Microelectronics Reliability xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/microrel

Analysis of time-dependent dielectric breakdown induced aging of SRAM cache with different configurations

Rui Zhang, Taizhi Liu, Kexin Yang, Linda Milor *

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

ARTICLE INFO

Article history:
Received 21 May 2017
Received in revised form 11 June 2017
Accepted 20 June 2017
Available online xxxx

Keywords:
SRAM
Time dependent dielectric breakdown (TDDB)
Monte Carlo simulation
Reliability
Performance
Configuration

ABSTRACT

Time dependent dielectric breakdown degrades the reliability of SRAM cache. A novel methodology to estimate SRAM cache reliability and performance is presented. The performance and reliability characteristics are obtained from activity extraction and Monte Carlo simulations, considering device dimensions, process variations, the stress probability, and the thermal distribution. Based on the reliability-performance estimation methodology, caches with various settings on associativity, cache line size, and cache size are analysed and compared. Experiments show that there exists a contradiction between performance and reliability for different cache configurations. Understanding the variation of performance and reliability can provide SRAM designers with insight on reliability-performance trade-offs for cache system design.

1. Introduction

As a dominant part of Systems-on-Chip (SoC), Static Random Access Memory (SRAM), which takes up half or more of the die area and most of the transistors in modern microprocessors, has pushed the evolution of semiconductor technology nodes for a long time. In SRAM design, transistors are scaled as much as possible to achieve goals on bit density and on performance. The smaller transistors are more vulnerable to reliability failure mechanisms such as Hot Carrier Injection (HCI), Bias Temperature Instability (BTI), and time dependent dielectric breakdown (TDDB). There exist many research papers which have investigated how BTI and HCI affect circuit lifetime and performance [1–7]. In this paper, the performance-reliability trade-off in cache systems is analysed by comparing hit rate (a measure of cache performance) and TDDB induced reliability (failure probability) of SRAM data caches with different configurations embedded in a state-of-art microprocessor. SRAM data cache configurations include varying associativity, cache line size, and cache size.

TDDB is one of the dominant failure mechanisms in modern VLSI systems. It occurs when conducting paths are formed in the Front-End-Of-The-Line (FEOL) dielectric (gate dielectric), Middle-Of-The-Line (MOL) dielectric, and Back-End-Of-The-Line (BEOL) dielectric. According to the location, TDDB is referred to as GTDDB, MTDDB, and BTDDB, respectively. Many studies have characterized GTDDB and BTDDB with modelling and experimental observations [8–11]. The process of breakdown for GTDDB is also reasonably well understood, via the percolation model. And the path to hard breakdown can be modelled with resistive conduction paths in timing analyses [12,13]. Similar models can be applied to BTDDB.

On the other hand, since MTDDB is a recently discovered wearout mechanism for technology nodes beyond 28 nm, few studies have thoroughly explored the reliability issues related to MTDDB. In [14–19], experiments on device-level reliability due to MTDDB have been designed and performed. The characterization methodology for product-level reliability issues related to MTDDB has been presented in [20,21]. In our work, GTDDB and MTDDB are combined to estimate the reliability of an SRAM data cache with different configurations, while considering the features of a FinFET SRAM cell and stress probabilities under various workloads.

As technology node scales to the deep sub-100 nm regime, device structures including thin-body silicon-on-insulator (SOI) MOSFETs, double-gate MOSFETs, triple-gate MOSFETs, and FinFETs have been proposed to overcome scaling induced issues, such as threshold voltage (V_{th}) lowering, leakage increase, drain-induced barrier lowering (DIBL), and subthreshold swing (SS) degradation. Among these novel devices, the FinFET is considered as the most promising substitute for planar MOSFETs beyond the 32 nm technology node.

This work considers an SRAM based on FinFETs, which is designed with the Predictive Technology Model (PTM) and NCSU 15 nm FreePDK [22,23]. It analyses the optimization of the data SRAM configurations in terms of the trade-off between performance and lifetime, where

* Corresponding author.
E-mail address: linda.milor@ece.gatech.edu (L. Milor).

http://dx.doi.org/10.1016/j.microrel.2017.06.040
0026-2714/

lifetime is limited by TDDB. Such analysis can provide SRAM designers with insight on performance-reliability trade-offs for cache system design.

This paper is arranged as follows. Section 2 summarizes SRAM stability and a device-level lifetime model due to GTDDB and MTDDB. Section 3 presents the lifetime assessment framework. Section 4 contains the performance-reliability analysis for different cache configurations. Section 5 concludes this paper.

2. Time-dependent dielectric breakdown models and SRAM reliability

2.1. GTDDB model

The breakdown of a gate dielectric consists of three phases. In the first phase, non-overlapping defect sites (traps) which don't conduct current begin to form in the gate oxide layer of a transistor. Then, with the increase of traps density, the path from gate to channel starts to conduct a small leakage current.

As the traps further accumulate, the transistor enters the third phase which represents its catastrophic failure and the end of life. The failure mechanisms in the second and third phases are known as soft oxide breakdown (SBD) and hard oxide breakdown (HBD).

The overall time-to-failure due to GTDDB depends on activity and temperature. For ultra-thin (less than 5 nm) gate oxide, the time-to-failure (TTF) due to GTDDB can be modelled with Weibull distribution [24] with characteristic lifetime:

$$\eta_{GTDDB} = A_{GTDDB}\left(\frac{F}{WL}\right)^{\frac{1}{\beta_{GTDDB}}}\cdot\frac{V^{a+bT}}{\alpha_{GTDDB}}\cdot e^{\left(\frac{c}{T}+\frac{d}{T^2}\right)} \quad (1)$$

where W and L are gate width and length, respectively, η is the time-to-failure for 63.2% of the sample devices, β_{GTDDB} is the Weibull shape parameter, α_{GTDDB} is the probability of stress, F is cumulative-failure percentile at use conditions, T is temperature, V is the gate voltage, and a, b, c, d and A_{GTDDB} are fitting parameters.

2.2. MTDDB model

Since the circuit supply voltage doesn't scale at the same rate with technology advancement, the aggressive shrinking of the insulator between the gates and the diffusion contacts leads to a greater electric field in the dielectric between the gate and contact. The greater electric field would seriously impact the lifetime of advanced VLSI. The characteristic lifetime due to MTDDB for each dielectric segment with vulnerable length L can be expressed as [20]

$$\eta_{MTDDB} = A_{MTDDB}\frac{L^{\frac{-1}{\beta_{MTDDB}}}}{\alpha_{MTDDB}}\cdot e^{-\gamma E^m}\cdot e^{-\frac{E_A}{kT}} \quad (2)$$

where α_{MTDDB} is the stress probability, β_{MTDDB} is the Weibull shape parameter, γ is the field acceleration factor, k is Boltzmann's constant, E_A is the activation energy, T is temperature, and E is the electric filed through the dielectric segment. $E = V/S_{MTDDB}$, where V and S_{MTDDB} are supply voltage and the space between a gate and a contact, respectively. A_{MTDDB} is a constant which depends on dielectric material properties. m is 1 for the E model, and ½ for the $\sqrt{E}$ model.

2.3. SRAM reliability

In the Leon3 microprocessor [25], the storage blocks, including a window-based register file unit (RF), trace buffer (TB), separate data (D-cache) and instruction (I-Cache) caches, and cache tag storage units (Dtags and Itags), are composed of SRAM cells. We have focused on the lifetime of the L1 data cache due to its high temperature and activity.

The FinFET SRAM cell is implemented with six fin-shaped field-effect transistors (6T). In the 6T structure, four transistors (two PFETs and two NFETs) form two inverters that store data for each bit. The remaining transistors (two NFETs) enable read/write operations under the control of the wordline (WL) signal. In the NCSU FREEPDK15 library, the layer stack for the FinFET includes standard BEOL layers above the Metal1 layer, FEOL layers, which consist of the active and gate layer, and MOL layers, involving AIL1, AIL2, and GIL [22,23]. The MOL layers are proposed to overcome performance degradation and resistance concerns between connected layers in advanced technology nodes. In this study, SRAM cell refers to the structure below the Metal1 (BEOL) layer as shown in Fig. 1. Therefore, GTDDB and MTDDB are the TDDB mechanisms that cause SRAM cell breakdown.

GTDDB occurs in the gate dielectric of transistors, while MTDDB occurs in the dielectric segments between the gate and the drain/source contact when there exists a voltage difference between the gate and the drain/source.

3. Lifetime estimation methodology

Time-to-failure due to GTDDB and MTDDB is a function of device dimensions, process variations, the stress probability, and the thermal distribution. Device dimensions are adopted from FREEPDK15. In this section, the steps for extracting the stress probability and the thermal profile are introduced. Then the SRAM lifetime distribution is characterized while considering process variations.

3.1. Extraction of activity and thermal profile

Activity extraction is realized with FPGA emulation which operates million times faster than RTL or SPICE simulations. I/O ports adapting modules and counters for counting I/O activity and data cache hit rate are added to microprocessor netlists before emulating the leon3 on an FPGA. Then the revised Leon3 system is compiled and downloaded to Altera DE2 FPGA. SRAM activity and state profiles are tracked and output when the emulated Leon3 microprocessors is running standard benchmarks [26]. Next, I/O activity and the gate-level netlist of the Leon3 are used for activity propagation to each net of the microprocessor system. Leon3 microprocessor executed six representative benchmarks in Mibench: Basicmath, Qsort, SHA, CRC32, FFT, and Dijkstra [26]. The extracted duty-cycle distributions of SRAM cells in a 2-way 32 KB data cache are presented in Fig. 2.

Based on the net activity and RC information from the layout, the power consumed by each component of the Leon3 core is obtained for thermal estimation. Then the thermal distribution is calculated in COMSOL, considering the self-heating effects of FinFETs [27]. Fig. 3 shows the temperature distribution when the leon3 processor is running a standard benchmark. The temperature variation throughout microprocessor is taken into account for SRAM lifetime characterization.

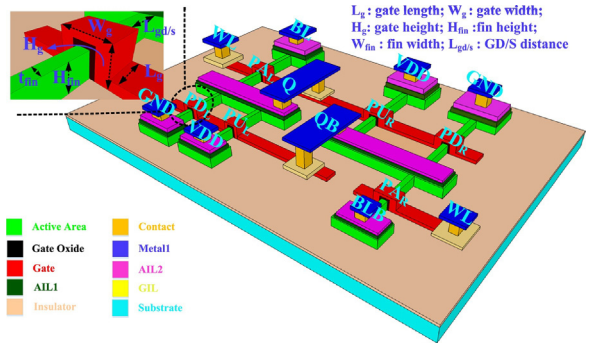

Fig. 1. Layout of a FinFET SRAM cell.

Please cite this article as: R. Zhang, et al., Analysis of time-dependent dielectric breakdown induced aging of SRAM cache with different configurations, Microelectronics Reliability (2017), http://dx.doi.org/10.1016/j.microrel.2017.06.040