾驶工作负载的 ONTAP AI 参考架构：解决方案设计

关于 NVIDIA

NVIDIA 于 1999 年发明了 GPU ，这激发了 PC 游戏市场的增长，重新定义了现代计算机图形，并实现了并行计算的变革。最近，GPU 深度学习又点燃了“下一个计算纪元”的现代人工智能；GPU 相当于计算机、机器人和无人驾驶汽车的大脑，可帮助理解和认知世界。
有关详细信息，请访问 www.nvidia.cn。

关于 NetApp

在充满综合人才的世界里，NetApp 是您的存储专家。我们只专注于一件事情，那就是帮助您充分利用数据的价值。NetApp 将值得信赖的企业级数据服务引入云中，并将云的简单灵活性引入数据中心。我们行业领先的解决方案支持各种客户环境以及世界上最大规模的公有云。

NetApp 是一家以云为主导、以数据为中心的软件企业，唯有 NetApp 可以帮助构建符合您需求的独特 Data Fabric，简化并连接您的云，以及随时随地安全地为合适的人员提供正确的数据、服务和应用程序。
www.netapp.com/cn/

NetApp