

iSTART

July 2026

iSTART iReport

芯测科技电子报第 16 期



Contents

News

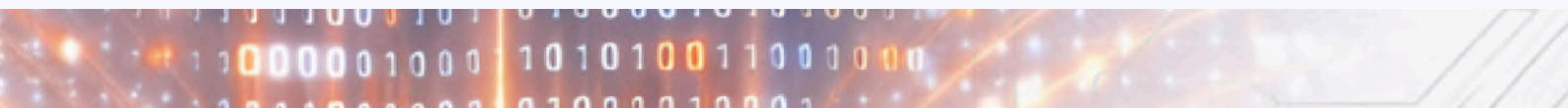
- AI 芯片良率成竞争关键 芯测科技以内存质量管理技术抢攻新商机2
- 黄仁勋点名 AI 工厂时代来临 芯测科技布局高可靠度芯片测试迎接新商机.....4
- 华为「韬定律」掀先进封装战局 芯测科技已抢先布局 IEEE 18386
- 用 AI 解决 AI！AI 运算带动内存复杂度提升 MBIST 更需要智能化算法8
- 芯测科技「内存测试算法推荐系统」取得新型专利 重新定义 SoC 测试开发流程10

IC Tech Insights

- 外挂内存涨价 DDI 架构正在改变？内嵌 SRAM 价值重新浮现12
- AI 推论市场扩大，SRAM/Cache/Buffer 质量成为量产新门坎14
- GPU 之后的竞争焦点：AI 系统效能正在重新定义16
- CoWoS 与 HBM：AI 时代内存产业的关键引擎18
- 从 TurboQuant 看 AI 芯片设计的新隐忧20

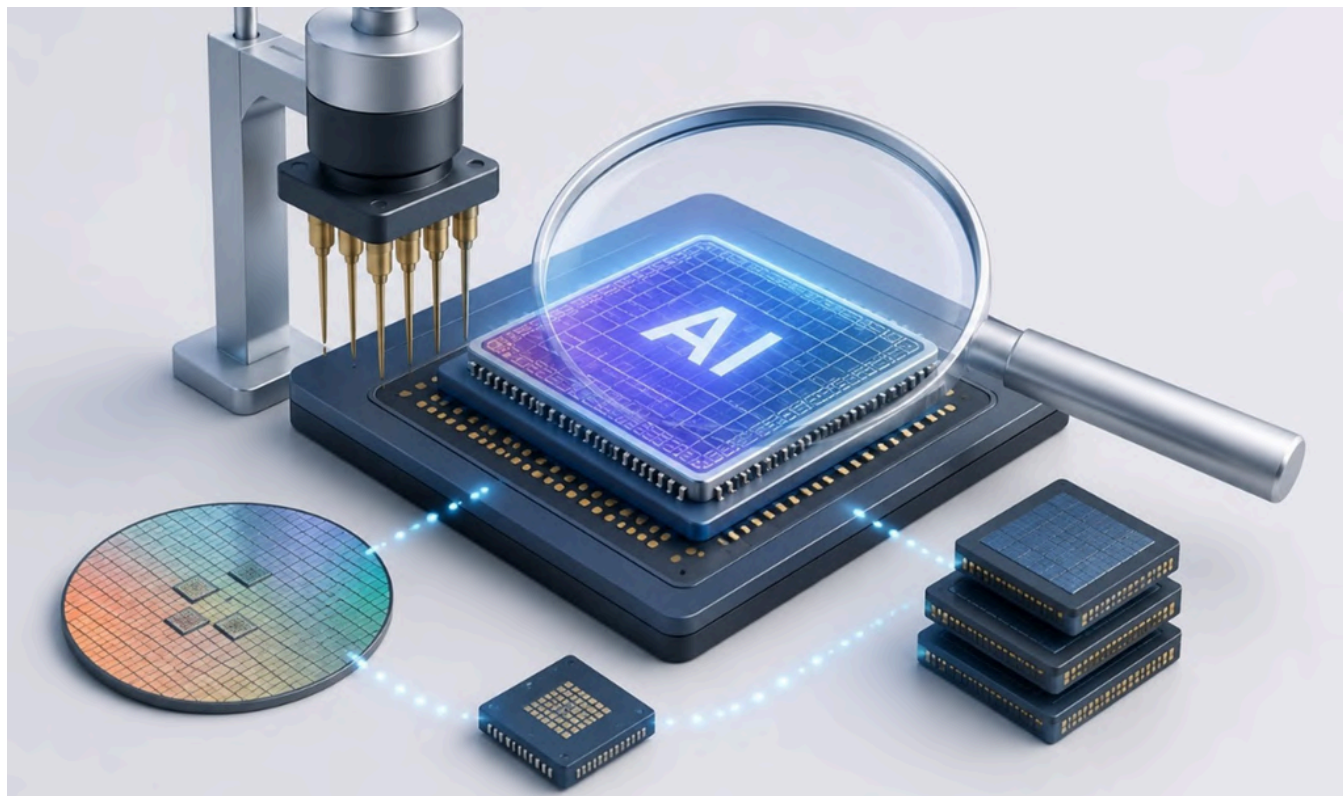
iSTART Class

- 从测试到智慧防护：内存质量管控工具的全面进化22
- 打造可客制化内存测试方法 灵活应变快速演进的半导体时代22
- AI 家电与智能显示时代：从语音控制到 CIM 的芯片测试挑战22
- 智慧座舱 (In-vehicle AI)：汽车电子的下一个算力战场22
- 当 MRAM 遇上 AI：下一代智能芯片的关键内存革命22
- NPU 运作需求下的 SRAM 重要性测试策略22
- LPU 使用 SRAM 的原因与测试关键22





AI 芯片良率成竞争关键 芯测科技以内存质量管理技术抢攻新商机



AI 芯片的竞争，正在从算力竞赛走向良率竞赛。当高效能运算芯片的设计愈来愈复杂，单颗芯片价值也持续攀升，任何微小缺陷都可能带来巨大损失。如何在量产阶段维持稳定质量与高良率，成为芯片厂的重要课题。

在影响芯片良率的众多因素中，内存可靠度是最关键的一环。AI 芯片内部大量使用 SRAM 作为快取、数据缓冲与运算暂存空间，直接影响数据存取效率、延迟与功耗表现。随着内存容量持续增加，SRAM 测试与修复的重要性也同步提升。

然而先进制程下的 SRAM 缺陷已不再只是传统的开路或短路问题，还可能出现写入异常、数据保持失效、动态错误，以及制程变异造成的边界型失效。这些缺陷不仅更难侦测，也更容易成为影响 AI 芯片良率的隐性风险。

芯测科技长期深耕内存测试、修复与质量管理技术，透过核心平台 START™ v5，协助客户在芯片设计时间就建立完整的内存质量防护机制。

START™ v5 提供完整的内存内建自我测试与内建自我修复架构，透过内建测试机制，系统能快速找出失效内存单元，提高缺陷检出能力，降低漏测风险。搭配内建修复技术后，系统可进一步



启用备援列或备援字段，自动修复损坏的内存区块，使原本可能报废的晶粒恢复正常运作，提升整体良率。

对高价值 AI 芯片而言，这项能力带来的效益尤其明显。即使良率仅提升少量百分比，也可能转化为可观的成本节省与产能提升。

除了基础测试与修复能力，芯测科技的用户自定义算法（User-Defined Algorithms, UDA）功能也提供高度弹性，让客户能依据不同内存架构、制程特性与应用需求，自行建立专属测试流程。由于 AI 芯片中的内存设计高度客制化，固定式测试算法已难以满足所有需求，更灵活的测试能力因此成为重要优势。

在测试效率方面，芯测科技的测试评估机制可协助分析不同算法的缺陷覆盖率与测试时间，帮助客户找出最佳测试组合，在测试质量与测试成本之间取得平衡。

此外，芯测科技的智能内存算法推荐技术 MART（MBIST Algorithm Recommendation Tool），能根据内存类型、制程节点与历史失效特征，自动推荐合适的测试算法组合，协助工程团队缩短开发时间，加快产品导入时程。

当 AI 芯片成本持续攀升，市场竞争的核心已不再只是制程领先，而是更有效控制缺陷、提升良率并确保质量稳定。芯测科技透过 START™ v5、内存测试与修复、客制化算法及智能化测试技术，持续协助客户提升产品可靠度与量产效率，成为 AI 半导体产业升级的重要技术伙伴。



黄仁勋点名 AI 工厂时代来临 芯测科技布局高可靠度芯片测试迎接新商机



NVIDIA 创办人暨执行长黄仁勋在 NVIDIA GTC Taipei 2026 主题演讲中指出，AI 产业已正式从「能不能用」迈向「能不能获利、能不能规模化」的新阶段，并将 AI Factory 定义为下一代数位基础建设。随着 Agentic AI、Physical AI、AI-native PC 与边缘运算快速发展，芯片系统对可靠度、可用率与长时间稳定运作的要求正大幅提升。

黄仁勋并提出「Compute is Revenue」观点，认为未来 AI 数据中心不只是运算中心，而是生产 AI Token 的工厂。当运算能力直接对应营收产出时，芯片失效、系统停机或可靠度不足所造成的损失，将直接影响企业获利能力的商业问题。

在此趋势下，专注于内存测试、修复与 DFT 技术的芯测科技（iSTART-TEK）正站上 AI 基础建设升级的重要位置。

近年来，AI 芯片、HPC 芯片与车用 SoC 持续增加 SRAM 容量，在先进制程中，SRAM 面积占比甚至超过芯片总面积的一半。随着 AI 工厂规模扩大，内存缺陷对系统稳定度与运算可用率的影响也更加显著。芯测科技所提供的 START™ v5 平台、SRAM BIST/BISR 解决方案以及 MART（MBIST Algorithm Recommendation Tool）技术，可协助芯片开发团队提升测试覆盖率、强化修复能力，并降低量产阶段的潜在风险。



此外，黄仁勋也特别强调 Agentic AI 将成为下一代运算模式，未来 AI 将能自主规划任务、调用工具并长时间执行工作流程。NVIDIA 更与 Cadence 合作推动芯片设计 Super Agent，预示 RTL 生成、验证、除错与设计流程将逐步导入 AI 工作流。

对此，芯测科技认为，未来 DFT、BIST、BISR、Repair Analysis 及 Diagnosis 所产生的大量测试数据，将有机会成为 AI 辅助设计与 AI 辅助除错的重要基础数据。透过将测试数据与 AI 分析能力结合，不仅能提升芯片开发效率，也能进一步优化产品质量与量产良率。

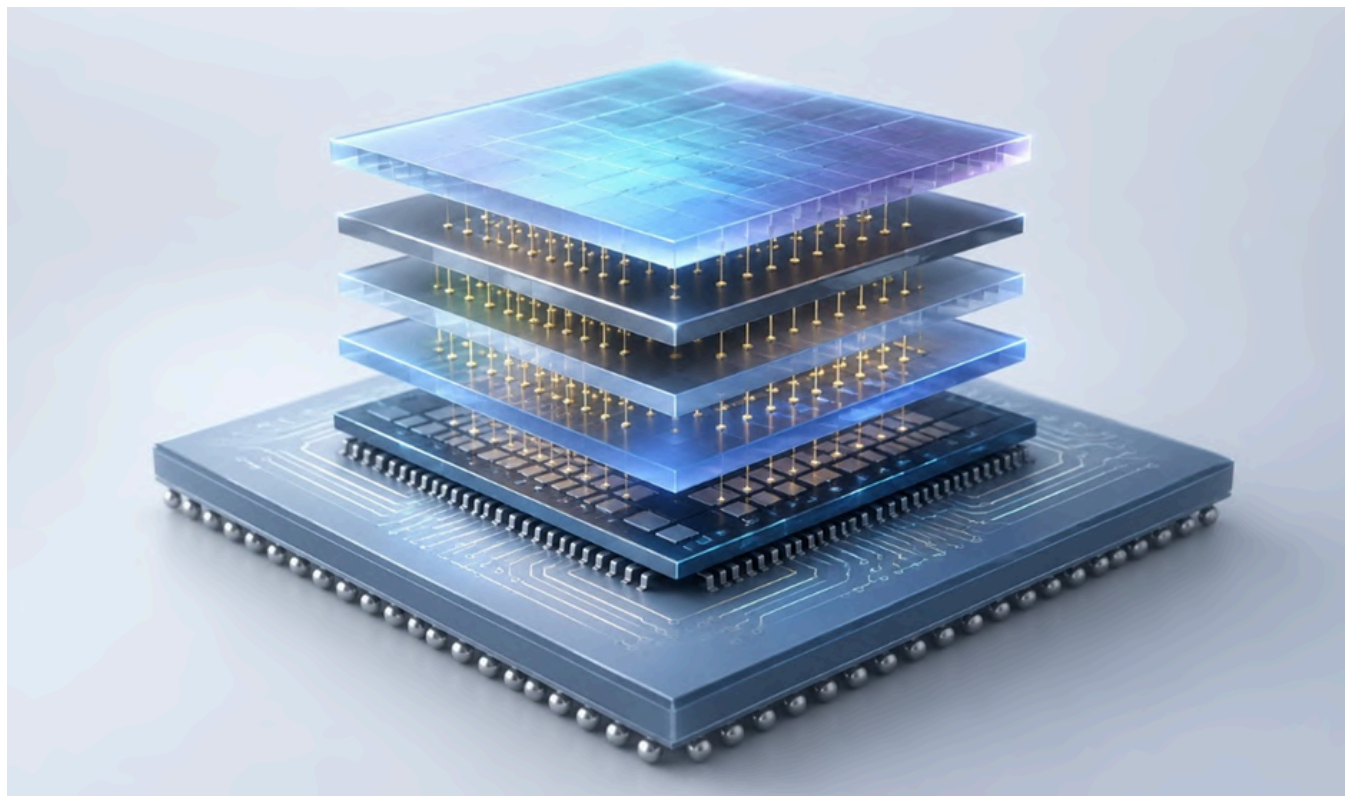
另一方面，随着 Physical AI、自动驾驶、机器人与边缘 AI 应用兴起，市场对功能安全 (Functional Safety)、长时间稳定运作以及高可靠度验证的需求也快速增加。黄仁勋指出，未来车用 AI 与机器人系统将需要更完整的测试验证流程，而 DFT、BIST、BISR 及诊断能力的重要性将持续提升。

芯测科技表示，当 AI 产业竞争从单纯追求算力，转向追求系统可靠度、可用率与长期稳定运作时，测试技术的价值也将从成本中心逐渐转变为竞争力来源。未来公司将持续深化内存测试算法、修复分析及先进 DFT 技术布局，协助客户打造符合 AI 工厂、Edge AI 与 Physical AI 时代需求的高可靠度芯片产品。

在黄仁勋描绘的 AI 基础建设蓝图中，芯片毫无疑问是创造营收的核心资产；而确保这些芯片稳定运作的测试与修复技术，也正成为 AI 时代不可或缺的关键基础能力。



华为「韬定律」掀先进封装战局 芯测科技已抢先布局 IEEE 1838



AI 运算需求持续爆发，半导体产业正逐步从「晶体管微缩」走向「系统级堆栈整合」的新竞争阶段。近期市场关注华为提出的「韬定律（Tao Law）」概念，强调透过先进封装、Chiplet 与多晶粒堆栈技术，突破传统摩尔定律在制程微缩上的限制。这也代表未来 AI 芯片的核心竞争力，已从单一芯片的晶体管密度，延伸为如何透过 3D-IC 与异质整合架构，提升整体系统效能、带宽与能效。

然而，当 AI 芯片逐步迈向 HBM、高速逻辑芯片与 Chiplet 的垂直堆栈架构后，测试与修复问题也开始成为先进封装时代的关键挑战。由于多颗die 封装完成后的价值极高，一旦在后段才发现内存或互连缺陷，将可能造成巨大的报废成本。因此，如何在 3D-IC 架构中维持可测试性（Testability）与可修复性（Repairability），已成为 AI 芯片产业链不可忽视的重要议题。

专注于内存测试与修复技术的芯测科技（iSTART-TEK）指出，随着产业从摩尔定律逐步转向堆栈整合架构，IEEE 1838 将成为 3D-IC 与异质整合芯片不可或缺的测试标准。芯测科技目前已率先完成 IEEE 1838 与内存测试、修复架构的整合，是目前少数能完整支持 IEEE 1838 的内存测试与修复 EDA 工具供货商之一。



IEEE 1838 是专为 3D-IC 与多 die 堆栈架构制定的测试标准，主要用于解决 die-to-die 测试存取、堆栈后测试信道管理、以及异质整合架构中的可测试性问题。在传统 SoC 架构中，测试主要针对单一 die；但在 HBM、Logic Die、AI Accelerator 与 Chiplet 大量堆栈的架构下，测试路径与故障定位的复杂度大幅提高，也让 IEEE 1838 逐渐成为先进封装的重要基础架构。

芯测科技表示，当产业开始以堆栈技术挑战摩尔定律时，测试架构也势必要同步升级。尤其在 CoWoS、3D-IC 与 HBM 等高阶封装架构中，一旦缺乏标准化测试机制，将难以有效管理多层 die 的测试存取与故障诊断。因此，未来只要走向多 die 堆栈与异质整合，IEEE 1838 几乎将成为必要选项。

芯测科技早在 3D-IC 尚未全面商用前，便已投入 IEEE 1838 相关技术开发，并将其整合进既有内存测试与修复平台中，完成与 MBIST、BISR 架构的深度串接。透过 IEEE 1838 架构，系统可在多层堆栈完成后，仍维持对 SRAM 与 HBM 的测试与修复能力，包含测试路径规划、故障定位、修复控制与测试存取管理等功能。

为提升带宽与降低延迟，HBM 与 GPU、NPU、AI Accelerator 的垂直堆栈正快速普及，也让内存测试的重要性同步提升。芯测科技的 IEEE 1838 解决方案，即是针对此类异质整合与高密度堆栈情境所设计，可支持多颗 die 的测试存取与内存修复流程，协助客户在先进封装架构下维持高良率与高可靠度。

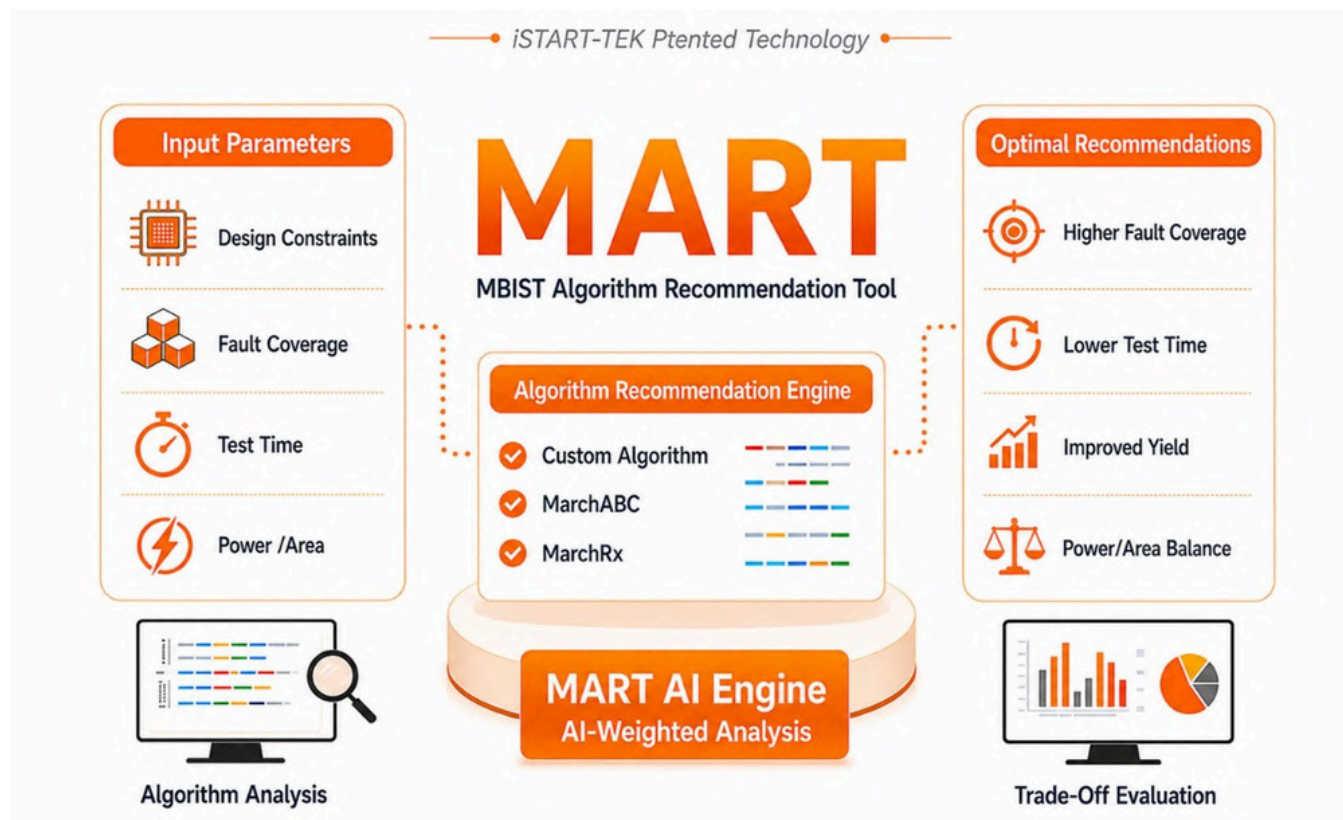
此外，在 CoWoS 与 3D-IC 架构下，一颗整合 HBM 与高阶逻辑芯片的 AI 芯片，其制造与封装成本相当高昂。若封装完成后才发现内存缺陷，没有有效修复机制将可能导致整颗模块报废。芯测科技透过 IEEE 1838 与内存修复技术的整合，让客户即使在堆栈完成后，仍可进行精准故障诊断与修复，大幅降低先进封装报废风险，并提高最终产品良率与投资报酬率（ROI）。

市场预期，在 HBM、Chiplet、CoWoS 与异质整合架构持续扩张下，IEEE 1838 将逐步成为 AI 芯片产业链的重要测试标准，而具备完整内存测试与修复能力的 EDA 工具供货商，也将在下一波 AI 半导体竞争中取得关键位置。

芯测科技强调，目前公司已具备完整的 IEEE 1838 内存测试与修复解决方案，可协助客户因应 AI Chiplet、HBM 与 3D-IC 时代下的测试与良率挑战，提前布局下一代先进封装市场。



用 AI 解决 AI ! AI 运算带动内存复杂度提升 MBIST 更需要智能化算法



生成式 AI 与 AI agent 快速发展，企业投资焦点正从单一加速器延伸至各类 AI 处理器与服务器 CPU。这些芯片的共通特征，是内建大量 SRAM 与 cache 以支撑数据密集运算。当内存容量与架构持续放大，测试与修复难度同步提高，MBIST 算法的选择逐渐成为影响量产良率、DPPM 与开发时程的重要决策。

在车用 AI、ADAS、智能座舱与 Edge AI 装置逐步普及的情况下，芯片质量门坎持续提高。产品不仅需要更高测试覆盖率，还必须在测试时间、功耗、面积与量产节奏之间取得平衡。工程团队在规划内存测试策略时，往往面临多重条件交互影响的复杂情境。传统以查表与经验为主的算法选择流程，正逐渐成为设计与测试流程中的隐形瓶颈。此时，若有智能化的算法推荐工具，将能有效突破测试效率的困境。

芯测科技开发的 MBIST 测试算法推荐工具（MBIST Algorithm Recommendation Tool, MART），进化为 AI Test Algorithm Copilot，将原本高度依赖人工判断的 MBIST 算法选择流程转为 AI 辅助决策。系统可整合 memory type、memory size、port 架构、目标 DPPM 区间、failure information、fail bitmap、repair result 与 test time budget 等关键参数，主动推荐最适合的 March 测试组合、UDA 测试元素与诊断流程，协助工程团队快速建立完整测试策略。



这项能力的核心，在于将过去「条件符合」的筛选模式，升级为多目标加权后的优化推荐。透过 AI 权重机制，MART 可在质量、测试时间与设计成本之间取得平衡，使 MBIST 策略更贴近实际产品需求。工程师不再需要反复查询文件与比对条件，即可在前期快速收敛测试方向，显著降低决策成本并缩短开发周期。

当 AI 芯片与车规市场持续追求低 DPPM 与高可靠度，测试策略优化已从工程细节升级为竞争力的一部分。能否更快找到最佳测试组合，往往直接影响产品上市时程与量产质量。面对内存规模与复杂度不断提升的产业趋势，MBIST 算法的选择正在从经验导向走向数据与 AI 导向。MART 的出现，象征内存测试正式进入正式迈入 AI 辅助决策的新阶段，为芯片设计团队提供更高效率与更高质量的决策工具。

iSTART



芯测科技「内存测试算法推荐系统」取得 新型专利 重新定义 SoC 测试开发流程



内存测试与修复领导厂商芯测科技今日宣布，其自主研发之「内存测试算法推荐系统」正式取得台湾新型专利。此系统即为芯测于2026年推出的 MBIST 检测算法推荐工具（MBIST Algorithm Recommendation Tool, MART）透过 AI 协作，自动化决策流程，显著提升内存测试开发效率。

随着半导体制程迈向奈米级微缩，系统单芯片（SoC）内的内存容量与复杂度呈几何倍数成长，如何从成百上千种测试算法中找出最佳解，已成为设计人员的巨大挑战。传统上，内存测试算法的选择极度依赖资深工程师的经验，若选择不当，不仅可能漏掉潜在缺陷，还会造成测试时间过长，拉高生产成本。芯测科技本次获专利之系统，正是为了解决这项产业长期存在的低效率问题。

此系统首先透过「提问产生模块」与「接口模块」，依序产生提问信息并接收相对应的响应，藉此精确撷取内存架构与应用场景的各项特征参数；随后，「算法推荐模块」会基于这些参数决定最适用的测试算法，并由「配置文件产生模块」自动生成对应的配置信息。

透过这套工具，工程师能快速获得优化测试算法与相关配置文件，有效缩短开发时程并提升内存测试的效率。

芯测科技表示，这项专利的取得，代表公司不只能提供强大的测试工具，更能快速协助客户做出正确的测试决策。透过参数化的引导，即便是经验尚浅的工程师，也能在极短时间内配置出专业级的内存测试方案。这项技术的核心价值，在于将复杂的测试经验转化为数字化的逻辑判断，从源头优化 SoC 的测试涵盖率与生产效率。



阅读更多

**芯测科技将推出完整 DFT 解决方案
强化先进芯片可测试性与量产质量**

[Read more](#)

**DDI 功能持续升级 带动 SRAM 容量成长
MBIST 与 MBISR 重要性同步提升**

[Read more](#)

**芯测科技内存测试与修复技术助力多元芯片量产
累计出货突破三亿颗**

[Read more](#)

iSTART



外挂内存涨价 DDI 架构正在改变？ 内嵌 SRAM 价值重新浮现



近期内存市场价格波动再次成为半导体产业焦点。尤其在 DRAM、利基型内存价格持续走升的情况下，许多依赖外挂 memory 的系统设计正面临新的成本压力。对显示驱动芯片（Display Driver IC, DDI）产业而言，影响层面不仅是采购成本问题，更牵涉下一代产品的架构选择。

在小尺寸显示应用，例如智能手机 OLED、穿戴装置或车载小型面板中，DDI 通常倾向将显示缓冲区直接做在芯片内部，以内嵌 SRAM 作为 frame buffer。这类架构的优势包括低延迟、低功耗、封装简化，且有助于装置轻薄化。

但到了大尺寸面板，例如电视、高阶显示器或大型车载中控屏幕，情况就不同了。当分辨率提升至 4K、8K，影像数据量暴增，若完全依靠内嵌内存，SRAM 面积可能大幅膨胀，导致芯片裸晶尺寸过大，进一步推高成本并影响良率。因此，这类 DDI 过去多采用外挂 DRAM 或 SDRAM，以较低的单位内存成本换取容量弹性。

然而市场条件正在改变。外挂内存的最大优势一直是容量成本低，但其弱点也越来越明显。首先是供应链风险。外部内存价格容易受到景气循环、产能调整与供需失衡影响，造成 BOM 成本难以预估。其次是系统功耗与延迟问题。数据在 DDI 与外部内存间频繁搬移，不仅增加 I/O 功耗，也会形成带宽瓶颈。



相较之下，内嵌 SRAM 虽然占用芯片面积较大，却能提供极高带宽与极低延迟，尤其在高刷新率、高画质显示需求持续增加的情况下，其技术价值正在上升。值得注意的是，当外挂内存价格上涨，而 SRAM 成本相对稳定时，企业势必会重新评估架构的成本组合。

过去企业可能认为增加 SRAM 成本太高，因此选择外挂内存；但当外挂内存不再便宜，甚至带来供应风险时，直接提升 DDI 内嵌 SRAM 容量，反而可能成为更具策略性的选择。

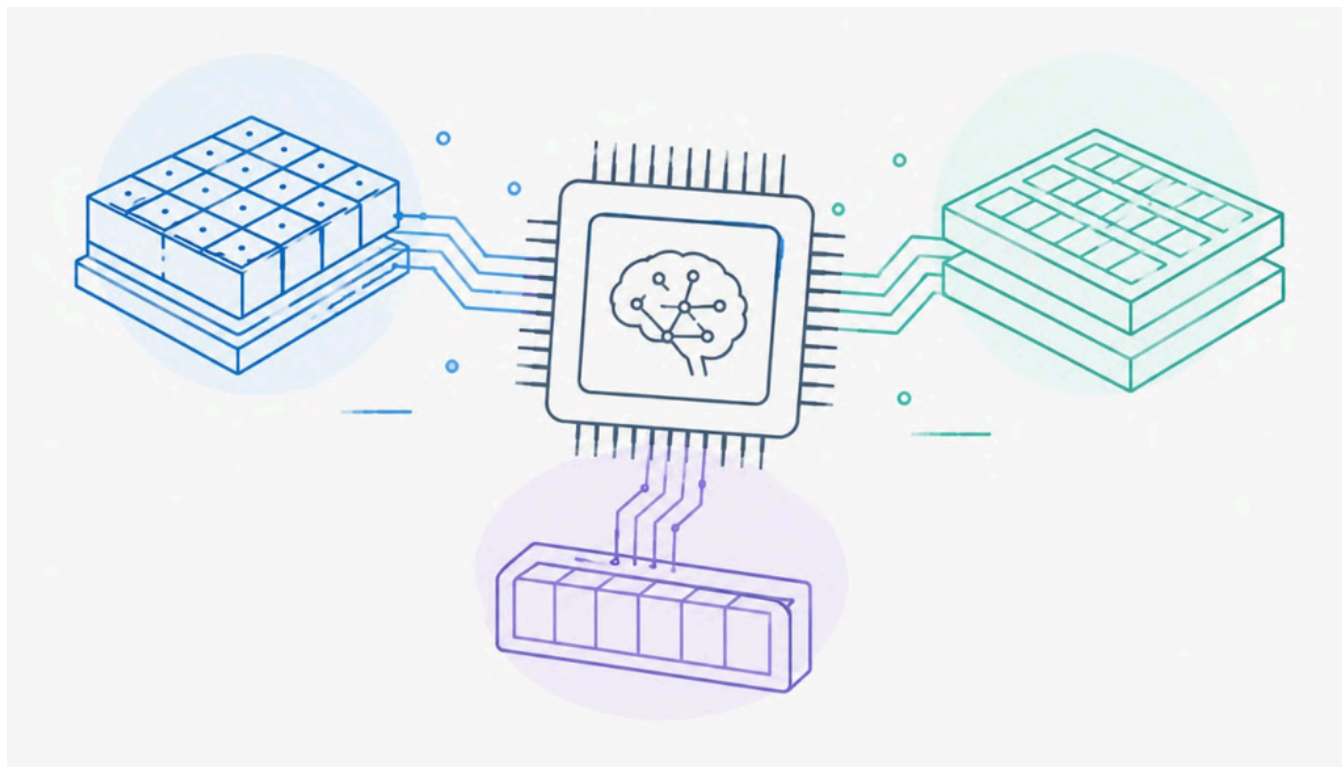
当然，内嵌 SRAM 也会面临其他的挑战。SRAM 容量增加，意味着芯片上内存面积占比持续上升。在先进制程中，许多 SoC 的 SRAM 面积占比甚至已超过 50%。内存越大，制程缺陷造成 fail bit 的机率越高，进而影响整体良率。因此 DDI 架构若要往高 SRAM 容量发展，关键在于后端是否具备足够成熟的内存测试与修复能力。

透过 MBIST (Memory Built-In Self-Test) 与 Memory BISR (Memory Built-In Self-Repair) 技术，芯片可在晶圆测试阶段提前找出 SRAM 中的缺陷区块，并透过备援机制完成修复，有效提升整体良率。企业不必因为担心良率下降而限制 SRAM 容量，而能透过先进的内存测试与修复技术，在提升效能的同时维持量产稳定性。

这也是芯测科技的核心优势所在。芯测科技长期深耕 Memory Test、Repair 与 Diagnosis 技术，提供完整的 MBIST、MBISR 与内存测试算法平台，协助客户在设计时间就建立高覆盖率的 SRAM 测试策略。面对 DDI、AI 芯片与车用 SoC 中不断增加的 SRAM 容量需求，芯测科技能协助客户有效提升内存良率、降低量产风险，并加速产品导入市场。



AI 推论市场扩大， SRAM/Cache/Buffer 质量成为量产新门坎



当 AI 产业从大型模型训练走向推论与 Edge AI 部署，芯片设计的重点也开始改变。过去 AI 芯片竞争集中在算力规模与平行运算能力，如今真正影响系统效率的，往往是芯片内部数据能否被快速、稳定且低功耗地搬移与存取。这也使得 AI 推论架构中的核心数据流系统——SRAM、Cache 与 Buffer 的重要性快速提升。

其中，SRAM 主要负责高速数据暂存与实时运算支持。在 AI 推论过程中，大量模型权重、中间运算结果与启动数据都需要被反复读写，若完全依赖外部 DRAM，不仅延迟增加，功耗也会大幅上升。因此，许多 AI 加速器与 Edge AI 处理器都会内建大量 SRAM 作为本地内存，以维持高吞吐量与低延迟特性。

然而，当 SRAM 容量与数量持续增加，内存缺陷风险也同步上升。先进制程下常见的弱位、数据保持不稳、读取干扰与电阻性缺陷，可能在高频推论或长时间运作下逐渐放大，进一步影响 AI 模型稳定性与整体 DPPM 表现。

除了 SRAM 之外，Cache 架构也正快速成为 AI 芯片设计核心。AI 运算与传统 CPU 工作负载最大的差异之一，在于数据重复利用率极高。同一组模型权重与特征数据，可能在短时间内被大量运算单元重复存取。若 Cache 效率不足，系统便需要频繁回读外部内存，不仅拖慢推论速度，也会增加功耗与带宽压力。



因此，许多 AI SoC 开始导入多层式 Cache、共享 Cache 与分布式 Cache 架构，希望提升数据重复利用效率。但随着 Cache 结构日益复杂，数据一致性、时序变异与存取稳定性问题也更加明显。特别是在高温、低电压或长时间 AI 运作情境下，Cache 可靠度已逐渐成为影响 AI 系统稳定性的关键因素之一。

与此同时，Buffer 受到的重视也日趋显著。AI 推论过程涉及大量数据流调度，不同运算模块之间需要持续进行数据交换与同步，因此 Buffer 架构会直接影响整体数据流效率。从 input buffer、weight buffer 到 feature buffer，Buffer 除了是暂存空间，更负责平衡不同运算节点间的数据流速率。

若 Buffer 设计或质量不稳定，容易产生数据壅塞、时序错配或数据遗失问题，进而影响整体吞吐量。尤其在 Edge AI 与实时推论应用中，Buffer 延迟与稳定性更直接关系到系统反应速度与实时性。

当 AI 芯片中的 SRAM、Cache 与 Buffer 规模同步成长，内存测试与修复策略也必须进一步升级。传统固定式内存测试流程，已难以全面对应不同内存架构与 AI 工作负载行为。针对这类需求，芯测科技（iSTART-TEK）推出的 MART、UDA 与 TEC 技术，正提供更具弹性与客制化能力的 SRAM 测试架构。

MART（MBIST Algorithm Recommendation Tool）是一套 AI 驱动的分析系统，可根据应用类型、DPPM 目标、功耗、效能与面积限制，协助使用者快速筛选适合的 SRAM 测试算法，可有效简化算法选择流程与测试规划难度。

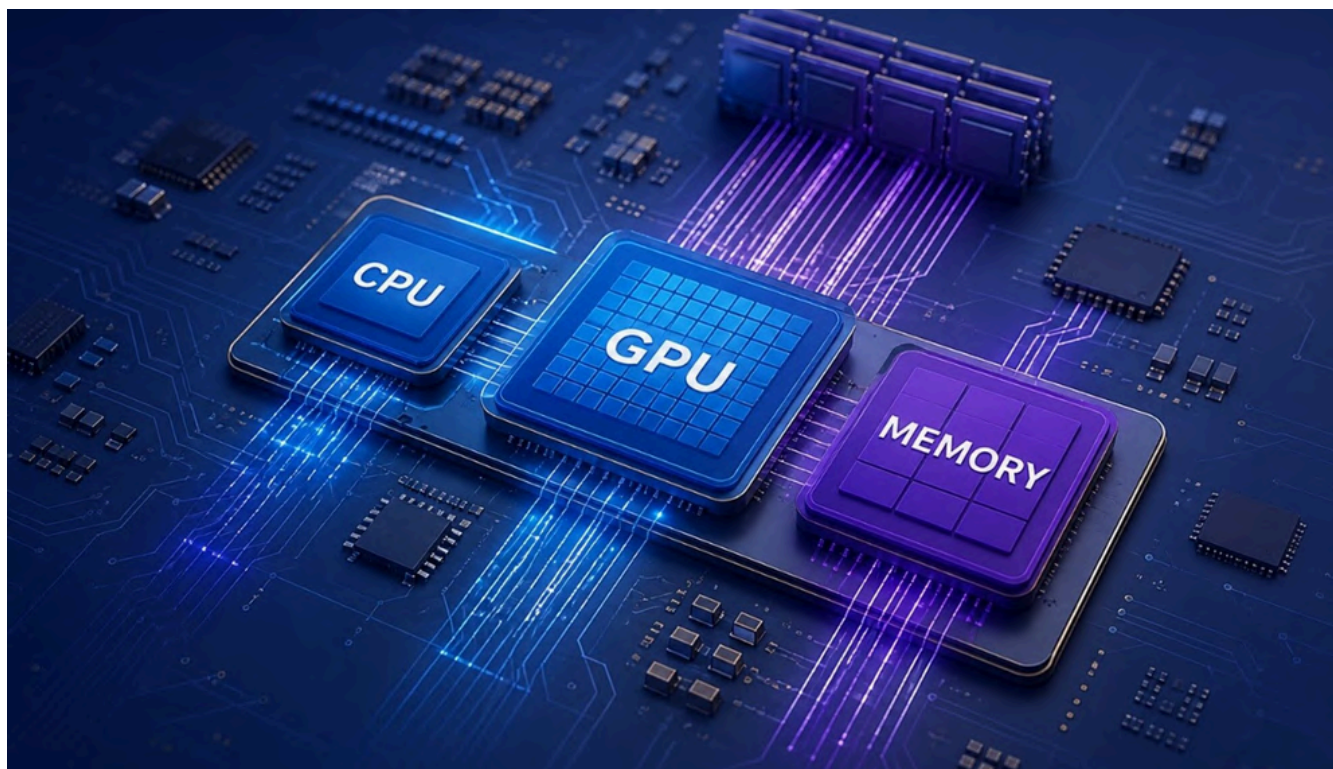
UDA（User-Defined Algorithms）则提供类似积木概念的平台架构，让使用者能自行定义基础测试元素，并组合成不同的内存测试算法。透过模块化方式，工程师可依不同 SRAM、Cache 与 Buffer 特性，建立更符合产品需求的测试流程，提升测试弹性与缺陷覆盖能力。

TEC（Testing Elements Change）可让测试工程师在 CP 与 FT 阶段，依据不同测试环境重新调整或组合 SRAM 测试算法。由于 SoC 测试常涉及极端电压与温度条件，不同环境可能对应不同内存缺陷型态，TEC 可协助工程师快速客制化测试元素，建立更符合特定需求的替代算法。

当 AI 产业开始进入大规模推论与长时间运作时代，SRAM、Cache 与 Buffer 已不再只是配角，而是决定 AI 芯片效能、功耗、良率与可靠度的核心基础。未来 AI 芯片的竞争，也将越来越取决于 Memory subsystem 本身的质量与测试能力。



GPU 之后的竞争焦点： AI 系统效能正在重新定义



生成式 AI 与大型语言模型快速发展，使「算力」长期被视为产业核心，而市场也持续聚焦在 GPU。然而，当 AI 应用从模型训练走向大规模推论、从云端延伸至边缘装置、并进一步演进为 AI Agent 与多模型协作架构时，系统瓶颈正逐步转移。AI 产业正从 GPU-centric 架构，走向 CPU+GPU+内存协同运算的新阶段。

在推论与 AI Agent 的情境下，系统复杂度显著提升。CPU 负责任务排程、逻辑控制与数据前处理，GPU 承担模型推论核心运算，而内存则成为数据与模型状态流动的关键节点。AI Agent 需要持续呼叫工具、维持短期与长期记忆、并进行多步骤推理，使整体运作更接近「系统工程」。

在这样的架构中，数据搬移成本快速上升，逐渐成为效能瓶颈。模型权重与中间特征需频繁在 CPU、GPU 与各类加速器间流动，而推论结果也需实时回写与整合。当模型规模与 agent 任务复杂度同步提升时，内存带宽、延迟与快取效率，成为决定系统吞吐与用户体验的关键因素。

因此，内存正从支持角色转变为核心关键。无论是 HBM、LPDDR，或是处理器内部的 SRAM 与 Cache，都构成 AI 系统效能的底层基础。特别是在 AI SoC 中，SRAM 不仅负责高速暂存，更直接影响延迟、能耗与整体运算效率，使内存设计逐渐成为与运算单元同等重要的竞争维度。



然而，随着内存在芯片中的比重提升，其可靠度问题也同步放大。制程变异、老化效应与工作负载压力，都可能导致 SRAM fault 或隐性错误累积，使得「可测试性」与「可修复性」成为 AI 芯片量产与长期运行的关键能力；这也是芯测科技长期专注地技术领域。

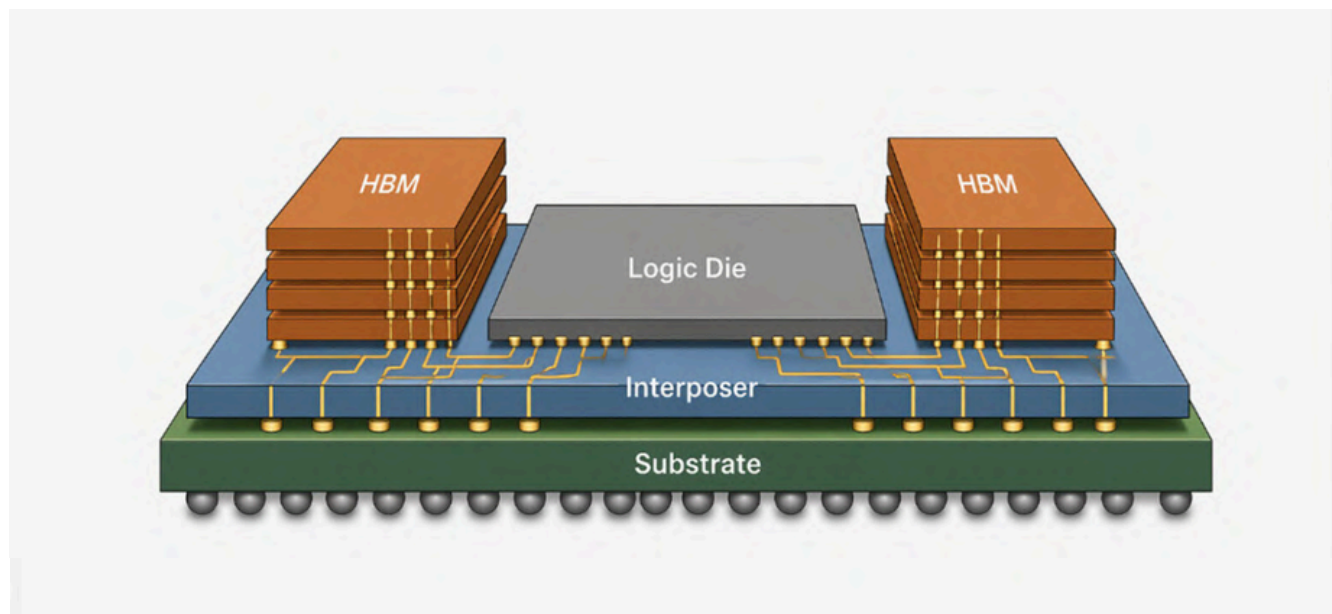
芯测科技聚焦于 AI 处理器内部存储器的完整解决方案，透过 MBIST 与 MBISR 架构，结合 UDA（User-Defined Algorithms）提升故障诊断效率与覆盖率，使内存问题能及早被精准定位并修复，降低后段风险与 field failure 机率。

同时，芯测科技的 MART（MBIST Algorithm Recommendation Tool）进一步将内存测试与修复流程智能化，可依不同内存型态、缺陷特征与应用场景，自动推荐最佳测试与修复策略，提升测试效率并缩短开发周期。在 AI 芯片高度客制化与快速迭代的趋势下，这类工具化能力正成为提升量产效率的重要基础。

整体而言，AI 基础建设的竞争已不再只是 GPU 算力的比拚，而是转向 CPU、GPU 与内存协同效率的系统级竞争。当运算架构走向高度整合与异质化，内存的稳定性与可修复性将直接影响 AI 系统的规模化能力与可靠度。在这样的产业转折点上，内存测试与修复技术，已成为支撑 AI 时代持续扩张的重要基础。



CoWoS 与 HBM： AI 时代内存产业的关键引擎



近年 AI 运算需求爆发，大型语言模型、生成式 AI 与高效能运算（HPC）工作负载，需要在短时间内存取海量权重与中间数据，使内存带宽与容量成为系统效能的核心指针。也因此，「数据搬运」逐渐成为半导体产业的瓶颈。在这样的背景下，台积电所研发的 CoWoS（Chip-on-Wafer-on-Substrate）先进封装技术与 HBM（High Bandwidth Memory）形成高度共生关系，重新定义 AI 服务器与高阶运算平台的架构。

首先，CoWoS 的最大贡献在于打破传统内存带宽瓶颈。过去 CPU 或 GPU 与外部 DRAM 之间依赖封装外的电路板连接，讯号传输距离长、功耗高且带宽受限。CoWoS 透过硅中介层（interposer），让逻辑芯片与多颗 HBM 在单一封装内紧密整合，将数据传输距离缩短到毫米等级，大幅提升 I/O 密度与传输效率。这种设计让 AI 加速器能以极高带宽存取内存，成为现代 AI GPU 与加速器的标准配置。随着大型模型对 KV Cache（键值快取）的需求急遽增加，高带宽与大容量内存已从「加分项」变成「必要条件」。

其次，CoWoS 推动了 HBM 技术世代的快速演进。AI 模型参数量持续成长，使单一封装内的内存容量需求不断攀升，从 HBM2、HBM2e 到 HBM3、HBM3e，堆栈层数与带宽持续提高。未来 AI 芯片甚至可能整合更多堆栈颗粒，以支持更大模型与更复杂的推论任务。换言之，先进封装能力已成为决定内存规格升级速度的重要因素。



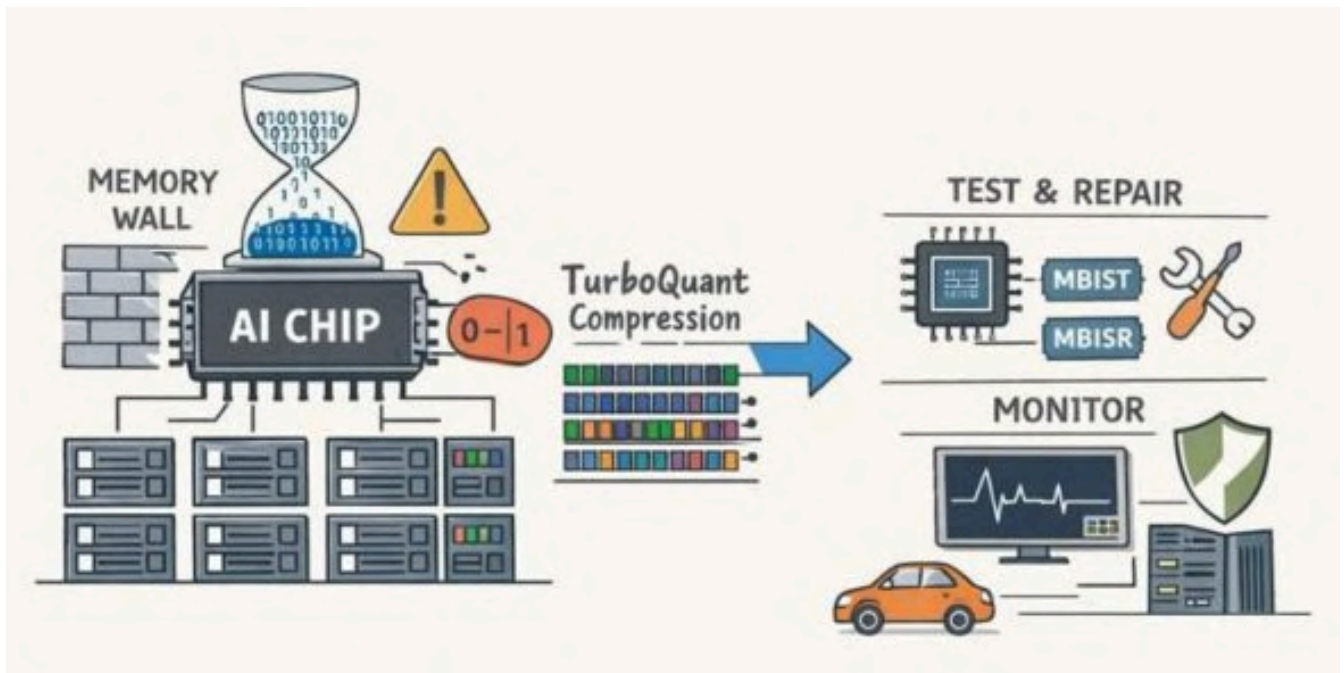
这股需求也改变了整体内存市场的供需结构。由于 HBM 属于高毛利产品，DRAM 制造商将大量产能转向 HBM 生产，导致传统 PC 与消费性电子所使用的 DRAM 供给相对收缩。另一方面，CoWoS 产能本身具有高度技术门槛与设备限制，AI 芯片的出货量受制于先进封装产能，使 HBM 供不应求的情况可能延续至 2027-2028 年。即使业界持续探索降低内存需求的技术，例如优化快取机制或新型数据压缩方法，对于顶级 AI 运算而言，高带宽内存的地位仍难以被取代。

展望未来，封装技术仍将持续向更高整合度演进。继 CoWoS-S、CoWoS-R 与 CoWoS-L 之后，混合键合（Hybrid Bonding）被视为下一波关键技术，透过更细微的互连间距与更高的 I/O 密度，使内存与逻辑芯片几乎达到「芯片级」的连接。这不仅能进一步降低功耗，也能显著提升整体系统效能，为下一代 AI 运算铺路。

在这样的产业趋势下，HBM 不仅是高效能运算的核心组件，也成为良率与可靠度挑战最严峻的内存类型之一。堆栈式结构与先进封装使测试与修复难度大幅提高，如何在出货前确保内存质量与长期可靠度，成为产业竞争的关键。芯测科技长期投入内存测试与修复技术，并拥有 ISO 26262 TCL1 认证的车规级测试算法与修复方案，协助芯片在先进封装与高带宽架构下维持高良率与高可靠度，成为 AI 时代内存质量把关的重要支撑力量。



从 TurboQuant 看 AI 芯片设计的新隐忧



在 AI 模型参数呈指数级增长的今天，「内存墙（Memory Wall）」已成为所有芯片设计者的共同敌人。近期某国际云端巨头发表的 TurboQuant 技术，无疑在产业内投下了一颗震撼弹。这项技术揭示了未来五年 AI 芯片发展的核心趋势：在有限的空间与带宽内，如何榨取极致的数据价值。

数据密度越高容错空间越小

TurboQuant 的核心任务在于优化大型语言模型（LLM）推论时的 KV Cache。它可将数据进行高达 6 倍的压缩，这意味着原本需要六台服务器才能跑的模式，现在可能只需要一台。

然而当我们沉浸于算法带来的「软件红利」时，半导体产业链的另一端却面临着全新的挑战。

这是一个物理学上的必然：当数据被极度压缩，每一个位所承载的信息权重就越高。在未压缩的数据中，一个位的翻转（Bit Flip）可能只是一次微小的噪声；但在 TurboQuant 这种高倍率量化压缩后，关键位的错误可能会导致整个推论结果的逻辑崩溃。

另外，为了配合此类压缩技术，AI 芯片内部整合了更密集的 SRAM 与高效能内存架构。随着制程微缩至 3 奈米甚至更先进节点，内存单元的良率与长期稳定性，已成为决定 AI 芯片能否量产并商用的关键。



从「好用」到「耐用」：不可或缺的底层守护

当产业焦点都在关注 AI 如何减少内存用量时，芯片设计者的下一步则是思考：这些被极度压榨的内存，如何稳住表现、不要出差错？这正是芯测科技（iSTART-TEK）在生态系中所扮演的角色。

在这种高效能运算（HPC）与 AI 芯片的演进路径上，芯测科技提供的内存测试与修复（DFT/MBIST/BISR）方案，实际上是在背后支撑TurboQuant 这类软件红利的的安全基座。

- 1.测试（Test）：芯测透过 START™ v5 等先进工具，在芯片设计时间便植入强大的 MBIST（内存自我测试）功能，精准筛选出缺陷。
- 2.修复（Repair）：藉由 MBISR（内存自我修复）技术，在发现位错误时自动启用备用空间进行修复，将原本可能报废的芯片化腐朽为神奇，直接提升产量良率。
- 3.监控（Monitor & Security）：在芯片运作的生命周期中，透过监控方案实时掌握内存健康状况，这在追求高度可靠性的 HPC 与车用领域尤为重要。

「软硬结合」才是 AI 真正的未来

TurboQuant 的发表，让我们看到了 AI 算法突破瓶颈的决心。但这场内存革命的下半场，必然会回归到硬件底层的可靠度竞争。从解决「如何节省空间」，到芯测科技解决「如何确保质量」，唯有内存的绝对稳定，软件端的压缩技术才能真正发挥实力。



Episode 51: 从测试到智慧防护：内存质量管控工具的全面进化

[Watch the video](#)

Episode 50: 打造可客制化内存测试方法 灵活应变快速演进的半导体时代

[Watch the video](#)

Episode 49: AI 家电与智能显示时代：从语音控制到 CIM 的芯片测试挑战

[Watch the video](#)

Episode 48: 智慧座舱 (In-vehicle AI)：汽车电子的下一个算力战场

[Watch the video](#)

Episode 47: 当 MRAM 遇上 AI：下一代智能芯片的关键内存革命

[Watch the video](#)

Episode 46: NPU 运作需求下的 SRAM 重要性与测试策略

[Watch the video](#)

Episode 45: LPU 使用 SRAM 的原因与测试关键

[Watch the video](#)