Helios 出貨、Vera Rubin 量產、GW 級資料中心——OCP 2026 峰會的平台時程與台灣 ODM 位置

從 AMD、NVIDIA 與 hyperscaler 的一手說法，看機櫃平台時程、電力與可靠度瓶頸，以及台灣 ODM 從「組裝」走向「交付層級」的證據強度

報告日期：2026–10–19

資料基礎：OCP Global Summit 2026（2026–10 舉行）約 27 場與本主題直接相關的簡報與影片字幕（AMD、NVIDIA、Broadcom、Meta、Google、Microsoft、Oracle、McKinsey、SemiAnalysis、Cisco / Supermicro、鴻海鴻佰、緯穎、廣達 QCT、華擎、和碩、英業達等），並以 Morgan Stanley 出貨追蹤作交叉參照。

說明：本報告以原廠與講者自述為主；自測、模擬、預測與讀圖估值均標明口徑。台廠敘述逐項標示證據強度（實際製造數據 / 合作開發 / 展示 / 僅 logo / 推論）。部分場次僅有影片字幕，字幕辨識可能有誤，相關處已降級標註。本報告不構成投資建議。

摘要

七個核心結論

1. **時程更新是第一重點。**AMD 於主題演講稱 Helios 機櫃自 2026-09 起量產出貨，首批 OpenAI、Meta、Anthropic、Microsoft、Oracle 已安裝，Q4 至 2027 上半年放量；NVIDIA 稱 Vera Rubin 已 full production（字幕：每週數千櫃）；Rubin Ultra NVL576 CPO 為 2027。這些皆為原廠自述，ODM 端口徑較保守（和碩稱 2026Q4 量產就緒、2027Q1 全面量產），三方口徑須分層閱讀。
2. **平台設計走向開放、乙太與液冷。**Helios 為 72 GPU、雙寬 Open Rack Wide、乙太 scale-up；AMD 架構師稱這是「最後幾代全電氣 scale-up 之一」。Broadcom 稱乙太 scale-up 機櫃未來約 6 個月內多款量產。Agentic AI 另帶出 CPU 機櫃新品類（AMD Helion、NVIDIA Vera CPU 機櫃、Arm AGI CPU）。
3. **需求端瓶頸是電力與設備交期，不是晶片。**McKinsey 估 2025–2030 累計資本支出約 US$7.3T、美國 2030 年電力缺口約 15 GW、發電機交期約 100 週；Meta Prometheus 已運轉 1 GW 以上、Hyperion 規劃 5 GW；Open DC 規格 v0.7 定 2027–2028 基線。
4. **叢集可靠度成為新 KPI。**Meta 兩個匿名平台中，舊平台網路相關中斷合計約 48%、新平台記憶體 22.5%、GPU 本體僅 1–2%；暗示網路與 HBM，而非運算晶粒，是可靠度瓶頸。
5. **Token 經濟口徑不可混用。**NVIDIA 稱 Vera Rubin 對 GB300 每 MW 最高 30 倍、每百萬 token 成本最低 45 倍；SemiAnalysis 第三方預覽（預發行軟體）最大 7.19 倍，且隨互動性區間大幅變動。
6. **台灣 ODM 競爭力從「組裝」轉向「交付層級」**（L10 托盤→L11 機櫃→L12 叢集→L13 設施）。證據最硬的是鴻海（自述 VR 自動化 94%、測試 FPY 80%，對 GB 的 39%）與和碩（自述 L11 / L12 已部署）；緯穎、廣達、華擎多為展示；緯創、神達、仁寶僅在 ODM logo 名單。
7. **信心：時程高（有原廠一手）、ODM 份額中（多為展示，缺客戶與出貨量）。**主要觀察事件為 Helios Q4 出貨量與 ODM 分配、和碩 2027Q1 放量、Rubin Ultra 量產時點與 Open DC v1.0。

投資重點表

重點	內容	相關標的	信心
平台時程提前確認	Helios 已量產出貨、Vera Rubin 已 full production；放量集中 2026Q4–2027H1	AMD、NVIDIA；ODM 鴻海、緯穎、和碩、廣達	高（原廠）
交付層級拉長	L10→L13 成共通語言，良率與自動化成為 ODM 差異化指標	鴻海（2317）、和碩（4938）	中
CPU 機櫃新品類	Helion（2026-11）、Vera CPU 機櫃（稱已出貨）、Arm AGI CPU；ODM 名單廣但無量	緯穎、廣達、華擎、鴻海、神達、英業達	低–中
電力與交期瓶頸	設備交期 2–3 倍、美國缺口約 15 GW，部署節奏受電力而非晶片限制	電力與液冷設備鏈（本報告不展開）	中
可靠度新 KPI	網路與記憶體占中斷大宗，利好高可靠互連與測試能力	具 L10–L12 測試能力之 ODM	中（單一來源）
Token 經濟口徑	廠商 30x / 45x 對第三方最大 7.19x，需求推論不宜採廠商高標	—	高

第 1 章 平台時程總表

OCP 2026 是原廠把「預期」改寫成「已出貨」的一週。下表整理各平台時點、來源與性質。**性質欄是關鍵**：「量產」指原廠稱已開始出貨，並不等於 ODM 端已大量交付；「預測」為原廠或第三方的未來陳述。

項目	時點	來源（OCP 2026 場次）	性質
AMD Helios 機櫃	2026-07 發表；2026-09 起量產出貨；Q4 至 2027H1 放量	AMD 主題演講〈Building the Next Era of AI Infrastructure〉	量產（AMD 自述）
AMD Helion Agentic 機櫃	2026-11 供貨	同上	預告
AMD EPYC「Verano」	2027H1	同上	預測
NVIDIA Vera Rubin NVL72	簡報 full production；字幕稱每週數千櫃	NVIDIA 主題演講〈AI Factories at Scale〉（2026-10-12）	量產（NVIDIA 自述）
Vera Rubin（和碩口徑）	2026Q4 量產就緒、2027Q1 全面量產	和碩〈Scaling AI Factory for the Rubin Era〉（影片字幕）	公司自述預測
Rubin 量產起點（外部引用）	2026.9；Rubin Ultra 2027	鴻海鴻佰〈From Chips to Compute〉引 TrendForce	外部預測
Vera CPU MGX 機櫃	簡報標 Shipping now；字幕稱 Meta 2027 年初建首座叢集	NVIDIA 主題演講	首批出貨（自述）
Spectrum-6 CPO 乙太網	2026 年出貨	NVIDIA 主題演講（字幕）	預告
Vera Rubin Ultra NVL576 CPO	2027（字幕「next year」；簡報未標年）	NVIDIA 主題演講	預測
Microsoft Maia 300	2026-10 發表；2027 多 GW 規模	Microsoft 主題演講〈Intelligence Infrastructure〉（僅字幕）	發表 / 預測（字幕，降一級）
Google TPU 8t / 8i	2026	Google〈From Chip to Grid〉TPU 世代時間軸	Google 時間軸
乙太 scale-up 機櫃量產	未來約 6 個月（至約 2027-04）	Broadcom 主題演講〈Scaling AI Infrastructure Through Open Innovation〉	預測
Meta Prometheus / Hyperion	1 GW 以上已運轉 / 5 GW 未來數年	Meta 主題演講〈Concrete to Code〉	營運 / 預測
Open DC 規格 v0.7	基線 2027、目標 2028+；11 月起月會，v1.0 未給日期	〈Next-Generation ML Infrastructure: The Open DC Architecture〉	草案

1.1 NVIDIA 與 AMD 的「量產」，與 ODM 的「量產」不是同一件事

議題	說法 A	說法 B	判讀

Vera Rubin 量產定義	NVIDIA：full production、美國與台灣工廠托盤量產、每週數千櫃（字幕）	和碩：2026Q4 量產就緒、2027Q1 全面量產；鴻海引 TrendForce：Rubin 2026.9 起	前者偏晶片 / 托盤層，後者偏機櫃放量或自家產線。三者並列，不採單一年份。作為量級參照，Morgan Stanley 2026-10-12 的追蹤顯示 GB200 / GB300 NVL72 等效 9 月出貨約 6.5K 台（口徑為 Blackwell 世代、月度），與「每週數千櫃」量級相近但世代不同，不可直接相加。
Rubin Ultra 時點	和碩字幕把 VR 與 Rubin Ultra 同列 2026Q4 / 2027Q1	NVIDIA 字幕與鴻海引 TrendForce：2027	和碩該句疑為字幕誤轉寫，本報告以 2027 為準。
Helios 量產	AMD：2026-09 已出貨，首批客戶已安裝	HPE、緯穎場次僅展示，未給出貨日期或數量	可寫「已開始量產出貨」，不寫數量。
Vera CPU 機櫃	NVIDIA：Shipping now、單櫃 256 顆 CPU（字幕）	先前外部預期為 2026Q4 量產	「已出貨」可能指首批客戶樣品，放量仍視 2026Q4。

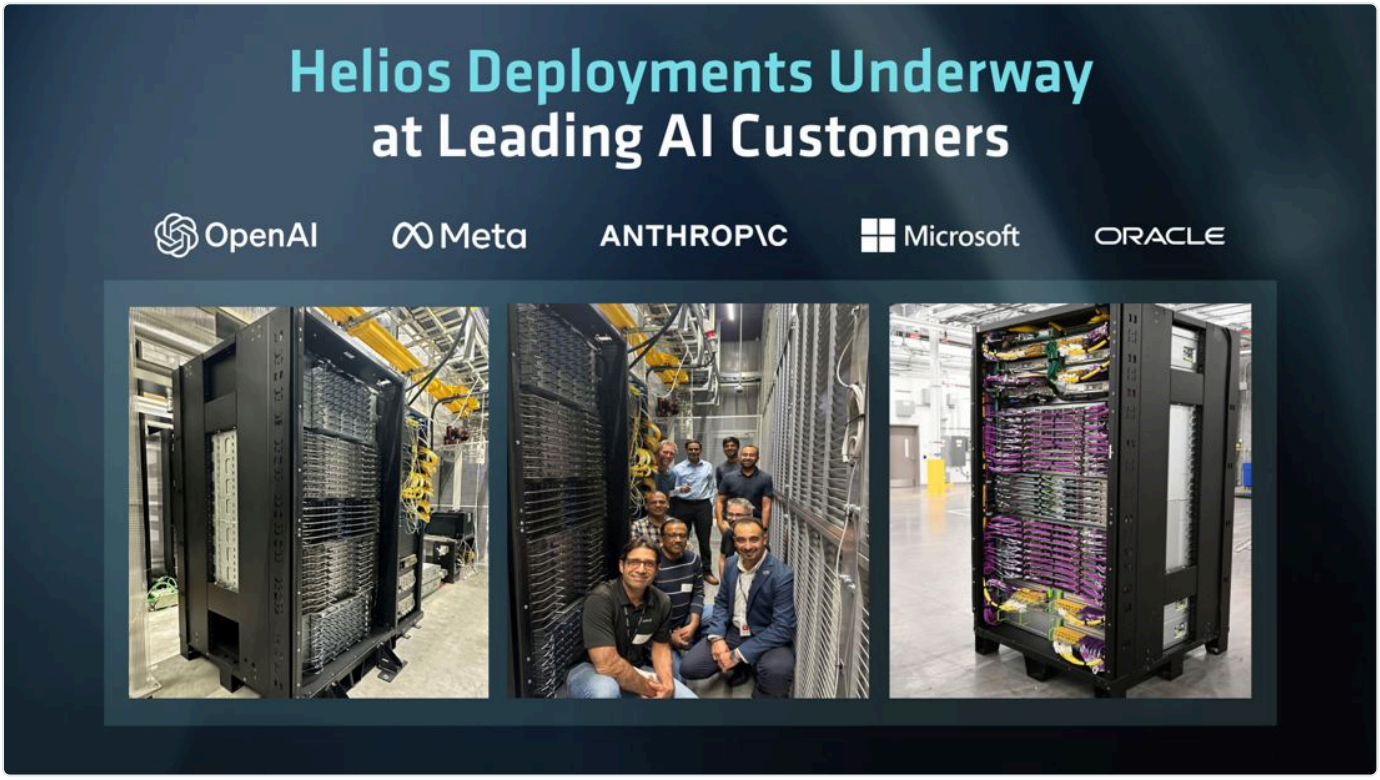


圖 1-1：AMD 公布的 Helios 部署頁。列出 OpenAI、Meta、Anthropic、Microsoft、Oracle 五家客戶標誌，下方為三張機房現場機櫃照。這是 AMD 的自述，未提供數量。來源：OCP 2026，AMD 主題演講〈Building the Next Era of AI Infrastructure〉。

第 2 章 AMD Helios 與 Helion：設計細節、網路、散熱與 ODM

2.1 規格與設計邏輯

Helios 是把 72 顆 MI455X 視為單一加速器的機櫃級產品。AMD 公布的機櫃規格為：FP4 2.9 EF、FP8 1.48 EF、HBM4 容量 31 TB、HBM4 頻寬 1.7 PB/s、scale-up 頻寬 260 TB/s、scale-out 頻寬 43 TB/s（均為 AMD 規格口徑）；字幕補充整櫃重量近 7,000 磅、單一運算托盤約 160 磅。AMD 另稱 tokens / 美元最多高 30%、DeepSeek R1 MLPerf 單 GPU 17,000 tokens/s（AMD 自測、offline 設定），兩者皆為廠商口徑。

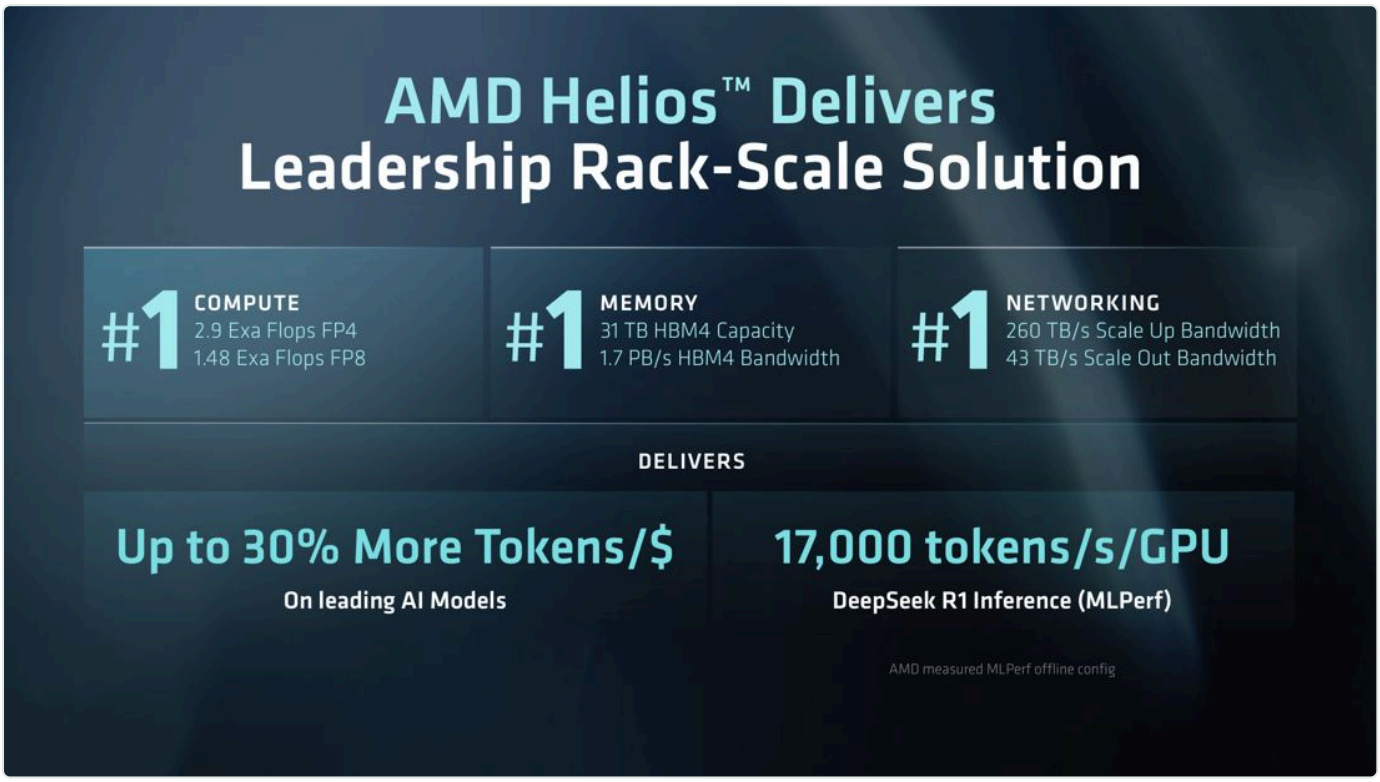


圖 2-1：Helios 機櫃級規格卡。左至右為運算、記憶體、網路三組規格，下方為 AMD 自稱的 tokens / 美元與 MLPerf 數字（註明為 AMD 自測 offline 設定）。來源：OCP 2026，AMD 主題演講。

設計取捨方面，AMD 在 Helios 設計訪談中說明：高頻寬 load/store 直連要求高密度，因此走液冷與盲插（運算、交換托盤滑入即接通電、冷卻與通訊）；選擇 Open Rack Wide 是因為額外寬度帶來更多前面板網路埠、內部空間並降低樓板載重；客戶可選電源電壓、水流與水溫（無冰水機 40–45°C）、電源冗餘（N+N 或 4 取 3）與是否加超級電容。AMD 並稱整套規格已公開一年多，方便供應鏈（連接器、電源、冷卻）預先建立，客戶可自選 ODM。

2.2 乙太 scale-up：已從概念進入量產

Helios 的 scale-up 以乙太承載 AMD 的 UALink 式 load/store（UALoE）：簡報顯示 18 個運算托盤、72 顆 MI455X 經 12 顆 Tomahawk 6 級（簡報寫 TH6，512-lane）交換器互連，每 GPU scale-up 頻寬 1.8 TB/s 單向。AMD 架構師（Steve Scott）稱這是「最後幾代全電氣 scale-up 之一」，之後會導入光學、擴大 scale-up 域；但訪談未給年份，也未提 MI500。Broadcom 在主題演講稱 5 家以上 hyperscaler 已在跑乙太 scale-up，機櫃內銅背板現世代 1.04 Pbps（72 XPU、約 250 kW、TH6 量產中），次世代 2.08 Pbps 僅標「Possible」；跨櫃改用 OCI 光學（200G 連結 7 pJ/bit、500 m，對照可插拔 23 pJ/bit）。以上為廠商說法，量產性仍待 ODM 出貨驗證。

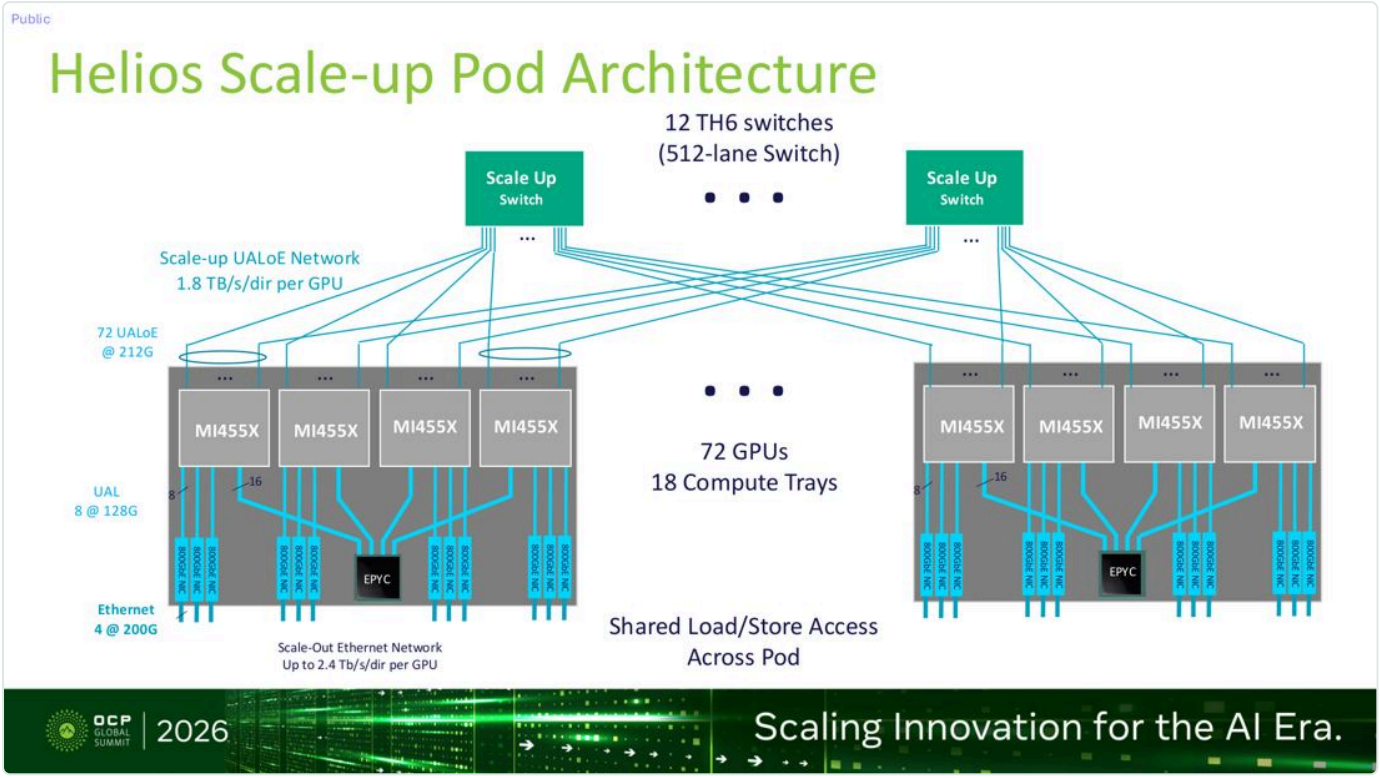


圖 2-2：Helios scale-up pod 拓撲。72 顆 MI455X 分布在 18 個運算托盤，經多顆 scale-up 交換器互連，每個托盤另有 EPYC 與 800GbE NIC 走 scale-out。注意：本頁標示每 GPU scale-out 最高 2.4 Tb/s，與同一簡報他頁（1.6T plane 組網）數字不一致，本報告不引用每 GPU 網路頻寬，僅採 AMD 機櫃規格卡。來源：OCP 2026，AMD / HPE〈Building AI Rack System with Open Standards〉。

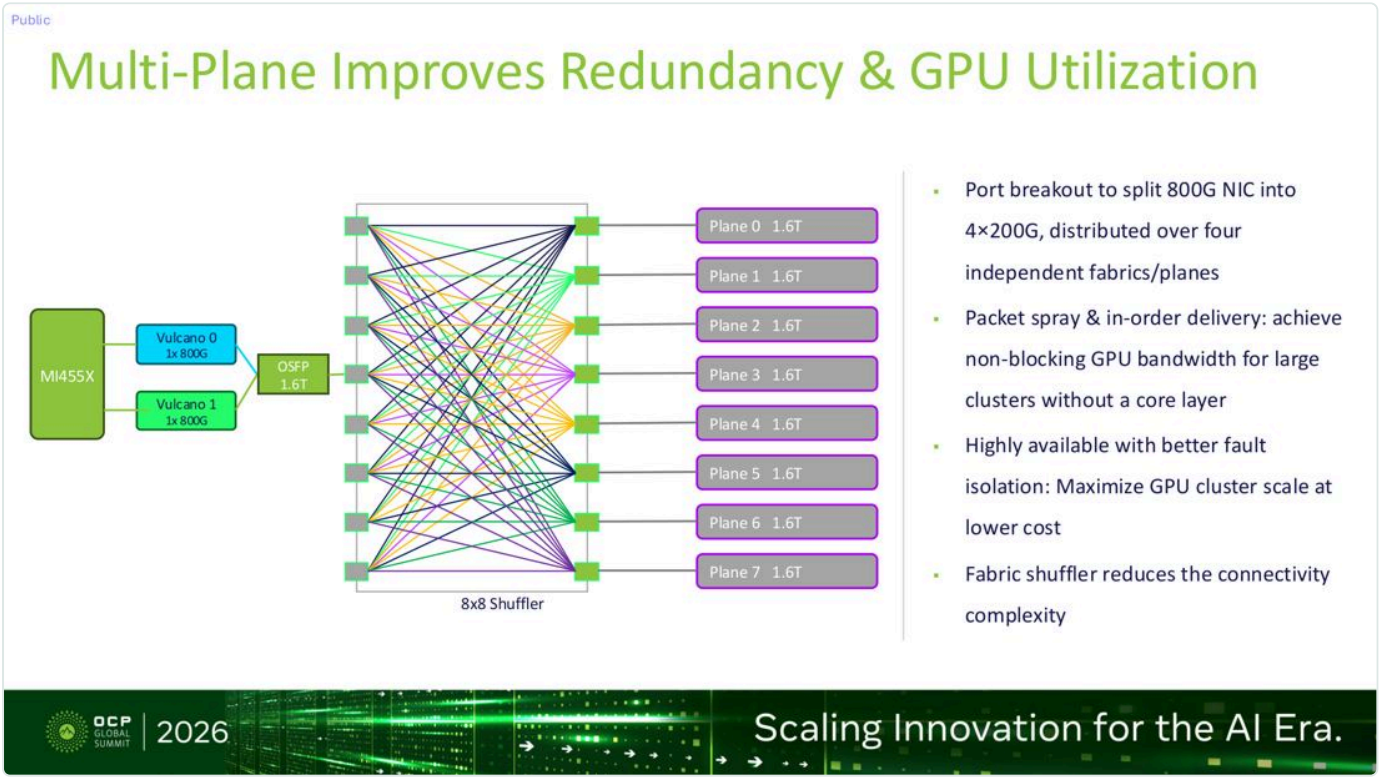


圖 2-3：Helios 多平面 scale-out 組網。每顆 MI455X 接兩張 Vulcano 800G NIC，經 1.6T OSFP 與 8x8 shuffler 分散到 8 個 1.6T plane，以 packet spray 提升冗餘與 GPU 利用率。來源：OCP 2026，AMD / HPE。

2.3 散熱：Helios 並非全液冷

nVent 在〈RDHX cooling for AMD Helios〉中指出，Helios 單櫃約 246–250 kW，約 89% 熱量由液冷帶走，其餘約 11%（約 27 kW）仍排入空氣，需背門熱交換器（RDHX）處理；AMD 與 nVent 對應的 RDHX 規

格為雙開門、48 VDC、可排 25–45 kW。因此 Helios 是「液冷＋背門」混合設計，機房端仍需空氣側熱管理。這是 nVent 的產品說明，nVent 自身有銷售誘因，數字以其簡報為準。

2.4 Helion：CPU 為主的 Agentic 機櫃

AMD 同日發表 Helion Agentic AI Rack（ORv3 44OU，最多 32,768 顆 Zen 6 核心、最多 16 張 MI350P、最多 102 Tbps 網路；字幕曾寫 32 張 MI350P，以簡報 16 張為準），2026–11 供貨並貢獻 OCP。簡報供貨生態 logo 含 Dell、HPE、Lenovo、Supermicro、Cisco，以及台灣的緯穎、廣達、鴻海、神達、英業達、華碩、華擎、技嘉——**僅為 logo，未載明各家角色、訂單或數量**。



圖 2–4：Helion Agentic Rack 與供貨生態 logo。標示 2026 年 11 月供貨，下方 13 家系統商標誌。僅為 logo，不代表訂單。來源：OCP 2026，AMD 主題演講。

第 3 章 NVIDIA Vera Rubin、Rubin Ultra 與 MGX 標準化

3.1 Vera Rubin 與製造性

NVIDIA 在主題演講中稱，Vera Rubin NVL72 對 GB300 NVL72 每 MW token 吞吐 2x / 10x / 最高 30x（隨互動性而異）、每百萬 token 成本最低降 45x（NVIDIA 自述，基準為 DeepSeek-V4-Pro 與 SemiAnalysis AgentX 資料，取點未明列；與第三方數字的比較見第 7 章）。製造性方面，GB200 運算托盤組裝約 90 分鐘，Vera Rubin 托盤約 1 分鐘；美國與台灣工廠已量產托盤（NVIDIA 自述）。

3.2 MGX 1.2：ODM 的共同底座

NVIDIA 將第三代 MGX 規格（1.2）貢獻 OCP：無線纜、無水管、100% 液冷的運算托盤；5,000 A 液冷匯流排；45°C 液冷盲插歧管；latch 可承受 6,000 磅配合力；機櫃加 100 磅鋼補強。MGX 的意義在於把機櫃機構與液冷介面標準化，降低 ODM 各自設計的成本，同時也降低 ODM 之間的差異化空間。

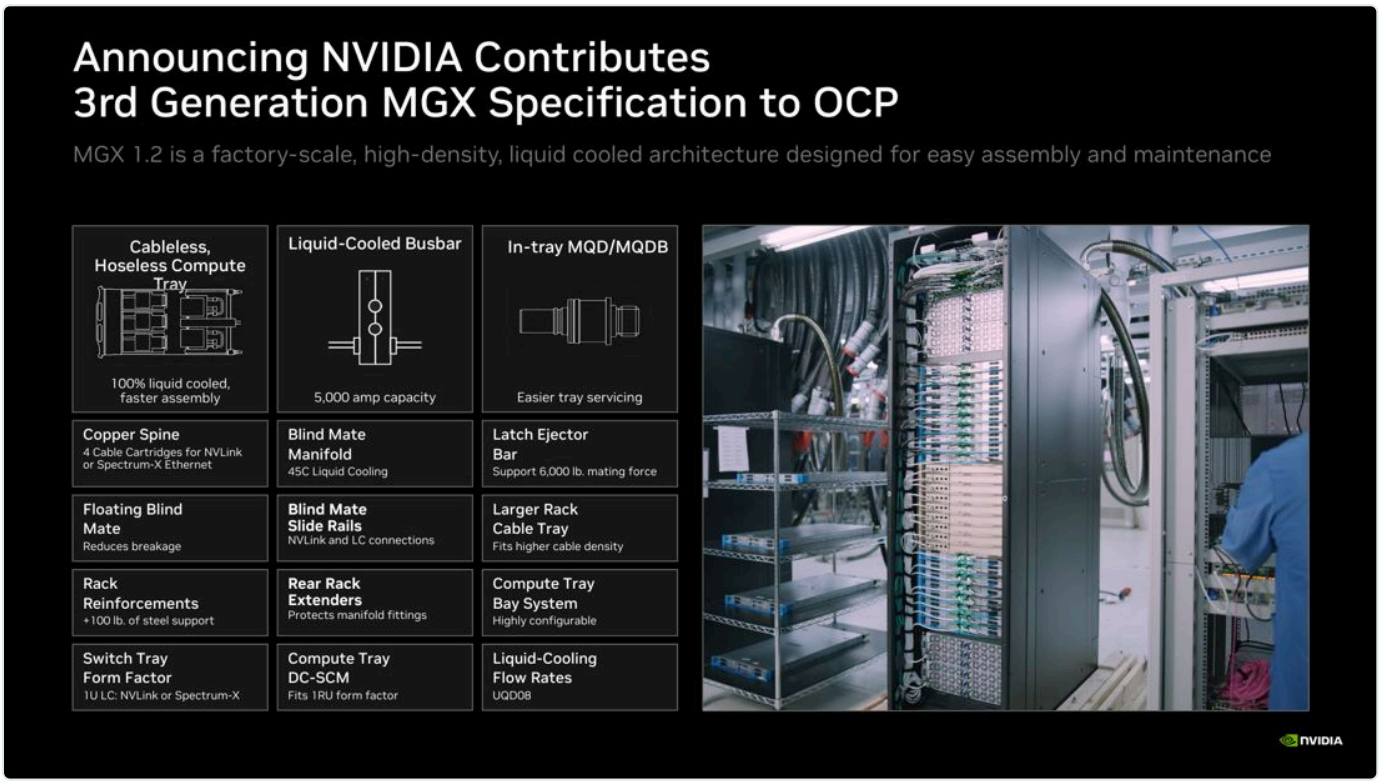


圖 3-1：MGX 1.2（第三代）規格要點與機櫃實景。左為十二項設計重點（無線纜運算托盤、液冷匯流排 5,000 amp、盲插歧管 45C、latch 6,000 lb 等），右為機房內機櫃現場照。來源：OCP 2026，NVIDIA 主題演講〈AI Factories at Scale〉。

3.3 Vera CPU 機櫃

NVIDIA 同時推出 Vera CPU MGX 機櫃（簡報標 Shipping now，字幕稱單櫃 256 顆 CPU、1U 液冷托盤、搭 BlueField 管理 agent sandbox）。ODM / OEM logo 共 19 家，含鴻海、仁寶、和碩、緯創、緯穎、廣達、英業達、神達、華碩、華擎、技嘉、微星，亦為僅 logo。與 AMD Helion 一同出現，顯示 agentic 工作負載把 CPU 機櫃推成獨立品類；但規模尚無數字（Vera CPU 單櫃 256 顆與 Helion 單櫃 32,768 核的口徑也未比對）。



圖 3–2：Vera CPU MGX 機櫃與合作系統商 logo。左為機櫃（標 Shipping now），中為無線纜全液冷托盤，右為四種可客製托盤，下方為合作系統商標誌。僅為 logo。來源：OCP 2026，NVIDIA 主題演講。

3.4 Rubin Ultra NVL576 CPO：2027

Vera Rubin Ultra NVL576 為 576 GPU 單一 NVLink 域、NVLink 7 CPO 交換托盤。簡報圖為八座機櫃以光纖在頂部串接，未標示 Kyber；字幕稱「next year」，鴻海引 TrendForce 亦為 2027。此為預測，且僅一場提及。



圖 3–3：Vera Rubin Ultra NVL576 CPO。左為運算托盤與 NVLink 7 CPO 交換托盤，右為八座機櫃經頂部黃色光纖串接的整體配置。簡報未標示量產年份。來源：OCP 2026，NVIDIA 主題演講。

3.5 800 VDC：設施先行、機櫃分步

NVIDIA 稱 Visionbay.ai 正在建置首座 800 VDC AI 工廠（16 MW 部署單元，簡報未給完工日）；和碩則稱設施端最快 2027 年中可部署，VR 與 Rubin Ultra 以 sidecar 為主。三者屬不同層級（設施電壓、機櫃電壓、Kyber 機櫃），本報告不合併為單一年份。

第 4 章 自研晶片與客製機櫃

Microsoft Maia 300：Microsoft 主題演講（僅有影片字幕，簡報檔被誤置，故本章數字降一級信心）稱 Maia 300 於 2026–10 發表，推論效能世代對世代最高 7 倍，Maia 系統 2027 年擴至多 GW 規模；Maia 200 為現行世代。此與市場先前對 Maia 300 2026 年即有數十萬顆的預估有出入，宜視為舊預估被時程推遲。Microsoft 同場另提 MRC（多平面乙太）可用兩層網路擴至 10 萬顆以上 GPU、Mt Diablo 電力架構（±400V 或 800V）與第三代熱交換單元。

Google TPU：Google 的 TPU 世代時間軸顯示 TPU 8t 與 8i 於 2026 年（僅一行，無規格）。Google 另在晶片到電網的場次主張 LVDC（0–800V）作為過渡路徑。

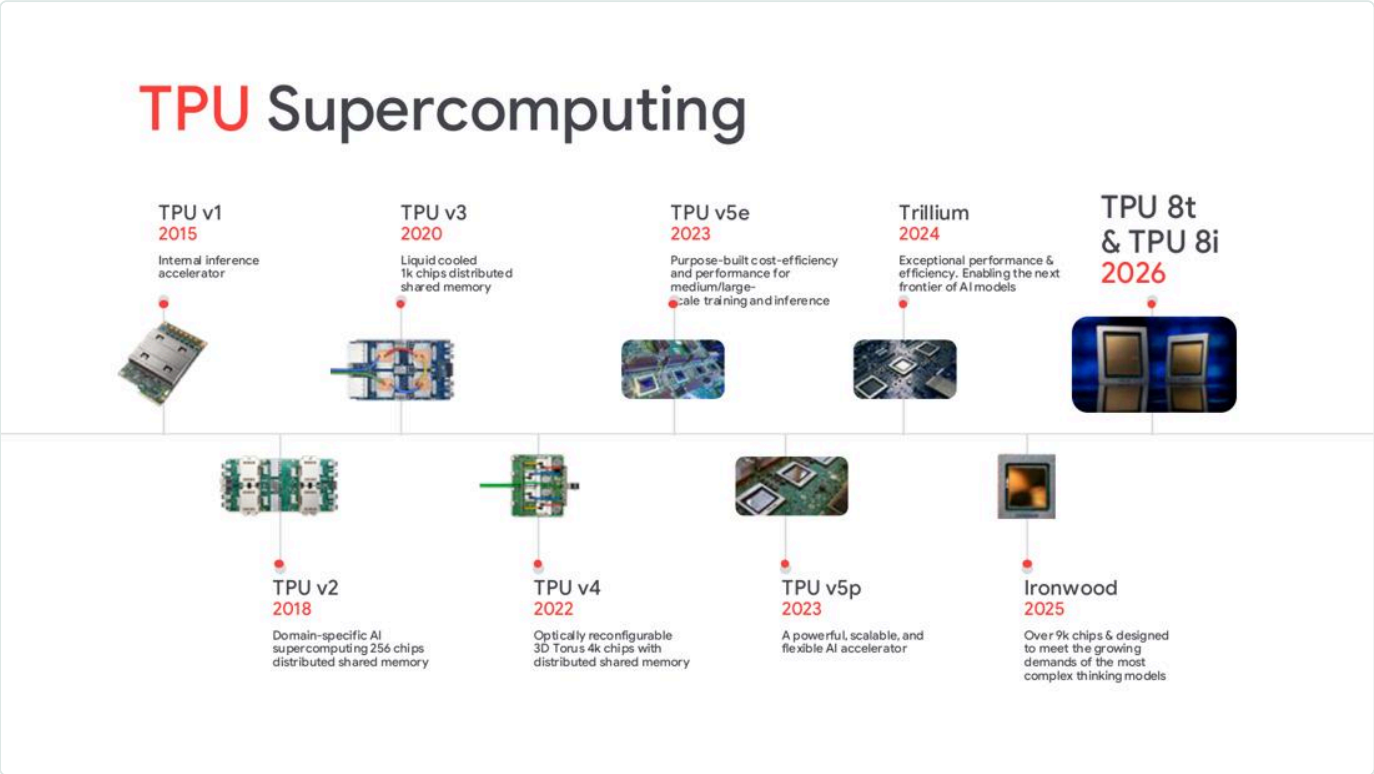


圖 4–1：Google TPU 世代時間軸。由 2015 年 v1 至 2025 年 Ironwood（逾 9,000 顆晶片組成），最右為 2026 年 TPU 8t 與 8i。來源：OCP 2026，Google〈From Chip to Grid〉。

Meta 自研晶片缺席：Meta 兩場演講完全未提 MTIA 或自研晶片，本報告不就此推論。Meta 的叢集以 NVIDIA 與 AMD 平台為主，並採用 Spectrum–6 102.4T 交換器；Meta 同時是 Helios 共同工程夥伴。

Astera Labs 開放推論機櫃：Astera 提出以 PCIe 為基礎的 Orion 機櫃（Scorpio X–Series 320 lane 交換器），讓客戶自帶 XPU，單跳可連至 160 GPU；其自測稱 8 GPU 點對點相較兩顆小交換器可多 50% tokens/s（Astera 自測）。此路線與 NVLink、UALink、乙太三條路線並列，且晶片級供應商並非台廠。

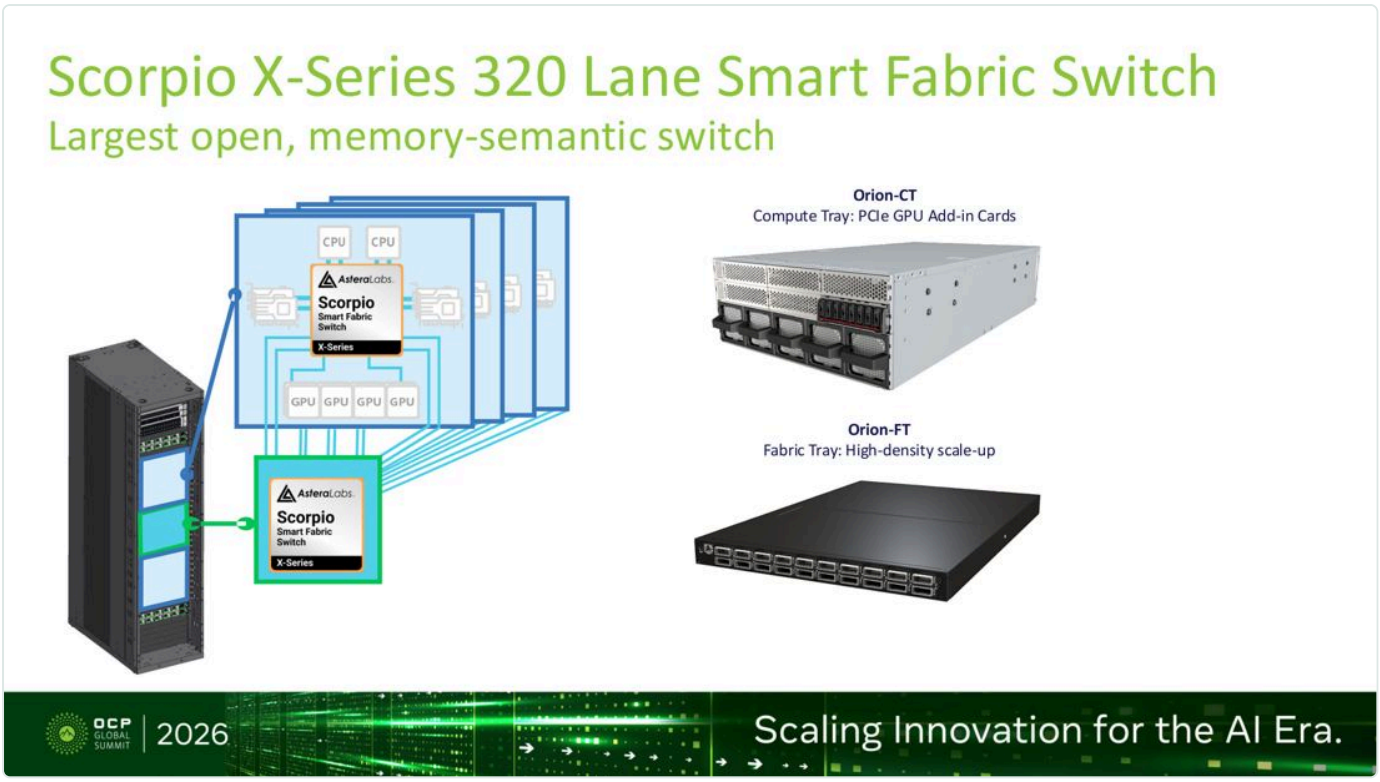


圖 4–2：Astera Labs Scorpio X-Series 與 Orion 機櫃。左為機櫃與交換器連接示意（運算托盤內 CPU＋GPU，與中間的 fabric 托盤），右為 Orion–CT 運算托盤與 Orion–FT 組構托盤實物。來源：OCP 2026，Astera Labs〈Bring Your Own XPU〉。

第 5 章 GW 級資料中心：Meta、Open DC 規格與需求瓶頸

5.1 Meta：Prometheus 與 Hyperion

Meta 稱 Prometheus 為 1 GW 以上叢集且已運轉，採含帳篷式廠房的多棟配置，建築間以光纖連接（簡報標示距離自 321 m 至 3,800 m），並以 Backend Aggregation（BAG）乙太 super spine 彙整；Hyperion 規劃未來數年達 5 GW。字幕曾寫 Prometheus 為「數萬顆 GPU」，與 1 GW 規模不符，疑為誤轉寫，GPU 數未給，不推算。

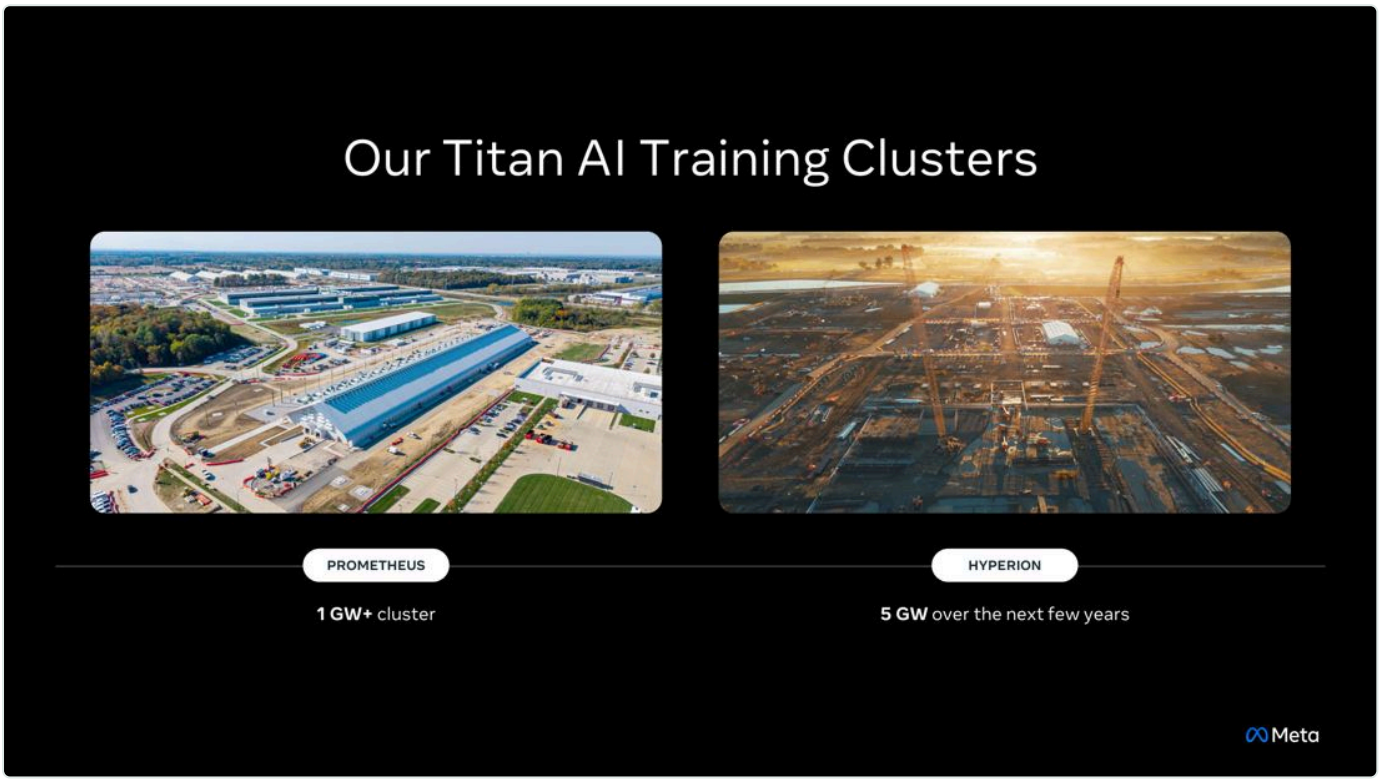


圖 5-1：Meta Titan AI 訓練叢集。左為 Prometheus（標 1 GW+ cluster）空拍，可見長條形帳篷式廠房；右為 Hyperion（標 5 GW over the next few years）施工空拍。來源：OCP 2026，Meta 主題演講〈Concrete to Code〉。

5.2 Open DC：2027–2028 基線

Google、Meta、AMD、NVIDIA 共同提出 Open DC 規格（v0.7 草案）：2027 年基線 Peak IT 密度 125 kW/ft、叢集 12 MW 下限；2028+ 目標 350 kW/ft（液冷 335）、叢集 24 MW 偏好；機櫃重量上限 8,000 磅。單位為每呎列長，與 Arista 創辦人 Bechtolsheim 以面積計的 10 kW/sqft 超密度構想（單一來源、講者自行估算）基準不同，方向一致但不可直接比較。

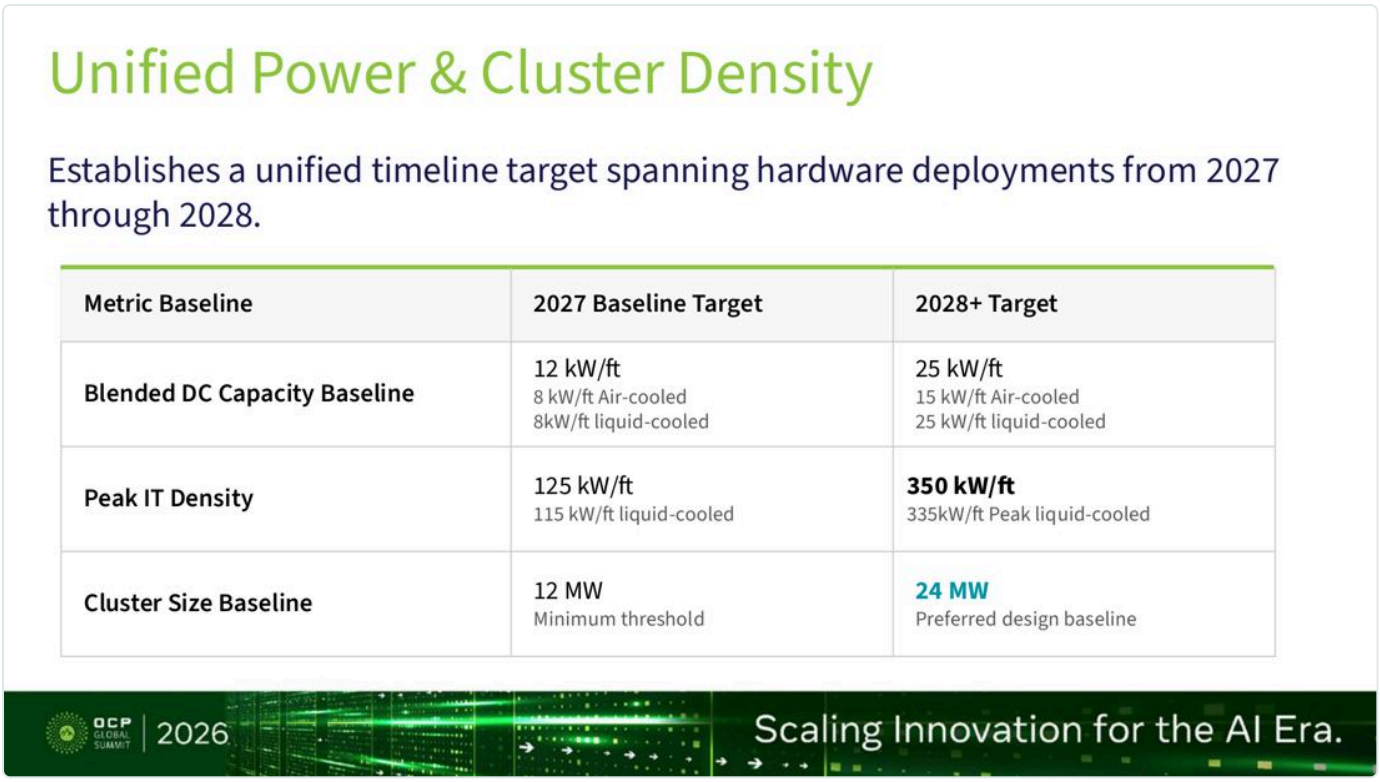


圖 5–2：Open DC 密度與叢集規模基線表。混合容量 12→25 kW/ft、Peak IT 125→350 kW/ft、叢集 12 MW→24 MW（2027 基線對 2028+ 目標）。為草案，v1.0 未給日期。來源：OCP 2026，Google / AMD / Meta / NVIDIA〈Next-Generation ML Infrastructure: The Open DC Architecture〉。

5.3 McKinsey：資本支出、缺口與交期

McKinsey 預估 2030 年全球資料中心 IT 算力需求在保守 / 基準 / 加速情境為 171 / 219 / 298 GW；2025–2030 累計資本支出約 US$7.3T（設計與製造約 4.4T，其中伺服器 3.5T；能源約 1.3T；建造約 1.0T；軟體約 0.6T，口徑為全價值鏈、不含營運商自身資本支出）。美國 2030 年需求 121 GW、供給 106 GW，缺口約 15 GW。此為顧問模型，**不可與券商對單一業者的資本支出（如 Meta 的年度 capex）並列比較**。McKinsey 並稱 OCP 標準化可縮短 3–6 個月，屬單一來源說法。

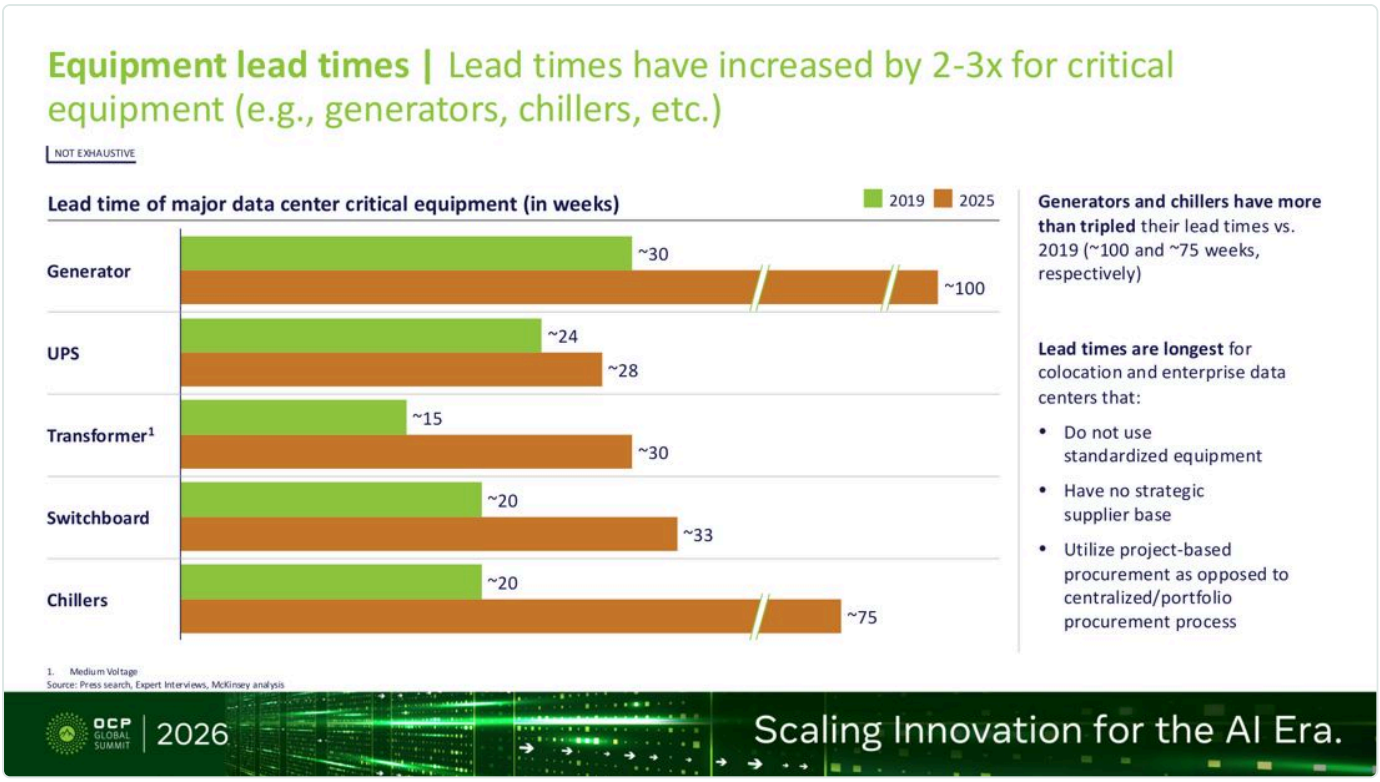


圖 5-3：關鍵設備交期，2019 年對 2025 年（週）。發電機約 30→100、冰水機約 20→75、變壓器（中壓）約 15→30、配電盤約 20→33、UPS 約 24→28。標示為「非窮盡」，來源為新聞搜尋、專家訪談與 McKinsey 分析。來源：OCP 2026，McKinsey〈From Compute Demand to Capacity Delivered〉。

第 6 章 叢集可靠度與 L10–L12 製造測試

6.1 Meta：中斷歸因

Meta 以兩個匿名平台（Platform–A、Platform–B，身分未具名，不可推斷為 GB200 或 Rubin）公開 2026 年的訓練中斷歸因。舊平台的橫向（scale–out）網路 23.6% 與縱向（scale–up）網路 24.2%，合計約 48%；新平台記憶體 22.5%（HBM 不可校正錯誤率為舊平台 4.8 倍）；GPU 本體與韌體僅 2.4% / 1.4%；軟體、工具與 pre–run 檢查等非硬體項合計約 26% / 34%。2026 年每週中斷率相對起始基準下降 75%（A）與 87%（B）（Meta 自述，圖讀終點約 25% / 13%；A 在窗口前已走過一段改善週期，B 為新平台，兩者不宜比較改善幅度）。

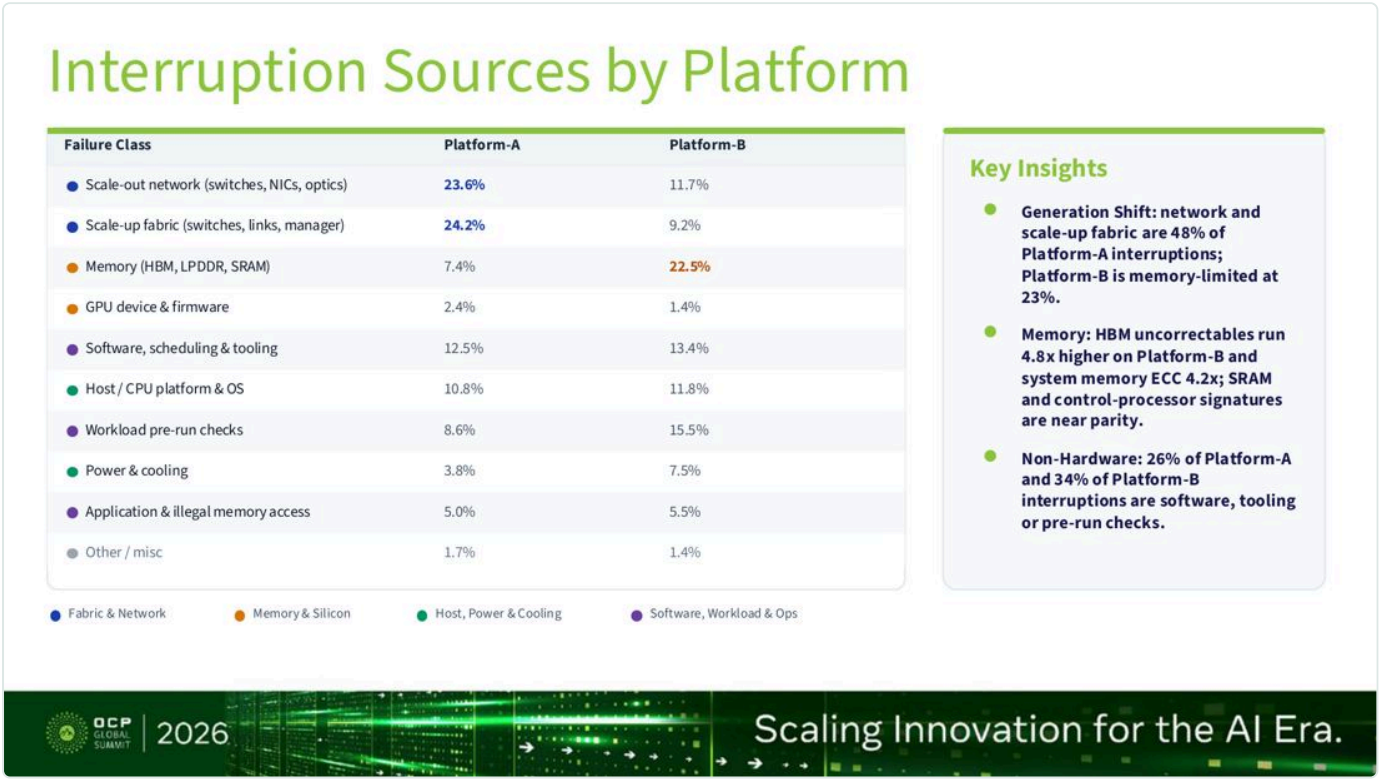


圖 6–1：兩個平台的中斷來源占比。十類來源在 Platform–A 與 Platform–B 的百分比，右側為 Meta 的關鍵洞察。來源：OCP 2026，Meta〈Achieving AI Cluster Reliability at Scale〉。

含義（推論）：可靠度瓶頸在互連與記憶體，而不是運算晶粒。對 ODM 而言，L10–L12 測試能否在出廠前攔截網路與記憶體缺陷，將直接影響客戶在叢集端的中斷率，這是「交付層級」價值的來源，但 Meta 並未將中斷歸因到任何 ODM。

6.2 交付層級：L10 到 L13

三個場次共用同一套語言：L10＝運算托盤，L11＝整櫃，L12＝叢集，和碩另延伸至 L13＝設施（含電力、液冷、空調與平面規劃）。



圖 6-2：鴻海 GB 對 VR 的自動化指標。自動化覆蓋 94%（對 56%）、NPI 前置 3.5 個月（對 4.5 個月）、組裝 FPY 97%（對 91%）、測試 FPY 80%（對 39%），左右為產線與機器人實景。均為鴻海自述，客戶未具名。來源：OCP 2026，鴻海鴻佰〈From Chips to Compute: Delivering AI Infra at GW Speed〉。

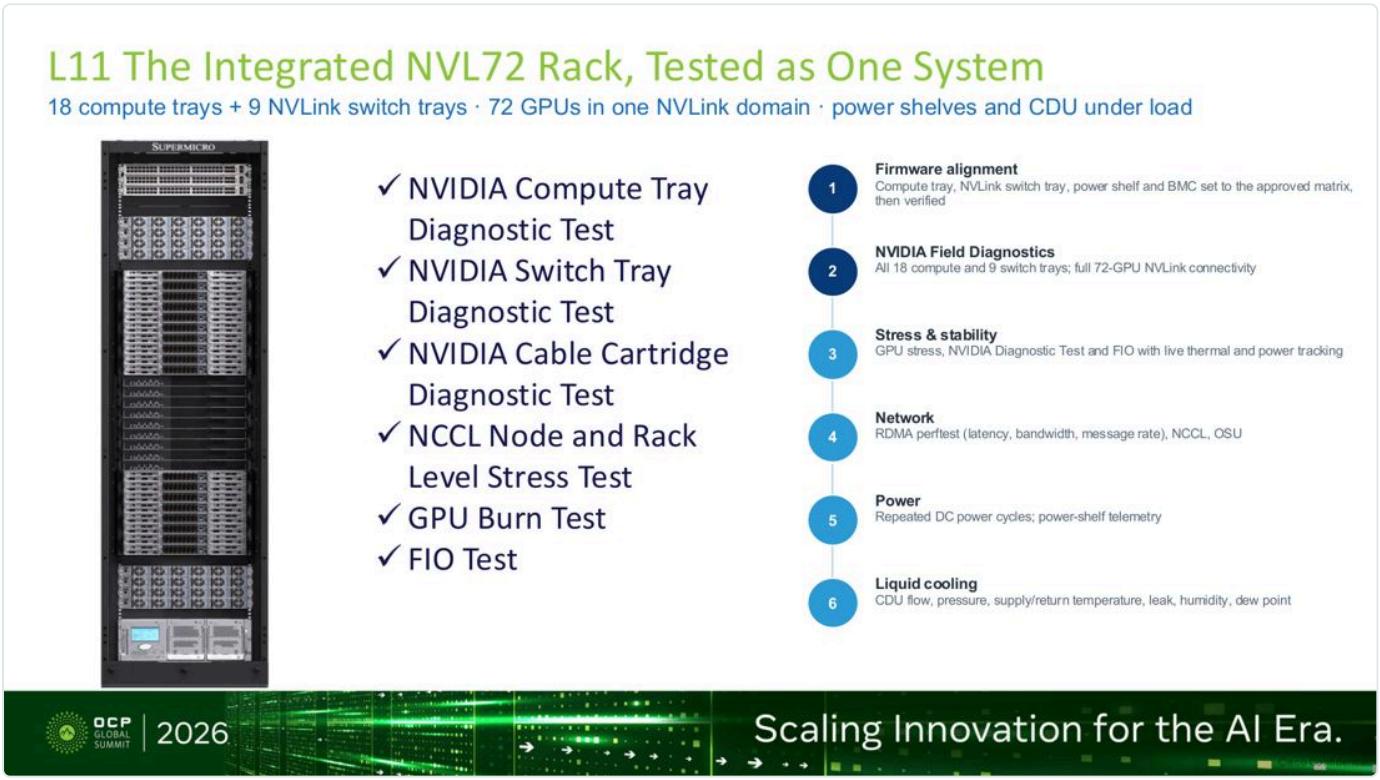


圖 6-3：Supermicro 的 L11 整櫃測試六步驟。韌體對齊、現場診斷、壓力穩定、網路、電源循環、液冷（CDU 流量、壓力、溫度、漏液、露點）。此為範例流程，非台廠。來源：OCP 2026，Cisco / Supermicro〈From Tray to Training Job: L10–L12 Manufacturing Test〉。

指標	鴻海（自述）	和碩（自述，字幕）	Supermicro（範例）

自動化	VR 94%（GB 56%）；L10 測試自動化 100%	GB300 約 50% 以上；VR 目標 90%+（目標，未驗證）；自動線 18–20 台 / 小時對人工 6–7	—
良率	VR 組裝 FPY 97%、測試 FPY 80%（GB 為 91%、39%）	GB300 L10 通過率 75%、L11 90%+；VR 目標 95%（目標）	L11 範例報告 18 台托盤通過 100%，總時長 4 小時 3 分（樣本）
規模	全球 1,000+ 測試位、100+ MW；單台 L10 測試 12 小時	2026 L11 / L12 已部署；2027 起 L13	—

口徑警示：鴻海為 FPY（一次通過率），和碩為通過率，Supermicro 為單一範例報告，三者定義不同，僅鴻海與和碩可概略並列，Supermicro 不納入比較。

鴻海另稱，前三大 CSP（未具名）的 VR 專案中，六道關卡共約 39 天，NPI 開發縮短 17 週。其引用 TrendForce 前五大 CSP 每年新增 AI 算力 2025 年 8 GW、2026E 18 GW、2027E 26 GW（口徑為 TrendForce 估算，單位定義待確認）。

第 7 章 Tokenomics：廠商說法與第三方

NVIDIA 與 SemiAnalysis 在同一週給出不同尺度的數字，原因是取點與軟體成熟度不同。

口徑	數字	條件
NVIDIA 自述（主題演講、Tokenomics）	每 MW token 吞吐 2x / 10x / 最高 30x；每百萬 token 成本最低 45x	Vera Rubin NVL72 對 GB300 NVL72；隨互動性（TPS/user）變動；取點與基準模型未於簡報明列；成本 45x 為曲線最大差距處
SemiAnalysis 預覽（第三方）	75 tok/s/user 1.45x；100 為 2.09x；125 為 4.29x；150 為 7.19x；170 為 5.56x；200 為 2.72x	DeepSeek V4 Pro 1.6T、agentic 負載；對「最佳 GB300 引擎」；Rubin 為預發行軟體、結果可能改變；模型化全含功耗 VR200 3.3 kW、GB300 2.12 kW

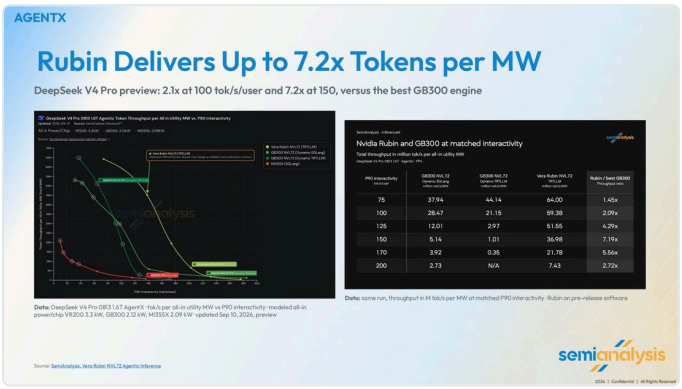


圖 7-1：SemiAnalysis 的 Rubin 對 GB300 每 MW token 比較。左為曲線，右為各互動性下的倍數表（最大 7.19x@150）。來源：OCP 2026，SemiAnalysis〈From Silicon to Services〉。

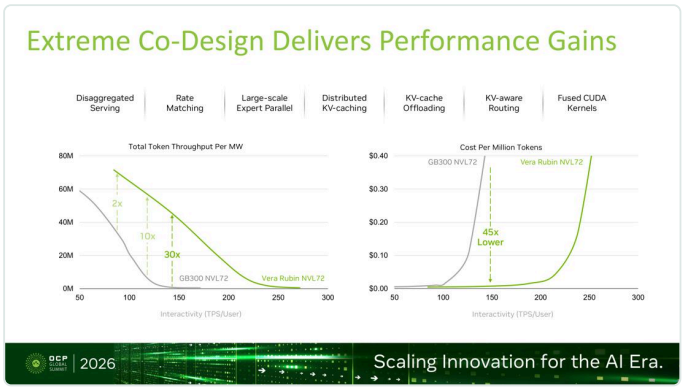


圖 7-2：NVIDIA 的每 MW token 與每百萬 token 成本曲線。標示 2x / 10x / 30x 與 45x lower。來源：OCP 2026，NVIDIA〈AI Tokenomics Explained〉。

判讀：兩者方向一致（Rubin 顯著優於 GB300），但尺度差距大，且倍數在不同互動區間劇烈變動（7.19x 後在 200 tok/s/user 降至 2.72x），不可把 30x 或 45x 當作平均改善。NVIDIA 的 45x 引用同一 SemiAnalysis 基準，但取點與軟體優化不同。對台股而言，需求推估宜以第三方區間為保守情境。SemiAnalysis 同場的單晶片 TCO（SemiAnalysis 口徑）：GB300 約 US$2.31 / 小時、B300 2.26、MI355X 1.5、TPU7 1.03，僅供口徑參考。

第 8 章 台灣 ODM 逐家證據表

證據強度分級：實際製造數據＝公司自述的產線或測試數據（客戶多未具名，非外部驗證）；合作開發＝與原廠或夥伴共同開發並有技術細節；展示＝展位或簡報展出產品，無客戶、時程或數量；僅 logo＝只出現在夥伴名單；推論＝本報告判斷，非來源明說。

公司	本次峰會揭露	證據強度	不能推論的事
鴻海（2317）/ 鴻佰	VR NVL72 製造與測試：自動化 94%、組裝 FPY 97%、測試 FPY 80%、L10 全自動、全球 1,000+ 測試位；前三大 CSP 專案 NPI 縮短 17 週；Helion、Vera MGX logo	實際製造數據（自述）；Helion 僅 logo	客戶未具名，不能對應到特定訂單或市占；數據為公司自述
和碩（4938）	L11 / L12 已部署；2026 超過 10,000 XPU；2027 確認訂單逾 200k（字幕，單位疑為 XPU，客戶未具名）；2027 起 L13；據點台灣、墨西哥、德州；Pegaverse 數位孿生稱桃園廠建置縮短約 40%；ASRock 為子公司	實際製造數據（自述，字幕）	200k 單位與客戶待查；先前券商估其 GB 機櫃市占低於 3%，與自述規模落差未解；「Rubin Ultra 2026Q4 量產就緒」疑為字幕誤轉寫，不採用
緯穎（6669）	展出 Helios 72 GPU 機櫃、Vera Rubin NVL72、機櫃級 CPO（與 Ayar Labs、創意、Corning；32 托盤、64 ASIC、>800 TB/s，實驗室示範）、6 kW 雙面冷板、鑽石微通道（稱熱阻較銅降 38%）、輕量化液冷（減重 >30%）；稱產能擴增逾 2 倍（未給基準）；AMD 官方名單確認其為 Helios 系統夥伴	展示（多場曝光最多）；Helios ODM 身分 合作開發	無出貨量、客戶未具名；營收 30.6B 簡報標 USD，疑為 NT$ 筆誤，不引用
廣達（2382）	QCT 展示 Vera Rubin POD 五類機櫃（VR NVL72、Vera CPU 機櫃、STX 儲存、Groq 3 LPX、SPX 網路）與數位孿生；Helion、Vera logo	展示	無時程、客戶與數字
華擎 ASRock Rack	Arm AGI CPU（最多 136 核）36 節點 ORv3 機櫃展示；MGX 機櫃；Helion、Vera logo	展示	無客戶與出貨
英業達（2356）	自家工廠 AI 代理（瑕疵檢測、週期時間、SOP 合規）；Helion、Vera logo	內部製造應用；僅 logo（產品面）	不是供貨訊號，無出貨、客戶、數字
緯創（3231）	Vera MGX logo；無專場；不在 Helion 名單	僅 logo	缺乏任何證據，勿推為 Helios 或 VR 份額
神達（3706）	Helion、Vera MGX logo	僅 logo	—
仁寶（2324）、技嘉、微星、華碩	Vera MGX logo；技嘉、華碩亦在 Helion logo	僅 logo	—
創意（3443）	與緯穎共同展示機櫃級 CPO 夥伴	展示	實驗室示範，無量產時程

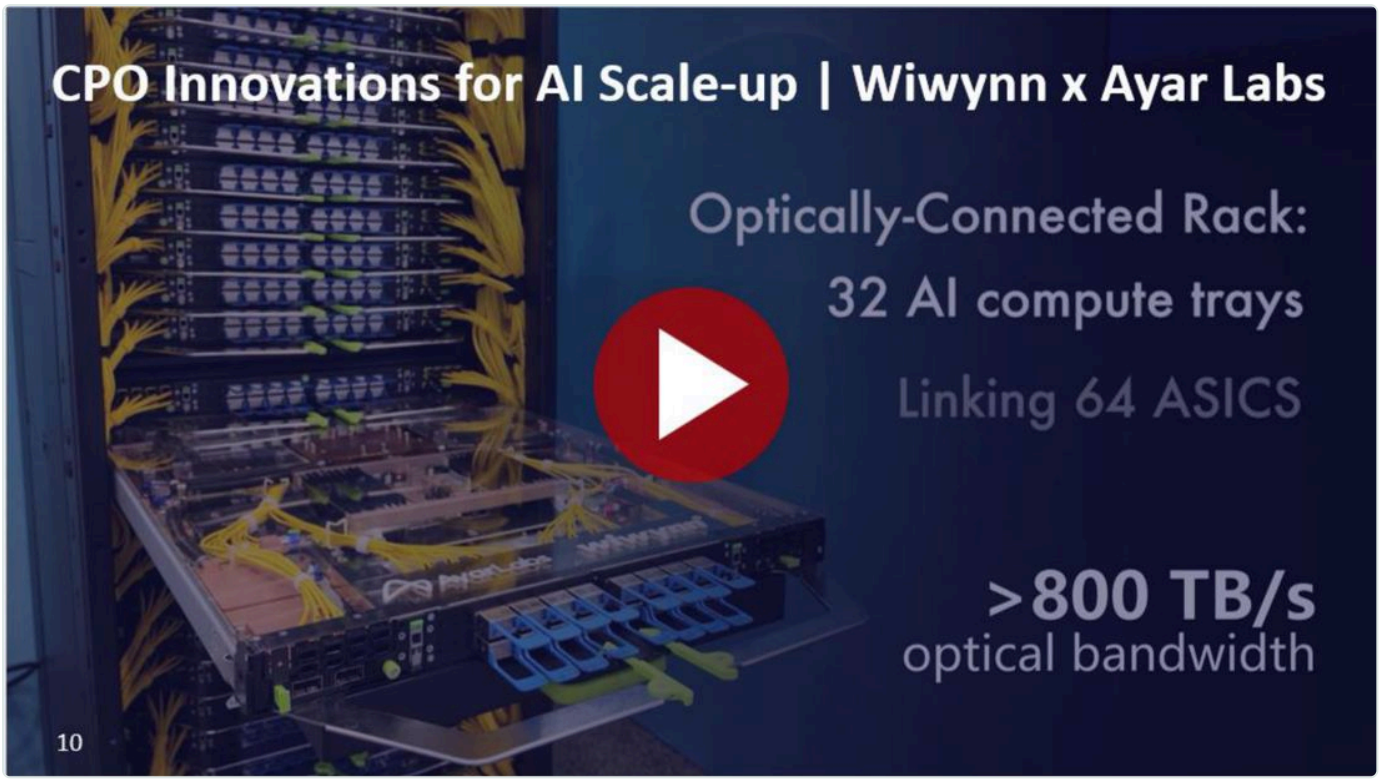


圖 8–1：緯穎與 Ayar Labs 的光互連機櫃（影片截圖）。標示 32 個 AI 運算托盤連接 64 顆 ASIC、光頻寬超過 800 TB/s。為現場實驗室示範，無年份與客戶。來源：OCP 2026，緯穎〈Building the AI Factory〉。



圖 8–2：QCT（廣達）Vera Rubin POD 的五類機櫃。Vera Rubin NVL72、Vera CPU Rack、Vera BlueField–4 STX 儲存、Groq 3 LPX、Spectrum–6 SPX 網路。為產品展示。來源：OCP 2026，廣達 QCT〈Accelerate your AI journey with QCT〉。

8.1 非台廠競爭者

Helios 的機櫃系統商並非台廠獨占：HPE 為 Helios 機櫃系統商，Sanmina、Celestica、Supermicro、Dell、Lenovo 亦在名單或測試流程中；Cisco Nexus One 在 L12 網路與 Spectrum–6 競合；Broadcom TH6 與 NVIDIA 交換晶片為交換核心，台廠價值在機櫃整合而非晶片。

第 9 章 台股供應鏈映射與觀察清單

下表依證據強度排序。「角色」中的推論標明為推論。本表為研究觀察框架，非買賣建議。

公司（代號）	角色	證據強度	觀察指標
鴻海（2317）	VR NVL72 製造與 L10 測試；Helion、Vera MGX logo	實際製造數據（自述）	Vera Rubin 出貨節奏與月度機櫃追蹤中的市占；測試 FPY 是否持續改善；美國與墨西哥產能
和碩（4938）	L11 / L12、2027 起 L13；Vera、HGX Rubin、AMD MI355 與後續	實際製造數據（自述，字幕）	2027Q1 全面量產是否兌現；200k 訂單單位與客戶釐清；德州與墨西哥產能；L13 首案
緯穎（6669）	Helios ODM；VR 機櫃；CPO 機櫃、鑽石冷板與輕量化液冷（展示）	合作開發＋展示	Helios Q4 出貨量與 ODM 分配揭露；產能擴增 >2 倍的基準與時點；CPO 機櫃是否有客戶；營收幣別更正
廣達（2382）	Vera Rubin POD 五類機櫃（展示）	展示	VR NVL72 出貨占比；Vera CPU 機櫃是否有量
華擎 / 和碩持股	Arm AGI CPU 機櫃	展示	首個 Arm 機櫃客戶；與 Helion、Vera CPU 的競合
英業達（2356）	工廠 AI（效率）；Helion、Vera logo	內部應用 / 僅 logo	是否出現 Helios 供貨實證；製造效率是否轉化為毛利
緯創（3231）、神達（3706）、仁寶（2324）	CPU 機櫃 ODM 名單	僅 logo	是否出現專場或出貨揭露；推論：目前不宜據此調整份額假設
創意（3443）	機櫃級 CPO 夥伴（展示）	展示	CPO 機櫃量產時程；客戶揭露

觀察事件清單

事件	為什麼重要
Helios Q4 出貨量與 ODM 分配揭露	AMD 已稱出貨但無數量；ODM 名單與份額是台廠受惠與否的關鍵
Vera Rubin ODM 出貨節奏（和碩 2027Q1）	檢驗 ODM 口徑與 NVIDIA 口徑的落差
Rubin Ultra 量產時點	2027 為預測，且和碩字幕互相矛盾
Maia 300 ODM	Microsoft 2027 多 GW 規劃若屬實，ODM 分配尚未揭露
Helion 11 月供貨	檢驗 CPU 機櫃 logo 名單是否轉成訂單
Open DC v1.0	決定 2027–2028 機櫃與管路尺寸，影響 ODM 設計
各 ODM 美國與墨西哥產能	交付層級拉長後，在地產能是接單條件

第 10 章 風險與反證條件

風險	反證條件 / 觀察訊號
Helios 放量不如預期	Q4 之後 AMD 或 ODM 未提出貨數量，或客戶仍停留在「已安裝」少量階段；以「首批已安裝」推「大量出貨」即為過度外推
電力瓶頸延後部署	發電機約 100 週、美國缺口約 15 GW；若部署受電力限制，機櫃出貨時間會被設施而非 ODM 決定；800V 設施最快 2027 年中
ODM 毛利受交付層級拉長拖累	L11–L13 延伸需要更多資本、庫存與在地人力；自動化與良率提升若無法轉成報價優勢，毛利可能被稀釋（推論，尚無財務數據佐證）
標準化壓縮差異化	MGX 與 Open DC 若降低進入門檻，美系 ODM 與 OEM（Supermicro、Dell、HPE、Sanmina、Celestica）可更快追趕
展示被誤當訂單	緯穎、廣達、華擎皆屬展示；logo 名單不含角色與數量
字幕與口徑誤差	和碩 200k、Rubin Ultra 時點、Microsoft Maia 300 皆依字幕；須待原簡報或財報核對
廠商 token 經濟高估	若第三方（如 SemiAnalysis）正式版數字低於預覽，需求與資本支出假設需下修

來源清單

OCP Global Summit 2026 場次（簡報、影片字幕）：

- AMD 主題演講〈Building the Next Era of AI Infrastructure〉；AMD / HPE〈Building AI Rack System with Open Standards〉；AMD〈Inside AMD Helios: Engineering AI at Rack Scale〉（影片字幕）；nVent〈RDHX cooling for AMD Helios〉
- NVIDIA 主題演講〈AI Factories at Scale: Advancing Open Infrastructure for Agentic AI〉；NVIDIA〈AI Tokenomics Explained〉
- Broadcom 主題演講〈Scaling AI Infrastructure Through Open Innovation〉
- Microsoft〈Intelligence Infrastructure: Scaling AI for the Frontier Era〉（僅字幕）；Google〈From Chip to Grid〉與〈Intelligent Infrastructure for the Intelligence Era〉；Oracle〈Scaling AI Infrastructure Responsibly Through Open Standards〉
- Meta〈Concrete to Code: Engineering AI Infrastructure at Gigawatt Scale〉與〈Achieving AI Cluster Reliability at Scale〉；Google / AMD / Meta / NVIDIA〈Next-Generation ML Infrastructure: The Open DC Architecture〉
- McKinsey〈From Compute Demand to Capacity Delivered〉；SemiAnalysis〈From Silicon to Services〉；Arista（Bechtolsheim）〈The Gigawatt AI Data Center Era〉（僅字幕）；Astera Labs〈Bring Your Own XPU: An Open Rack for Inference Performance and TCO〉
- 鴻海鴻佰〈From Chips to Compute: Delivering AI Infra at GW Speed〉；Cisco / Supermicro〈From Tray to Training Job: L10–L12 Manufacturing Test〉；和碩〈Scaling AI Factory for the Rubin Era〉與〈From Digital Twin to Production Readiness〉（影片字幕）；緯穎〈Building the AI Factory〉、緯穎 / Ayar Labs 光學 scale-up 機櫃與 Expo Hall 展位導覽；廣達 QCT〈Accelerate your AI journey with QCT〉；華擎〈Scaling AI Inference Infrastructure from Silicon to System〉；英業達 / NVIDIA〈Inventec Autonomous AI Factory Agents〉

券商與外部資料：Morgan Stanley，NVL72 機櫃出貨追蹤，2026-10-12（僅作出貨量級參照）；TrendForce（經鴻海簡報引用）。