新光投顧

〔投資研究報告〕

2026/5/13

HBM 開始排座位：AI 先拿貨，終端再排隊

目錄

高頻寬記憶體改寫記憶體業的供給排序

2026/5/13 新光投顧

高頻寬記憶體過去被市場視為記憶體景氣循環中的高價產品，投資重點多半放在平均售價、產品組合與毛利率改善。這個框架已經不夠。高頻寬記憶體正在從「高階記憶體品項」，升級為人工智慧算力供應鏈的產能優先權制度。取得產能的客戶，可以把晶片設計、先進封裝與資料中心建置往前推進；未取得產能的客戶，即使晶片平台準備完成，也可能卡在可交付算力之前。

產品世代的校準很重要。輝達 Vera Rubin 平台採用的是 HBM4；官方技術資料指出，Rubin 圖形處理器最高配置 288GB HBM4，記憶體頻寬最高可達 22TB/s，且 HBM4 相較 HBM3E 介面寬度翻倍。Hopper 主要使用 HBM3 / HBM3E，Blackwell 使用 HBM3E，Vera Rubin 則進入 HBM4 世代。這代表高頻寬記憶體的投資判讀不能停留在「高價記憶體」四個字，而要追蹤世代切換、客戶驗證、晶圓配置與封裝能力。

合約型態也在改變。TrendForce 引述韓媒報導指出，三星與 SK 海力士正從一年期短約，轉向與全球大型科技客戶簽訂三至五年的長期供應協議；部分協議還討論預付款、最低供應量與價格保護機制。這代表大型人工智慧客戶正在把記憶體供應提前鎖進晶片設計與資料中心建置計畫，而非只在季度報價中採購。

這輪變化的本質，不只是高頻寬記憶體漲價。更值得重估的是晶圓配置與客戶排序。當高頻寬記憶體、伺服器記憶體與企業級儲存取得較高產能優先權，手機、個人電腦、邊緣裝置與模組通路面臨的壓力，將從短期缺貨轉向長期順位下降。記憶體產業的主軸，正由庫存循環進入配給規則重寫。

目錄

一、 從報價表到配給表：HBM 開始決定誰先拿到算力

供應席位先於報價。

傳統記憶體市場的核心變數，是庫存與價格。原廠擴產、通路拉貨、終端庫存、現貨報價與合約價，共同構成景氣循環。但高頻寬記憶體帶來新的交易方式。大型人工智慧客戶不再只看季度報價，而是透過多年期協議提前取得供應席位，將記憶體產能納入整個算力交付計畫。

SK 海力士的公開訊號應精準解讀。較明確的重點集中在高頻寬記憶體產能：公司高層曾表示 2025 年高頻寬記憶體產能已銷售一空，2026 年產能也預計快速被客戶鎖定；同時，SK 海力士已供應 12 層堆疊 HBM3E，並開始提供 HBM4 樣品。這支持「高頻寬記憶體產能緊俏」的判斷，但不宜直接擴大成「所有傳統動態隨機存取記憶體與快閃記憶體產能全面售罄」。

分配權取代單純漲價。

高頻寬記憶體的稀缺來自兩個層面。需求端，人工智慧加速器需要極高頻寬與堆疊容量，每一代平台對容量、頻寬與功耗的要求同步提高。供給端，高頻寬記憶體需要先進動態隨機存取記憶體晶粒、矽穿孔、堆疊、先進封裝、測試與良率控制，無法像一般記憶體產品那樣快速切換產出。

因此，原廠手中的稀缺資產不只是報價能力，更是晶圓與封裝資源的分配權。當客戶願意用長約、預付款、價格保護或最低採購量交換未來供應，記憶體產業的商業模式會比過去更接近容量預訂制度。報價仍重要，但供應權本身開始具有金融價值。

設計階段已鎖記憶體。

TrendForce 指出，自研人工智慧晶片專案中，記憶體與先進封裝等關鍵規格與數量，常在設計階段就被提前鎖定。這是長期供應協議加速普及的原因之一。這會改變記憶體公司的談判位置。過去記憶體原廠面對終端景氣循環，容易在下行週期被迫降價；現在，若高頻寬記憶體被寫進客戶平台設計、資料中心上線與多年資本開支計畫，原廠取得更長營收能見度。但長約不等於無風險收入保證。若客戶的人工智慧資本開支或商業變現不如預期，延後拉貨與重新議價仍可能發生。

二、 AI 插隊效應：手機、PC 與模組廠的成本壓力重排

人工智慧記憶體擠壓先進產能。

高頻寬記憶體的產能排擠，不只發生在單一產品線。它需要先進晶粒與高階封裝測試資源；伺服器記憶體與企業級儲存，也因資料中心需求取得更高產能順位。原廠在有限資本支出與有限先進產線下，會優先配置給毛利率高、客戶承諾強、平台黏著度高的產品。

這使手機、個人電腦與部分消費性終端的供給順位下降。即使這些終端需求沒有同步大幅成長，也可能因人工智慧伺服器搶占產能而承受成本上升。記憶體價格上行不再只代表終端需求轉強，也可能代表非人工智慧終端被擠到較後面的產能隊伍。

行動記憶體壓力已經浮現。

TrendForce 對 2026 年第二季行動記憶體的觀察指出，智慧手機用行動記憶體合約價延續強勢；2026 年上半年連續兩季高額漲幅，使品牌成本壓力難以消化，智慧手機產量面臨下修，原先談定的長期採購位元量也可能無法達成。報告同時提到，在人工智慧伺服器持續搶占產能、動態隨機存取記憶體高價難解的背景下，品牌需要從軟體與系統架構減少記憶體需求，並強化雲端服務以對沖升容壓力。

這段資料的含義很清楚。手機品牌面臨的壓力，是供應順序改變後的結構性成本轉嫁問題。若人工智慧伺服器長期占據高階動態隨機存取記憶體資源，行動記憶體、個人電腦記憶體與通路模組的供需彈性會下降。

終端分化大於全面缺貨。

成本轉嫁能力將決定下游表現。高階手機、人工智慧個人電腦與高階邊緣裝置，仍可透過更高記憶體容量、更強端側人工智慧功能與品牌溢價吸收部分成本。中低階手機、入門筆電、白牌裝置與通路模組，更容易被成本上升壓縮毛利。

模組廠的壓力尤其直接。若原廠優先服務高頻寬記憶體、伺服器記憶體與大型客戶長約，模組廠取得穩定料源的難度上升。採購成本上行、客戶議價能力有限、庫存週轉風險增加，將使模組廠的盈利品質更不穩定。這輪記憶體漲價對

供應鏈各層的含義不同：上游原廠拿到分配權，中游模組廠面臨料源與毛利夾擊，下游品牌則依產品定位決定成本轉嫁成敗。

三、 記憶體走向 K 型價格：高階缺產能，低階拚供給

先進規格與成熟規格分道而行。

非人工智慧終端面臨的成本壓力，不能簡化成全面性記憶體缺貨。更精準的框架，是記憶體規格進入 K 型分化。DDR5、LPDDR5X、高頻寬記憶體與高容量伺服器記憶體，受到人工智慧伺服器與高階終端需求拉動，且與先進晶圓、先進封裝與高階測試資源存在競爭。這一側的成本壓力，對人工智慧個人電腦、旗艦手機與高階邊緣裝置最明顯。

成熟規格則呈現另一種邏輯。DDR4、LPDDR4X 與部分成熟節點記憶體，受到韓系、美系原廠轉產與減供影響，短期可能出現供給收縮與價格上行；但長鑫存儲等中國記憶體廠，在補貼、在地化供應鏈與行動記憶體需求支撐下，正加速擴大 LPDDR4X 與相關成熟規格市占。TrendForce 指出，長鑫存儲在 2026 年位元產出增速續居全球第一，其主力產品為行動記憶體；中國補貼政策與韓美原廠轉產造成的 LPDDR4X 供應缺口，也有助於其擴大市占。

成熟規格不是單純便宜。

成熟規格的定價由兩股力量拉扯：韓美原廠退出或減供推升短期價格，中國供給擴張則提高中期價格競爭與庫存重估風險。對終端品牌而言，這是供給來源、價格穩定度與庫存週轉風險同時上升。

高階終端的問題，是買得到但成本變高。人工智慧個人電腦、旗艦手機、高階邊緣裝置，需要 DDR5、LPDDR5X 或更高階規格，這些產品與人工智慧伺服器在先進產能與供應順位上競爭，品牌端承受較高物料成本與轉嫁壓力。
中低階終端的問題，則是規格遷移與供給來源重組。使用 DDR4、LPDDR4X 的入門手機、低階個人電腦與白牌設備，可能在短期遭遇韓美原廠減供造成的價格壓力，但中期會面對中國供應商擴產與價格競爭。採購策略變得更複雜：短期怕缺料，長期怕跌價；高階料怕買不到，低階料怕庫存重估。

模組廠承受雙重夾擊。

模組廠的壓力因此更立體。高階規格方面，DDR5、LPDDR5X 與伺服器記憶體

料源順位較低，採購成本高、供應不穩；成熟規格方面，若中國供給快速釋放，庫存評價與價格競爭風險上升。模組廠可能同時面臨「高階買得貴、低階跌得快」的雙重夾擊。

這也修正了「手機與個人電腦物料成本全面上升」的說法。更精準的表述應是：高階終端承受先進規格成本上升，中低階終端承受成熟規格供給重組與價格波動，模組廠則同時暴露在高階料源不足與低階庫存跌價風險之下。

四、 配給制度也有破口：長約、良率與推論架構的三道折價

長約不是鐵板收入。

高頻寬記憶體長期供應協議可以提高營收能見度，但它不是基礎設施領域常見的照付不議合約。記憶體歷史中，長約在上行週期看似保護原廠；一旦下游需求反轉，客戶延後拉貨、重新議價、調整規格或支付有限違約金的案例並不少見。

目前的不同之處，在於人工智慧客戶集中度高。少數大型雲端服務商與晶片客戶掌握需求端議價權。當人工智慧應用商業化順利，長約會強化原廠供應優先權；若資本開支被迫放緩，長約的保護效果可能低於市場假設。買方模型中，高頻寬記憶體長約應視為能見度提升，而非無風險收入保證。

較保守的做法，是追蹤預付款比例、最低採購量、價格下限、取消條款、延後拉貨條款與客戶集中度。合約結構越接近共同投資與產能保留，原廠保障越高；若只是框架協議或彈性採購承諾，保護力要打折。

良率突破會釋放供給。

高頻寬記憶體的供給瓶頸，部分來自堆疊、矽穿孔、先進封裝與測試良率。當良率偏低時，晶圓投入量對有效產出有放大損耗，也會擠壓傳統記憶體產能。但三星、SK 海力士與美光正在投入大量資本支出，擴充矽穿孔、先進封裝與測試能力。若製程學習曲線在 2026 年後跨過拐點，有效產出可能快速增加。

這是高頻寬記憶體稀缺敘事的第一個技術折價。只要高毛利產品持續存在，記憶體產業歷史上常見的結果是過度投資。即使高頻寬記憶體技術門檻高於傳統

產品，新增產能與良率改善仍會逐步釋放供給。一旦稀缺性下降，分配權溢價會被壓縮，原廠可能重新被市場放回景氣循環股估值框架。

因此，投資人不能只追蹤高頻寬記憶體需求，也要追蹤有效產出。關鍵指標包括堆疊層數良率、矽穿孔產能、先進封裝產能、測試產能、HBM4 放量時程，以及預訂產能到實際出貨的轉化率。

推論架構可能降低排擠強度。

第三個折價來自人工智慧架構演進。訓練階段高度依賴高頻寬記憶體，這是物理與系統設計上的結果。但人工智慧週期逐步從訓練走向推論，雲端大廠自研晶片也會重新優化記憶體架構。

輝達 Vera Rubin 官方技術資料本身就提供了一個例子。Vera CPU 搭配最高 1.5TB LPDDR5X，並透過第二代 NVLink-C2C 與 Rubin 圖形處理器的 HBM4 形成一致性記憶體池，讓應用可在低功耗 LPDDR5X 與高頻寬 HBM4 之間降低資料搬移與提升執行效率。這並沒有削弱高頻寬記憶體的地位，但說明未來平台會更積極混用不同記憶體層級，而非所有增量都由高頻寬記憶體單獨承擔。對推論應用而言，成本、功耗、延遲與容量配置同樣重要。部分邊緣推論、自研加速器與專用推論架構，可能更多採用 LPDDR5X、未來 LPDDR6、大容量片上靜態隨機存取記憶體，或透過壓縮、快取與模型架構優化降低對高頻寬記憶體的邊際需求。若推論增量不再與高頻寬記憶體需求一比一綁定，對整體動態隨機存取記憶體產能的排擠效應會低於最樂觀假設。

看盤順序要重排。

後續觀察指標應分成四組。

第一組是高頻寬記憶體合約與長約品質，包括價格區間、預付款、最低採購量、價格下限、取消條款與客戶集中度。

第二組是晶圓與封裝配置，包括高頻寬記憶體、伺服器記憶體、行動記憶體與個人電腦記憶體的產能占比，以及先進封裝與測試產能。

第三組是下游報價，包括 LPDDR5X、未來 LPDDR6、個人電腦記憶體、企業級固態硬碟、DDR4、LPDDR4X 與通路模組價格。

第四組是成本轉嫁，包括智慧手機出貨目標、人工智慧個人電腦搭載容量、模組廠毛利率與通路庫存。

總結來看，高頻寬記憶體已經是人工智慧供給優先權制度。它改變了晶圓配置、合約模式、伺服器與行動記憶體供給排序，也改變了手機、個人電腦、模組廠與邊緣裝置的成本結構。這套制度在 2025 至 2026 年具備高確定性，但投資人仍需對四個風險保持紀律：長約保護力有限、良率突破可能釋放供給、推論架構可能降低高頻寬記憶體的邊際排擠效果，以及中國成熟規格供給擴張帶來的 K 型定價。最值得重估的公司，是掌握高頻寬記憶體分配權、世代切換能力與客戶長約品質的原廠；最容易承壓的環節，則是高階料源順位下降、低階庫存跌價風險上升、且缺乏成本轉嫁能力的終端與模組供應鏈。