经验证的 NetApp 架构

采用 NVIDIA DGX A100 系统和 Mellanox 频谱以太网交换机的 NetApp ONTAP AI

NVA 设计

NetApp 公司 David Arnette 和 Sung-Han Lin
2020年12月| NVA-1153-DESIGN

摘要

本文档介绍了经过 NetApp 验证的机器学习（ML）和人工智能（AI）工作负载架构，该架构使用 NetApp® AFF A800 存储系统，NVIDIA DGX ® A100 系统和 NVIDIA® Mellanox® Spectrum ® SN3700V 200 Gb 以太网交换机。此设计采用基于融合以太网的 RDMA（RoCE）作为计算集群互连网络结构，可为客户提供完全基于以太网的架构来处理高性能工作负载。本文档还包括所实施架构的基准测试结果。

合作方

目录

表格目录

插图目录

内容提要

本文档包含适用于机器学习（ML）和人工智能（AI）工作负载的 NetApp ONTAP ® AI 参考架构的验证信息。此设计是使用 NetApp AFF A800 全闪存存储系统，八个 DGX A100 系统和 SN3700V 交换机实施的，用于计算集群互连和存储连接。该系统的运行和性能已通过行业标准基准工具的验证，并已证明可提供卓越的训练性能。您还可以轻松、独立地将计算和存储资源从半机架配置扩展到多机架配置，并凭借可预测的性能满足任何机器学习工作负载要求。

计划摘要

经验证的 NetApp 架构计划为客户提供针对特定工作负载和使用情形的参考配置和规模估算指导。这些解决方案包括：

- 经过全面测试
- 旨在最大程度地降低部署风险
- 旨在加快上市速度

本文档主要面向 NetApp 及其合作伙伴的解决方案工程师以及客户的战略决策者。本节介绍了用于确定支持经验证的工作负载所需的特定设备，布线和配置的架构设计注意事项。

NetApp ONTAP AI 解决方案

NetApp ONTAP AI 参考架构由 DGX A100 系统和 NetApp 云互联存储系统提供支持，由 NetApp 和 NVIDIA 开发并验证。它为 IT 组织提供了一种架构，可以：

- 消除复杂设计
- 支持独立扩展计算和存储
- 支持从小规模起步，然后无缝扩展
- 提供广泛的存储选项，满足各种性价比需求

NetApp ONTAP AI 将 DGX A100 系统和 NetApp AFF A800 存储系统与一流的网络紧密集成在一起。NetApp ONTAP AI 可消除复杂设计，避免盲目猜测，从而简化人工智能 (AI) 的部署。贵企业可以从小规模起步然后进行无中断扩展，同时还能智能地管理从边缘到核心再到云以及反向的数据传输。

图 1 显示了采用 DGX A100 系统的 ONTAP AI 系列解决方案的多种变体。AFF A800 系统性能已通过多达八个 DGX A100 系统的验证。通过向 ONTAP 集群添加存储控制器对，该架构可以扩展到多个机架，以支持多个 DGX A100 系统和数 PB 的存储容量，并实现线性性能。采用这种方法时，可以根据数据湖大小、所使用深度学习 (DL) 模型以及所需性能指标灵活地独立调整计算与存储的比例。

图 1）采用 NVIDIA DGX A100 系统的 NetApp ONTAP AI 系列。

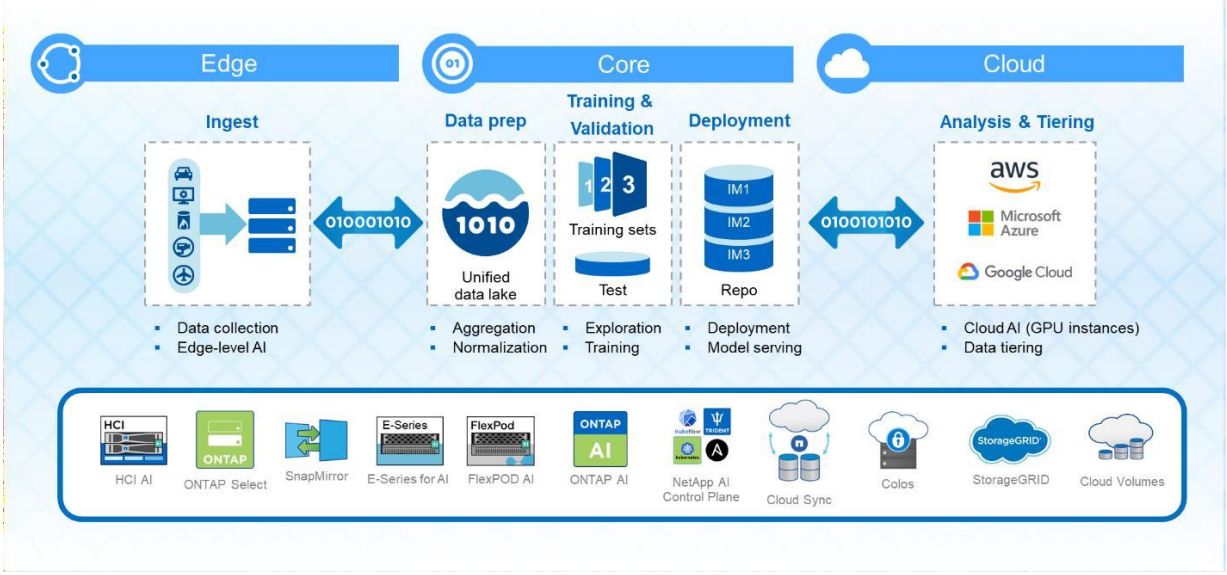
每个机架中 DGX-1 服务器和 AFF 系统的数量取决于所使用机架的电源和散热规格。系统的最终位置取决于计算流体动力学分析、气流控制和数据中心设计。

深度学习数据管道

深度学习是企业在竞争日益激烈的市场中发现欺诈行为，改善客户关系，优化供应链以及提供创新产品和服务的引擎。深度学习模型的性能和准确性可通过扩大神经网络的规模和复杂性以及提高训练模型所使用数据的数量和质量而得到大幅改进。

在具备海量数据集的情况下，搭建一个能够让您灵活地跨环境部署的基础架构是非常重要的。概括地讲，端到端深度学习部署由数据传输所经历的三个阶段组成：边缘（数据载入）、核心（训练集群和数据湖）和云（归档、分层和开发/测试）。这在数据跨数据管道所有三个阶段的物联网 (IoT) 等应用中非常典型。图 2 简要介绍了每个阶段的组成要素：

图 2）边缘 - 核心 - 云数据管道的组成部分。



以下列表介绍了在其中一个或多个领域中发生的一些活动。

- **载入**数据载入通常发生在边缘，例如，从自动驾驶汽车或销售点 (POS) 设备上捕获数据流。根据使用情形，可能需要在载入点或其附近部署 IT 基础架构。例如，零售商可能需要在每个商店占用少量空间来整合多台设备中的数据。
- **数据预处理**预处理是在训练前对数据进行清理使其规范化的必要步骤。预处理发生在数据湖中，数据湖可能是以 Amazon S3 层形式存在于云环境中，也可能以文件存储或对象存储形式存在于内部存储系统中。
- **培训和验证。**在关键的深度学习训练阶段，通常会定期将数据从数据湖复制到训练集群中。此阶段所用服务器使用 GPU 并行进行计算，从而形成巨大的数据处理能力。满足原始 I/O 带宽需求对于保持高 GPU 利用率至关重要。
- **部署。**经过训练的模型在测试后再部署到生产环境中。或者，可以将它们馈送回数据湖，以进一步调整输入权重；亦或在物联网应用中，也可以将这些模型部署到智能边缘设备。
- **分析和分层。**新的基于云的工具快速上市，因此可能需要在云中执行更多分析或开发工作。来自过去迭代的冷数据可能会无限期保存。许多人工智能团队倾向于将冷数据以对象存储形式归档到私有云或公共云中。根据计算要求，某些应用程序可以很好地将对象存储用作主数据层。

根据应用程序的不同，DL 模型可处理大量结构化和非结构化数据。这种不同会对底层存储系统提出各种不同的要求，包括存储数据的大小以及数据集中文件的数量。

高级存储要求包括：

- 能同时存储和检索数百万个文件
- 存储和检索多种数据对象，例如图像、音频、视频和时序数据
- 以低延迟提供并行高性能以满足 GPU 处理速度要求
- 实现跨边缘、核心和云的无缝数据管理和数据服务

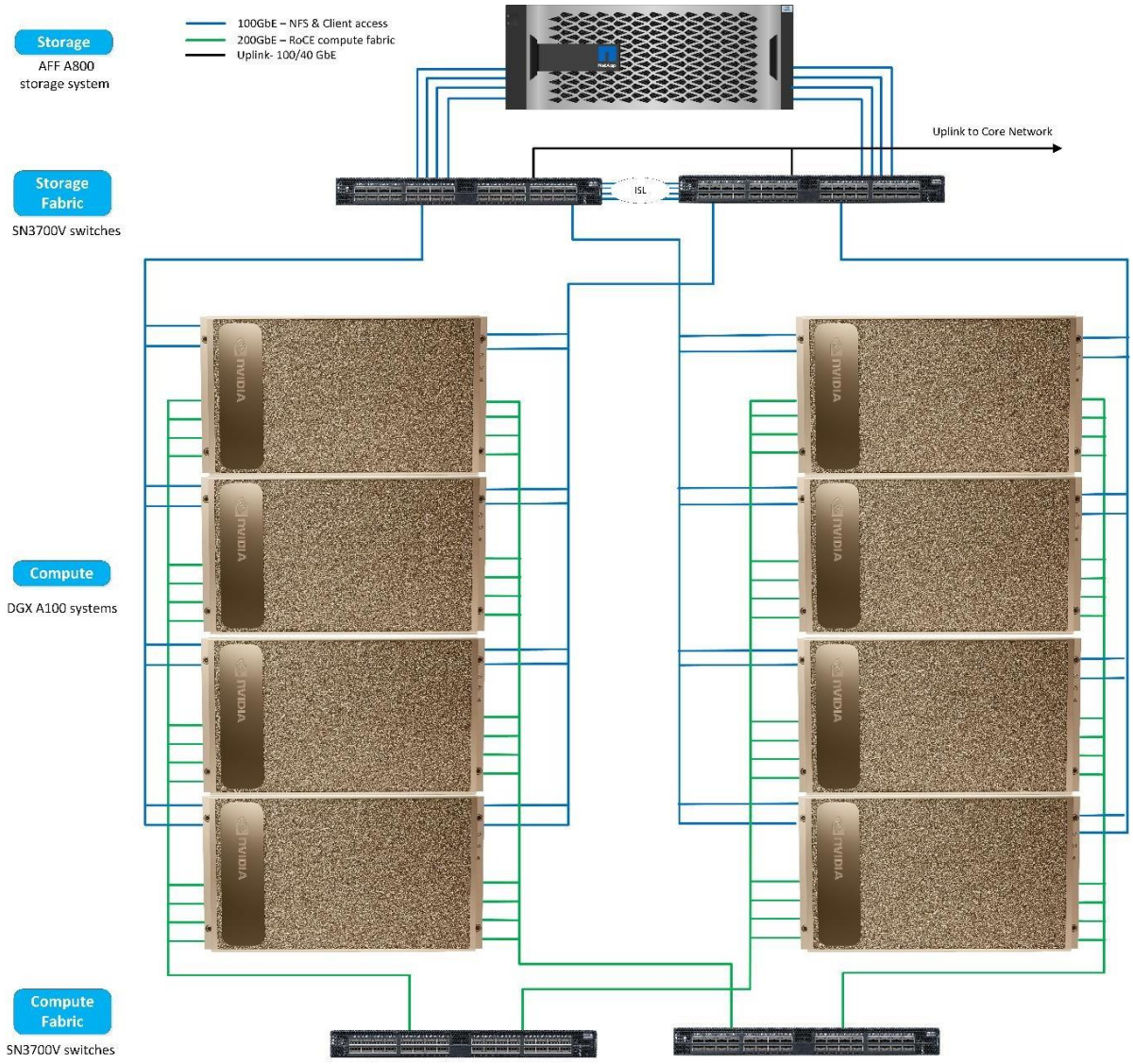
AFF 系统与 NetApp ONTAP 的卓越云集成功能以及软件定义的功能相结合，可支持跨边缘、核心和云的完整深度学习数据管道。本文档重点介绍数据管道的训练和推理组件的解决方案。

解决方案概述

深度学习系统所采用的算法属于计算密集型算法，而且特别适合 NVIDIA GPU 的架构。深度学习算法中执行的计算涉及并行运行的巨量矩阵乘法。利用 DGX 系统的单个和集群 GPU 计算架构的进步使其成为高性能计算（HPC），DL，视频处理和分析等工作负载的首选平台。要在这些环境中最大限度地提高性能，需要一个支持性基础架构，包括存储和网络，以使 GPU 能够始终为数据提供支持。因此，必须能以超低延迟和高带宽访问数据集。

此参考架构已通过一个 NetApp AFF A800 系统，八个 DGX A100 系统，两个 NVIDIA Mellanox Spectrum SN3700V 200 GB 以太网交换机（用于计算网络结构）以及另外两个 SN3700V 交换机（用于存储和客户端访问）的验证。图 3 显示了基础解决方案架构。

图 3）NetApp ONTAP AI 验证的架构。



NVIDIA DGX A100 系统

DGX-1 服务器是一个专门为深度学习工作流构建且软硬件全面集成的统包系统。每个 DGX A100 系统都由八个 NVIDIA A100 GPU 提供支持，这些 GPU 配置在采用 NVIDIA NVLink ® 和 NVIDIA NVSwitch ® 技术的混合立方体网状拓扑中。此配置可为 DGX A100 系统中的 GPU 间通信提供超高带宽，低延迟网络结构。这种拓扑结构对于多 GPU 训练必不可少，消除了基于 PCIe 的互连无法随 GPU 数量增长而实现性能线性提升的瓶颈。DGX A100 系统还配备了高带宽，低延迟的网络互连，用于通过 RoCE 和 InfiniBand 进行多节点集群。

NVIDIA NGC 总部

DGX A100 系统采用 NVIDIA NGC，这 是一种基于云的容器注册表，用于 GPU 加速软件。NGC 为当今针对 NVIDIA GPU 进行了优化的最常见深度学习框架（例如 Caffe2、TensorFlow、PyTorch、MXNet 和 TensorRT 等）提供容器。这些容器集成了框架或应用程序、必需驱动程序、库和通信原语，并由 NVIDIA 在整个堆栈上进行了优化，可提供最高 GPU 加速性能。NGC 容器采用 NVIDIA CUDA 工具包，其中包含 NVIDIA CUDA 基础线性代数子程序库 (CUDA Basic Linear Algebra Subroutines Library, cuBLAS)、NVIDIA CUDA 深度神经网络库 (CUDA Deep Neural Network Library, cuDNN) 等。NGC 容器还包含用于多 GPU 和多节点总体通信原语的 NVIDIA 集体通信库 (NVIDIA Collective Communications Library,

NCCL)，从而在进行深度学习训练时建立拓扑结构意识。NCCL 支持在一个 DGX A100 系统内的 GPU 之间以及在多个 DGX A100 系统之间进行通信。

NetApp AFF 系统

借助 NetApp AFF 一流的存储系统，IT 部门可以通过行业领先的性能，卓越的灵活性，云集成和一流的数据管理功能满足企业级存储需求。AFF 系统专为闪存设计，有助于加速、管理和保护业务关键型数据。

NetApp AFF A800 系统是行业首款端到端 NVMe 解决方案。对于 NAS 工作负载而言，一个 AFF A800 系统支持吞吐量为 25 GB/秒的顺序读取和 100 万次 IOPS 的小型随机读取，同时保持低于 500 微秒的延迟。

AFF A800 系统支持以下功能：

- 在包含 24 个节点的集群中实现最高 300 GB/秒的巨大吞吐量和 1140 万次 IOPS
- 100GbE 和 32 Gb FC 连接
- 最多 30 TB 固态驱动器（SSD），具有多流写入功能
- 在 2U 驱动器架中实现 2 PB 的高密度容量
- 从 200 TB（2 个控制器）扩展到 9.6 PB（24 个控制器）
- NetApp ONTAP 9.4，具有一整套数据保护和复制功能，可提供行业领先的数据管理

其他 NetApp 存储系统（例如 AFF A700，AFF A400 和 AFF A220）以较低的成本为较小的部署提供了较低的性能和容量选项。

NetApp ONTAP 9

ONTAP 9 是 NetApp 推出的最新一代存储管理软件，可帮助您打造现代化基础架构并向云就绪数据中心过渡。ONTAP 利用行业领先的数据管理功能，通过一套工具来管理和保护数据，而无论数据位于何处。还可以自由地将数据迁移到任何需要的地方，无论是边缘、核心还是云端。ONTAP 9 包含许多功能，不仅可以跨不同混合云架构简化数据管理，加速提供并保护关键数据，而且能帮助打造适应未来需求的基础架构。

简化数据管理

数据管理对于企业 IT 运营至关重要，通过得当的管理才能将适当的资源用于应用程序和数据集。ONTAP 具有以下功能，可简化操作并降低总运营成本：

- **实时数据缩减和扩展的重复数据删除。**数据压缩为 ML/DL 工作负载中经常使用的字母数字数据带来了主要优势。数据缩减可减少存储块内的空间浪费，而重复数据删除可大幅提升有效容量。
- **最低、最高和自适应服务质量 (QoS)。**精细的 QoS 控制有助于在高度共享的环境中保持关键应用程序的性能水平，并允许生产和开发部门共享基础架构，同时保证资源分配。

- **ONTAP FabricPool。**此功能支持将冷数据自动分层到公共云和私有云存储中，其中包括 Amazon Web Services (AWS)、Azure 和 NetApp StorageGRID® 解决方案。有关 FabricPool 的详细信息，请参见 TR-4598。

加快数据访问速度并提供数据保护

ONTAP 可提供卓越的性能和数据保护，而且可通过以下方式扩展这些功能：

- **高性能和低延迟。**ONTAP 尽可能以最低的延迟提供最高的吞吐量。
- **数据保护。**ONTAP 提供适用于所有平台的内置数据保护功能和通用管理功能。
- **NetApp 卷加密。**ONTAP 提供同时支持板载和外部密钥管理的本机卷级加密。
- **多租户和多因素身份验证。**ONTAP 支持以最高的安全性级别共享基础架构资源。

适应未来需求的基础架构

ONTAP 9 可帮助您满足瞬息万变的严苛业务需求：

- **无缝扩展和无中断运行。**ONTAP 支持向现有控制器和横向扩展集群无中断添加容量。您可以升级到 NVMe 和 32 Gb FC 等最新技术，而无需进行代价高昂的数据迁移或中断。
- **云连接。**ONTAP 是云互联支持最广泛的存储管理软件，在所有公共云中均提供适用于软件定义的存储 (ONTAP Select) 和云原生实例 (NetApp Cloud Volumes Service) 的选项。
- **与新兴应用程序相集成。**ONTAP 使用支持现有企业应用程序的相同基础架构为下一代平台和应用程序提供企业级数据服务。

NetApp FlexGroup 卷

培训数据集通常是一个由许多文件组成的大型集合，这些文件可能包含数十亿个文件。此类文件可能包括文本、音频、视频以及其他形式的非结构化数据，必须进行存储和处理才能并行读取。存储系统必须存储许多小文件，并且必须并行读取这些文件，以便执行顺序和随机 I/O

FlexGroup 卷（图 4）是由多个成员卷组成的单一命名空间，对存储管理员而言，就像 NetApp FlexVol® 卷一样加以管理和使用。FlexGroup 卷中的文件分配给各个成员卷，而且不会跨卷或节点进行条带化。它们支持以下功能：

- FlexGroup 卷可为高元数据工作负载提供高达 20 PB 的容量和可预测的低延迟。
- 它们在同一命名空间中最多支持 4000 亿个文件。
- 它们支持跨 CPU、节点、聚合和成员 FlexVol 卷并行运行 NAS 工作负载。

NetApp Trident

NetApp 提供的 Trident 是适用于 Docker 和 Kubernetes 的开源动态存储配置程序。Trident 与 NGC 和 Kubernetes 或 Docker Swarm 等常见业务流程协调程序相结合，支持您将深度学习 NGC 容器映像无缝部署到 NetApp 存储上，从而获得企业级人工智能容器部署体验。此类部署包括自动化流程编排、用于测试和开发的克隆、使用克隆进行升级测试、用于保护和满足合规性要求的副本以及针对 NGC AI 和 DL 容器映像的更多数据管理用例。

NVIDIA Mellanox 网络

NVIDIA Mellanox Spectrum 交换机—深度学习工作负载的理想选择

网络连接是 DL 基础架构的一个关键部分，负责高效，高效地在端点之间移动大量数据。具有一致性能，智能负载平衡和全面遥测功能的频谱以太网交换机是 DL 工作负载的理想网络元素。

稳定一致的性能

频谱以太网交换机可为 GPU 和 GPU 存储通信提供高带宽和稳定一致的低延迟数据路径。Spectrum 与 DGX A100 系统中的 NVIDIA Mellanox ConnectX ® 适配器一起，实施了一种紧密而高效的 ECN（显式拥塞通知）机制，可缓解瞬时拥塞并流畅的流量突发，从而最大程度地提高网络吞吐量。

智能负载平衡

网络是一种共享资源，必须在不同的流和端点之间公平地共享其带宽。数据包缓冲架构是交换机影响性能和流量公平的基本属性之一。Spectrum 交换机采用灵活且完全共享的缓冲区架构，可确保所有端口的性能均均衡，即使混合使用不同的端口速度也是如此。市场上的许多高速交换机都使用碎片化数据包缓冲区。缓冲区碎片化的交换机存在计划问题，可以优先为某些端口 / 流量提供更多带宽，而成本则其他端口 / 流量。这种流量不平衡会导致性能出现更多变化，进而影响分布式 DL 性能。

全面的遥测

要从深度学习基础架构中获得高投资回报，必须改善正常运行时间，并主动监控网络。传统的集中处理通过 SNMP 或流式传输获取的遥测数据的方法可能会以太网速度快速变得昂贵得令人无法承受。NVIDIA Mellanox What Just Happened®（WJH）利用硅级功能，在问题发生后立即快速识别并导出有关问题的粒度信息。由于此功能内置在平台中，因此中央数据收集器只会收集与问题描述相关的数据。WJH 可使主动式监控以太网速度实现可扩展性和实用性。借助 WJH，客户可以显著缩短解决问题描述的平均时间，并更好地规划容量。

技术要求

本节介绍 了用于解决方案验证 一节所述测试的硬件和软件。

硬件要求

表 1 列出了用于验证此解决方案的硬件组件。

表 1）硬件要求。

硬件	数量
DGX A100 系统	8
AFF A800 存储系统	1 个高可用性 (HA) 对，包括 48 个 1.92 TB NVMe SSD
SN3700V 以太网交换机	2 用于计算集群互连
	2 用于存储，客户端访问和带外管理

软件要求

表 2 列出了用于验证解决方案的软件组件。

表 2）软件要求。

软件	版本
ONTAP 存储 OS	9.7P6
网络交换机操作系统（全部）	Cumulus Linux 4.2.1
DGX 操作系统	4.99.10
Docker 容器平台	19.03.8[
容器版本	nvcr.io/nvidia/mxnet:20.06-py3 — MLPerf 测试 tensorflow:20.05-tf2-py3 —其他测试
OFED 版本	5.0-2.1.8
NCCL 测试版本	https://github.com/NVIDIA/nccl-tests/tree/ec1b5e22e618d342698fda659efdd5918da6bd9f
FIO 版本	3.1

解决方案架构

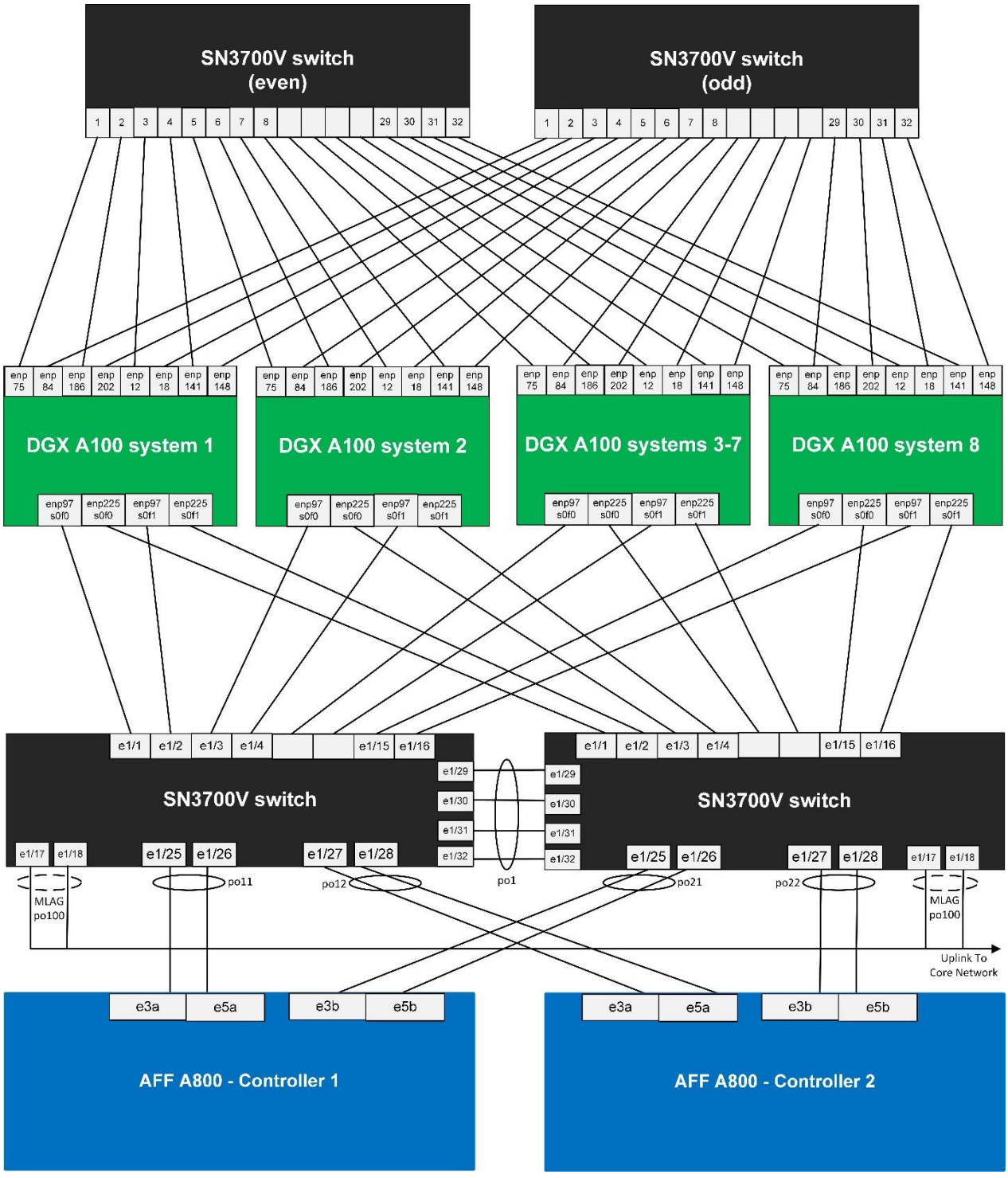
此架构经过验证，可满足运行深度学习工作负载的要求。有了这个验证结果，数据科学家可以在经过预先验证的基础架构上部署深度学习框架和应用程序，因此有助于消除风险，让企业集中精力从数据中获得有价值的洞察。此架构还可以为其他 HPC 工作负载提供出色的存储性能，而且无需对基础架构进行任何修改或调整。

网络拓扑结构和交换机配置

此参考架构利用单独的网络结构进行计算集群互连和存储访问。计算集群网络使用一对 SN3700V 以太网交换机，这些交换机作为独立的冗余网络结构运行。每个 DGX A100 系统都使用八个 200 Gbps 单端口 ConnectX-6 卡连接到交换机，其中偶数端口连接到一个交换机，奇数端口连接到另一个交换机。为 RoCE 配置了计算网络结构交换机，以便为 GPU 到 GPU 通信提供尽可能低的延迟。

另外，还使用两个 SN3700V 交换机来提供 NFS 存储连接以及对 DGX A100 系统的带内管理和客户端访问。这些 SN3700V 交换机配置了多机箱链路聚合（MLAG），以便在交换机发生故障时能够聚合带宽并进行透明故障转移。为以太网配置的两个双端口 ConnectX-6 卡用于从每个 DGX A100 系统向每个 SN3700V 交换机提供两个端口。每个卡的一个端口配置为专用于存储访问的绑定，每个卡上的另一个端口配置为绑定，用于带内管理和客户端访问。每个 AFF A800 存储系统都使用每个控制器的四个 100GbE 端口进行连接，并与每个交换机建立一个双端口 LACP 绑定，以便在存储控制器之间平衡工作负载分布。图 4 显示了整体网络拓扑。

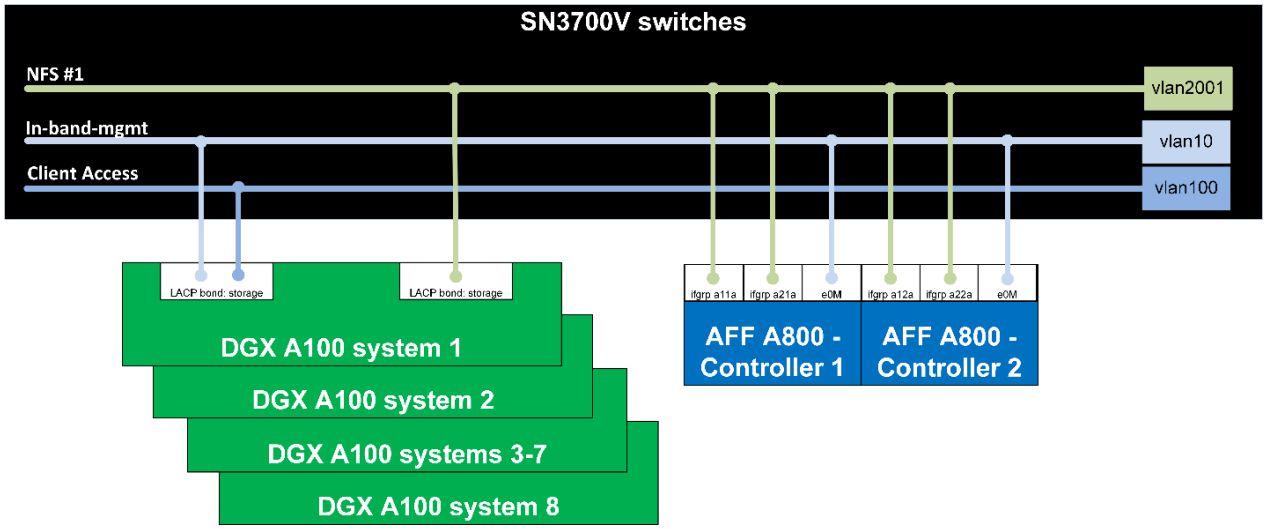
图 4）网络交换机端口配置。



以太网网络配置有多个 VLAN，用于隔离特定流量类型。NFS 存储流量，带内管理和客户端访问均具有专用 VLAN，可为每种流量类型提供适当的最大传输单元（Maximum Transmission Unit，MTU）和其他设置。例如，NFS 存储流量需要 MTU 9000，而其他典型以太网流量则使用 MTU 1500。

图 5 显示了主机和存储系统控制器的 VLAN 连接。请注意，AFF A800 存储系统控制器具有单独的 1GbE 管理接口，这些接口插入到单独的管理交换机中。

图 5）DGX-1 和存储系统端口的 VLAN 连接。

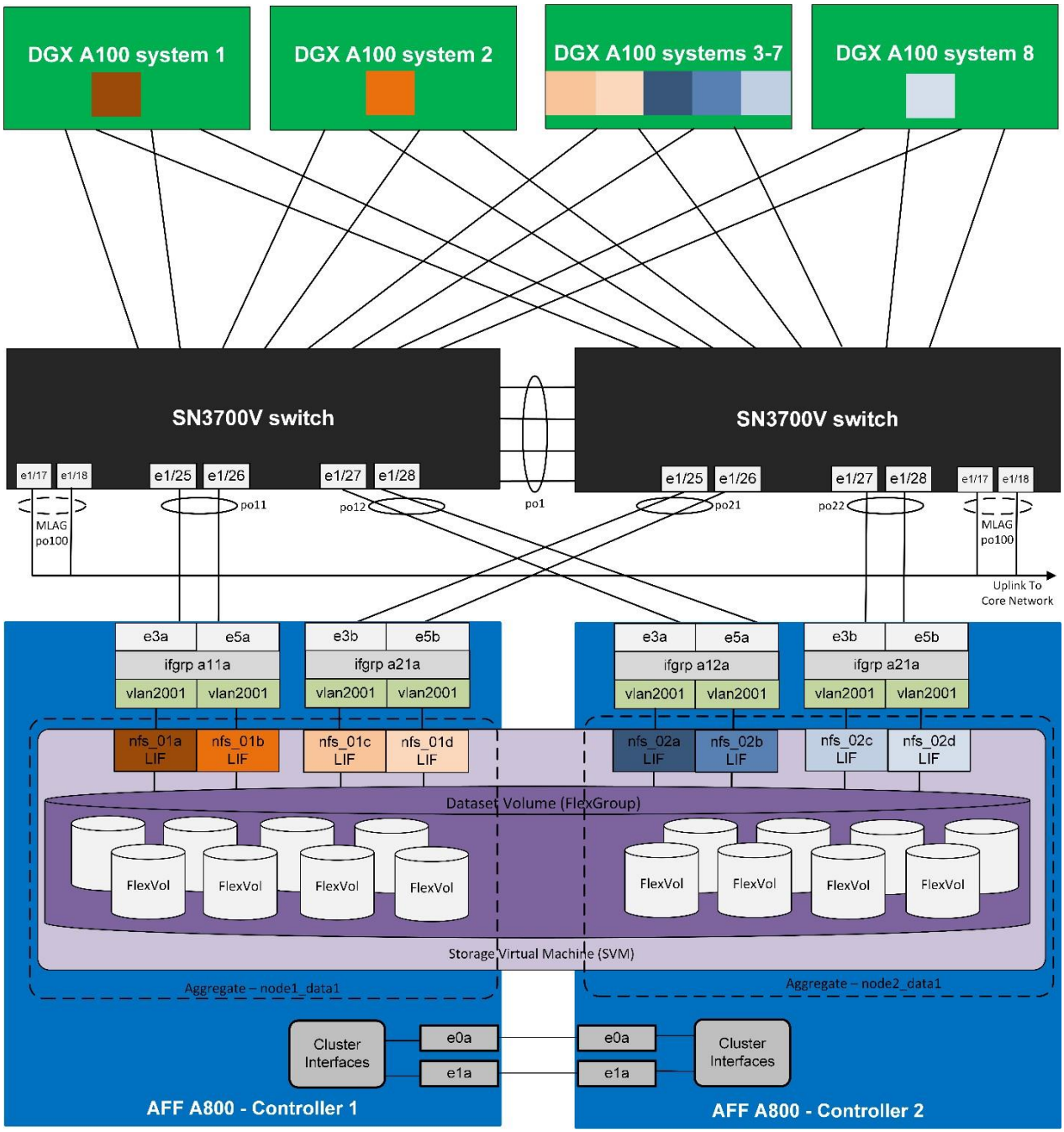


存储系统配置

为了满足此架构中任何潜在工作负载的存储网络需求，除了存储集群互连所需的板载端口外，每个存储控制器还配置了四个 100GbE 端口。图 6 显示了存储系统配置。在每个控制器上为每个交换机配置了一个双端口 LACP 接口组（图 6 中的 ifgrp）。这些接口组可提供与每个交换机高达 200 Gb/秒的弹性连接以用于数据访问。为 NFS 存储访问配置了两个 VLAN，两个存储 VLAN 都从交换机中继到这些接口组中的每个接口组。这种配置支持通过多个接口从每个主机并发访问数据，这样可增加可用于每个主机的潜在带宽。

从存储系统进行的所有数据访问均通过从专用于此工作负载的 Storage Virtual Machine (SVM) 的 NFS 访问提供。该 SVM 配置了总共四个逻辑接口 (LIF)，其中每个存储 VLAN 上两个 LIF。每个接口组托管一个 LIF，这样即是每个控制器上的每个 VLAN 一个 LIF，每个 VLAN 一个专用接口组。不过，两个 VLAN 均被中继到每个控制器上的两个接口组。此配置支持每个 LIF 故障转移到同一个控制器上的另一个接口组，这样两个控制器在发生网络故障时均可保持活动状态。

图 6）存储系统配置。

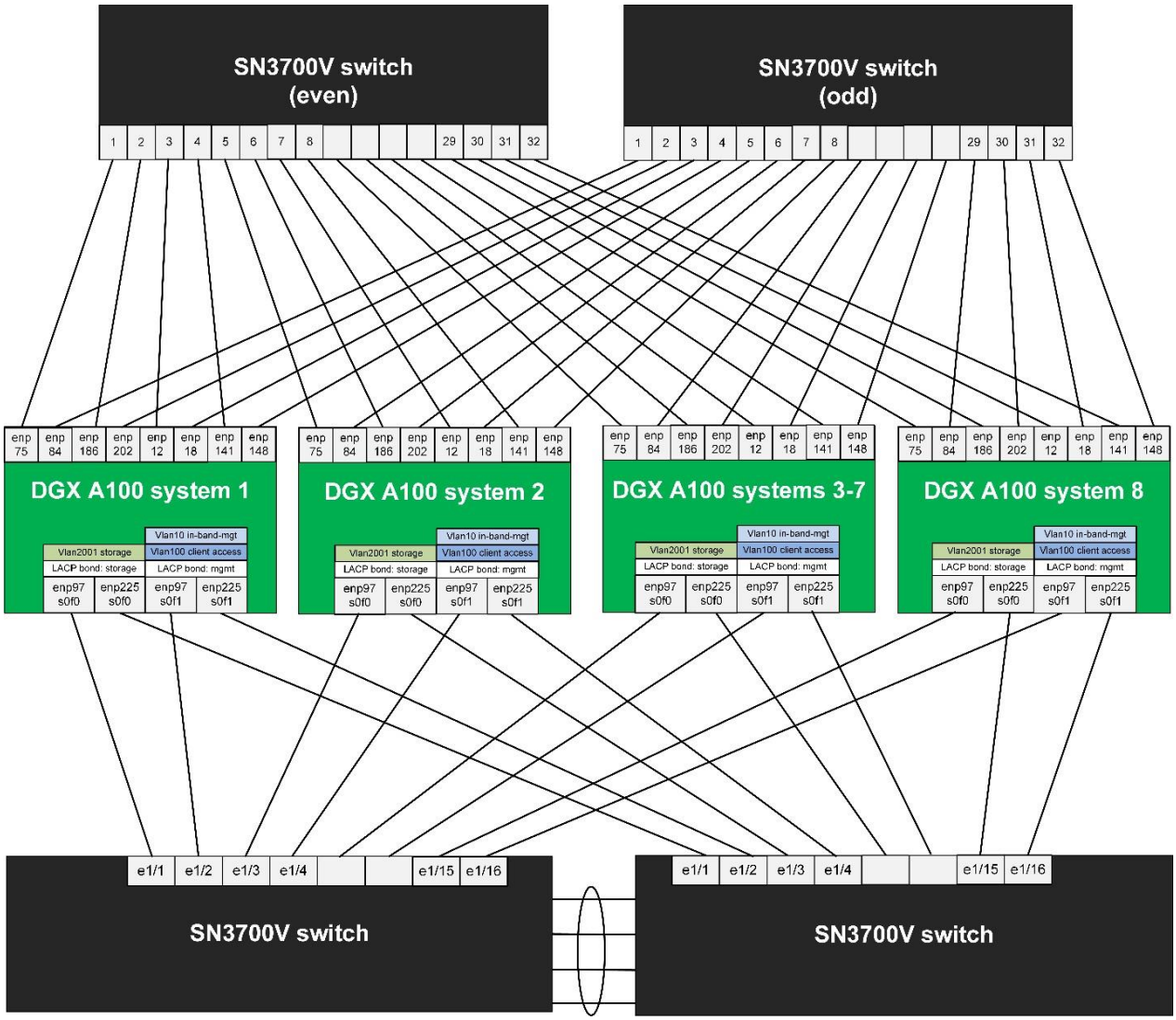


对于逻辑存储配置，此解决方案使用 FlexGroup 卷提供一个单一的存储池，该存储池分布在存储集群中的各节点上。每个控制器均托管一个含 46 个磁盘分区的聚合，每个磁盘由两个控制器共享。在数据 SVM 上部署 FlexGroup 时，便会在每个聚合上配置许多 FlexVol 卷，这些卷再结合成为 FlexGroup。通过这种方式，存储系统可提供单一存储池，其容量最高可扩展到阵列的最大容量，而且可通过同时利用阵列中的所有 SSD 提供出色的性能。NFS 客户端可通过为该 SVM 配置的任何 LIF 访问作为单个挂载点的 FlexGroup。您只需在存储集群中添加更多节点即可增加容量和客户端访问带宽。请注意，要使控制器或 FlexGroup 卷达到全部性能，不需要多个 IP 地址，但它们可以更好地在网络中实现哈希和负载分布。

主机配置

对于网络连接，每个 DGX A100 系统都配置有八个用于计算集群连接的 ConnectX-6 单端口网络接口卡和两个用于存储和客户端访问连接的 ConnectX-6 双端口卡。对于 InfiniBand 和以太网，这些卡支持高达 200 GB 的链路速度。在此参考架构中，为 200 GB RoCE 配置了八个单端口卡，并将其连接到一对 SN3700V 交换机，以实现计算集群连接。双端口卡上的端口连接到另一对 SN3700V 交换机，用于存储和客户端网络连接。图 7 显示了 DGX A100 系统的网络端口和 VLAN 配置。

图 7）DGX A100 系统的网络端口和 VLAN 配置。



对于以太网存储网络，主机端和交换机端的两个物理端口分别配置为 LACP 端口通道和 MLAG。另外两个端口配置为另一个 LACP 绑定，用于带内管理和客户端访问流量。由于 AFF A800 存储系统具有高性能功能，因此在此测试中已禁用主机端 NFS 文件系统缓存。

DGX OS 4.99 及更高版本使用 Linux 5.3 内核，其中包括 NFS nConnect 功能，可显著提高 NFSv3 存储性能。通过 nConnect，一个 NFS 挂载可以利用多个 TCP 会话来增加可用带宽，从而可能达到最大线速。此架构已通过 nConnect 的验证，可简化主机配置，同时提供与先前多个挂载配置相当的性能。下面列出了此测试中使用的特定主机端挂载参数：

- `nConnect=8`。为每个挂载的卷创建八个 TCP 会话以提高整体性能。
- `rsize=262144，wsize=262144`。将最大读写传输大小设置为 256k。ONTAP 支持高达 1 MB 的 NFS 传输大小，但测试表明，256 K 可以以最低延迟提供最大吞吐量。

解决方案验证

此参考架构已通过综合基准实用程序和深度学习基准测试进行验证，以确定系统的基线性能和操作。本节所述的每个测试都使用 技术要求中列出的特定设备和软件执行。

基础架构验证

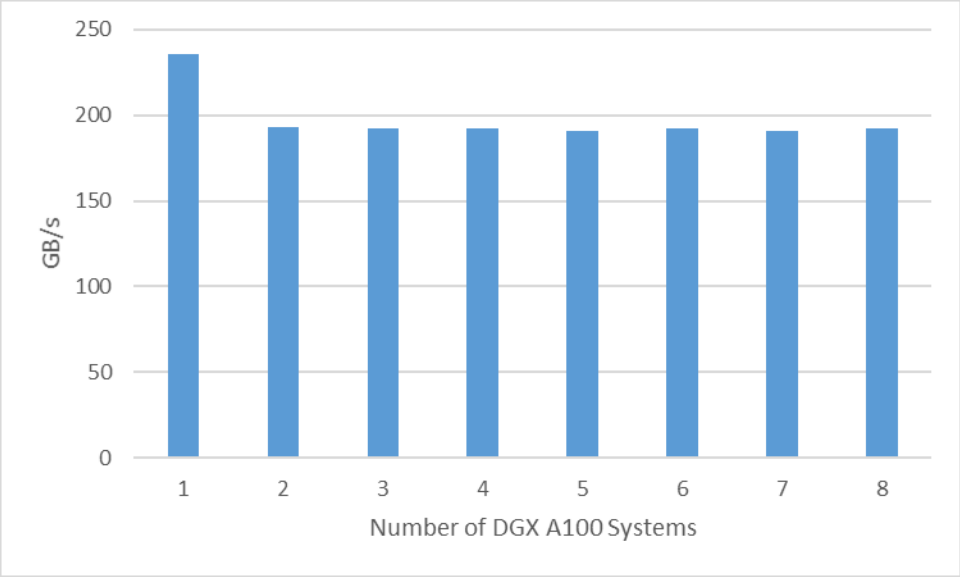
我们对一个，两个和四个 DGX A100 系统执行了以下测试，以验证所部署基础架构的基本操作和性能：

- NVIDIA nvsm 压力测试。此测试套件可对许多重要的 DGX A100 系统执行通过 / 未通过验证。所有系统都应报告此组中的测试的通过状态。
- NVIDIA NCCL all_reduce 性能
- FiO 带宽测试
- 每秒 FIO I/O 操作数（IOPS）测试
- 以下各节介绍了每个测试的详细信息和结果。

NVIDIA NCCL all_reduce 性能测试

此测试将验证 GPU 之间互连的性能。对于单节点系统，瓶颈应是 GPU 之间的 NVIDIA NVLink 连接。对于多节点系统，瓶颈应是 DGX A100 系统之间的以太网或 InfiniBand 连接。此测试使用所有八个可用物理连接测量系统之间的总带宽。
图 8 显示了 NCCL all_reduce 性能测试的结果。

图 8）NCCL 带宽测试结果。



FiO 带宽和 IOPS 测试

这些测试旨在使用合成 I/O 生成器工具 FIO 测量存储系统性能。我们使用了两种单独的配置，一种经过优化可提供最大带宽，另一种经过优化可实现 IOPS。每个配置都在运行时同时执行 100% 读取和 100% 写入，并创建 FIO 使用的文件作为一个单独的步骤，以便将这些活动与实际测试结果隔离开来。以下是这些测试的特定 FIO 配置参数：

- IOEngine = posixaio
- 直接 = 1
- 块大小 = 1024 k 用于带宽测试，4 k 用于 IOPS 测试
- numjobs = 120 用于带宽测试，180 用于 IOPS 测试
- iodepth = 32
- 大小 = 4194304k

利用这些测试中使用的工作负载参数，可以使用三个 DGX A100 系统使每个存储控制器饱和。图 9 显示了对多达八个 DGX A100 系统进行 FIO 带宽测试的结果。在此配置中，每个控制器上挂载四个主机，因此性能会线性扩展，直到第四个主机上的三个主机和平台达到单控制器最大约 22 GBps。接下来的四个主机将挂载到 HA 对中的第二个控制器上，并且对于这两个控制器，其行为相似，但最大不超过 45 GBps。

图 9）FIO 带宽测试结果（Gbps）。

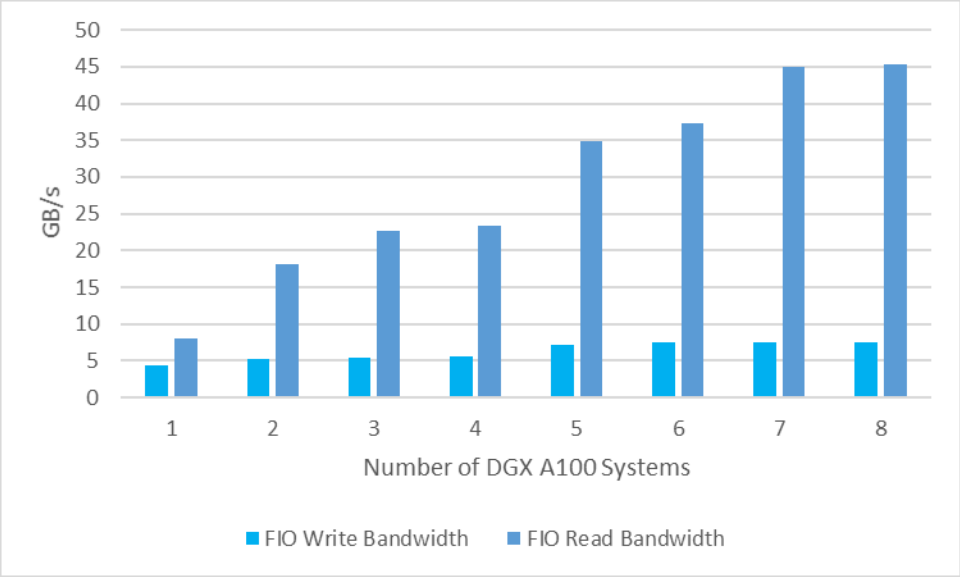
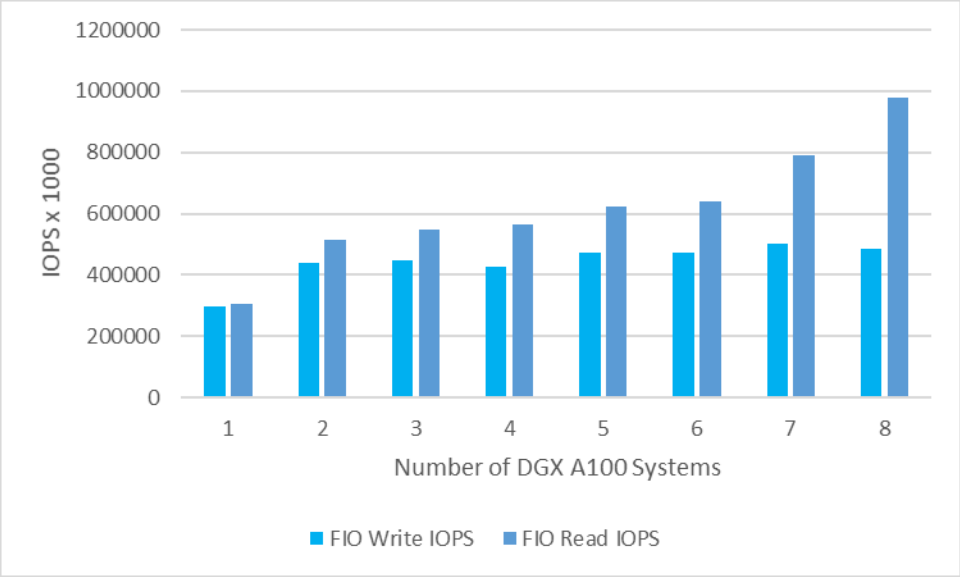


图 10 显示了 FIO IOPS 测试的结果。

图 10）FIO IOPS 测试结果（操作数 / 秒）。



深度学习工作负载验证

已使用 MLPerf Training v0.7 RESNET-50 基准测试验证了深度学习工作负载在已部署基础架构上的运行情况。此测试使用 MLPerf v0.7 测试标准验证使用 RESNET-50 模型的系统的性能，并使用在 MLPerf v0.7 测试规范中指定的参数和数据集。

下一节介绍了此测试的具体详细信息和结果。

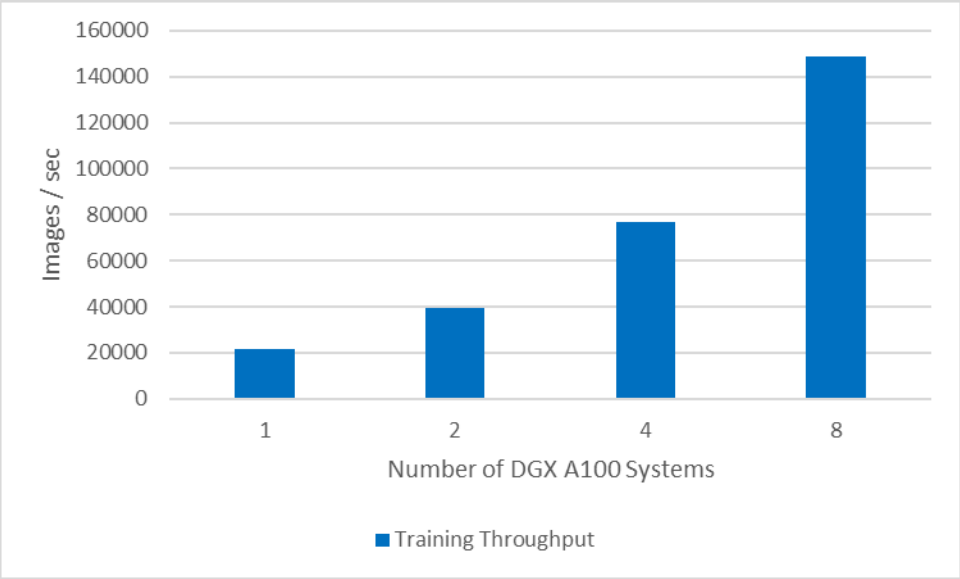
MLPerf 培训 v0.7 RESNET-50

此参考架构使用 MLPerf Training v0.7 基准测试，用于验证已部署基础架构上的深度学习工作负载的运行情况。MLPerf 是对各种神经网络的行业标准基准实施，用于验证深度学习基础架构的性能。此测试使用了采用 RESNET-50 的 MXNet 实施以及采用 IORecord 格式的 ImageNet 数据集来验证模型训练性能。Dali 用于加快数据的载入和预处理速度，Horovod 用于在多个 DGX A100 系统之间分发培训。随着工作负载的扩展（扩展能力较弱），显示的结果会使每个系统的批处理大小保持一致，即 408 个映像。

用于这些测试的基本容器映像是 NGC 中的 20.06 MXNet 映像。MLPerf 基准测试有意不针对任何特定硬件实施进行优化，因此，可以通过调整并发性等参数来提高这些测试中的整体系统性能。

图 11 显示了训练运行持续时间为 45 个时长的每秒平均图像数。

图 11）MLPerf 培训 v0.7 平均每秒映像数。



解决方案规模估算指导

此架构旨在供意欲采用 NVIDIA DGX-1 服务器和 NetApp AFF 系统实施高性能计算 (HPC) 基础架构的客户及合作伙伴作为参考。

如此验证所示，AFF A800 系统可轻松支持由八个 DGX A100 系统生成的深度学习培训工作负载。对于具有更高存储性能需求的更大规模部署，则可以在 NetApp ONTAP 集群中增加更多 AFF A800 系统。ONTAP 9 在一个集群中最多支持 12 个 HA 对（24 个节点）。借助此解决方案中验证的 FlexGroup 技术，一个 24 节点集群可以在一个卷中提供 20 PB 以上的吞吐量，并可提供高达 300 Gbps 的吞吐量。虽然此验证中使用的数据集相对较小，但 ONTAP 9 可以通过线性性能可扩展性扩展到令人惊叹的容量，因为每个 HA 对的性能均可与本文档中验证的级别相当。

AFF A400 等其他 NetApp 存储系统可为较小的部署提供较低的性能和容量选项，并且成本较低。根据此测试的结果，AFF A400 存储系统可以支持一个或两个 DGX A100 系统以及所测试的工作负载。由于 ONTAP 9 支持混合模式集群，因此您可以从更小占用空间起步，然后随着您的容量和性能需求的增长在集群中添加更多或更大的存储系统。

结论

DGX A100 系统是下一代深度学习平台，需要同样高级的存储和数据管理功能。通过将 DGX A100 与 NetApp AFF 系统相结合，可以几乎任意规模地实施这一经过验证的架构，从与 AFF A400 存储系统配对的单个 DGX A100 到 12 节点 AFF A800 集群上可能有 48 个 DGX A100 系统。AFF 与 NetApp ONTAP 的卓越云集成功能以及软件定义的功能相结合，可为成功实施深度学习项目提供跨边缘、核心和云的完整数据管道。

声明

作者对我们尊敬的 NVIDIA 和 NetApp 同事为本技术报告所做的贡献表示感谢。我们要对以自己的真知灼见为本白皮书的研究带来巨大帮助的所有人士表示真诚的赞赏和谢意。

从何处查找其他信息

如欲更详细地了解本文档所述的信息，请参见下列资源：

- NVA-1153-Deploy：采用 NVIDIA DGX A100 系统和 Mellanox Spectrum 以太网交换机的 NetApp ONTAP AI：
  www.netapp.com/pdf.html?item=/media/21789-nva-1153-deploy.pdf

NetApp AFF 系统：

- AFF 产品规格
  https://www.netapp.com/cn/media/ds-3582.pdf
- NetApp 借助 AFF 展现闪存优势
  https://www.netapp.com/us/media/ds-3733.pdf
- ONTAP 9.x 文档
  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286
- NetApp FlexGroup 技术报告
  https://www.netapp.com/cn/media/tr-4557.pdf

NetApp 互操作性表：

- NetApp 互操作性表工具
  http://support.netapp.com/matrix

NetApp Trident ：

- https://netapp.io/persistent-storage-provisioner-for-kubernetes/
- https://netapp-trident.readthedocs.io/en/stable-v19.04/kubernetes/index.html
- https://github.com/netapp/trident

NVIDIA DGX A100 系统

- NVIDIA DGX A100 系统
  https://www.nvidia.com/zh-cn/data-center/dgx-a100/
- NVIDIA Tesla A100 Tensor 核心 GPU
  https://www.nvidia.com/zh-cn/data-center/dgx-a100/
- NVIDIA GPU Cloud
  https://www.nvidia.com/en-us/gpu-cloud/

NVIDIA Mellanox 网络：

- NVIDIA Mellanox Spectrum SN3000 系列交换机
  https://www.mellanox.com/products/ethernet-switches/sn3000

机器学习框架：

- TensorFlow：适合所有人的开源机器学习框架
  https://www.tensorflow.org/
- Horovod：Uber 适用于 TensorFlow 的开源分布式深度学习框架
  https://eng.uber.com/horovod/
- 在容器运行时生态系统中启用 GPU
  https://devblogs.nvidia.com/gpu-containers-runtime/

数据集与基准测试：

- ImageNet
  http://www.image-net.org/
- MLPerf 培训和推理基准
  https://mlperf.org/

版本历史

版本	日期	文档版本历史
1.0 版	2020年11月	初始版本

要验证您的特定环境是否支持本文档所述的确切产品和功能版本，请参见 NetApp 支持站点上的互操作性表工具 (IMT)。NetApp IMT 中定义的产品组件和版本可用于构建 NetApp 所支持的配置。具体的配置结果取决于每个客户如何依照所发布规格进行安装。

版权信息

商标信息