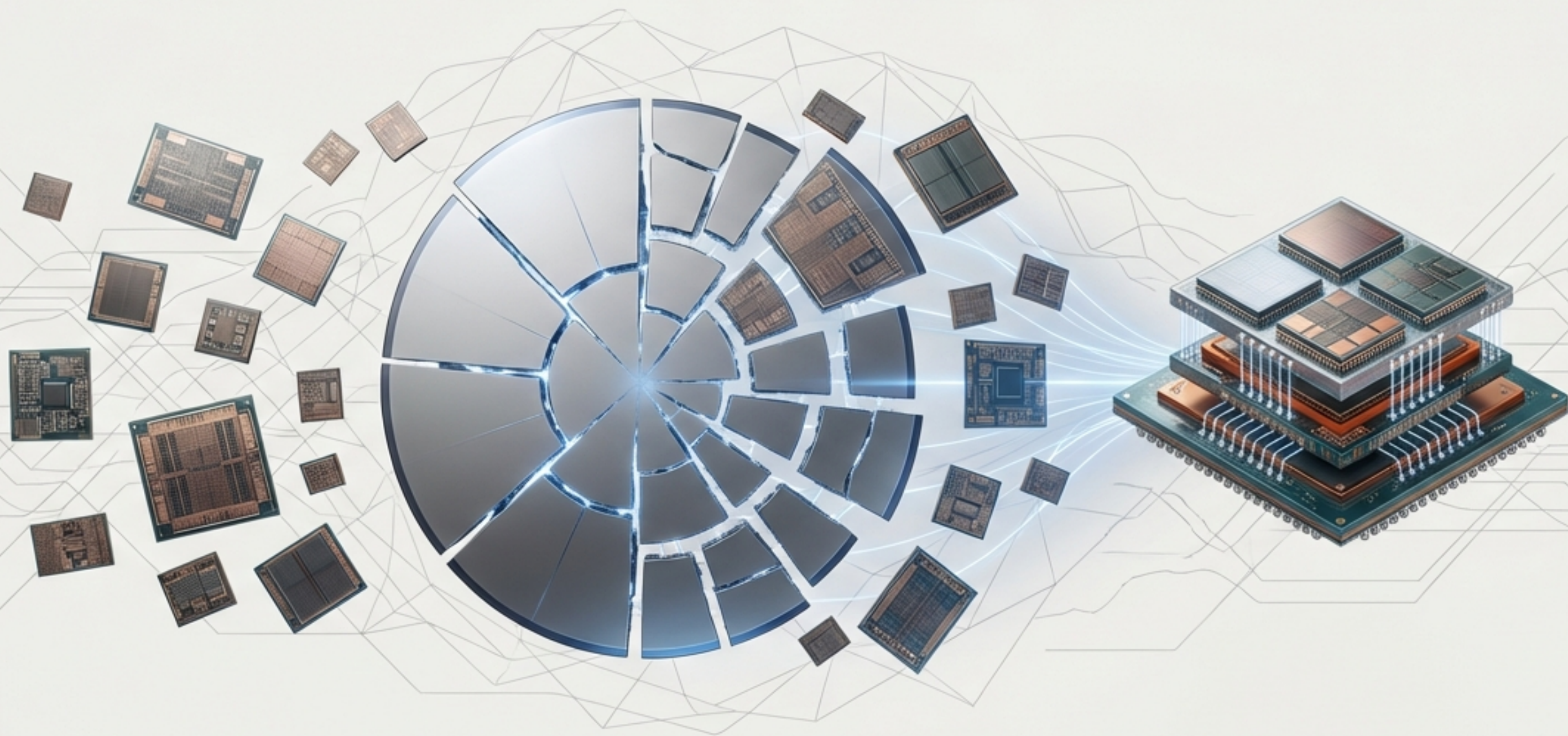


重塑運算：人工智慧如何解構與重建半導體世界

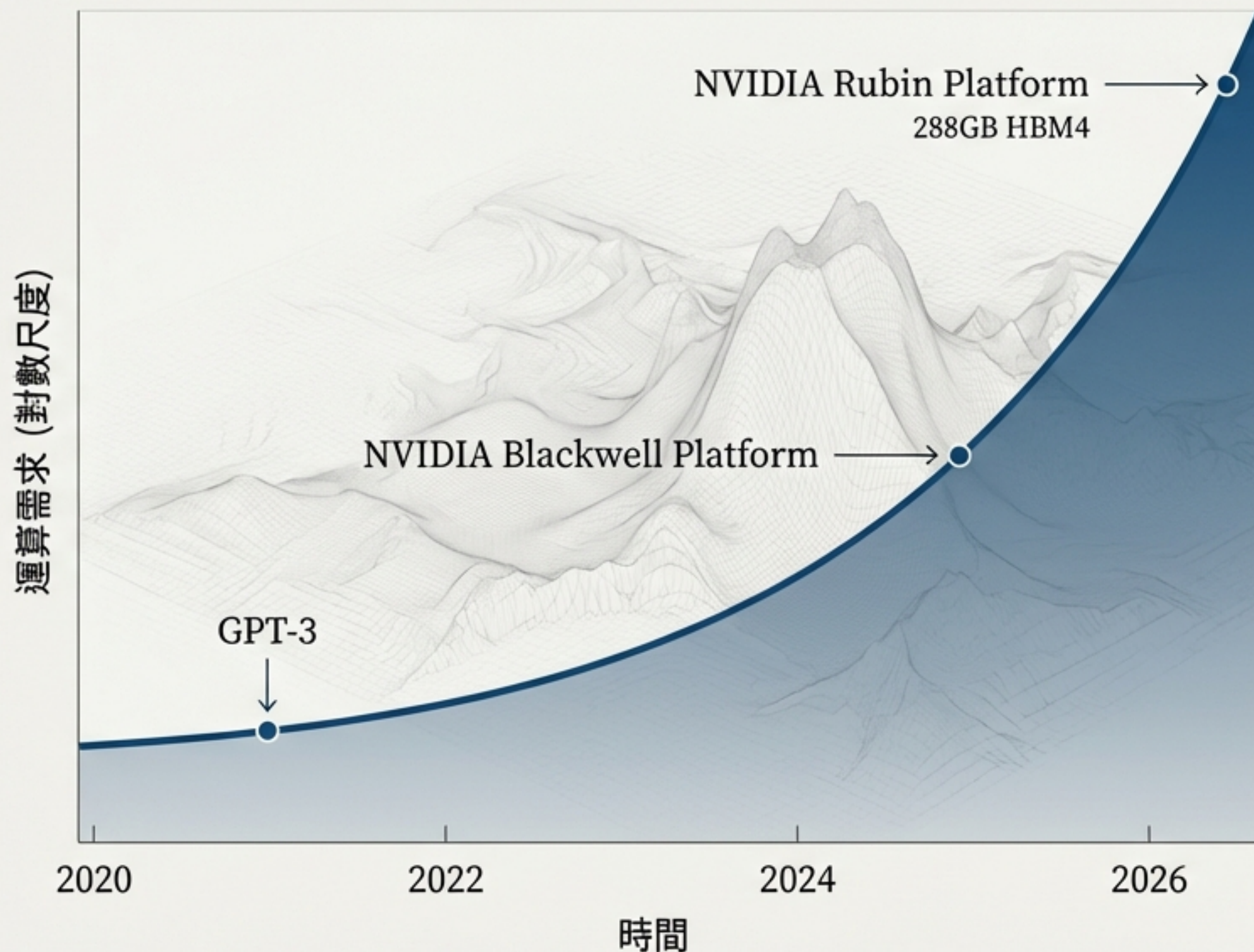
一場由資料中心到邊緣、由晶片到系統的全面革命



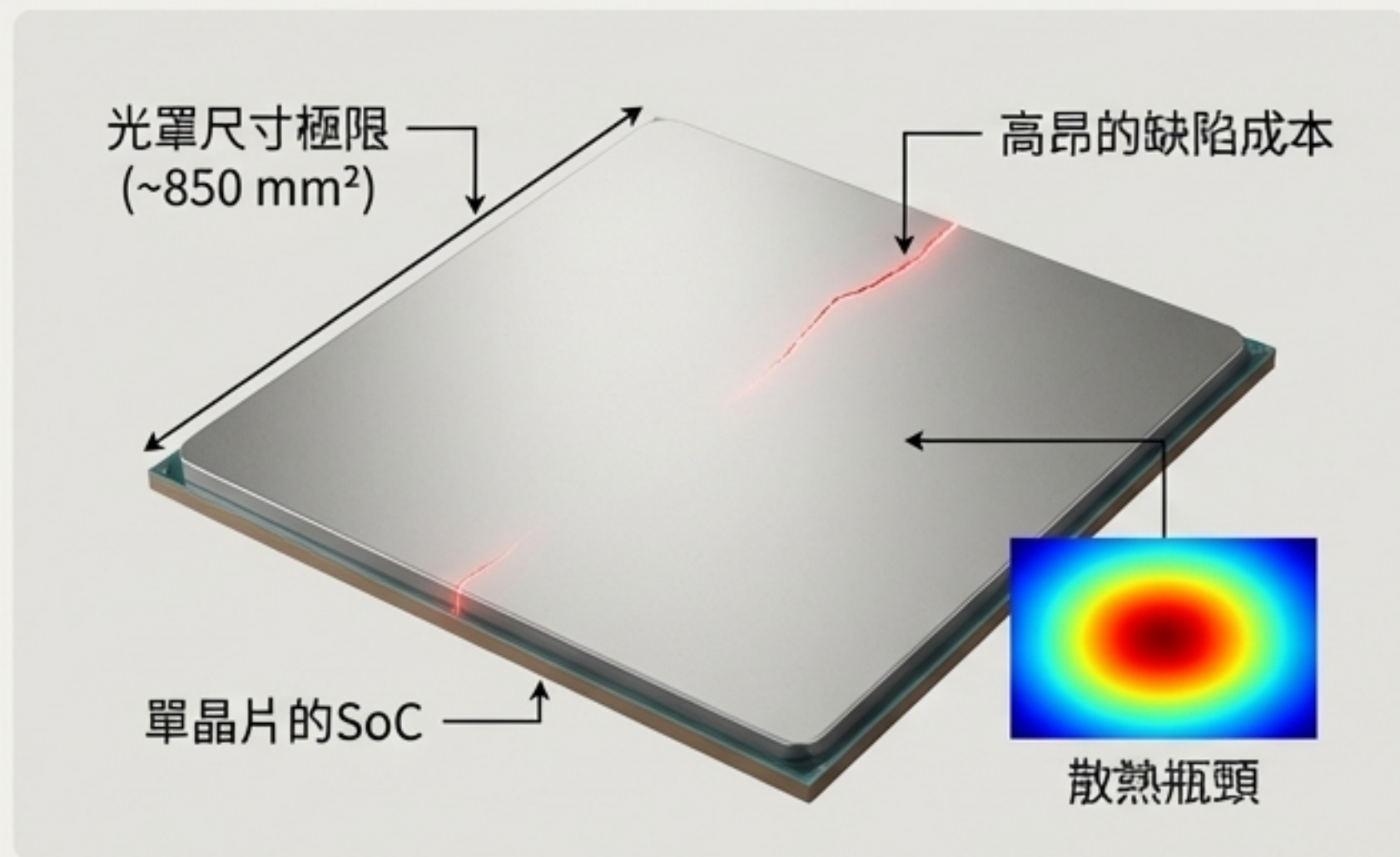
人工智慧的貪婪：基礎模型引爆的運算需求已達物理極限

大型 AI 模型，如語言翻譯和圖像生成，對計算能力、記憶體頻寬和儲存吞吐量的需求呈指數級增長。

- 隨著模型規模和複雜性增加，記憶體頻寬和容量已成為訓練和推論階段的關鍵效能瓶頸。
- **具體數據：** NVIDIA 的次世代平台對記憶體的需求呈倍數增長。例如，單顆 Rubin R100 超級晶片預計將配備高達 288GB 的 HBM4 記憶體。
- **結論：**傳統的運算架構已無法滿足需求，一場根本性的變革勢在必行。

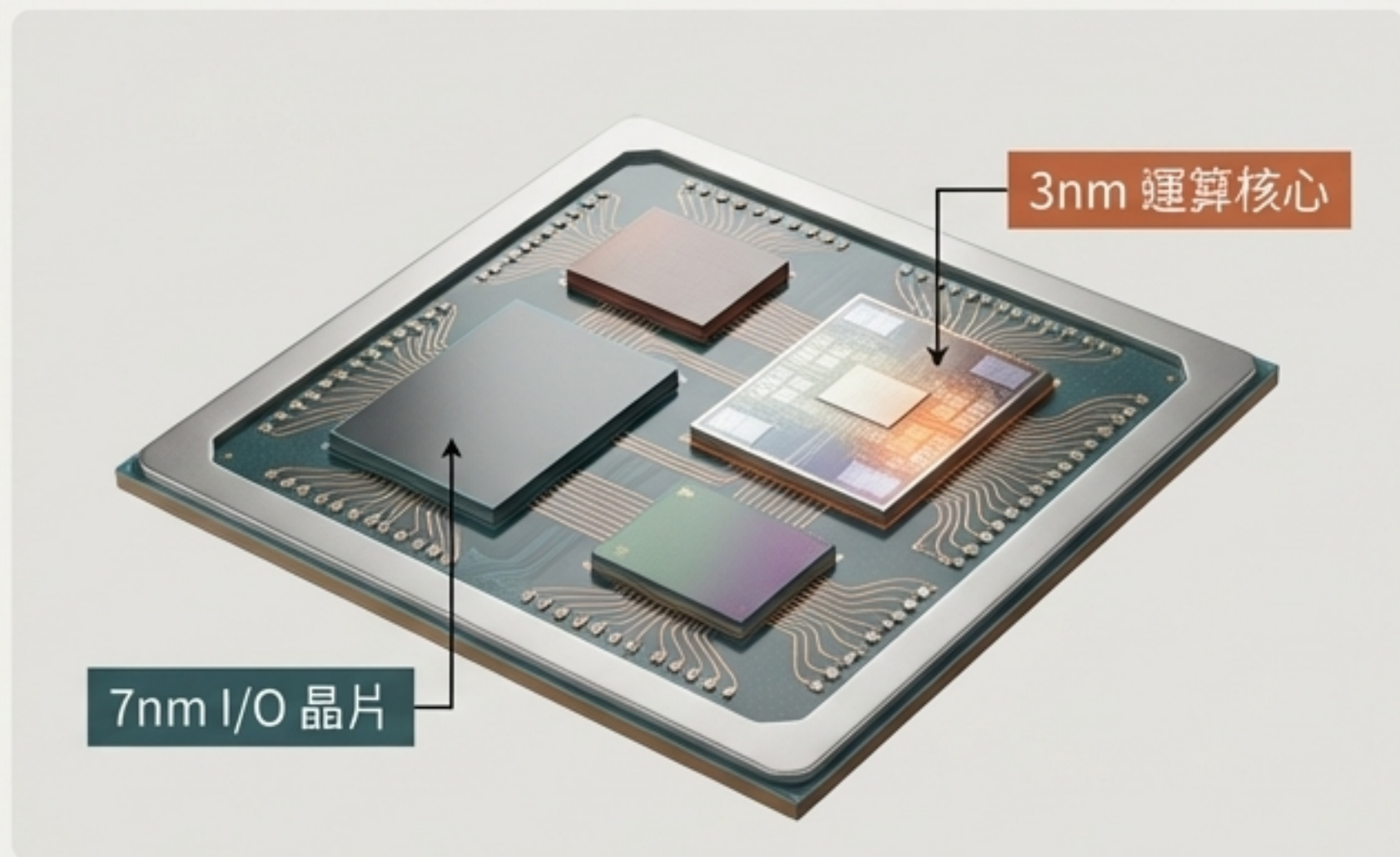


打破摩爾定律的物理枷鎖：Chiplet 為何是未來？



單晶片的挑戰

光罩尺寸極限、先進製程的低良率與高成本、以及散熱瓶頸限制了單晶片的發展。

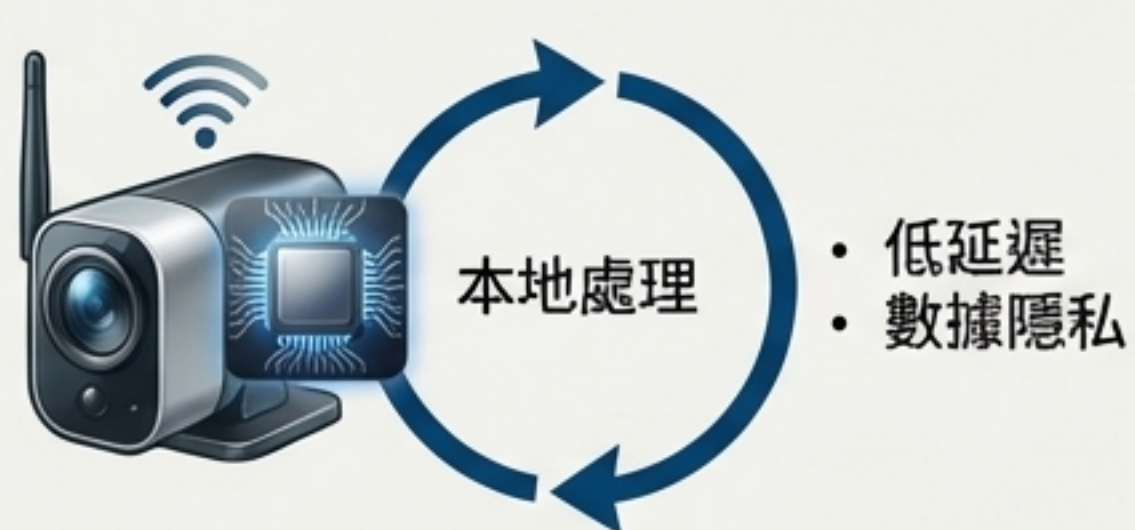
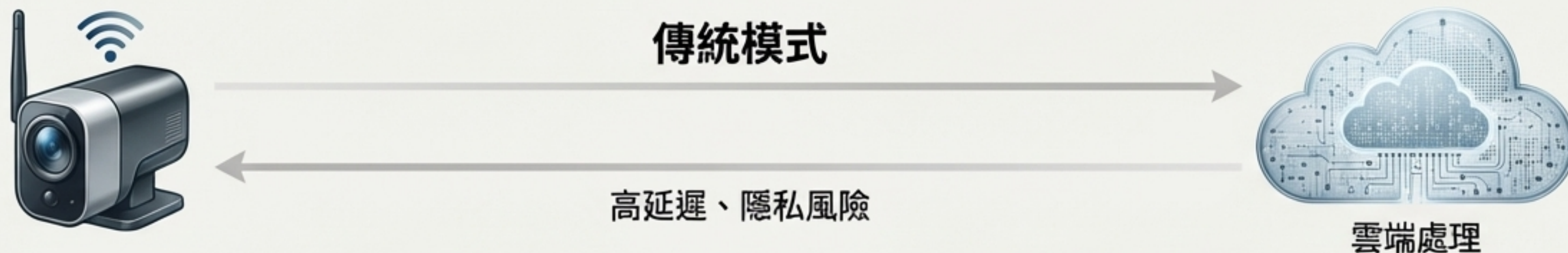


Chiplet 的解決方案

透過拆分設計，實現了提高良率、異質整合與模組化，突破物理限制。

業界實踐：AMD Zen CPU & MI300 APU, Intel Ponte Vecchio, NVIDIA Rubin GPU

智慧更靠近資料源：邊緣 AI 的興起與挑戰



具體應用案例



工業

透過持續監控設備，進行預測性維護，最大限度地減少停機時間。



醫療

在醫院中即時監控病患生命體徵，並在本地處理數據，保護病患隱私。

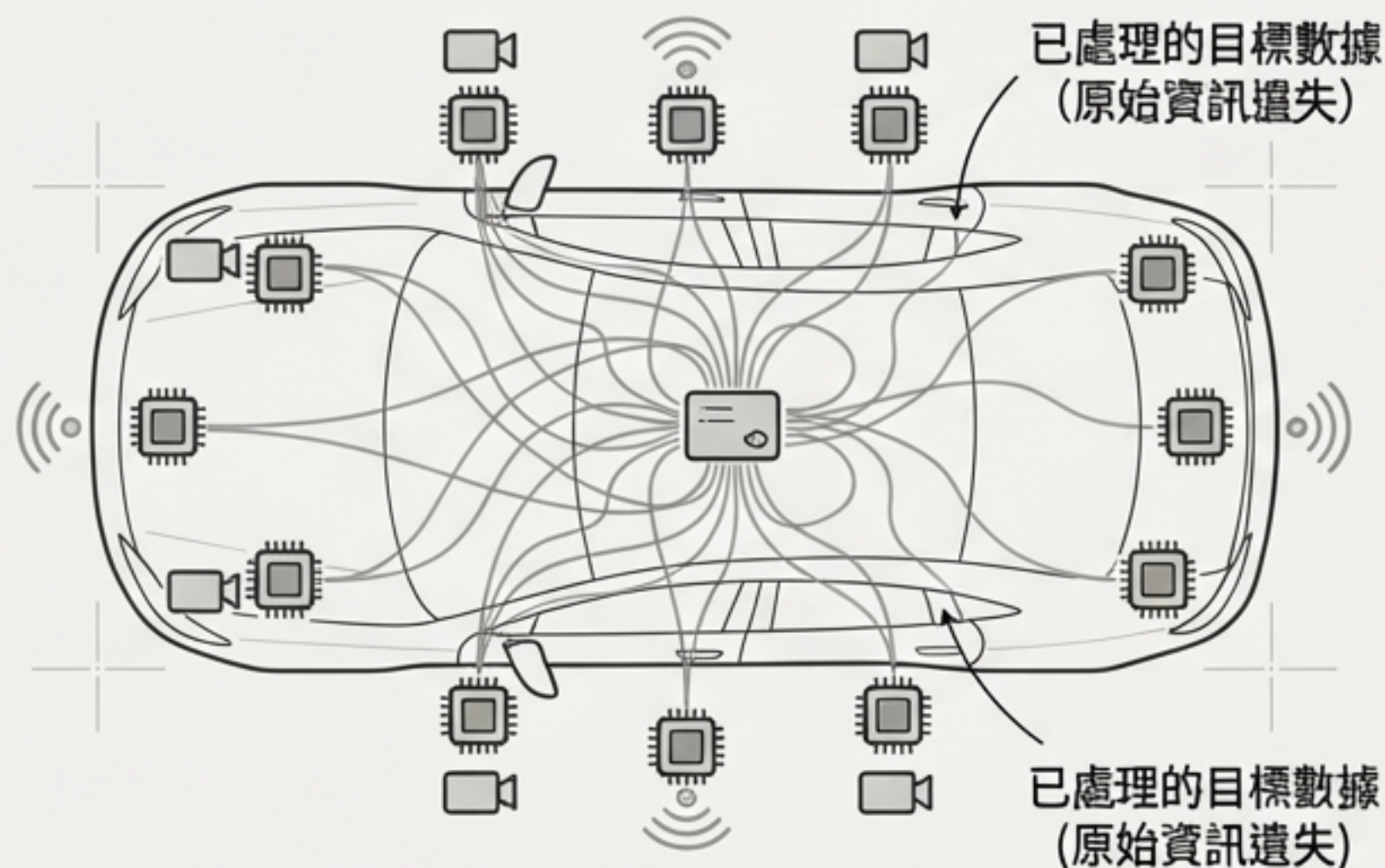


零售

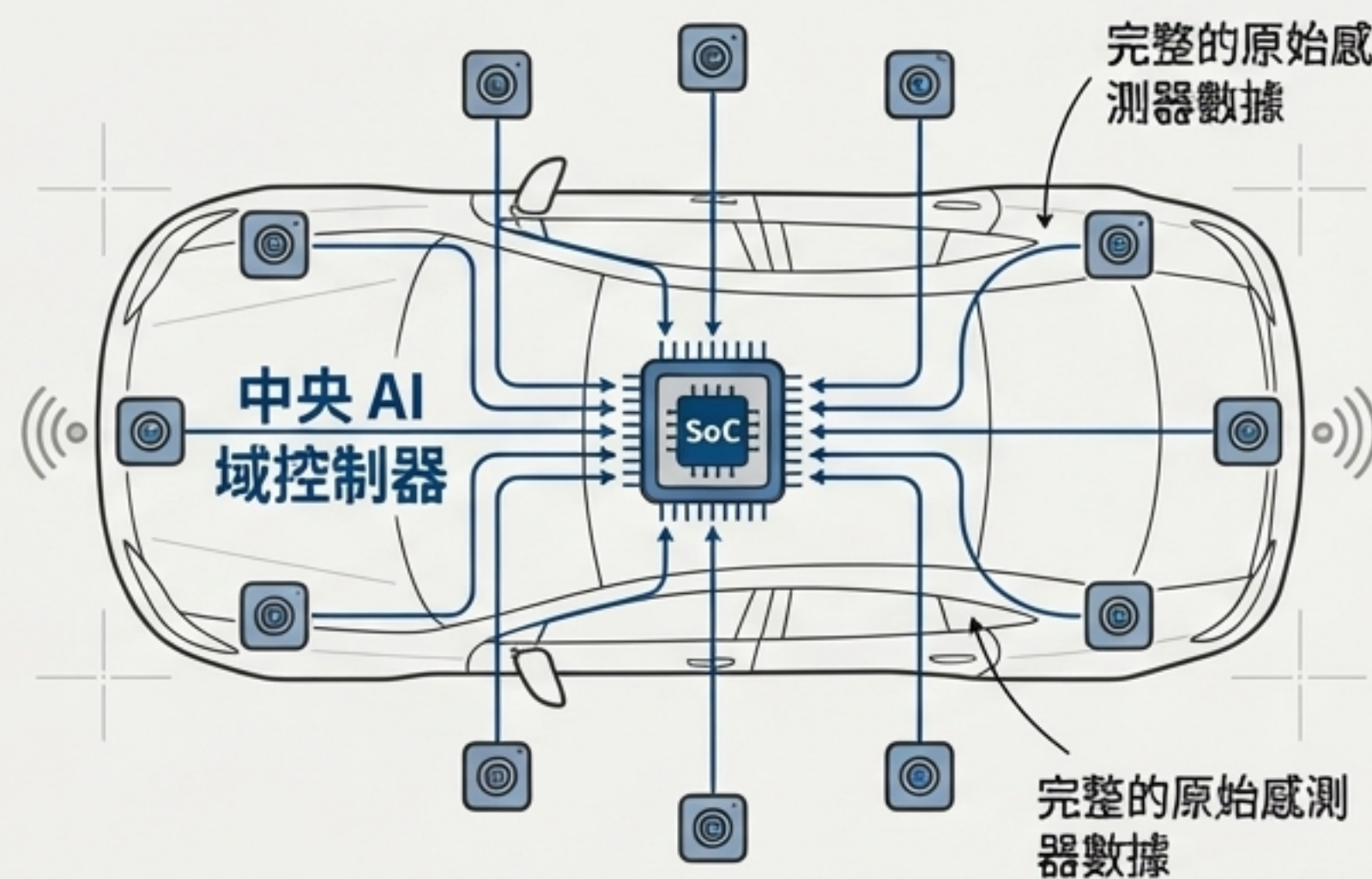
自動化視覺監控，追蹤庫存水平並分析購物者活動趨勢。

案例研究：汽車的「中央大腦」革命

過去分散式架構（後期融合）



未來的集中式架構（深度融合）



- **最佳的感知能力**

保留所有原始數據，可根據天氣和駕駛條件動態調整感測器權重。

- **降低成本與複雜性**

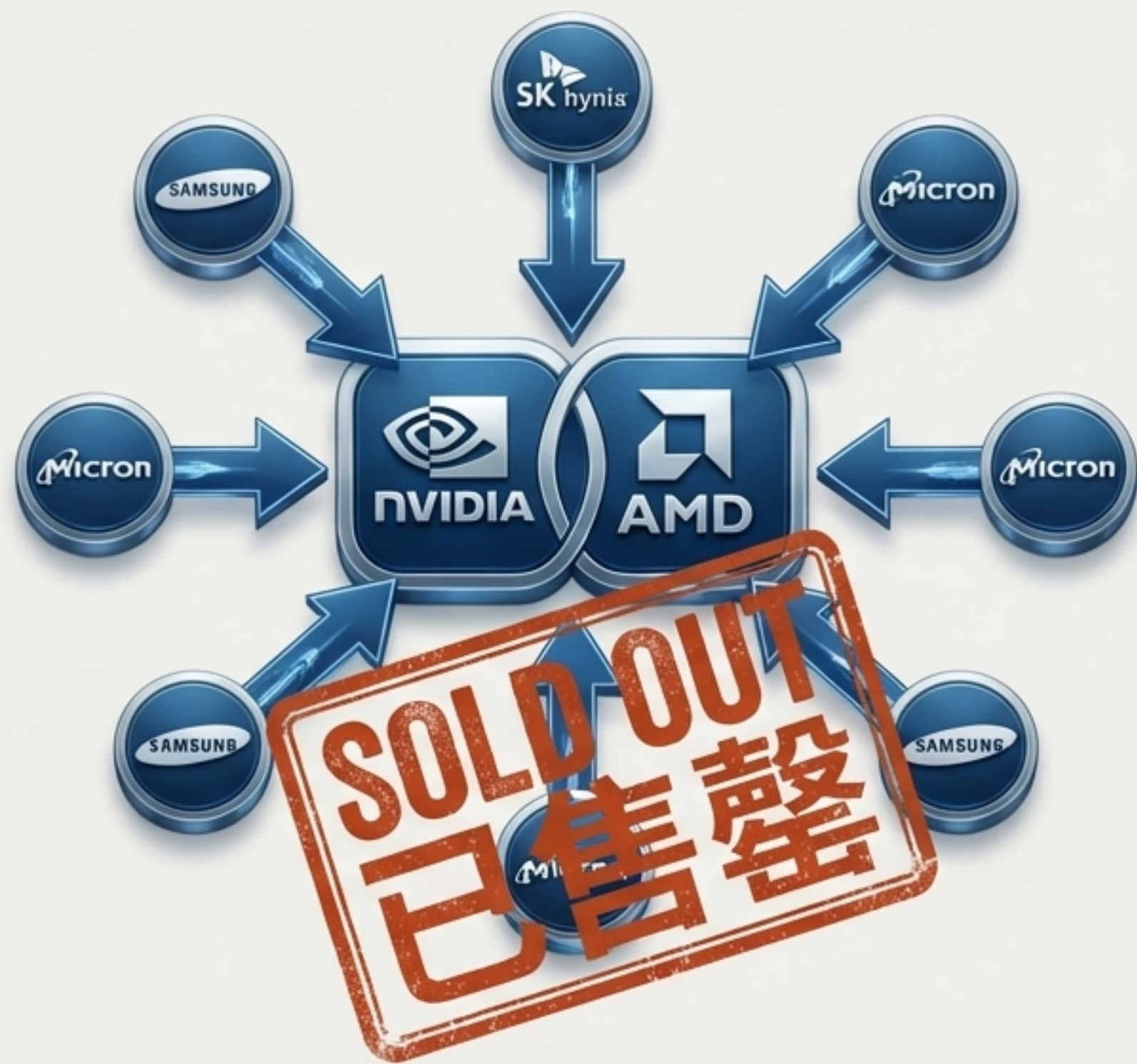
感測器模組簡化，無需內建運算單元，從而縮小尺寸、功耗和成本。

- **簡化 OTA 更新**

只需更新單一域控制器，而非多個分散式模組，降低了更新的複雜性。

HBM 記憶體「已售罄」：AI 時代最關鍵的戰略資源

- AI 加速器對記憶體頻寬的巨大需求，已使 HBM 成為唯一的解決方案，導致了長達數年的供需嚴重失衡。
- 市場動態與競爭格局：
 - 美光 (Micron)：憑藉 HBM3e 的功耗效率優勢，獲得 NVIDIA 的「首選供應商」地位。
 - SK 海力士 (SK Hynix)：憑藉 MR-MUF 技術，在 HBM3/3E 時代持續領先。
 - 三星 (Samsung)：解決 12 層堆疊良率問題後，預計於 2026 年初強勢回歸。
- 技術展望：業界正準備迎接 HBM4，該技術將利用「混合鍵合」實現更高密度和速度。



速度、功耗與可靠性之爭：SRAM vs. 新興記憶體

SRAM (靜態隨機存取記憶體)

優勢：存取速度達次奈秒級，快取首選。

挑戰：功耗高、密度低、易受輻射導致軟錯誤。

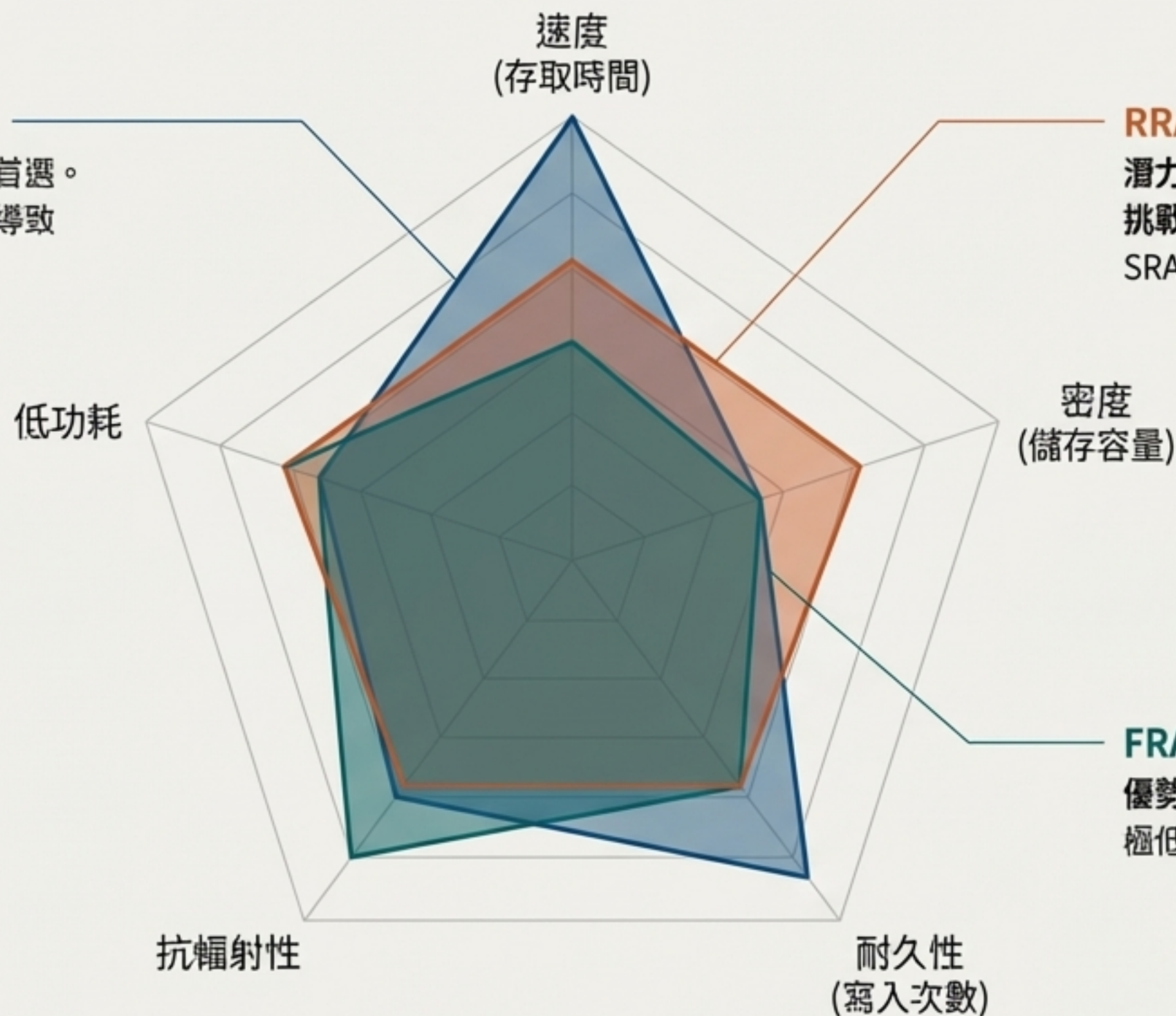
RRAM (電阻式記憶體)

潛力：非揮發性，密度高於 SRAM。

挑戰：寫入耐久性有限，速度仍不及 SRAM。

FRAM (鐵電記憶體)

優勢：結合速度與非揮發性，功耗極低，具有抗輻射特性。



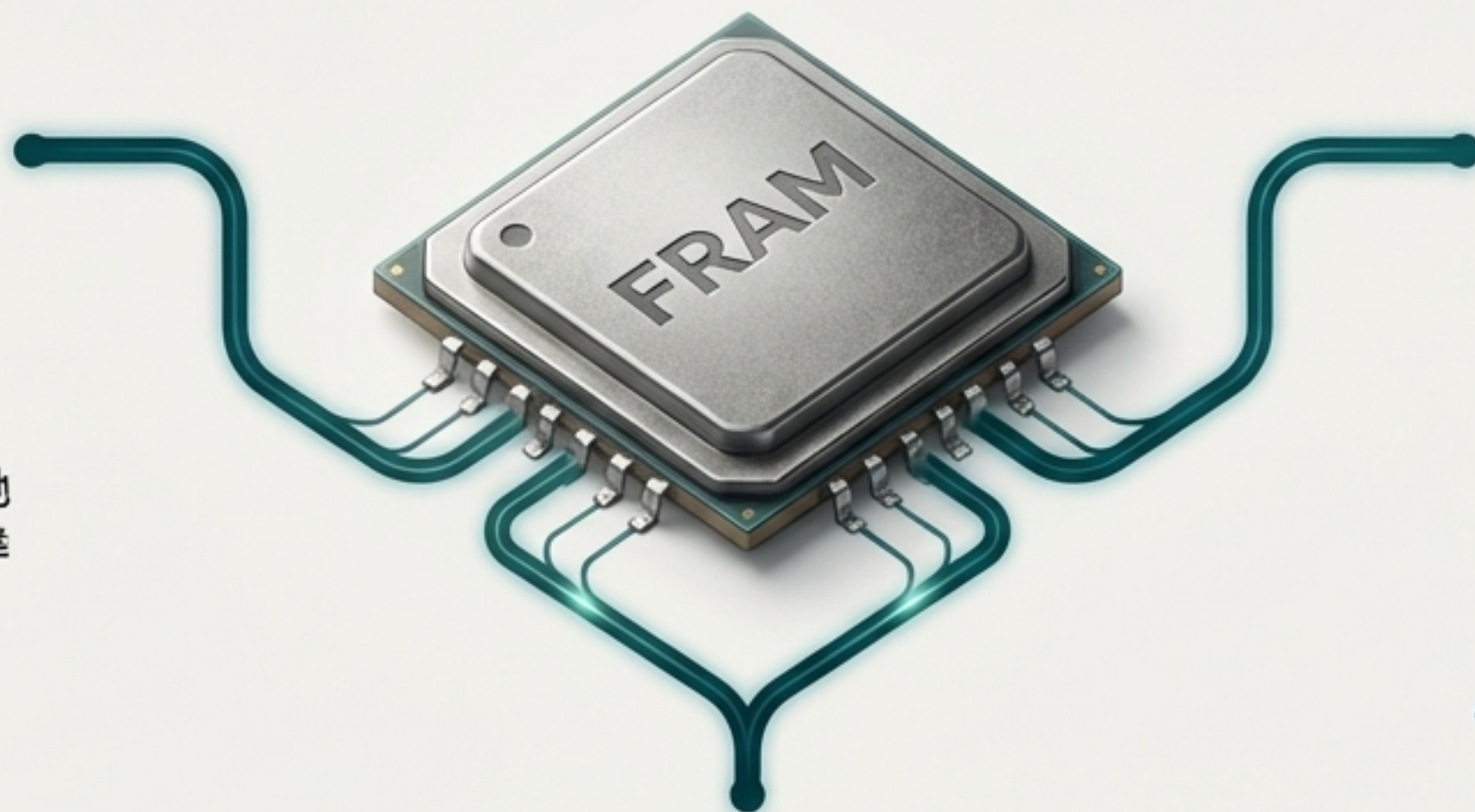
在極端環境中確保可靠性：鐵電記憶體的应用潛力



航太應用

NASA 正在測試將 FRAM 用於近地軌道衛星，以應對輻射導致的標準記憶體損壞問題。

抗輻射



醫療應用

可承受醫療設備滅菌過程中的嚴苛伽瑪射線輻射，適用於智慧注射器、植入式設備等。

耐滅菌



汽車應用

在高速車輛監控系統中，提供快速、可靠的資料擷取，確保斷電時關鍵數據不遺失。

高可靠性

核心原理：FRAM 利用鐵電材料的極化狀態儲存數據，而非電荷，使其對輻射引起的電荷洩漏不敏感。

群雄逐鹿：GPU、CPU 與 ASIC 的新戰場



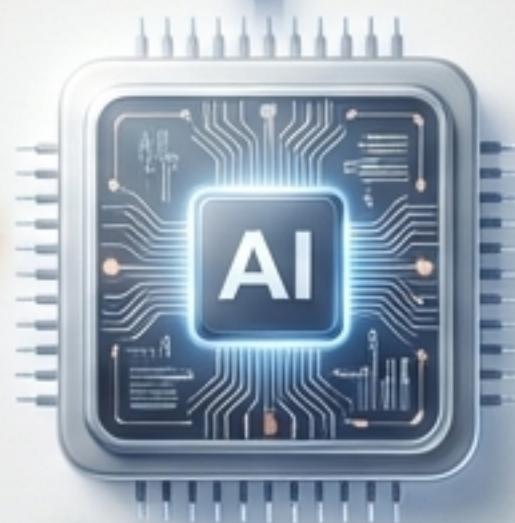
NVIDIA：憑藉 Blackwell Ultra GB300 和 Rubin 平台鞏固霸主地位。CUDA 生態系統是其最強大的護城河。



AMD：推出 Instinct MI350 系列，憑藉 CDNA 4 架構和 ROCm 軟體平台持續追趕。



Intel：透過 Gaudi 加速器和 18A 製程，試圖在 AI 領域取得突破。將於 2026 CES 推出 Panther Lake。



AI 運算

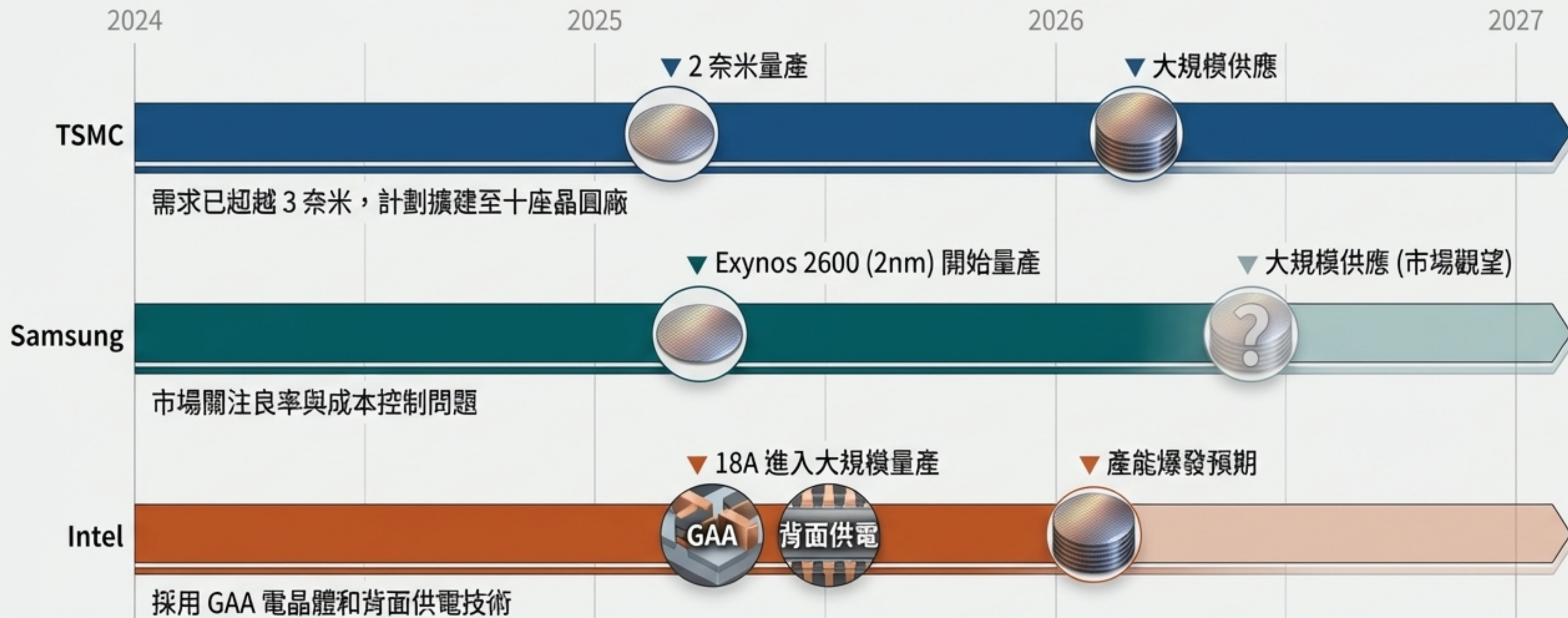


ASIC 陣營 (Google TPU)：在特定模型上展現更高運算速度和能效，體現「軟硬體整合」優勢。



RISC-V：正式進入 AI 運算核心戰場，應用場景擴展至智慧汽車和資料中心。

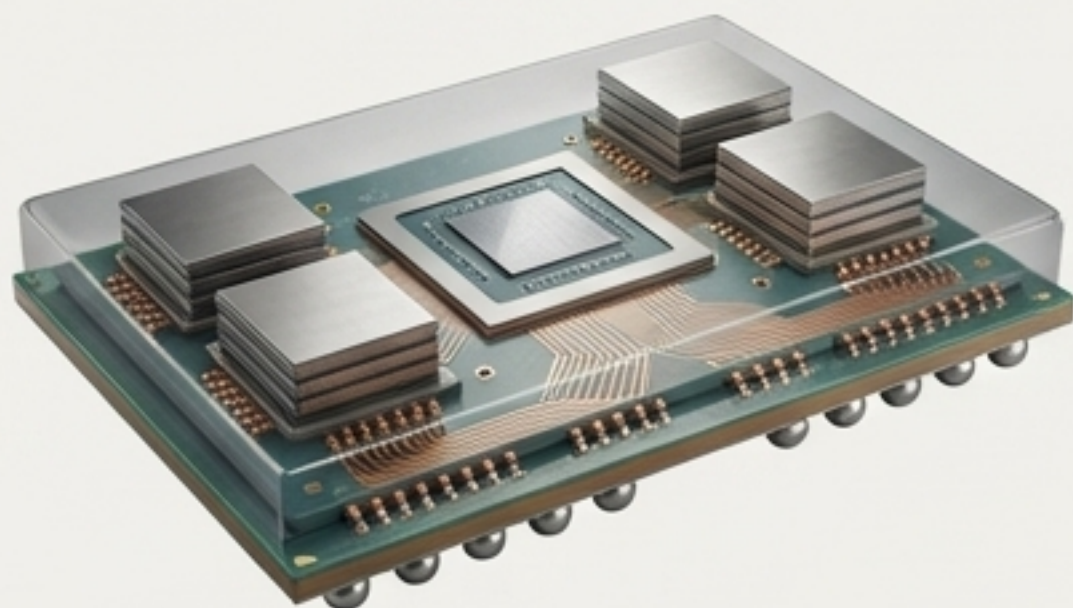
矽的終極競速：2 奈米製程的量產之路



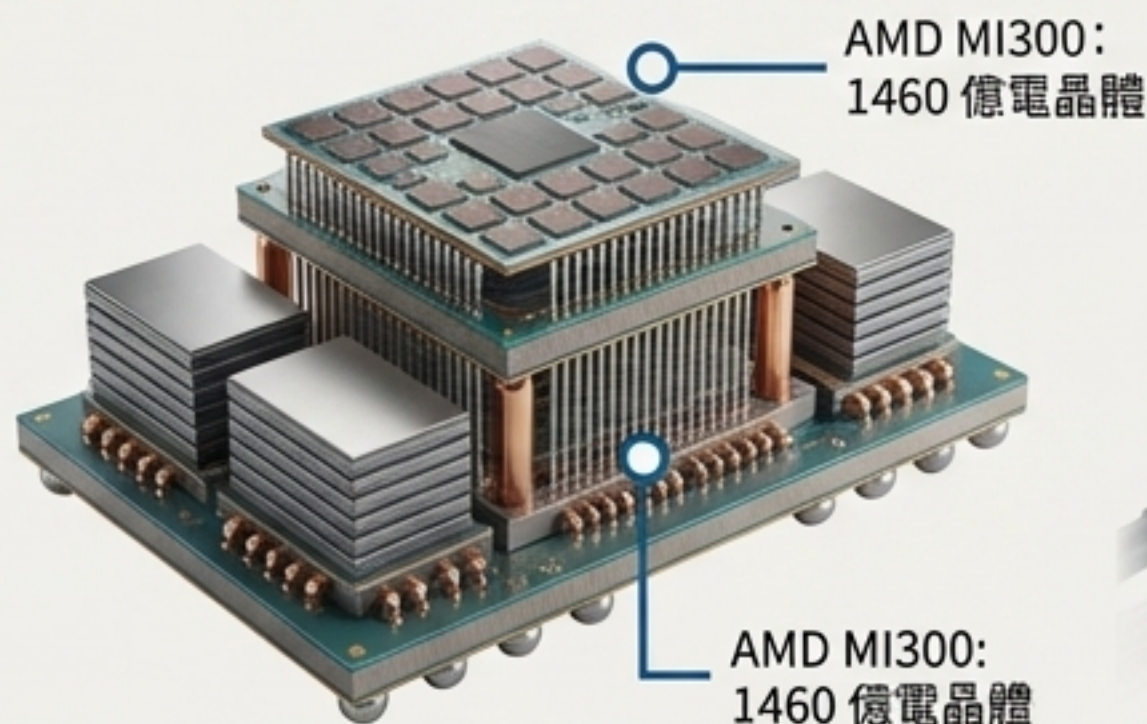
關鍵觀察點：「量產」不等於「大規模供應」。真正的產能爆發預計要到 2026 年才會到來。

超越摩爾定律的新路徑：封裝即系統

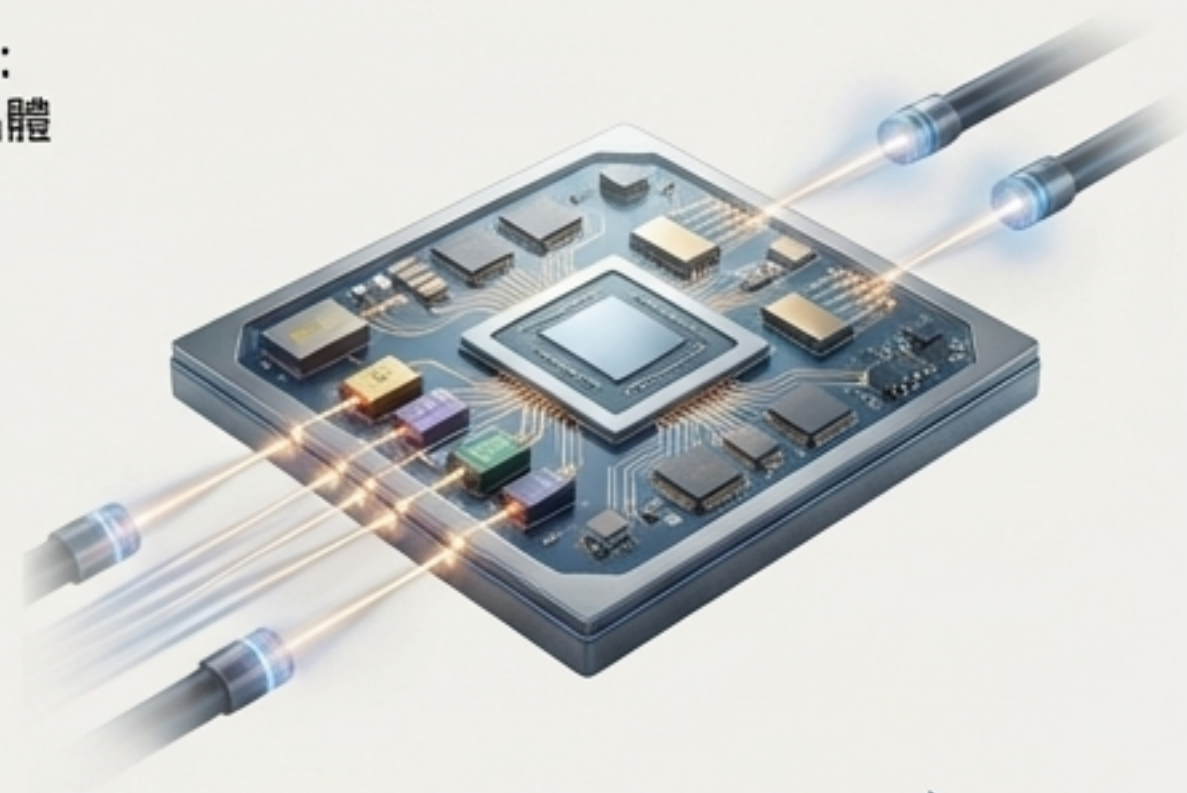
2.5D CoWoS



3D 堆疊



共同封裝光學 (CPO)



封裝技術演進

TSMC CoWoS 產能擴張

- 為滿足 AI 晶片需求，預測至 2026 年底月產能將達 12-13 萬片晶圓。

3D 堆疊的極致應用

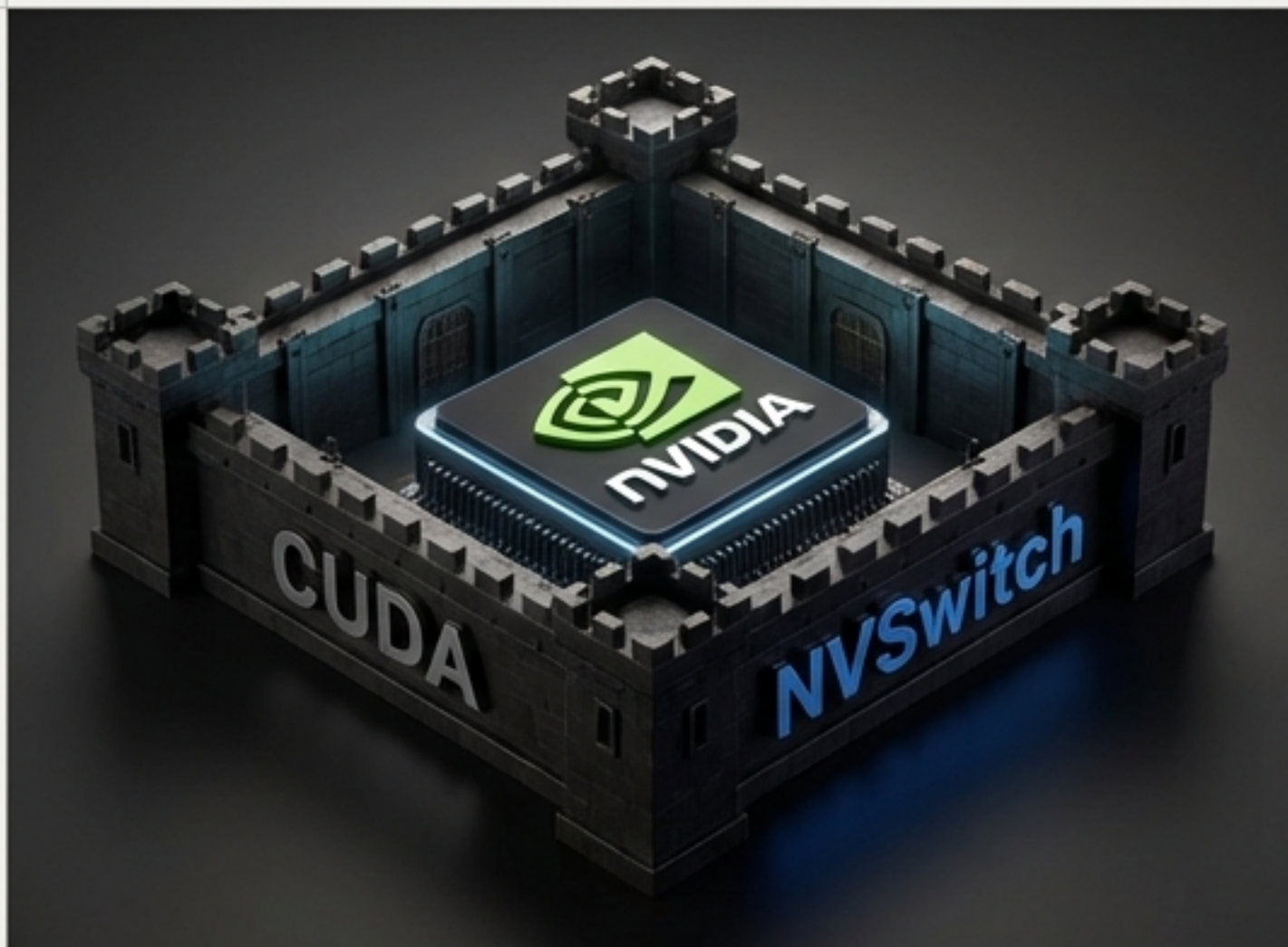
- 將 CPU、GPU 和 HBM 垂直整合，實現前所未有的電晶體密度。

未來趨勢 – CPO

- 為解決 1.6T 及更高傳輸速率的瓶頸，將光學元件與晶片共同封裝。

開放或封閉？UCIe 標準能否打破 NVIDIA 的硬體壟斷？

封閉生態系



開放生態系



UCIe 的潛在影響

- 催生第三方創新：專門提供壓縮、安全或推論等功能的 Chiplet 新創公司將有機會崛起。
- 增強供應鏈韌性：允許從多家晶圓廠採購 Chiplet，降低對單一供應商的依賴。
- 為超大規模資料中心業者提供彈性：可根據自身需求組合最佳 Chiplet，以降低成本並實現客製化。

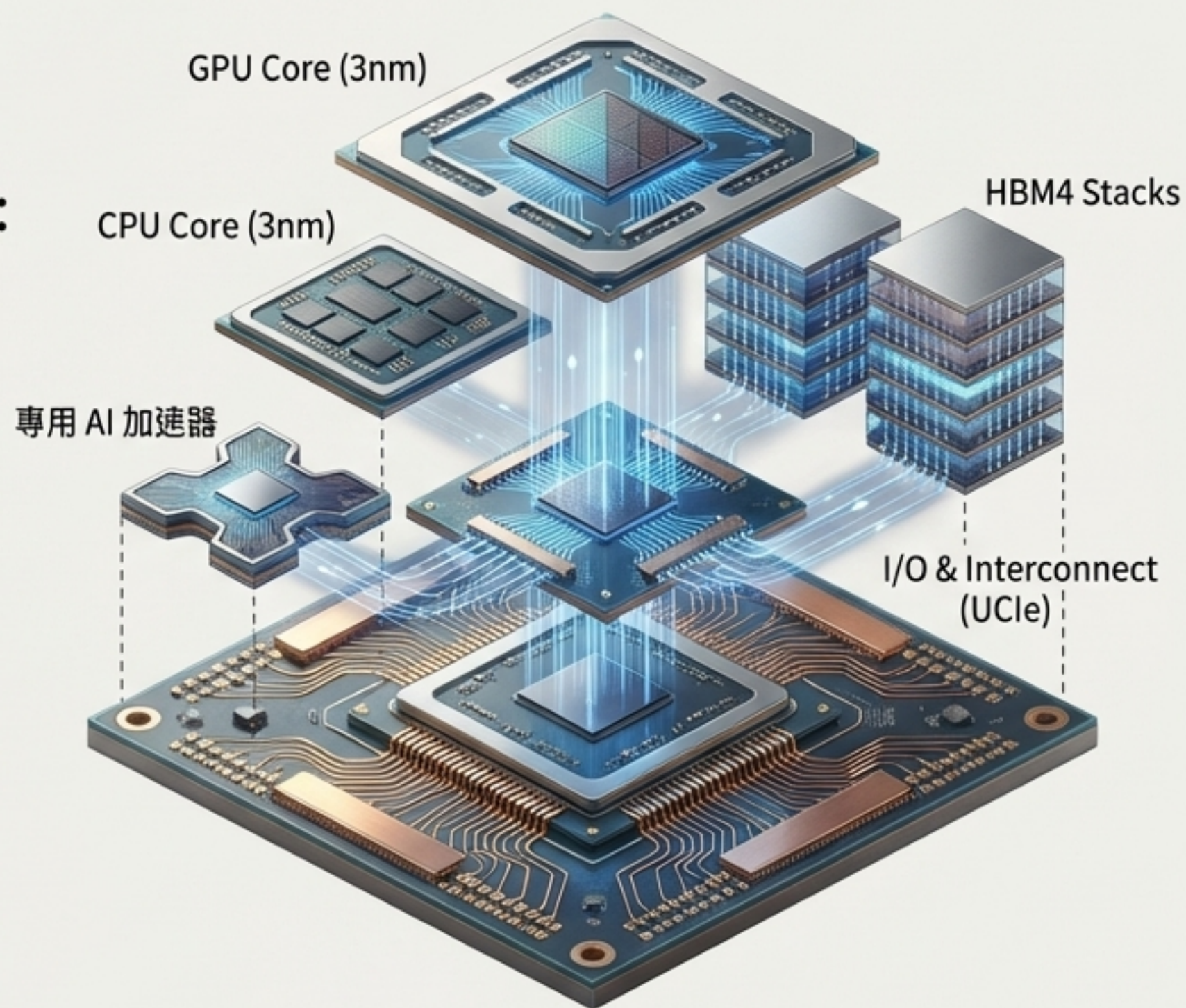
挑戰：NVIDIA 目前對 UCIe 持抵制態度，行業需團結在開放標準周圍以推動變革。

迎接兆級電晶體、異質整合與客製化的新紀元

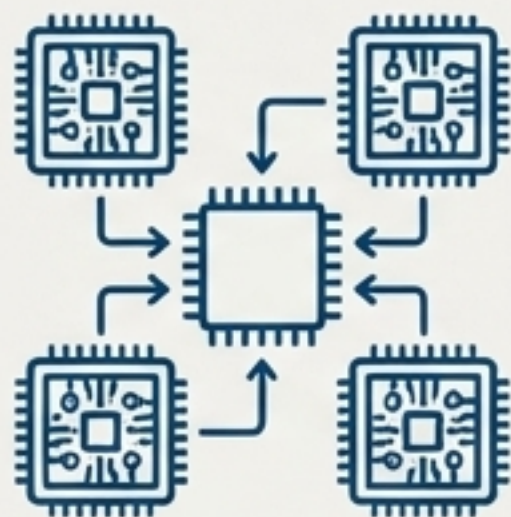
綜合所有趨勢，到 2030 年，主流的高效能運算系統將具備三大特徵：

1. **兆級電晶體封裝 (Trillion-Transistor Packages)：**
透過 Chiplet 和 3D 堆疊，將 CPU、GPU、超過 10 個 HBM 堆疊和專用加速器整合在一個封裝內。
2. **開放的多供應商生態：**UCIe 等標準將使「隨插即用」的客製化封裝成為可能。
3. **極致的專用化：**針對特定 AI 工作負載的專用 Chiplet 將被整合進主流平台。

結論：未來的創新單位不再是單一的晶片，而是整個「封裝」。競爭的焦點將從誰能製造最快的 GPU，轉向誰能最有效地整合不同技術，以構建最高效能的系統。

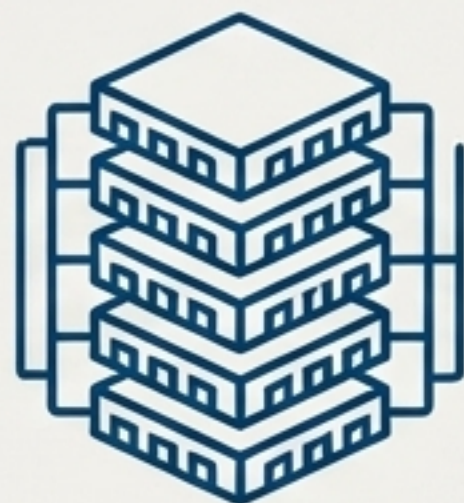


總結：重塑運算的三大核心力量



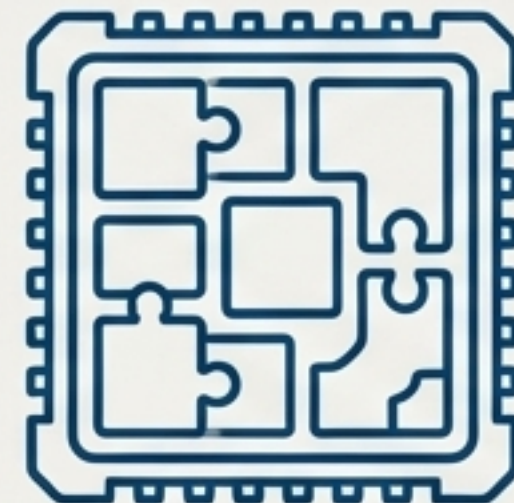
架構革命 (Architecture Revolution)

Chiplet 正在取代單晶片，成為高效能運算的新基礎。這種「解構」使得超越光罩極限、整合異質功能成為可能。



記憶體為王 (Memory is King)

HBM 已成為 AI 效能的關鍵瓶頸與市場焦點。同時，FRAM 等新興記憶體憑藉其獨特的可靠性，為邊緣和極端環境應用開闢了新道路。



整合至上 (Integration is Everything)

競爭的終極戰場已從單一晶片轉移至先進封裝和開放生態系的構建。未來的勝利者將是那些最擅長整合不同技術以打造最佳系統的公司。

謝謝