



NetApp和聯想合作的邊緣 **AI** 推理 NetApp artificial intelligence solutions

NetApp
August 18, 2025

目錄

NetApp和聯想合作的邊緣 AI 推理	1
TR-4886：邊緣 AI 推理 - NetApp與聯想 ThinkSystem - 解決方案設計	1
總結	1
介紹	1
結論	5
致謝	6
在哪裡可以找到更多信息	6

NetApp和聯想合作的邊緣 AI 推理

TR-4886：邊緣 AI 推理 - NetApp與聯想 ThinkSystem - 解決方案設計

Sathish Thyagarajan，NetApp Miroslav Hodak，聯想

本文檔介紹了一種運算和儲存架構，用於在滿足新興應用場景的邊緣環境中在NetApp儲存控制器和 Lenovo ThinkSystem 伺服器上部署基於 GPU 的人工智慧 (AI) 推理。

總結

高級駕駛輔助系統 (ADAS)、工業 4.0、智慧城市和物聯網 (IoT) 等一些新興應用情境需要在接近零延遲的情況下處理連續資料流。本文檔介紹了一種運算和儲存架構，用於在滿足這些要求的邊緣環境中的NetApp儲存控制器和 Lenovo ThinkSystem 伺服器上部署基於 GPU 的人工智慧 (AI) 推理。本文檔還提供了業界標準 MLPerf 推理基準的效能數據，評估了配備NVIDIA T4 GPU 的邊緣伺服器上的各種推理任務。我們研究了離線、單流和多流推理場景的效能，並表明具有經濟高效的共享網路儲存系統的架構性能高，並為多個邊緣伺服器的資料和模型管理提供了中心點。

介紹

越來越多的公司在網路邊緣產生大量數據。為了從智慧感測器和物聯網數據中獲得最大價值，組織正在尋找能夠實現邊緣運算的即時事件流解決方案。因此，運算要求高的工作越來越多地在資料中心之外的邊緣執行。人工智慧推理是這一趨勢的驅動因素之一。邊緣伺服器為這些工作負載提供了足夠的運算能力，尤其是在使用加速器時，但有限的儲存空間通常是一個問題，尤其是在多伺服器環境中。在本文檔中，我們展示瞭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使 AI 推理工作負載受益。

本文檔描述了邊緣 AI 推理的參考架構。它將多台聯想 ThinkSystem 邊緣伺服器與NetApp儲存系統結合，以創建易於部署和管理的解決方案。它旨在成為各種情況下實際部署的基準指南，例如具有多個攝影機和工業感測器的工廠車間、零售交易中的銷售點 (POS) 系統或識別自動駕駛汽車視覺異常的全自動駕駛 (FSD) 系統。

本文檔涵蓋由 Lenovo ThinkSystem SE350 Edge 伺服器和入門級NetApp AFF和 EF 系列儲存系統組成的運算和儲存配置的測試和驗證。參考架構為 AI 部署提供了高效且經濟的解決方案，同時也透過NetApp ONTAP 和NetApp SANtricity資料管理軟體提供全面的資料服務、整合資料保護、無縫可擴充性和雲端連接資料儲存。

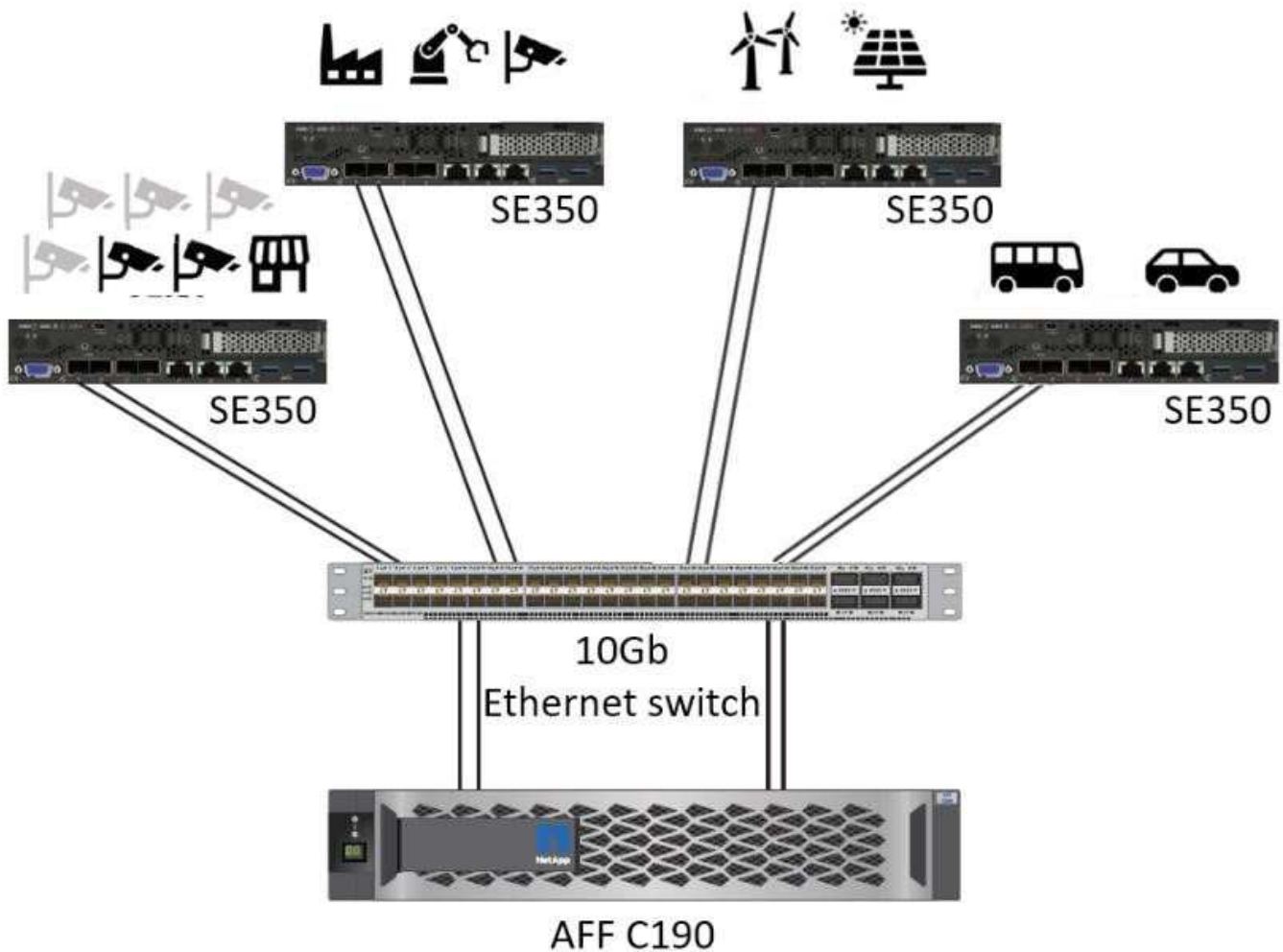
目標受眾

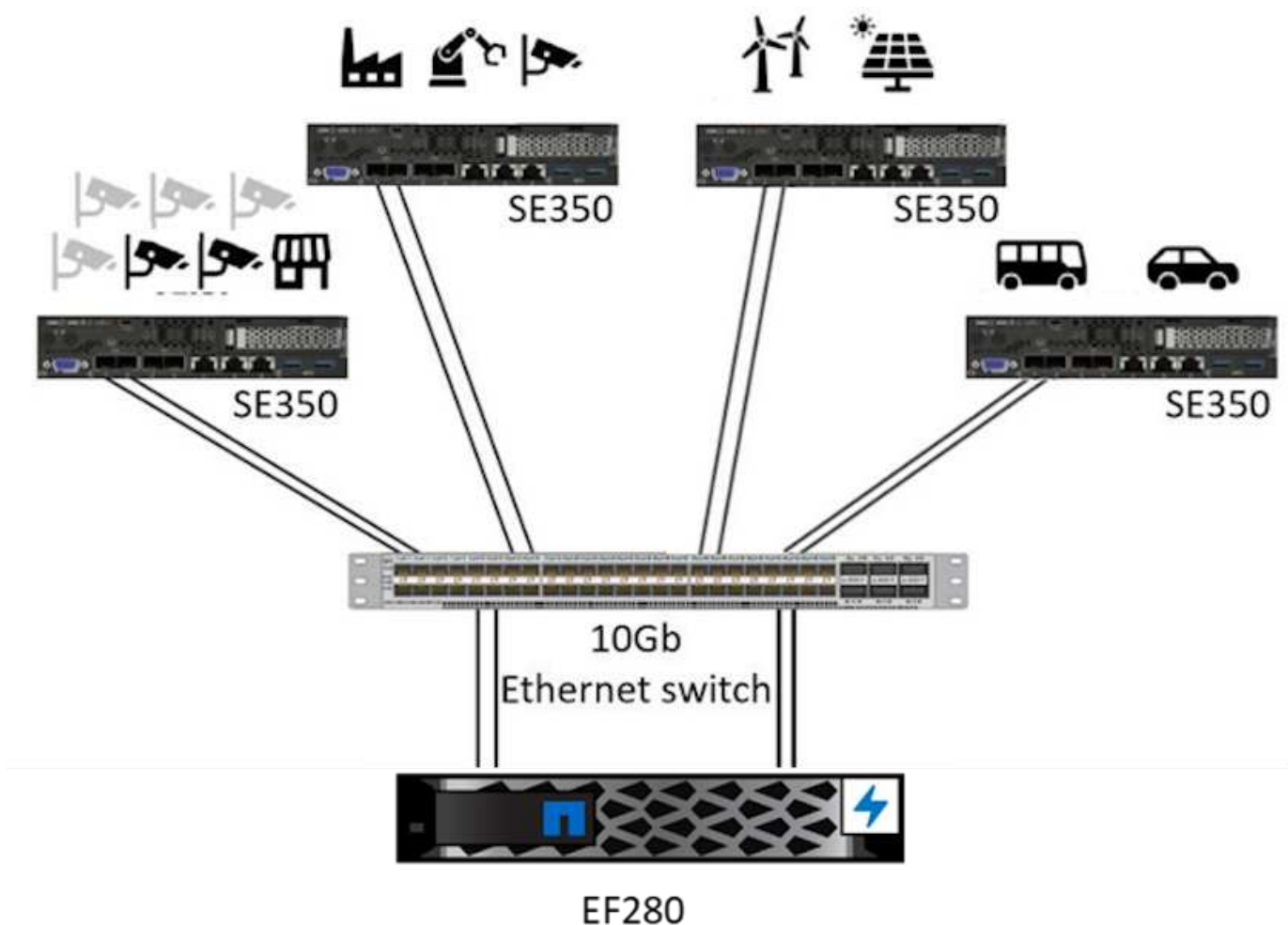
本文檔適用於以下受眾：

- 希望將邊緣 AI 產品化的商業領袖和企業架構師。
- 資料科學家、資料工程師、人工智慧/機器學習 (ML) 研究人員和人工智慧系統開發人員。
- 為 AI/ML 模型和應用程式的開發設計解決方案的企業架構師。
- 資料科學家和人工智慧工程師正在尋找部署深度學習 (DL) 和 ML 模型的有效方法。
- 負責邊緣推理模型的部署和管理的邊緣設備管理員和邊緣伺服器管理員。

解決方案架構

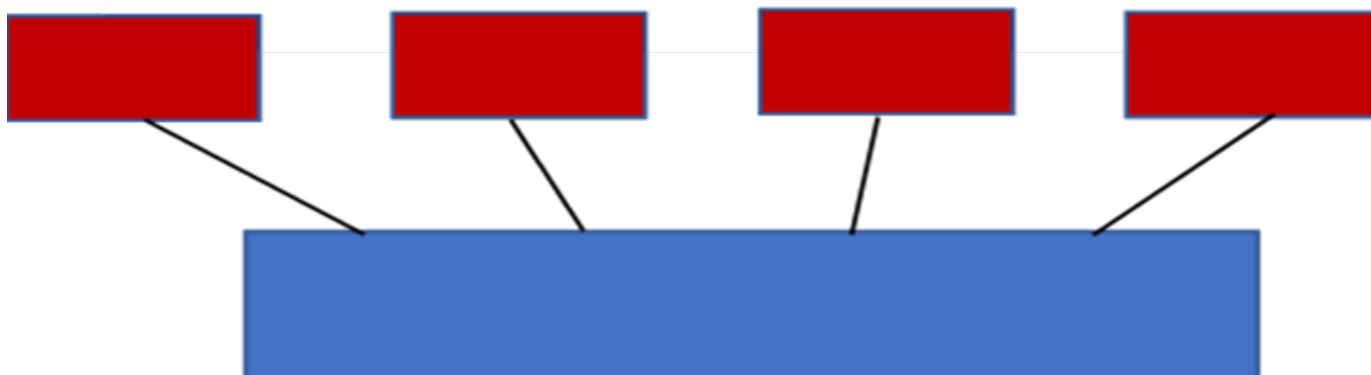
這款聯想 ThinkSystem 伺服器與 NetApp ONTAP 或 NetApp SANtricity 儲存解決方案旨在利用 GPU 和傳統 CPU 的處理能力來處理大型資料集上的 AI 推理。此驗證展示了高效能和最佳資料管理，其架構使用單一或多個 Lenovo SR350 邊緣伺服器與單一 NetApp AFF 儲存系統互連，如以下兩圖所示。





下圖中的邏輯架構概覽展示了此架構中運算和儲存元素的角色。具體來說，它顯示了以下內容：

- 邊緣運算設備對從攝影機、感測器等接收的資料進行推理。
- 具有多種用途的共享儲存元素：
 - 為推理模型和執行推理所需的其他數據提供一個中心位置。計算伺服器直接存取儲存空間並透過網路使用推理模型，而無需在本機複製它們。
 - 更新的模型推送到這裡。
 - 將邊緣伺服器接收的輸入資料存檔以供日後分析。例如，如果邊緣設備連接到攝影機，則儲存元件會保存攝影機擷取的視訊。



紅色的	藍色的
聯想計算系統	NetApp AFF儲存系統
邊緣設備對來自相機、感測器等輸入進行推理。	共享儲存保存推理模型和來自邊緣設備的數據以供後續分析。

NetApp和聯想的解決方案有以下主要優勢：

- 邊緣 GPU 加速運算。
- 部署由共享儲存支援和管理的多個邊緣伺服器。
- 強大的資料保護，滿足低恢復點目標 (RPO) 和復原時間目標 (RTO)，且不會遺失資料。
- 使用NetApp Snapshot 副本和克隆優化資料管理，以簡化開發工作流程。

如何使用此架構

本文檔驗證了所提出的架構的設計和效能。但是，我們還沒有測試某些軟體層級的部分，例如容器、工作負載或模型管理以及與雲端或資料中心內部的資料同步，因為它們特定於部署場景。這裡存在多種選擇。

在容器管理層面，Kubernetes 容器管理是一個不錯的選擇，無論是完全上游版本（Canonical）還是適合企業部署的修改版本（Red Hat）都得到了很好的支援。這["NetApp AI 控制平面"](#)它採用NetApp Trident和新添加的["NetApp DataOps 工具包"](#)為資料科學家和資料工程師提供內建的可追溯性、資料管理功能、介面和工具，以便與NetApp儲存整合。Kubeflow 是 Kubernetes 的 ML 工具包，它提供了額外的 AI 功能，並在 TensorFlow Serving 或NVIDIA Triton Inference Server 等多個平台上支援模型版本控制和 KFServing。另一個選擇是NVIDIA EGX 平台，它提供工作負載管理以及對支援 GPU 的 AI 推理容器目錄的存取。然而，這些選項可能需要付出大量努力和專業知識才能投入生產，並且可能需要第三方獨立軟體供應商 (ISV) 或顧問的協助。

解決方案領域

AI 推理和邊緣運算的主要優勢在於設備能夠高品質且無延遲地計算、處理和分析數據。本文檔中描述了太多的邊緣運算用例範例，但以下是一些突出的用例：

汽車：自動駕駛汽車

經典的邊緣運算範例是自動駕駛汽車（AV）中的高級駕駛輔助系統（ADAS）。無人駕駛汽車中的人工智慧必須快速處理來自攝影機和感測器的大量數據才能成為成功的安全駕駛員。花費太長時間來解讀物體和人之間的差異可能意味著生死，因此能夠盡可能靠近車輛處理資料至關重要。在這種情況下，一個或多個邊緣運算伺服器處理來自攝影機、雷達、光達和其他感測器的輸入，而共享儲存保存推理模型並儲存來自感測器的輸入資料。

醫療保健：病患監護

人工智慧和邊緣運算的最大影響之一是它能夠增強對家庭護理和重症監護病房 (ICU) 中慢性病患者的持續監測。監測胰島素水平、呼吸、神經活動、心律和胃腸功能的邊緣設備的數據需要即時分析，並且必須立即採取行動，因為採取行動來挽救某人的生命的時間有限。

零售：無收銀員支付

邊緣運算可以為人工智慧和機器學習提供支持，幫助零售商減少結帳時間並增加客流量。無收銀系統支援各種組件，例如：

- 身份驗證和存取。將實體購物者連接到已驗證的帳戶並允許進入零售空間。

- 庫存監控。使用感測器、RFID 標籤和電腦視覺系統來幫助購物者確認選擇或取消選擇商品。

在這裡，每個邊緣伺服器處理每個結帳櫃檯，共享儲存系統作為中央同步點。

金融服務：自助服務終端的人員安全與詐欺預防

銀行機構正在使用人工智慧和邊緣運算來創新和創造個人化的銀行體驗。使用即時數據分析和人工智慧推理的互動式自助服務終端現在不僅能夠讓 ATM 機幫助客戶提款，還能透過攝影機擷取的影像主動監控自助服務終端，以識別對人類安全的風險或詐欺行為。在這個場景中，邊緣運算伺服器和共享儲存系統連接到互動式資訊亭和攝影機，幫助銀行使用人工智慧推理模型收集和處理資料。

製造業：工業 4.0

第四次工業革命（工業 4.0）已經開始，同時也出現了智慧工廠和 3D 列印等新興趨勢。為了迎接資料主導的未來，大規模機器對機器 (M2M) 通訊和物聯網被整合在一起，以提高自動化程度，而無需人工干預。製造業已經高度自動化，添加人工智慧功能是長期趨勢的自然延續。人工智慧可以實現自動化操作，這些操作可以藉助電腦視覺和其他人工智慧功能實現。您可以自動化品質控制或依賴人類視覺或決策的任務，以便對工廠車間裝配線上的材料進行更快的分析，幫助製造工廠滿足所需的 ISO 安全和品質管理標準。在這裡，每個運算邊緣伺服器都連接到監控製造過程的感測器陣列，並根據需要將更新的推理模型推送到共用儲存。

電信：鏽蝕檢測、塔台檢查和網路最佳化

電信業使用電腦視覺和人工智慧技術處理影像，自動檢測鏽蝕並識別含有腐蝕並因此需要進一步檢查的手機訊號塔。近年來，使用無人機影像和人工智慧模型來識別塔的不同區域以分析鏽蝕、表面裂縫和腐蝕的情況越來越多。人們對人工智慧技術的需求持續成長，這些技術可以有效地檢查電信基礎設施和手機訊號塔，定期評估其性能是否下降，並在需要時及時修復。

此外，電信領域的另一個新興用例是使用人工智慧和機器學習演算法來預測資料流量模式、檢測支援 5G 的設備以及自動化和增強多輸入多輸出 (MIMO) 能源管理。無線電塔使用 MIMO 硬體來增加網路容量；然而，這會帶來額外的能源成本。部署在蜂窩基地台的「MIMO 睡眠模式」的 ML 模型可以預測無線電的有效使用情況，並有助於降低行動網路營運商 (MNO) 的能源消耗成本。人工智慧推理和邊緣運算解決方案可協助 MNO 減少往返資料中心的資料量、降低 TCO、優化網路營運並提高最終用戶的整體效能。

結論

人工智慧驅動的自動化和邊緣運算是幫助商業組織實現數位轉型並最大限度提高營運效率和安全性的領先方法。透過邊緣運算，資料處理速度更快，因為它不必往返於資料中心。因此，與資料中心或雲端來回發送資料相關的成本就會降低。當企業必須使用部署在邊緣的人工智慧推理模型近乎即時地做出決策時，更低的延遲和更高的速度會很有幫助。

NetApp 儲存系統提供與本機 SSD 儲存相同或更好的效能，並為資料科學家、資料工程師、AI/ML 開發人員以及業務或 IT 決策者提供以下優勢：

- 在人工智慧系統、分析系統和其他關鍵業務系統之間輕鬆共享資料。這種資料共享減少了基礎設施開銷，提高了效能，並簡化了整個企業的資料管理。
- 獨立可擴展的計算和存儲，以最大限度地降低成本並提高資源利用率。
- 使用整合的 Snapshot 副本和複製來簡化開發和部署工作流程，以實現即時且節省空間的使用者工作區、整合版本控制和自動部署。
- 實現災難復原和業務連續性的企業級資料保護。本文檔中介紹的 NetApp 和聯想解決方案是一種靈活的橫向擴

展架構，非常適合邊緣企業級 AI 推理部署。

致謝

- 俊傑Falkanger，聯想 HPC 和人工智慧解決方案高級經理
- NetApp技術行銷工程師 Dave Arnette
- Joey Parnell，NetApp E 系列 AI 解決方案技術主管
- Cody Harryman，NetApp品質保證工程師

在哪裡可以找到更多信息

要了解有關本文檔中描述的信息的更多信息，請參閱以下文檔和/或網站：

- NetApp AFF A系列陣列產品頁面

["https://www.netapp.com/data-storage/aff-a-series/"](https://www.netapp.com/data-storage/aff-a-series/)

- NetApp ONTAP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軟體資料表

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

- NetApp容器持久性儲存 — NetApp Trident

["https://netapp.io/persistent-storage-provisioner-for-kubernetes/"](https://netapp.io/persistent-storage-provisioner-for-kubernetes/)

- MLPerf

- ["https://mlcommons.org/en/"](https://mlcommons.org/en/)
- ["http://www.image-net.org/"](http://www.image-net.org/)
- ["https://mlcommons.org/en/news/mlperf-inference-v11/"](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"](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"](https://github.com/tensorflow/benchmarks)

- 聯想 ThinkSystem SE350 邊緣伺服器

["https://lenovopress.com/lp1168"](https://lenovopress.com/lp1168)

- 聯想 ThinkSystem DM5100F 統一快閃儲存陣列

["https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"](https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array)

版權資訊

Copyright © 2025 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。