

iSTART

July 2026

iSTART iReport

芯測科技電子報第 16 期



Contents

News

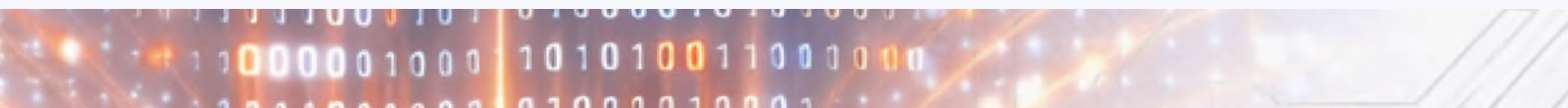
- AI 晶片良率成競爭關鍵 芯測科技以記憶體品質管理技術搶攻新商機2
- 黃仁勳點名 AI 工廠時代來臨 芯測科技布局高可靠度晶片測試迎接新商機4
- 華為「韜定律」掀先進封裝戰局 芯測科技已搶先布局 IEEE 18386
- 用 AI 解決 AI！AI 運算帶動記憶體複雜度提升 MBIST 更需要智慧化演算法8
- 芯測科技「記憶體測試算法推薦系統」取得新型專利 重新定義 SoC 測試開發流程 ...10

IC Tech Insights

- 外掛記憶體漲價 DDI 架構正在改變？內嵌 SRAM 價值重新浮現12
- AI 推論市場擴大，SRAM/Cache/Buffer 品質成為量產新門檻14
- GPU 之後的競爭焦點：AI系統效能正在重新定義16
- CoWoS 與 HBM：AI 時代記憶體產業的關鍵引擎18
- 從 Google TurboQuant 看 AI 晶片設計的新隱憂20

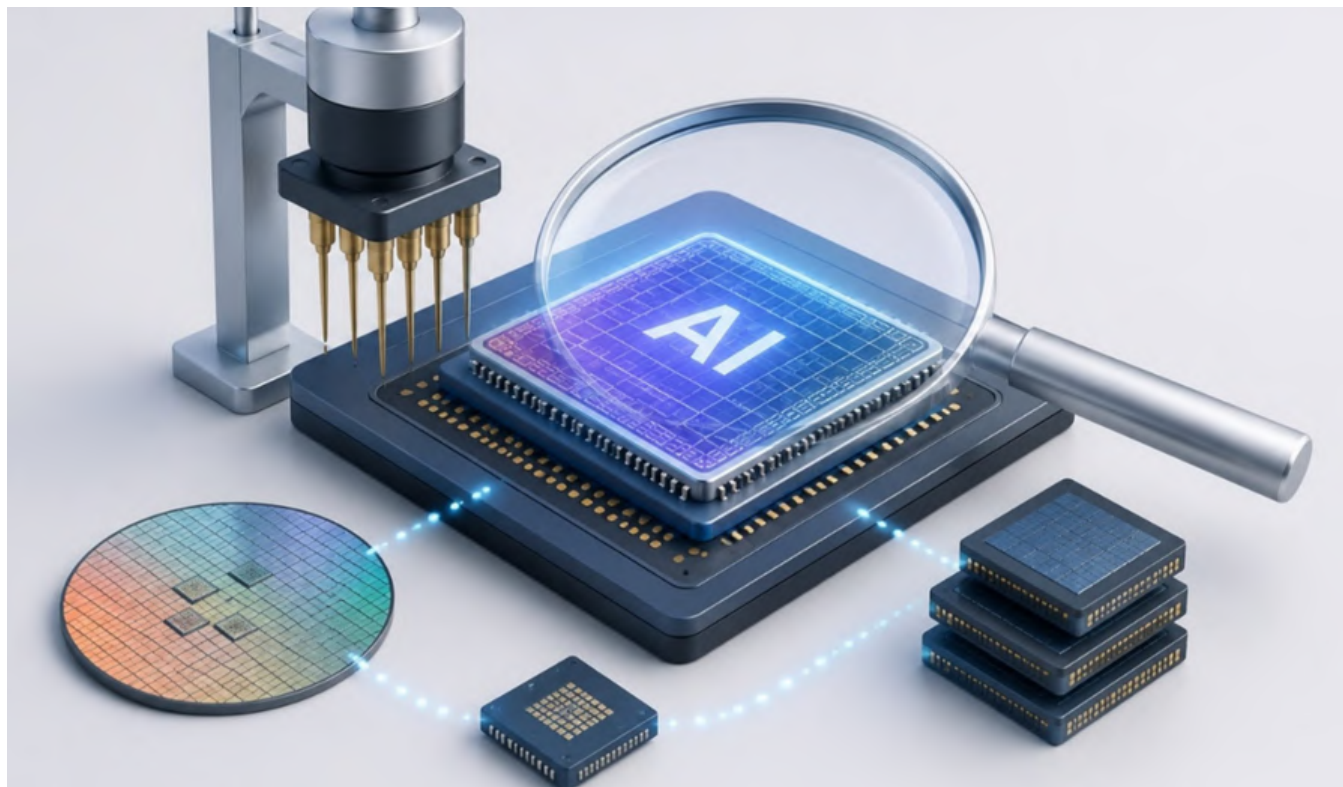
iSTART Class

- 從測試到智慧防護：記憶體品質管控工具的全面進化22
- 打造可客製化記憶體測試方法 靈活應變快速演進的半導體時代22
- AI 家電與智慧顯示時代：從語音控制到記憶體內運算的晶片測試挑戰22
- 智慧座艙 (In-vehicle AI)：汽車電子的下一個算力戰場22
- 當 MRAM 遇上 AI：下一代智慧晶片的關鍵記憶體革命.....22
- NPU 運作需求下的 SRAM 重要性與測試策略22
- LPU 使用 SRAM 的原因與測試關鍵22





AI 晶片良率成競爭關鍵 芯測科技以記憶體品質管理技術搶攻新商機



AI 晶片的競爭，正在從算力競賽走向良率競賽。當高效能運算晶片的設計愈來愈複雜，單顆晶片價值也持續攀升，任何微小缺陷都可能帶來巨大損失。如何在量產階段維持穩定品質與高良率，成為晶片廠的重要課題。

在影響晶片良率的眾多因素中，記憶體可靠度是最關鍵的一環。AI 晶片內部大量使用 SRAM 作為快取、資料緩衝與運算暫存空間，直接影響資料存取效率、延遲與功耗表現。隨著記憶體容量持續增加，SRAM 測試與修復的重要性也同步提升。

然而先進製程下的 SRAM 缺陷已不再只是傳統的開路或短路問題，還可能出現寫入異常、資料保持失效、動態錯誤，以及製程變異造成的邊界型失效。這些缺陷不僅更難偵測，也更容易成為影響 AI 晶片良率的隱性風險。

芯測科技長期深耕記憶體測試、修復與品質管理技術，透過核心平台 START™ v5，協助客戶在晶片設計階段就建立完整的記憶體品質防護機制。

START™ v5 提供完整的記憶體內建自我測試與內建自我修復架構，透過內建測試機制，系統能快速找出失效記憶體單元，提高缺陷檢出能力，降低漏測風險。搭配內建修復技術後，系統可



進一步啟用備援列或備援欄位，自動修復損壞的記憶體區塊，使原本可能報廢的晶粒恢復正常運作，提升整體良率。

對高價值 AI 晶片而言，這項能力帶來的效益尤其明顯。即使良率僅提升少量百分比，也可能轉化為可觀的成本節省與產能提升。

除了基礎測試與修復能力，芯測科技的使用者自定義演算法 (User-Defined Algorithms, UDA) 功能也提供高度彈性，讓客戶能依據不同記憶體架構、製程特性與應用需求，自行建立專屬測試流程。由於 AI 晶片中的記憶體設計高度客製化，固定式測試演算法已難以滿足所有需求，更靈活的測試能力因此成為重要優勢。

在測試效率方面，芯測科技的測試評估機制可協助分析不同演算法的缺陷覆蓋率與測試時間，幫助客戶找出最佳測試組合，在測試品質與測試成本之間取得平衡。

此外，芯測科技的智慧記憶體演算法推薦技術 MART (MBIST Algorithm Recommendation Tool)，能根據記憶體類型、製程節點與歷史失效特徵，自動推薦合適的測試演算法組合，協助工程團隊縮短開發時間，加快產品導入時程。

當 AI 晶片成本持續攀升，市場競爭的核心已不再只是製程領先，而是更有效控制缺陷、提升良率並確保品質穩定。芯測科技透過 START™ v5、記憶體測試與修復、客製化演算法及智慧化測試技術，持續協助客戶提升產品可靠度與量產效率，成為 AI 半導體產業升級的重要技術夥伴。

黃仁勳點名 AI 工廠時代來臨 芯測科技布局高可靠度晶片測試迎接新商機



NVIDIA 創辦人暨執行長黃仁勳在 NVIDIA GTC Taipei 2026 主題演講中指出，AI 產業已正式從「能不能用」邁向「能不能獲利、能不能規模化」的新階段，並將 AI Factory 定義為下一代數位基礎建設。隨著 Agentic AI、Physical AI、AI-native PC 與邊緣運算快速發展，晶片系統對可靠度、可用率與長時間穩定運作的要求正大幅提升。

黃仁勳並提出「Compute is Revenue」觀點，認為未來 AI 資料中心不只是運算中心，而是生產 AI Token 的工廠。當運算能力直接對應營收產出時，晶片失效、系統停機或可靠度不足所造成的損失，將直接影響企業獲利能力的商業問題。

在此趨勢下，專注於記憶體測試、修復與DFT技術的芯測科技（iSTART-TEK）正站上AI基礎建設升級的重要位置。

近年來，AI 晶片、HPC 晶片與車用SoC持續增加 SRAM 容量，在先進製程中，SRAM 面積占比甚至超過晶片總面積的一半。隨著 AI 工廠規模擴大，記憶體缺陷對系統穩定度與運算可用率的影響也更加顯著。芯測科技所提供的 START™ v5 平台、SRAM BIST/BISR 解決方案以及 MART（MBIST Algorithm Recommendation Tool）技術，可協助晶片開發團隊提升測試覆蓋率、強化修復能力，並降低量產階段的潛在風險。



此外，黃仁勳也特別強調 Agentic AI 將成為下一代運算模式，未來 AI 將能自主規劃任務、調用工具並長時間執行工作流程。NVIDIA 更與 Cadence 合作推動晶片設計 Super Agent，預示 RTL 生成、驗證、除錯與設計流程將逐步導入 AI 工作流。

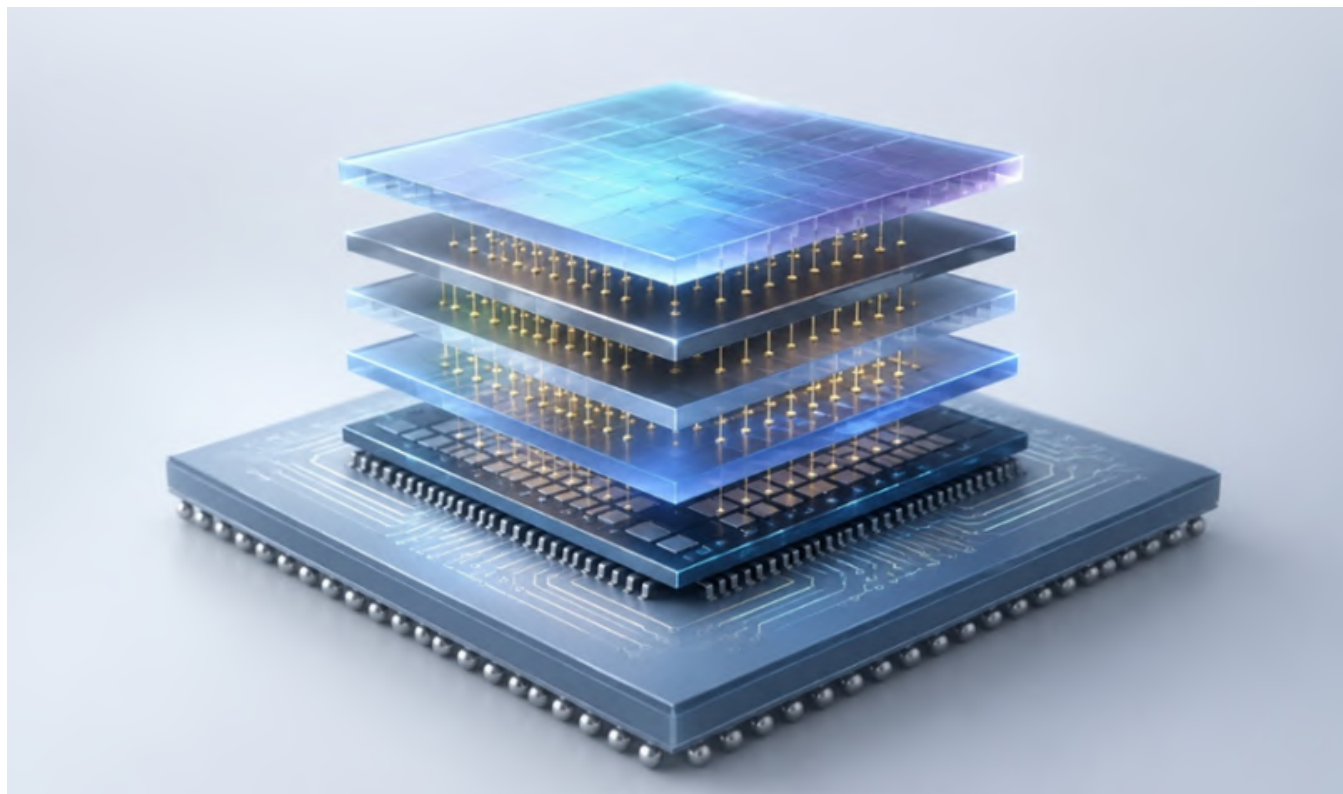
對此，芯測科技認為，未來 DFT、BIST、BISR、Repair Analysis 及 Diagnosis 所產生的大量測試資料，將有機會成為 AI 輔助設計與 AI 輔助除錯的重要基礎數據。透過將測試資料與 AI 分析能力結合，不僅能提升晶片開發效率，也能進一步優化產品品質與量產良率。

另一方面，隨著 Physical AI、自動駕駛、機器人與邊緣 AI 應用興起，市場對功能安全 (Functional Safety)、長時間穩定運作以及高可靠度驗證的需求也快速增加。黃仁勳指出，未來車用 AI 與機器人系統將需要更完整的測試驗證流程，而 DFT、BIST、BISR 及診斷能力的重要性將持續提升。

芯測科技表示，當 AI 產業競爭從單純追求算力，轉向追求系統可靠度、可用率與長期穩定運作時，測試技術的價值也將從成本中心逐漸轉變為競爭力來源。未來公司將持續深化記憶體測試演算法、修復分析及先進 DFT 技術布局，協助客戶打造符合 AI 工廠、Edge AI 與 Physical AI 時代需求的高可靠度晶片產品。

在黃仁勳描繪的 AI 基礎建設藍圖中，晶片毫無疑問是創造營收的核心資產；而確保這些晶片穩定運作的測試與修復技術，也正成為 AI 時代不可或缺的關鍵基礎能力。

華為「韜定律」掀先進封裝戰局 芯測科技已搶先布局 IEEE 1838



AI 運算需求持續爆發，半導體產業正逐步從「電晶體微縮」走向「系統級堆疊整合」的新競爭階段。近期市場關注華為提出的「韜定律（Tao Law）」概念，強調透過先進封裝、Chiplet 與多晶粒堆疊技術，突破傳統摩爾定律在製程微縮上的限制。這也代表未來 AI 晶片的核心競爭力，已從單一晶片的電晶體密度，延伸為如何透過 3D-IC 與異質整合架構，提升整體系統效能、頻寬與能效。

然而，當 AI 晶片逐步邁向 HBM、高速邏輯晶片與 Chiplet 的垂直堆疊架構後，測試與修復問題也開始成為先進封裝時代的關鍵挑戰。由於多顆 die 封裝完成後的價值極高，一旦在後段才發現記憶體或互連缺陷，將可能造成巨大的報廢成本。因此，如何在 3D-IC 架構中維持可測試性（Testability）與可修復性（Repairability），已成為 AI 晶片產業鏈不可忽視的重要議題。

專注於記憶體測試與修復技術的芯測科技（iSTART-TEK）指出，隨著產業從摩爾定律逐步轉向堆疊整合架構，IEEE 1838 將成為 3D-IC 與異質整合晶片不可或缺的測試標準。芯測科技目前已率先完成 IEEE 1838 與記憶體測試、修復架構的整合，是目前少數能完整支援 IEEE 1838 的記憶體測試與修復 EDA 工具供應商之一。



IEEE 1838 是專為 3D-IC 與多 die 堆疊架構制定的測試標準，主要用於解決 die-to-die 測試存取、堆疊後測試通道管理、以及異質整合架構中的可測試性問題。在傳統 SoC 架構中，測試主要針對單一 die；但在 HBM、Logic Die、AI Accelerator 與 Chiplet 大量堆疊的架構下，測試路徑與故障定位的複雜度大幅提高，也讓 IEEE 1838 逐漸成為先進封裝的重要基礎架構。

芯測科技表示，當產業開始以堆疊技術挑戰摩爾定律時，測試架構也勢必要同步升級。尤其在 CoWoS、3D-IC 與 HBM 等高階封裝架構中，一旦缺乏標準化測試機制，將難以有效管理多層 die 的測試存取與故障診斷。因此，未來只要走向多 die 堆疊與異質整合，IEEE 1838 幾乎將成為必要選項。

芯測科技早在 3D-IC 尚未全面商用前，便已投入 IEEE 1838 相關技術開發，並將其整合進既有記憶體測試與修復平台中，完成與 MBIST、BISR 架構的深度串接。透過 IEEE 1838 架構，系統可在多層堆疊完成後，仍維持對 SRAM 與 HBM 的測試與修復能力，包含測試路徑規劃、故障定位、修復控制與測試存取管理等功能。

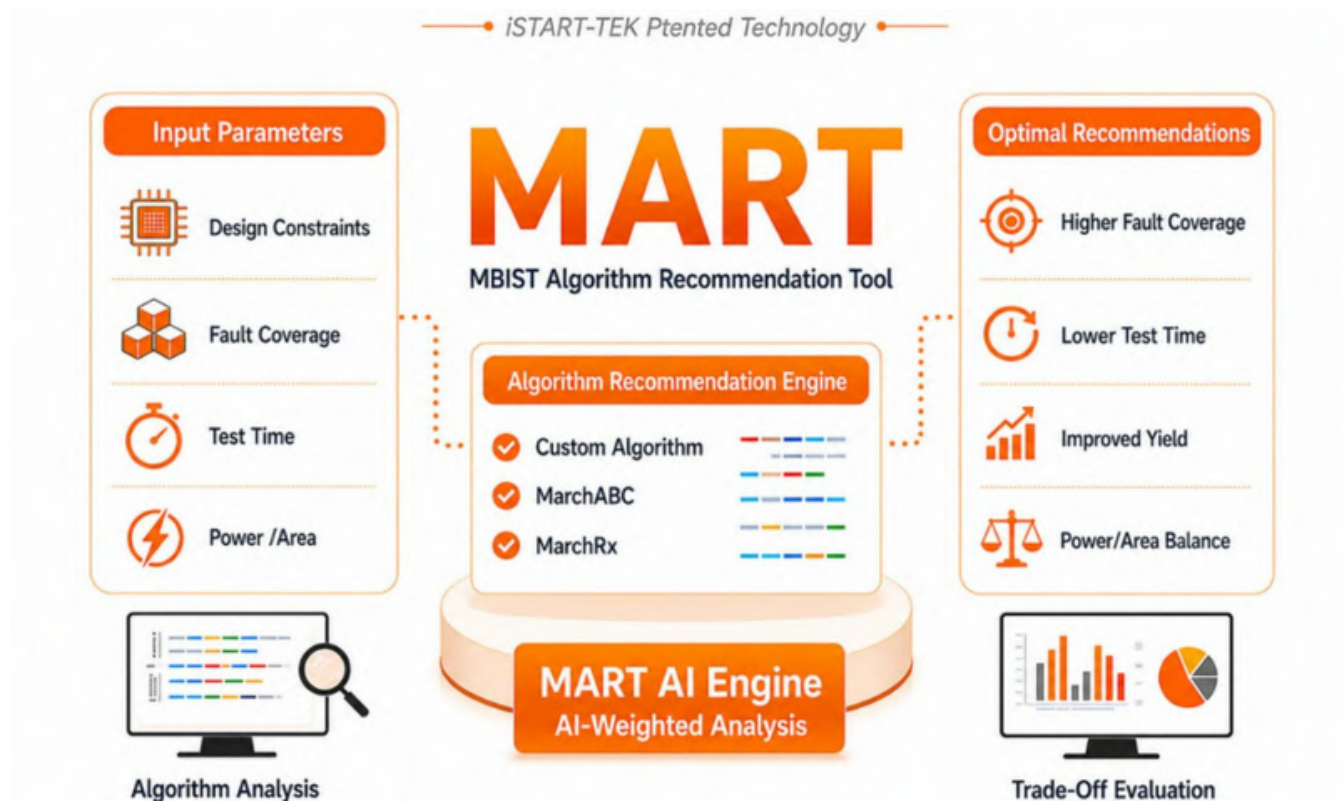
為提升頻寬與降低延遲，HBM 與 GPU、NPU、AI Accelerator 的垂直堆疊正快速普及，也讓記憶體測試的重要性同步提升。芯測科技的 IEEE 1838 解決方案，即是針對此類異質整合與高密度堆疊情境所設計，可支援多顆 die 的測試存取與記憶體修復流程，協助客戶在先進封裝架構下維持高良率與高可靠度。

此外，在 CoWoS 與 3D-IC 架構下，一顆整合 HBM 與高階邏輯晶片的 AI 晶片，其製造與封裝成本相當高昂。若封裝完成後才發現記憶體缺陷，沒有有效修復機制將可能導致整顆模組報廢。芯測科技透過 IEEE 1838 與記憶體修復技術的整合，讓客戶即使在堆疊完成後，仍可進行精準故障診斷與修復，大幅降低先進封裝報廢風險，並提高最終產品良率與投資報酬率（ROI）。

市場預期，在 HBM、Chiplet、CoWoS 與異質整合架構持續擴張下，IEEE 1838 將逐步成為 AI 晶片產業鏈的重要測試標準，而具備完整記憶體測試與修復能力的 EDA 工具供應商，也將在下一波 AI 半導體競爭中取得關鍵位置。

芯測科技強調，目前公司已具備完整的 IEEE 1838 記憶體測試與修復解決方案，可協助客戶因應 AI Chiplet、HBM 與 3D-IC 時代下的測試與良率挑戰，提前布局下一代先進封裝市場。

用 AI 解決 AI！AI 運算帶動記憶體複雜度 提升 MBIST 更需要智慧化演算法



生成式 AI 與 AI agent 快速發展，企業投資焦點正從單一加速器延伸至各類 AI 處理器與伺服器 CPU。這些晶片的共通特徵，是內建大量 SRAM 與 cache 以支撐資料密集運算。當記憶體容量與架構持續放大，測試與修復難度同步提高，MBIST 演算法的選擇逐漸成為影響量產良率、DPPM 與開發時程的重要決策。

在車用 AI、ADAS、智慧座艙與 Edge AI 裝置逐步普及的情況下，晶片品質門檻持續提高。產品不僅需要更高測試覆蓋率，還必須在測試時間、功耗、面積與量產節奏之間取得平衡。工程團隊在規劃記憶體測試策略時，往往面臨多重條件交互影響的複雜情境。傳統以查表與經驗為主的演算法選擇流程，正逐漸成為設計與測試流程中的隱形瓶頸。此時，若有智慧化的演算法推薦工具，將能有效突破測試效率的困境。

芯測科技開發的 MBIST 測試演算法推薦工具（MBIST Algorithm Recommendation Tool, MART），進化為 AI Test Algorithm Copilot，將原本高度依賴人工判斷的 MBIST 演算法選擇流程轉為 AI 輔助決策。系統可整合 memory type、memory size、port 架構、目標 DPPM 區間、failure information、fail bitmap、repair result 與 test time budget 等關鍵參數，主動推薦最適合的 March 測試組合、UDA 測試元素與診斷流程，協助工程團隊快速建立完整測試策略。



這項能力的核心，在於將過去「條件符合」的篩選模式，升級為多目標加權後的最佳化推薦。透過 AI 權重機制，MART 可在品質、測試時間與設計成本之間取得平衡，使 MBIST 策略更貼近實際產品需求。工程師不再需要反覆查詢文件與比對條件，即可在前期快速收斂測試方向，顯著降低決策成本並縮短開發週期。

當 AI 晶片與車規市場持續追求低 DPPM 與高可靠度，測試策略優化已從工程細節升級為競爭力的一部分。能否更快找到最佳測試組合，往往直接影響產品上市時程與量產品質。面對記憶體規模與複雜度不斷提升的產業趨勢，MBIST 演算法的選擇正在從經驗導向走向資料與 AI 導向。MART 的出現，象徵記憶體測試正式進入正式邁入 AI 輔助決策的新階段，為晶片設計團隊提供更高效率與更高品質的決策工具。

iSTART



芯測科技「記憶體測試算法推薦系統」 取得新型專利 重新定義 SoC 測試開發流程



記憶體測試與修復領導廠商芯測科技（iSTART-TEK，股票代碼：6786）今日宣布，其自主研發之「記憶體測試算法推薦系統」正式取得台灣新型專利。此系統即為芯測於2026年推出的MBIST 檢測演算法推薦工具（MBIST Algorithm Recommendation Tool, MART）透過 AI 協作，自動化決策流程，顯著提升記憶體測試開發效率。

隨著半導體製程邁向奈米級微縮，系統單晶片（SoC）內的記憶體容量與複雜度呈幾何倍數成長，如何從成百上千種測試演算法中找出最佳解，已成為設計人員的巨大挑戰。傳統上，記憶體測試演算法的選擇極度依賴資深工程師的經驗，若選擇不當，不僅可能漏掉潛在缺陷，還會造成測試時間過長，拉高生產成本。芯測科技本次獲專利之系統，正是為了解決這項產業長期存在的低效率問題。

此系統首先透過「提問產生模組」與「介面模組」，依序產生提問資訊並接收相對應的回應，藉此精確擷取記憶體架構與應用場景的各項特徵參數；隨後，「算法推薦模組」會基於這些參數決定最適用的測試算法，並由「配置檔產生模組」自動生成對應的配置資訊。

透過這套工具，工程師能快速獲得最佳化測試算法與相關配置檔，有效縮短開發時程並提升記憶體測試的效率。

芯測科技表示，這項專利的取得，代表公司不只能提供強大的測試工具，更能快速協助客戶做出正確的測試決策。透過參數化的引導，即便是經驗尚淺的工程師，也能在極短時間內配置出專業級的記憶體測試方案。這項技術的核心價值，在於將複雜的測試經驗轉化為數位化的邏輯判斷，從源頭優化 SoC 的測試涵蓋率與生產效率。



閱讀更多

芯測科技將推出完整 **DFT** 解決方案 強化先進晶片可
測試性與量產品質

[Read more](#)

DDI 功能持續升級 帶動 **SRAM** 容量成長 **MBIST**
與 **MBISR** 重要性同步提升

[Read more](#)

芯測科技記憶體測試與修復技術助力多元晶片量產
累計出貨突破三億顆

[Read more](#)

iSTART

外掛記憶體漲價 DDI 架構正在改變？ 內嵌 SRAM 價值重新浮現



近期記憶體市場價格波動再次成為半導體產業焦點。尤其在 DRAM、利基型記憶體價格持續走升的情況下，許多依賴外掛 memory 的系統設計正面臨新的成本壓力。對顯示驅動晶片（Display Driver IC, DDI）產業而言，影響層面不僅是採購成本問題，更牽涉下一代產品的架構選擇。

在小尺寸顯示應用，例如智慧手機 OLED、穿戴裝置或車載小型面板中，DDI 通常傾向將顯示緩衝區直接做在晶片內部，以內嵌 SRAM 作為 frame buffer。這類架構的優勢包括低延遲、低功耗、封裝簡化，且有助於裝置輕薄化。

但到了大尺寸面板，例如電視、高階顯示器或大型車載中控螢幕，情況就不同了。當解析度提升至 4K、8K，影像資料量暴增，若完全依靠內嵌記憶體，SRAM 面積可能大幅膨脹，導致晶片裸晶尺寸過大，進一步推高成本並影響良率。因此，這類 DDI 過去多採用外掛 DRAM 或 SDRAM，以較低的單位記憶體成本換取容量彈性。

然而市場條件正在改變。外掛記憶體的最大優勢一直是容量成本低，但其弱點也越來越明顯。首先是供應鏈風險。外部記憶體價格容易受到景氣循環、產能調整與供需失衡影響，造成 BOM 成本難以預估。其次是系統功耗與延遲問題。資料在 DDI 與外部記憶體間頻繁搬移，不僅增加 I/O 功耗，也會形成頻寬瓶頸。



相較之下，內嵌 SRAM 雖然占用晶片面積較大，卻能提供極高頻寬與極低延遲，尤其在高刷新率、高畫質顯示需求持續增加的情況下，其技術價值正在上升。值得注意的是，當外掛記憶體價格上漲，而 SRAM 成本相對穩定時，企業勢必會重新評估架構的成本組合。

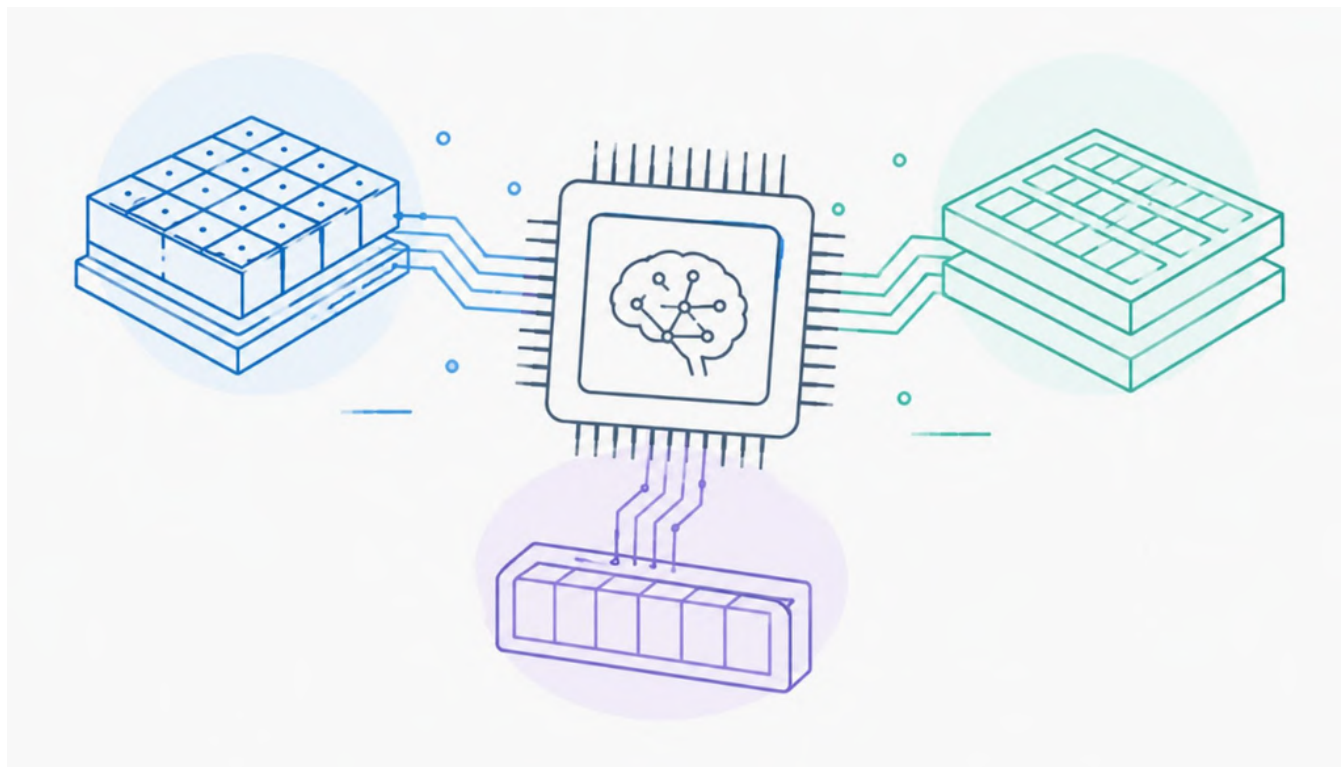
過去企業可能認為增加 SRAM 成本太高，因此選擇外掛記憶體；但當外掛記憶體不再便宜，甚至帶來供應風險時，直接提升 DDI 內嵌 SRAM 容量，反而可能成為更具策略性的選擇。

當然，內嵌 SRAM 也會面臨其他的挑戰。SRAM 容量增加，意味著晶片上記憶體面積占比持續上升。在先進製程中，許多 SoC 的 SRAM 面積占比甚至已超過 50%。記憶體越大，製程缺陷造成 fail bit 的機率越高，進而影響整體良率。因此 DDI 架構若要往高 SRAM 容量發展，關鍵在於後端是否具備足夠成熟的記憶體測試與修復能力。

透過 MBIST (Memory Built-In Self-Test) 與 Memory BISR (Memory Built-In Self-Repair) 技術，晶片可在晶圓測試階段提前找出 SRAM 中的缺陷區塊，並透過備援機制完成修復，有效提升整體良率。企業不必因為擔心良率下降而限制 SRAM 容量，而能透過先進的記憶體測試與修復技術，在提升效能的同時維持量產穩定性。

這也是芯測科技的核心優勢所在。芯測科技長期深耕 Memory Test、Repair 與 Diagnosis 技術，提供完整的 MBIST、MBISR 與記憶體測試演算法平台，協助客戶在設計階段就建立高覆蓋率的 SRAM 測試策略。面對 DDI、AI 晶片與車用 SoC 中不斷增加的 SRAM 容量需求，芯測科技能協助客戶有效提升記憶體良率、降低量產風險，並加速產品導入市場。

AI 推論市場擴大， SRAM/Cache/Buffer 品質成為量產新門檻



當 AI 產業從大型模型訓練走向推論與 Edge AI 部署，晶片設計的重點也開始改變。過去 AI 晶片競爭集中在算力規模與平行運算能力，如今真正影響系統效率的，往往是晶片內部資料能否被快速、穩定且低功耗地搬移與存取。這也使得 AI 推論架構中的核心資料流系統——SRAM、Cache 與 Buffer 的重要性快速提升。

其中，SRAM 主要負責高速資料暫存與即時運算支援。在 AI 推論過程中，大量模型權重、中間運算結果與啟動資料都需要被反覆讀寫，若完全依賴外部 DRAM，不僅延遲增加，功耗也會大幅上升。因此，許多 AI 加速器與 Edge AI 處理器都會內建大量 SRAM 作為本地記憶體，以維持高吞吐量與低延遲特性。

然而，當 SRAM 容量與數量持續增加，記憶體缺陷風險也同步上升。先進製程下常見的弱位元、資料保持不穩、讀取干擾與電阻性缺陷，可能在高頻推論或長時間運作下逐漸放大，進一步影響 AI 模型穩定性與整體 DPPM 表現。

除了 SRAM 之外，Cache 架構也正快速成為 AI 晶片設計核心。AI 運算與傳統 CPU 工作負載最大的差異之一，在於資料重複利用率極高。同一組模型權重與特徵資料，可能在短時間內被大量運算單元重複存取。若 Cache 效率不足，系統便需要頻繁回讀外部記憶體，不僅拖慢推論速度，也會增加功耗與頻寬壓力。



因此，許多 AI SoC 開始導入多層式 Cache、共享 Cache 與分散式 Cache 架構，希望提升資料重複利用效率。但隨著 Cache 結構日益複雜，資料一致性、時序變異與存取穩定性問題也更加明顯。特別是在高溫、低電壓或長時間AI運作情境下，Cache 可靠度已逐漸成為影響 AI 系統穩定性的關鍵因素之一。

與此同時，Buffer 受到的重視也日趨顯著。AI 推論過程涉及大量資料流調度，不同運算模組之間需要持續進行資料交換與同步，因此 Buffer 架構會直接影響整體資料流效率。從 input buffer、weight buffer 到 feature buffer，Buffer 除了是暫存空間，更負責平衡不同運算節點間的資料流速率。

若 Buffer 設計或品質不穩定，容易產生資料壅塞、時序錯配或資料遺失問題，進而影響整體吞吐量。尤其在 Edge AI 與即時推論應用中，Buffer 延遲與穩定性更直接關係到系統反應速度與即時性。

當 AI 晶片中的 SRAM、Cache 與 Buffer 規模同步成長，記憶體測試與修復策略也必須進一步升級。傳統固定式記憶體測試流程，已難以全面對應不同記憶體架構與 AI 工作負載行為。

針對這類需求，芯測科技（iSTART-TEK）推出的 MART、UDA 與 TEC 技術，正提供更具彈性與客製化能力的 SRAM 測試架構。

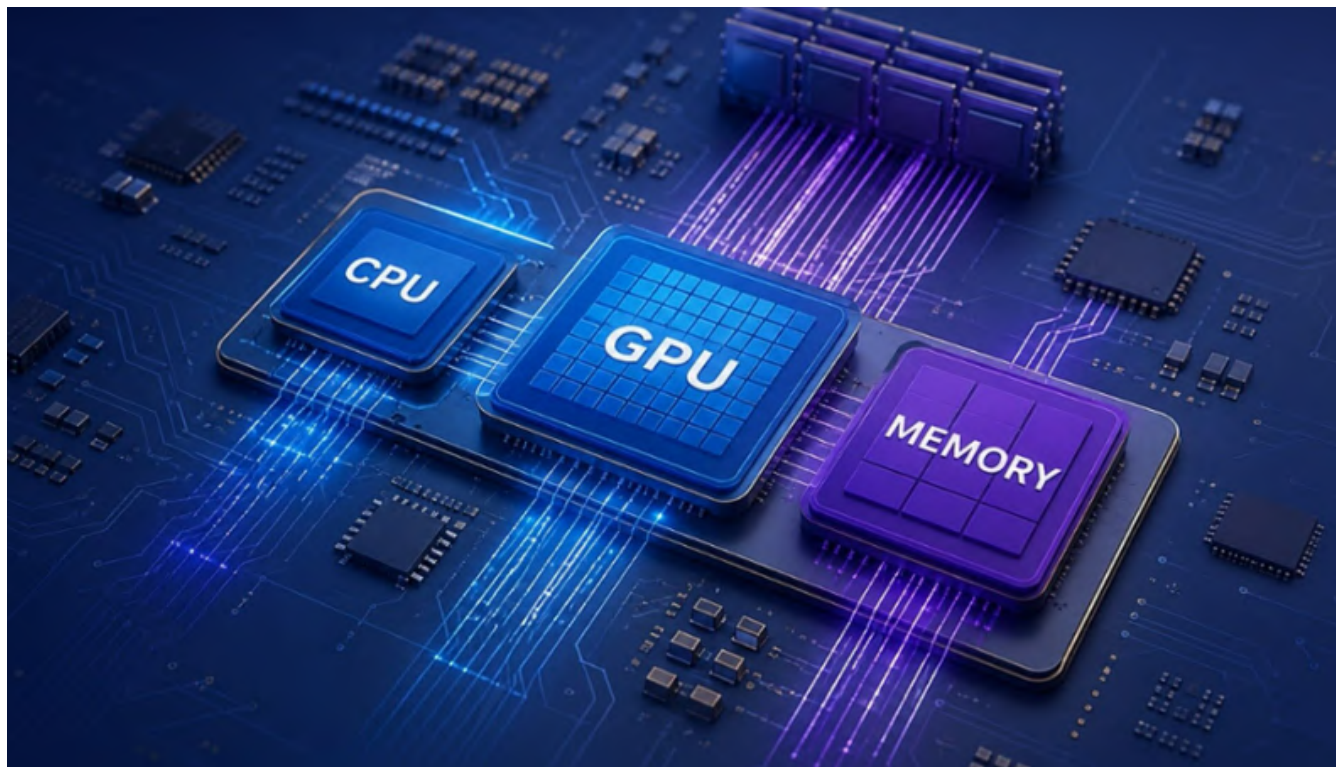
MART（MBIST Algorithm Recommendation Tool）是一套 AI 驅動的演算法分析系統，可根據應用類型、DPPM 目標、功耗、效能與面積限制，協助使用者快速篩選適合的 SRAM 測試演算法，可有效簡化演算法選擇流程與測試規劃難度。

UDA（User-Defined Algorithms）則提供類似積木概念的平台架構，讓使用者能自行定義基礎測試元素，並組合成不同的記憶體測試演算法。透過模組化方式，工程師可依不同 SRAM、Cache 與 Buffer 特性，建立更符合產品需求的測試流程，提升測試彈性與缺陷覆蓋能力。

TEC（Testing Elements Change）可讓測試工程師在 CP 與 FT 階段，依據不同測試環境重新調整或組合 SRAM 測試演算法。由於 SoC 測試常涉及極端電壓與溫度條件，不同環境可能對應不同記憶體缺陷型態，TEC 可協助工程師快速客製化測試元素，建立更符合特定需求的替代演算法。

當AI產業開始進入大規模推論與長時間運作時代，SRAM、Cache 與 Buffer 已不再只是配角，而是決定 AI 晶片效能、功耗、良率與可靠度的核心基礎。未來 AI 晶片的競爭，也將越來越取決於 Memory subsystem 本身的品質與測試能力。

GPU 之後的競爭焦點： AI 系統效能正在重新定義



生成式 AI 與大型語言模型快速發展，使「算力」長期被視為產業核心，而市場也持續聚焦在 GPU。然而，當 AI 應用從模型訓練走向大規模推論、從雲端延伸至邊緣裝置、並進一步演進為 AI Agent 與多模型協作架構時，系統瓶頸正逐步轉移。AI 產業正從 GPU-centric 架構，走向 CPU+GPU+記憶體協同運算的新階段。

在推論與 AI Agent 的情境下，系統複雜度顯著提升。CPU 負責任務排程、邏輯控制與資料前處理，GPU 承擔模型推論核心運算，而記憶體則成為資料與模型狀態流動的關鍵節點。AI Agent 需要持續呼叫工具、維持短期與長期記憶、並進行多步驟推理，使整體運作更接近「系統工程」。

在這樣的架構中，資料搬移成本快速上升，逐漸成為效能瓶頸。模型權重與中間特徵需頻繁在 CPU、GPU 與各類加速器間流動，而推論結果也需即時回寫與整合。當模型規模與 agent 任務複雜度同步提升時，記憶體頻寬、延遲與快取效率，成為決定系統吞吐與使用者體驗的關鍵因素。

因此，記憶體正從支援角色轉變為核心關鍵。無論是 HBM、LPDDR，或是處理器內部的 SRAM 與 Cache，都構成 AI 系統效能的底層基礎。特別是在 AI SoC 中，SRAM 不僅負責高速暫存，更直接影響延遲、能耗與整體運算效率，使記憶體設計逐漸成為與運算單元同等重要的競爭維度。



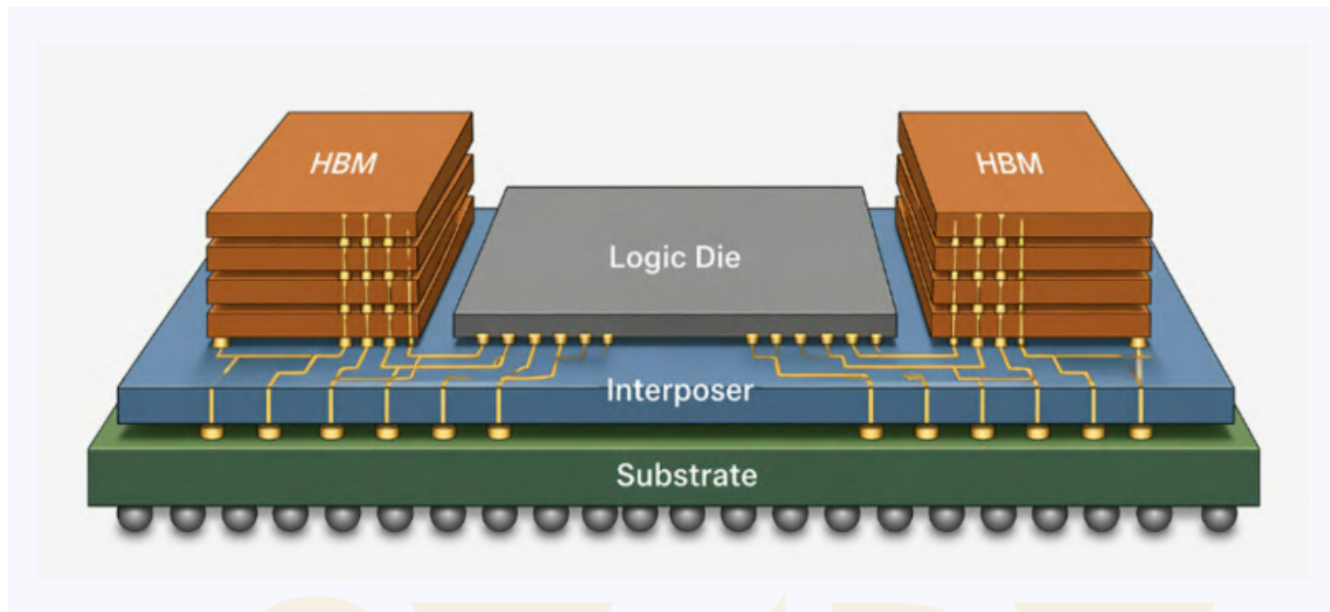
然而，隨著記憶體在晶片中的比重提升，其可靠度問題也同步放大。製程變異、老化效應與工作負載壓力，都可能導致 SRAM fault 或隱性錯誤累積，使得「可測試性」與「可修復性」成為 AI 晶片量產與長期運行的關鍵能力；這也是芯測科技長期專注地技術領域。

芯測科技聚焦於 AI 處理器內部記憶體的完整解決方案，透過 MBIST 與 MBISR 架構，結合 UDA（User-Defined Algorithms）提升故障診斷效率與覆蓋率，使記憶體問題能及早被精準定位並修復，降低後段風險與 field failure 機率。

同時，芯測科技的 MART（MBIST Algorithm Recommendation Tool）進一步將記憶體測試與修復流程智慧化，可依不同記憶體型態、缺陷特徵與應用場景，自動推薦最佳測試與修復策略，提升測試效率並縮短開發週期。在 AI 晶片高度客製化與快速迭代的趨勢下，這類工具化能力正成為提升量產效率的重要基礎。

整體而言，AI 基礎建設的競爭已不再只是 GPU 算力的比拚，而是轉向 CPU、GPU 與記憶體協同效率的系統級競爭。當運算架構走向高度整合與異質化，記憶體的穩定性與可修復性將直接影響 AI 系統的規模化能力與可靠度。在這樣的產業轉折點上，記憶體測試與修復技術，已成為支撐 AI 時代持續擴張的重要基礎。

CoWoS 與 HBM： AI 時代記憶體產業的關鍵引擎



近年 AI 運算需求爆發，大型語言模型、生成式 AI 與高效能運算（HPC）工作負載，需要在短時間內存取海量權重與中間資料，使記憶體頻寬與容量成為系統效能的核心指標。也因此，「資料搬移」逐漸成為半導體產業的瓶頸。在這樣的背景下，台積電所研發的 CoWoS（Chip-on-Wafer-on-Substrate）先進封裝技術與 HBM（High Bandwidth Memory）形成高度共生關係，重新定義 AI 伺服器與高階運算平台的架構。

首先，CoWoS 的最大貢獻在於打破傳統記憶體頻寬瓶頸。過去 CPU 或 GPU 與外部 DRAM 之間依賴封裝外的電路板連接，訊號傳輸距離長、功耗高且頻寬受限。CoWoS 透過矽中介層（interposer），讓邏輯晶片與多顆 HBM 在單一封裝內緊密整合，將資料傳輸距離縮短到毫米等級，大幅提升 I/O 密度與傳輸效率。這種設計讓 AI 加速器能以極高頻寬存取記憶體，成為現代 AI GPU 與加速器的標準配置。隨著大型模型對 KV Cache（鍵值快取）的需求急遽增加，高頻寬與大容量記憶體已從「加分項」變成「必要條件」。

其次，CoWoS 推動了 HBM 技術世代的快速演進。AI 模型參數量持續成長，使單一封裝內的記憶體容量需求不斷攀升，從 HBM2、HBM2e 到 HBM3、HBM3e，堆疊層數與頻寬持續提高。未來 AI 晶片甚至可能整合更多堆疊顆粒，以支援更大模型與更複雜的推論任務。換言之，先進封裝能力已成為決定記憶體規格升級速度的重要因素。

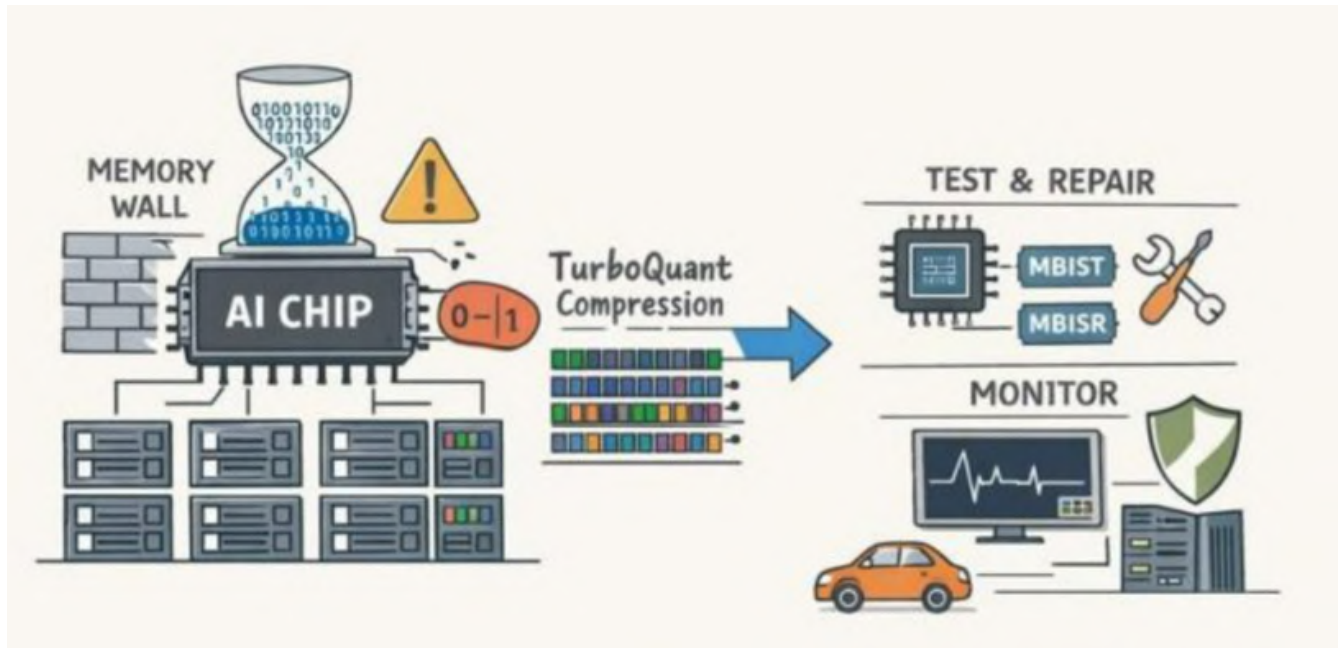


這股需求也改變了整體記憶體市場的供需結構。由於 HBM 屬於高毛利產品，DRAM 製造商將大量產能轉向 HBM 生產，導致傳統 PC 與消費性電子所使用的 DRAM 供給相對收縮。另一方面，CoWoS 產能本身具有高度技術門檻與設備限制，AI 晶片的出貨量受制於先進封裝產能，使 HBM 供不應求的情況可能延續至 2027-2028 年。即使業界持續探索降低記憶體需求的技術，例如優化快取機制或新型資料壓縮方法，對於頂級 AI 運算而言，高頻寬記憶體的地位仍難以被取代。

展望未來，封裝技術仍將持續向更高整合度演進。繼 CoWoS-S、CoWoS-R 與 CoWoS-L 之後，混合鍵合（Hybrid Bonding）被視為下一波關鍵技術，透過更細微的互連間距與更高的 I/O 密度，使記憶體與邏輯晶片幾乎達到「晶片級」的连接。這不僅能進一步降低功耗，也能顯著提升整體系統效能，為下一世代 AI 運算鋪路。

在這樣的產業趨勢下，HBM 不僅是高效能運算的核心元件，也成為良率與可靠度挑戰最嚴峻的記憶體類型之一。堆疊式結構與先進封裝使測試與修復難度大幅提高，如何在出貨前確保記憶體品質與長期可靠度，成為產業競爭的關鍵。芯測科技長期投入記憶體測試與修復技術，並擁有 ISO 26262 TCL1 認證的車規級測試演算法與修復方案，協助晶片在先進封裝與高頻寬架構下維持高良率與高可靠度，成為 AI 時代記憶體品質把關的重要支撐力量。

從 Google TurboQuant 看 AI 晶片設計的新隱憂



在 AI 模型參數呈指數級增長的今天，「記憶體牆（Memory Wall）」已成為所有晶片設計者的共同敵人。近期 Google 發表的 TurboQuant 技術，無疑在產業內投下了一顆震撼彈。這項技術揭示了未來五年 AI 晶片發展的核心趨勢：在有限的空間與頻寬內，如何榨取極致的數據價值。

數據密度越高 容錯空間越小

Google TurboQuant 的核心任務在於優化大型語言模型（LLM）推論時的 KV Cache。它能將數據進行高達 6 倍的壓縮，這意味著原本需要六台伺服器才能跑的模式，現在可能只需要一台。

然而當我們沉浸於演算法帶來的「軟體紅利」時，半導體產業鏈的另一端卻面臨著全新的挑戰。

這是一個物理學上的必然：當數據被極度壓縮，每一個位元所承載的資訊權重就越高。在未壓縮的數據中，一個位元的翻轉（Bit Flip）可能只是一次微小的雜訊；但在 TurboQuant 這種高倍率量化壓縮後，關鍵位元的錯誤可能會導致整個推論結果的邏輯崩潰。

另外，為了配合此類壓縮技術，AI 晶片內部整合了更密集的 SRAM 與高效能記憶體架構。隨著製程微縮至 3 奈米甚至更先進節點，記憶體單元的良率與長期穩定性，已成為決定 AI 晶片能否量產並商用的關鍵。



從「好用」到「耐用」：不可或缺的底層守護

當產業焦點都在關注 AI 如何減少記憶體用量時，晶片設計者的下一步則是思考：這些被極度壓榨的記憶體，如何穩住表現、不要出差錯？這正是芯測科技（iSTART-TEK）在生態系中所扮演的角色。

在這種高效能運算（HPC）與 AI 晶片的演進路徑上，芯測科技提供的 記憶體測試與修復（DFT/MBIST/BISR）方案，實際上是在背後支撐TurboQuant 這類軟體紅利的的安全基座。

- 1.測試（Test）： 芯測透過 START™ v5 等先進工具，在晶片設計階段便植入強大的 MBIST（記憶體自我測試）功能，精準篩選出缺陷。
- 2.修復（Repair）： 藉由 MBISR（記憶體自我修復）技術，在發現位元錯誤時自動啟用備用空間進行修復，將原本可能報廢的晶片化腐朽為神奇，直接提升產量良率。
- 3.監控（Monitor & Security）： 在晶片運作的生命週期中，透過監控方案即時掌握記憶體健康狀態，這在追求高度可靠性的 HPC與車用領域尤為重要。

「軟硬結合」才是 AI 真正的未來

Google TurboQuant 的發表，讓我們看到了 AI 演算法突破瓶頸的決心。但這場記憶體革命的下半場，必然會回歸到硬體底層的可靠度競爭。從 Google 解決「如何節省空間」，到芯測科技解決「如何確保品質」，唯有記憶體的絕對穩定，軟體端的壓縮技術才能真正發揮實力。



**Episode 51: 從測試到智慧防護：記憶體品質
管控工具的全面進化**

[Watch the video](#)

**Episode 50: 打造可客製化記憶體測試方法
靈活應變快速演進的半導體時代**

[Watch the video](#)

**Episode 49: AI 家電與智慧顯示時代：從語音
控制到記憶體內運算的晶片測試挑戰**

[Watch the video](#)

**Episode 48: 智慧座艙 (In-vehicle AI)：汽車
電子的下一個算力戰場**

[Watch the video](#)

**Episode 47: 當 MRAM 遇上 AI：下一代智慧
晶片的關鍵記憶體革命**

[Watch the video](#)

**Episode 46: NPU 運作需求下的 SRAM 重要性
與測試策略**

[Watch the video](#)

Episode 45: LPU 使用 SRAM 的原因與測試關鍵

[Watch the video](#)