AI在Edge大會上推斷- NetApp採用Lenovo ThinkSystem -解決方案設計

NetApp Solutions

NetApp
April 12, 2024

This PDF was generated from https://docs.netapp.com/zh-tw/netapp-solutions/ai/ai-edge-introduction.html on April 12, 2024. Always check docs.netapp.com for the latest.

目錄

AI在Edge大會上推斷- NetApp採用Lenovo ThinkSystem -解決方案設計 . . . 1
TR-4886：Edge的AI推斷-採用Lenovo ThinkSystem的NetApp解決方案設計 . . . 1
技術總覽 . . . 5
測試計畫 . . . 11
測試組態 . . . 12
測試程序 . . . 14
測試結果 . . . 15
架構規模調整選項 . . . 20
結論 . . . 21

AI在Edge大會上推斷- NetApp採用Lenovo ThinkSystem -解決方案設計

TR-4886：Edge的AI推斷-採用Lenovo ThinkSystem的NetApp解決方案設計

Sathish Thyagarajan、NetApp Miroslav Hodak、聯想

摘要

許多新興的應用程式案例（例如進階驅動程式輔助系統（ADAS）、Industry 4.0、智慧城市和物聯網（IoT））、都需要在接近零延遲的情況下處理持續的資料流。本文件說明運算與儲存架構、可在符合這些需求的邊緣環境中、於NetApp儲存控制器和Lenovo ThinkSystem伺服器上部署GPU型人工智慧（AI）提示。本文件也提供業界標準MLPerf推斷基準測試的效能資料、以評估配備NVIDIA T4 GPU之邊際伺服器上的各種推斷工作。我們會調查離線、單一串流和多重串流推斷案例的效能、並顯示採用具成本效益的共享網路儲存系統的架構具有高效能、可為多個邊緣伺服器提供資料和模型管理的集中點。

簡介

公司越來越多地在網路邊緣產生大量資料。為了從智慧型感應器和IoT資料中獲得最大價值、企業組織正在尋求能夠實現邊緣運算的即時事件串流解決方案。因此、運算需求嚴苛的工作會越來越多地在資料中心以外的邊緣執行。AI推斷是這種趨勢的驅動因素之一。邊緣伺服器可為這些工作負載提供足夠的運算能力、尤其是在使用加速器時、但有限的儲存設備通常是個問題、尤其是在多伺服器環境中。在本文件中、我們將說明如何在邊緣環境中部署共享儲存系統、以及它如何在不影響效能的情況下、為AI推斷工作負載帶來好處。

本文件說明邊緣AI推斷的參考架構。它結合多部Lenovo ThinkSystem邊緣伺服器與NetApp儲存系統、打造易於部署和管理的解決方案。本指南旨在作為在各種情況下實際部署的基礎指南、例如裝有多部攝影機和工業感測器的工廠場地、零售交易中的銷售點（POS）系統、或是能識別自主車輛視覺異常狀況的全自駕（FSD）系統。

本文件涵蓋由Lenovo ThinkSystem SE350 Edge Server和入門級NetApp AFF 支援中心與EF系列儲存系統所組成之運算與儲存組態的測試與驗證。參考架構為AI部署提供高效率且具成本效益的解決方案、同時也提供全方位的資料服務、整合式資料保護、無縫擴充性、以及採用NetApp ONTAP 功能與NetApp SANtricity 功能的NetApp資料管理軟體所提供的雲端連線資料儲存設備。

目標對象

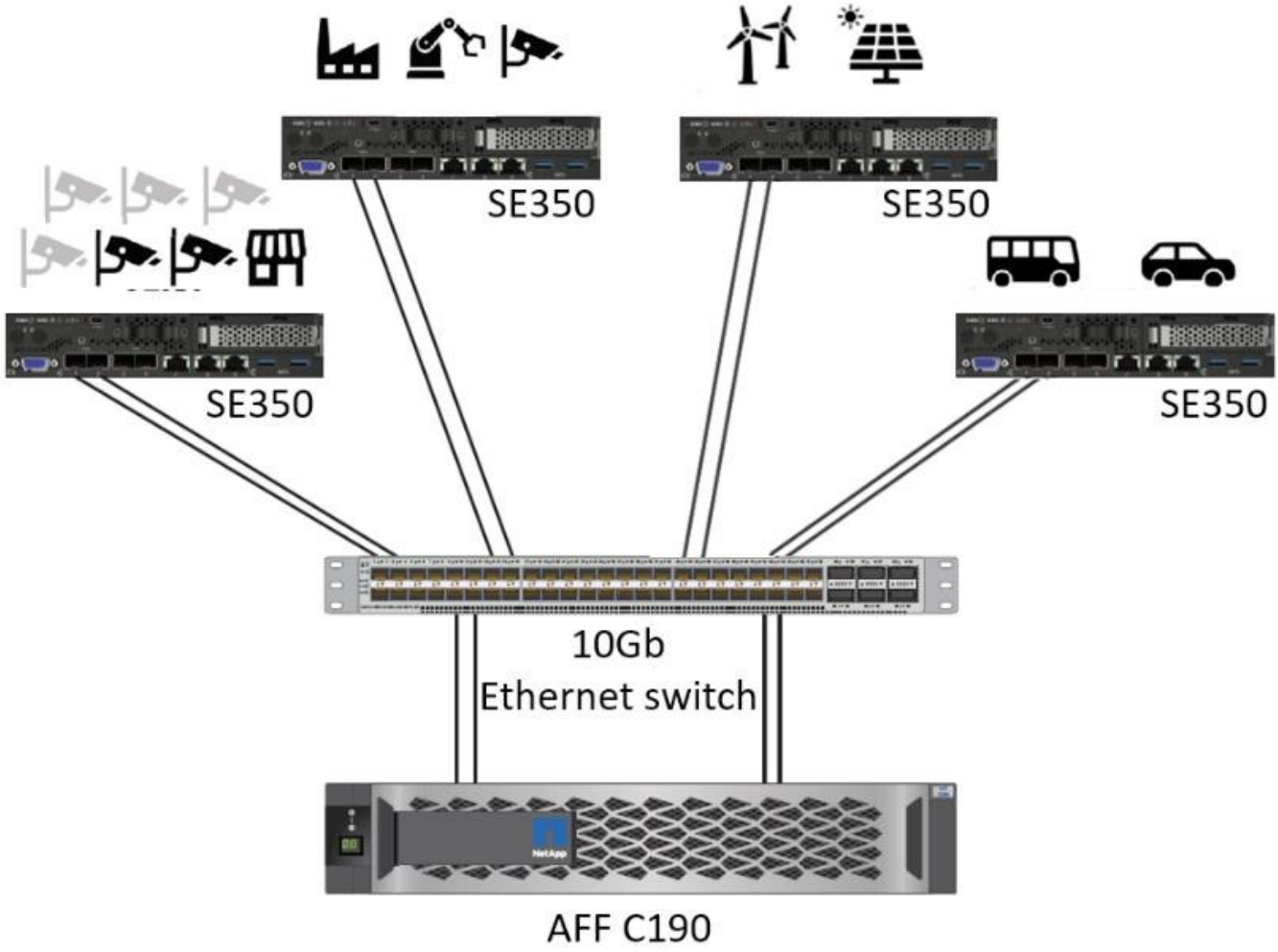
本文適用於下列對象：

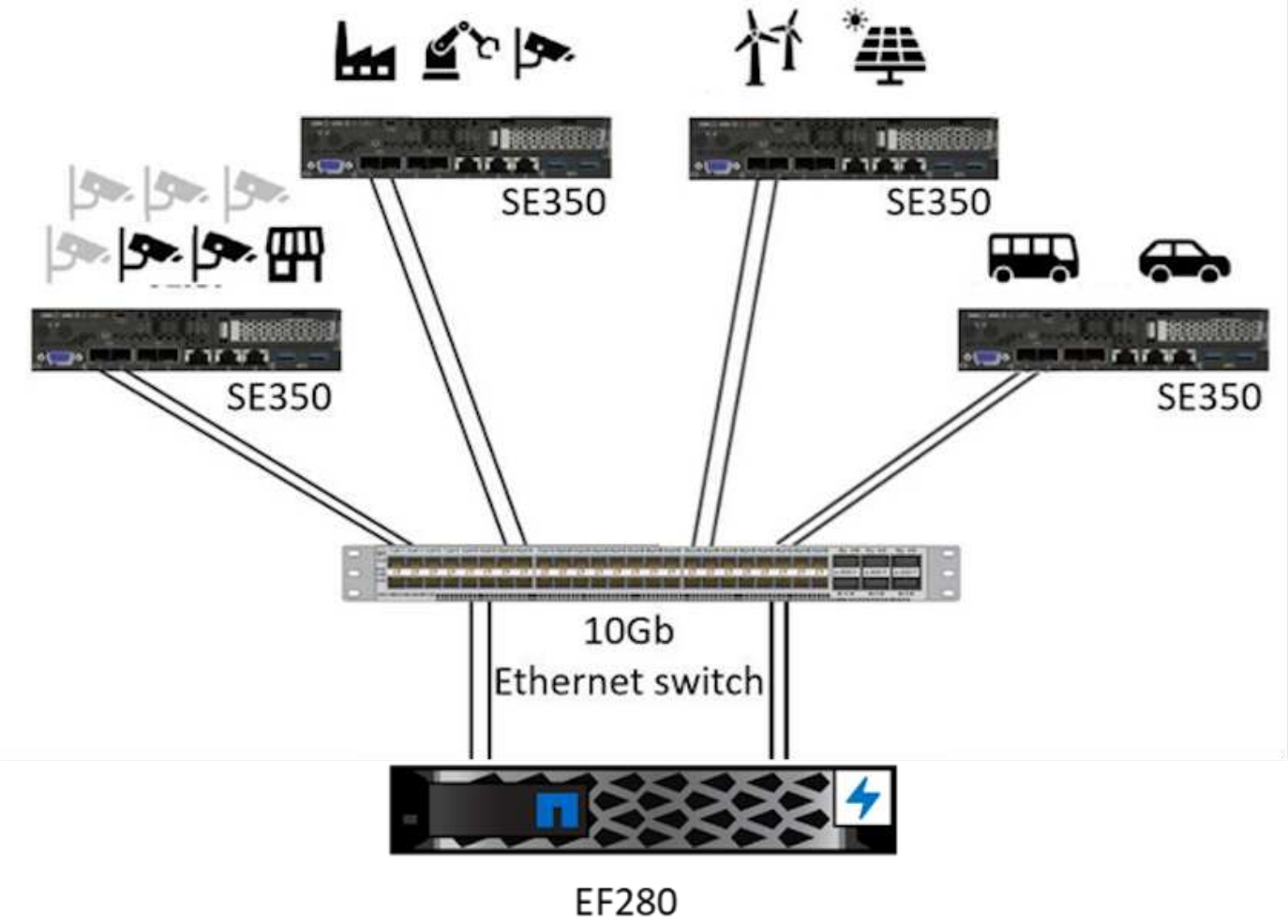
- 想要將邊緣AI產品化的企業領導者和企業架構設計師。
- 資料科學家、資料工程師、AI /機器學習（ML）研究人員、以及AI系統開發人員。
- 設計解決方案以開發AI/ML模型和應用程式的企業架構設計師。
- 資料科學家和AI工程師正在尋求有效方法來部署深度學習（DL）和ML模型。
- 負責部署及管理Edge推斷模型的Edge裝置管理員和Edge Server管理員。

解決方案架構

這款Lenovo ThinkSystem伺服ONTAP 器和NetApp的可靠性或NetApp SANtricity 的不二之選儲存解決方案、是

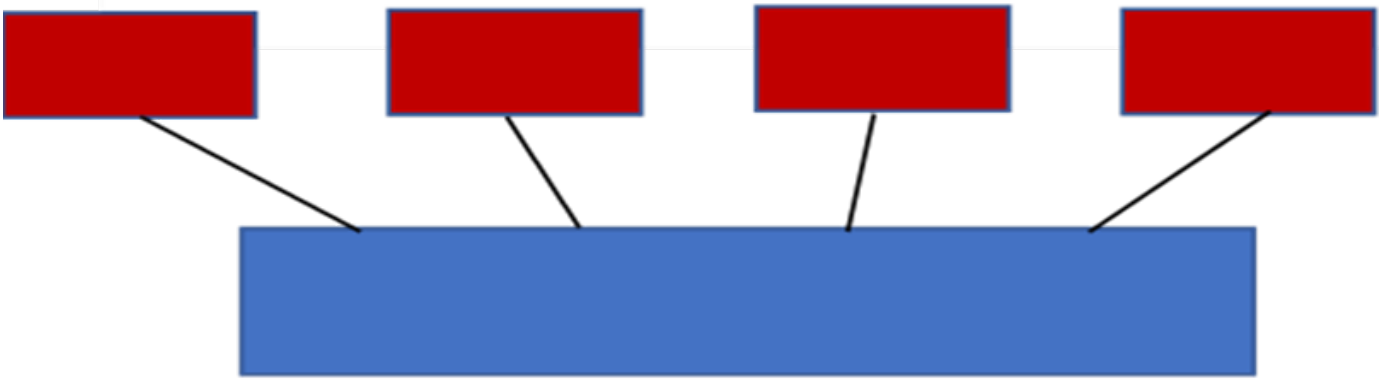
專為處理大型資料集的AI推斷、使用GPU的處理能力、以及傳統的CPU。這項驗證可展現高效能與最佳資料管理、其架構使用單一或多部Lenovo SR350邊緣伺服器、並與單AFF 一NetApp效益儲存系統互連、如下圖所示。





下圖中的邏輯架構總覽顯示此架構中運算和儲存元素的角色。具體而言、它顯示下列內容：

- 邊緣運算裝置會對從攝影機、感應器等接收的資料執行推斷。
- 一種共享儲存元件、可滿足多種用途：
  - 為執行推斷所需的推斷模型及其他資料提供集中位置。運算伺服器可直接存取儲存設備、並在整個網路上使用推斷模型、而不需要在本機複製。
  - 更新的機型將在此推播。
  - 歸檔邊緣伺服器接收的輸入資料、以供日後分析。例如、如果邊緣裝置連接至攝影機、儲存元件會保留攝影機拍攝的影片。



紅色	藍色
聯想運算系統	NetApp AFF 產品儲存系統
邊緣裝置會對來自攝影機、感應器等的輸入執行推斷。	共享儲存設備、保留來自邊緣裝置的推斷模型和資料、以供日後分析。

此NetApp與Lenovo解決方案提供下列主要效益：

- GPU加速邊緣運算。
- 部署多部邊緣伺服器、從共享儲存設備進行備份與管理。
- 強大的資料保護功能、可滿足低恢復點目標（RPO）和恢復時間目標（RTO）、而且不會遺失資料。
- 利用NetApp Snapshot複本與複製技術來最佳化資料管理、以簡化開發工作流程。

如何使用此架構

本文件驗證所建議架構的設計與效能。不過、我們尚未測試某些軟體層級的元件、例如容器、工作負載、或模型管理、以及與內部部署的雲端或資料中心的資料同步、因為它們是針對部署案例而設計。這裡有多種選擇。

在容器管理層級、Kubernetes Container管理是一個不錯的選擇、無論是完整上游版本（Canonical）或是適合企業部署的修改版本（Red Hat）、都能獲得良好支援。。 "NetApp AI Control Plane" 使用NetApp Trident和新增的 "NetApp DataOps工具套件" 提供內建的可追蹤性、資料管理功能、介面和工具、讓資料科學家和資料工程師能夠與NetApp儲存設備整合。Kubernetes的ML工具套件Kubefflow提供額外的AI功能、並支援TensorFlow服務或NVIDIA Triton Inference伺服器等多種平台上的模型版本設定和KFServing。另一個選項是NVIDIA EGX平台、可提供工作負載管理功能、並可存取支援GPU的AI推斷容器目錄。不過、這些選項可能需要投入大量心力和專業知識才能投入生產、而且可能需要協力廠商獨立軟體廠商（ISV）或顧問的協助。

解決方案領域

AI推斷和邊緣運算的主要優點在於、裝置能夠以高品質、無延遲的方式來運算、處理及分析資料。本文件中有太多邊緣運算使用案例可說明、但以下是幾個主要案例：

汽車：自主車輛

典型的邊緣運算圖示為自動車輛（AV）的進階駕駛輔助系統（ADAS）。無磁碟汽車的AI必須迅速處理來自攝影機和感應器的大量資料、才能成為成功的安全駕駛者。在物件與人之間解讀太久可能意味著生命或死亡、因此盡可能將資料處理在車輛附近是非常重要的。在此情況下、一或多個邊緣運算伺服器會處理來自攝影機、雷達、LIDAR及其他感應器的輸入、而共享儲存設備則會保留推斷模型、並儲存來自感應器的輸入資料。

醫療：病患監控

AI和邊緣運算最大的影響之一、就是能夠強化對在家中照護和深切治療單位（ICU）長期病患的持續監控能力。來自監測胰島素水平、呼吸、神經活動、心臟心律和腸胃功能的邊際裝置的資料、需要即時分析資料、但必須立即採取行動、因為行動的時間有限、無法拯救某人的生命。

零售：無收銀員付款

邊緣運算可以推動AI和ML、協助零售商縮短結帳時間、並增加腳步流量。無收銀機系統支援各種元件、例如：

- 驗證與存取：將實體購物者連線至已驗證的帳戶、並允許存取零售空間。
- 庫存監控：使用感測器、RFID標籤和電腦視覺系統、協助確認購物者選擇或取消選擇商品。

在這裡、每個邊緣伺服器都會處理每個結帳櫃位、而共享儲存系統則是中央同步點。

金融服務：資訊站的人類安全與防範詐騙

銀行組織正使用AI和邊緣運算技術來創新及創造個人化的銀行體驗。互動式資訊站使用即時資料分析和AI推斷功能、現在可讓ATM不僅協助客戶提領資金、還能透過從攝影機擷取的影像主動監控資訊站、以識別人類安全或詐騙行為的風險。在此案例中、邊緣運算伺服器和共享儲存系統會連線至互動式多媒體資訊站和攝影機、以AI推斷模式協助銀行收集和處理資料。

製造業：產業**4.0**

第四次產業革命（產業4.0）已經開始、也伴隨著智慧工廠和3D列印等新興趨勢。為了準備迎接資料導向的未來、我們整合了大型機器對機器（M2M）通訊與IoT、無需人為介入、即可提升自動化程度。製造業已高度自動化、加上AI功能、自然是長期趨勢的延續。AI可實現自動化作業、並藉由電腦願景和其他AI功能來實現自動化。您可以自動化品質控制、或是仰賴人的願景或決策來執行更快速分析工廠內組裝線路上的材料、以協助製造廠符合所需的ISO安全與品質管理標準。在此處、每部運算邊緣伺服器都會連接到監控製造流程的感測器陣列、並視需要將更新的推斷模型推送至共享儲存設備。

電信：除鏽偵測、塔式檢查及網路最佳化

電信產業使用電腦願景和AI技術來處理影像、這些影像會自動偵測出各種生生除、並識別含有侵蝕的電池塔、因此需要進一步檢查。近年來、利用無人機影像和AI模型來識別塔內的不同區域、以分析鐵鏽、表面碎裂和侵蝕。AI技術的需求持續成長、可有效檢查電信基礎架構和電池塔、定期評估是否有老化問題、並在需要時立即修復。

此外、電信領域的另一項新興使用案例是使用AI和ML演算法來預測資料流量模式、偵測具備5G功能的裝置、以及自動化及強化多重輸入和多重輸出（MIMU）能源管理。在無線塔上使用的是多重可擴充網路容量的MIMO-硬體、但也會增加能源成本。部署於行動站台的「MIMO睡眠模式」有多種ML模式、可預測無線電的有效使用、並有助於降低行動網路營運者（MNO）的能源消耗成本。AI推斷與邊緣運算解決方案可協助MNO減少資料中心來回傳輸的資料量、降低TCO、最佳化網路作業、並改善終端使用者的整體效能。

技術總覽

本節說明此AI解決方案的技術基礎。

NetApp AFF 產品

最先進的NetApp AFF ®不中斷儲存系統可讓AI在邊緣部署、以領先業界的效能、優異的靈活度、雲端整合及同級最佳的資料管理功能、滿足企業儲存需求。專為Flash所設計的NetApp AFF 支援功能、有助於加速、管理及保護業務關鍵資料。

- 入門級NetApp AFF 支援儲存系統是以FAS2750硬體和SSD Flash媒體為基礎
- HA組態中有兩個控制器

NetApp入門級AFF 的C190儲存系統支援下列功能：

- 磁碟機數量上限為24 x 960GB SSD
- 兩種可能的組態：
  - 乙太網路（10GbE）：4個10GBASE-T（RJ-45）連接埠
  - 統一化（16GB FC或10GbE）：4個統一化目標介面卡2（UTA2）連接埠
- 最大有效容量為50.5 TB

> 對於NAS工作負載、單一入門級AFF 的432C190系統可支援4.4GBps的連續讀取處理量、以及230K IOPS、以1毫秒或更短的延遲時間進行小型隨機讀取。

NetApp AFF 解決方案-A220

NetApp也提供其他入門級儲存系統、為大規模部署提供更高的效能與擴充性。對於NAS工作負載、單一入門級AFF 的AS20220系統支援：

- 連續讀取的處理量為6.2GBps
- 375K IOPS可在延遲1毫秒或更短時間內進行小型隨機讀取
- 最大磁碟機數為144x 960GB、3.8TB或7.6TB SSD
- 可用容量可擴充至大於1PB的部分AFF

NetApp AFF 產品-A250

- 最大有效容量為35PB、最大橫向擴充為2-24個節點（12個HA配對）
- 效能比AFF 不支援的A220高出≥45%
- 440k IOPS隨機讀取@1ms
- 以最新的NetApp ONTAP 產品資訊版本ONTAP 為基礎：《NetApp產品資訊

- 利用兩個25GB乙太網路進行HA和叢集互連

NetApp E系列EF系統

EF系列是入門級與中階All Flash SAN儲存陣列系列產品、可加速資料存取、並利用NetApp SANtricity 的Ef2軟體、更快從其中獲取價值。這些系統同時提供SAS和NVMe Flash儲存設備、並提供經濟實惠的極致IOPS、100微秒內的回應時間、以及高達44GBps的頻寬、是混合式工作負載和嚴苛應用程式（例如AI推斷和高效能運算（HPC）的理想選擇。

下圖顯示NetApp EF280儲存系統。

NetApp EF280

- 32GB/16GB FC、25GB/10Gb iSCSI及12Gb SAS支援
- 最大有效容量為96個磁碟機、總計1.5PB
- 10Gbps的處理量（循序讀取）
- 300K IOPs（隨機讀取）
- NetApp EF280是NetApp產品組合中成本最低的All Flash Array（AFF）

NetApp EF300

- 24個NVMe SSD磁碟機、總容量為367TB
- 擴充選項總計240x NL-SAS HDD、96x SAS SSD或組合
- 100GB NVMe / IB、NVMe / RoCE、iSER / IB和SRP/IB
- 32GB NVMe/FC、FCP
- 25GB iSCSI
- 20Gbps（循序讀取）
- 670K IOPs（隨機讀取）

> 如需詳細資訊、請參閱 "NetApp EF系列NetApp EF系列All Flash Array EF600、F300、EF570及EF280資料表"。

NetApp ONTAP 產品9.

NetApp最新一代的儲存管理軟體《支援》（NetApp）版本、可讓企業將基礎架構現代化、並移轉至雲端就緒的資料中心。ONTAP利用領先業界的資料管理功能ONTAP 、無論資料位於何處、只要使用一組工具、即可管理及保護資料。您也可以自由地將資料移至任何需要的位置：邊緣、核心或雲端。NetApp 9.8.1包含多項功能、可簡化資料管理、加速及保護關鍵資料、並在混合雲架構中提供新一代基礎架構功能。ONTAP

簡化資料管理

資料管理對企業IT營運至關重要、因此應用程式和資料集必須使用適當的資源。包含下列功能、可簡化及簡化作業、並降低總營運成本：ONTAP

- *即時資料壓縮與擴充重複資料刪除技術。*資料壓縮技術可減少儲存區塊內的空間浪費、重複資料刪除技術則可大幅提升有效容量。這適用於本機儲存的資料、以及分層至雲端的資料。
- *服務品質（AQO）的最低、最大及可調適性。*精細的服務品質（QoS）控制、有助於維持高共享環境中關鍵應用程式的效能等級。
- * NetApp FabricPool 功能可將冷資料自動分層至公有雲和私有雲儲存選項、包括Amazon Web Services（AWS）、Azure和NetApp StorageGRID 等儲存解決方案。如需FabricPool 更多有關資訊、請參閱 "TR-4598"。

加速並保護資料

供應優異的效能與資料保護、並以下列方式擴充這些功能：ONTAP

- *效能與較低的延遲。ONTAP *
- 資料保護 ONTAP 效能*功能提供內建的資料保護功能、並可在所有平台上進行通用管理。
- * NetApp Volume Encryption（NVE）* ONTAP 。*支援內建和外部金鑰管理、提供原生Volume層級的加密功能。
- 多租戶與多因素驗證。 ONTAP 支援以最高安全層級共享基礎架構資源。

符合未來需求的基礎架構

下列功能可協助滿足嚴苛且不斷變化的業務需求：ONTAP

- 無縫擴充與不中斷營運。 ONTAP 支援在不中斷營運的情況下、為現有控制器和橫向擴充叢集增加容量。客戶可以升級至最新技術、例如NVMe和32GB FC、而不需進行昂貴的資料移轉或中斷運作。
- * Cloud connection。ONTAP *效能不只是雲端連線能力最強的儲存管理軟體、ONTAP Select 還可在Cloud Volumes Service 所有公有雲中選擇軟體定義儲存（英文）和雲端原生執行個體（NetApp版本）。
- 與新興應用程式整合 ONTAP 。*此功能使用支援現有企業應用程式的相同基礎架構、為新一代平台和應用程式（例如自動駕駛車輛、智慧城市和產業4.0）提供企業級資料服務。

NetApp SANtricity

NetApp SANtricity 旨在為E系列混合式Flash與EF系列All Flash陣列提供領先業界的效能、可靠性與簡易性。針對繁重的工作負載應用、包括資料分析、視訊監控、備份與還原、實現E系列混合式Flash與EF系列All Flash陣列的最大效能與使用率。利用此功能、您可以在儲存設備保持連線的同時、完成組態調整、維護、容量擴充及其他工作。SANtricity此外、透過易於使用的內建System Manager介面、還能提供優異的資料保護、主動監控及認證安全功能。SANtricity若要深入瞭解、請參閱 "NetApp E系列SANtricity 《軟體資料表》"。

效能最佳化

效能最佳化SANtricity 的功能性支援功能可為您的所有資料分析、視訊監控及備份應用程式、提供高IOPs、高處理量及低延遲的資料。加速高IOPS、低延遲應用程式及高頻寬、高處理量應用程式的效能。

最大化正常運作時間

在儲存設備保持連線的情況下、完成所有管理工作。在不中斷I/O的情況下調整組態、執行維護或擴充容量透過自動化功能、線上組態、最先進的動態磁碟資源池（DPP）技術等功能、實現同級最佳的可靠性。

高枕無憂

透過易於使用的內建System Manager介面、即可提供優異的資料保護、主動監控及認證安全性。SANtricity簡化儲存管理工作。獲得所有E系列儲存系統進階調校所需的靈活度。隨時隨地管理NetApp E系列系統。我們隨裝即用的網路介面可簡化您的管理工作流程。

NetApp Trident

"Trident" NetApp是適用於Docker和Kubernetes的開放原始碼動態儲存協調工具、可簡化持續儲存的建立、管理和使用。Kubernetes原生應用程式Trident直接在Kubernetes叢集內執行。Trident可讓客戶將DL Container映像無縫部署到NetApp儲存設備、並為AI Container部署提供企業級體驗。Kubernetes使用者（例如ML開發人員和資料科學家）可以建立、管理及自動化協調與複製、以充分利用NetApp技術所提供的NetApp進階資料管理功能。

NetApp BlueXP 複製與同步

"BlueXP 複製與同步" 是一項NetApp服務、可快速且安全地同步資料。無論您需要在內部部署 NFS 或 SMB 檔案共用之間傳輸檔案、 NetApp StorageGRID 、 NetApp ONTAP S3 、 NetApp Cloud Volumes Service 、 Azure NetApp Files 、 Amazon Simple Storage Service （ Amazon S3 ）、 Amazon Elastic File System （ Amazon EFS ）、 Azure Blob 、 Google Cloud Storage 、 或者 IBM Cloud Object Storage 、 BlueXP 複製與同步功能可快速安全地將檔案移至所需的位置。資料傳輸完成後、即可在來源和目標上完全使用。BlueXP 複製與同步功能會根據預先定義的排程、持續同步資料、只移動資料量、因此資料複寫所花的時間和金錢將降至最低。BlueXP 複製與同步是一種軟體即服務（ SaaS ）工具、設定與使用極為簡單。BlueXP 複製與同步所觸發的資料傳輸是由資料代理人執行。您可以在 AWS 、 Azure 、 Google Cloud Platform 或內部部署中部署 BlueXP 複製和同步資料代理人。

Lenovo ThinkSystem伺服器

聯想ThinkSystem伺服器採用創新的硬體、軟體和服務、可解決客戶目前的挑戰、並提供革命性的符合用途的模組化設計方法、以因應未來的挑戰。這些伺服器運用同級最佳的業界標準技術、搭配與眾不同的Lenovo創新技術、為x86伺服器提供最大的靈活度。

部署Lenovo ThinkSystem伺服器的主要優勢包括：

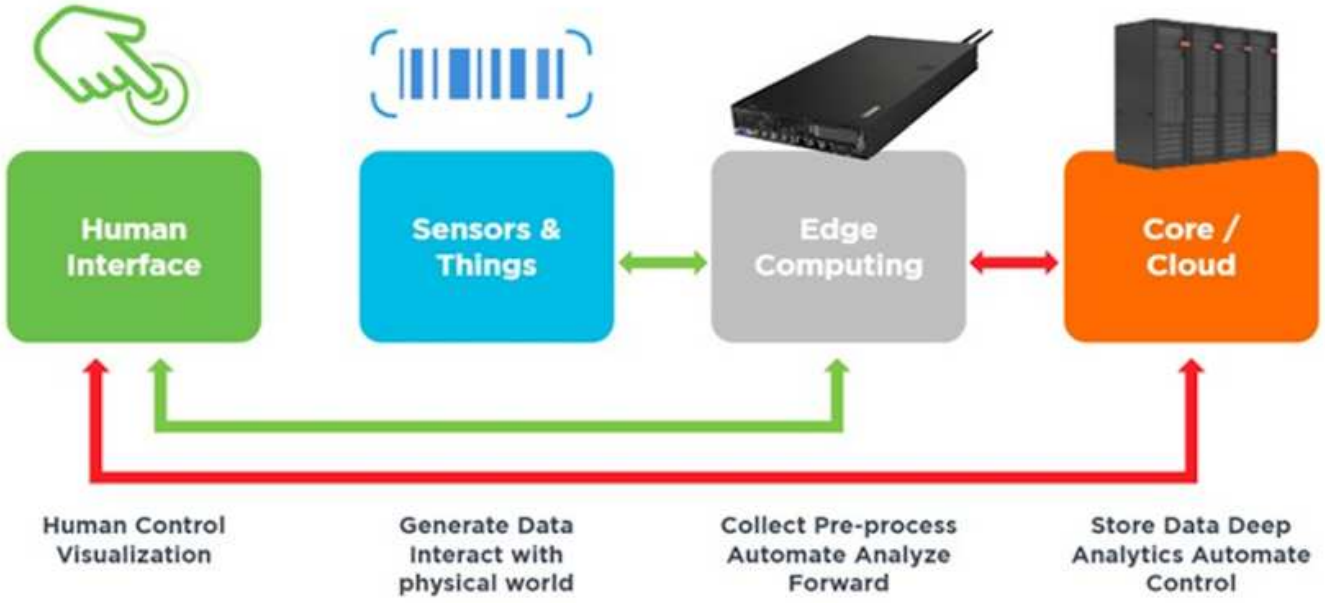
- 可隨企業成長而擴充的模組化設計
- 領先業界的恢復能力、可節省非計畫性停機所需的數小時成本
- 快速Flash技術可降低延遲、縮短回應時間、並即時進行更聰明的資料管理

在AI領域、Lenovo正採取實用的方法、協助企業瞭解ML和AI在工作負載方面的優勢、並將其納入其中。Lenovo客戶可在Lenovo AI創新中心中探索及評估Lenovo AI產品、充分瞭解其特定使用案例的價值。為了縮短實現價值所需的時間、這種以客戶為中心的方法可讓客戶針對已準備好使用並針對AI最佳化的解決方案開發平台、提供概念驗證。

Lenovo ThinkSystem SE350 Edge Server

邊緣運算可讓您在將IoT裝置的資料傳送到資料中心或雲端之前、先在網路邊緣分析資料。下圖所示的Lenovo ThinkSystem SE350、是專為邊緣部署的獨特需求所設計、採用精巧、堅固耐用且環境強化的尺寸規格、著重於靈活度、連線能力、安全性及遠端管理功能。

SE350採用Intel Xeon D處理器、可靈活支援邊緣AI工作負載的加速、專為因應資料中心以外各種環境的伺服器部署挑戰而打造。



MLPerf

MLPerf是領先業界的基準測試套件、可用來評估AI效能。涵蓋多個應用AI領域、包括影像分類、物件偵測、醫療成像和自然語言處理（NLP）。在此驗證中、我們使用推斷v0.7工作負載、這是完成此驗證時最新一次的MLPerf推斷。。 "MLPerf推斷v0.7" 套件包含四項資料中心與邊緣系統的新基準測試：

- 由Transformers（Bert）提供的* Bert.*雙向編碼器表示、可利用Srog資料集微調以回答問題。
- * DLRM-*深度學習建議模式（DLRM）是個人化的推薦模式、經過訓練可最佳化點閱率（CTR）。
- * 3D U-Net.* 3D U-Net架構是針對腦部腫瘤分割（Brat）資料集進行訓練。
- * RNN-T*重複神經網路換能器（RNN-T）是一種自動語音辨識（ASR）模式、訓練內容為LbridSpeech的子集。MLPerf推斷結果與程式碼均可公開取得、並在Apache授權下發佈。MLPerf推斷具有Edge部門、可支援下列案例：
- *單一串流。*此案例模擬回應能力是關鍵因素的系統、例如在智慧型手機上執行離線AI查詢。個別查詢會傳送至系統、並記錄回應時間。結果會報告所有回應的90個百分位數延遲。
- *多重串流*此基準測試適用於處理多個感應器輸入的系統。在測試期間、查詢會以固定時間間隔傳送。強制實施QoS限制（允許的最大延遲）。測試會報告系統在符合QoS限制的情況下可以處理的串流數量。
- *離線。*這是最簡單的案例、涵蓋批次處理應用程式、而指標是每秒取樣的處理量。系統可取得所有資料、基準測試會測量處理所有樣本所需的時間。

Lenovo已針對本文件所使用的伺服器「T4」（SE350）發表MLPerf推斷分數。請參閱以下網址的結果："https://mlperf.org/inference-results-0-7/" 在項目#0.7-145的「Edge、Closed Division」一節中。

測試計畫

本文件遵循MLPerf推斷v0.7 "程式碼"（MLPerf推斷v1.1） "程式碼"和 "規則"。我們執行MLPerf基準測試、設計用於在邊緣進行推斷、如下表所定義。

區域	工作	模型	資料集	**QSL**大小	品質	多重串流延遲限制
願景	映像分類	Resnet50v1.5	ImageNet（224x224）	1024.	99%的FP32	50ms
願景	物件偵測（大型）	SSD - ResNet34	可可（1200 x 1200）	64	99%的FP32	66毫秒
願景	物件偵測（小型）	SSD：移動Netsv1	可可（300x300）	256	99%的FP32	50ms
願景	醫療影像分割	3D UNIT	2019年（全年無休、全年無休、每天24小時160）	16	99%和99.9%的FP32	不適用
演講	語音對文字	RNNT	Lithpech開發-乾淨	2513	99%的FP32	不適用
語言	語言處理	Bert	Squad v1.1	10833	99%的FP32	不適用

下表列出Edge基準測試案例。

區域	工作	案例
願景	映像分類	單一串流、離線、多重串流
願景	物件偵測（大型）	單一串流、離線、多重串流
願景	物件偵測（小型）	單一串流、離線、多重串流

區域	工作	案例
願景	醫療影像分割	單一串流、離線
演講	語音對文字	單一串流、離線
語言	語言處理	單一串流、離線

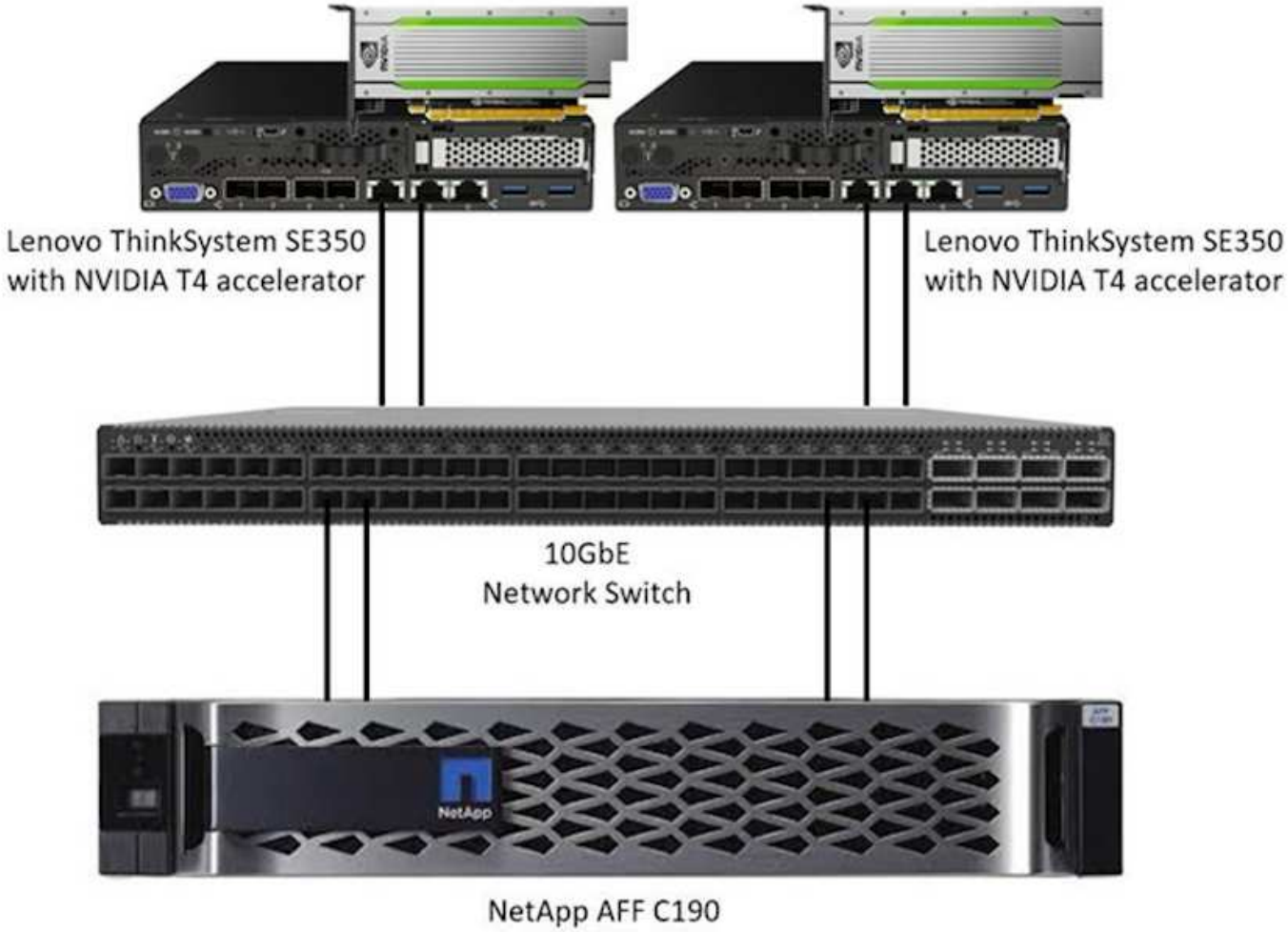
我們使用此驗證所開發的網路儲存架構來執行這些基準測試、並將結果與先前提交給MLPerf之邊緣伺服器上本機執行的結果進行比較。比較的目的是判斷共享儲存設備對推斷效能的影響程度。

測試組態

下圖顯示測試組態。我們使用NetApp AFF 的C190儲存系統和兩部Lenovo ThinkSystem SE350伺服器（每部伺服器都有一個NVIDIA T4加速器）。這些元件是透過10GbE網路交換器來連接。網路儲存設備擁有驗證/測試資料集和預先訓練的模型。伺服器提供運算能力、而且儲存設備可透過NFS傳輸協定存取。

本節說明已測試的組態、網路基礎架構、SE350伺服器、以及儲存資源配置詳細資料。下表列出解決方案架構的基礎元件。

解決方案元件	詳細資料
Lenovo ThinkSystem伺服器	• 2部SE350伺服器各配備1個NVIDIA T4 GPU卡
	• 每部伺服器都包含一個Intel Xeon D-2123IT CPU、其中四個實體核心執行速度為2.20GHz和128GB RAM
入門級NetApp AFF 產品資料儲存系統（HA配對）	• NetApp ONTAP ®9軟體 • 24個960GB SSD • NFS傳輸協定 • 每個控制器一個介面群組、其中四個邏輯IP位址用於掛載點



下表列出儲存組態：AFF 搭載2RU、24個磁碟機插槽的__LW_C190。

控制器	**Aggregate**	流通量**FlexGroup**	**Aggregatesize**	體積化	作業系統掛載點
控制器1.	Aggr1.	/netapplenovo_AI_FG	8.42TiB	15TB	/NetApp_聯想_FG
控制器2.	Aggr2.		8.42TiB		

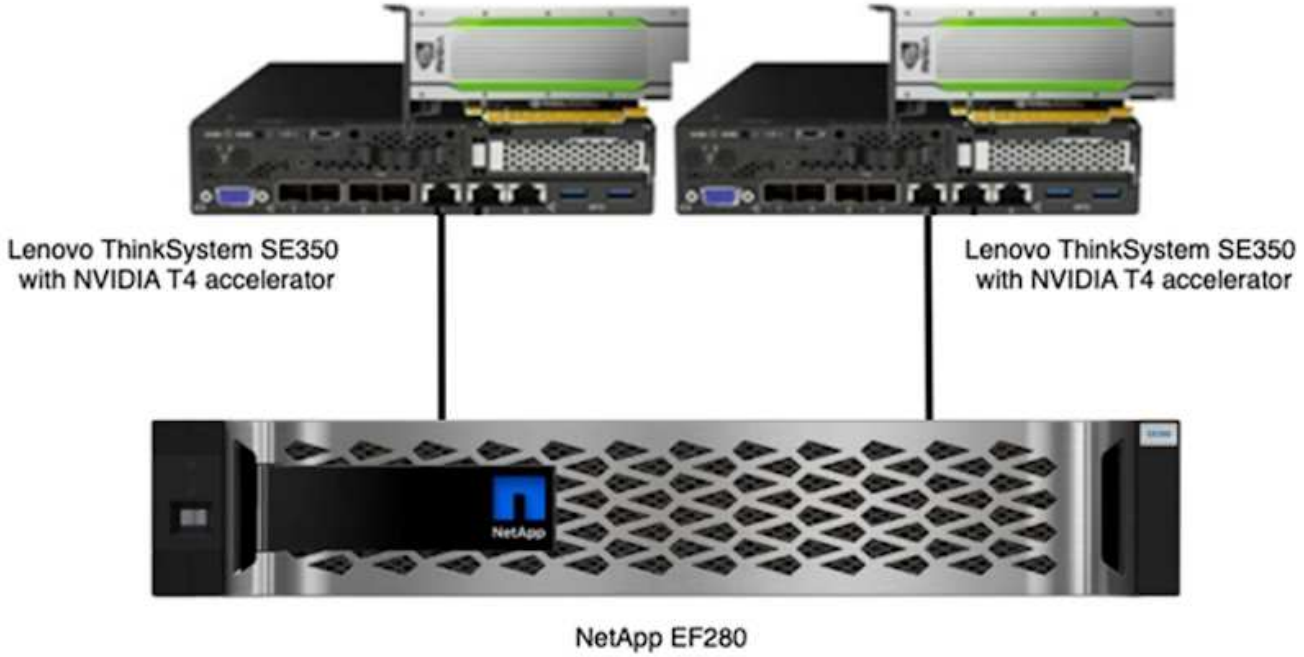
/netapp聯想_AI _fg資料夾包含用於模型驗證的資料集。

下圖顯示測試組態。我們使用NetApp EF280儲存系統和兩個Lenovo ThinkSystem SE350伺服器（每個伺服器都有一個NVIDIA T4加速器）。這些元件是透過10GbE網路交換器來連接。網路儲存設備擁有驗證/測試資料集和預先訓練的模型。伺服器提供運算能力、而且儲存設備可透過NFS傳輸協定存取。

下表列出EF280的儲存組態。

控制器	**Volume**群組	**Volume**	體積化	**DDPsize**	連線方法
控制器1.	DDP1	Volume 1	8.42TiB	16TB	SE350-1至iSCSI LUN 0

控制器	Volume群組	Volume	體積化	DDPsize	連線方法
控制器2.		Volume 2	8.42TiB		SE350-2至iSCSI LUN 1



測試程序

本節說明驗證此解決方案所使用的測試程序。

作業系統與AI推斷設定

在支援NVIDIA GPU的情況下、我們使用Ubuntu 18.04搭配NVIDIA驅動程式和Docker、並使用MLPerf AFF "程式碼" 在Lenovo提交給MLPerf Inference v0.7的過程中提供。

對於EF280、我們使用Ubuntu 20.04搭配NVIDIA驅動程式和Docker、支援NVIDIA GPU和MLPerf "程式碼" 在Lenovo提交給MLPerf Inference v1.1的過程中提供。

若要設定AI推斷、請遵循下列步驟：

1. 下載需要登錄的資料集、ImageNet 2012驗證集、Criteo TB資料集、以及Brat 2019訓練集、然後解壓縮檔案。
2. 建立至少1TB的工作目錄、並定義環境變數「MarlFF_Scrate_path」（指目錄）。

   您應該在共用儲存設備上共用此目錄、以供網路儲存設備使用、或在測試本機資料時共用本機磁碟。

3. 執行make「prebuild」命令、此命令會建立及啟動泊塢視窗容器、以執行必要的推斷工作。

   ⓘ 下列命令都是從執行中的Docker容器中執行：

- 下載預先訓練的AI模型以執行MLPerf推斷工作：「make download_model」
- 下載其他可免費下載的資料集：「make download_data」
- 預先處理資料：製作「preprocure_data」
- RUN：「make build」。
- 在運算伺服器中建置專為GPU最佳化的推斷引擎：「make general_引擎」
- 若要執行推斷工作負載、請執行下列（一個命令）：

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

AI推斷會執行

執行了三種類型的執行：

- 使用本機儲存設備的單一伺服器AI推斷
- 使用網路儲存設備的單一伺服器AI推斷
- 使用網路儲存設備的多伺服器AI推斷

測試結果

我們執行多項測試、以評估所建議架構的效能。

共有六種不同的工作負載（影像分類、物件偵測[小]、物件偵測[大]、醫療影像、語音對文字、 和自然語言處理[NLP]）、您可以在三種不同的情況下執行：離線、單一串流和多重串流。

> 最後一個案例只能用於影像分類和物件偵測。

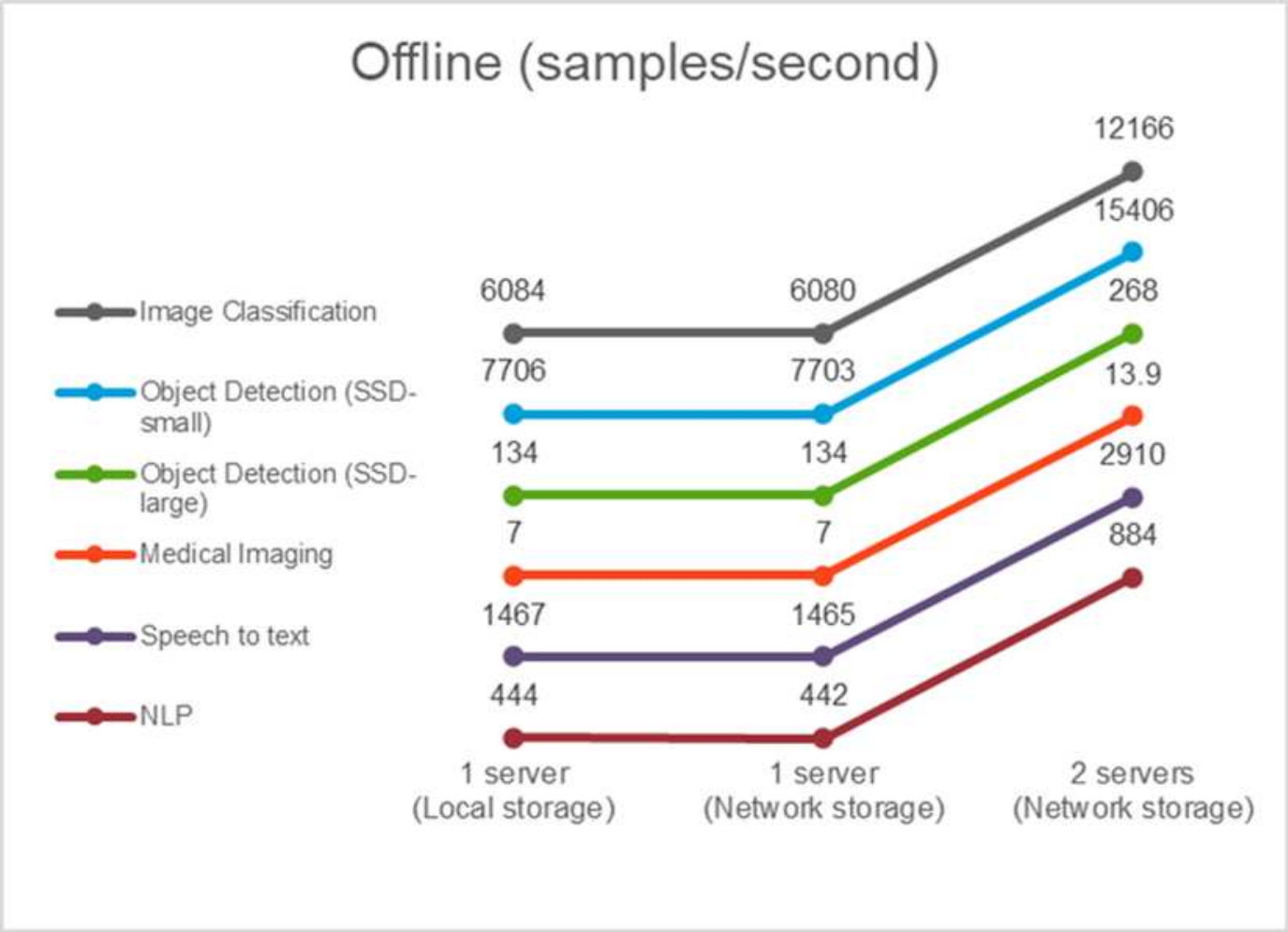
這可提供15種可能的工作負載、這些工作負載都是在三種不同的設定下進行測試：

- 單一伺服器/本機儲存設備
- 單一伺服器/網路儲存設備
- 多伺服器/網路儲存設備

以下各節將說明結果。

AI在離線情境中推斷AFF 以供參考

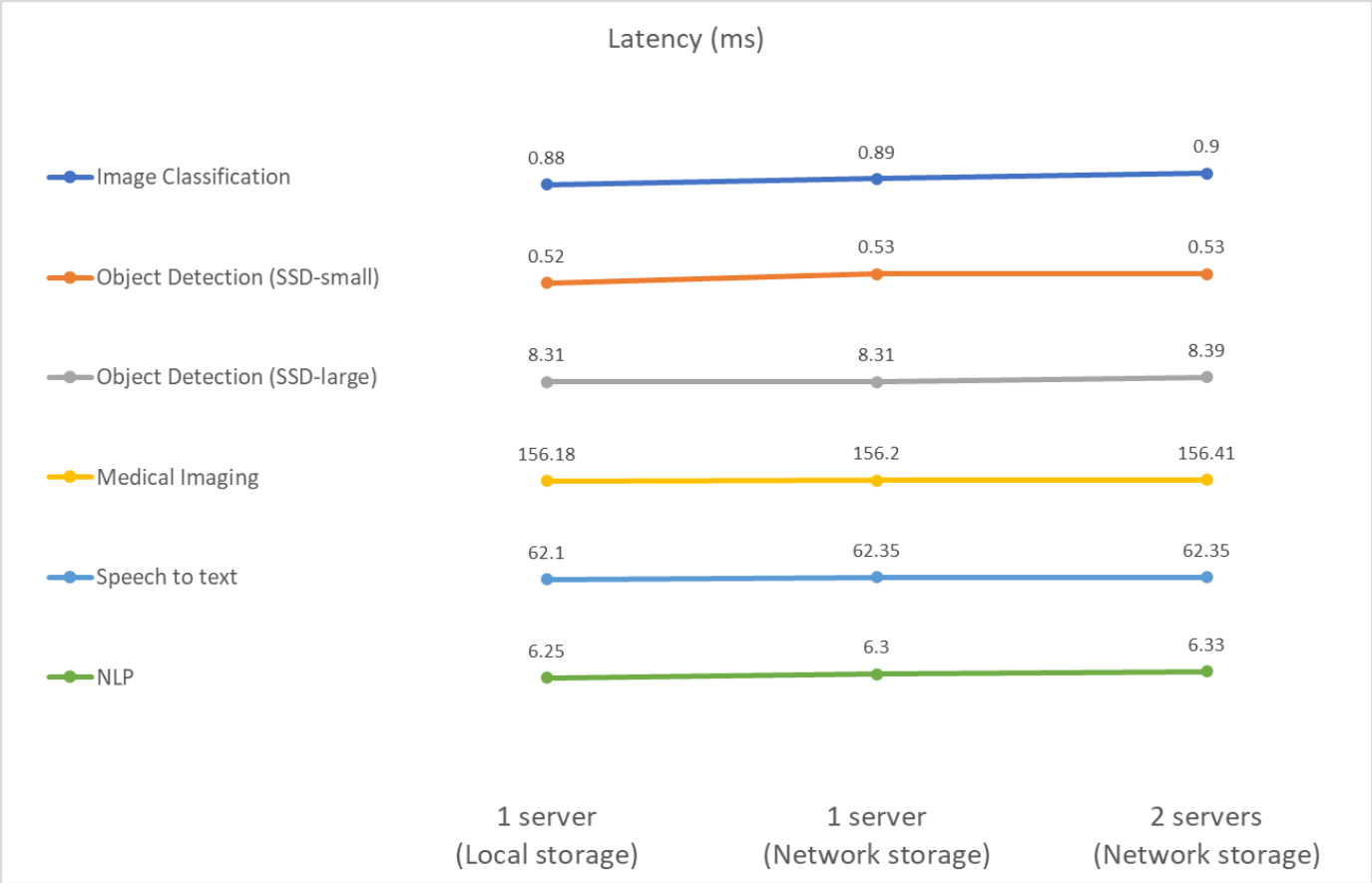
在此案例中、所有資料都可供伺服器使用、並測量處理所有樣本所需的時間。我們會將頻寬報告為每秒取樣數、做為測試結果。當使用多部運算伺服器時、我們會報告所有伺服器的總頻寬。下圖顯示所有三種使用案例的結果。在兩部伺服器的案例中、我們會報告兩部伺服器的總頻寬。



結果顯示、網路儲存設備不會對效能造成負面影響、因為這項變更很小、而且對於某些工作、找不到任何項目。新增第二部伺服器時、總頻寬可能會精確加倍、或是最糟的情況下、變更量會低於1%。

AI在單一串流情境中推斷**AFF** 以供參考

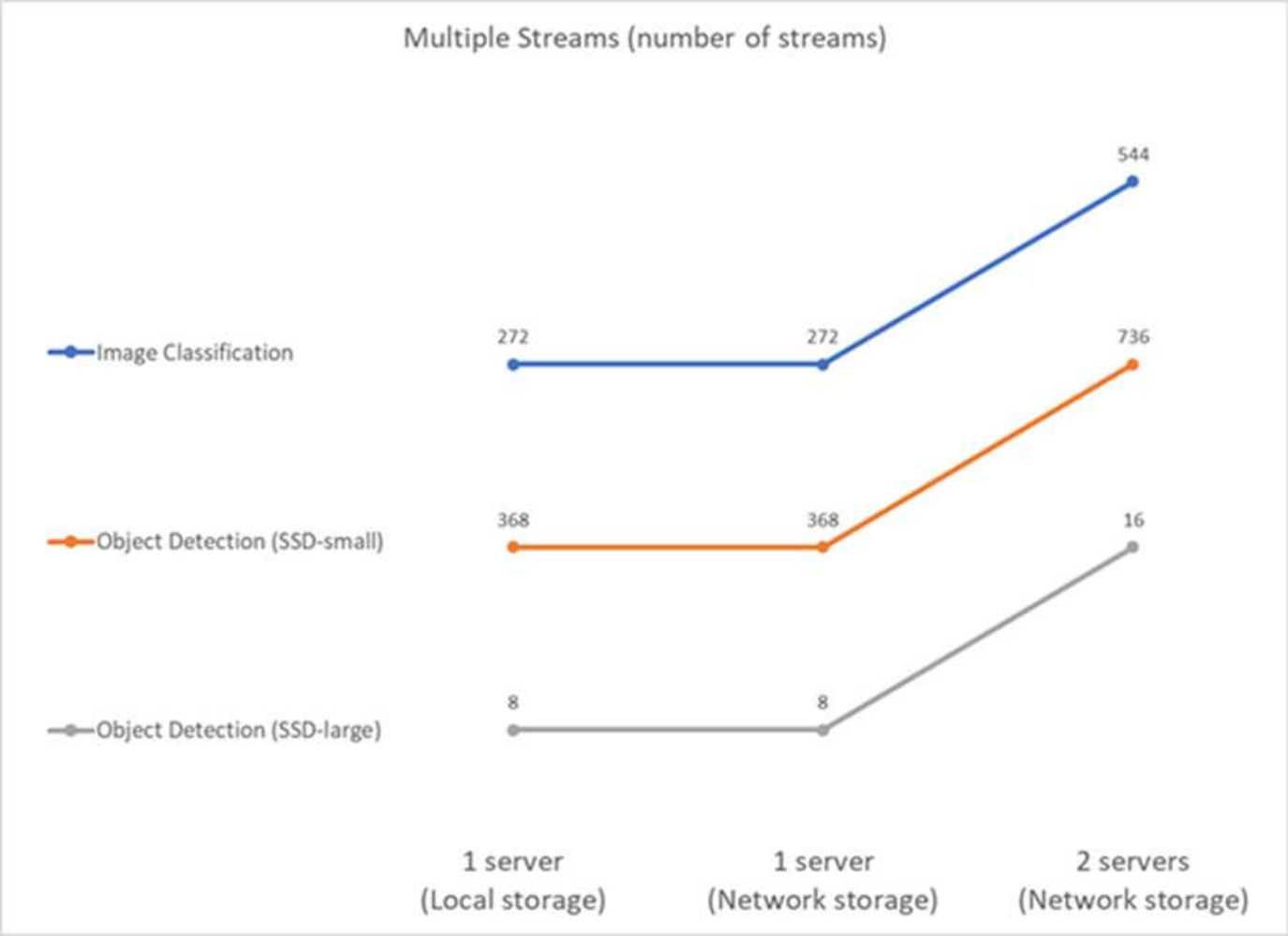
此基準測試可測量延遲。針對多個運算伺服器案例、我們會報告平均延遲。工作套件的結果如下圖所示。在兩部伺服器的案例中、我們會報告兩部伺服器的平均延遲。



結果再次顯示、網路儲存設備足以處理工作。在單一伺服器案例中、本機與網路儲存設備之間的差異最小或無。同樣地、當兩部伺服器使用相同的儲存設備時、兩部伺服器的延遲時間會維持不變、或是變更量會非常小。

AI在多重串流情境中推斷**AFF** 以供參考

在這種情況下、結果是系統可以處理的串流數量、同時滿足QoS限制。因此、結果永遠是整數。對於多部伺服器、我們會報告在所有伺服器上總計串流的總數。並非所有的工作負載都支援這種情況、但我們已經執行了這些工作負載。我們的測試結果摘要如下圖所示。在兩部伺服器的案例中、我們會報告來自兩部伺服器的串流總數。



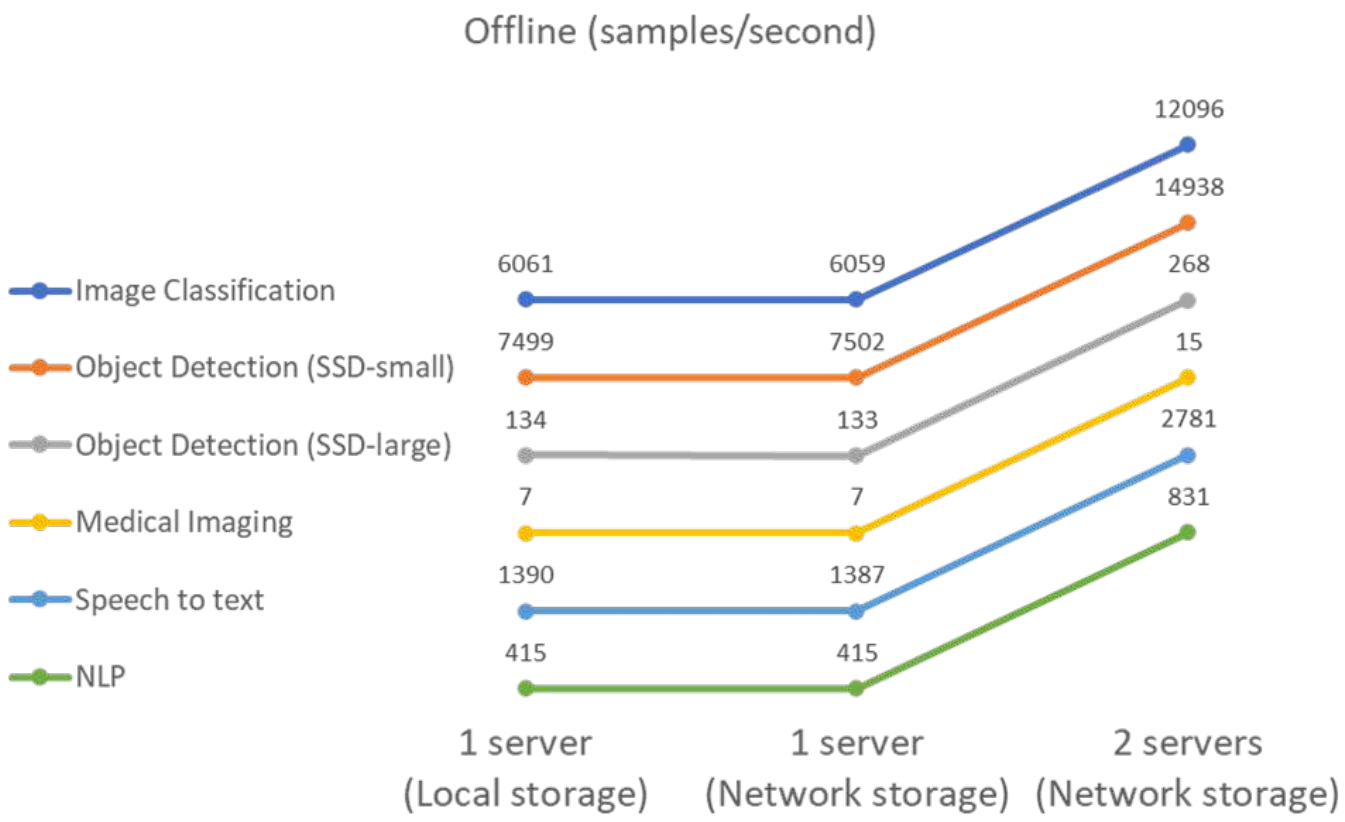
結果顯示設定的完美效能：本機與網路儲存設備的結果相同、而新增第二台伺服器的資料流數量是建議設定所能處理的資料流數量的兩倍。

EF的測試結果

我們執行多項測試、以評估所建議架構的效能。共有六種不同的工作負載（影像分類、物件偵測[小]、物件偵測[大]、醫療影像、語音對文字、 和自然語言處理（NLP]）、這是在兩種不同的情況下執行：離線和單一串流。以下各節將說明結果。

EF離線情境中的AI推斷

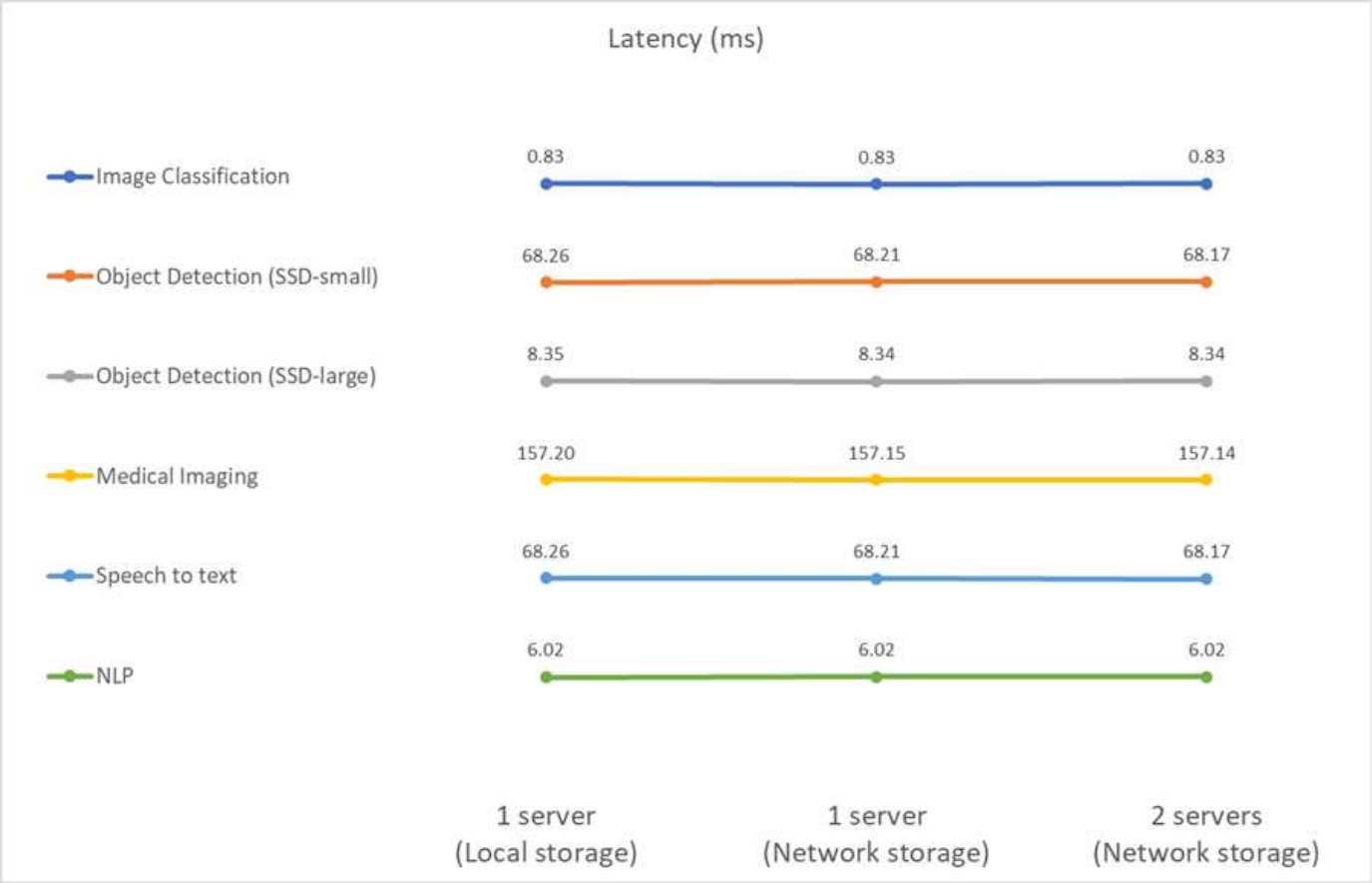
在此案例中、所有資料都可供伺服器使用、並測量處理所有樣本所需的時間。我們會將頻寬報告為每秒取樣數、做為測試結果。對於單一節點執行、我們會報告兩部伺服器的平均執行量、而對於兩部伺服器執行、則會報告所有伺服器的總頻寬總和。使用案例的結果如下圖所示。



結果顯示、網路儲存設備不會對效能造成負面影響、因為這項變更很小、而且對於某些工作、找不到任何項目。新增第二部伺服器時、總頻寬可能會精確加倍、或是最糟的情況下、變更量會低於1%。

EF單一串流案例中的**AI**推斷

此基準測試可測量延遲。在所有情況下、我們都會報告執行中所有伺服器的平均延遲時間。會提供工作套件的結果。



結果再次顯示、網路儲存設備足以處理工作。在單一伺服器的情況下、本機與網路儲存設備之間的差異最小或無。同樣地、當兩部伺服器使用相同的儲存設備時、兩部伺服器的延遲時間會維持不變、或是變更量會非常小。

架構規模調整選項

您可以調整用於驗證的設定、以符合其他使用案例。

運算伺服器

我們使用Intel Xeon D-2123IT CPU、這是SE350支援的最低CPU層級、有四個實體核心和60W TDP。雖然伺服器不支援更換CPU、但可以訂購更強大的CPU。支援的最高CPU是Intel Xeon D-2183IT、採用16核心、100W執行速度為2.20GHz。如此可大幅提升CPU的運算能力。雖然CPU本身並不是執行推斷工作負載的瓶頸、但它有助於資料處理及其他與推斷相關的工作。目前、NVIDIA T4是唯一適用於邊緣使用案例的GPU、因此目前無法升級或降級GPU。

共享儲存設備

在測試與驗證方面AFF 、NetApp RandomC190系統的儲存容量上限為50.5 TB、循序讀取的處理量為4.4GBps、小型隨機讀取的處理量為230K IOPS、經證實非常適合邊緣推斷工作負載。

不過、如果您需要更多儲存容量或更快的網路速度、您應該使用NetApp AFF 知識庫A220或 "NetApp AFF 產品-A250" 儲存系統：此外、NetApp EF280系統的最大容量為1.5PB、頻寬為10Gbps、也用於本解決方案驗證。如果您偏好以更高頻寬提供更多儲存容量、 "NetApp EF300" 可以使用。

結論

AI導向的自動化與邊緣運算是一項領先業界的方法、可協助企業組織實現數位化轉型、並將作業效率與安全性最大化。利用邊緣運算技術、資料處理速度更快、因為資料中心不需要往返資料中心。因此、將資料來回傳送至資料中心或雲端的相關成本會降低。如果企業必須使用部署在邊緣的AI推斷模型、以近乎即時的方式做出決策、則降低延遲並加快速度將會更為有利。

NetApp儲存系統提供與本機SSD儲存設備相同或更好的效能、並為資料科學家、資料工程師、AI/ML開發人員、企業或IT決策者提供下列效益：

- 輕鬆在AI系統、分析和其他關鍵業務系統之間共享資料。這種資料共享可降低基礎架構的負荷、改善效能、並簡化整個企業的資料管理。
- 獨立擴充的運算與儲存設備、可將成本降至最低、並改善資源使用率。
- 利用整合式Snapshot複本與複製、簡化開發與部署工作流程、提供即時且節省空間的使用者工作區、整合式版本控制、以及自動化部署。
- 企業級的資料保護功能、可實現災難恢復與營運不中斷。本文所述的NetApp與Lenovo解決方案是靈活的橫向擴充架構、非常適合在邊緣部署企業級AI推斷。

感謝

- J.法爾坎格、資深聯想HPC與AI解決方案經理
- NetApp技術行銷工程師Dave Arettette
- NetApp E系列AI解決方案技術主管Joey Parnell
- NetApp品質保證工程師Cody Harman

何處可找到其他資訊

若要深入瞭解本文件所述資訊、請參閱下列文件和/或網站：

- NetApp AFF A系列陣列產品頁面

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

- NetApp ONTAP 產品資料管理軟體：ONTAP 9資訊庫

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

- TR-4727：NetApp EF系列簡介

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

- NetApp E系列SANtricity 產品介紹

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

- NetApp容器持續儲存設備：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 複製與同步

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"
- TensorFlow基準測試

  "https://github.com/tensorflow/benchmarks"
- Lenovo ThinkSystem SE350 Edge Server

  "https://lenovopress.com/lp1168"
- Lenovo ThinkSystem DM5100F統一化Flash儲存陣列

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

版權資訊

Copyright © 2024 NetApp, Inc. 版權所有。台灣印製。非經版權所有人事先書面同意，不得將本受版權保護文件的任何部分以任何形式或任何方法（圖形、電子或機械）重製，包括影印、錄影、錄音或儲存至電子檢索系統中。

由 NetApp 版權資料衍伸之軟體必須遵守下列授權和免責聲明：

此軟體以 NETAPP「原樣」提供，不含任何明示或暗示的擔保，包括但不限於有關適售性或特定目的適用性之擔保，特此聲明。於任何情況下，就任何已造成或基於任何理論上責任之直接性、間接性、附隨性、特殊性、懲罰性或衍生性損害（包括但不限於替代商品或服務之採購；使用、資料或利潤上的損失；或企業營運中斷），無論是在使用此軟體時以任何方式所產生的契約、嚴格責任或侵權行為（包括疏忽或其他）等方面，NetApp 概不負責，即使已被告知有前述損害存在之可能性亦然。

NetApp 保留隨時變更本文所述之任何產品的權利，恕不另行通知。NetApp 不承擔因使用本文所述之產品而產生的責任或義務，除非明確經過 NetApp 書面同意。使用或購買此產品並不會在依據任何專利權、商標權或任何其他 NetApp 智慧財產權的情況下轉讓授權。

本手冊所述之產品受到一項（含）以上的美國專利、國外專利或申請中專利所保障。

有限權利說明：政府機關的使用、複製或公開揭露須受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中的「技術資料權利 - 非商業項目」條款 (b)(3) 小段所述之限制。

此處所含屬於商業產品和 / 或商業服務（如 FAR 2.101 所定義）的資料均為 NetApp, Inc. 所有。根據本協議提供的所有 NetApp 技術資料和電腦軟體皆屬於商業性質，並且完全由私人出資開發。 美國政府對於該資料具有非專屬、非轉讓、非轉授權、全球性、有限且不可撤銷的使用權限，僅限於美國政府為傳輸此資料所訂合約所允許之範圍，並基於履行該合約之目的方可使用。除非本文另有規定，否則未經 NetApp Inc. 事前書面許可，不得逕行使用、揭露、重製、修改、履行或展示該資料。美國政府授予國防部之許可權利，僅適用於 DFARS 條款 252.227-7015(b)（2014 年 2 月）所述權利。

商標資訊

NETAPP、NETAPP 標誌及 http://www.netapp.com/TM 所列之標章均為 NetApp, Inc. 的商標。文中所涉及的所有其他公司或產品名稱，均為其各自所有者的商標，不得侵犯。