INVITED: Optimizing device reliability effects at the intersection of physics, circuits, and architecture

Deepashree Sengupta, Vivek Mishra, and Sachin S. Sapatnekar*
Department of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN, USA.

ABSTRACT

Over the years, there has been tremendous progress in developing new methods for modeling and diagnosing reliability at the level of individual transistors and interconnects. The thrust to propagate these models to higher levels of abstraction to predict the reliability of larger circuits is much more recent. This paper addresses the intersection of physics, circuits, and architecture for reliability modeling and optimization that must come together for cross-layer optimization. For various device reliability phenomena, this paper shows how physical models can be leveraged at the circuit level, or circuit models at the architecture level, to deliver composite solutions that comprehend chip-level design goals.

Keywords

Reliability, bias temperature instability, hot carriers, oxide breakdown, electromigration, cross-layer optimization.

1. INTRODUCTION

Aging-related effects, which cause process drifts or fatal errors in chips over their lifetimes, have become increasingly important in recent years. The classical bathtub curve [1] pictorially explains the effect of temporal variations: after a steep initial failure rate, the number of failures levels off for a while before rising again. Process variations form a special case on this curve, corresponding to the static variations baked into the chip at time zero, while aging variations are represented on the rest of the curve. Many of these variations have strong sensitivity to the on-chip temperature. In recent years, the (formerly) flat region of the bathtub curve in intermediate years has seen by a greater propensity to failure. Therefore, it is essential to consider aging as a first-class objective during all levels of design.

This paper surveys methods for analyzing and optimizing reliability effects, and begins by providing an overview of reliability models at the physical and gate levels. This is followed by a discussion of circuit-level reliability issues: the expression of supposedly catastrophic errors at the device level as parametric faults at the circuit level, interactions between reliability mechanisms, and finally, presilicon and post-silicon circuit level aging mitigation. Finally, we discuss cross-layer system-level optimization where architectural methods are combined with circuit methods, using appropriate device-level models, to improve system reliability.

2. PHYSICS: RELIABILITY MODELS

Bias temperature instability: Bias temperature instability (BTI) is a phenomenon that causes threshold voltage shifts over long periods of time, eventually causing the circuit to fail to meet its specifications. The degradation is caused when a voltage bias is applied across the gate node of a transistor, and is sensitive to the temperature. A PMOS transistor in an inverter experiences negative BTI stress when its gate node is at logic 0, and the resulting increase in the threshold voltage is partially reversed when the voltage stress is removed (i.e., a logic 1 is applied). A similar phenomenon of positive BTI affects the threshold voltage of NMOS devices when they are stressed, and relaxes the degradation on the removal of stress. There are two theories for BTI, based on the reaction-diffusion (R-D) model [2, 3] and charge trapping (CT) [4, 5], with the latter being related to the phenomenon of $1/f^2$ random telegraph noise [6], with fast shifts and large variations.

Empirically, BTI degrades the threshold voltage at the rate of t^n, where t is the stress time and $n \sim 0.1 - 0.2$. The impact of BTI on gate delay shifts can be determined by determining the stress probability (SP), i.e., the probability that a signal is at a stressing level (e.g., logic 0 for a PMOS). The effective stress time, t, is computed by multiplying the age of the circuit by the stress probability, SP.

Hot carrier injection: Hot carrier injection (HCI) effects in MOSFETs are caused by the acceleration of carriers (electrons/holes) under lateral electric fields in the channel, to the point where they gain enough energy and momentum to cause damage, degrading mobilities and threshold voltages. At the device level, the HCI rate increases as $t^{1/2}$, where t is the time variable. Since the proportionality constant is relatively small, in the short term, HCI is overshadowed by BTI effects, where the exponent of t is smaller but the proportionality constant is larger. However, particularly for longer lifetime parts, the impact of HCI can be significant.

The traditional theory of HCI mechanisms was based on a field-driven model where the peak energy of carriers was determined by the lateral field of the channel, based on the theory of the so-called lucky electron model [7], but this does not capture HCI in scaled technologies. Newer energy-

*This work was supported in part by NSF awards CCF-1017778 and CCF-1162267 and by SRC contract 2012-TJ-2234.

DAC '16, June 05 - 09, 2016, Austin, TX, USA

ISBN 978-1-4503-4236-0/16/06…$15.00
http://dx.doi.org/10.1145/2897937.2905016

driven theories [8,9] have been introduced to explain carrier-induced degradation for short-channel devices at low Vdds. These include the effects of electrons of various energies, from high-energy channel hot carriers to medium-energy carriers to low-energy channel cold carriers, and degradation arises chiefly from medium- and low-energy carriers.

For large-scale circuit analysis, some scalable approaches for timing analysis under HC effects have been developed. The approach in [10] applies a duty factor to capture the effective stress time for HC effects, modeling the duty factor to be proportional to the transition time. The method in [11] uses the new energy-driven theories described above, and defines an age gain per transition using quasistatic characterization. Using abstractions based on the SP and activity factor (AF), the effective age is computed, and is used to determine device degradation.

Time-dependent dielectric breakdown: Time-dependent dielectric breakdown (TDDB) in gate oxides, illustrated in Fig. 1, is an irreversible reliability phenomenon that results in a sudden discontinuous increase in the conductance of the gate oxide at the point of breakdown, as a result of which the current through the gate insulator increases significantly. This is of concern as oxide thicknesses become thinner with technology scaling, increasing breakdown susceptibility.

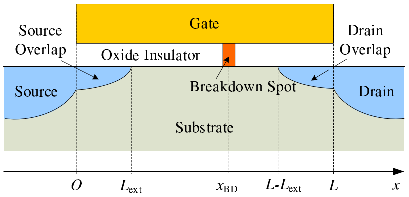


Figure 1: Schematic of gate oxide breakdown.

The time to breakdown can be modeled statistically using a Weibull distribution [12]. At the circuit level, the effect of TDDB on memory has been addressed in [13]. For logic circuits, a conventional area-scaling based method is presented in [14]. However, as will be discussed in Section 3.1, logic circuits are inherently resilient, and the area-scaling model can be pessimistic by about half an order of magnitude [15, 16].

Electromigration: When a current flows through an on-chip wire over a long period of time, it can cause a physical migration of atoms in the wire, particularly if the current density is high. The current conducting electrons can form an electron wind, which leads to momentum exchange with the constituent atoms of metal. This effect will lead to a net flux of metal atoms in the direction of electron flow (opposite of current direction), creating voids (depletion of material) upstream and hillocks (accumulation of material) downstream at locations of atomic flux divergence. Electromigration can cause uneven redistribution of resistance, dielectric cracking, and undesired open circuits.

3. CIRCUITS: ANALYSIS/OPTIMIZATION

3.1 Catastrophic vs. parametric faults

Traditionally, BTI and HCI have been considered to be parametric faults that can alter the performance of a circuit but not its functionality. By degrading the threshold voltage and drive current of a transistor, these phenomena result in a reduction in the speed of a circuit. On the other hand, TDDB and EM are often regarded as catastrophic faults, in that one failure can render a circuit nonfunctional.

In this section, we show that the dichotomy between parametric and catastrophic faults is not as stark when one considers the impact of these faults at the circuit level. Specifically, we point out that the weakest-link approach that has been used for reliability analysis of TDDB and EM, whereby the system fails when a single part fails, is far too pessimistic and therefore results in overdesign.

For the case of TDDB, consider the scenario [15] where a fault is induced in a transistor, as shown in an NMOS transistor in Cell n of Fig. 2. The breakdown is modeled using the resistors R_d and R_s whose values depend on the location of the breakdown on the transistor gate, i.e., whether it occurs closer to the source or the drain of the transistor, as illustrated in Fig. 1. This structure induces a resistive divider in the circuit, whereby the PMOS transistors in Cell m try to drive its output to logic 1, while the breakdown attempts to bring it down to logic 0, and the stronger of the two wins. Specifically, a breakdown that is close to the source provides an easy path for the output of Cell m to be incorrectly discharged – a catastrophic fault, but one that is in the middle of the transistor or close to the drain may be likely to preserve the logic 1, but slow down the output transition for Cell m – a parametric fault.

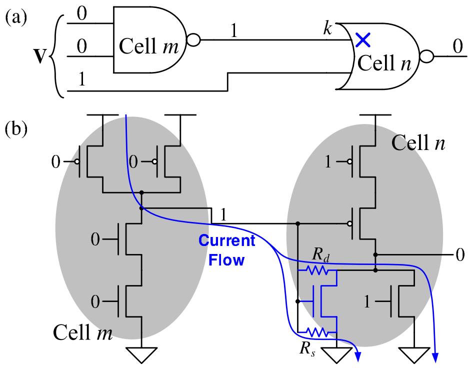


Figure 2: Cell-level analysis of the breakdown case.

A second example relates to the effect of electromigration in power grids or clock grids. A break in a wire can result in an open-circuit, but in a highly-redundant circuit, such as a power grid or a meshed clock grid, alternative pathways for carrying current can help mitigate the impact of this open-circuit. Specifically, it has been shown that if one considers circuit-level metrics such as the IR drop in power grid [17] or the clock skew in clock meshes [18], electromigration may cause small variations in these parameters, tantamount to a parametric shift, rather than a catastrophic failure.

3.2 Interactions between aging phenomena

To a great extent, reliability effects have been handled independently in the literature [19] since their root causes are typically orthogonal: BTI is caused by stress during the non-switching state, while HCI relates to switching events; BTI, HCI, and gate oxide TDDB are device-level effects while EM is an interconnect effect, and so on. While this is largely true, it is possible to see relationships between the impact of these aging phenomena at the circuit level.

As an example, consider the effect of AC EM degradations in a wire. This degradation is related to the average current density in a wire, accounting for a recovery factor (which determines the electron wind force) and the RMS value (which determines Joule heating, which further degrades EM). For an EM-susceptible wire, a breakdown becomes more likely as the wire ages. However, the device is driven by a gate whose transistors also degrade with time, implying that the current carried through the wire reduces with time.

	Normalized J (a.u)	Normalized TTF (a.u)
With BTI	1	1
Without BTI	0.99	1.04

Table 1: How driver BTI affects wire EM.

Table 1 shows a preliminary result for a 32X sized inverter (INVX_32) connected to a minimum width, 200um wire in 16nm technology. The normalized average EM current density, J, and the normalized EM time to failure (TTF) for the wire are shown for two cases: when BTI on the driver is ignored, and when it is considered. Even though the threshold voltage of the device shifts significantly over this period, the impact on the average current is very slight. This can be explained by the fact that switching current flows only over a small fraction of an entire cyclee therefore, in computing the average current, any changes in the nonzero currents are attenuated by the long periods of zero current. Despite the small reduction in average current density, there is a visible impact on the mean TTF.

3.3 Presilicon design for device aging

Since the delay of a circuit increases due to aging-induced parametric shifts in the drive current, a simple way to inoculate a circuit against failure is to provide it with sufficient delay margin to incorporate the impact of aging. An example is illustrated in Fig. 3 [20]. The topmost curve shows the results for the case where the circuit is designed to meet specifications at time zero, but no aging margin is maintained; clearly, this cannot be guaranteed to meet timing over the part lifetime. The next lower curve shows the case where the stress probabilities (SPs) for all gates are known, in which case the delay margin can be adjusted exactly to ensure that the circuit meets specifications throughout its lifetime (in this case, 10 years, as shown by the dotted line to the right). However, predicting the SP for all manufactured parts over all workloads is difficult, and therefore, it is common to use a worst-case SP over all workloads. This yields a larger delay margin, shown by the bottom-most curve.

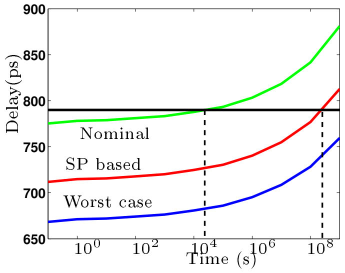


Figure 3: Circuit delay degradation under BTI.

The impact of variations on BTI depends on the specific mechanism being considered. While variations under the R-D mechanism are relatively small [21–23], they can be significant under the CT mechanism. Charge trapping and detrapping at each defect are random events that are characterized by the capture and emission time constants. The statistical variation in $\Delta V_{\rm th}$ depends not only on the statistics of these random events, but also the distribution of defects within a device, which can vary randomly [5, 24, 25]. Under these statistical perturbations, the variation of device lifetime can be extremely large for devices with a smaller number of defects N, as illustrated in Fig. 4.

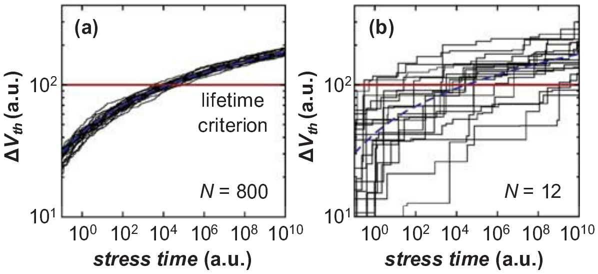


Figure 4: [26]: (a) Narrow distribution of lifetime in large devices where randomness averages out. (b) Large variation of lifetime in small devices where stochasticity predominates.

The impact of CT-induced variations has been examined in [27], where it is found that that although device-level variations can be large, their impact at the circuit level is significantly attenuated and is reasonable. There are two reasons for this. First, a typical critical path contains a large number of stages, which attenuates the impact of variations (recall that the ratio of variance to mean for a sum of i.i.d. random variables decreases with the number of terms in the sum). Second, transistors on critical paths are typically upsized, and as seen in Fig. 4, the $\Delta V_{\rm th}$ variations for larger devices are significantly smaller than for small devices.

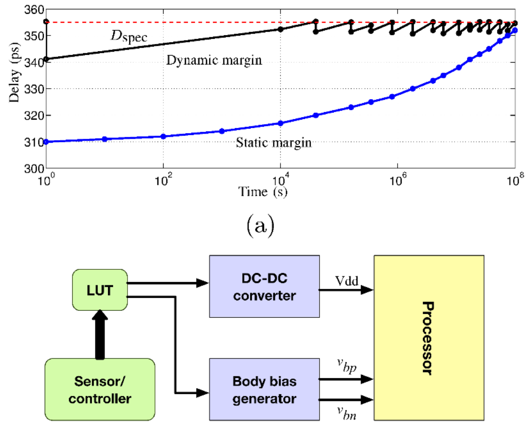
Some approaches have suggested the use of setting up specific sleep states that are designed to minimize BTI degradation through input vector control. However, the gains of such methods are relatively small.

3.4 Post-silicon design for device aging

Adaptive post-silicon techniques are an effective means for protecting circuit functionality from BTI degradation. Published methods based on adaptive margins have used time sensors [28], history sensors that track usage patterns [29], or surrogate sensor circuits [30–36]. Optimization knobs include adjustments in the clock frequency, supply voltages, and body biases [29, 37, 38]. Unlike a static margin, with its large power overhead in the early life due to the large margin, dynamic margins, illustrated in Fig. 5(a), use just enough margin and limit the power overhead at each instant.

A schematic of such a system in shown in Fig. 5(b). Based on sensor inputs, circuit performance may be dynamically recovered by, for example, changing the supply voltages and body biases through a look-up table. The simplest sensor is a simple time-based sensor, where worst-case aging is assumed and the circuit is compensated at regular time intervals.

An alternative is to use data from surrogate sensors, built in at the presilicon phase and tested at the post-silicon stage, to adaptively provide on-the-fly compensation to mitigate the effects of aging. These sensors range from simple inverter chain or ring oscillator (ROSC) circuits [30–34] to more complex circuits [35,36]. A particularly promising concept is the notion of silicon odometers [31].



(a)

(b)

Figure 5: (a) A control scheme for runtime circuit adaptation through Vdd and biases (v_{bn}, v_{bp}) (b) A dynamic schedule for speeding up an aged circuit.

To some extent, surrogate sensors may successfully capture the environment faced by the circuit. If they are placed close to the circuit and have a similar connection to the power grid, they can capture the thermal and supply voltage environment, and undergo similar shifts due to systematic or spatially-correlated process variations. However, surrogate sensors are unable to reflect aging in the circuit with complete accuracy due to the structural differences between the near-critical paths of the circuit and the sensor, including differences in the stressing patterns due to structural differences in the depth and type of gates on the critical path.

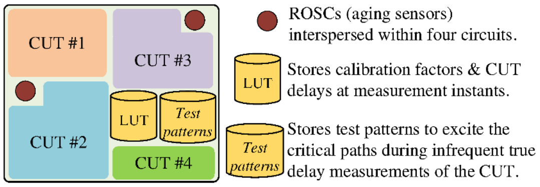

CUT #1
CUT #3
CUT #2
LUT
Test patterns
CUT #4
ROSCs (aging sensors) interspersed within four circuits.
LUT
Stores calibration factors & CUT delays at measurement instants.
Test patterns
Stores test patterns to excite the critical paths during infrequent true delay measurements of the CUT.

Figure 6: ROSCs interspersed within multiple circuits along with LUT of degradation ratio and test pattern storage block for direct CUT measurement.

A scheme based on this idea [39, 40] is depicted in Fig. 6, which shows multiple silicon odometer ROSC sensors interspersed within four circuit blocks, referred to as circuits under test (CUTs). The granularity, in terms of number and locations, of the ROSCs within a chip reflects a trade-off between area overhead and accuracy. Measurements from these surrogate aging sensors are translated to circuit delay degradations by multiplying them with a *degradation ratio* that is stored in a look-up table (LUT).

The scheme in [39] obtains the degradation ratios using an Upper-bound on $\mathcal{F}_{Max}$ (UofM) model that estimates a safe maximum frequency, $\mathcal{F}_{Max}$, for an aging CUT. This model accounts for the possibility that critical paths may change over the lifetime of a chip due to nonuniform delay degradation on various circuit paths, and therefore the maximum CUT delay, $D^C(t)$ is piecewise smooth. The UofM model is a smooth envelope, $D^C_{est}(t)$, for the CUT delay, shown in Fig. 7(a), that provides a tight upper-bound on the circuit delay. The degradation ratio can be multiplied by the measured ROSC degradation to obtain the CUT degradation.

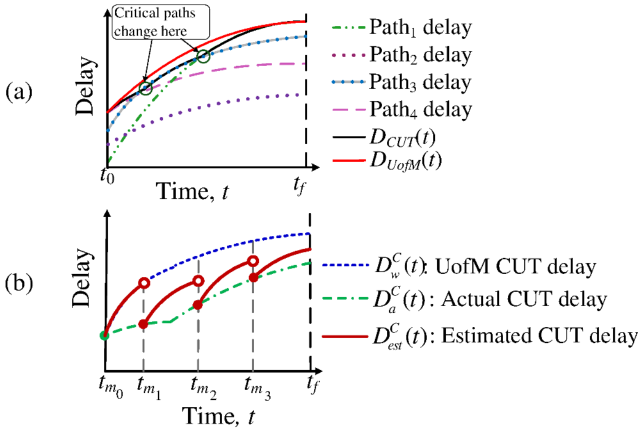


Figure 7: (a) CUT delay as maximum of path delays under aging. (b) Aging estimate with intermediate CUT measurements.

The UofM approach is based on worst-case workload assumptions and is agnostic to how the part is operated in the field, which influences its SP (for BTI) and activity factor (for HCI). The ReSCALE [40] approach is an alternative that reduces this pessimism by performing inexpensive and infrequent direct delay measurements on the CUT (it is shown that 2-4 measurements over the lifetime of a chip are adequate) to *recalibrate* the degradation ratios in the LUT. The delay measurement circuitry for the CUT requires the application of stimuli, stored in the *Test patterns* block, and schemes to measure the CUT delay under these stimuli using methods such as the Path-RO [41], delay shift circuits [42, 43], or techniques described in [29].

Fig. 7(b) shows the actual delay, $D^C_a(t)$, for a circuit under a real workload. The UofM prediction, $D^C_w(t)$, is pessimistic, and by recalibrating the degradation ratio at times $\{t_{m_1}, t_{m_2}, t_{m_3}\}$, a closer estimate, $D^C_{est}(t)$, is obtained.

4. SYSTEMS: OPTIMIZATION

At the system level, the optimization of reliability effects reduces to a management problem. One class of approaches [44–48] views the problem as one of operating the system within parameters that will guarantee lifetime reliability, or scheduling tasks to maximize reliability. Another class actively uses system-level flexibilities to limit the impact of aging. In particular, the latter class uses the notion of recovery that is associated with some aging phenomena, particularly the dominant device phenomenon of BTI. We will focus mostly on such methods in our discussion.

Wear-leveling: Wearout in a chip can influence various parts of the system unevenly, and the wear-leveling methods create internal system modifications that are likely to even out the impact of aging. This could be achieved by rebalancing the stress probabilities in a circuit to even out the wearout on pull-up and pull-down paths in a circuit.

As a concrete example, consider the core of an SRAM cell, which is a pair of back-to-back inverters. If the same logic value is stored for an extended period in the cell, then it degrades the PMOS of one inverter, while the PMOS in the other inverter is not stressed. This causes uneven wearout, degrading the signal-to-noise margin (SNM) of the cell.

If the contents of the SRAM cell are *deliberately* flipped periodically, then we ensure that each inverter is degraded more evenly. For values that do not normally change over long periods, such an inversion forces wear-leveling, while values that do change frequently automatically induce wear-leveling and are unaffected by this transform. In principle, a specific cell may coincidentally switch almost exactly at the same times when the contents are flipped, so that it holds a constant value throughout – but failure due to this improbable instance can be handled by standard error-correction mechanisms. It is shown in [49] that 80% of performance (read stability) can be restored through this procedure.

A modified version of this scheme that alters the SRAM cell to enhance recovery is presented in [50]. For arithmetic circuits in the datapath of a microprocessor, a similar idea to balance the SP values has been proposed in [51]. The Penelope approach [52] uses a set of available resources, such as idle pipelines, cache blocks, registers, and ports to storage structures, and writes targeted values into these to induce wear-leveling. An alternative method deactivates memory units on a rotating basis to enable recovery cycles [53].

Circadian rhythms: Under the human circadian rhythm, a period of sleep between periods of work represents a stage of relaxation that decelerates aging and allows more vigorous exertions during the cycle of wakefulness. A similar argument is made for an inanimate circuit in [54, 55].

Under traditional models, increasing the supply voltage, Vdd, of a circuit accelerates aging. However, this elevated voltage, referred to as Greater than NOMinal Operation (GNOMO), also allows tasks to be completed more quickly, and therefore, a period of sleep can be *deliberately* inserted into the task schedule, when Vdd is gated and the circuit recovers from BTI degradation [56, 57]. We explore this tradeoff in Fig. 8, for the case where a different baseline voltage, $V_{dd,n} = 0.8$V, is used, and several $V_{dd,g}$ values are considered. A higher value of $V_{dd,g}$ implies greater degradation during the compute period, and a larger idle time.

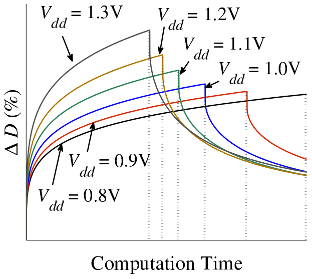


Figure 8: The delay degradation patterns of MCNC benchmark alu4 for $V_{dd} \in$ [0.8V, 1.3V].

If we extend this idea to running a workload on a processor, let us say that the completion time at the nominal and GNOMO supply voltage is t_n and t_g, respectively; typically these will correspond to a few milliseconds, i.e., millions of clock cycles. At iso-performance, the period of sleep under GNOMO is chosen to be $t_n - t_g$. Power savings are achieved through inactivity during the sleep period and from reduced delay margining overheads under this reduced aging.

At the microarchitecture level, several issues must be accounted for. First, the granularity of sleep and wakeup cycles is bounded both above and below. If the transitions to sleep cycles are too frequent, the overhead of wakeup time (typically tens of cycles) must be considered; if they are too infrequent (typically over 10 million cycles), then the rate of change is above the thermal time constant and the peak temperature for the chip may rise – further worsening reliability. Second, if the processor goes to sleep midway through a computation, state information may be lost and the program may have to go back to an earlier checkpoint. This can be avoided by storing state elements (registers and caches) in drowsy mode. Third, at the GNOMO Vdd value, the processor runs at a faster frequency, but external peripherals are unchanged, implying that an off-chip operation costs a larger number of clock cycles. This is either a cost that must be absorbed, or can be overcome through adapted latency-hiding methods. Another alternative is to reschedule sleep cycles by using, for example, predictors to overlap sleep periods with stall cycles, or even changing code to leverage sleep cycles by bunching together memory accesses into periods where the processor can sleep.

At the circuit level, GNOMO enables a reduction of about 25%-40% in delay degradation, translating to lower guardbands that can result in 1.7× to 2.7× lower power. At the microarchitecture level, counting all overheads, up to 13.5% system-level power savings are demonstrated.

The basic GNOMO approach does not require the detection of idle times since the idle times are *generated*, and not *detected*, and are hence predictable by construction. The method can be supplemented by detecting idle times that dynamically occur during workload execution (due to cache misses, branch mispredictions, etc.). GNOMO bears superficial similarity to race-to-halt (RTH) methods [58, 59], which perform a computation as fast as possible and then enter sleep mode to save leakage power. However, RTH methods do not explicitly consider aging, and nor do they specifically leverage recovery in circuit speed, as is the case for GNOMO.

An enhancement of the GNOMO idea is presented in [60]. Instead of merely putting the circuit to sleep and allowing passive recovery, the circuit is rejuvenated using active self-healing, by applying negative Vdd and high temperature.

Disposable cores: The BubbleWrap approach [61], applied to a homogeneous manycore processor, leverages the spread in core performance and lifetime due to process variations. It classifies cores into two categories. Throughput cores, which consume the least power at the target frequency, are used to run the parallel sections of the application at normal Vdd values, achieving high throughput. Expendable cores are dedicated to run sequential segments at an elevated Vdd, achieving high single-thread performance. However, such a core also ages rapidly: once it fails specifications, it is discarded and replaced by another expendable core.

5. CONCLUSION

Device-related aging effects have become increasingly important in recent years, and this paper provides a flavor for solutions that bring together device modeling, circuit analysis and optimization, and system optimization techniques. Future progress will rest on continuing to break down these barriers to provide true cross-layer solutions.

6. REFERENCES

[1] J. M. Carulli, Jr. and T. J. Anderson, “The impact of multiple failure modes on estimating product field reliability,” *IEEE Des. Test*, vol. 23, no. 2, pp. 118 – 126, 2006.

[2] M. A. Alam and S. Mahapatra, “A comprehensive model of PMOS NBTI degradation,” *Microelectronics Reliab.*, vol. 45,

no. 1, pp. 71 – 81, Jan. 2005.

[3] S. V. Kumar *et al.*, "An analytical model for negative bias temperature instability (NBTI)," in *Proc. ICCAD*, 2006, pp. 493 – 496.

[4] T. Grasser *et al.*, "Recent advances in understanding the bias temperature instability," in *Proc. IEDM*, 2010, pp. 4.4.1 – 4.4.4.

[5] B. Kaczer *et al.*, "Origin of NBTI variability in deeply scaled pFETs," in *Proc. IRPS*, 2010, pp. 26 – 32.

[6] T. Matsumoto *et al.*, "Impact of random telegraph noise on CMOS logic circuit reliability," in *Proc. ASP-DAC*, 2015.

[7] S. Tam *et al.*, "Lucky-electron model of channel electron injection in MOSFETs," *IEEE Trans. Electron Devices*, vol. D-31, no. 9, pp. 1116 – 1125, Sep. 1984.

[8] S. E. Rauch, III and G. La Rosa, "The energy-driven paradigm of NMOSFET hot-carrier effects," *IEEE T. Device Mater. Rel.*, vol. 5, no. 4, pp. 701 – 705, Dec. 2005.

[9] C. Guerin *et al.*, "The energy-driven hot-carrier degradation modes of nMOSFETs," *IEEE T. Device Mater. Rel.*, vol. 7, no. 2, pp. 225 – 235, 2007.

[10] D. Lorenz *et al.*, "Aging analysis of circuit timing considering NBTI and HCI," in *Proc. IOLTS*, 2009, pp. 3 – 8.

[11] J. Fang and S. S. Sapatnekar, "The impact of hot carriers on timing in large circuits," in *Proc. ASP-DAC*, 2012, pp. 591 – 596.

[12] E. Y. Wu *et al.*, "CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics," *IBM J. Res. Dev.*, vol. 46, no. 2/3, pp. 287 – 298, March–May 2002.

[13] F. Ahmed and L. Milor, "Analysis and on-chip monitoring of gate oxide breakdown in SRAM cells," *IEEE T. VLSI Syst*, vol. 20, no. 5, pp. 855 – 864, May 2012.

[14] K. Chopra *et al.*, "A statistical approach for full-chip gate-oxide reliability analysis," in *Proc. ICCAD*, 2010, pp. 698 – 705.

[15] J. Fang and S. S. Sapatnekar, "Scalable methods for the analysis and optimization of gate oxide breakdown," in *Proc. ISQED*, 2010, pp. 638 – 645.

[16] J. Fang and S. S. Sapatnekar, "Accounting for inherent circuit resilience and process variations in analyzing gate oxide reliability," in *Proc. ASP-DAC*, 2011, pp. 689 – 694.

[17] V. Mishra and S. S. Sapatnekar, "The impact of electromigration in copper interconnects on power grid integrity," in *Proc. DAC*, 2013, pp. 88:1–88:6.

[18] P. Jain *et al.*, "Stochastic and topologically aware electromigration analysis for clock skew," in *Proc. IRPS*, 2013, pp. 3D.4.1–3D.4.8:6.

[19] H. Amrouch *et al.*, "Towards interdependencies of aging mechanisms," in *Proc. ICCAD*, 2014, pp. 478–485.

[20] S. V. Kumar *et al.*, "NBTI-Aware Synthesis of Digital Circuits," in *Proc. DAC*, 2007, pp. 370 – 375.

[21] S. E. Rauch, "The statistics of NBTI-induced V_T and β mismatch shifts in pMOSFETs," *IEEE T. Device Mater. Rel.*, vol. 2, no. 4, pp. 89–93, Dec. 2002.

[22] K. Kang *et al.*, "Estimation of statistical variation in temporal NBTI degradation and its impact on lifetime circuit performance," in *Proc. ICCAD*, 2007, pp. 730 – 734.

[23] B. Vaidyanathan *et al.*, "Intrinsic NBTI-variability aware statistical pipeline performance assessment and tuning," in *Proc. ICCAD*, Nov. 2009, pp. 164–171.

[24] H. Reisinger *et al.*, "The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress," in *Proc. IRPS*, May 2010, pp. 7–15.

[25] M. Toledano-Luque *et al.*, "Response of a single trap to AC negative bias temperature stress," in *Proc. IRPS*, Apr. 2011, pp. 4A.2.1–4A.2.8.

[26] B. Kaczer *et al.*, "Atomistic approach to variability of bias temperature instability in circuit simulations," in *Proc. IRPS*, 2011, pp. XT.3.1 – XT.3.5.

[27] J. Fang and S. S. Sapatnekar, "Understanding the impact of transistor-level BTI variability," in *Proc. IRPS*, 2012, pp. CR2.1 – CR2.6.

[28] S. V. Kumar *et al.*, "Adaptive techniques for overcoming performance degradation due to aging in digital circuits," in *Proc. ASP-DAC*, 2009, pp. 284 – 289.

[29] E. Mintarno *et al.*, "Self-tuning for maximized lifetime energy-efficiency in the presence of circuit aging," *IEEE T. Comput. Aid D.*, vol. 30, no. 5, pp. 760–773, 2011.

[30] T. H. Kim *et al.*, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," *IEEE J. Solid-St. Circ.*, vol. 4, no. 4, pp. 874 – 880, Apr. 2008.

[31] J. Keane *et al.*, "An all-in-one silicon odometer for separately monitoring HCI, BTI, and TDDB," *IEEE J. Solid-St. Circ.*, vol. 45, no. 4, pp. 817 – 829, Apr. 2010.

[32] K. K. Kim *et al.*, "On-chip aging sensor circuits for reliable nanometer MOSFET digital circuits," *IEEE T. Circuits-II*, vol. 57, no. 10, pp. 798–802, 2010.

[33] T. Iizuka *et al.*, "Buffer-ring-based all-digital on-chip monitor for PMOS and NMOS process variability and aging effects," in *Proc. DDECS*, 2010, pp. 167–172.

[34] T. B. Chan *et al.*, "DDRO: A novel performance monitoring methodology based on design-dependent ring oscillators," in *Proc. ISQED*, 2012, pp. 633–640.

[35] Q. Liu and S. S. Sapatnekar, "Synthesizing a representative critical path for post-silicon delay prediction," in *Proc. ISPD*, 2009, pp. 183–190.

[36] S. Wang *et al.*, "Representative critical reliability paths for low-cost and accurate on-chip aging evaluation," in *Proc. ICCAD*, 2012, pp. 736–741.

[37] S. V. Kumar *et al.*, "Adaptive techniques for overcoming performance degradation due to aging in CMOS circuits," *IEEE T. VLSI Syst*, vol. 19, no. 4, pp. 603–614, 2011.

[38] L. Zhang and R. P. Dick, "Scheduled voltage scaling for increasing lifetime in the presence of NBTI," in *Proc. ASP-DAC*, 2009, pp. 492 – 497.

[39] D. Sengupta and S. S. Sapatnekar, "Predicting circuit aging using ring oscillators," in *Proc. ASP-DAC*, 2014, pp. 430–435.

[40] D. Sengupta and S. S. Sapatnekar, "ReSCALE: Recalibrating sensor circuits for aging and lifetime estimation under BTI," in *Proc. ICCAD*, 2014, pp. 492–497.

[41] X. Wang *et al.*, "Path-RO: A novel on-chip critical path delay measurement under process variations," in *Proc. ICCAD*, 2008, pp. 640–646.

[42] M. Agarwal *et al.*, "Circuit failure prediction and its application to transistor aging," in *IEEE VLSI Test Symp.*, 2007, pp. 277–286.

[43] Y. Li *et al.*, "CASP: Concurrent autonomous chip self-test using stored test patterns," in *Proc. DATE*, 2008, pp. 885 – 890.

[44] J. Srinivasan *et al.*, "The case for lifetime reliability-aware microprocessors," in *Proc. ISCA*, 2004, pp. 276 – 287.

[45] A. Tiwari and J. Torrellas, "Facelift: Hiding and Slowing Down Aging in Multicores," in *Proc. MICRO*, 2008, pp. 129 – 140.

[46] F. Oboril *et al.*, "Reducing NBTI-induced processor wearout by exploiting the timing slack of instructions," in *Proc. CODES/ISSS*, 2012.

[47] A. Rahimi *et al.*, "Aging-aware compiler-directed VLIW assignment for GPGPU architectures," in *Proc. DAC*, 2013.

[48] D. Gnad *et al.*, "Hayat: Harnessing dark silicon and variability for aging deceleration and balancing," in *Proc. DAC*, 2015.

[49] S. V. Kumar *et al.*, "Impact of NBTI on SRAM read stability and design for reliability," in *Proc. ISQED*, 2006, pp. 210 – 218.

[50] T. Siddiqua and S. Gurumurthy, "Enhancing NBTI recovery in SRAM arrays through recovery boosting," *IEEE T. VLSI Syst*, vol. 20, no. 4, pp. 616–629, Apr. 2012.

[51] E. Gunadi *et al.*, "Combating aging with the Colt duty cycle equalizer," in *Proc. MICRO*, 2010, pp. 103 – 114.

[52] J. Abella *et al.*, "Penelope: The NBTI-aware processor," in *Proc. MICRO*, 2007, pp. 85 – 96.

[53] J. Shin *et al.*, "A proactive wearout recovery approach for extending microarchitectural redundancy to extend cache SRAM lifetime," in *Proc. ISCA*, 2008, pp. 353 – 362.

[54] S. Gupta and S. S. Sapatnekar, "GNOMO: Greater-than-NOMinal Vdd Operation for BTI mitigation," in *Proc. ASP-DAC*, 2012, pp. 271 – 276.

[55] S. Gupta and S. S. Sapatnekar, "Employing circadian rhythms to enhance power and reliability," *ACM T. Des. Automat. El.*, vol. 18, no. 3(38), Jul. 2013.

[56] A. Calimera *et al.*, "NBTI-aware clustered power gating," *ACM T. Des. Automat. El.*, vol. 16, no. 1, Nov. 2010.

[57] A. Calimera *et al.*, "Design techniques for NBTI-tolerant power-gating architectures," *IEEE T. Circuits-II*, vol. 59, no. 4, pp. 249 – 253, 2012.

[58] G. Dhiman *et al.*, "Analysis of dynamic voltage scaling for system level energy management," in *Proc. Hotpower*, 2008.

[59] R. Efraim *et al.*, "Energy aware race to halt: A down to earth approach for platform energy management," *IEEE Comput. Archit. Lett.*, vol. 13, no. 1, pp. 25–28, Jan 2014.

[60] X. Guo *et al.*, "Modeling and experimental demonstration of accelerated self-healing techniques," in *Proc. DAC*, 2014.

[61] U. R. Karpuzcu *et al.*, "The BubbleWrap many-core: Popping cores for sequential acceleration," in *Proc. ISCA*, 2009, pp. 447 – 458.