iSTART

July 2026

iSTART iReport

Issue No.16

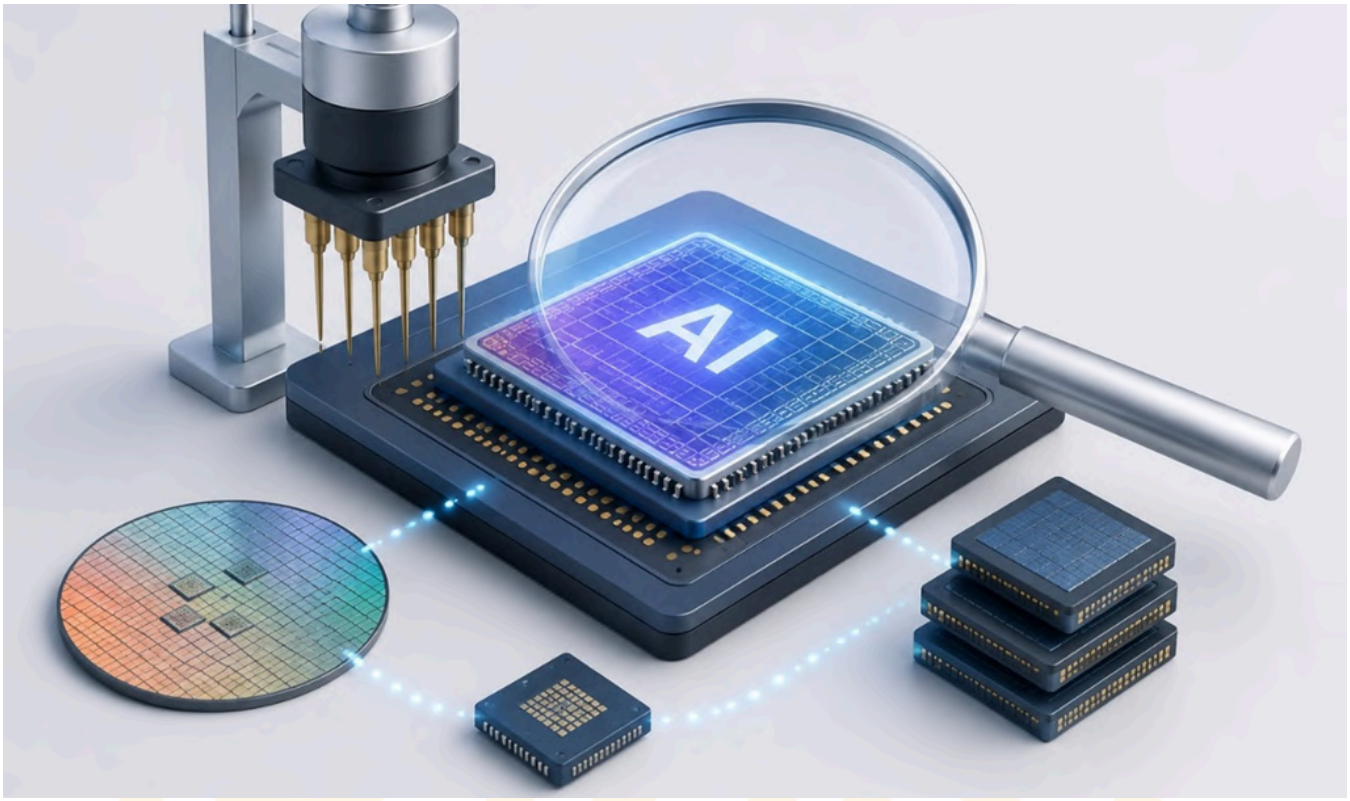
Contents

News

IC Tech Insights

iSTART Class

AI Chip Yield Becomes a Key Competitive Factor — iSTART-TEK Captures New Opportunities with Memory Quality Management Technology



Competition in AI chips is shifting from a race for computing power to a race for manufacturing yield. As high-performance computing chips become increasingly complex in design, the value of each individual chip continues to rise. Even minor defects can lead to significant losses. Maintaining stable quality and high yield during mass production has therefore become a critical challenge for chipmakers.

Among the many factors affecting chip yield, memory reliability is one of the most crucial. AI chips rely heavily on SRAM for cache, data buffering, and temporary storage during computation, directly impacting data access efficiency, latency, and power consumption. As memory capacity continues to grow, the importance of SRAM testing and repair is rising accordingly.

However, SRAM defects in advanced process nodes are no longer limited to traditional open or short circuit failures. Issues such as write failures, data retention failures, dynamic faults, and marginal failures caused by process variation are becoming increasingly common. These defects are not only harder to detect, but also more likely to become hidden risks that affect AI chip yield.

iSTART-TEK has long specialized in memory testing, repair, and quality management technologies. Through its core platform, START v5, the company helps customers establish comprehensive memory quality protection mechanisms at the chip design stage.

START v5 provides a complete Memory Built-In Self-Test (MBIST) and Memory Built-In Self-Repair (MBISR) architecture. Through built-in testing mechanisms, the system can quickly identify faulty memory cells, improving defect detection capability while reducing the risk of test escapes. Combined with built-in repair technology, the system can further activate redundant rows or columns to automatically repair damaged memory blocks, allowing dies that might otherwise be discarded to function normally again and improving overall yield.

For high-value AI chips, the benefits of this capability are especially significant. Even a small percentage increase in yield can translate into substantial cost savings and higher production output.

In addition to core testing and repair capabilities, iSTART-TEK's User-Defined Algorithms (UDA) feature provides high flexibility, enabling customers to build customized testing flows based on different memory architectures, process characteristics, and application requirements. Since memory designs in AI chips are highly customized, fixed testing algorithms are no longer sufficient for all scenarios. More flexible testing capabilities have therefore become a major competitive advantage.

In terms of testing efficiency, iSTART-TEK's test evaluation mechanism helps analyze defect coverage and test time across different algorithms, allowing customers to identify the optimal testing combination and achieve a balance between test quality and test cost.

In addition, iSTART-TEK's intelligent memory algorithm recommendation technology, MART (MBIST Algorithm Recommendation Tool), can automatically recommend suitable test algorithm combinations based on memory type, process node, and historical failure characteristics. This helps engineering teams shorten development time and accelerate product deployment.

As AI chip costs continue to rise, the core of market competition is no longer defined solely by process leadership. Instead, success increasingly depends on how effectively companies can control defects, improve yield, and ensure stable product quality. Through START v5, memory testing and repair technologies, customized algorithms, and intelligent testing solutions, iSTART-TEK continues to help customers enhance product reliability and mass production efficiency, becoming an important technology partner in the advancement of the AI semiconductor industry.

Jensen Huang Highlights the Arrival of the AI Factory Era; iSTART-TEK Expands High-Reliability Chip Testing Opportunities

NVIDIA Founder and CEO Jensen Huang stated during his keynote at NVIDIA GTC Taipei 2026 that the AI industry has officially entered a new phase, shifting from “Can it work?” to “Can it generate profit and scale effectively?” He defined the AI Factory as the next generation of digital infrastructure. As Agentic AI, Physical AI, AI-native PCs, and edge computing continue to advance, semiconductor systems are facing significantly higher requirements for reliability, availability, and long-term operational stability.

Huang also introduced the concept of “Compute is Revenue,” emphasizing that future AI data centers will no longer be merely computing facilities but factories that produce AI tokens. When computing capacity directly translates into revenue generation, chip failures, system downtime, or insufficient reliability become business issues that can directly impact profitability.

Against this backdrop, iSTART-TEK, a company specializing in memory testing, repair, and Design-for-Test (DFT) technologies, is well positioned to benefit from the ongoing evolution of AI infrastructure.

In recent years, AI chips, HPC processors, and automotive SoCs have continued to increase their SRAM capacity. In advanced process nodes, SRAM can account for more than half of a chip's total die area. As AI factories expand in scale, the impact of memory defects on system stability and computing availability becomes increasingly significant. iSTART-TEK's START™ v5 platform, SRAM BIST/BISR solutions, and MART (MBIST Algorithm Recommendation Tool) help chip developers improve test coverage, enhance repair capabilities, and reduce potential risks during mass production.

Huang also highlighted Agentic AI as the next generation of computing, where AI systems will be capable of autonomously planning tasks, utilizing tools, and executing workflows over extended periods. NVIDIA's collaboration with Cadence on semiconductor design Super Agents further signals the gradual adoption of AI-driven workflows across RTL generation, verification, debugging, and design processes.

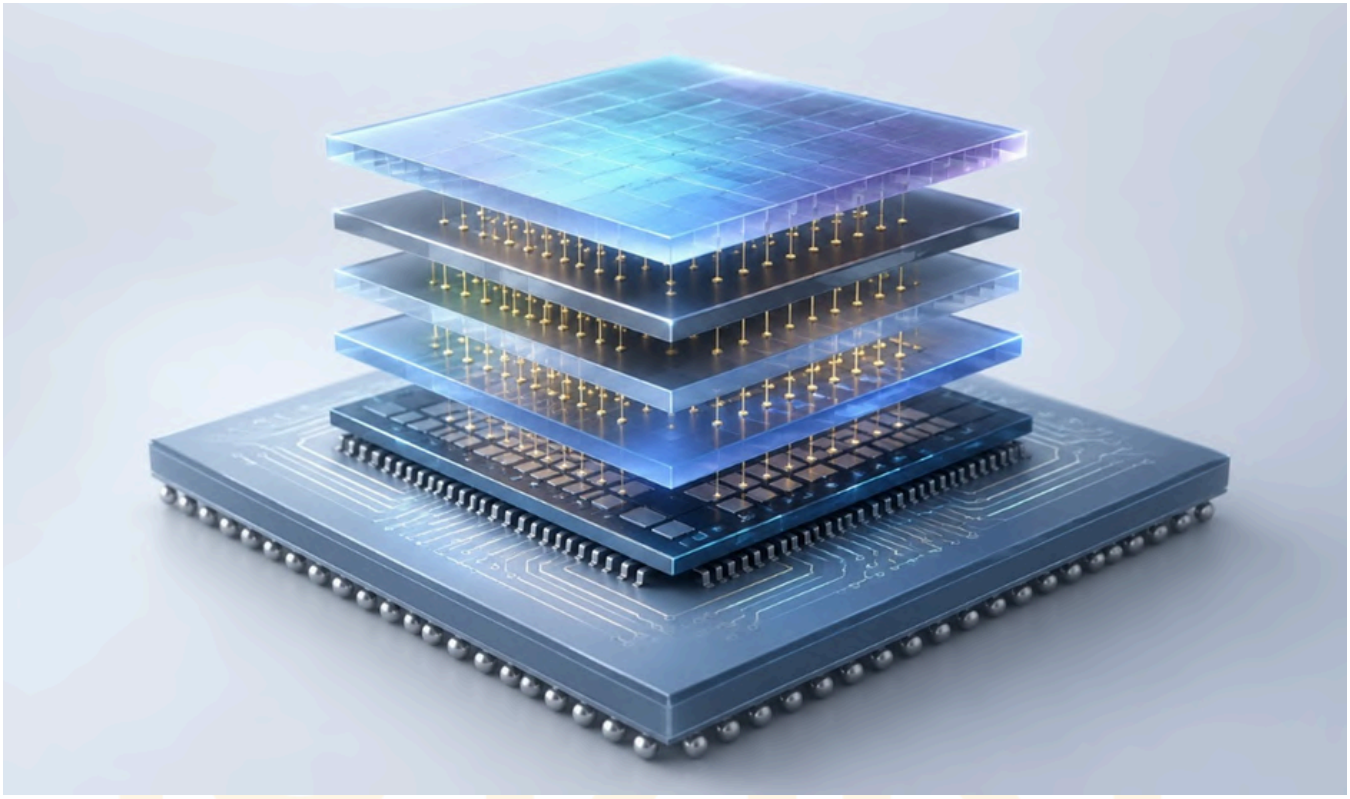
According to iSTART-TEK, the vast amount of data generated by DFT, BIST, BISR, repair analysis, and diagnosis technologies could become valuable foundational data for AI-assisted design and AI-assisted debugging. By combining test data with AI analytics, semiconductor companies can not only improve development efficiency but also further optimize product quality and manufacturing yield.

Meanwhile, the rise of Physical AI, autonomous driving, robotics, and edge AI applications is driving increasing demand for functional safety, long-term operational stability, and high-reliability validation. Huang noted that future automotive AI and robotic systems will require more comprehensive testing and verification processes, making DFT, BIST, BISR, and diagnostic capabilities increasingly important.

iSTART-TEK believes that as the AI industry shifts its focus from raw computing performance toward system reliability, availability, and long-term operational stability, testing technologies will evolve from being viewed as cost centers to becoming key sources of competitive advantage. The company plans to continue advancing its memory testing algorithms, repair analysis technologies, and advanced DFT solutions to help customers develop highly reliable semiconductor products for the AI Factory, Edge AI, and Physical AI era.

In the AI infrastructure vision outlined by Jensen Huang, semiconductor chips are undoubtedly the core assets that generate revenue. Equally important are the testing and repair technologies that ensure those chips operate reliably, making them an indispensable foundation of the AI era.

Huawei's "Tao Law" Sparks a New Advanced Packaging Race — iSTART-TEK Has Already Positioned Itself with IEEE 1838



As AI computing demand continues to surge, the semiconductor industry is gradually shifting from a focus on transistor scaling to a new era of system-level stacking and integration. Recently, the industry has been paying close attention to Huawei's proposed "Tao Law" concept, which emphasizes the use of advanced packaging, chiplets, and multi-die stacking technologies to overcome the limitations of traditional Moore's Law in process scaling. This signals that the future competitiveness of AI chips will extend beyond transistor density on a single die and increasingly depend on how effectively 3D-IC and heterogeneous integration architectures can improve overall system performance, bandwidth, and power efficiency.

However, as AI chips increasingly adopt vertical stacking architectures involving HBM, high-performance logic dies, and chiplets, testing and repairability have emerged as critical challenges in the advanced packaging era. Given the high value of multi-die packages, discovering memory or interconnect defects after packaging can result in substantial scrap costs. As a result, maintaining testability and repairability within 3D-IC architectures has become an essential issue across the AI semiconductor supply chain.

iSTART-TEK, a company specializing in memory testing and repair technologies, notes that as the industry transitions from Moore's Law toward stacking-based integration, IEEE 1838 is becoming an indispensable testing standard for 3D-ICs and heterogeneous integration devices. The company has already completed the integration of IEEE 1838 into its memory testing and repair architecture, making it one of the few EDA solution providers capable of delivering comprehensive IEEE 1838 support for memory testing and repair applications.

IEEE 1838 is a testing standard specifically developed for 3D-IC and multi-die stacked architectures. It addresses key challenges such as die-to-die test access, post-stack test path management, and testability within heterogeneous integration environments. While traditional SoC testing primarily targets a single die, the widespread adoption of stacked architectures involving HBM, logic dies, AI accelerators, and chiplets has significantly increased the complexity of test access and fault localization, making IEEE 1838 an increasingly important foundation for advanced packaging technologies.

According to iSTART-TEK, as the industry uses stacking technologies to challenge the limits of Moore's Law, testing infrastructures must evolve accordingly. In advanced packaging architectures such as CoWoS, 3D-IC, and HBM, the absence of standardized testing mechanisms makes it difficult to efficiently manage multi-die test access and fault diagnosis. Consequently, IEEE 1838 is expected to become a fundamental requirement for future multi-die and heterogeneous integration designs.

Long before 3D-IC technology entered mainstream commercialization, iSTART-TEK had already invested in IEEE 1838-related development and integrated it into its memory testing and repair platform, achieving deep integration with MBIST and BISR architectures. Through the IEEE 1838 framework, the system can preserve SRAM and HBM testing and repair capabilities even after multi-layer stacking is completed, including test path planning, fault localization, repair control, and test access management.

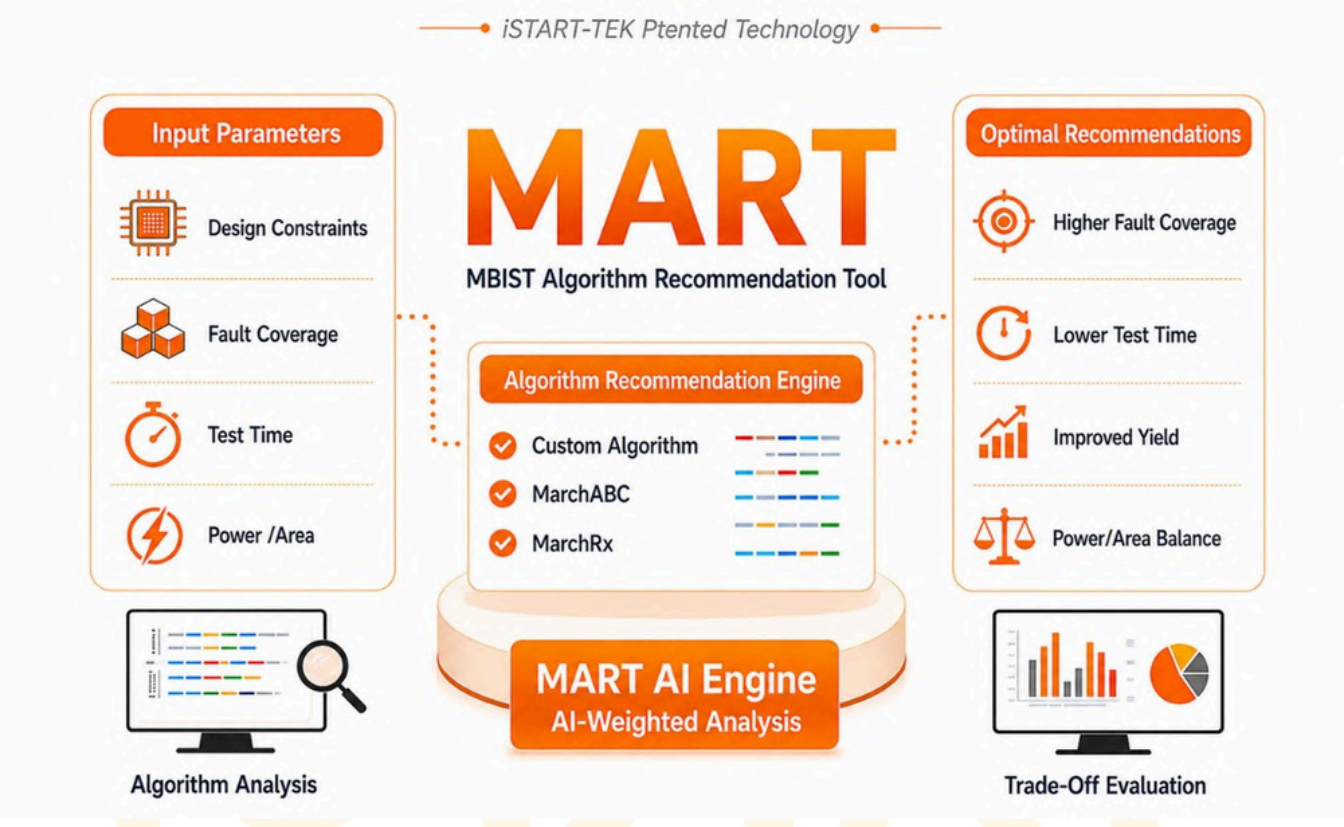
To increase bandwidth and reduce latency, vertical stacking of HBM with GPUs, NPUs, and AI accelerators is rapidly becoming the industry norm, further elevating the importance of memory testing. iSTART-TEK's IEEE 1838 solution is specifically designed for such heterogeneous integration and high-density stacking scenarios. It supports multi-die test access and memory repair workflows, helping customers maintain high yield and reliability in advanced packaging environments.

In addition, AI devices integrating HBM and high-performance logic dies through CoWoS and 3D-IC architectures represent a significant manufacturing and packaging investment. If memory defects are discovered after packaging and no effective repair mechanism is available, the entire module may need to be discarded. By combining IEEE 1838 with advanced memory repair technologies, iSTART-TEK enables customers to perform precise fault diagnosis and repair even after die stacking has been completed, significantly reducing advanced packaging scrap risks while improving final yield and return on investment (ROI).

Industry observers expect IEEE 1838 to become an increasingly important testing standard throughout the AI semiconductor ecosystem as HBM, chiplets, CoWoS, and heterogeneous integration architectures continue to expand. EDA vendors capable of providing comprehensive memory testing and repair capabilities are expected to play a critical role in the next phase of AI semiconductor competition.

iSTART-TEK emphasizes that it already offers a complete IEEE 1838 memory testing and repair solution, enabling customers to address the testing and yield challenges associated with AI chiplets, HBM, and 3D-IC technologies while preparing for the next generation of advanced packaging opportunities.

Using AI to Solve AI Challenges! AI Computing Drives Memory Complexity Growth, Making Smarter MBIST Algorithms Essential



The rapid rise of generative AI and AI agents is shifting enterprise investment focus beyond standalone accelerators toward a broader range of AI processors and server CPUs. A common characteristic of these chips is the integration of large amounts of SRAM and cache to support data-intensive computing. As memory capacity and architecture continue to scale, the complexity of testing and repair also increases accordingly. The selection of MBIST algorithms is gradually becoming a critical factor that directly impacts production yield, DPPM, and development schedules.

With the growing adoption of automotive AI, ADAS, intelligent cockpit systems, and Edge AI devices, chip quality requirements continue to rise. Products now demand not only higher test coverage, but also a careful balance among test time, power consumption, silicon area, and mass production schedules. When planning memory test strategies, engineering teams often face highly complex scenarios involving multiple interdependent constraints. Traditional algorithm selection methods based primarily on lookup tables and engineering experience are increasingly becoming hidden bottlenecks in the design and testing flow. In this context, intelligent algorithm recommendation tools can provide an effective way to overcome testing efficiency challenges.

The MBIST Algorithm Recommendation Tool (MART), developed by iSTART-TEK, has evolved into an AI Test Algorithm Copilot, transforming the traditionally experience-driven MBIST algorithm selection process into an AI-assisted decision-making workflow. The system can integrate critical parameters including memory type, memory size, port architecture, target DPPM range, failure information, fail bitmap, repair result, and test time budget to proactively recommend the most suitable March test combinations, UDA test elements, and diagnostic flows, enabling engineering teams to rapidly establish comprehensive testing strategies.

The core value of this capability lies in upgrading the traditional “condition matching” approach into a multi-objective optimized recommendation mechanism. Through AI-based weighting methodologies, MART can balance product quality, test time, and design cost simultaneously, allowing MBIST strategies to better align with real-world product requirements. Engineers no longer need to repeatedly search through documentation or manually compare testing conditions. Instead, they can quickly converge on optimal testing directions during the early development stage, significantly reducing decision-making overhead and shortening development cycles.

As AI chips and automotive-grade semiconductor markets continue to pursue ultra-low DPPM and higher reliability, test strategy optimization has evolved from a pure engineering detail into a key competitive advantage. The ability to identify the optimal testing combination more quickly can directly affect both product time-to-market and mass production quality. Facing the ongoing increase in memory scale and architectural complexity, MBIST algorithm selection is transitioning from an experience-driven methodology toward a data-driven and AI-assisted approach. The emergence of MART symbolizes the official transition of memory testing into a new era of AI-assisted decision-making, providing chip design teams with a more efficient and higher-quality decision support platform.

iSTART-TEK's "Memory Test Algorithm Recommendation System" Secures Utility Model Patent, Redefining SoC Test Development Workflow

Memory test and repair leader iSTART-TEK today announced that its self-developed "Memory Test Algorithm Recommendation System" has officially obtained a utility model patent in Taiwan. This system is implemented in iSTART-TEK's MBIST Algorithm Recommendation Tool (MART), launched in 2026, and leverages AI-assisted automated decision-making to significantly enhance memory test development efficiency.

As semiconductor processes scale down to the nanometer level, memory capacity and complexity within System-on-Chips (SoCs) are growing exponentially, making it a major challenge for designers to identify the optimal algorithm among hundreds or thousands of test options. Traditionally, memory test algorithm selection has relied heavily on the experience of senior engineers. Incorrect choices can not only miss potential defects but also prolong test times and increase production costs. The patented system from iSTART-TEK addresses this long-standing industry inefficiency.

The system first employs a "Question Generation Module" and an "Interface Module" to sequentially generate inquiries and receive corresponding responses, accurately capturing key parameters of memory architecture and application scenarios. Based on these parameters, the "Algorithm Recommendation Module" determines the most suitable test algorithm, while the "Configuration File Generation Module" automatically produces the corresponding configuration information.

Using this tool, engineers can quickly obtain optimized test algorithms and related configuration files, effectively shortening development cycles and improving memory testing efficiency.

iSTART-TEK stated that securing this patent demonstrates the company's capability not only to provide powerful testing tools but also to help customers make accurate test decisions quickly. Through parameterized guidance, even less-experienced engineers can configure professional-grade memory test solutions in a very short time. The core value of this technology lies in transforming complex test experience into digital logic decisions, optimizing SoC test coverage and production efficiency from the outset.

Additional Highlights

iSTART-TEK to Launch a Comprehensive DFT Solution to Enhance Testability and Mass Production Quality of Advanced Chips

Read more

DDI Functional Upgrades Drive SRAM Capacity Growth, Increasing the Importance of MBIST and MBISR

Read more

---

iSTART-TEK Memory Test and Repair Technologies Enable Mass Production Across Diverse Chip Applications, Surpassing 300 Million Units Shipped

Read more

Rising External Memory Prices: Is DDI Architecture Changing? The Value of Embedded SRAM Is Re-emerging

Recent fluctuations in memory pricing have once again become a major focus in the semiconductor industry. As DRAM and specialty memory prices continue to rise, many system designs that rely on external memory are facing new cost pressures. For the Display Driver IC (DDI) industry, the impact extends beyond procurement costs—it also affects architectural decisions for next-generation products.

In small-display applications such as smartphone OLED panels, wearable devices, and compact automotive displays, DDIs typically integrate display buffers directly on-chip, using embedded SRAM as the frame buffer. This architecture offers several advantages, including low latency, low power consumption, simplified packaging, and support for thinner and lighter devices.

However, the situation differs for large displays such as TVs, high-end monitors, and large automotive central control screens. As resolutions increase to 4K and 8K, image data volume grows dramatically. If these systems rely entirely on embedded memory, the required SRAM area may expand significantly, resulting in larger die sizes, higher costs, and reduced yield. As a result, DDIs for these applications have traditionally adopted external DRAM or SDRAM to achieve greater memory capacity at a lower cost per bit.

However, market conditions are changing. The primary advantage of external memory has always been its low capacity cost, but its weaknesses are becoming increasingly apparent. The first issue is supply chain risk. External memory pricing is highly sensitive to economic cycles, capacity adjustments, and supply-demand imbalances, making BOM costs difficult to predict. The second issue is system power consumption and latency. Frequent data transfers between the DDI and external memory not only increase I/O power consumption but also create bandwidth bottlenecks.

In contrast, although embedded SRAM occupies more die area, it delivers extremely high bandwidth and ultra-low latency. As demand for higher refresh rates and better display quality continues to rise, its technical value is becoming increasingly significant. More importantly, when external memory prices rise while SRAM costs remain relatively stable, companies are likely to reassess the cost trade-offs of different memory architectures.

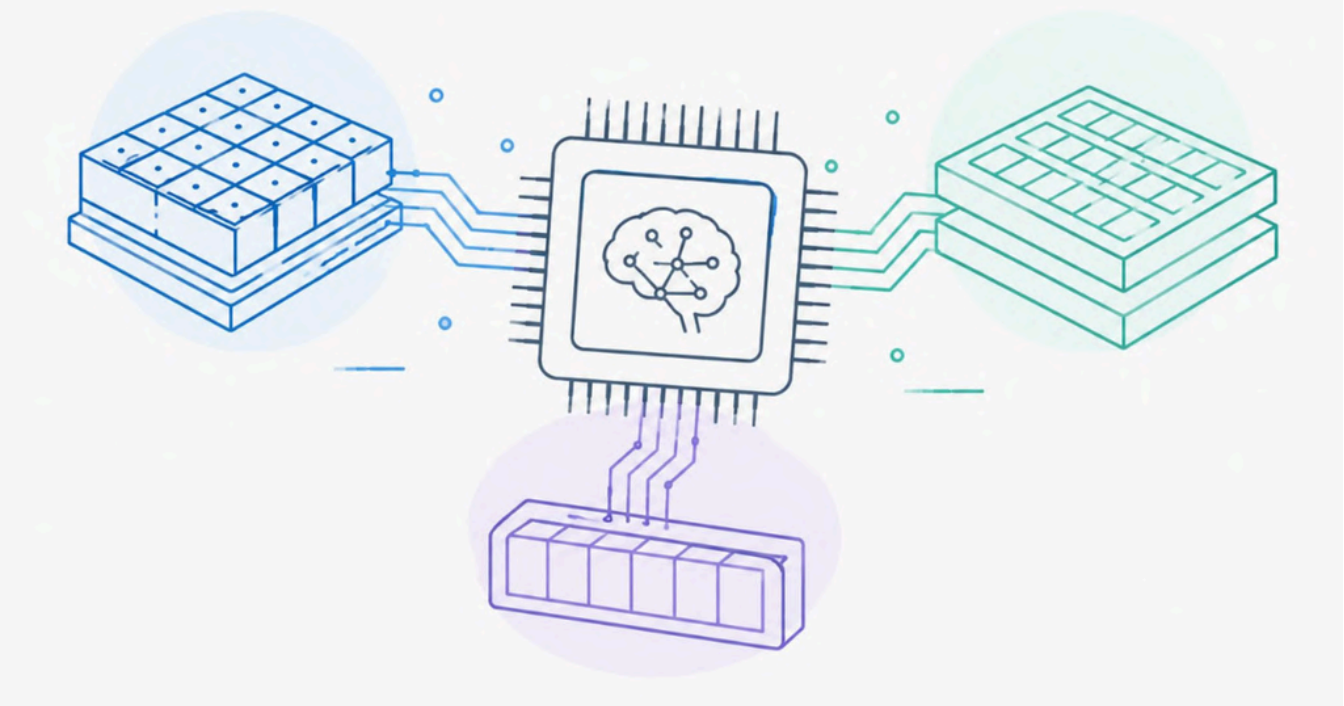
In the past, many companies considered embedded SRAM too expensive and therefore preferred external memory. But when external memory is no longer inexpensive—and may even introduce supply risks—increasing the embedded SRAM capacity within DDI may become a more strategic choice.

Of course, embedded SRAM also introduces new challenges. As SRAM capacity increases, the memory area ratio within the chip continues to grow. In advanced process nodes, SRAM already accounts for more than 50% of die area in many SoCs. The larger the memory, the higher the probability that process defects will create fail bits, directly impacting overall yield. Therefore, if DDI architectures are moving toward larger SRAM capacity, a critical factor is whether the backend flow includes sufficiently mature memory test and repair capabilities.

Through MBIST (Memory Built-In Self-Test) and Memory BISR (Memory Built-In Self-Repair) technologies, chips can identify defective SRAM regions during wafer testing and repair them using redundancy mechanisms, effectively improving overall yield. This enables companies to increase SRAM capacity without being constrained by yield concerns, allowing them to improve performance while maintaining manufacturing stability.

This is where iSTART-TEK demonstrates its core strength. With deep expertise in memory test, repair, and diagnosis technologies, iSTART-TEK provides comprehensive MBIST, MBISR, and memory test algorithm platforms to help customers establish high-coverage SRAM test strategies during the design stage. As SRAM capacity continues to grow in DDIs, AI chips, and automotive SoCs, iSTART-TEK helps customers improve memory yield, reduce production risk, and accelerate time-to-market.

AI Inference Market Expansion Makes SRAM, Cache, and Buffer Quality a New Production Threshold



As the AI industry shifts from large-scale model training toward inference and Edge AI deployment, the priorities in chip design are also changing. In the past, competition among AI chips was largely centered on compute scale and parallel processing capability. Today, however, what often determines system efficiency is whether data inside the chip can be moved and accessed quickly, reliably, and with low power consumption. This has significantly increased the importance of the core data-flow subsystems in AI inference architectures—SRAM, cache, and buffer.

Among them, SRAM is primarily responsible for high-speed data storage and real-time computational support. During AI inference, large volumes of model weights, intermediate computation results, and activation data must be repeatedly read and written. If the system relies entirely on external DRAM, latency increases significantly and power consumption rises accordingly. As a result, many AI accelerators and Edge AI processors integrate large amounts of SRAM as on-chip local memory to maintain high throughput and low-latency performance.

However, as SRAM capacity and quantity continue to increase, the risk of memory defects also rises. Common issues in advanced process nodes—such as weak bits, unstable data retention, read disturbance, and resistive defects—may gradually become more severe under high-frequency inference or prolonged operation, potentially impacting AI model stability and overall DPPM performance.

In addition to SRAM, cache architecture is rapidly becoming a core component of AI chip design. One of the biggest differences between AI workloads and traditional CPU workloads is the extremely high level of data reuse. The same set of model weights and feature data may be repeatedly accessed by numerous compute units within a short period of time. If cache efficiency is insufficient, the system must frequently fetch data from external memory, which not only slows inference speed but also increases power consumption and memory bandwidth pressure.

To address this, many AI SoCs are adopting multi-level cache, shared cache, and distributed cache architectures to improve data reuse efficiency. However, as cache structures become increasingly complex, issues related to data coherence, timing variation, and access stability also become more prominent. Especially under high-temperature, low-voltage, or long-duration AI operating conditions, cache reliability is gradually becoming one of the key factors affecting AI system stability.

At the same time, buffer is receiving increasing attention. AI inference involves intensive data-flow scheduling, requiring continuous data exchange and synchronization among different compute modules. As a result, buffer architecture directly affects overall data-flow efficiency. From input buffers and weight buffers to feature buffers, buffers serve not only as temporary storage but also as mechanisms for balancing data-flow rates between different compute nodes.

If buffer design or quality is unstable, issues such as data congestion, timing mismatch, or data loss may occur, ultimately reducing overall throughput. This is particularly critical in Edge AI and real-time inference applications, where buffer latency and stability directly impact system responsiveness and real-time performance.

As SRAM, cache, and buffer capacities continue to scale in AI chips, memory testing and repair strategies must also evolve. Traditional fixed memory testing flows are no longer sufficient to comprehensively address the diverse memory architectures and workload behaviors found in modern AI systems.

To meet these challenges, iSTART-TEK's MART, UDA, and TEC technologies provide a more flexible and customizable SRAM testing framework.

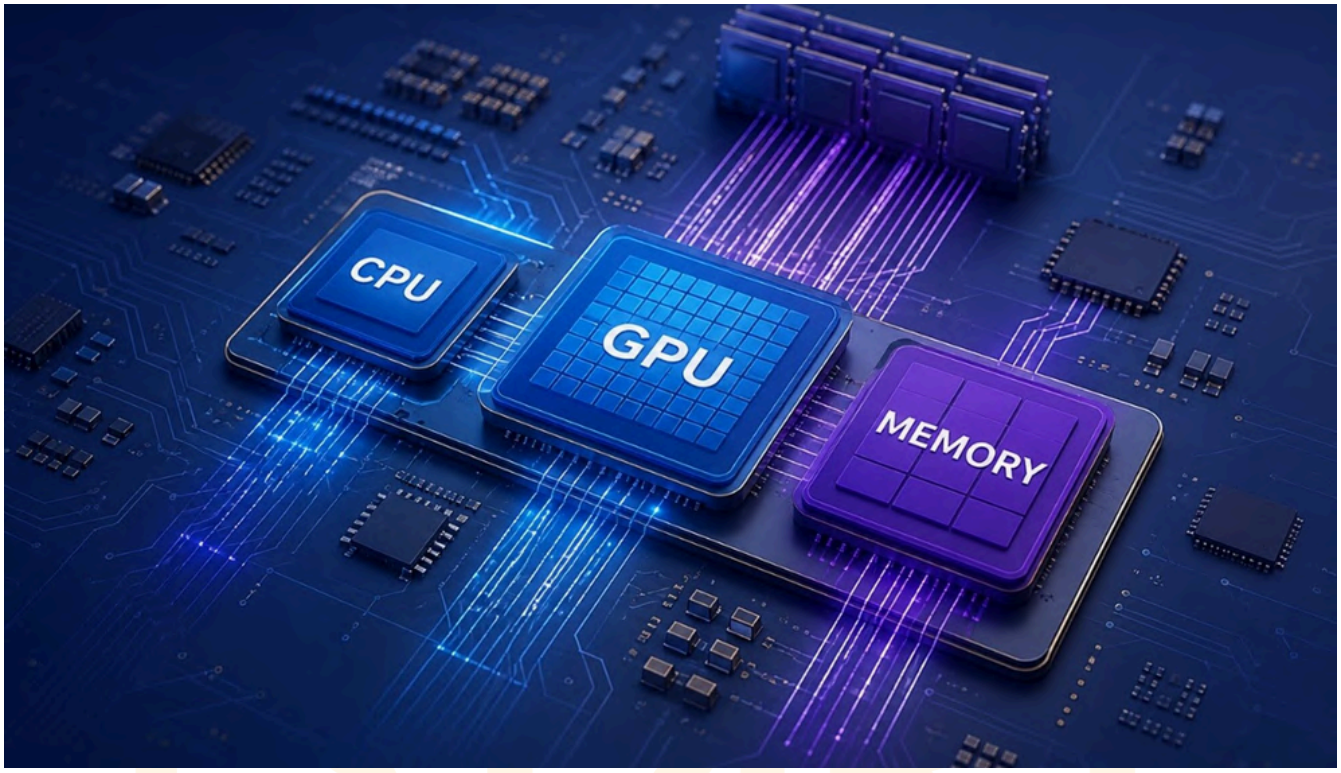
MART (MBIST Algorithm Recommendation Tool) is an AI-driven algorithm analysis system that helps users quickly identify suitable SRAM testing algorithms based on application type, DPPM targets, power consumption, performance requirements, and area constraints. It effectively simplifies algorithm selection and test planning.

UDA (User-Defined Algorithms) provides a modular, building-block-like platform that allows users to define fundamental testing elements and combine them into customized memory testing algorithms. Through this modular approach, engineers can develop testing flows tailored to the specific characteristics of different SRAM, cache, and buffer architectures, improving testing flexibility and defect coverage.

TEC (Testing Elements Change) enables test engineers to adjust or reassemble SRAM testing algorithms during CP and FT stages according to different testing environments. Since SoC testing often involves extreme voltage and temperature conditions, different environments may correspond to different types of memory defects. TEC helps engineers rapidly customize testing elements and build alternative algorithms optimized for specific requirements.

As the AI industry enters an era of large-scale inference and long-duration operation, SRAM, cache, and buffer are no longer secondary components. They have become fundamental elements that determine AI chip performance, power efficiency, yield, and reliability. Future competition in AI chips will increasingly depend on the quality and testability of the memory subsystem itself.

GPU After the Bottleneck Shift: AI System Performance Is Being Redefined



The rapid advancement of generative AI and large language models has long positioned “compute power” as the core of the industry, with market attention heavily centered on GPUs. However, as AI applications evolve from model training to large-scale inference, extend from cloud to edge devices, and further progress toward AI agents and multi-model collaboration architectures, system bottlenecks are gradually shifting. The AI industry is moving away from a GPU-centric paradigm toward a new stage of CPU + GPU + memory co-optimized computing.

In inference and AI agent scenarios, system complexity increases significantly. CPUs are responsible for task scheduling, logical control, and data preprocessing, GPUs handle core model inference computation, while memory becomes the critical hub for data and model state movement. AI agents continuously invoke tools, maintain short-term and long-term memory, and perform multi-step reasoning, making the overall system behave more like an integrated system engineering problem rather than pure acceleration.

Within this architecture, data movement cost is rapidly rising and is becoming a key performance bottleneck. Model weights and intermediate features must frequently travel between CPUs, GPUs, and various accelerators, while inference outputs must also be synchronized and integrated in real time. As model scale and agent task complexity increase simultaneously, memory bandwidth, latency, and cache efficiency become decisive factors for system throughput and user experience.

As a result, memory is shifting from a supporting role to a core component. Whether it is HBM, LPDDR, or on-chip SRAM and cache, all form the foundational layer of AI system performance. In particular, within AI SoCs, SRAM not only serves as high-speed storage but also directly impacts latency, power efficiency, and overall computational performance, making memory design as critical as compute units in system-level competition.

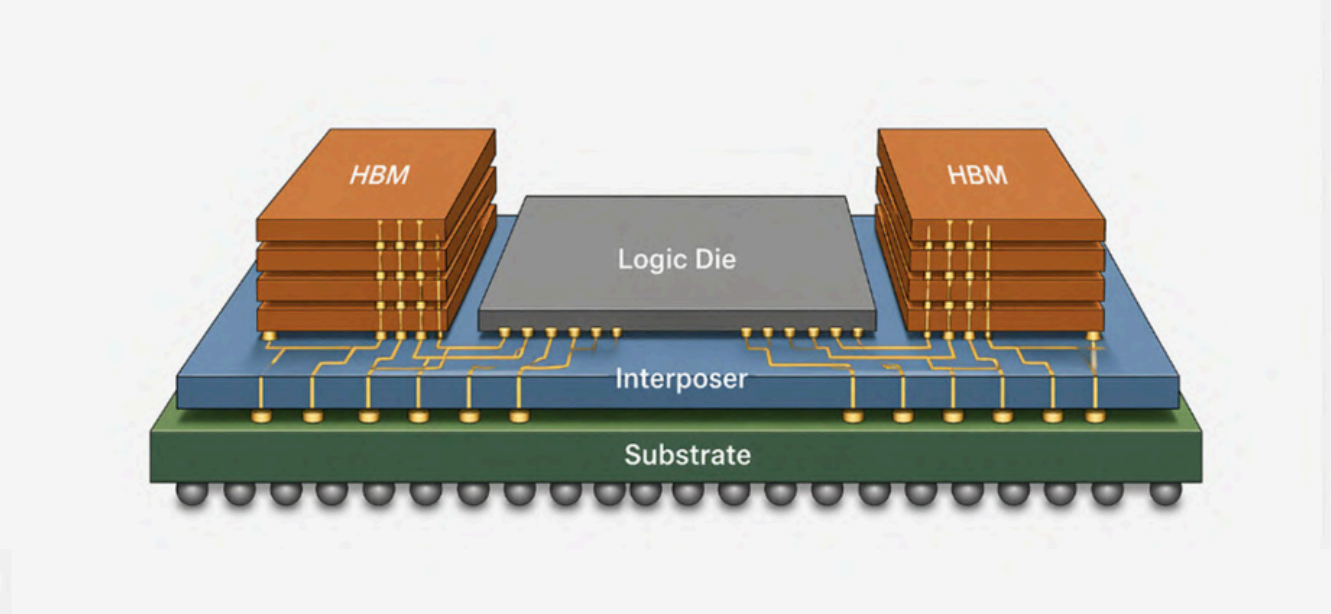
However, as the proportion of memory in chips continues to grow, reliability challenges are also amplified. Process variation, aging effects, and workload stress can all lead to SRAM faults or latent error accumulation. This elevates “testability” and “repairability” into essential capabilities for both mass production and long-term system operation—an area that ISTART-TEK has long been focused on.

ISTART-TEK provides comprehensive solutions for on-chip memory in AI processors. Through MBIST and MBISR architectures, combined with User-Defined Algorithms (UDA), it improves fault diagnosis efficiency and coverage, enabling memory issues to be accurately identified and repaired at an early stage, thereby reducing downstream risks and field failure rates.

In addition, ISTART-TEK’s MART (MBIST Algorithm Recommendation Tool) further enhances intelligence in memory testing and repair workflows. It can automatically recommend optimal test and repair strategies based on different memory types, defect characteristics, and application scenarios, improving testing efficiency and shortening development cycles. In the context of highly customized and rapidly iterating AI chips, such tool-based capabilities are becoming a critical foundation for improving production efficiency.

Overall, AI infrastructure competition is no longer a contest of GPU compute power alone, but is shifting toward system-level efficiency across CPU, GPU, and memory. As computing architectures become more integrated and heterogeneous, the stability and repairability of memory will directly determine the scalability and reliability of AI systems. At this inflection point in the industry, memory test and repair technologies have become an essential foundation supporting the continued expansion of the AI era.

CoWoS and HBM: The Key Engines of the Memory Industry in the AI Era



In recent years, the explosive growth of AI computing demand has made memory bandwidth and capacity central performance metrics for large language models, generative AI, and high-performance computing (HPC) workloads. These applications must access massive volumes of model weights and intermediate data within extremely short timeframes, making “data movement” an increasingly critical bottleneck in the semiconductor industry. Against this backdrop, TSMC’s CoWoS (Chip-on-Wafer-on-Substrate) advanced packaging technology and High Bandwidth Memory (HBM) have formed a highly symbiotic relationship, redefining the architecture of AI servers and high-end computing platforms.

First, CoWoS plays a pivotal role in breaking traditional memory bandwidth limitations. In conventional architectures, CPUs or GPUs connect to external DRAM through printed circuit boards, resulting in long signal transmission distances, high power consumption, and limited bandwidth. By introducing a silicon interposer, CoWoS tightly integrates logic chips and multiple HBM stacks within a single package, shortening data transmission distances to the millimeter scale and dramatically increasing I/O density and efficiency. This architecture enables AI accelerators to access memory at extremely high bandwidth, making it the standard configuration for modern AI GPUs and accelerators. As the demand for KV cache in large models grows rapidly, high bandwidth and large memory capacity have shifted from being “nice-to-have” features to absolute necessities.

Second, CoWoS is accelerating the evolution of HBM technology generations. As AI model parameter counts continue to grow, the memory capacity required within a single package keeps increasing.

From HBM2 and HBM2e to HBM3 and HBM3e, stack height and bandwidth have consistently improved. Future AI chips may integrate even more memory stacks to support larger models and more complex inference workloads. In other words, advanced packaging capability has become a key factor determining the pace of memory specification upgrades.

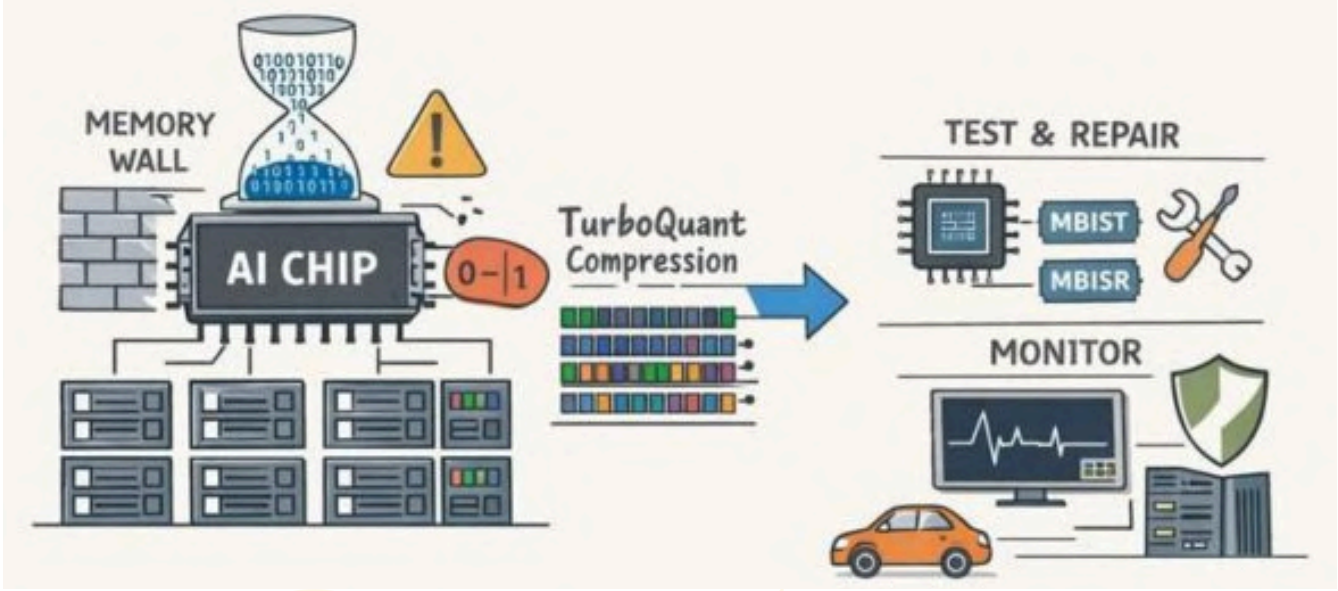
This demand surge is also reshaping the supply–demand structure of the overall memory market. Because HBM is a high-margin product, DRAM manufacturers are shifting substantial production capacity toward HBM, leading to relatively tighter supply for traditional PC and consumer electronics DRAM. Meanwhile, CoWoS capacity itself faces high technical barriers and equipment constraints, meaning AI chip shipments are limited by advanced packaging capacity. As a result, the HBM supply shortage may persist through 2027–2028. Even as the industry explores methods to reduce memory requirements—such as improved caching mechanisms and new data compression techniques—high-bandwidth memory remains irreplaceable for top-tier AI computing.

Looking ahead, packaging technologies will continue to evolve toward higher levels of integration. Following CoWoS-S, CoWoS-R, and CoWoS-L, hybrid bonding is widely viewed as the next key breakthrough. With finer interconnect pitch and higher I/O density, hybrid bonding brings memory and logic chips closer to “chip-level” interconnection. This advancement not only reduces power consumption but also significantly boosts overall system performance, paving the way for the next generation of AI computing.

Within this industry trend, HBM is not only a core component of high-performance computing but also one of the most challenging memory types in terms of yield and reliability. Its stacked structure and advanced packaging significantly increase testing and repair complexity.

Ensuring memory quality and long-term reliability before shipment has become a critical competitive factor. iSTART-TEK has long been dedicated to memory testing and repair technologies and holds ISO 26262 TCL1-certified automotive-grade test algorithms and repair solutions, helping chips maintain high yield and reliability under advanced packaging and high-bandwidth architectures, and serving as a key enabler of memory quality assurance in the AI era.

New Concerns in AI Chip Design: Insights from Google TurboQuant



In an era where AI model parameters are growing exponentially, the “Memory Wall” has become a common adversary for all chip designers. Google’s recent unveiling of TurboQuant has undoubtedly sent shockwaves through the industry. This technology reveals a core trend for AI chip development over the next five years: how to extract maximum data value within limited space and bandwidth.

Higher Data Density, Lower Margin for Error

The core mission of Google TurboQuant is to optimize the KV Cache during Large Language Model (LLM) inference. It can compress data by up to 6x, meaning a model that previously required six servers might now only need one.

However, while we immerse ourselves in the “software dividend” brought by algorithms, the other end of the semiconductor supply chain faces a brand-new challenge.

It is a physical inevitability: as data is extremely compressed, the informational weight carried by each individual bit increases. In uncompressed data, a Bit Flip might only result in minor noise; however, under high-ratio quantization like TurboQuant, an error in a critical bit could lead to a total logical collapse of the inference result.

Furthermore, to accommodate such compression techniques, AI chips are integrating denser SRAM and high-performance memory architectures. As process nodes shrink to 3nm and beyond, the yield and long-term stability of memory units have become the deciding factors in whether an AI chip can be mass-produced and commercially deployed.

From “Functional” to “Durable”: The Indispensable Underlying Guardian

While the industry focuses on how AI can reduce memory consumption, the next step for chip designers is to consider: how can these highly strained memories maintain stable performance without failing? This is exactly the role iSTART-TEK plays within the ecosystem.

In the evolutionary path of High-Performance Computing (HPC) and AI chips, the Memory Testing and Repair (DFT/MBIST/BISR) solutions provided by iSTART-TEK serve as the safety foundation supporting software dividends like TurboQuant.

- **Testing:** Through advanced tools like START v5, iSTART-TEK embeds powerful MBIST (Memory Built-In Self-Test) functions during the chip design phase to accurately filter out defects.
- **Repair:** Utilizing MBISR (Memory Built-In Self-Repair) technology, the system automatically activates redundant space to repair bit errors when detected. This transforms potentially scrapped chips into functional ones, directly boosting production yield.
- **Monitoring & Security:** Throughout the chip’s lifecycle, monitoring solutions provide real-time health status of the memory. This is particularly crucial in HPC and automotive sectors where high reliability is non-negotiable.

The True Future of AI: The Convergence of Software and Hardware

The release of Google TurboQuant demonstrates the determination of AI algorithms to break through existing bottlenecks. However, the second half of this memory revolution will inevitably return to a competition over underlying hardware reliability.

From Google solving “how to save space” to iSTART-TEK solving “how to ensure quality,” it is only through absolute memory stability that software-side compression technologies can truly realize their full potential.

Episode 51: From Testing to Intelligent Protection: The Comprehensive Evolution of Memory Quality Control Tools — Watch the video

Episode 50: Building Customizable Memory Test Methodologies for a Rapidly Evolving Semiconductor Era — Watch the video

Episode 49: AI Appliances and the Smart Display Era: From Voice Control to CIM Chip Testing Challenges — Watch the video

Episode 48: In-Vehicle AI: The Next Computing Battleground in Automotive Electronics — Watch the video

Episode 47:When MRAM Meets AI: The Key Memory Revolution for Next-Generation Intelligent Chips — Watch the video

Episode 46: The Importance of SRAM and Testing Strategies under NPU Operational Requirements — Watch the video

Episode 45: Why LPUs Use SRAM and the Key Testing Considerations — Watch the video