

AI 晶片的運算能力競賽與晶片供應鏈分流

TAcc+ 新創分析團隊

當大數據與 AI 結合，想要高效執行演算法與處理呈指數成長的海量資料，必須倚賴強大的運算能力，針對 AI 演算所設計的晶片應運而生，且發展迅速。根據 MarketsandMarkets 報告，AI 晶片市場規模於 2023 年達到 512 億美元，預計至 2028 年可達 1,318 億美元，年複合成長率為 20.8%。

AI 晶片的運算能力競賽

ChatGPT 爆紅帶動 GenAI 應用快速發展，從文字、圖像、聲音到影像生成，提升了客服、美術創作、音樂創作、廣告製作...等各產業的生產力。

各式 AI 助理的發展潛力，讓眾多業者嗅到了殺手級應用的商機爆發可能，以 Microsoft 的 Azure、Amazon 的 AWS、Alphabet 的 Google Cloud 三大龍頭為代表的雲服務業者，紛紛加快與加大投資力度，也帶起一波 AI 伺服器需求大增的狂潮。

AI 晶片的運算能力軍備競賽正如火如荼進行中，幾乎每半年就有運算能力倍數成長的新晶片推出，目前在專業 AI 晶片市場，NVIDIA 一枝獨秀，市佔率高達 95%以上。

NVIDIA 創辦人兼執行長黃仁勳 2024 年 3 月 19 日於 2024 NVIDIA GTC 大會中分享，訓練語言模型時所需要的每秒浮點運算速度 (FLOPS, Floating-point operations per second)，自 2017 年 Google 推出 Transformer 模型後，頂尖企業推出的模型對於運算速度的成長要求相較過往更加快速，至 2023 年底推出的 GPT-MoE 模型，需要的每秒浮點運算速度已來到 1.8 兆次。

黃仁勳也舉例：2021 年使用 NVIDIA GPU 的超級電腦 SELENE，規格為 4,480 個 A100 型 GPU，運算效能為每秒 3 百萬兆浮點運算 (exaFLOPS)，以及 112TB/s 連線頻寬。2023 年的超級電腦 Eos，規格為 10,752 個 H100 型 GPU，運算效能為每秒 43 百萬兆浮點運算，以及 1,100TB/s 連線頻寬，僅 2 年，整體運算速度就提升超過 10 倍！

NVIDIA 推出新一代 GPU 架構 Blackwell，瞄準劃時代的 GenAI 應用需求，希望 Blackwell 晶片成為推動這場新工業革命的引擎。Blackwell 使用台積電 4 奈米製程生產，配備 2,080 億個電晶體。NVIDIA 將 2 顆 Blackwell 架構的

GPU 加上 1 顆 ARM 架構的 Grace CPU，組成超級 AI 晶片「GB200」，在 AI 的訓練速度上比前一代 H100 GPU 晶片快 4 倍，推論能力提升 30 倍以上，成本和能源消耗則能降低 25 倍。

老對手 AMD 是主要競爭者。NVIDIA 在 2022 年 3 月推出 H100 GPU 晶片後，AMD 於 2023 年 6 月發表適用於 AI 運算的 GPU 晶片 MI300X，以及由 GPU 和 CPU 組成的「APU」MI300A。MI300X 擁有 1,530 億個電晶體、192GB 的記憶體容量、每秒 5.3TB 的記憶體頻寬，AMD 宣稱相較 H100，電晶體數量約 2 倍，記憶體容量 2.4 倍，記憶體頻寬 1.6 倍。

幾個月後，NVIDIA 在 2024 年 3 月推出 Blackwell GPU 和 GB200，效能大幅超越 AMD 產品，且對外宣布未來推出新款晶片的速度，將由每二年一次，縮短為每年一次。

黃仁勳於 2024 年 COMPUTEX 演講中宣告，2025 年將推出 Blackwell Ultra，2026 年推出下一代 GPU 平台 Rubin，2027 年續推 Rubin Ultra，可看出每 2 年推出新 GPU 平台，隔年推出升級版的進化模式。AMD 執行長蘇姿丰也於 2024 年第二季財報公佈之際，對外表示 AMD 將每年推出新款 AI 晶片，預計 2024 年第四季推出 MI325X，2025 年推出 MI350，2026 年推出 MI400。

上述雙方除了在 AI 晶片的硬體功能表現上互別苗頭，更在整體運算效能和成本效益上做競爭。業內專家 Doug O' Laughlin 指出，現階段 NVIDIA 能夠主宰 AI 晶片市場，是因為同步提供顧客最完整的解決方案，包括最好的晶片、最佳的網路設備以及最優秀的軟體 CUDA。針對此，AMD 也開始將旗下軟體 ROCm 設定為開源，並提供轉換工具以利顧客將 CUDA 轉換為 ROCm。AMD 執行長蘇姿丰受訪時表示，GenAI 所需要的高運算能力，須由高效能晶片、大容量記憶體與軟體來支持。

二家 GPU 晶片大廠搶佔 AI 晶片市場之際，CPU 晶片龍頭 Intel 不甘示弱，於 2023 年 12 月發表開發的 AI 晶片 Gaudi3。

Intel 於 2019 年收購 Habana Labs 後，隨即投入 Gaudi 晶片的開發。執行長 Pat Gelsinger 在發佈會上表示 AI 個人電腦是 AI 伺服器需求暴增後的下一波主角，特別介紹專為 Windows 筆電和個人電腦設計的 Core Ultra 晶片，以及第五代 Xeon 伺服器晶片。

這二款晶片都內含提升圖形處理能力、可加速執行 AI 程式的專用 AI 零件

「神經處理單元 (NPU)」。Core Ultra 能強化每台電腦離線時以單機處理較小規模 AI 運算任務的能力；Xeon 常與 GPU 搭配，用於訓練和部署 GenAI 系統。

為了提升端點裝置的 AI 算力，Intel 透過 2016 年收購新創企業 Movidius，取得其產品「視覺運算處理器 (VPU)」，能以低功耗執行 AI 運算。初期推出類似於隨身碟大小的裝置，以 USB 接頭外接於各類端點裝置，提升該裝置的 AI 運算能力。

之後 Intel 也於 Meteor Lake 系列處理器中導入 VPU 設計，提升 AI 運算效能，同時降低 CPU 的工作負載，減少功耗。由此看出，Intel 正同時佈局資料中心導入 AI 雲服務，以及離線端點裝置導入 AI 功能的算力需求。

其他科技巨頭 Alphabet、IBM、Meta、Microsoft、Amazon 等則積極搶進 AI 晶片設計領域，主要目的是另闢取得 AI 晶片的管道，開發出更符合自己需求、更具成本效益的解決方案。

例如，Alphabet 於 2023 年 12 月推出其最先進的 AI 模型「Gemini」，能夠處理視訊、音訊、文字等不同形式的資訊；此模型使用的新技術加入到旗下 AI 助理產品「Gemini」的 2024 年新版本中。

Alphabet 也發表新一代自行開發的客製化 AI 晶片「TPU v5p」，比上一代 AI 晶片的訓練速度加快近 2 倍，更能配合新版 AI 模型之運算需求。

IBM 於 2023 年 10 月發表模擬人類大腦運作的 AI 推論晶片「NorthPole」，最大特點在於所有的記憶體都在晶片上，不需另外連結記憶體，擺脫因資料傳輸速度跟不上資料處理速度所造成的瓶頸 (Von Neumann Bottleneck)，適用於需要即時處理大量資料的邊緣運算應用場景，如自駕車導航裝置、機器人操作、路況監控、衛星監控、資安偵測等。

除了科技巨頭，也有不少新創企業投入開發 AI 晶片，茲列舉如下：

● SambaNova Systems

看準傳統 CPU 和 GPU 架構之運算能力提升的成本效益逐漸下降，在高階 GPU 價格持續攀升且供貨不及下，為資料中心業者另闢提升算力的解方。

SambaNova 聚焦研發資料分析和 AI 的軟硬體設備、電腦處理器，並使用 AI 演算法提升客製化晶片的設計能力，追求讓晶片在運算能力、功耗、尺寸控

制等各面向效能超越 GPU。SambaNova 已獲得 Alphabet、軟銀、Intel 的投資，其產品也於 2024 年 1 月獲頂尖半導體製造商 Analog Devices 採用。

●Cerebras Systems

與 NVIDIA、AMD、Intel 等公司將晶圓切割為小晶片的方式不同，Cerebras 運用整片晶圓製成巨型晶片，建立一種專為 AI 運算最佳化的新型電腦系統，相較現今將大量的電力和成本花費在重新連結晶片上的主流設計，可以降低互聯和網路成本與功耗。

Cerebras 已獲得 Open AI 執行長 Sam Altman 與 Benchmark 等創投的投資，2023 年 7 月 20 日對外宣佈與阿拉伯聯合大公國科技控股集團聯手打造 AI 超級電腦。2024 年 3 月 13 日也推出新款 WSE-3 晶圓級 AI 晶片，並對標 H100 晶片，宣稱效能相當於 62 顆 H100。據聞，Cerebras Systems 已提出首次公開發行 (IPO) 申請，規劃 2024 年下半年上市。

●Graphcore

致力於開發專為訓練 AI 而設計的晶片產品「IPU」，適用於小規模的 AI 模型運算。其創辦人認為，雖然 GPU 比 CPU 更適用於 AI 運算，仍不是專為 AI 所設計。Graphcore 已獲得微軟和紅杉資本等投資，但受到新設計難打入既有供應鏈，以及 ChatGPT 帶起大型語言模型 (LLM) 之運算需求。

GPU 相較 IPU 更適合 LLM，加上美中晶片戰爭也限制了公司在中國市場的業務發展，導致產品長時間未獲市場青睞。2024 年公司開始尋求被外部收購之機會，並於 2024 年 7 月被軟銀集團收購。

●Groq

聚焦設計專門用於 AI 推論的晶片，且可使用客製化的編譯器 (compiler，一種電腦程式，可將某種程式語言寫成的原始碼轉換成另一種程式語言) 支援各種語言模型。

Groq 開發的晶片可在低耗電下，加速 AI、機器學習和高效能運算領域的複雜工作負載，相較其他競品擁有更高的成本效益。Groq 近年陸續收購 AI 解方公司 Maxeler Technologies、Definitive Intelligence，並於 2024 年成立新部門 Groq Systems，積極拓展客戶與開發者社群。

●Mythic

以數位類比轉換器 (DAC)、類比數位轉換器 (ADC) 和支援記憶體內運算的嵌入式快閃記憶體，來實現演算法加速。期望透過成熟製程，開發專為 AI 推論設計的加速器晶片，帶來比使用數位架構的競品更佳的性價比、低功耗優勢。

近年因 NVIDIA 在 AI 晶片市場的主導地位強勢，使創投業者對於 AI 晶片新創的投資風險意識上升，Mythic 執行長 Dave Rick 表示籌資曾一度遇到困難，幸而在 2024 年 8 月順利獲得 Cisco、Samsung、BlackRock 領投的 D 輪 6 億 4 千萬美元投資。

地緣政治引發晶片供應鏈分流

臺灣自 1960 年代推動輕工業出口，1970 年代發展基礎工業與重工業，1980 年代推動經濟自由化與發展技術密集產業，1990 年代推動亞太營運中心與發展資訊產業。1991 年蘇聯解體宣告冷戰結束，全球政經局勢由政治集團間的對抗衝突走向開放和解，國際間的貿易量快速成長，世界各地產業也隨之串聯與整併，漸漸形成以全球為整體的產業鏈和分工體系，各國業者依據自身發展優勢，透過全球運籌整合資源，完成商品製造與販售。

臺灣在上述主、客觀條件成熟之下，逐步發展為全球高科技產業供應鏈的重要聚落。

依據研究機構 TrendForce 調研結果，截至 2024 年第一季，全球前十大晶圓代工業者之營收狀況如下表。臺灣企業共有四家進入前十大，合計市佔率達 69.6%，台積電是領導企業，市佔率達 62.3%，顯示臺灣在全球半導體產業專業分工體系中的晶片製造項目，佔有舉足輕重的地位。

表 1、全球前十大晶圓代工業者營收與市佔率

排名	公司名稱	營收（單位：百萬美元）			市佔率	
		2024Q2	2024Q1	季增率	2024Q2	2024Q1
1	台積電	20,819	18,847	10.5%	62.3%	61.7%
2	三星	3,833	3,357	14.2%	11.5%	11.0%
3	中芯國際	1,901	1,750	8.6%	5.7%	5.7%

4	聯電	1,756	1,737	1.1%	5.3%	5.7%
5	格羅方德	1,632	1,549	5.4%	4.9%	5.1%
6	華虹集團	708	673	5.1%	2.1%	2.2%
7	高塔半導體	351	327	7.3%	1.1%	1.1%
8	世界先進	342	306	11.6%	1.0%	1.0%
9	力積電	320	316	1.2%	1.0%	1.0%
10	合肥晶合	300	310	-3.2%	0.9%	1.0%
前十大合計		31,962	29,172	9.60%	96%	96%

資料來源：TrendForce

2018 年起，美中貿易糾紛導致雙方展開貿易戰，全球政經局勢由開放和解走向政治集團間的對抗衝突。2020 年初 COVID-19 肆虐全球，疫情嚴重與各國管制邊境導致全球供應鏈大亂。

緊接著，2022 年 2 月 21 日俄國與烏克蘭多年來的零星衝突突然轉為全面開戰，衝擊全球能源與糧食供應。2023 年巴拿馬運河遭逢 70 年來最嚴重乾旱，重創全球船運運能；10 月 7 日爆發以哈戰爭導致紅海航線中斷，迫使貨船改走好望角或麥哲倫海峽，令貨運成本大增。

東亞地區的南海、台海、東海、日本海周邊政治局勢也趨於緊張。在貿易戰、疫情、戰爭與天災導致供應鏈中斷的風險不斷上升之下，許多國家為了因應變局，紛紛研擬政策強化本國供應鏈，尤其是美中競爭所聚焦的高科技產業，更為重中之重。

2023 年 10 月 14 日，台積電創辦人張忠謀出席公司運動會時，回憶他在 2019 年的運動會曾示警：隨著美科技戰的延燒，台積電將因地緣政治成為各方爭取合作的對象。四年後再回顧，不但印證其判斷，更直言過往的半導體產業全球化將逐漸消失，自由貿易倒退。各先進國家將轉以國家安全為優先考量，台積電成為各個國家競相爭取開設於「本土」的企業。

從競爭的角度來看，其他競爭者會積極利用本身的地緣政治優勢，試圖加速趕上或超越台積電，預期未來幾年，台積電將面對更為嚴峻的挑戰。現在累積的優勢未必能維持 20-30 年，但相信公司經營團隊能嚴陣以待、謹慎應對，繼續保持領先。

表 1 台積電於海外投資設立晶圓廠的時程規劃

地點	投產時程	設廠規劃	投資狀況
日本	2024 年	28/22/16/12 奈米製程	總支出約 3 兆日元。由台積電與 Sony、豐田等業者合資成立的 JASM 進行設廠。日本補助 1.2 兆。
日本	2027 年	40/7/6 奈米製程	
美國	2025 年	4 奈米製程	總支出超過 650 億美元。美國已直接補助 66 億，並可能提供貸款 50 億與 25% 的投資稅收抵免。
美國	2028 年	2 奈米製程	
美國	2029 年	2 奈米或更先進製程	
德國	2027 年	28/22/16/12 奈米製程	總支出超過 100 億歐元。由台積電與 BOSCH、Infineon、NXP 等業者合資成立的 ESMC 進行設廠。德國補貼 50 億，並斥資改善當地基礎建設。

資料來源：中央社

全球主要先進國家、科技品牌大廠正致力於晶片供應鏈在地化，並建立第二、第三供貨管道，藉此避開上述風險導致的斷鏈危機。

晶片供應鏈正由過往的全球化轉朝區域化分流，從表 2 整理的台積電海外設廠地點即可看出，東亞、北美、西歐會是第一波分流後的三大主要晶片產區，中國與印度等國也都在積極建立供應鏈，可預期未來至少會分流成 5 至 10 個區域供應鏈。

與晶片相關的上、下游業者，也在配合進行跨區、跨國設廠。例如，2023 年以來，受惠 AI 伺服器需求大增而股價飆漲的伺服器代工大廠廣達、緯創與美超微。

廣達在惠普等伺服器品牌商要求分散中國製造風險下，為滿足「中國+1」的分流規劃，選擇至泰國設廠，因而在 2024 年獲得惠普大訂單，未來也將持

續擴充產能，泰國將成為集團中僅次於臺灣的最大一條龍伺服器生產據點。

緯創董事長在 2024 年初的尾牙表示，公司正在墨西哥、越南建廠，後續也規劃於印度、美國加州設廠。

美超微透過與 NVIDIA 建立緊密的夥伴關係，業績大幅翻揚，股價自 2023 年初至 2024 年初大漲逾 6 倍。總裁暨執行長梁見後於受訪時表示，看到亞洲有很多大型電信及資料客戶的需求，公司正在持續擴大臺灣產能，將發展重心從美國移到臺灣。

參考資料：

- AI 晶片：前沿技術與創新未來。張臣雄。2021。
- AI 晶片為何如此重要 (上)。黃則普。2021。
- AI 黎明時刻 蘇姿丰領 AMD 正面對決。工商時報。2023。
- CPU vs. GPU：發揮兩者的最大效益。Intel。
- IBM 發表模擬人類大腦運作，號稱效能超越 4 奈米 GPU 的 AI 推論晶片 NorthPole。iThome。2023。
- Nvidia (NVDA) -全球最大圖像處理器公司。太股達人。2022。
- TrendForce：第二季供應鏈急單挹注晶圓代工利用率，全球前十大晶圓代工產值季增 9.6%、VIS 排名上升兩位。TrendForce。2024。
- 台積電評估德國設廠，外界預料首選薩克森矽谷。財經新報。2021。
- 直擊 | AMD 推 AI 新晶片，劍指輝達一哥地位！誰說沒客戶？微軟、Meta 都來了。數位時代。2023。
- 英特爾也強調 AI！Meteor Lake 系列處理器導入 VPU 設計。科技新報。2023。
- 英特爾發表新款 AI 晶片與 Nvidia 和 AMD 競爭。經濟日報。2023。
- 黃仁勳 2024 Nvidia GTC 完整演講，揭露在 AI 宇宙的最大野心。商周。2024。
- 黃仁勳旋風席捲全台「Jensanity」！獨家解析！黃仁勳 COMPUTEX 演講

在說什麼？想搭上 AI 趨勢就該看這部影片！曲建仲。2024。

- 業界不樂見一家獨大，兩虎相爭煙硝味濃；AMD 挑戰輝達，供應鏈下半年開新局。科技新報。2024。
- 輝達市值狂飆！AI 晶片戰，AMD 能打破 Nvidia 壟斷嗎？天下雜誌。2024。