Allianz Global Investors
安聯投資

2026年6月

從雲端到邊緣：人工智能機遇的下一波浪潮

隨著人工智能（AI）的發展重心由快速模型訓練，逐步轉向實際部署，市場關注正愈來愈集中於推理（inference），以及AI如何實現高速、高效且具規模化的成果輸出。

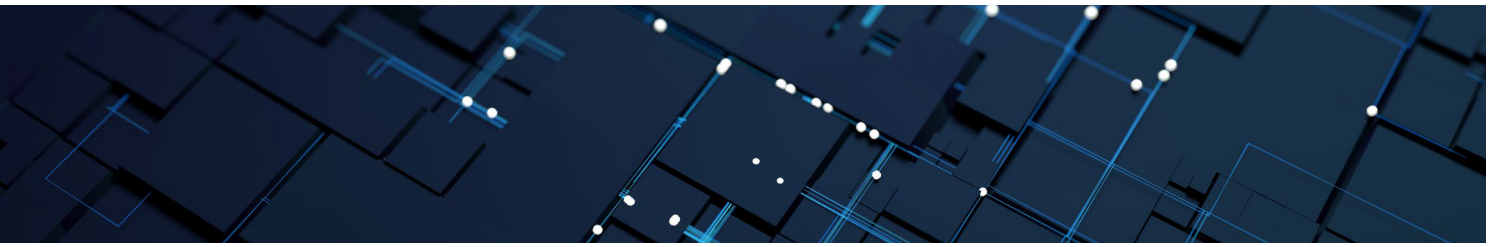
大型語言模型（LLMs）的訓練需要龐大的AI晶片集群，幾乎完全在雲端環境中進行，由超大規模雲端服務供應商（hyperscalers）提供運算資源，當中包括圖形處理器（GPU）及專用集成電路（ASIC）。相對而言，推理則是將模型所學知識應用於實際場景的過程，所採用的晶片組合不同，重點在於低延遲（即快速回應需求）及高吞吐量（即同時處理大量請求）。

過去數年，模型訓練發展迅速，主要優先追求速度而非成本效益。隨著大型語言模型的技術突破，我們正邁向AI發展的重要轉捩點，使推理及AI的商業應用逐步成為核心焦點。值得留意的是，我們到達這個轉捩點的速度相當之快。OpenAI創始成員之一、並提出「vibe coding」概念的Andrej Karpathy

Jeremy Gleeson
首席投資總監 - 全球科技股票

Sunil George
高級投資組合經理 - 全球科技股票

Brad Reynolds
投資組合經理 - 全球科技股票



於2025年10月曾表示，市場對編程代理（coding agents）的熱潮被「過度誇大」。然而僅僅三個月後，他便指出：

「大型語言模型代理（特別是Claude與Codex）的能力在2025年12月前後跨越了某種一致性門檻，並在軟件工程領域引發範式轉變……這是我在過去20年編程經驗中，對基本編程流程影響最大的一次變革，而且僅在短短數週內發生。」

這並非前沿科技首次被低估，但此次變革的速度尤為驚人。研究及顧問機構Gartner預測，到2026年，55%的AI優化基建即服務（IaaS）支出將用於推理工作負載，並在2029年上升至超過65%。這引發了一個關鍵問題：在推理發展的下一階段，應採用何種最適合的架構以達致最佳效率？其中，邊緣AI（edge AI）——即在本地或接近使用場景運行AI模型，而非完全依賴雲端——被視為具領先優勢的解決方案之一。

由內容傳輸邁向分散式AI基建

相比集中式雲端運算，邊緣AI（edge AI）利用終端設備及／或分布於地理位置上更接近用戶請求的節點網絡（points of presence）。從這一角度而言，這並非全新的架構，而是由多項過往被歸類為「物聯網」（Internet of Things）的元素所組成。

這些分散式網絡過去的主要客戶群集中於娛樂服務，例如網絡遊戲及媒體串流。然而，隨著AI持續快速發展，如何善用各種可用渠道分配運算資源，正變得前所未有地重要。

在此架構之下，其中一個備受關注的分支為由小型數據中心所組成的網絡，即內容傳輸網絡（Content Delivery Networks, CDNs）。由於CDN具備高度分散的特性，現時亦開始在AI推理（inference）方面發揮重要作用。

由內容傳輸網絡（CDN）轉型為邊緣AI供應商的趨勢，可見於近期Akamai與Anthropic達成的18億美元交易。這個傳統CDN在全球130個國家擁有超過4,300個節點（points of presence）。儘管其最初是為娛樂服務（如內容傳輸）而建立，但其網絡架構亦能自然延伸至邊緣的分散式推理應用。這一案例突顯，邊緣AI將如何與雲端AI形成互補，並逐步成為高效部署AI推理的重要工具。將推理轉移至邊緣所帶來的優勢包括：更低延遲、更高吞吐量、更佳可用性與私隱保障、減少頻寬瓶頸，以及由於本地設備推理的邊際成本較低，從而降低每次查詢的成本。

由於邊緣運算在計算能力方面存在限制，具高度複雜性的AI模型仍需於雲端處理。然而，對於較為專注的應用場景而言，若推理速度及高吞吐量為關鍵效能指標，則相關推理工作可於邊緣端進行優化。相關應用包括但不限於：

- 自動駕駛——受惠於快速推理及高吞吐量
- 穿戴式裝置／擴增實境（AR）／虛擬實境（VR）——受惠於持續可用性
- 工業機械人——受惠於可用性及較低的單次查詢成本
- 醫療保健——受惠於更佳私隱保障及減少頻寬瓶頸

未來或將出現一種分流且分層的推理架構：低成本且對延遲敏感的任務會於邊緣端執行，而更複雜的工作負載則保留於雲端，並根據運算強度進行分配與路由。隨著兩個層級的效率提升及能力擴展，競爭優勢或將由單純模型規模，轉向工作負載調度及成本優化能力。

人工智能價值鏈的新機遇

向邊緣端進行推論（inference）的轉變正創造新的投資機會，同時亦擴展既有的投資範疇。而其中許多機會均與**我們已識別的**全球科技主題一致。

最明顯的新機遇在於內容**傳遞網絡（CDN）向邊緣AI供應商的轉型**。隨著網絡服務收入增長加快，這些供應商將由商品化的頻寬提供者轉變為即時推論的關鍵基礎設施，從而受惠。此發展亦與其現有服務形成協同效應，因為不少供應商同時提供網絡安全及網絡功能。這一轉型亦可能帶來估值重評，而這種情況在其他高度受惠於AI的子行業中亦曾出現。

另一個既有的投資機會，亦是全球科技團隊在2025年的主題之一，即光學網絡行業，預期將再度受惠。該行業促成**分佈式網絡**之間的**高速連接**。隨著AI網絡由集中於超大規模雲端架構，擴展至涵蓋數以千計接入點（points of presence）的廣泛邊緣網絡，光學網絡將從中受益。光學網絡可支援推論所需的更高流量吞吐量及傳輸速度，特別是在分散式邊緣網絡環境下的推論需求。

此外，隨著AI逐步與現實世界互動，邊緣推論亦將進一步**推動模擬半導體（analog semiconductors）的復興**。每項裝置的模擬半導體含量增加，加上因邊緣推論而出現的更多分散式終端設備，將進一步帶動該行業的增長。這亦是全球科技團隊於2026年的主題之一，我們對此主題的發展前景仍然充滿信心。

最後，儘管雲端架構曾帶動記憶體行業出現以高頻寬記憶體（HBM）及動態隨機存取記憶體（DRAM）為核心、以提升吞吐量為重點的超級周期，邊緣推論或將**推動對不同類型記憶體需求**的擴展。鍵值快取（KV Cache）作為AI對話過程中的**短期工作記憶**，正突顯此趨勢的重要性；若缺乏該功能，AI在生成每一個新詞時將出現類似失憶的情況，無法記住提示內容的開端。KV Cache 須儲存在電腦中速度最快的記憶體（即DRAM）中，以便AI能即時存取，從而降低延遲。無論哪種記憶體最適合邊緣推論，整體趨勢很可能為**記憶體超級周期**提供額外動力，並推動需求更趨多元化。這亦是全球科技團隊於2026年的另一重要主題。

整體而言，上述發展顯示，邊緣推論不僅是一項技術變革，更代表AI生態系統的結構性擴展，從而在價值鏈各環節釋放新的效率、應用場景及投資機遇。

發行人:

香港 – 安聯環球投資亞太有限公司

MACS-P-20811613

ADM-W5661898