April 2026

iSTART iReport

芯測科技電子報第 15 期

芯測科技 IEEE 1838 解決方案到位
搶灘 AI Chiplet 高階封裝商機

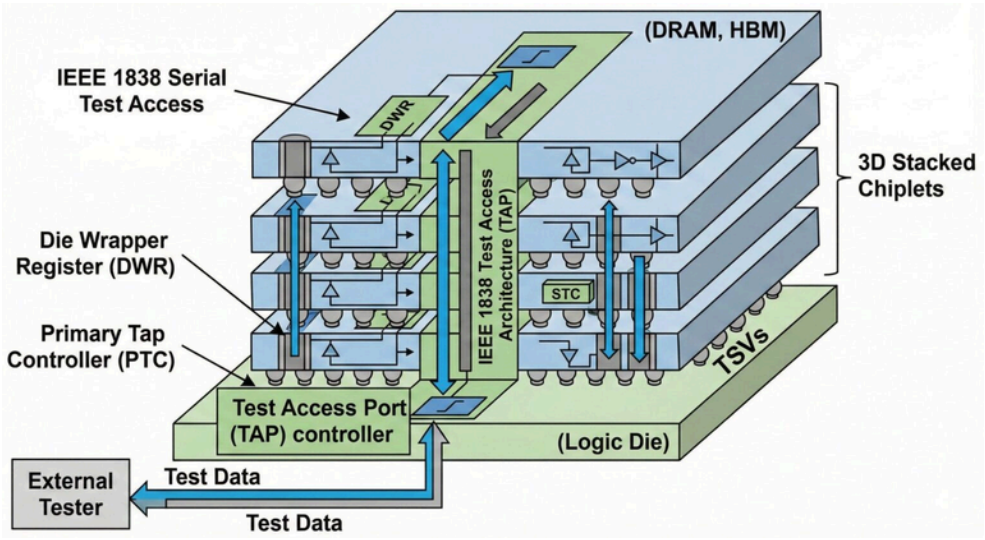
AI 運算需求快速攀升，半導體產業正全面邁向以 3D-IC 與 Chiplet 為核心的先進封裝時代。高頻寬記憶體 (HBM) 與邏輯晶片的垂直堆疊，已成為支撐 AI 加速器效能與能效的關鍵技術路線。其中以 CoWoS 為代表的 2.5D/3D 封裝架構，雖能大幅提升系統整體效能，卻也突顯了測試、修復與良率控管上前所未有的挑戰。

專注於記憶體測試與修復技術的芯測科技 (iSTART-TEK) ，宣布其 IEEE 1838 3D-IC 測試解決方案已進入成熟應用階段，是目前少數能完整對應 3D-IC 測試標準、並實際落實於先進封裝專案的 EDA 供應商之一，提前卡位 AI Chiplet 與 HBM 高階封裝商機。

率先佈局 IEEE 1838，建立 3D-IC 測試門檻

IEEE 1838 為專為 3D-IC 與異質整合架構制定的測試標準，涵蓋 die-to-die 測試存取、測試通道重組、以及堆疊後的可測試性設計。然而由於技術門檻高、整合複雜度大，目前市場上能成熟支援 IEEE 1838 的 EDA 工具選項相當有限。

芯測科技早在 3D-IC 尚未全面商用之前，便已投入相關架構與測試流程的研發，將 IEEE 1838 納入其記憶體測試與修復平台中，並完成與既有 BIST、BISR 機制的整合。此率先佈局不僅讓芯測科技在 3D-IC 測試領域建立技術護城河，也大幅提高後進者的進入門檻。



深度對應異質整合，滿足 HBM 與邏輯晶片堆疊需求

AI 晶片的效能瓶頸，已不再僅限於運算單元，還高度取決於 HBM 與邏輯晶片之間的資料傳輸效率。為此，HBM 與 GPU、AI 加速器、客製化邏輯晶片的垂直整合，成為主流設計方向，也讓測試架構必須同時理解「記憶體」與「邏輯」的堆疊關係。

芯測科技的 IEEE 1838 解決方案，即是針對 HBM 與邏輯晶片的異質整合情境所設計，可支援多顆 die 堆疊後的測試路徑規劃、測試存取管理，以及記憶體層級的故障定位與修復策略。透過此架構，客戶得以在 CoWoS 等先進封裝流程中，保有對 HBM 與 SRAM 的可測試性與可修復性，確保 AI 系統的效能與可靠度。

降低 CoWoS 報廢風險，放大先進封裝投資回報

在 CoWoS 與 3D-IC 架構下，一顆整合 HBM 與高階邏輯晶片的成品，往往代表極高的製造成本。一旦在封裝後才發現記憶體或互連缺陷，若無有效的測試與修復機制，報廢損失將相當可觀。

芯測科技透過 IEEE 1838 與其記憶體修復技術的整合，協助客戶在多 die 堆疊完成後，仍能進行精準的故障診斷與修復，大幅提升最終良率。這不僅直接降低先進封裝的報廢成本，也讓客戶在導入 CoWoS 與 AI Chiplet 架構時，能有效放大整體投資報酬率 (ROI) 。

芯測全流程 eFlash BIST IP 服務
把關 AI 伺服器關鍵元件品質

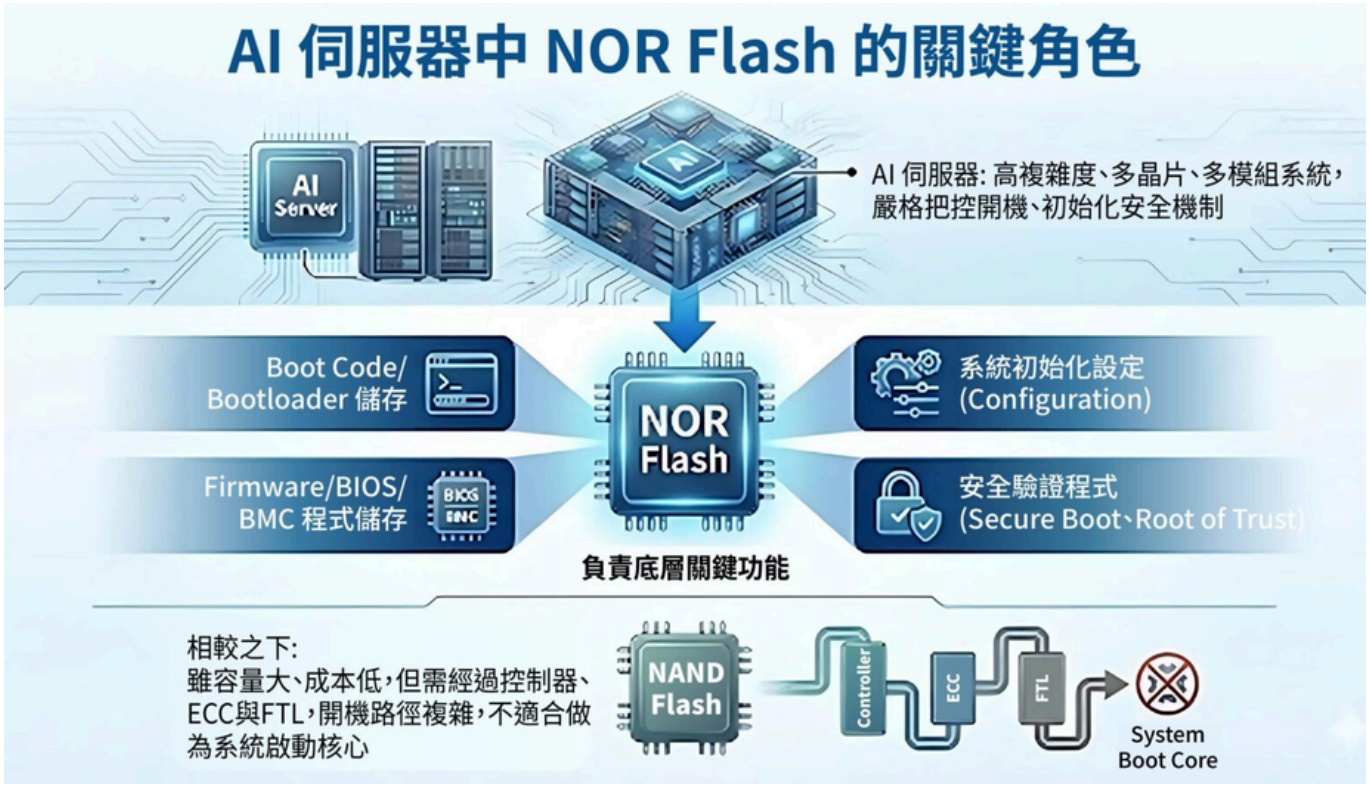
生成式 AI 與大型語言模型 (LLM) 爆發性成長，全球資料中心對於 AI 伺服器的效能要求已達前所未有的高度。然而在追求算力的同時，「系統穩定性」與「資料可靠性」成為了隱形的技術天花板。記憶體測試與修復領導廠商芯測科技分析，AI 伺服器中的關鍵元件 NOR Flash 雖看似微小，卻是啟動與安全的核心。

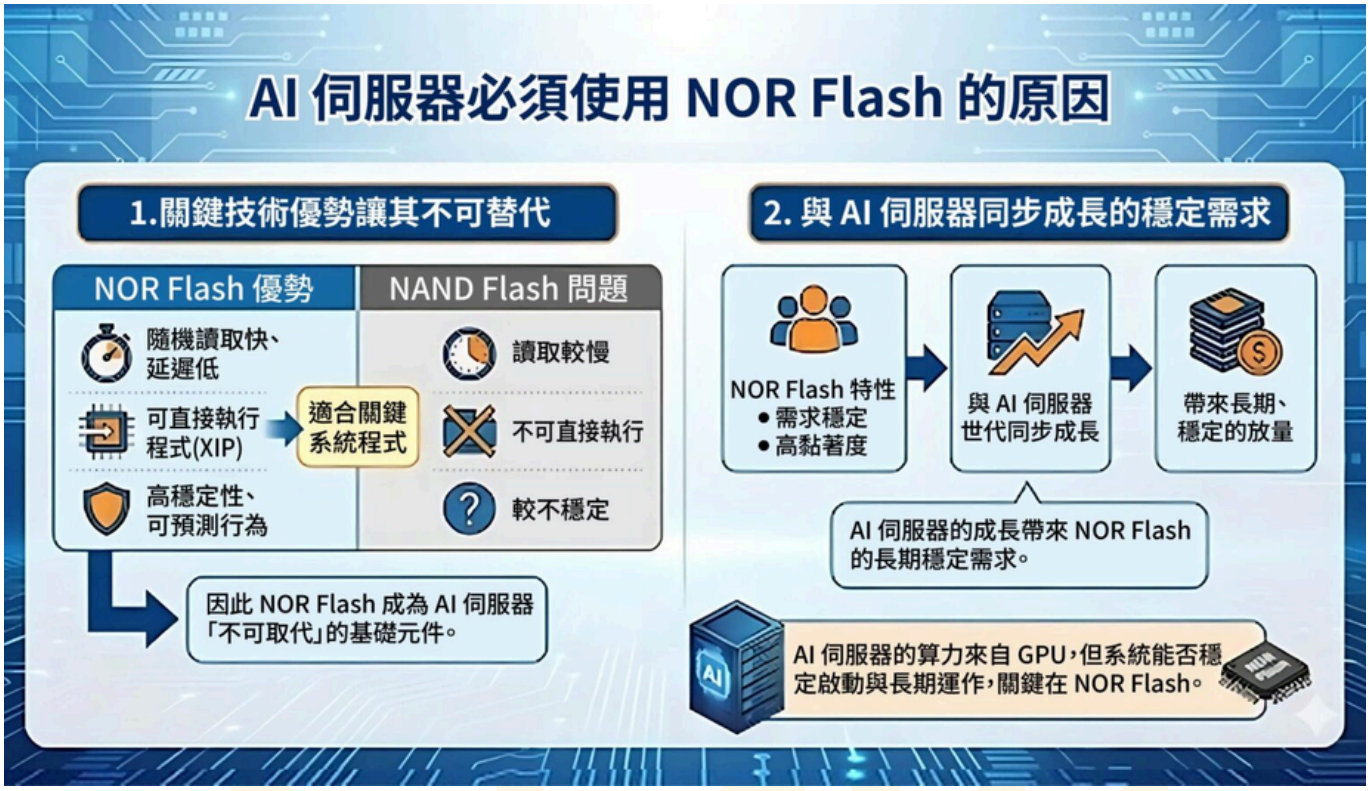
AI 伺服器的核心痛點——NOR Flash

在 AI 伺服器的架構中，NOR Flash 負責存儲 Boot Code 與 Bootloader，並具備低功耗與快速讀取功能。隨著系統複雜化，任何細微的記憶體故障都可能導致整台伺服器宕機，造成巨大的算力損失。

在多數 AI 伺服器架構中，NOR Flash 是啟動與安全機制中高度關鍵的元件。針對此趨勢，芯測提供的 **eFlash (NOR Flash) BIST IP** 正是為了解決以下設計痛點：

- **高檢驗標準：** 沿襲車用電子晶片的嚴苛檢驗標準，確保在高速運算的熱循環環境下依然穩定。
- **生命週期監控：** 透過專利測試演算法，於晶片生命週期不同階段進行健康狀態檢測。

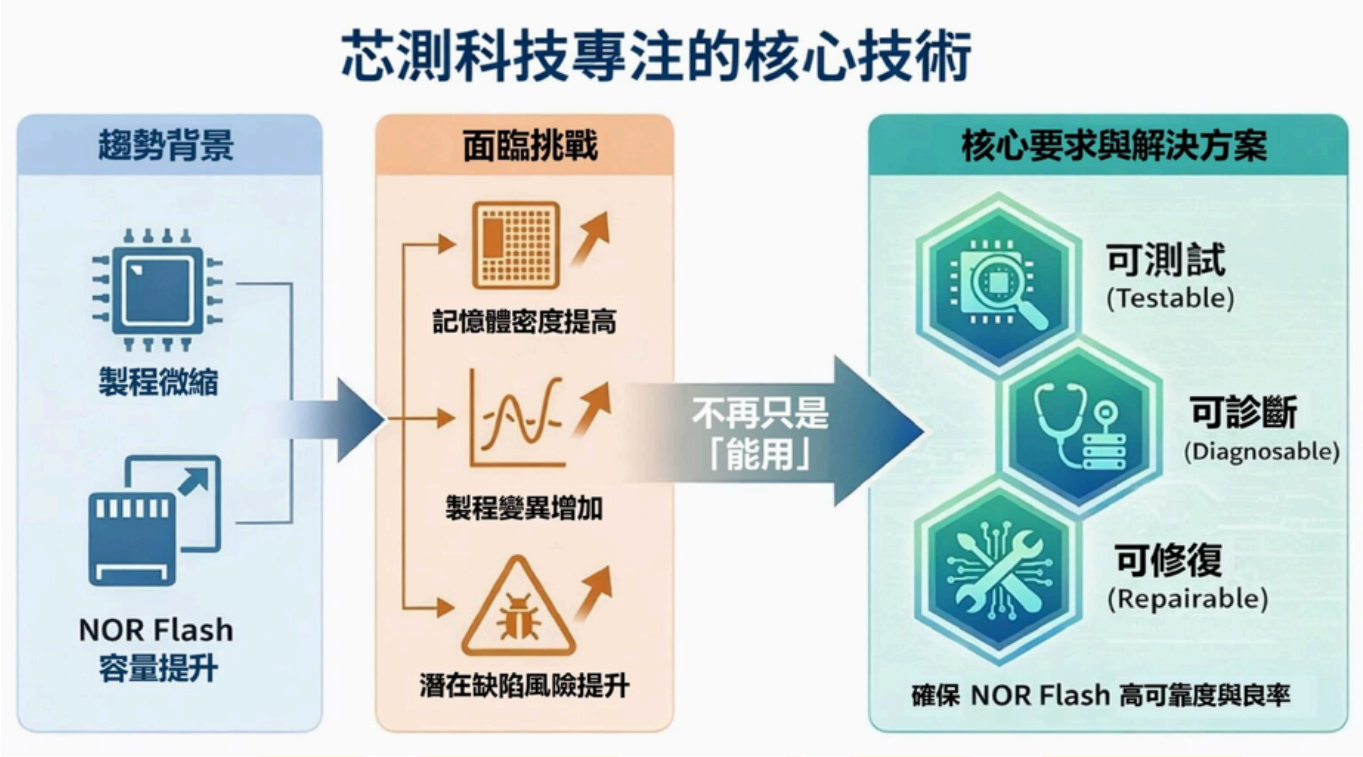




從 BIST 到 BISR 的完整技術鏈

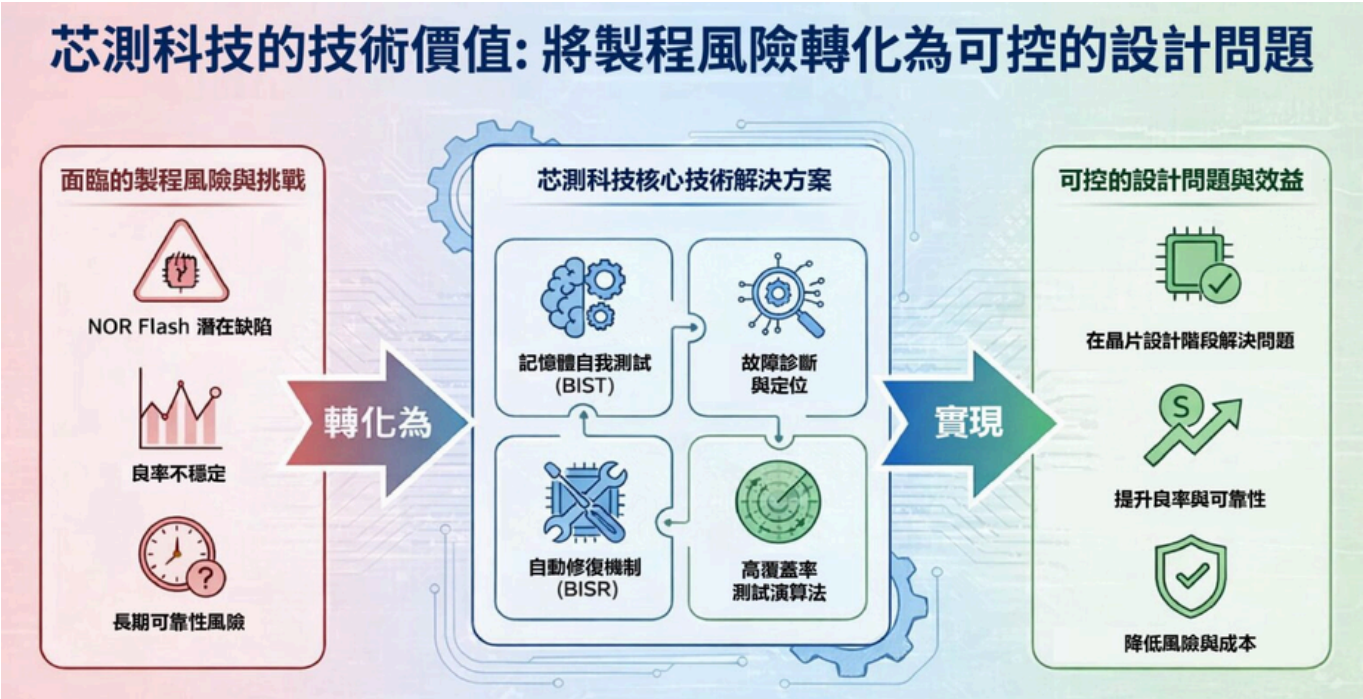
芯測科技的優勢在於將複雜的測試流程轉化為可控的設計效益。其核心技術解決方案包含四大模組：

1. 記憶體自我測試 (BIST)： 提供極高的故障檢出率。
2. 故障診斷與定位： 縮短工程師除錯時間，加速產品上市。
3. 自動修復機制 (BISR)： 透過硬體修復大幅提升晶片良率，這對昂貴的 AI 晶片尤為重要。
4. 高覆蓋率測試演算法： 針對先進製程 (如 eFlash 節點) 提供精準的測試覆蓋。



產業領先地位：從車規級實績到全方位設計服務

芯測科技之所以能快速切入 AI 供應鏈，得益於其長期在高安全關鍵領域的累積，並具備成功協助客戶完成 55nm eFlash 製程設計的實戰經驗，包括從前端 IP 評估、BIST 設計，到後端佈局繞線 (APR) 與 Turnkey 流片的全流程服務。這種全方位的交付能力，讓晶片設計公司能專注於 AI 架構創新，而將複雜的測試工作與品質門檻交由芯測把關。



NPU 熱潮下 Edge AI 正在重新定義 AI 運算架構

隨著 AI 應用從雲端走向終端設備，NPU (Neural Processing Unit) 成為近期產業討論的焦點。相較於傳統 CPU 與 GPU，NPU 以更低功耗與更高效率，讓 AI 模型能直接在邊緣端即時運行，不僅降低延遲，也減少對雲端的依賴，同時強化資料隱私。

在 Edge AI 發展早期即深耕此領域的耐能智慧 (Kneron) ，長期專注於邊緣 AI SoC 與 NPU 架構設計，透過「AI 晶片 × 邊緣運算 × 圖像演算法」的整合，推動智慧物聯、自動駕駛與智慧安防等應用落地。

芯測科技與耐能智慧有密切的合作經驗。在耐能智慧的技術交流與實務經驗驅動下，芯測科技持續精進於記憶體測試與修復解決方案，協助先進 AI SoC 在量產階段兼顧效能、良率與可靠度。

AI 推論新戰局：
SRAM 成為效能關鍵，如何搶佔先機？

輝達 (NVIDIA) GTC 大會再次點燃 AI 戰火！據媒體報導，輝達預計推出 V4 版本的 LPU，透過台積電 N3P 製程搭配 CoWoS-R 來打造，未來搭配 GPU、NVLink (高速互連) 整合為完整系統解決方案。

業者分析，隨生成式 AI 由訓練走向隨著 AI 推論需求爆炸式成長，傳統記憶體頻寬已面臨瓶頸，存取速度更快的 SRAM 正成為提升運算效能的「第一排」關鍵武器。當台積電與全球 IC 設計大廠紛紛強化 SRAM 佈局時，如何確保晶片中龐大 SRAM 陣容的可靠性與良率，就是勝出的核心。

在此浪潮下，芯測科技所提供最全面的記憶體測試與修復解決方案，將可讓企業奪得領先優勢：

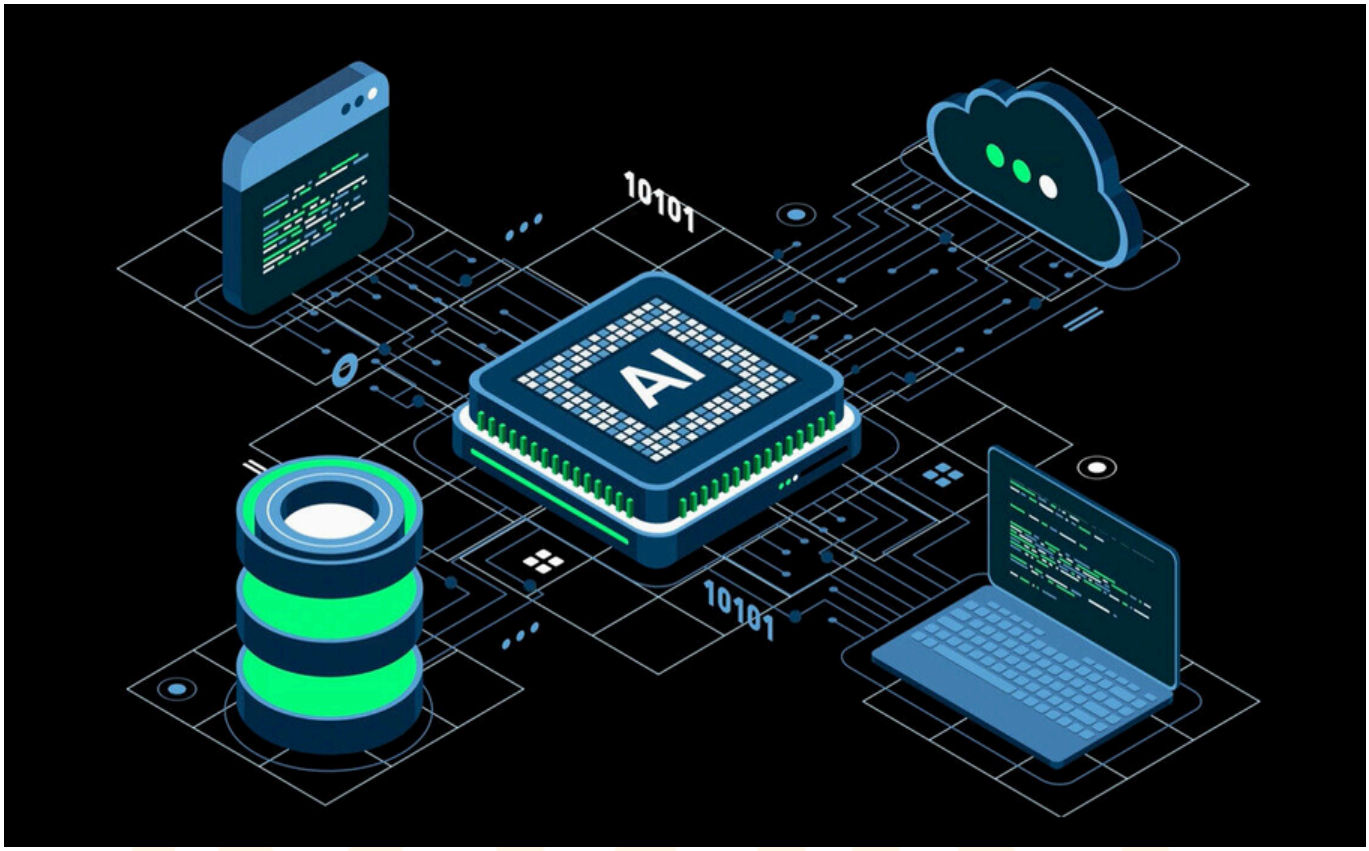
自動化開發平台：縮短記憶體測試電路設計時間，加速產品上市。

高效率修復技術：提升 SRAM 良率，將昂貴的先進製程成本降到最低。

應對複雜架構：ISO26262 TCL1 車規等級技術，可輕鬆因應 AI 晶片高規格的記憶體需求。

在 AI 晶片追求極致的道路上，芯測始終是最可靠的夥伴。

SRAM Repair 正成為 AI 浪潮下的關鍵成長動能



從先進製程授權、PUF 導入 AI 架構，到雲端 AI 與高效能運算需求擴張，產業資金與技術焦點正快速向 AI 相關應用集中。從近期的市場訊號可明顯看出：AI Data Center Processor、HBM 高頻寬記憶體、以及各類 AI 加速晶片，正在推升對嵌入式記憶體可靠度的要求。

先進製程 (3nm 及以下) 中，SRAM bit cell 尺寸持續縮小，製程變異、隨機缺陷與老化效應更加敏感。在 AI 推論與訓練架構中，大量的 on-chip SRAM、cache 與 buffer 被用於降低延遲與資料搬移成本，使得單顆 SoC 的 SRAM 容量顯著上升。SRAM 不再只是配角，而是直接影響良率、功耗與系統穩定度的核心元件。

在此趨勢下，SRAM 修復顯得日益重要，且 已從輔助角色，轉變為支撐成長的關鍵技術。當晶片內部記憶體規模擴大，任何微小缺陷都可能造成整顆晶片報廢；高效且精準的修復機制，成為控制成本與確保量產穩定性的必要條件。但在大容量與先進製程條件下，傳統記憶體測試方式面臨測試時間過長、pattern 冗餘與成本上升等問題。

因應此挑戰，芯測科技 START™ v5 (榮獲 2025 EE Awards Asia 年度最佳 EDA 工具) 以專利化演算法與 UDA 模組化架構，讓工程師可依製程與產品特性調整測試策略，在提升良率與可靠度的同時，優化測試效率與量產成本。

延伸閱讀：

- **從 TPU、LPU、NPU 到 CIM：為何先進 AI 運算都離不開 SRAM？**
  https://www.istart-tek.com/2026/01/08/10737/
- **當 TOPS 不再等於效能 AI 算力真正的瓶頸在哪裡？**
  https://www.istart-tek.com/2026/01/21/10806/

當記憶體成為稀缺資源，誰能掌握量產與成本的主控權？

DRAM 與 NAND Flash 進入了新一輪結構性漲價週期，記憶體已成為直接影響系統成本、交付節奏與供應風險的關鍵資源。這一輪漲價的背景，並非單純來自傳統終端需求回溫，而是 AI 推論、車用電子與邊緣運算，對高密度、高可靠度記憶體的長期消耗正在成為常態。

在此環境下，晶片設計與系統廠更實際面對的問題，是如何在供應受限的前提下，把現有記憶體資源轉化為可量產的產品。特別是在車用與 AI 應用中，記憶體失效不僅影響良率，更可能牽動整體配置與產品上市時程。

芯測科技的記憶體測試與修復解決方案，正是因應此產業型態轉變的最佳利器。透過完整的 BIST、BISR 與 Repair 架構，客戶可在不增加製程成本的前提下，提高可用記憶體比例，降低報廢率，緩解記憶體單價上升所帶來的壓力。

同時，芯測以 START™ v5 平台為基礎，提供 UDA (User-Defined Algorithm) 機制，讓客戶能依據實際 AI 運算負載或車用使用情境，客製化測試與修復演算法，而非受限於制式模板。這對於長時間運作、高存取頻率的應用場景尤為關鍵。

在記憶體供應趨於緊張、價格中樞上移的時代，測試與修復已成為影響成本控制與供應確定性的核心能力。透過芯測長年在此領域的耕耘，將可協助客戶有效且系統化的管控記憶體風險，取得最佳解決方案。

立即諮詢芯測科技：https://www.istart-tek.com/contact/

輝達 GTC 突顯 AI 時代的記憶體新賽局：從 SRAM 到 MLC NAND 的良率保衛戰

輝達 (NVIDIA) GTC 2026 大會甫落幕，全球科技產業的目光再次聚焦於「AI 推論 (Inference)」。執行長黃仁勳在演講中明確指出，我們已經進入了「代幣工廠 (Token Factory)」的時代，運算需求在短短兩年內暴增了百萬倍。

在這場運算風暴中，半導體架構正經歷一場前所未有的變革。過去我們關注的是運算核心的算力，但現在記憶體的表現才真正決定了 AI 代理人 (Agentic AI) 的反應速度和營運成本的表現。

異構推論新趨勢：當 GPU 遇上 LPU

在GTC相關報導和延伸討論中，最受矚目的技術轉向莫過於「解構式推論 (Disaggregated Inference)」。輝達整合了 Groq 團隊的技術，推出全新的 Groq LP30 晶片。這背後的邏輯在於解決「記憶體牆」的瓶頸：Vera Rubin GPU 擅長處理「預填 (Prefill)」階段，利用其強大算力處理海量上下文。Groq LPU 則專攻「代幣生成 (Decode)」階段。它的設計核心是捨棄傳統的高頻寬記憶體 (HBM)，改為搭載龐大的片上 SRAM。

為什麼是 SRAM？因為它的延遲極低、速度極快。當 AI 在進行即時對話或生成程式碼時，數據能直接在晶片內部快速流動，不必頻繁跳出晶片去存取外部記憶體。因此，SRAM 的容量與穩定性決定了 AI 晶片的競爭力。

2D NAND 的意外逆襲：MLC 價格飆漲的背後

正當先進製程領域瘋狂追求 SRAM 與 HBM 之際，成熟製程市場也爆發了震盪。根據產業最新情資，利基型記憶體市場正出現「價格翻倍」的奇異現象。

三星、美光等國際大廠為了騰出產能給利潤更高的 AI 應用 (如 300 層以上的 3D NAND)，正加速淡出低容量、舊製程的 2DNAND 產線。這導致了網通、數位電視與機上盒等設備必備的 MLC NAND 陷入嚴重斷鏈，2026 年的供應量預計減半，市場缺口高達 3 至 4 成。

此現象反映出，即便在非 AI 核心運算的利基設備中，記憶體的可靠度與穩定供應依然是企業生存的命脈。當面臨 NAND 史無前例的天價，如何確保每一顆出廠記憶體的完美品質，成為了 IC 設計公司與製造商獲利的關鍵。

晶片良率：AI 時代的隱形護城河

不論是 GTC 大會上備受推崇的 SRAM 核心 LPU，還是市場稀缺的 MLC NAND，都有一個共同的技術挑戰：良率與可靠性。

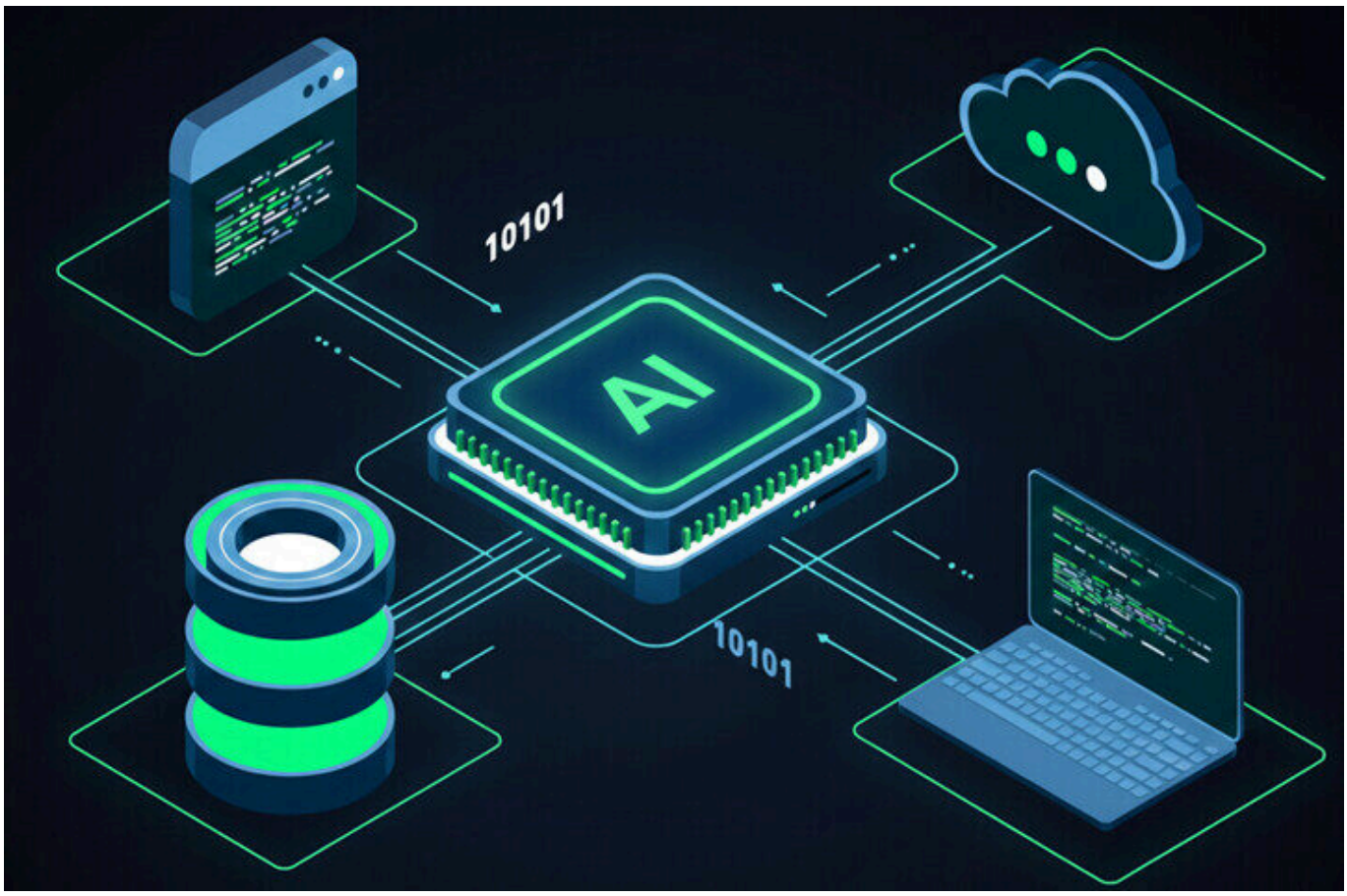
隨著先進製程邁向 N3P 甚至是更精細的維度，SRAM 在晶片中所佔的面積越來越大，發生故障的機率也隨之增加。如果一顆昂貴的 AI 晶片因為內部的 SRAM 損壞而報廢，那將是企業巨大的財務損失。這也是為什麼投資市場近期高度看好相關測試技術的原因——在追求極致效能的道路上，「測試與修復」已從以往的輔助角色，躍升為守護晶片設計品質的標配。

智慧化與高彈性記憶體測試修復技術 成為獲利關鍵

在這波 AI 推論與記憶體缺貨潮中，芯測科技憑藉著深耕多年的記憶體測試與修復技術，成為 SRAM 良率的守門員。如果說標準演算法是基礎防護，芯測的使用者自定義演算法平台（User-Defined Algorithm, UDA）就是一座靈活的研發基地。透過圖形化介面與獨家的 TEC（Testing Elements Change）技術，工程師無需編寫複雜程式碼，即可針對高溫、低壓等特殊應用場景，自行量身打造測試策略。無論是偵測複雜的 Leakage Defect 還是多樣化的記憶體缺陷（如 SAF、DRDF），UDA 都能實現全面性的防禦效果，確保晶片在任何嚴苛環境下都能穩定運作。

同時，面對先進製程下日益複雜的 SRAM 架構，芯測推出的 MART（MBIST Algorithm Recommendation Tool） 將 AI 技術導入測試流程，克服傳統表格查詢的繁瑣和缺乏彈性的測試選擇問題。使用者僅需透過簡單的互動問答，系統便會根據功耗、面積、良率等條件自動進行 AI 權重加權分析。這不僅大幅降低了決策成本與 DPPM（每百萬顆不良品數），更讓 BIST 演算法的挑選從「人工經驗」進化為「智慧化精準推薦」，協助客戶在激烈的 AI 晶片賽局中奪得上市先機。

先進製程與 AI 伺服器架構演進：獨立式 NOR Flash 在系統管理中的戰略轉型



在當前高效能運算 (HPC) 與 AI 伺服器的快速演進中，記憶體在系統架構中的角色正經歷變革。市場對資料移動成本、系統功耗與整體運行效率的考量日益增加，非揮發性記憶體已從單純的被動儲存元件，也逐步成為參與系統資源配置、安全管理與引導運作的關鍵環節。在此背景下，嵌入式快閃記憶體 (eFlash) 與獨立式 NOR Flash 的定位出現了明顯的消長。

過去在微控制器 (MCU) 與低功耗邊緣運算 SoC 中，開發者傾向於將 eFlash 直接整合於晶片內部，以維持高整合度並降低系統複雜度。然而，當半導體製程節點持續縮小至 28nm 以下，甚至進入 FinFET 先進製程時，eFlash 面臨了嚴峻的結構性限制。由於 eFlash 寫入時需要高電壓元件與特殊的製程模組，這與先進邏輯製程追求的低電壓、超薄閘極氧化層特點產生了物理上的互斥。

強行整合大容量 eFlash 通常需增加約 8–12 道額外光罩 (依製程平台與記憶體架構而異) ，顯著提升生產成本，更會面臨良率下降與可靠度控制難度增加的挑戰。在先進製程中，若持續追求 eFlash 的整合，已不再符合經濟效益。

這一技術瓶頸促使系統設計轉而採用外掛式儲存方案，使得獨立式 NOR Flash 在多種高效能應用中重新獲得重視。透過 QSPI、OSPI 或其他高速序列介面，SoC 可將韌體、啟動程式與關鍵設定參數儲存在外部的專用記憶體晶片中。這種分工模式使邏輯晶片能專注於高密度運算與 AI 加速，而將非揮發性儲存交由具備專業製程的記憶體廠商負責，提升了整體的製程彈性與供應鏈成本控制能力。

這種轉變在 AI 伺服器架構中尤為關鍵。儘管 AI 系統的主要運算與資料交換依賴於 HBM 或 DDR5 等揮發性記憶體，但負責系統底層管理的「基板管理控制器」(BMC) 則對 NOR Flash 有著極高的依賴。由於 AI 伺服器結構日益複雜，內部包含大量的 GPU、FPGA 與網路交換元件，BMC 必須透過獨立式 NOR Flash 來儲存核心韌體、執行遠端監測與故障診斷。同時為了因應資安威脅，系統通常需要內建安全啟動機制 (Root of Trust) ，這類加密簽章與安全引導程式亦需儲存在高可靠度的 NOR Flash 中。

此外，AI 伺服器對可靠度的標準遠高於一般消費性電子產品。任何啟動錯誤或韌體損毀都可能導致整個運算叢集停機，產生高額的營運損失。因此資料中心使用的 NOR Flash 在失效率 (DPPM) 控制、耐久度與在高溫環境下的資料保存時間方面，必須符合極為嚴格的驗證規範。這種高可靠度需求，使 NOR Flash 從單純的開機元件，轉變為確保系統穩定性與安全性的戰略組件。隨著 AI 系統導入更頻繁的韌體更新與多版本管理機制，NOR Flash 的讀寫循環壽命與資料完整性驗證變得越發重要。

從產業結構來看，這種變化促成了半導體供應鏈的重新分工。過去 eFlash 是 SoC 廠商的核心整合能力，但在 AI 與先進製程環境下，將儲存功能外移至專業記憶體供應商，反而有助於提升整體效率。邏輯設計公司可專注於運算架構與加速器設計，而記憶體廠商則在高可靠度市場中建立技術壁壘。總體而言，先進製程下的整合門檻，為外掛式 NOR Flash 創造了新的增長空間。AI 伺服器市場持續擴張，推升了高可靠度非揮發性儲存的需求，這不僅確立了 NOR Flash 在系統管理架構中不可或缺的地位，也使其成為支撐 AI 時代算力穩定運行的重要幕後功臣。

當 TOPS 不再等於效能 AI 算力真正的瓶頸在哪裡？

近幾年，AI 晶片的發表會幾乎都離不開一個關鍵字：TOPS。無論是資料中心 GPU、車用 SoC，還是邊緣 AI 處理器，算力數字一代比一代高，從數十 TOPS、數百 TOPS，快速跳到上千甚至上萬 TOPS。表面上看，AI 算力似乎已經不是問題，但在實際系統設計與應用中，效能瓶頸卻仍然頻繁出現。這也讓工程師開始重新思考一個問題：TOPS 真的是決定 AI 效能的核心指標嗎？

TOPS (Tera Operations Per Second) 代表晶片在特定運算精度條件下，每秒可完成的理論運算次數。多數 AI 晶片會以 INT8、INT4 等低精度運算作為標準，因為這類運算最符合推論場景，也最容易堆疊出漂亮的算力數字。然而，TOPS 本身並不等於實際應用效能，它更像是一顆引擎的最大馬力值，無法反映整個系統是否能長時間穩定地輸出這個效能。

以 NVIDIA 為例，近幾代資料中心 GPU 的 AI 算力早已進入萬 TOPS 等級。H100、B200 等產品在低精度 AI 運算模式下，理論算力極高，足以支撐大型語言模型與生成式 AI 的推論需求。在終端與邊緣市場，NVIDIA Jetson 系列、Qualcomm、MediaTek、Apple 以及 Google 等，也陸續推出具備數十到數百 TOPS 的 NPU SoC，用於影像辨識、語音處理與本地端 AI 推論。從規格表來看，AI 算力似乎已全面到位。

但實務上，AI 推論並不是單純的運算而已。每一次運算之前，都必須先從記憶體讀取權重與特徵資料，運算完成後，再將結果寫回記憶體。當模型規模變大、資料重用頻率提高，系統的效能往往受限於資料搬移，而非運算單元本身。這也是為什麼在許多應用中，即使晶片標示的 TOPS 很高，實際效能卻無法線性成長。

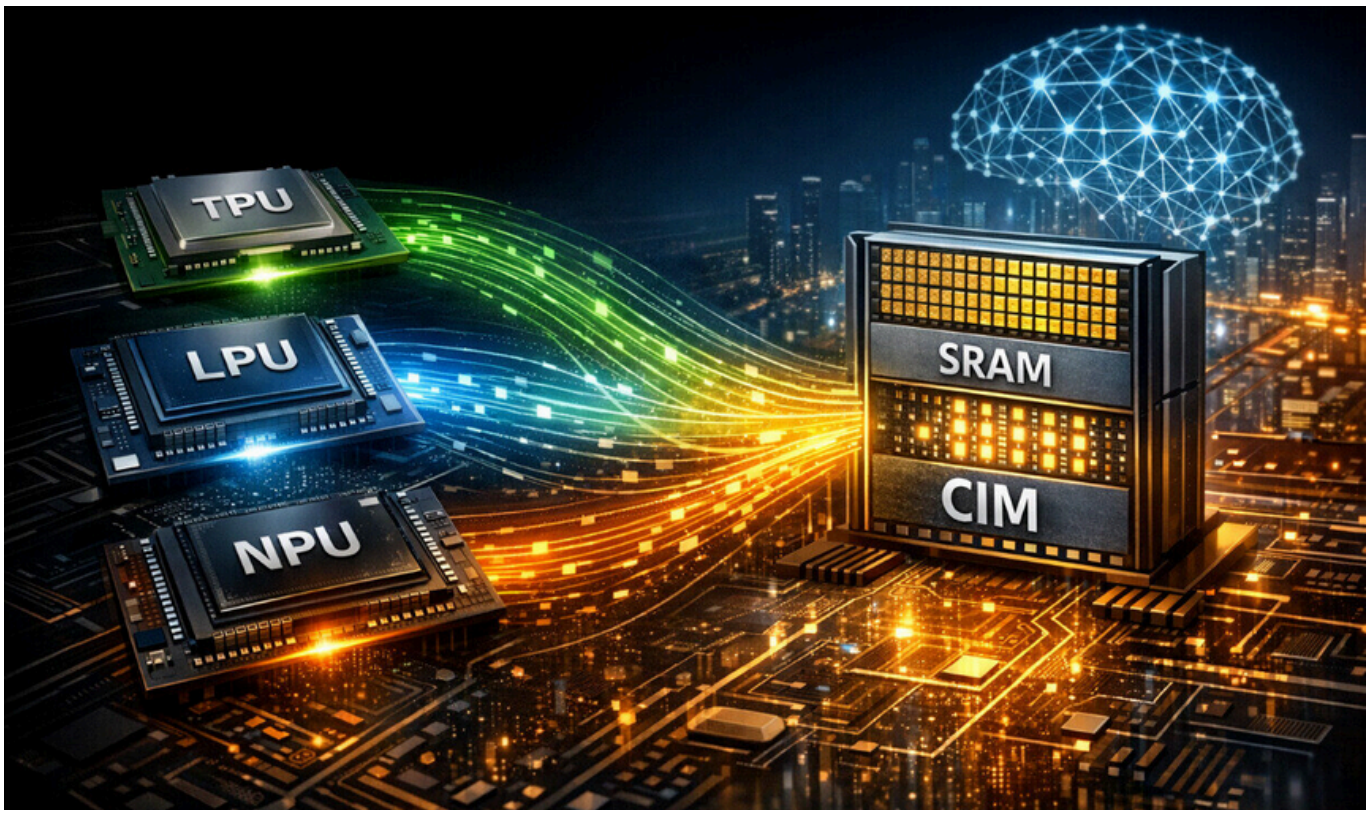
因此我們須回歸到記憶體架構問題。為了支撐高 TOPS 運算，各大 AI 晶片廠紛紛加強記憶體頻寬設計。資料中心 GPU 大量採用 HBM，透過堆疊式封裝與超高頻寬，縮短運算單元與外部記憶體之間的距離。而在 SoC 與 NPU 設計中，晶片內部 SRAM 的面積比例則持續攀升，成為不可忽視的關鍵資源。

SRAM 在 AI 晶片中的角色，並不只是暫存資料這麼簡單。它負責承接高頻、低延遲的資料存取需求，支援權重快取、特徵圖暫存以及中間運算結果保存。對於卷積神經網路、Transformer 等模型而言，資料的重複存取頻率極高，如果每一次都必須回到外部 DRAM，不僅延遲增加，功耗也會快速上升。因此，越來越多 AI 架構選擇將關鍵資料留在片上 SRAM 中，真正發揮 TOPS 所代表的算力。

因為 SRAM 扮演關鍵角色，其可靠度與測試覆蓋率的重要性也隨之放大。在高頻、長時間運作下，SRAM 容易面臨讀擾、耦合干擾、老化等問題，一旦發生錯誤，影響的不只是單一運算，而是整個 AI 推論結果的穩定性。這使得記憶體測試與修復機制，成為 AI 晶片設計中不可或缺的一環。

從 TOPS 的快速膨脹，到記憶體架構的持續進化，可以看出 AI 晶片競爭的重心正在轉移。算力仍然重要，但真正拉開差距的，往往是誰能讓資料流動得更順暢、記憶體更可靠。當 TOPS 不再只是宣傳數字，而能被完整釋放時，AI 晶片的價值，才算真正被發揮出來。

從 TPU、LPU、NPU 到 CIM：為何先進 AI 運算都離不開 SRAM？



近來 NVIDIA 併購 Groq 的消息引發高度討論，TPU、LPU、NPU 等各類 AI 專用處理器再次成為焦點。無論是雲端資料中心、邊緣 AI，或車用與工業應用，運算架構正快速從通用 CPU，轉向為特定工作負載量身打造的加速器。然而在這些名稱各異、定位不同的處理器背後，卻有一個共同且關鍵的核心 —— SRAM。

TPU、LPU、NPU 的共通基礎

TPU、LPU、NPU 的設計目標雖不相同，但皆高度聚焦於高吞吐量、低延遲與能效比。在實際架構中，大量的權重資料、特徵圖與中間運算結果，都必須被反覆且高速地存取。相較於 DRAM，SRAM 具備低延遲、高頻寬與可預測時序的特性，因此成為這類 AI 加速器中最不可或缺的記憶體元件。

在 TPU 中，SRAM 常被用於大型 on-chip buffer，以支援矩陣運算的資料重用；在強調低延遲推論的 LPU 架構中，SRAM 更是決定即時反應能力的關鍵；而在各式 NPU 設計裡，SRAM 的容量配置、存取並行度與可靠度，往往直接影響整體 AI 效能與功耗表現。

從「記憶體輔助運算」走向「記憶體即運算」

當 AI 模型規模持續放大，資料在運算單元與記憶體之間搬移所消耗的能量，已成為效能與功耗的主要瓶頸。這正是 Computing-In-Memory (CIM) 架構受到高度關注的原因。CIM 的核心概念，是將部分運算直接在記憶體陣列中完成，藉此大幅降低資料搬移次數，突破傳統馮紐曼架構的限制。

對於本就大量依賴 SRAM 的 TPU、LPU、NPU 而言，在特定運算場景下導入 SRAM-based CIM，是目前被廣泛認為有必要的演進路線之一。透過在 SRAM 內部或周邊電路中整合運算能力，AI 加速器得以在維持高可靠度的同時，顯著提升能效比，這也使 CIM 成為下一世代 AI 晶片的重要關鍵技術之一。

CIM 不只是效能 更關乎可靠與安全

CIM 架構所面臨的設計挑戰，並不僅止於效能。記憶體本身的變異性、老化效應，以及運算與儲存耦合後所帶來的測試與驗證難度，都讓 SRAM 的測試與修復能力變得前所未有的重要。特別是在車用與關鍵基礎設施等高可靠度應用中，CIM 若無完善的測試與修復機制，將難以真正量產與導入。

另一方面，後量子時代的資安需求，也讓 AI 與加密運算的結合日益緊密。能否在低功耗條件下，高效執行新一代密碼演算法，正成為智慧裝置與車用系統的重要課題。

AI 運算趨勢下必備的記憶體技術

在這樣的趨勢下，攸關 TPU、LPU、NPU 效能與成本的 SRAM 品質，以及 CIM 架構逐漸成為突破資料搬移瓶頸的重要方向。芯測科技長期深耕的 SRAM 測試與修復領域，以及 CIM 架構的研發，將可成為推動技術突破的關鍵引擎。

芯測致力於提供各類記憶體之測試與修復專用 EDA 工具與 IP，並針對授權客戶，提供涵蓋後端流程的一站式設計服務。近年來芯測更積極投入 CIM 架構的研發，重新定義 AI 運算的能效極限。 目前正在開發的 SRAM-based CIM 架構，具備 8-bit 運算精度與極低功耗，並具備良好延展性，可進一步支援 RRAM 架構設計，以滿足不同應用場景的需求。

面對量子時代帶來的資安挑戰，芯測亦率先導入格點密碼與 NTT 運算優化技術，透過 CIM 架構加速後量子密碼演算法，為智慧裝置與車用系統提供長效且穩固的安全保障。 除了 CIM 架構本身的開發，芯測科技同時也提供 CIM 的測試電路開發環境，讓先進記憶體運算技術不僅能「算得快」，更能「測得準、用得久」。在 TPU、LPU、NPU 持續演進的浪潮中，芯測科技以深厚的 SRAM 技術底蘊，成為推動 CIM 與 AI 晶片走向高效、可靠與節能的重要力量。

當 MRAM 遇上 AI：下一代智慧晶片的關鍵記憶體革命

隨著 AI 晶片算力持續提升，系統效能與功耗的瓶頸正逐漸從運算核心轉向記憶體架構本身。在 AI 推論與訓練過程中，大量模型權重與中間資料需頻繁於記憶體與運算單元之間搬移，其能耗與延遲往往高於實際運算本身。為降低資料搬移成本並提升效率，新一代 AI SoC 開始導入 **MRAM**，利用其兼具高速讀寫與非揮發特性，使關鍵資料可長時間常駐晶片內，進一步改善功耗與延遲表現，並支援邊緣 AI 與低功耗應用需求。

然而，MRAM 的磁性儲存機制對製程變異、材料特性與環境條件高度敏感，可能出現寫入翻轉失敗、磁阻值漂移、Retention 衰退或讀取邊界不穩等失效模式。在高頻讀寫與長時間 AI 工作負載下，這些潛在問題更容易被放大，若缺乏完善的測試與修復機制，將直接影響良率與產品可靠度。因此，在 AI SoC 中導入 MRAM 時，必須建立完整的 **BIST 與 BISR 測試修復架構**，並進行壓力測試、長時間保存測試與跨電壓溫度條件驗證，以確保量產品質並有效控制 DPPM。

芯測科技長期深耕記憶體測試與修復技術，自主開發的 **MRAM BIST IP** 已成功導入先進製程 SoC 並完成實際量產應用，具備支援車規等級與高可靠度需求的驗證能力。在高頻 AI 工作負載情境下，芯測解決方案能提供穩定的測試與修復支援，協助客戶兼顧效能、功耗與長期可靠度，並在新一代 AI 記憶體架構轉型中建立良率與量產競爭優勢。

觀看完整影片

NPU 運作需求下的 SRAM 重要性與測試策略

在 NPU (Neutral Processing Unit) 的實際設計中，SRAM 往往佔據晶片中最大面積、消耗最多功耗，同時也是最容易影響可靠度的關鍵元件。為了支援高並行度與高度資料重用的 AI 推論行為，NPU 會在運算單元周圍配置大量 on-chip SRAM，作為權重緩衝區、特徵圖暫存與局部資料存取空間，確保資料能就地取用並高速周轉。

然而，這樣的架構也使 SRAM 長時間處於高頻率、重複存取的極端使用情境。讀擾 (Read Disturb) 、耦合干擾、弱寫入、感測邊界不穩，以及資料保留 (Retention) 相關失效，在 NPU 中更容易被放大，成為影響良率與 DPPM 的主要風險來源。這也使得 NPU 的效能瓶頸，往往不在於算力，而是在記憶體子系統的設計與測試品質。

本集從 NPU 的實際運作需求出發，解析不同 SRAM 使用情境 (如權重緩衝、特徵圖暫存、累加結果與待命區域) 所對應的潛在失效風險，並說明為何 NPU 的 SRAM 測試不能僅依賴單一、靜態的測試演算法，而必須導入可依情境調整的測試與修復策略，才能有效降低量產風險。

針對此挑戰，芯測科技以 START™ v5 平台 為核心，結合 UDA (User-Defined Algorithm) 的圖形化與模組化設計，讓工程師可依不同 SRAM 結構與使用行為，自行組合最合適的 MBIST 測試流程。同時搭配 MART (MBIST 演算法推薦工具) ，將既有 SRAM 測試經驗系統化，協助在多重限制條件下快速選擇適當的測試演算法組合，降低決策成本，避免因測試策略不當而影響量產。

透過 UDA 與 MART 的整合應用，SRAM 測試不再只是形式上的驗證流程，而能真正貼近 NPU 的實際 AI 工作負載，在兼顧可靠度的同時，提升良率、控制 DPPM，並確保量產效率與產品競爭力。

觀看完整影片

LPU 使用 SRAM 的原因與測試關鍵

隨著邊緣 AI 與低功耗 AI 推論應用快速成長，LPU (Language Processing Unit) 在系統架構設計上，普遍大量採用 SRAM 作為主要 on-chip memory。其關鍵原因不僅在於速度，而是因為 LPU 推論行為對 記憶體存取延遲、穩定性與可預期性 有極為嚴苛的要求。本集將從 AI 推論流程出發，解析為何在影像辨識、語音觸發與感測器資料分析等即時應用中，記憶體行為往往比運算效能更容易成為系統瓶頸。

內容將深入比較 SRAM 與 DRAM 在存取延遲、refresh 機制、功耗行為與系統穩定性 上的差異，說明 DRAM 在 LPU 架構中可能帶來的不確定性風險，以及 SRAM 在低電壓、高頻存取環境下的關鍵價值。同時也說明，在現代 LPU SoC 中，大量、異質化的 SRAM bank 如何顯著提升記憶體測試與驗證的複雜度。

本集進一步探討 傳統 SRAM 測試方法與實際 LPU 推論存取情境之間的落差，並說明僅以基本讀寫測試，為何無法有效反映低電壓、實際負載下的潛在風險。最後，將介紹芯測科技如何透過 START™ v5 平台，結合 UDA (User-Defined Algorithm) 與 MART (MBIST Algorithm Recommendation Tool) ，協助設計團隊依 LPU 應用特性，建立更貼近實際推論行為的 MBIST 測試策略，兼顧可靠度、良率與量產效率。

觀看完整影片