Design and Adaptations to Mitigate Aging and Improve Reliability

Vignesh Thangavel
School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, Florida
vigneshthangavel@knights.ucf.edu

***Abstract*—As transistor downsizing continues beyond Moore's law, new challenges plague its operation, affecting its reliability with time and hence characterizing the aging behavior of the entire microarchitecture. With increased variability and unreliability owing to the breakdown of traditional bulk approximations, new constraints call for reconsideration and inclusion of a larger number of design goals. The major aging factors and some novel and insightful approaches to tackle and better characterize the same have been discussed in this paper. Key techniques used and tradeoffs in and between these different approaches with projections for future work have been presented to indicate their position in research.**

Keywords—Threshold Voltage, Negative Bias Temperature Instability (NBTI), Time Dependent Dielectric Breakdown(TDDB), Hot Carrier Injection(HCI), Oxide Breakdown (ODB), Electromigration (EM), Wear-Leveling, Multi-Core Architecture, FPGA, Aging, Reliability, Delays

I. UNDERSTANDING AGING

With continued miniaturization of transistors and circuits alike, a lot has been accomplished in terms of increased capabilities with reduced voltage supply and power requirements and this trend continues to motivate the same recursively. However, as the process size nears a few nm, several factors introduce variability in its design which manifest largely in their runtime behavior as aging effects. Aging generally refers to an undesired change in the behavior of the circuit, resulting in outputs different from that intended either in terms of unacceptable delays in producing the outputs or improper characterization of logic levels owing to increased threshold voltage required to achieve transistor switching. Thus, by its very definition (sharing a close analogy with the biological counterpart), it cannot be eliminated completely, but can surely be mitigated to effect graceful degradation. Aging unfortunately, can't be measured using a single or a couple of metrics and requires an understanding of the most important underlying factors.

A. Factors Characterizing Aging

Some of the most prominent factors that characterize aging are as follows:

1) Electromigration (EM)

Electromigration is the phenomenon where the high frequency flow of electrons causes momentum exchange with the metal ions in the interconnect medium containing both, resulting in mass transfer of the metal that forms hillocks and valleys of metal concentration in the interconnect, thereby affecting the conductivity and inductance properties adversely. The polycrystalline nature of the metal deposited results in mass build up or depletion at the boundaries called triple points, where three grains (each of size 100 nm approximately) meet.[4] This is evidenced by the fact that testing the lines under identical conditions (175°C, $2x10^6$A/cm2) led to polycrystalline line failure after 30 h, whereas the single crystal line showed no degradation after 26,000 h.[9] A critical length called Blech legth[5] exists, below which electromigration is inhibited in metals. This length is exceeded more often in architectures with greater miniaturization and thinner wires.

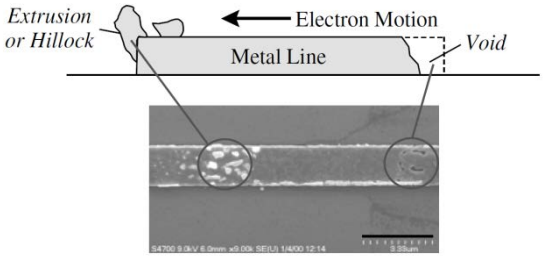

Figure 1: Electromigration in a silver line stressed with 23MA/cm2 at T= 160˚C. Image courtesy of [9].

Typical EM test methods use temperatures above 200˚C and current densities bove 10^6 A/cm^2, while normal ICs operate at maximum temperatures of 100-175˚C and line current densities of $5x10^5$ A/cm^2 or less. One popular test is the *standard wafer level EM acceleration test* (SWEAT), which uses special test structures and high current densities and hence results in a measurement time of 30-60s. The equations characterizing the effect on the mean time to failure (MTTF) is given as below (Black's equation[3]):

$$MTTF_{EM} \propto (J - J_{crit})^{-n} * e^{(Ea/kT)}$$
$$\propto J^{-n} * e^{(Ea/kT)}$$
$$\propto (Vdd \cdot f \cdot p)^{-n} * e^{(Ea/kT)}$$

Where J is the current density, J_{crit} is the critical/threshold current density above which EM begins, n is a material dependent constant and is 1.1 for silicon, Ea is the activation energy and equals 0.9 eV and other symbols have usual meanings. [2] Further the actual instantaneous current density J is modeled as:

$$J = C*V_{dd}*f*p/(W*H)$$

Where C is the capacitance and W and H are from the device geometry. [2] Electromigration occurs in two phases – incubation phase and catastrophic phase. In the incubation phase, interconnect characteristics remain mostly unchanged, while in the catastrophic phase a sharp rise in interconnect resistance is seen. Also, electromigration is usually found to follow log-normal distribution statistically speaking and is less affected by ac than it is by dc.

2) Time Dependent Dielectric Breakdown (TDDB)

As the name suggests, in this phenomenon, the gate oxide formed of silicon dioxide breaks down and results in the formation of a conductive path through the gate oxide. The Klein/Solomon model[6] characterizes oxide breakdown as a multistage event with a prolonged wear out period where *trap generation* occurs in the oxide, followed by the partial discharge event where *soft breakdown* due to accumulation of charge traps occurs, resulting in local distributions of high current density, and finally a catastrophic breakdown of the dielectric called *hard breakdown* is seen after multiple instances of soft breakdowns have occurred.[7][8] Oxide thickness, operating voltage and temperature are the most significant factors affecting TDDB and the dependence is encapsulated in the equation:

$$MTTF_{TDDB} \propto (1/V)^{(a-bT)} e^{(X+Y/T+ZT)/kT}$$

Where V is the operating voltage, T is the temperature, k is the Boltzmann constant and a, b, X, Y, Z are all fitting parameters based on Klein/Solomon model. [2] Dielectric and gate oxide breakdown generally follows Weibull disribution, statistically speaking. Oxide breakdown is affected more by ac than by dc. The interested reader may refer to [9] for supporting facts.

3) Negative Bias Temperature Instability(NBTI)

NBTI occurs primarily in p-channel MOS devices stressed with high temperature and large negative gate voltages and hence moderately large oxide electric fields (6MV/cm or less).[9] PBTI is generally considered to be recoverable though. However, with the advent of dielectrics with high dielectric constants and metal gate transistors, PBTI has started to gain equal importance as NBTI in FPGA architectures. [11] BTI manifests as an increase in threshold voltage (in magnitude) and a decreased in drain current and transconductance and is caused due to interface trapped charges and fixed oxide charges. Interface traps are created when the Si-H bonds break at the silicon-oxide interface, H then migrates, and dangling Si bonds result in degraded threshold voltages due to the degraded mobilities. Fixed charges add to this effect. This problem is particularly acute in technologies below 130nm which have thinner oxides. Gate oxides that scale to less than 4nm have a dominant effect on limiting circuit life time through NBTI.

One interesting property of NBTI is the huge dependence of threshold voltage variation with the switching behavior of the transistor of interest and upto 75% of previous NBTI-induced degradations can be annealed by biasing the pMOS gate at supply voltage (depending on the duty cycle and input patterns). [10] The most successful model describing NBTI and its effect on transistor degradation is the reaction-diffusion theory. The interested reader may refer to [10] for further details about modeling NBTI at the transistor, gate and circuit levels respectively and its effect on circuit performance in static and dynamic conditions.

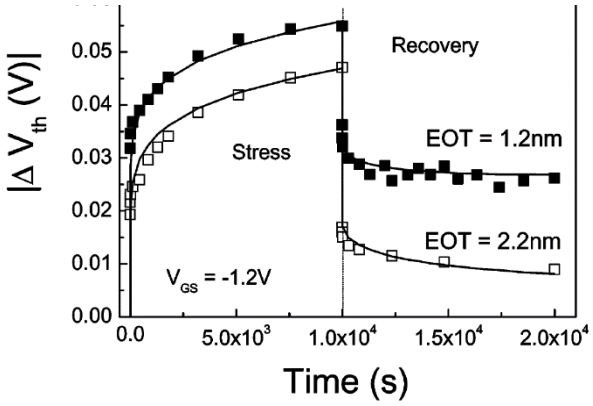

Figure 2: Variation in Vth for dynamic NBTI. The stress part corresponds to the case when a gate voltage is applied and the recovery occurs when gate voltage is zero. Image courtesy of [10].

4) Hot Carrier Injection (HCI)

Hot carriers refer to energetic carriers which are injected into the oxide when electrons in the channel enter the drain space charge region and cause impact ionization. These carriers can then be injected into the oxide as trapped charges, can flow through the oxide or generate interface traps. Alternately, they can flow to the substrate contact and generate photons which can propagate into the device and create electron hole pairs. The first mechanism results in voltage and mobility degradation while the second causes forward biasing of source-substrate junction, leading to further impact ionization and possibly snapback breakdown. Interface traps similar to the NBTI case are formed and measurement of substrate current for NMOS and gate current for PMOS indicate level of degradation due to HCI.

In addition to the above factors, there are also those like *stress induced leakage current (SILC)* and *electrostatic discharge (ESD).* [9] However, SILC is not expected to affect devices with oxides thinner than 5nm due to reduced trap generation rates in them. ESD can be avoided if human contact with the devices is prevented. Thus, these factors don't really hold much weight in this discussion.

B. *Effects on reliability and businesses*

Reliability is important, besides power and area constraints and would be a concern at par with these in future

technologies. The major challenges in reliability of CMOS with continuing miniaturization are as follows:

- Designs are getting increasingly susceptible to transient errors such as those induced by radiation. Error rates stay constant on per-bit basis, but total chip-level error rates grow with the scale of integration.
- Burn-in for screening early-life failures is becoming obsolete. Major burn-in challenges include power dissipation, cost, and serious concern about potentially reduced effectiveness in the future. Burn-in alternatives, e.g., and very low voltage (VLV) testing are also becoming ineffective owing to several reasons: high leakage, process variations, and reduced voltage margins.
- Device degradation (also referred to as aging), induced by degradation mechanisms such as bias temperature instability (BTI) is becoming increasingly important. While design margins are being squeezed to achieve high energy efficiency, expanded design margins are required to cope with aging. Hence, traditional speed or voltage margins to overcome degradation may become too expensive.

Typically, the failure times of devices are characterized by the terms Failure Rate, MTTF and MTBF.

Failure Rate: Failure rate as the name suggests is the number of failures per unit time.

Mean Time to Failure: It is the arithmetic average of failure times. This metric is used to characterize failures of devices that cannot be repaired and can fail only once.

Mean Time Between Failures: It is the arithmetic mean of the difference between failure times. This metric is used to characterize failures of devices that can be repaired and hence can fail multiple times.

Mean Time to Repair: It is the average time taken to repair a device after failing.

Generally, though MTBF = MTTF+MTTR. In aging literature, MTTF serves to be a good indicator of wear since aging of a device implies an inability to restore the previously available quality of functionality and hence in this sense an "irreparable failure" in quality. The phenomenon of aging, however is gradual and device performance statistics and signatures are good indicators of the degree of aging degradation to help decide on the lifetime and functionality of a device in a mission.

Failure rates of devices are plotted against time to study their reliability and one obtains what is popularly called the bathtub curve as shown below:

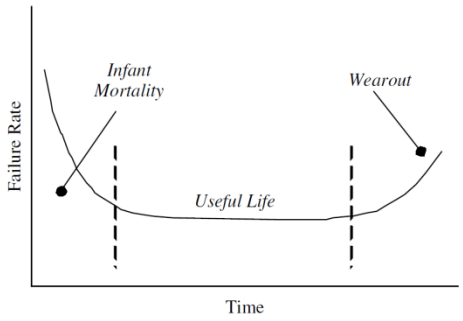

Figure 3: The bathtub curve showing failure rate and its effect on device lifetime. Image courtesy of [9].

One major motivation to plot bathtub curves is to identify the time period corresponding to the wearout phase and develop early warning systems to take remedial action to effect graceful degradation of the containing circuit. Doing so has several implications in identifying and formulating appropriate strategies to deal with unreliability due to aging.

Formulating the effect of wearout requires considering all the mechanisms mentioned above and manifests largely as signal propagation latency at the microarchitecural level, where both transistor/gate level delays and interconnects need to be estimated. In order to model interconnect delays in VLSI circuits, the fitted Elmore Delay [13] based on the Elmore Delay equation [12] is used. It relates the delay with load capacitance, geometry, driver impedance, etc. and is given as in [2] as:

$$\text{Delay}(r_d, c_l, l, w) = A \cdot r_d c_a l w + B \cdot r_d c_f l + C \cdot r_d c_l + D \cdot r c_a l^2/2 + E \cdot r c_f l^2/(2w) + F \cdot r l c_l /w$$

With r as the sheet resistance of interconnect wire, the terms dependent on r as γ and the ones independent of r as κ, the above can be re-written as:

$$\text{Delay}(r) = \kappa + r^* \gamma.$$

Effect of TDDB on signal latency is given in [8]. Similarly, effects of NBTI and HCI on signal latency can be used to calculate MTTF to characterize aging and account for the corresponding reliability.

The importance of hardware failures and their impact on businesses is highly significant. For instance, outages of high-profile websites like Amazon and Paypal were caused by hardware failures and their effects have been documented in [14] and [15] respectively. Some business customers have reported downtime costs of more than $1 million per hour. [16] Consumer expectations are always increasing in spite of the decrease in reliability of individual devices. This fact is supported by a recent poll conducted by Gartner Research which showed that more than 84% of organizations rely on systems older than 5 years and more than 50% use systems as old as ten years or more. [17] A broad taxonomy of approaches to mitigate aging and some of their pros and cons have been compared and contrasted in the next few sections.

II. Traditional Approaches

Traditionally, aging mitigation and issues of reliability have been addressed through testing and checking strategies and design margin considerations with the occasional use of canary circuits as failure predictors. Speed binning has also been practiced to help classify circuits based on their speed performance. Redundancy based approaches have dominated the sphere of approaches to ensure reliability and have eventually lead to the proliferation of CED based approaches. Last but not the least, parity checking and error detection codes have been used from time to time to protect against errors in small and often critical portions of circuits or to reinforce protection using other techniques. Traditional approaches are mostly a one-time static fix to problems of reliability and aging and hence have limited applicability. These approaches have been described below briefly.

Guard bands are typically designed by taking into consideration the worst case in terms of performance. Voltage and timing guard bands are generally decided during the design stage of the microarchitecture. The worst case delays are simulated corresponding to performance degradation due to variability factors at fabrication time and aging degraded performance which may occur during the expected normal life of the device. Based on the same, the processor is overdesigned such that delays of all logic paths are always ensured to be less than the operating clock period. Voltage guard bands are provided by taking into consideration stress probabilities at various nodes in the processor and allowable frequency variation with voltage and temperature and energy overheads to satisfy the design conditions throughout the reliable lifetime of the processor. Design is further fine-tuned by extraction of critical paths to allow for greater margins in these portions. This approach is relatively newer and enhances the effectiveness of guard banding technique. Some smart choices with regards to critical elements still continue to under the aegis of guard bands largely and would continue to do so. However, implementing guard band based overdesigning is faced with tremendous challenges due to process variations and cannot be uniformly implemented for entire circuits anymore.

Figure 4: Modular silicon verification hardware. Image courtesy of [44]

Silicon validation techniques typically are done pre-and-post fabrication and are accordingly called pre- and post-silicon validations. Pre-silicon strategies include design rule checks, RTL and timing synthesis to ensure that specifications are met. This stage however runs simulations that are several orders of magnitude slower than real silicon and hence several bugs and sometimes Trojans escape detection. Post-silicon strategies address these issues by providing for bug identification and localization and in addition uses techniques like burn-in testing, very low voltage testing to identify circuits at risk of early life failures. This is also accompanied by use of special structures classified as DFX (design for excellence) activities that includes a suite of DFT(Design for Testability), DFV (Design for Validation), DFD(Design for Debug), DFY(Design for Yield). These structures are used to validate, debug and test the fabricated circuits. Such special structures continue to be active areas of research and innovation in testing, but miniaturization challenges the effectiveness and scalability of these approaches.

Speed binning is a technique that economizes on fabrication cost where CPUs satisfying different speed specifications are manufactured by the same manufacturing process and then tested at decreasing frequencies/speeds of operation starting from the highest specified frequency to segregate those that perform as desired at the frequency of testing, from those that can't. The processors are thus 'binned' based on the highest frequency of normal operation/performance. Such binning can be done based on various parameters of operation. This method efficiently handles variations that manifest in terms of timing performance, but is done at design time only and holds quite some value in the face of process variations.

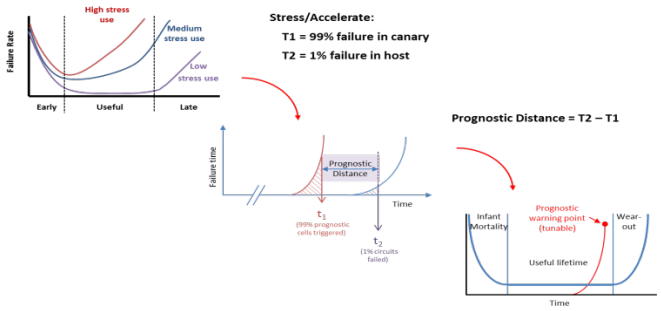

Figure 5: Principle of canary circuit operation. Image courtesy of [42]

Canary circuits are specially designed replicas of cores or critical components of the architecture which are by design and placement, carefully subjected to the same conditions as that of the component it is a replica of. They fail in advance of the circuit they are charged with protecting thereby providing an early indication of the wear out phase of that component. They are very effective as aging sensors, but have serious design issues owing to the requirement of replicating the conditions as in the core/component being monitored. In the face of high variability in newer technology generations, ensuring the same is increasingly difficult and hence spells a major limitation of this technique. The idea of canary logic is very powerful and its effective combination with certain approaches based on latency measurements for voltage scaling such as Razor[26] and parity checking sees it improve upon them. For instance [30] introduces canary flip flops to ease the clock tree design complexity of Razor while achieving the

same functionality while parity protection covers the cases where canary FF based prediction fails thereby offering much promise in such canary based approaches.

The core ideas of approaches discussed so far have focused on static design time approaches. Redundancy based approaches have since long been in use as the earliest and most intuitive dynamic runtime adaptation to aging and reliability factors affecting performance. Triple Modular Redundancy (TMR) helps mask one fault by voting, while NMR masks N-2 faults and since several modifications have been suggested to these, involving use GAs and is actively being researched for improvements in specific scenarios. Largely, pure redundancy based approaches are static fixes as they are designed at runtime and have limited coverage of faults. One approach that builds on lines inspired by canary circuits and the redundancy viewpoint, while offering a significant paradigm shift towards dynamic detection and recovery of *multiple* transient and permanent faults for entire microprocessor cores is DIVA (Dynamic Implementation Verification Architecture)[20] which also summarizes issues with traditional approaches to verification. DIVA adds a checker core in addition to the deeply speculative normal core minus the retirement stage (in the pipelined processor), dubbed DIVA core. Verification proceeds by comparing outputs in the main core and the pre computed outputs in the checker core to control pipeline flushing in case of faults. It is constrained by area and energy overheads but holds some promise if they are overcome. This work was extended to several processor cores to cover hard failure detection as well (as against DIVA's focus on soft errors alone) by Bower et. al[21]. The technique relied on maintaining counters for major architectural structures and association of every instance of incorrect execution detected by DIVA checker to a corresponding structure and decommissioning of the same when the number of faults for that counter exceeds a threshold and cold sparing is done. Work done by Shivakumar et. al[22] argues for the exploitation of existing redundancy within modern processors to increase fault tolerance yield through reconfiguration. Vijaykumar et. al[32, 33] have proposed approaches that exploit the idle and often redundant resources in a superscalar processor to effect time redundant computation by perform verification of computation during periods of low demand. This approach leverages on the work done by Slipstream group[36] on simultaneous redundant multithreading and instruction reuse[37]. Redundant computation also raises some adversities in raising workloads and sometimes being counter-effective in mitigating aging and instead causing it if the stress on resources is high enough. ReStore[38] is a variation on redundant computation which uses symptom detection to trigger replication in computation only when the probability of error is high. Thus, we see that *several flavors of redundancy have been explored and this continues to be an active area of research indicating a smooth, yet sometimes inspiring of radical transitions from traditional to non-traditional approaches*.

Concurrent error detection is inspired by redundancy and parity prediction based approaches to deal with multiple faults. Parity based protection has been in use for long in communication and holds special value in providing limited protection to aging related performance degradation for reliable operation. However, when combined with other approaches to protect more critical elements like say the MSB, the effectiveness of the host approach is substantially increased. Built in self-testing(BIST) community is inspired by coding based approaches like parity and attempts to reduce the cost and reliance on external test equipment by using test pattern generation only as exhaustive as necessary for operation and output analysis in the circuit under test. Roving STARs is a popular example for FPGAs that partitions FPGA into rows and columns and periodically tests a block by taking it offline and re-routing its function to available redundant resources. Often BISTs involve downtimes, but are nevertheless a growing area of research and have in several ways inspired many low overhead CED approaches. Since CED employs several different techniques depending on the situation at hand, one is confronted with the question of which CED technique to choose, which [43] addresses by comparing and contrasting identical and diverse duplication, parity prediction and error codes for different cases projecting diverse duplication as highly appropriate. CED is based on traditional methods but is very forward looking to inspire several modern approaches.

III. Modern Approaches and Recent Research

Soft error resilience is an area of growing importance owing to the increased susceptibility of more complex, smaller technology generations to not one, but often several timing errors which ripple through stages to adversely affect performance. This risk is not only limited to space application, but also at higher altitudes and even terrestrial radiation which can introduce a bit-flip to trigger several hard to detect malfunctions. Memory protection was traditionally the target for soft errors and expensive error detection codes helped with it. Recent research has established, with wide acceptance, an equal proneness in sequential elements as well. Also, traditional approaches to soft-error tolerance focused on SEUs(Single event upsets), while SEMUs(Single event multiple upsets) continue to grow in importance in smaller technologies and thus calls for more robust techniques for soft error resilience, especially for sequential elements. BISER (Built in Soft Error Resilience) based on the work done by Mitra et. al[24] uses a specially designed BISER latch (with an inbuilt redundant latch driven by the same clock) with a specially designed C-element that retains the previous value in case of a discrepancy in output. Weak keeper is also provided

to ensure scalability to serve a large fan-out.

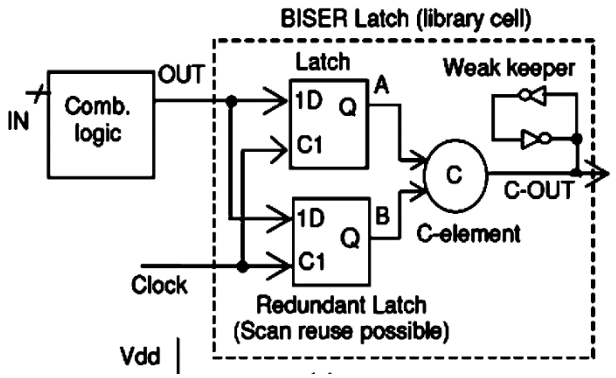

Figure 6: BISER latch operation. Image courtesy of [24]

LEAP(Layout Design Through Error Aware Positioning) [24]is a layout design principle that leverages positioning of NMOs and PMOS diffusion nodes to result in collection of opposite charges and act together to nullify the effect of the particle strike, which otherwise results in negative and positive transients respectively in NMOS and PMOS when it occurs in isolation. LEAP-DICE[42] builds on DICE(dual interlocked storage cell) that provides strong SEU tolerance and combines LEAP based layout to offer strong SEMU resilience as well to the DICE design. Power and area overheads are significant considerations in these approaches to soft error resilience. Arbitrary insertion of soft error resilience techniques can violate these constraints and hence cross layer implementation across circuits, architecture, applications, logic-different levels of system stack can ensure lower overheads and cost-effectiveness. Mitra's work[24] discusses some valuable insertion points which includes selective insertion of BISER and parity protection for different bits and the capacity to turn on or off the BISER capability. Reuse of post silicon validation structures has also been suggested. These approaches in soft error resilience are novel for the reasons discussed above and have much potential when combined effectively with the traditional counterparts.

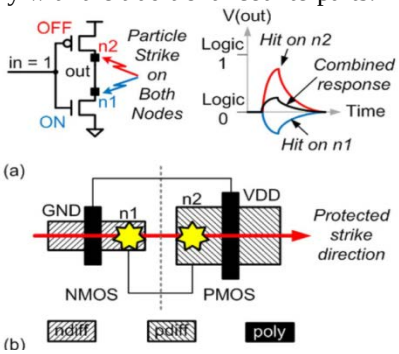

Figure 7: LEAP principle. Image courtesy of [24]

Circuit early life prediction is another area of active research owing to its usefulness in predicting malfunction before the errors occur. It also holds promise to be implemented temporally or spatially (using a few minimal circuit blocks to indicate failures in larger critical blocks). The effectiveness of this approach is based on the fact that circuits exhibit performance signatures that can be utilized to predict failures. For instance, transistor delay signatures can be used to predict gate oxide failures. Mitra[24] uses this by invoking a comparison between outputs generated by system clock and delayed clock by for a given circuit (inverter chain in this experiment) while under greater than normal stress exerted during self-repair/diagnostics mode, which is inserted between normal modes of operation to identify any delay fluctuations and also the corresponding leakage current profile to identify gate oxide ELF. This approach compares with that of canary circuits and is possibly inspired by the same. However, this requires consideration of downtimes and criticality of circuit when inserting downtimes for self-test and diagnostics, which is avoided with the use of canary circuits alone.

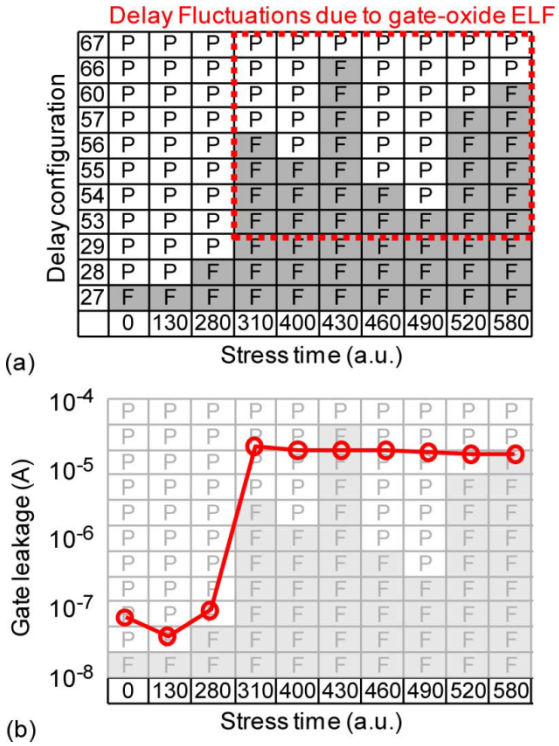

Figure 8:Gate oxide ELF prediction. Image courtesy of [24]

In order to optimize self-testing to implement approaches as discussed above, Mitra also presents CASP(Concurrent Autonomous Chip Self-Test and Diagnostics Using Stored Test Patterns)[24] which uses off chip storage (FLASH) with test compression algorithms and microarchitecture support in terms of resource reallocation and sharing to allocate function of one uncore component to another briefly, use of idled uncore components when the corresponding core component is under test, use of minimal redundancy for mission critical hardware and software support in terms of I/O request handling, OS migration and scheduling for the cores under test, thus orchestrating cross-layer optimization. These approaches serve SoCs and heterogeneous cores very well and hold promise for future designs.

In contrast to the diagnostics based approach aboveand their suitability for heterogeneous cores, dynamic voltage scaling approaches present a radical technique to handle aging in homogeneous multi-core circuits. Leakage currents prevent

scaling down of supply voltage inspite of the increasing number of transistors per unit area. Since power budget per unit area and the temperature for proper operation avoiding aging based degradation are both fixed, miniaturization sees a major challenge in delivering performance in terms of the number of cores that can be powered on simultaneously. This is referred to as the many core power wall. Design of aging guard bands also faces limitation as was discussed before. Work done by Karpuzcu et.al[25] presents BubbleWrap architecture inspired by dynamic voltage scaling for aging management(DVSAM), which tunes supply voltage continuously exploiting any remaining guard band to achieve one of service life/power/performance objectives. Using two sets of cores –throughput (focusing on low power and service life objectives) and expendable(focusing on high performance for short service life as its objective)- this approach achieves both goodput and sequential acceleration and creatively uses aging sensors to dynamically decide on the supply voltages of each core, functionality of each core and dynamic frequency adaptation to satisfy specifications and performance goals at hand. This approach is best described as a creative use of aging and builds upon the work done by Facelift group[41].

IV. Projections for the Future

Sources of variations and variation aware design by Borkar. [18].

The discreteness of dopant atoms in the transistor channel becomes more pronounced in smaller technology generations and hence introduces much variability in electrical characteristics even when identical manufacturing processes are used owing to the breakdown of bulk approximations. Sub-Wavelength lithography uses wavelengths of light larger than the technology (and feature) sizes, and is increasingly in use owing to the unavailability of extreme ultraviolet technology (13nm). This results in the introduction of several variations including line edge roughness. These are the chief static sources of variations in future technologies and affect aging patterns through the inherent variations they introduce. This causes different devices to age differently and newer mechanisms would be required to identify aging variations and innovative strategies built upon such understanding to mitigate the same. In addition to the above, heat flux variations across the die is more extreme for smaller circuits and results in various effects including creation of hotspots, more inductive and resistive voltage drops at unexpected places, dynamic supply voltage variations, variations in sub-threshold leakage and hence in power delivery demand across the die. These dynamic variations would call for runtime detection and adaptation mechanisms to ensure reliability over the device lifetime.

V. Conclusion

The chief aging mechanisms characterizing circuits and future circuit designs have been discussed. Traditionally used approaches and their limitations were identified and discussed and some recent research which builds on these fundamental approaches in the domains of selected microarchitectures, FPGAs and multi core architectures were discussed. With probabilistic and variability aware architectures gaining importance, hopefully aging and its variations would see mitigation in future design paradigms.

Acknowledgment

I would like to thank Dr. Ronald DeMara for guiding me through this study and in providing useful insights from time to time to better understand this dynamic area of research.

References

[1] Blome, J.; Shuguang Feng; Gupta, S.; Mahlke, S., "Self-calibrating Online Wearout Detection," Microarchitecture, 2007. *MICRO 2007. 40th Annual IEEE/ACM International Symposium on* , vol., no., pp.109,122, 1-5 Dec. 2007

[2] J. A. Blome, S. Feng, S. Gupta, and S. Mahlke. Online timing analysis for wearout detection. *In Proc. of the 2nd Workshop on Architectural Reliability (WAR)*, pages 51–60, 2006.

[3] J. R. Black. Mass transport of aluminum by momentum exchange with conducting electrons. In *Proc. of the 1967 International Reliability Physics Symposium*, Nov. 1967.

[4] F.M. d'Heurle and I. Ames, "Electromigration in Single-Crystal Aluminum Films," *Appl. Phys. Lett.* 16, 80–81, Jan. 1970.

[5] I. Blech*:* Electromigration in Thin Aluminum Films on Titanium Nitride*. Journal of Applied Physics*, Vol 47, pp. 1203-1208, April 1976

[6] P. Solomon. Breakdown in silicon oxide-a review. *Journal of Vacuum Science and Technology*, 14(5):1122–1130, Sept. 1977.

[7] D. Dumin. Oxide Reliability: A Summary of Silicon Oxide Wearout, Breakdown, and Reliability. *World Scientific Publishing Co. Pte. Ltd.*, 2002.

[8] A. Avellan and W. H. Krautschneider. Impact of soft and hard breakdown on analog and digital circuits. *IEEE Transactions on Device and Materials Reliability*, 4(4):676–680, Dec. 2004.

[9] Reliability and Failure Analysis. In: Schroder, Dieter K., (2006), 3rd edition, *Semiconductor Material and Device Characterization.* New Jersey: Wiley, pp 689-712.

[10] W. Wang *et al.*, "The impact of NBTI effect on combinational circuit: Modeling, simulation, and analysis," *IEEE Trans. Very Large Scale Integr.(VLSI) Syst.*, vol. 18, no. 2, pp. 173–183, Feb. 2010.

[11] Kiamehr, S.; Amouri, A.; Tahoori, M.B., "Investigation of NBTI and PBTI induced aging in different LUT implementations," *Field-Programmable Technology (FPT), 2011 International Conference on* , vol., no., pp.1,8, 12-14 Dec. 2011

[12] W. C. Elmore. The transient response of damped linear network with particular regard to wideband amplifiers. *Journal of Applied Physics*, 19(1):55–63, Jan. 1948.

[13] A. I. AbouSeido, B. Nowak, and C. Chu. Fitted elmore delay: A simple and accurate interconnect delay model. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 12(7):691–696, July 2004.

[14] Amazon: Outage Due to Hardware Not Hackers [Online]. Available: http://news.cnet.com/8301-1009_3-20025440-83.html

[15] Incisive Media Investments Ltd.,, London, U.K., "PayPal Outage Hits eBayMerchants,"[Online].Available: http://www.v3.co.uk/v3/news/2247234/paypal-outage-hits-ebay

[16] E. R. Alliance. Online survey results: 2001 cost of downtime, 2001.

[17] Gartner data systems conference, Dec. 2005.

[18] Borkar, S., "Designing reliable systems from unreliable components: the challenges of transistor variability and degradation," *Micro, IEEE , vol.25*, no.6, pp.10,16, Nov.-Dec. 2005

[19] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan. Temperature-aware microarchitecture: Modeling and implementation. *ACM Transactions on Architecture and Code Optimization*, 1(1):94–125, 2004.

[20] T. Austin. Diva: a reliable substrate for deep submicron microarchitecture design. In Proc. of the *32nd Annual International Symposium on Microarchitecture*, pages 196–207, 1999.

[21] F. A. Bower, D. J. Sorin, and S. Ozev. A mechanism for online diagnosis of hard faults in microprocessors. In Proc. of the *38th Annual International Symposium on Microarchitecture*, pages 197–208, 2005.

[22] P. Shivakumar, S. Keckler, C. Moore, and D. Burger. Exploiting microarchitectural redundancy for defect tolerance. In *Proc. of the 2003 International Conference on Computer Design*, page 481, Oct. 2003.

[23] Stott, E.; Cheung, P. Y K, "Improving FPGA Reliability with Wear-Levelling," *Field Programmable Logic and Applications (FPL), 2011 International Conference on* , vol., no., pp.323,328, 5-7 Sept.

[24] Mitra, S; Brelsford, K.; Young Moon Kim; Lee, H.-H.K.; Li, Y, "Robust System Design to Overcome CMOS Reliability Challenges," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on* , vol.1, no.1, pp.30,41, March 2011.

[25] Ulya R. Karpuzcu, Brian Greskamp, and Josep Torrellas. 2009. The BubbleWrap many-core: popping cores for sequential acceleration. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 42)*. ACM, New York, NY, USA, 447-458.

[26] T. Austin, D. Blaauw, T. Mudge, and K. Flautner. Making typical silicon matter with razor. *IEEE Computer*, 37(3):57–65, Mar. 2004.

[27] Hongyan Zhang; Bauer, L.; Kochte, M.A.; Schneider, E.; Braun, C.; Imhof, M.E.; Wunderlich, H.-J.; Henkel, J., "Module diversification: Fault tolerance and aging mitigation for runtime reconfigurable architectures," *Test Conference (ITC), 2013 IEEE International* , vol., no., pp.1,10, 6-13 Sept. 2013

[28] R. A. Ashraf, A. Alzahrani, and R. F. DeMara, "Exploring Spatial Redundancy to Mitigate Aging-Induced Timing Degradation," *ACM/EDAC/IEEE 51st Design Automation Conference (DAC)*, San Francisco, California, USA, June 1 – 5, 2014.

[29] N. Khoshavi, R. A. Ashraf, and R. F. DeMara, "Applicability of Power-Gating Strategies for Aging Mitigation of CMOS Logic Paths," *IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS-2014)* submitted on March 28, 2014.

[30] Kunitake, Y.; Sato, T.; Yasuura, H.; Hayashida, T., "Possibilities to miss predicting timing errors in canary flip-flops," *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on* , vol., no., pp.1,4, 7-10 Aug. 2011

[31] T. Kehl, "Hardware Self-Tuning and Circuit Performance Monitoring," *International Conference on Computer Design*, 1993.

[32] M. Gomaa and T. Vijaykumar. Opportunistic transient-fault detection. *In Proc. of the 32nd Annual International Symposium on Computer Architecture*, pages 172–183, June 2005.

[33] T. Vijaykumar, I. Pomeranz, and K. Cheng. Transient-fault recovery via simultaneous multithreading. *In Proc. of the 29th Annual International Symposium on Computer Architecture*, pages 87–98, May 2002.

[34] J. Ray, J. Hoe, and B. Falsafi. Dual use of superscalar datapath for transient-fault detection and recovery. *In Proc. of the 34th Annual International Symposium on Microarchitecture*, pages 214–224, Dec. 2001.

[35] J. Smolens, J. Kim, J. Hoe, and B. Falsafi. Efficient resource sharing in concurrent error detecting superscalar microarchitectures. *In Proc. of the 37th Annual International Symposium on Microarchitecture*, pages 256–268, Dec. 2004.

[36] V. Reddy, S. Parthasarathy, and E. Rotenberg. Understanding prediction-based partial redundant threading for low-overhead, high-coverage fault tolerance. *In 14th International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 83–94, Oct. 2006.

[37] A. Sodani and G. Sohi. Dynamic instruction reuse. *In Proc. of the 25th Annual International Symposium on Computer Architecture*, pages 194–205, June 1998.

[38] N. Wang and S. Patel. Restore: Symptom based soft error detection in microprocessors. *In International Conference on Dependable Systems and Networks*, pages 30–39, June 2005.

[39] B. Zhai et al. Energy Efficient Near-Threshold Chip Multi-Processing. *In Int. Symp. on Low power Electronics and Design*, 2007.

[40] K. Fan et al. Bridging the Computation Gap between Programmable Processors and Hardwired Accelerators. *In Int. Symp. on High Performance Computer Architecture*, February 2009.

[41] A. Tiwari and J. Torrellas. Facelift: Hiding and Slowing Down Aging in Multicores. *In Int. Symp. On Microarchitecture*, November 2008.

[42] Z. Qi and M. R. Stan. NBTI resilient circuits using adaptive body biasing. *In Proceedings of the 18th ACM Great Lakes symposium on VLSI, GLSVLSI '08*, pages 285{290, 2008.

[43] S. Gupta and S. Sapatnekar. Employing circadian rhythms to enhance power and reliability. *ACM Trans. Des. Autom. Electron. Syst.*, 18(3):38:1{38:23, July 2013.

[44] Kiamehr, S.; Amouri, A.; Tahoori, M.B., "Investigation of NBTI and PBTI induced aging in different LUT implementations," *Field-Programmable Technology (FPT), 2011 International Conference on* , vol., no., pp.1,8, 12-14 Dec. 2011

[45] Ridgetop Group's Sentinel Silicon Tehnology: http://ridgetopgroup.com/products/semiconductor/sentinelsilicon.php

[46] Mitra, S; McCluskey, E.J., "Which concurrent error detection scheme to choose ?," *Test Conference, 2000. Proceedings. International* , vol., no., pp.985,994, 2000

[47] Cosmic Circuits, silicon validation for AFE for Wireless MIMO: http://www.chipestimate.com/tech-talks/2010/05/18/Cosmic-Circuits-(a-Cadence-company)-Choosing-the-right-Analog-Front-End-Solution-for-Wireless-MIMO

[48] H. Lee et al., "LEAP: Layout design through error-aware placement for soft-error resilient sequential cell design," in Proc. Int. Rel. Phys. Symp., 2010, pp. 203–212.