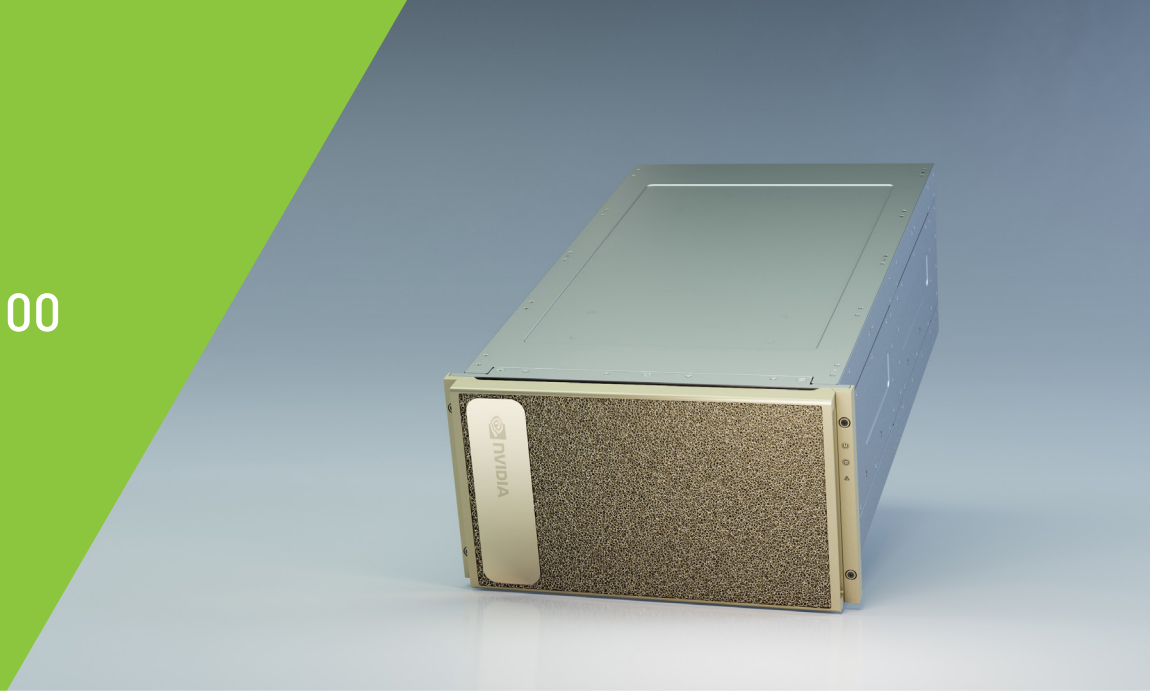
NVIDIA DGX A100
用於建構對話式 AI 應用程式



對話式人工智慧興起

- 預計至 2023 年，全球在對話式人工智慧（AI）上的支出將超過 138 億美元。[1] – IDC
- 採用對話式人工智慧（AI）的早期效益，為產生平均 30% 的商業價值年成長率。[2] – Deloitte Digital

建構業界領先對話式人工智慧的挑戰

採用對話式人工智慧（AI）的成功關鍵，在於實現自然的擬人化互動。其需要具備情境感知、理解情感的能力以及同時進行對話的能力—且皆在短短幾毫秒內發生。此外，為了實現 ROI，開發人員必須擁有人工智慧專業知識、獲得大量特定產業或產品資料的能力，以及加快模型迭代和提高準確度的基礎架構與工具。NVIDIA® DGX™ A100 提供了實現技術領先的對話式人工智慧所需要的高效能人工智慧基礎架構，讓企業能更容易建置具有超人般語言理解能力的人工智慧助理、通訊應用程式和聊天機器人。

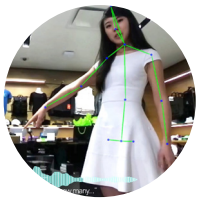
善用現成工具與最佳化模型

NVIDIA Jarvis 是為了建構端對端 GPU 最佳化對話式人工智慧服務而開發的應用程式框架，其中包含數百個預先經過訓練的模型。Jarvis 具有針對語音、視覺和自然語言理解（NLU）的最佳化端對端管道。這些模型使用了 **NVIDIA DGX 系統**，並在開放和專有資料集上進行訓練，訓練時間超過 100,000 小時。開發人員可以用簡單的 API 搭配 DGX 系統上的 **NVIDIA NeMo** 工具套件，以特定領域的資料進行微調。他們也可以使用 NeMo，從零開始建構和訓練最先進的模型，例如 Quartznet、Jasper、BERT、Tacotron2 與 WaveGlow。建立出融合語音與視覺的多模態技能，即可在對話式人工智慧應用程式上實現更自然的互動。

實現最高準確度和擬人式對話

NVIDIA DGX A100 搭載八個 NVIDIA A100 Tensor 核心 GPU —為有史以來技術最領先的資料中心加速器。第三代 Tensor 核心已進行最佳化，可加快矩陣乘法的計算速度，且成為人工智慧訓練和推論的關鍵。Tensor Float（TF32）精度可支援更大的量級和精度，相較於前幾代的人工智慧技術，其效能提高了 10 倍—且不需要變更任何程式碼。由於自然語言處理（NLP）的輸入矩陣通常大而稀疏，當可用單字的數量僅佔字典中的一小部分時 3，在各種常見的人工智慧網路中，A100 GPU 中新的結構稀疏功能可將運算速度加快 2 倍。第三代 NVIDIA NVLink®、NVIDIA NVSwitch™ 和 NVIDIA Mellanox InfiniBand 可在所有 GPU 之間提供超高頻寬和低延遲連線，並能擴充多個 DGX A100 系統，以訓練最大的自然語言處理模型。

視訊會議翻譯、轉錄
每天 2 億場會議



商場智慧助理
1200 萬間零售商店

客服中心
每天 5 億通電話

車載智慧助理
每年 7500 萬輛新車

智慧喇叭
每年售出 1.5 億台

提供互動式回應

即時效能和擬人式對話的延遲閾值小於 300 毫秒（ms）。利用 NVIDIA Jarvis 對話式人工智慧框架，開發人員可以將技術最領先的推論模型最佳化，並提供在 150 ms 內執行的即時服務（在 CPU 的平台上僅需要 25 秒），使效能提升 160 倍。在 NVIDIA Triton 推論伺服器上提供多種模型，利用 NVIDIA TensorRT 最佳化有效運作，並透過 Kubernetes 叢集上的 Helm chart，使用單一命令部署服務。利用**多執行個體 GPU（MIG）**創新，將傳輸量最大化，在單一 DGX A100 上支援多達 56 個同時推論伺服器－每個伺服器在硬體層級都完全隔離，且具有高頻寬記憶體、快取和運算核心。

通用對話式人工智慧平台：從原型到生產

NVIDIA DGX A100 是適用於人工智慧基礎架構的通用系統，範圍從分析到訓練再到推論。DGX A100 樹立了運算密度的新基準，將 5 petaFLOPS 的人工智慧效能融入 6U 規格中，以全能的單一整合式系統取代不靈活的傳統運算基礎架構。使用多達 8 個 NVIDIA A100 Tensor 核心 GPU，在 DGX A100 上訓練並微調大型模型（例如 Megatron-BERT），或將每一個 GPU 分成 7 個獨立的執行個體，以執行推論。此 MIG 創新讓使用者可以在同一個系統上同時混搭多個訓練和推論工作，善用專用資源實現最佳的利用率。

透過 DGXpert 專業建議，更快邁向成功

NVIDIA DGX A100 為完整的軟硬體平台，有數千位 NVIDIA 人工智慧專家作為後盾，並以獲得全球最大 DGX 試驗場－ NVIDIA DGX SATURNV 的知識作為發展基礎。擁有 DGX A100，即可直接造訪由人工智慧專家所組成的全球團隊 **NVIDIA DGXpert**，取得平台中所提供的規範性指導和設計專業知識，協助加快人工智慧轉型。這將可確保關鍵性應用程式的快速啟用及流暢運作，並大幅縮短獲得深入見解所需的時間。

大規模實現最高準確度

S&P Global 創新中心 **Kensho** 使用 NVIDIA 的對話式人工智慧框架，開發出適用於金融和商業的語音辨識解決方案。他們利用在 DGX A100 系統架構設計叢集 NVIDIA DGX SuperPOD™ 上訓練的模型，將轉錄財報電話和金融語音的準確度提高 20%，完全超越了商業解決方案。

簡易三大步驟

步驟 1
運用 **NGC** 的預先訓練模型（在 DGX 系統上訓練 100,000 小時）

步驟 2
重新訓練並微調您在 DGX A100 上的資料模型

步驟 3
透過 Kubernetes 叢集上的 Helm chart，以僅僅一行程式碼輕鬆部署服務

深入瞭解 NVIDIA DGX A100：**www.nvidia.com/dgxa100**

探索 NVIDIA Jarvis 的對話式人工智慧：**developer.nvidia.com/nvidia-jarvis**

1 David Schubmehl。IDC Worldwide Artificial Intelligence Software Platforms Forecast，2020-2024。2020 年 6 月。Market Forecast - Doc # US45724520

2 Deloitte。Conversational AI:The Next Wave of Customer and Employee Experiences。2019 年第 4 季。

3 Luis Filipe Kopp、José Barbosa da Silva Filho、Claudio Miceli de Farias 及 Priscila Machado Vieira Lima。Modeling Sparse Data as Input for Weightless Neural Network。2019 年 4 月。

ABC Partner

NVIDIA