

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，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。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，並驅動 AI 工作負載由集中式雲端，走向結合私有雲、邊緣節點與終端設備的混合式部署。

在此背景下，企業不再僅依賴大規模資料中心提供的 AI 服務，而是基於成本控管、資料主權、延遲與可靠性等考量，開始選擇最佳的 AI 基礎設施部署。同時，LLM 與 AI 平台業者亦逐步由雲端 API 服務，轉向對企業的直接服務模式，進一步加速企業端 AI 算力需求的地化。

這一系列變化，使 AI 硬體不再只追求極致效能的單一規格競賽，而是進入一個以部署場域、工作負載與應用情境為導向的「規格重整」階段。企業對 AI 的需求，還要考慮到隱私安全、即時回應，還有複雜應用下的高額 API 算力支出，在此情況下，超大規模資料中心是否還能獨攬 AI 算力？私有雲、邊緣裝置與終端設備是否瓜分算力市場？既有以訓練為務的雲端基礎建設正面臨硬體檢視，因應推論所需的高階 AI 伺服器架構正進行處理器、記憶體、網路與加速器的重新配置。

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，並深入探討其對 AI 硬體配置與供應鏈分工所帶來的影響。研究重點不僅在於趨勢判讀，更著眼於實際可落地的硬體規格演進路徑。

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、規格定義與市場布局的重要參考，精準對接企業 AI 時代下持續擴大的基礎設施商機。



蕭聖倫

蕭聖倫 Jim Hsiao

DIGITIMES
資深分析師

目錄

1. LLM 掀起全球新一波 AI 市場爆發	10
1.1. AI 技術核心 LLM 發展趨勢	11
1.2. 企業軟體導入生成式 AI 的發展趨勢	18
2. 企業 AI 業者服務布局與策略	33
2.1. 企業 AI 服務「重資本、深知識」的市場特性	33
2.2. 企業 AI 服務市場重點供應商格局	37
2.3. 企業 AI 服務市場現況與未來趨勢	47
3. 生成式 AI 應用成熟引領硬體發展多元走向	54
3.1. 模型訓練的規模擴張法則逐漸收斂 市場越益重視模型推論的效率	54
3.2. LLM 推論表現仍與模型規模顯著相關 短中期仍依賴大規模雲端叢集	58
3.3. 推論快速發展對當代 AI 伺服器架構帶來一定壓力	61
3.4. Dynamo 及 Rubin CPX 為 NVIDIA 針對推論需求推出的改良方案	65
3.5. AI 硬體推論能力仍需強化 最大挑戰來自記憶體	68
4. 主要企業 AI 服務供應商硬體布局	71
4.1. GOOGLE	71
4.2. 亞馬遜	77
4.3. 微軟	81
4.4. 甲骨文	87
4.5. Meta	93
4.6. xAI	97
4.7. OpenAI 與 Anthropic	99
4.8. 高階 AI 伺服器未來 3 年成長趨勢	105
分析師團隊	109
ABOUT DIGITIMES 介紹	110
免責聲明	111
著作權聲明	111

圖目錄

圖 1	AI 技術熱潮發展歷程	10
圖 2	AI 涵蓋技術的關係圖	12
圖 3	2026 年 LLM 發展四大關鍵趨勢	13
圖 4	LLM 參數量 M 型化趨勢	14
圖 5	2025 年全球指標業 LLM 功能特色	16
圖 6	AI 技術發展趨勢	17
圖 7	2024~2030 年全球 LLM 市場規模變化	18
圖 8	2028 年生成式 AI 潛在應用市場預測	20
圖 9	2025~2028 年生成式 AI「對話機器人」應用市場熱區趨勢	22
圖 10	2025~2028 年生成式 AI「軟體開發」應用市場熱區趨勢	24
圖 11	2025~2028 年生成式 AI「圖像生成與處理」應用市場熱區趨勢	25
圖 12	2025~2028 年生成式 AI「影像生成與處理」應用市場熱區趨勢	26
圖 13	2025~2028 年生成式 AI「企業營運資訊自動化」應用市場熱區趨勢	28
圖 14	2025~2028 年生成式 AI「製程自動化」應用市場熱區趨勢	29
圖 15	生成式 AI 應用領域運算位置屬性評估	30
圖 16	企業導入生成式 AI 軟體的內外部需求條件與軟硬體的挑戰	32
圖 17	企業 AI 服務關鍵資源與市場特性	35
圖 18	企業選擇 AI 服務部署方案的考量因素	37
圖 19	企業 AI 服務市場重點供應商類型與代表業者	38
圖 20	大型 CSP AI 服務的方案與布局	41
圖 21	企業軟體業者 AI 服務的方案與布局	42
圖 22	企業 IT 大廠 AI 服務的方案與布局	44
圖 23	AI 模型新創的企業 AI 服務與特點	46
圖 24	企業 AI 服務供應商布局及雲端算力比重對照	47
圖 25	六大企業 AI 服務目前主流、未來 3 年與潛力架構變化示意圖	49
圖 26	企業 AI 服務供應商相對優勢和潛在應用領域	50
圖 27	企業 AI 服務市場兩大趨勢與驅動因素	52
圖 28	企業 AI 服務市場現況與未來變化	53
圖 29	訓練 LLM 所需算力預測	55
圖 30	OpenAI o1 模型透過訓練或思考推論所取得的 AIME 答題正確率變化	56
圖 31	決定推論運算硬體效能的兩大要素與其反向關係	57
圖 32	NVIDIA Hopper 與 Blackwell 系統的推論算力演進	58
圖 33	五大雲端巨頭 2022~2027 資料中心資本支出年增率預測	59
圖 34	AI 負載對 AI 伺服器關鍵組件的負荷要求狀況	63
圖 35	NVIDIA 傳統推論處理方式與採用 Dynamo 推論的比較	65
圖 36	Rubin CPX 與 Rubin200 GPU 主要規格比較	67
圖 37	HBM 與 HBF 封裝架構比較	69
圖 38	3D 記憶體-邏輯晶片堆疊架構	70
圖 39	Google AI IaaS 歷年關鍵服務內容與硬體建置變化	72
圖 40	Google AI SaaS 歷年關鍵服務內容變化	74
圖 41	Google 加速器採用現況與未來 3 年展望	76

圖 42	Google 2023~2028 各運算平台高階 AI 伺服器出貨量預估	76
圖 43	亞馬遜 AI IaaS 歷年關鍵服務內容與硬體建置變化	79
圖 44	亞馬遜加速器採用現況與未來 3 年展望	80
圖 45	亞馬遜 2023~2028 各運算平台高階 AI 伺服器出貨量預估	81
圖 46	微軟 AI IaaS 歷年關鍵服務內容與硬體建置變化	82
圖 47	微軟 AI SaaS 歷年關鍵服務內容變化	84
圖 48	微軟 AI PaaS 歷年關鍵服務內容變化	85
圖 49	微軟加速器採用現況與未來 3 年展望	86
圖 50	微軟 2023~2028 各運算平台高階 AI 伺服器出貨量預估	87
圖 51	甲骨文 AI IaaS 歷年關鍵服務內容與硬體建置變化	89
圖 52	甲骨文 AI PaaS 歷年關鍵服務內容變化	90
圖 53	甲骨文 AI SaaS 歷年關鍵服務內容變化	91
圖 54	甲骨文加速器採用現況與未來 3 年展望	92
圖 55	甲骨文 2023~2028 各運算平台高階 AI 伺服器出貨量預估	93
圖 56	傳統推薦系統與採用 GEM 模型推薦系統的比較	94
圖 57	Meta 加速器採用現況與未來 3 年展望	95
圖 58	Meta 2023~2028 各運算平台高階 AI 伺服器出貨量預估	96
圖 59	xAI 訓練 LLM 所需算力預測	98
圖 60	xAI 2023~2028 各運算平台高階 AI 伺服器出貨量預估	99
圖 61	ChatGPT 與 Claude 各種方案與差異化功能重點比較	101
圖 62	OpenAI 加速器採用現況與未來 3 年展望	102
圖 63	OpenAI 2023~2028 各運算平台高階 AI 伺服器出貨量預估	103
圖 64	Anthropic 加速器採用現況與未來 3 年展望	104
圖 65	Anthropic 2023~2028 各運算平台高階 AI 伺服器出貨量預估	105
圖 66	全球高階 AI 伺服器 2023~2028 年出貨量預估	105
圖 67	主要業者全球高階 AI 伺服器 2023~2028 年出貨量佔比預估	107
圖 68	主要加速器平台全球高階 AI 伺服器 2023~2028 年出貨量佔比預估	108

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

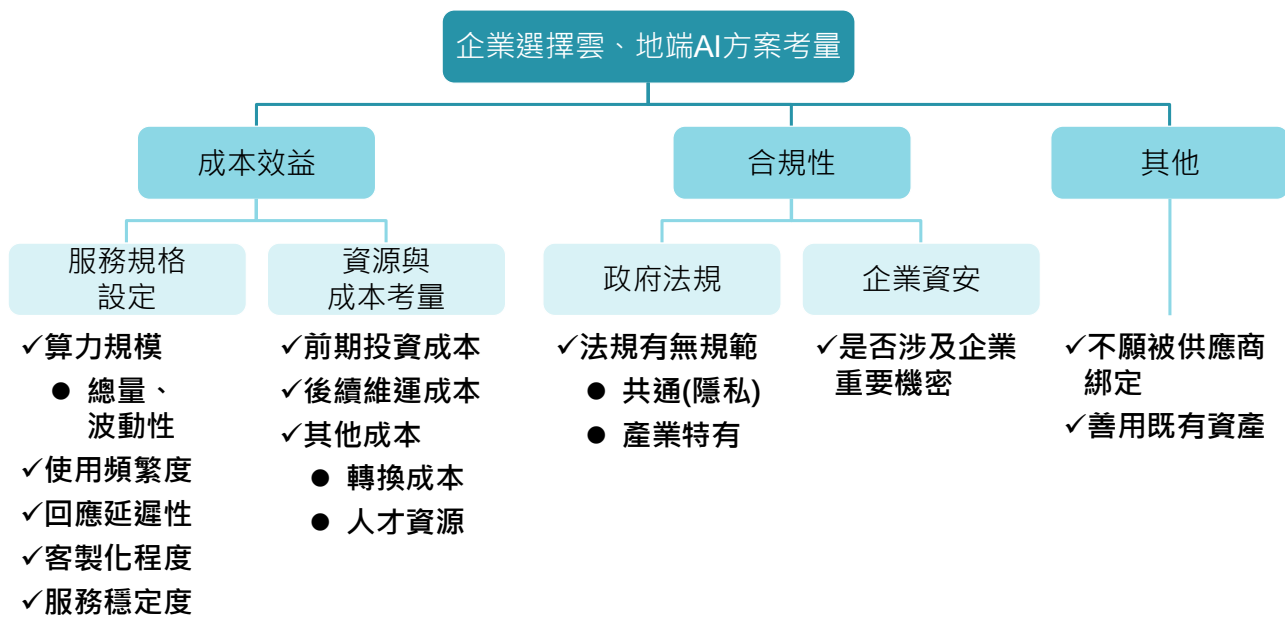
應用領域	影響算力配置位置因素							運算位置屬性
	延遲時間	資料公開性	任務複雜度	資料規模	算力需求	標準化程度	目前上雲程度	
對話機器人	低	中	中	高	高	高	高	雲端
企業營運 資訊自動化	中	低	高	高	高	中	低	混合
製程自動化	低	低	中	中	中	低	低	邊緣
軟體開發	中	中	低	低	低	高	高	混合
圖像生成與處理	中	高	低	低	中	中	高	混合
影像生成與處理	高	高	中	中	高	低	高	雲端

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 diffusion 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 - Claude 5 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 年下半年推出，Rubin Ultra 於 2027 年下半年推出。

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



張衍 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，提供量身定製的研究服務。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橡股份有限公司（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

大橡股份有限公司（DIGITIMES）所屬網站與平面刊物（DIGITIMES 科技網、智慧應用、橡經閣、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橡股份有限公司（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。