從 HBM 到 SSD：AI 推論記憶體階層重組

OCP 2026 峰會的證據與台股位置

報告日期：2026–10–19

資料基礎：2026 年 OCP Global Summit（2026–10）中記憶體與客製晶片相關的 13 場簡報與演講，涵蓋 Samsung、SK hynix、Marvell、Meta、Micron、創意（GUC）、聯發科；輔以公開的券商與產業研究說法作交叉對照。

閱讀提醒：本報告中的效益數字絕大多數為廠商自行測試、模擬或預測，各家模型、硬體與基線不同，只能比較方向，不可直接比較大小；台廠受惠多數為推論，報告內逐項標註證據強度。

本報告為研究整理，不構成投資建議。

核心結論

1. **KV cache 把記憶體問題從「單顆 HBM 容量」改寫成「跨機櫃的分層容量與寫入耐久」問題。**HBM、DRAM、CXL 記憶體池、SSD 的四到五層架構，是 Samsung、SK hynix、Marvell、Micron 各場演講共同的方向；但各家效益數字全是自測，模型、硬體與基線都不同，本報告只比較方向，不比較大小（第 1、2 章，表 1）。
2. **SSD 這一層的門檻是耐久與資料放置，不是容量。**SK hynix、Samsung、Meta 三場不約而同指向 FDP（Flexible Data Placement）、ZNS（Zoned Namespace）與 Placement ID：KV cache 推高的是 NAND 寫入量與企業級 SSD 的規格要求，也就是對控制器韌體能力的要求（第 4 章）。這是本報告最不顯而易見的結論；但 Samsung 與 SK hynix 都用自家控制器，外部控制器廠的機會要打折（推論）。
3. **CXL 記憶體池仍在試點（POC）階段。**沒有任何一場給出量產日；Samsung 的 2027 年次世代池規格是路線目標；Meta 雖已大規模用 CXL，但用在通用容量型負載，不是 LLM KV cache（第 2 章）。
4. **HBF（High Bandwidth Flash）時程以 SK hynix 自家簡報為準：**規格 1.0 預期 2027 年 1 月、HBF Cube 預計 2028 年第 1 季可取得；券商高盛（GS）的「2027 年商業樣品」是較樂觀的並列說法（第 5 章）。
5. **客製 HBM 與晶粒間（die-to-die）互連的路線分歧，沒有收斂。**Marvell 推 SBD、創意推自有 GLink、聯發科仍用 UCIe / Mlink；HBM5 每堆疊約 90–100W 的熱密度，是下一個瓶頸（第 6、7 章）。
6. **台股有一手證據的只有兩家：聯發科與創意。**聯發科公司簡報顯示 2026 年資料中心營收由 10 億美元劃改為 20 億美元，並展示 ASIC 托盤搭配 NVSwitch 托盤的機櫃設計；創意提出完整的 GLink 規格與 HBM4E 16Gbps PHY 就緒說法。兩家的客戶皆未具名。群聯、慧榮、威剛、宜鼎、南亞科的受惠都只是推論（第 8、10 章）。
7. **整體信心：中。**方向有多場一致支持；量化全為廠商自測；台廠受惠多為推論。

投資重點表

重點	投資含義	相關標的	證據	信心
記憶體階層化（HBM→DRAM→CXL→SSD）	需求敘事由 HBM 擴散到 DRAM 容量、企業級 SSD 與 NAND；但分層效益為自測，要看量產與客戶導入才算數	三大記憶體廠（Samsung、SK hynix、Micron）；台廠為推論	多場一致	中
SSD 耐久與資料放置成為規格門檻	eSSD 控制器與韌體（FDP、ZNS、遙測）能力成為差異點；垂直整合者吃掉大部分機會	群聯、慧榮（推論）；Samsung、SK hynix（自有控制器）	推論	低至中
CXL 池仍在 POC	不宜將 CXL 寫成已放量；先看 Samsung 2027 路線與 Meta 的 LLM 用例是否出現	Marvell（Structera）、Samsung、SK hynix	展示 / POC	中
HBF 2027–2028 才可能可取得	NAND 近記憶體化是中期題材，非 2026–27 年財報題材	SK hynix、SanDisk（聯盟成員）	自家時程	中
客製 HBM 與 D2D 路線分歧	IP 廠（Marvell、創意、聯發科）各押不同標準；誰成為事實標準尚無定論	創意（3443）、聯發科（2454）、台積電（2330）	展示 / 自述	中
聯發科資料中心 ASIC 上修	公司自述 2026 年營收預期由 10 億上修為 20 億美元，首個量產設計 2026 年第 4 季出貨；客戶未具名	聯發科（2454）	展示 / 自述	中

1　為什麼記憶體變成階層問題

過去談 AI 記憶體，焦點是 HBM 的頻寬與容量。OCP 2026 的記憶體場次把另一件事講得很直接：生產環境的 LLM 服務為了重用已算過的提示前綴（prefix），會把 KV cache 保留一段時間（TTL），保留越久、可重用的狀態越多、工作集（working set）也越大。SK hynix 的簡報引用 Anthropic 提供 5 分鐘與 1 小時兩種快取 TTL、OpenAI 延伸快取最長 24 小時的公開做法，並指出 OpenAI、Google Cloud 與 DeepSeek 的服務都已把 KV 往儲存層搬。

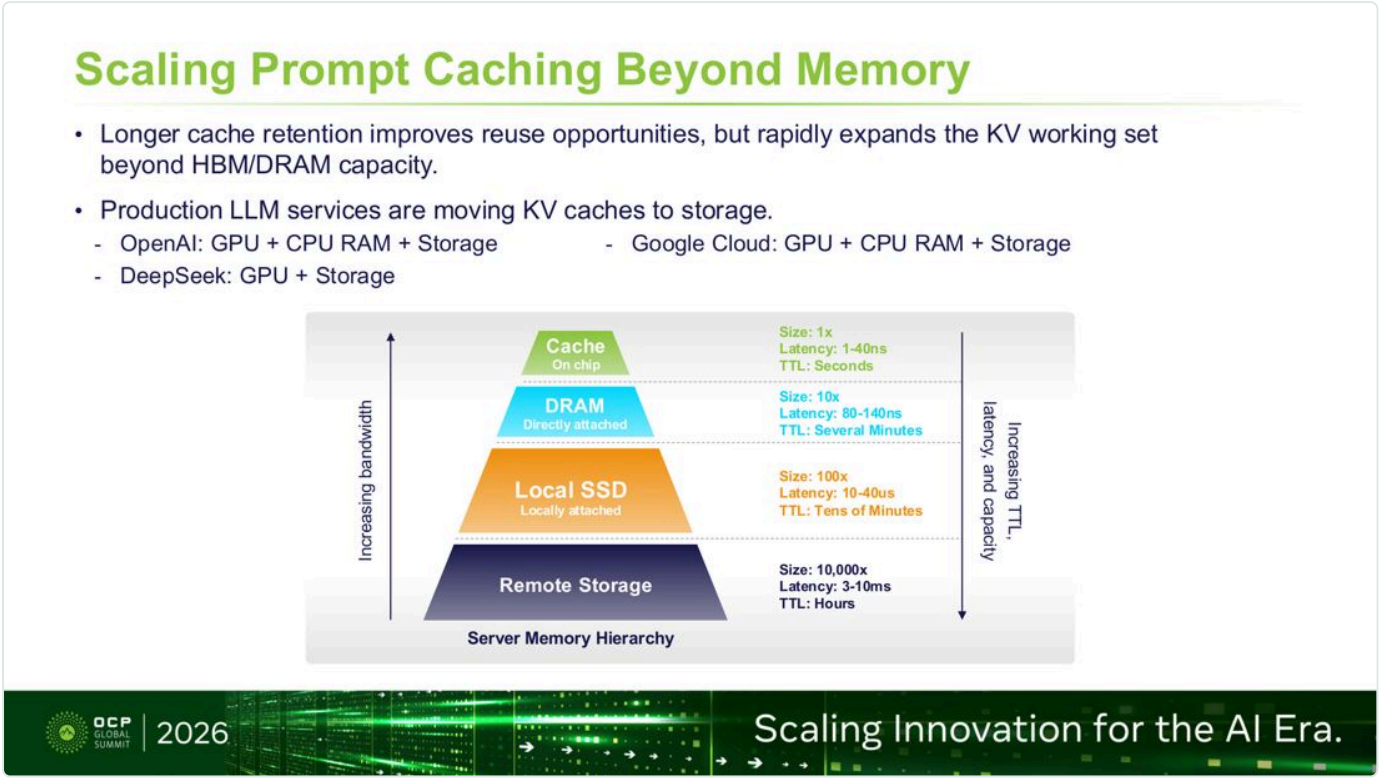


圖 1：伺服器記憶體階層金字塔。由上而下為晶片內快取、直連 DRAM、本機 SSD、遠端儲存；容量約 1x / 10x / 100x / 10,000x，延遲約 1–40ns / 80–140ns / 10–40µs / 3–10ms，可保存時間由秒、數分鐘、數十分鐘到小時。（SK hynix 的示意值。）
來源：OCP 2026，SK hynix 場次〈Storage as a Performance Tier for LLM Serving〉

工作集到底有多大

SK hynix 以 Qwen3-14B（BF16）為例推算：保留 10 分鐘 TTL 的聊天型負載，每顆 GPU 的 KV cache 需求最高約 781GB。對照 NVIDIA B200 伺服器上每 GPU 的 HBM 140GB（模型本體占 50GB）、主機 DRAM 375GB，超出 HBM 加 DRAM 的部分只能落在 1–10TB 的 SSD 層。這是「單一模型、單一資料集（Qwen-Bailian 匿名資料集）」的推算，不是普遍數字，但它說明了分層的必要性。

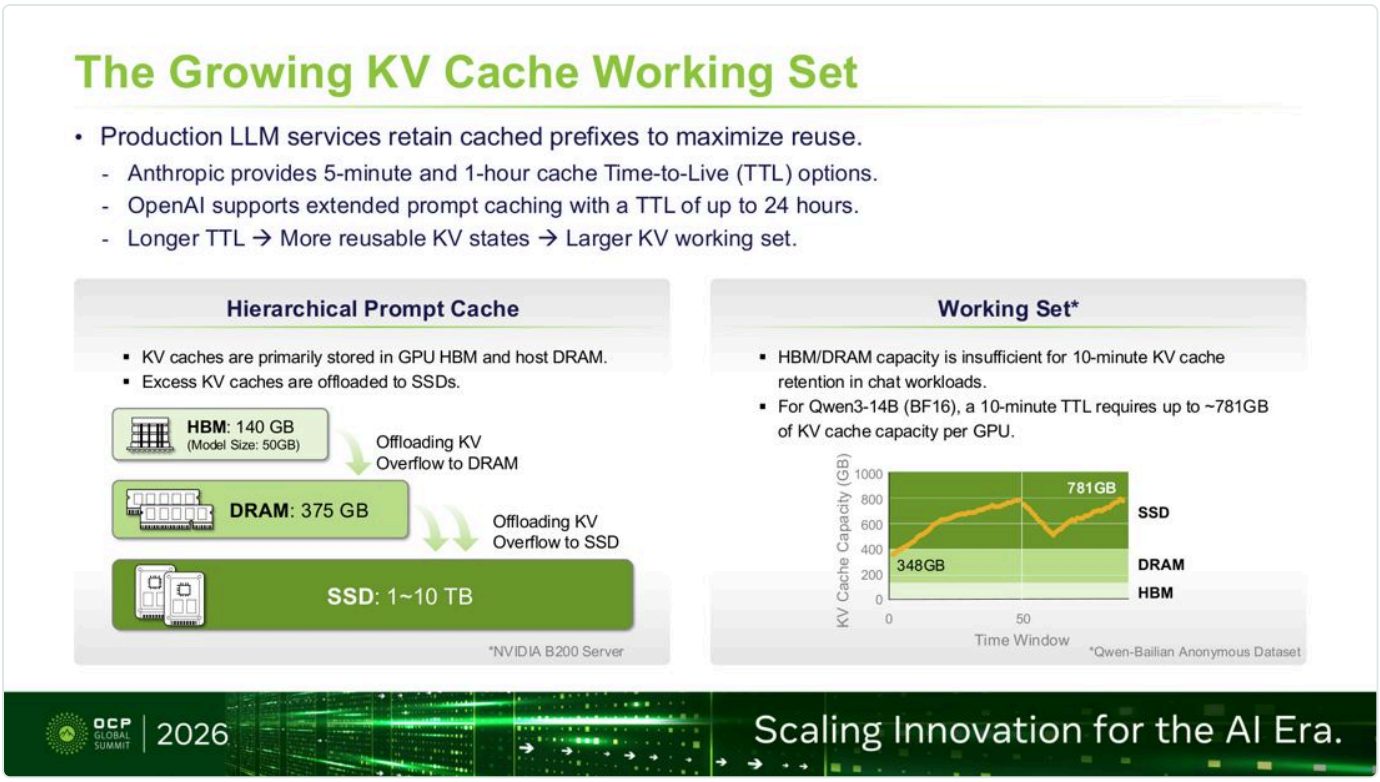


圖 2：KV cache 工作集隨時間增長。左：分層提示快取，HBM 140GB、DRAM 375GB，超出者溢出至 SSD（1–10TB）。右：Qwen3-14B 以 10 分鐘 TTL，容量需求由 348GB 起、峰值約 781GB，超出 HBM 與 DRAM 的部分落在 SSD。
來源：OCP 2026，SK hynix 場次〈Storage as a Performance Tier for LLM Serving〉；推算值，口徑為單一模型。

各家切法不同，但都是四到五層

來源	階層劃分	備註
Marvell	G1 GPU HBM；G1.5 光學記憶體機櫃（PFMA）；G2 系統 DRAM（1–2TB）；G2.5 CXL 網路記憶體（數十 TB、數百 ns）；G3 本機 SSD；G3.5 高效能快閃；G4 共享物件 / 檔案（數百 TB）	廠商分層模型
SK hynix	晶片內快取、DRAM、本機 SSD、遠端儲存四階（圖 1）	以 TTL 區分
Samsung（代理式 AI 場次）	熱（活躍 session：HBM / 本機 DRAM / CXL）、溫（system prompt、多輪歷史）、冷（文件、程式庫：區塊 / 物件）	影片場次，依講者口述整理
Samsung（2030 願景）	L0 SRAM、L0.5 zHBM / 3D DRAM、L1 HBM5、L1.5 光學 HBM、L2 LPDDR / SOCAMM、L2.5 CXL 池 / zNAND-O、L3 KV SSD、L3.5 乙太網路 SSD、L4 TLC / QLC SSD	願景，無規格（第 9 章）
Micron	以「容量層」與「頻寬層」兩族群建模 KV 卸載	建模，不是量測產品

容量要跨過門檻才有效

Micron 的建模給出一個對 NAND 與 SSD 很重要的提醒：KV 卸載層的有效重用率並不隨容量線性上升，而是有門檻。在 Llama 3-70B、GH200 的設定下，容量約 400GB 起重用率趨近飽和；並行數拉高到 64 時，活躍 KV 工作集超過容量，有效重用率崩落、首字延遲（TTFT）急升。換句話說，「買不到門檻的容量」等於沒有效益；今天足夠的記憶體層，也可能在生產並行度下失效。

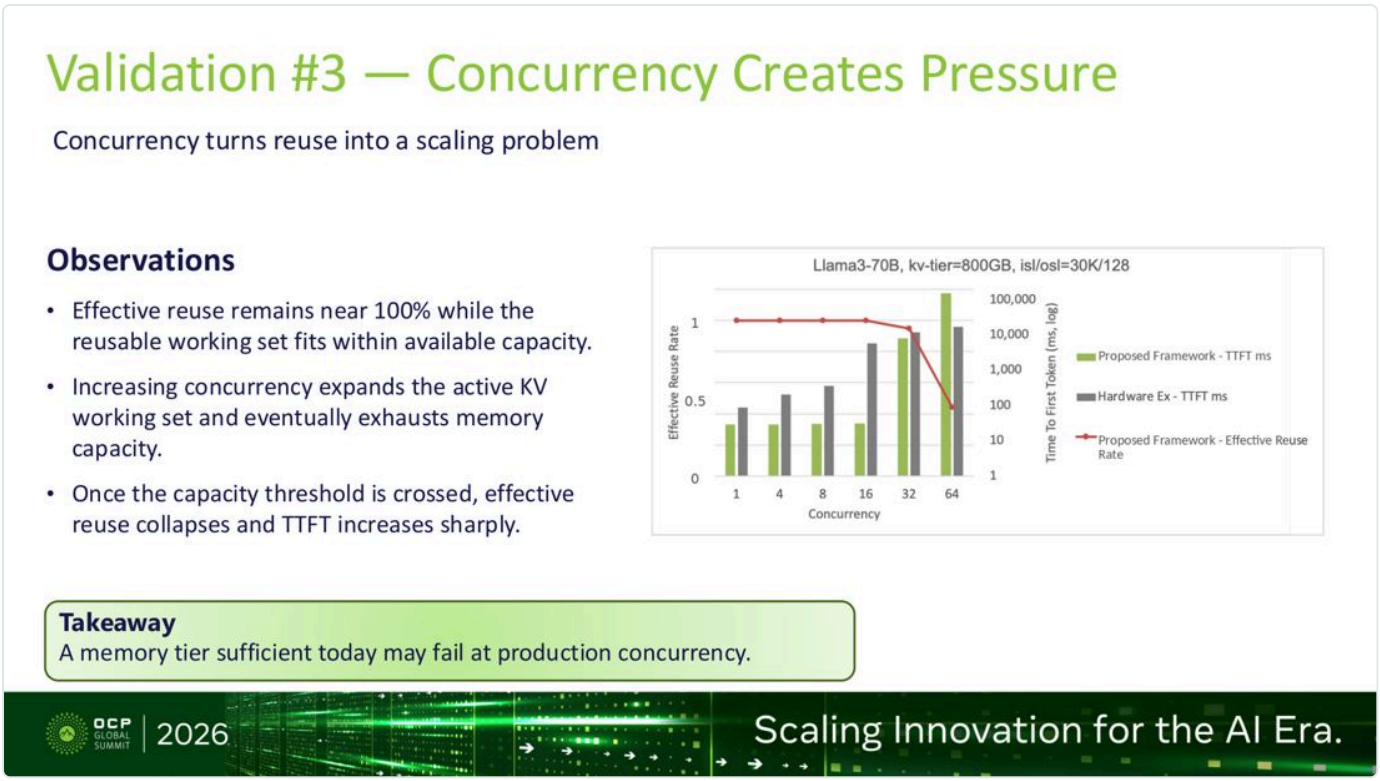


圖 3：並行數對有效重用率與 TTFT 的影響。Llama3-70B、KV tier 800GB、輸入 / 輸出 30K / 128 tokens；有效重用率（紅線）在並行 1–16 維持約 1，並行 32 略降、64 大幅崩落；TTFT（對數軸）自並行 16 起由數百毫秒急升至萬毫秒級。
來源：OCP 2026，Micron 場次〈Tiered Memory for LLM Inference — A Modeling Framework for KV-Cache Offload〉；建模與圖讀值。

需求端數字的證據等級

Samsung 場次口述了「每 token 讀取量兩年增 6 倍」「單一雲端業者每月產生約 140EB 的 KV cache」「400 個 1M-token 並行 session 約 18TB KV」等需求數字，來源為講者口述，未見原始出處驗證，本報告只當作敘事背景，不作為估算依據。

2　CXL 記憶體池與記憶體中心機櫃

KV 卸載的第一站是 CXL 記憶體池：用 CXL 把大量 DDR 記憶體做成可被多台伺服器共享的池，容量大於單機 DRAM、延遲低於 SSD 與遠端儲存。SK hynix 的 MARS（Memory-centric AI Rack System）是最完整的架構展示：GPU 機櫃與 CPU 機櫃內部以 CXL 連到機櫃內的記憶體池，另有獨立的記憶體機櫃，內含多組 CXL 或 CXL 混合記憶體池（含 DRAM 加 NAND），運算機櫃與記憶體機櫃之間以 RDMA 網路相連。

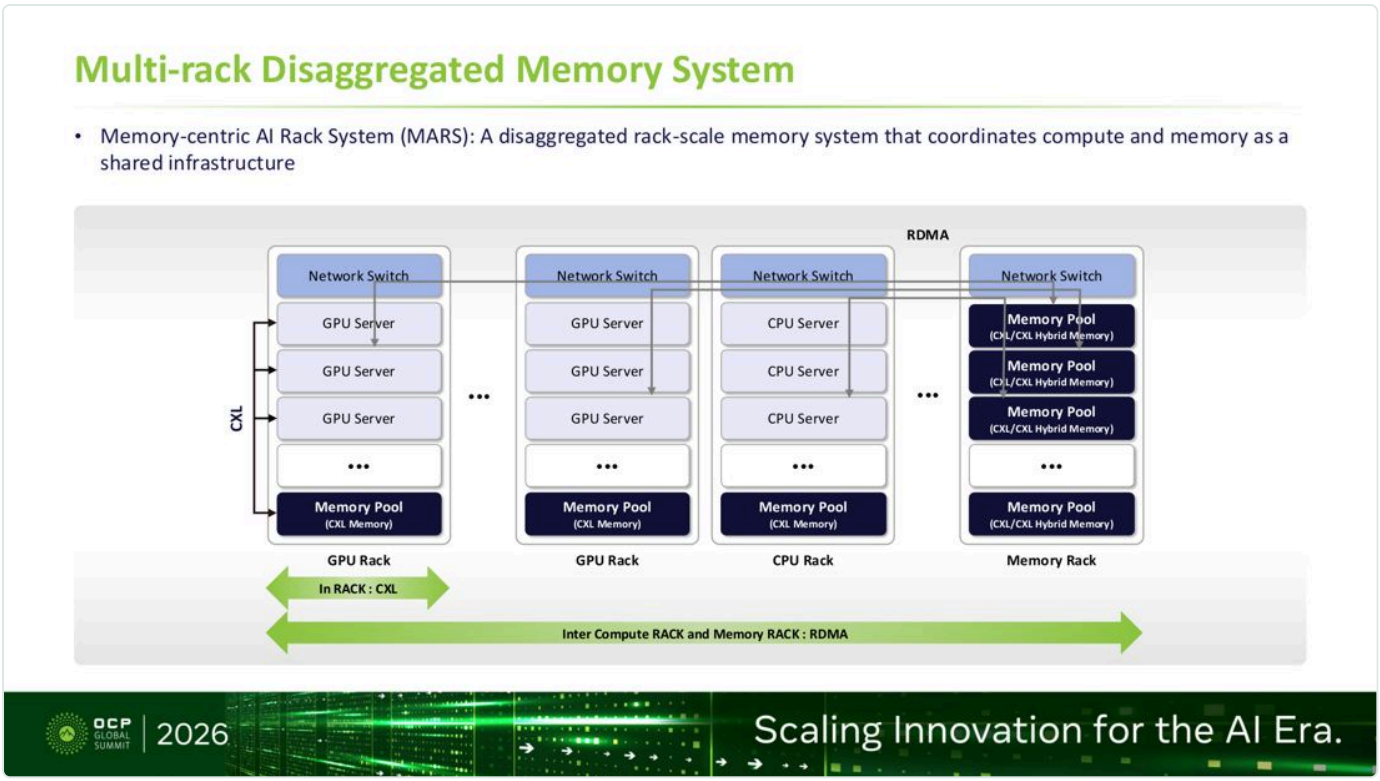


圖 4：MARS 多機櫃解耦記憶體系統。GPU 機櫃與 CPU 機櫃內以 CXL 連接機櫃內記憶體池；獨立的記憶體機櫃內為多個 CXL / CXL 混合記憶體池；運算機櫃與記憶體機櫃之間經網路交換器以 RDMA 連接。
來源：OCP 2026，SK hynix 場次〈Introduction to Memory-centric AI Rack System (MARS)〉

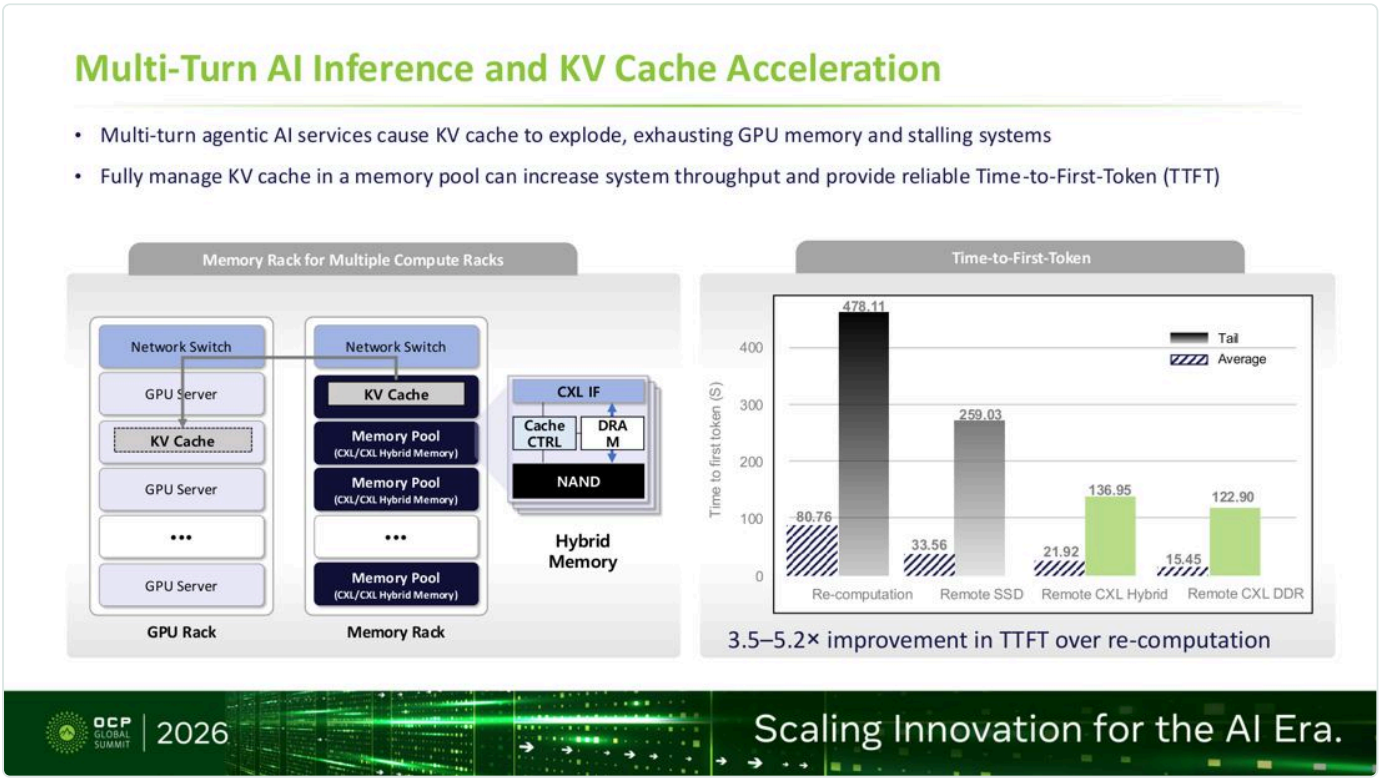


圖 5：MARS 的 KV cache 加速與 TTFT 結果（自測）。以秒為單位，平均值（斜線柱）／尾端值（實柱）：重算 80.76 / 478.11；遠端 SSD 33.56 / 259.03；遠端 CXL 混合 21.92 / 136.95；遠端 CXL DDR 15.45 / 122.90。簡報標示相對重算改善 3.5–5.2 倍。模型與硬體配置簡報未載。
來源：OCP 2026，SK hynix 場次〈Introduction to Memory-centric AI Rack System (MARS)〉

三家自測對照：方向一致，數字不可互比

下表把各場自測結果並列，每列附模型、硬體與基線。三場都顯示「卸載到 CXL 池優於重算或 RDMA」，但負載、模型與基線都不同，絕對值與倍數不能互比。

表 1　各家 KV 卸載自測結果對照（全為廠商自測或建模）

來源場次	模型 / 負載	硬體	基線	結果	口徑備註
SK hynix MARS	多輪長 context，模型未載	配置未載	GPU 重算	TTFT 平均 80.76 秒降至 15.45 秒（CXL DDR）；遠端 SSD 33.56 秒；混合 21.92 秒	單位為秒；平均與尾端兩組易混
Samsung 代理式分層	機器人 / 視覺語言模型多輪	CXL 主機加 Blackwell GPU（POC）	RDMA	TTFT 650ms 降至 200ms；吞吐 7 升至 14 req/s	影片字幕；「最多 25 倍」為相對重算的最大值，條件未述，不作代表值
Samsung 高擴展記憶體	vLLM 實測	CMM-D 池，10–80TB、782GB/s	未詳	TTFT 降 38%、吞吐 3.2 倍；CMM-D 擴展使 TPS 5.6 倍（不增 GPU）	影片字幕，基線未詳
Marvell	Llama-3.1-8B，lmbench 長文件問答，峰值 16 users	H200＋Structera A	HBM 重算	TTFT 5,933ms（重算）/ 4,243ms（SSD）/ 2,162ms（DRAM）/ 1,199ms（Structera A）；吞吐 31,394 / 44,318 / 80,813 / 139,497 tok/s	小模型單機；早期結果
SK hynix 運算共享池	名為 Arcee 的流量樣本	搭配 Structera CXL	Mooncake	TTFT 398ms 對 709ms；E2E 2,084ms 對 4,766ms	簡報文字寫 1.4x，與圖上絕對值不符（見第 3 章）
SK hynix 儲存場	Qwen3-14B，100% 命中微基準與 ShareGPT	H100、Gen5 NVMe；vLLM＋LMCache	記憶體（DRAM）	SSD 未優化為記憶體的 1.26–1.92 倍；平行度調校後 1.07–1.14 倍	比的是 SSD 對 DRAM，方向與上列不同
Micron	Llama3-70B，輸入 / 輸出 30K / 128	GH200（建模）	重用率 0.1 到 1	TTFT 約 60.10 秒降至 5.77 秒；容量 400GB 起飽和	建模框架，部分為圖讀值

Meta 的實況：CXL 已上線，但不是為了 LLM KV

Meta 的演講提供了少見的營運面視角：CXL 記憶體擴充已在機隊導入，用於通用容量型負載，簡報列出的效益包括快取類負載吞吐約 +25%（資料保留 5–10 倍）、分散式快取平均查詢延遲 −29%、資料倉儲每台 executor 數 +25%、開發機每台 VM 數 +33%（機隊 −15%）、ML 參數伺服器伺服器數 −25% 且吞吐 +12%。簡報未給 CXL 版本、廠商、配置與規模，也強調大規模部署需要標準化的除錯與遙測（失效模式的可觀測性）。因此 Meta 是「CXL 可營運」的證據，不是「CXL 用於 LLM KV cache」的證據。

時程現況

本批簡報沒有任何一場給出 CXL 池的量產日期：SK hynix MARS 為「客製試點」；SK hynix CMM-Ax 與 Niagara 2.0 為展示；Samsung 的 CMM-D MD310 可供客戶 POC、MD320（LPDDR 基）開發中，2027 年次世代池（20TB、最高 5TB/s）是路線圖。CXL 池現況應寫為 POC 與試點。

CXL 頻寬論：廠商反論與商用現況

論點 A	論點 B
SK hynix 發表於學術期刊的論文論點：CXL 頻寬僅約 HBM 的 0.2%，要真正接近 HBM 需走光學互連的記憶體池。	Marvell 自測（Llama-3.1-8B、單張 H200）：CXL 卡 139k tok/s 對 DRAM 81k tok/s；Samsung 2030 願景同時保留 CXL 池（L2.5）與光學 HBM（L1.5）。
判讀：兩者並存，不互相裁定。廠商是以容量與利用率論證，而非頻寬；商用卡仍為 CXL 2.0 / PCIe 5.0（單卡約 200GB/s 級，如 CMM-Ax 3.0）。Marvell 的結果是 8B 小模型、單機的廠商自測。	

3　近記憶體運算：在池內做 KV 選擇與 tokenization

CXL 池不只是被動的容量。SK hynix 的 CMM-Ax 3.0 把運算放進記憶體卡：CXL 2.0 Type 3 加 PCIe 5.0 x16、512GB DDR5-6400（約 200GB/s）、16 顆 Arm Neoverse V2 核心，並提供 Attention API、向量搜尋等軟體層。實驗使用 Marvell 的 Structera CXL 控制器，做法是把 KV 評分放進池中：預填（prefill）階段每算完一層，就把該層完整 KV 傳進 CXL 池，由 CMM-Ax 在池內做第一階段評分（選出重要的 KV 索引）；解碼階段只依選中的索引（例如 {1, 4, 5, 6}）把部分 KV 取回 GPU，完整 KV 留在池內。

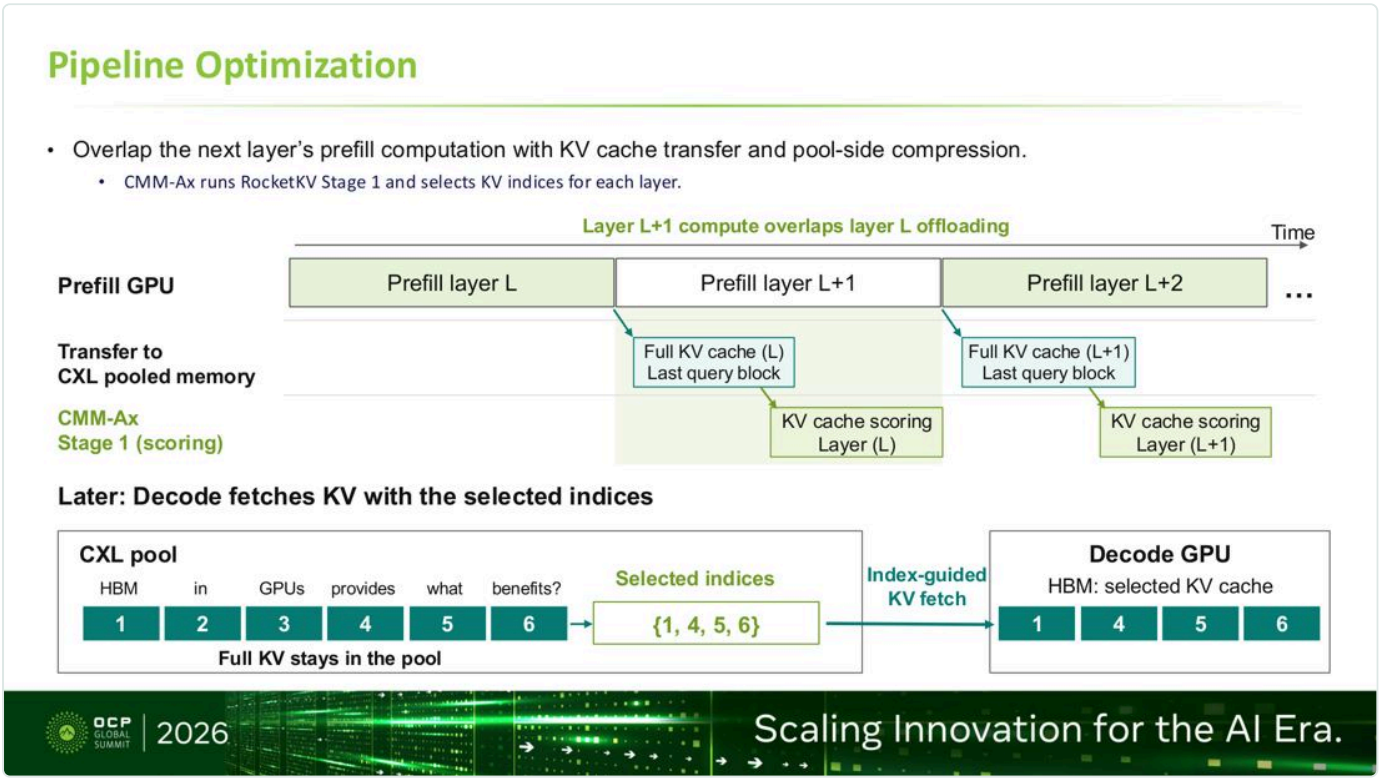


圖 6：管線最佳化。Prefill GPU 計算第 L+1 層，與第 L 層 KV 傳至 CXL 池並行；CMM-Ax 在池內做 KV 評分；Decode GPU 依選中索引只取回部分 KV，完整 KV 留在池內。
來源：OCP 2026，SK hynix 場次〈Agentic AI Serving System with Computational Shared Memory Pool〉

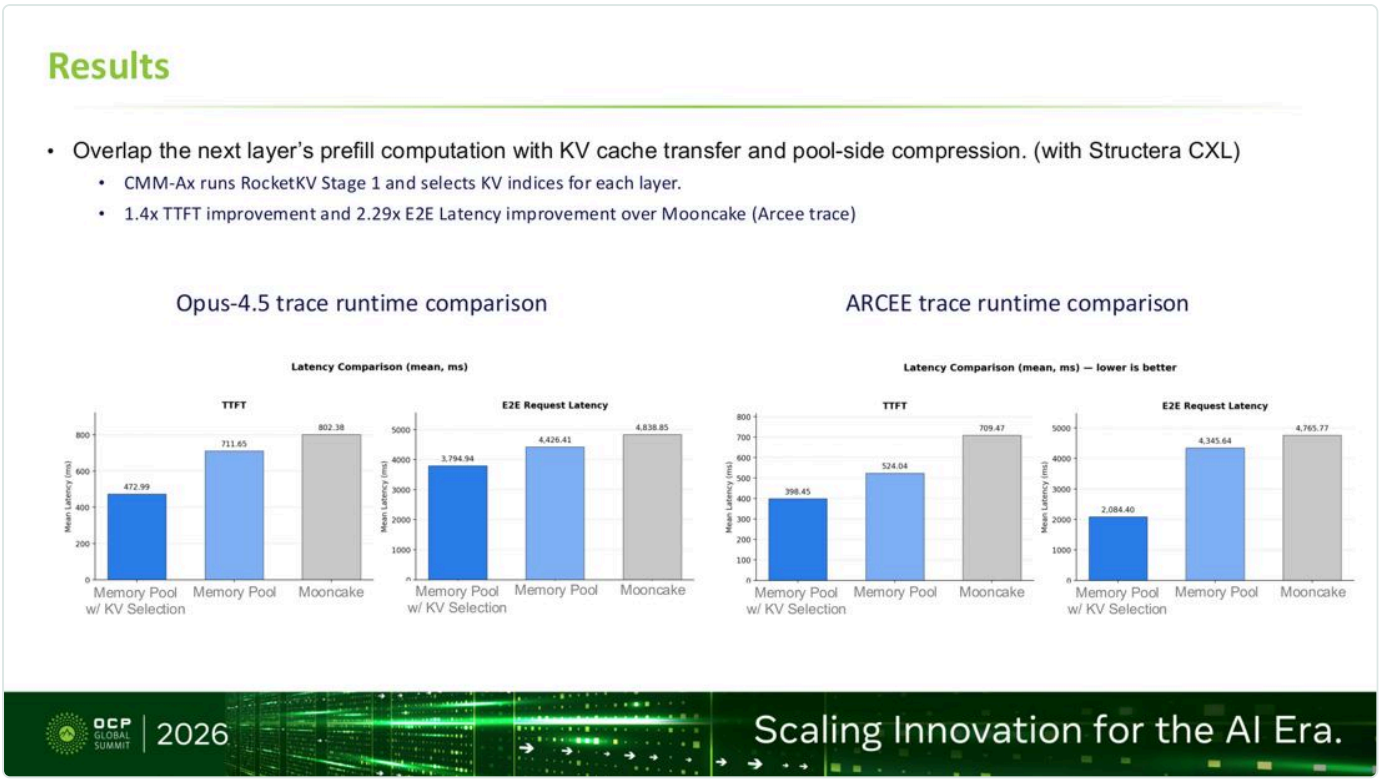


圖 7：KV 選擇池對無選擇池與 Mooncake 的延遲比較（自測，搭配 Structera CXL）。Arcee 流量樣本 TTFT 為 398 / 524 / 709ms（KV 選擇池 / 一般池 / Mooncake），E2E 為 2,084 / 4,346 / 4,766ms；名為「Opus-4.5」的流量樣本為 473 / 712 / 802ms。「Opus-4.5」「Arcee」是流量樣本名稱，不代表實測該模型。
來源：OCP 2026，SK hynix 場次〈Agentic AI Serving System with Computational Shared Memory Pool〉

> **圖文不一致**
> 簡報文字寫「TTFT 改善 1.4 倍、E2E 改善 2.29 倍」。以圖上絕對值計算，Arcee 的 TTFT 為 709 對 398ms，約 1.78 倍；與一般池（無 KV 選擇）比為 524 對 398ms，約 1.32 倍；E2E 4,766 對 2,084ms 約 2.29 倍，與文字相符。本報告引用圖上絕對值，不使用 1.4 倍。

同一場次還把 tokenization 搬進 CXL 池：25.6K 輸入、前綴命中率 96% 的條件下，tokenization 時間由約 130ms 降至 6ms（約 −95%），TTFT 由約 345ms 降至約 221ms（約 −36%）。Marvell 場次則探討壓縮卸載：以 FP8 壓縮 KV，TTFT 由 4,864ms 降至 3,374ms（2.0 倍）；CacheGen 路線標示 5.6 倍但實作未完成。這些都是「讓池子更聰明」的方向，目前同樣停在實驗與早期結果。

項目	內容	證據等級
CMM-Ax 3.0（SK hynix）	CXL 2.0、512GB DDR5-6400、200GB/s、16 核 Arm Neoverse V2	展示（FMS、OCP、SC 2025 出展）
Marvell Structera（X / A / P / S）	Structera S：CXL 2.0 交換、2TB/s、port-to-port 小於 100ns、最多 100TB 池；PFMA：1TB HBM3 加 16/32TB DDR5、CXL 3.1、小於 350ns	展示，無出貨時程；PFMA 數字為規格
池內 tokenization 卸載	130 降至 6ms；TTFT 345 降至 221ms	廠商自測

4　SSD 當 KV 層：耐久與資料放置才是門檻

寫入量把容量算式改寫了

KV cache 的寫入是連續而龐大的。SK hynix 以 Qwen3-14B 推算，每顆 GPU 每日 KV 寫入量最高約 57.5TB。若要在該寫入量下滿足耐久限制，各種 NAND 的最小 SSD 容量為：SLC 約 0.29TB、TLC 約 14.4TB、QLC 約 28.8TB；透過預留空間（OP）可把 TLC 降至 4.8TB，透過 ZNS 可把 QLC 降至 8.8TB。簡報同時直言，部分企業級 SSD 無法達到所需的每日全碟寫入次數（DWPD）。換句話說，決定能不能用 QLC 的是寫入放大（WAF）與資料放置，不是標稱容量。

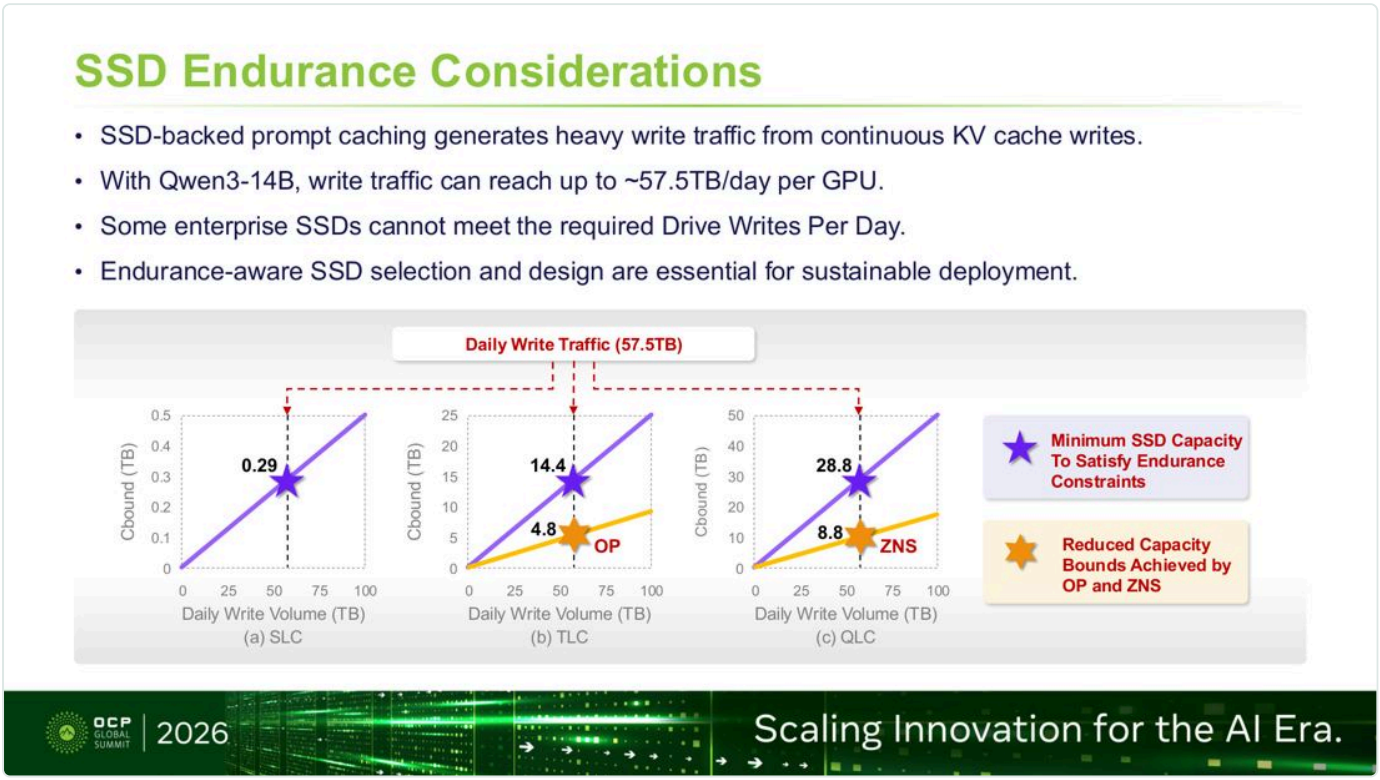


圖 8：KV 寫入量下的 SSD 最小容量需求。每 GPU 每日寫入 57.5TB；SLC / TLC / QLC 滿足耐久所需最小容量分別為 0.29 / 14.4 / 28.8TB，OP 與 ZNS 可降至 4.8TB（TLC）與 8.8TB（QLC）。Qwen3-14B 推算值。
來源：OCP 2026，SK hynix 場次〈Storage as a Performance Tier for LLM Serving〉

三場指向同一組解法

場次	說法	口徑	性質
SK hynix	每 GPU 日寫入最高約 57.5TB；QLC 需 28.8TB，ZNS 可降至 8.8TB；提 OP、ZNS、FDP 與 KV SSD 研究方向	Qwen3-14B 推算	研究主張，未宣布產品
Samsung（高擴展記憶體）	KV 與代理負載需 DWPD 3 以上，QLC 單獨撐不住；FDP 使 WAF 由 3–4 降到 1，多租戶混合負載實測 3.4 降至 1.1；混合媒體（同一碟劃出 SLC 分區）	Samsung 實驗室；影片字幕	自測；PM1763（Gen6）、PM1773（E3.S、16–256TB、0.6 DWPD）為發表產品
Meta	QLC 因 WAF 約 3–5 倍，耐久僅 TLC 的約 20%；以 Placement ID 把有效 DWPD 由 0.17 提升到 0.96	條件未載	機隊實況；無導入比例

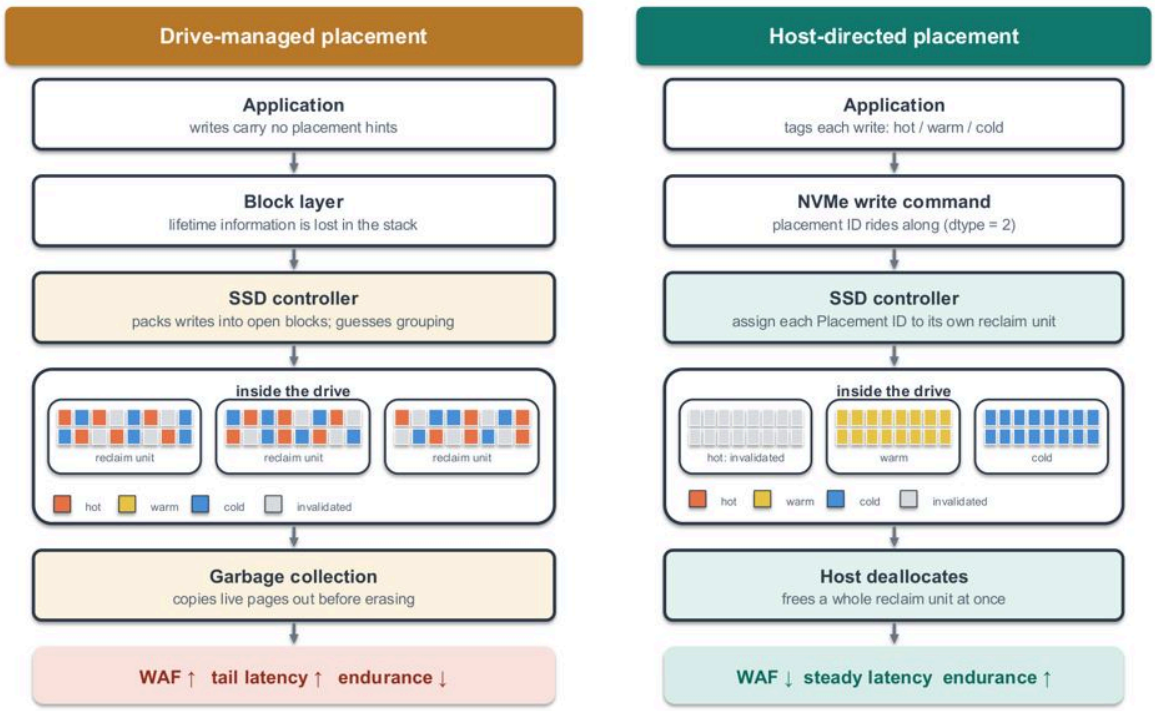


圖 9：資料放置——由碟自行猜測 vs 由主機標記。左：應用寫入不帶提示，SSD 控制器自行猜測分組，垃圾回收（GC）搬移存活頁，導致 WAF 上升、尾延遲上升、耐久下降。右：應用把每筆寫入標為 hot / warm / cold，NVMe 寫入命令攜帶 Placement ID（dtype=2），各進入獨立的回收單元，主機一次釋放整個回收單元，WAF 下降、延遲平穩、耐久提升。
來源：OCP 2026，Meta 場次〈QLC @ Hyperscale〉

Meta 的 QLC 實況：高密度難取得，品質差異大

Meta 列出的 QLC 規格範圍：單碟 2026 年最大 122.88–245.76TB（TLC 為 61.44TB、HDD 為 26TB）、0.25–0.6 DWPD、持續寫入 2.0–3.5GB/s、2U 24 碟可達 2.95PB。導入初期失效率表（單位簡報未標）顯示廠商 A 總計 2.9、廠商 B 總計 11.7、全部 QLC 為 3.7，主要差異在媒體品質；兩家廠商皆未具名，245.76TB 是否已出貨也未說明。QLC 定位在 TLC 與 HDD 之間的 warm 層，並非 LLM KV 主戰場。因此，Meta 是「QLC 可營運但供應與品質仍有落差」的證據，不能引用為大幅追加訂單的證據。

> **本章的非顯而易見結論。**SSD 層的競爭不在 NAND 容量，而在「誰能讓 QLC 活過 KV 的寫入量」。FDP / ZNS / Placement ID 把部分責任由 SSD 控制器搬到主機軟體與標準介面；對控制器廠而言，韌體對這些功能的支援與 GC / WAF 遙測能力，會變成進入資料中心 eSSD 的規格門檻（Meta 明確要求遙測與標準化除錯）。但 Samsung（PM17xx 自家控制器）與 SK hynix 為垂直整合，外部控制器廠的機會主要在非三大廠 NAND 的客戶端——此為推論，簡報未點名任何控制器廠。

5　HBF 與時程

HBF（High Bandwidth Flash）把 NAND 核心晶粒（簡報圖為四層）疊在 Base / Logic Die 上，與處理器並排放在中介層上，目標是以約 HBM 一半的頻寬、約 10 倍的容量，補 HBM 與 eSSD 之間的缺口。SK hynix 以卡級口徑比較：頻寬 0.5 倍、容量 14 倍（HBF 卡 3TB 對 Base 卡 288GB）、每位元能耗 3.7 倍；eSSD 則為頻寬小於 0.1 倍、容量大於 500 倍、每位元能耗 6.9 倍。

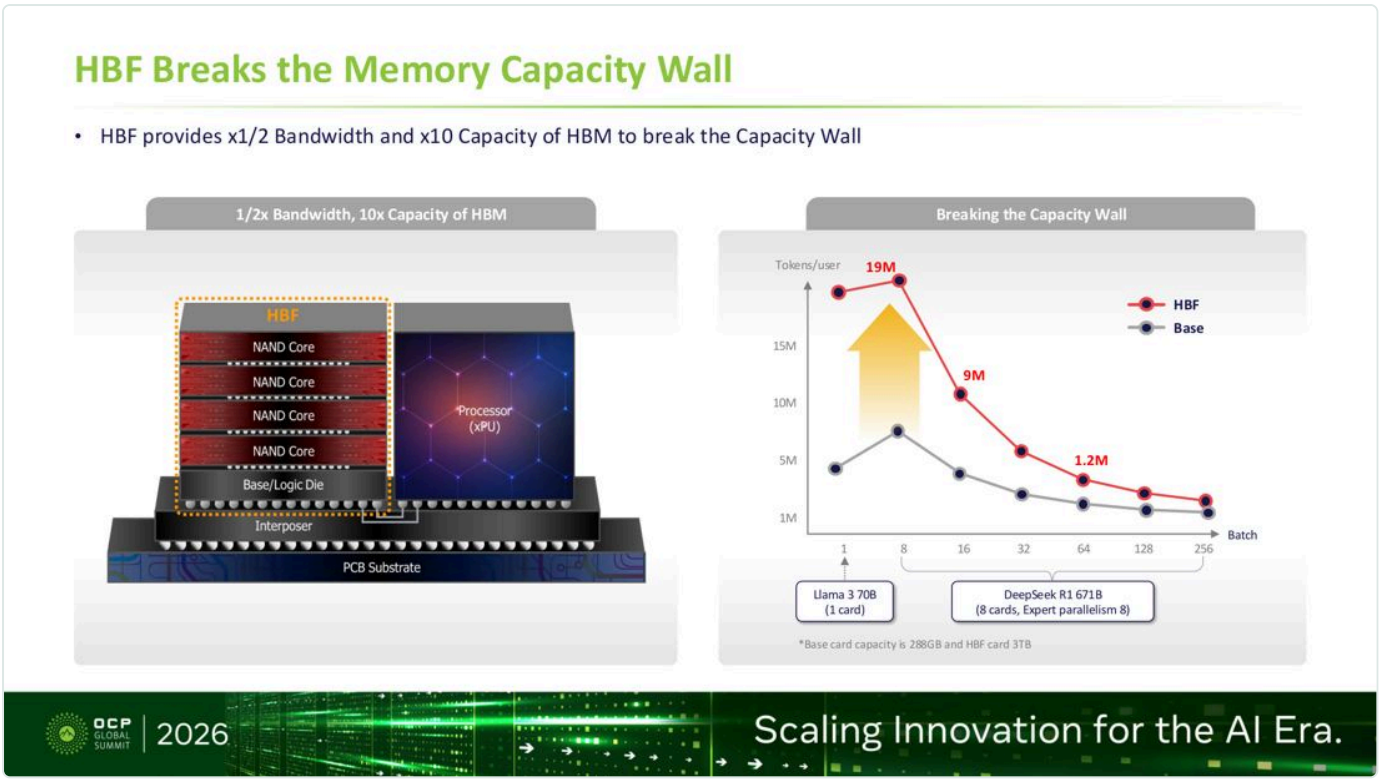


圖 10：HBF 結構與每使用者 tokens 對批次大小。左：HBF 為四層 NAND Core 疊在 Base / Logic Die，與處理器（xPU）並排於中介層。右：HBF 在批次 8 達峰值約 19M tokens/user（批次 1 略低），批次 16 約 9M、批次 64 約 1.2M，皆高於 Base；註記 Base 卡 288GB、HBF 卡 3TB；Llama 3 70B 單卡、DeepSeek R1 671B 八卡。
來源：OCP 2026，SK hynix 場次〈Introduction to Memory-centric AI Rack System (MARS)〉；廠商模擬值。

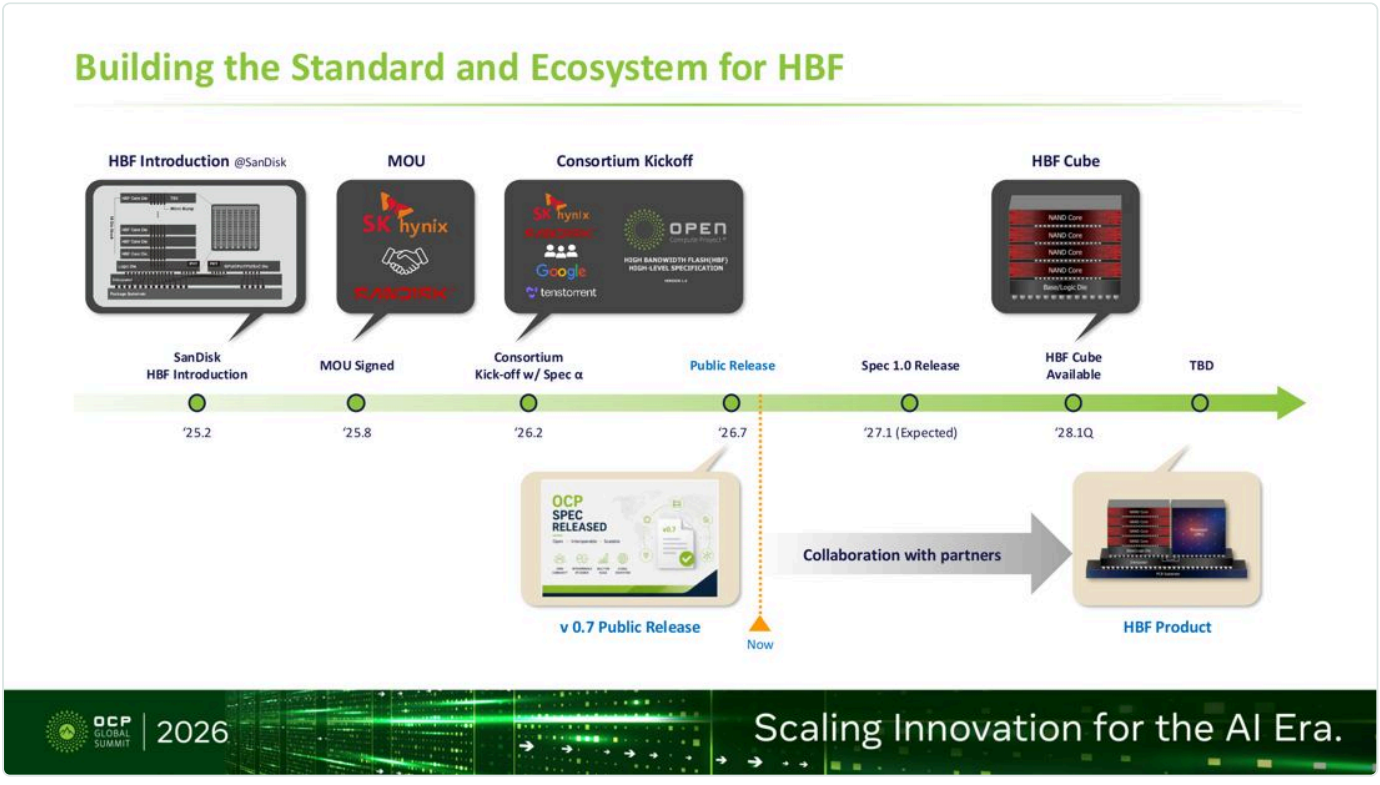


圖 11：HBF 標準與生態系時程。2025 年 2 月 SanDisk 發表 HBF；2025 年 8 月 SK hynix 與 SanDisk 簽 MOU；2026 年 2 月 OCP 聯盟啟動（含 Spec α，標示 SK hynix、SanDisk、Google、Tenstorrent）；2026 年 7 月 v0.7 公開（標示「現在」）；2027 年 1 月 Spec 1.0（預期）；2028 年第 1 季 HBF Cube 可取得。
來源：OCP 2026，SK hynix 場次〈Introduction to Memory-centric AI Rack System (MARS)〉

議題	說法 A	說法 B	判讀
HBF 樣品 / 量產時程	SK hynix 簡報：Spec 1.0 預期 2027 年 1 月、HBF Cube 2028 年第 1 季可取得	高盛（GS）：商業樣品 2027 曆年開始出貨；凱基：量產在 2028 年之後	可並存：規格 1.0 與樣品 / Cube 是不同里程碑；以 SK hynix 自家時程為「可取得」基準，GS 為較樂觀說法，並列保留
HBF 容量比	SK hynix：卡級 14 倍（3TB 對 288GB）	另有研究以堆疊級比較：HBM4 約 36GB 對 HBF 約 512GB	堆疊級與卡級口徑不同，不可互比

6　客製 HBM 與 base die 分工、HBM5 的熱

Marvell 的論述是：HBM 已占 XPU 約 30% 的晶粒面積、超過 50% 的 XPU 成本，且許多 GPGPU / ASIC 的記憶體物料成本占 60–80%。HBM4 把介面加倍到 2048 位元（32 個 64 位元通道），PHY 面積、功耗與晶粒邊緣（beachfront）都在膨脹，與運算邏輯搶同一段邊緣；因此限制加速器擴展的越來越是記憶體 I/O，而非核心。

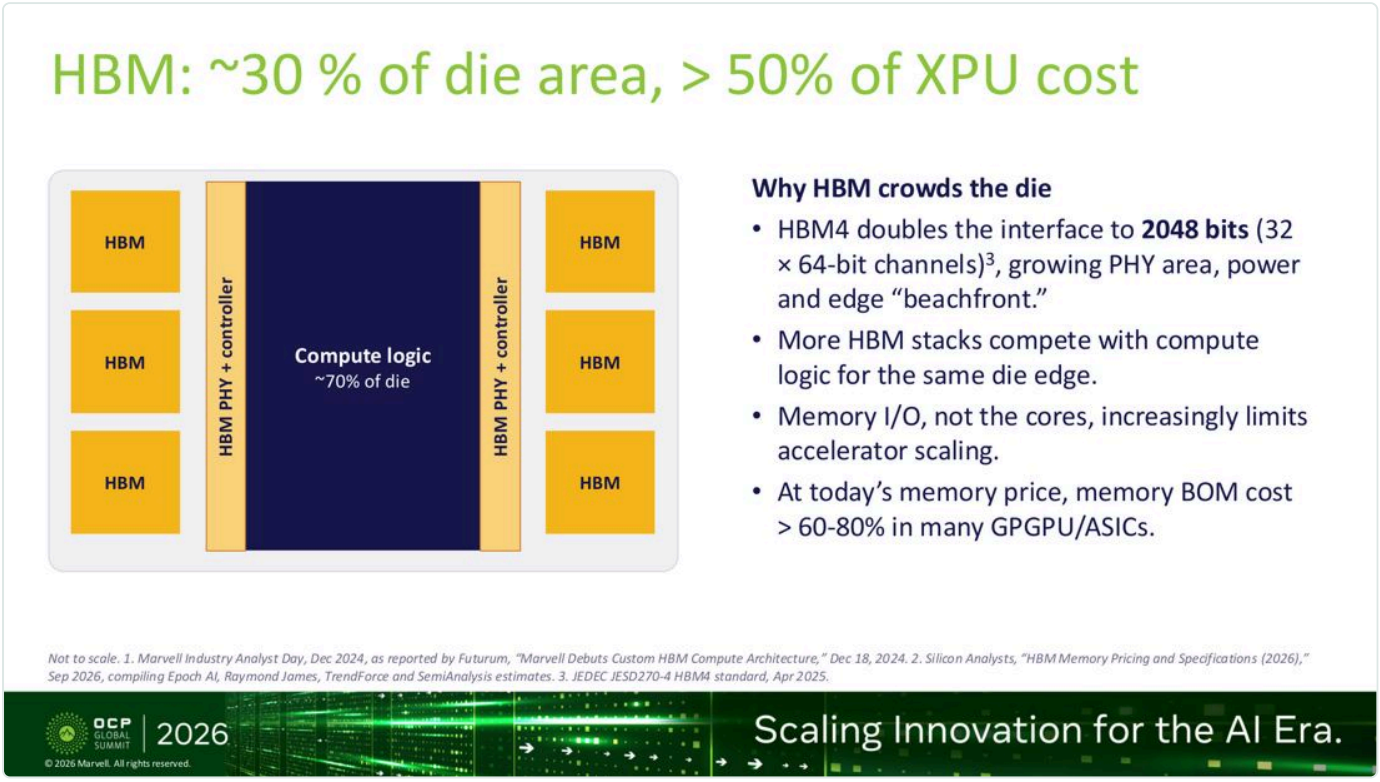


圖 12：HBM 占晶片面積約 30%、占 XPU 成本超過 50%。示意圖（非按比例）：中央為運算邏輯（約 70% 晶粒），兩側為 HBM PHY 加 controller，外側各三顆 HBM。
來源：OCP 2026，Marvell 場次〈Custom HBM as a System〉；廠商論述，簡報註明引自分析師研究。

客製 HBM（cHBM）的主張是：把 PHY 與 controller 搬進客製的 base die，可回收至多約 25% 的面積、換成最多約 30% 更多的運算；代價是必須同時解 D2D 功耗與封裝熱。兩項面積數字與其他研究的差異，只是基準不同，並非互相矛盾：

議題	說法 A	說法 B	判讀
HBM 面積占比	Marvell：HBM PHY 加 controller 約占 XPU 30% 面積	SemiAnalysis：Rubin 約 16% 晶粒面積	對象與基準不同（XPU 整顆 vs 特定晶粒、PHY 範圍不同），不互相覆蓋
客製 HBM 面積節省	Marvell：整體至多回收約 25% 面積	ECTC 論文：host PHY 面積 −60%	−60% 是 PHY 本身，25% 是整顆晶粒；不衝突
HBM5 規格與時程	Marvell：最高 28Gbps、每堆疊約 100W，熱問題約 2028 年（8 層升 16 層）浮現；next-gen 每堆疊 5.3–7.2TB/s（二手引用）	創意：24–32Gbps、2029 年量產、32Gbps 約 90W、PHY 面積 14 降至 8mm²	皆為預測，並列；量產以創意 2029 為較明確的一項

HBM5 的熱是下一個瓶頸

Marvell 預估 HBM5 每堆疊約 100W，最熱點在 base die 上的 D2D PHY；功率密度由 2021 年約 0.48 W/mm² 升到 2026 年預估 1.25 W/mm²，I/O 超過 28Gbps 時預估超過 2.0 W/mm²（Marvell 自行估算，含估計成分）。創意的估算為 HBM4E 16Gbps 約 60W、HBM5 32Gbps 約 90W。兩家的熱數字同量級，但同為預測。解法方向是在 PHY 旁加矽質熱路徑：SK hynix 的 iHBM 在 D2D PHY 上方加矽塊（簡報標示熱阻約降 30%），產業構想另有在核心晶粒堆疊旁加 Heat Path Block。

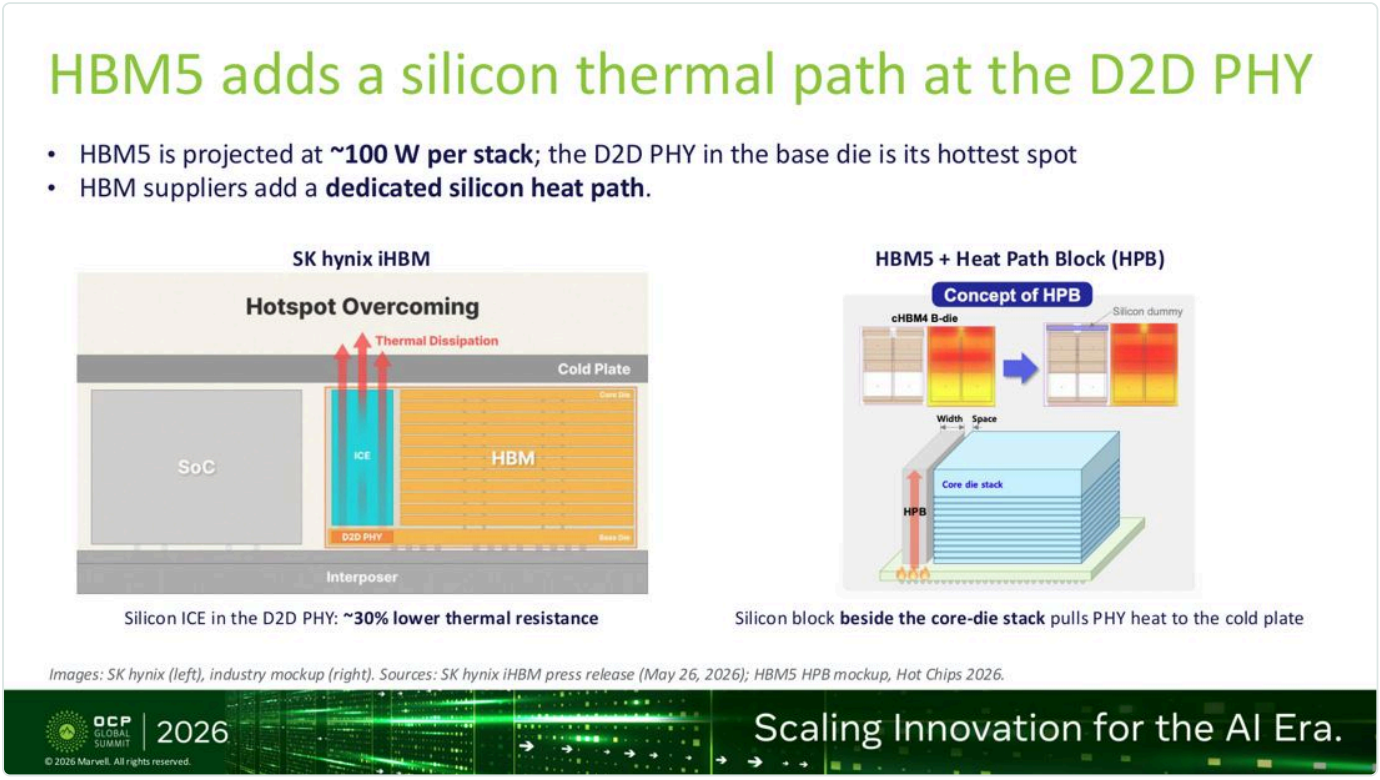


圖 13：HBM5 在 D2D PHY 處新增矽質散熱路徑。左：SK hynix iHBM 剖面，SoC 旁的 HBM 堆疊在 D2D PHY 上方有矽塊連到冷板（標示熱阻約降 30%）；右：HBM5 加 Heat Path Block 構想，核心晶粒堆疊旁加矽塊把 PHY 熱量導向冷板。文字稱 HBM5 約 100W / 堆疊，D2D PHY 為最熱點。
來源：OCP 2026，Marvell 場次〈Custom HBM as a System〉；圖源為 SK hynix 新聞稿與業界 Hot Chips 2026 模型，預測性質。

7　die-to-die 與 3D 整合：路線之爭

客製 HBM 之後的問題是 base die 與運算晶粒之間怎麼連。三家 IP / 晶片廠各有主張，全是 IP 賣方立場，標準尚未收斂：

- **Marvell：**認為 UCIe 並非為 HBM 級頻寬密度設計，推出 SBD（同時雙向）：每條線 64Gbps、PHY 面積約 15%、功耗最多降 75%（尖峰流量降 42%）、密度大於 7TB/s（簡報引自 2025 年 8 月新聞稿）。
- **創意：**主推自有 GLink，認為 UCIe 架構複雜、帶來額外功耗與延遲，多數場景不必遵循標準；GLink-LL 速率 4–8Gbps，功耗 0.05–0.10 pJ/bit，延遲 1–3ns，BER 小於 1E-30、不需 CRC 與重送。
- **聯發科：**簡報仍採 UCIe 與自家 Mlink 作封裝內 die-to-die 互連。

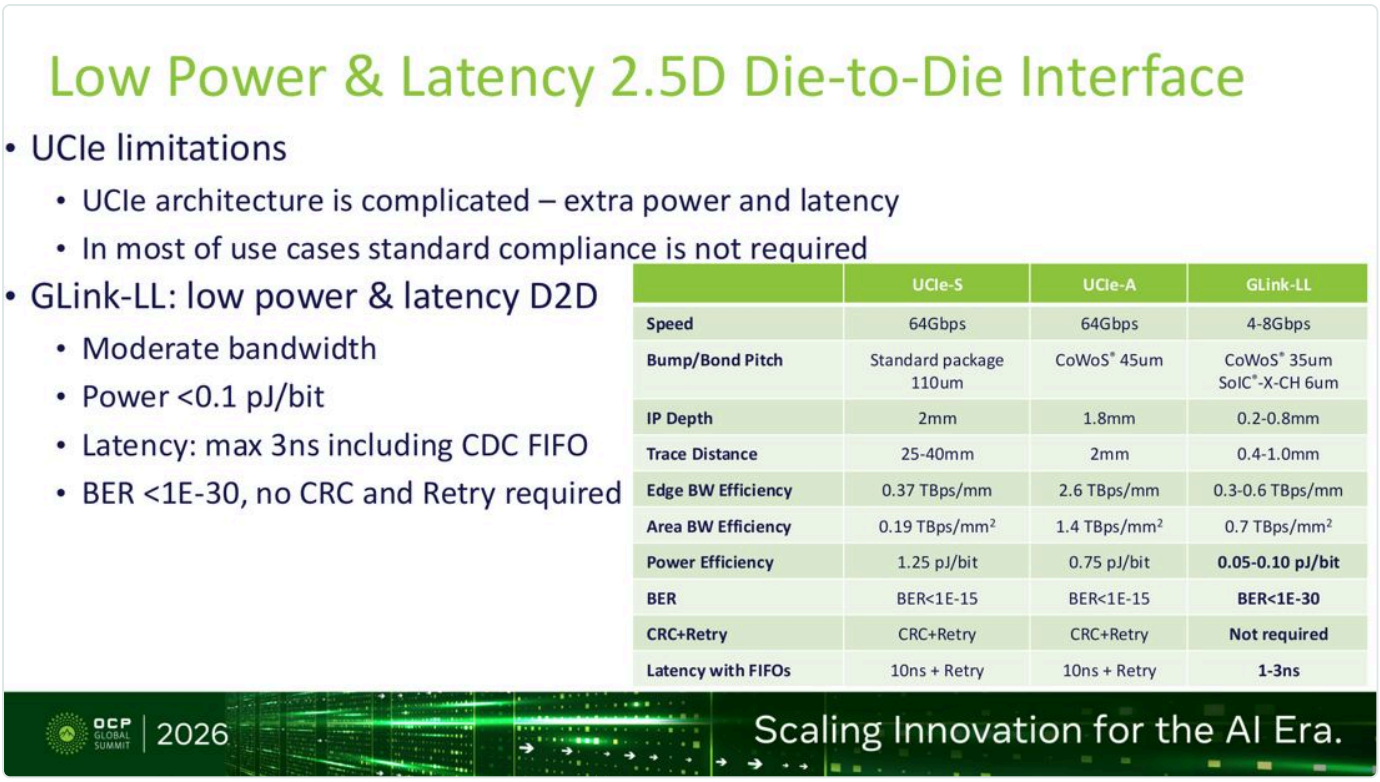


圖 14：2.5D 晶粒間介面比較（UCIe-S / UCIe-A / GLink-LL）。速率 64 / 64 / 4–8Gbps；凸塊間距 110μm / CoWoS 45μm / CoWoS 35μm 與 SoIC-X-CH 6μm；邊緣頻寬效率 0.37 / 2.6 / 0.3–0.6 TBps/mm；功耗 1.25 / 0.75 / 0.05–0.10 pJ/bit；BER 1E-15 / 1E-15 / 1E-30；GLink-LL 不需 CRC 與重送，延遲 1–3ns（UCIe 為 10ns 加重送）。
來源：OCP 2026，創意（GUC）場次〈Scaling up AI Compute, Memory and Connectivity〉；創意自行比較，由 IP 賣方提供。

GLink-3D 方面，創意列出四代：1.0（DDR，N6 / N5，0.2 pJ/bit，9 Tbps/mm²）、2.0（N2P，0.03 pJ/bit，50 Tbps/mm²）、2.1（SDR，N3P / N5，40 Tbps/mm²）皆標示 silicon proven，並稱已用於客戶 AI ASIC（客戶未具名，無法獨立驗證）；3.0（N2P，80 Tbps/mm²）仍在執行階段。但同一份簡報也將 N2P 列為設計挑戰，「silicon proven」的範圍宜保守解讀。

DRAM on Compute：把 DRAM 直接疊在運算晶粒上

創意進一步提出 DRAM on Compute Logic：HBM 以熱壓接合（TCB）做 3D 堆疊，間距約 20μm（HBM4E），因 TSV 密度低而在 bank 之間共享 TSV；DRAM on Logic 則用混合鍵合（5–8μm 間距），介面密度高 10–15 倍，每個 bank 有獨立寬匯流排。創意以「全光罩尺寸四層 DRAM」對「六顆 16Gbps HBM4E」比較：總頻寬 32 對 24 TBps、延遲 30 對 200ns、功耗 0.5 對 2.5 pJ/bit，但記憶體密度（四層 110 對十二層 310 MB/mm²）較低、最高溫 125°C 對 95°C，且成熟度與成本仍偏向 HBM。

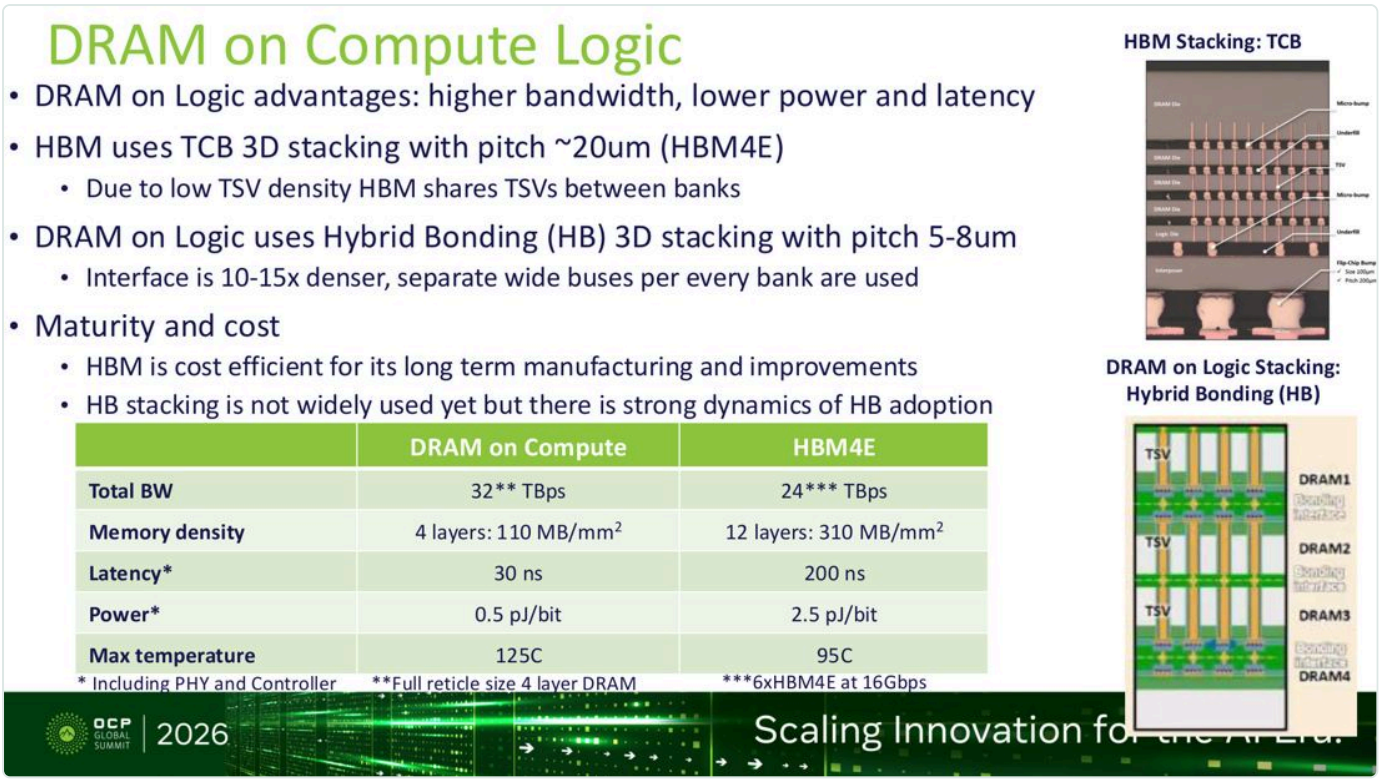


圖 15：DRAM on Compute Logic 與 HBM4E 比較。左下表：32 對 24 TBps 總頻寬、110 對 310 MB/mm² 密度（四層對十二層）、30 對 200ns 延遲、0.5 對 2.5 pJ/bit、125 對 95°C；右側為 TCB 微凸塊剖面（上）與混合鍵合剖面（下）。
來源：OCP 2026，創意（GUC）場次〈Scaling up AI Compute, Memory and Connectivity〉；創意自行比較，密度比較基準層數不同。

創意場次另含兩項關聯內容：TSMC COUPE（共封裝光學，CPO）光引擎的測試流程與含 CPO 的 GPU 托盤實物（托盤上可見 AyarLabs 與緯穎 wiwynn 標示）；以及 System on Wafer（SoW）——一片晶圓大於 40 倍光罩尺寸、最多 64 顆 HBM4E、TDP 20kW、峰值供電 30kW。這兩項屬於展示與規格，本報告不當作訂單。

8　資料中心 ASIC：聯發科的證據

聯發科的簡報是這批簡報中台廠最具體的一手證據。簡報用劃線更正的方式顯示：2026 年資料中心營收由「1 Billion（USD）」劃掉改為「2 Billion（USD）」；TAM 為 700–800 億美元，年度由 2028 劃掉改為 2027；聯發科預期市占 15–20%。時間軸則列出 2026 年「首個量產設計 Q4 出貨」，下方 tape-out 節點為 Intel I16、TSMC N3P、N2P 與 A14（A14 落在 2026 欄為排版判讀）。



圖 16：聯發科資料中心營收與 TAM（公司簡報，含手改痕跡）。2026 年由 1 Billion（USD）劃掉改為 2 Billion（USD）；TAM 700–800 億美元，年度標示 2028 被劃掉、改為 2027；聯發科市占 15%–20%；底部為晶圓、封裝、XPU、運算板、托盤與機櫃產品示意。來源：OCP 2026，聯發科場次〈The Silicon Heartbeat of the AI Economy〉；公司自述的預測。

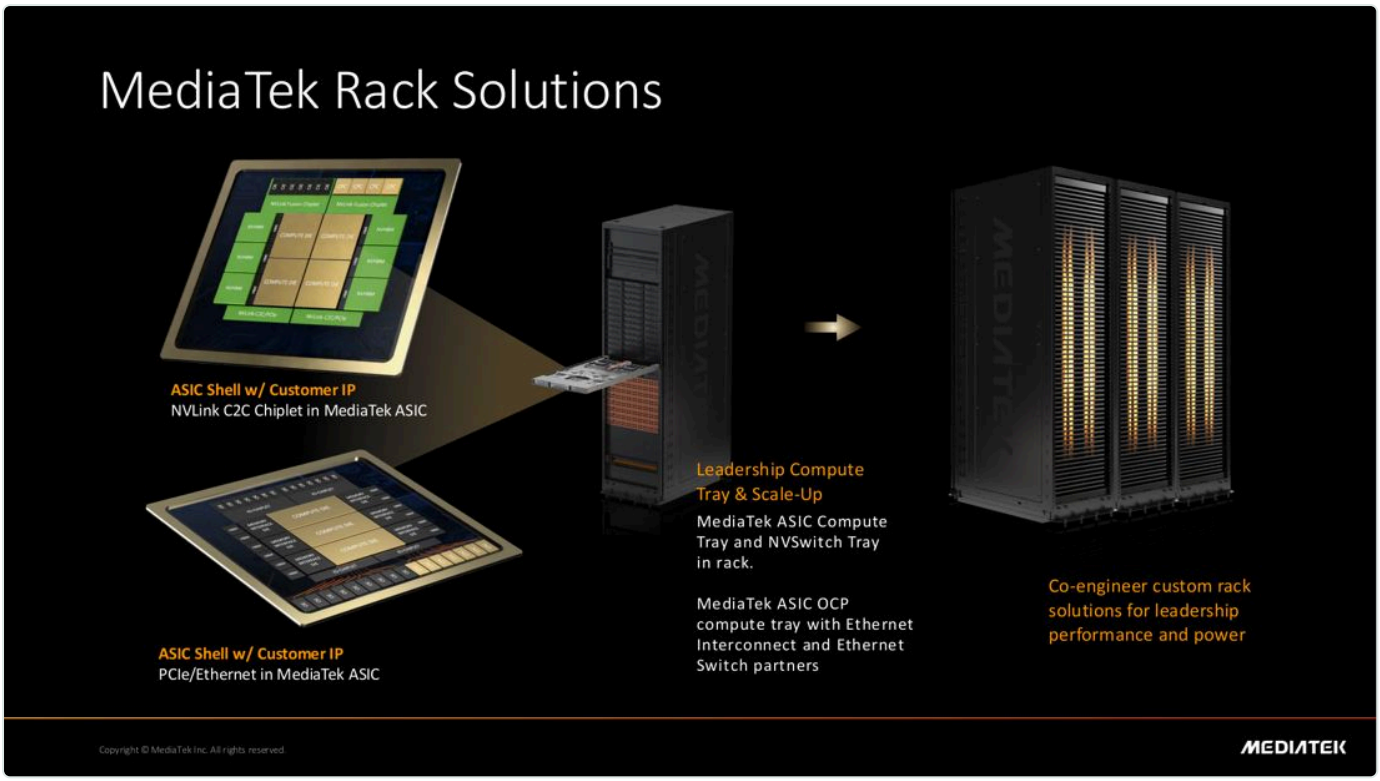


圖 17：聯發科機櫃方案（展示圖）。上方為兩種 ASIC shell 封裝圖：含客戶 IP 的 NVLink C2C Chiplet 版，與含客戶 IP 的 PCIe / 乙太網路版；中間為「MediaTek ASIC 運算托盤加 NVSwitch 托盤」同櫃設計，另有 OCP 運算托盤搭乙太網路互連；右側為共同設計客製機櫃。
來源：OCP 2026，聯發科場次〈The Silicon Heartbeat of the AI Economy〉；展示圖，客戶未具名。

項目	簡報所示	證據強度與限制
2026 資料中心營收	20 億美元（由 10 億上修）	展示 / 自述 公司自述預測，非財報；僅有簡報上的手改痕跡
首個量產設計	2026 年第 4 季出貨	展示 / 自述 客戶未具名；不可連到任何特定客戶或產品
NVLink Fusion ASIC shell	含客戶 IP 的 NVLink C2C Chiplet 版 ASIC 外殼	展示 2025 年起與 NVIDIA 合作；非訂單
機櫃共同設計	ASIC 托盤搭 NVSwitch 托盤同櫃；可共同設計客製機櫃	展示 展示圖，非訂單；與某券商「不推自家機櫃」的說法有出入，偏向「可談包整機架」
記憶體方案	SRAM、LPDDR / GDDR、近記憶體運算、標準與客製 HBM、3D 記憶體堆疊五類	展示 能力列示，無客戶與時程
互連	封裝內 UCIe / Mlink；板上高速 SerDes PCIe 與 uLED；板外乙太網路 SerDes、CPC 連接器與 CPO	展示 與創意 GLink、Marvell SBD 路線分歧
與台積電	列為早期技術合作夥伴（先進節點、DTCO、CPO、SoW）	合作開發 簡報自述

一項不應引用的說法

某券商報告曾轉述「聯發科 3.5D 約 10 倍 die-to-die 能效」。本報告回查聯發科簡報的第 11、12 頁，未找到此數字，因此不當作簡報事實引用，僅標記為口述或其他來源。

9　Samsung 2030 願景（短）

Samsung 的 2030 年推論資料中心願景把上述所有層級放進一張圖：L0 SRAM / zHBM、L1 HBM5 / 3D DRAM、L1.5 光學 HBM、L2 LPDDR / SOCAMM、L2.5 CXL 池 / zNAND-O、L3 KV SSD、L3.5 乙太網路 SSD、L4 TLC / QLC SSD，由光連結把 Memory Pod 接在一起，容量朝 exabyte 乃至 zettabyte 邁進。要點是：這是願景，圖中雖有各層延遲與容量的量級（例如 L2.5 為數百 ns、數百 GB 至 1TB），但沒有任何產品規格與年份，圖中 L0.5、L2.5 還有重複標示。可用來理解「業界想把記憶體層級拆成多細」，不可當作產品時程。同一家公司的產品面較具體（CMM-D、PM1763 / PM1773、FDP、zHBM 對 HBM5 頻寬 4–8 倍、能效 3 倍的廠商值），但除 2027 年 CXL 池路線外也都無量產日期。

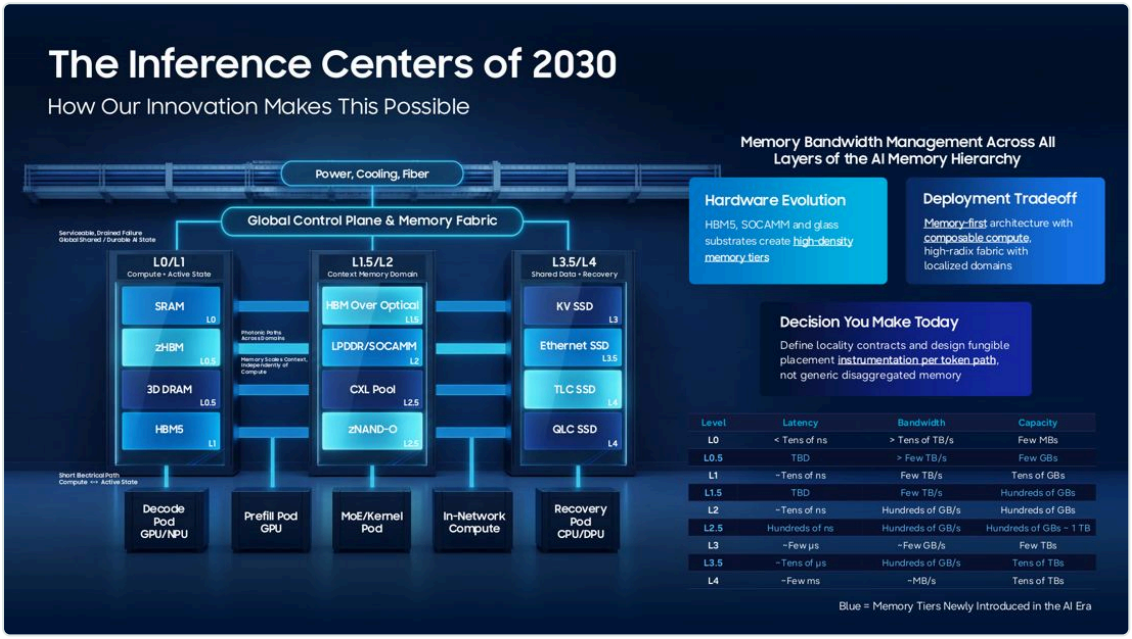


圖 18：Samsung 2030 推論資料中心記憶體階層（願景）。L0–L4 各層與對應 Decode / Prefill / MoE / 網路內運算 / Recovery Pod；右下表列各層延遲、頻寬、容量的量級；藍色為 AI 時代新增的記憶體層。
來源：OCP 2026，Samsung 場次〈Inference Data Center 2030: Memory Pods Fabric with Rightsized Compute〉；願景，無規格。

10　台股供應鏈映射與觀察清單

本章把上述證據映射到台股。分級標準：實際出貨（有客戶與出貨證據）、合作開發、展示、僅 logo、推論（簡報未點名，由架構邏輯推得）。本批簡報中，沒有任何一家台廠達到「實際出貨」的證據等級。

公司	代號	角色	證據強度	為何相關 / 為何打折	觀察指標
聯發科	2454	資料中心 ASIC、NVLink Fusion ASIC shell、客製機櫃共同設計	展示 加公司自述	簡報有營收上修痕跡與首個量產設計 Q4 出貨；客戶全未具名、機櫃為展示圖，不得連到特定客戶	2026 Q4 首個量產設計出貨；法說對資料中心營收的正式口徑
創意電子	3443	GLink-LL / 3D、HBM4E 16Gbps PHY / controller、DRAM on Compute、IVR、COUPE 測試、SoW 平台	展示 加公司自述	規格最完整；稱 GLink-3D 已用於客戶 AI ASIC（客戶未具名）；HBM5 規劃 2029 年量產；但與 Marvell SBD、UCIe 路線競爭	COUPE 與 GLink-3D 的客戶揭露；HBM4E 導入案
台積電	2330	CoWoS、SoIC-X、COUPE、SoW，先進節點	合作開發	聯發科與創意簡報均列為平台與夥伴；無新衝突	COUPE tape-out 與 SoW 進度
緯穎	6669	GPU 托盤展示件	僅 logo / 展示	僅出現在創意簡報的托盤照片標示（wiwynn）；不構成訂單	含 CPO 托盤是否進入出貨
群聯	8299	eSSD 控制器（FDP、ZNS、遙測、平行度）	推論	FDP / 遙測 / 平行度成為規格門檻，對韌體強的控制器廠有利；但簡報未點名，Samsung / SK hynix 自有控制器壓縮空間	FDP 在 eSSD 規格的普及；非三大廠 NAND 客戶的企業級導入
慧榮（Silicon Motion，美股 SIMO）	—	eSSD 控制器韌體（Meta 要求 FDP、GC / WAF 遙測）	推論	同群聯邏輯；簡報未點名，勿推論其受益或受損	同上
威剛	3260	DRAM / SSD 模組與通路	推論	高密度 QLC 取得難（Meta 失效率差距大）、Samsung PM1773 自有品牌，對模組廠為雙面；無證據	企業級 QLC 供貨與產品線
宜鼎	5289	工控 SSD / DRAM 模組	推論	同威剛；為雙面影響，簡報未點名	企業級 SSD 與 CXL 模組產品
南亞科	2408	標準 DDR5 / LPDDR	推論	受益容量擴張敘事（CXL DIMM、SOCAMM2）；zHBM / 客製 HBM 為三大廠領域，打折	DDR5 / LPDDR 在 CXL 與 SOCAMM2 的導入

非台廠的競爭與意涵

- **Marvell**：同時是客製 HBM / D2D IP（SBD）與 CXL（Structera）的供應者，與創意在 D2D 與 HBM 介面 IP 直接競爭；它的「UCIe 非為 HBM 級密度設計」觀點，同時挑戰創意 GLink 與聯發科 UCIe 路線。SK hynix 的實驗使用 Marvell 的 Structera，Samsung 與 Marvell、Liquid 共建 CXL appliance，其在 CXL 池側的曝光度高於其他控制器廠（本批未見 Astera Labs 場次，不代表其缺席）。
- **三大記憶體廠**：Samsung、SK hynix、Micron 均往 CXL 池、客製 HBM base die 與 KV SSD 擴張，台廠的控制器與模組空間被垂直整合壓縮（推論）。

- **NVIDIA：**CMX 被 Samsung 稱為 KV 快閃層（每 GPU 16TB，口述），並透過 NVLink Fusion 把聯發科 ASIC 拉進同機櫃生態。

觀察清單

追蹤指標	預期時點	為什麼重要
HBF 規格 1.0 發布	2027 年 1 月（預期）	確認 SK hynix 時程是否延後，並比對 GS 的 2027 曆年樣品說法
Samsung CMM-D MD320 進度	未給日期	CXL 池由 POC 走向量產的最近訊號；2027 年次世代 20TB 池是否如期
創意 COUPE 與 GLink-3D 客戶揭露	COUPE tape-out 約 2026 年底（券商說法）	由「自述已用於客戶」變成可查證的客戶 / 量產
聯發科首個量產設計出貨與營收修正	2026 年第 4 季	驗證 2 億美元級別的上修是否轉成財報
Meta QLC 導入規模	—	目前無規模與廠商；若出現追加訂單或規模揭露，才有條件評估 eSSD 控制器與模組廠
FDP 在 eSSD 規格中的普及	—	決定外部控制器廠是否真有進入門檻差異

風險與反證條件

風險 / 反證	為何會推翻論點	看什麼訊號
CXL 池量產延後或僅停留在 POC	主論點「分層向 CXL 池擴張」被推遲，Marvell、Samsung 池規格落空	2027 年 Samsung 池是否出樣；SK hynix MARS 是否由「客製試點」進入客戶部署
SSD 耐久無法滿足，KV 停留在 DRAM / CXL	若 QLC 在 KV 寫入量下無法通過耐久，SSD 層的容量需求落空	各家 DWPD 規格與實測；FDP 實際 WAF 數字能否在客戶環境重現
HBM 容量提升削弱分層需求	HBM4E / HBM5 容量與 zHBM 若快速成長，部分 KV 可不外溢	下一代 GPU 的 HBM 容量；容量門檻（Micron 約 400GB）與並行度的實況
垂直整合排擠外部控制器	Samsung、SK hynix 自有控制器吃掉 eSSD 增量，群聯、慧榮的機會落空	三大廠 eSSD 在 hyperscaler 的占比；非三大廠客戶的 FDP 導入
D2D 標準收斂到 UCIe	若 UCIe 成為事實標準，創意 GLink、Marvell SBD 的差異化縮小	客戶 ASIC 採用 GLink / SBD / UCIe 的揭露；UCIe 新版對 HBM 級密度的回應
聯發科營收修正不如預期	簡報僅為公司自述預測，客戶未具名；出貨與認列可能延遲	Q4 法說的資料中心營收口徑；首個量產設計是否如期出貨
數字口徑風險	自測數字缺乏獨立驗證，且模型、硬體、基線各異	第三方測試或客戶公開的實測

來源清單

OCP 2026 峰會場次（2026 年 10 月）：

- SK hynix：〈Introduction to Memory-centric AI Rack System (MARS)〉（Hoshik Kim）；〈Storage as a Performance Tier for LLM Serving〉（Joohee Lee、Donguk Moon）；〈Agentic AI Serving

System with Computational Shared Memory Pool〉（Jongryool Kim）

- Samsung：〈Tiering the Agent's Memory: From CXL Pooling to SSD-Backed KV Cache Offloading〉（影片場次，依字幕整理）；〈Keep Your Compute Running at Full Throttle with Highly Scalable Memory〉（影片場次，依字幕整理）；〈Inference Data Center 2030: Memory Pods Fabric with Rightsized Compute〉
- Marvell：〈XPU Scale Composable Smart Memory Fabric〉；〈Custom HBM as a System: Co-designing the Base Die, Die-to-Die Power, and Package Thermals for AI Accelerators〉
- Meta：〈CXL in Hyperscale Environment: Fleet-Scale Challenges, Failure Modes & the Path to Better Debuggability〉；〈QLC @ Hyperscale〉
- Micron：〈Tiered Memory for LLM Inference — A Modeling Framework for KV-Cache Offload〉
- 創意（GUC）：〈Scaling up AI Compute, Memory and Connectivity〉
- 聯發科：〈The Silicon Heartbeat of the AI Economy: The Efficiency Paradox in Next-Gen Data Center Silicon〉

交叉對照的券商與產業研究（僅取方向、不引用原文）：高盛（HBF 樣品時程）、凱基（HBF 量產時程）、摩根士丹利（聯發科 3.5D 能效說法）、SemiAnalysis（HBM 面積占比）、永豐（聯發科機櫃方向）、UBS（COUPE tape-out 時程）；ECTC 論文（host PHY 面積）。

資料處理說明：Samsung 兩場為影片場次，無投影片，數字依講者口述字幕整理，部分名詞與型號未經原廠覆核；SK hynix、Marvell、Meta、Micron、創意、聯發科、Samsung 2030 場次以簡報頁面為準。圖片均取自上述簡報頁面，僅作研究說明引用。