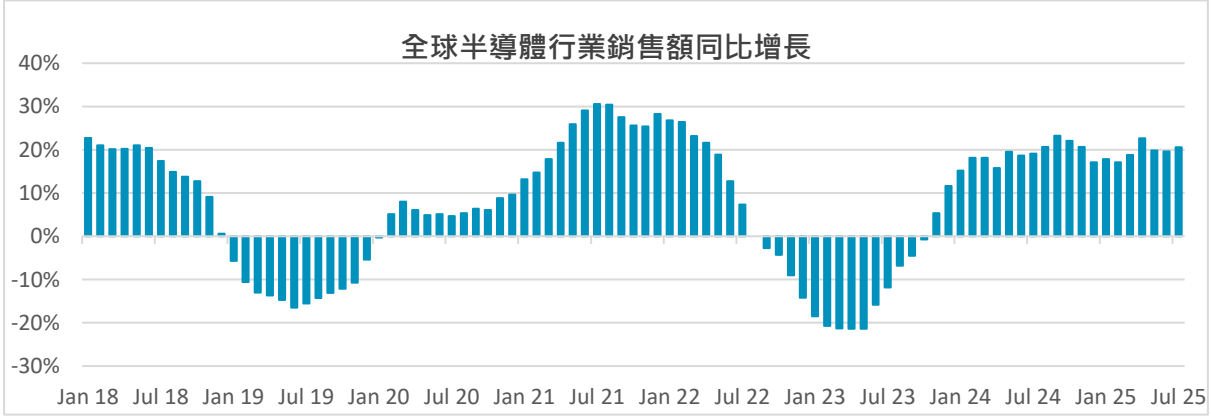
半導體如何塑造新一代 AI 創新格局

15-10-2025

人工智能（AI）依然是當今最具革命性的技術，並由一眾半導體公司主導和引領突破性進展。在邏輯和記憶晶片需求的帶動下，全球半導體市場預計今年將延續 2024 年的強勁復甦，全年增長 18%，總收入達到 8,000 億美元，其中美洲及亞太區預期將會錄得最高增長。數據中心的擴張繼續帶動顯著增長，專注於 AI 和半導體創新的公司尤其受惠。



資料來源：美國半導體行業協會。截至 2025 年 9 月 5 日。

AI 維持強勁增長

儘管面臨關稅上升及經濟阻力增加等挑戰，隨著 AI 驅動的變現機會逐步浮現，超大規模雲端服務商（hyperscalers）的資本支出繼續維持上升趨勢。2025 年第二季度，全球數據中心資本支出按年飆升 43%。微軟、亞馬遜和 Google 均表示，AI 工作負載需求持續超出現有基礎設施的承載能力，預計全年將繼續擴大新增容量以應對需求增長。OpenAI 擬斥資約 4,000 億美元，聯同甲骨文及軟銀集團，在美國興建五個全新數據中心據點。這些新據點最終將具備 7 吉瓦（gigawatts）的電力容量，顯著提升 OpenAI 的運算能力，以支援其各項服務。Meta 亦正在建設多個多吉瓦級數據中心集群，以加速實現其 AI 目標。其中首個數據中心預計將於明年啓用，屆時其 2026 年的資本支出將進一步上升。Meta 將 AI 作為其廣告策略重心，並計劃在明年年底前讓品牌客戶能夠全面使用 AI 工具來設計和制定具針對性的廣告活動。根據客戶的預算，這些新工具將可生成整個廣告，包括圖片、影片和文字，並發送給目標受眾。

過往，AI 的需求主要集中於訓練工作負載，尤其是前沿 AI 模型。在繼續投入資源建設更龐大 AI 模型的同時，科技巨頭亦將轉為重新分配更多的投資於推理（inference）領域。推理是指已完成訓練的 AI 模型處理新數據，從而產生灼見、進行預測或支援決策的階段。訓練模型基本上是一次性支出，對模型進行指示（推理）生成詞元（token），而每個詞元皆附帶成本。在 Google I/O 2025 大會主題演講中，Alphabet 行政總裁桑德爾·皮查伊（Sundar Pichai）透露，2025 年 4 月，Alphabet 所有產品和應用程序編程接口（API）處理高達 480 萬億個詞元，較去年同期激增 50 倍。詞元（token）量快速攀升反映 AI 模型的應用日益普及，預示對運算能力將有更大需求，並帶動對晶片的需求增長。

AI 推理時代

隨著新推理模型面世，投資也加速轉向推理領域。傳統的 AI 模型雖然能快速回應，並擅於識別模式，卻往往無法理解更廣泛的背景，且難以處理複雜的推理任務。推理模型旨在將複雜的問題分解為更小、更易處理的步驟，再透過明確的邏輯推理解決問題。此類模型能夠展示其推理過程，並遵循結構更加清晰的思考流程，因而在處理用戶查詢時需要更長的運算時間。這些模型在推理過程中需要使用明顯較多的運算資源進行推理，以處理複雜的問題。從基本模式識別演進至結構化推理對 AI 的發展至關重要，可釋放 AI 有效應對各種現實世界複雜挑戰的潛能。隨著 AI 應用範圍快速擴大，對推理能力的需求亦將相應激增。

AI 代理的崛起浪潮

AI 代理旨在為機構的運作模式帶來革命性的轉變，令生產力和營運效率實現突破性增長。這些智能系統在設計上能夠透過理解目標、制定決策並採取行動來獨立執行任務，從而達成預設目標。雖然人類負責定義期望結果，但 AI 代理能自主選擇達成有關目標所需的最佳行動。AI 代理的應用範圍廣泛，從支援學術研究、簡化線上購物流程，以至規劃度假行程皆可。客戶服務、銷售與營銷以及資訊科技及網路安全是未來六個月內最常部署或計劃引入 AI 代理的三大業務領域。隨著企業逐步將 AI 代理整合至各種營運場景，對運算基礎設施的需求正在急劇上升。

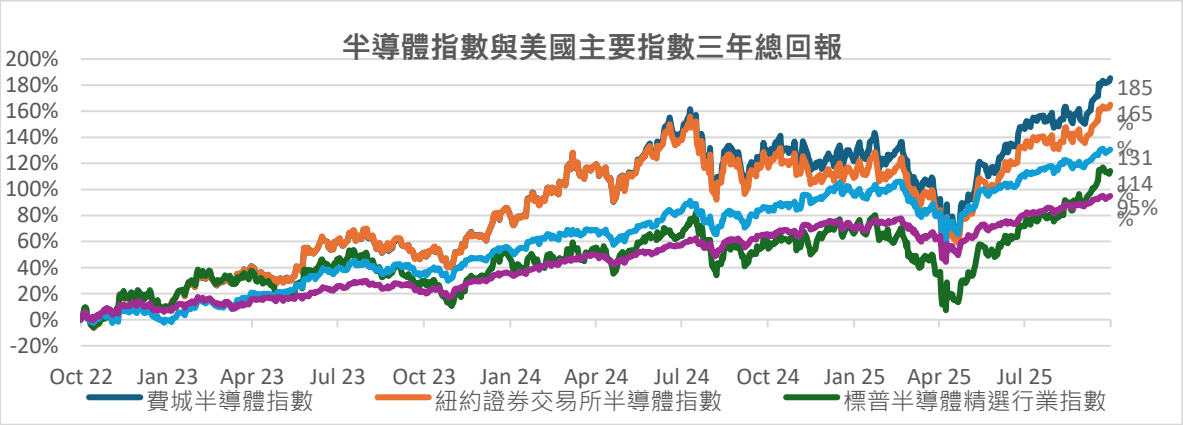
客製化 AI 晶片的崛起

超大規模雲端服務商日益專注於特殊應用晶片（ASIC）基礎設施，以滿足激增的 AI 需求。ASIC 專為特定工作負載打造，不但在執行相關任務時的效率遠高於圖形處理器（GPU），更可大幅降低成本。儘管開發 ASIC 初期涉及龐大的投資金額，但一旦前期成本被吸收，預計應用這些晶片運行生成式 AI 工作負載的長期成本將會下降。例如，Google 於 2025 年 4 月推出專為推理工作負載而設的第七代 Tensor Processing Unit（TPU）「Ironwood」。過去 Google 的自家 TPU 僅限於內部使用，如今公司正擴大

對外開放，以加快雲端業務的增長。邁威爾科技（Marvell Technology）預計客製化運算裝置市場規模將於 2028 年增長至 554 億美元，規模超過 2023 年的八倍。

費城半導體指數（SOX™）— 半導體行業的領先指標

納斯達克的費城半導體指數（SOX）涵蓋前 30 大主要從事半導體設計、分銷、製造和銷售的美國上市公司股票及美國存託憑證（ADR）。過去三年，SOX 的總回報率高達 185%，較紐約證券交易所半導體指數高出 20 個百分點，也是標普半導體精選行業指數回報的 1.6 倍。



資料來源：納斯達克全球指數、彭博。截至 2025 年 9 月 30 日。

總結

為支援日趨複雜的 AI 模型架構，市場對進階訓練能力的需求持續上升，同時，AI 推理已儼然成為關鍵的增長動力。詞元數量急升反映 AI 模型的應用正持續擴大。AI 代理的普及亦勢必改變多個行業的格局，同時大幅推高運算的需求。此動態發展突顯了半導體產業在推動下一波 AI 創新浪潮中擔當重要的角色。納斯達克的費城半導體指數（SOX）在截至 2025 年 9 月底的三年總回報為 185%。

蔡朗賢 (David Tsoi)　CFA 聯盟

蔡朗賢是納斯達克亞太區指數研究部主管，負責亞太地區的指數研究工作。該團隊提供全部納斯達克指數系列的研究，包括全球和美國的廣泛市場、主題、多因子和期權指數。加入納斯達克前，他曾於 Global X ETFs 主導超過 20 只 ETF 的產品開發，當中包括全球首只亞洲綠色債券 ETF，以及亞太區最大規模的備兌買權（covered call）ETF。在此之前，他任職於香港證監會，參與資產管理行業的政策制定以及產品審批，包括香港首只 ESG ETF 的審批。

蔡朗賢擁有多項專業資格，包括特許金融分析師（CFA）、特許另類投資分析師（CAIA）、金融風險管理師（FRM）、ESG 分析師（CESGA）和國際公認反洗錢師（CAMS）認證。他畢業於香港中文大學，獲得工商管理學士一級榮譽學位。

關於「CFA 聯盟」

「CFA 聯盟」是由《CFA 指揮室》過往逾十年以來的節目嘉賓組成。《CFA 指揮室》是由香港特許金融分析師學會（CFA Society Hong Kong）與新城財經台共同策劃的投資者教育計劃，通過訪問金融業及商界的專業人士，提升大眾對業界和各類投資的認識。

*作者的言論純屬個人意見，並不代表本會立場。內容只供參考及作教育用途，並非投資意見，亦非對任何產品或服務的建議、認許或推介。

文章由 香港特許金融分析師學會 CFA Society Hong Kong 協助提供