

科技

AI資本支出一飛衝天 定期定額是王道



報告摘要

- 人工智慧正在從「訓練」典範轉移至「推論」
- 外資近期上調台積電資本支出展望，帶動了半導體設備廠商漲勢
- 記憶體將持續受惠AI由訓練移至推論的典範轉移

2026年01月21日

主題趨勢報

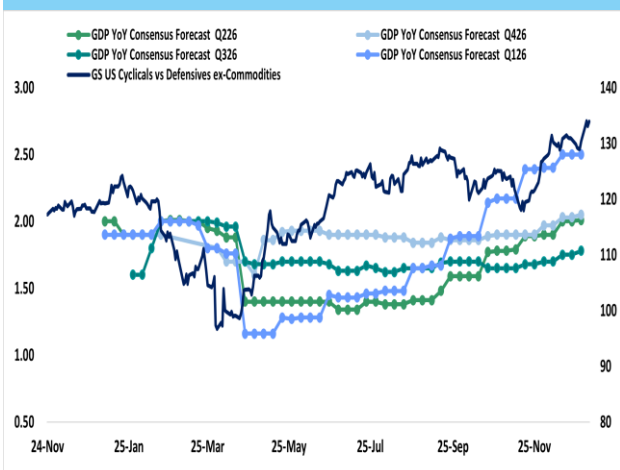
投顧觀點

2026 年開局，美股科技板塊展現了歷史罕見的極致分化，資金高度集中於 AI 基礎設施、記憶體與半導體設備（由CES以及外資上調台積電資本支出預期所帶動），反觀軟體與互聯網板塊，因投資人對 2026 年估值支撐存疑，整體表現疲軟。過去「科技七巨頭」齊漲的格局已瓦解，轉變為鮮明的「零和博弈」。這顯示在缺乏全面增量資金的情況下，市場正在巨頭之間進行激烈的輪動與倉位調整。在風險偏好上，市場呈現強烈的「順週期」特徵。儘管部分總體經濟數據（如 ISM）偏弱，但「循環股對防禦股」的相對表現卻創下自 2025 年 5 月以來最佳，這呼應了市場對美國上半年「減稅、寬鬆與關稅衝擊減輕」的正面經濟展望。這種樂觀情緒引發「追逐Beta」，也就是買進投機性股票的效應，資金溢出至太空、量子運算及無人機等高波動風險資產。隨著市場焦點從 CES 的題材炒作迅速轉向即將到來的 Q4 財報季，市場資金可能重新作出布局。

操作建議

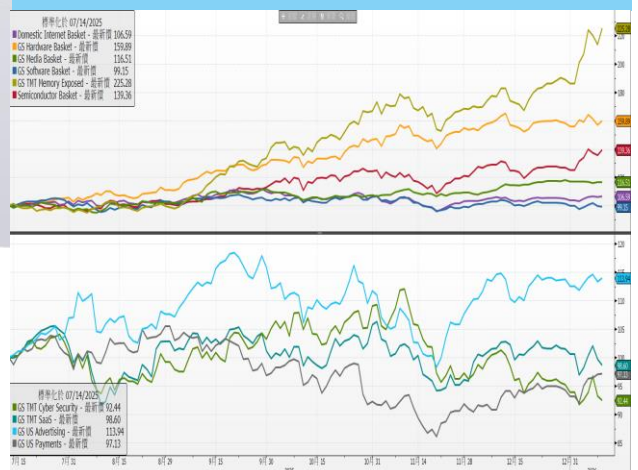
半導體、記憶體、設備商等AI相關基礎建設的長期基本面展望無虞，台積電上調資本支出以及輝達在CES上演講皆強調「供給端仍遠遠未滿足當前對算力的需求」。但不可否認市場在短線的預期已相當滿，同時相關題材類股，估值基本交易在2027年的預期水準，故短線上宜謹慎、切勿過於積極的追高，而長線則維持拉回買進的觀點，故建議長期投資人可透過定期定額的方式進行分批佈局。

循環股對防禦股的相對表現創下自 2025 年 5 月以來佳



資料來源：Bloomberg · 永豐投顧整理

AI相關基礎建設類股過去一年表現強勢



資料來源：Bloomberg · 永豐投顧整理

投資聲明:本研究報告由永豐商業銀行提供，資料來源為永豐證券投資顧問股份有限公司，所載資料僅供參考，無法保證內容完整性或預測正確性，亦不作任何保證，資料內容若有變更，將不另行通知，客戶投資前應慎審考量本身風險承受度，並應自行承擔投資風險。投資均有風險，任何投資商品過去績效亦不代表未來績效之保證，且有可能受市場波動導致本金損失；若申購以外幣計價的投資商品應瞭解匯率變動風險亦有可能導致本金之損失。未經永豐商業銀行及永豐證券投資顧問股份有限公司同意，不得以任何形式沿用、複製或轉載本文件之內容。

2026年01月21日

主題趨勢報

外資近期上調台積電資本支出展望，帶動了半導體設備廠商漲勢

根據高盛最新報告指出，台積電預計於 2026 至 2028 年間投入超過 1,500 億美元資本支出，其中 2027 年資本支出將顯著跳升至 540 億美元，此一數據修正顯示 AI 驅動的先進製程需求已具備長期能見度。資本支出的加速將導致半導體設備、先進封裝 (CoWoS) 及廠務工程產業出現結構性供需變化。此一評級調升帶動了半導體設備相關廠商在今年初的漲勢，可從兩點切入討論：

一、資本支出趨勢分析

高盛預期，台積電 2026 年資本支出預估上修至 460 億美元，2027 年進一步擴大至 540 億美元。此一資本投入規模確立了產業週期的延續性。

1. 設備交期與訂單積壓：資本支出的高強度延續意味著半導體設備商的接單能見度 (Visibility) 已從標準的 6-12 個月延伸至 24-36 個月。這消除了市場對於 2025 年設備出貨觸頂後的衰退疑慮。
2. 製程升級的資本密集度：隨著 2nm 預計於 2026 年貢獻 7.5% 營收 (超越 3nm 初期表現)，先進製程設備 (如 EUV) 的營收佔比將持續提升，設備業者的產品組合優化，有助於維持高毛利水準。

二、CoWoS 產能擴張

報告預估 CoWoS 產能於 2026 年與 2027 年將分別成長 89% 與 81%。分析其驅動因素，除了終端需求 (AI GPU/ASIC) 增加外，技術規格的變遷構成另一項關鍵變數。

- 晶片尺寸與產出率的負相關：隨著晶片尺寸變大導致產出下降，AI 晶片逼近光罩極限，單片晶圓可切割的晶片數量減少，且封裝複雜度提升。
- 設備需求脫鉤：為了維持同等的終端晶片產出量，所需的封裝與檢測設備數量必須呈現超額增長。這意味著設備需求的成長率將高於終端晶片的出貨成長率。
- 檢測剛需化：隨著晶片堆疊層數增加，良率成本極高，製程中的檢測頻率將大幅提升，帶動自動光學檢測與 X-ray 檢測設備進入高成長期。

2026年01月21日

主題趨勢報

人工智慧正在從「訓練」典範轉移至「推理」

本次 CES 2026 所釋放出的核心訊號非常清楚：人工智慧 (AI) 正從過去「大量訓練、快速生成內容」的階段，正式邁向以「推理」與「自主行動」為核心的新時代。

在這樣的趨勢下，NVIDIA 透過 Vera Rubin 平台的推出，為 AI 運算開啟了全新的成長路徑。即使在「摩爾定律放緩」的情況下，NVIDIA 仍透過整體系統設計，持續拉開與競爭對手的差距，進一步強化其長期競爭優勢，且隨著 AI 從訓練到推論的過度，記憶體將成此輪最為受惠的關鍵產業。

A. Vera Rubin 平台：全面投產 (Full Production)

NVIDIA 再次明確傳達一個方向：AI 的競爭，已不再只是單一晶片效能的比拼，而是整套運算系統的全面整合能力。這代表 NVIDIA 不只是提供「更快的 GPU」，而是打造一個從運算、記憶體到網路都高度整合的 AI 基礎架構，讓客戶能以更低成本，運行更複雜的 AI 應用。

Vera Rubin 平台預計將於 2026 年下半年 (2H26) 正式量產出貨。需要特別強調的是，Rubin 並不是一顆單獨的晶片，而是一整套 AI 運算平台，設計目標並非追求帳面規格，而是實際運作效率。平台主要由多個關鍵元件組成，包括：

- Vera CPU：新一代 CPU，大幅提升資料傳輸效率。
- Rubin GPU：專為 AI 推理優化，相較 Blackwell 的推理效能提升 5 倍，訓練效能提升 3.5 倍。
- NVLink 6 Switch：高速互連與網路系統，確保大量資料能快速流動、不塞車。
- 關鍵數據：Rubin 雖然電晶體數量僅增加約 1.6 倍，但透過整體系統的最佳化設計，實現了 Token 生成成本每年降低 10 倍、生成量增加 5 倍的經濟效益，這代表 AI 的「使用成本曲線」正在被重新拉低，也意味著 AI 應用將更快擴散到企業、政府與各種商業場景中。

B. 摩爾定律失效後的「TCO 暴力美學」：單位運算重新定義

- 邏輯核心：傳統摩爾定律每年僅能提供 15-25% 的電晶體密度成長，但 AI 模型規模與 Token 需求是指數級暴增，若依照舊路徑，資料中心將無利可圖。
- NVIDIA 的解法：透過 Vera Rubin 平台進行「極端協同設計 (Extreme Co-design)」。
 - NVIDIA 不再單賣 GPU，而是販售 "The Rack" (機架) 甚至 "The Data Center" 作為最小運算單位。
 - 客戶評估的指標從「單晶片性價比」轉向「整座資料中心的總擁有成本 (TCO)」。Rubin 透過整合 CPU、GPU、Switch 與光通訊，實現了「成本降低 10 倍」的經濟效益。

2026年01月21日

主題趨勢報

C. 從「生成 (Generating)」到「推論 (Reasoning)」：推論市場的價值重估

- CES 確立了 Test-time Scaling (測試時擴展) 的主流化。AI 不再只是像 Chat GPT 一樣「檢索並生成」答案，而是像人類一樣「花時間思考、規劃、自我反思」後才回答。
- 推論算力大於訓練算力：這類「會思考的 AI」導致推論過程的運算量呈現數量級跳躍。推論 (Inference) 不再是輕量級負載，而是高強度的持續運算。

D. 架構革命：推理上下文記憶體儲存平台

這是本次 CES 2026 在技術層面最關鍵、但也最容易被忽略的一項突破。NVIDIA 重新定義 AI 的「記憶體使用方式」，讓 AI 不再受限單一晶片的記憶體容量；過去 AI 在推理過程中使用的「上下文記憶 (KV Cache)」，多被視為臨時資料，用完即丟、無法共享，但隨著 AI 開始處理：

- 更長的對話內容
- 更複雜的多步推理任務

這種做法已明顯成為瓶頸，為了解決這個問題，NVIDIA 透過 BlueField-4 DPU 搭配 企業級 SSD，建立一個可共享的上下文記憶體儲存池。其結果是每顆 GPU 能夠存取的推理記憶體容量，從原本約 1TB，大幅提升至約 16TB，有效打破過去「單一 GPU 記憶體上限」的限制，讓 AI 推理可以像雲端服務一樣橫向擴展，進而支援更複雜、能長時間運作的 AI 代理 (Agent) 應用。

E. 記憶體產業趨勢：從 HBM 到 SSD 的全面受惠

隨 AI 運算方式改變，記憶體的重要性明顯提升，且不只高階產品受惠，整條產業鏈都有機會。

- HBM (高頻寬記憶體) :HBM 仍是 AI 核心運算不可或缺的關鍵零組件，需求持續強勁。由於供給有限，市場預期短缺情況將延續至 2026 年，價格維持高檔。
- SSD / NAND Flash: SSD 是本次最明確的受惠者。新架構使 AI 大量使用 SSD 來存放「思考與記憶內容」，企業級 SSD 從單純儲存設備，升級為影響 AI 效率的重要元件，需求有望成長。
- DDR 記憶體: DDR 則負責系統協調與管理。隨著 AI 系統規模擴大，DDR 使用量也將增加。