Lustre 搭配 NetApp E 系列儲存系統

NetApp artificial intelligence solutions

NetApp
August 31, 2026

目錄

Lustre 搭配 NetApp E 系列儲存系統

Lustre 與 NetApp E 系列儲存設備－簡介

Lustre 搭配 NetApp E 系列儲存設備，是一個經過驗證的平行檔案系統解決方案，適用於 AI 和 HPC 工作負載，由可重複的 HA 伺服器對和 E 系列 All Flash 陣列建置區塊組成。

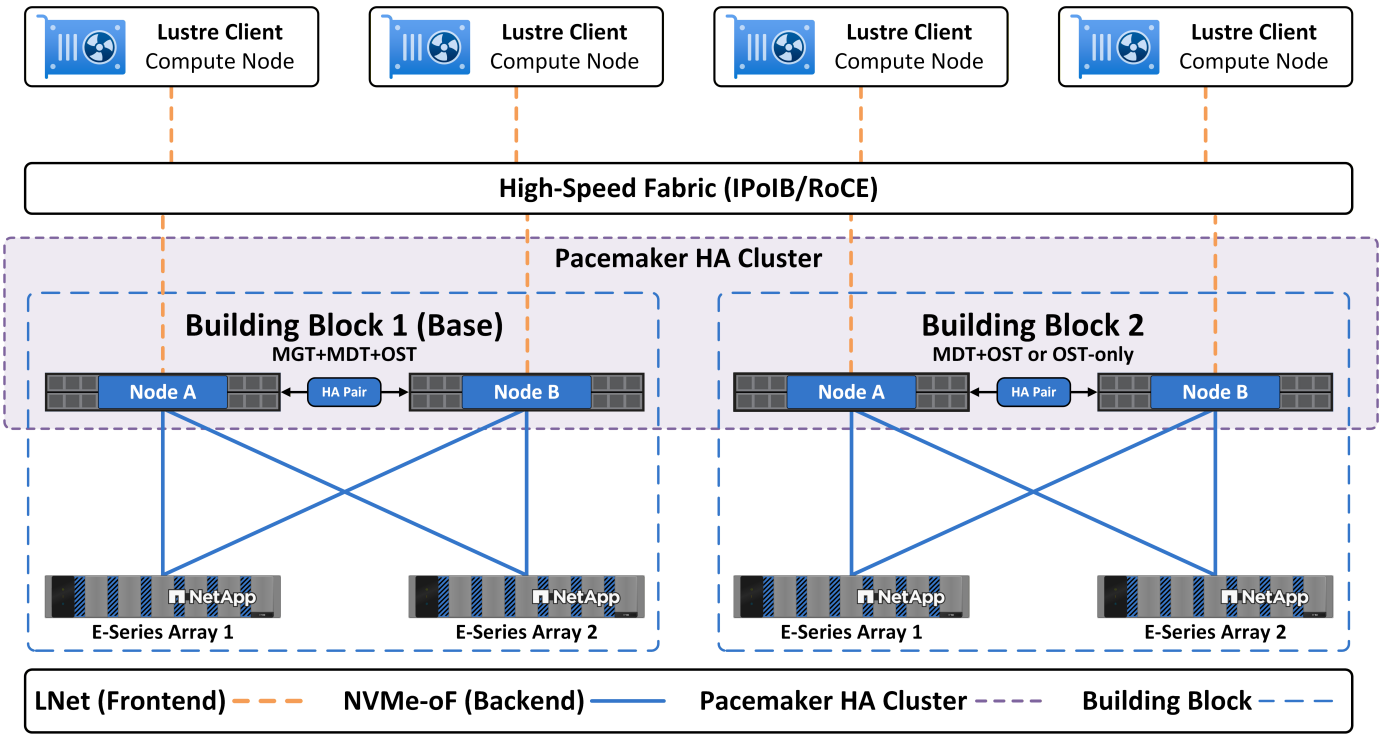
解決方案概述

Lustre 與 NetApp E 系列儲存設備結合，將開源的 Lustre 平行檔案系統與 NetApp E 系列區塊儲存設備整合，為資料密集型工作負載交付易擴充的高效能基礎架構。每個建置區塊包含兩個配置為高可用度 （HA） 的 OSS/MDS 伺服器節點和兩個 E 系列 All Flash 陣列。後端 NVMe-oF 連線將區塊 I/O 傳輸至陣列，而獨立的 LNet 架構則連接用戶端與 OSS/MDS 伺服器節點，以存取平行檔案系統。

基礎建置區塊承載管理伺服器 (MGS)、初始中繼資料目標 (MDT) 和物件儲存目標 (OST)。隨著工作負載需求的成長，其他建置區塊會向基礎 MGS 註冊，以擴充中繼資料容量、資料容量和彙總處理量。

下圖展示了一個範例部署，其中包含多個建置區塊和透過 LNet 連接的 Lustre 用戶端。

Lustre 搭配 NetApp E 系列儲存解決方案總覽



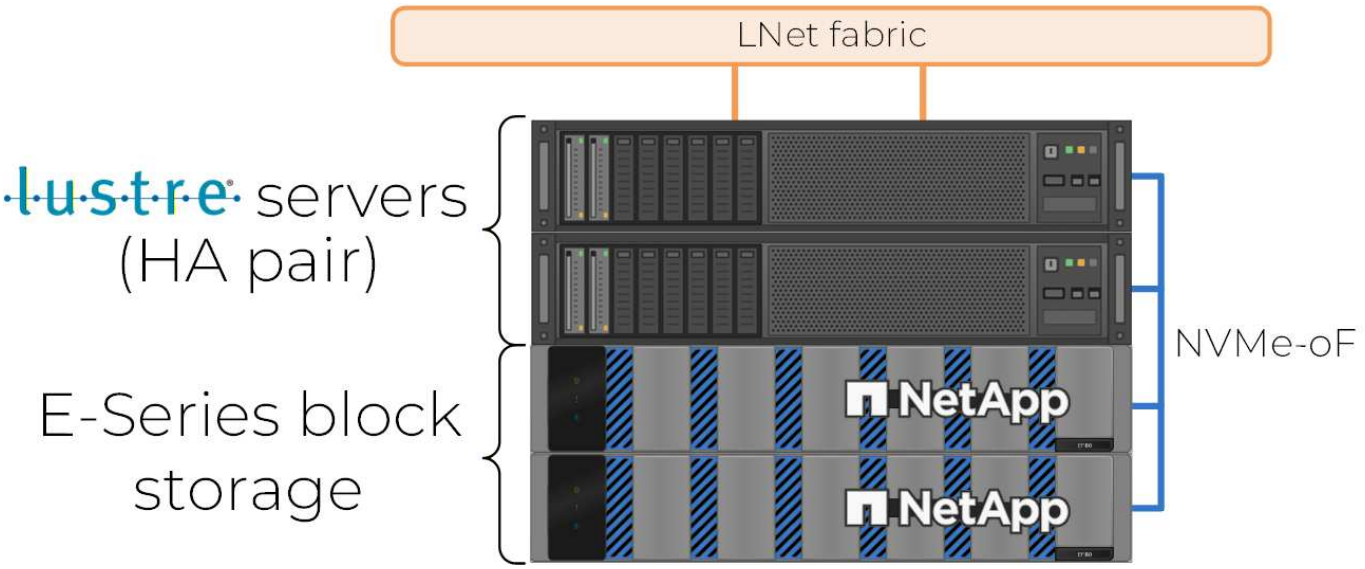
Lustre 搭配 NetApp E 系列儲存系統解決方案架構

了解 Lustre 與 NetApp E 系列儲存系統如何利用建置區塊、高可用度、網路拓撲和用戶端，以便您規劃容量、效能和容錯移轉。

建置區塊設計

建置區塊是 NetApp E 系列 Lustre 解決方案的基本易擴充單元。每個建置區塊包含兩個配置為高可用度 (HA) 配對的伺服器節點，提供 Lustre 物件儲存伺服器 (OSS) 和中繼資料伺服器 (MDS) 功能；兩個 NetApp E 系列 All Flash 陣列；伺服器與陣列之間的專用後端 NVMe over Fabrics (NVMe-oF) 連線；以及用於用戶端和節點間通訊的專用前端 Lustre 連網 (LNet) 連線。一個 Pacemaker 和 Corosync HA 叢集可包含來自一個或多個建置區塊的伺服器配對。此 NetApp `ansible-lustre`集合使用 Ansible 自動化來部署和設定 Lustre 服務。

Lustre 建置區塊



建置區塊可根據檔案系統的成長需求進行擴充。每個檔案系統都需要一個基礎建置區塊。基礎建置區塊在管理目標 (MGT) 上託管管理伺服器 (MGS)，並提供初始中繼資料目標 (MDT) 和物件儲存目標 (OST) 容量。其他建置區塊可擴充檔案系統，並將其 MDT 和 OST 目標註冊至基礎建置區塊的 MGS。這種模組化方法可讓容量和效能根據工作負載需求獨立擴充。

NetApp 對於大多數部署，建議使用以下建置區塊組態方案。列出的 MGS、MDT 和 OST 數量是建議的預設值，這些值已經過最佳化，可最大化 E 系列儲存陣列的效能和容量。您可以根據工作負載需求，在 Ansible 清單中調整目標數量、Volume 大小和 E 系列磁碟機分配。例如，需要更大中繼資料儲存容量的部署，可以在相同的建置區塊硬體中資源配置更多 MDT 和更少的 OST。

下表總結了建置區塊類型和建議的目標數量。

類型	**MGS**	**MDT**	**OST**	使用
基礎	1	8	32	檔案系統的第一個建置區塊。承載 MGS 以及初始 MDT 和 OST 容量。
MDT+OST	0	8	32	新增中繼資料和資料容量。註冊到基礎建置區塊 MGS。
僅限 OST	0	0	32	僅增加資料容量。已註冊到基礎建置區塊 MGS。

- 磁碟機類型和資源池佈局： E 系列支援傳統 RAID Volume 群組和動態磁碟資源池（DDP）。對於 TLC

NVMe 磁碟機，OST 儲存設備使用 RAID 6 Volume 群組，MGS 和 MDT 儲存設備使用 RAID 1 Volume 群組。對於 QLC NVMe 磁碟機，共享的磁碟機資源池使用 DDP。

- 目標經銷： 每個建置區塊中的 Lustre 目標在兩個 OSS/MDS 伺服器節點和兩個 E 系列陣列之間保持均衡分佈。這種佈局將 I/O 分散到伺服器 CPU 和陣列控制器上，以提高 NVMe-oF 路徑效能並維持備援。請參閱 "目標經銷和 NVMe-oF 連線功能" 中的 "硬體組件"。

支援的作業系統、Lustre 版本、陣列韌體和相關建置區塊元件列於"NetApp 互作業性矩陣工具（IMT）"的 **E** 系列 **Lustre** 下。如需詳細的 Volume 數量、磁碟機佈局和規模指南，請參閱"硬體組件"。如需軟體元件相關資訊，請參閱"軟體元件"。

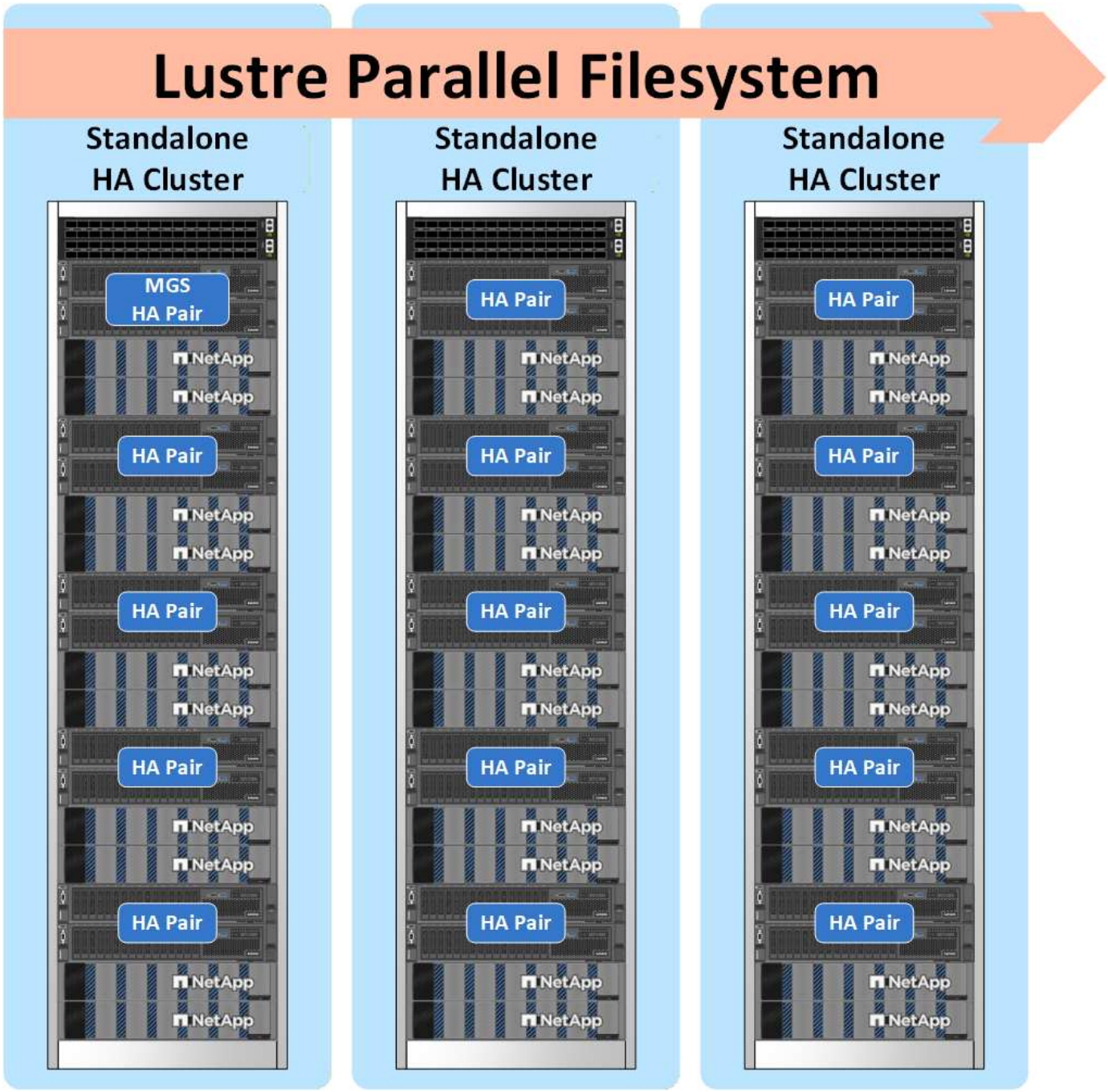
規模化建議

Lustre 檔案系統透過新增建置區塊來進行擴充，當中繼資料容量、資料容量或兩者都成為瓶頸時，即可進行擴充。根據檔案數量和中繼資料 IOPS 來擴充 MDT；根據容量和彙總處理量需求來擴充 OST。

1. 每個檔案系統一個基本建置區塊： 部署一個基本建置區塊；它承載 MGS 和初始 MDT 和 OST 目標。
2. 附加建置區塊設定檔： 當現有 MDT 容量充足且資料容量或處理量必須增加時，新增僅包含 OST 的建置區塊。當中繼資料效能或命名空間容量成為瓶頸時，新增包含 MDT 和 OST 的建置區塊。
3. **HA** 叢集限制： 將每個 Pacemaker 和 Corosync HA 叢集限制為五個建置區塊（十個 Lustre 伺服器節點）。將較大的部署平均拆分為多個 HA 叢集，以避免大型組態中的資源限制。

下圖展示了最多五個建置區塊堆疊在一個 42U 機架中的情況（每個機架一個 Pacemaker HA 叢集）。多個機架可以參與單一 Lustre 檔案系統；只有基礎建置區塊承載 MGS。

Lustre 平行檔案系統機架擴充



有關規模範例，請參見"尺寸指南"。

HA 架構

Lustre 與 NetApp E 系列儲存採用共享磁碟高可用度 （HA） 架構，並與 NetApp E 系列儲存系統整合。Pacemaker 管理 HA 資源，Corosync 提供叢集成員關係和郵件功能。它們共同管理每個建置區塊中兩個 OSS/MDS 伺服器節點之間的 Lustre 目標容錯移轉。多重路徑 NVMe-oF 將兩個 OSS/MDS 伺服器節點連接到相同的 E 系列 Volume，因此任何一個節點都可以在需要時接管目標服務。

每個 MGS、MDT 和 OST 都設定為具有相依性的 HA 資源，位於 Pacemaker 資源群組中。Pacemaker 可確保資源以正確的順序啟動和停止、保持共置於相同節點上，且一次只能在一個節點上執行。兩個節點同時存取相同目標可能會毀損底層檔案系統。

- 目標容錯移轉： 兩個 OSS/MDS 伺服器節點在叢集中互為對等節點，但每個 Lustre 目標在同一時間僅在一個節點上處於作用中。Pacemaker 監控資源會監控每個目標及其相依性的健全狀況，並在目標於其目前節點上無法使用時觸發容錯移轉。發生故障時，Pacemaker 會以正確的順序在合作夥伴節點上重新啟動受影響的資源群組，服務恢復後，用戶端會以透明方式重新連接至目標的新位置。由於每個目標僅在單一節點上啟動，因此檔案系統永遠不會面臨來自兩個節點的並行寫入。
- 共享儲存設備存取： 多重路徑 NVMe-oF 路徑將每個 OSS/MDS 伺服器節點連接到建置區塊中的所有 E 系列控制器，因此每個目標 Volume 都可以從任一節點存取。如果某個伺服器節點發生故障，其合作夥伴節點已擁有到相同 Volume 的作用中路徑，無需重新設定儲存設備連線即可接管目標的所有權。如果某個陣列控制器或路徑發生故障，多重路徑會將 I/O 重新導向至倖存的控制器。這種雙重備援使 Lustre 目標可在 OSS/MDS 伺服器節點或陣列控制器之間容錯移轉，而不會失去對後端 Volume 的存取。
- **STONITH** 隔離： 當發生故障時，Pacemaker 有時無法與故障節點通訊以確認其目標已停止。在從其他位置重新啟動這些目標之前，Pacemaker 會隔離故障節點（理想情況下是透過斷電），以確保其完全關閉。隔離功能可防止腦裂情況，即兩個節點同時存取相同目標並導致底層檔案系統毀損，從而在容錯移轉期間保持資料完整性。NetApp 建議 `fence_redfish`在配備支援 `Redfish` 的基板管理控制器 （BMC） 的伺服器上使用此功能。其他隔離代理，例如 `fence_apc`，也受支援。

有關叢集管理和隔離組態，請參閱 "HA 使用者指南" 中的 "ansible-lustre 集合"。

網路架構

此解決方案為儲存設備後端流量和 LNet 前端流量使用分隔的網路路徑。

- 後端 （**NVMe-oF**）： OSS/MDS 伺服器節點透過 NVMe-oF 使用 NVMe/InfiniBand 或 NVMe/RoCE 連接到 E 系列陣列。NVMe-oF 路徑承載 Lustre 目標服務和 E 系列 Volume 之間的區塊 I/O。有關 EF80 六 HCA 後端佈線，請參閱 "機架和纜線 Lustre 硬體"。有關跨節點和陣列的慣用目標放置位置，請參閱 "目標經銷" 中的 "硬體組件"。
- 前端（**LNet**）： Lustre 用戶端和 OSS/MDS 伺服器節點使用 `@o2ib`網路類型與 NVIDIA OFED 堆疊，透過 LNet 進行通訊。InfiniBand 和 RoCE 都是經過驗證的前端傳輸協定。ansible-lustre 集合會將所有前端介面設定為多軌 LNet，以彙總多個連接埠來提升頻寬，並為用戶端和節點間流量提供路徑備援。
- 管理和叢集： 帶外管理網路提供伺服器管理、BMC 存取和陣列管理。STONITH 隔離代理（例如 `fence_redfish`）使用此網路存取伺服器 BMC 並移除故障節點的電源。Corosync 也可以透過此網路作為前端架構以外的附加環來執行叢集通訊。設定具有多個環的 Corosync 可為叢集郵件增加備援，並降低單一網路故障導致仲裁中斷的風險。

Lustre 用戶端

Lustre 用戶端載入用戶端核心模組，建立與 OSS/MDS 伺服器節點的 LNet 連線，並裝載檔案系統，為應用程式提供單一、一致且符合 POSIX 標準的命名空間。I/O 以並行方式分散於各作用中的 OST。用戶端節點位於建置區塊外部，並透過前端 LNet 架構使用檔案系統。此解決方案的用戶端作業系統未列於 IMT 中；請使用 "Lustre 支援對照表" 以查看已部署 Lustre 版本所支援的已測試用戶端發行版及核心版本（例如，Red Hat Enterprise Linux 9、SUSE Linux Enterprise Server 15 和 Ubuntu 24.04）。

用戶端節點需要連線至與 OSS/MDS 伺服器節點相同的 LNet 架構和網路類型。如有需要，LNet 路由器可以橋接分隔的 LNet 子網路或網路架構。

NetApp 提供適用於 Rocky Linux 9.8 和 Red Hat Enterprise Linux 9.8 的預編譯 Lustre 伺服器 RPM 套件。NetApp 不提供預先編譯的用戶端軟體套件。請從 "netapp-lustre 儲存庫" 中的 Lustre 原始程式碼，為用戶端作業系統和核心建置用戶端軟體套件。

安裝用戶端套件後，"ansible-lustre 集合" 中的選用 `lustre_client` 角色會設定用戶端網路介面和 LNet、在選

取時啟用多軌 LNet、驗證連線功能，並管理持續性 systemd 裝載單元。此角色不會建置 Lustre。僅在填入 `lustre_client_packages` 時，它才會安裝用戶端套件。如果需要，可以手動設定用戶端。如需逐步程序，請參閱 "部署解決方案" 和 "lustre_client 角色文件"。

Lustre 與 NetApp E 系列儲存設備 - 硬體元件

當您使用 NetApp E 系列儲存設備調整 Lustre 規模及訂購時，請針對伺服器、NetApp E 系列陣列、儲存配置及連網使用這些硬體需求。有關軟體元件，請參閱 "軟體元件"。

伺服器要求

此解決方案採用自帶伺服器（BYOS）模式。每個建置區塊需要兩個符合或超過以下規格的 OSS/MDS 伺服器節點。

節點規格

下表列出每個 OSS/MDS 節點的最低伺服器要求。

成分	要求
數量	每個建置區塊包含 2 個節點
CPU	AMD EPYC 或 Intel Xeon，32 核心或更高
記憶	256 GB DDR5（最低配置）
網路 HCA	6 個雙埠 200Gb HCA（參見HCA 連線功能）
PCIe 插槽	2× PCIe Gen5 x16 和 4× PCIe Gen5 x8 （請參閱 PCIe 插槽要求）
啟動磁碟機	2 × 磁碟機採用 RAID 1（建議使用軟體或硬體 RAID）

NetApp 使用 Lenovo ThinkSystem SR665 V3 伺服器驗證了此解決方案。其他廠商的同等伺服器，只要符合這些要求，亦可支援。

HCA 連線功能

下表列出每個 OSS/MDS 伺服器節點的 HCA、LNet 前端連接埠和 NVMe-oF 路徑要求。

每個節點的 **HCA**	每個 **Lustre** 節點的 **LNet** 連接埠	每個 **Lustre** 節點的 **NVMe-oF** 路徑	範例 **HCA**
6（12 個連接埠）	4	總共 8 個（每個陣列 4 個）	MCX755106AS-HEAT（雙埠 200Gb PCIe gen5 卡）

NetApp 建議每個節點使用六個 HCA，以最大化 EF80 儲存設備陣列的頻寬。

PCIe 插槽要求

EF80 建置區塊中的每個 OSS/MDS 伺服器節點包含六個雙埠 HCA：兩個用於 LNet，四個用於 NVMe-oF。插槽寬度決定了每個 HCA 的可用頻寬，因此請確認每個插槽和轉接卡的電氣寬度，而不僅僅是卡片本身的寬度。安裝在 Gen5 x8 插槽中的 Gen5 x16 卡將以 x8 模式運作。

- **LNet HCAs**： 安裝在 PCIe Gen5 x16 插槽中，以達到完整的前端處理量。
- **NVMe-oF HCAs:** 安裝在 PCIe Gen5 x8 或 PCIe Gen5 x16 插槽中。
- **NUMA** 平衡： 將 HCA 平均分配到兩個 NUMA 區域，以便每個區域擁有兩個 LNet 介面和四個 NVMe-oF 介面。

下表列出 EF80 建置區塊中每個 OSS/MDS 伺服器節點的最小插槽數。

流量類型	每個節點的 **HCA**	最小插槽寬度	每個 **NUMA** 區域的插槽
LNet	2	PCIe Gen5 x16	1
NVMe-oF	4	PCIe Gen5 x8	2

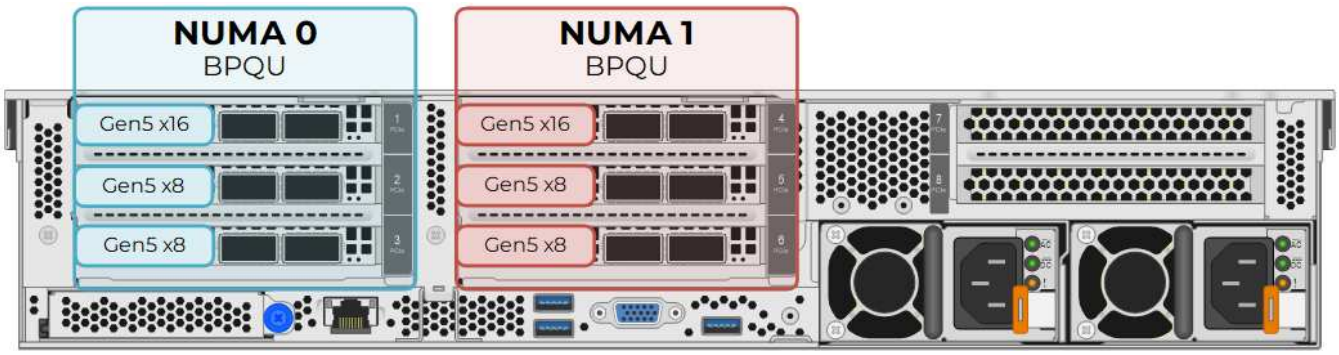
您也可以選擇將雙埠 HCA 上的連接埠分割，使一個連接埠承載 LNet 流量，另一個連接埠承載 NVMe-oF 流量。將所有四個 HCA 安裝在 PCIe Gen5 x16 插槽中，以確保每個 LNet 連接埠都能達到完整處理量，然後再新增兩個 HCA 安裝在 Gen5 x8 或 Gen5 x16 插槽中，用於剩餘的 NVMe-oF 連接埠。

▼ **Lenovo ThinkSystem SR665 V3** 的擴充座配置範例

以下兩種立管組合均符合 EF80 建置區塊的插槽要求。

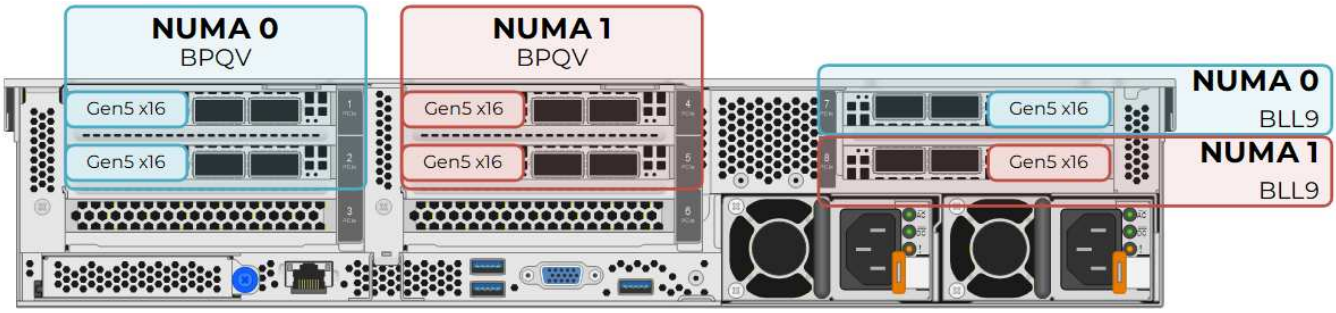
最小插槽寬度

在轉接卡位置 1 和 2 安裝 BPQU 轉接卡。每個 BPQU 轉接卡提供一個 PCIe Gen5 x16 插槽和兩個 PCIe Gen5 x8 插槽。這些轉接卡總共提供兩個用於 LNet 的 Gen5 x16 插槽和四個用於 NVMe-oF 的 Gen5 x8 插槽，均勻分佈在 NUMA 區域內。



所有第五代 **x16** 插槽

在擴充卡插槽 1 和 2 中安裝 BPQV 擴充卡，在擴充卡插槽 3 中安裝 BLL9 擴充卡。每個 BPQV 擴充卡提供兩個 PCIe Gen5 x16 插槽。BLL9 擴充卡提供 NUMA 區域 0 的插槽 7 和 NUMA 區域 1 的插槽 8，這兩個插槽均為 PCIe Gen5 x16。此組合提供六個 Gen5 x16 插槽，每個 NUMA 區域三個，並支援在同一 HCA 的各個連接埠之間分割 LNet 和 NVMe-oF 流量。



EF80 標準建置區塊

EF80 是此解決方案版本經過驗證的標準平台。

陣列規格

下表列出建置區塊（兩個陣列）的 EF80 陣列規格。

成分	規格
模型	NetApp EF80
外形尺寸	2U 基本機箱，24 個內部 NVMe SSD 插槽
控制器	雙控制器（A 和 B）
磁碟機	每個陣列配備 24 個 NVMe SSD
I/O 連接	在經過驗證的 Lustre 設計中，每個陣列配備 8 個 200Gb NVMe/IB 或 NVMe/RoCE 主機連接埠

EF80 平台在每個控制器中安裝三個雙埠主機 I/O 模組後，每個陣列最多支援十二個主機連接埠。經過驗證的 Lustre 設計在插槽 1 和 2 中使用主機 I/O 模組，每個陣列可實現八個主機連接埠。

每個 EF80 陣列也使用專用的插槽 4 I/O 模組進行控制器間鏡射。將控制器 A 的連接埠 4a 連接到控制器 B 的連接埠 4a，並將控制器 A 的連接埠 4b 連接到控制器 B 的連接埠 4b。這些連線提供快取鏡射和 I/O 傳輸，不用於主機 NVMe-oF 流量。請參閱 "連接 EF50 和 EF80 控制器間鏡射連接線"。

有關所有型號的完整 EF 系列規格，請參閱"NetApp EF 系列 All Flash Array 產品型錄"。

磁碟機佈局（每個陣列 **24** 個磁碟機）

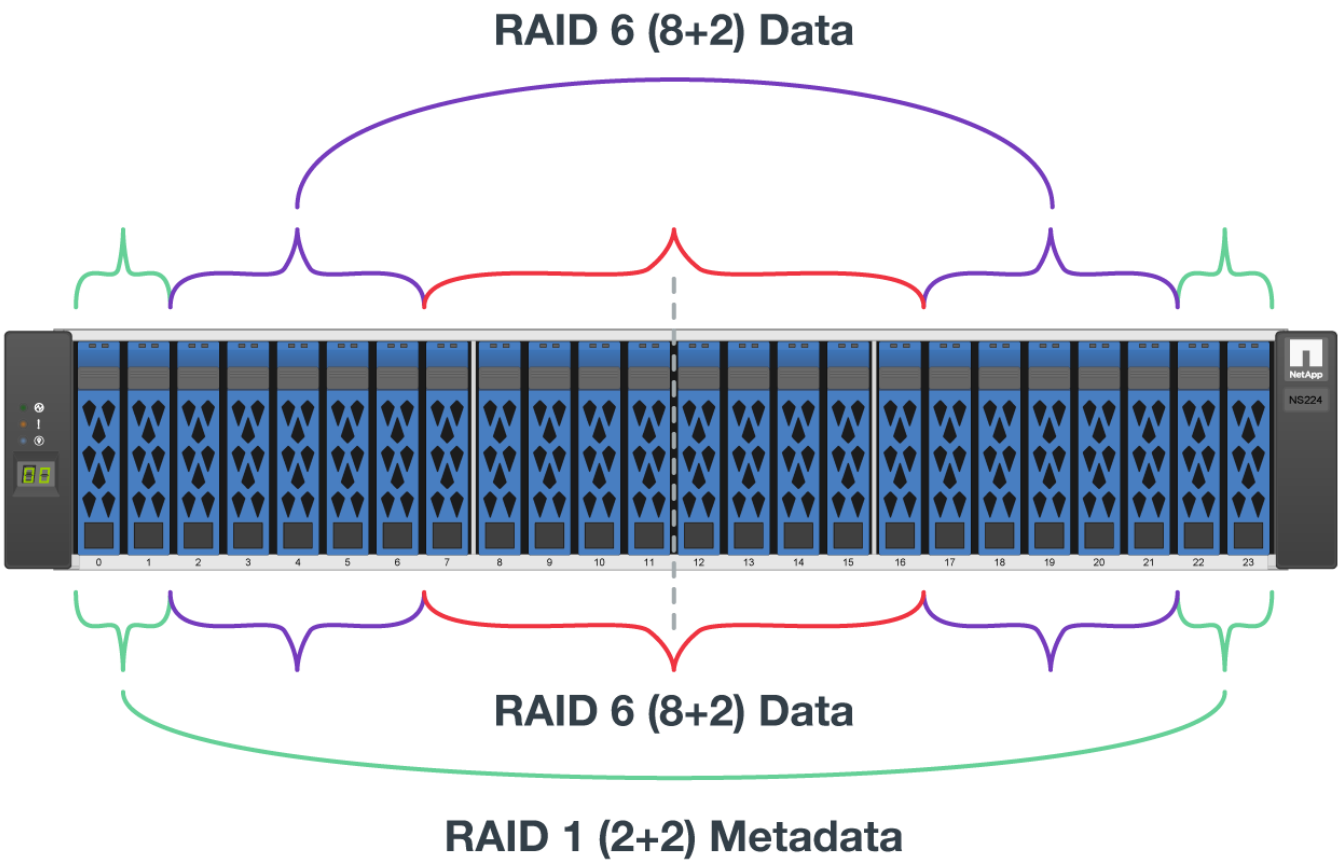
基本建置區塊中的每個 EF80 陣列都使用全部 24 個 NVMe 磁碟機插槽：

- 4 個磁碟機組成 RAID 1，用於 MGS/MDT 儲存設備（共享的 Volume 群組或 DDP 分配）
- 10 個磁碟機組成 RAID 6 陣列，用於 OST 儲存設備（第一個 OST 資源池）
- 10 個磁碟機組成 RAID 6 陣列，用於 OST 儲存設備（第二個 OST 資源池）

此版面配置適用於使用 3.84 TB、7.68 TB 或 15.3 TB 磁碟機並採用傳統 Volume 群組的情況。當使用 30.7 TB 或 61.4 TB Capacity Flash （QLC） 磁碟機時，請在所有 24 個磁碟機上資源配置單一動態磁碟資源池，並在資源池中為 MGS 和 MDT 磁碟區建立 RAID 1 磁碟區，同時為 OST 使用預設的 RAID 6 磁碟區。

目前不建議在此解決方案中使用 1.92 TB 的磁碟機。請使用上面列出的經過驗證的磁碟機容量之

—。



Volume 與目標計數（基礎建置區塊）

下表列出了基礎建置區塊的 MGS、MDT 和 OST 數量。

目標類型	每個 **BB** 的數量	**RAID /** 資源池	附註
MGS	1	RAID 1	僅基礎建置區塊；陣列 1
MDT	8	RAID 1	每個陣列 4 個
OST	32	RAID 6 (TLC) 或 DDP (QLC)	每個陣列 16 個

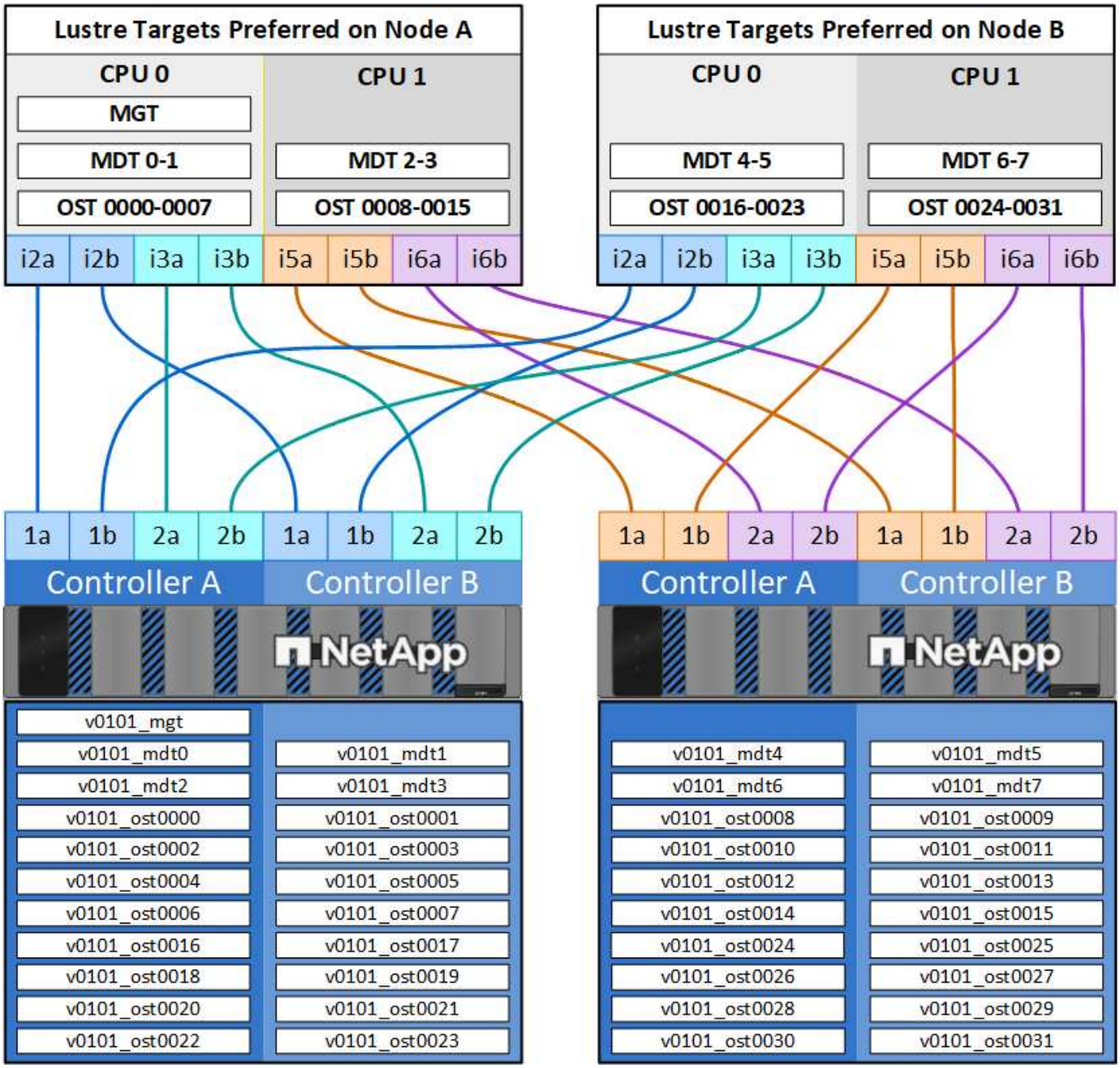
在 Ansible 清單中，請使用以下 Volume 規模調整指南作為起點。

體積	建議大小	附註
MGS	5–10 GiB	僅組態資料
MDT	RAID 1 容量 ÷ MDT 數量	每顆約含 1–2 個 TiB（典型值）
OST	RAID 6 或 DDP 容量 ÷ 每個資源池的 OST 數量	隨磁碟機容量而變化；參見"尺寸指南"

主要建置區塊 **Volume** 分佈

下圖展示了在基礎 EF80 建置區塊中，Lustre 目標的首選放置位置以及 OSS/MDS 伺服器節點和 E 系列陣列之間的 NVMe-oF 連接。圖中將管理目標標記為 **MGT**（管理目標）；一個 MGT 承載本文件其他部分提及的 MGS（管理伺服器）服務。

基礎建置區塊目標經銷（**EF80**）



下表列出 Volume 在兩個陣列中的分佈情況，以及每個目標偏好的伺服器。

目標類型	陣列 **1**	陣列 **2**	伺服器 **1**	伺服器 **2**	附註
MGS	1	0	1	0	僅限基礎建置區塊
MDT	4	4	4	4	每個伺服器優先選擇來自每個陣列的 2 個 MDT

目標類型	陣列 **1**	陣列 **2**	伺服器 **1**	伺服器 **2**	附註
OST	16	16	16	16	每個 RAID 6 資源池 8 個 × 每個陣列 2 個資源池，或等效的 DDP 分配
Volume 總計	**21**	**20**	**21**	**20**	

儲存資源池選擇：**TLC** 與 **QLC**

根據陣列中 NVMe 磁碟機的容量選擇 E 系列資源池類型：

- **3.84 TB**、**7.68 TB** 和 **15.3 TB** 磁碟機： OST Volume 使用 **RAID 6 Volume** 群組，MGS 和 MDT Volume 使用 **RAID 1 Volume** 群組，或共享磁碟機資源池使用 DDP。
- **30.7 TB** 和 **61.4 TB** 容量的 **Flash** （**QLC**） 磁碟機： 僅使用 動態磁碟資源池 （**DDP**） 。在 DDP 內為 MGS 和 MDT 儲存設備建立 RAID 1 Volume。從 DDP 為 OST 儲存設備建立 RAID 6 Volume。

有關按磁碟機大小和佈局劃分的每個建置區塊可用容量估算，請參閱"尺寸指南"。

網路需求

後端（**NVMe-oF**）

- NVMe/InfiniBand 或 NVMe/RoCE，位於每個 OSS/MDS 節點與每個 E 系列陣列之間
- NVMe/RoCE 後端介面上的 MTU 9000（典型）
- 每個 Lustre 節點到儲存設備陣列有八條路徑（每個陣列四條路徑）
- EF80 每個節點使用六個 HCA（NVMe-oF 使用 i2、i3、i5 和 i6；LNet 使用 i1 和 i4） 。將節點 A 連接到標籤結尾為 `a`的控制器連接埠，例如 `1a`和 `2a`。將節點 B 連接到標籤結尾為 `b`的控制器連接埠，例如 `1b`和 `2b`。請參閱"EF80 六 HCA 後端佈線"。

前端（**LNet**）

- IPoIB 或 RoCE（適用於 LNet (`@o2ib` 網路類型）
- RoCE 前端介面的 MTU 為 9000（典型值）
- 當有四個前端連接埠可用時，建議使用多軌 LNet（EF80：i1 和 i4 HCA，兩個連接埠）
- 專用 InfiniBand 或 RoCE 交換器架構，連接 OSS/MDS 伺服器節點和 Lustre 用戶端
- 建議在 RoCE 架構上使用無損 RoCE （PFC） 。

管理

- 伺服器 BMC 和陣列管理連接埠的頻外管理網路
- 專用 Corosync 網路或共享的前端架構（取決於站台）

Lustre 搭配 **NetApp E** 系列儲存系統 **-** 軟體元件

查看用於部署 Lustre 與 NetApp E 系列儲存設備的軟體堆疊和 Ansible 自動化。有關伺服

器、儲存設備和網路的資訊，請參閱 "硬體組件"。

建置區塊軟體

下表列出 E 系列陣列和 OSS/MDS 伺服器節點所使用的軟體元件。請以 "NetApp 互作業性矩陣工具（IMT）" 作為官方來源，確認支援的版本和元件組合。部署前，請搜尋 **E-Series Lustre**，並確認作業系統、Lustre 、SANtricity 作業系統、DOCA-OFED、HCA 韌體和 HA 套件的版本。

成分	功能
SANtricity OS	在每個 E 系列陣列上執行，並提供 SANtricity System Manager、雙主動控制器高可用度（具備自動 I/O 路徑故障移轉）、自動負載平衡和路徑監控、動態磁碟資源池和傳統 RAID、自動磁碟機重建、鏡像快取保護、Data Assurance、主動式磁碟機健全狀況監控以及線上軟體、韌體和組態變更。
Lustre	為用戶端提供單一符合 POSIX 標準的命名空間，同時將中繼資料和檔案資料分散至多個目標，以實現並行存取。管理伺服器 （MGS） 儲存檔案系統組態，中繼資料目標 （MDT） 管理命名空間，物件儲存目標 （OST） 將檔案資料儲存在 E 系列 Volume 上。此分離式設計可透過新增建置區塊來擴充效能和容量。
Red Hat Enterprise Linux (RHEL) 或 Rocky Linux	為 Lustre 目標服務、高可用度叢集以及儲存設備和網路連線功能提供核心、系統服務和套件架構。支援的作業系統儲存庫亦提供 Pacemaker、Corosync、fence 代理程式、NVMe-oF 和多重路徑套件，這些套件已作為解決方案堆疊共同進行測試。
高可用度叢集服務（ Pacemaker、Corosync 和 fence 代理程式）	這些服務協同工作，為 OSS/MDS 伺服器節點上的 Lustre 目標提供協調一致的高可用度。它們維護叢集成員關係和仲裁，監控資源，協調有序故障轉移，並在啟動其他位置的共享目標之前隔離無回應的節點。這種協調機制可防止對 E 系列 Volume 的腦裂存取，並在故障期間保護檔案系統的完整性。
NVIDIA DOCA-OFED	為 InfiniBand 和 RoCE 架構提供 HCA 驅動程式、RDMA 程式庫及管理公用程式。同一軟體堆疊支援對 E 系列陣列的低延遲後端 NVMe-oF 存取，以及與 Lustre 用戶端的高處理量前端 LNet 通訊。

NetApp 為 Rocky Linux 9.8 和 Red Hat Enterprise Linux 9.8 提供預先建置的 Lustre 伺服器 RPM。

Ansible 自動化

Lustre 搭配 NetApp E 系列儲存系統的部署與交付，是透過 Ansible 自動化從具備對 E 系列陣列及 OSS/MDS 伺服器節點管理存取權的控制節點來實現。Ansible 清單對所需的解決方案進行建模，包括陣列、伺服器、儲存設備分配、網路介面和高可用度叢集資源。

NetApp E 系列 Lustre Ansible 集合利用 SANtricity 和 Host 集合協調端對端部署。這些集合協同運作，為 E 系列儲存設備進行資源配置和映射、準備伺服器作業系統、設定 NVMe-oF 和 LNet 連線、部署 Lustre 目標並建立 Pacemaker 資源。此自動化功能使組態可重複執行，並簡化了在現有檔案系統中新增建置區塊的程序。

下表總結了 Ansible 控制節點上使用的主要軟體套件。如需目前的軟體套件相依性，請參閱 "NetApp E 系列 Lustre Ansible collection"。

套件	依賴或目的
Python	為 Ansible 和所需的 Python 模組提供執行環境。
ansible-core	需要 Python，並執行 NetApp E 系列 Ansible 集合。

套件	依賴或目的
其他 Python 套件	`cryptography`， `ipaddr`， `netaddr`， `passlib`， `resolvelib`， `jinja2`，和 `pyyaml`
NetApp E 系列 Lustre Ansible collection	設定 Lustre、LNet、NVMe-oF 主機連線功能和 Pacemaker 資源。
NetApp E 系列 SANtricity Ansible 集合	透過 SANtricity Web Services API 設定 E 系列陣列、儲存設備資源池、Volume 及主機對應。
NetApp E 系列主機 Ansible 集合	為 OSS/MDS 伺服器節點準備 LNet 和 NVMe-oF 連線至 E 系列陣列。

Lustre 用戶端軟體

Lustre 用戶端軟體使系統能夠透過前端 LNet 架構存取平行檔案系統。每個用戶端都需要與其執行中的核心以及伺服器上所部署的 Lustre 版本相容的 Lustre 核心模組和 Lustre 用戶端公用程式。

使用 "Lustre 支援對照表" 尋找與您的 Lustre 版本相容的用戶端作業系統和核心。NetApp 不提供預先編譯的 Lustre 用戶端軟體套件。請從 "netapp-lustre 儲存庫" 中 NetApp 提供的 Lustre 原始碼，為用戶端作業系統和核心建立用戶端軟體套件。安裝 Lustre 用戶端軟體套件後， `lustre_client` 角色可以設定網路介面、LNet，並為 Lustre 檔案系統建立持久的 systemd 裝載單元。

有關用戶端部署步驟，請參閱"部署解決方案"。

Lustre 搭配 NetApp E 系列儲存設備 - 規模調整指引

使用 NetApp E 系列儲存建置區塊，根據容量和中繼資料需求來調整 Lustre 的規模，決定何時新增容量，並審查影響效能的因素。

容量規模調整

一個包含兩個 EF80 陣列的基本建置區塊提供了以下 Lustre 目標佈局。如需偏好的目標放置和 NVMe-oF 路徑，請參閱 "主要建置區塊 Volume 分佈" 中的 "硬體組件"。

成分	計數	典型 **Volume** 大小（每個目標）
MGS	1	5-10 GiB
MDT	8	大小因磁碟機容量和 RAID 佈局而異
OST	32	大小因磁碟機容量和 RAID 佈局而異

每個 EF80 陣列使用 24 個 NVMe 磁碟機。支援的容量包括 3.84 TB、7.68 TB 和 15.3 TB，支援傳統 Volume 群組（MGS/MDT 使用 RAID 1，OST 使用 RAID 6）或 DDP；以及 30.7 TB 或 61.4 TB 容量的 Flash（QLC）磁碟機，僅支援 DDP。目前不建議在此解決方案中使用 1.92 TB 磁碟機。有關磁碟機插槽佈局和資源池選擇，請參閱"硬體組件"。

下表列出了 EF80 建置區塊（兩個陣列，48 個磁碟機）的近似可用容量。陣列可用容量與最終 Lustre 檔案系統可用容量不同。MDT inode 分配、日誌、過度資源配置和庫存 Volume 百分比都會減少應用程式可用的容量。

磁碟機容量	佈局（每個陣列）	近似可用容量（建置區塊）
3.84 TB	RAID 1 (2+2) + RAID 6 (10) + RAID 6 (10)	138.08 TB
3.84 TB	DDP（24 個磁碟機，2 個預留）	133.63 TB
7.68 TB	RAID 1 (2+2) + RAID 6 (10) + RAID 6 (10)	276.35 TB
7.68 TB	DDP（24）	267.47 TB
15.3 TB	RAID 1 (2+2) + RAID 6 (10) + RAID 6 (10)	552.88 TB
15.3 TB	DDP（24）	535.15 TB
30.7 TB	僅限 DDP	1,070.35 TB
61.4 TB	僅限 DDP	2,162.70 TB

將中繼資料和資料分別設定大小，然後在 Ansible 清單中設定 Volume 大小。部署範本使用 ldiskfs 格式化目標；MDT 使用 Lustre 預設的 inode 比率，並設定 OST 範本 `-i 4096`。

- 中繼資料 **(MDT)**： 根據預期檔案數量確定 MDT 的規模。預留大約兩倍於預期檔案數量的成長空間。MDT 容量過小會導致即使 OST 容量充足，也無法建立新檔案。
- *資料（OST）：*根據可用容量和處理量需求決定 OST 的大小。
- **MGS:** 僅使用較小的管理目標（5-10 GiB）進行檔案系統組態。

根據所選磁碟機容量調整清單中的 MDT 和 OST Volume 大小。僅當站台需要不同的 inode 比率時，才在 `format_options.mkfsoptions`中 `eseries_lustre_filesystem_mdt`或 `eseries_lustre_filesystem_ost`中進行變更。有關 inode 比率的背景資訊，請參閱"Lustre Wiki：Lustre 調校"。有關何時新增建置區塊，請參閱[規模化]。

規模化

當容量或中繼資料負載需要時，新增建置區塊：

- 僅限 **OST**： 檔案數量和中繼資料負載已在現有 MDT 容量範圍內，但您需要更大的資料容量或彙總處理量。此方案新增 32 個 OST 和兩個 OSS/MDS 伺服器節點。
- **MDT+OST**： 檔案數量或中繼資料速率接近 MDT 容量上限，或您需要更高的中繼資料處理量和資料容量。增加 8 個 MDT 和 32 個 OST。

在現有 MDT 上使用 `mdt.*.md_stats`，以確認中繼資料是否已飽和。每個 Pacemaker/Corosync 叢集最多只能包含五個建置區塊（十個 OSS/MDS 伺服器節點）。對於更大規模的部署，請將建置區塊平均分配至多個 HA 叢集。請參閱"解決方案架構"以了解擴充規則。

以下範例展示了一個 HA 叢集中具有基本建置區塊和僅 OST 建置區塊的檔案系統。

建置區塊	類型	伺服器	陣列	**MGS**	**MDT**	**OST**
BB1	基礎	2	2	1	8	32
BB2	僅限 OST	2	2	0	0	32
總計		4	4	1	8	64

BB2 使用 `mgsnode=`將其 OST 目標註冊到 BB1 中的 MGS。BB2 的 OST 索引是全域分配的（例如，OST0032 到 OST0063），因此沒有索引與 BB1 衝突。

表現

正式驗證使用 Lustre 用戶端的 IOR、mdtest 和 fio 來表徵相對處理量、IOPS 和中繼資料行為，並驗證高可用度容錯移轉。這些測試不提供效能保證數據。綜合基準測試衡量的是最佳情況下的效能，可能無法反映實際應用程式效能。

效能大致與建置區塊的數量成正比。實際結果取決於磁碟機類型和資源池佈局、NVMe-oF 路徑數量、LNet 架構和用戶端數量，以及資料 I/O 與中繼資料 I/O 的比例。

請聯絡您的 NetApp 帳戶團隊，以取得特定工作負載的規模和效能建議。

有關磁碟機佈局和 Volume 統計，請參閱 "硬體組件"。有關部署步驟，請參閱 "部署解決方案"。有關現場級庫存指南，請參閱 "自訂 Lustre Ansible 庫存"。

部署解決方案

將 **Lustre** 與 **NetApp E** 系列儲存設備搭配部署

依序完成硬體、環境準備、清單、Ansible 部署和驗證任務，即可透過 NetApp E 系列儲存設備部署 Lustre。

先決條件

在開始部署之前，請確保以下事項：

- 您遵循了"尺寸指南"中的規模和建置區塊指南，並選擇了符合"硬體組件"的硬體。
- 您已確認伺服器、HCA、E 系列、作業系統、Lustre、SANtricity 作業系統和 DOCA-OFED 組合已列於 "NetApp 互通性矩陣工具" 的 **E-Series Lustre** 清單中。
- 計畫建置區塊的所有 OSS/MDS 伺服器節點和 E 系列陣列均已在現場，可安裝到機架。
- 您擁有伺服器 BMC、陣列管理連接埠以及將作為 Ansible 控制節點的主機的憑證和網路存取。

> NetApp 支援使用 "NetApp E 系列 Lustre Ansible collection" 部署 Lustre HA 叢集和 Lustre 用戶端。請勿手動設定 Lustre 目標、LNet、NVMe-oF 主機連線功能或 Pacemaker 資源。

部署工作流程

請依序完成以下任務：

1. "機架和纜線 Lustre 硬體".

   將 Lustre 伺服器和 EF80 陣列安裝到機架上，然後連接後端 NVMe-oF 和前端 LNet 架構。

2. "準備 Lustre 部署環境".

   設定陣列和伺服器，然後在控制節點上安裝 Ansible 及其相依性。

3. "自訂 Lustre Ansible 庫存".

   選擇站台架構模板，並填入伺服器、陣列、網路、建置區塊和憑證。

4. "部署 Lustre HA 叢集和用戶端".

   準備伺服器作業系統，安裝 NVIDIA DOCA-OFED，執行主要部署 playbook，並設定 Lustre 用戶端。

5. "驗證並擴充 Lustre 部署".

   生產環境使用前，請確認叢集健全狀況、掛載點和容量。當檔案系統需要其他建置區塊時，請擴充現有清單。

相關資訊

- "NetApp E 系列 Lustre Ansible collection"
- "NetApp Lustre 來源儲存庫"
- "HA 管理使用者指南"
- "效能調校指南"
- "Lustre 用戶端部署範本"

機架和纜線 **Lustre** 硬體

將 Lustre 伺服器和 NetApp EF80 陣列安裝到機架上，然後為每個建置區塊連接後端 NVMe-oF 和前端 LNet 架構。

開始之前

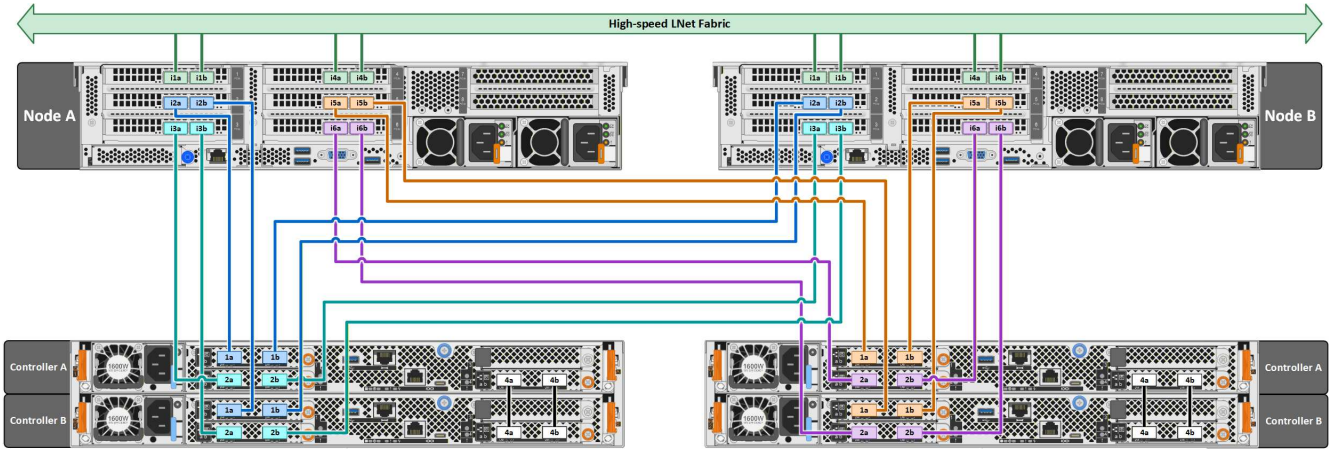
查看 "硬體組件" 中的 HCA 計數、PCIe 插槽要求、路徑要求和 MTU 指南。

步驟

1. 根據建置區塊拓撲，將所有 Lustre 伺服器和 E 系列儲存陣列安裝到機架上並連接電纜。

2. 對於每個 EF80 陣列，將控制器 A 的連接埠 4a 連接到控制器 B 的連接埠 4a，並將控制器 A 的連接埠 4b 連接到控制器 B 的連接埠 4b。這些專用連接提供控制器間的鏡射，不用於主機流量。參見"連接 EF50 和 EF80 控制器間鏡射連接線"。
3. 根據 EF80 六 HCA 拓撲，在伺服器和陣列之間連接後端 NVMe-oF 網路。

EF80 六 HCA 後端佈線（RoCE 或 InfiniBand）



4. 將每台 Lustre 伺服器和 Lustre 用戶端上的四個前端 LNet 介面（i1 和 i4 HCA）連接到專用 InfiniBand 或 RoCE 交換器架構。
5. 使用 RoCE 時，請將架構設定為無損操作並啟用優先權流程控制 （PFC）。Ansible playbook 會在建置區塊中的所有介面上設定無損 RoCE。

完成後

"準備伺服器、陣列和 Ansible 控制節點".

準備 Lustre 部署環境

在每個 EF80 陣列和 Lustre 伺服器上設定管理存取和基本設定，然後準備 Ansible 控制節點來部署解決方案。

準備 E 系列陣列

步驟

1. 在每個控制器上設定管理介面，並確認 SANtricity System Manager 是否可以存取。
2. 設定 Ansible 清單將使用的陣列名稱和管理憑證。
3. 確認陣列處於最佳狀態，所有硬碟均已到位且報告正常。

有關設定管理連線的說明，請參閱"E 系列文件"。

準備 Lustre 伺服器

步驟

1. 在每台伺服器上設定基板管理控制器（BMC），以實現頻外存取，並設定 UEFI 系統設定以獲得最佳效能。
2. 安裝支援的作業系統（RHEL 或 Rocky Linux），然後設定主機名稱和管理網路連接埠。

3. 準備所有韌體和軟體套件，以符合 **E** 系列 **Lustre** 解決方案的要求，如 "軟體元件" 中所述。

準備 **Ansible** 控制節點

指定一個可透過管理網路存取所有 Lustre 伺服器和 E 系列陣列的虛擬或實體 Linux 系統，作為 Ansible 控制節點。

以下步驟使用 Ubuntu 24.04。有關其他 Linux 發行版的說明，請參閱"Ansible 安裝文件"。

步驟

1. 安裝 Python、pip 和 Python 虛擬環境套件：

```
sudo apt-get install python3 python3-pip python3-setuptools python3-venv
```

2. 建立並啟動 Python 虛擬化環境：

```
python3 -m venv ~/pyenv
source ~/pyenv/bin/activate
```

3. 在已啟動的虛擬環境中安裝 Ansible 和所需的 Python 套件：

```
pip install "ansible-core>=2.19" cryptography jinja2 ipaddr netaddr \
  passlib pyyaml resolvelib
```

4. 安裝 NetApp E 系列 Lustre Ansible 集合。此集合的安裝也會安裝其相依集合，包括 SANtricity 和 Host 集合：

```
ansible-galaxy collection install netapp_eseries.lustre
```

5. 驗證已安裝的 Ansible、Python 和集合版本：

```
ansible --version
python3 --version
ansible-galaxy collection list netapp_eseries.lustre
```

6. 設定從控制節點到每個 Lustre 伺服器的免密碼 SSH 連線。請將 `<ip_or_hostname>`替換為伺服器的管理 IP 位址或主機名稱：

```
ssh-keygen
ssh-copy-id <ip_or_hostname>
```

對每個 Lustre 伺服器重複上述步驟。 `ssh-copy-id`

> 請勿為 E 系列陣列設定無密碼 SSH。Ansible 透過 SANtricity Web Services API 使用清單中定義的憑證來管理這些陣列。

完成後

"自訂 Ansible 清單".

自訂 **Lustre Ansible** 庫存

建立 EF80 部署範本的工作複本，並針對您的伺服器、陣列、網路和 Lustre 建置區塊自訂其清單。

將選定的庫存保留在其架構和 Platform 目錄下。保留 shared `playbooks` 目錄和 `passwords.yml` 檔案於 `building_blocks` 的根目錄。

選擇部署範本

首先選擇與站台網路架構相符的範本：

- **InfiniBand:** "IB_H2_IB_DAC_A2_EF80/lustre_building_block_cluster_1"
- **RoCE:** "ROCE_H2_ROCE_DAC_A2_EF80/lustre_building_block_cluster_1"

設定站台專屬設定

在清單中設定特定站台的值，以及本機儲存庫和 udev 規則設定。表格中的行引用指向 `ansible-lustre release/1.0.0` 標籤中的 RoCE EF80 範本檔案。此 InfiniBand 範本使用相同的變數名稱和結構，但 LNet 介面變數除外，該變數已在表格中註明。

欄位	**Ansible** 變數	模板檔案 (`release/1.0.0`)	範圍	附註
檔案系統名稱	`eseries_lustre_filesystem_mgt.format_options.fsname`, `eseries_lustre_filesystem_mdt.format_options.fsname`, `eseries_lustre_filesystem_ost.format_options.fsname`	"ha_cluster.yml#L76", "#L90", "#L110"	全集群範圍	巢狀於 MGT、MDT 和 OST 映射的 `format_options`之下。最多 8 個字元。在所有三個位置使用相同的值。頂層 `fsname`金鑰無效。

欄位	Ansible 變數	模板檔案 (release/1.0.0)	範圍	附註
用於目標註冊的 MGS NID	eseries_lustre_filesystem_mdt.format_options.mgsnode, eseries_lustre_filesystem_ost.format_options.mgsnode	"ha_cluster.yml#L98", "#L118"	全集群範圍	嵌套在 MDT 和 OST 下 format_options。包含 HA 配對中兩個節點的前端 LNet NID。首先列出設定為慣用 MGS 擁有者的節點的 NID，然後列出合作夥伴節點的 NID。
Pacemaker 叢集名稱	eseries_lustre_pacemaker_cib_properties_overrides.cluster_properties.cluster-name	"ha_cluster.yml#L228"	全集群範圍	嵌套於 Pacemaker CIB 屬性覆蓋項下。HA 叢集的唯一名稱。
BMC 圍欄位址	eseries_lustre_pacemaker_fencing_agents.fence_redfish.ip	"ha_cluster.yml#L211", "#L214", "#L217", "#L220"	每個 Lustre 伺服器重複此操作	每個伺服器節點的 Redfish BMC IP 位址。
提醒郵件收件人	eseries_lustre_alert_email_list	"ha_cluster.yml#L233"	全集群範圍	HA 資源和故障通知的接收者。
郵件網域	eseries_lustre_mail_service_overrides.conf.mydomain	"ha_cluster.yml#L240"	全集群範圍	用於發送警示電子郵件的 Postfix 網域。
NTP 時間來源	eseries_lustre_time_sync_overrides.server_pools	"ha_cluster.yml#L250"	全集群範圍	叢集節點的時間來源。指定一個 NTP 資源池或一個或多個獨立的 NTP 伺服器。包含必要的 `pool`或 `server`指令以及諸如 `iburst`之類的選項。
介面 PCI 到名稱 udev 映射	eseries_ip_udev_rules	"ha_cluster.yml#L172"	每個 HCA 插槽佈局重複	將每個 PCI 插槽對應到一個持久性介面名稱（L182 處的值）。使用 `lspci -nnvmm`收集。
本機套件儲存庫（可選）	eseries_common_yum_repos	"ha_cluster.yml#L256", "#L261"	全集群範圍	對於實體隔離安裝，請取消註解區塊並設定儲存庫 url （L264）。
伺服器管理 IP	ansible_host	"host_vars/lustre_01.yml#L10"	每個 Lustre 伺服器重複此操作	每個伺服器一個 `host_vars/lustre_NN.yml`檔案。

欄位	**Ansible** 變數	模板檔案 (`release/1.0.0`)	範圍	附註
NVMe-oF 後端介面	`eseries_nvme_roce_interfaces`（RoCE）/ `eseries_nvme_ib_interfaces`（InfiniBand）	"RoCE host_vars/lustre_01.yml#L21", "InfiniBand host_vars/lustre_01.yml#L17"	每個 Lustre 伺服器重複此操作	用於 E 系列陣列的 NVMe-oF 儲存設備流量的後端介面位址。
LNet 叢集介面	`eseries_roce_interfaces`（RoCE）／ `eseries_ipoib_interfaces`（InfiniBand）	"RoCE host_vars/lustre_01.yml#L47", "InfiniBand host_vars/lustre_01.yml#L38"	每個 Lustre 伺服器重複此操作	前端 LNet 介面位址（L56 處的 RoCE 值；L47 處的 InfiniBand 值）。
Corosync 節點 IP	`eseries_lustre_corosync_node_ips`	"host_vars/lustre_01.yml#L58"	每個 Lustre 伺服器重複此操作	列出 HA 叢集中的所有節點 IP 位址（L62 層的值）。
Lustre 目標服務 NID	`lustre_target_service_nodes`	"host_vars/lustre_01.yml#L64"	每個 Lustre 伺服器重複此操作	包含 HA 配對兩個節點的前端 LNet NID (`@o2ib`），以便故障轉移後目標服務仍可連線（L67 中的值）。
陣列管理 IP	`ansible_host`	"host_vars/netapp_01.yml#L13"	每個 E 系列陣列重複上述步驟	一個控制器的管理 IP 位址；API URL 由該 IP 位址衍生而來。
建置區塊群組成員資格	`mgs` / `mds_NN` / `oss_NN`	"lustre_inventory.yml"	每個建置區塊	每個清單主機名稱必須與其對應的 `host_vars` 檔案的主檔名相符。例如，主機 `lustre_01` 使用 `host_vars/lustre_01.yml`，而 `netapp_01` 使用 `host_vars/netapp_01.yml`。僅在第一個建置區塊中包含 MGS 群組。在基礎建置區塊和每個 MDT+OST 擴充建置區塊中包含 MDS 群組；僅在僅限 OST 的擴充建置區塊中省略 MDS 群組。使用 `mgsnode` 向現有的 MGS 註冊擴充目標。

對每個主機和建置區塊重複上述設定。為每個 Lustre 伺服器建立一個 `host_vars/lustre_NN.yml`檔案，為每個陣列建立一個 `host_vars/netapp_NN.yml`檔案，並為每個額外的建置區塊重複清單群組。

部署 playbook 會從共享的 `building_blocks/passwords.yml`檔案載入陣列、Pacemaker 和 BMC 憑證。部署前，請使用 Ansible Vault 填入此檔案並對其加密，或在執行時使用 `--extra-vars`安全地置換這些值。請勿將明文憑證提交至來源控制。

部署範本中的 IP 位址、主機名稱、介面名稱和 Volume 規模調整均為範例。在執行任何 playbook 之前，請修改環境的所有清單檔案。

完成後

"部署 Lustre HA 叢集和用戶端".

部署 **Lustre** 高可用度叢集和用戶端

準備伺服器作業系統，安裝 NVIDIA DOCA-OFED，部署 Lustre HA 叢集，並使用 NetApp Ansible playbooks 設定 Lustre 用戶端。

在開始之前，請完成並確保"自訂 Lustre Ansible 庫存"中所述的物品清單。

部署 **Lustre HA** 叢集

步驟

1. 從工作複本的 `building_blocks/playbooks`目錄中，準備 Lustre 伺服器上的作業系統。將 `<inventory_path>`替換為您自訂的庫存目錄路徑：

```
ansible-playbook -i <inventory_path>/lustre_inventory.yml
deploy_lustre_os/deploy_lustre_os_playbook.yml
```

2. 在每台 Lustre 伺服器上安裝 NVIDIA DOCA-OFED。針對上一步安裝的核心建置核心模組。請按照適用於您核心的步驟進行操作"NVIDIA DOCA-Host 安裝與升級"。以下範例展示了在 RHEL 或 Rocky Linux 上安裝 DOCA-OFED 的過程。

   a. 安裝 DOCA 主機儲存庫套件並重新整理套件快取。請將 `<doca_host_package>`替換為您的作業系統和架構對應的 DOCA 主機套件：

```
dnf install -y wget tar
wget https://www.mellanox.com/downloads/DOCA/<doca_host_package>.rpm
rpm -i <doca_host_package>.rpm
dnf clean all
dnf makecache
```

   b. 為正在執行的核心建立 DOCA 核心模組。該指令碼會列印其所建立的套件路徑：

```
dnf install -y doca-extra
/opt/mellanox/doca/tools/doca-kernel-support
```

   c. 安裝指令碼建立的核心模組儲存庫套件，然後重新整理套件快取。將 `<doca_kernel_repo>`替換為上一步中的路徑：

```
rpm -Uvh <doca_kernel_repo>.rpm
dnf makecache
```

d. 安裝 DOCA-OFED 使用者空間、核心和 NVMe 套件。將 `<kernel_version>`替換為目前執行中的核心版本：

```
dnf install -y doca-ofed-userspace
dnf install -y --disablerepo=doca doca-kernel-<kernel_version>
dnf install -y kmod-mlnx-nvme
```

3. 執行主要部署 playbook：

```
ansible-playbook -i <inventory_path>/lustre_inventory.yml
lustre_playbook.yml
```

調整 `--forks`以依部署規模控制 `Ansible` 並行設定的主機數量。例如，對於較大的叢集，請新增 `--forks 20`。

此腳本配置 E 系列 Volume 群組或動態磁碟資源池，以及 Volume，設定 NVMe-oF 主機連線功能，格式化並裝載 Lustre 目標，設定 LNet，套用效能調校，以及設定 Pacemaker HA 資源。

包含一個建置區塊的部署將形成一個雙節點 Pacemaker 叢集。在容錯移轉時，節點 A 將獲得額外的投票權。若要在雙節點叢集中實現仲裁，請考慮設定 qdevice。當部署包含多個建置區塊時，每個建置區塊都會向更大的 Pacemaker 叢集貢獻一對雙節點伺服器，直到達到已記錄的叢集限制。請檢閱"HA 管理使用者指南"中的隔離和仲裁組態，以便叢集能夠在節點故障期間安全地恢復目標。

部署 **Lustre** 用戶端

透過自訂用戶端部署範本並執行用戶端 playbook，可在任何需要檔案系統存取的系統上部署 Lustre 用戶端。

步驟

1. 從 "Lustre 支援對照表" 中選擇相容的用戶端作業系統和核心。如果尚未安裝，請從 "netapp-lustre" 建置並安裝適用於該核心的 Lustre 用戶端軟體套件。
2. 建立範本目錄的工作複本 `clients`（包括其共享的 `passwords.yml`檔案），並選擇與您的架構相符的範本：
   - **InfiniBand:** "IB_lustre_cluster_clients"
   - **RoCE:** "ROCE_lustre_cluster_clients"
3. 自訂 `lustre_client_inventory.yml`、`host_vars`和 `group_vars`以供您的用戶端使用。表格中的行引用指向 `ansible-lustre`release/1.0.0`標籤中的 RoCE 用戶端範本檔案；InfiniBand 用戶端範本使用相同的變數，除非另有說明。

欄位	**Ansible** 變數	模板檔案 (`release/1.0.0`)	範圍	附註
用戶端主機	`lustre_clients`	"lustre_client_inventory.yml #L4"	每個用戶端	列出每個用戶端主機名稱。

欄位	**Ansible** 變數	模板檔案 (`release/1.0.0`)	範圍	附註
用戶端 SSH 使用者名稱和密碼	`ssh_client_user` / `ssh_client_become_pass`	"passwords.yml#L2"	全集群範圍	僅當 SSH 使用者不是 `root`時，才設定 become 密碼。
LNet 介面	`eseries_lustre_lnet`	"group_vars/all.yml#L7"	全集群範圍	用戶端 LNet 介面名稱（L13 的值）。
裝載 MGS NID	`lustre_client_mounts.mgsnode`	"group_vars/all.yml#L17"	全集群範圍	用於裝載檔案系統的 MGS NID （L20 中的值）。
檔案系統名稱	`lustre_client_mounts.fsname`	"group_vars/all.yml#L21"	全集群範圍	必須與叢集 `fsname`相符。
裝載點	`lustre_client_mounts.mount_point`	"group_vars/all.yml#L22"	全集群範圍	用戶端裝載目錄。
用戶端管理 IP	`ansible_host`	"host_vars/client_01.yml#L4"	每個用戶端重複	每個用戶端一個 `host_vars/client_NN.yml` 檔案。
儲存架構介面	`eseries_roce_interfaces`	"host_vars/client_01.yml#L10"	每個用戶端重複	前端介面位址（第 15 行的值）。InfiniBand 模板在此處使用 `eseries_ipoib_interfaces`。

對每個用戶端重複上述設定。為每個用戶端建立一個 `host_vars/client_NN.yml`**檔案，並將相應的主機新增至** `lustre_client_inventory.yml`。填入共享的 `clients/passwords.yml`檔案，並在執行用戶端 playbook 之前使用 Ansible Vault 對其加密。

4. 從用戶端模板目錄執行用戶端 playbook：

```
ansible-playbook -i lustre_client_inventory.yml
lustre_client_playbook.yml
```

用戶端 playbook 設定用戶端網路介面和 LNet，並可為 Lustre 檔案系統建立持久的 systemd 裝載單元。

完成後

"驗證部署".

驗證並擴充 **Lustre** 部署

在生產環境使用前，請確認叢集健全狀況、Lustre 掛載點和檔案系統容量，然後在需要時使用相同的清單新增建置區塊。

驗證部署

步驟

1. 從 Lustre 伺服器確認所有 Pacemaker 資源均已在預期節點上啟動：

```
pcs status
```

2. 從 Lustre 用戶端確認 Lustre 檔案系統已掛載：

```
mount -t lustre
```

3. 從 Lustre 用戶端確認檔案系統容量，並確認 MDT 和 OST 目標可見：

```
lfs df -h
```

有關使用 IOR、mdtest 和 fio 進行效能驗證的資訊，請參見"尺寸指南"。

對於受控目標移轉或 HA 容錯移轉測試，請遵循 "HA 管理使用者指南" 中的程序。

擴充檔案系統

若要增加容量或中繼資料效能，請使用額外的建置區塊擴充現有清單。請勿為新建置區塊建立分隔的清單。劇本會自動分配唯一的 MDT 和 OST 索引，並將新目標註冊到現有的 MGS 中。

步驟

1. 將新的建置區塊主機、陣列和資源群組（MDT+OST 或僅 OST）新增至您用於原始部署的相同清單和 `host_vars`和 `group_vars`檔案。
2. 使用更新後的清單重新執行主要部署 Playbook。

請參閱 "解決方案架構" 以取得擴充規則，以及參閱 "尺寸指南" 以取得何時新增每種建置區塊類型的指導。

Lustre 搭配 NetApp E 系列儲存設備 - 相關資源

在規劃或擴充 Lustre 與 NetApp E 系列儲存設備部署時，請使用以下連結取得相關文件、工具和軟體。有關規模調整和部署步驟，請參閱 "尺寸指南" 和 "部署解決方案"。

NetApp 資源

- "NetApp 互通性矩陣工具"。搜尋 **E-Series Lustre**。
- "NetApp EF 系列 All Flash Array 產品型錄"
- "NetApp SANtricity System Manager 文件"
- "NetApp ansible-lustre 集合"。有關庫存欄位指導，請參閱"自訂 Lustre Ansible 庫存"。
- "NetApp Lustre 來源儲存庫"。NetApp 提供適用於 Rocky Linux 9.8 和 Red Hat Enterprise Linux 9.8 的預先

建置 Lustre 伺服器 RPM。從 NetApp 來源為用戶端作業系統和核心建置 Lustre 用戶端套件。

Lustre 社群

- "Lustre 檔案系統"
- "Lustre 支援對照表"適用於用戶端作業系統和核心
- "Lustre Wiki：Lustre 調校"

相關 NetApp 人工智慧基礎架構解決方案

- "NetApp上的 BeeGFS 與 E 系列存儲"
- "部署 IBM Spectrum Scale 與 NetApp E 系列儲存設備"
- "搭載NetApp EF 系列的NVIDIA DGX SuperPOD"