Processor-level Reliability Simulator for Time-Dependent Gate Dielectric Breakdown

Chang-Chih Chen, Taizhi Liu, Soonyoung Cha and Linda Milor

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA USA

changchih@gatech.edu and linda.milor@ece.gatech.edu

Abstract— Time-dependent gate dielectric breakdown (TDDB) is a leading reliability concern for modern microprocessors. In this paper, a framework is proposed to analyze the impact of TDDB on state-of-art microprocessors and to estimate microprocessor lifetimes due to TDDB. Our methodology finds the detailed electrical stress and temperature of each device within a microprocessor system running a variety of standard benchmarks. Combining the electrical stress profiles, thermal profiles, and device-level models, we perform timing analysis on the critical paths of a microprocessor using our methodology to characterize microprocessor performance degradation due to TDDB and to estimate the lifetime distribution of logic blocks. In addition, we study DC noise margins in conventional 6T SRAM cells as a function of TDDB degradation to estimate memory lifetime distributions. The lifetimes of memory blocks are then combined with the lifetimes of logic blocks to provide an estimate of the system lifetime distribution.

Keywords—microprocessor; reliability; modeling; timing analysis; aging; SRAM; cache; gate oxide breakdown; time-dependent dielectric breakdown

I. INTRODUCTION

Increasing operating temperatures and electrical fields, combined with the scaling of dimensions, have contributed to faster aging due to wearout. Not only does this result in the shortening of microprocessor lifetimes, it leads to faster wearout resultant performance degradation with operating time. Microprocessor lifetime is a function of both device and backend wearout. Since previous work has studied the impact of backend wearout mechanisms, bias temperature instability and hot carrier injection on microprocessors [1]–[5], this paper focuses on time-dependent gate dielectric breakdown (TDDB). TDDB is a failure mechanism in MOSFETs, when the gate dielectric breaks down as a result of long-term application of electric field. The breakdown is caused by the formation of conducting paths through the gate dielectric to the substrate due to electron tunneling currents.

The analysis of TDDB is different than the analysis of backend wearout mechanisms, such as electromigration and stress-induced voiding. Backend wearout mechanisms result in open and short circuits, which cause system failure directly. Evidence indicates that electromigration and stress-induced voiding failures are sudden failures [6],[7]. Hence it is sufficient to model the time-to-failure of components of the system and to combine them statistically. TDDB, on the other hand, causes a gradual weakening of the devices. The weakening is both random and a function of stress and temperature. However, unlike backend wearout mechanisms, the relationship between degradation and the circuit

performances must be taken into account to determine the lifetime distribution.

Device lifetime is a function of two kinds of stress: electrical and thermal. An increase in either of the two results in decreased device reliability. The increase in device densities has been achieved through a reduction in device dimensions, which means that the devices undergo increased electrical stress during their lifetime. The resulting increase in operating frequency, as well as device densities, has led to greater thermal stress, which also increases with each new generation. A decrease in device reliability and the increase in system complexity translate into systems whose lifetime characterization is both challenging due to the large number of devices that degrade simultaneously and extremely critical.

The physics describing device failure mechanisms has matured as a result of years of refinement to existing theories. However, the extension of these models to large and complex microprocessor systems is not straightforward. Microprocessor system reliability analysis requires techniques to extend the results gathered from small test structures to large complex microprocessors. Such an endeavor requires methods to manage the deluge of data that comes with analyzing large numbers of devices degrading at different rates.

The purpose of this paper is to present a methodology to assess microprocessor lifetimes and circuit performances due to TDDB by developing the link between the device-level wearout models and the circuit performance requirements while taking into account realistic use scenarios [8]. The lifetime results are not based on degradation limits selected by process engineers, but instead by the time-to-failure of circuit performances.

Since TDDB is activity and temperature dependent, the proposed framework determines the average electrical stress and temperature of each device within the system by running a variety of standard benchmarks. Microprocessors contain both logic and SRAM components. Hence, both types of blocks are considered in this work.

We combine the electrical stress profiles, thermal profiles and device-level models, and apply statistical timing analysis (incorporating process variations) to characterize microprocessor performance degradation of logic blocks. Similarly, electrical and thermal stress profiles are combined with device-level models to determine the memory cell time-to-failure. The memory cell lifetime distributions are combined, taking into account the use of error correcting codes, to determine the overall memory lifetime. This work on computing the lifetime distributions for both logic and memory enables a designer to make any updates in the design

to enhance reliability prior to committing a design to manufacture.

It is well understood that when the gate dielectric layer abruptly loses its insulating properties, it's called hard breakdown (HBD), and HBD can be detected as a large jump in the current vs. time curve. Prior to hard breakdown, TDDB experiences soft breakdown (SBD) where the leakage current in the gate dielectric slightly increases with time, while the gate dielectric still remains its insulation property. SBD is observed as a gradual weakening of the devices. This furthermore can cause failures of the system as the circuit performances degrade when the degradation is sufficient to cause timing errors and to degrade the static noise margins beyond performance specifications. In this work, as in prior work [9], SBD is modeled by inserting a gate-to-source resistance (R_{G2S}) or gate-to-drain resistance (R_{G2D}) in a target gate in order to create the current leakage path in the circuit. The values of the resistance are a function of the number of SBD paths. The number of SBD paths is determined with a percolation model to count the number of conduction paths and the time to soft breakdown (SBD) for each stress scenario. The number of soft breakdown paths translates into a resistance, which in turn impacts timing and memory performances. This paper presents a methodology to determine logic and memory circuit tolerance to soft breakdown events for the first time, in addition to the computation of the corresponding lifetime distributions.

Timing analysis is implemented in [10], including the updating of path selection throughout the aging process. However, prior work has involved smaller circuits and assumptions about stress distributions for each device [10],[11]. In this work we have used emulation to handle large systems running actual benchmarks to determine the actual activity of circuits and memory cells under realistic use scenarios. The results from emulation are used to update timing analysis and the analysis of memory performances based on actual usage patterns. Additionally, a methodology to compute the lifetime distribution of memory blocks is presented for the first time, and memory and logic lifetimes are combined to obtain the system lifetime.

This work not only accounts for activity and temperature, but also accounts for the fact that processors are not in operation at all times. Modern processors operate at varying frequencies and voltages in response to demands to conserve power. Our processor examples contain only a single operating model. Given this constraint, realistic use conditions include operation modes, standby, and periods of time when the processor is turned off, as illustrated in Fig. 1. This paper takes these use scenarios into account, which are based on large scale user monitoring [8]. These use scenarios are based on actual average usage patterns for different applications. Cycling between states is not considered, since this work focuses on wearout mechanisms affecting the processor, rather than the package.

Fig. 1. The use scenarios provided by Intel are shown [8].

The rest of the paper is organized as follows. Section 2 presents the device-level wearout models that we have used in this study. Section 3 gives the overview of our system-level aging assessment framework, describing the methodology to determine model parameters through FPGA emulation. Section 4 describes our methodology to evaluate performance degradation of a microprocessor due to TDDB and presents the degradation and lifetime results for logic blocks of the microprocessor. In Section 5, we present analysis of SRAM noise margins under TDDB degradation and estimate SRAM lifetime. Section 6 combines the lifetimes of memory and logic blocks together. Section 7 determines variation in lifetime as a function of supply voltage and ambient temperature, and Section 7 concludes the paper.

II. DEVICE-LEVEL WEAROUT MODELS

We first model TDDB at the device level and then abstract the models to the system level. TDDB is one of the key reliability issues for CMOS devices. Stress induced leakage current (SILC) is induced by trap-assisted tunneling mechanisms where electrons pass from the cathode to the anode via defect sites (neutral traps) in the gate dielectric by the electrical field [12]-[15]. When the gate dielectric experiences partial breakdown, it is known as soft breakdown (SBD) [12],[16].

Experimental observations indicate that the mean time to failure is a function of the total gate oxide surface area, temperature, and gate voltage due to the weakest-link character of gate dielectric breakdown [17]. The conventional model provides a time-to-failure for HBD, as a Weibull distribution with two parameters, η , and the shape parameter, β , where [18]

$$\eta = A \left(\frac{1}{WL} \right)^{1/\beta} \frac{V_G^{a+bT}}{\alpha} e^{-1/\beta} \exp \left(\frac{c}{T} + \frac{d}{T^2} \right) \quad (1)$$

and where W and L are device width and length, respectively, α is the fraction of time that the gate is under stress, T is temperature, V_G is the gate voltage, and a , b , c , d , and A are fitting constants.

Fig. 2. Defect generation in the gate dielectric layer based on a 2D percolation model for SBD and HBD paths.

However, circuits may fail prior to HBD, due to SBD. To determine the impact of SBD it is important to study the degradation process to understand when and if circuits fail due to SBD. Circuits have been known to operate during breakdown [19]. In order to model circuit performance degradation under SBD, time-dependent resistance models [9],[20] and time-dependent leakage current models [21] have been proposed for SPICE simulation. This work uses time-dependent resistance models.

Using emulation, described in the next section, the devices are partitioned into groups that experience equivalent stress and temperature. More specifically, for an nMOS device, the time under stress is the time that the gate has the supply voltage applied. This time depends on the input patterns and the propagation of these patterns to each MOSFET.

For each group of devices, the next step is to determine the number of devices experiencing different numbers of soft breakdown paths. This is done using the percolation model.

Fig. 2 shows the percolation model (PM) concept that involves placing neutral traps randomly within the gate dielectric and analyzing the number of resistive conduction paths in a 2D matrix representing the gate dielectric layer [22]. We have expanded the model to 3D to count an accurate number of conduction paths.

In the percolation model, the defect generation rate depends only on the gate voltage (V_G) and temperature (T). Therefore, during electrical stress, the trap density in the gate dielectric increases with stress time t as a power law in the anode hole injection model. Stress is converted to a number of traps [23], [24]:

$$N_{trap}(t, V_G) = A \exp(BV_G) t^\zeta \cdot \tau_{ox} WL \cdot \exp(-\theta T), \quad (2)$$

where A , B , and ζ are fitting constants, and τ_{ox} , W , and L are oxide thickness, gate width, and length, respectively, and $\theta \approx 0.01^\circ\text{C}^{-1}$ [24].

Fig. 3 shows the PM simulation results and the probability distribution of the time of occurrence of conduction paths in the oxide layer as a function of gate size (Fig. 3(a)). It also shows the number of SBD paths as a function of time (Fig. 3(b)). Then, as we can see in Fig. 3(b), if we know the stress duration of an applied gate voltage to a MOSFET, the probability of a fixed number of conduction paths can be estimated.

Fig. 3. Time distribution of defect generation in SiO₂. (a) The probability distribution of the time of occurrence of the k^{th} SBD path for different gate sizes. (b) The probability distribution of the number of SBD paths for a fixed gate size as a function of time.

Fig. 4. Probability of the k^{th} SBD path for a fixed gate size and as a function of time.

The results in Fig. 3 indicate the probability that at least n paths are observed in the oxide. To find the probability that there are exactly n paths, it is necessary to subtract the $n - 1^{\text{st}}$ curve from the n^{th} curve. A result is illustrated in Fig. 4.

Fig. 5. SBD resistance as a function of the number of SBD paths.

Note that the number of breakdown paths in the gate depends on the operating conditions. Hence, given a time under operation, the emulator determines the time under stress for each group of devices. This is used to look up probabilities of different numbers of SBD paths. The number of devices in the group multiplies the probabilities to estimate the expected number of devices with each number of breakdown events in the group.

The time under stress for the i^{th} device is a function of bias. Let α_i , where $0 \leq \alpha_i \leq 1$, be the fraction of time under stress for the i^{th} device. Then, $t_{stress,i} = t\alpha_i$, where t is the time under operation. Let $p_i = f_{SBD(n)}(t_{stress,i})$ be a probability of n SBD paths for the i^{th} device. Therefore, $p_i = f_{SBD(n)}(t\alpha_i)$. If the group of devices has N devices, then Np_i devices are randomly selected to have n breakdown paths at time t . Each sample randomly selects the devices experiencing SBD.

Next, for each breakdown path, the SBD leakage resistance is calculated with the QPC model [25],

$$R_{SBD} \cong V_G / \left[\frac{4e}{h\alpha} N \cdot \exp(-\xi\Phi) \cdot \sinh\left(\frac{ae(V_G - V_0)}{2}\right) \right], \quad (3)$$

where $\Phi = 3 \sim 4eV$, $V_0 = 0 \sim 0.5V$, $\xi = 2 \sim 3eV^{-1}$, h is Plank's constant, e is the electron charge, and N is number of SBD conduction paths. The location of SBD, gate-to-drain vs. gate-to-source, is randomly selected. The resistance as a function of the number of SBD paths is illustrated in Fig. 5.

The impact of SBD on a ring oscillator is illustrated in Fig. 6. It shows the waveform comparison of the case with no SBD path, one SBD path, two SBD paths and three SBD paths. It can be seen that more SBD paths result in larger delays, while not degrading signal swing.

III. AGING ASSESSMENT FRAMEWORK

The time-to-failure of TDDDB is a function of device stress and the thermal profile. A framework for the acquisition of spatial and temporal thermal/electrical stress of a system was constructed.

Fig. 6. The impact of SBD on ring oscillator performance.

Fig. 7. The schematic of the electrical/thermal aging assessment framework. Yellow blocks indicate tools, while blue blocks indicate data.

Running RTL or SPICE simulations of a complete microprocessor to extract the activity profile of each net is not feasible in most cases, since it may take a few months to finish simulating a single benchmark. On the other hand, simulating microprocessors with standard benchmarks on an FPGA takes only a few minutes. Our electrical aging assessment framework is schematically described in Fig. 7 and provides an efficient way to acquire electrical and thermal profiles for any digital system for use in system-level reliability analysis.

The RTL has been synthesized and loaded to the FPGA with the Xilinx ISE (Integrated Software Environment) [26]. Once the FPGA is programmed, the activity can be collected by placing counters at the I/O ports to track the state probabilities and the toggle rates of the ports during application runtime, as illustrated in Fig. 8.

Fig. 8. The system used to collect activity profile of the microprocessor contains an FPGA board that implements the microprocessor system and exports data on the activity profile to a PC.

Since the I/O ports for each unit can be found on the top of each module, a script-based tool automatically adds some counters into the RTL. These counters are attached to the I/O ports of each functional unit. There are two types of counters: a state counter to record the duration of logic “1” and logic “0” and an edge counter which records the number of transitions. Also, the activity transmission unit is inserted into the RTL automatically, to transmit the activity profiles back to the host computer through the RS-232 protocol. The complexity of this RTL revision process is $O(n)$, where n is the number of I/O ports. Since the complexity is linear, the RTL revision process is scalable and can be implemented for large systems. Our current work focuses on implementing a microprocessor on a single FPGA, so the revised RTL is executable as long as the FPGA has enough resources (gates) to support large systems. A set of standard benchmarks [27] was used as the applications for analysis.

Because the FPGA does not have enough resources to place counters on all nets within the design, only the I/O activities of each block are captured. The I/O activities and the netlist were then used for activity propagation to each net in the design, depending on its logic behaviour, for a complete stress/transition probability profile of the internal nodes of the microprocessor under study. This is done using PrimeTime [28] on a block-by-block basis. Thus, we have the probability of a transition occurring in any cell and the probability at each state, i.e., the probability at logic “1”. It is this probability at logic “1” and logic “0” that is needed to compute the probability that each device is under TDDb stress. Figs. 9 and 10 show the distributions of the stress probability and the transition rate, respectively, when the microprocessor is running a set of standard benchmarks. It can be seen that the distribution of the state probability and transition rate are almost uniform.

Fig. 9. The distribution of the transition rate for the microprocessor is shown while running a standard benchmark.

Fig. 10. The distribution of the state probability for the microprocessor is shown while running a standard benchmark.

Some errors are introduced due to use of the propagations of the transition rate and state probability to internal nodes within a block, as needed to calculate the transition rate and state probability of internal nets. However, the errors are small enough to be acceptable. The propagation of transition rate and state probability was verified by comparing the exact transition numbers and state periods of randomly selected nets from the microprocessor with the ones calculated by propagations. The results, as illustrated in Fig. 11, show that the percent errors for more than 90% of the selected samples are less than 10% for transition rate. More than 80% of the selected samples have errors that are less than 15% for state probability. The high errors are mostly from the nets in deeper locations of the circuit that are far from the I/Os. Since errors are cumulative, activity propagation to deeper nets leads to a larger difference between the real transition rate and state probability and the calculated ones.

Fig. 11. Percent error distribution of the randomly selected nets.

The temperature variation throughout the microprocessor is also taken into account when modeling TDDb. The netlist was used for layout generation. The RC information from the layout, together with the net activities, was used for the extraction of the power profile and the consequent thermal profile, through the power simulator [28] and the thermal simulator [29], respectively.

Fig. 12 shows the static temperature distribution when an example microprocessor system is running a standard benchmark in active mode and when it is in standby mode. The static temperature is set to be the environmental temperature (30°C) when the system is in the off mode. The static temperature is the temperature when the system reaches a stable status, when the heat generated by task execution and the heat dissipated by the cooler are balanced. None of the benchmarks that were considered in this study exhibited thermal runaway. The thermal transients associated with switching between active, standby, and off states were assumed to have a negligible impact, since the transition time was assumed to be small compared to the time in any state.

Then, combining the RC parasitics, the thermal profile, and the calculated probability of current flow and voltage stress, we can use device level models to characterize the impact of TDDb on the microprocessors under study through the performance degradation analysis flow proposed in the next section to determine the wearout profile of the logic blocks and memory blocks.

IV. IMPACT OF TDDb ON MICROPROCESSOR LOGIC BLOCK RELIABILITY

A. Performance Degradation Analysis Flow

To characterize the impact of TDDb on logic circuits, the signal edge degradation caused by TDDb needs to be studied. Fig. 13 illustrates the data flow and structure for modeling logic circuit performance degradation due to TDDb.

We begin with extracting all the paths through the microprocessor system under study via block-based statistical timing analysis [30].

Fig. 12. The static temperature distributions for an example microprocessor in (a) standby mode and (b) operation mode.

Fig. 13. The schematic of the proposed flow for performance degradation analysis is shown. Yellow blocks indicate tools, while blue blocks indicate data.

To study the impact of TDDb on the microprocessor, the device models described in Section 2 determine the gate-to-source resistance (R_{G2S}) and gate-to-drain resistance (R_{G2D}) of each device within the extracted critical paths of the microprocessor system. The R_{G2S} and R_{G2D} variations are functions of the electrical stress and thermal profiles acquired from the framework illustrated in Fig. 7. Note that the electrical stress profiles acquired from Fig. 7 include only the activity for each net in the microprocessor. Activity propagation was implemented to obtain the electrical stress profiles for each transistor in the standard cells within the paths.

To include the additional delay caused by TDDb for each gate within each path for further timing analysis, a standard cell library has been built to model the delay drift caused by the R_{G2S}/R_{G2D} variations of each cell, in addition, to other

variations from process and environmental parameters. The gate delays of the standard cells can be modeled via first-order linear regression as

$$D = d_0 + \sum_{i=1}^n d_i \Delta R_{G2S/G2Di} + \sum_{i=n+1}^{(m+1)n} d_i \Delta X_{i-n} + d_{(m+1)n+1} \text{Slope} + d_{(m+1)n+2} C_{load}, \quad (4)$$

where D is the gate delay of a cell; n is the number of transistors in the cell; m is the number of process parameters, $\Delta R_{G2S/G2Di}$ are the R_{G2S} and R_{G2D} variations in transistor i due to TDDB, ΔX_i is the threshold voltage shift in transistor i (ΔV_{thi}) due to process variations, ΔX_i is the shift in channel length (ΔL) due to process variations; Slope is the input slope of the input waveform to the cell; C_{load} is the loading capacitance of the cell; d_0 is a constant term; and d_i , $i=1,2,\dots,(m+1)n+2$, are sensitivity coefficients. Linear regression in eqn. (4) is just an example, our implementation involves the use of regression splines and a Pi-model load [30].

Taking into account the additional delays caused by TDDB, the gate-level netlist, the timing constraint file which defines system timing speculations, and the RC parasitics, we sort all paths to find the critical ones for the microprocessor system for further analysis. As time increases and TDDB degrades, device characteristics and the paths are resorted to determine a new set of critical paths.

The extracted critical paths, together with all acquired RC parasitic elements in the cells and in the interconnects in the paths, are then simulated with SPICE-like accuracy using statistical path-based timing analysis [30]. Process parameter variations for the devices within the extracted critical paths are taken into account in the timing simulations by running Monte Carlo simulations with random values for process parameters.

B. Logic Wearout Simulation Results

The well-known open-source LEON3 IP core processor with superscalar abilities was studied [31]. The microprocessor logic units consist of a 32-bit general purpose integer unit (IU), a 32-bit multiplier (MUL), a 32-bit divider (DIV), and a memory management unit (MMU). Storage blocks include a window-based register file unit (RF), separate data (D-Cache) and instruction (I-Cache) caches and cache tag storage units (Dtags and Itags). The microprocessor includes around 240k gates.

A set of standard benchmarks [27] were run on the microprocessor. The runtime for executing each benchmark on the system is around one to three minutes. The electrical stress and thermal profiles for the system were collected using the framework described in Section 3.

Process variations were taken into account through Monte Carlo simulations. The process model includes die-to-die variation in channel length and random within die variation in threshold voltage, with standard deviations of 10%. It is assumed that the dominant cause of channel length variation is lithography and etch control, where all devices on a chip receive approximately the same shift. It is assumed that the dominant cause of the initial threshold voltage variation is random dopant fluctuations due to limits in implant control.

Fig. 14. The delay distributions of the critical paths of the microprocessor due to TDDB for different use scenarios and for two different stress times (10^5 seconds and 10^9 seconds). The paths are ordered from shortest to longest delay and plotted by percentile.

The electrical and thermal profiles, process variation models, together with the lifetime models from Section 2, were then used for performance degradation analysis to analyze the impact of TDDB on microprocessor reliability.

Fig. 14 shows the delay distributions of the critical paths of the LEON3 microprocessor due to TDDB for the different use scenarios in Fig. 1 and for two different stress times (10^5 seconds and 10^9 seconds). As the stress time increases, the critical-path delay also increases which further degrades the circuit performance. The results show that the critical paths of the microprocessor degrade differently while undergoing longer stress time. The delay distributions of the critical paths not only provide us with the degradation rates of the microprocessor system due to TDDB, but also indicate variation in degradation.

Since TDDB results in an increase in data path delays, a system fails when there are timing violations, such as setup time and hold time violations occurring in the critical paths. Timing violation analysis of each critical path within a system is system frequency dependent because there are bigger delay margins when a system has a lower operating frequency, and there are smaller delay margins when a system has a higher operating frequency.

By weighting the lifetimes of operation, standby and off mode in accordance with Fig.1, we have estimated the lifetimes of the microprocessor under study based on different operating frequencies for different use scenarios, as shown in Fig. 15. Random variation for TDDB is not just due to the process variations, but also due to random selection of devices with SBD. The lifetime is limited by SBD at higher operating frequencies and by HBD at lower operating frequencies.

Besides LEON3, the 32-bit RISC microprocessor [32] which includes around 73k gates was also studied.

Fig. 15. The estimated lifetimes of the microprocessor due to TDDB for different use scenarios and different system frequencies. Dotted lines show the 90% confidence boundaries when considering process variation.

Fig. 16. The estimated lifetimes of the RISC microprocessor due to TDDB for different benchmarks and different system frequencies. Dotted lines show the 90% confidence boundaries when considering process variations.

Fig. 16 shows the estimated lifetime due to TDDB for the RISC microprocessor based on different operating frequencies. Similar to the results for the LEON3, the results clearly indicate that the estimated system lifetimes decrease as the system frequency increases, and gaming has the shortest lifetime. The results also show that the lifetime distribution at any operating frequency is best fit with a Weibull distribution

V. PERFORMANCE DEGRADATION ANALYSIS FOR MEMORY BLOCKS

This work has considered the memory arrays for both the LEON3 and RISC microprocessors, all with the same cell design. LEON3 includes D- and I-Caches that contain 1024 32-bit words, which consist of 32,768 SRAM cells, tag storage units that contain 128 28-bit words, which consist of 3584 SRAM cells, and a RF that contains 256 32-bit words, which consists of 8192 SRAM cells. The RISC microprocessor has a 512 word $\times$ 69 bit memory, which consists of 35,328 SRAM cells. Since each SRAM cell within the memory experiences

different electrical stress and temperature when the microprocessor is executing benchmarks, each SRAM cell has different gate leakage currents due to TDDB.

As an SRAM cell undergoes AC and DC stress, the cell becomes increasingly skewed as one device degrades more than the others. This leads to impaired noise immunity.

SRAMs are characterized with several performance metrics. These include the read and retention static noise margins (SNMs), write margin, the read current (I_{READ}), and the minimum retention voltage ($V_{\text{dd-min}}$). The static noise margins are defined as the minimum DC noise voltage necessary to change the state of an SRAM cell. The read SNM is measured with the access transistors turned on, while the access transistors are off for the retention SNM. The write margin is the minimum voltage needed to flip the state of the cell, with the access transistors turned on. $V_{\text{dd-min}}$ is the minimum voltage in which the SRAM retains its state. Finally, the read current, which is inversely proportional to access time, is the current flow through pull-down devices when performing a read operation.

A set of standard benchmarks were run on the microprocessor system under study. The electrical stress and thermal profiles for the memory were collected by the aging assessment framework described in Section 3. The use scenarios in Fig. 1 further modify the stress profile of each cell. The electrical and thermal profiles, process parameter variations, together with the lifetime models from Section 2, were then used to analyze the performance of each SRAM cell within the memory.

Fig. 17 shows the degradation of the read SNM, the write margin, the read current, and $V_{\text{dd-min}}$ of the memory due to TDDB for different use scenarios. To better understand the impact of TDDB on the memory for different performance metrics, the performances are normalized with respect to their specifications and nominal fault free performances, i.e.,

$$Y = (X - X_{\text{spec}}) / (X_{\text{fault-free}} - X_{\text{spec}}) \quad (5)$$

where X is the performance parameter under study, $X_{\text{fault-free}}$ is the nominal performance, and X_{spec} is the design specification for fault-free operation. It can be seen that the read current (I_{READ}) is the most affected performance by TDDB. The minimum retention voltage, read stability, and write margin are relatively unaffected.

In order to manage the data volume, we first compute the degradation in each of the SRAM performances for a set of stress conditions associated with normal operation. We partition the stress into 21 states. The number of SRAM cells in different stress states is illustrated in Fig. 18 for the memory of the LEON3 microprocessor (five memory blocks) and for the memory in the RISC processor.

Then, we analyze the use scenarios to determine the degradation of each cell performance in each stress state. The lifetime distribution is computed for each stress state for each use scenario. The lifetime distributions of a cell for four representative stress states for a single use scenario are illustrated in Fig. 19.

Fig. 17. The performance metrics of the memory for different use scenarios under TDDB.

Fig. 18. The distributions of the number of cells in each stress state for 5 different memory blocks in LEON3 microprocessor.

Fig. 19. Lifetime distributions for a cell for a variety of cell states for the general usage scenario.

Fig. 20. The lifetime distributions of the SRAM cells in the Data Cache Block of the LEON3 microprocessor due to TDDB.

The lifetime distributions in Fig. 19 is best fit with Weibull distribution. The characteristic lifetimes from the Weibull distributions, combined with the stress-state distribution in Fig. 18, are used to determine the lifetime distribution of all the SRAM cells in each memory block for different use scenarios, as illustrated in Fig. 20 for one of the blocks. The other blocks are similar.

The fitted Weibull distributions are used to find the probability of failure of an SRAM cell, PF_{bit} . Suppose there are no error correcting codes, then the word size is $N=32$ for the D-cache, I-cache and the RF blocks of the LEON3, $N=28$ for the Dtags and Itags blocks of the LEON3, and $N=69$ for the memory block of the RISC microprocessor. If we assume that every SRAM cell in a word has the same stress state, then the lifetime distributions of words in different stress states can be derived from Fig. 19. The probability of failure of a word is calculated as

$$PF_{word} = 1 - (1 - PF_{bit})^N \quad (6)$$

where PF_{word} is the probability of failure of a word, PF_{bit} is the probability of failure of a bit, and N is the number of bits in one word. Since PF_{bit} changes as a function of time, PF_{word} also changes as a function of time. Since different stress states have different values of PF_{bit} , the calculations of PF_{word} are repeated for 21 different stress states.

Then the probability of failure of the SRAM block is also obtained as a function of time.

$$PF_{SRAM} = 1 - \prod_{i=1}^{21} (1 - PF_{word_i})^{N_{word_i}} \quad (7)$$

where PF_{SRAM} is the probability of failure of the whole memory block, PF_{word_i} is the probability of failure of a word for the stress state i , and N_{word_i} is the number of words in stress state i .

Fig. 21. The failure probability under different use scenarios for the Data Cache block of the LEON3 microprocessor due to TDDb with and without redundancy.

Fig. 22. The failure probability under different use scenarios for the memory of the RISC microprocessor due to TDDb with and without redundancy.

Error correcting codes (ECCs) can ensure that a memory system can tolerate faults. BCH codes [33] require seven additional bits per word and can correct one bit per word. The relationship between failures of single bits, P_{fail} , and the failure of the word is modeled with a binomial distribution. For a word containing N bits, then the probability of failure of a word, F_{word} , is

$$PF_{word} = 1 - (1 - PF_{bit})^N - N * PF_{bit}(1 - PF_{bit})^{N-1}. \quad (8)$$
 The word size containing the ECC codes is $N=39$ for the D-cache, I-cache, and RF blocks of the LEON3, $N=35$ for the Dtags and Itags blocks of the LEON3, and $N=76$ for the memory block of the RISC microprocessor.

If the memory uses error correcting codes, PF_{word} for different stress states is calculated using (8) and PF_{SRAM} is obtained using (7). PF_{SRAM} of each memory block (five memory blocks in LEON3 and one memory block in the RISC microprocessor) as a function of time is shown in Figs. 21 and 22.

Fig. 23. The estimated lifetimes of the microprocessor due to TDDb for different use scenarios and different system frequencies, combining logic and memory blocks. It is assumed that the memory blocks do not include ECCs. Dotted lines show the boundaries when considering process variation.

VI. COMBINATION OF LOGIC AND MEMORY LIFETIME

The overall lifetime distribution of the microprocessor due to TDDb is obtained by combining the lifetime of the logic blocks and the lifetime of the memory blocks. We fit both the lifetimes of each block with Weibull distributions. Let $\eta_1, \eta_2, \eta_3, \dots, \eta_n$ be the characteristic lifetimes of n blocks. Let $\beta_1, \beta_2, \beta_3, \dots, \beta_n$ be the corresponding shape parameters. Then, the characteristic lifetime of the system, η_{chip} , is the solution of [1],[3]-[5]

$$1 = \sum_{i=1}^n \left(\frac{\eta_{chip}}{\eta_i} \right)^{\beta_i}. \quad (7)$$

Similarly, the shape parameter of the system, β_{chip} , is the solution of [1],[3]-[5]

$$\beta_{chip} = \sum_{i=1}^n \beta_i \left(\frac{\eta_{chip}}{\eta_i} \right)^{\beta_i}. \quad (8)$$

The overall lifetime for the LEON3 combines five memory blocks and logic. The result is a function of frequency, because the logic lifetime is a function of frequency. The results with and without ECCs are shown in Figs. 23 and 24. It can be seen that the use scenario and ECCs make a substantial impact on lifetime. The similar results for the RISC processors are shown in Figs. 25 and 26.

VII. IMPACT OF SUPPLY VOLTAGE AND AMBIENT TEMPERATURE ON TDDb LIFETIMES

TDDb is activity, voltage and temperature dependent. The working environment (VDD and ambient temperature) dramatically impacts TDDb lifetimes. In this section, we study the impact of VDD and temperature on TDDb lifetimes, and then present some insights on how to improve TDDb reliability.

Fig. 24. The estimated lifetimes of the microprocessor due to TDDB for different use scenarios and different system frequencies, combining logic and memory blocks. It is assumed that the memory blocks include ECCs. Dotted lines show the boundaries when considering process variation.

Fig. 25. The estimated lifetimes of the RISC microprocessor due to TDDB for different benchmarks and different system frequencies, combining the logic and the memory block. It is assumed that the memory block does not include ECCs. Dotted lines show the boundaries when considering process variations.

We have repeated the calculations of TDDB lifetimes at the four different corners of VDD and temperature shown in Table 1. The ΔVDD and ΔT in Table 1 denote the percentage change compared to nominal values of VDD and temperature, respectively. For our examples, the nominal VDD is 1.2V and the nominal temperature is 299K. As seen from Table I, the worst-case test corner is Corner 1, where VDD is 1.32V and temperature is 359K (82.8°C).

Fig. 26. The estimated lifetimes of the RISC microprocessor due to TDDB for different benchmarks and different system frequencies, combining the logic and the memory block. It is assumed that the memory block includes ECCs. Dotted lines show the boundaries when considering process variations.

Table 1 The VDD and Temperature Corners		
Corners	ΔVDD	ΔT
1	+10%	+20%
2	-10%	+20%
3	-10%	-20%
4	+10%	-20%

Fig. 27. The average overall lifetimes of the LEON3 microprocessor due to TDDB at operating frequency 290MHz for different use scenarios and for different test corners in Table 1 are shown.

The average overall lifetimes of the LEON3 microprocessor due to TDDB for different use scenarios and for different test corners are shown in Fig. 27. Corner 1, which is the worst-case test condition, has the shortest lifetime. At Corner 1, where VDD is 1.32V and the temperature is 359K (82.8°C), the TDDB lifetimes are less than one year for all use scenarios except for Office Work, whose lifetime is close to one year. From this observation, it can be seen that the simultaneous stress of both VDD and temperature dramatically shortens TDDB reliability. Also, from the comparison

between Corner 2 and Corner 4, the TDDB lifetimes are more sensitive to V_{DD} than to temperature.

From the perspective of TDDB reliability, if the chip is used in a high temperature environment, adaptively decreasing the V_{DD} is essential to improving TDDB lifetime. Therefore, adaptive voltage scaling is a helpful technique and can compensate for the actual operating conditions of the chip. Also, efficient thermal sink and power management are good techniques to avoid thermal runaway and to improve TDDB lifetime. Since TDDB degrades circuit delays of logic blocks, adaptive frequency scaling is also beneficial to TDDB reliability.

VIII. CONCLUSION

This paper presents a simulation workflow to analyze state-of-art microprocessor reliability under TDDB degradation. Taking into account the detailed thermal and electrical stress profiles of microprocessor systems while running real-world applications, a methodology is proposed to accurately assess microprocessor lifetime based on TDDB. This paper presents a way to establish the link between the device-level wearout models, the electrical stress profile, the thermal profile, and system performances for both logic and memory blocks. For the logic blocks, device degradation under realistic use conditions is linked to timing analysis. For memory, device degradation is linked to critical SRAM performances, namely the read static noise margin, write margin, read current, and the minimum retention voltage.

The proposed methodology achieves system-level reliability analysis due to TDDB. We achieve this by analyzing both logic blocks and memory arrays. This work is beneficial to microprocessor designers by providing them the perspective on reliability. It can enable the designers to make any updates in the design to enhance reliability prior to committing a design to manufacture. Our work can be extended to analyze the lifetimes of homogeneous or heterogeneous multi-microprocessor systems.

ACKNOWLEDGEMENT

The authors thank the Defense Advanced Research Projects Agency under grant HR0011-11-1-0011 for support. The authors would also like to thank the National Science Foundation under Award #116786 for financial support.

REFERENCES

- [1] C.-C. Chen and L. Milor, "System-level modeling and microprocessor reliability analysis for backend wearout mechanisms," *Proc. Design Automation and Test in Europe*, pp. 1615-1620, 2013.
- [2] C.-C. Chen, S. Cha, T. Liu, and L. Milor, "System-level modeling of microprocessor reliability degradation due to BTI and HCI," *IEEE International Reliability Physics Symposium*, pp. CA.8.1-CA.8.9, 2014.
- [3] M. Bashir, C.-C. Chen, L. Milor, D.H. Kim, and S.K. Lim, "Backend dielectric reliability full chip simulator," *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 8, pp. 1750-1762, August 2014.
- [4] C.-C. Chen, F. Ahmed, D. H. Kim, S. K. Lim, L. Milor, "Backend dielectric reliability simulator for microprocessor system," *Microelectronics Reliability*, vol. 52, pp. 1953-1959, 2012.
- [5] C.-C. Chen, M. Bashir, L. Milor, D.H. Kim, and S.K. Lim, "Simulation of system backend dielectric reliability," *Microelectronics Journal*, 2014.
- [6] R.G. Filippi, R.-C. Wang, A. Brendler, P.S. McLaughlin, J. Poulin, B. Redder, and J.R. Lloyd, "The effect of a threshold failure time and bimodal behavior on the electromigration lifetime of copper interconnects," *Proc. IEEE Int. Reliability Physics Symp.*, 2009, pp. 444-451.
- [7] H.W. Yao, K.-Y. Yang, P. Justison, M. Rayasam, O. Aubel, and J. Poppe, "Stress migration model for Cu interconnect reliability analysis," *J. Applied Physics*, vol. 110, no. 7, p. 73504, Oct. 2011.
- [8] R. Kwasnick, A.E. Papathanasiou, M. Reilly, A. Rashid, B. Zaknoon, and J. Falk, "Determination of CPU use conditions," *Proc. Int. Reliability Physics Symp.*, 2011, pp. 2C.3.1-2C.3.6.
- [9] M. Choudhury, V. Chandra, K. Mohanram, and R. Aitken., "Analytical model for TDDB-based performance degradation in combinational logic," in *Design, Automation & Test in Europe Conf.*, 2010, pp. 423-428.
- [10] J. Fang and S.S. Sapatnekar, "The impact of BTI variations on timing in digital logic circuits," *IEEE Trans. Device & Materials Reliability*, vol. 13, no. 1, pp. 277-286, March 2013.
- [11] V. Huard, C. Parthasarathy, C. Guerin, T. Valentin, E. Pion, M. Mammasse, N. Planes, and L. Camus, "NBTI degradation: From transistor to SRAM arrays," *Proc. Int. Reliability Physics Symp.*, 2008, pp. 289-300.
- [12] K. Okada, "Analysis of the relationship between defect site generation and dielectric breakdown utilizing A-mode stress induced leakage current," *IEEE Trans. on Electron Devices*, vol. 47, no. 6, pp. 1225-1230, 2000.
- [13] R. Degraeve, J.L. Ogier, R. Bellens, P.J. Roussel, G. Groeseneken, and H.E. Maes, "A new model for the field dependence of intrinsic and extrinsic time-dependent dielectric breakdown," *IEEE Trans. Electron Devices*, vol. 45, no. 2, pp. 472-481, 1998.
- [14] S. Takagi, N. Yasuda, and A. Toriumi, "Experimental evidence of inelastic tunneling and new I-V model for stress-induced leakage current," *IEDM Tech. Dig.*, 1996., pp. 323-326.
- [15] R. Degraeve, B. Kaczer, A. De Keersgieter, and G. Groeseneken, "Relation between breakdown mode and breakdown location in short channel NMOSFETs and its impact on reliability specifications," *IEEE Int. Reliability Physics Symp.*, 2001, pp. 360-366.
- [16] B.P. Linder, D.J. Frank, J.H. Stathis, and S.A. Cohen, "Transistor-limited constant voltage stress of gate dielectrics," in *Symp. VLSI Technol. Dig.*, 2001, pp. 93-94.
- [17] E. Y. Wu, E. J. Nowak, A. Vayshenker, W. L. Lai, and D. L. Harmon, "CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics," *IBM Journal of Research and Development*, vol. 46, 2002.
- [18] X. Li, J. Qin, and J.B. Bernstein, "Compact modeling of MOSFET wearout mechanisms for circuit-reliability simulation," *IEEE Trans. Devices and Materials Reliability*, vol. 8, no. 1, pp. 98-121, 2008.
- [19] B. Kaczer, R. Degraeve, M. Rasras, K. Van de Mieroop, P.J. Roussel, and G. Groeseneken, "Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability," *IEEE Trans. Electron Devices*, vol. 49, no. 3, pp. 500-506, March 2002.
- [20] T. Nigam, A. Kerber, and P. Peumans, "Accurate model for time-dependent dielectric breakdown of high-k metal gate stacks," *IEEE Int. Reliability Physics Symposium*, 2009, pp. 523-530.
- [21] S.Y. Kim, G. Panagopoulos, C.-H. Ho, M. Katoozi, E. Cannon, and K. Roy, "A Compact SPICE Model for Statistical Post-Breakdown Gate Current Increase Due to TDDB," *IEEE Int. Reliability Physics Symposium*, 2013.
- [22] J.H. Stathis, "Percolation models for gate oxide breakdown," *J. Appl. Phys.*, vol. 86, no. 10, pp. 5757-5766, Nov. 1999.
- [23] M. Houssa, P.W. Mertens, and M.M. Heyns, "Relation between stress-induced leakage current and time-dependent dielectric breakdown in ultra-thin gate oxides," *Semiconductor Sci. Technol.*, vol. 14, no. 10, pp. 892-896, 1999.
- [24] S. Lombardo, J. H. Stathis, B. P. Linder, K. L. Pey, F. Palumbo, and C. H. Tung, "Dielectric breakdown mechanisms in gate oxides," *Journal of Applied Physics*, vol. 98, no. 12, p. 121301, Dec. 2005.
- [25] E. Miranda and J. Suñé, "Analytic modeling of leakage current through multiple breakdown paths in SiO₂ films," *IEEE Int. Reliability Physics Symp.* 2001, pp. 367-379.
- [26] Xilinx ISE, <http://www.xilinx.com/products/design-tools/ise-design-suite>
- [27] Mibench benchmark: <http://www.eecs.umich.edu/mibench>
- [28] PrimeTime: www.synopsys.com
- [29] HotSpot: www.lava.cs.virginia.edu/HotSpot

- [30] T. Liu, C.-C. Chen, and L. Milor, "Accurate standard cell characterization and statistical timing analysis using multivariate adaptive regression splines," *Proc. Int. Symp. on Quality Electronic Design*, 2015.
- [31] LEON3 processor: www.gaisler.com
- [32] J. Draper, J. Sondeen, S. Mediratta, and I. Kim, "Implementation of a 32-bit RISC processor for the data-intensive architecture processing-in-memory chip", *Proc. IEEE Int. Conf. on Application-Specific Systems, Architectures and Processors*, 2002, pp.163 – 172, 17-19 July 2002.
- [33] B. Sklar and F.J. Harris, "The ABCs of Linear Block Codes," *IEEE Signal Processing Magazine*, pp. 14-35, July 2004.