边缘人工智能推理—采用联想 ThinkSystem 的 NetApp —解决方案设计

NetApp Solutions

NetApp
September 10, 2024

This PDF was generated from https://docs.netapp.com/zh-cn/netapp-solutions/ai/ai-edge-introduction.html on September 10, 2024. Always check docs.netapp.com for the latest.

目录

边缘人工智能推理—采用联想 ThinkSystem 的 NetApp —解决方案设计 1
TR-4886：《前沿人工智能推理—采用联想 ThinkSystem 的 NetApp —解决方案设计》 1
技术概述 5
测试计划 11
测试配置 12
测试操作步骤 14
测试结果 15
架构规模估算选项 20
结论 21

边缘人工智能推理—采用联想 ThinkSystem 的 NetApp —解决方案设计

TR-4886： 《前沿人工智能推理—采用联想 ThinkSystem 的 NetApp —解决方案设计》

NetApp 公司 Sathish Thyagarajan ，联想公司 Mirosav Hodak

本文档介绍了一种计算和存储架构，用于在边缘环境中针对新兴应用程序场景在 NetApp 存储控制器和联想 ThinkSystem 服务器上部署基于 GPU 的人工智能（ AI ）推理。

摘要

高级驾驶辅助系统（ ADAS ），行业 4.0 ，智能城市和物联网（ IoT ）等多种新兴应用场景需要在近乎零的延迟下处理持续数据流。本文档介绍了一种计算和存储架构，用于在边缘环境中为 NetApp 存储控制器和联想 ThinkSystem 服务器部署基于 GPU 的人工智能（ AI ）推理。本文档还提供了行业标准 MLPerf 推理基准测试的性能数据，用于评估配备 NVIDIA T4 GPU 的边缘服务器上的各种推理任务。我们对脱机，单流和多流推理情形的性能进行了调查，并显示具有经济高效的共享网络存储系统的架构具有高性能，为多个边缘服务器的数据和模型管理提供了一个中心。

简介

企业越来越多地在网络边缘生成海量数据。为了从智能传感器和物联网数据中获得最大价值，企业正在寻找支持边缘计算的实时事件流式解决方案。因此，在数据中心以外的边缘执行计算要求苛刻的作业的情况越来越多。AI 推理是这一趋势的推动因素之一。边缘服务器可以为这些工作负载提供足够的计算能力，尤其是在使用加速器时，但有限的存储通常是问题描述，尤其是在多服务器环境中。在本文档中，我们将介绍如何在边缘环境中部署共享存储系统，以及该系统如何在不影响性能的情况下为 AI 推理工作负载带来优势。

本文档介绍了边缘 AI 推理的参考架构。它将多个联想 ThinkSystem 边缘服务器与一个 NetApp 存储系统相结合，创建了一个易于部署和管理的解决方案。本指南旨在为各种情形下的实际部署提供一个基线指南，例如，在工厂车间安装多个摄像头和工业传感器，零售交易中的销售点（ POS ）系统或用于识别自动驾驶车辆中的视觉异常的完全自驾（ FSD ）系统。

本文档介绍了对由联想 ThinkSystem SE350 边缘服务器和入门级 NetApp AFF 和 EF 系列存储系统组成的计算和存储配置的测试和验证。这些参考架构可为 AI 部署提供高效且经济高效的解决方案，同时还可通过 NetApp ONTAP 和 NetApp SANtricity 数据管理软件提供全面的数据服务，集成数据保护，无缝可扩展性以及云连接数据存储。

目标受众

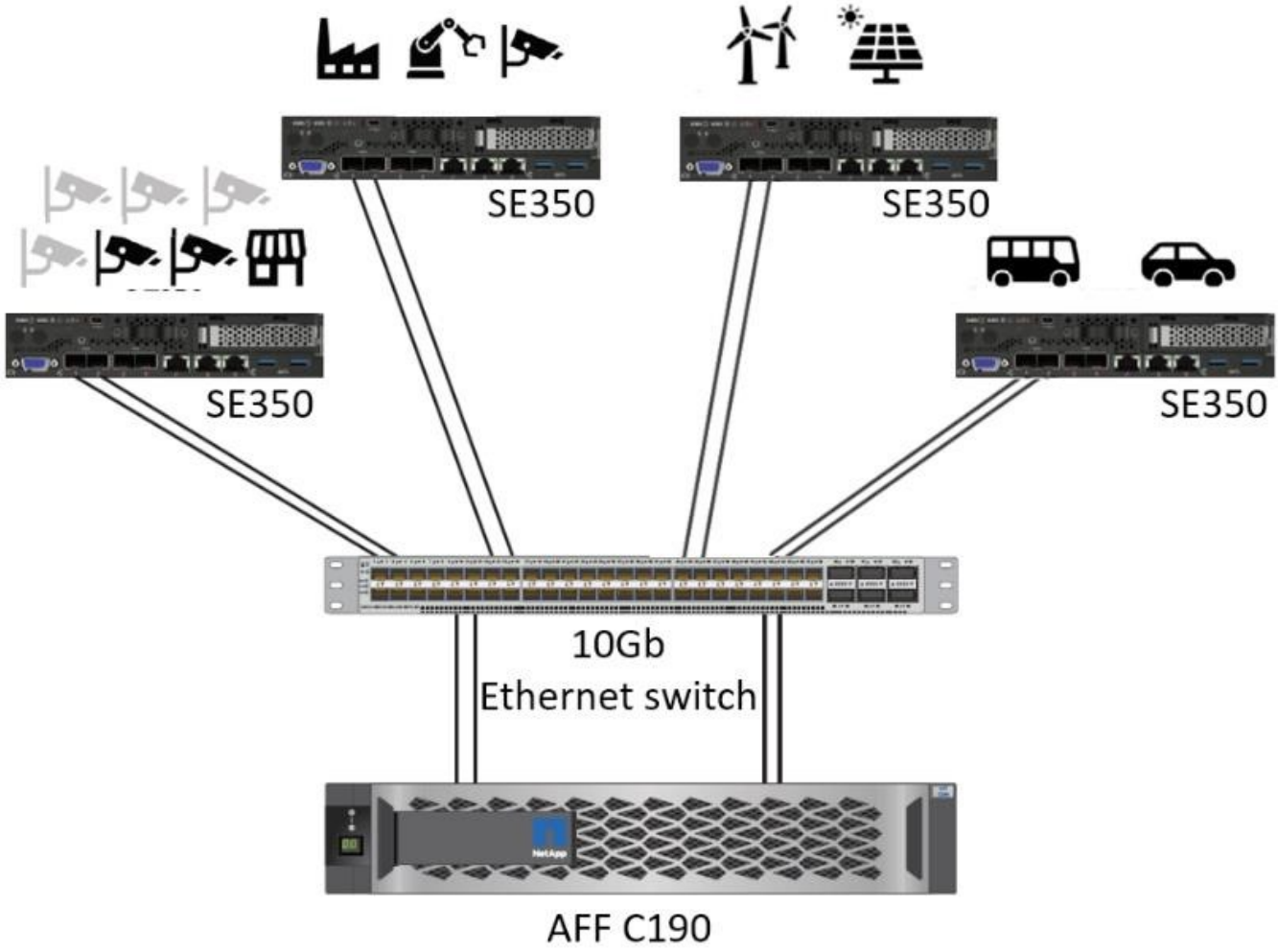
本文档面向以下受众：

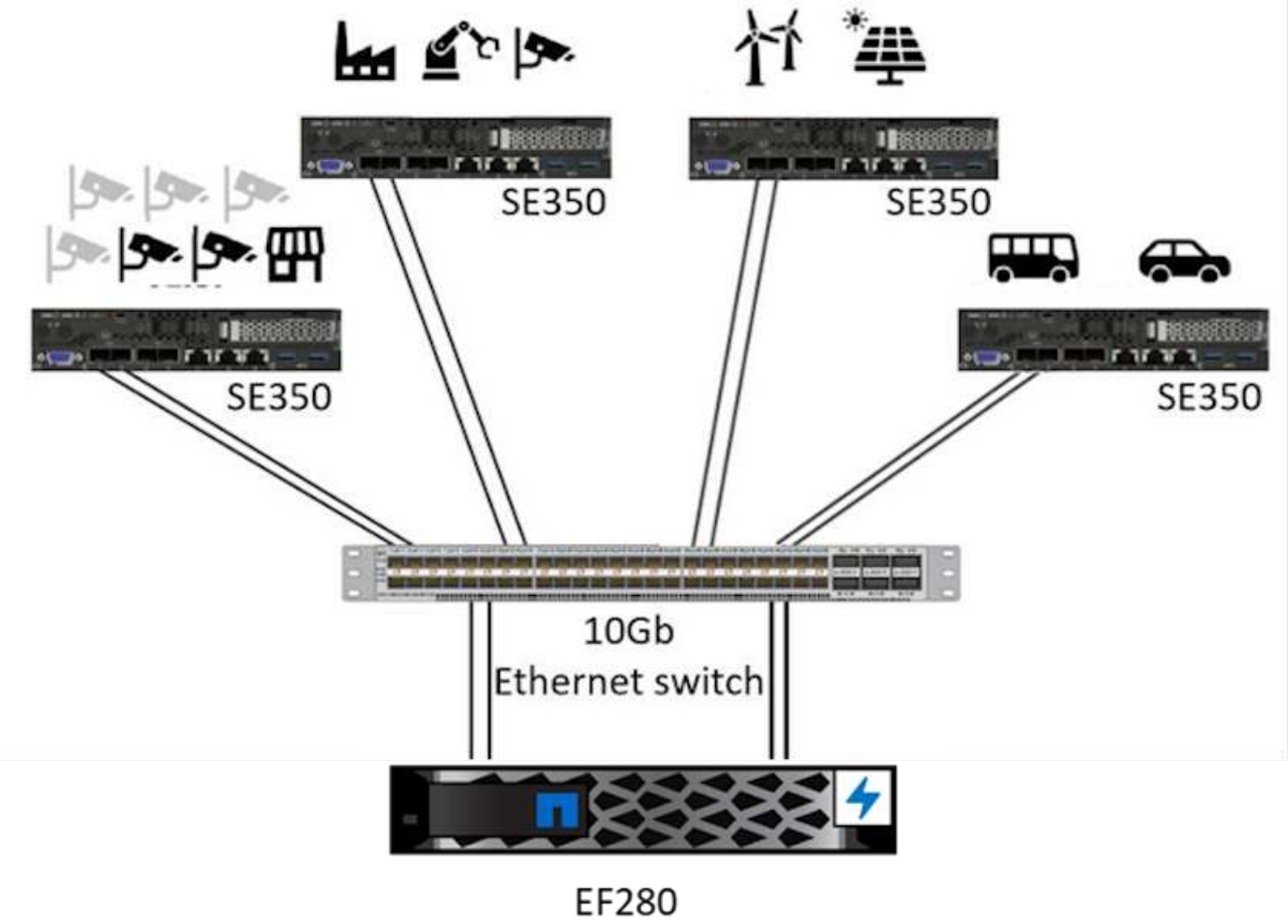
- 希望在边缘将 AI 产品化的业务主管和企业架构师。
- 数据科学家，数据工程师，人工智能 / 机器学习（ AI / 机器学习， ML ）研究人员和人工智能系统开发人员。
- 设计用于开发 AI/ML 模型和应用程序的解决方案的企业架构师。
- 数据科学家和 AI 工程师正在寻找部署深度学习（ DL ）和 ML 模型的高效方法。

- 边缘设备管理器和边缘服务器管理员，负责部署和管理边缘推理模型。

解决方案架构

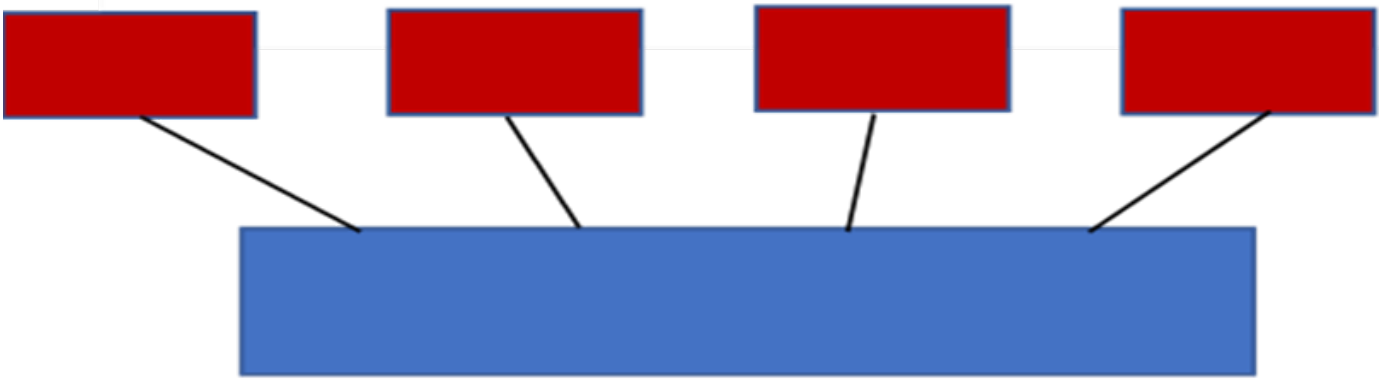
此联想 ThinkSystem 服务器和 NetApp ONTAP 或 NetApp SANtricity 存储解决方案旨在利用 GPU 与传统 CPU 的处理能力处理大型数据集上的 AI 推理。此验证通过一个架构来展示高性能和最佳数据管理，该架构使用一个或多个与单个 NetApp AFF 存储系统互连的联想 SR350 边缘服务器，如以下两个图所示。





下图中的逻辑架构概述显示了此架构中计算和存储元素的角色。具体而言，它显示以下内容：

- 边缘计算设备对从摄像机，传感器等接收的数据执行推理。
- 一种可用于多种用途的共享存储元素：
  - 为推理模型和执行推理所需的其他数据提供一个中央位置。计算服务器可以直接访问存储，并在网络中使用推理模型，而无需将其复制到本地。
  - 此处推送更新的型号。
  - 对边缘服务器接收的输入数据进行归档，以供日后分析。例如，如果边缘设备连接到摄像机，则存储元素会保留摄像机捕获的视频。



红色	蓝色
联想计算系统	NetApp AFF 存储系统
边缘设备对来自摄像机，传感器等的输入执行推理。	共享存储，用于存放边缘设备的推理模型和数据，以供日后分析。

NetApp 和联想解决方案具有以下主要优势：

- GPU 加快了边缘计算速度。
- 部署从共享存储提供支持和管理的多个边缘服务器。
- 强大的数据保护功能，可满足低恢复点目标（ RPO ）和恢复时间目标（ RTO ）的要求，而不会丢失任何数据。
- 利用 NetApp Snapshot 副本和克隆优化数据管理，以简化开发工作流。

如何使用此架构

本文档将验证建议架构的设计和性能。但是，我们尚未测试某些软件级别的组件，例如容器，工作负载或模型管理以及与内部云或数据中心的数据同步，因为它们是特定于部署情形的。此处有多种选择。

在容器管理级别， Kubernetes 容器管理是一个不错的选择，完全上游版本（ Canonical ）或适用于企业部署的修改版本（ Red Hat ）均支持此功能。。 "NetApp AI 控制平台" 使用 NetApp Trident 和新添加的 "NetApp DataOps 工具包" 为数据科学家和数据工程师提供内置可追溯性，数据管理功能，接口和工具，以便与 NetApp 存储集成。Kubeflow 是适用于 Kubernetes 的 ML 工具包，可在 TensorFlow Serving 或 NVIDIA Triton 推理服务器等多个平台上提供额外的 AI 功能以及对型号版本控制和 KFServing 的支持。另一个选项是 NVIDIA EGX 平台，它可提供工作负载管理以及对支持 GPU 的 AI 推理容器目录的访问。但是，这些选项可能需要大量的精力和专业知识才能投入生产，并且可能需要第三方独立软件供应商（ ISV ）或顾问的协助。

解决方案区域

AI 推理和边缘计算的主要优势是设备能够在无延迟的情况下以高质量计算，处理和分析数据。本文档中需要介绍的边缘计算用例太多，但下面是几个突出的示例：

汽车：自动驾驶汽车

典型的边缘计算图示位于自动驾驶汽车（ AV ）中的高级驾驶辅助系统（ ADAS ）中。无人驾驶汽车中的 AI 必须快速处理来自摄像机和传感器的大量数据，才能成为成功的安全驱动器。在对象和人类之间进行解释所花费的时间过长可能意味着生命或死亡，因此能够尽可能接近车辆处理数据至关重要。在这种情况下，一个或多个边缘计算服务器处理来自摄像机，雷达， LIDAR 和其他传感器的输入，而共享存储则保存推理模型并存储来自传感器的输入数据。

医疗保健：患者监控

AI 和边缘计算的最大影响之一是，它能够在家庭护理和集中护理部门（ ICU ）中增强对慢性病患者的持续监控。监控 Insulin 级别，呼吸，神经活动，心率和消化系统功能的边缘设备中的数据需要即时分析数据，必须立即对这些数据执行操作，因为拯救生命的行动时间有限。

零售：无收银员付款

边缘计算可以为 AI 和 ML 提供支持，帮助零售商缩短结账时间并增加流量。无收银员系统支持各种组件，例如：

- 身份验证和访问。将物理购物者连接到经过验证的帐户并允许访问零售空间。
- 清单监控。使用传感器， RFID 标签和计算机视觉系统帮助确认买家选择或取消选择商品。

  此处，每个边缘服务器都会处理每个签出计数器，而共享存储系统则充当一个中央同步点。

金融服务：信息亭的人员安全和防止欺诈

银行组织正在使用 AI 和边缘计算来创新和打造个性化的银行体验。利用实时数据分析和人工智能推理的交互式信息亭现在不仅可以帮助客户提取资金，还可以通过从摄像机捕获的图像主动监控信息亭，以识别对人类安全或欺诈行为的风险。在这种情况下，边缘计算服务器和共享存储系统连接到交互式信息亭和摄像机，以帮助银行使用 AI 推理模型收集和处理数据。

制造业：行业 **4.0**

第四次工业革命（工业 4.0 ）已经开始，同时也出现了 Smart Factory 和 3D 打印等新兴趋势。为迎接数据主导的未来，大规模机器到机器（ M2M ）通信和物联网已集成在一起，可提高自动化程度，无需人工干预。制造业已经高度自动化，增加 AI 功能是长期趋势的自然延续。AI 可实现自动化操作，借助计算机视觉和其他 AI 功能，可以实现自动化操作。您可以自动执行质量控制或依赖于人类远见或决策的任务，以便更快地分析工厂车间内装配线上的材料，从而帮助制造工厂满足所需的 ISO 安全和质量管理标准。此处，每个计算边缘服务器都连接到一组传感器，用于监控制造过程，更新的推理模型会根据需要推送到共享存储。

电信：防锈检测，塔式检查和网络优化

电信行业使用计算机视觉和人工智能技术来处理图像，这些图像可自动检测到是否存在防腐问题，并识别含有防腐问题的单元塔，因此需要进一步检查。近年来，使用无人机图像和 AI 模型来识别塔中不同的区域，以分析防腐，表面裂纹和防腐的情况有所增加。对 AI 技术的需求继续增长，这些技术可以高效地检查电信基础架构和单元塔，定期评估其降级情况，并在需要时及时修复。

此外，电信领域另一个新兴的使用情形是，使用 AI 和 ML 算法预测数据流量模式，检测支持 5G 的设备，以及自动化和增强多输入和多输出（ MIMG ）能源管理。在无线电塔上使用了 MIMO 硬件来增加网络容量，但这会带来额外的能源成本。在单元站点上部署的 ML 型号的 "MIMO 休眠模式 " 可以预测是否高效使用了无线电，并有助于降低移动网络运营商（ MNO ）的能耗成本。AI 推理和边缘计算解决方案可帮助 MNO 减少来回传输到数据中心的数据量，降低 TCO ，优化网络运营并提高最终用户的整体性能。

技术概述

本节介绍此 AI 解决方案的技术基础。

NetApp AFF 系统

一流的 NetApp AFF 存储系统支持在边缘进行 AI 推理部署，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满足企业级存储需求。NetApp AFF 系统专为闪存设计，可帮助加速，管理和保护业务关键型数据。

- 入门级 NetApp AFF 存储系统基于 FAS2750 硬件和 SSD 闪存介质
- HA 配置中的两个控制器

NetApp 入门级 AFF C190 存储系统支持以下功能：

- 最大驱动器数为 24 个 960 GB SSD
- 两种可能的配置：
  - 以太网（10GbE）： 4 个 10GBASE-T （RJ-45）端口
  - 统一（16 Gb FC 或 10GbE）： 4 个统一目标适配器 2 （UTA2）端口
- 最大有效容量为 50.5 TB

> 对于 NAS 工作负载，一个入门级 AFF C190 系统支持连续读取吞吐量为 4.4 GBps ，延迟为 1 毫秒或更短的小型随机读取吞吐量为 230 K IOPS 。

NetApp AFF A220

NetApp 还提供了其他入门级存储系统，可为大规模部署提供更高的性能和可扩展性。对于 NAS 工作负载，一个入门级 AFF A220 系统支持：

- 顺序读取的吞吐量为 6.2 GBps
- 在延迟不超过 1 毫秒的情况下进行小型随机读取时，可实现 375000 IOPS
- 最大驱动器数为 144 个 960 GB ， 3.8 TB 或 7.6 TB SSD
- AFF A220 可扩展到 1 PB 以上的有效容量

NetApp AFF A250

- 最大有效容量为 35 PB ，最大横向扩展为 2-24 个节点（12 个 HA 对）
- 与 AFF A220 相比，性能提高≥ 45%
- 440 ， 000 次 IOPS 随机读取 @1 毫秒
- 基于最新的 NetApp ONTAP 版本 ONTAP 9.8 构建
- 利用两个 25 Gb 以太网实现 HA 和集群互连

NetApp E 系列 EF 系统

EF 系列是一组入门级和中端全闪存 SAN 存储阵列，可通过 NetApp SANtricity 软件加快数据访问速度，并帮助您更快地从中获得价值。这些系统提供 SAS 和 NVMe 闪存存储，并为您提供经济实惠的至极高 IOPS ， 100 微秒以下的响应时间以及高达 44 GBps 的带宽，使其成为混合工作负载以及 AI 推理和高性能计算（ HPC ）等要求苛刻的应用程序的理想之选。

下图显示了 NetApp EF280 存储系统。

NetApp EF280

- 支持 32 Gb/16 Gb FC ， 25 Gb/10 Gb iSCSI 和 12 Gb SAS
- 最大有效容量为 96 个驱动器，总容量为 1.5 PB
- 10 Gbps 吞吐量（顺序读取）
- 30 万次 IOPS （随机读取）
- NetApp EF280 是 NetApp 产品组合中成本最低的全闪存阵列（ AFA ）

NetApp EF300

- 24 个 NVMe SSD 驱动器，总容量为 367 TB
- 扩展选项，总容量为 240 个 NL-SAS HDD ， 96 个 SAS SSD 或两者的组合
- 100 GB NVMe/IB ， NVMe/RoCE ， iSE/IB 和 SRP/IB
- 32 GB NVMe/FC ， FCP
- 25 GB iSCSI
- 20 Gbps （顺序读取）
- 670K IOPS （随机读取）

有关详细信息，请参见 "NetApp EF 系列 NetApp EF 系列全闪存阵列 EF600 ， F300 ， EF570 和 EF280 产品规格"。

NetApp ONTAP 9.

ONTAP 9.8.1 是 NetApp 推出的最新一代存储管理软件，可帮助企业打造现代化的基础架构并过渡到云就绪数据中心。借助行业领先的数据管理功能，无论数据位于何处， ONTAP 都可以通过一组工具来管理和保护数据。您还可以将数据自由移动到需要的任何位置：边缘，核心或云。ONTAP 9.8.1 提供了许多功能，可简化数据管理，加快和保护关键数据，并在混合云架构中实现下一代基础架构功能。

简化数据管理

数据管理对于企业 IT 运营至关重要，以便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。 * 数据缩减可减少存储块中浪费的空间，重复数据删除可显著提高有效容量。此适用场景数据存储在本地，并分层到云。
- * 最低，最高和自适应服务质量（ AQoS ）。 * 细粒度服务质量（ 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。
- * NetApp FabricPool 。 * 此功能可将冷数据自动分层到公有和私有云存储选项，包括 Amazon Web Services （ AWS ）， Azure 和 NetApp StorageGRID Storage 解决方案。有关 FabricPool 的详细信息，请参见 "TR-4598"。

加速和保护数据

ONTAP 9 可提供卓越的性能和数据保护，并通过以下方式扩展这些功能：

- * 性能和低延迟。 * ONTAP 可提供尽可能高的吞吐量，并尽可能降低延迟。
- * 数据保护。 * ONTAP 提供内置数据保护功能，并在所有平台之间进行通用管理。
- * NetApp 卷加密（ NVE ）。 * ONTAP 提供原生卷级加密，并支持板载和外部密钥管理。
- * 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。

Future-Proof 基础架构

ONTAP 9 具有以下功能，可满足不断变化的苛刻业务需求：

- * 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。客户可以升级到 NVMe 和 32 Gb FC 等最新技术，而无需进行成本高昂的数据迁移或中断。
- * 云连接。 * ONTAP 是云连接最广泛的存储管理软件，可在所有公有云中选择软件定义的存储（ ONTAP Select ）和云原生实例（ NetApp Cloud Volumes Service ）。
- * 与新兴应用程序集成。 * ONTAP 使用支持现有企业级应用程序的相同基础架构，为下一代平台和应用程序（例如自动驾驶汽车，智能城市和行业 4.0 ）提供企业级数据服务。

NetApp SANtricity

NetApp SANtricity 旨在为 E 系列混合闪存和 EF 系列全闪存阵列提供行业领先的性能，可靠性和精简性。为繁重工作负载应用程序（包括数据分析，视频监控以及备份和恢复）实现 E 系列混合闪存和 EF 系列全闪存阵列的最高性能和利用率。借助 SANtricity ，可以在存储保持联机状态的同时完成配置调整，维护，容量扩展和其他任务。SANtricity 还提供卓越的数据保护，主动监控和认证安全性—所有这些功能均可通过易于使用的机载 System Manager 界面进行访问。要了解更多信息，请参见 "NetApp E 系列 SANtricity 软件产品规格"。

性能优化

经过性能优化的 SANtricity 软件可为您的所有数据分析，视频监控和备份应用程序提供具有高 IOPS ，高吞吐量和低延迟的数据。提高高 IOPS ，低延迟应用程序和高带宽，高吞吐量应用程序的性能。

最大限度地延长正常运行时间

在存储保持联机的情况下完成所有管理任务。在不中断 I/O 的情况下调整配置，执行维护或扩展容量借助自动化功能，联机配置，最先进的动态磁盘池（ Dynamic Disk Pool ， DPP ）技术等实现同类最佳的可靠性。

请轻松休息

SANtricity 软件可通过易于使用的机载 System Manager 界面提供卓越的数据保护，主动监控和认证安全性。简化存储管理任务。获得对所有 E 系列存储系统进行高级调整所需的灵活性。随时随地管理您的 NetApp E 系列系统。我们基于 Web 的盒装界面简化了您的管理工作流。

NetApp Trident

"Trident" NetApp 是适用于 Docker 和 Kubernetes 的开源动态存储编排程序，可简化永久性存储的创建，管理和使用。Trident 是 Kubernetes 原生应用程序，直接在 Kubernetes 集群中运行。借助 Trident ，客户可以将 DL 容器映像无缝部署到 NetApp 存储上，并为 AI 容器部署提供企业级体验。Kubernetes 用户（例如 ML 开发人员和数据科学家）可以创建，管理和自动化流程编排和克隆，从而充分利用 NetApp 技术提供的 NetApp 高级数据管理功能。

NetApp BlueXP复制和同步

"BlueXP复制和同步" 是一项 NetApp 服务，用于快速安全地同步数据。无论您是需要在内部NFS还是SMB文件共享、NetApp StorageGRID、NetApp ONTAP S3、NetApp Cloud Volumes Service、Azure NetApp Files 、Amazon Simple Storage Service (Amazon S3)、Amazon Elelic File System (Amazon EFS)、Azure Blb 、Google Cloud Storage、 或者IBM Cloud Object Storage、BlueXP Copy and Sync可将文件快速安全地移动到您需要的位置。数据传输完成后，即可在源和目标上完全使用。BlueXP复制和同步功能会根据预定义的计划持续同步数据、仅移动增量、从而最大限度地减少数据复制所需的时间和资金。BlueXP Copy and Sync是一款软件即服务(SaaS)工具、设置和使用极其简单。由BlueXP复制和同步触发的数据传输由数据代理执行。您可以在AWS、Azure、Google Cloud Platform或内部部署BlueXP Copy and Sync数据代理。

联想 **ThinkSystem** 服务器

联想 ThinkSystem 服务器采用创新的硬件，软件和服务，可解决客户当前面临的挑战，并提供一种循序渐进的模块化设计方法来应对未来的挑战。这些服务器利用同类最佳的行业标准技术以及联想的差异化创新技术，为 x86 服务器提供最大的灵活性。

部署联想 ThinkSystem 服务器的主要优势包括：

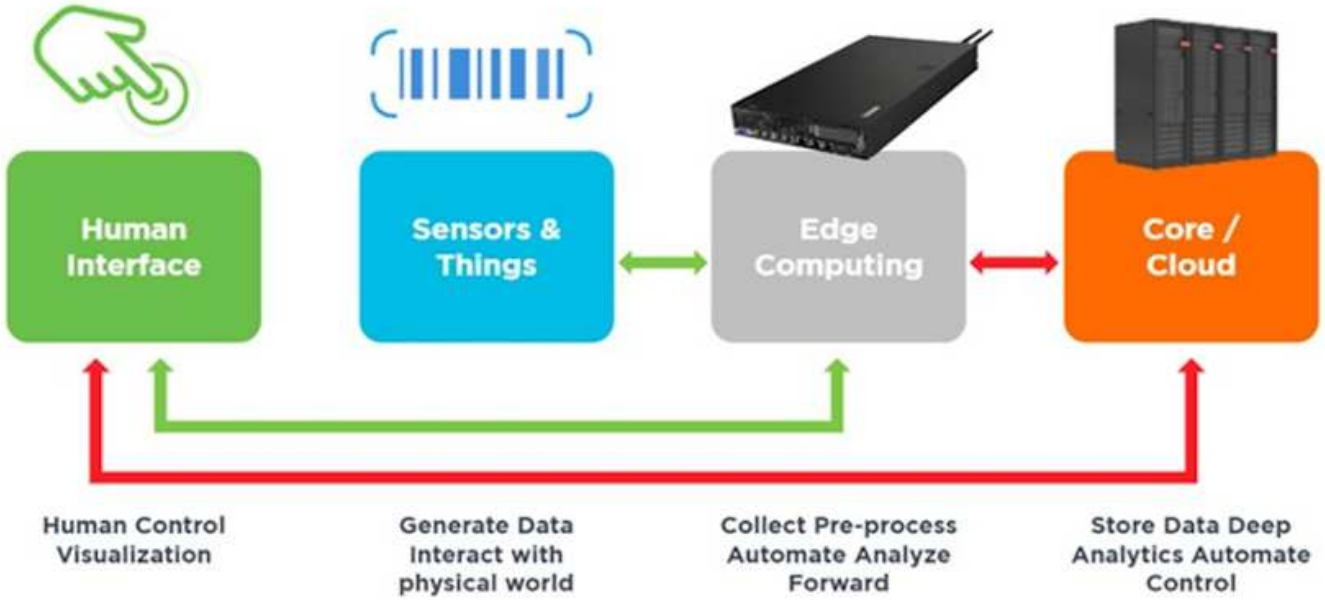
- 高度可扩展的模块化设计，可随业务发展而扩展
- 行业领先的弹性，可节省数小时的成本高昂的计划外停机时间
- 快速闪存技术可实现更低的延迟，更快的响应速度以及更智能的实时数据管理

在 AI 领域，联想正在采取切实可行的方法帮助企业了解 ML 和 AI 的优势并将其用于工作负载。联想客户可以在联想 AI 创新中心探索和评估联想 AI 产品，以充分了解其特定用例的价值。为了缩短实现价值的时间，这种以客户为中心的方法可以为客户提供解决方案开发平台的概念验证，这些平台已准备就绪，可供 AI 使用并进行优化。

联想 **ThinkSystem SE350** 边缘服务器

通过边缘计算，可以在将 IoT 设备中的数据发送到数据中心或云之前在网络边缘对其进行分析。如下图所示，联想 ThinkSystem SE350 专为满足边缘部署的独特要求而设计，采用紧凑的加固型环境加固外形，重点关注灵活性，连接性，安全性和远程易管理性。

SE350 采用 Intel Xeon D 处理器，可以灵活地支持边缘 AI 工作负载的加速，专为应对数据中心以外各种环境中服务器部署的挑战而构建。



MLPerf

MLPerf 是用于评估 AI 性能的行业领先基准套件。它涵盖应用 AI 的许多方面，包括图像分类，对象检测，医学影像和自然语言处理（ NLP ）。在此验证中，我们使用了推理 v0.7 工作负载，这是此验证完成时 MLPerf 推理的最新迭代。。 "MLPerf 推理 v0.7" 套件包括四个适用于数据中心和边缘系统的新基准：

- Transformers （ Bert ）提供的 * 双向编码器表示法经过微调，可使用 squad 数据集回答问题。

- * 深度学习建议模式（ DLRM ）是一种个性化和建议模式，经过培训可优化点击率（ CTR ）。
- * 3D U-Net.* 3D U-Net 架构接受过有关脑肿瘤分段（ Bras ）数据集的培训。
- * RNN-T.* 经常性神经网络传感器（ RNNP-T ）是一种自动语音识别（Automatic Speech Recognition ，ASR ）模型，该模型经过部分 LibriSpeech 的训练。MLPerf 推理结果和代码已公开发布，并已获得 Apache 许可证。MLPerf 推理具有一个 Edge 分支，可支持以下情形：
- * 单一流。 * 此场景模拟响应能力是关键因素的系统，例如在智能手机上执行脱机 AI 查询。各个查询将发送到系统并记录响应时间。结果将报告所有响应的 90 百分位延迟。
- * 多流。 * 此基准测试适用于处理来自多个传感器的输入的系统。在测试期间，系统会按固定的时间间隔发送查询。会施加 QoS 限制（允许的最大延迟）。此测试将报告系统在满足 QoS 限制的情况下可以处理的流数量。
- * 脱机。 * 这是涉及批处理应用程序的最简单情形，指标是每秒样本吞吐量。系统可以使用所有数据，基准测试可测量处理所有样本所需的时间。

联想已发布了本文档中使用的服务器 SE350 与 T4 的 MLPerf 推理得分。有关结果，请参见 "https://mlperf.org/inference-results-0-7/" 在条目 #0.7-145 的 " 边缘，封闭分区 " 一节中。

测试计划

本文档遵循 MLPerf 推理 v0.7 "code"， MLPerf 推理 v1.1 "code"，和 "rules"。我们运行了专为在边缘进行推理而设计的 MLPerf 基准测试，如下表所述。

区域	任务	型号	数据集	**QSL** 大小	质量	多流延迟限制
愿景	映像分类	Resnet50v1.5	ImageNet （ 224x224 ）	1024	FP32 的 99%	50 毫秒
愿景	对象检测（大型）	SSD - ResNet34	可可可（ 1200 x 1200 ）	64	FP32 的 99%	66 毫秒
愿景	对象检测（小型）	SSD — MobileNetsv1	可可可（ 300 x 300 ）	256	FP32 的 99%	50 毫秒
愿景	医学影像分段	3D UNET	Brats 2019 （ 224x224x160 ）	16.	FP32 的 99% 和 99.9%	不适用
语音	语音到文本	RNNT	Libraispeech 开发清理	2513	FP32 的 99%	不适用
language	语言处理	Bert	Sad v1.1	10833	FP32 的 99%	不适用

下表介绍了 Edge 基准测试场景。

区域	任务	场景
愿景	映像分类	单流，脱机，多流
愿景	对象检测（大型）	单流，脱机，多流
愿景	对象检测（小型）	单流，脱机，多流
愿景	医学影像分段	单个流，脱机
语音	语音到文本	单个流，脱机

区域	任务	场景
language	语言处理	单个流，脱机

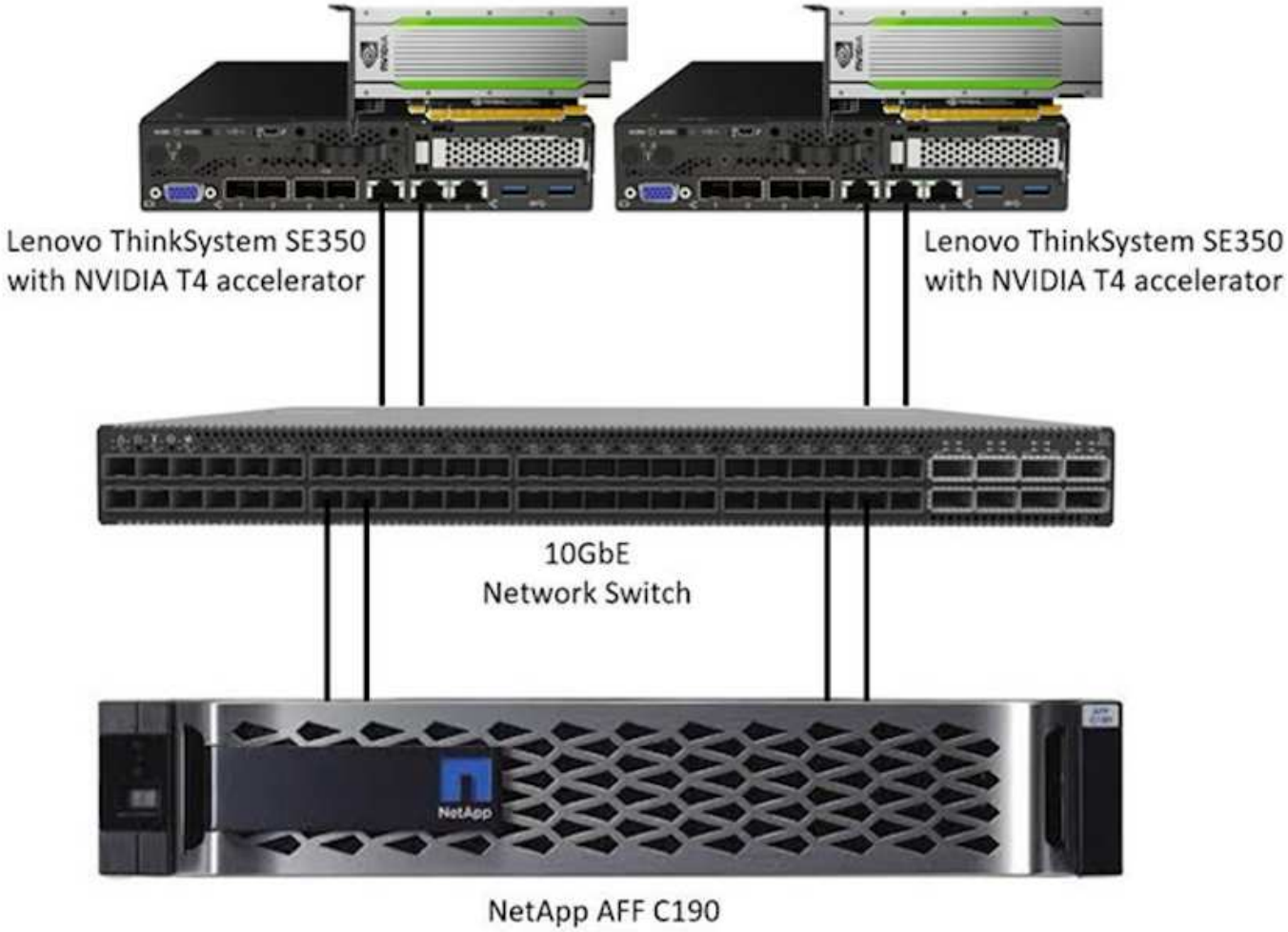
我们使用在此验证中开发的网络存储架构执行了这些基准测试，并将结果与先前提交给 MLPerf 的边缘服务器上本地运行的结果进行了比较。比较结果是，确定共享存储对推理性能有多大影响。

测试配置

下图显示了测试配置。我们使用了 NetApp AFF C190 存储系统和两个联想 ThinkSystem SE350 服务器（每个服务器都有一个 NVIDIA T4 加速器）。这些组件通过 10GbE 网络交换机进行连接。网络存储包含验证 / 测试数据集和经过预先培训的模型。这些服务器可提供计算功能，并可通过 NFS 协议访问存储。

本节介绍测试的配置，网络基础架构， SE350 服务器以及存储配置详细信息。下表列出了解决方案架构的基本组件。

解决方案组件	详细信息
联想 ThinkSystem 服务器	• 2 个 SE350 服务器，每个服务器具有一个 NVIDIA T4 GPU 卡
	• 每个服务器都包含一个 Intel Xeon D-2123IT CPU ，其中四个物理核心运行速度为 2.20GHz ，并具有 128 GB RAM
入门级 NetApp AFF 存储系统（ HA 对）	• NetApp ONTAP 9 软件 • 24 个 960 GB SSD • NFS 协议 • 每个控制器一个接口组，挂载点有四个逻辑 IP 地址



下表列出了存储配置：具有 2RU 的 AFF C190 ， 24 个驱动器插槽。

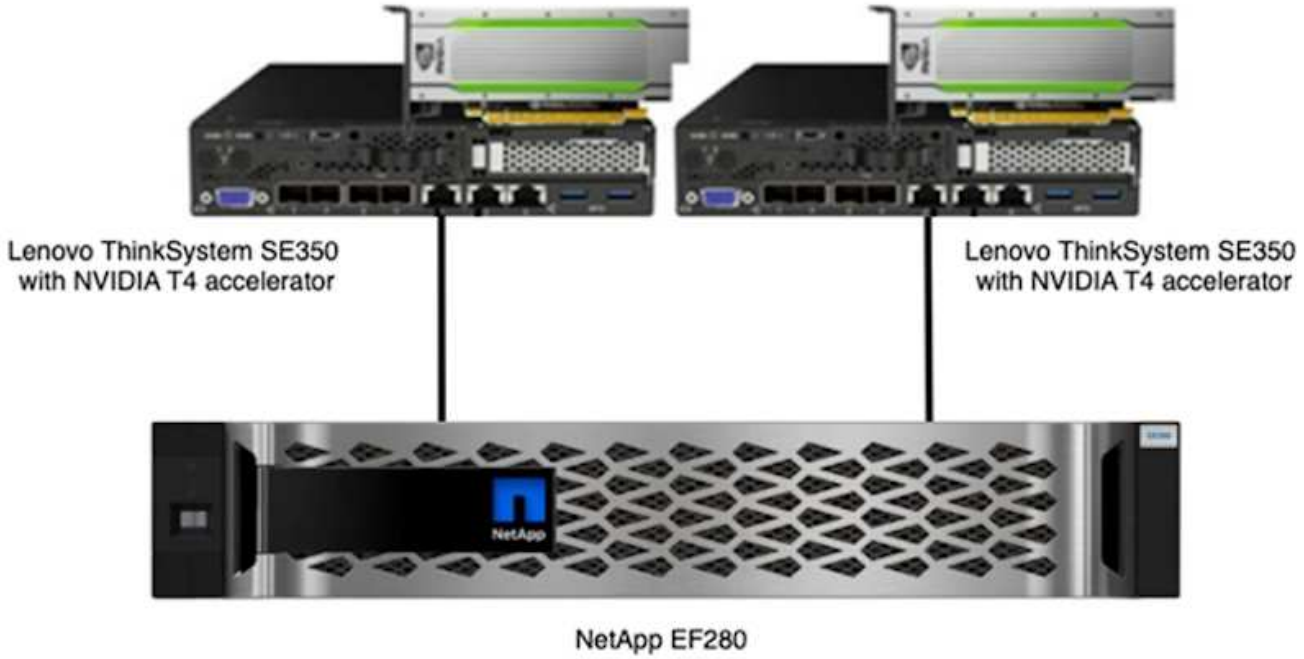
控制器	聚合	**FlexGroup** 卷	聚合大小	卷大小	操作系统挂载点
控制器 1	aggr1.	/netapplenovo_AI_fg	8.42 TiB	15 TB	/NetApp_Lenovo ， fg
控制器 2.	Aggr2.		8.42 TiB		

/netappLenovo AI_fg 文件夹包含用于模型验证的数据集。

下图显示了测试配置。我们使用了 NetApp EF280 存储系统和两个联想 ThinkSystem SE350 服务器（每个服务器都有一个 NVIDIA T4 加速器）。这些组件通过 10GbE 网络交换机进行连接。网络存储包含验证 / 测试数据集和经过预先培训的模型。这些服务器可提供计算功能，并可通过 NFS 协议访问存储。

下表列出了 EF280 的存储配置。

控制器	卷组	**Volume**	卷大小	**DDP** 大小	连接方法
控制器 1	DDP1	卷 1	8.42 TiB	16 TB	SE350-1 到 iSCSI LUN 0
控制器 2.		第 2 卷	8.42 TiB		SE350-2 到 iSCSI LUN 1



测试操作步骤

本节介绍用于验证此解决方案的测试过程。

操作系统和 AI 推理设置

对于 AFF C190 ，我们使用支持 NVIDIA GPU 的 Ubuntu 18.04 和 Docker ，并使用 MLPerf "code" 在联想提交到 MLPerf 推理 v0.7 的过程中提供。

对于 EF280 ，我们将 Ubuntu 20.04 与 NVIDIA 驱动程序结合使用，并将 Docker 与 NVIDIA GPU 和 MLPerf 结合使用 "code" 在联想提交到 MLPerf 推理 v1.1 的过程中提供。

要设置 AI 推理，请执行以下步骤：

1. 下载需要注册的数据集， ImageNet 2012 验证集， Criteo TB 数据集和 Brats 2019 训练集，然后解压缩这些文件。
2. 创建一个至少包含 1 TB 的工作目录，并定义环境变量 `MLPERF_scrating_path` 引用该目录。

   在使用本地数据进行测试时，您应在共享存储上共享此目录以供网络存储使用情形使用，或者在本地磁盘上共享此目录。

3. 运行 make `prebuild` 命令，该命令可为所需的推理任务构建并启动 Docker 容器。

   ⓘ 以下命令全部从正在运行的 Docker 容器中执行：

   - 为 MLPerf 推理任务下载经过预先培训的 AI 型号： `make download_model`
   - 下载可免费下载的其他数据集： `make download_data`

- 预处理数据： make `preprocess_data`
- 运行： `m`构建 。
- 针对计算服务器中的 GPU 优化的构建推理引擎： `make generate_engine`
- 要运行推理工作负载，请运行以下命令（一个命令）：

```
make run_harness RUN_ARGS="--benchmarks=<BENCHMARKS>
--scenarios=<SCENARIOS>"
```

AI 推理运行

执行了三种类型的运行：

- 使用本地存储的单服务器 AI 推理
- 使用网络存储的单服务器 AI 推理
- 使用网络存储的多服务器 AI 推理

测试结果

我们会运行大量测试来评估建议的架构的性能。

有六种不同的工作负载（图像分类，对象检测（小），对象检测（大），医学影像，语音到文本， 和自然语言处理（ NLP ），您可以在三种不同的情形下运行：脱机，单流和多流。

> 最后一种情形仅适用于映像分类和对象检测。

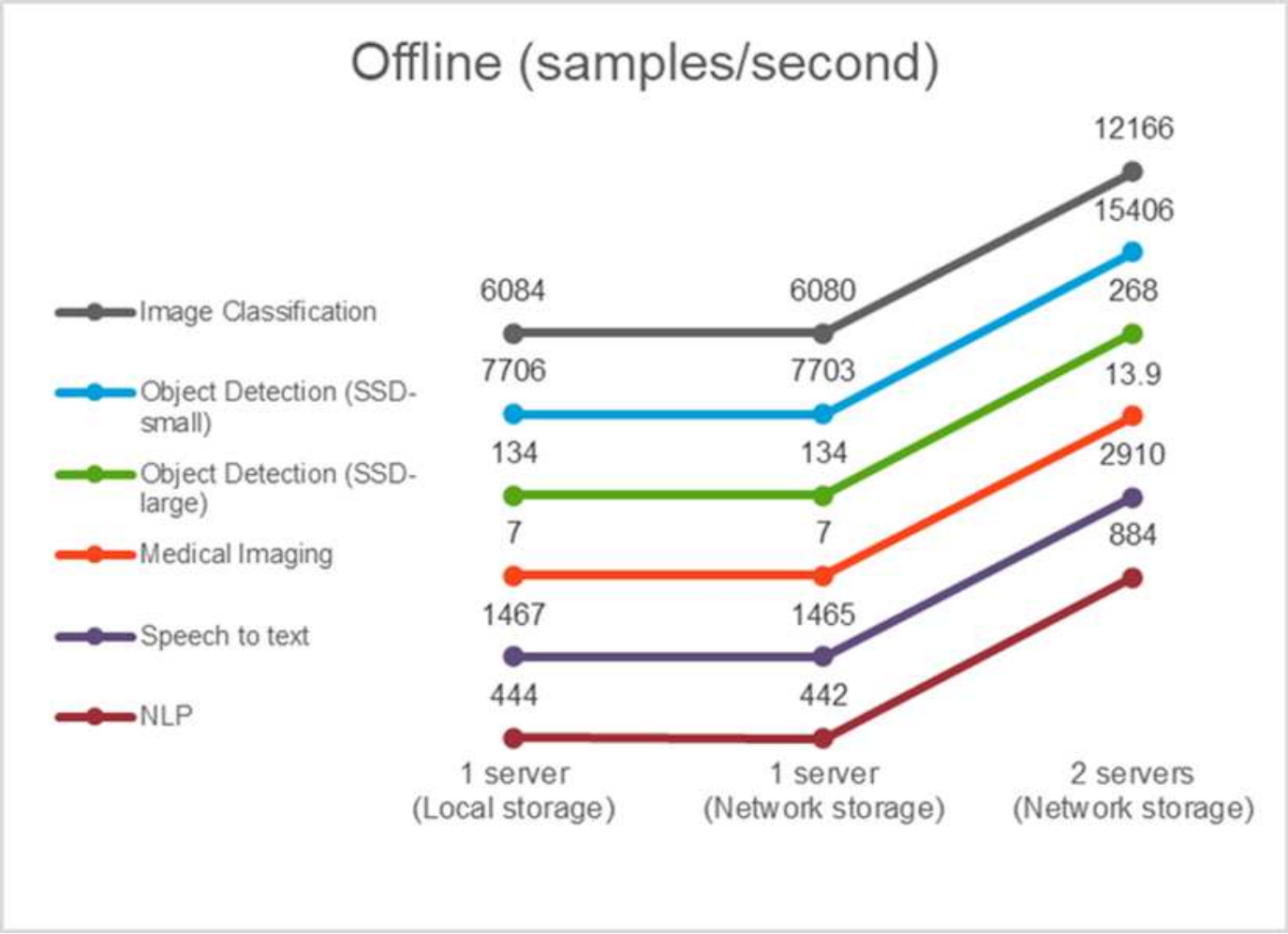
这样可以提供 15 个可能的工作负载，这些工作负载都在三种不同的设置下进行了测试：

- 单个服务器 / 本地存储
- 单个服务器 / 网络存储
- 多服务器 / 网络存储

以下各节将介绍这些结果。

AFF 脱机情形中的 AI 推理

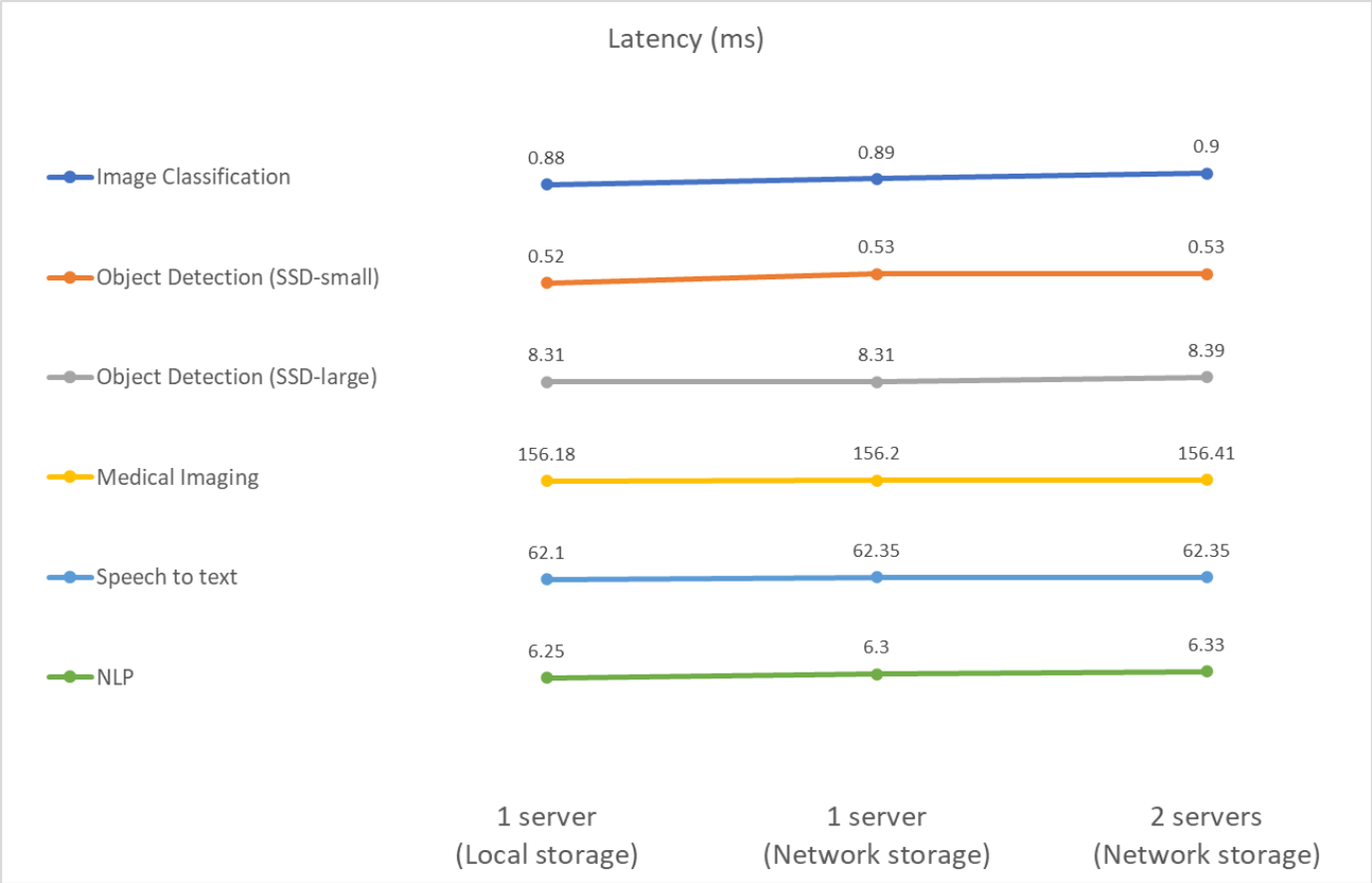
在这种情况下，服务器可以使用所有数据，并测量了处理所有样本所需的时间。我们会将带宽报告为每秒样本数作为测试结果。如果使用了多个计算服务器，则会报告所有服务器的总带宽总和。下图显示了所有这三种使用情形的结果。对于双服务器情形，我们会报告两个服务器的总带宽。



结果显示，网络存储不会对性能产生负面影响，更改极少，对于某些任务，未找到任何结果。添加第二台服务器时，总带宽恰好是两倍，或者最差情况下，更改率小于 1% 。

在 AFF 的单个流方案中进行 AI 推理

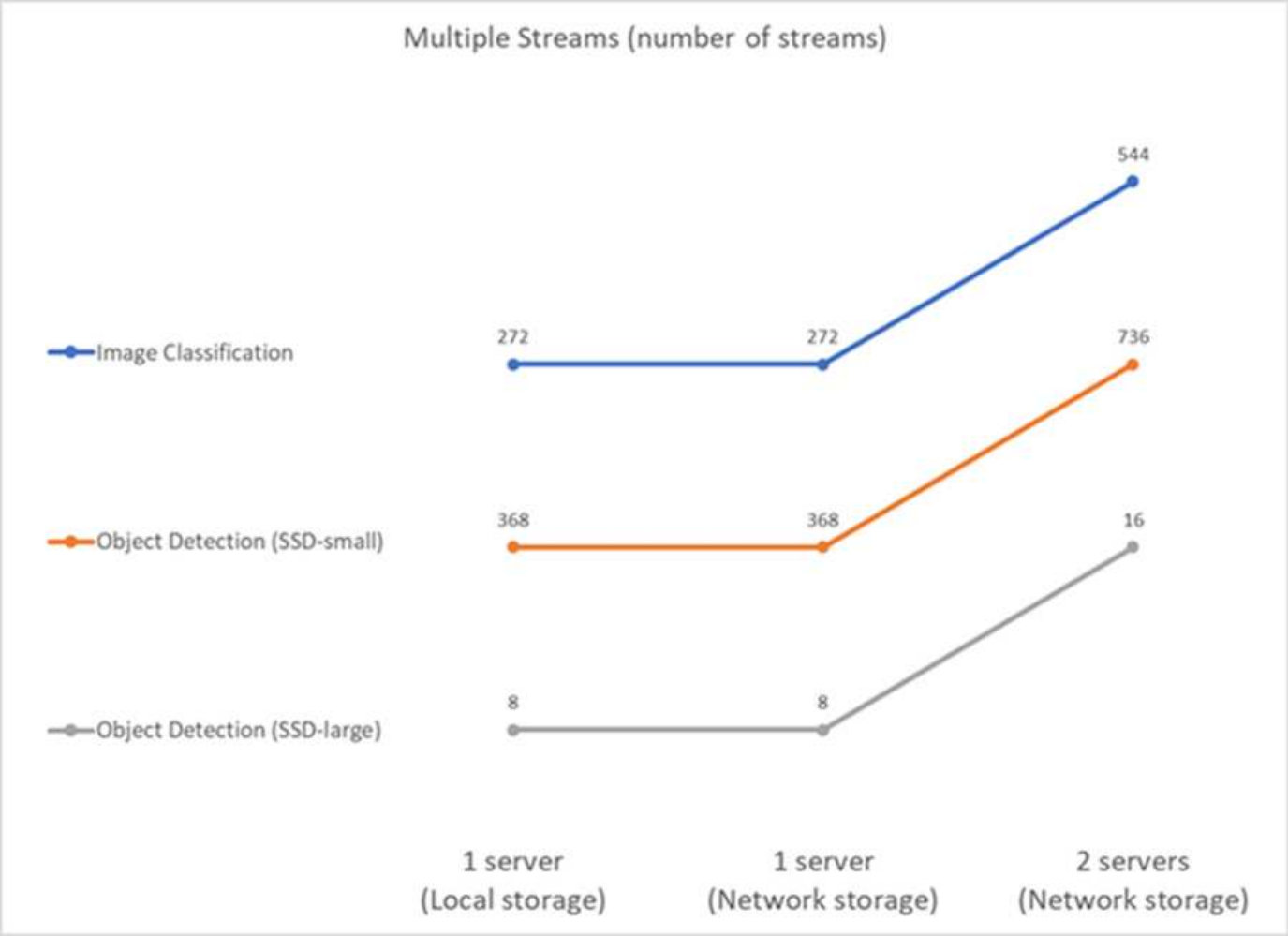
此基准测试可测量延迟。对于多个计算服务器案例，我们会报告平均延迟。下图显示了这组任务的结果。对于双服务器案例，我们会报告两个服务器的平均延迟。



结果再次表明，网络存储足以处理这些任务。在一台服务器的情况下，本地存储与网络存储之间的差别很小或没有差别。同样，当两个服务器使用相同的存储时，两个服务器上的延迟保持不变或变化量非常小。

在 AFF 的多流方案中进行 AI 推理

在这种情况下，结果是系统在满足 QoS 限制的情况下可以处理的流数量。因此，结果始终为整数。对于多个服务器，我们会报告所有服务器上的流总数。并非所有工作负载都支持此方案，但我们已执行了这些工作负载。下图总结了我们的测试结果。对于双服务器案例，我们会报告两个服务器的流总数。



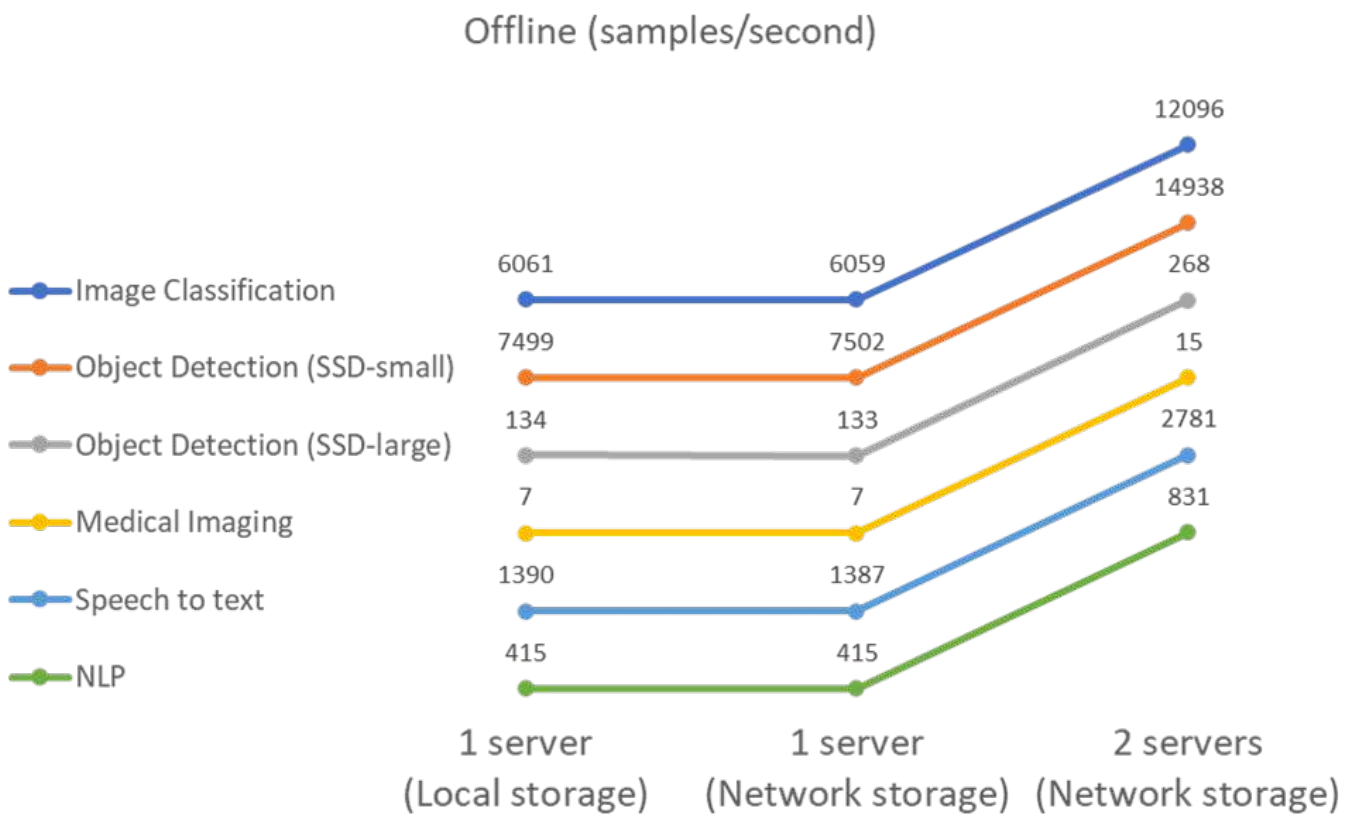
结果显示了设置的完美性能—本地存储和网络存储的结果相同，添加第二个服务器会使建议设置可以处理的流数量增加一倍。

EF 的测试结果

我们会运行大量测试来评估建议的架构的性能。有六种不同的工作负载（图像分类，对象检测（小），对象检测（大），医学影像，语音到文本， 和自然语言处理（ NLP ））， 这两种情况下运行：脱机和单流。以下各节将介绍这些结果。

EF 脱机情形中的 AI 推理

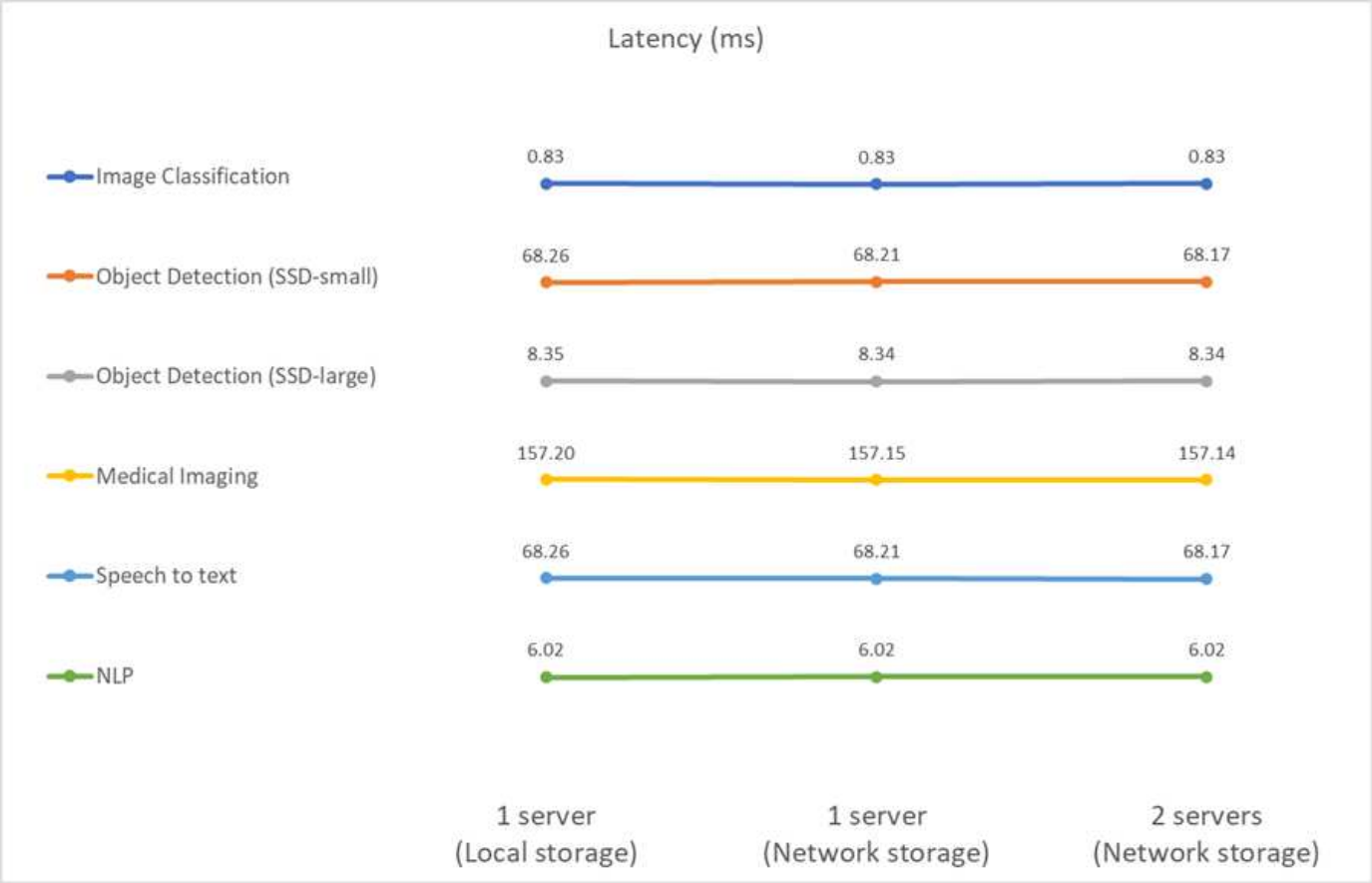
在这种情况下，服务器可以使用所有数据，并测量了处理所有样本所需的时间。我们会将带宽报告为每秒样本数作为测试结果。对于单节点运行，我们会报告两个服务器的平均值，而对于两个服务器运行，我们会报告所有服务器的总带宽总和。下图显示了使用情形的结果。



结果显示，网络存储不会对性能产生负面影响，更改极少，对于某些任务，未找到任何结果。添加第二台服务器时，总带宽恰好是两倍，或者最差情况下，更改率小于 1% 。

在一个流场景中对 **EF** 进行 **AI** 推理

此基准测试可测量延迟。对于所有情况，我们都会报告运行中涉及的所有服务器的平均延迟。系统将提供此任务套件的结果。



结果再次显示，网络存储足以处理这些任务。在一台服务器的情况下，本地存储与网络存储之间的差别很小或没有差别。同样，当两个服务器使用相同的存储时，两个服务器上的延迟保持不变或变化量非常小。

架构规模估算选项

您可以调整用于验证的设置，以适合其他使用情形。

计算服务器

我们使用的是 Intel Xeon D-2123IT CPU ，这是 SE350 支持的最低 CPU 级别，具有四个物理核心和 60 瓦 TDP 。虽然服务器不支持更换 CPU ，但可以订购功能更强大的 CPU 。支持的最大 CPU 是 Intel Xeon D-2183IT ，16 核， 100 瓦，运行频率为 2.20 GHz 。这样可以显著提高 CPU 计算能力。虽然 CPU 本身并不是运行推理工作负载的瓶颈，但它有助于处理数据以及执行与推理相关的其他任务。目前， NVIDIA T4 是唯一可用于边缘使用情形的 GPU ；因此，目前无法升级或降级 GPU 。

共享存储

在测试和验证中，我们在本文档中使用了 NetApp AFF C190 系统，该系统的最大存储容量为 50.5 TB ，连续读取的吞吐量为 4.4 GBps ，小型随机读取的 IOPS 为 230k ，经证实非常适合边缘推理工作负载。

但是、如果您需要更多存储容量或更快的网络速度、则应使用NetApp AFF A220或NetApp AFF A250存储系统。此外， NetApp EF280 系统的最大容量为 1.5 PB ，带宽为 10 Gbps ，也用于此解决方案验证。如果您希望使用更多的存储容量和更高的带宽、 "NetApp EF300" 可以使用。

结论

AI 驱动的自动化和边缘计算是帮助业务组织实现数字化转型并最大限度地提高运营效率和安全性的领先方法。借助边缘计算，数据处理速度更快，因为它不必往返于数据中心。因此，将数据来回发送到数据中心或云的相关成本会降低。如果企业必须使用部署在边缘的人工智能推理模型近乎实时地做出决策，则降低延迟和提高速度将非常有用。

NetApp 存储系统可提供与本地 SSD 存储相同或更好的性能，并为数据科学家，数据工程师， AI/ML 开发人员以及业务或 IT 决策者带来以下优势：

- 在 AI 系统，分析和其他关键业务系统之间轻松共享数据。这种数据共享可减少基础架构开销，提高性能并简化整个企业的数据管理。
- 可独立扩展的计算和存储，最大限度地降低成本并提高资源利用率。
- 利用集成的 Snapshot 副本和克隆简化开发和部署工作流，实现瞬时且节省空间的用户工作空间，集成版本控制和自动化部署。
- 企业级数据保护，可实现灾难恢复和业务连续性。本文档中介绍的 NetApp 和联想解决方案是一种灵活的横向扩展架构，非常适合在边缘进行企业级 AI 推理部署。

致谢

- J.Falkanger ，高级联想 HPC & AI 解决方案经理
- NetApp 技术营销工程师 Dave Arnette
- NetApp E 系列 AI 解决方案技术主管 Joey Parnell
- NetApp QA 工程师 Cody Harryman

从何处查找追加信息

要了解有关本文档中所述信息的更多信息，请参见以下文档和 / 或网站：

- NetApp AFF A 系列阵列产品页面

  "https://www.netapp.com/data-storage/aff-a-series/"

- NetApp ONTAP 数据管理软件— ONTAP 9 信息库

  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286

- TR-4727 ：《 NetApp EF 系列简介》

  https://www.netapp.com/pdf.html?item=/media/17179-tr4727pdf.pdf

- NetApp E 系列 SANtricity 软件产品规格

  https://www.netapp.com/pdf.html?item=/media/19775-ds-3171-66862.pdf

- 适用于容器的 NetApp 持久存储— NetApp Trident

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- MLPerf
  - "https://mlcommons.org/en/"
  - "http://www.image-net.org/"
  - "https://mlcommons.org/en/news/mlperf-inference-v11/"
- NetApp BlueXP复制和同步

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"
- TensorFlow 基准测试

  "https://github.com/tensorflow/benchmarks"
- 联想 ThinkSystem SE350 边缘服务器

  "https://lenovopress.com/lp1168"
- 联想 ThinkSystem DM5100F 统一闪存存储阵列

  "https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"

版权信息

版权所有 © 2024 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。 美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

商标信息

NetApp、NetApp 标识和 http://www.netapp.com/TM 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。