搭載NVIDIA DGX 系統的NetApp AIPod

NetApp artificial intelligence solutions

NetApp
February 12, 2026

目錄

搭載NVIDIA DGX 系統的NetApp AIPod

NVA-1173 NetApp AIPod與NVIDIA DGX 系統 - 簡介

POWERED BY

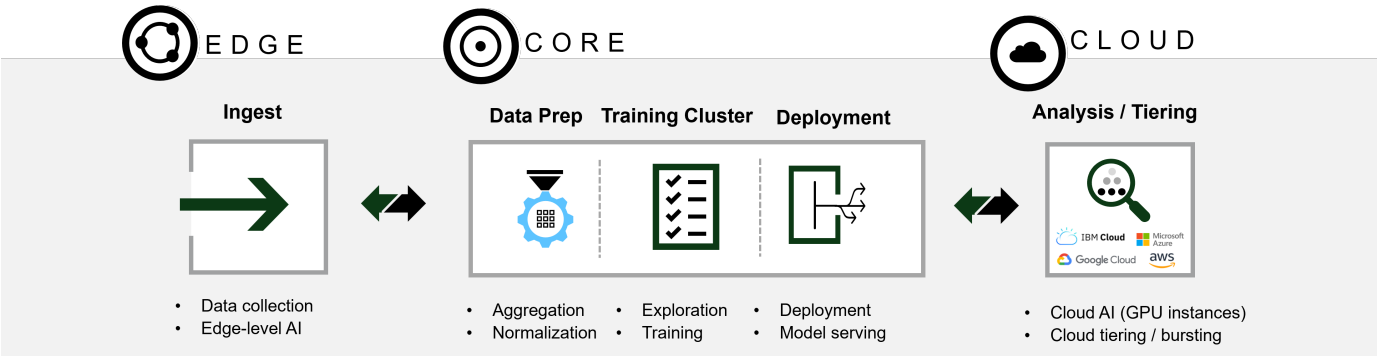
NVIDIA

NetApp解決方案工程

執行摘要

NetApp™ AIPod配備NVIDIA DGX™ 系統和NetApp雲端連接儲存系統，透過消除設計複雜性和猜測，簡化了機器學習 (ML) 和人工智慧 (AI) 工作負載的基礎設施部署。基於NVIDIA DGX BasePOD™ 設計，旨在為下一代工作負載提供卓越的運算效能，搭載NVIDIA DGX 系統的AIPod增加了NetApp AFF儲存系統，使客戶能夠從小規模開始並無中斷地發展，同時智慧地管理從邊緣到核心再到雲端的資料。 NetApp AIPod是NetApp AI 解決方案產品組合的一部分，如下圖所示。

NetApp 人工智慧解決方案組合



本文檔描述了AIPod參考架構的關鍵組件、系統連接和配置資訊、驗證測試結果和解決方案規模指導。本文檔適用於有興趣為 ML/DL 和分析工作負載部署高效能基礎架構的NetApp和合作夥伴解決方案工程師以及客戶策略決策者。

搭載NVIDIA DGX 系統的 NVA-1173 NetApp AIPod - 硬體組件

本節重點介紹具有NVIDIA DGX 系統的NetApp AIPod的硬體組件。

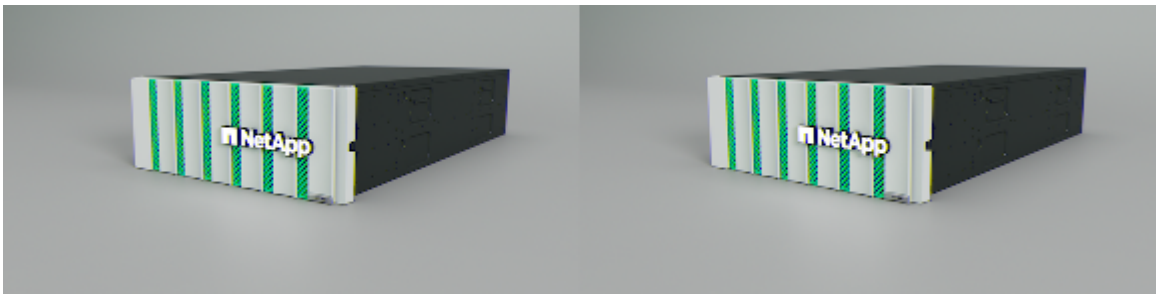
NetApp AFF儲存系統

NetApp AFF最先進的儲存系統使 IT 部門能夠透過業界領先的效能、卓越的靈活性、雲端整合和一流的資料管理來滿足企業儲存需求。 AFF系統專為快閃記憶體設計，有助於加速、管理和保護關鍵業務資料。

AFF A90儲存系統

由NetApp ONTAP資料管理軟體提供支援的NetApp AFF A90提供內建資料保護、可選的反勒索軟體功能以及支援最關鍵業務工作負載所需的高效能和彈性。它消除了對關鍵任務操作的中斷，最大限度地減少了效能調整，並保護您的資料免受勒索軟體攻擊。它提供：• 業界領先的效能 • 不折不扣的資料安全性 • 簡化的無中斷升級

NetApp AFF A90儲存系統



業界領先的性能

AFF A90可輕鬆管理深度學習、人工智慧和高速分析等新一代工作負載以及 Oracle、SAP HANA、Microsoft SQL Server 和虛擬化應用程式等傳統企業資料庫。它使關鍵業務應用程式保持最高速度運行，每個 HA 對高達 2.4M IOPS，延遲低至 100µs，並且性能比以前的NetApp型號提高高達 50%。借助 NFS over RDMA、pNFS 和會話中繼，客戶可以使用現有的資料中心網路基礎設施實現下一代應用程式所需的高水準網路效能。客戶還可以透過對 SAN、NAS 和物件儲存的統一多協定支援進行擴展和成長，並透過統一的單一ONTAP資料管理軟體為本地或雲端資料提供最大的靈活性。此外，還可以透過Active IQ和Cloud Insights提供的基於 AI 的預測分析來優化系統健康狀況。

不妥協的資料安全

AFF A90系統包含一整套NetApp整合和應用程式一致的資料保護軟體。它提供內建資料保護和尖端反勒索軟體解決方案，用於預防和攻擊後復原。可以阻止惡意檔案寫入磁碟，並且可以輕鬆監控儲存異常以取得洞察。

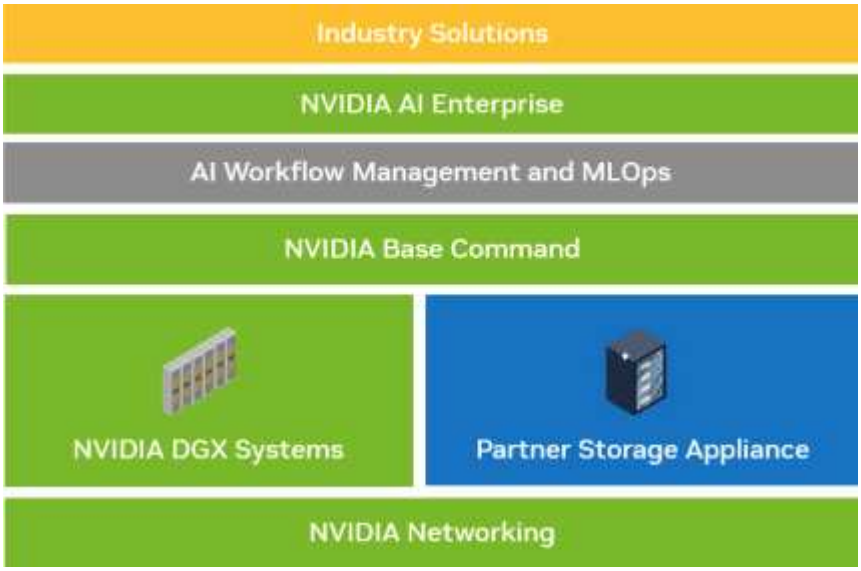
對於現有的 A800 客戶來說， AFF A90可以作為無中斷機殼內升級。 NetApp憑藉其先進的可靠性、可用性、可維護性和可管理性 (RASM) 功能，可輕鬆更新並消除關鍵任務操作的中斷。此外，由於ONTAP軟體會自動為所有系統元件應用韌體更新， NetApp進一步提高了營運效率並簡化了 IT 團隊的日常活動。

對於最大的部署， AFF A1K系統提供最高的效能和容量選項，而其他NetApp儲存系統（如AFF A70和AFF C800）則以較低的成本為較小的部署提供選項。

NVIDIA DGX BasePOD

NVIDIA DGX BasePOD是由NVIDIA硬體和軟體元件、MLOps 解決方案以及第三方儲存組成的整合解決方案。利用NVIDIA產品和經過驗證的合作夥伴解決方案的橫向擴展系統設計最佳實踐，客戶可以實現高效且易於管理的 AI 開發平台。圖 1 突顯了NVIDIA DGX BasePOD的各個元件。

NVIDIA DGX BasePOD 解決方案



NVIDIA DGX H100 系統

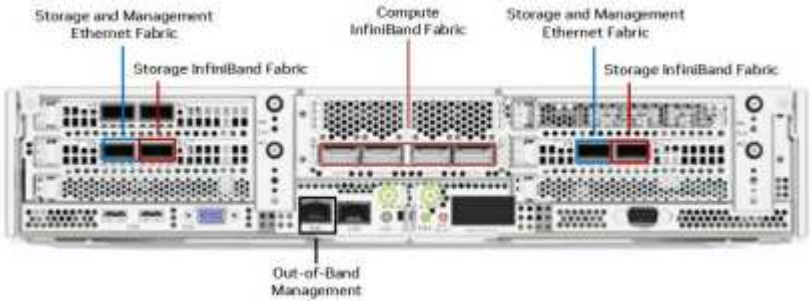
NVIDIA DGX H100™ 系統是 AI 的強大引擎，由NVIDIA H100 Tensor Core GPU 的突破性效能加速。

NVIDIA DGX H100 系統

DGX H100 系統的主要規格如下：• 八個NVIDIA H100 GPU。 • 每個 GPU 配備 80 GB GPU 內存，總計 640GB。 • 四個NVIDIA NVSwitch 晶片。 • 雙 56 核心 Intel Xeon Platinum 8480 處理器，支援 PCIe 5.0。 • 2 TB DDR5 系統記憶體。 • 四個 OSFP 端口，服務八個單端口NVIDIA ConnectX™-7（InfiniBand/乙太網路）適

配器和兩個雙端口NVIDIA ConnectX-7（InfiniBand/乙太網路）適配器。 • 兩個 1.92 TB M.2 NVMe 硬碟用於 DGX OS，八個 3.84 TB U.2 NVMe 硬碟用於儲存/快取。 • 最大功率10.2 kW。 DGX H100 CPU 托盤的後連接埠如下所示。四個 OSFP 連接埠為 InfiniBand 計算結構的八個 ConnectX-7 適配器提供服務。每對雙連接埠 ConnectX-7 適配器為儲存和管理結構提供平行路徑。帶外端口用於BMC存取。

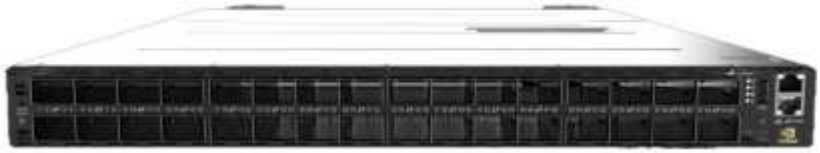
NVIDIA DGX H100 後面板



NVIDIA網絡

NVIDIA Quantum-2 QM9700 交換機

NVIDIA Quantum-2 QM9700 InfiniBand 交換器



具有 400Gb/s InfiniBand 連接的NVIDIA Quantum-2 QM9700 交換器為NVIDIA Quantum-2 InfiniBand BasePOD 配置中的運算結構提供動力。 ConnectX-7 單埠適配器用於 InfiniBand 計算結構。每個NVIDIA DGX 系統與每個 QM9700 交換器都有雙重連接，從而在系統之間提供多條高頻寬、低延遲路徑。

NVIDIA Spectrum-3 SN4600 交換機

NVIDIA Spectrum-3 SN4600 交換器

NVIDIA Spectrum™-3 SN4600 交換器總共提供 128 個連接埠（每個交換器 64 個），為 DGX BasePOD 的帶內管理提供冗餘連接。 NVIDIA SN4600 交換器可提供 1 GbE 到 200 GbE 之間的速度。對於透過乙太網路連接的儲存設備，也使用NVIDIA SN4600 交換器。 NVIDIA DGX 雙埠 ConnectX-7 轉接器上的連接埠用於內建管理和儲存連線。

NVIDIA Spectrum SN2201 交換機

NVIDIA Spectrum SN2201 交換機

NVIDIA Spectrum SN2201 交換器提供 48 個端口，可為帶外管理提供連接。帶外管理為 DGX BasePOD 中的所有元件提供整合的管理連線。

NVIDIA ConnectX-7 轉接器

NVIDIA ConnectX-7 適配器

NVIDIA ConnectX-7 轉接器可提供 25/50/100/200/400G 的吞吐量。 NVIDIA DGX 系統使用單埠和雙埠 ConnectX-7 轉接器，為具有 400Gb/s InfiniBand 和乙太網路的 DGX BasePOD 部署提供靈活性。

NVA-1173 NetApp AIPod與NVIDIA DGX 系統 - 軟體元件

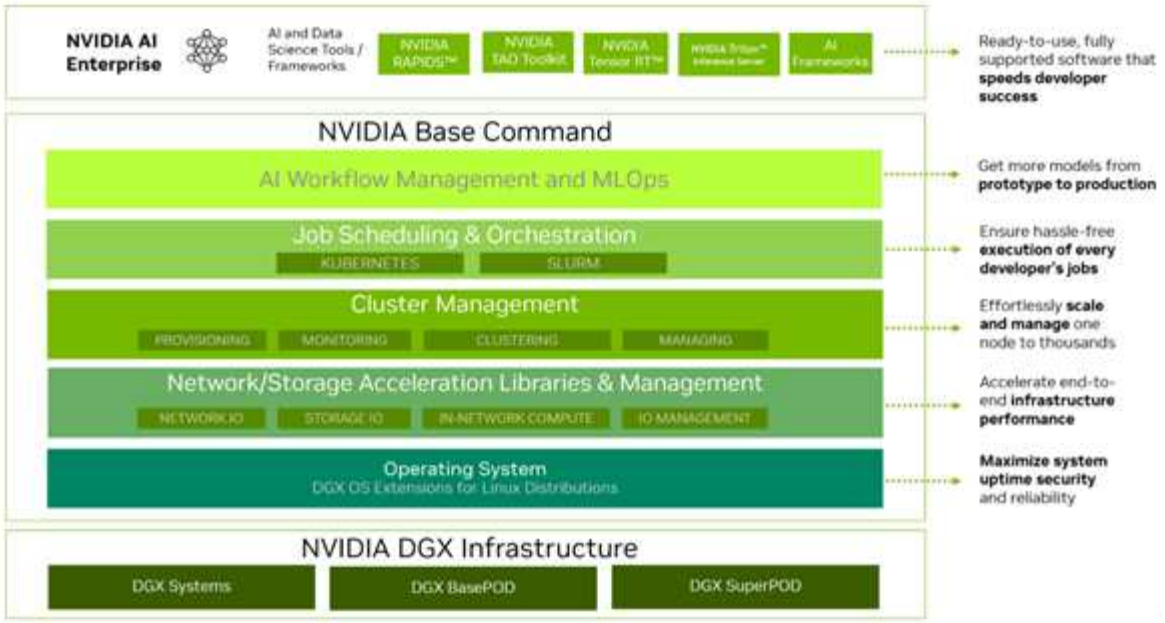
本節重點介紹具有NVIDIA DGX 系統的NetApp AIPod的軟體元件。

NVIDIA軟體

NVIDIA基本指令

NVIDIA Base Command™ 為每個 DGX BasePOD 提供支持，使組織能夠充分利用NVIDIA軟體創新的最佳成果。企業可以透過經過驗證的平台充分發揮其投資潛力，該平台包括企業級編排和叢集管理、加速運算、儲存和網路基礎設施的程式庫以及針對 AI 工作負載優化的作業系統 (OS)。

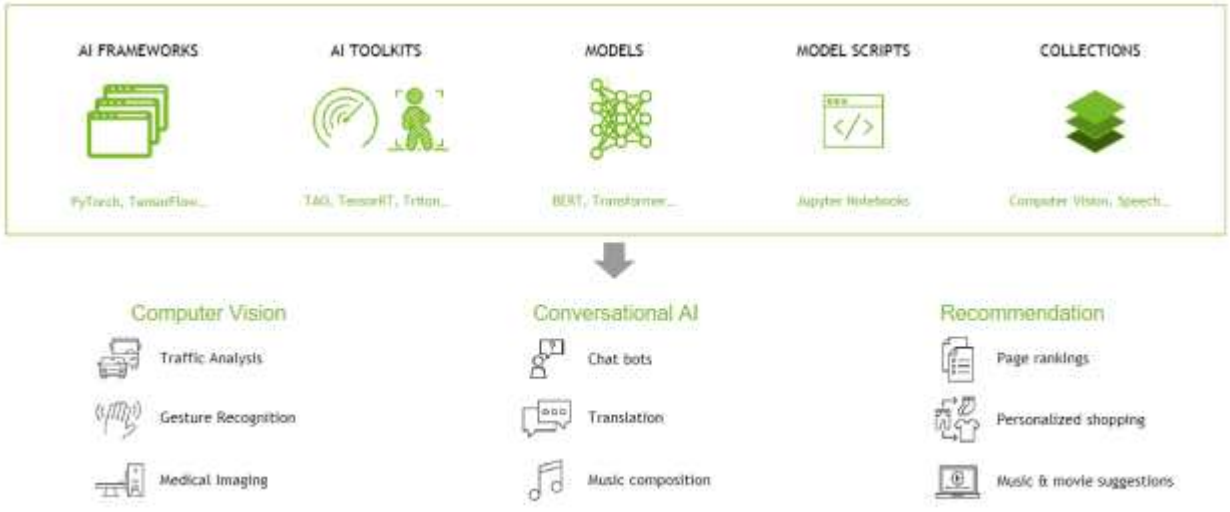
NVIDIA BaseCommand 解決方案



NVIDIA GPU 雲端 **(NGC)**

NVIDIA NGC 提供的軟體可滿足具有不同 AI 專業水平的資料科學家、開發人員和研究人員的需求。 NGC 上託管的軟體會針對一組常見漏洞和暴露 (CVE)、加密和私鑰進行掃描。它經過測試和設計，可擴展到多個 GPU，在許多情況下，可擴展到多節點，確保用戶最大限度地利用其在 DGX 系統上的投資。

NVIDIA GPU 雲端



NVIDIA AI 企業版

NVIDIA AI Enterprise 是一個端對端軟體平台，可讓每個企業都能夠使用生成式 AI，為在NVIDIA DGX 平台上優化的生成式 AI 基礎模型提供最快、最高效的運行時。憑藉生產級的安全性、穩定性和可管理性，它簡化了生成式 AI 解決方案的開發。 NVIDIA AI Enterprise 包含在 DGX BasePOD 中，企業開發人員可以存取預訓練模型、最佳化框架、微服務、加速庫和企業支援。

NetApp軟體

NetApp ONTAP

ONTAP 9 是NetApp最新一代儲存管理軟體，它支援企業實現基礎架構現代化並過渡到雲端就緒資料中心。ONTAP利用業界領先的數據管理功能，只需一套工具即可管理和保護數據，無論數據位於何處。您也可以將資料自由移動到任何需要的地方：邊緣、核心或雲端。 ONTAP 9 包含眾多功能，可簡化資料管理、加速和保護關鍵數據，並支援跨混合雲架構的下一代基礎架構功能。

加速並保護數據

ONTAP提供卓越等級的效能和資料保護，並透過以下方式擴展這些功能：

- 性能和更低的延遲。 ONTAP以最低的延遲提供最高的吞吐量，包括支援使用 NFS over RDMA、平行 NFS (pNFS) 和 NFS 會話中繼的NVIDIA GPUDirect Storage (GDS)。
- 資料保護。 ONTAP提供內建資料保護功能和業界最強大的反勒索軟體保障，並在所有平台上實現通用管理。
- NetApp磁碟區加密 (NVE)。 ONTAP提供原生磁碟區級加密，同時支援板載和外部金鑰管理。
- 儲存多租戶和多因素身份驗證。 ONTAP支援以最高等級的安全性共用基礎架構資源。

簡化資料管理

資料管理對於企業 IT 營運和資料科學家至關重要，以便將適當的資源用於 AI 應用程式和訓練 AI/ML 資料集。以下有關NetApp技術的附加資訊超出了本次驗證的範圍，但可能與您的部署相關。

ONTAP資料管理軟體包括以下功能，可簡化操作並降低總營運成本：

- 快照和複製支援 ML/DL 工作流程的協作、平行實驗和增強資料治理。
- SnapMirror可在混合雲和多站點環境中實現無縫資料移動，並在所需的時間和地點提供資料。
- 內聯資料壓縮和擴展重複資料刪除。資料壓縮減少了儲存區塊內部浪費的空間，重複資料刪除顯著增加了有效容量。這適用於本地儲存的資料和分層到雲端的資料。
- 最小、最大和自適應服務品質 (AQoS)。細粒度的服務品質 (QoS) 控制有助於維持高度共享環境中關鍵應用程式的效能水準。
- NetApp FlexGroups 支援在儲存叢集中的所有節點上分散數據，為超大資料集提供龐大的容量和更高的效能。
- NetApp FabricPool。提供冷資料到公有和私有雲儲存選項的自動分層，包括 Amazon Web Services (AWS)、Azure 和NetApp StorageGRID儲存解決方案。有關FabricPool的更多信息，請參閱 "TR-4598：FabricPool最佳實踐"。
- NetApp FlexCache。提供遠端磁碟區快取功能，可簡化檔案分發、減少 WAN 延遲並降低 WAN 頻寬成本。FlexCache支援跨多個站點的分散式產品開發，以及從遠端位置加速存取公司資料集。

面向未來的基礎設施

ONTAP具有以下功能，可協助滿足嚴苛且不斷變化的業務需求：

- 無縫擴展和無中斷操作。 ONTAP支援在線為現有控制器和橫向擴展叢集新增容量。客戶可以升級到最新技術，例如 NVMe 和 32Gb FC，而無需昂貴的資料遷移或中斷。
- 雲端連線。 ONTAP是與雲端連接最緊密的儲存管理軟體，在所有公有雲中均提供軟體定義儲存（ONTAP

Select）和Google Cloud NetApp Volumes Volumes ）的選項。

- 與新興應用程式的整合。 ONTAP使用支援現有企業應用的相同基礎架構，為下一代平台和應用（如自動駕駛汽車、智慧城市和工業 4.0）提供企業級資料服務。

NetApp DataOps 工具包

NetApp DataOps Toolkit 是一款基於 Python 的工具，可簡化由高效能、橫向擴展NetApp儲存支援的開發/培訓工作區和推理伺服器的管理。 DataOps Toolkit 可以作為獨立實用程式運行，並且在利用NetApp Trident自動化儲存作業的 Kubernetes 環境中更有效。主要功能包括：

- 快速配置由高效能、橫向擴充NetApp儲存支援的新的高容量 JupyterLab 工作區。
- 快速配置由企業級NetApp儲存支援的全新NVIDIA Triton 推理伺服器實例。
- 近乎即時地克隆高容量的 JupyterLab 工作區，以實現實驗或快速迭代。
- 用於備份和/或可追溯性/基準的高容量 JupyterLab 工作區的近乎即時的快照。
- 近乎即時地配置、複製和快照高容量、高效能資料磁碟區。

NetApp Trident

Trident是一個完全支援的開源儲存編排器，適用於容器和 Kubernetes 發行版（包括 Anthos）。Trident可與整個NetApp儲存產品組合搭配使用，包括NetApp ONTAP，並且還支援 NFS、NVMe/TCP 和 iSCSI 連線。Trident允許最終用戶從其NetApp儲存系統配置和管理存儲，而無需儲存管理員的干預，從而加速 DevOps 工作流程。

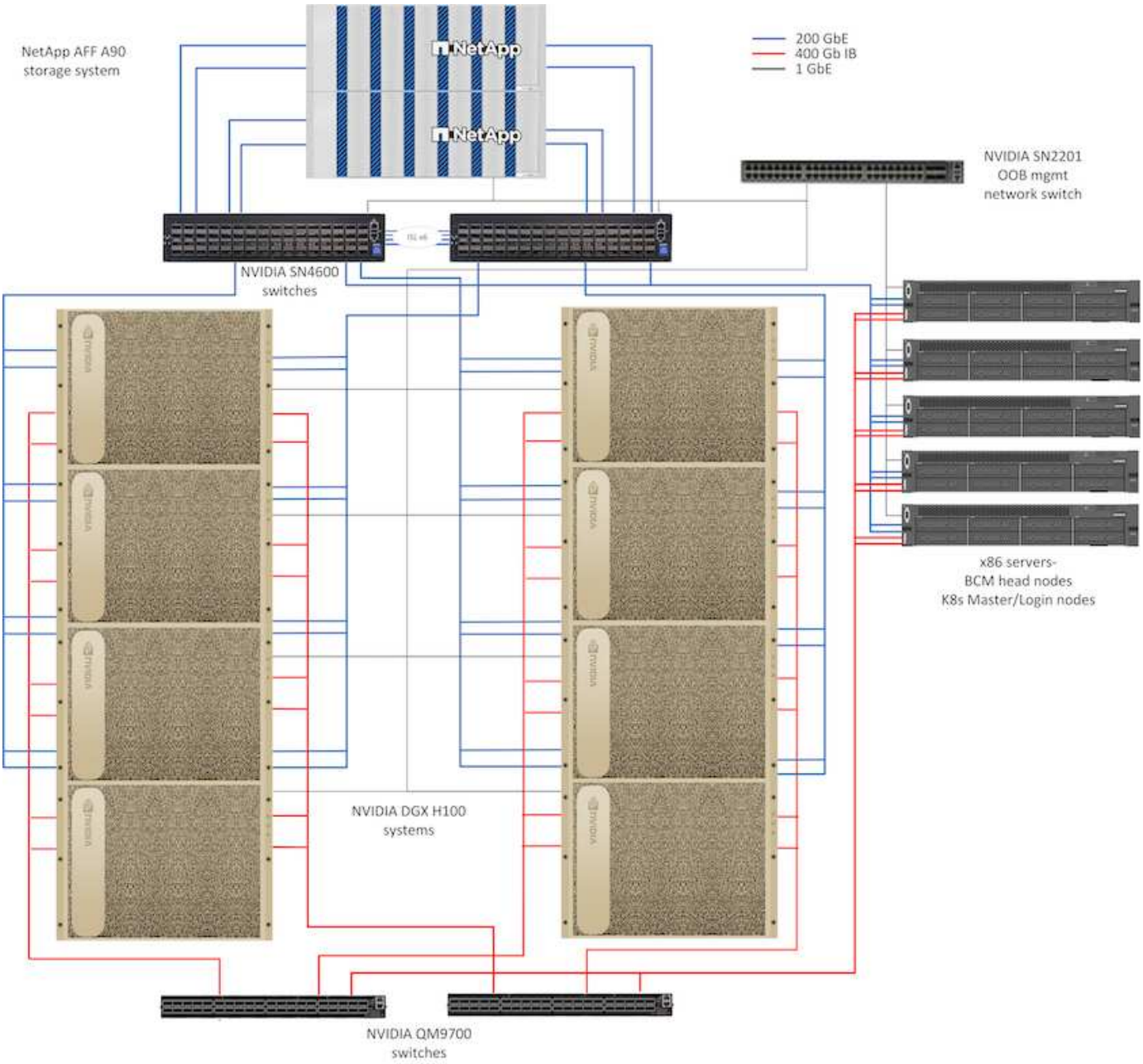
NVA-1173 NetApp AIPod與NVIDIA DGX H100 系統 - 解決方案架構

本節重點在於採用NVIDIA DGX 系統的NetApp AIPod的架構。

搭載 **DGX** 系統的**NetApp AIPod**

此參考架構利用單獨的結構進行計算叢集互連和儲存訪問，並在計算節點之間實現 400Gb/s InfiniBand (IB) 連接。下圖展示了NetApp AIPod與 DGX H100 系統的整體解決方案拓撲。

NetApp AIpod 解決方案拓撲



網路設計

在此配置中，計算叢集結構使用一對 QM9700 400Gb/s IB 交換機，它們連接在一起以實現高可用性。每個 DGX H100 系統使用八個連接連接到交換機，其中偶數連接埠連接到一個交換機，奇數連接埠連接到另一個交換器。

對於儲存系統存取、帶內管理和用戶端訪問，使用一對 SN4600 乙太網路交換器。交換器之間透過交換器間連結連接，並配置多個VLAN來隔離各種流量類型。在特定 VLAN 之間啟用基本 L3 路由，以在同一交換器上的用戶端和儲存介面之間以及交換器之間啟用多條路徑，從而實現高可用性。對於更大的部署，可以透過根據需要為主幹交換器添加額外的交換器對以及為其他葉子交換器添加額外的交換器對，將以太網網路擴展為葉子-主幹配置。

除了計算互連和高速乙太網路之外，所有實體設備還連接到一個或多個 SN2201 乙太網路交換機，以進行帶外管理。請參閱"部署詳細信息"頁面以取得有關網路配置的更多資訊。

DGX H100 系統的儲存存取概述

每個 DGX H100 系統都配備了兩個雙埠 ConnectX-7 轉接器用於管理和儲存流量，並且對於此解決方案，每個卡上的兩個連接埠都連接到同一個交換器。然後將每個卡的一個連接埠配置為 LACP MLAG 綁定，並將一個連接埠連接到每個交換機，並且在帶內管理、用戶端存取和用戶級儲存存取的 VLAN 都託管在此綁定上。

每張卡上的另一個連接埠用於連接AFF A90儲存系統，並且可以根據工作負載要求以多種配置使用。對於使用 NFS over RDMA 來支援NVIDIA Magnum IO GPUDirect Storage 的配置，連接埠單獨使用，且 IP 位址位於單獨的 VLAN 中。對於不需要 RDMA 的部署，儲存介面也可以配置 LACP 綁定，以提供高可用性和額外的頻寬。無論是否使用 RDMA，用戶端都可以使用 NFS v4.1 pNFS 和會話中繼掛載儲存系統，以實現對叢集中所有儲存節點的平行存取。請參閱"部署詳細信息"頁面以取得有關客戶端配置的更多資訊。

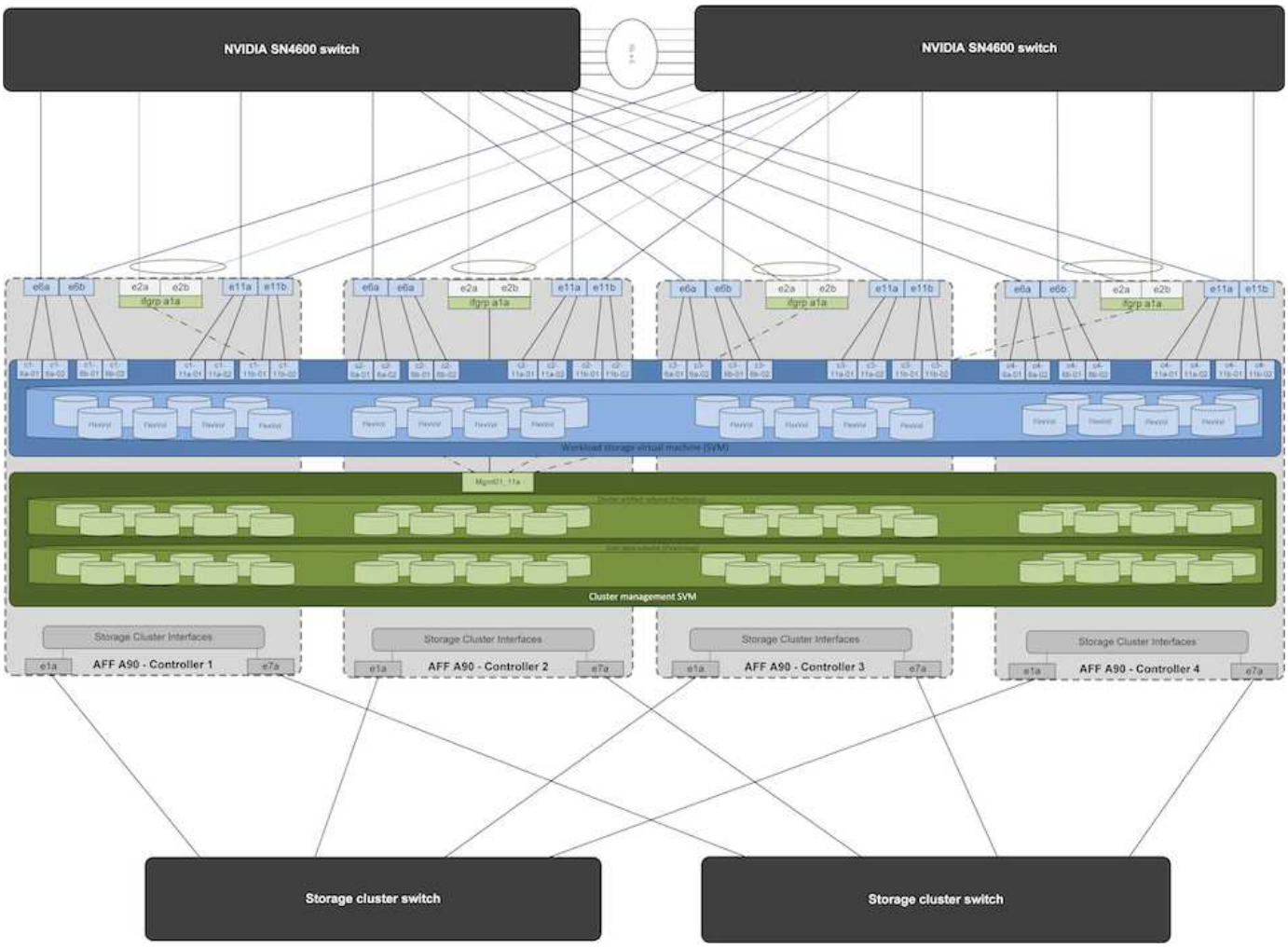
有關 DGX H100 系統連接的詳細信息，請參閱"NVIDIA BasePOD 文檔"。

儲存系統設計

每個AFF A90儲存系統使用每個控制器的六個 200 GbE 連接埠進行連接。每個控制器的四個連接埠用於從 DGX 系統存取工作負載數據，每個控制器的兩個連接埠配置為 LACP 介面群組，以支援從管理平面伺服器存取叢集管理工件和使用者主目錄。儲存系統的所有資料存取均透過 NFS 提供，其中有一個專用於 AI 工作負載存取的儲存虛擬機器 (SVM) 和一個專用於叢集管理用途的單獨 SVM。

管理 SVM 只需要一個 LIF，該 LIF 託管在每個控制器上配置的 2 連接埠介面組上。其他FlexGroup磁碟區在管理 SVM 上進行配置，以容納叢集管理構件，如叢集節點映像、系統監控歷史資料和最終使用者主目錄。下圖顯示了儲存系統的邏輯配置。

NetApp A90 儲存叢集邏輯配置



管理平面伺服器

此參考架構還包括五個基於 CPU 的伺服器，用於管理平面。其中兩個系統用作NVIDIA Base Command Manager 的頭節點，用於叢集部署和管理。其他三個系統用於提供額外的叢集服務，例如 Kubernetes 主節點或利用 Slurm 進行作業排程的部署的登入節點。利用 Kubernetes 的部署可以利用NetApp Trident CSI 驅動程式為AFF A900儲存系統上的管理和 AI 工作負載提供具有持久性儲存的自動設定和資料服務。

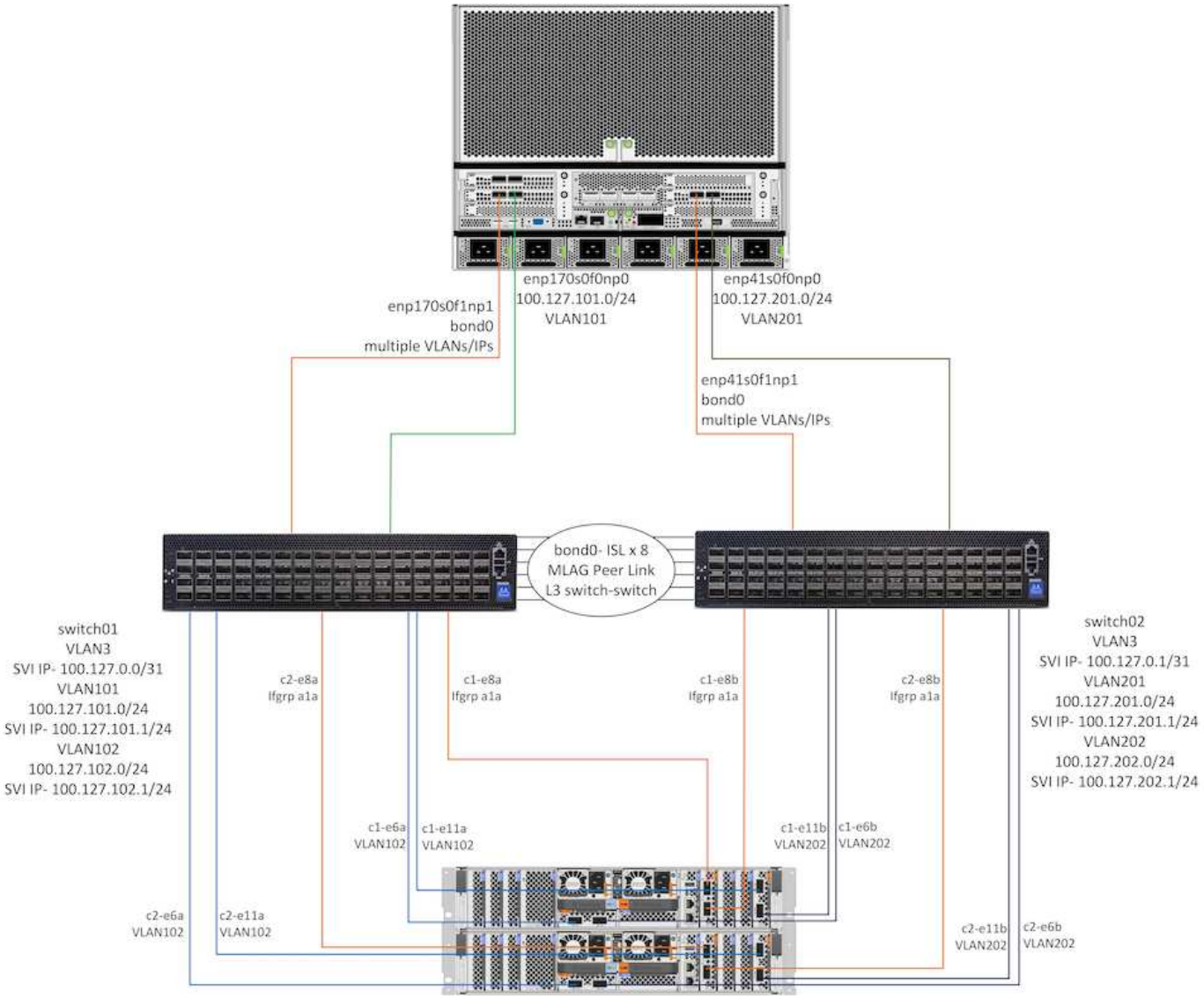
每台伺服器都實體連接到 IB 交換器和乙太網路交換機，以實現叢集部署和管理，並透過管理 SVM 配置 NFS 掛載到儲存系統，以儲存前面所述的叢集管理工件。

NVA-1173 NetApp AIPod與NVIDIA DGX 系統 - 部署詳情

本節介紹驗證此解決方案期間所使用的部署細節。使用的 IP 位址僅供參考，請依部署環境進行修改。有關此配置的實現中使用的特定命令的更多信息，請參閱相應的產品文檔。

下圖顯示了 1 個 DGX H100 系統和 1 個 HA 對AFF A90控制器的詳細網路和連接資訊。以下部分中的部署指南是基於此圖中的詳細資訊。

NetApp AIpod 網路配置



下表顯示了最多 16 個 DGX 系統和 2 個AFF A90 HA 對的範例佈線分配。

交換器和連接埠	裝置	設備連接埠
交換器1埠1-16	DGX-H100-01 至 -16	enp170s0f0np0，插槽1埠1
交換器1埠17-32	DGX-H100-01 至 -16	enp170s0f1np1，插槽1埠2
交換器1埠33-36	AFF-A90-01 至 -04	端口 e6a
交換器1埠37-40	AFF-A90-01 至 -04	端口 e11a
交換器1埠41-44	AFF-A90-01 至 -04	端口 e2a
交換器1埠57-64	ISL 到交換器 2	埠 57-64
交換器2埠1-16	DGX-H100-01 至 -16	enp41s0f0np0，插槽2埠1
交換器2埠17-32	DGX-H100-01 至 -16	enp41s0f1np1，插槽 2 埠 2
交換器2埠33-36	AFF-A90-01 至 -04	埠 e6b
交換器2埠37-40	AFF-A90-01 至 -04	埠 e11b

交換器和連接埠	裝置	設備連接埠
交換器2埠41-44	AFF-A90-01 至 -04	埠 e2b
交換器2埠57-64	ISL 到交換器 1	埠 57-64

下表顯示了本次驗證中使用的各個組件的軟體版本。

裝置	軟體版本
NVIDIA SN4600 交換機	Cumulus Linux v5.9.1
NVIDIA DGX 系統	DGX 作業系統 v6.2.1（Ubuntu 22.04 LTS）
Mellanox OFED	24.01
NetApp AFF A90	NetApp ONTAP 9.14.1

儲存網路配置

本節概述乙太網路儲存網路配置的關鍵細節。有關配置 InfiniBand 計算網路的信息，請參閱"NVIDIA BasePOD 文檔"。有關交換器配置的詳細信息，請參閱"NVIDIA Cumulus Linux 文檔"。

設定 SN4600 交換器的基本步驟概述如下。此程序假定佈線和基本交換器設定（管理 IP 位址、許可證等）已完成。

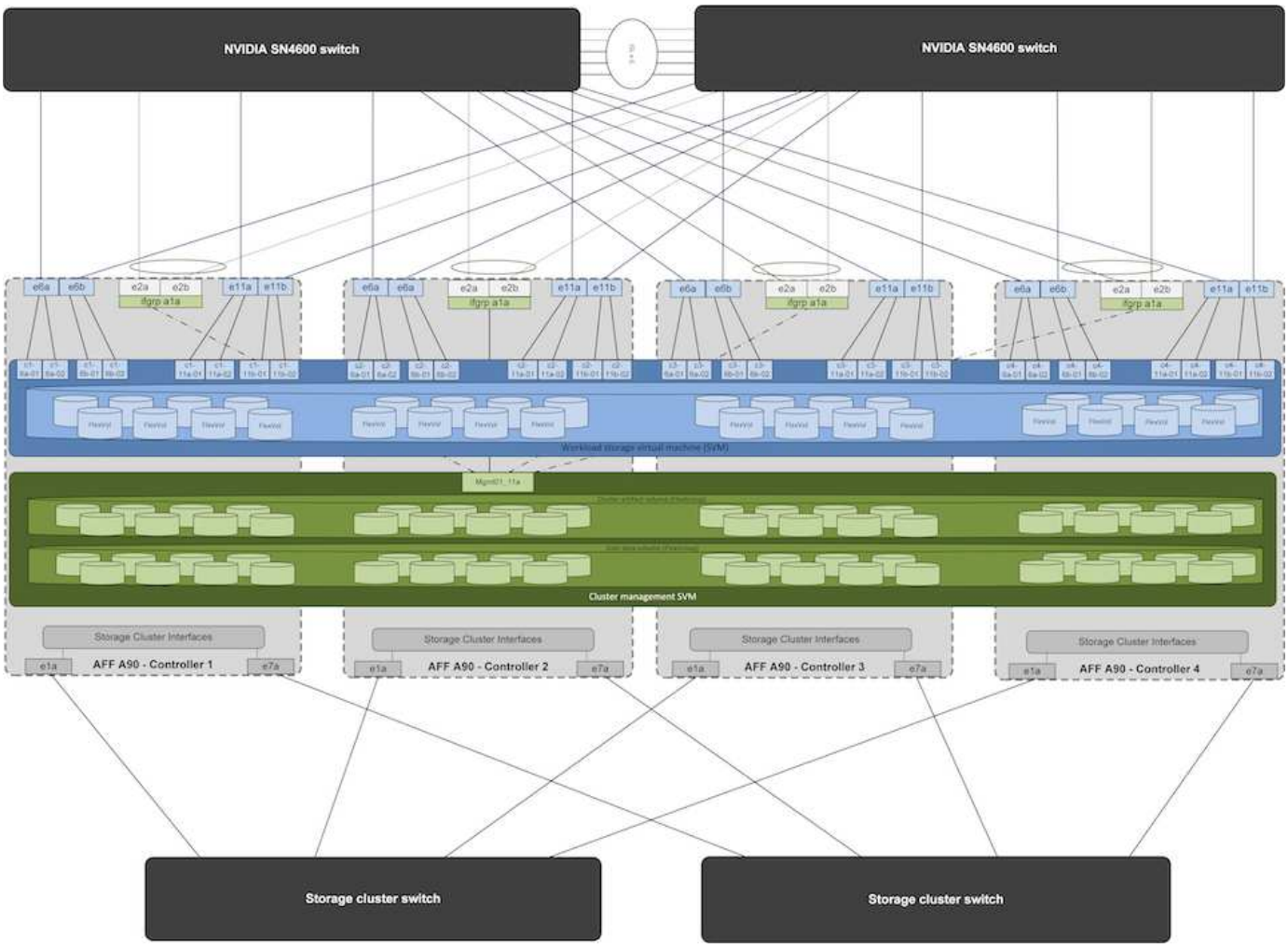
1. 配置交換器之間的 ISL 綁定以啟用多鏈路聚合 (MLAG) 和故障轉移流量
   - 本次驗證使用了 8 條鏈路，為測試的儲存配置提供了足夠的頻寬
   - 有關啟用 MLAG 的具體說明，請參閱 Cumulus Linux 文件。
2. 為兩台交換器上的每對用戶端連接埠和儲存連接埠配置 LACP MLAG
   - 每個交換器上的連接埠 swp17 用於 DGX-H100-01（enp170s0f1np1 和 enp41s0f1np1），連接埠 swp18 用於 DGX-H100-02，等等（bond1-16）
   - 每個交換器上的連接埠 swp41 用於AFF-A90-01（e2a 和 e2b），連接埠 swp42 用於AFF-A90-02，等等（bond17-20）
   - nv 設定介面 bondX 鍵成員 swpX
   - nv 設定介面 bondx 綁定 mlag id X
3. 將所有連接埠和 MLAG 綁定新增至預設橋接域
   - nv 設定 int swp1-16,33-40 橋接域 br_default
   - nv 設定 int bond1-20 橋接域 br_default
4. 在每台交換器上啟用 RoCE
   - nv 設定 roce 模式無損
5. 配置 VLAN - 2 個用於客戶端端口，2 個用於儲存端口，1 個用於管理，1 個用於 L3 交換器到交換機
   - 開關 1-
     - VLAN 3 用於在用戶端 NIC 發生故障時進行 L3 交換器到交換器的路由
     - 每個 DGX 系統上的儲存連接埠 1 的 VLAN 101（enp170s0f0np0，slot1 連接埠 1）

        - 每個AFF A90儲存控制器上的連接埠 e6a 和 e11a 的 VLAN 102
        - VLAN 301 用於使用 MLAG 介面對每個 DGX 系統和儲存控制器進行管理
    - 開關 2-
        - VLAN 3 用於在用戶端 NIC 發生故障時進行 L3 交換器到交換器的路由
        - 每個 DGX 系統上的儲存連接埠 2 的 VLAN 201（enp41s0f0np0，slot2 連接埠 1）
        - 每個AFF A90儲存控制器上的連接埠 e6b 和 e11b 的 VLAN 202
        - VLAN 301 用於使用 MLAG 介面對每個 DGX 系統和儲存控制器進行管理
6. 根據需要將實體連接埠指派給每個 VLAN，例如客戶端 VLAN 中的用戶端連接埠和儲存 VLAN 中的儲存連接埠
    - nv 設定 int <swpX> 橋接域 br_default 存取 <vlan id>
    - MLAG 連接埠應保持為中繼端口，以根據需要在綁定介面上啟用多個 VLAN。
7. 在每個 VLAN 上設定交換器虛擬介面 (SVI) 以充當網關並啟用 L3 路由
    - 開關 1-
        - nv 設定 int vlan3 ip 位址 100.127.0.0/31
        - nv 設定 int vlan101 ip 位址 100.127.101.1/24
        - nv 設定 int vlan102 ip 位址 100.127.102.1/24
    - 開關 2-
        - nv 設定 int vlan3 ip 位址 100.127.0.1/31
        - nv 設定 int vlan201 ip 位址 100.127.201.1/24
        - nv 設定 int vlan202 ip 位址 100.127.202.1/24
8. 建立靜態路由
    - 同一交換器上的子網路將自動建立靜態路由
    - 當客戶端連結發生故障時，交換器到交換器的路由需要額外的靜態路由
        - 開關 1-
            - nv 設定 VRF 預設路由器靜態 100.127.128.0/17 通過 100.127.0.1
        - 開關 2-
            - nv 設定 VRF 預設路由器靜態 100.127.0.0/17 透過 100.127.0.0

儲存系統配置

本節介紹此解決方案的 A90 儲存系統配置的關鍵細節。有關ONTAP系統配置的更多詳細信息，請參閱"ONTAP文件"。下圖顯示了儲存系統的邏輯配置。

NetApp A90 儲存叢集邏輯配置



配置儲存系統的基本步驟概述如下。此過程假設基本儲存叢集安裝已經完成。

1. 在每個控制器上配置 1 個聚合，所有可用分割區減去 1 個備用分割區
   - aggr create -node <節點> -aggregate <節點>_data01 -diskcount <47>
2. 在每個控制器上配置 ifgrps
   - 網路連接埠 ifgrp create -node <節點> -ifgrp a1a -mode multimode_lacp -distr-function port
   - 網路連接埠 ifgrp add-port -node <節點> -ifgrp <ifgrp> -ports <節點>:e2a,<節點>:e2b
3. 在每個控制器上的 ifgrp 上設定 mgmt vlan 端口
   - 網路連接埠 vlan 建立 -節點 aff-a90-01 -連接埠 a1a -vlan-id 31
   - 網路連接埠 vlan 建立 -節點 aff-a90-02 -連接埠 a1a -vlan-id 31
   - 網路連接埠 vlan 建立 -節點 aff-a90-03 -連接埠 a1a -vlan-id 31
   - 網路連接埠 vlan 建立 -節點 aff-a90-04 -連接埠 a1a -vlan-id 31
4. 建立廣播域
   - 廣播域創建-廣播域vlan21-mtu 9000-連接埠aff-a90-01：e6a，aff-a90-01：e11a，aff-a90-02：e6a，aff-a90-02：e11a，aff-a90-03：e6a，aff-a90-03：e11a，aff-a90-03：e6a，aff-a90-03：e11a，affa：96a-6a
   - 廣播域創建-廣播域vlan22-mtu 9000-埠aaff-a90-01：e6b，aff-a90-01：e11b，aff-a90-02：e6b，aff-

a90-02：e11b，aff-a90-03：e6b，aff-a90-02：e11b，aff-a90-03：e6b，aff-a90-03：e11baff-a90-03：e6b，aff-a90-03：e11b，affa

- 廣播域創建-廣播域vlan31-mtu 9000-端口aff-a90-01:a1a-31，aff-a90-02:a1a-31，aff-a90-03:a1a-31，aff-a90-04:a1a-31

5. 建立管理 SVM *
6. 配置管理 SVM
   - 創建 LIF
     - net int create -vserver basepod-mgmt -lif vlan31-01 -home-node aff-a90-01 -home-port a1a-31 -address 192.168.31.X -netmask 255.255.255.0
   - 創建FlexGroup磁碟區-
     - 卷創建-vserver basepod-mgmt-volume home-size 10T-auto-provision-as flexgroup-junction-path /home
     - 卷創建-vserver basepod-mgmt-volume cm-size 10T-auto-provision-as flexgroup-junction-path /cm
   - 制定出口政策
     - 匯出政策規則建立-vserver basepod-mgmt-policy default-client-match 192.168.31.0/24-rorule sys-rwrule sys-superuser sys
7. 建立資料SVM*
8. 配置資料 SVM
   - 配置 SVM 以支援 RDMA
     - vserver nfs 修改-vserver basepod-data -rdma 已啟用
   - 創建 LIF
     - net int create -vserver basepod-data -lif c1-6a-lif1 -home-node aff-a90-01 -home-port e6a -address 100.127.102.101 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-6a-lif2 -home-node aff-a90-01 -home-port e6a -address 100.127.102.102 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-6b-lif1 -home-node aff-a90-01 -home-port e6b -address 100.127.202.101 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-6b-lif2 -home-node aff-a90-01 -home-port e6b -address 100.127.202.102 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-11a-lif1 -home-node aff-a90-01 -home-port e11a -address 100.127.102.103 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-11a-lif2 -home-node aff-a90-01 -home-port e11a -address 100.127.102.104 -netmask 255.255.255.0
     - net int create-vserver basepod-data-lif c1-11b-lif1-home-node aff-a90-01-home-port e11b-address 100.127.202.103-netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c1-11b-lif2 -home-node aff-a90-01 -home-port e11b -address 100.127.202.104 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c2-6a-lif1 -home-node aff-a90-02 -home-port e6a -address 100.127.102.105 -netmask 255.255.255.0
     - net int create -vserver basepod-data -lif c2-6a-lif2 -home-node aff-a90-02 -home-port e6a -address 100.127.102.106 -netmask 255.255.255.0

- net int create -vserver basepod-data -lif c2-6b-lif1 -home-node aff-a90-02 -home-port e6b -address 100.127.202.105 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-6b-lif2 -home-node aff-a90-02 -home-port e6b -address 100.127.202.106 -netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-11a-lif1 -home-node aff-a90-02 -home-port e11a -address 100.127.102.107 -netmask 255.255.255.0
- net int create-vserver basepod-data-lif c2-11a-lif2-home-node aff-a90-02-home-port e11a-address 100.127.102.108-netmask 255.255.255.0
- net int create -vserver basepod-data -lif c2-11b-lif1 -home-node aff-a90-02 -home-port e11b -address 100.127.202.107 -netmask 255.255.255.0
- net int create-vserver basepod-data-lif c2-11b-lif2-home-node aff-a90-02-home-port e11b-address 100.127.202.108-netmask 255.255.255.0

9. 配置 LIF 以進行 RDMA 訪問
   - 對於使用ONTAP 9.15.1 的部署，實體資訊的 RoCE QoS 設定需要ONTAP CLI 中不可用的作業系統層級指令。請聯絡NetApp支援以取得 RoCE 支援連接埠配置的協助。 NFS over RDMA 功能正常
   - 從ONTAP 9.16.1 開始，實體介面將自動配置適當的設定以實現端對端 RoCE 支援。
   - net int 修改-vserver basepod-data -lif * -rdma-protocols roce
10. 在資料 SVM 上配置 NFS 參數
    - nfs 修改 -vserver basepod-data -v4.1 已啟用 -v4.1-pnfs 已啟用 -v4.1-trunking 已啟用 -tcp-max-transfer -size 262144
11. 創建FlexGroup卷
    - 卷創建-vserver basepod-data-volume資料-size 100T-auto-provision-as flexgroup-junction-path /data
12. 建立導出策略
    - 匯出政策規則建立-vserver basepod-data-policy default-client-match 100.127.101.0/24-rorule sys-rwrule sys-superuser sys
    - 匯出政策規則建立-vserver basepod-data-policy default-client-match 100.127.201.0/24-rorule sys-rwrule sys-superuser sys
13. 創建路線
    - 路由新增-vserver basepod_data-目的地100.127.0.0/17-網關100.127.102.1度量20
    - 路由新增-vserver basepod_data-目的地100.127.0.0/17-網關100.127.202.1度量30
    - 路由新增-vserver basepod_data-目的地100.127.128.0/17-網關100.127.202.1度量20
    - 路由新增-vserver basepod_data-目的地100.127.128.0/17-網關100.127.102.1度量30

用於 **RoCE** 儲存存取的 **DGX H100** 配置

本節介紹 DGX H100 系統配置的關鍵細節。許多配置項目可以包含在部署到 DGX 系統的 OS 映像中，或在啟動時由 Base Command Manager 實作。這裡列出它們以供參考，有關在 BCM 中配置節點和軟體映像的更多信息，請參閱"BCM 文件"。

1. 安裝其他軟體包
   - ipmitool

- python3-pip

2. 安裝 Python 套件
   - 波羅米科
   - matplotlib
3. 軟體包安裝後重新配置 dpkg
   - dpkg——配置-a
4. 安裝 MOFED
5. 設定 mst 值以進行效能調整
   - mstconfig -y -d <aa:00.0,29:00.0> 設定 ADVANCED_PCI_SETTINGS=1 NUM_OF_VFS=0 MAX_ACC_OUT_READ=44
6. 修改設定後重置適配器
   - mlxfwreset -d <aa:00.0,29:00.0> -y 重置
7. 在 PCI 裝置上設定 MaxReadReq
   - setpci -s <aa:00.0,29:00.0> 68.W=5957
8. 設定 RX 和 TX 環形緩衝區大小
   - ethtool -G <enp170s0f0np0,enp41s0f0np0> rx 8192 tx 8192
9. 使用 mlnx_qos 設定 PFC 和 DSCP
   - mlnx_qos -i <enp170s0f0np0,enp41s0f0np0> --pfc 0,0,0,1,0,0,0,0 --trust=dscp --cable_len=3
10. 為網路連接埠上的 RoCE 流量設定 ToS
    - echo 106 > /sys/class/infiniband/<mlx5_7,mlx5_1>/tc/1/traffic_class
11. 在適當的子網路上為每個儲存 NIC 設定一個 IP 位址
    - 100.127.101.0/24 用於儲存 NIC 1
    - 100.127.201.0/24 用於儲存 NIC 2
12. 配置帶內網路連接埠進行 LACP 綁定（enp170s0f1np1、enp41s0f1np1）
13. 為每個儲存子網路的主路徑和次路徑配置靜態路由
    - 路由新增 –net 100.127.0.0/17 gw 100.127.101.1 metric 20
    - 路由新增 –net 100.127.0.0/17 gw 100.127.201.1 metric 30
    - 路由新增 –net 100.127.128.0/17 gw 100.127.201.1 公制 20
    - 路由新增 –net 100.127.128.0/17 gw 100.127.101.1 公制 30
14. 掛載 /home 卷
    - 安裝-o vers = 3，nconnect = 16，rsize = 262144，wsize = 262144 192.168.31.X：/home /home
15. 掛載/資料卷
    - 安裝資料卷時使用了以下安裝選項-
      - vers=4.1 # 啟用 pNFS 來並行存取多個儲存節點
      - proto=rdma # 將傳輸協定設為 RDMA，而不是預設的 TCP

- max_connect=16 #啟用 NFS 會話中繼來聚合儲存連接埠頻寬
- write=eager # 提高緩衝寫入的寫入效能
- rsize=262144,wsize=262144 # 將 I/O 傳輸大小設為 256k

NVA-1173 NetApp AIPod與NVIDIA DGX 系統 - 解決方案驗證與規模調整指南

本節重點介紹採用NVIDIA DGX 系統的NetApp AIPod的解決方案驗證與尺寸調整指引。

解決方案驗證

使用開源工具 FIO 透過一系列合成工作負載驗證了此解決方案中的儲存配置。這些測試包括讀寫 I/O 模式，旨在模擬執行深度學習訓練作業的 DGX 系統產生的儲存工作負載。使用同時運行 FIO 工作負載的 2 插槽 CPU 伺服器叢集來驗證儲存配置，以模擬 DGX 系統叢集。每個客戶端都配置了前面描述的相同網路配置，並添加了以下詳細資訊。

以下安裝選項用於此驗證：

版本=4.1	啟用 pNFS 來並行存取多個儲存節點
原型=rdma	將傳輸協定設為 RDMA，而不是預設的 TCP
連接埠=20049	為 RDMA NFS 服務指定正確的連接埠
最大連線數=16	啟用 NFS 會話中繼來聚合儲存連接埠頻寬
寫=渴望	提高緩衝寫入的寫入效能
rsize=262144,wsize=262144	將 I/O 傳輸大小設定為 256k

此外，客戶端的 NFS max_session_slots 值配置為 1024。由於此解決方案是使用 NFS over RDMA 進行測試的，因此儲存網路連接埠配置了主動/被動結合。本次驗證使用了以下債券參數：

模式=主動備份	將綁定設定為主動/被動模式
primary=<介面名稱>	所有客戶端的主介面分佈在交換器上
mii-監控間隔=100	指定監控間隔為100ms
故障轉移 mac 策略=活動	指定活動鏈路的 MAC 位址是綁定的 MAC。這是 RDMA 透過綁定介面正確運行所必需的。

儲存系統配置如下，包括兩個 A900 HA 對（4 個控制器），每個 HA 對連接兩個 NS224 磁碟架，每個磁碟架有 24 個 1.9TB NVMe 磁碟機。如架構部分所述，所有控制器的儲存容量使用FlexGroup磁碟區進行組合，並且所有用戶端的資料分佈在叢集中的所有控制器上。

儲存系統規模指南

NetApp已成功完成 DGX BasePOD 認證，經測試的兩個 A90 HA 對可以輕鬆支援由 16 個 DGX H100 系統組成的叢集。對於具有更高儲存效能需求的大型部署，可以將額外的AFF系統新增至NetApp ONTAP叢集中，單一叢集中最多可包含 12 個 HA 對（24 個節點）。使用本解決方案中所述的FlexGroup技術，24 節點叢集可以在單一命名空間中提供超過 79 PB 和高達 552 GBps 的吞吐量。其他NetApp儲存系統（例如AFF A400、A250 和

C800）以較低的成本為較小規模的部署提供較低的效能和/或更高的容量選項。由於ONTAP 9 支援混合模型集群，客戶可以從較小的初始佔用空間開始，並隨著容量和效能需求的增長向集群添加更多或更大的儲存系統。下表粗略估計了每個AFF型號支援的 A100 和 H100 GPU 的數量。

NetApp 儲存系統規模調整指南

		Throughput[2]	Raw capacity *(typical[3] / max)*	Connectivity	# NVIDIA A100 GPUs supported[4]	# NVIDIA H100 GPUs supported[5]
NetApp® AFF A1K	1 HA pair[1]	56 GB/s	368TB / 14.7PB	200 GbE	1-160	1-80
	12 HA pairs	672 GB/s	**4.4PB / 176.4PB**		**1920**	**960**
AFF A90	1 HA pair	46 GB/s	368TB / 6.6PB	200 GbE	1 – 128	1-64
	12 HA pairs	552 GB/s	**4.4PB / 79.2PB**		**1536**	**768**
AFF A70	1 HA pair	21 GB/s	368TB / 6.6PB	200 GbE	1-48	1-24
	12 HA pairs	252 GB/s	**4.4PB / 79.2PB**		**576**	**288**

NVA-1173 NetApp AIPod與NVIDIA DGX 系統 - 結論及其他訊息

本節包含更多關於具有NVIDIA DGX 系統的NetApp AIPod的資訊的參考。

結論

DGX BasePOD 架構是下一代深度學習平台，需要同樣先進的儲存和資料管理功能。透過將 DGX BasePOD 與NetApp AFF系統結合， NetApp AIPod與 DGX 系統架構幾乎可以在任何規模上實現。結合NetApp ONTAP卓越的雲端整合和軟體定義功能， AFF可為成功的 DL 專案提供涵蓋邊緣、核心和雲端的全方位資料管道。

附加資訊

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 ONTAP資料管理軟體 — ONTAP資訊庫

  "https://docs.netapp.com/us-en/ontap-family/"

- NetApp AFF A90儲存系統-

  https://www.netapp.com/pdf.html?item=/media/7828-ds-3582-aff-a-series-ai-era.pdf

- NetApp ONTAP RDMA 資訊-

  "https://docs.netapp.com/us-en/ontap/nfs-rdma/index.html"

- NetApp DataOps 工具包

  "https://github.com/NetApp/netapp-dataops-toolkit"

- NetApp Trident

"概況"

- NetApp GPUDirect 儲存部落格-

  "https://www.netapp.com/blog/ontap-reaches-171-gpudirect-storage/"

- NVIDIA DGX BasePOD

  "https://www.nvidia.com/en-us/data-center/dgx-basepod/"

- NVIDIA DGX H100 系統

  "https://www.nvidia.com/en-us/data-center/dgx-h100/"

- NVIDIA網絡

  "https://www.nvidia.com/en-us/networking/"

- NVIDIA Magnum IO™ GPUDirect® 存儲

  "https://docs.nvidia.com/gpudirect-storage"

- NVIDIA基本指令

  "https://www.nvidia.com/en-us/data-center/base-command/"

- NVIDIA基礎指令管理器

  "https://www.nvidia.com/en-us/data-center/base-command/manager"

- NVIDIA AI 企業版

  "https://www.nvidia.com/en-us/data-center/products/ai-enterprise/"

致謝

本文檔由NetApp解決方案和ONTAP工程團隊（David Arnette、Olga Kornievskaia、Dustin Fischer、Srikanth Kaligotla、Mohit Kumar 和 Raghuram Sudhaakar）編寫。作者也要感謝NVIDIA和NVIDIA DGX BasePOD工程團隊的持續支持。