iSTART

April 2026

iSTART iReport

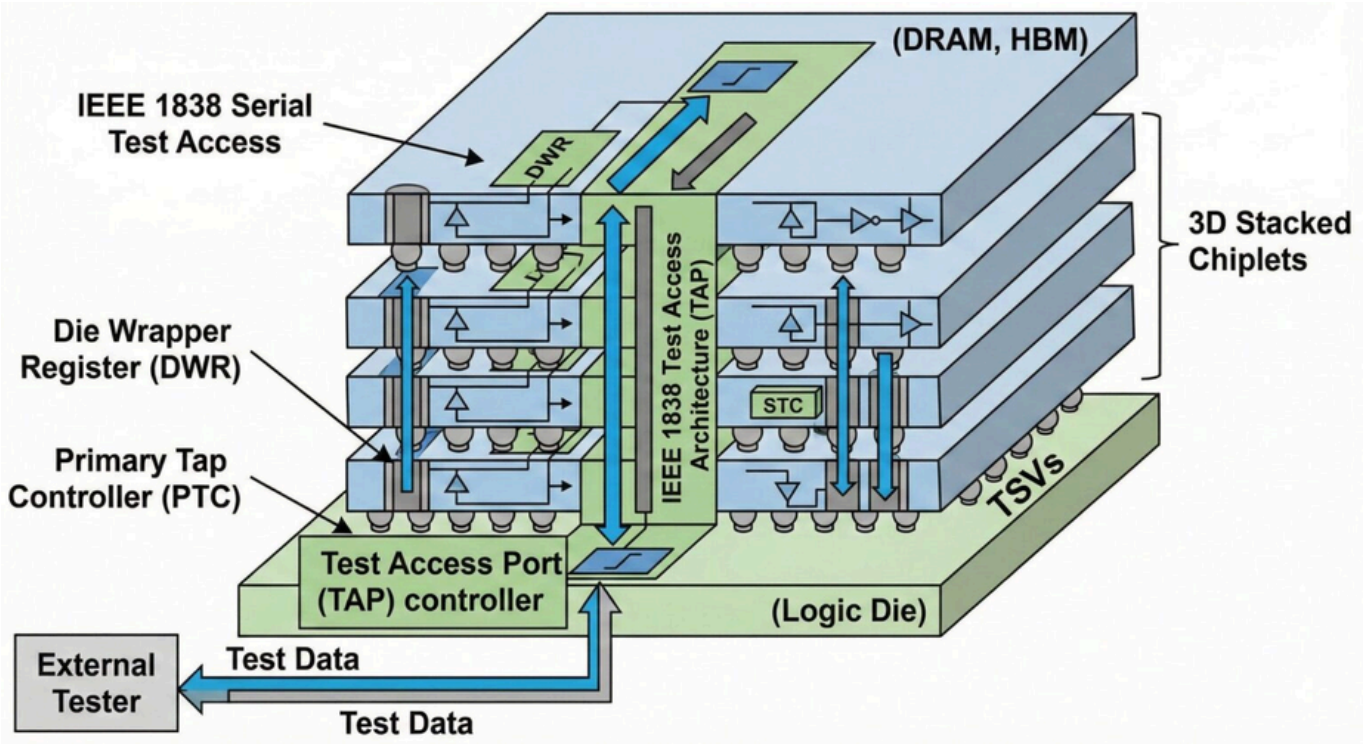
Issue No. 15

iSTART-TEK's IEEE 1838 Solution Ready, Advancing into the AI Chiplet Advanced Packaging Opportunity

As AI computing demand continues to surge, the semiconductor industry is rapidly transitioning toward an advanced packaging era centered on 3D-IC and Chiplet architectures. The vertical stacking of High Bandwidth Memory (HBM) and logic dies has become a critical technology path for delivering the performance and energy efficiency required by AI accelerators. Among these approaches, 2.5D/3D packaging architectures represented by CoWoS significantly enhance overall system performance, but they also introduce unprecedented challenges in testing, repair, and yield management. iSTART-TEK, a company specializing in memory test and repair technologies, announced that its IEEE 1838 3D-IC testing solution has entered a mature application stage. The company is now one of the few EDA vendors capable of fully supporting 3D-IC testing standards and deploying them in real-world advanced packaging projects, enabling early positioning in the rapidly growing AI Chiplet and HBM advanced packaging market.

Early Deployment of IEEE 1838 Establishes a High Barrier for 3D-IC Testing

IEEE 1838 is a testing standard specifically developed for 3D-IC and heterogeneous integration architectures, covering die-to-die test access, test path reconfiguration, and post-stack design-for-testability. Due to its high technical threshold and integration complexity, the number of EDA tools in the market that can maturely support IEEE 1838 remains very limited. iSTART-TEK began investing in related architectures and testing flows well before 3D-IC entered large-scale commercial adoption. By incorporating IEEE 1838 into its memory test and repair platform and completing integration with existing BIST and BISR mechanisms, iSTART-TEK has established a strong technological moat in the 3D-IC testing domain, while significantly raising the entry barrier for latecomers.



Deep Alignment with Heterogeneous Integration for HBM and Logic Stacking

The performance bottleneck of AI chips is no longer confined to compute units alone, but increasingly depends on data transfer efficiency between HBM and logic dies. As a result, vertical integration of HBM with GPUs, AI accelerators, and custom logic chips has become the mainstream design direction, requiring test architectures that can simultaneously address both memory and logic stacking relationships.

iSTART-TEK's IEEE 1838 solution is designed specifically for heterogeneous integration scenarios involving HBM and logic dies. It supports post-stack test path planning, test access management, as well as memory-level fault localization and repair strategies. Through this architecture, customers can maintain testability and repairability for HBM and SRAM within advanced packaging flows such as CoWoS, ensuring the performance and reliability of AI systems.

Reducing CoWoS Scrap Risk and Maximizing ROI in Advanced Packaging

Under CoWoS and 3D-IC architectures, a finished device integrating HBM and high-end logic dies represents a substantial manufacturing cost. If memory or interconnect defects are only discovered after packaging, the lack of effective test and repair mechanisms can result in significant scrap losses.
By integrating IEEE 1838 with its memory repair technologies, iSTART-TEK enables customers to perform precise fault diagnosis and repair even after multi-die stacking is completed, substantially improving final yield. This not only directly reduces scrap costs in advanced packaging, but also allows customers to effectively maximize overall return on investment when adopting CoWoS and AI Chiplet architectures.

Amid the Rise of NPUs, Edge AI Is Redefining AI Computing Architectures

As AI applications move from the cloud to end devices, Neural Processing Units (NPUs) have become a major focus of industry discussion. Compared with traditional CPUs and GPUs, NPUs deliver higher efficiency and lower power consumption, enabling AI models to run directly on edge devices. This not only reduces latency and dependence on the cloud, but also strengthens data privacy.

Kneron, an early pioneer in Edge AI, has long focused on edge AI SoC and NPU architecture design. By integrating AI chips, edge computing, and image algorithms, Kneron continues to accelerate real-world applications across smart IoT, autonomous driving, and intelligent security.
iSTART-TEK has extensive experience collaborating with Kneron.
Driven by technical exchanges and practical expertise from Kneron, iSTART-TEK continues to advance its memory test and repair solutions, helping next-generation AI SoCs achieve optimal performance, yield, and reliability in mass production.

The New Frontier of AI Inference: SRAM is the Key to Performance—Are You Ready?

NVIDIA GTC has once again ignited the AI revolution! According to recent reports, NVIDIA is expected to unveil the V4 LPU, leveraging TSMC's N3P process and CoWoS-R packaging. Integrated with GPUs and NVLink, this creates a powerhouse system solution for the next generation of computing.

As Generative AI shifts from training to massive inference demands, traditional memory bandwidth is hitting a bottleneck. High-speed SRAM has emerged as the "front-line" weapon to break through these performance limits. With TSMC and global IC design giants doubling down on SRAM architecture, the true competitive edge lies in ensuring the reliability and yield of these massive SRAM arrays.

iSTART-TEK provides the industry's most comprehensive memory testing and repair solutions to give you the leading advantage:

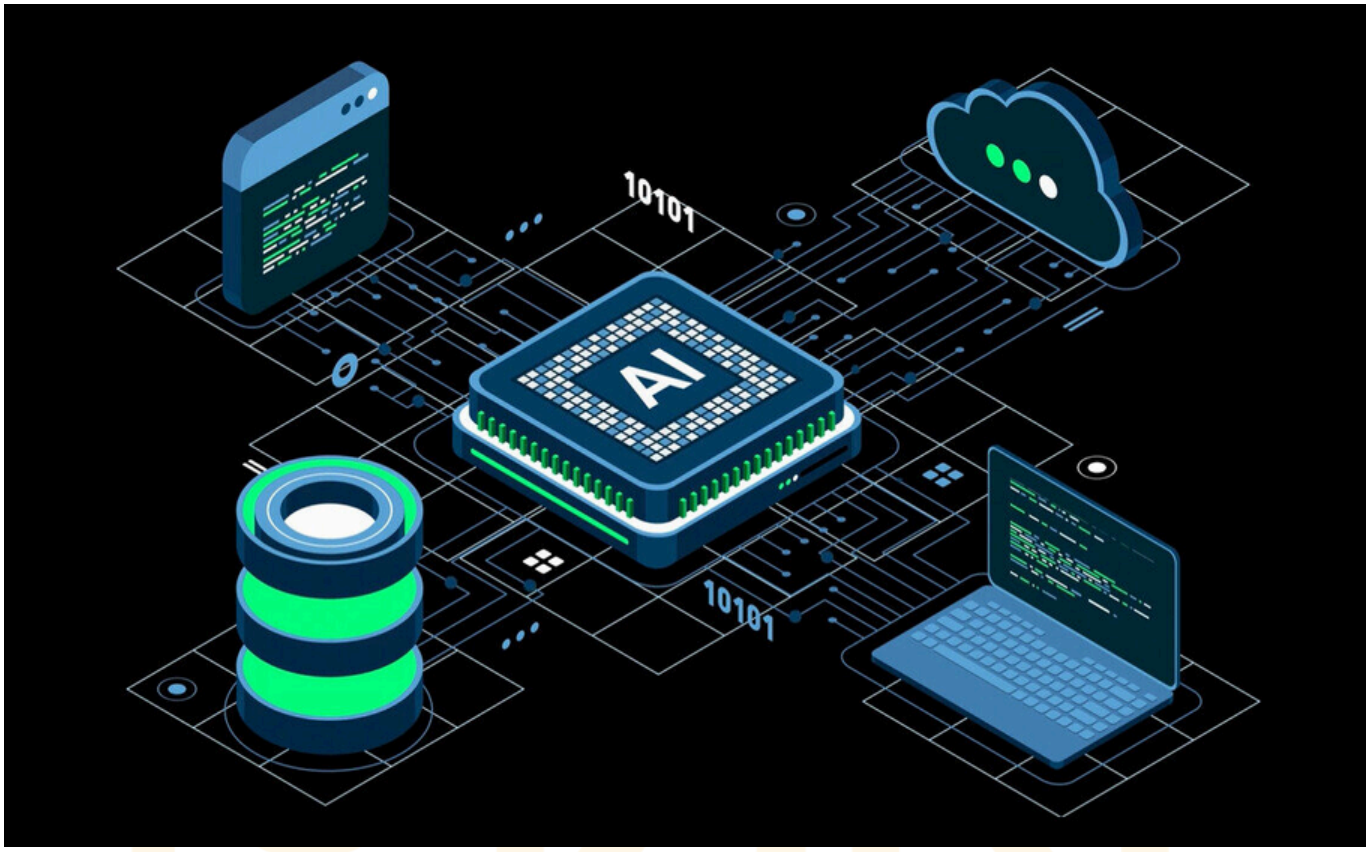
Automated Development Platform: Drastically reduces memory test circuit design time and accelerates Time-to-Market .

High-Efficiency Repair Technology: Optimizes SRAM yield, minimizing the high costs associated with advanced process nodes.

Support for Complex Architectures: Featuring ISO 26262 TCL1 automotive-grade certification, easily meeting the high-spec memory demands of AI chips.

On the road to ultimate AI performance, is your most reliable partner.

SRAM Repair Emerging as a Critical Growth Driver Under the AI Wave



From advanced process licensing and the integration of PUF (Physical Unclonable Functions) into AI architectures, to the expanding demand for Cloud AI and High-Performance Computing (HPC), industrial capital and technical focus are rapidly consolidating toward AI-related applications. Recent market signals clearly indicate that AI Data Center Processors, HBM (High Bandwidth Memory), and various AI Accelerator chips are driving up the reliability requirements for embedded memory.

In advanced manufacturing nodes (3nm and below), the size of SRAM bit cells continues to shrink, making them more sensitive to process variations, random defects, and aging effects. In AI inference and training architectures, massive amounts of on-chip SRAM, caches, and buffers are utilized to reduce latency and data movement costs, leading to a significant increase in SRAM capacity per SoC. SRAM is no longer a supporting player; it has become a core component directly impacting yield, power consumption, and system stability.

Industry Trend

Under this trend, SRAM Repair has become increasingly vital, evolving from an auxiliary role into a cornerstone technology for growth. As internal memory scales expand, even a minute defect can lead to the scrapping of an entire chip. Efficient and precise repair mechanisms have become essential for controlling costs and ensuring mass production stability. However, under the conditions of high capacity and advanced processes, traditional memory testing methods face challenges such as excessive testing time, redundant patterns, and rising costs.

To address these challenges, iSTART-TEK's START™ v5 (Winner of the 2025 EE Awards Asia – EDA Tool of the Year) utilizes patented algorithms and a UDA (User Defined Algorithm) modular architecture. This allows engineers to adjust testing strategies based on specific process and product characteristics, optimizing testing efficiency and production costs while simultaneously enhancing yield and reliability.

Extended Reading:

- From TPU, LPU, NPU to CIM: Why Advanced AI Computing Cannot Do Without SRAM
https://lnkd.in/gKj45ucs
- When TOPS No Longer Equals Performance: Where Is the True Bottleneck in AI Compute?
https://lnkd.in/gH-w8NGx

When Memory Becomes a Scarce Resource, Who Controls Mass Production and Cost?

As DRAM and NAND Flash enter a new cycle of structural price increases, memory has become a critical resource that directly impacts system cost, delivery schedules, and supply risk. This round of price increases is not simply driven by a recovery in traditional end-market demand, but by the fact that long-term consumption of high-density, high-reliability memory is becoming the norm across AI inference, automotive electronics, and edge computing.

Under these conditions, the more immediate challenge facing chip designers and system manufacturers is how to convert available memory resources into products that can be manufactured at scale under constrained supply. Particularly in automotive and AI applications, memory failures not only affect yield but can also disrupt overall system configurations and product launch schedules.

Industry Trend

iSTART-TEK's memory test and repair solutions are specifically designed to address this shift in industry dynamics. Through a comprehensive BIST, BISR, and repair architecture, customers can increase usable memory capacity, reduce scrap rates, and alleviate the pressure caused by rising memory unit costs—without adding process cost.

Built on the START™ v5 platform, iSTART-TEK further provides the UDA (User-Defined Algorithm) framework, enabling customers to customize test and repair algorithms based on actual AI workloads or automotive usage scenarios, rather than being constrained by fixed, one-size-fits-all patterns. This capability is particularly critical for applications characterized by long operating durations and high access frequencies.

In an era of tightening memory supply and a rising price baseline, test and repair have become core capabilities for cost control and supply assurance. Leveraging its long-standing expertise in this field, iSTART-TEK helps customers systematically manage memory-related risks and achieve optimal solutions.

Contact us at https://lnkd.in/ga5FxpeU

NVIDIA GTC Highlights a New Memory Battlefield in the AI Era: From SRAM to MLC NAND Yield Defense

NVIDIA GTC 2026 has just concluded, once again drawing global attention to "AI inference." During the keynote, CEO Jensen Huang clearly stated that we have entered the era of the "Token Factory," where computational demand has surged by a millionfold in just two years.

Amid this wave of computation, semiconductor architecture is undergoing an unprecedented transformation. In the past, the focus was on compute power, but today, memory performance has become the decisive factor in determining the responsiveness and operational cost of Agentic AI.

A New Trend in Disaggregated Inference: When GPUs Meet LPUs

Among the discussions surrounding GTC, the most notable technological shift is "disaggregated inference." NVIDIA has integrated technology from the Groq team to introduce the new Groq LP30 chip. The underlying logic is to overcome the "memory wall" bottleneck.

The Vera Rubin GPU excels at the "prefill" stage, leveraging massive compute power to process large-scale context. In contrast, the Groq LPU specializes in the "decode" stage, focusing on token generation. Its core design abandons traditional high-bandwidth memory (HBM) in favor of massive on-chip SRAM.

Why SRAM? Because it offers extremely low latency and high speed. During real-time AI interactions or code generation, data can flow rapidly within the chip without frequently accessing external memory. Therefore, the capacity and stability of SRAM directly determine the competitiveness of AI chips.

The Unexpected Comeback of 2D NAND: Behind the Surge in MLC Prices

While advanced process nodes aggressively pursue SRAM and HBM, the mature-node memory market is experiencing unexpected disruption. According to the latest industry insights, niche memory segments are seeing a peculiar phenomenon of "price doubling."

Major players such as Samsung and Micron are accelerating their exit from low-capacity, legacy 2D NAND production lines to reallocate capacity toward higher-margin AI applications, such as 300+ layer 3D NAND. As a result, essential MLC NAND used in networking devices, digital TVs, and set-top boxes is facing severe supply shortages. Supply in 2026 is expected to drop by half, with a market gap reaching 30% to 40%.

This phenomenon underscores that even in non-core AI applications, memory reliability and stable supply remain critical to business survival. In the face of unprecedented NAND price hikes, ensuring the flawless quality of every shipped memory unit has become a key determinant of profitability for IC design houses and manufacturers.

Chip Yield: The Hidden Moat in the AI Era

Whether it is the SRAM-centric LPU highlighted at GTC or the increasingly scarce MLC NAND, both share a common technical challenge: yield and reliability.
As process technologies advance to N3P and beyond, SRAM occupies an increasingly larger portion of chip area, raising the probability of defects. If an expensive AI chip is scrapped due to faulty SRAM, the financial loss can be enormous. This is why the investment market has recently shown strong interest in testing technologies. In the pursuit of extreme performance, "test and repair" has evolved from a supporting role into a fundamental safeguard of chip design quality.

Intelligent and Flexible Memory Test & Repair Technologies Become Profit Drivers

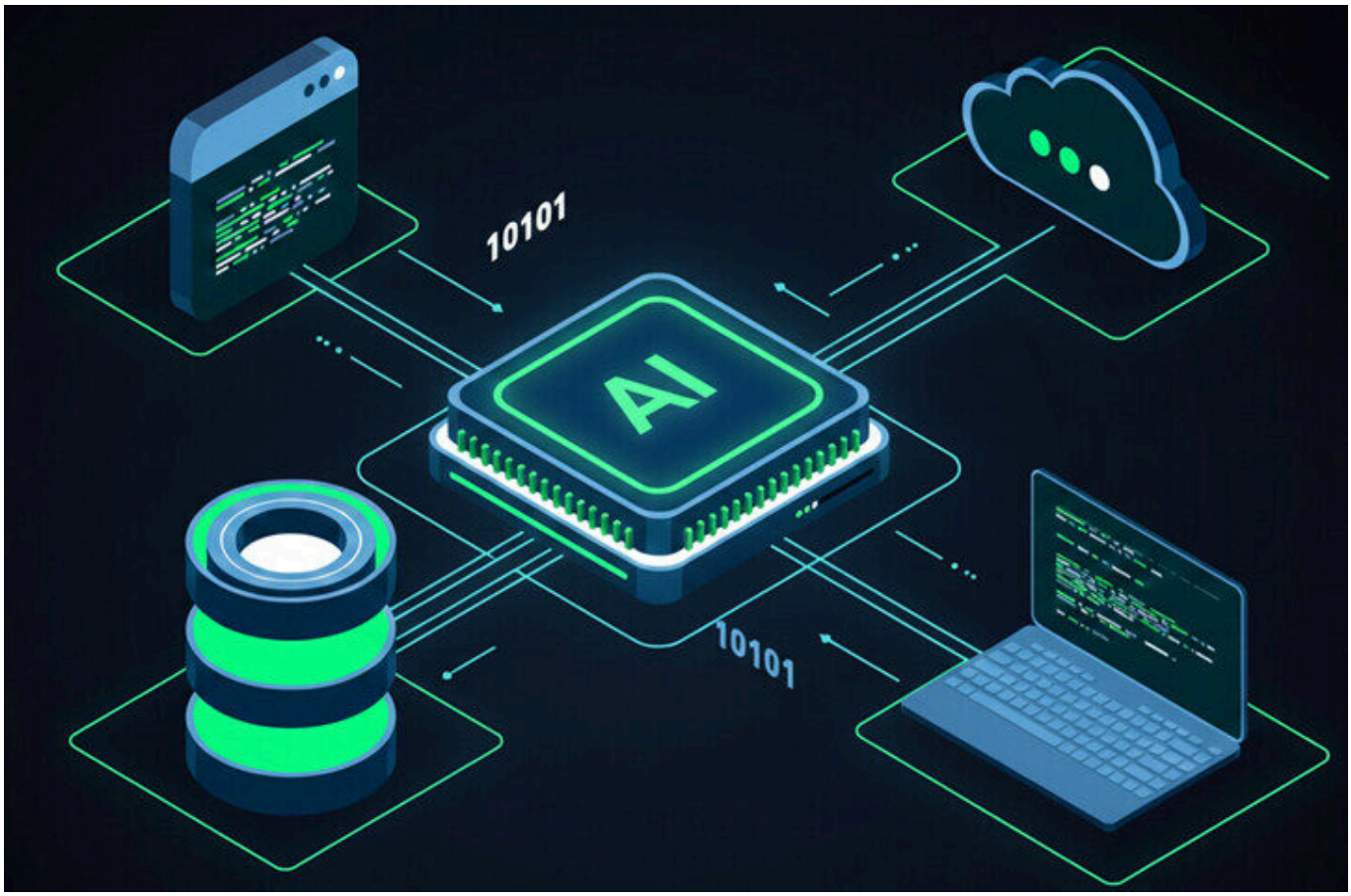
Amid the surge of AI inference demand and memory shortages, iSTART-TEK has become a key guardian of SRAM yield through its long-established expertise in memory test and repair technologies.
If standard algorithms represent the baseline defense, iSTART-TEK's User-Defined Algorithm (UDA) platform serves as a flexible R&D environment. Through a graphical interface and proprietary Testing Elements Change (TEC) technology, engineers can design customized test strategies for specific conditions—such as high temperature or low voltage—without writing complex code.

Whether detecting complex leakage defects or diverse memory fault types (such as SAF and DRDF), UDA enables comprehensive coverage, ensuring stable operation of chips even under harsh conditions.

At the same time, to address the growing complexity of SRAM structures in advanced nodes, iSTART-TEK has introduced MART (MBIST Algorithm Recommendation Tool), which incorporates AI into the test flow. This overcomes the limitations of traditional lookup tables and rigid test selection processes.

With simple interactive inputs, the system performs AI-weighted analysis based on factors such as power, area, and yield. This significantly reduces decision-making costs and DPPM (defective parts per million), transforming BIST algorithm selection from experience-driven judgment into intelligent, precision-based recommendation—helping customers secure time-to-market advantages in the highly competitive AI chip landscape.

Advanced Process and AI Server Architecture Evolution: The Strategic Transformation of Discrete NOR Flash in System Management



In the context of the rapid evolution of high-performance computing (HPC) and AI servers, the role of memory within system architectures is undergoing a significant transformation. Market demands increasingly emphasize data movement costs, system power consumption, and overall operational efficiency. Non-volatile memory (NVM) has evolved from a purely passive storage component into a critical element participating in system resource allocation, security management, and boot operations. Against this backdrop, there has been a clear shift in the positioning of embedded flash memory (eFlash) and standalone NOR Flash.

Historically, in microcontrollers (MCUs) and low-power edge computing SoCs, developers preferred to integrate eFlash directly on-chip to maintain high integration levels and reduce system complexity. However, as semiconductor process nodes shrink below 28nm and even enter advanced FinFET processes, eFlash faces severe structural limitations. Writing to eFlash requires high-voltage devices and specialized process modules, which physically conflict with advanced logic processes designed for low-voltage operation and ultra-thin gate oxides. Forcibly integrating large-capacity eFlash typically requires approximately 8–12 additional masks (depending on the process platform and memory architecture), significantly increasing production costs while also challenging yield and reliability control. In advanced process nodes, continuing to pursue on-chip eFlash integration is no longer economically viable.

This technical bottleneck has driven system designs toward external storage solutions, bringing discrete NOR Flash back into focus for various high-performance applications. Using QSPI, OSPI, or other high-speed serial interfaces, SoCs can store firmware, boot programs, and critical configuration parameters in external dedicated memory chips. This division of labor allows logic chips to focus on high-density computation and AI acceleration, while NVM is handled by memory vendors with specialized processes, improving overall process flexibility and supply chain cost management.

This shift is particularly critical in AI server architectures. Although the primary computation and data exchange in AI systems rely on volatile memories such as HBM or DDR5, the Baseboard Management Controller (BMC), responsible for low-level system management, relies heavily on NOR Flash. As AI server structures become increasingly complex, with numerous GPUs, FPGAs, and network switching components, the BMC uses discrete NOR Flash to store core firmware, perform remote monitoring, and conduct fault diagnostics. To address cybersecurity threats, systems typically implement secure boot mechanisms (Root of Trust), with encrypted signatures and secure boot code stored in highly reliable NOR Flash.

Furthermore, AI servers demand reliability standards far beyond those of consumer electronics. Any boot failure or firmware corruption can halt an entire compute cluster, incurring significant operational losses. Consequently, NOR Flash used in data centers must meet stringent verification standards regarding failure rate (DPPM), endurance, and data retention under high-temperature conditions. These high-reliability requirements have transformed NOR Flash from a simple boot component into a strategic element that ensures system stability and security. With the adoption of more frequent firmware updates and multi-version management in AI systems, the read/write cycle endurance and data integrity verification of NOR Flash become increasingly critical.

From an industry structure perspective, this change has prompted a redivision of labor within the semiconductor supply chain. While eFlash was once a core integration capability for SoC vendors, moving storage functions to specialized memory suppliers in the context of AI and advanced process nodes improves overall efficiency. Logic design companies can focus on computation architectures and accelerator design, while memory vendors build technological barriers in high-reliability markets. Overall, the integration threshold in advanced process nodes has created new growth opportunities for external NOR Flash. The expanding AI server market drives demand for high-reliability non-volatile storage, establishing NOR Flash as an indispensable component in system management architectures and a key enabler of stable compute performance in the AI era.

When TOPS No Longer Equals Performance: Where Is the True Bottleneck in AI Compute?

In recent years, almost every AI chip launch has revolved around a single keyword: TOPS. Whether in data center GPUs, automotive SoCs, or edge AI processors, each generation boasts higher compute numbers—from tens or hundreds of TOPS to thousands or even tens of thousands. On the surface, AI compute seems no longer a problem, but in real-world system design and applications, performance bottlenecks still frequently occur. This has led engineers to reconsider a key question: is TOPS really the core metric that determines AI performance?

TOPS (Tera Operations Per Second) represents the theoretical number of operations a chip can perform per second under a specific precision. Most AI chips report INT8, INT4, or other low-precision operations as the standard because these are most relevant to inference scenarios and allow for impressive compute numbers. However, TOPS alone does not equate to real-world performance—it is more like the maximum horsepower of an engine and does not reflect whether the system can sustain that performance over time.

Take NVIDIA as an example: the AI compute of recent generations of data center GPUs has already reached the tens of thousands of TOPS. Products like the H100 and B200 offer extremely high theoretical compute in low-precision AI modes, sufficient for running large language models and generative AI inference. In the endpoint and edge markets, NVIDIA's Jetson series, Qualcomm, MediaTek, Apple, and Google have also released NPU SoCs with tens to hundreds of TOPS for image recognition, speech processing, and on-device AI inference. From the datasheet perspective, AI compute appears to be fully addressed.

In practice, however, AI inference is not just about computation. Every operation requires reading weights and feature data from memory beforehand and writing results back after computation. As model sizes grow and data reuse increases, system performance is often limited by data movement rather than the compute units themselves. This explains why, in many applications, actual performance does not scale linearly with the chip's stated TOPS.

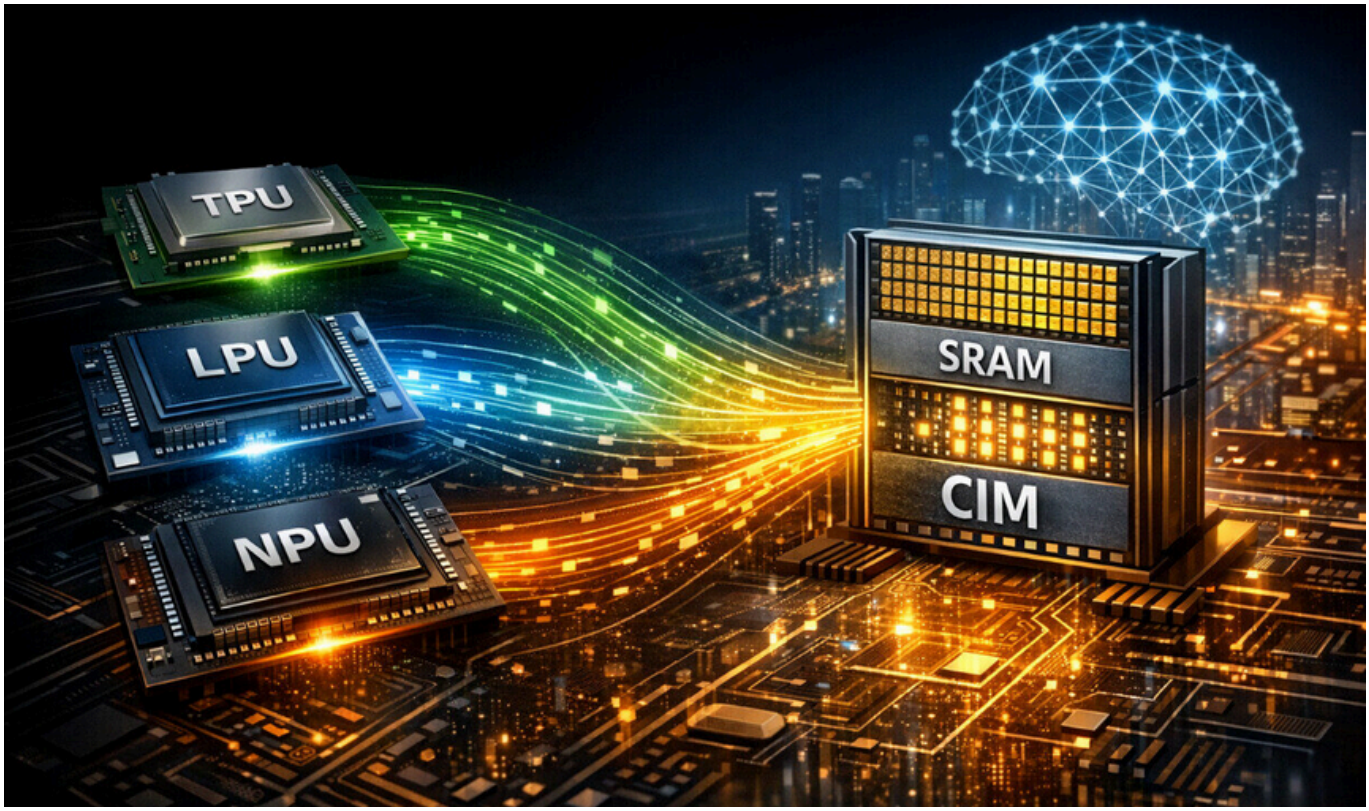
This brings the focus back to memory architecture. To support high-TOPS computation, AI chip vendors have been enhancing memory bandwidth. Data center GPUs increasingly adopt HBM, leveraging stacked packaging and ultra-high bandwidth to shorten the distance between compute units and external memory. In SoC and NPU designs, the proportion of on-chip SRAM continues to rise, becoming a critical resource.

SRAM in AI chips is more than just temporary storage. It handles high-frequency, low-latency data access, supporting weight caching, feature map buffering, and intermediate result storage. For CNNs, Transformers, and similar models, repeated data access is extremely frequent. If every access requires going back to external DRAM, latency increases and power consumption rises sharply. Therefore, many AI architectures choose to keep critical data in on-chip SRAM to fully realize the compute potential represented by TOPS.

Because SRAM plays a key role, its reliability and test coverage become increasingly important. Under high-frequency, long-duration operation, SRAM is prone to issues like read disturb, coupling interference, and aging. Any error can affect not just a single operation but the stability of the entire AI inference result. As a result, memory testing and repair mechanisms have become indispensable in AI chip design.

From the rapid growth of TOPS to the continuous evolution of memory architecture, it is clear that the competitive focus of AI chips is shifting. Compute power remains important, but the true differentiator is often who can make data flow more smoothly and memory more reliable. When TOPS is no longer just a marketing number but can be fully realized, the true value of an AI chip is finally unleashed.

From TPU, LPU, NPU to CIM: Why Advanced AI Computing Cannot Do Without SRAM



Recent news of NVIDIA's acquisition of Groq has sparked widespread discussion, putting AI-specific processors such as TPU, LPU, and NPU back into the spotlight. Whether in cloud data centers, edge AI, or automotive and industrial applications, computing architectures are rapidly shifting from general-purpose CPUs to accelerators tailored for specific workloads. Yet behind these differently named and positioned processors lies a common and critical core: SRAM.

The Common Foundation of TPU, LPU, and NPU

Although TPU, LPU, and NPU have different design goals, they all focus on high throughput, low latency, and energy efficiency. In practice, large volumes of weight data, feature maps, and intermediate results must be repeatedly and rapidly accessed. Compared with DRAM, SRAM offers low latency, high bandwidth, and predictable timing, making it an indispensable memory component in these AI accelerators.

From "Memory-Assisted Computing" to "Computing-in-Memory"

As AI models continue to scale, the energy consumed by moving data between compute units and memory has become a major bottleneck for performance and power. This is precisely why Computing-in-Memory (CIM) architectures have attracted significant attention. The core idea of CIM is to perform part of the computation directly within the memory array, drastically reducing data movement and overcoming the limitations of the traditional von Neumann architecture.

For TPUs, LPUs, and NPUs that already heavily rely on SRAM, incorporating SRAM-based CIM in specific computational scenarios is widely regarded as a necessary evolutionary step. By integrating computing capabilities within or around the SRAM, AI accelerators can significantly improve energy efficiency while maintaining high reliability, making CIM a key technology for next-generation AI chips.

CIM Goes Beyond Performance: Reliability and Security Matter

The design challenges of CIM go beyond mere performance. Memory variability, aging effects, and the added complexity of coupled compute-and-store operations all make SRAM testing and repair more critical than ever. In high-reliability applications such as automotive and critical infrastructure, CIM cannot be truly mass-produced and deployed without comprehensive test and repair mechanisms.

Meanwhile, post-quantum security requirements increasingly demand the integration of AI with cryptographic computation. The ability to execute next-generation cryptographic algorithms efficiently under low-power conditions has become a critical requirement for smart devices and automotive systems.

Essential Memory Technologies for the AI Computing Trend

In this context, SRAM quality—critical for the performance and cost of TPU, LPU, and NPU—as well as CIM architectures, is increasingly recognized as a crucial path to overcoming data-movement bottlenecks. iSTART-TEK's long-standing expertise in SRAM testing and repair, coupled with its CIM architecture development, positions the company as a key driver of technological advancement.

iSTART-TEK is dedicated to providing specialized EDA tools and IP for testing and repairing various types of memory, offering authorized customers a one-stop design. In recent years, iSTART-TEK has also actively invested in CIM architecture R&D, redefining energy efficiency limits for AI computing. The SRAM-based CIM architecture currently under development supports 8-bit compute precision with ultra-low power consumption and offers strong scalability, with further compatibility for RRAM-based designs to meet diverse application needs.

Facing the security challenges of the quantum era, iSTART-TEK has pioneered the integration of lattice-based cryptography and NTT computation optimization techniques, accelerating post-quantum cryptographic algorithms via the CIM architecture to provide long-term, robust security for smart devices and automotive systems. Beyond CIM architecture development itself, iSTART-TEK also provides a CIM test circuit development environment, ensuring that advanced memory computing technologies not only compute fast, but also test accurately and operate reliably. Amid the ongoing evolution of TPU, LPU, and NPU architectures, iSTART-TEK's deep SRAM expertise has become a critical force in driving CIM and AI chips toward higher efficiency, reliability, and energy savings.

When MRAM Meets AI: The Key Memory Revolution for Next-Generation Intelligent Chips

As AI chip performance continues to advance, system bottlenecks are increasingly shifting from compute cores to memory architecture. In AI inference and training workloads, massive volumes of model weights and intermediate data are repeatedly transferred between memory and compute units. The energy and latency consumed by these data movements often surpass those of the actual computations. To reduce data transfer overhead and improve efficiency, next-generation AI SoCs are introducing MRAM, leveraging its high-speed read/write capability and non-volatile nature to keep critical data resident on-chip, further enhancing power efficiency, reducing latency, and supporting edge AI and low-power applications.

However, MRAM's magnetic storage mechanism is highly sensitive to process variations, material characteristics, and environmental conditions, potentially leading to write switching failures, resistance drift, retention degradation, or read-margin instability. Under high-frequency access and sustained AI workloads, these risks are amplified, making robust test and repair mechanisms essential to ensure yield and long-term reliability. Implementing MRAM in AI SoCs therefore requires a complete BIST and BISR framework, combined with stress testing, long-term retention validation, and verification under varying voltage and temperature conditions to control DPPM effectively.

iSTART-TEK has long specialized in memory test and repair technologies. Its self-developed MRAM BIST IP has been successfully integrated into advanced process SoCs, supporting automotive-grade reliability requirements. In high-frequency AI workload scenarios, iSTART-TEK's solution provides stable test and repair support, enabling customers to balance performance, power efficiency, and long-term reliability while establishing a competitive advantage in next-generation AI memory architectures.

Watch now

The Importance of SRAM and Testing Strategies under NPU Operational Requirements

In practical **NPU (Neutral Processing Unit)** designs, SRAM often occupies the largest portion of chip area, consumes the most power, and has the greatest impact on overall reliability. To support highly parallel AI inference workloads with intensive data reuse, NPUs integrate large amounts of on-chip SRAM around compute units as weight buffers, feature-map storage, and local data-access resources, enabling low-latency and high-throughput operation.

However, this architecture also places SRAM under extreme usage conditions, with long-duration, high-frequency, and repetitive access patterns. Failure modes such as read disturb, coupling interference, weak writes, sensing-margin instability, and retention-related data loss are more easily amplified in NPUs, becoming major contributors to yield loss and elevated DPPM. As a result, NPU performance bottlenecks often arise not from compute capability, but from the design and test quality of the memory subsystem.

This episode examines real NPU operating requirements to analyze how different SRAM usage scenarios—such as weight buffers, feature-map storage, accumulation buffers, and idle regions—introduce distinct reliability risks. It also explains why NPU SRAM testing cannot rely on a single, static test algorithm, and instead requires context-aware, configurable test and repair strategies to effectively mitigate mass-production risk.

To address these challenges, iSTART-TEK leverages the START™ v5 platform, combining the graphical and modular design of UDA (User-Defined Algorithms) to allow engineers to flexibly construct MBIST test flows tailored to different SRAM structures and access behaviors. In parallel, MART (MBIST Algorithm Recommendation Tool) systematizes accumulated SRAM test knowledge, enabling rapid selection of suitable algorithm combinations under multiple constraints, reducing decision cost and preventing production risks caused by improper test strategies.

Through the integrated use of UDA and MART, SRAM testing evolves beyond a basic pass/fail process, aligning more closely with real NPU AI workloads while simultaneously improving yield, controlling DPPM, and ensuring mass-production efficiency and product competitiveness.

Watch now

Why LPUs Use SRAM and the Key Testing Considerations

As edge AI and low-power AI inference applications continue to expand, LPUs (Language Processing Units) increasingly adopt SRAM as their primary on-chip memory. The key reason is not speed alone, but the stringent requirements LPU inference places on memory access latency, stability, and predictability.

This episode explains why, in real-time applications such as image recognition, voice activation, and sensor data analysis, memory behavior often becomes a more critical bottleneck than compute performance. It compares SRAM and DRAM in terms of latency, refresh behavior, power characteristics, and system stability, and highlights the uncertainty risks DRAM may introduce in LPU architectures.

The episode further discusses how modern LPU SoCs integrate large, heterogeneous SRAM banks operating under low-voltage and high-frequency conditions, significantly increasing test complexity. It concludes by introducing how iSTART-TEK's START™ v5 platform, together with UDA and MART, enables application-aware MBIST strategies that better reflect real inference behavior—helping design teams balance reliability, yield, and mass-production efficiency.

Watch now