新光投顧

〔投資研究報告〕

2026/3/5

AI ASIC 浪潮：GPU 壟斷的第一道裂縫

目錄

一、理解 AI 算力的核心單位：什麼是「算子」

二、GPU 與 ASIC 的根本差異：算子彈性

三、雲端企業為何投入 AI ASIC

四、真正的護城河：軟體與系統互連

五、結論：GPU 壟斷尚未瓦解，但算力市場開始分層

AI 算力市場正在從單一架構走向分工架構

2026/3/5 新光投顧

過去兩年人工智慧產業的核心敘事幾乎完全圍繞 GPU 展開。隨著生成式 AI 需求快速成長，大型語言模型訓練與推論所需的算力大幅增加，使 GPU 成為資料中心最重要的加速器。NVIDIA 憑藉 CUDA 軟體生態與高效能 GPU 架構，在 AI 算力市場建立了極高的市占率。

然而隨著 AI 資本支出持續擴張，大型雲端服務商逐漸開始發展自研 AI 晶片，客製化 ASIC 的出現讓市場開始討論 GPU 壟斷是否會鬆動。Google TPU、Amazon Trainium 以及 Meta 與 Microsoft 的 AI 晶片，都代表著雲端企業試圖提高算力自主性的策略。

從產業演進的角度觀察，這一趨勢更可能代表 AI 算力市場開始出現**架構分工**。GPU 仍將主導高彈性運算需求，而 ASIC 則逐步承接部分固定型工作負載。這種分工將使 AI 晶片市場由單一架構轉向多架構並存。

目錄

一、 理解 AI 算力的核心單位：什麼是「算子」
二、 GPU 與 ASIC 的根本差異：算子彈性
三、 雲端企業為何投入 AI ASIC
四、 真正的護城河：軟體與系統互連
五、 結論：GPU 壟斷尚未瓦解，但算力市場開始分層

一、 理解 AI 算力的核心單位：什麼是「算子」

要理解 GPU 與 ASIC 的競爭關係，必須先理解 AI 計算的基本單位——**算子（Operator）**。

在深度學習模型中，所有運算都是由一系列基本計算操作組成，這些基本操作即為算子。例如：

- 矩陣乘法（Matrix Multiplication）
- Softmax

- Layer Normalization
- 各種激活函數（ReLU、GELU 等）

大型模型其實就是大量算子依照特定順序組合而成的計算流程。以 Transformer 架構為例，注意力（Attention）計算與矩陣乘法等算子會在每一層模型中反覆出現，形成整個模型的主要算力負載。

從算力需求的角度，可以將 AI 計算簡化為：

$$Compute = \sum (Operator_i \times Frequency_i)$$

也就是說，AI 所需算力取決於不同算子的計算量與出現頻率。晶片設計的核心任務，其實就是如何更有效率地執行這些算子。
這也是 AI 硬體架構設計的基礎。

二、 GPU 與 ASIC 的根本差異：算子彈性

GPU 與 ASIC 之間最核心的差異，其實在於對算子的支援方式。

GPU 屬於**通用型算子處理架構**。其設計允許多種類型的算子在同一硬體平台上執行，因此能夠適應不同 AI 模型的需求。當模型架構發生變化時，GPU 仍能透過軟體更新與算子最佳化維持相容性。

相對而言，ASIC 通常針對少數核心算子進行硬體最佳化。例如某些 ASIC 會專門針對矩陣乘法或特定推論流程設計，藉此提高能效並降低成本。這種設計使 ASIC 在特定工作負載上效率更高，但也降低了其對新模型架構的適應能力。

AI 模型仍在快速演進的階段，算子結構並未完全固定。當模型架構改變時，通用架構的 GPU 往往具有更長的生命週期。從工程角度看，GPU 的通用性其實是一種對模型不確定性的保險。

因此，ASIC 滲透率的提升往往意味著某些 AI 應用的算子結構已經趨於穩定。

三、 雲端企業為何投入 AI ASIC

大型雲端服務商之所以積極發展 ASIC，核心原因在於 AI 運算成本。

當推論需求達到極大規模時，通用 GPU 的成本與功耗會成為主要負擔。如果能針對固定工作負載設計專用晶片，就能在能效與成本上取得顯著優勢。

因此，雲端企業傾向將以下工作負載轉向 ASIC：

- 大規模推論
- 固定型 AI 服務
- 高度可預測的算子流程

相反地，GPU 仍然主導：

- 模型訓練
- 新模型測試
- 高度變動的工作負載

這使 AI 算力市場逐漸形成分層結構，而非單純替代關係。

四、 真正的護城河：軟體與系統互連

GPU 平台的優勢不只來自硬體本身，更來自完整的軟體生態。
NVIDIA 建立的 CUDA 平台包含大量最佳化算子庫，例如 cuDNN 與 TensorRT。這些工具能將深度學習框架中的算子轉換為 GPU 最佳化運算，使開發者能快速部署與調整模型。

從 GPU 環境遷移到 ASIC 平台，不僅涉及硬體更換，也意味著：

- 軟體工具重寫
- 算子庫重新最佳化
- 工程團隊重新建立

這些成本往往遠高於晶片本身的差價。

另一方面，AI 系統規模不斷擴大，使晶片之間的高速互連成為新的性能瓶頸。高速互連技術（如 NVLink 與 InfiniBand）決定了數千顆加速器能否高效協作。未來 AI 算力競爭將越來越偏向系統層級，而不只是單顆晶片性能。

五、 結論：GPU 壟斷尚未瓦解，但算力市場開始分層

ASIC 的興起將對整個半導體供應鏈產生多重影響。

首先，先進製程需求仍將持續增加。無論 GPU 或 ASIC，大多依賴最先進製程節點，因此台積電仍將是 AI 晶片製造的核心。

其次，ASIC 與 GPU 都需要高階封裝技術。當 ASIC 需求增加時，先進封裝產能可能在不同晶片架構之間重新分配，進而影響整體 AI 晶片出貨節奏。

此外，不同晶片架構對記憶體需求也可能不同。GPU 通常需要大量 HBM，而部分 ASIC 設計可能採用其他記憶體架構。若 ASIC 占比提高，HBM 需求曲線也可能受到影響。

這些變化將逐步改變 AI 供應鏈的利潤結構。

AI 算力市場正從單一架構主導的階段，逐步轉向多架構並存的格局。

GPU 仍然是目前最重要的 AI 運算平台，其通用性與軟體生態在短期內難以被取代。然而隨著 AI 應用逐漸成熟，部分穩定型工作負載將轉向 ASIC 執行，使市場開始出現新的分工。

這種分工不必然削弱 GPU 需求，但確實代表 AI 算力市場正在出現結構變化。未來 AI 晶片競爭將不只是硬體性能之爭，而是算子結構、軟體生態與系統互連能力的綜合競賽。

GPU 的壟斷地位尚未結束，但其市場結構的第一道裂縫已經出現。