



並行 NFS ONTAP 9

NetApp
January 08, 2026

目錄

並行 NFS	1
簡介	1
了解ONTAP中的平行 NFS (pNFS)	1
了解ONTAP中的 pNFS 架構	2
ONTAP中的 pNFS 用例	8
ONTAP中的 pNFS 部署策略	13
規劃	14
pNFS部署計劃	14
pNFS調優與效能最佳實踐	16
pNFS 指令、統計資料和事件日誌	20

並行 NFS

簡介

了解ONTAP中的平行 NFS (pNFS)

並行 NFS 於 2010 年 1 月作為 RFC 標準 RFC-5661 推出，旨在透過分離元資料和資料路徑，允許客戶端直接存取 NFSv4.1 伺服器上的檔案資料。這種直接存取方式透過資料本地化、CPU效率和操作並行化帶來效能優勢。2018 年編寫了後續的 RFC，涵蓋 pNFS 佈局類型 (RFC-8434)，其中定義了文件、區塊和物件佈局的標準。ONTAP利用 pNFS 操作的檔案佈局類型。



從 2024 年 7 月開始，先前以 PDF 形式發布的技術報告內容已與ONTAP產品文件整合。ONTAP NFS 儲存管理文件現在包含了來自 *TR-4063: NetApp ONTAP 中的平行網路檔案系統 (pNFS)* 的內容。

多年來，NFSv3 一直是 NFS 協定的標準版本，幾乎適用於所有用例。然而，該協定存在一些局限性，例如缺乏狀態性、權限模型簡陋以及基本的鎖定功能。NFSv4.0 (RFC 7530) 在 NFSv3 的基礎上進行了一系列改進，隨後的 NFSv4.1 (RFC 5661) 和 NFSv4.2 (RFC 7862) 版本進一步改進了這些改進，增加了並行 NFS (pNFS) 等功能。

NFSv4.x 的優勢

NFSv4.x 相較於 NFSv3 有下列優勢：

- 由於 NFSv4 僅使用單一連接埠 (2049) 進行操作，因此對防火牆友善。
- 進階且積極的快取管理，例如 NFSv4.x 中的委託機制
- 採用密碼學的強大 RPC 安全選項
- 字元國際化
- 複合操作
- 僅支援 TCP
- 有狀態協定 (不像NFSv3那樣是無狀態的)
- 全面整合 Kerberos，實現高效率的身份驗證機制
- NFS轉診
- 支援與 UNIX 和 Windows 相容的存取控制
- 基於字串的使用者和群組標識符
- pNFS (NFSv4.1)
- 擴展屬性 (NFSv4.2)
- 安全標籤 (NFSv4.2)
- 稀疏檔案操作 (FALLOCATE) (NFSv4.2)

有關 NFSv4.x 的更多信息，包括最佳實踐和功能詳情，請參閱 "NetApp技術報告4067：NFS最佳實務與實作指南"。

相關資訊

- "NFS組態總覽"
- "NFS 管理概述"
- "資料區管理FlexGroup"
- "NFS 主幹總覽"
- <https://www.netapp.com/pdf.html?item=/media/19370-tr-4523.pdf>
- "NetApp技術報告4616：ONTAP NFS Kerberos in Some with Microsoft Active Directory"

了解ONTAP中的 pNFS 架構

pNFS 架構由三個主要元件組成：支援 pNFS 的 NFS 用戶端、為元資料操作提供專用路徑的元資料伺服器以及提供檔案本地化路徑的資料伺服器。

用戶端存取 pNFS 需要與 NFS 伺服器上的資料和元資料路徑建立網路連線。如果 NFS 伺服器包含用戶端無法存取的網路接口，則伺服器可能會向客戶端通告無法存取的資料路徑，從而導致服務中斷。

元資料伺服器

當用戶端使用 NFSv4.1 或更高版本發起掛載，且 NFS 伺服器上啟用了 pNFS 時，pNFS 中的元資料伺服器就會建立。完成此操作後，所有元資料流量都將透過此連接發送，並在掛載期間保持在此連接上，即使介面遷移到另一個節點。

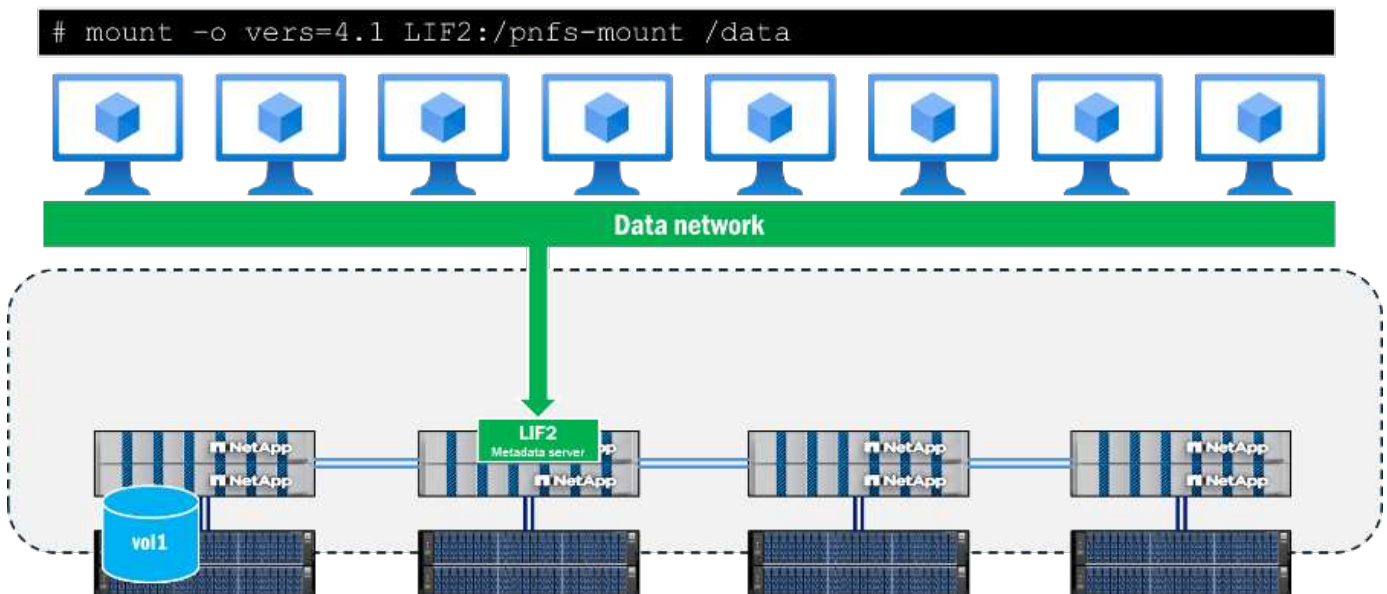


圖 1. 在ONTAP的 pNFS 中建立元資料伺服器

pNFS 支援是在掛載呼叫期間確定的，具體是在 EXCHANGE_ID 呼叫中確定的。在 NFS 操作下方的封包擷取中可以看到這個標誌。當 pNFS 標誌 EXCHGID4_FLAG_USE_PNFS_DS 和 EXCHGID4_FLAG_USE_PNFS_MDS 如果設定為 1，則該介面符合 pNFS 中的資料和元資料操作條件。

```

  Operations (count: 1)
    Opcode: EXCHANGE_ID (42)
      Status: NFS4_OK (0)
      clientid: 0x004050a97100001c
      seqid: 0x00000001
    flags: 0x00060100, EXCHGID4_FLAG_USE_PNFS_DS, EXCHGID4_FLAG_USE_PNFS_MDS, EXCHGID4_FLAG_BIND_PRINC
      0... .. = EXCHGID4_FLAG_CONFIRMED_R: Not set
      .0.. .. = EXCHGID4_FLAG_UPD_CONFIRMED_REC_A: Not set
      .... ..1.. .. = EXCHGID4_FLAG_USE_PNFS_DS: Set
      .... ..1.. .. = EXCHGID4_FLAG_USE_PNFS_MDS: Set
      .... ..0 .. = EXCHGID4_FLAG_USE_NON_PNFS: Not set
      .... ..1 .. = EXCHGID4_FLAG_BIND_PRINC_STATEID: Set
      .... ..0. = EXCHGID4_FLAG_SUPP_MOVED_MIGR: Not set
      .... ..0 = EXCHGID4_FLAG_SUPP_MOVED_REFERER: Not set

```

圖 2. pNFS掛載的資料包捕獲

NFS 中的元資料通常包括檔案和資料夾屬性，例如檔案句柄、權限、存取和修改時間以及所有權資訊。元資料還可以包括建立和刪除呼叫、連結和取消連結呼叫以及重新命名。

在 pNFS 中，還有一部分元資料呼叫是 pNFS 特性特有的，這些呼叫將在後續章節中進行更詳細的介紹。"RFC 5661"。這些呼叫用於協助確定 pNFS 合格設備、設備到資料集的映射以及其他所需資訊。下表列出了 pNFS 特有的元資料操作。

營運	說明
佈局獲取	從元資料伺服器取得資料伺服器對應。
佈局委員會	伺服器提交佈局並更新元資料映射。
佈局返回	返回佈局；如果資料被修改，則傳回新佈局。
取得設備資訊	客戶端取得儲存叢集中資料伺服器的最新資訊。
取得設備列表	客戶端請求提供參與儲存叢集的所有資料伺服器的清單。
CB_LAYOUTRECALL	如果偵測到衝突，伺服器會從客戶端呼叫資料佈局。
CB_RECALL_ANY	將所有佈局傳回給元資料伺服器。
CB_NOTIFY_DEVICEID	設備 ID 發生任何變更時發出通知。

資料路徑資訊

元資料伺服器建立並開始資料操作後，ONTAP開始追蹤符合 pNFS 讀取和寫入操作條件的設備 ID，以及將叢集中的磁碟區與本機網路介面關聯起來的設備對應。當對掛載點執行讀取或寫入操作時，就會發生此程序。元資料調用，例如 GETATTR，不會觸發這些設備映射。因此，運行一個 `ls` 掛載點內的命令不會更新映射。

可以使用ONTAP CLI 的高階權限查看設備和映射，如下所示。

```

::*> pnfs devices show -vserver DEMO
(vserver nfs pnfs devices show)
Vserver Name      Mapping ID      Volume MSID      Mapping Status
Generation
-----
DEMO              16             2157024470      available        1

::*> pnfs devices mappings show -vserver SVM
(vserver nfs pnfs devices mappings show)
Vserver Name      Mapping ID      Dsid             LIF IP
-----
DEMO              16             2488             10.193.67.211

```



這些指令中沒有磁碟區名稱。相反，將使用與這些磁碟區關聯的數字 ID：主集 ID (MSID) 和資料集 ID (DSID)。若要尋找與映射相關的體積，您可以使用 `volume show -dsid [dsid_numeric]` 或者 `volume show -msid [msid_numeric]` 具備 ONTAP CLI 的高階權限。

當客戶端嘗試讀取或寫入位於遠離元資料伺服器連接的節點上的檔案時，pNFS 將協商適當的存取路徑，以確保這些操作的資料本地性，並且客戶端將重定向到已發布的 pNFS 設備，而不是嘗試穿越叢集網路來存取該檔案。這有助於降低 CPU 佔用率和網路延遲。

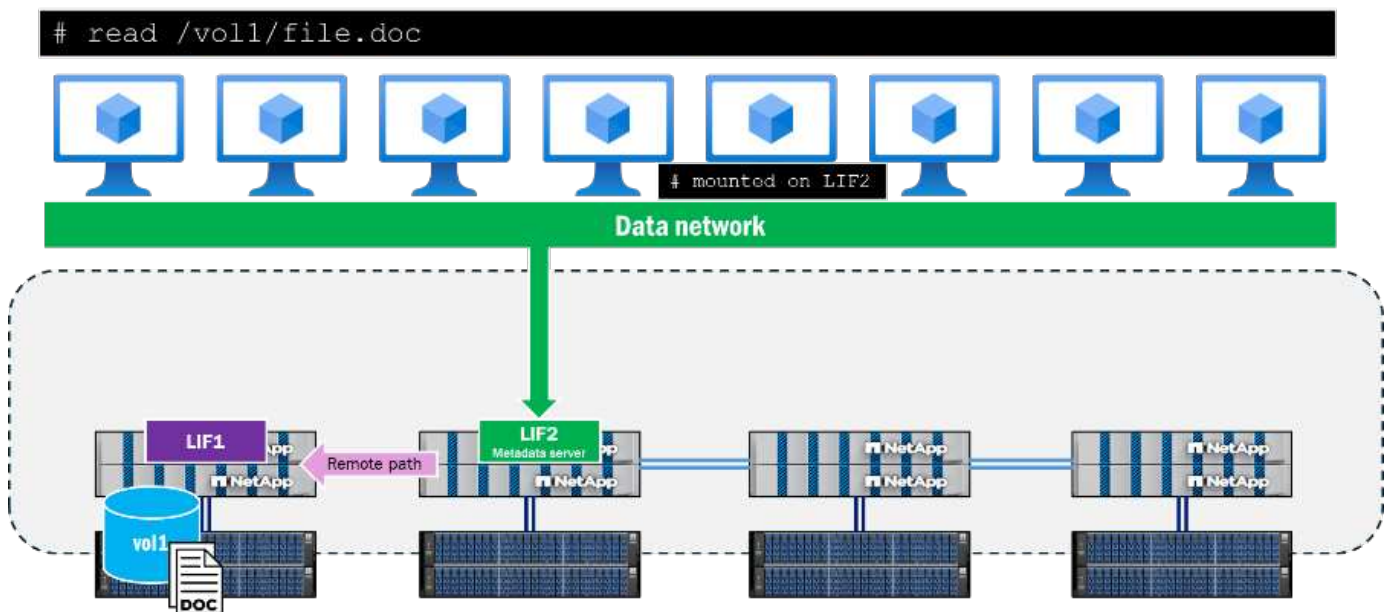


圖 3. 使用 NFSv4.1 的遠端讀取路徑，無需 pNFS

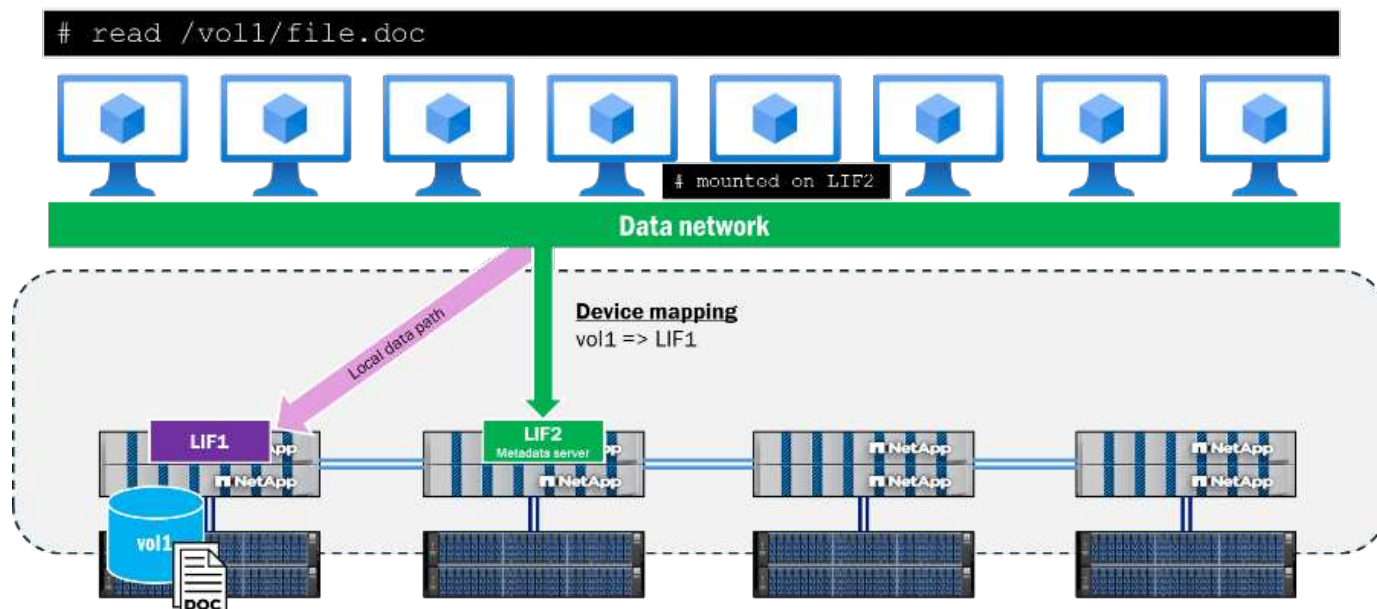


圖 4. 使用 pNFS 的本地化讀取路徑

pNFS 控制路徑

除了 pNFS 的元資料和資料部分之外，還有一個 pNFS 控制路徑。控制路徑由 NFS 伺服器用於同步檔案系統資訊。在 ONTAP 集群中，後端集群網路會定期複製，以確保所有 pNFS 設備和設備映射保持同步。

pNFS 設備群組工作流程

以下描述了客戶端請求讀取或寫入磁碟區中的檔案後，pNFS 設備如何在 ONTAP 中填入。

1. 客戶端請求讀取或寫入；執行 OPEN 操作並檢索檔案句柄。
2. 執行 OPEN 操作後，用戶端透過元資料伺服器連接，以 LAYOUTGET 呼叫的方式將檔案句柄傳送至儲存裝置。
3. LAYOUTGET 向客戶端傳回有關文件佈局的信息，例如狀態 ID、條帶大小、文件段和裝置 ID。
4. 然後客戶端取得裝置 ID，並向伺服器發送 GETDEVINFO 呼叫以檢索與該裝置關聯的 IP 位址。
5. 儲存裝置會傳送回覆，其中包含用於本機存取該裝置的關聯 IP 位址清單。
6. 用戶端透過儲存設備傳回的本機 IP 位址繼續進行 NFS 通訊。

pNFS 與 FlexGroup 體積的交互

ONTAP 中的 FlexGroup 磁碟區將儲存呈現為跨越叢集中多個節點的 FlexVol volume 組成部分，從而允許工作負載利用多個硬體資源，同時保持單一掛載點。由於多個具有多個網路介面的節點與工作負載交互，因此在 ONTAP 中看到遠端流量穿越後端叢集網路是很自然的結果。

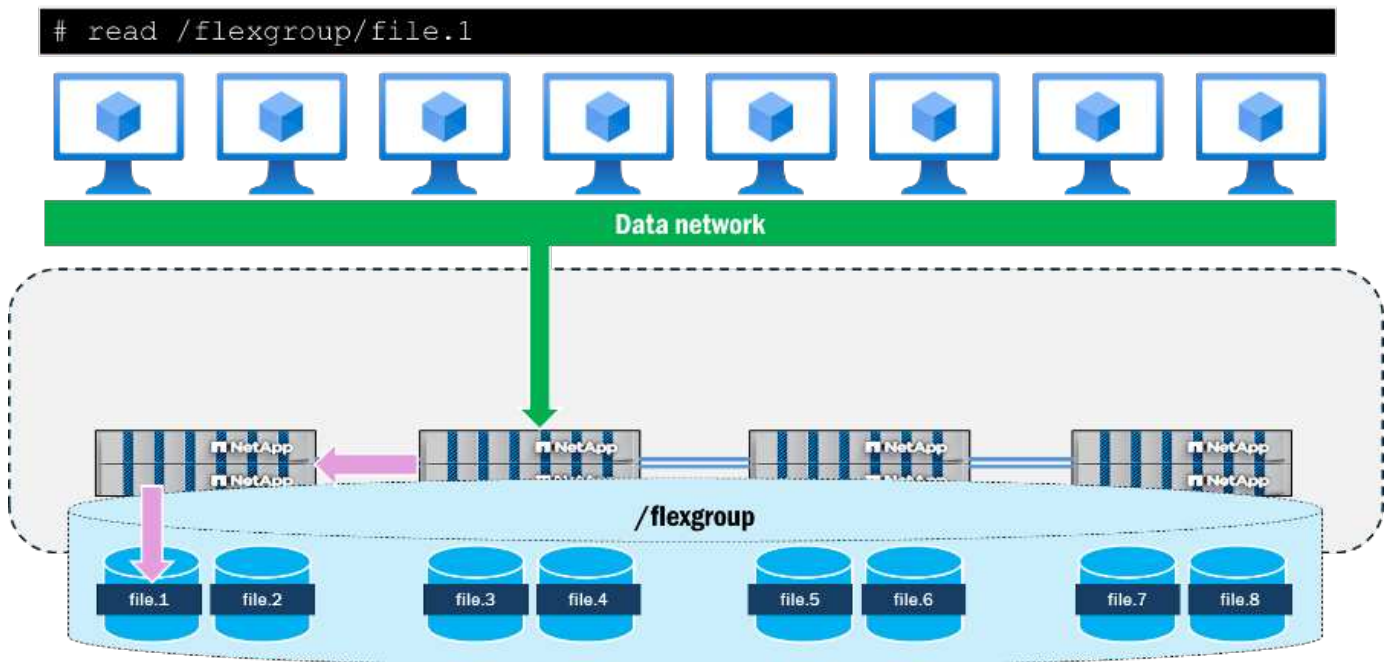


圖 5. 在不使用 pNFS 的情況下，透過FlexGroup磁碟區存取單一文件

使用 pNFS 時，ONTAP會追蹤FlexGroup磁碟區的檔案和磁碟區佈局，並將它們對應到叢集中的本機資料介面。例如，如果包含正在存取的檔案的組成磁碟區位於節點 1 上，則ONTAP將通知用戶端將資料流量重新導向至節點 1 上的資料介面。

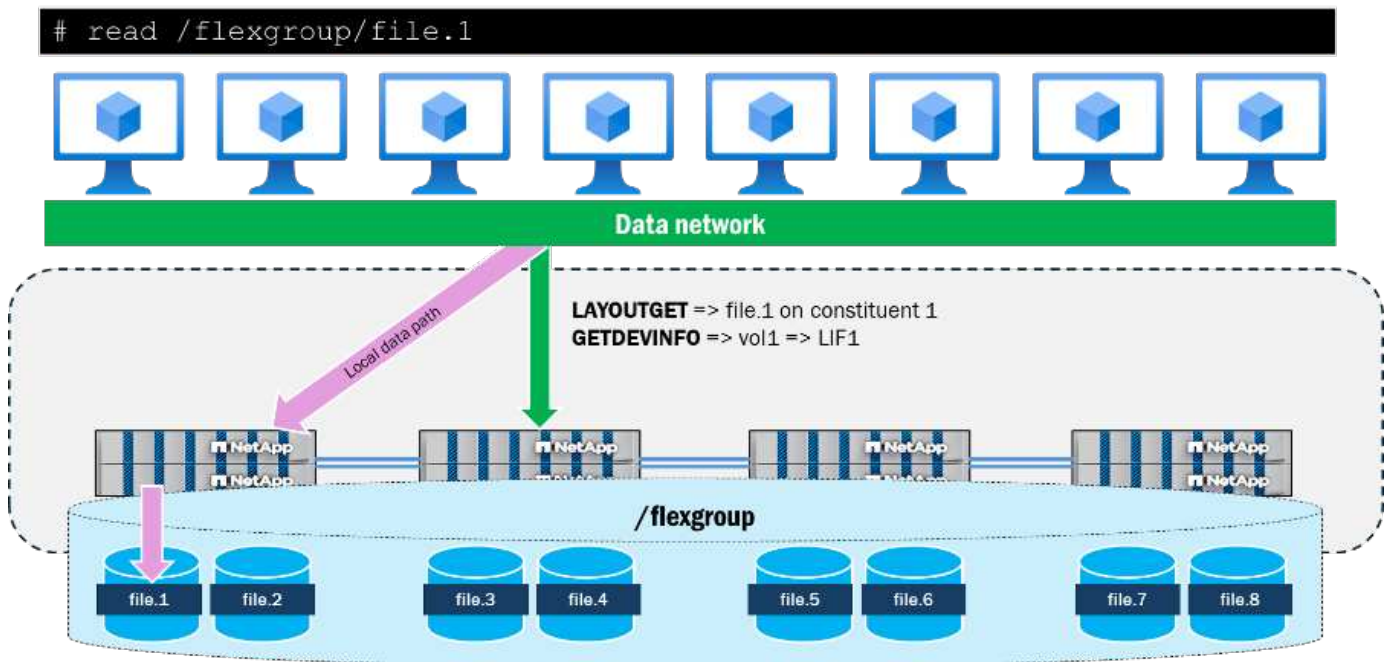


圖 6. 在FlexGroup區中使用 pNFS 進行單一檔案存取

pNFS 也提供了從單一客戶端呈現平行網路路徑到檔案的功能，而沒有 pNFS 的 NFSv4.1 則不具備此功能。例如，如果用戶端想要使用 NFSv4.1（不使用 pNFS）從相同掛載點同時存取四個文件，則所有文件都將使用相同的網路路徑，ONTAP叢集將改為向這些文件發送遠端請求。掛載路徑可能會成為操作的瓶頸，因為它們都遵循一條路徑到達同一個節點，並且還要處理元資料操作以及資料操作。

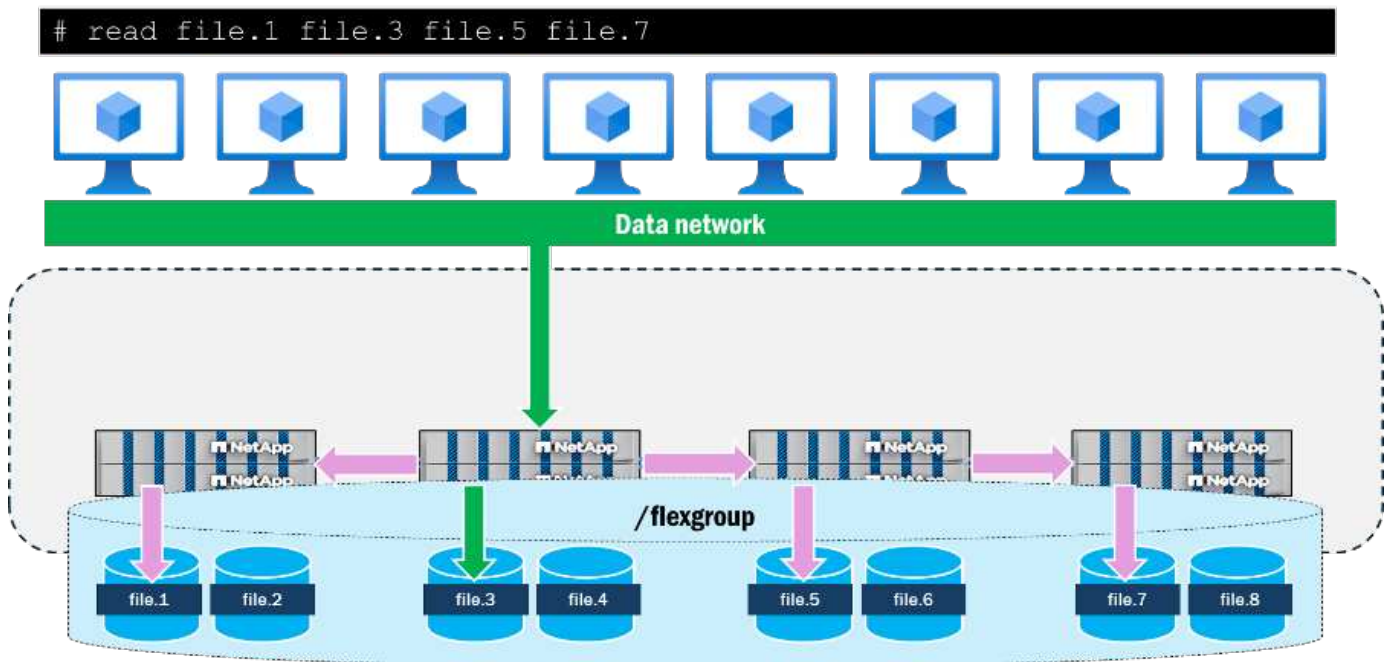


圖 7. 在不使用 pNFS 的情況下，FlexGroup 磁碟區中可同時存取多個檔案。

當使用 pNFS 從單一客戶端同時存取相同的四個檔案時，用戶端和伺服器會協商每個節點上的檔案本機路徑，並使用多個 TCP 連線進行資料操作，而掛載路徑則作為所有元資料操作的位置。這樣既可以利用本地檔案路徑來降低延遲，也可以透過使用多個網路介面來提高吞吐量，前提是客戶端可以發送足夠的資料來使網路飽和。

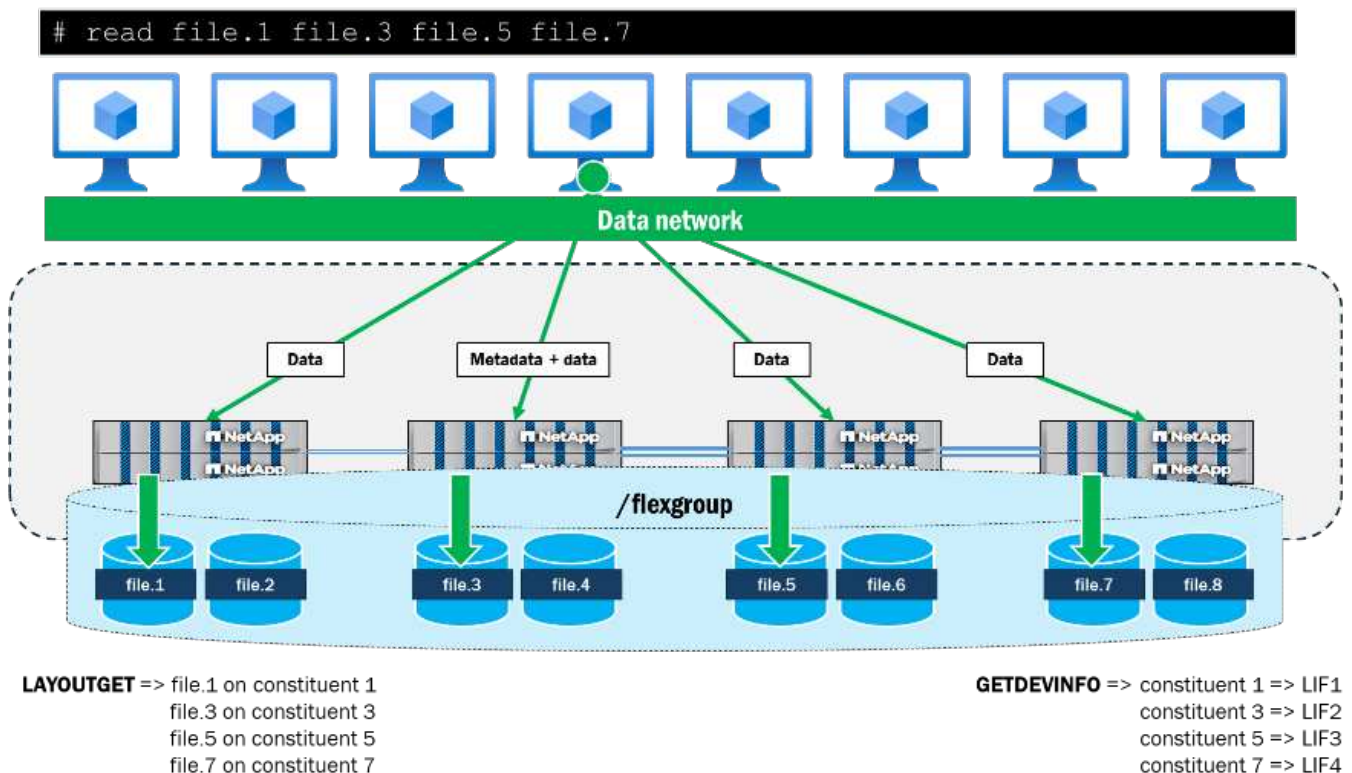


圖 8. 在FlexGroup卷中使用 pNFS 實現多個文件同時訪問

以下顯示的是在單一 RHEL 9.5 用戶端上進行簡單測試運行的結果，其中使用 dd 並行讀取四個 10GB 檔案（全部位於兩個ONTAP叢集節點的不同組成磁碟區上）。使用 pNFS 時，每個檔案的整體吞吐量和完成時間都提高了。當不使用 pNFS 而使用 NFSv4.1 時，掛載點本機檔案和遠端檔案之間的效能差異比使用 pNFS 時更大。

測試	每個檔案的吞吐量 (MB/s)	每個文件的完成時間
NFSv4.1：無 pNFS	• 文件 1–228 (本地) • 文件.2–227 (本地) • 文件.3–192 (遠端) • 文件.4–192 (遠端)	• 文件 1–46 (本地) • 文件.2–46.1 (本地) • 文件.3–54.5 (遠端) • 文件.4–54.5 (遠端)
NFSv4.1：與 pNFS	• 文件 1–248 (本地) • 文件.2–246 (本地) • 文件.3–244 (本地透過pNFS) • 文件.4–244 (本地透過pNFS)	• 文件.1–42.3 (本地) • 文件.2–42.6 (本地) • 文件.3–43 (本地透過pNFS) • 文件 4–43 (本地透過 pNFS)

相關資訊

- ["資料區管理FlexGroup"](#)
- ["NetApp技術報告 4571：FlexGroup最佳實踐"](#)

ONTAP中的 pNFS 用例

pNFS 可以與ONTAP 的各種功能一起使用，以提高效能並為 NFS 工作負載提供更大的靈活性。

使用 nconnect 的 pNFS

NFS 為一些較新的客戶端和伺服器引入了一種新的掛載選項，該選項提供了一種在掛載單一 IP 位址的同時提供多個 TCP 連線的方法。這提供了一種機制，可以更好地並行化操作，繞過 NFS 伺服器和客戶端的限制，並有可能為某些工作負載提供更高的整體效能。ONTAP 9.8 及更高版本支援 nconnect，前提是用戶端支援 nconnect。

當使用 nconnect 連接 pNFS 時，將使用 nconnect 選項並行連接 NFS 伺服器通告的每個 pNFS 裝置。例如，如果 nconnect 設置為 4，並且有 4 個符合 pNFS 條件的接口，則每個掛載點創建的連接總數最多為 16 個（4 個 nconnect x 4 個 IP 位址）。

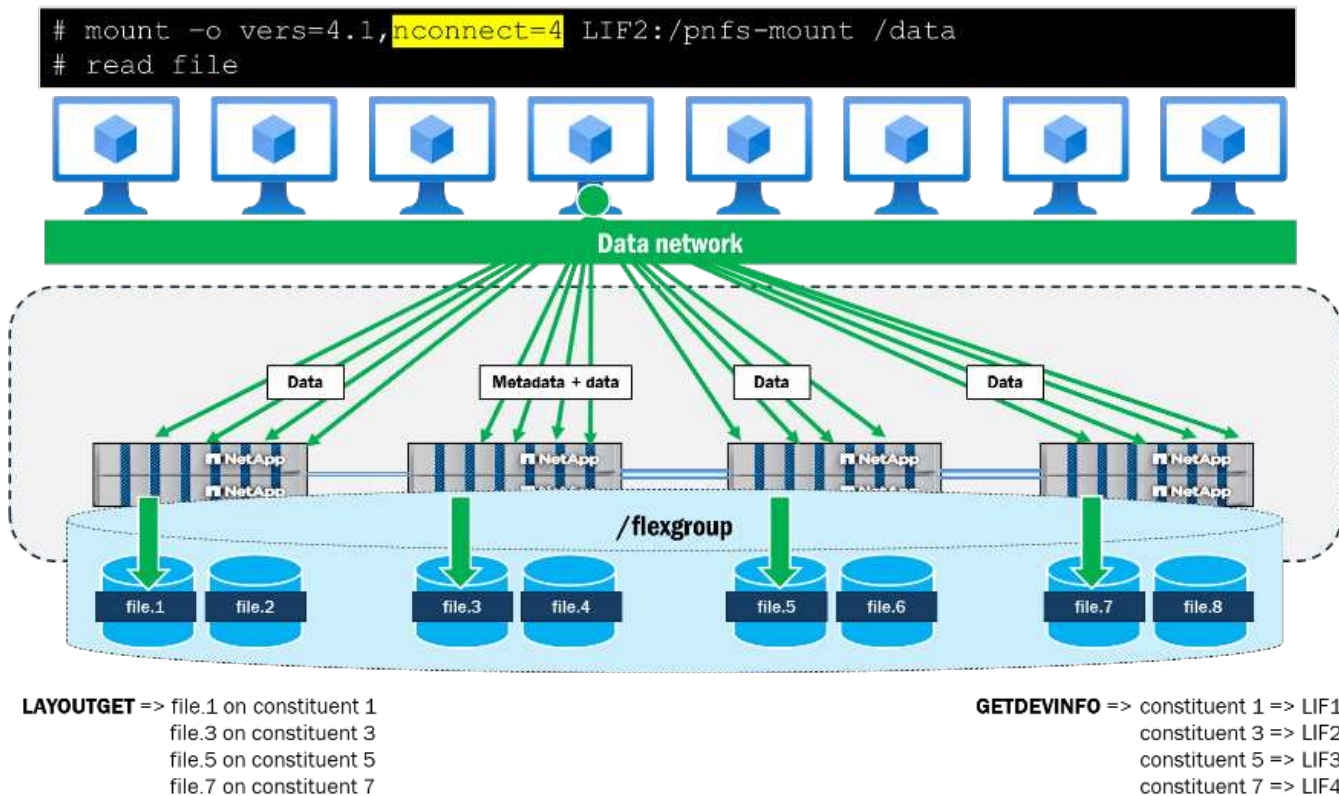


圖 9. pNFS 的 nconnect 設定為 4

["了解更多關於ONTAP對 NFSv4.1 的支持"](#)

pNFS 與 NFSv4.1 會話中繼

NFSv4.1 會話中繼 ("RFC 5661，第 2.10.5 節") 是指在客戶端和伺服器之間使用多個 TCP 連線來提高資料傳輸速度。ONTAP 9.14.1 新增了對 NFSv4.1 會話中繼的支持，必須與也支援會話中繼的用戶端一起使用。

在ONTAP中，會話中繼可以跨叢集中的多個節點使用，以提供額外的吞吐量和連接冗餘。

會話中繼可以透過多種方式建立：

- 透過掛載選項自動發現：大多數現代 NFS 用戶端可以透過掛載選項建立會話中繼（請查看您的作業系統供應商的文件），該選項會向 NFS 伺服器發出訊號，以便將有關會話中繼的資訊傳送回用戶端。此資訊以 NFS 資料包的形式出現。fs_location4 稱呼。

所使用的掛載選項取決於客戶端的作業系統版本。例如，Ubuntu Linux 的各種版本通常使用 max_connect=n 表示要使用會話中繼。在 RHEL Linux 發行版中，trunkdiscovery 使用了掛載選項。

Ubuntu 範例

```
mount -o vers=4.1,max_connect=8 10.10.10.10:/pnfs /mnt/pnfs
```

RHEL 範例

```
mount -o vers=4.1,trunkdiscovery 10.10.10.10:/pnfs /mnt/pnfs
```



如果您嘗試使用 `max_connect` 在 RHEL 發行版上，它將被視為 `nconnect`，會話中繼將無法如預期般運作。

- 手動建立：您可以透過將每個單獨的 IP 位址掛載到相同的匯出路徑和掛載點來手動建立會話中繼。例如，如果同一個節點上有兩個 IP 位址（10.10.10.10 和 10.10.10.11），用於匯出路徑，`/pNFS` 您需要運行兩次 `mount` 指令：

```
mount -o vers=4.1 10.10.10.10:/pNFS /mnt/pNFS
mount -o vers=4.1 10.10.10.11:/pNFS /mnt/pNFS
```

對所有要參與主幹的介面重複此程序。



每個節點都有自己的會話主幹。主幹不經過節點。



使用 pNFS 時，僅使用會話中繼或 `nconnect`。同時使用兩者會導致不良行為，例如只有元資料伺服器連線才能獲得 `nconnect` 的好處，而資料伺服器只能使用單一連線。

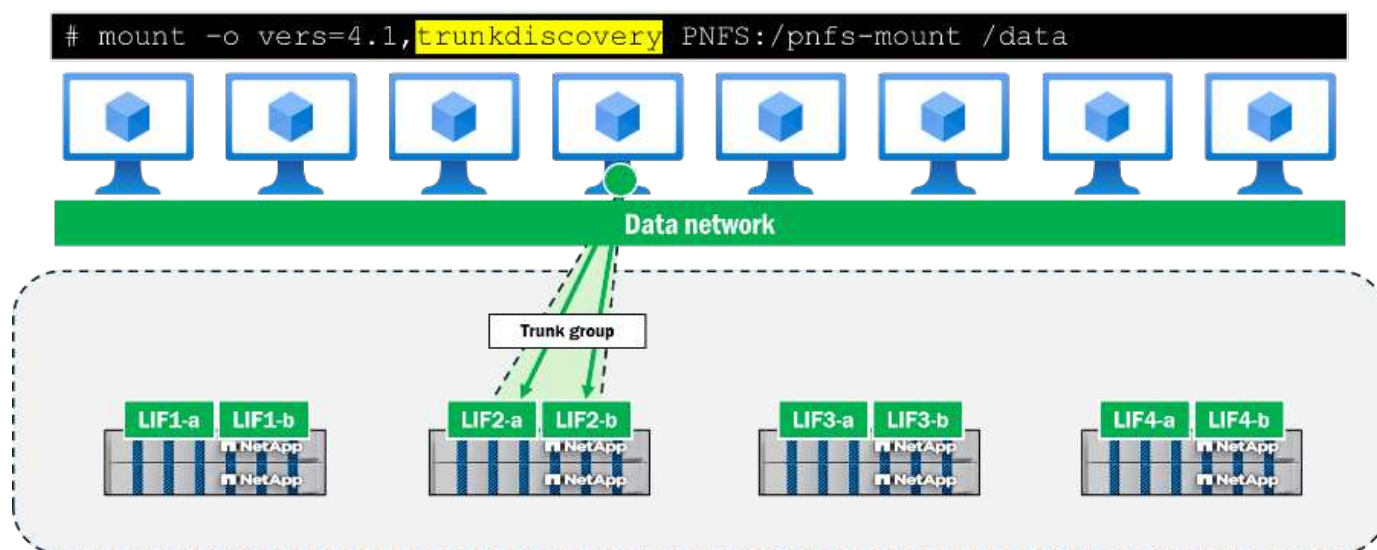
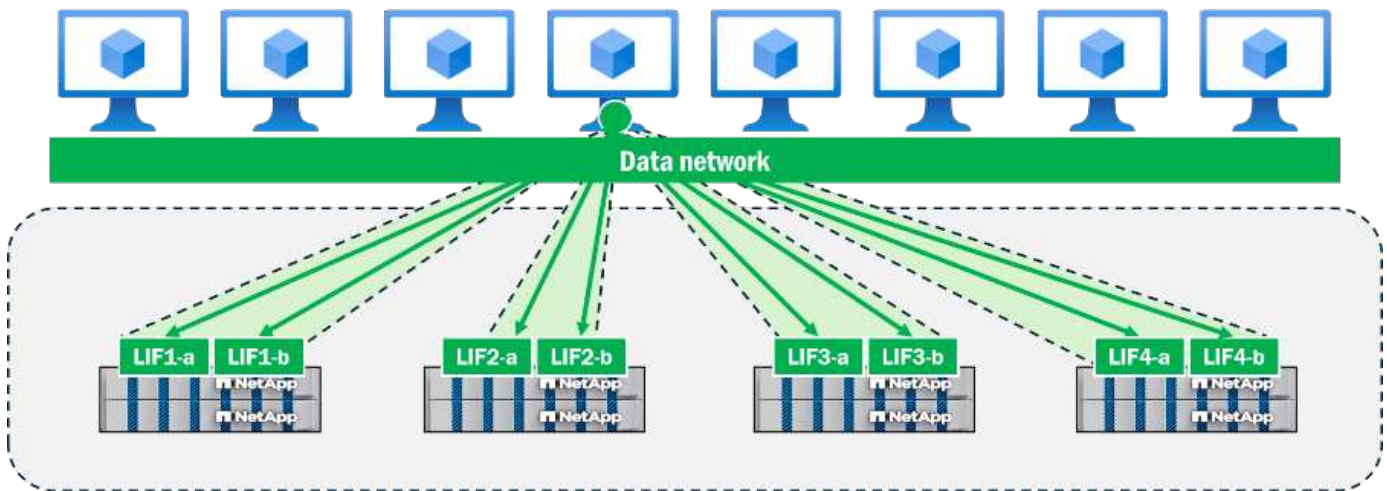


圖 10. ONTAP 中的 NFSv4.1 會話中繼

pNFS 可以為叢集中的每個參與節點提供本機路徑，並且當與會話中繼一起使用時，pNFS 可以利用每個節點的會話中繼來最大限度地提高整個叢集的吞吐量。

```
# mount -o vers=4.1, trunkdiscovery PNFS:/pnfs-mount /data
```



什麼時候 `trunkdiscovery` 如果使用，則會利用附加的 `GETATTR` 呼叫 (`FS_Locations`) 來取得 NFS 伺服器節點上所列的會話中繼接口，該節點是掛載介面所在的位置。一旦這些位址返回，後續的掛載點就會指向返回的位址。這可以從掛載過程中的資料包擷取中看出。

198	1.219372			NFS	246	V4	Call (Reply In 199)	GETATTR FH: 0x787f5cf1
199	1.219579			NFS	238	V4	Reply (Call In 198)	GETATTR


```

  ▾ Opcode: SEQUENCE (53)
    Status: NFS4_OK (0)
    sessionid: 7100001e004090a90000000000000409
    seqid: 0x00000000
    slot id: 0
    high slot id: 63
    target high slot id: 63
    > status flags: 0x00000000
  ▾ Opcode: PUTFH (22)
    Status: NFS4_OK (0)
  ▾ Opcode: GETATTR (9)
    Status: NFS4_OK (0)
  ▾ Attr mask: 0x01000100 (FSID, FS_Locations)
    ▾ reqd_attr: FSID (8)
      > fattr4_fsid
    ▾ reco_attr: FS_Locations (24)
      ▾ fattr4_fs_locations
        pathname components: 0
        ▾ fs_location4
          num: 1
        ▾ fs_location4
          ▾ servers
            num: 1
            ▾ server: 
              length: 14
              contents: 
              fill bytes: opaque data
            pathname components: 0
  
```

圖 11. 掛載期間的 NFS 會話主幹發現：資料包捕獲

["了解更多關於NFS中繼的信息"](#)

pNFS 與 NFSv4.1 轉診

NFSv4.1 參考提供了一種初始掛載路徑重新導向模式，該模式會在發出掛載請求時將用戶端引導至磁碟區的位

置。NFSv4.1 轉介功能在單一 SVM 內運作。此功能嘗試將 NFS 掛載點定位到與資料磁碟區位於相同節點上的網路介面。如果該介面或磁碟區在掛載到客戶端時移動到另一個節點，則資料路徑將不再本地化，直到建立新的掛載點為止。

pNFS 不會嘗試定位掛載路徑。相反，它會使用掛載路徑建立一個元資料伺服器，然後根據需要動態地本地化資料路徑。

NFSv4.1 引用可以與 pNFS 一起使用，但此功能沒有必要。啟用 pNFS 的轉診功能不會產生明顯效果。

"啟用或停用 NFSv4 引用"

pNFS與高階容量平衡的交互

"進階容量平衡" ONTAP會將檔案資料的一部分寫入FlexGroup磁碟區的各個組成磁碟區（單一FlexVol磁碟區不支援此功能）。隨著檔案的成長，ONTAP決定開始將資料寫入不同的組成磁碟區上的新的多部分 inode，該組成磁碟區可能位於同一節點上，也可能位於不同的節點上。對這些多 inode 檔案進行的寫入、讀取和元資料操作對客戶端是透明的，不會造成乾擾。先進的容量平衡技術改善了FlexGroup各組成磁碟區之間的空間管理，從而提供了更穩定的效能。

pNFS 可以根據 NFS 伺服器中儲存的檔案佈局訊息，將資料 I/O 重新導向到本機網路路徑。當一個大檔案被分成多個部分，分佈在多個組成磁碟區中，而這些磁碟區可能跨越叢集中的多個節點時，ONTAP中的 pNFS 仍然可以為每個檔案部分提供本機流量，因為ONTAP也維護了所有檔案部分的檔案佈局資訊。讀取檔案時，資料路徑的局部性會根據需要而改變。

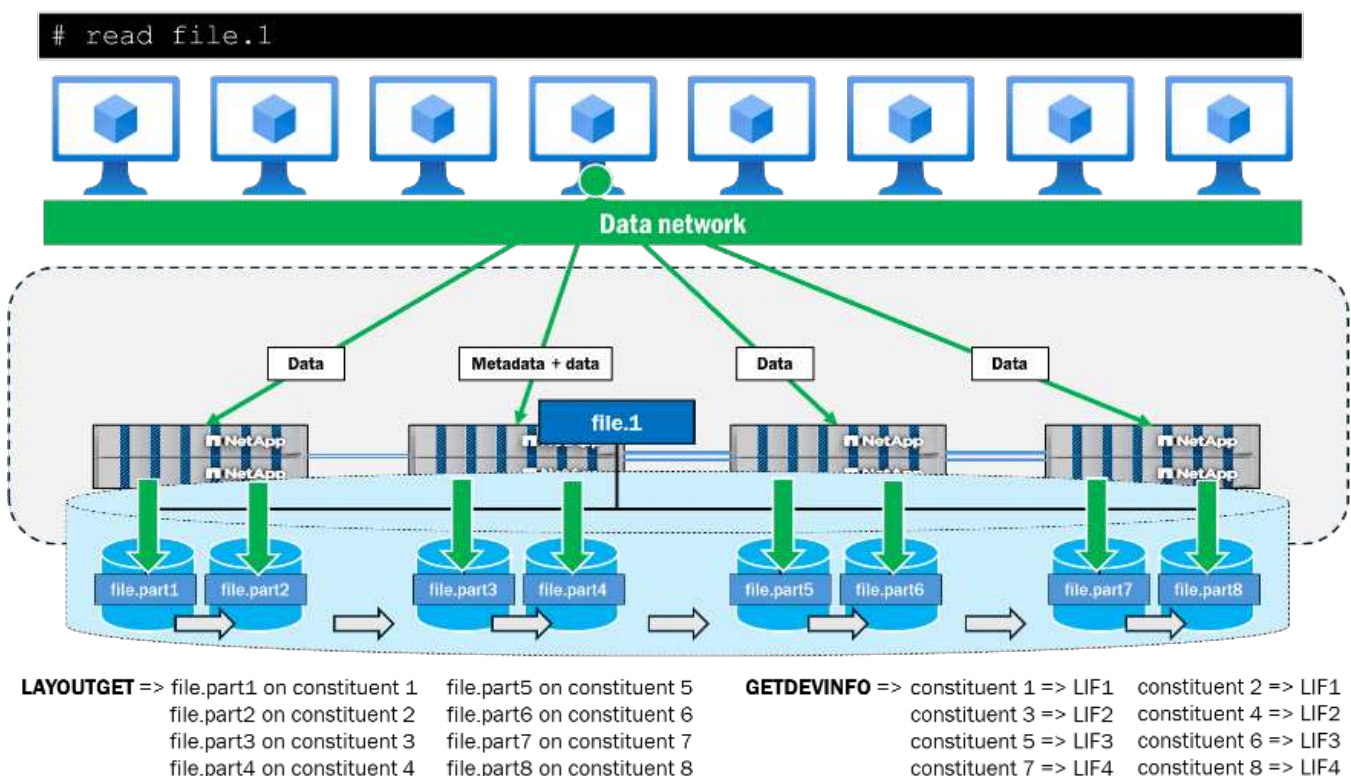


圖 12. 基於 pNFS 的高階容量平衡

相關資訊

- "FlexGroup卷配置"

ONTAP中的 pNFS 部署策略

pNFS 的引入是為了改進傳統的 NFS，它分離元資料和資料路徑，提供資料本地化，並支援平行操作。

傳統NFS的挑戰與pNFS的優勢

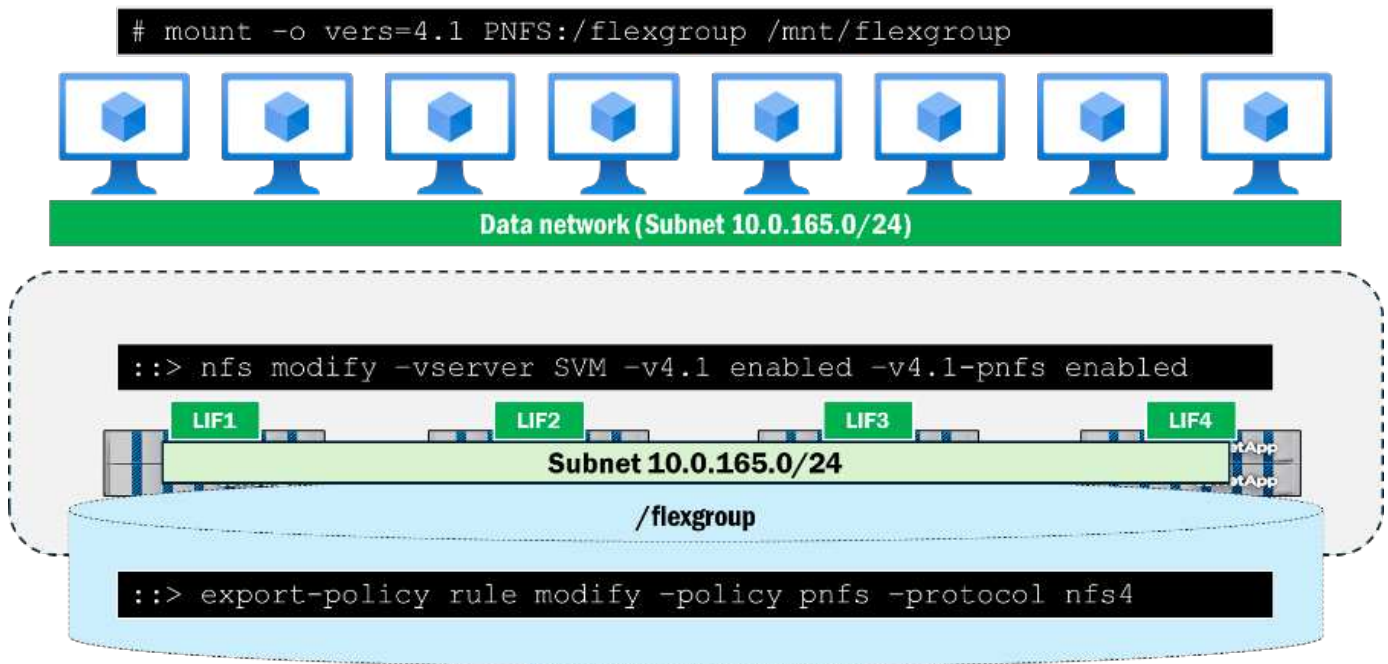
下表列出了傳統 NFS 面臨的挑戰，並解釋了ONTAP中的 pNFS 如何解決這些挑戰。

挑戰	pNFS 獲益
元資料和資料走同一條路徑 在傳統的 NFS 中，元資料和資料走同一條路徑，這可能會導致網路和 CPU 飽和，因為一條路徑會連接到叢集中的單一硬體節點。當許多用戶嘗試存取同一個 NFS 匯出時，這種情況會更加嚴重。	元資料和資料路徑分離，資料路徑並行化 透過分離 NFS 流量的元資料和資料路徑，並為資料路徑提供多條網路路徑，ONTAP叢集中的 CPU 和網路資源得到最大化利用，從而提高了工作負載的擴展性。
工作負載分配挑戰 在ONTAP NAS 叢集中，最多可以有 24 個節點，每個節點都可以有自己的一組資料捲和網路介面。每個磁碟區都可以承載自己的工作負載，或是工作負載的子集；而使用FlexGroup卷，為了簡化操作，工作負載可以存在於存取單一命名空間的多個節點上。當客戶端掛載 NFS 匯出時，網路流量將在單一節點上建立。當被存取的資料位於叢集中的單獨節點上時，就會發生遠端流量，這會增加工作負載的延遲和管理的複雜性。	本地並行資料結構路徑 由於 pNFS 將資料路徑與元資料分離，並根據磁碟區在叢集中的位置提供多個並行資料路徑，因此可以透過減少叢集中網路流量的距離以及利用叢集中的多個硬體資源來降低延遲。此外，由於ONTAP中的 pNFS 會自動重新導向資料流量，管理員無需管理多個匯出路徑和位置。
NFS 掛載點遷移 掛載點建立後，卸載並重新掛載磁碟區會造成乾擾。ONTAP提供了在節點之間遷移網路介面的功能，但這會增加管理開銷，並且會對使用 NFSv4.x 的有狀態 NFS 連線造成乾擾。重新定位掛載點的部分原因與資料本地性挑戰有關。	自動路徑重定位 使用 pNFS 時，NFS 伺服器會維護一個網路介面和磁碟區位置的表。當客戶端透過 pNFS 中的元資料路徑請求資料結構時，伺服器將向客戶端提供最佳化的網路路徑，然後客戶端將使用該路徑進行資料操作。這大大降低了工作負載的管理開銷，並且在某些情況下可以提高效能。

配置要求

在NetApp ONTAP中設定 pNFS 需要以下步驟：

- 支援 pNFS 且掛載於 NFSv4.1 或更高版本的 NFS 用戶端
- ONTAP中的 NFS 伺服器已啟用 NFSv4.1 (`nfs modify -v4.1 enabled` (預設為關閉))
- ONTAP中的 NFS 伺服器已啟用 pNFS (`nfs modify -v4.1-pnfs enabled` (預設為禁用))
- 每個節點至少需要一個可路由至 NFS 用戶端的網路介面。
- SVM 中具有允許 NFSv4 導出策略和規則的資料卷



滿足上述配置要求後，pNFS 即可自行運作。

相關資訊

- ["NFS 組態"](#)
- ["ONTAP對 NFSv4.1 的支持"](#)
- ["pNFS 的網路介面連接"](#)

規劃

pNFS部署計劃

在您的環境中部署 pNFS 之前，請確保滿足先決條件，並了解互通性要求和配置限制。

先決條件

在ONTAP中啟用和使用 pNFS 之前，請確保符合以下要求：

- NFS伺服器上已啟用NFSv4.1或更高版本
- 至少一個 ["每個節點都存在資料 LIF。"](#) 在託管 NFS 伺服器的 SVM 叢集中
- 全部 ["SVM 中的資料 LIF 是可路由的"](#) 致 NFS 用戶端
- NFS 用戶端支援 pNFS（2014 年及以後的大多數現代 Linux 發行版均支援）。
- 客戶端與SVM中所有資料LIF之間的網路連線正常
- 所有資料 LIF 的 DNS 解析（如果使用主機名稱）均已正確配置。
- ["資料量FlexGroup"](#) 已配置（建議使用以獲得最佳效果）
- ["NFSv4.x ID 域匹配"](#) 客戶與ONTAP之間
- ["NFS Kerberos"](#)（如果使用）在 SVM 中的所有資料 LIF 上啟用。

最佳實務總結

在您的環境中部署 pNFS 時，請遵循以下最佳實務：

- 使用 "資料量FlexGroup" 為了獲得最佳性能和容量擴展性
- 確保所有 "SVM中的網路介面是可路由的。" 致客戶
- "停用NFSv4.0" 確保客戶端使用 NFSv4.1 或更高版本
- 分散掛載點到多個網路介面和節點
- 使用輪詢 DNS "負載平衡元資料伺服器"
- 核實 "NFSv4.x ID 域匹配" 在客戶端和伺服器上
- 執行 "網路介面遷移" 和 "儲存故障轉移" 維護窗口期
- 使能夠 "NFS Kerberos" 如果使用 Kerberos 安全機制，則所有資料 LIF 都會受到影響。
- 避免使用 "NFSv4.1 轉介" 使用 pNFS 時
- 測試 "nconnect 設定" 謹慎操作，避免超出 TCP 連線限制
- 考慮 "會話中繼" 作為替代方案 "n 連接"（請勿同時使用兩者）
- 核實 "客戶端作業系統供應商支持" 部署前先對 pNFS 進行處理

互通性

ONTAP中的 pNFS 設計用於與符合 RFC 標準的 NFS 用戶端搭配使用。以下幾點要考慮：

- 最現代 "2014 年及以後的 Linux 發行版" 支援 pNFS（RHEL 6.4、Fedora 17 及更高版本）
- 請向您的用戶端作業系統供應商確認是否支援 pNFS。
- pNFS 可與FlexVol和 "資料量FlexGroup"
- pNFS 支援 NFSv4.1 和 "NFSv4.2"
- pNFS 可以與...一起使用 "NFS Kerberos"（krb5、krb5i、krb5p），但性能可能會受到影響
- pNFS 可以與...一起使用 "n 連接" 或者 "會話中繼"（但不能同時進行）
- pNFS 無法在 "NFSv4.0"

限制

以下限制適用於ONTAP中的 pNFS：

- "TCP 連線限制" 每個節點的具體限制因平台而異（請查看NetAppHardware Universe以了解特定限制）。
- 最大檔案大小：取決於磁碟區類型和ONTAP版本
- 最大文件數：最多 2000 億個文件 "資料量FlexGroup"
- 最大容量：高達 60 PB "資料量FlexGroup"
- "網路介面數量"每個節點至少需要一個資料 LIF；負載平衡可能需要更多。

使用時 "nconnect 與 pNFS"請注意，TCP 連線數會迅速成長：

- 每個使用 nconnect 的客戶端掛載都會為每個資料 LIF 建立多個 TCP 連線。
- 由於許多客戶使用了較高的 nconnect 值，["TCP 連線限制"](#) 可以超過
- TCP連線數超過限制會阻止建立新連接，直到現有連線被釋放為止。

相關資訊

- ["pNFS 的網路介面連接"](#)
- ["啟用或停用 NFSv4.1"](#)
- ["ONTAP對 NFSv4.1 的支持"](#)
- ["ONTAP對 NFSv4.2 的支持"](#)
- ["NetApp Hardware Universe"](#)

pNFS調優與效能最佳實踐

在ONTAP中使用 pNFS 時，請遵循以下注意事項和最佳實踐，以獲得最佳結果。

容量類型推薦

ONTAP中的 pNFS 可以與FlexVol捲和FlexGroup磁碟區一起使用，但為了獲得最佳的整體效果，請使用FlexGroup磁碟區。

FlexGroup卷提供：

- 單一掛載點可以跨越叢集中的多個硬體資源，同時允許 pNFS 本地化資料流量。
- 海量儲存容量（高達 60 PB）和高檔案數量（高達 2000 億個檔案）
- 支援多部分文件以實現容量平衡和潛在的性能優勢
- 支援對單一工作負載的捲和硬體進行並行訪問

["了解FlexGroup磁碟區管理"](#)

客戶推薦

並非所有 NFS 用戶端都支援 pNFS，但大多數現代客戶端都支援。RHEL 6.4 和 Fedora 17 是最早支援 pNFS 的用戶端（大約在 2014 年），因此可以合理地假設過去幾年發布的客戶端版本都完全支援此功能。ONTAP 對 NFS 的支援立場是「如果用戶端支援該功能且符合 RFC 標準，而我們也支援該功能，那麼這種組合就是受支援的」。但是，最佳實踐是確保客戶端作業系統供應商支援 pNFS。

成交量變動

ONTAP能夠以非中斷的方式在同一叢集中的節點或聚合之間移動捲，從而提供容量和效能平衡的靈活性。當ONTAP中發生磁碟區移動時，pNFS 設備對應會自動更新，以便在必要時通知客戶端使用新的捲到介面關係。

["了解如何移動卷"](#)

網路介面遷移

ONTAP能夠將網路介面遷移到同一叢集中的不同節點，從而實現效能平衡和維護靈活性。與磁碟區遷移類似，當ONTAP中發生網路介面遷移時，pNFS 設備對應會自動更新，以便在必要時通知用戶端使用新的捲到介面關係。

然而，由於 NFSv4.1 是有狀態協議，網路介面遷移可能會對正在積極使用 NFS 掛載的客戶端造成乾擾。最佳實踐是在維護視窗期間進行網路介面遷移，並通知客戶可能出現的網路中斷。

儲存故障轉移/恢復

pNFS 遵循與 NFSv4.1 相同的儲存故障轉移考量。這些內容在下文有詳細介紹。"[NetApp技術報告4067：NFS最佳實務與實作指南](#)"一般來說，任何涉及 pNFS 的儲存故障轉移/復原都應該在維護視窗內進行，由於協定的狀態性，可能會出現儲存中斷。

元資料工作負載

元資料操作的規模很小，但數量可能很大，具體取決於工作負載（您是否正在建立大量檔案？）您是否正在執行“查找”命令？）以及文件總數。因此，元資料呼叫量高的工作負載會佔用 NFS 伺服器大量的 CPU 資源，並可能導致單一連線瓶頸。pNFS（以及一般的 NFSv4.x）並不適合對效能要求較高的元資料工作負載，因為該協定版本的狀態性、鎖定機制以及某些安全特性會對 CPU 使用率和延遲產生負面影響。這些工作負載類型（例如高 GETATTR 或 SETATTR）通常在 NFSv3 中表現較好。

元資料伺服器

pNFS 中的元資料伺服器是在 NFS 匯出首次掛載時建立的。安裝點一旦建立，就會一直保持原位，直到重新安裝或移動資料介面為止。因此，最佳實務是確保多個客戶端存取同一磁碟區時掛載到 SVM 的不同節點和資料介面。這種方法可以實現跨節點和 CPU 資源的元資料伺服器負載平衡，同時最大限度地利用叢集中的網路介面。實現此目的的一種方法是建立輪詢 DNS 設置，這在下文中有所介紹。"[NetApp技術報告 4523：ONTAP中的DNS 負載平衡](#)"。

NFSv4.x ID 域

NFSv4.x 透過多種方式提供安全功能（詳見下文）。"[NetApp技術報告4067：NFS最佳實務與實作指南](#)"）。NFSv4.x ID 網域是實作 NFS 匯出時客戶端和伺服器必須就 ID 網域進行驗證的方法之一。ID 網域不匹配的副作用之一是，使用者或群組會顯示為匿名使用者（本質上是被壓制），以防止未經授權的存取。對於 NFSv4.x（以及 pNFS），最佳實踐是確保客戶端和伺服器上的 NFSv4.x ID 網域匹配。

n 連接

如前所述，ONTAP中的 nconnect 可以幫助提高某些工作負載的效能。對於 pNFS，需要了解的是，雖然 nconnect 可以透過大幅增加與儲存系統的 TCP 連線總數來提高效能，但當許多客戶端利用掛載選項時，它也可能透過使儲存上的 TCP 連接過載而造成問題。NetAppHardware Universe涵蓋了每個節點的 TCP 連線限制。

當節點的 TCP 連線數超過限制時，在現有連線釋放之前，不允許建立新的 TCP 連線。這可能會為可能遭遇暴風雪的地區帶來複雜情況。

下表顯示了使用 nconnect 的 pNFS 如何可能超出 TCP 連線限制：

客戶數量	nconnect 值	每個掛載點每個節點的潛在 TCP 連線總數
1.	4.	4.

客戶數量	nconnect 值	每個掛載點每個節點的潛在 TCP 連線總數
100	4.	400
1000	8.	8000
10000.	8.	80000
10000.	16.	160000 ¹

¹ 超出大多數ONTAP單節點 TCP 連線限制

NFSv4.1 工作階段主幹

ONTAP中的會話中繼可用於提高 NFSv4.x 掛載的吞吐量和路徑彈性。與 pNFS 一起使用時，叢集中的每個節點都可以建立會話中繼。但是，會話中繼需要每個節點至少兩個接口，而 pNFS 需要每個節點至少有一個接口才能按預期工作。此外，SVM 中的所有介面都必須能夠路由到 NFS 用戶端。同時使用 nconnect 時，會話中繼和 pNFS 無法正常運作。將 nconnect 和會話中繼視為互斥的功能。

"了解NFS中繼"

網路介面連接

pNFS 需要叢集中每個節點都具有可路由的網路介面才能正常運作。如果與託管 pNFS 的 NFS 伺服器位於同一 SVM 中，且存在其他無法路由到 NFS 用戶端的網路接口，則ONTAP仍會在設備映射中向客戶端通告這些接口。當 NFS 用戶端嘗試透過不同子網路中的介面存取資料時，將無法連接，導致服務中斷。最佳實踐是，在使用 pNFS 時，只允許 SVM 中的網路介面可供客戶端存取。



預設情況下，pNFS 要求 SVM 中的任何資料 LIF 都可路由到 NFS 用戶端上的接口，因為 pNFS 設備清單將填入 SVM 中的任何資料 LIF。因此，可能會選擇不可路由的資料 LIF，這可能會造成中斷情況。最佳實踐是，僅在使用 pNFS 時配置可路由資料 LIF。

從ONTAP 9.18.1 RC1 及更高版本開始，您可以按子網路指定哪些介面符合 pNFS 流量的條件，從而允許混合使用可路由介面和不可路由介面。請聯絡NetApp支援部門以取得有關指令的資訊。

NFSv4.0

NFSv4.0 是ONTAP NFS 伺服器中可以與 NFSv4.1 一起啟用的選項。但是，pNFS 不能在 NFSv4.0 上運作。如果 NFS 伺服器啟用了 NFSv4.0，則用戶端可能會在不知情的情況下掛載該協定版本，並且無法利用 pNFS。因此，在使用 pNFS 時，最佳實踐是明確停用 NFSv4.0。NFSv4.1 仍需啟用，並且可以獨立於 NFSv4.0 運作。

NFSv4.1 轉介

NFSv4.1 引用會將客戶端的掛載路徑本地化到擁有該磁碟區的節點上的網路介面。pNFS 在地化資料路徑，而掛載路徑則成為元資料伺服器。

雖然這兩個功能可以實際一起使用，但將 NFSv4.1 引用與 pNFS 一起使用可能會導致在同一節點上堆疊多個元資料伺服器，從而降低將元資料伺服器分散到多個叢集節點的能力，這種效果並不理想。在使用 pNFS 時，如果元資料伺服器沒有均勻分佈在叢集中，那麼單一節點的 CPU 可能會被元資料請求淹沒，從而造成效能瓶頸。

因此，在使用 pNFS 時，最好避免使用 NFSv4.1 引用。相反，應將掛載點分散到叢集中的多個網路介面和節點上。

"了解如何啟用或停用 NFSv4 轉儲"

NFS Kerberos

使用 NFS Kerberos，可以對 krb5 進行身份驗證加密，並進一步對封包進行 krb5i 和 krb5p 加密。在 SVM 中，此功能是基於每個網路介面啟用的，詳情請參見[此處]。"[NetApp 技術報告 4616：ONTAP NFS Kerberos in Some with Microsoft Active Directory](#)"。

由於 pNFS 可以將資料流量重定向到 SVM 中的節點和網路接口，因此 SVM 中的每個網路介面都必須啟用 NFS Kerberos 並使其正常運作。如果 SVM 中的任何網路介面未啟用 Kerberos，則 pNFS 在嘗試存取這些介面上的資料磁碟區時將無法正常運作。

例如，當在具有兩個網路介面（只有一個啟用了 Kerberos）的 pNFS 支援的 SVM 上使用並行 dd 運行讀取測試時，位於啟用了 Kerberos 的介面上的檔案效能良好，而位於未啟用 Kerberos 的介面上的節點上的檔案始終無法完成讀取。當兩個介面都啟用 Kerberos 時，所有檔案都能如預期運作。

只要在 SVM 的所有網路介面上啟用了 NFS Kerberos，NFS Kerberos 就可以與 pNFS 一起使用。請記住，由於封包的加密/解密，NFS Kerberos 可能會造成效能損失，因此最佳實踐是使用工作負載徹底測試 pNFS 與 NFS Kerberos，以確保任何效能損失都不會對工作負載造成過大的影響。

下面是一個在 RHEL 9.5 用戶端上使用 pNFS 並結合 krb5（身份驗證）和 krb5p（端對端加密）時的平行讀取效能範例。Krb5p 在這項測試中表現下降了 70%。

克爾貝羅斯口味	MB/秒	完成時間
KRB5	• File1-243 • File2-243 • File3-238 • File4-238	• File1-43 • File2-43.1 • File3-44 • File4-44.1
krb5p	• File1-72.9 • File2-72.8 • File3-71.4 • File4-71.2	• File1-143.9 • File2-144.1 • File3-146.9 • File4-147.3

["了解 NFS 中的 Kerberos，以實現強大的安全性"](#)

NFSv4.2

NFSv4.2 已加入 ONTAP 9.8 中，是目前可用的最新 NFSv4.x 版本 (RFC-7862)。NFSv4.2 沒有明確的選項來啟用/停用它。相反，它是與 NFSv4.1 一起啟用/停用的。（-4.1 enabled）。如果用戶端支援 NFSv4.2，則在掛載命令期間，除非另有指定，否則它將協商支援的最高 NFS 版本。minorversion=2 安裝選項。

ONTAP 中的 NFSv4.2 支援以下功能：

- 安全標籤（MAC 標籤）
- 延伸屬性
- 稀疏文件操作（FALLOCATE）

pNFS 隨 NFSv4.1 引入，但也受 NFSv4.2 及其相關功能的支援。

pNFS 指令、統計資料和事件日誌

這些 ONTAP CLI 指令專門用於 pNFS。您可以使用它們進行配置、故障排除和收集統計資料。

啟用 NFSv4.1

```
nfs modify -vserver SVM -v4.1 enabled
```

啟用 pNFS

```
nfs modify -vserver SVM -v4.1-pnfs enabled
```

顯示 pNFS 設備（進階權限）

```
pnfs devices show -vserver SVM
```

Vserver Name Generation	Mapping ID	Volume MSID	Mapping Status	
-----	-----	-----	-----	
SVM	17	2157024470	notavailable	2
SVM	18	2157024463	notavailable	2
SVM	19	2157024469	available	3
SVM	20	2157024465	available	4
SVM	21	2157024467	available	3
SVM	22	2157024462	available	1

顯示 pNFS 裝置對映（高階權限）

```
pnfs devices mappings show -vserver SVM
```

Vserver Name	Mapping ID	Dsid	LIF IP
-----	-----	-----	-----
SVM	19	2449	10.x.x.x
SVM	20	2512	10.x.x.y
SVM	21	2447	10.x.x.x
SVM	22	2442	10.x.x.y

擷取 **pNFS** 特有的效能計數器（進階權限）

```
statistics start -object nfsv4_1 -vserver SVM -sample-id [optional-name]
```

查看 **pNFS** 特有的效能計數器（進階權限）

```
statistics show -object nfsv4_1 -vserver SVM
```

查看 **pNFS** 特有計數器清單（進階權限）

```
statistics catalog counter show -object nfsv4_1 -counter *layout*|*device*
```

Object: nfsv4_1

Counter	Description
-----	-----
getdeviceinfo_avg_latency	Average latency of NFSv4.1 GETDEVICEINFO operations.
getdeviceinfo_error	The number of failed NFSv4.1 GETDEVICEINFO operations.
getdeviceinfo_percent	Percentage of NFSv4.1 GETDEVICEINFO operations.
getdeviceinfo_success	The number of successful NFSv4.1 GETDEVICEINFO operations.
getdeviceinfo_total	Total number of NFSv4.1 GETDEVICEINFO operations.
getdevicelist_avg_latency	Average latency of NFSv4.1 GETDEVICELIST operations.
getdevicelist_error	The number of failed NFSv4.1 GETDEVICELIST operations.
getdevicelist_percent	Percentage of NFSv4.1 GETDEVICELIST operations.
getdevicelist_success	The number of successful NFSv4.1 GETDEVICELIST operations.

getdevicelist_total operations.	Total number of NFSv4.1 GETDEVICELIST operations.
layoutcommit_avg_latency operations.	Average latency of NFSv4.1 LAYOUTCOMMIT operations.
layoutcommit_error operations.	The number of failed NFSv4.1 LAYOUTCOMMIT operations.
layoutcommit_percent operations.	Percentage of NFSv4.1 LAYOUTCOMMIT operations.
layoutcommit_success operations.	The number of successful NFSv4.1 LAYOUTCOMMIT operations.
layoutcommit_total operations.	Total number of NFSv4.1 LAYOUTCOMMIT operations.
layoutget_avg_latency operations.	Average latency of NFSv4.1 LAYOUTGET operations.
layoutget_error operations.	The number of failed NFSv4.1 LAYOUTGET operations.
layoutget_percent operations.	Percentage of NFSv4.1 LAYOUTGET operations.
layoutget_success operations.	The number of successful NFSv4.1 LAYOUTGET operations.
layoutget_total operations.	Total number of NFSv4.1 LAYOUTGET operations.
layoutreturn_avg_latency operations.	Average latency of NFSv4.1 LAYOUTRETURN operations.
layoutreturn_error operations.	The number of failed NFSv4.1 LAYOUTRETURN operations.
layoutreturn_percent operations.	Percentage of NFSv4.1 LAYOUTRETURN operations.
layoutreturn_success operations.	The number of successful NFSv4.1 LAYOUTRETURN operations.
layoutreturn_total operations.	Total number of NFSv4.1 LAYOUTRETURN operations.

查看 NFS 的活動網路連接

您可以使用以下命令驗證是否與 SVM 建立了多個 TCP 連線：network connections active show 命令。

例如，如果要查看 NFS 會話中繼，請尋找來自相同客戶端在每個節點的不同介面上的連線：

```
cluster::*> network connections active show -node cluster-0* -vserver PNFS
```

CID	Ctx	Vserver Name	Interface Name:Local	Port	Remote Host:Port	Protocol/Service
Node: node-01						
2304333128	14	PNFS	data1:2049		ubuntu22-224:740	TCP/nfs
2304333144	10	PNFS	data3:2049		ubuntu22-224:864	TCP/nfs
2304333151	5	PNFS	data1:2049		ubuntu22-226:848	TCP/nfs
2304333167	15	PNFS	data3:2049		ubuntu22-226:684	TCP/nfs
Node: node-02						
2497668321	12	PNFS	data2:2049		ubuntu22-224:963	TCP/nfs
2497668337	18	PNFS	data4:2049		ubuntu22-224:859	TCP/nfs
2497668344	14	PNFS	data2:2049		ubuntu22-226:675	TCP/nfs
2497668360	7	PNFS	data4:2049		ubuntu22-226:903	TCP/nfs

查看已連接客戶端的 **NFS** 版本信息

您也可以使用以下方式查看 NFS 連線：`nfs connected-clients show` 命令。請注意，此處顯示的客戶名單是過去 48 小時內有活躍 NFS 流量的客戶。空閒的 NFS 用戶端（即使仍然掛載）可能要等到存取掛載點時才會顯示出來。您可以透過指定以下參數來篩選這些客戶端，使其僅顯示最近造訪過的用戶端：`-idle-time` 特徵。

例如，要查看過去 10 分鐘內 pNFS SVM 中有活動的客戶：

```
cluster::*> nfs connected-clients show -vserver PNFS -idle-time <10m>
```

```
Node: node-01
```

Vserver	Data-IP	Local	Remote	Client-IP	Protocol	Volume	Policy	Idle-Time	Reqs	Reqs	Trunking
10.x.x.a	nfs4.2	PNFS_root	default	9m 10s 0	149	false	10.x.x.a	nfs4.2			
FG_0001	default	9m 10s 135847	0	false	10.x.x.b	nfs4.2	PNFS_root	default	8m 12s 0	157	false
10.x.x.b	nfs4.2	FG_0001	default	8m 12s 52111	0	false					

相關資訊

- ["了解ONTAP中的平行 NFS \(pNFS\)"](#)

版權資訊

Copyright © 2026 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。