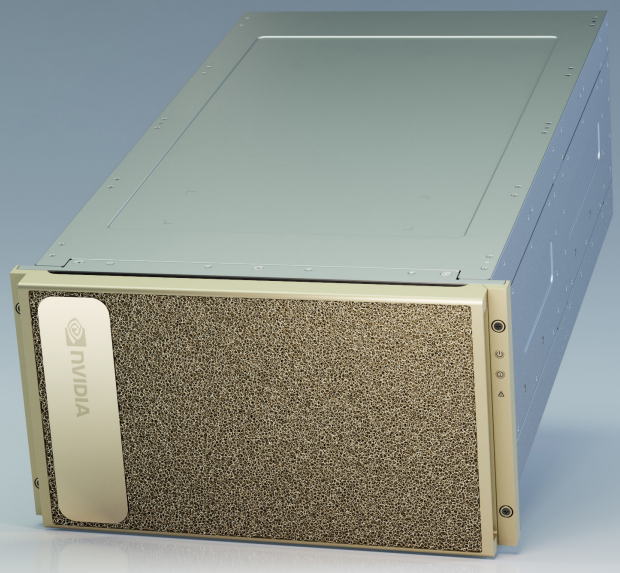
NVIDIA DGX A100
適用於 AI 基礎架構的通用系統



擴充企業 AI 的挑戰

每家企業都必須利用人工智慧（AI）轉型，才能在充滿挑戰的時代生存並茁壯。然而，企業需要改進傳統方法的 AI 基礎架構平台，其採用緩慢的運算架構，這些架構因分析、訓練和推論工作負載而孤島化。舊方法產生複雜問題、增加成本、限制規模速度且不適用於現代 AI。企業、開發人員、資料科學家和研究人員需要整合所有 AI 工作負載的新平台，以簡化基礎架構並加快 ROI。

適用於所有 AI 工作負載的通用系統

NVIDIA DGX™ A100 是適用於所有 AI 工作負載的通用系統 — 從分析到訓練以至推論。DGX A100 樹立運算密度的新基準，將 5 petaFLOPS 的 AI 效能融入 6U 規格，以單一整合式系統取代傳統的運算基礎架構。DGX A100 還能提供前所未有的效能，利用 NVIDIA A100 Tensor 核心 GPU 的多執行個體 GPU 功能精細分配運算能力，讓管理員能夠分配適合特定工作負載的資源。如此可確保最大、最複雜的工作，以及最小、最簡單的工作都受到支援。利用來自 NGC 的最佳化軟體執行 DGX 軟體堆疊，將密集的運算能力與完整的工作負載彈性結合，使 DGX A100 成為單節點部署，和透過 NVIDIA DeepOps 部署的大規模 Slurm 和 Kubernetes 叢集的理想選擇。

NVIDIA DGXpert 直接支援

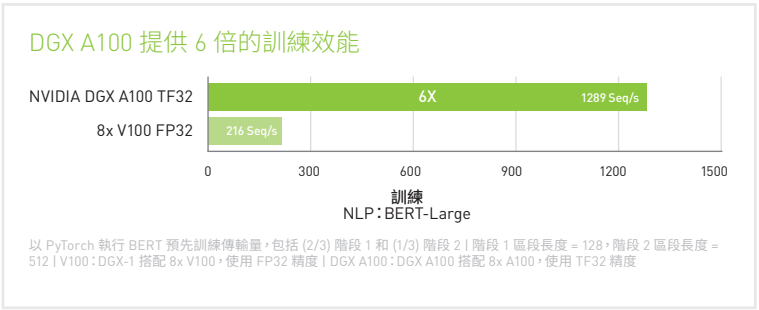
NVIDIA DGX A100 不僅只是伺服器，而是完整的軟硬體平台，此平台以從全球最大的 DGX 試驗場 — NVIDIA DGX SATURNV 獲得的知識為基礎，並且以數千名 NVIDIA DGXpert 為後盾。DGXpert 是精通 AI 的專家，提供規範性指導和設計專業知識以加快 AI 轉型。其在過去十年內累積豐富的知識和經驗，協助您將 DGX 投資的價值最大化。DGXpert 可確保關鍵應用程式快速啟用並維持流暢運作，大幅縮短取得洞見的時間。

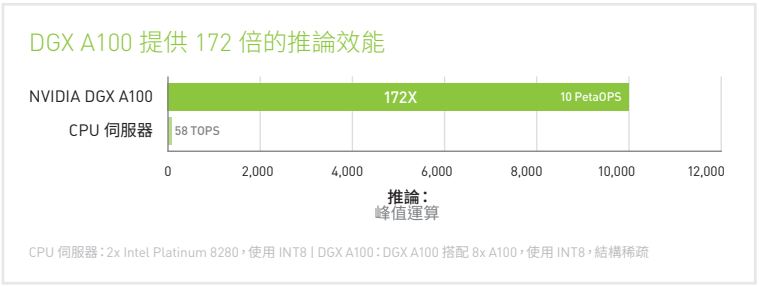
系統規格

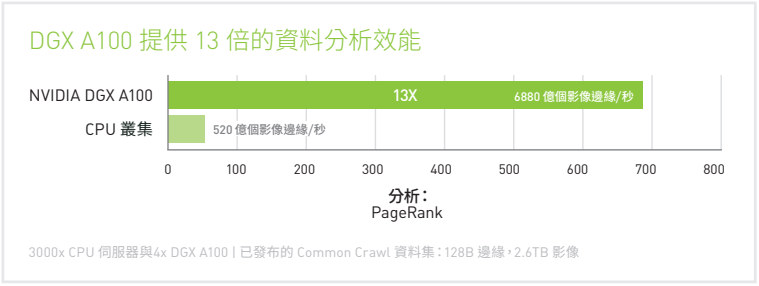
GPU	**8x NVIDIA A100 Tensor 核心 GPU**
GPU 記憶體	**共 320 GB**
效能	**5 petaFLOPS AI** **10 petaOPS INT8**
NVIDIA NVSwitches	**6**
系統電源用途	**最大 6.5kW**
CPU	**雙 AMD Rome 7742，共 128 核心，2.25 GHz（基礎）、3.4 GHz（最大提升）**
系統記憶體	**1TB**
網路	**8x 單連接埠 Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand** **1x 雙連接埠 Mellanox ConnectX-6 VPI 10/25/50/100/200Gb/s 乙太網路**
儲存裝置	**OS: 2x 1.92TB M.2 NVME drives** **作業系統：2x 1.92TB M.2 NVME 磁碟機** **內部儲存空間：15TB（4x 3.84TB）U.2 NVME 磁碟機**
軟體	**Ubuntu Linux OS**
系統重量	**271 lbs (123 kgs)**
含包裝系統重量	**315 lbs (143kgs)**
系統尺寸	**高度 10.4 in（264.0 mm）** **寬度最大 19.0 in（482.3 mm）** **長度最大 35.3 in（897.1 mm）**
工作溫度範圍	**5°C 至 30°C（41°F 至 86°F）**

最快解決時間

NVIDIA DGX A100 搭載八個 NVIDIA A100 Tensor 核心 GPU，為使用者提供無與倫比的加速，並且針對 NVIDIA CUDA-X™ 軟體和端對端 NVIDIA 資料中心解決方案堆疊，全面進行最佳化。NVIDIA A100 GPU 帶來新的精度 TF32，其運作方式與 FP32 相同，同時為 AI 提供比上一代高 20 倍的 FLOPS，而且最棒的是，不必修改任何程式碼即可獲得此加速。使用 NVIDIA 的自動混合精度時，A100 只需使用 FP16 精度時增加一行程式碼，即可進一步將效能提升 2 倍。A100 GPU 還具有領先同級的每秒 1.6 兆位元組數（TB/s）記憶體頻寬，比上一代增加超過 70%。此外，A100 GPU 擁有更多晶片內建記憶體，包括比上一代大將近 7 倍的 40MB Level 2 快取，將運算效能最大化。DGX A100 也首次搭載下一代 NVIDIA NVLink™，將 GPU 至 GPU 直接頻寬加倍，達到每秒 600 十億位元組數（GB/s），比 PCIe Gen 4 高將近 10 倍，以及新的 NVIDIA NVSwitch，比上一代快 2 倍。此前所未有的效能提供最快解決時間，讓使用者解決過去無法克服或難以克服的挑戰。







世界上最安全的企業用 AI 系統

NVIDIA DGX A100 為 AI 企業提供最穩健的安全態勢，以多層方法保護所有主要硬體和軟體元件。包含基板管理控制器（BMC）、CPU 機板、GPU 機板、自我加密磁碟機和安全啟動的 DGX A100 內建安全性，讓 IT 專注於啟用 AI，而不必花時間進行威脅評估和緩解。

以 Mellanox 造就無與倫比的資料中心擴充性

NVIDIA DGX A100 擁有所有 DGX 系統中最快的 I/O 架構，是大型 AI 叢集的基礎構件，例如可擴充 AI 基礎架構的企業藍圖 — NVIDIA DGX SuperPOD™。DGX A100 搭載 8 個單連接埠 Mellanox ConnectX-6 VPI HDR InfiniBand 介面卡（用於叢集化）、以及 1 個雙連接埠 ConnectX-6 VPI 乙太網路介面卡（用於儲存和網路），全部都能達到 200Gb/s。DGX A100 將龐大的 GPU 加速運算效能與最先進的網路軟硬體最佳化結合，可擴充至數百或數千個節點以解決最大的挑戰，例如對話式 AI 和大規模影像分類

與可靠的資料中心領導者共同打造經驗證的基礎架構解決方案

我們與儲存和網路技術的領導供應商合作，提供結合 NVIDIA DGX POD™ 參考架構優點的各種基礎架構解決方案。我們透過 NVIDIA 合作夥伴網路提供完全整合且可立即部署的解決方案，使 IT 的資料中心 AI 部署變得更簡單、更快速。

欲深入瞭解 NVIDIA DGX A100，請瀏覽 **www.nvidia.com/zh-tw/data-center/dgx-a100/**

NVIDIA