From Lab to Data Center: Accelerating Silicon Readiness Through Reliability Innovation

Sujata Paul (Email: sujatapaul@microsoft.com; Phone: +1-919-334-8542)
Ray Talacka, Georgia Modoran, Anis Rahman, Shridhar Dixit, Serena Wang, Joey Lee, Saurabh Agrawal
Microsoft, 1045 La Avenida St, Mountain View, CA 94043

Abstract*— As Microsoft scales its in-house silicon efforts with compute and AI accelerators like Cobalt and Maia, reliability engineering has emerged as a cornerstone of successful productization. This paper emphasizes the critical need for comprehensive reliability testing – including High Temperature Operating Life (HTOL), Electrostatic Discharge (ESD), Latch-Up (LU), Fuse Test and Soft Error Rate (SER) – to ensure these chips meet the stringent demands of hyperscale AI data canters. We highlight the importance of an Intelligence-Based Qualification methodology that rigorously targets high-risk failure mechanisms, ensuring Cobalt and Maia achieve the reliability and quality required for mission-critical environments. Cobalt and Maia are designed for high-throughput, low-latency performance in Microsoft's Azure cloud infrastructure, where reliability is critical to provide the best possible customer experience – it is essential for sustaining uptime, guaranteeing performance consistency, and enabling long-term deployment viability. Failures in the field could lead to cascading service disruptions across cloud services, increased power demands from redundancy, and added support costs. Microsoft's silicon reliability strategy is therefore anchored in rigorous and innovative qualification methods focused on intelligent testing. By integrating advanced test infrastructure and data-driven validation into the development lifecycle, Microsoft ensures that Cobalt and Maia can be confidently scaled to power the next generation of AI and cloud computing with minimal risk of unexpected failures.

I. Introduction

Microsoft's recent development into custom silicon design for cloud infrastructure – exemplified by the compute processor Cobalt100 and the AI accelerator Maia100 – underscores the critical role of reliability engineering in bringing these chips from lab to data center. These in-house designs deliver cutting-edge performance for high-throughput, low-latency workloads (e.g., AI inference and cloud services running on Azure's custom ARM64-based servers). In fact, Cobalt100 processors are already powering major products like Microsoft Teams and Azure VMs do with industry-leading price-performance. In such environments, even minor hardware failures can have outsized impacts, causing service downtimes or degrading user experience. Thus, reliability is elevated from a mere checklist item to a strategic necessity in product development.

Reliability and Quality: It is useful to distinguish these concepts. Quality refers to conformance to specifications at the time of deployment (often called "time-zero" quality). It ensures a chip is defect-free and performs as designed when it ships. Reliability refers to the ability of that chip to continue functioning within specs over its intended lifetime (often 5–10 years for data center hardware) under normal operating conditions. In other words, quality is about delivering a flawless product upfront, while reliability is about sustaining flawless operation over time. Both are vital for cloud services: high initial quality means fewer early failures in the field, and high reliability means low failure rates as years pass.

Even a single unmitigated chip failure in the field can ripple through dependent services – leading to cascading disruptions, triggering failover protocols that raise overall power usage, and incurring significant maintenance costs. For example, if an AI accelerator like Maia in a server fails unexpectedly, it could interrupt critical AI tasks, force workloads to switch to backup hardware (reducing efficiency), and require a costly server replacement. At the scale of Microsoft's global cloud (with millions of chips deployed), such incidents must be exceedingly rare to maintain the promised 24/7 uptime. This is why Microsoft treats reliability engineering as a high priority from the earliest design stages: reliability isn't just about avoiding warranty costs; it is about protecting the integrity and availability of cloud services that enterprises and consumers rely on daily.

Extrinsic vs. Intrinsic Failures: A helpful concept in reliability engineering is the "bathtub curve," which describes product failure rates over time (Fig.1). New hardware often faces a higher early-life failure rate due to manufacturing defects or other extrinsic issues – analogous to how some new cars might have initial defects. Over time, if these early defects are weeded out, the product enters a long period of low, steady failure rate (the "useful life" plateau). Eventually, as the product ages and wears out, failure rates increase again towards the end of life due to intrinsic wear-out mechanisms (e.g., material degradation).

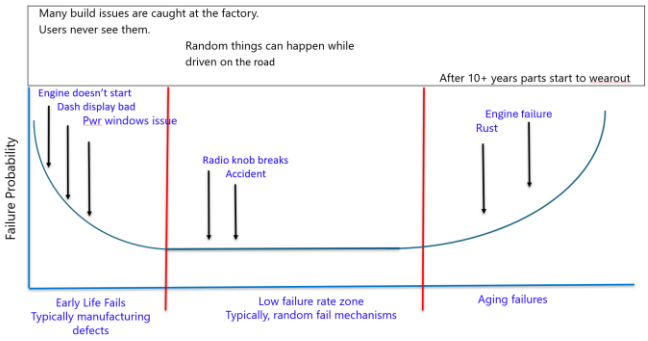

Fig. 1. Bathtub Curve -Defect Scenario (Example of New Car Issues)

Reliability engineering aims to suppress both the early and late failures: through robust manufacturing quality control (to eliminate extrinsic failures or "infant mortalities") and through design-for-reliability (DFR) plus stress testing (to ensure intrinsic failures won't occur until well beyond the intended service life). In practice, this means new chips undergo screenings like burn-in (to catch latent defects before shipping) and incorporate margins and protective features in design so that phenomena like transistor aging or electromigration do not cause issues until far past their expected usage period.

To achieve these goals for Cobalt and Maia, Microsoft has adopted an "Intelligence-Based Qualification" methodology – a data-driven, targeted approach to reliability qualification. Instead of uniformly testing every aspect of the chip to extremes (which would be time-consuming and costly), this methodology focuses on the highest-risk failure mechanisms and scenarios identified via design analysis and prior experience. For instance, if analysis shows that memory arrays are particularly sensitive to cosmic radiation, additional testing effort is spent on Soft Error Rate evaluation. If high-current power circuitry is a potential weak point, tests like electromigration stress or thermal cycling might be emphasized. This intelligent focusing of test resources ensures comprehensive coverage of likely failure modes while optimizing time-to-market. It also balances often tricky trade-offs between exposing "observable" failures in the lab and not inducing unrealistic stress that could create "latent" damage. The result is a qualification program that gives high confidence in real-world reliability without unnecessary test redundancy.

The following sections of this paper detail how reliability and quality are engineered and validated for Microsoft's in-house silicon. In Section II, we describe the key reliability tests – HTOL, ESD, Latch-Up, Fuse Test and SER – and the methodologies employed. Section III summarizes how this reliability efforts ensure that Microsoft's chips meet the demands of hyperscale deployment. Finally, Section IV (Scope of Future Work) highlights upcoming challenges and directions in reliability engineering as silicon technology and usage models evolve. Throughout, we maintain a cross-disciplinary perspective, explaining not just the technical procedures but also why they matter for the overall goal of dependable, scalable cloud infrastructure.

II. Reliability Test Methodology

The four primary categories are: (A) High Temperature Operating Life (HTOL), (B) Electrostatic Discharge (ESD), (C) Latch-Up (LU), and (D) Fuse Test and (E) Soft Error Rate (SER) testing. In addition, numerous standard production tests (voltage stress, burn-in, etc.) are used on the manufacturing line to ensure quality. Here we focus on the dedicated reliability tests that simulate years of use and rare events within a compressed timeframe.

A. High Temperature Operating Life (HTOL) Testing: HTOL is an accelerated life test designed to provoke aging-related failures by operating chips under extreme conditions for extended periods. The aim is to simulate multiple years of normal operation in a matter of weeks. In HTOL, a batch of devices is powered up and run at elevated temperature (typically 125 °C or higher, well above the usual data center junction temperature of ~65 to 105 °C) and often under elevated voltage (e.g., 5–10% above nominal) continuously for hundreds or thousands of hours. These harsh conditions accelerate time-dependent failure mechanisms such as Bias Temperature Instability (BTI), Hot Carrier Injection (HCI), Time-Dependent Dielectric Breakdown (TDDB), and Electromigration (EM):

BTI causes transistors to gradually slow down (threshold voltage increases) with prolonged static bias at temperature.

HCI causes transistors to degrade (carrier-induced damage) due to dynamic bias also leading to speed loss over time.

TDDB refers to the gate oxide wearing out; when it fails, a transistor's gate oxide suddenly breaks down, causing an abrupt failure (like a short).

Electromigration affects metal interconnects carrying current; over time, atoms can migrate, causing a wire to become more resistive or even open circuit (gradual or sudden loss of connectivity).

Microsoft uses custom high-temperature ovens and burn-in boards for HTOL experiments. Dozens of chips are mounted on burn-in boards that can withstand 125 °C, and these boards provide stable power, clocking, and I/O to the devices (Fig.2). Importantly, the chips are not idle, they execute specially crafted dynamic stress test patterns continuously during HTOL. These patterns maximize switching activity in various blocks of the chip to generate internal heat and stress (for example, a pattern may repeatedly write and read all bits of an SRAM to test for any bit cell that might start failing after multiple cycles). They are also designed to periodically exercise different circuits to ensure a broad coverage of the silicon. (Fig.3). By doing so, HTOL not only accelerates wear-out mechanisms, but also can expose any latent manufacturing flaws that weren't caught in initial tests (those would manifest as early failures under stress).

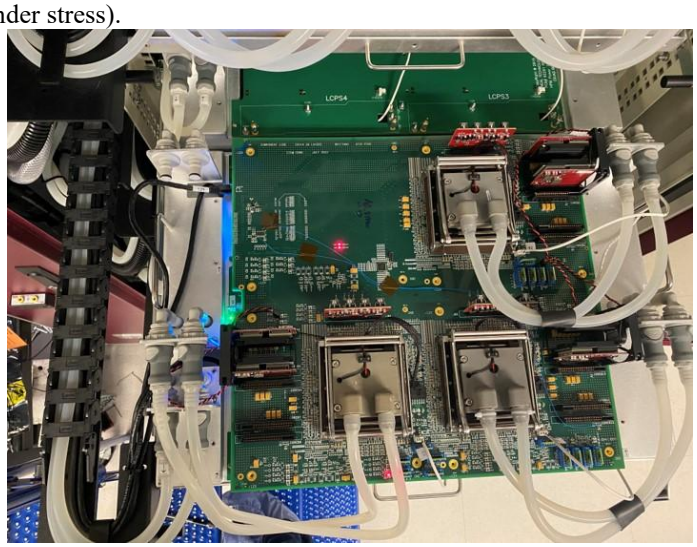

Fig. 2. Example of HTOL Board, 3 DUTs per board with customized socket with a liquid cooling system.

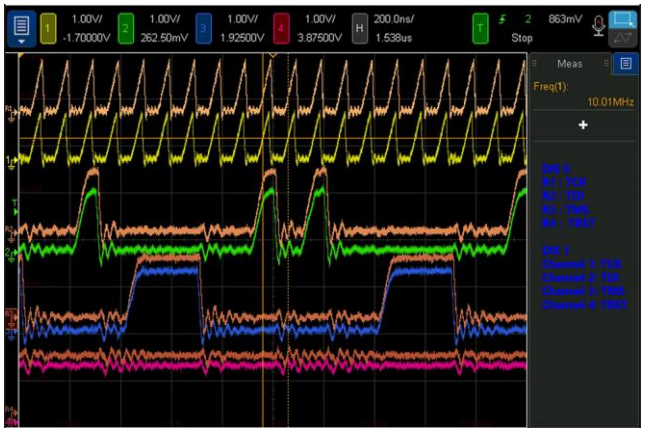

Fig. 3. Dynamic stress pattern, clocks activity

Throughout the HTOL run key parameters are monitored. We log supply currents, and in more sophisticated setups, on-board sensors track each device's temperature and even voltage rails (Fig.4). This way, if a device fails or enters an abnormal state (e.g., drawing excessive current due to an incipient short), we can detect it immediately and shut it down to prevent secondary damage (and note the time of failure for analysis).

Fig. 4. Example of Temperature, Voltage and Current monitoring

We also remove devices at intermediate intervals (e.g., 168 hours, 500 hours) for brief testing to measure any parametric drift. One common metric is the minimum operating voltage (Vmin) of the chip or certain subsystems – essentially, how much voltage is needed for the device to meet timing. As transistors age (BTI and HCI effects), Vmin tends to increase (Fig.5). By collecting this data, we can estimate how much Vmin degradation parts are expected to see in data center use conditions. Such information is crucial as it provides input on the required design margin a core needs for aging to function at its original frequency at end of life. In the Cobalt CPU case, the designs showed sufficient margin, and the drift was determined to be acceptable. If the drift were larger than expected, it could indicate a reliability risk or the need to adjust operating conditions (like slightly increasing voltage over life or derating clock frequency) to maintain performance – something we want to avoid if possible.

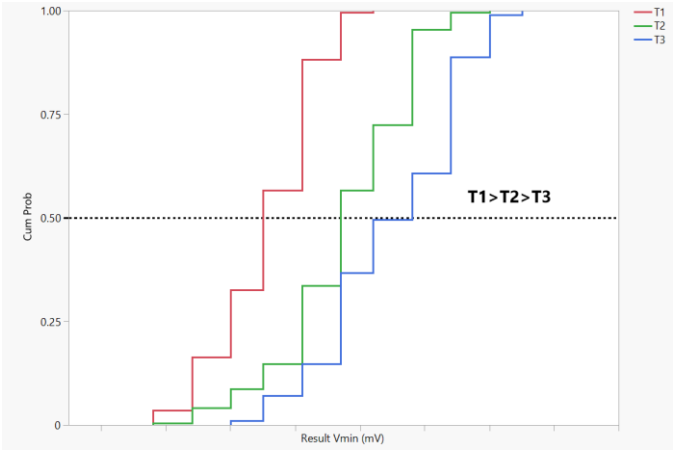

Fig. 5. Example of Vmin degradation plot

HTOL thus serves two purposes: (1) It flushes out any "infant mortality" failures that escaped the manufacturing tests (2) It confirms that intrinsic wear-out mechanisms will not cause failures within the product's intended lifetime. By examining any failures and the parameter shifts in surviving devices, reliability engineers verify that the product's lifespan is supported with appropriate margins. For Cobalt and Maia, devices in HTOL exhibited minimal parametric shifts and zero catastrophic failures across all units tested, giving high confidence that field failure rates due to intrinsic aging will be extremely low. In summary, HTOL provides evidence that these chips can "go the distance" in the data center – running hot and hard for years – without wearing out.

B. Electrostatic Discharge (ESD) Testing: Every handling of a chip – whether during manufacturing, assembly onto a circuit board, or maintenance in the field – carries a risk of electrostatic discharge. ESD is the sudden flow of static electricity between objects of different potential, and can occur, for instance, when a person touches a chip or when a charged device comes into contact with ground. A tiny spark that might be harmless to a human (we often only notice ESD above ~3,000 V) can instantly destroy or damage the nanometer-scale transistors in a modern IC if not properly protected. To guard against this, chips include on-die ESD protection structures at their pads: essentially special diodes or clamp circuits that shunt large currents away from sensitive circuitry whenever a high voltage spike occurs on an I/O or power pin. However, these protection elements themselves add capacitance and constraints to I/O design. Designing robust ESD protection without compromising high-speed signal integrity is a significant challenge – excessive protection can slow down I/O performance, whereas insufficient protection can lead to field failures. Microsoft follows industry-standard ESD qualification tests to ensure our chips can survive real-world handling. Two main models are used:

The Human Body Model (HBM) simulates a person discharging static to the chip. In the test, a 100pF capacitor (representing the capacitance of a human body) charged to a

certain voltage (e.g., 2000 V) is discharged through a 1.5kΩ resistor into the pin under test (Fig.6). This yields a relatively slow rise-time pulse (~100 ns) with a peak current of ~1.3 A at 2 kV. Each pin of the device (for all pin combinations) is zapped with both positive and negative polarity pulses. Our goal is for the device to pass at least the JEDEC standard requirement JS001 (typically ±1000 V HBM).

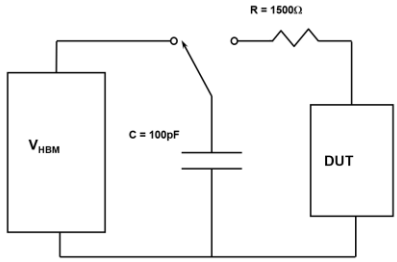

Fig. 6. HBM Circuit Schematic Currently run on the Thermo Fischer Tester

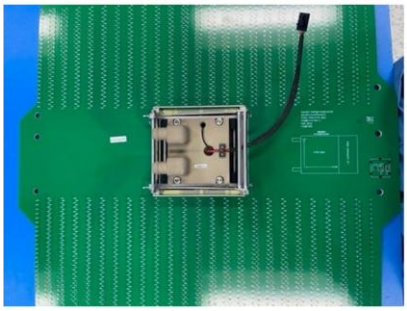

Fig. 7. Picture of ESD HBM Board for Thermo Fischer Tester (MK4)

Fig. 8. Picture of ESD Tester (Thermo Fischer Testers for HBM and CDM)

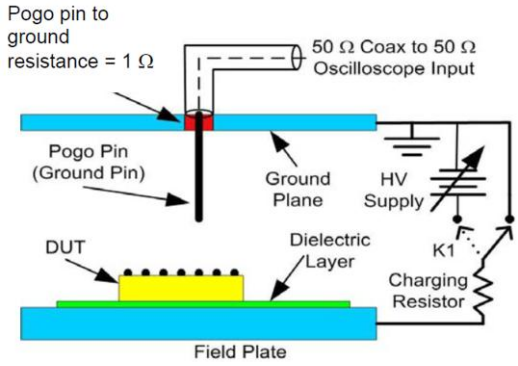

Fig. 9 CDM Testing Configuration / Schematic

The Charged Device Model (CDM) simulates the device itself acquiring charge and then discharging (for example, a chip sliding through a handler and touching ground). In CDM testing, the chip is placed on a field plate and charged, then one pin is rapidly grounded. CDM pulses are extremely fast (rise times ~ns) and high peak current, but short duration. The stress is localized to the pin that discharges (Fig.9). CDM is highly dependent on package form factor; larger packages can store more charge. We test to at least ±500 V CDM, a common requirement for robust devices (some smaller packages have lower targets).

We use an automated ESD tester for these experiments (Fig.8). After each zap at a given stress level, the device is fully tested. The "pass" criterion is that the device shows no degradation or damage (no increase in leakage currents, no shift in digital/analog parametrics, and obviously still functions correctly). The testing steps up voltage until either the spec level is passed, or a failure is observed. For Cobalt and Maia, all units passed HBM and CDM specifications comfortably, with no functional or parametric changes. Many pins in fact survived significantly higher levels indicating our on-chip protection design had margin beyond the minimum requirements. This ensures that during manufacturing and assembly – when chips might be exposed to static from machinery or human operators – they are very unlikely to suffer ESD damage.

Notably, our ESD design philosophy was to include "just enough to be safe but not so much that it slows down the transistors." The test results validate this approach: the chips meet stringent ESD robustness without any measurable impact on high-speed signalling performance. For instance, high-speed lanes on Cobalt were able to run at full speed, and we still achieved 1kV ESD-HBM requirements on those pins by using innovative clamp circuits that present very low capacitance under normal operation but activate during an ESD event. ESD testing is somewhat "invisible" to end users (who assume a device can tolerate normal handling), but it is an essential reliability item – without it, a simple static zap when inserting a cable could kill a server. By qualifying these devices to industry-standard ESD immunity, we significantly reduce the risk of field returns or failures due to handling mishaps or environmental static. Additionally, we enforce strict ESD-safe procedures in our factories and data centers (e.g., grounding straps, ionizers), but the chip-level robustness provides a safety net.

C. Latch-Up (LU) Testing: Latch-up is a potentially catastrophic phenomenon that can occur in CMOS integrated circuits, wherein a parasitic structure inadvertently forms a low-impedance path between power and ground, causing a huge current draw. It is typically triggered by a current or voltage spike and, once triggered, often persists ("latches") until power is cycled, often resulting in permanent damage due to overheating. Latch-up originates from the inherent parasitic bipolar transistors in the CMOS substrate. Modern chip processes and layouts include guard rings and trench isolation to greatly reduce latch-up risk, and technologies like Silicon-on-Insulator further mitigate it. However, our chips use a

conventional bulk CMOS process for optimal performance and cost, so latch-up prevention relies on design measures and needs to be verified by testing.

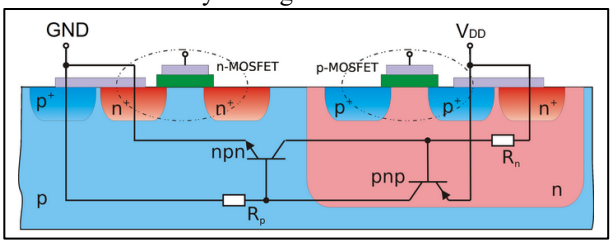

Fig. 10 Parasitic (PNP/NPN) Structures in CMOS: The Root of Latch-Up

Fig. 11 Example LU set up in ATE (Automatic Test Equipment) 93K environment

We perform JESD78-compliant latch-up testing, which involves two main types of stresses: over-voltage (Vtest) and over-current (Itest):

For over-voltage, each power supply pin (e.g., VDDCORE, VDDIO) is raised to 1.5× its nominal voltage while the chip is at the maximum operating temperature (around 85–125 °C). This tests whether any junction isolation breaks down and triggers parasitic conduction when the supply overshoots.

For over-current, each I/O pin is sequentially tested by injecting a current of ±100 mA into the pin (for a specified duration) while the chip is biased normally. Essentially, we force a large current into the pin (for positive and negative polarity) to simulate conditions like an overshoot on an input or misconnection that might forward-bias diodes into the substrate.

After each stress, we check if the chip is still functioning and, importantly, whether the supply current has returned to normal. A latch-up event would typically manifest as the device drawing excessive current even after the stress is removed. In our testing on Cobalt and Maia, no latch-up events were observed at the stress limits (all pins passed ±100 mA injection, and all power rails passed 1.5×Vmax). The devices remained fully operational after each pulse, with currents nominal. This indicates that the on-chip protections (guard rings around I/O structures, careful well engineering, etc.) are effective. We even tested beyond spec on a few sample units to gauge margin (e.g., some supplies tolerated ~1.6×Vmax before any anomalous behaviour).

This result is significant for reliability because latch-up can be a single-point catastrophic failure – if a chip ever latched up in the field (say due to an extreme transient on the power supply or an I/O excursion outside valid range), it could physically burn out or at least force a system reboot. By passing latch-up testing, we ensure that under all defined operating conditions (and well beyond), the chips will not succumb to this failure mode. It gives peace of mind that events like power spikes, or plugging in a mismatched cable that injects current, will not send the chip into a destructive short-circuit state. Essentially, the devices are latch-up immune within a generous safety margin. This allows system engineers to design power delivery and I/O interfaces without resorting to extraordinary measures to prevent any minor overshoot – the chips have inherent resilience.

D. Fuse Testing (One-Time Programmable Fuses): Modern high-performance chips like Cobalt and Maia include arrays of one-time programmable (OTP) eFuses that store vital configuration and security information. These metal-link fuses are “blown” (permanently opened) to encode data such as calibration settings, feature enables/disable flags, unique chip identifiers, and security keys. The reliability of these fuses is critically important because both testability and security depend on them functioning as intended throughout the product’s life. For example, certain test modes are only accessible when specific fuses are intact and conversely are locked out by blowing those fuses after manufacturing. Similarly, sensitive security features (like disabling JTAG debug or permanently provisioning encryption keys) are implemented via fuses; a failure in a fuse bit could either accidentally enable a backdoor or erase a key. Therefore, in addition to standard qualification of the fuse IP done by the foundry, Microsoft performs specialized fuse reliability testing focused on two potential failure mechanisms: read disturb and fuse regrowth.

Read Disturb Testing: Read disturb refers to the risk that repeatedly reading an intact (unblown) fuse could inadvertently program it (i.e., blow it) due to the stress of the read operations themselves. Each fuse read involves applying a voltage across the fuse and sensing current, which, if done thousands of times at elevated conditions, might cause electromigration (EM) in the fuse’s metal line. Over many cycles, EM could accumulate enough atomic migration to break the line, effectively “blowing” a fuse that was meant to remain intact. In normal use, fuses are read only a handful of times – typically at chip power-on to configure the device, on the order of a few hundred reads over a device’s lifetime (e.g., a server reboot infrequently). However, to ensure a robust reliability margin, we perform an accelerated read disturb test: we repeatedly read all fuse bits tens of thousands of times under worst-case conditions.

Specifically, we executed a fuse-read sequence that significantly exceeds the expected number of read cycles encountered in the field, applying elevated voltage and temperature to simulate worst-case stress conditions. A representative fuse block—such as the system fuse macro

containing critical configuration bits—was selected to typify the most vulnerable fuse geometry. The test was conducted in two phases: first, unprogrammed fuses were repeatedly read to confirm stability; then, a checkerboard pattern was programmed and subjected to the same stress to verify that programmed bits remained intact and unprogrammed bits did not inadvertently blow.

The procedure was run on an Automated Test System (ATE) using specialized microcode to maximize stress. We applied a stringent pass criterion, flagging any unexpected fuse state changes for investigation. No failures were observed, confirming that the fuse read circuitry and metal structures are robust against electromigration effects. These results validate that the fuse design maintains integrity well beyond typical usage scenarios, aligning with industry expectations for long-term reliability.

Fuse Regrowth Testing: The counterpart to read disturb is the risk of fuse regrowth – a phenomenon where a fuse that has been deliberately blown might "heal" or partially reform, reconnecting the circuit. Regrowth can occur via mechanisms like thermal diffusion or electrochemical migration of metal atoms across the tiny gap in a blown fuse. If a fuse ever regrew, it could be disastrous: a security fuse intended to permanently disable a feature might inadvertently re-enable it, or calibration data could corrupt.

To assess fuse integrity over time, we perform a high-temperature storage test using devices programmed with a checkerboard fuse pattern—alternating blown and unblown fuses. This setup allows us to monitor both fuse states under stress. The devices are placed in an oven and baked at elevated temperature for an extended duration, far beyond typical operating conditions, to accelerate potential metal diffusion or recrystallization effects that could cause regrowth in blown fuses. Periodically, the bake is paused and fuse states are read out using standard procedures under elevated conditions to detect any early signs of conductivity returning. After the full bake duration, a final read is performed to verify fuse states. The pass criterion requires all blown fuses to remain open circuits; any instance of a blown fuse regaining conductivity is considered a failure. This test ensures that fuse programming remains permanent and reliable throughout the product's lifetime.

The results were excellent: no regrowth was observed in any device — all blown fuses remained in their blown (open) state after baking at high temp and long duration. This indicates that the fuse programming procedure (the electrical "blow" pulse) creates a sufficiently large and stable gap in the metal that even prolonged high-temperature exposure cannot bridge. Essentially, once our fuses are blown, they stay blown. The test also implicitly validates that we are not under-blowing or over-stressing the fuses: an under-programmed fuse (insufficient energy) might leave ragged metal filaments that could reconnect, whereas an over-programmed fuse (excess energy) could damage surrounding dielectric. The lack of regrowth failures suggests our fuse blow process (developed per foundry guidelines) strikes the right balance, cleanly severing the link without collateral damage (Fig.13, Fig.14.).

Context and Security Considerations: These fuse tests complement the other reliability qualifications by addressing non-volatile, one-time programming elements of the chip. While HTOL, ESD, etc., ensure the transistor circuits and interconnects remain reliable, fuse testing ensures the integrity of on-chip configuration memory. Notably, security-critical fuses (such as those controlling JTAG lockout or ROM patches) are typically a subset of the overall fuse macro, and they benefit from the same blanket of testing. By demonstrating that fuses neither accidentally program nor unprogrammed under extreme conditions, we establish that security settings latched in fuses at manufacturing will persist correctly for the device's lifetime. Testability is also enhanced: many of these fuses are used to enable special test modes or store test results (like repair data, voltage binning information, etc.). Our fuse reliability tests give confidence that such data will not be lost or altered. For instance, some fuse bits store per-chip voltage-frequency calibration; thanks to the regrowth tests, we know that calibration values won't drift or revert due to fuse instability.

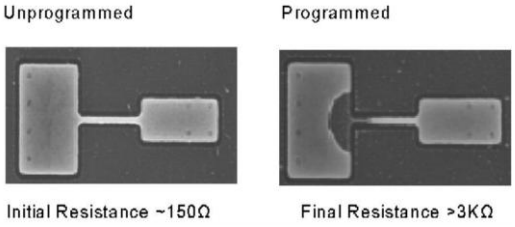

Fig. 13 Example of Fuse programming uses high Current to "break" the Metal resulting in an Open or higher resistance

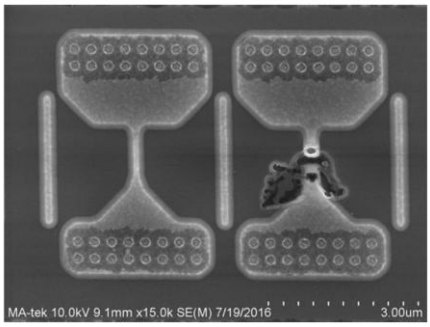

Fig. 14 Example of Incomplete fusing / Fuse regrowth, right picture shows excessive Break

In summary, fuse testing is a specialized but crucial part of our overall reliability qualification. By vigorously testing for read disturb and regrowth, we verified that the one-time programmable elements of Cobalt and Maia are stable and permanent under all foreseeable conditions. Unblown fuses remain intact even after extreme read stress, and blown fuses remain irreversible, showing no tendency to heal. These results, combined with foundry qualification data, give high assurance that the chips' configuration and security settings (as defined by fuse states) will be maintained correctly over time. Thus, the integrity of features controlled by fuses – from disabling test access to encoding cryptographic keys – is preserved, bolstering the overall reliability and trustworthiness of the platform.

E. Soft Error Rate (SER) Testing: Unlike the previous issues which cause permanent faults, soft errors are transient, random faults induced typically by external radiation. High-energy cosmic particles (primarily neutrons generated by cosmic rays colliding with the atmosphere) or trace radioactive elements in chip packaging can flip bits or induce spurious currents in a chip without damaging it permanently. The device continues to operate, but a memory bit might change from 0 to 1, or a logic state might glitch. If not mitigated, soft errors can lead to incorrect computations or system crashes. In large-scale deployments, the sheer number of devices and operations means even rare events become important; hence we quantify susceptibility in terms of SER, often given in FIT (Failures in Time), where 1 FIT = 1 failure in 10^9 device-hours.

Microsoft's SER testing involved exposing our chips to a calibrated neutron radiation beam to accelerate the occurrence of soft errors. We partnered with an external facility that produces a high-energy neutron beam with a flux many orders of magnitude above natural sea-level background. In essence, a few hours in the beam can simulate decades' worth of cosmic ray exposure. During this test, chips (particularly the memory-intensive parts like Maia's large on-die SRAMs) were powered and running self-test routines. For example, we load memory with a known pattern and continuously read and verify it while under bombardment, so any bit flips are immediately caught by a mismatch. We also include error-correcting code (ECC) logic on memories, so we can log whether a bit flip was corrected by ECC or if it resulted in an uncorrectable error.

The outcomes of SER testing are twofold: (1) Measure the raw error rate of various structures (e.g., how often per hour a single-bit error occurs in a standard sized SRAM under a given neutron flux), and (2) Validate that our error mitigation circuits (ECC, parity, etc.) work as intended. The data from accelerated tests is then scaled to real-world conditions using standard models for our chips, we observed a certain number of bit flips in the beam, all of which were corrected by ECC or detected by parity. From this, we calculated that a single Cobalt CPU or Maia accelerator has an SER on the order of a few hundred FITs for uncorrected errors – roughly meaning an average server might experience an uncorrectable memory error once in several decades of operation due to cosmic rays. This is in line with industry norms and is further mitigated at the system level (e.g., our servers use ECC memory and system-level retries, so even if a CPU cache bit flips, higher-level error handling usually prevents any user-visible impact).

An important aspect is distinguishing between SDC (Silent Data Corruption) and DUE (Detected Uncorrectable Error). An SDC means a wrong answer occurs with no indication (dangerous in, say, financial calculations), whereas a DUE triggers, for example, a machine check that will crash the process or system (safe, as it alerts that something went wrong). In our testing, we pushed the design to see if any SDCs occur. Thanks to strong ECC and parity mechanisms on all large arrays and critical flops, we observed no silent data corruptions – every radiation-induced event was either corrected on the fly or at least detected. This gives confidence that in the field, soft errors will either be invisible (corrected) or at worst cause a controlled fail-stop (like a server reboot), but not undetected wrong computations. Furthermore, the empirical data matched closely with our pre-silicon SER modelling, confirming that our design hardening (e.g., adding ECC on caches) was effective. We were even able to characterize the SER of complex logic (like a custom ARM core) – something rarely published – and found it to meet expectations.

One concrete outcome was the calculation of the Architecture Vulnerability Factor (AVF) for various structures. For example, not all storage bits in a processor are equally likely to cause an error if flipped; many flips might occur in bits that are not actively used or that get overwritten before being utilized. The testing helped calibrate these factors. The team found that certain critical sections of logic had very low AVF, meaning protective measures (such as redundancy or error detection in those sections) were sufficient. Additionally, the fact that Microsoft could measure logic SER in hardware for a complex modern chip (a first-of-its-kind achievement) means they can now validate and improve their radiation hardening approaches. The SER results fed back into design guidelines for next-generation products to ensure even those rarest of events are accounted for. In summary, the SER testing confirmed that the likelihood of an undetected, harmful soft error in any of these chips is extremely low. In a large deployment, correctable errors will occur infrequently and can be handled by ECC without impacting user-level services, and uncorrectable errors (which might cause a machine reboot, for example) are expected to be so rare that they have an insignificant impact on overall system availability.

Summarizing the SER experiment: by subjecting the chips to a neutron beam, we effectively fast-forwarded through billions of device-hours of operation. The measured soft error rate is extremely low, and more importantly, the architecture's resilience (via ECC, redundancy) ensures that the effective SER (as seen by the system) is even lower, well within acceptable limits for our cloud reliability targets. This means that even at the scale of thousands of accelerators and millions of processor cores, the likelihood of cosmic-ray-induced outages or corruptions will be very small. From a cross-disciplinary viewpoint, this is crucial because it ties hardware reliability to software correctness and overall service reliability – we can guarantee a level of computational integrity that, combined with system software measures, keeps cloud services computing accurately and continuously.

III. CONCLUSION

In this paper, we presented Microsoft's approach to accelerating silicon readiness through a focus on reliability innovation, as applied to the Cobalt and Maia in-house processors. Reliability engineering has been shown to be a foundational element in successfully moving these chips from the lab into large-scale production and operation in Azure data centers. The comprehensive qualification regimen – spanning HTOL, ESD, latch-up, Fuse test and SER tests – ensured that every major reliability concern is addressed proactively.

The results demonstrate that these chips meet or exceed the stringent reliability requirements of hyperscale cloud

computing. In practical terms, this means they will operate continuously with high availability: they can handle years of heavy workloads without performance loss, endure the rigors of manufacturing and field use without electrical or mechanical failure, and maintain data integrity in the face of environmental radiation. The Intelligence-Based Qualification methodology proved effective in guiding the reliability effort, allowing Microsoft to concentrate on critical failure mechanisms and thereby ensure robustness while maintaining an efficient development timeline.

By integrating advanced test infrastructures (like custom HTOL ovens and neutron beam testing capabilities) into our development pipeline, we were able to identify and eliminate potential reliability issues early. Moreover, the knowledge gained from these reliability tests forms a feedback loop to design. For instance, learning from the first-generation products has already influenced design adjustments and guardband policies for the next generation of accelerators. This commitment to data-driven validation at every stage – design, prototyping, and product qualification – sharply reduces the risk of unexpected failures after deployment.

In conclusion, Microsoft's silicon reliability program has enabled the deployment of Cobalt, Maia, and other custom silicon with confidence. The phrase "from lab to data center" signifies that a chip isn't truly successful until it performs flawlessly at scale in real customer scenarios; through rigorous reliability innovation, we bridged that gap. The resulting products empower Microsoft's AI and cloud services with cutting-edge performance without compromising on the reliability that enterprise customers demand. This work reinforces the principle that as we push the boundaries of silicon performance, an equally strong emphasis on reliability and quality is not just advisable but essential.

IV. FUTURE WORK

Looking ahead, there are several industry-wide challenges and future work areas in silicon reliability that Microsoft's team is actively exploring:

Scaling to Larger Dies: Each new generation of datacenter silicon (CPUs, GPUs, AI accelerators) tends to have increased die size and transistor count. Ensuring the same (or better) reliability on a die that might be twice as large is non-trivial. A larger chip has more potential points of failure (simply by area) and higher power density. Future work involves improving fabrication processes to keep defect levels extremely low and enhancing on-chip redundancy so that even if one small part fails, it does not affect the whole device. Advanced monitoring circuits on large dies may help predict failures before they happen. Maintaining reliability will require innovation in how we design self-checks and redundancy into these massive chips.

Multi-Chip Packaging Reliability: The industry is moving towards multi-chip modules and 2.5D/3D integration (for example, placing high-bandwidth memory stacks close to compute dies, or partitioning a large chip into smaller chiplets on an interposer). While this improves performance, it introduces new reliability concerns at the package level. Differences in thermal expansion coefficients between materials can cause stress as the device heats and cools, potentially leading to cracks or delamination over time. Managing the mechanical and thermal reliability of multi-die packages is a future focus. This includes extensive temperature cycling tests, improved underfill materials to absorb stress, and designing interconnects (micro-bumps, through-silicon vias) that can tolerate strain. The qualification process will expand to cover these package-induced failure modes comprehensively.

Silent Data Corruption (SDC) in Advanced Nodes: As transistor geometries shrink to ever smaller nanometre scales, we face new kinds of subtle failure modes. Systematic defects or marginal design conditions that were insignificant before can now sometimes cause logic errors that are not caught by traditional tests. These could manifest as very rare firmware or software errors (SDCs) that are hard to trace. Future reliability efforts will invest in techniques to detect and mitigate such issues. This might involve designing critical logic with duplication and compare (to catch an internal error), using parity bits more widely, or developing on-line self-test capabilities that periodically check the health of a circuit. Additionally, working closely with process engineers to identify any particular weak structures (for instance, a certain type of transistor that is more prone to variability) will be important. The objective is to ensure that as devices become more complex and operate closer to physical limits, we do not see an increase in unexplained computation errors.

High-Speed I/O and ESD Co-design: One emerging challenge is that as I/O speeds climb (serdes lanes of 100+ Gbps, for example), the traditional methods of protecting those I/O pins (like big ESD diodes or RC networks) become less feasible due to the added capacitance or latency they introduce. However, the need for ESD robustness remains. Future work will include developing innovative ESD protection techniques that are compatible with high-speed operation. This could mean using new device structures that remain transparent during normal operation but activate during an ESD event or leveraging package-level solutions (like ESD grounding paths in connectors). Another area is adjusting standards for what levels of ESD are required for ultra-high-speed interfaces and finding a balance that ensures reliability without stifling performance. Microsoft and others in the industry are researching materials and circuit topologies that can achieve ESD protection with minimal parasitic effects, which will be crucial for the next generations of networking and accelerator chips.

In summary, while the current reliability program has been successful, the work is never truly done. Scaling to the next nodes and architectures will bring new reliability questions that must be met with equally novel solutions. Microsoft's future reliability efforts will likely broaden to include not just chip-level, but system-level and package-level considerations, employing more automation (perhaps AI-driven analysis of reliability data) to predict and prevent failures. By continuing to prioritize reliability engineering and staying ahead of emerging challenges like multi-chip packaging and SDC, we aim to ensure that even as our silicon pushes new performance frontiers, it remains rock-solid in the demanding environment

of the data center. The ongoing commitment to reliability innovation will enable Microsoft's hardware to remain dependable as the foundation for cloud services, thereby extending the "lab to data center" success we have achieved with Cobalt and Maia to all future silicon endeavours.

ACKNOWLEDGMENT

The author would like to thank the Silicon Reliability Engineering team at Microsoft for their invaluable assistance with the extensive testing reported in this paper. In particular, we are grateful to the engineers and technicians who carried out the qualification experiments, ensuring the integrity of the results. Finally, we want to acknowledge our colleagues who offered constructive feedback on early drafts of this paper, helping to refine our analysis and presentation.

REFERENCES

[1] JESD22-A108: High Temperature Operating Life, Rev. D, Dec. 2016.
[2] JS-001: Electrostatic Discharge Sensitivity Testing – Human Body Model (HBM) – Device Level, Rev. 2024.
[3] JS-002: Electrostatic Discharge Sensitivity Testing – Charged Device Model (CDM) – Device Level, Rev. 2025.
[4] JESD78E, IC Latch Up Test, Revision of JESD78D, November 2011.
[5] JESD89-2A: Test Method for Alpha Source Accelerated Soft Error Rate, Revision of JESD89-2, November 2004.
[6] JESD22-A115: Fuse Test Method*, Rev. A, Dec. 2006.