Breakdown of Traditional Test Assumptions in AI Silicon

Chetan Arvind Patil
Contributing Writer | EPDT

Semiconductor testing has traditionally focused on post-fabrication validation - with structural fault models, scan-based coverage and parametric measurements used to determine device quality. This approach assumes that logic correctness, timing closure and defect screening at wafer sort and final test are enough to guarantee system functionality.

However, AI silicon brings new architectural and physical challenges beyond these assumptions. Devices now include heterogeneous compute elements and high-bandwidth memory stacks. They use advanced packaging structures (such as 2.5D interposers and 3D hybrid bonding). High-speed die-to-die interconnects run at multi-Gbit data rates. Such features create new failure modes - including interconnect integrity loss, timing changes across dies, plus packaging-induced stress. Traditional test methods cannot fully detect these issues.

AI accelerators also operate under workload-driven conditions that change validation requirements and the AI workloads do not behave like deterministic logic systems. Instead, they use variable data distributions and mixed-precision formats (like FP8, BF16, and INT8). They also show highly parallel patterns, which stress memory bandwidth and interconnects. As a result, devices might pass conventional tests, but exhibit accuracy loss, stability issues, or varying performance in real-world deployments.

Furthermore, packages with several hundred W power densities generate dynamic heat gradients and rapid voltage fluctuations. These factors affect timing margins and reliability. Logic, interconnect, memory, power and workload domains interact in complex ways. Such interactions are difficult to observe or reproduce after silicon production, underscoring the limits of traditional test flows.

Limitations of post-silicon test in AI accelerators

Post-silicon testing historically relied upon well-defined observability and controllability through design-for-test (DfT) structures - enabling detection of structural defects and parametric deviations. This approach proves effective when failures are localised and can be exposed through scan chains or built-in self-test. However, in AI silicon many critical failure modes arise from system-level interactions that are not directly observable through conventional test mechanisms.

As examples, signal integrity degradation across die-to-die interconnects, latency variations in memory access paths and synchronisation mismatches between distributed compute elements often manifest only under realistic traffic conditions. These conditions are difficult to emulate during production test, where patterns are optimised for coverage and throughput instead of system behaviour. Consequently, certain interaction-driven failures remain hidden during traditional validation flows.

Transition to heterogeneous integration further complicates defect isolation. Known-good-die methodologies assume that individually tested dies will function correctly when integrated into a package. In practice, interface mismatches, thermal coupling and mechanical stress introduce failure mechanisms that only appear after assembly. These integration-induced effects reduce conventional screening techniques' effectiveness - shifting the burden of validation toward system-level testing.

Another limitation lies in the inability to accurately replicate deployment environments during test. AI accelerators operating in data centre settings experience workload bursts, dynamic frequency scaling and sustained high utilisation which create complex power and thermal profiles. Production test environments, constrained by throughput and infrastructure, cannot fully reproduce these conditions. In addition, fragmentation of test data across simulation, emulation, wafer sort and system-level validation prevents effective correlation, making it difficult to trace failures to their root causes and to apply insights consistently across design and manufacturing.

Domain	Traditional Approach	Shift-Left in AI Silicon	Impact
Workload Validation	Functional vectors applied post-silicon.	Workload-aware modelling during architecture and design.	Aligns silicon behaviour with real application performance.
Interconnect Test	Structural and parametric checks at test stages.	Pre-silicon co-validation of die-to-die and memory interfaces.	Reduces integration-related failures and improves yield.
Power and Thermal	Measured during late-stage validation.	Modelled during design partitioning and floor planning.	Prevents late-stage redesign and improves reliability.
Observability	DfT structures focused on logic coverage.	System-level instrumentation planned during design.	Enhances debug capability across lifecycle.
Test Data	Generated and analysed at manufacturing stages.	Continuous data flow from simulation to field.	Enables predictive analytics and faster root-cause analysis.
Validation Scope	Device-level correctness.	System and workload-level behaviour.	Bridges gap between silicon test and deployment conditions.

Table 1: The differences that a shift-left is making

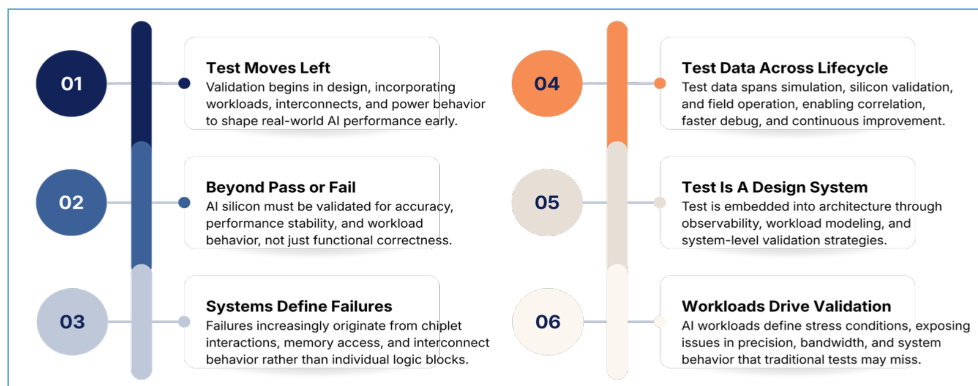

Figure 1: AI silicon is driving test from production to design

This transition reflects a broader mindset change. Instead of focusing solely on whether devices pass predefined tests, the emphasis moves toward ensuring that systems behave correctly under realistic operating conditions. By incorporating these considerations early in design processes, potential issues can be identified and addressed before they propagate into later stages. Alignment between silicon validation and real-world deployment requirements is thus improved and the likelihood of late-stage failures, costly redesigns and performance inconsistencies is significantly reduced.

What shift-left really means in AI silicon

Shift-left in AI silicon is not simply the early execution of existing test methodologies. It represents a redefinition of test as an integral component of system design. This begins with the incorporation of workload-aware validation into architectural planning, where AI accelerators are evaluated against representative models and data patterns during early design stages to ensure that performance, accuracy and stability targets are met under realistic conditions.

Equally important is the expansion of observability beyond structural coverage. Designers must introduce instrumentation that provides visibility into interconnect behaviour, memory transactions and power dynamics. This requires planning for telemetry and monitoring capabilities during the design phase, rather than relying solely on post-silicon debug tools.

Cross-domain co-validation is another essential element. Logic, interconnect, power and thermal effects are tightly coupled in AI systems, which necessitates integrated simulation and analysis frameworks. By validating these interactions early, designers can identify and mitigate issues that would otherwise emerge late in the development cycle.

Data continuity also becomes a foundational requirement. Test data must be treated as a lifecycle asset, with consistent metrics and formats enabling analysis across simulation, silicon validation and field operation. This continuity supports faster root-cause identification and enables predictive insights that improve both design and manufacturing outcomes. The shift-left imperative becomes tangible when examining how specific aspects of test evolve across the lifecycle (as outlined in Table 1).

Long-term impact

The shift-left imperative represents a structural transformation in semicond development. As AI workloads and system architectures continue to evolve, increasingly influences fundamental design decisions - including com partitioning, interconnect topology, memory hierarchy, power delivery strate etc. In this context, test is no longer downstream of design, but becom constraint and a guide for architectural trade-offs.

This transformation also reshapes cost optimisation. Early, workload-aware validation enables better informed trade-offs between design complexity, test time and manufacturing yield - reducing the likelihood of costly redesigns and late-stage corrections. In parallel, integrating test data across the lifecycle, from simulation and emulation to silicon validation and field operation, creates a continuous feedback loop. This closed-loop approach allows insights from real-world behaviour to feed subsequent design iterations, improving both product quality and development efficiency.

The definition of correctness in AI systems also evolves in this framework. Validation extends beyond binary pass/fail outcomes to include performance consistency, numerical accuracy and long-term reliability under representative workloads. This requires validation methodologies accounting for variability, dynamic operating conditions and systemlevel interactions across compute, memory and interconnect domains. In AI silicon, test is no longer an endpoint, but a design-time system shaping how devices behave under real-world conditions and bridging the gap between silicon functionality and system-level performance.

Use our in-house mods operation
to meet your project's requirements
Learn more:
hammondmfg.com/mods
uksales@hammfg.com • 01256 812812