Vikas Nagaraj[1]

Optimizing DFT Test Coverage for AI Accelerators and Compute Chips

JES Journal of Electrical Systems

Abstract: - The number of transistors, heterogeneous cores, and chiplet-based integrations packed into AI accelerators and compute chips at advanced nodes has steep test and reliability requirements. This paper provides a practical and straightforward design of Design for Test (DFT) coverage optimisation when available resources are limited in terms of cost and time. It begins with the principles of scan, traditional fault models, and access standards (IEEE 1149.1/1500/1687). It proceeds to current developments, including hierarchical DFT, pattern compression, cell-sensitive faults, and power/thermal-sensitive testing. Specific AI silicon issues, such as massive parallel operation, large memory hierarchies, multiple clock/power domains, and 3D packaging, are converted into actual techniques: full/partial scan with at-speed capture, MBIST/LBIST, boundary scan reuse, TSV and micro-bump testing, and stress-controlled scheduling. An overview of the key steps includes the initial plans for RTL/floorplan, coverage target (Example. DPPM), fault model, ML-assisted ATPG, hierarchical IP wrappers, chiplet/3D integration, coverage grading, and production correlation. Cases of improvements in the yield and test time in accelerator and chiplet processors at high volume are seen. In the future, multi-stage access networks will be necessary for heterogeneous integration and wafer-to-wafer stacking. In-field/online DFT will transition to continuous health monitoring and self-repair, and adaptive pattern generation will become possible through AI-based automation of defect analytics. This leads to a practical roadmap for high-yield, reliable, and cost-effective testing of next-generation AI computing.

Keywords: Design for Test (DFT); AI accelerators; MBIST/LBIST; chiplets and 3D integration; power- and thermal-aware testing.

Article Received: 23 July, 2025 **Accepted:** 21 September, 2025 **Published:** 26 October, 2025

1. Introduction

The rapid development of AI accelerators and high-performance compute chips is reshaping data centers, automotive systems, and mainstream consumer electronics. These designs integrate billions of transistors, heterogeneous compute cores, and chiplet-based 2.5D/3D packaging at advanced process nodes such as 5 nm and 3 nm. While these innovations enable unprecedented computational density, they also create steep challenges for testing and long-term reliability. Hidden manufacturing defects, accelerated ageing effects, and complex multi-die interconnects can lead to costly rework, yield loss, or field failures if not addressed early in the design cycle. To mitigate these risks, the semiconductor industry relies on Design for Test (DFT)—a design discipline that embeds controllability and observability into the chip architecture. DFT is far more than a late-stage enhancement; it provides the structural hooks through which production testers can quickly locate faults, isolate weak components, and guarantee product quality. Techniques such as scan-based testing, memory and logic built-in self-test (MBIST/LBIST), boundary-scan standards (IEEE 1149.1/1500/1687), and chiplet/3D-specific test networks allow manufacturers to detect defects efficiently and maintain yield and reliability. When planned from the earliest stages of RTL development and floorplanning, DFT also shortens time-to-market by avoiding late design changes and routing conflicts.

This paper presents a practical and cost-conscious roadmap for optimizing DFT test coverage under aggressive power, performance, and schedule constraints. It combines foundational principles with current advances such as hierarchical DFT, pattern compression, cell-aware and power/thermal-sensitive fault modeling, and AI-assisted ATPG (automatic test pattern generation). The approach transforms AI-specific design challenges—including massive parallelism, large memory hierarchies, multi-clock and multi-power domains, and dense 3D interconnects—into actionable techniques like partial/full scan with at-speed capture, MBIST/LBIST, TSV and micro-bump testing, and stress-controlled scheduling.

[1] MTS at Advanced Micro Device(AMD), San Jose, California, USA

Email: vikas.jodigattenagaraj@gmail.com

This paper is organized into various chapters, each addressing a critical element of DFT coverage optimization. Literature Review outlines classical scan design, traditional fault models, and key access standards, then examines modern trends such as hierarchical DFT, pattern compression, cell-aware faults, and power/thermal-sensitive testing as applied to AI-class silicon and chiplet packaging. DFT Challenges in AI Accelerators and Compute Chips analyzes the implications of massive parallelism, deep memory hierarchies, multi-domain clocking, advanced 5 nm/3 nm process nodes, and 3D interconnect defect mechanisms for controllability, observability, and reliability. Key DFT Techniques and Strategies details the integrated use of scan-based testing, MBIST/LBIST, boundary scan, hierarchical and chiplet-specific test networks, and power- and thermal-aware methods to achieve high fault coverage within strict PPA budgets. Methodology for Optimizing DFT Test Coverage proposes a complete design flow—from early RTL/floorplan co-optimization and DPPM target setting to advanced fault modeling, machine-learning-based ATPG, hierarchical IP wrappers, chiplet/3D integration, coverage grading, and pilot-to-volume production correlation. Balancing Coverage, Cost, and Time examines the economic and engineering trade-offs among coverage goals, silicon overhead, tester memory, and pattern count, including the use of compression and at-speed testing. Case Studies and Best Practices illustrate how hierarchical DFT and multi-stage interconnect testing improve yield and reduce time-to-market in high-volume AI accelerators and chiplet-based processors. Future Trends explores forthcoming needs such as heterogeneous integration with wafer-to-wafer stacking, in-field and online self-test with self-repair, and AI-driven defect analytics for adaptive pattern generation. Recommendations distill practical design guidelines to achieve robust and efficient test coverage, while Conclusion synthesizes the key insights, emphasizing that early, AI-assisted DFT integration is essential for high-yield, cost-effective, and future-ready AI compute silicon.

2. Literature Review

2.1 Foundational DFT principles

Design for Test originates from the primitive scan design techniques and fault model analysis, aiming to gain an understanding of how real silicon can fail. The stuck-at faults were characterized by classic works, in which a node is slightly stuck at logic 0 or 1 and transition faults that isolate speed-dependent problems. To generate these faults, Automatic Test Pattern Generation (ATPG) algorithms were created to generate effective input vectors, which comprise the standard of industrial test flows. These standards include IEEE 1149.1 (also known as JTAG), which established a standardized boundary-scan architecture to be able to access and control internal signals via a simple serial interface. Subsequently, IEEE 1500 qualified test access to embedded cores, and IEEE 1687 (IJTAG) offered an on-chip network protocol that was adaptable for connecting to multiple test instruments [19]. To complement these hardware-oriented developments, system design is increasingly incorporating data-driven considerations, particularly in cases where large-scale databases and high-throughput applications are involved. Similar to the trade-offs applied in test architectures, to guarantee consistency in distributed data systems, such as MongoDB, there must be a balance between performance and reliability, which is no different, as noted by [9]. His article on the subject of MongoDB and the concept of data consistency, "Bridging the gap between performance and reliability," highlights the importance of solid access protocols and clearly defined fault-handling strategies as means of ensuring predictable system behavior in real-life circumstances [8]. It is these initial concepts, scan insertion, fault modelling, and portable test access, that remain fundamentals, not alone on the account of their offering of clear visibility and controllability despite the complexity of the design they are intended for, but because their assertion of structured reliability speaks well to current anxieties about data integrity and system resilience. As illustrated in the figure below, the design integrates a combinational circuit with memory elements to capture current state and determine next state, reflecting foundational DFT principles of scan insertion, fault modeling, and structured test access ensuring reliability and controllability

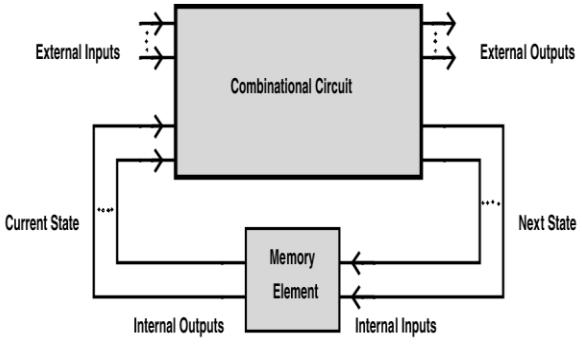


Figure 1: design-for-testability

2.2 superior methods and trends.

As process technology increased to FinFET and gate-all-around nodes, new physical effects and power constraints demanded more complex DFT methods [10]. Hierarchical DFT breaks down large systems into self-contained blocks, making it simple to generate patterns and simultaneously test subsystems in parallel. Embedded Deterministic Test (EDT) and X-masking are test compression methods that minimise tester vectors and coverage without loss, thereby reducing test cost and time. Cell-aware fault modelling goes beyond the mere stuck-at models, taking into consideration faults inside standard cells, which can be very critical in cases where lithography and material constraints form minimal internal shorts or opens. Power- and thermal-conscious DFT uses pattern sequencing and variable-voltage power gating to reduce IR drop and local heating, to match pattern-test strategies to the energy envelopes of more advanced nodes. Simultaneously, the 2.5D and 3D packaging also introduced new forms, such as through-silicon vias and micro-bumps, which require special test access and stress testing to provide reliable vertical interconnects.

2.3 DFT of AI and machine learning accelerators.

Unlike general-purpose processors, AI accelerators have high rates of massively parallel compute arrays, large on-chip memories, and data movement networks. Studies have indicated that systolic arrays together with the tensor cores have specific fault sensitivities, such as at memory interfaces and clock distribution points. Old scan and built-in self-test (BIST) are still in use; however, new flows adopt fault models and patterns to these data-centric structures. The other new direction is to use machine learning to DFT itself: predictive models to determine probable hotspots of faults and use that to drive pattern prioritization. This is reminiscent of predictive analytics techniques used in other technology domains, such as big data management and event-driven systems. These AI-aided techniques have the benefit of saving time and computational effort in Automatic Test Pattern Generation (ATPG), shifting the effort to the most critical areas of the design.

These methods have strategic and operational implications that are consistent with the thought in other fields. DFT of AI accelerators can consider adopting both conventional and AI-based testing techniques to enhance fault coverage and minimize single-point weaknesses, just as dual sourcing techniques in supply-chain management can be used to make the supply-chain more resilient and less prone to risk [11]. Even current-day DFT may incorporate automated, security-conscious fault detection to ensure that testing and validation are kept up to date with rapid design iteration and threats, as part of integrating security into CI/CD pipelines via Develops practices, including SAST, DAST, and SCA [17]. Collectively, these views imply that future versions of DFT to support AI and machine learning accelerators must encompass predictive AI-based analytics and further stratified and secure operations so that design integrity, test efficiency, and supply-chain robustness can all improve concurrently.

2.4 Industry best practices and white papers.

TSMC, Samsung, and Intel foundries have already released their guidelines on yield learning, defect screening and reliability criteria, which can affect DFT planning since the initial RTL drafts [12]. These are guidelines that cover targets of coverage of various defect types, a rule of thumb of power testing and design rules to prevent DFT structures that cannot be compatible with advanced lithography. Vendors such as Synopsys, Cadence and Siemens EDA augment them with tool flows incorporating hierarchical DFT, advanced compression and AI-based optimization. Their work offers practical recipes for incorporating scan chains, MBIST controllers, and scan cells at boundaries, while maintaining low overhead and timing closure predictability.

2.5 Gaps identified

Although this has been made, there are still significant gaps. No standard, widely used methodology has yet been established that optimizes work with AI accelerators, which primarily utilize enormous memory, extensive interconnects, and heterogeneous chiplet-based packaging. Current flows work with one or the other of the following: logic scan, memory BIST or die-to-die link testing, but do not provide an overall framework between RTL and package assembly up to in-field operation. Scaffable DFT enabling wafer-to-wafer stacking, dynamic voltage and frequency scaling, and in-system self-test is on the increase. Such gaps will need to be addressed by integrating basic scan and access standards with adaptive, data-driven approaches that are capable of evolving with the rapidly changing AI hardware ecosystem.

3. DFT Challenges in AI Accelerators and Compute Chips

The Design for Test (DFT) of AI accelerators and compute chips cannot be optimized by simply using standard test flows. These machines are a combination of unprecedented architectural size, advanced technology nodes, narrow performance-to-power-area (PPA) budgets, and complex physical designs. All these factors present

distinct challenges in the quest to achieve high test coverage while maintaining a low manufacturing cost and adhering to a product schedule.

As illustrated in the figure below, institutions and organizations drive technological innovations that commercialize through enterprises and entrepreneurs, benefiting rural users and creating rural prosperity, thereby inspiring further innovation—mirroring DFT's iterative cycle of continuous testing and optimization in AI accelerators.

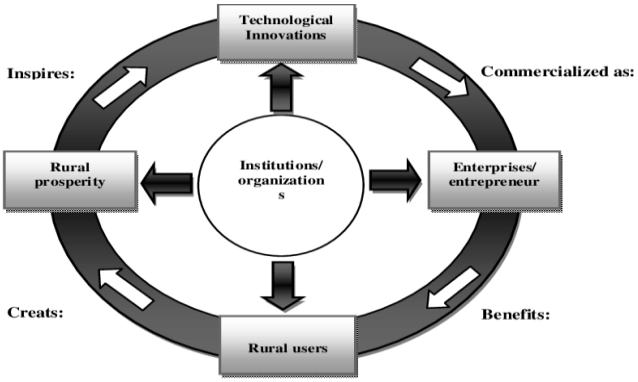


Figure 2: Local Technological Innovation cycle

3.1 Architectural complexity

The current AI accelerators are based on the principle of massive parallelism, which involves thousands of compute cores and deep memory hierarchies to perform matrix multiplication and neural network tasks. SRAM and external DRAM controllers need to transfer terabytes of data per second, and hundreds of voltage and frequency domains are independent to achieve dynamic power and performance requirements. It makes controllability and observability complex from a DFT viewpoint. Scan chains should be able to cross over several asynchronous clock areas and power islands without excessive routing overhead and over-timing.

It is observed that traditional single-domain test assumptions are violated, and designers must implement hierarchical scan insertion, multi-clock synchronization, and advanced pattern sequencing [14]. Large memories should also have an internal self-test (MBIST) capable of testing defects in large address spaces and different retention conditions, with minimal additional silicon area consumption and with the DFT logic itself testable. Proper scheduling of DFTs and patterns of tests is necessary to control timing, minimize cross-clock domain conflicts, and ensure efficiency in very heterogeneous accelerator designs, just as efficient notification scheduling has been used to improve responsiveness and performance in other applications, such as healthcare, by scheduling complex dependencies [29].

As illustrated in the table below, AI accelerators overcome DFT challenges—massive parallelism, multi-domain complexity, and large memory—using hierarchical scan insertion, MBIST, cross-domain synchronization, and optimized scheduling, achieving reliable fault coverage and efficiency while minimizing silicon overhead and timing conflicts.

Table 1: DFT Challenges and Solutions in AI Accelerator Architectural Complexity

Aspect	Description	Techniques/Methods	Benefits
Massive Parallelism	Thousands of compute cores and deep memory hierarchies create heavy controllability and observability demands.	Hierarchical scan insertion, multi-clock synchronisation	Ensures test coverage without excessive routing or timing issues.
Data Transfer Demands	SRAM and DRAM controllers move terabytes per second, requiring robust DFT integration.	Advanced pattern sequencing, scan crossing across clock/power domains	Maintains observability across heterogeneous subsystems.
Multiple Voltage/Frequency Domains	Independent domains complicate testing across asynchronous regions.	Cross-domain scan chains, synchronisation logic	Reduces timing conflicts and enables reliable multi-domain testing.

Aspect	Description	Techniques/Methods	Benefits
Large Memory Structures	High-capacity memories require efficient defect detection.	Memory Built-In Self-Test (MBIST), retention testing	Detects memory-related faults with minimal area overhead.
DFT Logic Overhead	Added test circuitry must also be validated.	Self-test of DFT logic, efficient area allocation	Ensures reliability without excessive silicon cost.
Scheduling Complexity	Multiple test patterns must be managed efficiently across diverse components.	Test scheduling, dependency-based sequencing (inspired by healthcare notification scheduling)	Optimises timing, reduces conflicts, and enhances test efficiency.

3.2 Advanced technology nodes

The transition to processes based on 5 nm and 3 nm process technologies, along with the introduction of gate-all-around (GAA) transistors, has altered the nature of failure modes that are more effectively addressed through standard failure models. Fin and nanosheet sizes may be varied, which may result in minor leakage or delay faults. Meanwhile, the chiplet-based designs and 3D integration stack several dies through-silicon vias (TSVs) or micro-bumps. These vertical interconnections cause further stress and novel defects, such as voids, cracks, and thermal-induced openings, which would be difficult to identify using conventional stuck-at or transition tests. Signal integrity is further challenged by high aspect ratio vias and dense interposer routing, necessitating special interconnect tests and monitors for DFT. Additionally, the mechanical and thermal contacts of stacked dies imply that failures may develop with time. Thus, there is a need to develop speed- and thermally conscious test solutions that remain efficient after packaging.

3.3 PPA limitations and time constraint.

AI accelerators must be able to deliver high performance while minimizing power consumption and silicon area. Any further addition of scan cells, observation points, or other BIST controllers will take up area and potentially introduce timing overhead. There is always a balance between high coverage and a low impact on power, performance, and cost (PPA). The speed of operation would also constrain the allowance for adding additional logic to critical paths, and consequently, at-speed testing is also challenging.

Besides that, there is commercial pressure to have shorter design cycles and a quicker time to market. DFT planning should thus begin early in the RTL stage and maintain close contact with physical design to prevent late rework in DFT, which can be costly in terms of delaying tape-out. The optimization of any iterative test pattern needs to be able to fit within compressed schedules, and pattern compression methods need to be chosen wisely so that they do not add enormous development burdens to the tester's memory constraints. The problems are similar to the fault-tolerant design concepts employed in event-driven systems, focusing on resilience and active risk mitigation to ensure that performance is not affected by stress [6]. They further compare the approaches towards scalability and cost in microservices architectures, where the need to balance between infinite scalability and financial limits requires early planning and ongoing optimization to achieve both technical and business goals [5].

3.4 Defect mechanisms

The higher the density of the integration, the higher the probability of physical defects other than the traditional transistor-level failures. Interconnect and via defects, e.g. voids, openings created by electro migration, or shorts, increase with a smaller metal pitch and greater current densities. The 3D stacked designs, especially TSVs and micro-bumps, are prone to mechanical stress, thermal cycle and contamination, which may cause latent defects that cannot be detected using conventional wafer sort. Reliability in the long-term is also an issue: effects of ageing, dielectric breakdown over time and migration due to stress may produce in-field failures months or years after manufacture. DFT should not only be able to identify manufacturing defects, but also support continuous reliability testing, as well as in-system testing, which can be used to detect faults that may develop slowly. To achieve this, experts will need sophisticated fault models, specific stress patterns, and potentially on-chip sensors that can generate early warnings without incurring prohibitive overhead [4].

The techniques required to treat these changing defect mechanisms reflect issues in other areas of technology. As an illustration, best practices in containerization using Docker and Kubernetes require precise

planning and continuous observability to isolate malfunctions and maintain operational stability during peak scaling pressures [18]. Similarly, the scalable SaaS implementation governance of the enterprise operation reflects the scale of systematic control and responsive control required to maintain the same performance and reliability as the system expands and advances [32]. Overall, AI accelerators and compute chips unite the architectural scale, new materials, and aggressive schedules that challenge the longstanding DFT methodologies. These obstacles necessitate early design, adaptive fault modelling, and innovative access and compression methods that can ensure quality without compromising the severe limitations of next-generation silicon.

As illustrated in the figure below, void growth is accelerated by rising current density and temperature, leading to narrowed conduction paths, increased electro migration flux, and temperature-driven diffusion, emphasizing the need for sophisticated DFT fault models and proactive in-system reliability monitoring

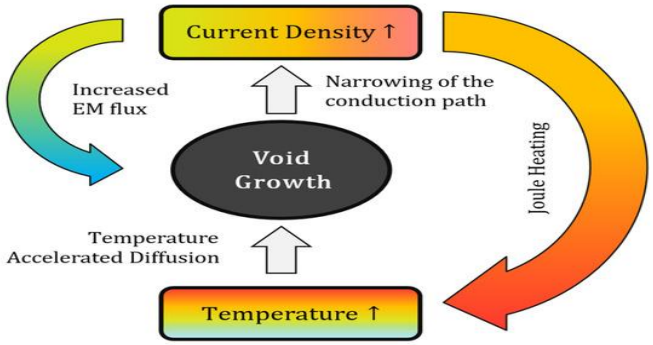


Figure 3: Current density–temperature–EM flux interdependence accelerates void growth.

4. Key DFT Techniques and Strategies

In order to obtain a high test coverage of AI accelerators and compute chips, design teams use a mixture of established and modern Design for Test (DFT) methods. These techniques are cooperative in revealing cooperative structural and timing defects, managing multiple clock and power domains, and ensuring spaces that the chip remains testable as density and performance needs grow.

4.1 Scan-based testing

The basis of the logic test has been scan-based testing—a full-scan design. The bosses connect to chains; each flip-flop is turned into a scan cell, and it can be easily controlled and monitored at test time by observing internal nodes. Partial insertion is applied when complete insertion would incur too much area and timing overhead, concentrating on risky or poorly controlled aspects. In contrast, other low-risk aspects remain in their normal state. Multicore asynchronous clock domain accelerators of AI must be synchronised carefully. Lock-up latches, multiple capture clocks, and crossing isolation over clock domain techniques are used to ensure that capture and shifting occur without timing violations. Real-world faults that only manifest in real operational conditions, such as transition and path delay faults, can only be detected with at-speed testing, which launches and captures signals at actual operating speeds. Pattern compression can be incorporated into the scan flow and is used to ensure that the volume of data remains within manageable bounds, while also reducing the time taken by the tester without compromising coverage.

4.2 Built-in self-test (BIST)

BIST incorporates a special test logic on the chip, which provides patterns and responses analysis without excessive reliance on external testers. AI designs are critical in two aspects. Memory BIST (MBIST) is used on large SRAM arrays and embedded DRAM blocks, which occupy the majority of the area in neural network accelerators. MBIST controllers are capable of using a set of algorithms, March tests, coupling tests, and retention checks to identify stuck-at, transition, and soft errors, and also repair redundancy to increase yield. The use of logic BIST (LBIST) is used when the digital block under consideration is complex and the random-pattern testing can be applied successfully. The conventional ATPG patterns are unlikely to scale. LBIST stresses the logic at-speed with on-chip pseudo-random pattern generators and signature analysers, which have proven mostly expensive and challenging to do with other stress testing methods, especially with high-frequency AI cores. Phase shifting and careful seeding are used to overcome any random-pattern resistance and also to ensure that the critical faults are exercised. Dynamic control of test strategies and the ability to cope with a variety of fault conditions in situ are indicative of a dynamic memory inference network, where flexible, adaptable architectures quickly and efficiently deal with variable data and lines of reasoning [26].

4.3 Boundary scan and standards

Boundary scan is a standardised access to chip I/Os and embedded cores, and makes board- and system-level testing easy. IEEE 1149.1 (JTAG) is the standard that specifies the fundamental architecture of boundary scan cells and the Test Access Port (TAP) controller, allowing serial access to on-chip pins and logic. IEEE 1500 builds on these principles to embedded cores where wrappers and control protocols render each block of IP independently testable and reusable. IEEE 1687 (IJTAG) creates a high-level, flexible network of test tools in the chip, enabling plug-and-play sensors, monitors, and core-level FT capabilities. These standards enable consistent test access and simple retargeting of patterns between block level and the entire system in AI and compute chips that combine a wide variety of IP, sourced from multiple different sources [16].

These standards offer flexibility and end-to-end control that reflect best practices in other aspects of technology. An example involving AI-enhanced security and inventory optimisation involves constant automation to preserve integrity and accommodate changing needs in CI/CD-based retail systems, ensuring that vulnerabilities and resource needs are met in a timely manner [21]. Likewise, orchestrated resource management and modular scaling are concepts relevant to cloud cost optimization and sustainability in Kubernetes environments, ideas familiar to hierarchical and reusable IEEE 1500 and IJTAG-based test networks [25]. In addition, the safe real-time exchange of data between heterogeneous platforms, as displayed in the Internet of Things and healthcare integration with Aerospike and Salesforce Marketing Cloud, is similar to the necessity of reliable, standardized interfaces in integrating test information and instructions across various IP sources [9]. With the integration of these standards in the initial design phase, manufacturers are now able to design DFT architectures that are flexible, secure and inexpensive, despite increasing system complexity and third-party IP integration.

As illustrated in the figure below, boundary scan architecture incorporates a TAP controller, instruction and boundary registers, and serial data paths (TDI/TDO), enabling standardized, hierarchical test access defined by IEEE 1149.1, IEEE 1500, and IEEE 1687 for complex AI and compute chips

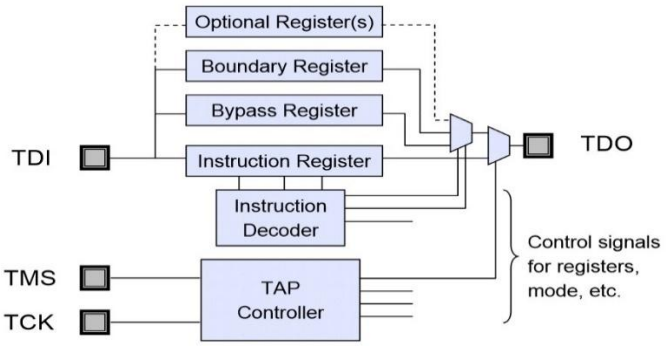


Figure 4: Boundary-Scan Standard and the associated Test Access Port

4.4 DFT of chiplets and 3D.

Several of the latest high-performance processors are based on multi-die architectures and chiplet designs, connected by interposers, micro-bumps, or through-silicon vias (TSVs). These interconnects require specialised testing strategies. Pre-bond testing is a test that verifies individual dies in probe-accessible DFT structures like partial scan chains and loopback links. Post-bond tests are aimed at TSV continuity, micro-bump integrity and interposer routing and frequently with built-in interconnect BIST or die-to-die boundary scan chains. IEEE 1838 and improvements on IEEE 1687 give recommendations on how to structure such multi-die test networks. Early package-level routing is necessary in order to allow TSV test chains and interconnect monitors to be installed without any signal integrity or power distribution impact.

4.5 Power- and thermal-aware DFT

AI accelerators are run close to power and thermal boundaries, and therefore, test-induced stress becomes a significant issue. Large areas can be switched concurrently to produce excessive IR drop, local hotspots, or even irreversible damage, due to test patterns. Power- and thermal-aware DFT can overcome these risks by utilising patterns, sequencing tests to prevent activity simultaneously, and on-chip sensors to monitor activity in real-time. Power gating can isolate non-tested domains, whereas clock gating can reduce dynamic switching. Intelligent scheduling, as well as dynamic scaling of voltage during test, balances coverage and stress. These methods safeguard yield and reliability, and permit in-depth detection of delay and bridging faults, which require high-speed activity.

These strategies form a stratified DFT approach that is highly applicable to the complexity of AI and computing silicon. Scan-based testing offers universal logic visibility, BIST provides autonomy and speed, boundary scan and standards facilitate system-level integration, and chiplet-specific techniques offer reliability across multiple dies via safety and cost-effective testing in both manufacturing and in-field environments. Hotspots or test scheduling optimisation through predictive analytics is similar to the methods applied to fuel business intelligence and develop DevOps efficiency, where forecasting via data can be used to address risk and become more responsive [20]. Similarly, methods for constantly tracking and responding to physical conditions are similar to innovations in the field of telematics, which manage fleets. Real-time tracking and communication of assets enhance efficiency and reliability in a dynamic working environment [23].

5. Methodology for Optimizing DFT Test Coverage

This section provides a workable, step-by-step methodology for achieving high and cost-effective Design for Test (DFT) coverage of AI accelerators and compute chips. The methodology combines research results with best engineering practices, ensuring that DFT objectives align with the complexity of silicon, production cycles, and long-term durability requirements. 5.1 Architecture and early planning. Effective DFT must be designed at the outset, not as an afterthought. DFT Co-optimisation with RTL development and floor planning. Co-optimization of DFT with RTL development and floor planning allows test structures to be integrated naturally without interrupting timing or area budgets. At this point, the targets of coverage should be established according to the reliability and quality needs of the product, including the target defect parts per million (DPPM). These targets are usually highly aggressive in the case of AI accelerators, which commonly serve data centre or automotive applications. Sign-off metrics used in DFT, including minimum scan coverage, memory BIST coverage, and boundary scan completeness, should be included in important design milestones along with functional verification and timing closure. Premature design also makes the boundaries of clock and power domains clear, in which the location of lock-up latches, power gating cells, and scan compression logic are determined so that testability can be stable as the design matures.

As illustrated in the figure below, Synopsys SystemVerilog Catalyst services—DFT Ultra, DFT Compiler, DFT Max, and Tetra Max—provide integrated tool support for early DFT planning and optimization, enabling high coverage, cost-effective testing, and seamless co-optimization with RTL and floor planning.

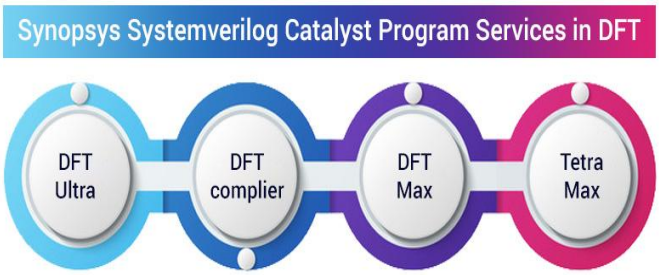


Figure 5: solutions-for-optimal-dft-design-for-testability-in-lower-technology-node

5.2 Fault modelling and pattern generation.

Adequate test coverage is based on accurate fault modelling. Although conventional stuck-at and transition models are still necessary, AI-specific models offer more powerful alternatives that include bridging faults, cell-aware faults, and path delay faults. These include imperceptible manufacturing flaws and timing-related problems that may arise within nanoscale nodes or extremely parallel data-paths. It is then possible to use advanced Automatic Test Pattern Generation (ATPG) tools that generate patterns to address these fault models, as well as using compression methods to minimize the number of vectors to be stored on the tester and to minimise the total test time.

More recent developments include machine-learning-assisted ATPG, where predictive code is used to identify areas of high risk based on layout characteristics, past defects, or silicon feedback. Such a narrow attention makes the process of producing patterns more efficient and allows for preserving coverage even though the size of the design increases. It is becoming increasingly common across other branches of AI to combine predictive and data-driven analysis with pattern generation, a trend that follows similar developments in other fields, such as the use of large language models to generate more accurate and contextual descriptions of images [30]. Equally, the ability of deep learning to respond to complex questions about pictures based on natural language evidence

illustrates how multi-modal reasoning can be used to reduce the choice of functional patterns and maximize test coverage in complex designs [31].

As illustrated in the table below, fault modelling and pattern generation integrate conventional models, AI-specific faults, advanced and machine-learning-assisted ATPG, and data-driven AI trends, enabling adaptive testing, reduced test time, and comprehensive coverage for increasingly complex semiconductor designs.

Table 2: Fault Modelling and Pattern Generation Approaches

Aspect	Description	Examples/Methods	Benefits
Conventional Models	Traditional fault models still used for baseline coverage.	Stuck-at faults, Transition faults	Provide essential foundation for test coverage.
AI-Specific Models	Capture subtle defects and complex timing issues in nanoscale and parallel systems.	Bridging faults, Cell-aware faults, Path delay faults	Address imperceptible manufacturing flaws and timing-related problems.
Advanced ATPG Tools	Generate test patterns targeting both conventional and AI-specific fault models.	Pattern generation with compression techniques	Reduce number of vectors stored; minimize test time.
Machine-Learning-Assisted ATPG	Predictive models highlight high-risk design areas.	Layout analysis, historical defect data, silicon feedback	Improves efficiency and maintains high coverage in larger designs.
Integration with AI Trends	AI methods enhance test strategies by learning from data patterns.	Predictive and data-driven analysis, multi-modal reasoning	Enables adaptive testing aligned with complexity growth.
Parallel in Other AI Fields	Similar techniques applied beyond hardware testing.	Large language models, deep learning for image and language reasoning	Shows potential for narrowing functional patterns and maximizing coverage.

5.3 Hierarchical and IP-level DFT.

Tensor cores, systolic arrays, and high-bandwidth memory controllers are examples of IP blocks that are combined in large AI accelerators [22]. Reusable DFT wrappers can be used to encapsulate all these IP blocks so that each can be individually tested and then integrated effortlessly at the top. Integration standards, such as IEEE 1687 (IJTAG), make this process easy since they outline an adaptable network that links embedded devices and core wrappers to the tester outside. The hierarchical design has the least amount of test time, allows IP reuse to scale across multiple projects, and enables top-level pattern regeneration without forcing a re-processing of a delayed change to a particular block.

5.4 Chiplet and 3D DFT integration.

As chiplet designs and 3D-stacked designs become the trend in the field of high-performance computing, DFT should not be limited to the die boundary. In the initial package design, designers are advised to design special TSV and interconnect test chains as well as insert boundary scan cells that can be accessed after assembling the package. Inter-die loopback connections and internal self-test circuits allow at-speed testing of die-to-die links to verify that defects in the form of micro-bump openings, shorts or thermal stress cracks are identified. Adherence to standards, such as IEEE 1838, or the use of extensions of IEEE 1687, offers a simple framework upon which such multi-die test requirements can be managed, facilitating post-bond and pre-bond tests.

5.5 Analysis of validity and coverage.

After the DFT logic and test patterns are implemented, extensive validation is carried out to verify that the intended coverage is attained. The tools of coverage grading detect untestable logic, helping to add control or observation points and refine ATPG patterns. This cycle has to be finished prior to the tape-out to prevent the expensive silicon re-spins. At this stage, power and thermal simulation is also necessary to ensure that running of

tests will not result in an IR drop or hot spots that may diminish yield or cause damage to the device. Comparison of the simulation with prototype silicon data during bring-up further provides confidence.

5.6 Pilot and volume ramp

The final stage involves linking design-time test coverage to the actual manufacturing outcome. In the early silicon bring-up phase, ATPG coverage vs wafer probe yield and packaged part yield is used to ensure the test program is effective at detecting real defects. Feedback on this stage will lead to minor modifications of patterns and test limits prior to full production. As production increases to high volume, continued monitoring of defect data enables continuous optimisation of screening patterns and adaptive control of test limits, thereby maintaining high yield and reliability while minimizing tester time loss. Such a systematic approach, beginning with an initial design phase, rich fault modelling, hierarchical and chiplet-sensitive DFT, and feedback on coverage assurance and manufacturing, is what enables AI accelerators and compute chips to achieve very high-quality objectives. It provides high coverage without compromising PPA budgets or schedules, and it lays a basis for continued improvement as process technologies, as well as AI workloads, change.

6. Balancing Coverage, Cost, and Time

Design for Test (DFT) test coverage optimisation in AI accelerators and compute chips is both a technical issue and a business decision to make. The key objective is to ensure that all manufactured devices are of high quality, with the cost of production, silicon overhead and schedule impact being controlled. Since these items can be delivered in large volumes and serve as mission-critical applications, the tradeoff between cost and coverage should be planned and continuously maintained. This balance is only possible through cross-functional collaboration between design, DFT, and manufacturing teams, in a manner that allows the engineering tradeoffs to benefit the overall product economics. The most apparent is the tradeoff between the volume of patterns on test and tester resources. To achieve high fault coverage, ATPG patterns are typically needed in large amounts in order to identify stuck-at, transition, bridging, and cell-aware faults. Patterns which demand more memory in the tester and more time to apply also demand more expensive equipment to test, and can cause bottlenecks on high-priced automated test equipment. In large-scale AI chips, any minor increment in test time will translate to millions of dollars of operating cost per year. This problem reverberates the computational and storage tradeoffs faced when comparing individual image captioning algorithms in large-scale applications [33]. Pattern compression methods, such as Embedded Deterministic Test (EDT) and more modern methods like X-masking, have been used to alleviate this burden by minimising the number of bits that need to be stored and shifted when a test is executed. Compression permits very high coverage with a smaller number of tester cycles but has its costs: it obligates the introduction of decompressor logic on chip, and the insertion must be very careful to eliminate timing or power problems. The choice between silicon area and routing overhead versus tester time savings will require a decision on a particular compression ratio.

The other important decision is whether to conduct at-speed testing and slower functional testing. Delays, faults, and timing defects can only be detected at at-speed, and this is especially important with AI accelerators when operating at multi-gigahertz frequencies. However, with high frequency, these switching patterns activate more nodes at once, and increase dynamic power consumption, as well as the probability of local heat or IR-drop. They are also able to extend the total test time by requiring a few extra setup and capture cycles. Slower tests minimise the silicon stress and save on time, but can overlook the edge cases. Each product shall thus be a tradeoff between the coverage at full speed and the real constraints of cost, power, and tester throughput [31]. Tradeoffs are caused by coverage targets themselves, which are typically expressed in terms of defects per million (DPM). The parts in safety-critical or data-centers might be so safety-critical or data-driven that the highest levels of DPPM are needed, and they thus demand more rigorous testing and longer tester time. A slightly higher DPPM may be welcomed in consumer electronics or cost-sensitive applications to reduce costs and accelerate ramp-up. These choices are business-based decisions in nature, although they have to be echoed in DFT planning at the earliest phases to prevent surprises at the end of the development.

As illustrated in the figure below, SoC compiler workflows integrate RTL, design specifications, and third-party ATPG to insert optimal test points, enabling tradeoff analysis between coverage and silicon area to achieve high DFT coverage with minimal overhead and balanced production economics.

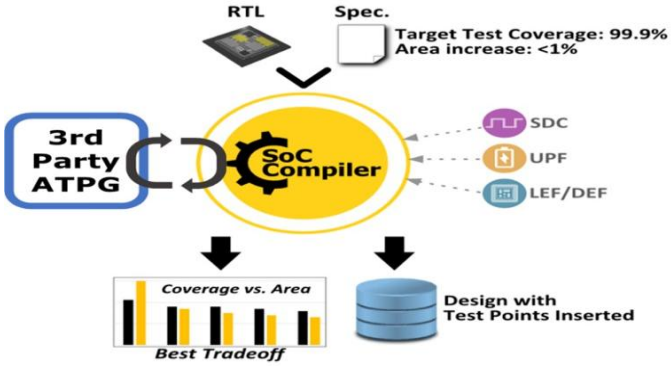


Figure 6: Lowering-the-dft-cost-for-large-socs-with-a-novel-test-point-exploration-implementation-methodology

6.2 Design impact

The overhead of adding the capabilities of DFT is inescapable. Scan chains, observation points, control logic and hard-built-in self-test (BIST) circuitry claim silicon area and may cause additional capacitance on important nets. This may result in routing congestion, deteriorated leakage or slower maximum clock frequency unless carefully planned early. In the case of AI accelerators that base their sales approach on the idea of performance per watt, the overhead can be a significant factor.

Design teams overcome these effects by using partial scan and selective observation techniques that target the most at-risk logic, without scanning the less risky areas. Hierarchical DFT also localises the test structures such that each block or chiplet can be tested in isolation and does not overload the entire system. When using power-sensitive techniques like domain-based power gating and clock gating at test, dynamic switching activity is minimized and IR-drop is capped so that test modes can grow at safe operating levels. These methods enable adequate test coverage with minimal impact on power and timing budgets. Floor planning and early sign-off metrics are extremely important [34]. Teams can incorporate DFT into the routing process of scan chains by locating compression logic off the critical paths and verifying power assignment during worst-case test conditions. This preemptive measure keeps the overheads of areas and time within reasonable limits, which do not affect product specifications, market competitiveness, or tape-out schedules. The tradeoff between coverage, cost, and time of AI accelerators and compute chips is a multi-dimensional process. It requires early design and clever pattern control, energy, and place-aware design methods to fulfil business needs and objectives without compromising quality.

7. Case Studies and Best Practices

Practical implementations of large compute processors and AI accelerators demonstrate that appropriate DFT planning has a direct positive impact on yield, reliability, and cost-effectiveness. The detailed illustrations below illustrate how businesses have utilised hierarchical DFT, memory BIST, boundary scan, and advanced interconnect testing to achieve high success within the production process. They also highlight important practices that demonstrate it is a proactive design practice, rather than a late-fix discipline.

7.1 High-volume AI accelerator

An example of a state-of-the-art AI accelerator targeted at data center inference demonstrates that hierarchical DFT can be used to achieve huge designs. This chip had thousands of compute units and multi-megabank SRAM memory to support matrix multiplication and deep-learning workloads. Since the beginning of the RTL phase, DFT was regarded by the design team as a central aspect of architecture. Every major compute block (ten cores, vector engines, and specialized memory subsystems) was packaged with its own scan and memory BIST (MBIST) logic, which is similar to the focus on early, feedback-driven iteration that facilitates successful growth in design-oriented fields, including AI-powered career coaching [15].

With reusable DFT wrappers built around these blocks, engineers would be able to develop and test patterns at the block level, way before the entire system was built. This enabled the independent generation of patterns and reuse, making it easy to update or configure subsequent versions of the product in the event of a block update. It further eased timing closure by isolating scan chains and test control signals within each block, and minimized routing congestion on the top level. Another important aspect was the aggressive compression of test patterns. Embedded Deterministic Test (EDT) and X-masking decreased the number of patterns which needed to be stored on the tester. This reduced the memory requirements of testers and shortened the test application time

without compromising coverage. Because each block was independently exercisable, pattern generation was performed in parallel by several teams, making the schedule shorter.

This hierarchical flow of DFT was used during production, and block-level MBIST and system-level scan chains were used to provide close stuck-at and transition fault coverage. At-speed testing proved that logic paths that were delay sensitive under real operating conditions worked correctly. Consequently, there was a quantifiable increase in the first-pass yield of early silicon bring-up over earlier designs of comparable scale. The reduced retest rate and fewer minutes per device taken to perform the test had a direct impact on decreasing manufacturing costs and creating a cost-efficient ramp to high-volume production. This case illustrates that by incorporating DFT at the architectural level, as opposed to the physical design level, it is possible to achieve a higher level of coverage, as well as less time to market.

As illustrated in the figure below, an AI accelerator ecosystem integrates project launch, specialized data types, IP libraries, toolkits, system integration, and Catapult HLS, supporting hierarchical DFT planning that enables early block-level testing, efficient pattern reuse, and accelerated high-volume production.

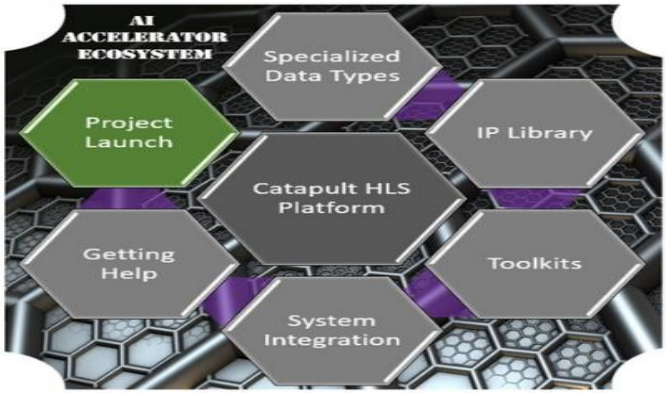


Figure 7: providing-an-ai-accelerator-ecosystem

7.2 Chiplet-based compute processor.

A second case study is based on a multi-die compute processor that places logic chiplets, memory chiplets, and I/O chiplets on a silicon interposer. This architecture reduced the cost of manufacturing and enhanced yield by enabling the fabrication of individual chiplets at the optimal manufacturing node. Nonetheless, it gave rise to a new category of DFT problems based on the thousands of micro-bumps and through-silicon via (TSV) interconnects that attach the chiplets into one package. The design team realised early on that the interconnect reliability was to play a vital role in the quality of the products in the long run. They used IEEE 1838 guidelines for 3D DFT and IEEE 1687 networks to connect embedded monitors at various levels of the stack. Inter-die loopback paths and boundaries scan cells were put strategically at the die interfaces so that the electrical test could be thoroughly done at each assembly stage. Each chiplet was tested post-bond through pre-bond tests, and interposer routing and TSV continuity and micro-bump integrity were verified through post-bond tests.

Along with structural testing, self-test circuits were built into test die-to-die connections in early silicon bring-up and burn-in [27]. These on-chip tests ran at full pace and realistic thermal loads, also revealing the presence of minor defects like micro-voids, stress-induced opens and contamination-related shorts that would otherwise not have been detected. Through repeated testing of interconnects before bonding, after bonding, and in system-level testing, the team successfully reduced the latent field failures and increased package-level reliability. This regimented, multi-stage approach enabled the product to be subjected to high data centre reliability standards without being overworked or delayed on schedule. It also showed the effectiveness of using boundary scan, loopback structures and BIST together with industry standards in ensuring that even complex assemblies of chiplets can be tested and maintained throughout their life cycle.

As illustrated in the table below, chiplet-based DFT combines IEEE 1838/1687 standards, inter-die loopbacks, pre- and post-bond testing, and built-in self-tests, ensuring reliable interconnects, reduced latent failures, and high data-center-grade reliability without delaying production schedules.

Table 3: DFT in Chiplet-Based Compute Processor

Focus Area	Description / Approach	Outcome / Benefit
Architecture	Multi-die processor with logic, memory, and I/O chiplets on silicon interposer.	Reduced manufacturing cost and improved yield.
Main DFT challenge	Reliability issues from thousands of micro-bumps and TSV interconnects.	Required new test strategies for inter-die connectivity.
Standards applied	IEEE 1838 for 3D DFT and IEEE 1687 for embedded monitor networks.	Enabled structured monitoring and thorough electrical test coverage.
Testing strategy	Inter-die loopback paths and boundary scan cells placed at die interfaces.	Allowed thorough testing at each assembly stage.
Pre-bond & post-bond testing	Pre-bond chiplet tests; post-bond verification of interposer routing, TSV continuity, and micro-bump integrity.	Ensured connection quality throughout assembly.
Self-test circuits (BIST)	Built-in self-tests during bring-up and burn-in, run at full speed and realistic thermal loads.	Detected micro-voids, stress-induced opens, and contamination-related shorts.
Multi-stage testing	Repeated testing before bonding, after bonding, and at system level.	Reduced latent field failures and improved long-term reliability.
Overall benefit	Regimented approach combining boundary scan, loopback, BIST, and standards.	Achieved high data centre reliability standards without delays in product schedule.

7.3 Key takeaways

As both case studies emphasise, effective DFT is an inter-functional project that must involve close Co-ordination among architecture, logic design, physical implementation, and product engineering units. Early co-optimisation ensures the capture of test requirements as basic design constraints, as opposed to being inserted late into the flow when they are expensive and disruptive. Notable enablers of scalability and reuse are hierarchical DFT and modular BIST. The teams can integrate, update and retarget test patterns effectively by wrapping IP blocks or chiplets with their own test logic and by following standards such as IEEE 1687 and IEEE 1838. The modular model is beneficial in AI accelerators and compute processors, which will undergo numerous revisions or derivative models. Of equal consideration is the feedback that exists between the data of production and DFT refinement. Online control of yield and defect trends guides changes in ATPG patterns, the choice of fault models, and revisions to test schedules. This is a dynamic process that enables the test program to keep pace with the manufacturing process, achieving high yield and reliability despite process drift or the emergence of new failure modes.

These practices, combined, indicate that DFT planning is not a purely technical process, but a strategic investment. This can be accomplished through early participation, top-down strategies, access via standards, and continuous data enhancement: AI accelerators and compute chips can run high test coverage without compromising cost, performance or time to market. Such lessons offer guidance on future projects that will need to fulfil the twin requirements of speedy innovation and unyielding product quality.

8. Future Trends

The high-paced development of semiconductor production and critical applications, which rely on the centrality of artificial intelligence, are changing how Design for Test (DFT) needs to be developed. The next-generation AI accelerators and compute processors will bring forward beyond the old-fashioned one-time testing of factories to the constantly evolving intelligent lifecycle management. Due to the increasing complexity of devices, it will be necessary to develop new strategies to ensure high yield, long-term reliability, and manage cost.

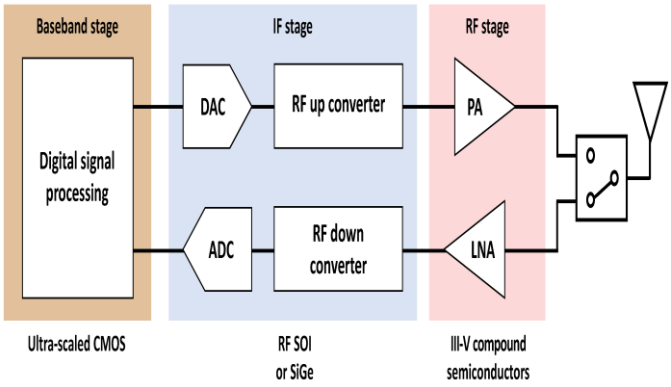
8.1 Heterogeneous integration and wafer-to-wafer stacking.

The single, monolithic dies are being replaced by heterogeneous integration in the industry. This model utilises logic, memory, analogue interfaces, and specialised AI accelerators, which are optimised for each process node to perform their specific role. These components can then be stacked in 3D or bonded together on the wafer. This architecture reduces signal lengths and offers improved bandwidth, enabling very high data rates required for large AI workloads. New test challenges accompany such advantages, however. Deeply buried internal layers in a 3D stack are inaccessible to direct probing, and failures in through-silicon vias (TSVs), micro-bumps or interposer routing can only be seen when the stack is under stress, e.g. temperature cycling or mechanical load. The traditional wafer probe techniques can no longer be considered adequate. DFT must then be further advanced in the future with additional access techniques.

A basis is given by distributed test networks using IEEE 1687 and the new IEEE 1838 standard [13]. These enable test data and instructions to access embedded instruments in every die or wafer layer, even after additional bonding. TSVs and interfaces with micro-bumps can incorporate built-in monitors that continuously assess the integrity of the electrical connection and monitor degradation over time. The test flows will be required to consist of several steps: pre-bond tests to check every die prior to assembly, mid-bond tests during assembly to eliminate early failures and post-bond tests to confirm the correctness of the assembled part.

Designers will design these multi-level test hooks more and more early in the floor planning and package design. Unnecessary interconnections and loopbacks will be strategically located in such a way that one failed via or bump does not mean loss to any product. The mechanical and thermal stresses would be considered early in layout and would minimise the chances of latent defects that would only be discovered after field operation. The difficulty in contending with these multi-layered test and integration flows is comparable to the problems of consolidating multiple ERP systems during large-scale data migration, where orchestration and redundancy are the primary factors that preserve data integrity and operational stability [2]. On the same note, the success of large language models in doing specialised ecommerce search problems via fine-tuning, and the notion that successive optimisation and built-in feedback can constantly enhance performance and reliability, are also applicable to DFT strategy improvement as stack density and interconnect complexity increase [28]. With this form of access and redundancy, manufacturers can achieve high yields and reliability and as stacking elements continue to increase, so does the interconnection density.

As illustrated in the figure below, heterogeneous integration combines ultra-scaled CMOS baseband processing, RF SOI/SiGe intermediate stages, and III-V compound semiconductor RF stages, enabling 3D stacking and wafer bonding that demand advanced IEEE 1687/1838 test access for reliable multi-die DFT.



8.2 In-field and online DFT

With the deployment of AI accelerators in systems with high stakes, such as autonomous vehicles and aerospace control, as well as in hyper scale cloud data centres, it is vital to have continuous, fault-tolerant operation [24]. This necessity is forcing DFT to move beyond its established role as a filter used only once during product production into a continuous health application throughout the product's lifetime. The next-generation DFT will incorporate on-chip sensors and controllers that will periodically perform self-tests while the system is in operation. The defects that these checks can identify include time-dependent dielectric breakdown or electro migration in interconnects, as well as degradation of memory cells that develops gradually. Upon the detection of an abnormality, inbuilt mechanisms like activating spare rows, error-correction adjustments, and automatic re-swerving of interconnect routes can also be activated.

Voltage, current, temperature, signal integrity monitors, and controls in real-time will indicate conditions such as sudden IR drops or local hotspots in a system, long before it encounters functional failure. This data may

be recorded and processed to plan a proactive maintenance, implement a mitigation based on software, or gracefully reduce performance in a manner that is not unexpected, unlike an unexpected outage. This development makes DFT a reliability management embedded. Instead of merely demonstrating that a chip is fault-free upon shipment, the test infrastructure becomes a proactive and permanent protection, guaranteeing a long service life and predictable performance. The necessity of organized control and active policy control in this respect is the role of data governance in reinforcing cooperation between ERP and master data management (MDM), where continuous monitoring and active control are essential to ensuring data integrity and compliance over a period [3]. This type of in-field and online DFT will be essential where a high level of service or safety is required, as even a momentary outage or unnoticed degradation can be severe [18].

8.3 AI-driven DFT automation

Machine learning and AI are also starting to transform the very process of the test. AI is capable of searching through large volumes of data and identifying subtle relationships that enable human engineers to overlook design layouts, process parameters, and historical yield or defect data. Machine-learning algorithms can be utilised during the design phase to identify high-risk areas within the machine, such as dense interconnects, complex clocking, or challenging lithography, and flag them for further control or observation. This helps direct DFT funds to where they will have the most significant impact. In automatic test pattern generation (ATPG), AI can be used to guide the creation of patterns, focusing on probable failure modes to reduce the total pattern count and tester memory footprint, with no harm and even increased fault coverage.

Real-time defect analytics can be implemented once production is initiated [1]. Test data of wafers and packages are processed in real time, and the output is used to modify pattern sets dynamically. When a given defect signature is identified as increasing on a lot or wafer, the tester will know automatically to focus on the patterns of similar faults. The entire test procedure can be self-optimising over time, similar to predictive maintenance in industrial automation. The AI-based automation offers the following benefits: increased speed in DFT development, reduced test application time, and improved yield on the first pass. It is also capable of ongoing learning, thus every new generation of AI accelerator utilises the knowledge from previous designs and production batches. With the increasing complexity of chips and the increase in the number of potential defects, these intelligent techniques will be necessary in order to control test time and cost without compromising quality.

As illustrated in the table below, AI-driven DFT automation leverages machine learning for risk identification, AI-guided ATPG, real-time defect analytics, and continuous learning, enabling efficient resource allocation, reduced pattern count, adaptive testing, and continual improvement across design generations.

Table 4: AI-Driven DFT Automation

AI/ML Application	Description	Benefits
Risk identification during design phase	Machine learning identifies high-risk areas (dense interconnects, complex clocking, lithography) and flags them for review.	Directs DFT resources effectively, improves design reliability.
AI-guided ATPG (Automatic Test Pattern Generation)	AI focuses pattern creation on probable failure modes, reducing total pattern count and tester memory footprint.	Reduced test data size, faster testing, increased fault coverage.
Real-time defect analytics in production	Wafer and package test data processed in real time; patterns dynamically modified to respond to defect signatures.	Self-optimising test flow, predictive maintenance-like adaptability.
Ongoing learning across generations	AI accelerators leverage knowledge from past designs and production batches.	Continuous improvement, reduced cost and time with each generation.

9. Recommendations

Based on the discussion of the challenges, methodologies, and future trends, a set of recommendations can be made to ensure that the design and product teams obtain high DFT test coverage without incurring any cost, performance overruns, or schedule slippage [7]. These guidelines summarise best practices demonstrated in industry case studies and future research, providing practical recommendations for implementation.

Several recommendations should be adhered to to attain high and cost-effective DFT test coverage. To begin with, DFT should be considered during the conceptual design and not as an architectural addition or a post-design workload. The processes of co-optimization must start at the RTL and floorplanning phases so that scan chains, MBIST controllers, and boundary scan networks become organically part of the design. Early quantifiable coverage targets, e.g., stuck-at coverage, transition coverage, and cell-aware fault coverage or DPPM targets, must be specified at early project stages and expressed in project milestones, timing closure, and power sign-off. The early integration reduces the last-minute changes and eliminates timing or routing issues. Second, exploit current fault modeling. More modern AI accelerators cannot be based only on classical stuck-at and transition models; more elaborate models, such as bridging, cell-aware, and path delay faults, will be needed to characterize defects that are typical of 5 nm and 3 nm technology.

With these and machine-learning-based ATPG, high-risk regions can be detected, and specific pattern generation to improve defect detection can be done, as well as managing test data volumes and tester time. 3rd, use hierarchical and normative DFT. IP block and chiplets wrapping with reusable test wrappers and embedded instruments, chiplet-based architecture, and 3D standards IEEE 1687 and IEEE 1838, can be used to provide pattern reuse, pattern development in parallel, and lower ATPG runtime, as well as flexibility in late design changes or derivatives. Fourth, come up with an effective chiplet and integration process. Many die or wafer-to-wafer stacks also need pre-bond, mid-bond, and post-bond testing, such as TSV, inter-die loopback structures, and boundary scan cells, to provide excellent die-to-die and interposer connectivity. Before assembly, it is necessary to perform early-stage DFT planning to identify interconnect defects that enhance yield and reliability in the long term. Fifth, test control and thermal stress. DFT methods that are both power-aware and thermal-aware, such as power gating, clock gating, intelligent pattern scheduling, and on-chip thermal sensors and voltage monitors, can help keep IR drop and hotspot creation within a safe range, as well as dynamically adjust test sequences. Sixth, develop a continuous feedback loop. By comparing tester data to silicon yield in pilot production and volume ramp, fault models and patterns can be refined in preparation for adaptive test programs, and new defect mechanisms can be identified early. Be ready to have in-field and AI-assisted DFT. On-chip sensors and self-test programs can constantly check the health of the device and take self-repair measures, including switching on spare rows or dynamically rerouting interconnects. Meanwhile, with an investment in AI-driven analytics to keep orchestrating patterns and dynamically optimizing test coverage, future design cycles will be shortened and allow long-term reliability in mission-critical deployments.

10. Conclusion

Design for Test (DFT) has emerged as a core need for designing reliable and affordable AI accelerators and compute chips. With increasing semiconductor manufacturing processes of 5 nm, 3 nm and smaller, and with multi-die and chiplet designs, the possibility of some latent manufacturing flaws increases exponentially. To satisfy both the high-performance and long-term reliability requirements of today's market, it is now more critical to incorporate DFT as a central component of the design and manufacturing process and not as a supplementary or late design task. Sound DFT is based on early planning. Starting with RTL and floorplanning, it is possible to add scan chains, built-in self-test (BIST) structures, and boundary scan networks to the natural hierarchy of the design. This early beginning enables quality goals, such as defect parts per million (DPPM), to be adequately defined, and the test logic can be defined within area and timing constraints. It also assists in the placement that is power and thermal conscious, so that the test modes do not exceed the safe operating limits. Such planned designs prevent congestion at routing in the last minute, and minimize the possibility of schedule slippage or expensive silicon re-spins.

Pattern generation and comprehensive fault modelling are also important. The current defect models are not limited by the conventional stuck-at and transition defects, but cover bridging defects, path delay defects, and cell-conscious faults within the conventional cells. State-of-the-art automatic test pattern generation (ATPG) tools, combined with pattern compression, can effectively target these complex faults with manageable tester memory and test time. Machine-learning-assisted ATPG introduces an additional optimization step, where risky areas are predicted and pattern sets are automatically modified to be the most effective. Complex AI accelerators require hierarchical and IP-level DFT. Individual cores, memories, and accelerator engine reusable test interfaces are easily integrated and reused in patterns. Standards like IEEE 1687 enable accessing embedded instruments at the plug-and-play level, and IEEE 1838 provides a systematic framework for 3D-stacked or chiplet-based designs. These standards enable the coverage of multi-die systems, including through-silicon vias (TSV) and micro-bump, as well as other advanced interconnects that are otherwise inaccessible to current probe technology.

Another dimension of critical dimension is power and thermal management. At-speed testing provides a method of exposing timing-related faults as well as establishing a sudden burst of current and a local hotspot. To address this, power- and thermal-aware DFT has introduced domain-based power gating, clock gating during test,

and pattern sequencing to limit simultaneous switching. These ensure yield protection and prevent overstress in both manufacturing and field testing. In the future, DFT will continue to evolve in response to emerging trends. New multi-level test access architectures and inherent redundancy will be necessary to maintain yield with heterogeneous integration and the further stacking of wafers. In-field and online DFT will transform into a built-in reliability checker, operating periodically through self-tests and deploying self-repair systems to facilitate the long service life of mission-critical systems, such as cloud data centres and autonomous vehicles. During this time, AI-based DFT automation will involve real-time defect analytics and adaptive pattern generation to decrease development cycles, reduce tester time, and increase incremental first-pass yield.

Ensuring that these practices, as well as innovations, produce a holistic road to high-yield, long-life AI compute products. Next-generation AI accelerators and computer chips can be designed to achieve aggressive performance and cost targets while still providing the robustness and reliability required by their high-performance applications. This can be achieved by implementing DFT as a key component of the design and manufacturing process, encompassing the initial RTL planning process, all the way through production ramp and in-field operation.

References;

[1] Aqlan, F., Ramakrishnan, S., & Shamsan, A. (2017, December). Integrating data analytics and simulation for defect management in manufacturing environments. In *2017 winter simulation conference (WSC)* (pp. 3940-3951). IEEE. https://doi.org/10.1109/WSC.2017.8248104

[2] Bonthu, C. (2025). The role of data governance in strengthening ERP and MDM collaboration. International Journal of Computational and Experimental Science and Engineering. https://ijcesen.com/index.php/ijcesen/article/view/3783

[3] Bonthu, C. (2025). Unifying multiple ERP systems: A case study on data migration and integration. Utilitas Mathematica. https://utilitasmathematica.com/index.php/Index/article/view/2785

[4] Canal, R., Hernandez, C., Tornero, R., Cilardo, A., Massari, G., Reghenzani, F., ... & Abella, J. (2020). Predictive reliability and fault management in exascale systems: State of the art and perspectives. *ACM Computing Surveys (CSUR)*, *53*(5), 1-32. https://doi.org/10.1145/3403956

[5] Chavan, A. (2023). Managing scalability and cost in microservices architecture: Balancing infinite scalability with financial constraints. Journal of Artificial Intelligence & Cloud Computing, 2, E264. http://doi.org/10.47363/JAICC/2023(2)E264

[6] Chavan, A. (2024). Fault-tolerant event-driven systems: Techniques and best practices. Journal of Engineering and Applied Sciences Technology, 6, E167. http://doi.org/10.47363/JEAST/2024(6)E167

[7] Chen, W., Ray, S., Bhadra, J., Abadir, M., & Wang, L. C. (2017). Challenges and trends in modern SoC design verification. *IEEE Design & Test*, *34*(5), 7-22. https://doi.org/10.1109/MDAT.2017.2735383

[8] Dhanagari, M. R. (2024). MongoDB and data consistency: Bridging the gap between performance and reliability. *Journal of Computer Science and Technology Studies, 6*(2), 183-198. https://doi.org/10.32996/jcsts.2024.6.2.21

[9] Dhanagari, M. R. (2025). *Bridging IoT and Healthcare: Secure, Real-Time Data Exchange with Aerospike and Salesforce Marketing Cloud*. *International Journal of Computational and Experimental Science and Engineering*, 11(4). https://ijcesen.com/index.php/ijcesen/article/view/3853/1161

[10] Duan, H. (2024). From MOSFET to FinFET to GAAFET: The evolution, challenges, and future prospects. *Appl. Comput. Eng*, *50*(1), 113-120. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://pdfs.semanticscholar.org/e725/53b1c19cdb34e7c771de4b1767d84d4756e8.pdf

[11] Goel, G., & Bhramhabhatt, R. (2024). Dual sourcing strategies. *International Journal of Science and Research Archive*, 13(2), 2155. https://doi.org/10.30574/ijsra.2024.13.2.2155

[12] Guthaus, M., Batten, C., Brunvand, E., Gaillardon, P. E., Manohar, R., Mazumder, P., ... & Stine, J. (2023). NSF Integrated Circuit Research, Education and Workforce Development Workshop Final Report. *arXiv preprint arXiv:2311.02055*. https://doi.org/10.48550/arXiv.2311.02055

[13] Hung, S. C., Bhoumik, P., Chaudhuri, A., Banerjee, S., & Chakrabarty, K. (2024). Design-for-Test Solutions for 3-D Integrated Circuits. *Integrated Circuits and Systems*, *1*(1), 3-17. https://doi.org/10.23919/ICS.2024.3419629

[14] Kalel, D. (2024). *Advanced Structural and Semi-Formal Verification Flow for Clock Domain Crossing (CDC) in Asynchronous Multiclock Systems* (Doctoral dissertation, Université Grenoble Alpes [2020-....]). https://theses.hal.science/tel-04695158/

[15] Karwa, K. (2023). AI-powered career coaching: Evaluating feedback tools for design students. Indian Journal of Economics & Business. https://www.ashwinanokha.com/ijeb-v22-4-2023.php

[16] Katsaros, K., Mavromatis, I., Antonakoglou, K., Ghosh, S., Kaleshi, D., Mahmoodi, T., ... & Simeonidou, D. (2024). AI-native multi-access future networks-the REASON architecture. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3507186

[17] Konneru, N. M. K. (2021). Integrating security into CI/CD pipelines: A DevSecOps approach with SAST, DAST, and SCA tools. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

[18] Konneru, N. M. K. (2021). Integrating security into CI/CD pipelines: A DevSecOps approach with SAST, DAST, and SCA tools. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

[19] Kumar, A. (2019). The convergence of predictive analytics in driving business intelligence and enhancing DevOps efficiency. International Journal of Computational Engineering and Management, 6(6), 118-142. Retrieved from https://ijcem.in/wp-content/uploads/THE-CONVERGENCE-OF-PREDICTIVE-ANALYTICS-IN-DRIVING-BUSINESS-INTELLIGENCE-AND-ENHANCING-DEVOPS-EFFICIENCY.pdf

[20] Kumar, S., Satheesh, N., Mahapatra, A., Sahoo, S., & Mahapatra, K. K. (2016, December). Securing IEEE 1687 standard on-chip instrumentation access using PUF. In *2016 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS)* (pp. 56-61). IEEE. https://doi.org/10.1109/iNIS.2016.024

[21] Malik, G. (2025). *AI-Driven Security and Inventory Optimization: Automating Vulnerability Management and Demand Forecasting in CI/CD-Powered Retail Systems.* International Journal of Computational and Experimental Science and Engineering (IJCESEN). https://ijcesen.com/index.php/ijcesen/article/view/3855/1153

[22] Mishra, A., Cha, J., Park, H., & Kim, S. (Eds.). (2023). *Artificial intelligence and hardware accelerators*. Berlin: Springer. https://link.springer.com/book/10.1007/978-3-031-22170-5

[23] Nyati, S. (2018). Transforming telematics in fleet management: Innovations in asset tracking, efficiency, and communication. International Journal of Science and Research (IJSR), 7(10), 1804-1810. Retrieved from https://www.ijsr.net/getabstract.php?paperid=SR24203184230

[24] Pasham, S. D. (2020). Fault-Tolerant Distributed Computing for Real-Time Applications in Critical Systems. The Computertech, 1-29. https://www.yuktabpublisher.com/index.php/TCT/article/view/142

[25] Pinnapareddy, N. R. (2025). Cloud cost optimization and sustainability in Kubernetes. Journal of Information Systems Engineering and Management. https://www.jisem-journal.com/index.php/journal/article/view/8895

[26] Raju, R. K. (2017). Dynamic memory inference network for natural language inference. International Journal of Science and Research (IJSR), 6(2). https://www.ijsr.net/archive/v6i2/SR24926091431.pdf

[27] Rendon, M. J. (2024). *12nm Finfet aging characterization through wear-out sensor design* (Doctoral dissertation, University of British Columbia). http://hdl.handle.net/2429/89220

[28] Samantapudi, R. K. R. (2025). *Advantages & impact of fine tuning large language models for ecommerce search. Journal of Innovation in Science, Engineering and Management (JISEM), 10*(45s). https://doi.org/10.52783/jisem.v10i45s.8898

[29] Sardana, J. (2022). The role of notification scheduling in improving patient outcomes. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

[30] Singh, V. (2022). Integrating large language models with computer vision for enhanced image captioning: Combining LLMS with visual data to generate more accurate and context-rich image descriptions. Journal of Artificial Intelligence and Computer Vision, 1(E227). http://doi.org/10.47363/JAICC/2022(1)E227

[31] Singh, V., Doshi, V., Dave, M., Desai, A., Agrawal, S., Shah, J., & Kanani, P. (2020). Answering Questions in Natural Language About Images Using Deep Learning. In *Futuristic Trends in Networks and Computing Technologies: Second International Conference, FTNCT 2019, Chandigarh, India, November 22–23, 2019, Revised Selected Papers 2* (pp. 358-370). Springer Singapore. https://link.springer.com/chapter/10.1007/978-981-15-4451-4_28

[32] Subham, K. (2025). Scalable SaaS implementation governance for enterprise sales operations. *International Journal of Computational and Experimental Science and Engineering.* https://ijcesen.com/index.php/ijcesen/article/view/3782

[33] Sukhadiya, J., Pandya, H., & Singh, V. (2018). Comparison of Image Captioning Methods. *INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH*, *6*(4), 43-48. https://rjwave.org/ijedr/papers/IJEDR1804011.pdf

[34] Tehranipoor, M., Zamiri Azar, K., Asadizanjani, N., Rahman, F., Mardani Kamali, H., & Farahmandi, F. (2024). Secure physical design. In *Hardware Security: A Look into the Future* (pp. 401-445). Cham: Springer Nature Switzerland. https://link.springer.com/chapter/10.1007/978-3-031-58687-3_9