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)

主要优势

部署简单

- 通过消除复杂的设计和避免盲目猜测来加快发展步伐
- 加速创新和实验
- 借助企业级数据服务和简单的技术更新来简化部署

提供业务所需的性能和可扩展性

- 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
- 一个命名空间即可处理超过 4000 亿个文件

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
- 提供人工智能专业知识和单点联系支持
- 借助 NetApp® Data Fabric 加速云集成

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。

中断会影响运营支出，降低数据科学家的工作效率。深度学习基础架构非常复杂，涉及大量软硬件依赖关系。维持深度学习基础架构的持续运行需要扎实的全堆栈人工智能专业知识。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 超级计算机和 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，加速训练和推理。

“深度学习是一场革命，几乎所有市场都深陷其中。由 NVIDIA DGX 超级计算机和 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 人工智能主管
Monty Barlow

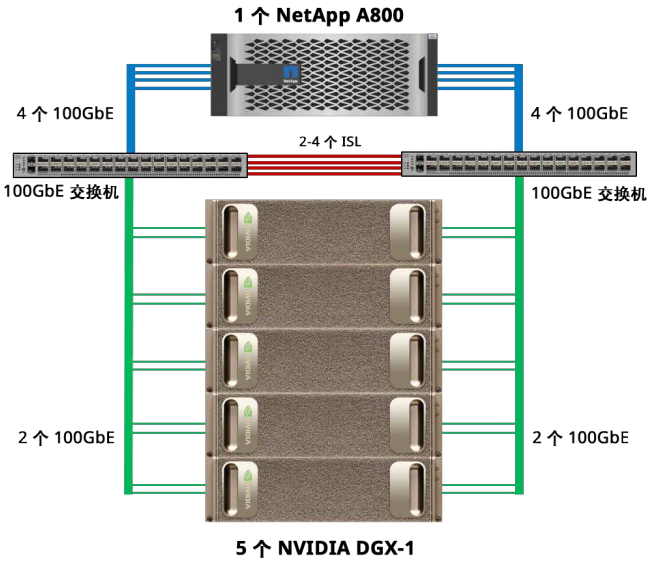

图 1) NetApp ONTAP AI 业已验证的架构。

简化设计和部署

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。但是，您可以利用 ONTAP AI 经过验证的参考架构来消除设计复杂性，避免盲目猜测，更快速地上手。Trident — NetApp 专为 Kubernetes 打造的存储配置程序 — 支持客户将 NVIDIA GPU Cloud (NGC) 容器映像无缝迁移至 NetApp 的企业级闪存存储，进一步提升 ONTAP AI 的部署速度。

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。一个 DGX-1 服务器即可提供超过 1 PFLOPS 的人工智能计算能力，相当于一个完整的采用传统 CPU 服务器的数据中心。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，来满足需求最苛刻的深度学习训练工作负载。

在使用 ImageNet 数据、搭载 NetApp AFF A800 系统和 NVIDIA DGX-1 服务器、以 1:4 存储计算比配置的条件下对 ONTAP AI 进行的测试中，实现了每秒 23000 个训练映像 (Training Images Per Second, TIPS) 的训练吞吐量和 60000 TIPS 的推理吞吐量。在此配置中，预期可以获得如下结果：超过 2 Gbps 的持续吞吐量（高峰时段 5 Gbps）、低于 1 毫秒的延迟，以及高于 95% 的 GPU 利用率。对于 NAS 工作负载，一个 AFF A800 系统支持吞吐量为 25 Gbps 的顺序读取和 100 万次 IOPS 的小型随机读取，同时保持低于 500 微秒的延迟。这些结果表明，性能余量可以支持更多的 DGX-1 服务器，能够满足需求增长。

提供业务所需的性能和可扩展性

ONTAP AI 支持客户从小规模入手，逐渐按需扩展；支持无中断向集群模式配置添加计算、存储和网络资源；支持以 1:1 存储计算比配置起步，并支持扩展至 1:5 甚至更高的比率配置，来满足数据增长需求。NetApp 的机架级架构支持企业从 AFF A220 入手，然后利用全闪存解决方案根据增长需要进行扩展，从数百 TB 到数十 PB，任您所选。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 NetApp 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，您可以求助单点联系支持以及我们的认证支持模式，获得问题解决方案和指导。

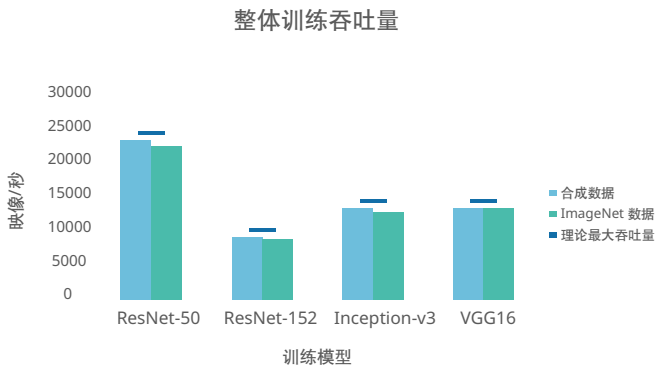

图 2) 所有模型的训练吞吐量。

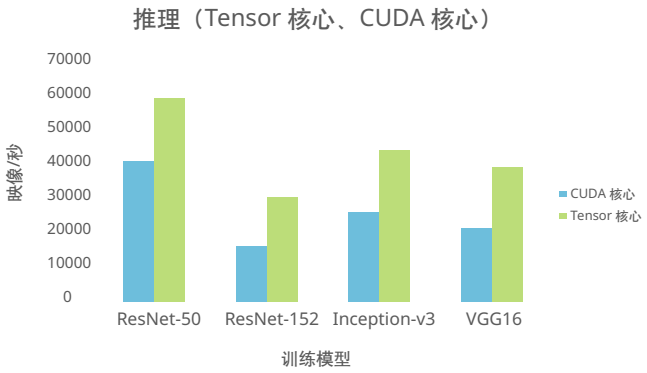

图 3) 所有训练模型的推理。

NetApp 与 NVIDIA：携手推动创新

ONTAP AI 的核心是 NVIDIA DGX-1 AI 超级计算机，一个专为深度学习打造的全面集成软硬件的系统。每台 DGX-1 服务器均由 8 个 Tesla V100 Tensor Core GPU 提供动力支持，配置在采用 NVIDIA NVLink 的混合式立方体网格拓扑结构中。DGX-1 与 NVLink 的结合，可以为多 GPU 训练所需的 GPU 间通信提供超高带宽和低延迟的优势，从而消除与基于 PCIe 的互连相关的瓶颈。DGX 平台采用经过优化的 NVIDIA GPU Cloud 深度学习软件堆栈技术，可以获得最大 GPU 加速深度学习性能。

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 能够以竞争友商解决方案 4 倍的速度更快地将数据馈送给 NVIDIA DGX-1 系统。[1]

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。

NetApp Data Fabric 提供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。ONTAP 借助业内领先的数据服务能力，通过一套工具管理和保护任意位置的数据，支持客户根据需要从边缘到核心再到云自由移动数据。

该解决方案集成 Cisco Nexus 3232C 100 Gb 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。现在，在使用 ONTAP AI 时，可以利用我们为 NVIDIA、NetApp 和 Cisco 业已验证的架构专设的单点联系支持来简化部署和管理。

解决方案组件

- NVIDIA DGX-1 服务器
- NetApp AFF A800 存储系统
- Cisco Nexus 3232C 网络交换机
- NVIDIA GPU Cloud 深度学习软件堆栈
- Trident — NetApp 的开源动态存储配置程序

关于 NVIDIA

NVIDIA（纳斯达克股票代码：NVDA）在 1999 年发明了 GPU，自此引发 PC 游戏市场的蓬勃发展，重新定义了现代计算机图形，并掀起了一波并行计算革命浪潮。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：**http://www.nvidia.com/dgx**。

关于 NetApp

NetApp 是混合云数据管理领域的权威企业。我们提供一系列混合云数据服务，旨在简化云端和内部环境中的应用程序及数据管理，加速推进数字化转型。NetApp 携手合作伙伴，赋予全球企业充分释放数据的全部潜能、增加客户接触点、扶植创新和优化企业运营的能力。有关详细信息，请访问 **www.netapp.com/cn**。#DataDriven

全国销售热线：4008-1818-11

NetApp 北京
北京市朝阳区东大桥路 9 号
侨福芳草地 C 座 6 层 606 室
邮编：100020
电话：86-10-59293000
传真：86-10-59293099

NetApp 上海
上海市静安区南京西路 338 号
天安中心 2503-2506 室
邮编：200003
电话：86-21-61328000
传真：86-21-61328001

NetApp 广州
广州市天河区天河路 385 号
太古汇 1 座 702 室
邮编：510620
电话：86-20-28317511
传真：86-20-28317515

NetApp 成都
成都市滨江东路 9 号
香格里拉办公室 18 楼
邮编：610021
电话：86-28-66065070
传真：86-28-66065071

NetApp 深圳
深圳市福田区中心四路 1 号
嘉里建设广场 3 座 6 楼 604 单元
邮编：518048
电话：86-755-82754900
传真：86-755-82754999

NetApp 杭州
杭州市西湖区学院路 28 号
德力西大厦 9 层
邮编：310012
电话：86-571-28091284
传真：86-571-28091277

NetApp 南京
南京市鼓楼区汉中路 2 号
亚太商务楼 8 层
邮编：210005
电话：86-25-66102617

NetApp 武汉
武汉市江岸区中山大道 1628 号
武汉天地企业中心 5 号 8 层
邮编：430014
电话：86-27-82206035
传真：86-27-82206177