DELL Technologies

AMD
together we advance_

2024 | 白皮書

透過 AI 賦予企業能力：進入選擇的時代

目錄

DELL Technologies

AMD
together we advance_

2024 | 白皮書

利用 AI 實現產業轉型的商機

在現今，AI 是您實現企業轉型以迎向未來創新的最佳機會。從 Accenture Vision Technology 2023 收集的資料顯示，98% 的全球高階主管認為，在未來三到五年內，AI 基礎模型將在其組織策略中發揮重要作用[1]。

AI 能夠提高工作效率、推動創新和改進決策流程，因此對零售、保健和金融服務等領域的企業非常有用。然而，儘管有這些優勢，但由於一些常見的誤解，在整合 AI 時仍然讓人感到障礙。

您需要一個 AI 開發人員團隊才能開始：

雖然資料科學方面的專業知識對於開發進階 AI 解決方案和瞭解基本原理仍然很有價值，但它不再是必要條件。簡單易用的 AI 工具數量大增，Hugging Face 等平台和工作專屬的模型降低了開發 AI 解決方案所涉及的大部分複雜性。

您需要在硬體上花費數千萬才能獲得結果：

這種誤解嚴重破壞了現今可用 AI 資源的多樣性。雖然這些眾所周知的資源通常功能強大且獲得良好支援，但它們不一定是每個企業最合適或最符合成本效益的選擇。

您需要不懈努力才能獲得加速器：

雖然加速器在繁重的 AI 工作負載上表現出色，但企業的 AI 應用程式可能並不需要那麼多運算能力。需要過長的等待時間才能獲得領先市場的加速器，也很不切實際。在許多情況下，AI 最佳化的 CPU 可實際提供即時產生 AI 輔助分析和決策所需的效能和效率，並且是一種更具成本效益的適應性解決方案。

[1] Accenture，2023 年 3 月 30 日，《Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds》(Accenture Technology Vision 2023：生成式 AI 將開創業務的大膽新未來，並融合實體世界和數位世界)，https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds

幸運的是，AI 局勢正在不斷發展。**Dell** 與 **AMD** 攜手合作，透過專為支援現今 AI 需求而設計的端對端基礎結構，讓更多使用者可使用 AI 技術和工具，以打破這些誤解。

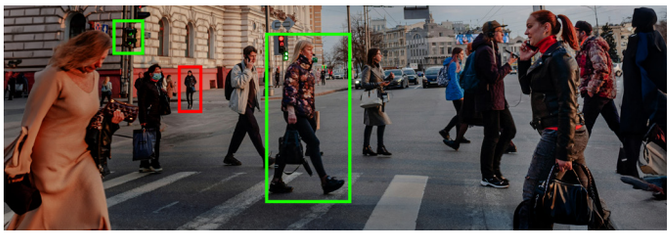
您可以開始使用已經過最佳化的模型、可靠的軟體堆疊，以及多功能的硬體系統，所有這些都可透過 Dell 和 AMD 的合作關係公開取得。利用 AI 不再需要使用越來越少的加速器、大量熟練的 AI 工程師或用於部署大規模雲端叢集的資源。

Dell 與 **AMD** 的合作提供統一的軟硬體生態系統，旨在讓開發人員建立端對端的 AI 解決方案，輕鬆有效地結合轉換學習、微調和推斷工作。在 **Hugging Face** 的支援下，我們現在擁有不斷成長的模型產品組合，可在搭載 AMD EPYC™ 處理器或 AMD Instinct™ MI300X 加速器的 Dell PowerEdge 伺服器上執行，讓開發人員能夠微調、套用轉換學習，並部署以進行推斷工作。對 AMD ROCm™ 和 AMD ZenDNN™ 的投資，以及與 PyTorch、Tensorflow 和 ONNX Runtime 框架的合作關係，是應用 AI 開發人員體驗 AI 普及化的基本實現要素。以下堆疊圖表詳細說明構成 Dell 和 AMD 統一 AI 生態系統的要素。

產業中的 AI

隨著資源的多樣化和對開放原始碼創新的重視，AI 正在遷移到許多不同的產業，包括客戶服務、金融和銀行、保健和零售等。然而，在這些產業中，AI 透過處理資料分析、自動化、個人化和預測分析等關鍵功能，共同使組織能夠釋放其專屬資料的潛力，並重新構想其 AI 工作流程。AMD ROCm 和 ZenDNN 程式庫還可加速這些 AI 工作流程，以近乎即時的方式提供結果。

深入瞭解 AI 究竟如何影響下列各個產業。



汽車業

AI 用於自動駕駛汽車的物體偵測、車道追蹤和決策。AI 還可以預測車輛元件何時可能發生故障，從而實現主動維護並減少停機時間。

製造業和工業

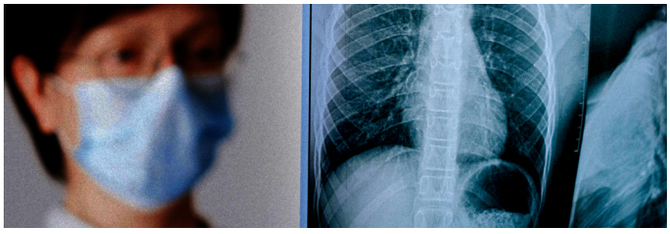
AI 可用於製造業和工業的預測性維護、品質控制、流程最佳化和供應鏈管理，從而提高效率並減少停機時間。

零售業

AI 可以分析客戶行為，以提供個人化的產品推薦，提高客戶忠誠度和銷售額。AI 還可以透過預測需求，以及盡可能減少存貨過多或缺貨情況，來最佳化庫存量。

金融服務

AI 可用於金融和銀行業，進行欺詐偵測、風險評估、客戶服務和投資分析，從而提高安全性和做出更明智的決策。



醫療

AI 可用於保健領域的各種應用，包括醫學影像分析、疾病診斷、個人化治療規劃和藥物探索，從而改善患者治療效果並降低成本。

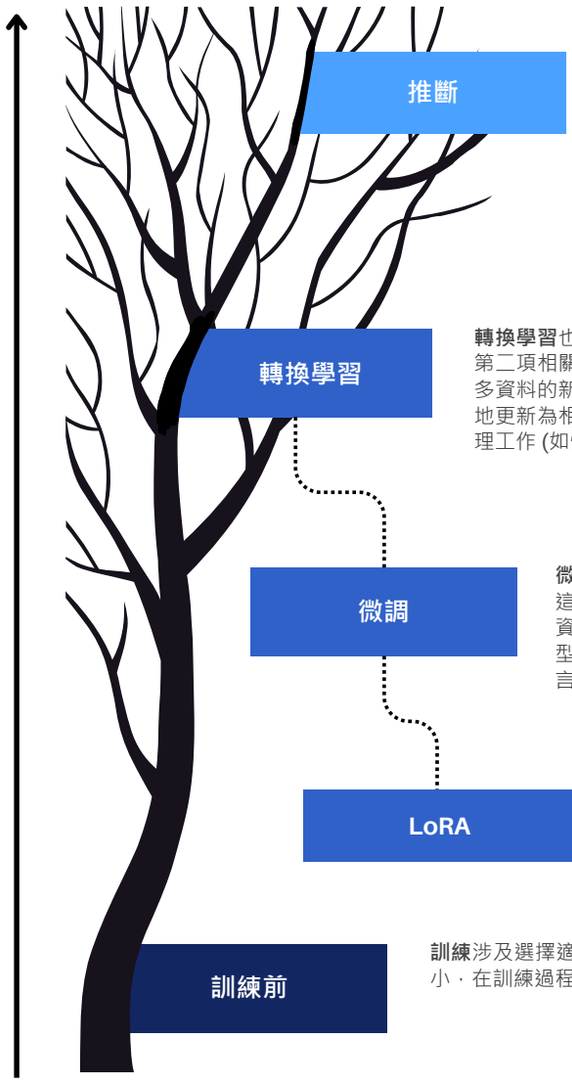
服務自動化

採用 AI 技術的聊天機器人可以處理客戶查詢並提供支援，減少對人力介入的需求。AI 還可以自動執行重複性工作，例如資料輸入或文件處理，進而提高效率並減少錯誤。

IT 決策者必須考慮的事項

入門：解析 AI

在瀏覽這些使用案例之前，讓我們更深入瞭解 AI 生命週期。AI (人工智慧) 生命週期是指開發、部署和維護 AI 系統所涉及的階段。雖然具體的方法和術語可能會有所不同，但典型的 AI 生命週期通常包括模型訓練和推斷。



推斷是使用經過訓練的 AI 模型，根據不明新資料進行預測、分類或決策的過程。在 AI 生命週期的此階段，將在生產環境中部署和使用經過訓練的模型，以執行其設計要執行的特定工作。在推斷過程中，模型會取得輸入資料，並使用從訓練階段學習的模式和關係來處理這些資料，然後在 GenAI 工作負載中產生輸出內容，例如預測、分類或文字。接著，此輸出內容可用於各種應用，例如向使用者推薦產品、偵測系統中的異常或識別影像中的物體。在自然語言處理中，擷取擴增生成 (RAG) 透過允許模型存取和利用外部知識來源，來提高部署模型時的準確性，使其能夠在更廣泛的背景關係中，更加瞭解自然語言輸入並改善相對應的回應。

轉換學習也稱為重新訓練，是一種機器學習技術。在重新訓練中，接受一項工作訓練的模型會重新用於第二項相關工作。大致的概念是將模型從具有大量可用標記訓練資料之工作中學到的知識，用於沒有太多資料的新工作。我們不是從頭開始學習過程，而是從解決相關工作中學到的模式開始，因此可以有效地更新為相關工作訓練的模型參數，使其更適合新工作。轉換學習主要用於電腦視覺技術和自然語言處理工作 (如情感分析)，以減少這些工作常見的大量運算能力需求。

微調在 AI 生命週期中是指使用新資料集或工作，進一步訓練預先訓練之機器學習模型的過程。這種技術通常用於轉換學習，讓使用特定工作的大型資料集訓練的模型配合相關工作或較小的資料集調整。微調允許將知識從預先訓練的模型轉換到新的工作或資料集，與從頭開始訓練模型相比，通常可以更快整合並提高效能。它廣泛用於各種 AI 應用，包括電腦視覺技術、自然語言處理和語音辨識。

LoRA (低階適應) 是一種用於微調深度學習模型的新技術，透過減少可訓練參數的數量並實現高效的工作切換來工作，最終使微調過程的運算密集度變得更低，並且更容易透過廣泛使用的硬體存取。

訓練涉及選擇適當的演算法或模型架構、調整模型參數，以及在內部評估模型的效能。根據所訓練模型的大小，在訓練過程之每個迭代中的參數更新，可能是一個資料和運算非常密集的過程。

圖 1：AI 生命週期

關鍵選擇

效能

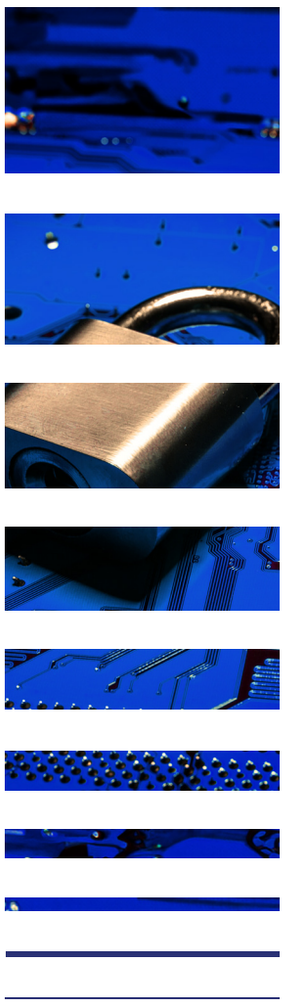
在許多這些實際應用中，即時或近乎即時的決策對於成功至關重要。例如，必須及時識別金融交易或保險索賠中的欺詐活動，以防止財務損失並保護商業資產。在製造業中，必須動態監控組裝線或出廠狀態中的瑕疵，以確保品質。實際上，處理推斷工作負載的處理器必須針對快速且高效處理傳入資料流最佳化。Dell PowerEdge 伺服器搭配 AMD EPYC 處理器是多功能的組合，非常適合處理邊緣推斷工作負載，以及涉及高效能運算、雲端運算及 Big Data 分析的工作。

資料安全性

資料安全性對於 AI 系統 (尤其是那些利用生成式 AI 的系統) 的成功至關重要，也是致力將 AI 納入其作業之技術領導者的關注要點。AI 系統通常依賴大量資料，其中可能包括敏感和機密資訊，如個人詳細資料、財務資料或專屬資訊。保護這些資料對於防止未經授權的存取或資料竊盜，以及確保 AI 模型和預測的精確性、可靠性和一致性至關重要。

機密運算是一種技術，可促進安全邊界中的資料處理，保護資料免受未經授權的當事方 (包括雲端供應商和其他使用者) 未經授權的存取或操作[2]。而在處理過程中會使用加密和其他安全措施隔離資料。AMD Infinity Guard 是整合至 AMD EPYC 處理器的一系列精密的安全功能，採用安全加密虛擬化 (SEV) 技術來支援機密運算，而 SEV 技術使用只有處理器知道的金鑰來加密虛擬機器 (VM)。這些服務旨在使用 AMD SEV-Secure Nested Paging (SEV-SNP) 提供硬體式信賴的執行環境，強化客體保護，以協助防範外部威脅。

聯合學習是維護資料安全性的另一種方法。它會跨分散的設備或伺服器訓練中央模型[3]。每個裝置不是將所有資料傳輸到一個中心位置，而是在本機訓練模型，並且僅共用模型更新。此方法可保護隱私權，並支援協同合作學習，而無需共用原始資料。Dell Technologies 的 Federated AI 平台可在收集資料集時，在網路邊緣的資料集上執行運算程序、AI 和 ML 演算法，僅透過網路將數學模型、中繼資料和查詢結果分享給其他邊緣裝置、資料中心或雲端。這種交換支援近乎即時地從大型分散式資料集中，擷取可據以行動的深入見解，而不會洩露資料和任何智慧財產權，從而強化了結果。



[2] Advanced Micro Devices, Inc.，2023 年 8 月 30 日，《AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings》(AMD 分享推動創新機密運算領導力雲端產品之技術的技術詳細資料) https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html
Advanced Micro Devices, Inc.，2021 年，》Data Center Solutions, Confidential Computing》(資料中心解決方案，機密運算) 解決方案簡介，https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf

[3] Analytics Vidhya，2023 年 12 月，《Federated Learning: A Beginner’ s Guide》(聯合學習：初學者指南)，https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data
Dell Technologies，2021 年，《A federated learning platform for real-time artificial intelligence》(適用於即時人工智慧的聯合學習平台) 解決方案簡介，https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf

擴充您的解決方案

平衡成本與創新

在成本和創新之間取得適當的平衡，可確保 AI 解決方案不僅在財務上可行，而且具有影響力，從而為企業和使用者帶來真正的價值。找到這種平衡的關鍵要素，在於識別既能解決您的使用案例，又能輕鬆整合至現有基礎結構的硬體。在現代 AI 硬體市場中，各行各業對加速器的需求增加，加上產能限制、物流挑戰和半導體短缺，種種因素導致了加速器短缺。

然而，CPU 已經是大多數資料中心的標準元件，與新增全新的加速器硬體相比，整合更簡單且更具成本效益。AI 最佳化 CPU 可以利用現有的軟體和工具，減少大量更新設備或重新訓練的需求。CPU 也可為 AI 以外的各種工作提供更大的彈性和效率，讓您更多樣化地運用資料中心內的資源。使用搭載 AMD EPYC 處理器的 Dell PowerEdge 伺服器更新資料中心，可支援您完成現有的工作負載，同時為 AI 推動的更多創新和效率進展做好準備。

簡化度和彈性

從長遠來看，AI 系統的簡化度和彈性對於建構有效、具適應性且可擴充的 AI 解決方案至關重要。存取一套可補強硬體的軟體框架和最佳化，藉此強化效能，而且無需花費額外的時間和精力進行跨平台整合。這些品質對於處理混合 AI 工作負載尤其重要，混合 AI 工作負載涉及不同類型的 AI 工作的組合，例如訓練、推斷和資料處理。

AMD 和 Dell Technologies 透過結合硬體與軟體解決方案，處理混合 AI 工作負載。AMD EPYC 處理器提供高效能運算能力，並具備同時多執行緒 (SMT) 和高核心數等功能，為 AI 工作負載提供高效率的並行處理。這些處理器針對 AI 工作進行了最佳化，可為訓練和推斷工作負載提供強大的效能。搭載 AMD EPYC 處理器的 Dell PowerEdge 伺服器，提供可擴充且彈性的平台來部署 AI 工作負載。此外，Dell OpenManage Software 套件提供的管理工具，可為混合 AI 工作負載最佳化資源分配和效能監控。

AMD 還提供 Unified Inference Frontend (UIF)，利用了現今每個軟體堆疊的效能強化版本，以及用於 AMD EPYC 處理器的 AMD ZenDNN 程式庫、用於 AMD Instinct Accelerators 的開放原始碼 AMD ROCm 堆疊，與用於 AMD 適應性 SoC 的軟體堆疊。AMD ROCm 也可搭配各種 AMD CPU 和加速器使用，包括專業和消費者等級產品。

確保可解釋性

可解釋的 AI 在確保人工智慧應用程式的透明度、可信度和有效性方面，發揮著關鍵作用。可解釋的 AI 提供了對 AI 模型如何做出決策的深入見解，揭示了潛在的因素和推斷過程。這種透明度對於獲得利益關係人的信任至關重要，尤其是在保健、金融和刑事司法等敏感領域，這些領域的決策直接影響到個人的生活。

人機迴圈 AI 系統利用人類智慧來強化 AI 效能，並減少演算法偏差。透過整合人類監督，這些系統可以更有效地處理複雜和模棱兩可的情況，確保 AI 解決方案符合倫理和社會規範。此外，人類的參與能夠根據現實世界的意見反應不斷完善和調整 AI 模型，促進迭代改善和長期可靠性。這些方法對於建構負責任、可說明、具包容性並可為社會帶來最佳利益的 AI 系統至關重要。

現實情境

Scalers AI 與 Dell 和 AMD 合作，展示搭載 AMD 處理器的 Dell PowerEdge 伺服器功能。瞭解如何將這些技術運用於零售和保健情境中的訓練、轉換學習和推斷。

零售業

Scalers AI 建構了零售庫存管理參考解決方案，該系統旨在透過實作物件偵測 AI 模型，監控和管理零售貨架上的庫存量。該參考解決方案利用 SSD_MobileNet_V2 模型來識別和辨識商店貨架上的產品，最終實現自動庫存盤點和庫存量的精確監控。該模型使用 SKU110K 影像資料集進行了轉換學習，其中包括來自 Roboflow 的 23,000 張影像。透過利用電腦視覺技術和機器學習演算法，系統可以偵測商品何時不足或缺貨，向商店員工提供警示，以便及時進貨或補貨。

此解決方案採用搭載 AMD EPYC 9354P 32 核心處理器的 Dell PowerEdge R7615 伺服器。

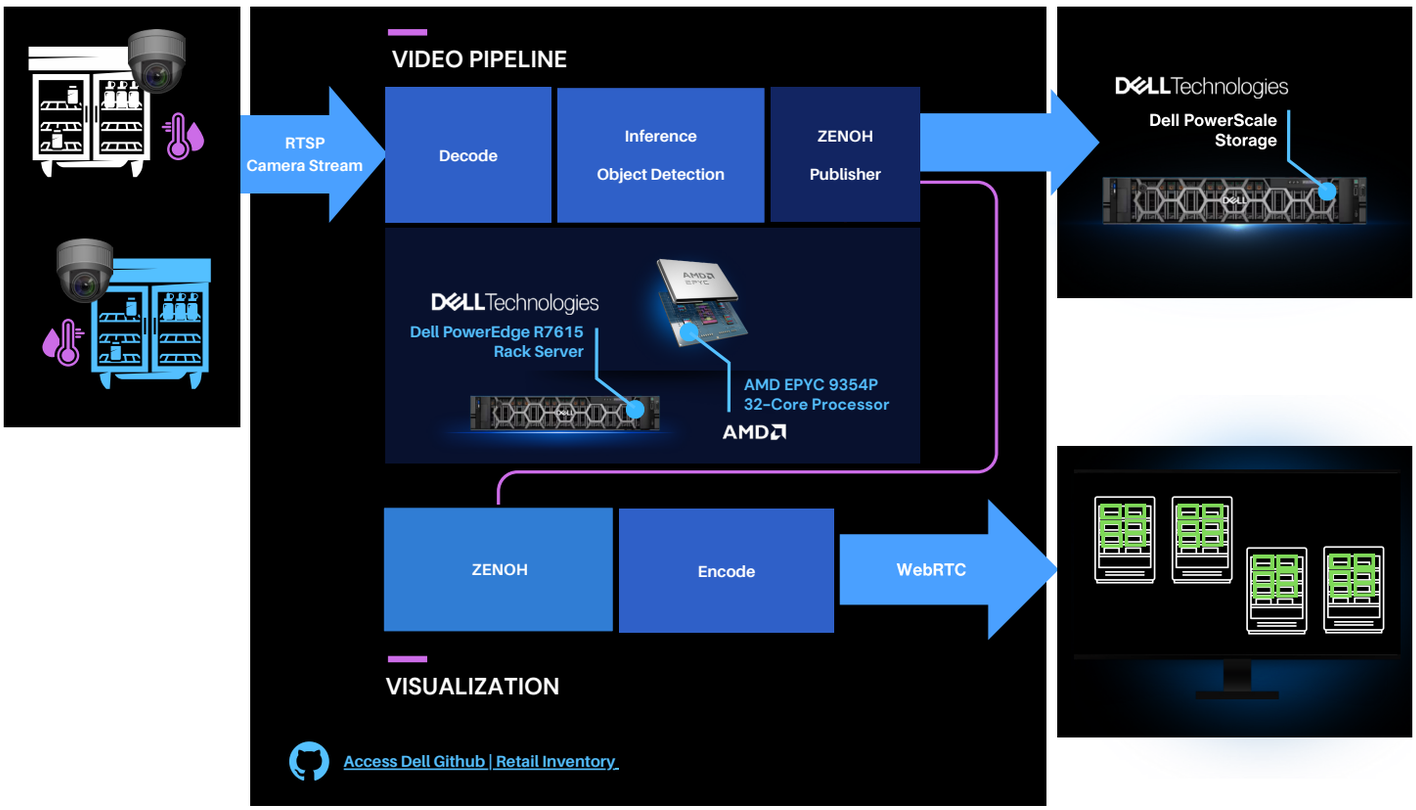


圖 2：零售庫存管理參考解決方案架構圖表

醫療保健

採用 AI 技術的醫學影像具有巨大的價值，因為它能夠透過提高診斷準確性和效率來強化保健，並針對肉眼可能難以偵測到的狀況，為保健專業人員提供精確深入見解。AI 可自動化分析醫學影像，進而縮短診斷時間，加快做出治療決策，最終改善病患治療成果。

Scalers AI 利用搭載 AMD EPYC 9554 64 核心處理器的 Dell PowerEdge R7625 伺服器的功能，建立採用 AI 技術的肺炎偵測醫學影像解決方案。該解決方案使用先進的演算法和機器學習技術來分析醫學影像，例如 X 光或 CT 掃描，有助於提高病患肺炎的診斷速度和準確性。最終，這引入了額外的電腦輔助審查層，帶來可幫助保健專業人員更有效地處理大量影像資料的潛力。

該參考解決方案利用 ResNet50 模型，分析從 NIH 臨床中心資料集取得的胸部 X 光影像。其主要目標是偵測肺炎的存在與否，本質上是執行二元分類。該模型使用來自 NIH 臨床中心資料集的 X 光 DICOM 資料集進行訓練，涉及使用 ResNet50 架構的轉換學習。

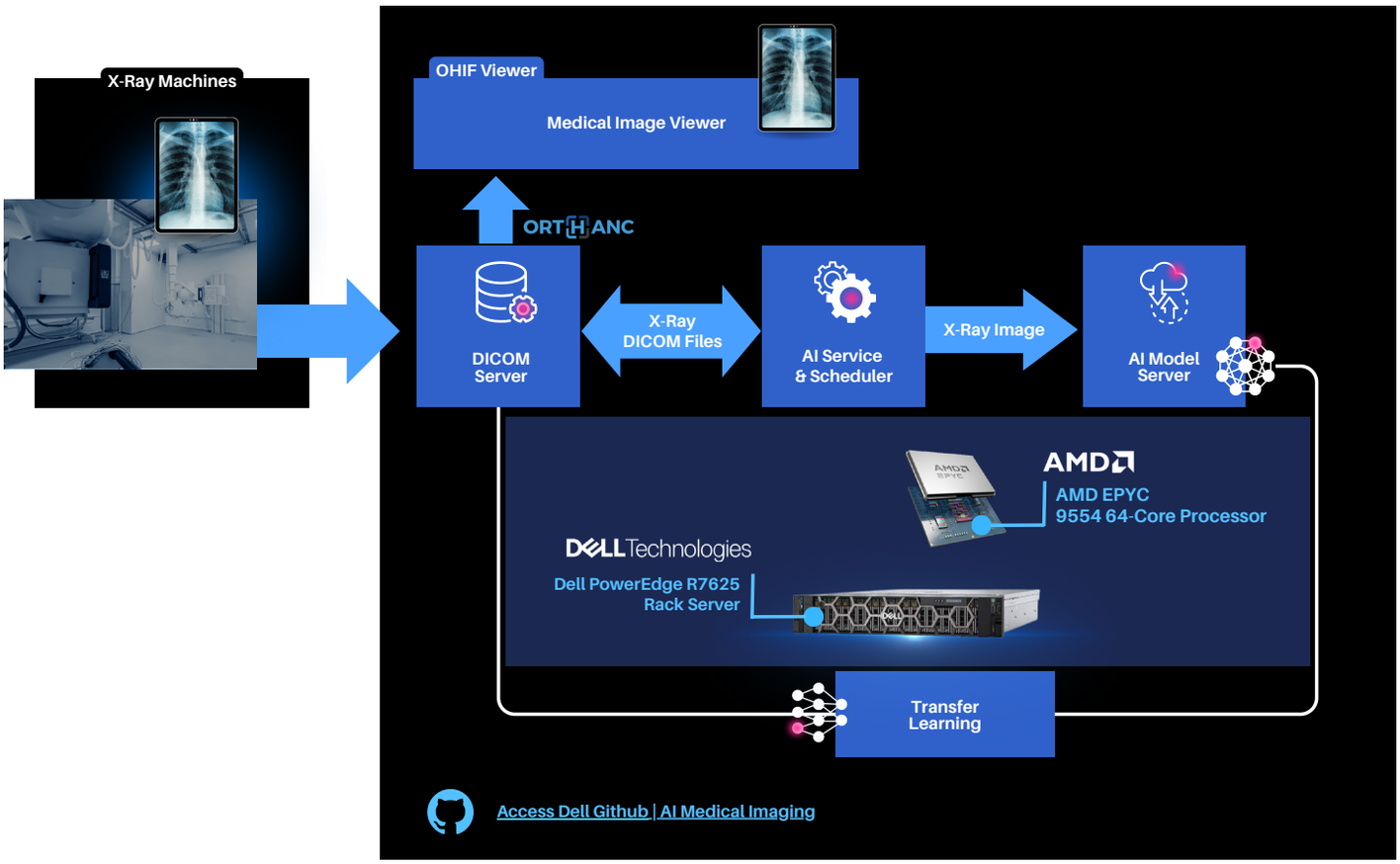
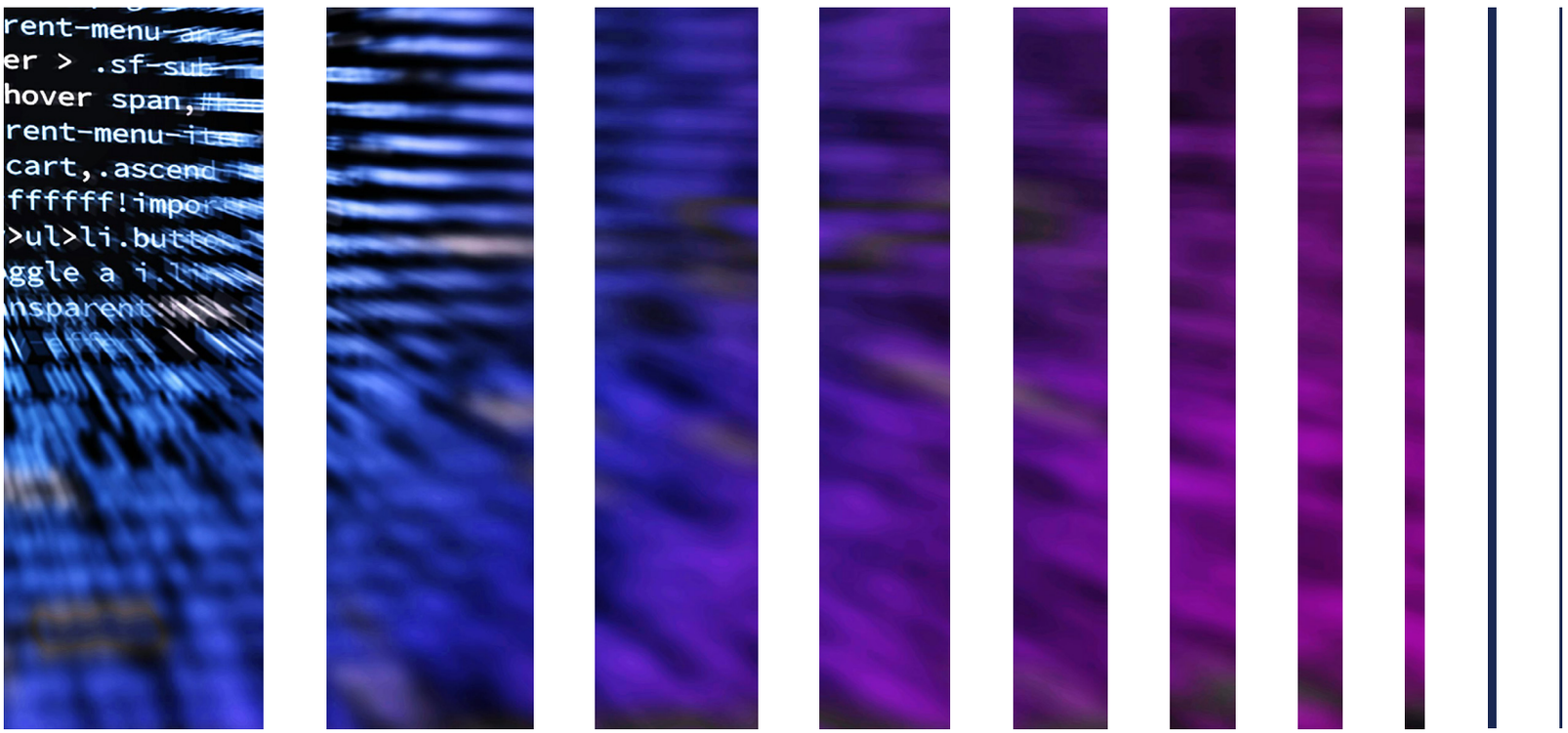


圖 3：AI 醫學影像解決方案架構圖表

我們的解決方案

適合所有人的 AI：DELL 與 AMD 將 AI 普及化

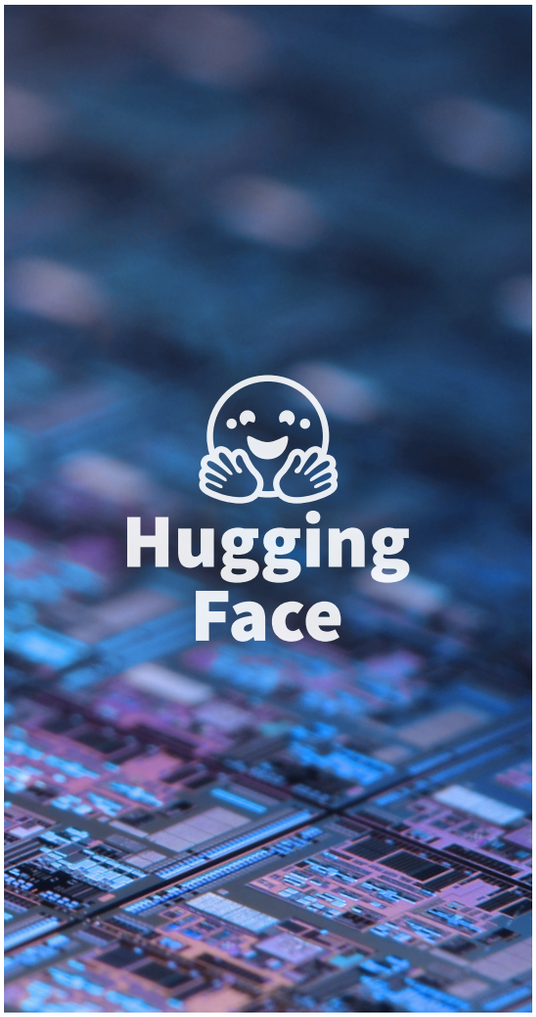
這項合作為 AI 的普及化奠定了基礎，而普及化對於促進 AI 生態系統的創新和包容性至關重要。為了達到此目的，Dell 和 AMD 提供搭載最先進的 AMD CPU 和加速器技術的強大伺服器組合，讓個人和組織能夠運用 AI，解決各自領域的獨特難題。搭載 AMD Instinct MI300X 加速器的 Dell PowerEdge 伺服器能夠處理大型 AI 工作負載，例如訓練和微調大型語言模型 (LLM)，而配備 AMD EPYC 處理器的 Dell PowerEdge 伺服器則擅長處理邊緣推斷工作負載。除了基礎硬體平台之外，AMD 還提供 ZenDNN 軟體庫，用於最佳化 AMD CPU 上的深度學習推斷，還提供 AMD ROCm 軟體庫，以改善 AMD Instinct Accelerators 的訓練、微調和推斷功能。所有這些選項在 AMD 的 Unified Inferencing Model (UIF) 中順暢整合，使用者可以透過該模型建構端對端的 AI 解決方案，其中包含軟體框架彈性選擇、軟體最佳化和硬體平台選擇。

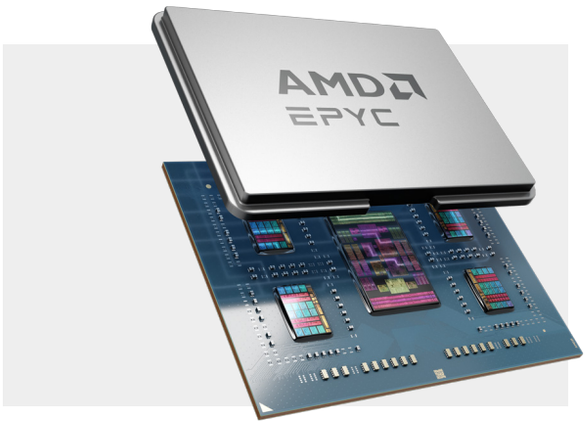


與 HUGGING FACE 協同合作

渴望採用 AI 的企業可以從直接透過 Hugging Face，利用針對其特定需求量身打造的預先存在模型或 AI 工作流程開始。Hugging Face 是一個致力於資料科學和機器學習的開放原始碼平台。AMD 已與 Hugging Face 達成合作，其共同目標是透過將 AMD 專屬的軟體最佳化新增到已經與 AMD 平台順暢整合的軟體庫和框架中，提供一流的轉換器效能。Hugging Face 正在與 AMD 的工程團隊積極合作最佳化關鍵模型以實現尖峰效能，將 AMD ROCm 整合到 Transformers 庫中，並改善 Optimum-AMD (專為 AMD 平台設計的程式庫)，以協助 Hugging Face 使用者以最少的程式碼變更加以利用。

Dell Technologies 最近也與 Hugging Face 合作，簡化企業在領先業界的 Dell 基礎結構產品和服務上，使用 Hugging Face 社群開發、微調及套用其專屬開放原始碼生成式 AI (Gen AI) 模型的程序。Dell 正在 Hugging Face 平台上開發新的入口網站，其中包含自訂的專用容器和指令碼，可協助使用者運用 Dell 的伺服器和資料儲存系統，安全、輕鬆地部署 Hugging Face 上可用的開放原始碼模型。企業現在可以充分利用 Hugging Face 資源，直接在搭載 AMD 處理器的 Dell PowerEdge 伺服器上部署模型，並使用自己的專屬資料建構端對端的 AI 解決方案。



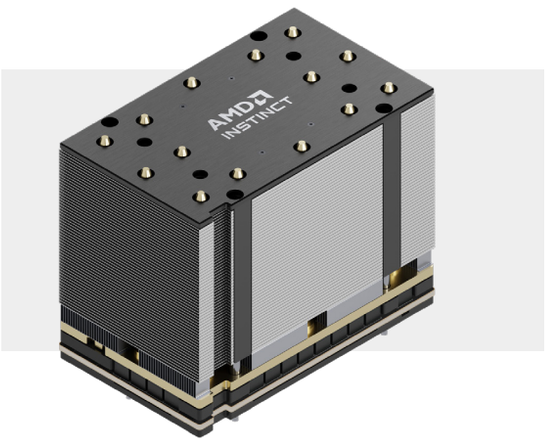


AMD EPYC 處理器

AMD 透過其 AMD EPYC 處理器，提供現代雲端型資料中心所需的技術進展。這些處理器是從頭開始設計的系統單晶片 (SoC)，可有效滿足目前和未來資料中心的需求。AMD EPYC 9000 系列處理器可為資料中心配備最多 128 個核心、256 個執行緒、12 個記憶體通道 (每個插槽支援最多 6 TB 的記憶體)，以及 128 個 PCIe Gen5 通道。這可以與業界領先的硬體嵌入式 x86 伺服器安全解決方案互相搭配。透過將基本的運算、記憶體、I/O 和安全資源整合到 SoC 中，AMD EPYC 處理器可產生頂級效能，並有助於降低總體擁有成本 (TCO)。

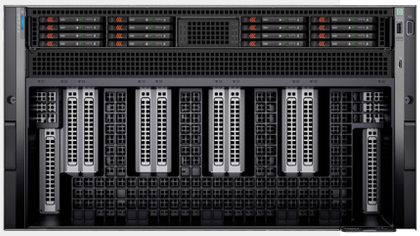
AMD INSTINCT MI300X 加速器

AMD Instinct MI300X 加速器奠基於尖端的 AMD CDNA 3 架構，可為最密集的 AI 和 HPC 應用程式提供領先業界的效率與效能。此加速器配備 304 個高效能運算單元，並具有 AI 專屬的功能，例如支援新資料類型、圖片和影片解碼，以及單一加速器上無與倫比的 192 GB HBM3 記憶體。



AMD ROCM 6 開放原始碼軟體平台

AMD ROCm 6 開放原始碼軟體平台經過最佳化，可充分發揮 AMD Instinct MI300X 加速器的高效能運算 (HPC) 和 AI 工作負載效能。此外也延伸了對 AMD Instinct MI300X 加速器的支援，確保與業界軟體架構相容。AMD ROCm 平台封裝了各種驅動程式、開發工具和 API，可促進從核心層級到最終使用者應用程式的加速器程式設計，並可根據您的特定要求量身打造。AMD ROCm 特別適用於高效能運算 (HPC)、人工智慧 (AI) 和科學運算中的應用。此外，AMD ROCm 平台還提供對多加速器運算的支援，包括用於伺服器與節點通訊的遠端直接記憶體存取 (RDMA)。



DELL POWEREDGE 伺服器產品組合

Dell 對 AMD 的投資在市場上創造了推動 AI 普及化的關鍵選擇，這從其搭載 EPYC 的四個伺服器平台，以及搭載 AMD Instinct MI300X 加速器的旗艦級 Dell PowerEdge XE9680 機架式伺服器可見一斑。最新一代的 Dell PowerEdge 伺服器採用 AMD EPYC 處理器，可同時提升商業靈敏性及上市時間，並可支援轉型工作負載，例如資料庫和分析、虛擬化、軟體定義的儲存方式、虛擬桌面基礎結構 (VDI)、容器化、高效能運算 (HPC)、AI 以及機器學習 (ML)。其單插槽 (單 CPU) 機架式伺服器能以符合成本效益的方式，兼顧效能與儲存容量，可隨著業務順暢增加；而其雙插槽 (雙 CPU) 機架式伺服器則具備廣泛功能，可因應更繁重的工作負載。

Dell PowerEdge XE9680 機架式伺服器是專為 AI 工作設計的強大資料處理機器。其支援八個加速器，非常適合用於機器學習 (ML)/深度學習 (DL) 訓練和推斷工作負載，特別是訓練大型語言模型 (LLM) 的工作負載。配備八個 MI300X 加速器，每個加速器具有 192GB 的 5.3 TB/s 高頻寬記憶體 (HBM3)，使每台伺服器的總 HBM3 容量為 1.5 TB，FP16 效能超過 21 petaflops，有了這些，搭載 AMD Instinct MI300X 加速器的 Dell PowerEdge XE9680 機架式伺服器，已準備好進一步將 Gen AI 存取功能延伸到企業。這使企業能夠訓練更大的模型、將資料中心足跡降至最低、降低總體擁有成本，並獲得競爭優勢。

摘要

AI 推動的快速創新步伐，正在以比其他任何技術轉型都快的速度，徹底改變資料中心工作負載。為了支持這些技術進展，Dell 和 AMD 正在努力建立一個更具包容性、創新性，且以合乎道德的方式開發的 AI 生態系統，鼓勵各行各業的開發人員在開放原始碼資源上協同合作，並推動現代的 Gen AI 創新。無論您的 AI 解決方案是透過 AMD EPYC 處理器或搭載 AMD Instinct Accelerators 的伺服器滿足效能需求，我們都能提供彈性，讓您在我們的硬體平台上執行 AI 工作負載，並發揮 Dell 和 AMD 所提供的優勢。

參考資料

AMD 影像：AMD.com，AMD 合作夥伴資源庫，**https://www.amd.com/en/partner/resources/resource-library.html**

Dell 影像：**Dell.com**