解決方案簡介

# NETAPP NOVUS：野心是您唯一的界限

## 利用地球上最快的儲存設備，突破擴充極限




## 本文摘要

NetApp® Novus 是一種革命性的橫向擴充儲存架構，專為下一代 AI 基礎架構而打造，可滿足新雲端、主權 AI 供應商、超大規模企業和建構千兆瓦級 AI 工廠的大型企業所需的速度、並行性和命名空間規模。GPU 的效能不僅受限於模型大小或網路頻寬，還受限於資料基礎架構能夠以多快的速度為其提供資料。NetApp Novus 透過解耦的中繼資料架構、直接並行資料存取和通用命名空間消除了這個瓶頸，協助 GPU 環境在擴充時保持生產力。

閒置的 GPU 不僅是運算未被充分利用。在 AI 工廠的規模下，資本、電力和場地空間產生的是熱量，而不是 AI 產出。NetApp Novus 改變了儲存格局，讓許多架構於 ONTAP 的資料資源作為單一資料環境運作，同時中繼資料服務可獨立擴充，以支援並行與命名空間的成長。專為 Zettabyte 級檔案系統與 AI 基礎架構效率而打造，NetApp Novus 可協助營運人員持續為 GPU 供應資料，減少與資料相關的停滯，在無碎片化的情況下進行擴充，並提升大規模 AI 的經濟效益。

## 問題

想像一下，成千上萬的 GPU，整齊地排列在機架上，隨時準備就緒，卻在消耗稀缺的資金、電力和場地空間。它們是現代資料中心中最昂貴的資產之一，但只有當它們執行有用的工作時才能創造價值：訓練模型、調校模型、提供推理服務或推進代理式 AI 工作流程。在千兆瓦級 AI 基礎架構中，限制因素並不總是模型、GPU 或網路。越來越重要的是，資料基礎架構能否保證每個 GPU 都能獲得足夠的資料。GPU 等待資料的每一秒都會造成生產力損失和投資浪費。在許多大型 AI 環境中，資料瓶頸可能會導致昂貴的 GPU 基礎架構嚴重未充分利用，即使對 AI 輸出的需求持續成長。問題不僅僅是團隊需要「更快的儲存設備」。更深層的問題是架構上的：傳統的儲存系統並非設計用於獨立擴充中繼資料並行性、頻寬、容量和命名空間作業，以支援數萬個高度平行的 GPU 用戶端。

每個儲存系統都有一個由其控制器和互連決定的上限。透過更多獨立的系統突破該上限，營運商可能是用一個瓶頸來換取解決另一個瓶頸：更多的命名空間、更多的放置決策、更多的重新裝載、更多的營運孤島，以及更多讓容量或效能受困的方式。原本應該像一個統一的資料環境，卻變成了一系列儲存設備孤島，應用程式團隊和基礎架構團隊只好自行將它們拼接起來。

這種碎片化既是效能問題，也是營運效率問題。AI 團隊不應需要成為儲存設備工程師來決定資料儲存在哪裡、裝載哪個叢集、如何重新平衡工作負載，或如何在碎片化的基礎架構中協調檢查點。在人工智慧工廠規模下，營運複雜性本身就成為了一種擴充限制。

即使團隊設法突破了一個上限，他們也會遇到另一個上限：電力。每一瓦用於強制資料基礎架構跟上進度的功耗，都意味著 GPU 在進行生產性工作時少了一瓦可用功耗。為了彌補架構限制而過度建置儲存設備，可能會增加電力、冷卻、實體佔用空間和成本，而此時 AI 操作員正試圖在有限的電力範圍內最大化輸出。

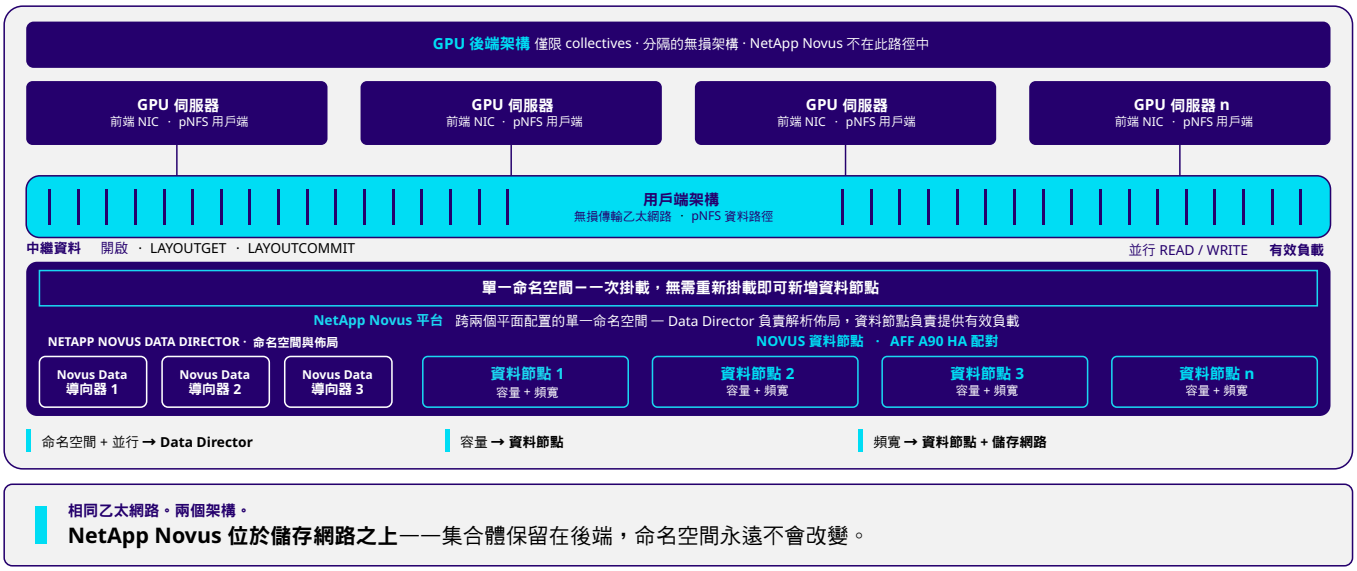
當儲存設備無法跟上需求時，GPU 就會閒置，電力就會浪費，原本用於產生智慧的機架會將稀缺的資金、電力和場地空間轉化為浪費的熱量，而不是 AI 輸出。在大規模 AI 中，使用率不僅僅是一項效能指標。這是經濟模型。

## 解決方案

NetApp Novus 將中繼資料層與儲存層解耦。中繼資料會追蹤跨聯合儲存系統的資料佈局，並將佈局傳遞給每個 GPU 用戶端，然後就不再干預。用戶端直接對基礎儲存設備進行讀寫操作，無需每次 I/O 操作都透過中繼資料進行往返。這就是突破單一儲存系統限制的關鍵。儲存層本身運行在獲證實的 NetApp ONTAP® 上，以線速提供資料，從而消除該瓶頸。這種方法可讓您以精細化的方式提升效能，並以線性方式增加效能。新增儲存系統，即可為聯合資源池增加效能容量。中繼資料也可以按照自己的路徑擴充，因此您可以獨立於儲存層來調整擴充規模。這意味著無需重新裝載，也不會產生碎片。

這一切都不需要新用戶端。NetApp Novus 透過大多數機組已經在執行的相同核心內 NFSv4.2 和 pNFS Flex Files 堆疊與 GPU 節點通訊，並具備 nconnect 以提供多連線處理量，以及 GPUDirect Storage 以提供完全跳過 CPU 的直接 GPU 到儲存設備 I/O。在數千台 GPU 伺服器的機組中，沒有任何專屬項目需要建置、合格與推出 —— NetApp Novus 就像 NFS 一樣裝載，而且每個用戶端都已經知道如何與之通訊。






**NetApp Novus 架構：**NetApp Novus 將中繼資料協調與資料搬移分離。GPU 用戶端從 NetApp Novus Data Director 中繼資料伺服器取得資料佈局，然後直接對基於 ONTAP 的儲存層執行 I/O，讓中繼資料並行性與儲存設備處理量能獨立擴充。

## NetApp Novus 的交付方式

NetApp Novus 作為整合式儲存解決方案交付，結合了解耦的中繼資料層與基於 ONTAP 的資料節點。在初始部署模型中，在 NetApp A90 系統上執行的 ONTAP 提供持久的資料服務基礎，而 pNFS 中繼資料伺服器服務則在合格的伺服器基礎架構上執行。中繼資料層向用戶端呈現命名空間和佈局；ONTAP 資料節點提供資料路徑。

這種方法為客戶提供針對 AI 和 GPU 即服務環境精心策劃的高效能部署模型，同時保有 NetApp Novus 的架構靈活度。目前的交付模式著重於基於 ONTAP 的 A90 資料節點和合格的 pNFS 中繼資料伺服器基礎架構。預計未來的部署模型將隨著時間擴充架構，包括軟體定義選項，其中中繼資料服務和基於 ONTAP 的資料服務都可以在合格的第三方硬體上執行。有關支援的配置、資格判斷詳情和部署要求，請參閱最新的 NetApp Novus 產品文件。

## 結論

人工智慧工廠的經濟效益取決於使用率。您的競爭對手擁有相同的 GPU 和相同的 AI 模型，但他們缺少的是能夠保持這些 GPU 高效運作的儲存層。NetApp Novus 消除了瓶頸，並將許多系統聯合成 Zettabyte 級檔案系統基礎架構，最終讓您最大的基礎架構投資產生價值和成果，而不是閒置。使用 NetApp Novus，每個 GPU 都能保持高效運作。

**與 NetApp Novus 架構師溝通**，將處理量和節點數量對應到您的 GPU 目標。

### 主要優勢

- **確保每個 GPU 都高效運作。**閒置的 GPU 是成本最高且無法銷售的資產。NetApp Novus 消除了儲存設備瓶頸，使更多運算能夠滿載運行，並將更多運算轉換為可計費工時。
- **在不造成命名空間碎片化的前提下擴充處理量。**NetApp Novus 將中繼資料協調與資料搬移分開，因此命名空間操作不會阻礙資料路徑。在共同命名空間下新增以 ONTAP 為基礎的資料資源以擴充彙總處理量，同時獨立擴充中繼資料層以支援更多用戶端、租戶和檔案操作。
- **提高人工智慧基礎架構效率。**NetApp Novus 旨在協助營運商在電力、場地空間、成本和營運的實際限制下進行擴充。基於 ONTAP 的資料階層提供有效率的儲存基礎，而軟體定義的中繼資料層則讓團隊僅在需要的地方新增中繼資料容量。未來的軟體定義資料服務將擴展這種靈活度，使營運商能夠圍繞適合其基礎架構的配置進行建置。
- **安全的多租戶和 QoS。**在單一 NetApp Novus 部署中執行多個租戶，彼此隔離，並採用租戶層級加密、原則型的資源分配，以及針對人員和 AI 代理程式的精細安全性控制。
- **以標準為基礎的用戶端。**無需建置、合格或修補專屬代理程式。NetApp Novus 裝載於 GPU 節點上已在執行的核心內 NFS 用戶端之上，因此無需進行用戶端推出專案，數千個機組即可準備好投入生產。

本解決方案簡介包含有關預期產品功能的前瞻性陳述。這些聲明反映了我們目前的計劃和預期，且主旨會有所變動。實際產品功能、功能和時間可能與所述有所不同，除適用法律要求外，我們不承擔交付任何特定功能或功能的義務。

本文件為機器翻譯，生成譯文僅供參考。若與英文版之間有任何矛盾或不一致之處，應以英文版內容為主。

**NetApp 台灣**

台北市 110 信義區松仁路 97 號 8 樓之 2 電話：886 2 8729 5000 傳真：886 2 8729 5050

聯絡業務人員

**關於 NetApp**

NetApp 是一家智慧型資料基礎架構公司，結合統一化資料儲存、整合式資料服務及 CloudOps 解決方案，將顛覆變動的世界轉化為每位客戶的大好商機。NetApp 打造無封閉環境的基礎架構，利用觀察能力及人工智慧來實現業界最理想的資料管理。作為唯一原生嵌入全球最大雲端平台的企業級儲存設備服務，我們的資料儲存設備可在各種環境中提供靈活度。我們的資料服務透過卓越的網路彈性、治理和應用程式敏捷度創造資料優勢，同時我們的 CloudOps 解決方案提供效能和效率的持續優化。無論資料類型、工作負載，或環境為何，您都可以利用 NetApp 來推動資料基礎架構轉型，實現更多商業契機。**www.netapp.com**

© 2026 NetApp, Inc. 版權所有。NETAPP、NETAPP 標誌及 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。SB-4532-0926-zhTW

4-22sep26