Reliable Cache Design with Detection of Gate Oxide Breakdown Using BIST

Fahad Ahmed and Linda Milor
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332
fahmed6@gatech.edu and linda.milor@ece.gatech.edu

***Abstract*— Scaling of device sizes has reduced gate oxide thickness to a few atomic layers, increasing the vulnerability of the gate oxide to breakdown. During breakdown, devices go through a gradual wearout process after an initial gate leakage increase leading to device failure. It is proposed that if wearout can be monitored, cache arrays with failing cells can be reliably operated after reconfiguration given available memory redundancy. Using experimentally verified gate oxide breakdown models, a detailed analysis of the effect of progressive gate oxide breakdown on the performance of a conventional 6T SRAM cell is presented for 45nm predictive technology. The DC margin trends (Read, Write and Retention) and access times (Read and Write) during wearout are analyzed, and a cell breakdown point due to degradation in each of these parameters is defined. A combination of these results is used to formulate a practical definition for the hard-breakdown point of a cell. Using an on-chip PVT (process, voltage, and temperature) tolerant monitoring scheme, it has been shown that gradual wearout in SRAM cells, due to gate oxide breakdown, is detectible, and cell failure can be predicted before its occurrence.**

I. Introduction

DEVICE length scaling has been accompanied by a continuous reduction in gate-oxide thickness with the aim of maintaining acceptable device performance. Reliability margins of these aggressively scaled devices have been reduced. This problem has been somewhat alleviated by the introduction of high-K dielectric gate stacks which have shown improvement in terms of reliability, but as the results in [1] indicate, gate oxide breakdown can still have a significant impact in determining the lifetime of a device.

Throughout the lifetime of a device, various oxide defects, called oxide traps, are formed generally in non-overlapping regions of the oxide. As their density increases, the traps may start overlapping and at a certain point a "critical defect density" is reached and a continuous defect path is formed through the oxide [2]. This shows up as an increase in the gate leakage current [3], and is generally known as soft breakdown (SBD). Since device scaling has also been accompanied by a continuous reduction in the supply voltage, SBD may not coincide with device failure. Although operational degradation is seen, device switching characteristics and logical functionality may be maintained even after SBD [4], leading to a significant increase in device reliability margins for digital circuits [3].

To understand the impact of gate oxide breakdown (GOBD) resultant device degradation on a functional block, it is imperative to understand the circuit level implications of SBD. A post-SBD transistor may show gate currents that are orders of magnitude higher than nominal values. With time gate leakage further increases until the device loses its transistor characteristics [5] and enters hard breakdown (HDB). At the circuit level, this region of gradual decay between SBD and HBD leads to weakening logic voltages, degrading drive currents, and increasing logic delays [5]. Predicting device failure is possible if this gradual degradation can be monitored en route to HBD.

In this work we show that cell failures in cache due to GOBD can be predicted before their occurrence, using circuit design techniques, and the exact location of the failing site can be pinpointed down to the failing cell with little hardware overhead. We start off with an exhaustive analysis of the effect of GOBD on the DC margins of a 6T SRAM cell. We also show that GOBD can cause timing failures in SRAM cells, because of access time degradation, leading to frequency dependent failure probabilities. From the combination of these results, we formalize a definition for the HBD point for a cell (HBD_{cell}). Using an on-chip PVT tolerant monitoring scheme, based on bit-line discharge detection (BDD), we show that GOBD can be monitored throughout the SBD to HBD_{cell} degradation process, enabling us to predict cell failure, due to GOBD, before its occurrence.

This work is organized as follows. In Section 2, we relate device failure to cell failure. In Section 3, we introduce the defect models that have been used for the analysis of GOBD. The gradual parametric degradation in an SRAM cell during the degradation process is analyzed in Section 4. In Section 5 we introduce our proposed GOBD monitoring scheme. Section 6 concludes the paper with the summary.

II. Relating Device Failure to Cell Failure

A conventional SRAM cell consists of three types of transistors (latch PMOS, latch NMOS and access NMOS), and each one is critical to correct memory operation. Due to the different roles each plays during read/write/retention, the gate size and the amount of time the dielectric is under stress varies significantly among the three types of transistors. Since gate dielectric breakdown is a strong function of gate area [2] and stress frequency [6], using the same wearout functions for all the transistors in the cell could lead to highly inaccurate predictions. To predict the cause of SRAM failure from among the transistors in the cell, probability density functions of wearout were modeled for each

978-1-4244-5028-2/09/$25.00 ©2009 IEEE

individual transistor. These functions are dependent on their respective operating conditions.

It has been shown that Weibull failure rate distributions:

$$P(t_{fail}) = 1 - \exp\left(-\left(\frac{t_{fail}}{\eta}\right)^{\beta}\right) \quad (1)$$

provide a good fit for the rate of dielectric breakdown of both PMOS and NMOS transistors [3]. $P(t_{fail})$ is the probability that failure will occur before time, t_{fail}, and η and β are the two parameters that characterize the Weibull distribution. Interpolating lifetimes from reliability stress test to normal operating conditions is usually done using exponential or power law models with the later being the accepted model for gate dielectrics, since the former violates the Weibull scaling law [7]. The power law relates 'N', the power law exponent, to the time to breakdown, and, V_G, the operating voltage, as follows:

$$T_{BD} = a \times V_G^{-N} \quad . \quad (2)$$

T_{BD} for NMOS transistors indicates a stronger voltage dependency because of a higher power law exponent, in comparison with PMOS devices [8]. Inserting the power law model in the Weibull distribution function results in:

$$P(t_{fail}) = 1 - \exp\left(-\left(\frac{t_{fail}}{aV_G^{-N}}\right)^{\beta}\right) \quad . \quad (3)$$

Another distinction between the NMOS and PMOS transistors in an SRAM cell is the relatively larger NMOS sizes. This factor can be conveniently integrated into the failure distribution function using the Weibull area scaling property [2]:

$$P(t_{fail}) = 1 - \exp\left(-\left(\frac{t_{fail}}{a_{\min} V_G^{-N} \left(1/Area\right)^{1/\beta}}\right)^{\beta}\right) \quad (4)$$

where a_{min} is calculated for a minimum sized device.

Equation (4) was used to model the time to dielectric failure for the PMOS and NMOS devices in the SRAM cell latch. The dependency of dielectric breakdown on the stress frequency was then incorporated into the failure distribution model for the NMOS devices used as access transistors in an SRAM cell.

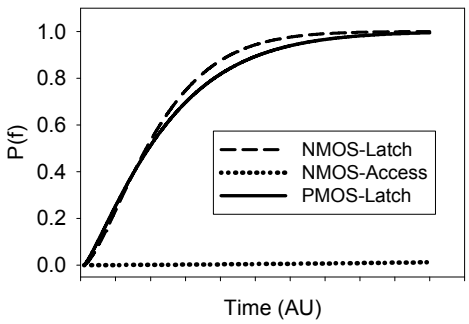

Fig. 1. The cumulative probability density functions for the failure probabilities of the devices in an SRAM cell.

The stress characteristics of the devices in the latch of an SRAM cell vary considerably from those of the access devices. While the devices in the latch may be under DC stress for large periods of time, access devices are off for most of their lifetime, only being activated if they reside in the column with the accessed cell. Even during the activation, these devices are not under DC stress like the devices in the latch, but are activated using uni-polar pulses with typically high frequencies. [6] analyzed the effect of dynamic stress on dielectric breakdown behavior and showed a very strong dependence of dielectric breakdown on the frequency of stress. A correction factor, *K(f),* was then introduced in the Weibull parameter η to account for the stress frequency variations for different devices:

$$P(t_{fail}) = 1 - \exp\left(-\left(\frac{t_{fail}}{K(f) a_{\min} V_G^{-N} \left(1/Area\right)^{1/\beta}}\right)^{\beta}\right) \quad (5)$$

This is the final form of the Weibull probability density function that was used in this study to model the cumulative failure probability of NMOS and PMOS transistors in the SRAM latch and the NMOS access devices. The cumulative probability density functions are shown in Figure 1. Our results indicate that the access devices in an SRAM cell are the least likely to cause cell failures among the three devices considered. Surprisingly, fairly comparable failure probabilities are seen for the devices in the latch. The stronger voltage dependence of the NMOS device is offset by the fact that latch NMOS devices are larger than latch PMOS devices. This result indicates that we should expect relatively similar figures for SRAM cell failures caused by PMOS devices and NMOS devices.

Experimental results from reliability test of caches however attribute over 90% of cell failures to failing NMOS devices [9]. This is in sharp contrast to the projected results using our model. This apparent inconsistency can be explained by relating the cell sensitivity to the latch PMOS and NMOS devices. In an SRAM cell, the NMOS device is stronger than the PMOS device and is more critical to correct cell operation. A read operation to a cell involves reading a '0', during which the relatively strong NMOS insures that the PMOS is unable to turn on and toggle the cell. The write operation, on the other hand, involves discharging a stored '1' to '0' which in turn means turning off the weak PMOS while turning on the stronger NMOS. Since GOBD translates to device weakening, it is proposed that cell failures due to latch PMOS failures can be ignored when compared to those due to latch NMOS failures.

To prove this assumption, cell stability tests were performed on a sample cache array. Using Monte Carlo sampling of equation (5), failure probabilities were related to breakdown gate resistances[10]. Each cell was submitted to sequential write-retention-read cycles. The initial value stored in the cell was the inverse of the value being written to insure cell toggling during the write operation. After the write operation, the word-lines were set low for two clock cycles to evaluate retention stability of the cell after which the stored data was read. The node voltage that was toggled

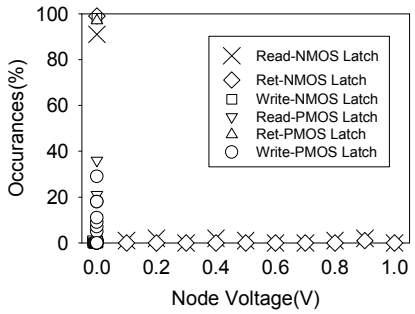

Fig. 2. Rise in '0' node during sequential write-retention-read cycles for PMOS and NMOS gate dielectric degradation.

to '0' after the write operation was tested after the write, retention and read operations to test cell stability.

Figure 2 shows the spread of the voltage rise for the node storing a '0' during sequential write-retention-read operations for degrading dielectrics of latch NMOS and PMOS devices. This result can now be used to understand the difference between the proposed device failure probability model and the experimental results for cell failures. Although the probabilities of failure for both NMOS and PMOS devices are comparable, the cell probability of failure due to the PMOS device turns out to be almost negligible. In fact all our failures occurred due to the NMOS device, even at very high gate leakage current values. Therefore gate dielectric degradation is not a significant reliability concern for degrading PMOS devices in SRAM cells.

Figure 3 shows the spreads of simulation results plotted against the gate resistance for both degrading NMOS and PMOS devices. The figure clearly indicates that SRAM cells are very resistant to PMOS degradation. Stable operation was observed for fairly low gate resistance values for PMOS devices. Since progressive oxide breakdown involves a gradual gate leakage increase after an initial rise, low cell operable gate resistance values for PMOS devices translates into a significant increase in cell lifetime under PMOS degradation. Degradation in NMOS devices turns out to be the main source of cell failures. If wearout failures in latch NMOS devices can be predicted, cache reliability can be significantly improved.

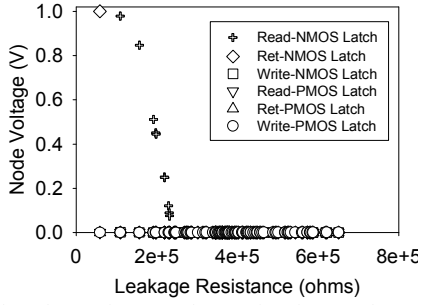

Fig. 3. Node voltage after a read operation for a node storing '0' for degrading PMOS and NMOS gate resistance.

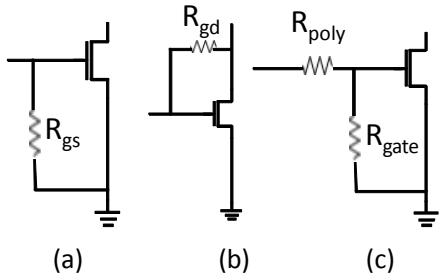

Fig. 4. (a) The gate-to-source leakage, (b) the gate-to-drain leakage and (c) the drive current degradation models for GOBD.

III. Modeling the GOBD Effect

Although changes in threshold voltage and device trans-conductance have also been observed as a result of a degrading gate dielectric [11], GOBD is generally modeled as an increase in gate leakage current. The leakage path could appear between the gate and the source/drain region or between the gate and substrate. Experimental results have shown that gate-to-substrate leakage usually has a minimal effect on circuit performance compared to gate-to-source/drain leakage [4]. Therefore, GOBD resulting in gate-to-substrate shorts has been ignored in this work. Moreover, in Section II we showed that although failure probabilities of PMOS and NMOS devices are comparable, the probability of cell failure due to PMOS degradation is very small when compared to cell failure due to NMOS degradation.

Figure 4 shows the NMOS gate leakage models used for our analysis. The R_{gs} model was used to model the gate-to-source leakage path, while the R_{gd} model was used to model the gate-to-drain leakage path.

Generally, the analysis of GOBD in SRAM cells has been done using the models shown in Figure 4(a) and 4(b). But these models do not account for the change in drive current caused by GOBD and hence cannot be used to estimate the access time shifts in the SRAM cells. The degradation in device drive current has previously been modeled in [5], and we have used the same model to study the effect of GOBD on cell access times using the R_{poly}-R_{gate} model shown in Figure 4(c).

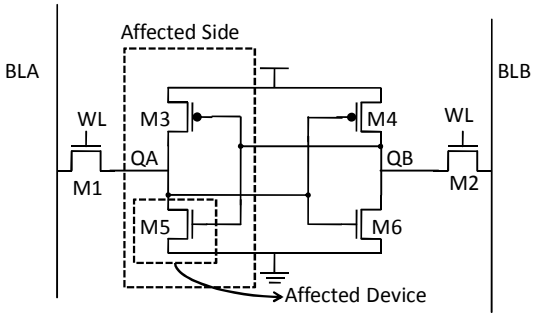

Fig. 5. A conventional 6T SRAM cell with the affected side and the affected transistor highlighted.

IV. SRAM STABILITY UNDER GOBD

In this section we study the stability of a conventional SRAM cell undergoing GOBD. Figure 5 shows the design of a conventional 6T SRAM cell. Unlike data paths, a cell could store the same voltage for long periods of time. The two NMOS devices, M5 or M6, in the latch at the centre of the cell could be under constant stress for years, which makes these devices especially susceptible to GOBD. Since GOBD generally has very low ppm defect levels, it can be safely assumed that M5 and M6 are unlikely to experience equal degradation. One of these transistors dominates the wearout characteristic of the cell. As shown in Figure 5, M5 is assumed to be the defective device and the inverter formed by M3 and M5 is referred to as the effected side of the cell. Experimental results in [12] indicate that the gate resistance values for SBD are spread between 100K to 800K ohms while those for HBD are in the range of a few kilo ohms. For our analysis, the gate resistance was varied from 10K to 1M ohms or until the failure point was reached, which ensures that the region between SBD and HBD is covered.

A. DC Margin Degradation-R_{gs} and R_{gd} Breakdown

M5 was replaced by the R_{gs} and R_{gd} GOBD models and the trend in read and retention SNM (static noise margin) was studied. The stability margins were extracted by fitting squares between the SNM curves and observing the diagonal length of the smaller of the two squares [13]. They are shown in Figure 6. For the write margin measurement, the bit-line on the side storing '1' was gradually raised from '0' upwards while the other side was kept at logic '1'. A single pulse with a pulse width of 250ps (equivalent to a memory system operating at 2 GHz) was applied to WL. Write margin was defined as the maximum voltage that was able to flip the cell when that single pulse is applied to WL. An improving write margin trend is seen for the case when QB is discharged by the gate-to-source leakage path. This is due to the weakening '1' and '0' at QB and QA, respectively, which increases the relative strengths of M6 and M3. For the case when QA is discharging, there is no degradation in the stored logic values, though the leakage does weaken M4 and M5, which makes it harder to toggle the cell and results in increased write margins. For the R_{gd} GOBD model, since a low impedance path between the gate and drain of an inverter results in degrading logic values, it becomes easier to overwrite the cell, and hence improved writability is observed.

B. Access Times (Read and Write)

Apart from an obvious increase in gate leakage, gate dielectric degradation also impacts the gate capacitance, hence affecting the induced channel. This degrades the drive current of the device [5]. This effect was modeled by [5] in their GOBD model shown in Figure 4(c). In this model, R_{poly} is the resistance of the polysilicon and R_{gate} is the resistance of the leakage path through the gate. The degraded drive current is modeled by the resistive divider network formed by R_{poly} and R_{gate}. Our simulated results, plotted in Figure 6, show that the read access time could exceed the maximum allowed time in the presence of GOBD, hence causing an access time failure.

Write access is defined as the time to successfully toggle the stored values of a cell. As shown in Figure 6, the write access time is mostly immune to GOBD. In fact a significant improvement was seen for the case when QB is discharged, a result in agreement with results for DC margins.

C. Cell Breakdown-HBD_{cell}

We have analyzed how different SRAM cell parameters vary as a function of a degrading gate dielectric and defined the points when a cell failure occurs due to each parameter. In this section we use those results to define a hard breakdown point for the cell. Specifically, Figure 6 shows the degradation of all the significant parameters as a function of the degrading leakage resistance.

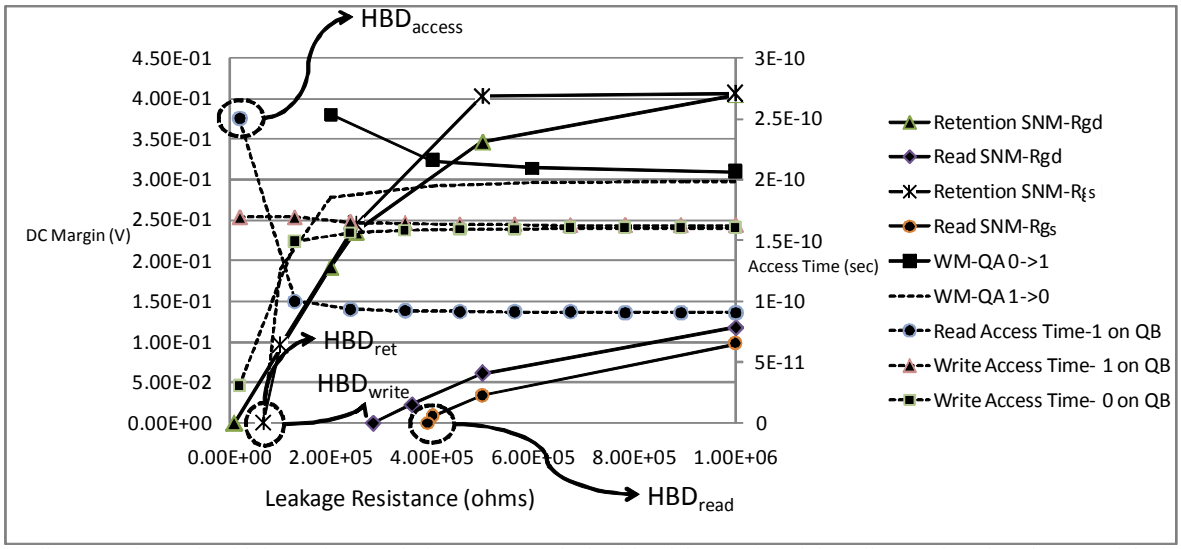

Fig. 6. Cell parametric trends with increasing gate leakage current. The hard breakdown point of the cell was selected as the max of the individual breakdown values.

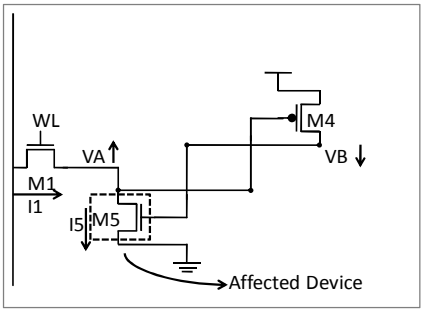

Fig. 7. Degrading stored logic values impact the bit-line discharge during read.

Since the write margins did not undergo any significant degradation under the R_{gd} defect model, they were not included in Figure 6. The case when node QA discharged is the only write operation adversely affected by GOBD, as seen in Figure 6. Similarly the write access times were unaffected to a large extent by GOBD with some cases showing slight improvement. The read access time for the case when QB stores a '1' is the most affected by GOBD and is the limiting factor in terms of access time for GOBD. HBD_{cell} can be defined by equation (6):

$$HBD_{cell}=MAX\ (HBD_{read},HBD_{ret},HBD_{write},HBD_{access}) \quad . \quad (6)$$

These points are highlighted in Figure 6. Clearly, cell stability under GOBD, in our case, is limited by read stability.

V. GOBD Monitoring Scheme-The Bit Line Discharge Detector (BDD)

A. BDD System Design

One of the effects of GOBD is degrading logic values. Our GOBD monitoring scheme works on the concept of 'degrading zeros'. As mentioned earlier, the read operation basically consists of discharging the bit-line next to the node that is storing a '0'. As the voltage value for a logical '0' degrades and rises, it affects the read current through the access transistor during the read operation.

Consider the case when QA stores a '0', as shown in Figure 7. The bit-line discharge current through M1 is dependent on I1 and I5, the currents through M1 and M5, respectively. As the gate leakage of M5 increases, the voltage VA at node QA increases, degrading current I1 due to channel length modulation. This also results in degradation in VB, the voltage at node QB, dependent on the resistive divider network formed by the 'on' state resistance of M4 and the leakage path through the gate of M5. As VB decreases, it exponentially degrades the current I5. As a result, there is degradation in the bit-line discharge voltage, as shown in Figure 8.

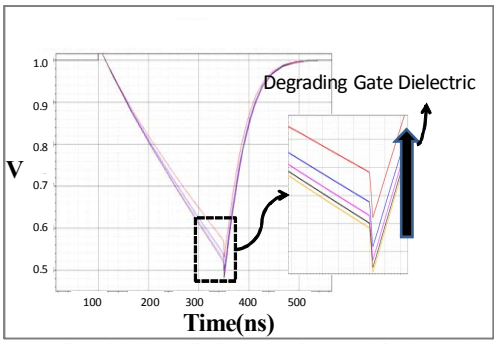

Fig. 8. Degrading bit-line discharge voltage with increasing gate leakage.

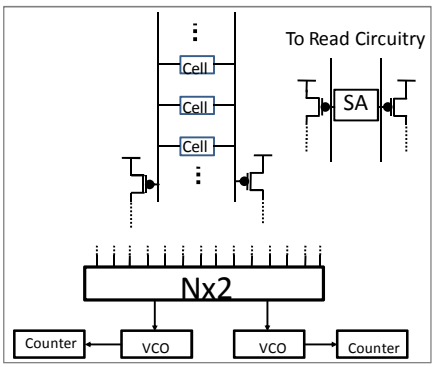

Fig. 9. The BDD. Although the system is attached directly to the bit-lines, it could alternatively be attached directly to the sense amplifier input to reduce complications in memory layout.

Figure 9 shows the BDD design for detecting and enhancing the change in the bit-line voltage. During test mode (TM) the cell goes through multiple read cycles. During each read cycle, the bit-line next to the node storing a '0' discharges partially, turning on the PMOS attached to it. The gate of the PMOS has absolutely no effect on the performance of the cell. Although it would fractionally increase the bit-line capacitance, this increase is negligible compared to the typically large bit-line capacitances. The PMOS, then, amplifies the discharge degradation and starts charging up the capacitor that controls the VCO frequency.

Since cache layouts are very regular, the bit-line PMOS devices can be moved out of the array and attached to the sense amplifiers, as shown in Figure 9, hence leading to a significant overhead reduction. An Nx2 MUX is introduced for further hardware overhead reduction. Every select line of the MUX connects two of the inputs to the two outputs. Since during a read operation, the signal activating the required column and row of the cache array is already generated, the same signal can be used for the MUX selection lines. Hence our scheme requires no extra control signals during TM. This means that during the idle time, the whole cache or a part of it could be monitored for GOBD by a few repetitive software generated read commands.

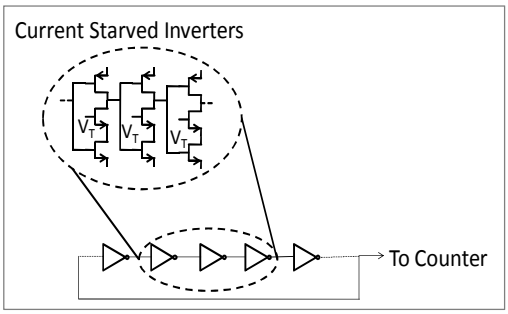

Fig. 10. Design of the broadband VCO.

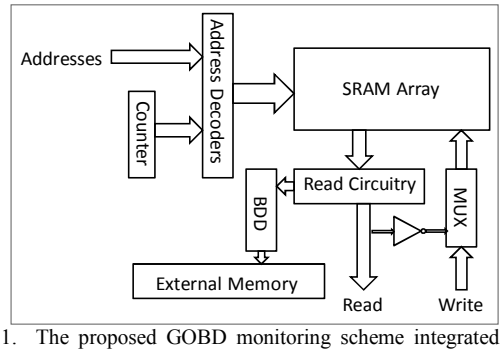

Fig. 11. The proposed GOBD monitoring scheme integrated into a memory system.

A ring oscillator consisting of current starved inverters is used as a broadband VCO, as shown in Figure 10. The voltage V_T controls the frequency of oscillation of the VCO. The drains of the bit-line PMOS devices are directly connected to V_T. As the array enters TM, V_T is connected to the drains of the bit-line PMOS devices. As the bit-line starts its repetitive discharge, the bit-line PMOS devices start charging up the intrinsic capacitance of the central NMOS in the current starved inverters. With increasing charge, V_T starts increasing, finally initiating oscillations in the VCO.

A significant result of our analysis of GOBD using the R_{gs} defect model is the fact that a gate-to-source leakage increase only affects one side of the cell. That means that a cell undergoing GOBD is detectible using our monitoring scheme only for one of the logic values stored. Generally, for frequently accessed cells this would not be a problem, since GOBD is a slow process and the probability of a degrading cell going undetected decreases with an increasing number of cell toggles.

However, to make the design more robust, a cell toggling scheme is presented in Figure 11. This ensures that a degrading dielectric is detected irrespective of the leakage path and the data stored in the cell.

B. BDD Simulation Results

The proposed system was run for a memory system having cells with both gate-to-source leakage degradation and gate-to-drain leakage degradation. Gate leakage resistances were varied from 1M to the calculated HBD_{cell} value. The results are shown in Figure 12.

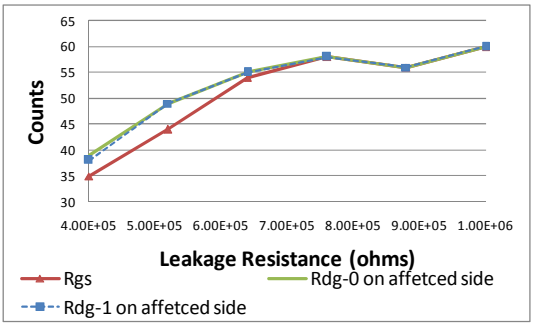

Fig. 12. Final count degradation of the GOBD monitoring system.

VI. SUMMARY

An exhaustive analysis has been done on the effects of GOBD on SRAM cell performance. We have shown how the DC margins and access times of the cell degrade with a degrading gate dielectric. A final breakdown point for the cell as been defined (HBD_{cell}), and using circuit design techniques, we have shown that cell failure can be detected.

ACKNOWLEDGMENT

The authors thank the Semiconductor Research Corporation for their financial support, under Tasks 1645.001 and 1645.002.

REFERENCES

[1] L. Sungjoo and D. L. Kwong, "TDDB and polarity-dependent reliability of high-quality, ultrathin CVD HfO₂ gate stack with TaN gate electrode," *IEEE Electron Device Letters,* vol. 25, pp. 13-15, 2004.

[2] J. Stathis, "Percolation models for gate oxide breakdown," *Journal of Applied Physics,* vol. 86, p. 5757, 1999.

[3] A. Kerber and et al, "Lifetime Prediction for CMOS Devices with Ultra Thin Gate Oxides Based on Progressive Breakdown," in *Proc. IEEE Int. Reliability Physics Symposium*, 2007, pp. 217-220.

[4] B. Kaczer *et al.,* "Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability," *IEEE Trans. on Electron Devices,* vol. 49, pp. 500-506, 2002.

[5] C. Tze Wee *et al.*, "Gate-Oxide Early Life Failure Prediction," in *Proc. IEEE VLSI Test Symposium,* 2008, pp. 111-118.

[6] P. Chaparala *et al.*, "Electric field dependent dielectric breakdown of intrinsic SiO₂ films under dynamic stress," in *Proc. IEEE Int. Reliability Physics Symposium*, 1996, pp. 61-66.

[7] P. E. Nicollian *et al.*, "The Current Understanding of the Trap Generation Mechanisms that Lead to the Power Law Model for Gate Dielectric Breakdown," in *Proc. IEEE Int. Reliability Physics Symposium,* 2007, pp. 197-208.

[8] M. Rohner, A. Kerber, and M. Kerber, "Voltage Acceleration of TBD and Its Correlation to Post Breakdown Conductivity of N- and P-Channel MOSFETs," in *Proc. IEEE Int. Reliability Physics Symposium Proceedings,* 2006, pp. 76-81.

[9] L. Yung-Huei *et al.*, "Prediction of Logic Product Failure Due To Thin-Gate Oxide Breakdown," in *Proc. IEEE Int. Reliability Physics Symposium,* 2006, pp. 18-28.

[10] R. Degraeve *et al.*, "Explaining `Voltage-Driven' Breakdown Statistics by Accurately Modeling Leakage Current Increase in Thin SiON and SiO2/High-K Stacks," in *Proc. IEEE Int. Reliability Physics Symposium*, 2006, pp. 82-89.

[11] A. Cester *et al.*, "Collapse of MOSFET drain current after soft breakdown," *IEEE Transactions on Device and Materials Reliability,* vol. 4, pp. 63-72, 2004.

[12] V. Ramadurai, N. Rohrer, and C. Gonzalez, "Sram Operational Voltage Shifts in the Presence of Gate Oxide Defects in 90 NM SOI," in *Proc. IEEE Int. Reliability Physics Symposium Proceedings,* 2006, pp. 270-273.

[13] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits,* vol. 22, pp. 748-754, 1987.