White Paper

將人工智慧技術應用到工作站帶來的意義

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

IDC 觀點

人工智慧（AI ）已經成為各行各業中突顯差異化的重要指標，而執行 AI 所需的硬體也日新月異。科技產業非常關注最先進的 AI 模型的指數成長趨勢，相關討論涉及數百億個參數、降低精準度、擴展記憶體、高效能運算 (HPC) 等 AI 訓練和推論需求，以及加速伺服器機架。但事實是，這麼大規模的 AI 運算需求極少，在企業中更是如此。

現在，許多企業都在努力推動 AI 計劃，包括不需要超級電腦的生成式 AI。其實，有許多 AI 開發以及部署在邊緣端的 AI 應用都是在功能強大的工作站進行。工作站對於 AI 開發和部署有諸多優勢： AI 科學家或開發人員再也不用花時間協商伺服器使用時間，即使資料中心不易獲得伺服器型繪圖處理器 (GPU)，工作站仍能加速繪圖處理器 (GPU)。相較於伺服器，工作站非常經濟實惠，只要一筆費用不高的一次性支出，不像雲端服務的費用會快速累積。另外，透過將機敏資料妥善儲存於工作站的方式，企業也將相當安心。此外，科學家或開發人員也不用擔心成本增加等議題，可以放心進行人工智慧模型實驗。

IDC 發現，在 AI 部署這個應用場域，邊緣環境的成長速度勝於企業內部或外部雲端環境。同樣地，工作站也成為 AI 推論的重要平台，甚至不需要繪圖處理器 (GPU)，即可在軟體最佳化的 CPU 上進行推論。部署在工作站的邊緣 AI 推論案例正快速成長，包括 AIOp、災害回應、放射學、石油和天然氣探勘、土地管理、遠距醫療、交通管理、製造廠監控和無人機。

本白皮書將聚焦在討論工作站對 AI 開發和部署扮演的角色與發揮的效益，並簡述 Dell 的 AI 工作站產品組合。

情勢概觀

AI 的爆炸性成長以及對資訊基礎設施的影響

全球組織投入 AI 專案的數量成長迅速。在各個產業中，許多任務都已經部分或全部透過 AI 模型驅動的軟體執行。在 IDC 追蹤的各層面 AI 數據中，有一個很值得參考的指標：企業和雲端服務提供商計畫花費在 AI 開發與執行的伺服器上開的支出。這項支出會在 2026 年以前達到 346 億美元，佔全球伺服器總支出的 22%。但伺服器不是全貌，許多 AI 準備、開發、測試以及部署都是在工作站上進行。無論企業規模，越來越多組織發現，只要在應用中增加一定的 AI 功能即可帶來嶄新商機，因此 AI 模型的實驗性專案急遽增加，而具備即時性、可用性且鄰近數據等優勢的強大耐用的工作站則成為最佳選擇之一。

AI 演算法已發展數十年，為什麼一夕之間變得如此流行？理由在於，大量、低廉且多元的(如非結構化和半結構化資料) 數據資料取得性越來越高，以及在平行模型增強的線性運算支援下，AI 演算法-類神經

網絡-的處理時間變得越來越可以被接受，意味著類神經網路可以自動學習、執行越來越難的任務，資料科學家也可以在開發方面取得顯著進展。雖然傳統的機器學習 (ML) 仍適用於文字或數字資料，但深度學習 (DL) 更擅長處理影片、音訊、語言等資料。

傳統的機器學習模型通常可以在最多具幾十個核心的工作站 CPU 上開發，但類神經網路需要輔助處理器才能在數千個核心上平行處理。主要原因是，在機器學習領域中，功能擷取和分類都是手動流程，但在深度學習領域中都是自動的，需要使用大型資料集，透過不斷重複來訓練模型。目前最常見的輔助處理器是繪圖處理器 (GPU)，但市面上也開始出現新創公司開發的新型人工智慧專用處理器。這種使用獨立的輔助處理器平行處理的加速方式徹底改變了伺服器和工作站市場，催生出 IDC 所謂的大規模平行運算。

2022 年，全球市場的加速伺服器產值達 218 億美元，到 2026 年將成長到 434 億美元，其中 57% 為執行 AI 的加速伺服器。與此同時，用於工作站的獨立繪圖處理器 (GPU) 銷售量將成長至 640 萬顆；IDC 估計，在 AI 的持續發酵下，截至 2026 年，用於科學或軟體工程目的的工作站市場將增加至近 20 億美元。

AI 開發階段

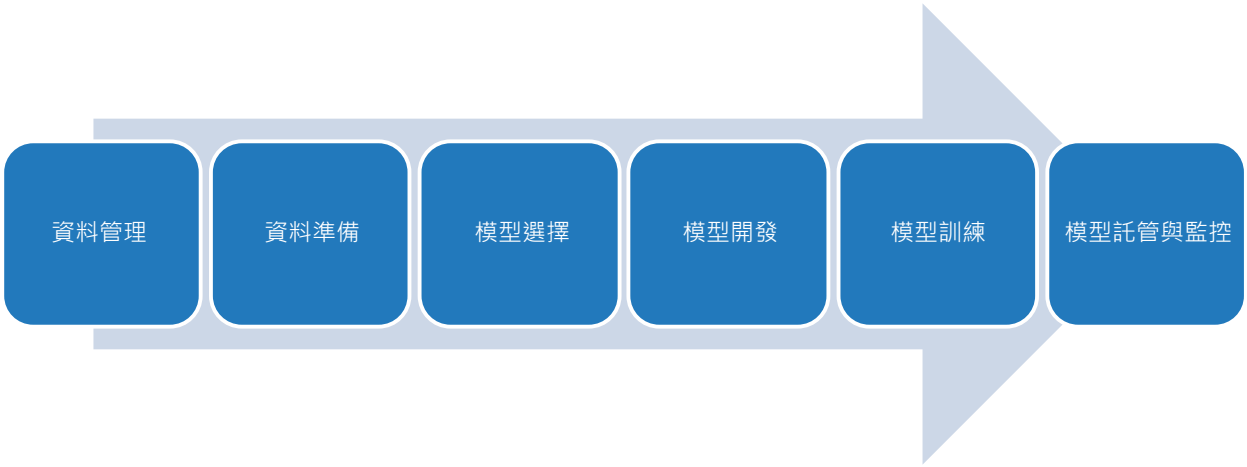
如前文所述，隨著資料類型和資料量增加，以及新的運算方式，類神經網路成為了可行的方式。這個方程式的第一部分是資料量和類型，根據一些說法，深度學習 AI 計劃中，高達 80% 的工作是以資料管理和準備為主。在模型設計和訓練開始之前，須先匯入、管理和準備資料。以下是 IDC 整理的 AI 開發階段 (請見圖 1)：

- **資料管理**：從組織匯入、生成、取得的跨資料中心、邊緣和雲端的大量資料中，辨別和管理與 AI 模型相關的資料 (而這些資料可以是任何類型，如驅動資料或串流資料，且其中有許多資料可能需要特定的治理方式)。
- **資料準備**：將資料 (檔案、區塊或物件) 儲存在資料倉庫或資料湖中，再加以清理，確保資料的完整和品質，接著將資料轉換成 AI 模型能使用的格式，如 Spark 或 Pandas 等工具
- **模型挑選**：根據錯誤率和效能，判斷最適合執行 AI 任務的模型
- **模型開發**：使用 XGBoost、LightGBM、GLM、Keras、TensorFlow、PyTorch、Caffe、RuleFit、FTRL、Snap ML、scikit-learn 或 H2O 等框架來設計 AI 模型
- **模型訓練**：透過足夠的處理器和/或輔助處理器核心在運算基礎設施上訓練模型以實現平行處理 (包括解釋、驗證和記錄模型決策的能力，以確保公平性、當責和透明性) (包括試作，即在模型上推論以測試訓練完的模型)。
- **模型託管和監控**：在生產環境中部署模型，以執行其設計目的之任務，通常稱為「AI 推論」，並監控其效能

工作站可搭配資料中心、雲端或邊緣基礎設施，在這六個階段中的任一階段發揮重要作用。

圖 1

AI 開發階段



資料來源：IDC，2023 年

在工作站上開發 AI 模型

工作站與個人電腦

眾所周知，個人電腦 (PC) 的功能不足以支援 AI 開發工作。資料科學家和 AI 開發人員通常會參與組織的重要策略專案，因此，保障其生產力不受侷限非常重要。工作站的效能往往比 PC 更可以被預測，因為工作站通常是由更高效能的元件打造而成，並針對執行軟體進行最佳化。

這些元件包括：

- **高階處理器**：例如 Intel Xeon 可擴充處理器。
- **強大的繪圖處理器 (GPU)**：例如 NVIDIA 的 RTX 專業繪圖處理器 (GPU)：NVIDIA RTX 6000 Ada。
- **更大的儲存空間**：部分工作站可提供高達 60TB 的儲存空間，I/O 速度也比 PC 的速度快得多。
- **更大的記憶體**：工作站可支援高達 6TB 的記憶體。
- **冷卻**：高效能元件會產生大量的熱能，資料科學家需要冷卻能力足夠的工作站來防止過熱並維持最佳效能。
- **網路介面卡 (NIC)**：對於要處理儲存於遠端伺服器上大型資料集的資料科學家而言，若要快速且有效率地傳輸資料，高速網路介面卡不可或缺。
- **顯示器**：高品質的顯示器有助於資料呈現的視覺效果，資料科學家應尋找兼具高解析度、高色彩準確度和大尺寸螢幕的顯示器。
- **錯誤校正碼 (ECC) 記憶體**：ECC 能偵測並校正最常見的內部資料損毀，有效預防長時間的 AI 訓練過程中，系統設備因為硬錯誤 (錯誤位元) 或軟錯誤 (位元翻轉，導致錯誤值) 而當機；ECC 還能確保結果的準確性，對於攸關性命的工作 (如醫療保健)是最重要的需求。

- **專業晶片**：例如 Intel Movidius 視覺處理器 (VPU)，這是一種平行的輔助處理器，適用於零售、安全和工業自動化等環境中的電腦視覺和邊緣人工智慧應用。FPGA 也能用於工作站，如財務應用程式等。
- **最佳化軟體**：舉 OneAPI 為例，OneAPI 是 Intel 的標準型程式設計模型，能簡化 CPU、繪圖處理器 (GPU)、FPGA 和其他加速器，或是 CUDA，即 NVIDIA 的平行運算平台兼應用程式設計介面，能在 GPU 上執行一般工作負載。

CPU 與 AI 的繪圖處理器 (GPU)

工作站可用於 AI 開發的各階段，工作站通常會有各種功能。儘管繪圖處理器 (GPU) 廣泛用於平行處理，但在工作站開發 AI 模型時，CPU 的作用也相當重要。CPU 和繪圖處理器 (GPU) 一樣，同樣能用於資料調處和開發傳統機器學習模型。CPU 也能用於資料探勘 (即使用資料集的可視化表現，了解資料特性的過程)。

深度學習訓練時，隨著繪圖處理器 (GPU) 在實際訓練過程中接管工作，主機的 CPU 作用會有所降低，但即使如此，CPU 仍是作業系統或 CUDA 等重要軟體的處理層，並肩負協作 GPU 之間或與其他晶片的角色。而且，在生產環境中使用工作站執行 AI 模型時，CPU 越來越常擔任 AI 推論引擎這樣的新角色。IDC 預計，在 2024 年以前，AI 推論的基礎設施支出將超過 AI 訓練的 AI 支出，其中會有很大一部分 (39%) 是花在主機的 CPU 上。

工作站與伺服器：共生關係

對大多數組織而言，因應 AI 開發需求彈性使用工作站、企業伺服器和雲端資源是最實用的方式，對 AI 專案的不同開發階段，工作站、伺服器和雲端資源是共生關係。

相較於資料中心伺服器，工作站的優勢在於，資料科學家可以在任何想要的地方工作，這在疫情期間是很重要的因素，但在正常情況下也是如此。資料科學家還可以在工作站上進行 AI 模型實驗，根據需要反覆進行迭代。現代工作站具備強大繪圖處理器 (GPU)，因此，不僅迭代過程更具互動性，可提供即時回饋和結果，無須請求存取伺服器，也不會有其他資料中心的限制。資料科學家還能透過工作站將運算彈性靈活的移動到離資料更近的地方，進而節省頻寬、減少網路壅塞，並增加資料產出量。此外，工作站可以針對不同需求設定，如傳統機器學習任務或深度學習密集型工作。

而且，儘管加速伺服器市場呈現顯著成長，但加速伺服器尚未廣泛用於企業資料中心。撰寫此白皮書時，平均僅 4% 的企業資料中心已完成加速，這表示許多組織仍無法在現有內部繪圖處理器 (GPU) 上開發或執行 AI。因此，加速工作站也是 AI 開發的一個可行替代方案。

高度加速的工作站現在已經相當強大，只要 AI 模型沒有過大，就能執行深度學習訓練，不用在伺服器上進行訓練。在有繪圖處理器 (GPU) 的工作站上訓練的模型可用 CPU 的推論功能，部署在沒有配備繪圖處理器 (GPU) 的工作站或伺服器。Intel 的 DL Boost 和 oneAPI 等軟體技術可加強驅動 CPU 上的 AI 推論，讓已部署到資料中心的未加速伺服器能支援 AI 應用。

工作站與雲端資源

雲端運算已徹底改變組織、基礎設施、資料和應用的實務。雲端憑藉近乎無限的擴充能力，讓開發人員能夠按需採用所需資源，或因限制減少而加快創新的步伐。從這點來看，雲端似乎是 AI 開發的完美方式。

但實際情況不見得是如此。IDC 表示，越來越多組織將部分工作負載從公有雲轉回內部基礎設施，原因如下：

- **雲端可用性**：任何仰賴雲端服務的使用者都體驗過服務中斷期間，無論是雲端提供商自身的問題，或是超大規模資料中心和終端使用者之間的網路中斷導致的服務中斷。這種情況下，使用者只能等待服務提供商解決問題，生產力也將因此停滯。
- **安全性和法規遵循**：在許多產業中，企業治理策略都會規定資料通訊和儲存位置，此做法會限制雲端服務的使用。歐洲的 GDPR 和美國的加州消費者隱私保護法 (California Consumer Privacy Act) 等許多政府法規，都會強制執行資料主權施行細則。
- **成本**：組織通常會低估雲端服務費用的成長速度，尤其是對需要高效能運算能力和大量儲存空間的工作負載。雲端成本除要計算所有類型的資源消耗費用，還必須含括將數據資料遷回企業基礎設施的費用。
- **試錯壓力**：多數 AI 專案都從大量實驗開始，失敗的模型也是開發過程的一部分；過程中，若雲端費用漸增，可執行成果卻遙遙無期，AI 科學家和開發人員的心理負擔將隨之增加。

工作站可以解決這些限制，同時持續使用雲端原生技術，如微服務型架構和 API 驅動的自動化服務，這讓工作站和資料中心伺服器的比較具備一些優勢：

- **工作不受距離限制**：消除對公有雲的依賴後，即可實現中斷連線的情況。許多高度安全環境不會和公共網路連接，而人 AI 工作站能專門滿足此需求，當然，也能夠降低對本地伺服器的昂貴網路連線需求。
- **資料局部性**：IoT 裝置或其他連線設備大幅增加，因此邊緣位置的資料呈指數成長。透過將運算資源和專用工作站放在一起，可以有效限制資料移動，解決許多法規遵循的要求。
- **自由實驗**：AI 模型的訓練和最佳化是一種迭代過程，包括各種試錯。開發人員應該有自由實驗的彈性，不用因為試錯可能產生的額外費用而綁手綁腳。工作站還能為訂製工具提供更大的彈性。

關於後面這點，比較工作站和雲端部署的價格相對容易，因為大多數雲端服務提供商提供的即時成本預估都是以終端使用者希望部署的任意配置為基礎計算出來的。以某大型雲端提供商為例，一台一般虛擬機 (VM) 配有 NVIDIA T4 和 375GiB SSD 儲存裝置例項，每週使用五天，每天八小時，這樣的成本是 140 美元。若將 VM、T4 和 SSD 增加一倍，則每月成本將達 365 美元。維持兩台 VM，但將 T4 再加倍到四個、將儲存空間增加至 4 個 375GiB，並在環境中完全用於執行訓練，則成本會增加至每月 2700 美元。而此可見， AI 開發的雲端成本容易飆升至每年數萬美元，大幅超過高階工作站的年折舊。

在工作站試作 AI

與內部伺服器和雲端相比，工作站在試作 AI 模型方面具備顯著優勢。資料中心的伺服器可能處於滿使用率的狀態，或是因過於重要，反而不適用於 AI 試作和測試。再者，如前所述，大量使用雲端資源作為測試環境時，雲端資源可能會導致成本迅速超支。工作站能為 AI 科學家或開發人員帶來自由，不用再受伺服器存取權限牽制，或是在試作階段煩惱雲端費用不斷增加。工作站的單次成本很低，可以隨時隨地自由試作，不需額外成本。

在工作站上部署 AI 模型

雖然在工作站開發 AI 模型的策略行之有年，但 IDC 發現，在工作站 (尤其是在邊緣) 部署 AI 模型的使用案例越來越多，也就是將 AI 投入生產環境，在工作站上執行推論。終端使用者已經發現邊緣的優勢，因此，越來越多人將邊緣當成伺服器的 AI 部署位置，從 2020 年到 2024 年，每年硬體支出增加兩倍以上，而工作站也相去不遠。

IDC 將邊緣定義為分散式運算的典範應用，將基礎設施和應用服務部署在集中式雲端和內部資料中心之外、盡可能靠近產生和使用資料的位置，如遠端辦公室和分公司，以及特定產業位置 (如工廠、倉庫、醫院和零售商店)。

越來越多資料和運算密集型工作負載部署在內部或邊緣位置。這麼做是為了減少公有雲的固有限制，例如減少上傳大型資料集所需的時間和 AI 訓練的變動成本，尤其是那些需要大量資料科學實驗的應用場景。

IDC 研究顯示，邊緣環境是快速成長的 AI 部署場域。2023 年，組織在邊緣 AI 運算的投資達 29 億美元，到 2026 年會成長至 69 億美元 (請見「2022-2026 年全球 AI 硬體預測：AI 運算和儲存市場成長強勁，IDC #US49671722，2022 年 9 月」)。此外，以邊緣環境作為工程和技術等 HPC 工作負載的部署選擇也越來越受到重視。企業目前在邊緣工作負載的投資近 10 億美元，到 2027 年將成長至 24 億美元 (請見「*2023-2027 年全球高效能運算伺服器預測：企業將超越 HPC 實驗室*，IDC #US50525123，2023 年 4 月」)。對於有些領域，部署 AI 工作站是很合理的選擇。

在邊緣工作站部署 AI 模型時，不見得像 AI 開發一樣需要高階繪圖處理器 (GPU)。輕量級繪圖處理器 (GPU) 即可處理 AI 推論，而且在某些案例中，根本不需要繪圖處理器 (GPU)。在這種情況下，CPU 可充分執行推論任務，特別是與 Intel DL Boost 等優化服務一起使用。Intel DL Boost 是 Intel 微處理器上的一組指令集功能，用於加速人工智慧工作負載，如 AI 推論。Intel 表示，支援 Intel DL Boost 的第四代 Gen Intel Xeon 可擴充處理器的 INT8 即時推論資料吞吐量較前一代處理器 (BERT-Large SQuAD)高出 1.45 倍，這讓工作站更適合部署在邊緣，因為邊緣位置的電力、行動力和溫度控制所需要的功耗更低。Intel Movidius Myriad (M2) 的功耗僅 12W，因此非常適合這一功耗範圍。

在工作站部署 AI 的使用案例

有些使用情境很適合在本地部署的工作站上部署 AI。這些情境的共同特徵是大量機器產生的時間序列資料和非結構化資料，如影片串流和映像。也有些情況是，領域專家系統需要透過人類解釋來增強 AI 模型。

範例含括：

- **AIOps**：隨著 IT 系統規模和複雜程度增加，越來越需要從被動意外管理轉為主動監控。當組織將基礎設施和應用服務分散至幾乎沒有技術人員的邊緣位置時，這點尤其重要。透過將正常效能基線模型化的方式，系統即可自動識別異常並啟動補救步驟。
- **災難應變**：在緊急情況下，第一個反應的急救人員必須迅速評估情況、追蹤關鍵設備與部署資源以幫助最需要的人。這項工作通常得在沒有網路連線的環境下進行，需要一個本地工作站負責聚合各種數據資料、根據 AI 模型進行推論，並自動與關鍵人員通訊。
- **放射學**：隨著影像處理技術的進步，單次掃描產生的資料量與日遽增，需要將資料留在現場才能即時分析。使用數百萬個預先範例訓練出來的 AI 模型能比人眼更準確地識別出模式，進而提高準確率。
- **石油和天然氣探勘**：上游的石油和天然氣公司透過彙整與分析遙測、地震和影像處理資料定位出自然資源儲量、選擇鑽井的位置，並且最佳化生產過程中的設備效能。這通常只能在可以使用昂貴的衛星通訊的地區才有辦法進行數據分析。
- **癌症研究和藥物開發**：研究型醫院和學術研究人員會使用 AI 和自然語言處理技術協助腫瘤科醫師找出最適合每一位病患的個人化癌症治療方式。他們還結合機器學習和電腦視覺，讓放射科醫師更清楚地了解患者腫瘤的情況。另外，他們也會使用演算法，進一步了解癌症如何擴散，以及找出最佳治療方式。

- **保險索賠評估**：人工索賠處理是勞動密集型工作，且容易出現人為失誤。透過可以評估索賠是否合法的 AI，保險機構不僅可以降低成本，理賠人員也可以將心力放在需要深入調查的案例。這樣可以在不降低準確性的情況下，增加營運的總資料吞吐量。
- **遠距照護**： AI 可根據穿戴式裝置設備的即時生命徵象，量身打造個人治療計劃，提高病患的康復率。這類資訊需要跟病患的就醫紀錄及類似病例的知識庫結合。這對於高度依賴遠距照護的鄉下地區尤其重要。
- **零售安全 (反竊盜)**：透過即時分析串流影片，將有助於預測可能導致違法活動的人類行為。這種做法通常需要將多個影片剪接在一起，才能追蹤個人在商店內的行動。為在最短時間內辨識重大事件，這種做法最好是在本地環境執行。
- **車流管控**：負責交通營運的行政機關越來越常使用 AI 協調紅綠燈和電子看板，以改善車流並確保市民安全。這種做法需要結合各種輸入，包括攝影機和路口感測器的遙測資料，才能將交通模式最佳化。
- **製造工廠監控**：對工廠管理人員來說，確保關鍵流程能正常運行和實踐生產計劃最為重要，這需要關鍵設備的預先維修、缺陷自動檢測和工廠供應鏈的內外最佳化予以支援，在這個應用範疇中，AI 可以協助操作人員提升效率，同時，兼顧安全標準。
- **無人機**：自動分析無人機拍攝的映像，以前所未有的規模監控各種情況。這種做法對天然氣和電力基礎設施的檢查、保險調查、搜救工作、精準農業、漁場及野生動物保護區維護，都帶來重大的影響。
- **日常辦公環境**：透過 Microsoft Copilot 等 AI 生產力工具逐步提升日常辦公環境效率。
- **可再生能源**：可再生能源站點 (如風力發電站、水力發電站或太陽能發電場) 需要即時監控、維護，以及在本地環境蒐集與分析系統設備產生的各種數據資料。

專為 AI 而生的 DELL 工作站

Dell 針對不同類型的 AI 開發與實施工作提供系列工作站產品服務，所有工作站產品都隸屬於 Data Science Workstation (DSW) 品牌。本章節將簡單描述規格，而後討論 AI 的人員和應用，如資料科學家和 Dell DSW 技術的優勢。這些 AI 資料科學工作站是專為資料科學家所設計。最新的 Precision Data Science Workstation 使用 AI 功能，對裝置進行微調，將資料科學家最常用的應用程式效能最佳化。如此一來，資料科學家便能更早完成最重要的工作。此外，Dell Precision 工作站經獨立 ISV 測試和認證，確保能支援 Dell 客戶完成日常任務所需的高效能應用程式。

Dell 工作站如何脫穎而出

Dell Precision 工作站是以 NVIDIA RTX GPU 為核心，旨在協助組織分析，以及為人工智慧應用提供強大的擴充能力和效能。Dell Technologies 提供全面的硬體解決方案，這些解決方案已經優化、可以執行先進的 AI 軟體：

- **強大的硬體配置**：Dell Precision 工作站提供一系列強大的硬體組態，包括多核心處理器、大容量 RAM 和多個繪圖處理器 (GPU) 選項。這些元件為 AI 任務提供必要的運算資源，並且提升訓練和推論的效率。
- **擴充能力和可自訂性**：Dell Precision 工作站可擴展且可自訂，使用者可以根據自身的特定 AI 需求，量身打造硬體組態。這樣的彈性能確保，工作站可以針對 AI 工作負載的特定需求進行最佳化。
- **認證和最佳化**：Dell 和 NVIDIA 協作，認證 Precision 工作站和 NVIDIA RTX GPU 的相容性和效能，包括 NVIDIA RTX 6000 Ada 世代顯示卡。此認證可確保使用 Dell Precision 工作站和 NVIDIA RTX GPU 執行 AI 任務時，實現流暢整合及最佳化效能。

- **強大的處理能力**：Dell Precision 工作站配有 Intel 處理器，可提供 AI 任務所需的運算能力。透過多核心處理器和高時脈速度，這些工作站可提供 AI 工作流程中訓練和推論所需的效能。
- **軟體和工具支援**：Dell Precision 工作站預先載入支援 AI 開發和部署的軟體和工具。其中包括使用 NVIDIA RTX GPU 的最佳化軟體堆疊、AI 框架和媒體櫃，讓使用者能更輕鬆展開 AI 計劃。

另外，以下章節討論的技術是 Dell 工作站脫穎而出的其他重要領域。

可靠的記憶體技術

Dell 以 ECC 為基礎，提供一項名為 Reliable Memory Technology Pro (RMT Pro) 的技術，該技術旨在極大程度的延長連續運作時間， 可以跟 ECC 記憶體搭配使用，藉此即時偵測和修正記憶體錯誤。根據 Dell，RTM Pro 消除大部分的記憶體錯誤，即使 DIMM 仍處於完全使用狀態，也能防止記憶體再次出現錯誤。系統重新開機後，RTM Pro 會隔離有錯誤的記憶體區域，並將之隱藏在作業系統之外，因此，AI 資料科學家和開發人員不會遇到反覆、持續崩潰的狀況，因為有錯誤的記憶體是可以被解決的，從而大幅提升生產力。

Dell Optimizer for Precision

Dell 還在其多數工作站中加入 Dell Optimizer for Precision，可以自動調整系統設定，讓工作站以最快速度執行各種流行的商業應用。這樣就能提升資料科學家或開發人員的生產力。這項工具還能幫 IT 人員建立處理器、儲存裝置、記憶體和顯示卡使用率的即時效能報告。目前還不能在 Linux 上執行 DOP，因此主要用於部署 AI，因為開發 AI 常透過 Linux 開放原始碼軟體完成。Dell Optimizer for Precision 還會提供 ExpressSign-in、Express Charge (手機版)、Intelligent Audio 和報告及分析工具，協助微調工作站。

挑戰/機會

對企業而言

IDC 發現人工智慧市場出現分歧。一方面，企業透過大規模採用 AI 等數據策略維持競爭力，例如，同行已使用名列百大超級電腦的企業級 AI 基礎設施產品完成優秀的工作內容。另一方面，企業每天都可以看到許多在資料中心或雲端伺服器上實驗的小型 AI 專案因為預算不足和硬體效能不佳而有所侷限。

對許多企業而言，第一種情況無關緊要，但第二種卻銘刻在心。他們所面臨的挑戰是，若不在雲端環境或繪圖處理器 (GPU) 型資料中心伺服器花費大量金錢，便無法提供適合的工具給自家的 AI 資料科學家和/或開發人員，讓他們即時執行 AI 訓練。IDC 認為，若為這些企業的科學家和開發人員提供強大的繪圖處理器 (GPU) 加速工作站，這些企業將受益匪淺。

對 Dell 而言

市場存在一種誤解，認為 AI 開發需要昂貴的加速伺服器軟體，甚至需要叢集伺服器。對擁有數十億參數的最大 AI 演算法而言，是這樣沒錯，但多數企業都不是要開發這類大規模演算法。企業已開始用 AI 計劃處理有用、有效且可管理的工作，而且許多企業都沒意識到，可以在工作站開發和部署這種一般規模的 AI 模型。Dell 面臨的挑戰是打破先入為主的觀念，讓市場了解工作站產品組合的可能性。

與此同時，Dell 須確保工作站能交付產品，且避免長期下來，成為技術瓶頸。這表示得不斷快速創新，才不會讓以正確方式使用工作站的終端使用者失望 (也就是不打算嘗試執行數十億參數演算法的使用者)。這也表示，對突然開始快速擴展或是演算法規模已經非常大的客戶來說，可以從工作站無縫銜接，過渡到 Dell 的 AI 伺服器系列。當然，Dell 不會錯過這個機會，無論客戶想要什麼規模的 AI 計劃，Dell 都能提供合適的解決方案。

結語

IDC 認為，工作站目前在許多使用案例中，尚未獲得充分的重視。工作站為 AI 科學家和開發人員提供強大的繪圖處理器 (GPU) 加速平台，此平台的資本支出 (CAPEX) 比伺服器低、OpEx 遠比雲端例項低，而且還能更自由地使用 AI 模型做實驗。若企業著手開發的 AI 計劃不需要參數達十億的演算法，或許可以考慮為 AI 團隊配備工作站，減少 AI 開發的限制並輕鬆地進行邊緣部署。

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com