



著作權聲明公告

本活動資料(包括但不限文字、圖片、影音等講義內容或其他文件)之著作權歸屬於「財團法人資訊工業策進會產業情報研究所(MIC)」所有，均受著作權法之保護及國際著作權法律的保障，僅授權報名本活動之個人非商業用途，並請註明引用出處及來源。

謹提醒，倘個人未取得書面授權同意，逕自透過電子郵件或LINE等媒體媒介轉載分享本活動資料，已經侵犯MIC的智慧財產權，將視情節之重大程度提出法律追訴，如經確認違法行為，不僅個人受罰，公司亦將負連帶賠償責任，並造成公司商譽之損害。

感謝您對於智慧財產權尊重與理解，如有意請求授權使用本活動資料，歡迎聯繫02-2378-2306；members@iii.org.tw，謝謝您！

全球資料中心市場發展趨勢與商機

魏傳虔

產業顧問兼副主任

產業情報研究所(MIC)

財團法人資訊工業策進會

2026/08/27

NVIDIA「The Five-Layer AI Cake」揭露AI應用新方向

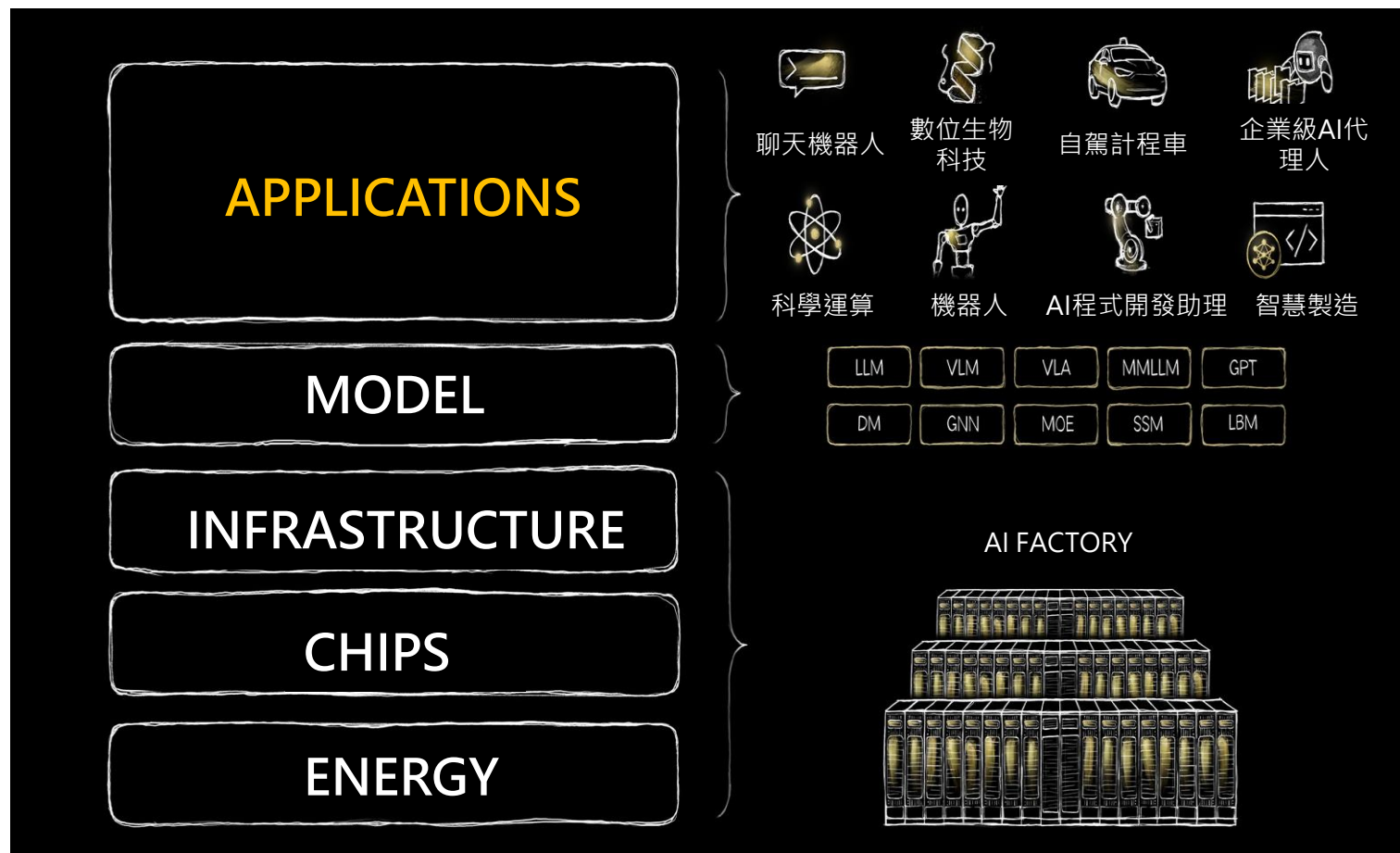
「應用」是經濟價值產生之處。機器透過鏡頭、感測器能夠看及感受周圍環境，並透過AI模型提供**解讀周圍環境能力**，使其能對現實世界進行反應，如獨立導航、車輛組裝、物流管理等

AI模型能理解多種類型的資訊，包括語言、生物、化學、物理、金融、醫療，以及真實世界本身。為實體AI提供思維能力並指導機器行動

基礎設施包括土地、電力輸送、散熱、建設工程、網路，及將數萬顆處理器協調運作為單一AI運算系統，這些系統稱之為「AI Factory」

晶片層的進展，決定了AI能以多快的速度擴展以及智慧成本能降到多低

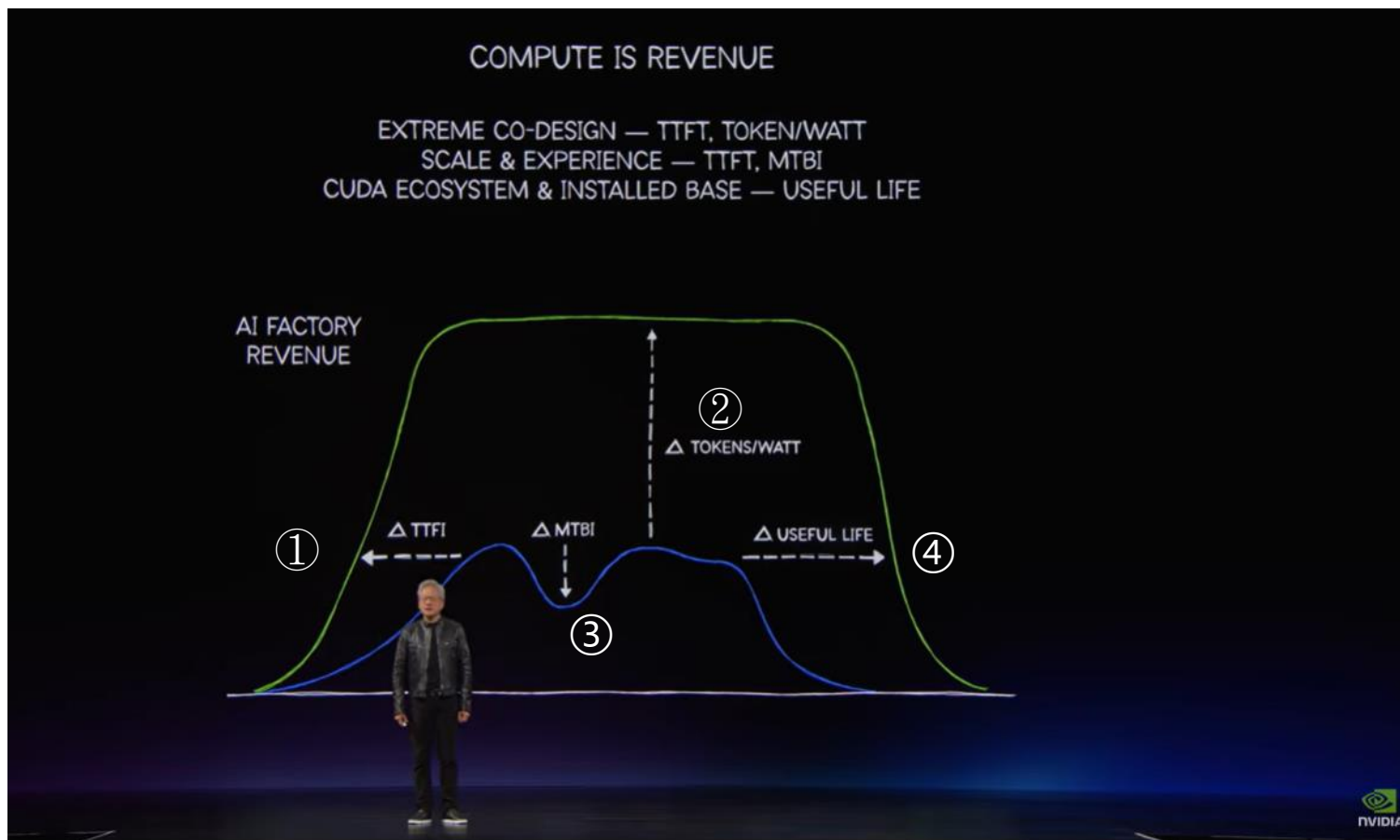
能源是AI基礎設施的第一原理，也是限制系統能產生多少智慧的根本約束



資料來源：NVIDIA，MIC整理，2026年8月

- ◆ 「五層AI蛋糕理論」揭示AI產業垂直整合邏輯：底層晶片與基礎設施為算力地基，中層平台與模型為技術核心，頂層應用則是最終價值實現。每一層的穩固與進化，皆是支撐上層開發與最終軟體體驗的關鍵

算力即營收，AI Factory仰賴整體系統整合與彈性



資料來源：NVIDIA · MIC整理 · 2026年8月

- ◆ 藍線代表AI資料中心成本趨勢線，綠線代表NVIDIA AI Factory能提供AIDC提高營收的競爭力
- ◆ AIDC更快上線、能源效率佳、基礎設施彈性規劃，助力創造更多營收

1. AI Factory部署速度 (TTFT)

代表從資料中心蓋好到能營運的時間
多久時間能讓AIDC跑出第一個token

2. 生產效率 (Tokens /Watt)

當晶片運算效率高，冷卻與電力系統
耗電越少，更多電力用於運算系統，
每一瓦電能產生更多的Token數

3. AI Factory可靠性 (MTBI)

平均中斷時間，代表系統或叢集在發生下
一次硬體故障、軟體錯誤或中斷任務之前，
平均能夠穩定運行的時間長度

4. 使用生命週期 (Useful Life)

確保資料中心架構彈性能因應未來系統需求

AI DC建置關鍵議題 從算力系統延伸到整體工程系統

AI運算任務變化

(AI Computing)

AI算力需求從訓練擴大到推論，加速運算架構異質整合

機櫃功率持續飆升

(AI Rack)

高密度機櫃成為出貨核心，推動AIDC系統整合

AI加速器互連架構

(Network)

為擴充AI加速器算力規模，及擴大記憶體池，互連技術與速度成為重要議題

液冷滲透率提升

(Liquid Cooling)

AI晶片功率提高且高密度機箱/機櫃增加，液冷系統採用率更高

電力系統升級

(Peak power)

為了確保高功率AI機櫃供電穩定，需升級機櫃及基礎設施供電系統

電網擴張出現瓶頸

(Utility)

電網容量限制已成為AIDC選址、擴張的最大挑戰

資料來源：各公司，MIC整理，2026年8月

- ◆ 為滿足AI運算任務變化所進行的系統優化，機櫃功率持續飆升影響伺服器與互連架構
- ◆ 支持運算系統的關鍵基礎設施，需要克服散熱與大電流的兩個物理議題，及強化電網韌性

大綱

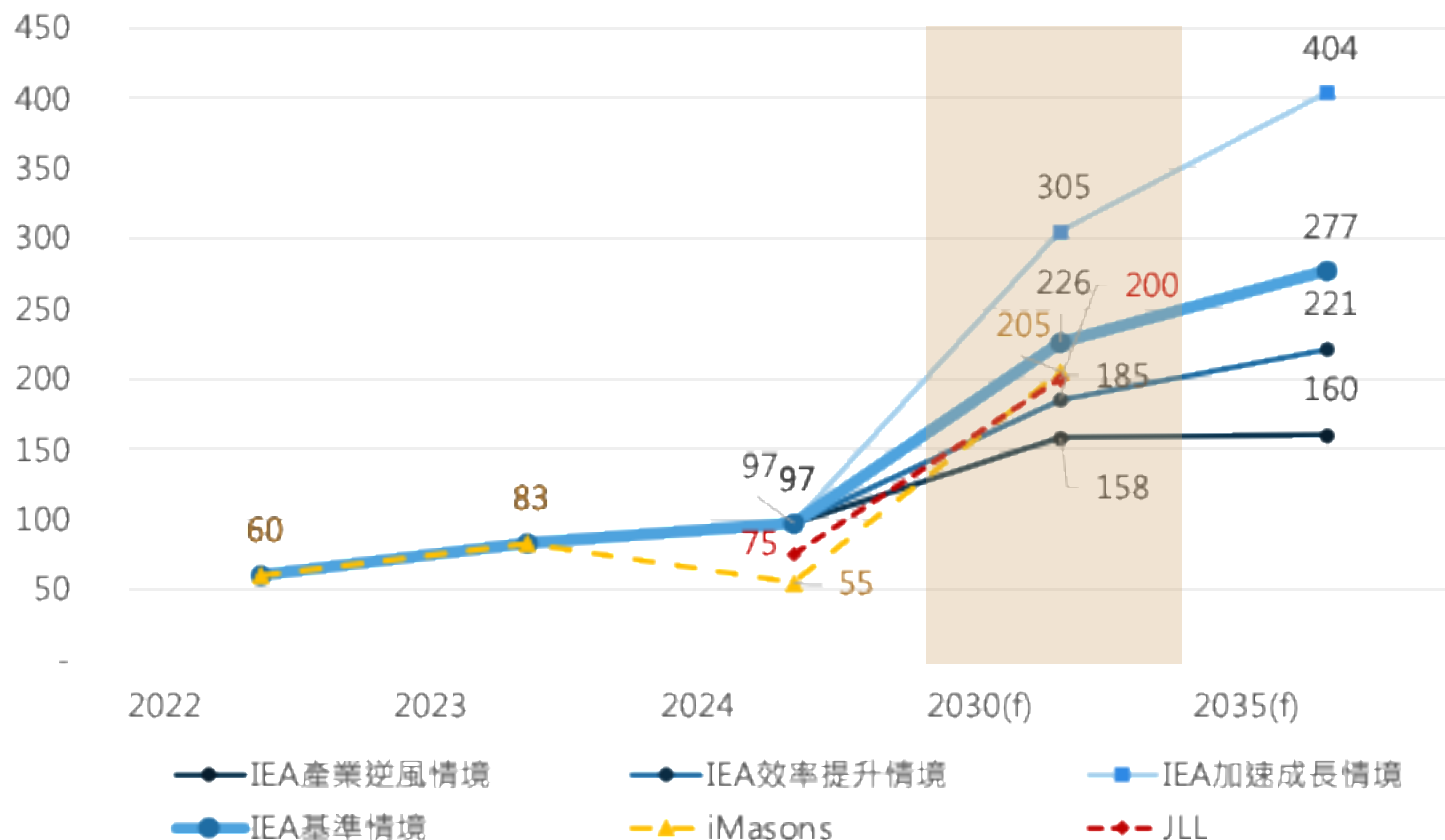
- ◆ 全球市場趨勢
- ◆ 關鍵議題分析
- ◆ 台灣產業機會
- ◆ 結論

全球市場趨勢

AI算力需求成長 資料中心電力規模在2030年將突破200 GW

單位：GW

國際機構對全球資料中心電力規模預測



AI為資料中心用電主要驅動力

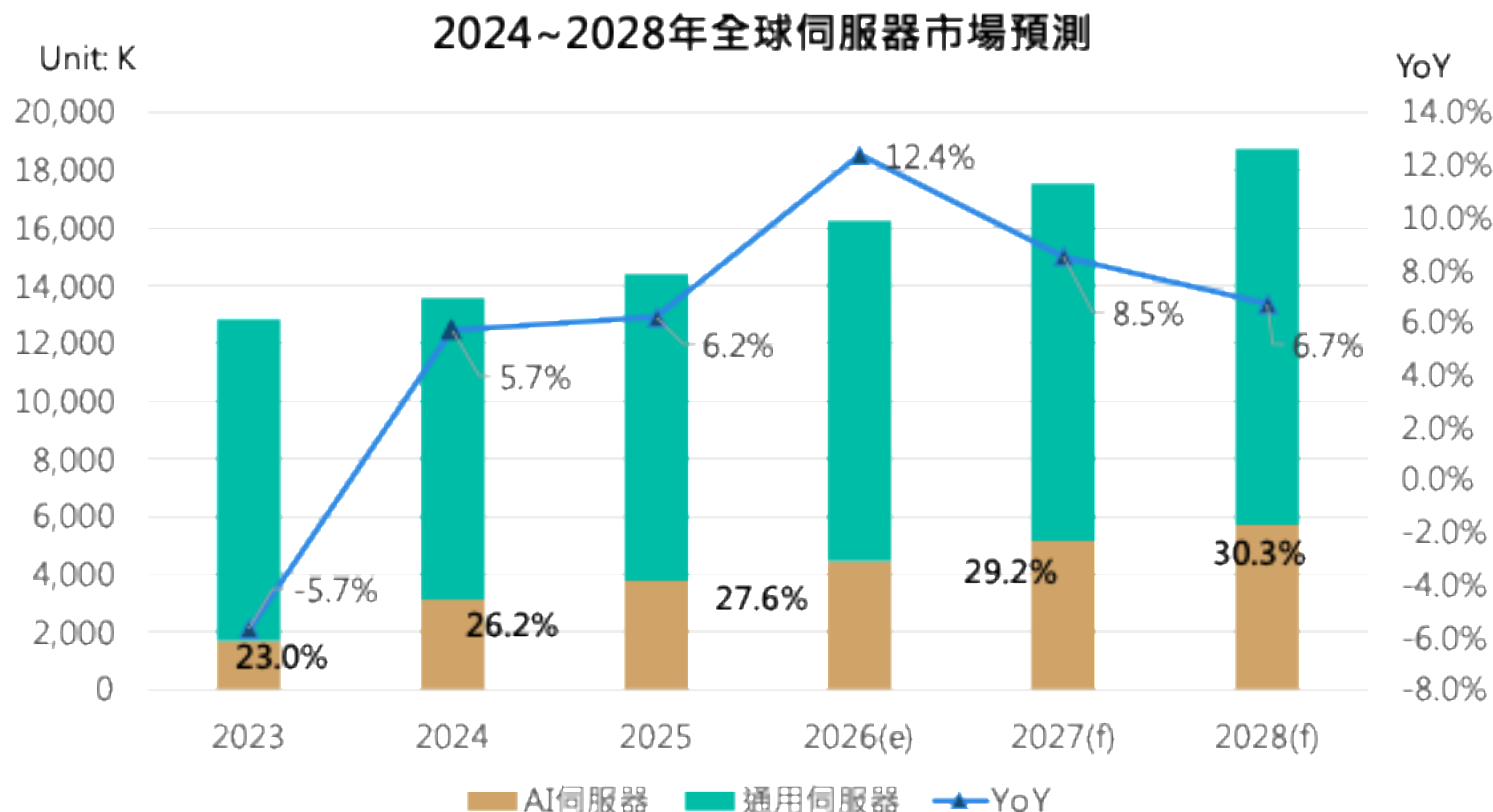
- ◆ 電力快速增長動力，主要來自於AI算力建置需求
- ◆ 依據現有狀況穩定推展，IEA認為2030年總電力容量可達226 GW，其他研究機構之預測也都在200 GW以上

AI推論自2026年開始進入擴張期

- ◆ 2024-2025年的電力增長，以AI訓練需求為主
- ◆ 2026年以後，進入到AI應用擴張階段，用電將以AI推論為主

資料來源：IEA、JLL、iMasons、MIC整理，2026年8月

2026年進入AI資料中心建設狂潮 推動整體伺服器出貨成長



資料來源：MIC，2026年8月

影響全球伺服器市場因素

雲端服務商

擴增美國GW級超大型資料中心

AI伺服器、基礎設施資本支出占比增加

自研晶片、客製化伺服器採購數量提升

NeoCloud

持續建造AI資料中心

各國本地GPU運算提供商數量增多

伺服器品牌商

企業端AI伺服器需求增加

AI伺服器帶動一般伺服器採購回升

- ◆ 2026年市場核心動能來自於雲端服務商、AI模型業者、主權AI加速投資資料中心算力設備，推動市場成長
- ◆ 在AI推論成長下，雲端服務商除了加速AI伺服器布局，同時也增加一般伺服器採購，支持AI Agent算力需求



AI伺服器推動品牌業者營收成長 以中大型AI專案為主

品牌業者



ODM Direct



布局AI DC
專案

Hyperscaler



AI模型商



Neocloud



AI主權 專案



企業採購

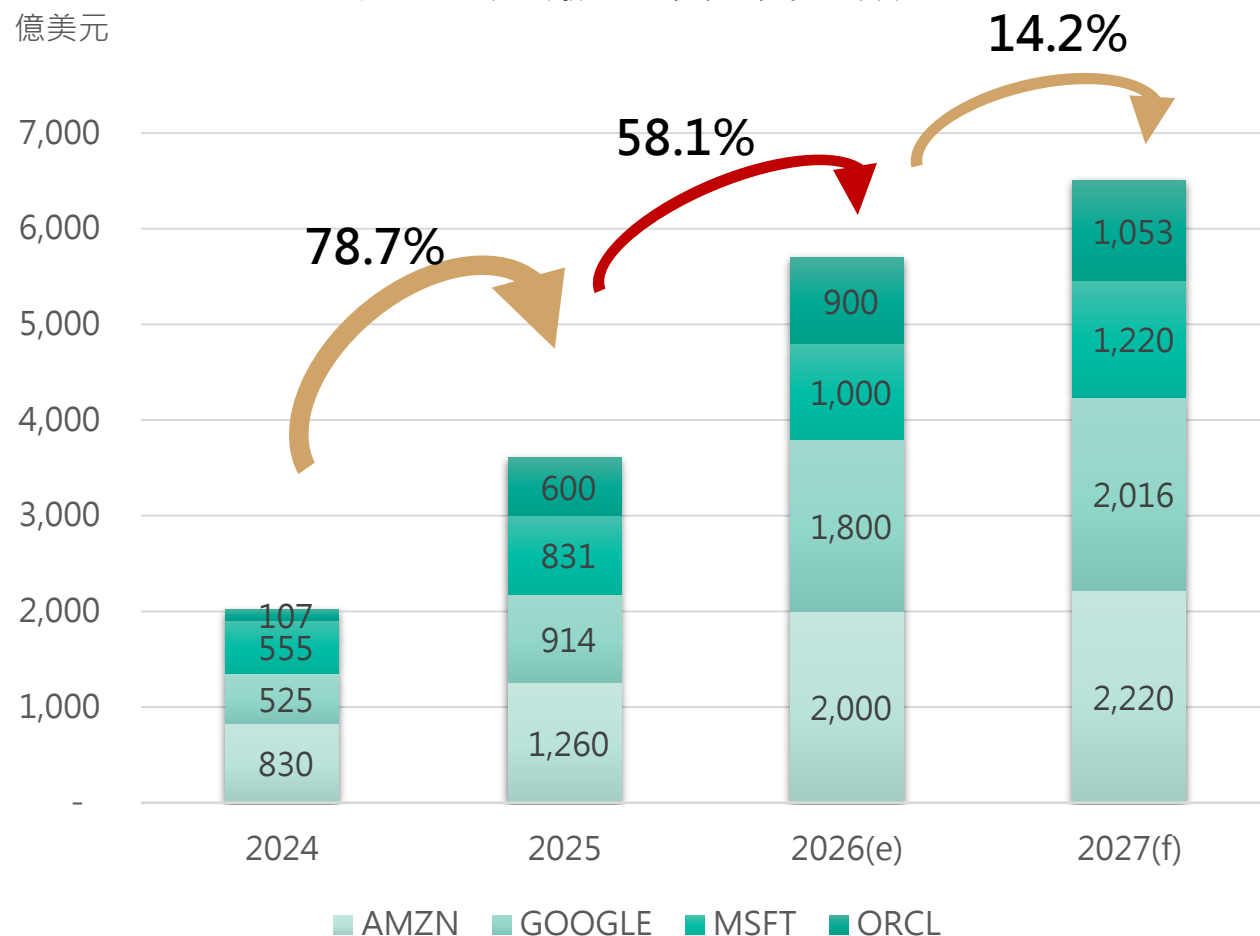
備註：僅列舉部分公司名稱

資料來源：各公司，MIC整理，2026年8月

- ◆ 隨著Hyperscaler持續布建超大規模AI DC，成為採購高階AI算力設備的主力客戶，持續推升ODM Direct出貨市占
- ◆ 當前AI伺服器採購客戶以中大型業者為主，除ODM Direct供貨CSP客戶，品牌商則以Neocloud、主權AI專案為主

AI算力供不應求 雲端服務商加速投入資料中心建設

前四大雲端服務商資本支出預測



算力供不應求、吸引大廠擴大資本支出

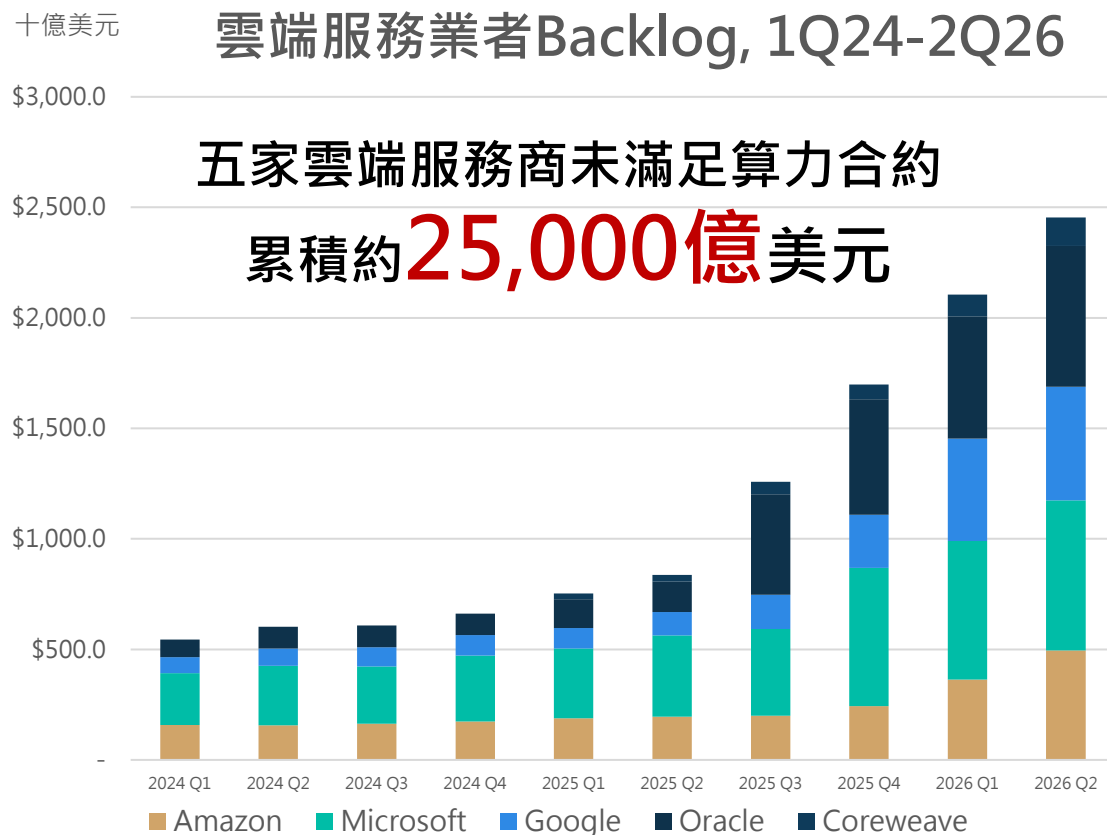
- ◆ 進入AI 基礎建設週期核心期：全球雲端服務商的資本支出（Capex）出現爆發性成長，以資料中心建設、自研晶片、網路布建為主要支出項目

超前準備 AIDC布局轉向供應主導模式

- ◆ AIDC將改變過往雲端資料中心建置模式，在市場需求浮現前，先行預備算力供給
- ◆ CSP正持續鎖定土地、電力、晶片，及更多關鍵零組件供給，為未來AI算力需求做準備

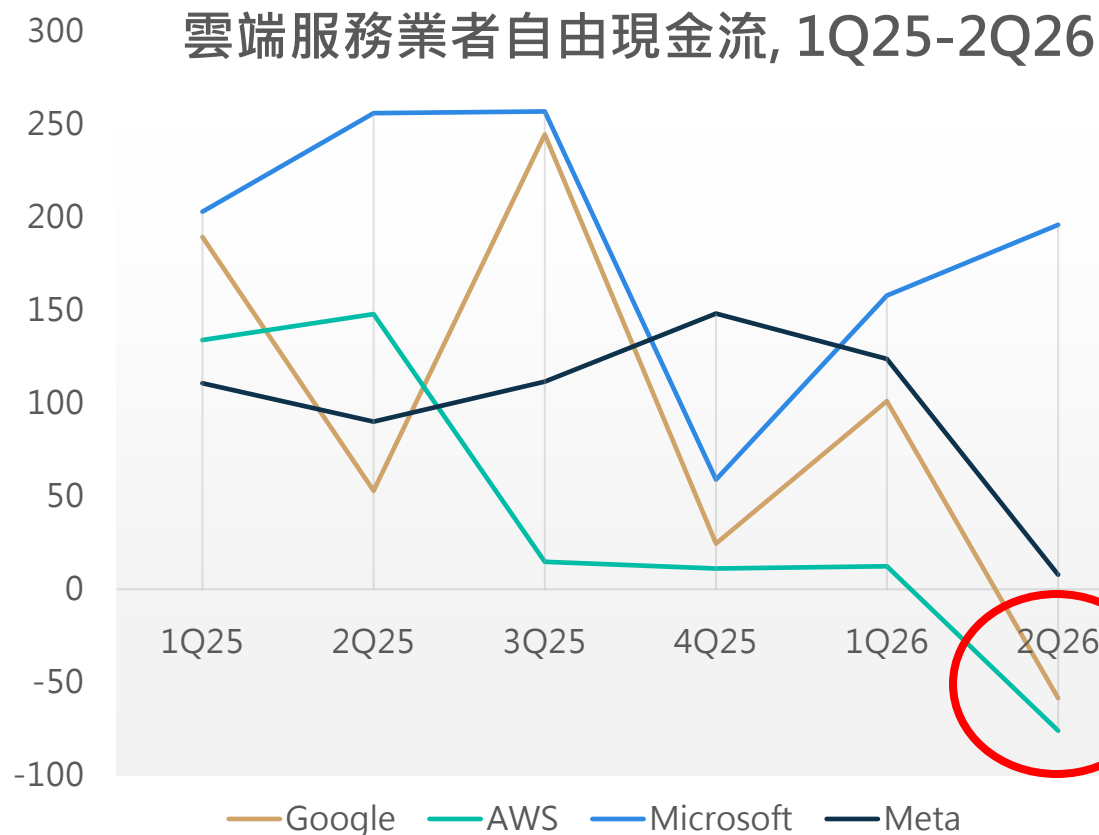
資料來源：各公司，MIC整理，2026年 8 月

算力需求超過供給速度，加速投入算力建設轉為經濟收入



資料來源：各雲端業者，MIC整理，2026年8月

- ◆ AI算力需求強勁：參考雲端服務商財報會議資訊，其累積龐大的客戶訂單，顯示AI需求強勁，現階段市場算力需求大於供給



資料來源：各雲端業者，MIC整理，2026年8月

- ◆ Google、AWS自由現金流在2Q 26首次轉負，顯示AI基礎建設支出快速增加，甚至超過營運現金流，增加外部融資需求，除以舉債方式補充資金，也將以算力租賃（減少自建比重）方式，提高資本支出彈性

因應AI運算任務多元化 產業巨頭調整發展戰略



倡導AI DC生態系

- 倡議AI資料中心開源生態系(Open Data Center for AI)
- 機櫃即伺服器(Rack-Scale Computing)

2025/10

2025/12

2026/3

GTC 2026- NVIDIA 高舉大推論時代

- 黃仁勳提出AI五層蛋糕論述，表示AI進入推論時代，AI工廠將因應上層AI應用需求而有不同的算力配置



成立新AI戰略部門 因應市場趨勢

- 部門三大區塊：AI模型、晶片發展，量子運算
- 軟硬設計一體化，統籌新一代AI DC建置規範，DC從原本IT基礎設施概念轉向AI工廠



算力需求鎖定推論

- Meta預計在2026下半年推出自研晶片 MTIA 300，將優先鎖定推論運算
- 除晶片外，也將打造自研晶片機櫃系統



推出自家CPU

- Arm首度發表自研資料中心CPU，將業務從IP授權擴展到晶片產品
- 鎖定資料中心及代理式AI應用

從Google與OpenAI的Token消耗量反映出AI推論時代來臨



資料來源：Google · 2026年8月

- ◆ 每月有超過**850萬**開發者使用Google模型建立新的應用程式和體驗
- ◆ 我們的模型API目前每分鐘大約可以處理**190億**個Token
- ◆ 過去12個月裡，超過**375家**Google Cloud客戶各自處理了超過**1兆**個Token，代表各行各業對AI的巨大需求

	用 OpenAI 打造應用的開發者人數	ChatGPT每週使用者數量	API 平台每分鐘Token數
2023	200萬	1億+	3億
2025/10	400萬	8億+	60億
2026/03	400萬	9億+	150億

資料來源：Open AI · 2026年6月

Token使用量 (2025/10)	家數	
1兆	31	 duolingo AI線上學習
1,000億	69	 perplexity AI搜尋
100億	54	 Genspark AI搜尋
		 Notion 學習筆記
		 ABRIDGE 醫療AI
		 salesforce CRM

關鍵議題分析

- AI晶片
- 機櫃平台

AI算力賽道升級 記憶體與頻寬成為推論時代的關鍵

LLM 對硬體的需求

- 算力(PFLOPs)
- 記憶體容量
- 記憶體頻寬
- 網路頻寬

運算輸出

- 低精度運算：可接受
- 可重現性：很重要
(每次運算得到相同結果)

模型訓練



- 核心：大算力導向
- 記憶體：容量 + 頻寬

讀寫量：兩者均多 (讀取數據、更新權重)



- 訓練成本：一段時間研發投入

AI推論



- 核心：記憶體頻寬

讀寫量：讀取 > 寫入 (固定權重)



- 推論成本：持續不斷支出，使用越多，支出越高

資料來源：Google · MIC整理，2026年8月

AI晶片功能分流 根據運算用途強化成本效率競爭

AI加速晶片	晶片特性	商業模式	關鍵優勢	
通用旗艦級晶片 - NVIDIA Rubin - AMD MI 400系列	支援通用運算，具高算力及大容量HBM	開放市場產品/系統與機櫃級解決方案	高通用性與生態完整 支援模型訓練、後訓練、推論，並搭配成熟軟體生態	算力核心
雲端自研晶片 - Google TPU 8t - Amazon Trainium3 - Microsoft Maia200 - Meta MTIA 300	- 定位為大規模AI模型訓練晶片，導入原生FP4 - 支援AI訓練及大規模推論，強調Token成本效益 - 針對生成式AI與大型模型的推論情境，進行極致最佳化 - 支援社群網站排名(ranking)、推薦及GenAI運算工作	內部自用/ 雲端算力租借/ 服務自家平台運算與模型運算	提高平台控制力 雲端服務商將晶片、軟體堆疊、模型服務、資料中心等強化成本效益，降低對外部GPU供應之依賴	平衡成本效益
專用低延遲推論晶片 - NVIDIA Groq 3 LPU - Google TPU 8i	- 記憶體採SRAM (500MB)，用途為在decode階段加速Token產出 - 記憶體包含SRAM(384 MB) 及 HBM(288GB)	開放市場產品 內部自用/ 雲端算力租借/	低延遲/吞吐最佳化 透過on-chip SRAM降低decode階段資料搬移時間，適合即時Token產生及長上下文推論	

資料來源：各公司，MIC整理，2026年8月

- ◆ AI晶片市場將依應用需求，分流為通用GPU、低延遲專用晶片、及CSP內部自研等AI加速晶片
- ◆ NVIDIA與Google推出推論專用晶片，搶攻市場需求擴增的推論及AI代理市場

算力應用多元強化異質架構需求 提升CPU在AI平台戰略角色



晶片商、雲端服務商布局自研CPU

	X86架構	Arm架構
晶片商	Intel Xeon	Nvidia Vera
	AMD Epyc	Arm AGI (Meta)
CSP		Amazon (Graviton)
		Google (Axion)
		Microsoft (Cobalt)
	--	

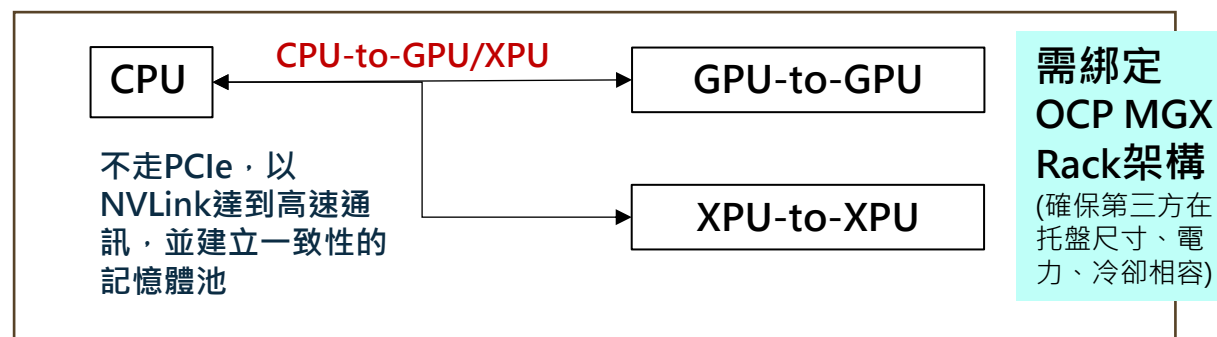
因Agentic AI、MoE、長上下文等運算服務增加，將可能增加資料中心內CPU的配比

資料來源：各公司，MIC整理，2026年8月

- ◆ AI工作任務型態更多元下，在Agentic AI運算架構下，需要更多CPU承擔任務調度、記憶體管理，及系統編排等，驅動資料中心由算力導向轉向算力+複雜任務調度系統整合
- ◆ 對採用客製化XPU廠商而言，導入NVLink Fusion能將其CPU建立起與NVIDIA GPU高速通訊的機會

NVIDIA Fusion提供異質運算架構生態整合

透過NVLink-C2C協議，第三方晶片接入NVIDIA運算架構



FUJITSU Fujitsu將其CPU與NVIDIA GPU互連

Qualcomm 利用NVLink Fusion將其CPU切入AI資料中心市場

amazon Amazon自研Graviton與Trainium4導入NVLink Fusion，並採用NVIDIA MGX機櫃架構，簡化資料中心各平台管理

關鍵議題分析

- AI晶片
- 機櫃平台

因應AI算力擴展需求 從單機伺服器往大規模機櫃平台發展

傳統伺服器機箱



4U RTX PRO 伺服器



2U NVIDIA MGX 伺服器



4U HGX 液冷伺服器

企業級大規模運算系統

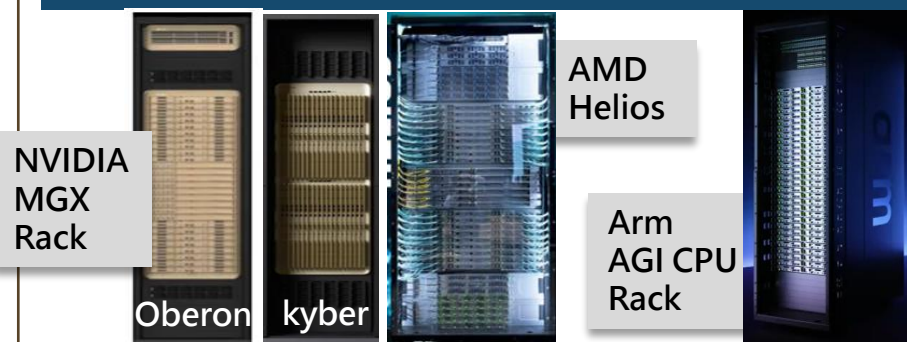


Supermicro
NVIDIA HGX

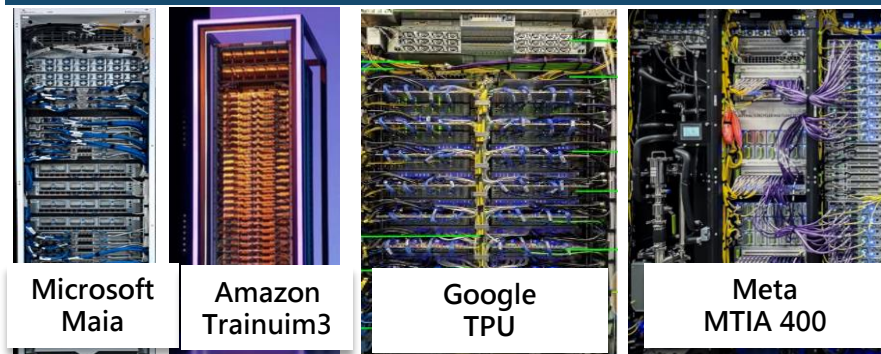


Asus
NVIDIA NVL72

晶片商機櫃級系統產品



雲端服務商AI機櫃

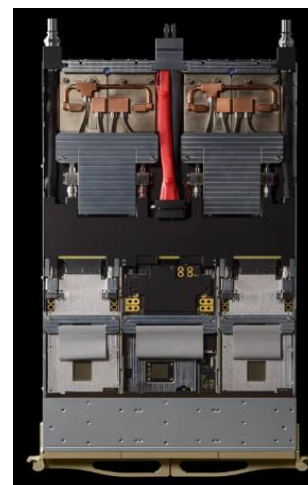


資料來源：各公司，MIC整理，2026年8月

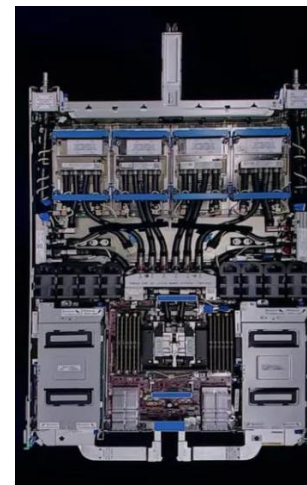
- ◆ 伺服器產品型態兩極化，高階大規模AI運算系統（包含模型訓練與重度推論）採機櫃級出貨，整合液冷散熱、櫃內電力匯流排與高速網路互連
- ◆ 通用型伺服器、中小型模型推論系統及邊緣運算設備等，因功耗相對較低，仍維持氣冷架構並可按單台出貨

高密度AI運算平台走向液冷 + 匯流排供電 + 模組化設計

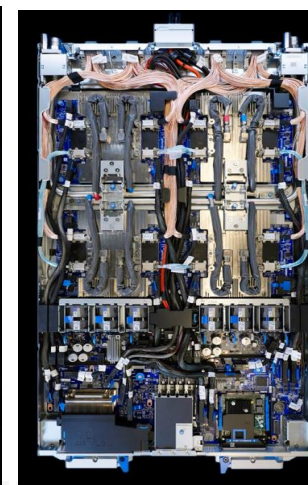
	CPU *數量	AI加速器 *數量	供電	散熱
NVIDIA NVL72	Vera (Arm) *2	Rubin. *4	匯流排	100% 液冷
AMD Helios	Venice (x86) *1	MI455. *4	匯流排	70% 液冷
Google TPU	--	TPUv7 *4	匯流排	70% 液冷
Amazon UltraServers	Graviton5 (Arm)*1	Trainium3 *4	匯流排	70% 液冷
Microsoft Maia	--	Maia 200*4	匯流排	70% 液冷
Meta MTIA	(TBD) *1	MTIA 400*4	匯流排	70% 液冷



NVIDIA Rubin



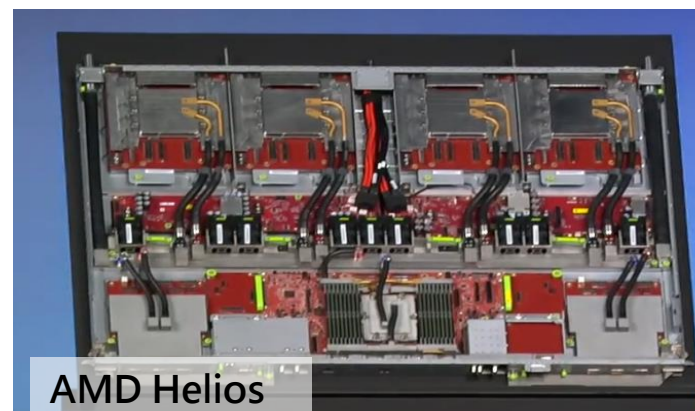
AWS Trainium3



Microsoft Maia



Meta MTIA



AMD Helios



Google Ironwood

資料來源：各公司，2026年8月

- ◆ 從各家托盤設計可見，伺服器競爭重心正由單機板卡性能，延伸到整櫃級的電力與散熱整合能力
- ◆ 產業價值將不只集中在AI加速器本身，而是同步往液冷模組、電源系統、機櫃整合與系統驗證移動

AI伺服器/機櫃密度提高 關鍵零組件規格同步拉升

	傳統伺服器	第一代AI伺服器 (A100/H100)	第一代AI機櫃 (NVIDIA Oberon NVL72)	下世代AI機櫃 (NVIDIA Kyber NVL576)
單機櫃功耗	5-15 KW	20-40 KW	120-140 KW	200+ KW
GPU TDP	無	700 W	Up to 1,100 W	Up to 1,800 W
IC載板技術	ABF載板	大尺寸ABF載板	大型封裝/高功率載板	更大封裝，未來材料可能從樹脂轉向玻璃
PCB層數(板級)	10-14層	20-26層	32+層	更高層數
配電電壓	12V DC	12V/48V DC	48V/54V DC 機櫃內配電	800V DC 機櫃內配電
供電架構	1.6-2.2 KW PSU	3.3 KW PSU	5.5 kW PSU電源架/匯流排配電	18.3 kW PSU電源架/匯流排配電
散熱技術	氣冷	氣冷/70%液冷	70%液冷/全液冷	全液冷
機構設計	19吋伺服器	HGX、MGX	MGX NVL72/托盤內無纜線	垂直Blade/機櫃密度更高

高功率：算力升級，電力密度與熱密度同步升級

- ◆晶片功率提高 → 單顆GPU/加速器更耗電，電流變大
- ◆功率密度提高 → IC及板端的耐壓、耐熱、可靠度要求上升
- ◆大電流產生更多熱 → 散熱必須轉向採用液冷
- ◆整櫃功率提升 → 需要更大功率的PSU、Power Shelf、Busbar、HVDC等供電架構

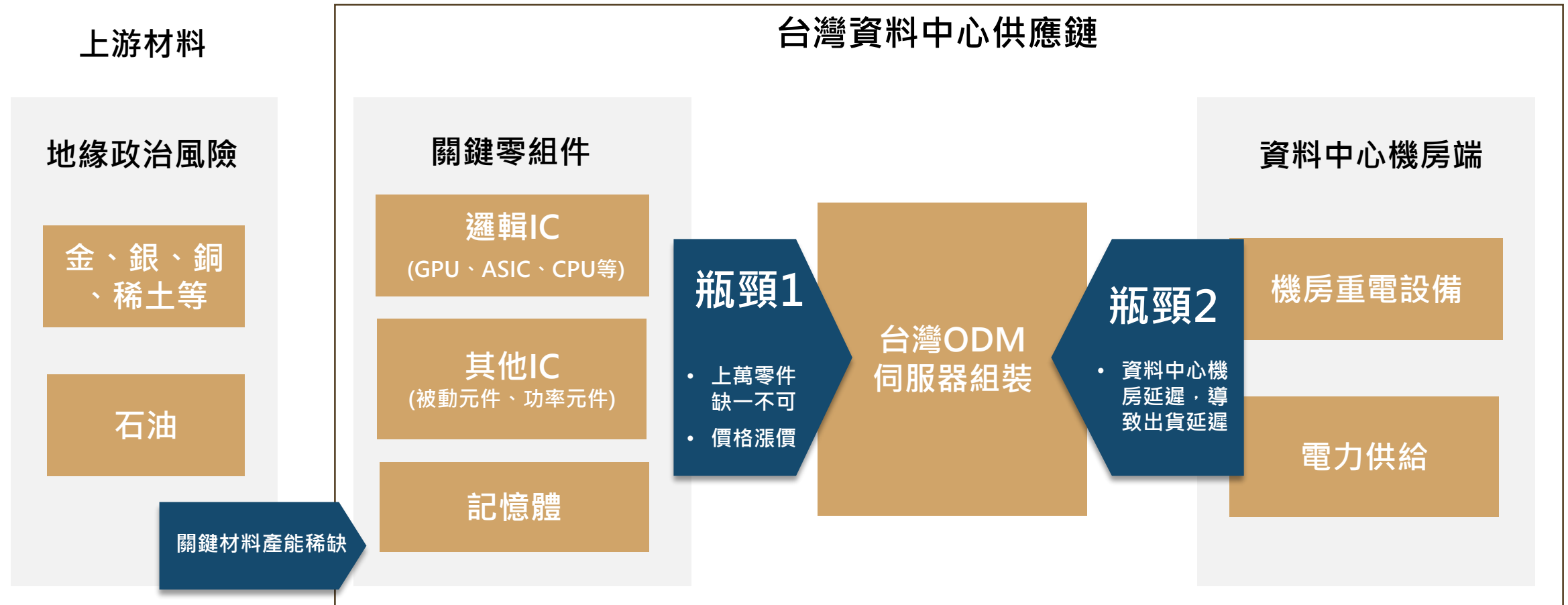
訊號高速化：打通AI資料傳輸瓶頸

- ◆GPU互連與資料吞吐需求提升，需要大頻寬
- ◆PCB朝大板、多層、高速低損耗發展
- ◆帶動板材、連接器、線材與高速傳輸零件升級

資料來源：各公司，MIC整理，2026年8月

台灣產業機會

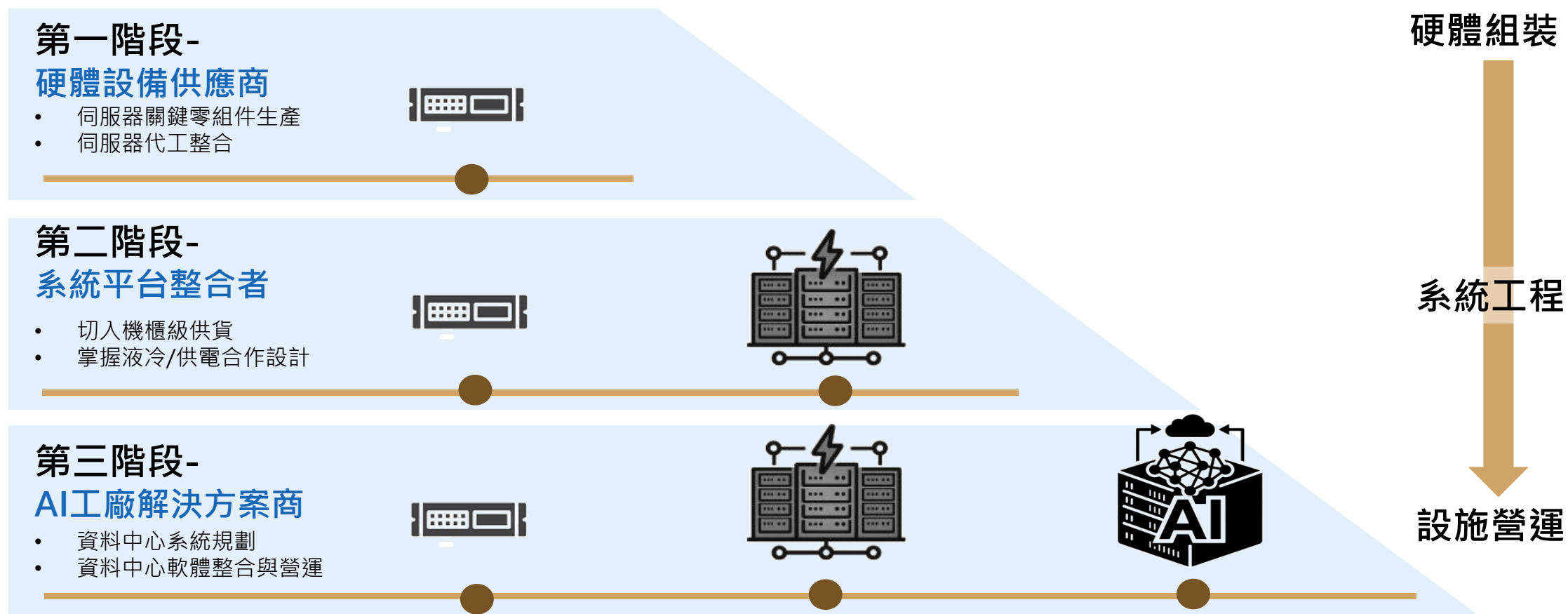
挑戰：市場需求不是問題 供應鏈供貨瓶頸影響AI DC落地時程



資料來源：MIC，2026年8月

- ◆ 台灣伺服器供應鏈部分，雲端服務商在2025年下半年開始加速擴增AI基礎設施，伺服器所需之記憶體、CPU、關鍵IC元件出現拉貨潮，導致關鍵零組件供給吃緊，價格隨之高漲
- ◆ 當瓶頸發生在資料中心端，伺服器出貨可能因為機房重電設備未到位及資料中心電力供給不及而延遲出貨

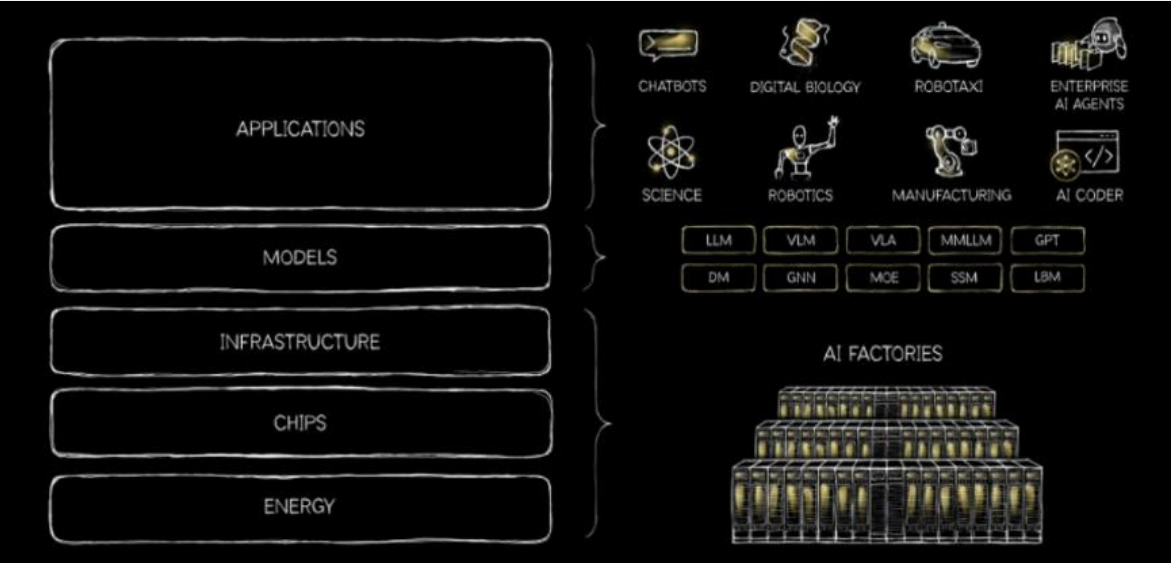
機會(1/2)：台灣伺服器產業轉型朝向AI DC整合方案布局



資料來源：華碩、技嘉，MIC整理，2026年8月

- ◆ 台灣伺服器業者的價值升級，正從既有的零組件供應與整機代工，延伸到液冷、供電、機櫃與資料中心基礎設施
- ◆ 從硬體出貨走向解決方案，切入AI DC規劃、整合相關軟體業者，並提供營運服務，擴大產業利基

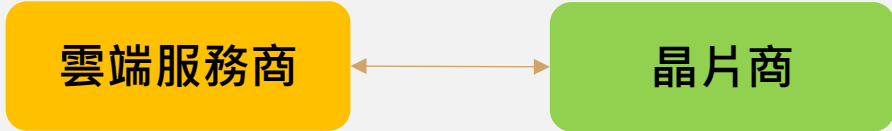
機會(2/2)：AI DC走向系統整合階段 台灣可切入硬體升級與設計



資料來源：NVIDIA，2026年3月

- ◆ AI平台競爭已從單點晶片效能，轉向晶片、網路互連、軟體堆疊與資料中心基礎設施的共同優化
- ◆ AI資料中心架構將成為影響AI部署速度、營運效益與擴充彈性之關鍵因素

資料中心架構分流



以自有運算任務及雲端服務需求為導向，進行系統垂直優化

以生態系為核心，推動跨客戶之通用平台系統

驅動因素

- AI部署密度提升，基礎設施成擴張瓶頸
- CSP自研晶片加速，平台架構分流擴大

競爭焦點

由單機硬體轉向系統整合

台灣伺服器產業升級機會

平台共設計夥伴

- 參與高密度AI基礎設施前期設計
- 切入機櫃、供電、散熱與維運架構設計

關鍵模組供應者

- 聚焦液冷、供電、機構與互連等高技術模組
- 在高功耗、高密度下建立不可替代的模組能力

全球製造與在地服務

- 因應各地部署需求，強化快速供貨與在地支援
- 提升台灣業者在國際大廠供應鏈中長期黏著度



資料來源：MIC，2026年8月

結論

結論

◆ AIDC競爭核心，已由單一伺服器轉向整體基礎設施

- 競爭焦點由AI晶片效能，轉向互連、軟體堆疊與資料中心基礎設施的整體系統優化
- AI基礎設施的設計能力，將直接影響AI模型部署速度、能源效率，及後續擴充彈性

◆ 伺服器產業型態，正由單機出貨轉向 Rack-scale與系統整合

- AI算力需求由訓練延伸到推論，推升高密度機櫃、液冷與模組化供電設計成為重點
- 市場價值也從加速器，延伸到液冷、供電、機構件、高速互連與硬體系統驗證

◆ 台灣伺服器產業應由製造導向，轉向高值化之共設計、整體解決方案

- 延伸既有製造優勢，切入平台設計、機櫃/供電/散熱架構設計及關鍵模組供應
- 同時台灣業者也積極切入各國AIDC市場，以硬體整合優勢提供AIDC整合方案，擴大產業覆蓋面



Thank you.

Your Gateway to Strategic Insights

魏傳虔 產業顧問兼副主任

chriswei@iii.org.tw

產業情報研究所

陳奕伶 資深產業分析師

ylchen@iii.org.tw

產業情報研究所

智慧財產權暨引用聲明

- ◆ 本活動所提供之講義內容或其他文件資料，均受著作權法之保護，非經資策會或其他相關權利人之事前書面同意，任何人不得以任何形式為重製、轉載、傳輸或其他任何商業用途之行為
- ◆ 本講義內容所引用之各公司名稱、商標與產品示意照片之所有權皆屬各公司所有
- ◆ 本講義全部或部分內容為資策會產業情報研究所整理及分析所得，由於產業變動快速，資策會並不保證本活動所使用之研究方法及研究成果於未來或其他狀況下仍具備正確性與完整性，請台端於引用時，務必注意發布日期、立論之假設及當時情境