Analytical Model for SRAM Dynamic Write-ability Degradation due to Gate Oxide Breakdown

Vikas Chandra
ARM R&D, San Jose, CA
vikas.chandra@arm.com

Robert Aitken
ARM R&D, San Jose, CA
rob.aitken@arm.com

***Abstract*—Progressive gate oxide breakdown is emerging as one of the most important source of stability degradation in nanoscale SRAMs, especially at lower supply voltages. Low voltage operation of SRAM arrays is critical in reducing the power consumption of embedded microprocessors, thus necessitating the lowering of V_{min}. However, the oxide breakdown undesirably increases V_{min} due to increase in dynamic write failures and eventually static write failures as the supply voltage decreases. In this work, we describe an analytical model based on the Kohlrausch-William-Watts (KWW) function to predict the degradation in the WL_{crit} as the oxide breakdown increases. The KWW model also accurately predicts the efficacy of the word-line boosting and Vdd lowering write-assist techniques in reducing WL_{crit}. Simulation results from an industrial low-power 32nm SRAM show that model is accurate to within 1% of SPICE across range of supply voltages and severity of oxide breakdown with orders of improvement in runtime.**

I. Introduction

Static random access memory (SRAM) is a critical part of most VLSI system-on-chip (SoC) applications. The SRAM bit cell design has to cope with stringent requirements on the cell area leading to minimum (or close to minimum) sized transistors. Reducing operating voltage, to reduce dynamic and leakage power consumption, is challenging because of increasing bit cell device variations and reliability concerns with each technology node [1], [11], [18]. The minimum operating voltage beyond which yield loss from device variability becomes excessive is known as V_{min}. It has been observed that write failure is often the major V_{min} limiter in nanometer processes [2].

The V_{min} challenge in SRAM is further exacerbated due to an important reliability concern - gate oxide breakdown. Progressive soft oxide breakdown (SBD) in CMOS devices is becoming one of the most important source of time dependent degradation [13]. Studies have shown that the rate of trap formation increases as the permittivity of the dielectric increases [10]. With the introduction of high-k gate dielectrics, the probability of having gate oxide breakdown during the device lifetime increases substantially. Since the V_{min} of an SRAM is dictated by write failure, it is imperative to be able to predict the write-ability in the presence of soft breakdown to avoid over-design and pessimistic margins. The SPICE-based simulation of soft breakdown effects on SRAM dynamic stability is computationally demanding. The complexity arises due to the fact that the dynamic write-ability metric (called WL_{crit}) [16] is a function of the severity of soft breakdown as well as the supply voltage [4]. Hence, it is necessary to develop analytical models that can accurately predict WL_{crit} in the presence of gate oxide breakdown.

This paper proposes an accurate analytical model based on the Kohlrausch-William-Watts (KWW) function [8], [15] to predict the WL_{crit} of an SRAM bit cell across a range of supply voltage and soft breakdown severity. There exists a critical breakdown resistance (R_{crit}) for a given supply voltage at which the SRAM write failure transitions from being dynamically limited to statically limited. The proposed analytical model addresses these issues by dividing the impact of a soft breakdown into two components: (i) increase in R_{crit} with reduction in supply voltage and (ii) increase in WL_{crit} with reduction in supply voltage and increase in the severity of soft breakdown. The KWW function based analytical model captures the behavior of a feedback system (for example, a bit cell) quite well in the presence of a perturbation (for example, oxide breakdown). The analytical model is also able to accurately predict the behavior of WL_{crit} with the various write-assist schemes. Results indicate that the maximum error in the WL_{crit} predicted by the analytical model as compared with SPICE is less than 1% over a large range of oxide breakdown defects and supply voltage values.

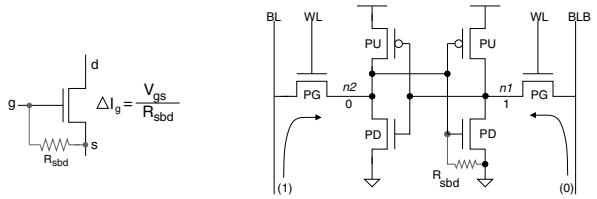
The rest of the paper is organized as follows. Section II describes the fault model of gate oxide breakdown. Section III introduces the concept of static and dynamic stability regions. Section IV describes the proposed analytical model based on the Kohlrausch-William-Watts (KWW) function. Section V analyzes the efficacy of commonly used write-assist techniques for oxide degraded SRAMs. Section VI summarizes the work and concludes.

II. Gate Oxide Breakdown Model

We only consider gate-to-diffusion (source or drain) breakdown since it represents the worst-case scenario for an SBD [6]. Breakdown to the channel can be modeled as a superposition of gate-to-drain and gate-to-source breakdowns. Since the probability of having more than one breakdown in an NMOS transistor is low [14], we consider only gate-to-source breakdown in this paper. Figure 1(a) shows the fault model for the gate oxide breakdown defect. The increase in gate leakage due to the oxide breakdown is modeled as a time-dependent resistive short between the gate and the source (or drain) depending on the location of the breakdown.

The time-dependent gate-to-source resistance model has also been experimentally verified [5]. It has been shown that the time to oxide breakdown in PMOS is an order of magnitude higher than in NMOS [7], [11]. Hence, the dynamic write-ability degradation model presented in this work only considers a soft breakdown in NMOS even though the model

978-3-9810801-7-9/DATE11/©2011 EDAA



(a) (b)

Fig. 1. (a) Gate oxide breakdown model for an NMOS transistor (b) Degradation location and state of the bit cell for worst case write-ability

also holds for a soft breakdown in PMOS. Figure 1(b) shows the breakdown location and the bit cell state which causes reduction in the write margin when an opposite value is being written. The degradation of the access NMOS transistors (PG) in Figure 1(b) is not considered since they stay "on" for a very small duration (only when the word-line is enabled).

III. Static and Dynamic Stability Regions

The static stability metric, SNM, defines the stability as the amount of noise needed to collapse a bi-stable system to a mono-stable system [9], [12], [17]. Figure 2 shows the SNM as a function of R_{sbd} across a range of supply voltage. It can be noted that the SNM approaches 0 as R_{sbd} approaches a critical breakdown resistance (defined as R_{crit}). Also, the value of R_{crit} increases as the supply voltage decreases. A lower supply voltage system has lesser noise immunity and hence will have higher R_{crit}. When the value of R_{sbd} approaches

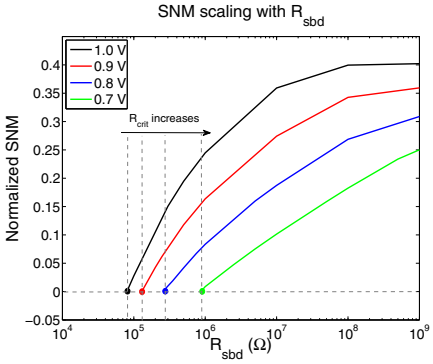
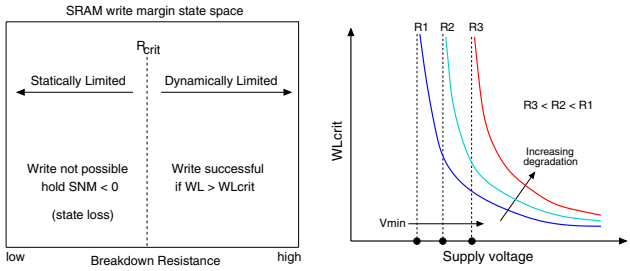


Fig. 2. SNM scaling with R_{sbd} and supply voltage

R_{crit}, the SNM becomes 0 and the feedback system is no longer stable. Also, the separatrix changes shape as R_{sbd} decreases [4]. From Figure 1(b), we can note that the stable state corresponds to $((n1, n2) = (1, 0))$. As the separatrix moves, the distance from the stable state to the separatrix also increases. By the definition of dynamic write-ability metric, a larger WL_{crit} will be required to cross the separatrix for a successful write operation as the severity of degradation increases (R_{sbd} decreases).

The two failure regions (or stability limit regions) in the write margin state space are separated by R_{crit} (Figure 3(a)). As long as the the breakdown resistance (R_{sbd}) is larger than R_{crit}, the bit cell write-ability is dynamically limited which means that the cell can be written by making the WL larger than WL_{crit}. The scaling of WL_{crit} with supply voltage is a



(a) (b)

Fig. 3. (a) Write margin state space with $Rcrit$ dividing the dynamic and static stability regions (b) WL_{crit} scaling with supply voltage and R_{sbd}

function of R_{sbd}, as explained earlier. As shown in Figure 3(b), V_{min} increases as the level of degradation increases (R_{sbd} decreases). As long as the supply voltage is higher than V_{min}, the bit cell is write-able given that WL is larger than WL_{crit} (dynamic stability region). However, the dynamic write margin decreases with decreasing R_{sbd}. In a limiting case, as R_{sbd} becomes equal to R_{crit}, both WL_{crit} and V_{min} asymptotically approach ∞ and the bit cell becomes statically unwrite-able. Essentially, it implies that with breakdown resistance of R_{crit} or lower the bit cell becomes unwrite-able irrespective of the width of the word-line pulse.

IV. Proposed Analytical Model

Figure 4 shows the change in the DC transfer curves and the separatrix as the gate oxide breakdown increases. The initial state of the system is $((n1, n2) = (1, 0))$ and after the cell is written, the state flips to $((n1, n2) = (0, 1))$. As R_{sbd} decreases (degradation increases) the system however moves away from the steady-state stable state as shown in the Figure 4. Also, the shape of the separatrix changes in such a way that

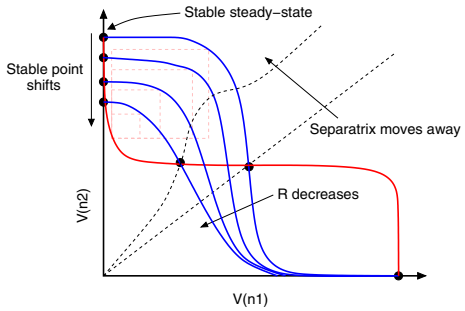


Fig. 4. DC transfer curves and separatrix as a function of R_{sbd}

it moves away from the initial state, thus requiring more energy to cross the separatrix. Essentially, the introduction of R_{sbd} in a bi-stable feedback system changes the stability dynamics and the system becomes *strained*. The strain, which is a function of R_{sbd}, is due to the deviation of the state from the steady-state stable state and the separatrix shape change. As can be noted from Figure 4, the final state of the system as well as the separatrix tend to move back to the steady-state stable state (or relax) as R_{sbd} increases. This tendency is akin to the *relaxation phase* found in many complex systems.

If we define a function θ which represents the *relaxation-coefficient* of the system, the sensitivity of θ with respect to

the severity of oxide degradation is a function of the current state of the system as well as R_{sbd} (denoted as R for brevity). Mathematically, this can be expressed as:

$$\frac{d\theta}{dR} = -C\frac{\theta}{R^{\alpha}}, \qquad \text{for } R_{crit} \leq R \leq \infty \tag{1}$$

Equation 1, when integrated between the R bounds and rearranged, leads to the following equation:

$$\theta(R) = e^{-\left(\frac{R-R_{crit}}{\rho}\right)^{\alpha}}, \text{ for } R_{crit} \leq R \leq \infty \tag{2}$$

where ρ describes the sensitivity to variations in R and α is the stretched exponential coefficient. Due to presence of α, the function $\theta(R)$ gets stretched, leading to a stretched exponential function. In physics, the stretched exponential function, also known as the Kohlrausch-William-Watts (KWW) function, is often used as a phenomenological description of relaxation in disordered systems. In a wide variety of complex system applications, including the modeling of the fluorescence imaging, polymer dynamics and muscle rheology, KWW functions have proven to be accurate in modeling the associated relaxation processes [8], [15]. In mathematics, the KWW function is also known as the complementary cumulative Weibull distribution.

The following properties hold for an arbitrary KWW function, $\theta(x)$ (of the generic form $e^{-\gamma(x)}$, $0 \leq x < \infty$):

- $\gamma(0) = 0$, and
- the derivative $\dot{\gamma}(x)$ (and hence $\theta(x)$) is monotonic.

In the system described in this work, the KWW function is the *relaxation-coefficient* defined by $\theta(R)$ such that $R_{crit} \leq R \leq \infty$. The first property describe above holds for $\theta(R)$, as defined by the following equations:

$$\begin{aligned} \theta(R) &= 0, \qquad \text{if } R = \infty \\ &= 1, \qquad \text{if } R = R_{crit} \end{aligned}$$

The second property also holds since the stability metric is monotonic with respect to change in R (Figure 2).

The dynamic write-ability metric, WL_{crit} is a function of the *relaxation-coefficient*, $\theta(R)$. When the system is relaxed (no oxide breakdown), $\theta(R) = 0$ and $WL_{crit} = WL_0$ (WL_0 is the steady-state WL_{crit} for $R = \infty$ at a given supply voltage). As R approaches R_{crit}, $\theta(R)$ approaches 0 and WL_{crit} approaches ∞. The following equation captures the behavior of WL_{crit} as a function of $\theta(R)$:

$$WL_{crit}(R) = \frac{WL_0}{1 - \theta(R)} \tag{3}$$

Substituting the value of $\theta(R)$ from Equation 2 in Equation 3, we get:

$$WL_{crit}(R) = \frac{WL_0}{1 - e^{-\left(\frac{R-R_{crit}}{\rho}\right)^{\alpha}}} \tag{4}$$

Figure 5 shows the scaling of WL_{crit} with R_{sbd} across the supply voltage range for both SPICE and the analytical model. The values of R_{sbd} range from 1GΩ (fresh oxide) to 10KΩ (hard breakdown). It can be observed that the maximum error between SPICE and the model is less than 1% across the whole supply voltage range. There are three variables in Equation 4, namely, α, ρ and R_{crit}. The value of the stretched exponential coefficient, α, ranges between 0.20 and 0.23 for all voltages and hence assumed to be a constant for a given technology node. The value of ρ increases as

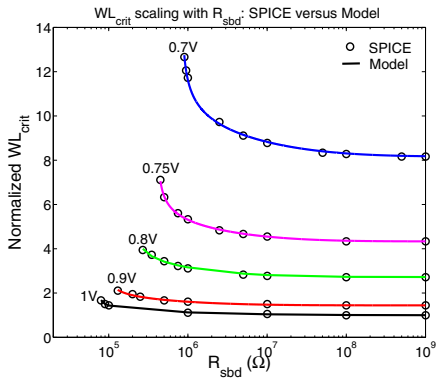
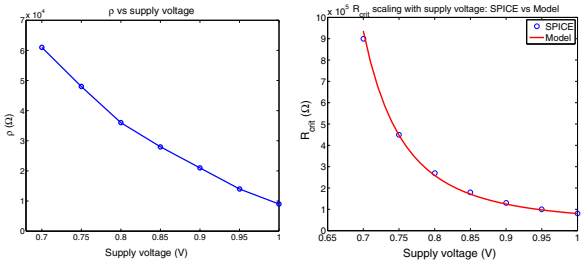


Fig. 5. WL_{crit} scaling with R_{sbd} and supply voltage: SPICE versus Model

the supply voltage decreases and the increase is super-linear. Since ρ signifies the sensitivity to variations in R, it implies that at lower supply voltages, WL_{crit} is more sensitive to changes in R. The value of R_{crit} also changes with supply voltage as shown earlier in Figure 2. The behavior of R_{crit} with respect to supply voltage can be modeled well by the Equation, $R_{crit}(V) = K.e^{\left(\frac{V_{nom}}{V}\right)^{\beta}}$. The value of β is 3.45 and hence the equation models a *compressed exponential*. The physical significance of the compressed exponential is not known. Figures 6(a) and 6(b) show the values of ρ and $Rcrit$ across the supply voltage range.



(a) (b)

Fig. 6. (a) ρ scaling with supply voltage (b) R_{crit} scaling with supply voltage: SPICE vs Model

A few observations can be made from Figure 5. As R_{sbd} reduces, WL_{crit} increases sharply with respect to supply voltage. Additionally, as the supply voltage reduces, the rate of increase of WL_{crit} with respect to R_{sbd} goes up sharply as well. As described earlier in the paper, when R_{sbd} equals R_{crit}, WL_{crit} asymptotically approaches ∞ and the feedback system is statically unstable. The analytical KWW model accurately captures these behaviors across the whole supply voltage range.

V. Modeling Efficacy of Write-Assist Techniques

Write-assist (WA) techniques are commonly used to reduce the V_{min} of SRAMs [3]. Since the increase in WL_{crit} is exponential as the supply voltage scales down, several WA

techniques have been proposed in literature to improve the scaling of WL_{crit} [3]. Of the various WA techniques, the two commonly used in current generation SRAMs are word-line (WL) boosting and Vdd lowering. In this work, we boosted the WL by 100 mV and decreased the Vdd by 100 mV for the two WA techniques respectively.

Figure 7 shows the scaling of WL_{crit} with respect to R_{sbd} for the nominal as well as the word-line boosting cases at 0.75V. As can be noted, the analytical KWW model described

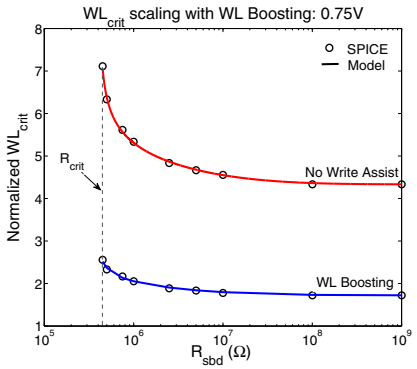


Fig. 7. WL_{crit} scaling for non-WA and WL boosting WA cases

by Equation 4 accurately predicts the behavior of the WA case as well. The WL_{crit} decreases substantially for all values of R_{sbd}, so the WL boosting WA is definitely effective. The improvement in WL_{crit} increases as the supply voltage scales down. However, it is interesting to note that the value of R_{crit} remains exactly same for both cases implying that the WL boosting assist does not change the boundary between dynamic and static stability regions. The value of α also remains unchanged but ρ decreases from 48 KΩ to 30 KΩ. This essentially means that the sensitivity of WL_{crit} to variations in R_{sbd} decreases with word-line boosting WA technique.

Figure 8 shows the impact of Vdd lowering WA on WL_{crit}. Again, it can be noted that the analytical KWW model based on Equation 4 accurately models the behavior of the Vdd lowering WA method. At lower degradation level (higher

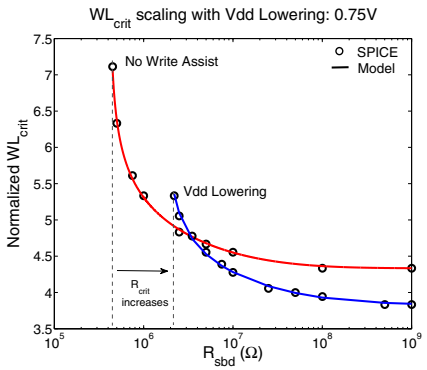


Fig. 8. WL_{crit} scaling for non-WA and Vdd lowering WA cases

R_{sbd}), the improvement due to the assist technique can be clearly seen (though not as much as the WL boosting assist method). As the degradation increases (R_{sbd} decreases), the benefits due to the Vdd lowering WA technique starts to dwindle. Eventually, at certain R_{sbd}, the WL_{crit} with the assist becomes larger than the nominal case and it can be mainly attributed to increase in R_{crit} (the value of R_{crit} increases from 0.412 MΩ to 1.76 MΩ). Since the voltage is lowered for the bit cell in the Vdd lowering WA technique, R_{crit} is a function of the lowered voltage. Hence, for the Vdd lowering WA technique, Equation 4 can be re-written as:

$$WL_{crit}(R, V, V_l) = \frac{WL_0(V)}{1 - e^{-\left(\frac{R - R_{crit}(V_l)}{\rho}\right)^{\alpha}}} \quad (5)$$

where V_l stands for the lower bit cell voltage. This implies that R_{crit} will increase for the Vdd lowering WA technique, which can be observed in Figure 8. The value of α remains unchanged (0.21) but ρ increases from 48 KΩ to 140 KΩ.

VI. Conclusions

This paper presented an accurate analytical model based on the Kohlrausch-William-Watts (KWW) function for analyzing the SRAM write operation in the presence of gate oxide breakdown. We showed that there exists a critical breakdown resistance, R_{crit}, which divides the stability space into dynamically and statically limited regions. As supply voltage decreases, the value of R_{crit} increases thus making the SRAM more susceptible to write failures. The KWW model also accurately models the efficacy of the word-line boosting and Vdd lowering write-assist techniques in reducing WL_{crit}. Simulation results show that the model is accurate to within 1% of SPICE across range of supply voltages and severity of oxide breakdown with orders of improvement in runtime.

References

[1] A. J. Bhavnagarwala *et al*, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE Journal of Solid-State Circuits*, Vol. 36, pp. 658-665, Apr. 2001.

[2] A. J. Bhavnagarwala *et al*, "Fluctuation limits & scaling opportunities for CMOS SRAM cells," *IEDM*, 2005.

[3] V. Chandra *et al*, "On the Efficacy of Write-assist Techniques in Low Voltage Nanoscale SRAMs," *DATE*, 2010.

[4] V. Chandra *et al*, "On the Impact of Gate Oxide Degradation on SRAM Dynamic and Static Write-ability," *ASP-DAC*, 2011.

[5] T. W. Chen *et al*, "Experimental study of gate-oxide early life failures," *IRPS*, 2009.

[6] R. Degraeve *et al*, "Relation between breakdown mode and breakdown location in short channel NMOSFETs and its impact on reliability specifications," *IRPS*, 2001.

[7] B. Kaczer *et al*, "Impact of MOSFET gate oxide breakdown on digital circuit operation and Reliability," *IEEE Transactions on Electron Devices*, Vol. 49, No. 3, pp. 500-506, Mar 2002.

[8] K. C. B. Lee *et al*, "Application of stretched exponential function to fluorescence lifetime imaging," *Biophysical Journal*, Vol. 81, pp. 1265-1275, 2001.

[9] J. Lohstroh, "Static and dynamic noise margins of logic circuits," *IEEE Journal of Solid-State Circuits*, vol. 14, 1979.

[10] T. Nigam *et al*, "Accurate model for time-dependent dielectric breakdown of high-k metal gate stacks," *IRPS*, 2009.

[11] R. Rodriguez *et al*, "Oxide breakdown model and its impact on SRAM cell functionality," *Simulation of semiconductor processes and devices*, 2003.

[12] E. Seevinck *et al*, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid-State Circuits*, 1987.

[13] J. H. Stathis, "Physical and predictive models of ultra thin oxide reliability in CMOS devices and circuits," *IRPS*, 2001.

[14] J. Sune and E. Y. Wu, "Statistics of successive breakdown events in gate oxides," *IEEE Electron Device Letters*, pp. 272-274, 2003.

[15] J. Trzmiel *et al*, "Properties of the relaxation time distribution underlying the Kohlrausch-William-Watts photoionization of the DX centers in $Cd_{1-x}Mn_xTe$ mixed crystals," *Journal of Condensed Matter Physics*, Vol. 21, 2009.

[16] J. Wang *et al*, "Analyzing Static and Dynamic Write Margin for Nanometer SRAMs," *ISLPED*, 2008.

[17] M. Wieckowski *et al*, "A Black Box Method For Stability Analysis of Arbitrary SRAM Cell Structures," *DATE*, 2010.

[18] K. Zhang *et al*, "Low Power SRAMs in nanoscale CMOS technologies," *IEEE Trans. Electron Devices,* Vol. 55, No. 1, pp. 145-151, Jan. 2008.