企業 AI 時代下
算力架構重塑與硬體演進

DIGITIMES

自主 AI 與多模態技術正帶動 AI 的企業需求漸起，並使推論成長性將高於訓練， 加上 AI 負載由雲端蔓延到地端，AI 算力逐漸正面臨架構重組。

DIGITIMES 觀察，「對話機器人」、「軟體開發」、「圖像生成」、「影像生成」，「企業營運資訊自動化」與「製程自動化」等 6 大 AI 應用逐漸導入企業，龐大的推論需求與雲邊端多元部署，使 AI 基礎設施正面臨「規格重整」。本報告將為供應鏈提供規格導航與硬體需求預測，做為定義下一代產品的絕佳參考，為您精準對接龐大的 AI 基礎設施商機。

No.8, 2026/3

SPECIAL REPORT

執行摘要

隨著生成式 AI 從研發階段邁入企業規模化導入，AI 基礎設施正出現關鍵結構轉折，並引爆下一波算力需求。相較於以雲端為核心、以模型訓練為主的第一波 AI 投資浪潮，企業端實際應用的擴散，正使推論算力的成長性逐步超越訓練算力，並提高企業自建地端平台的意願。雲端算力面臨架構重新檢視，以藉此提高推論效能，維持龐大企業 AI 市場上的主導地位。

大語言模型正走向兆參數量、思維鏈推論、多模態輸出、Agentic AI 代理等「四大趨勢」，引領對話機器人、軟體開發、圖像生成、影像生成，企業營運資訊自動化，與製程自動化等「六大 AI 應用」逐漸導入企業。多元應用帶動下，企業不再僅依賴大規模資料中心提供 AI 服務，而是基於成本控管、資料主權、延遲與可靠性等考量，來選擇最佳的 AI 基礎設施部署。

提供雲端算力的雲端服務商，為維持在 AI 算力的統御地位，一方面持續進行龐大資本投資，與強化 IaaS、PaaS AI 服務，來供企業無虞使用，另一方面也因應推論需求而重新檢視運算架構，藉此提高推論效率，降低企業的算力支出。部分雲端服務商更進一步擴大自家 SaaS 服務，以進一步鞏固雲端算力的需求，來提高算力投資的規模化門檻，阻止企業因應 AI 而自建算力的意圖。

在此情況下，雲端業者是否還能獨攬 AI 算力?長期支持 LLM 龐大訓練算力的 NVIDIA 平台，能否在推論當道的現今持續坐穩算力龍頭?雲端算力基於這波龐大企業 AI 需求，究竟能帶來多少高階 AI 伺服器的未來出貨?

本報告以企業 AI 導入為核心視角，系統性分析算力架構的重塑方向，並證明雲端算力在龐大企業 AI 需求中，如何維持既有地位，進而推導出主要的雲端業者、LLM 業者與算力平台業者，究竟誰能在其中獲利。

透過本報告，供應鏈業者可更清楚理解未來 AI 基礎設施的主流架構樣貌，作為下一代產品 Roadmap、規格定義、與選擇合作對象的重要參考，以精準對接企業 AI 時代下持續擴大的基礎設施商機。

蕭聖倫 Jim Hsiao

DIGITIMES
資深分析師

目錄

圖目錄

LLM 技術正從單一規模競爭向多元化、實用化方向的演進，並在 2026 年趨向四大關鍵方向，首先，模型參數量呈現 M 型化發展，較大型的模型持續提高參數以追求極致效能，而較小的模型則維持在 100 億個以下，強調效率與可部署性；第二為指標業者持續聚焦在分解複雜任務與強化推理能力，使模型能更精準處理多步驟問題；第三為多模態支援成為主流，模型不僅能處理文字，還可整合圖像、音頻、視頻等多種檔案格式的輸入與輸出；最後，Agentic AI 的興起，讓 LLM 從被動工具轉變為能自動化流程的 AI 代理。

目前絕大多數的生成式 AI 應用所需的算力，都是發生在雲端，生成式 AI 應用領域受多項因素，影響其算力配置的位置，這些因素（探討影響算力發生位置的變數）包含「延遲時間」、「資料公開性」、「任務複雜度」、「資料規模」、「算力需求」、「標準化程度」以及「目前上雲程度」等，這些因素共同決定是否將算力配置於雲端、混合、裝置，以優化效能與成本。

圖 生成式 AI 應用領域運算位置屬性評估

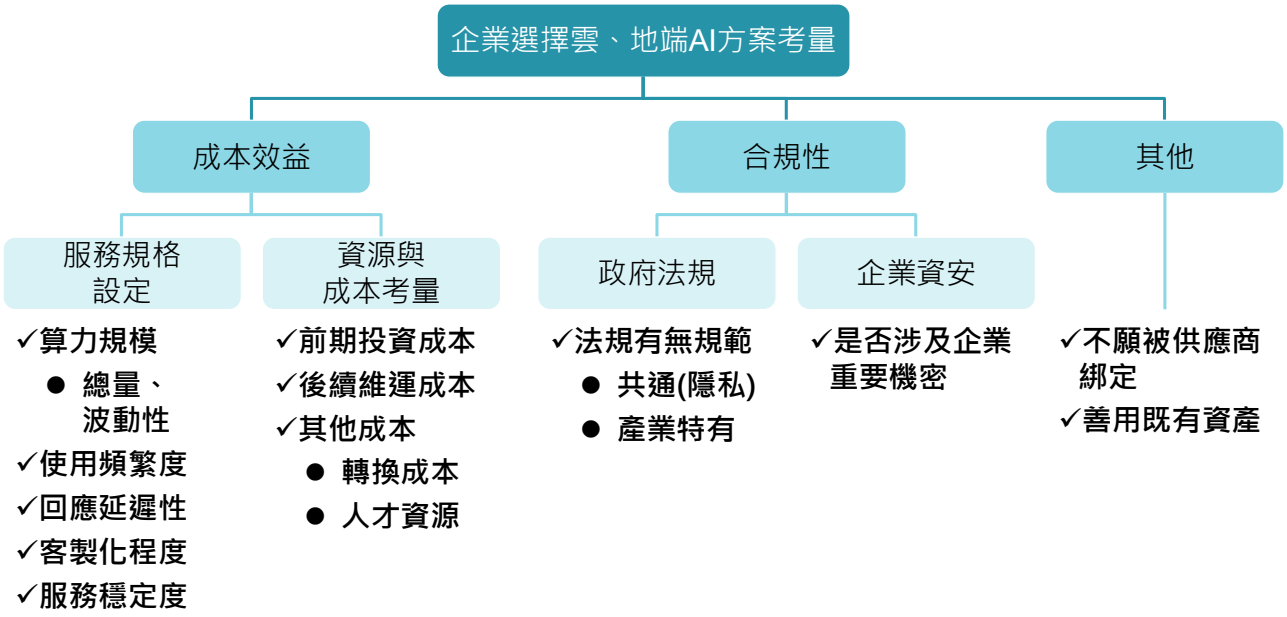
應用領域	影響算力配置位置因素							運算位置屬性
	延遲時間	資料公開性	任務複雜度	資料規模	算力需求	標準化程度	目前上雲程度	
對話機器人	低	中	中	高	高	高	高	雲端
企業營運資訊自動化	中	低	高	高	高	中	低	混合
製程自動化	低	低	中	中	中	低	低	邊緣
軟體開發	中	中	低	低	低	高	高	混合
圖像生成與處理	中	高	低	低	中	中	高	混合
影像生成與處理	高	高	中	中	高	低	高	雲端

資料來源：DIGITIMES，2026/3

全球 AI 產業正逐漸從技術研發走向服務落地，從企業用戶角度而言，研擬導入 AI 技術的投資計畫時，將率先面臨選擇「雲端」或「地端」部署方案的挑戰。而 AI 服務潛在價值大、

導入成本高的特性，企業將更謹慎考量更多因素，但考量因素大致可歸納為成本效益、合規性與其他（含企業自主性）等面向。這三大面向不僅影響企業用戶選擇 AI 服務的部署類型，甚至將影響廣大 AI 軟硬體供應鏈的商機變化。

圖　企業選擇 AI 服務部署方案的考量因素



資料來源：DIGITIMES，2026/3

AI 應用市場自 2026 年起可望明顯爆發，而此變化也將對支撐資料中心運算架構與關鍵硬體帶來全新的挑戰與機會。目前在頂尖 LLM 的規模擴張法則逐漸失靈，加上 CoT 推論的擴張效果開始超越訓練，目前 CoT 推論的效果仍需依賴動輒數兆參數的尖頂模型，因此主要仍在大規模的雲端叢集運行。

大型雲端業者在雲端 AI 運算的早期投入與競爭，將預先為未來數年的 AI 應用市場帶來充足且經濟的運算與開發資源，企業在進行 AI 應用開發時，將因成本較低及開發門檻較低，而優先選擇大型雲端業者所提供的 AI 雲端運算方案。

快速發展的多模態推論模型也更適合雲端推論，其圖像、音訊及影像等生成對運算資料型態較文字有不同要求。目前市場上主流多模態模型如 OpenAI Sora、Google Veo 3 的內部模型皆採 difussion transformer 架構，即結合擴散模型與 LLM 的核心架構 transformer 來完成影像模型的多模態生成，其架構較傳統語言模型更為複雜，在生成高畫質影像內容時，對加速器的運算要求將非常高，因此較適合以雲端叢集算力來處理其推論。

推論相關需求快速發展，但過去 AI 伺服器主要以訓練做為設計基礎，使產業檢討現行的硬體架構並開始進行改良，Dynamo 軟體方案及 2027 年初預計推出的 Rubin CP0X GPU 與相關硬體架構，即為 NVIDIA 對硬體推論優化的解答之一，然隨著各 AI 應用爆發帶來更大量的推論需求，AI 伺服器相關設計仍有改善空間，近期市場關注的改革標的主要是新型記憶體架構的導入。

亞馬遜 AI IaaS 相關的算力服務，主要透過 AWS 下的 EC2 （Elastic Compute Cloud）服務提供。AI PaaS 的服務則包括了提供模型操作與 Agent 建置服務的 Bedrock、模型建置服務的 SageMaker，及可以透過自然語言建立 App 的工具 AWS App Studio。AI SaaS 主要包括了商用級對話機器人 Amazon Q，電商購物助理 Rufus 等服務。除了自家 ASIC 算力，為了滿足 AWS 上大量採用 GPU 算力來訓練或推論生成式 AI 的客戶，亞馬遜仍同時規劃大量 GPU 算力，包括 NVIDIA 與超微方案，據 DIGITIMES 了解，在 UBB 架構 AI 伺服器的採購中，亞馬遜一直是領先的採購者。

圖　亞馬遜 AI IAAS 歷年關鍵服務內容與硬體建置變化

		2016~2022	2023~2025H1	2025H2~
AI IaaS	關鍵服務內容	**Anthropic算力供應** • 2021年開始合作，為P4d初期用戶，並將其用於訓練Claude 1。	**Anthropic算力供應** • Claude 3與Claude 3.5採AWS P5及TPU訓練，Trainium主要用於實驗測試與推論。 • Claude 4為首個在訓練過程中採用Trainium2的模型。	**Anthropic算力供應** • Claude5預期將完全以Trainium2與Trainium3訓練完成。 **OpenAI算力供應** • 投資500億，2027年將開始採用Trainium3與Trainium4算力。
AI算力建置		**NVIDIA** • 2017年P3 GPU VM提供V100算力。 • 2020年推出P4d大型叢集，4,000顆A100可同步運算，滿足第一波LLM訓練需求。 **Amazon** • 2019年推出第1代Inferentia，2022年第1代Trainium上線。	**Amazon** • 2023年推出第2代Inferentia，2024年底推出第2代Trainium。 **NVIDIA** • 2023年開始建置P5(H100)算力，2024年底推出P5e(H200)。 • 2025年中推出P6e-GB200 及P6-B200。	**Amazon** • Project Rainier將建置1M顆以上的Trainium2與Trainium3。 **NVIDIA** • P6e-GB300服務於2025年12月啟動 • Vera Rubin系統預計於2026年下半推出，Rubin Ultra於2027年下半推出。

資料來源：DIGITIMES，2026/3

分析師團隊

OUR TEAM

共同編撰分析師

蕭聖倫 資深分析師
伺服器

黃耀漢 分析師
人工智慧

陳冠榮 分析師
雲端服務

張珩 分析師
電腦運算

陳加鑫 分析師
伺服器

About DIGITIMES 介紹

【關於 DIGITIMES】DIGITIMES 成立於 1998 年，為大中華地區報導科技產業全球供應鏈、區域市場、科技應用及市場趨勢首屈一指的專業媒體平台，具備貫穿產業上中下游與終端市場的研究數據、產銷資料與專業評析，並提供諮詢服務為客戶帶來產業宏觀趨勢與注入前瞻價值。

DIGITIMES 研究服務掌握科技產業全球供應鏈，專注於資訊、消費性電子、通訊、半導體、汽車科技、人工智慧、物聯網及平面顯示器等領域，以及區域市場的研究報告。自成立以來，研究中心已發表超過 7,000 篇高影響力的報告。未來，研究中心將持續推動前沿科技研究，擴展內容和服務範圍，致力成為提供關鍵洞察和引領科技發展的先驅。

研究報告

研究報告涵蓋七大領域 23 個頻道與全球產業數據，每年發佈超過 300 篇報告，內容以分析全球及台灣產銷狀況、產業發展現況、產品技術趨勢、領導廠商策略及競爭態勢。包括區域及新興市場研究和關鍵零組件發展，即時提供客戶所需的產業情報，為台灣最專業且權威的產業分析服務。

到府簡報

以九大分類提供宏觀大勢/供應鏈布局、半導體、Display Trends 通訊產業趨勢/5G/B5G/、垂直應用/專網/O-RAN、NB/高效能運算（HPC）/Cloud、EV/未來車、AI、物聯網（IoT）、智慧應用/數位轉型等領域的研究報告為基礎，整合當前產業發展熱門議題，提供企業專屬的到府簡報服務。

系列論壇

以科技大勢為焦點的系列論壇，每年精心策劃四場圍繞當前熱門議題的精彩活動，探討最新科技趨勢與創新應用，此外，還有一場年度重磅論壇，科技大勢展望未來，解析未來科技發展方向與潛在機遇，幫助企業掌握先機，提升競爭力。論壇旨在促進科技與產業的深度融合，推動創新發展，共同迎接科技新時代的挑戰和機遇。

Special Report

每年推出四篇長篇報告，深入分析當前焦點產業，提供全面的產業脈絡、市場動態及技術演進。報告旨在為企業領袖、投資者和從業者提供權威的資訊和深刻的洞察，幫助他們掌握產業趨勢，做出明智決策。通過詳細的數據分析和專業見解，助企業在快速變化的市場中保持競爭優勢，洞悉未來發展機遇，驅動創新和增長。

顧問專案

根據企業的研究需求，訂定專屬研究範疇，提供量身定制的研究服務。專注於資訊、消費性電子、通訊、半導體、汽車科技、人工智慧、物聯網、平面顯示器等領域。以深入的產業分析和專業見解，助力企業洞悉市場趨勢，制定精確策略，在快速變化的科技環境中抓住機遇，實現創新和提升競爭力。

DIGITIMES： 瞭解更多

聯絡我們

有任何問題，歡迎隨時跟我們聯繫，我們很樂意為您服務。

服務時間：周一至周五 09:00~18:00

傳真：+886-2-8712-3366

客服專線：+886-2-8712-5398

客服信箱：member@digitimes.com

免責聲明

本公司提供之報告內容係根據本公司認可之資料來源，並基於特定日期所進行之判斷，惟由於產業倍速變動、資訊之不完整及其他不確定因素，本公司並不保證本研究報告內容於未來仍具備正確性與完整性，報告中所有的意見及預估，如有變更恕不另行通知。

本研究報告資訊，僅提供客戶做為一般參考，並非針對特定對象提供專屬之建議，使用者如有參考或內部引用時做為決策依據，應自行判斷衡量該資訊，並自負引用之結果。除顯係可歸責乙方之事由外，使用者不得因使用本研究報告資訊所造成之任何直接或間接之損害要求乙方負責。本報告之內容取材自據信為可靠之資料來源，但概不以明示或默示的方式，對資料之準確性、完整性或正確性作出任何陳述或保證。本研究報告載述意見進行更改與撤回不再另行通知使用者。本研究報告內容屬大椽股份有限公司（以下簡稱DIGITIMES）之著作權，嚴禁抄襲與仿造，具體詳請參閱本報告之著作權聲明。

著作權聲明

大椽股份有限公司（DIGITIMES）所屬網站與平面刊物（DIGITIMES 科技網、 智慧應用、 椽經閣、活動＋、電子時報等）上刊載的所有內容，包括但不限於文字報導、 照片、 影像、 插圖、錄音片、影音片、檔案、網站畫面的安排、網頁設計等素材，均受到中華民國著作權法、國際著作權法律及智慧財產權相關法律的保障，相關智慧財產權包括但不限於商標權、專利權、著作權、營業秘密與專有技術等。

網站與平面刊物內容的著作權為大椽股份有限公司（DIGITIMES）所有，或其他授權 DIGITIMES 使用的內容提供者所有。

使用者下載或拷貝網站與平面刊物的內容或服務僅限於供個人、非商業用途之使用，但不得以任何形式傳輸、重製、散布或提供予公眾。使用人利用時必須遵守著作權法的所有相關規定，不可變更、發行、播送、轉賣、重製、改作、散布、表演、展示或利用 DIGITIMES 所屬網站與平面刊物上局部或全部內容及服務以賺取利益。