A 32nm SRAM Reliability Macro for Recovery Free Evaluation of NBTI and PBTI

Pulkit Jain Ayan Paul Xiaofei Wang Chris H. Kim

Department of Electrical and Computer Engineering, University of Minnesota
200 Union Street Southeast, Minneapolis, MN 55455, USA (Email: pulkit.jain25@gmail.com)

Abstract

A scalable test structure for recovery free evaluation of the impact of NBTI and PBTI on read/write operation in a SRAM macro has been developed. A novel non-invasive methodology keeps the stress interrupts for measurements within a few microseconds, preventing unwanted BTI recovery, while providing a parallel stress-measure capability on 32kb sub-arrays. Measurement results in a 32nm high-κ/metal-gate silicon-on-insulator process show that proposed schemes provides 35mV better accuracy in read V_{MIN} and 10X accuracy in BFR.

Introduction

Bias temperature instability (BTI) is a primary reliability concern in sub-32nm SRAMs [1-3]. NBTI and PBTI under DC stress conditions prevailing in an SRAM cell leads to an increase in read V_{MIN} and a decrease in write V_{MIN} as illustrated in (Fig. 1).

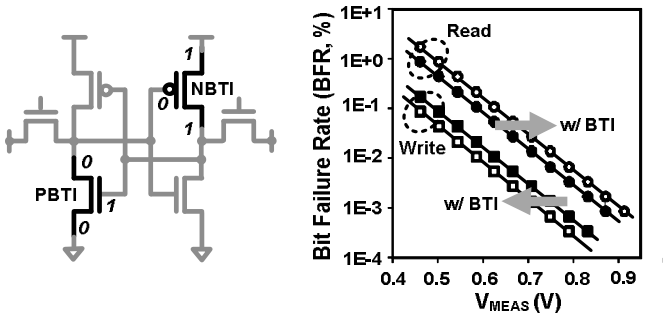


Fig. 1 (Left) SRAM static stress condition promote BTI stress in the two highlighted MOSFETs. (Right) Under the influence of BTI stress, SRAM read V_{MIN} worsens while write V_{MIN} improves.

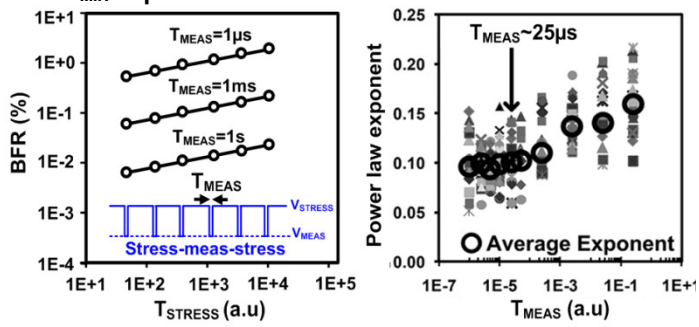


Fig. 2 (Left) Longer T_{MEAS} results in optimistic BTI data (= lower bitcell failure rate) due to the unwanted fast recovery. (Right) Power law exponents measured at different T_{MEAS} indicates a recovery time constant of ~25µs [4].

While there is a pressing need to do an in-situ statistical characterization of BTI on large memory arrays, the phenomenon of fast BTI recovery can lead to inaccurate results if the measurement time, T_{MEAS} is not in the microsecond scale (Fig. 2, [4-5]). In simple test circuits such as ring oscillators, there is flexibility to gate on/off stress applied to small blocks. However, the approach cannot be extended to SRAM/memory. Since, the supply rail is shared globally across all rows, the enormous data running into several megabits, has to be processed in parallel. Moreover, the entire data also needs to be readout off-chip as on-chip storage would be too costly in terms of area. Considering, a typical data acquisition frequency of few megahertz, such a fast measurement becomes problematic. This has been the main limitation for existing approaches ([3][6], table 1). Kim et al. [3], used off-chip control of supply during measurement to obtain the SRAM V_{MIN} during measurements, which takes few seconds to obtain the result, leading to extensive recovery in measurements. Recently, [6] proposed a BFR tracking approach with local data storage similar to this work for fast measurements. However, the overall approach was not scalable to full SRAM arrays and couldn't be used for progressive evaluation of BTI. Instead end-of-life estimation of degradation metric was provided, which has limited use for reliability modeling. Our work proposes the *first* known test structure for recovery free evaluation of NBTI and PBTI on read/write operation in a SRAM macro targeting a microsecond order T_{MEAS}. The main techniques proposed are (i) Pseudo-Reads with deferred Stressed Readout (PR-SR), and (ii) Flip-Latch-Restore with intermittent Scan out (FLR-S). Measurement results in a 32nm HKMG SOI process show a 35mV better accuracy in read V_{MIN} and a 10x more accurate Bitcell Failure Rate (BFR) estimation using a T_{MEAS} of 3µs.

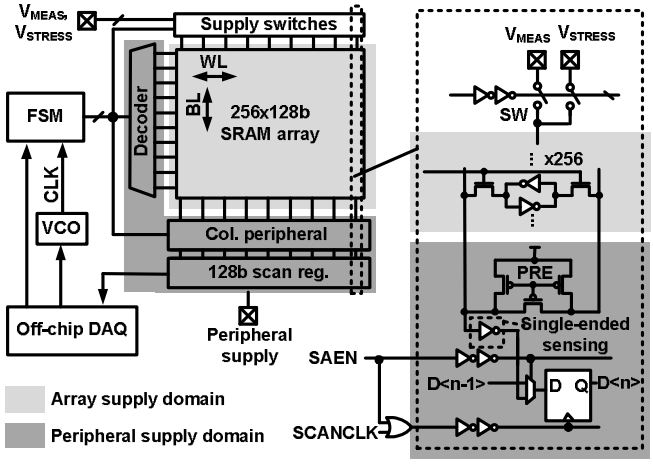


Fig. 3 SRAM reliability macro architecture. Bit-cell array is representative of a product sub-array and features a 128b scan and single-ended sensing for ease of test. BIST

functionality is realized by an on-chip finite state machine that administers the stress-measure-stress sequence.

Proposed SRAM Reliability Macro Design

Fig. 3 shows the proposed SRAM reliability macro. Overall, SRAM specific components are designed to be representative of a product sub-array. For reducing implementation complexity and pin count, we refrained from column multiplex or sense amplifier, and opted for a Single-Ended Sensing (SES) scheme with a slow scan based readout. A marker row with alternate hardwired '1' and '0's was used to verify correct address flow during dynamic operation. The complicated part of the BIST (Built In Self-Test), like controlling the supply switches for measurement and stress modes, measurement times, pulse width control, read/write commands, address sequencing, etc. were handled by the on-chip Finite State Machine (FSM) and voltage controlled oscillator. The slower timings like scans and BFR readout were handled by Labview® off-chip. On-chip supply switches were used on a column wise granularity with delayed firing of signals to reduce current spikes during supply switching and optimize the overall switching time.

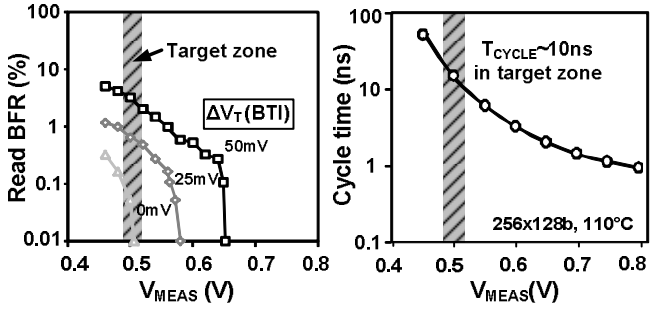


Fig. 4 Simulations of a 256x128b sub-array in 32nm SOI. (Left) Read BFR at different V_{MEAS} and BTI. V_{MIN0} is around 0.5V for target a BFR value of 0.01%. (Right) SRAM cycle time for different V_{MEAS}. Cycle time is ~10ns for the target V_{MEAS}.

Fig. 4 plots show simulated BFR at different operating voltages and BTI shifts. A target BFR of >0.01% from a 32kb subarray for smooth BFR trends mandates a small measurement voltage, V_{MEAS} ~0.5V with a corresponding SRAM cycle time of ~10ns.

Read Timing Sequence

Fig. 5 shows example timing diagrams of the conventional [3] and proposed methods. Prior to applying V_{STRESS}, all bitcells are initialized through a blanket write '0'. Next, the peripheral supply is externally lowered down to $V_{MEAS,}$ a level corresponding to a target read BFR. This completes the initialization step. Next, stress is applied in a stress-measure-stress routine with exponentially increasing stress intervals using an array supply of V_{STRESS}. In the short measure window, the array supply is lowered to V_{MEAS}, using on-chip switches with 20% of $T_{MEAS,}$ dedicated to supply switching. A pseudo-read burst consisting of up to 256 sequential WL perturbations follows next. If we consider an affected row, all cells on it that are 'weak' get a data flip, while others that are 'strong' retaining their original values. Thus pass/fail information corresponding to this measurement interrupt gets stored locally in that same cell. After this, the array supply is switched back to V_{STRESS} to prevent unwanted BTI recovery. We defer the full read and off-chip data acquisition in this stressed stage as the pass/fail info is retained. Due to the long stress periods, this can be done much slowly without interrupting the overall test procedure. Note that since the array operates at a high stress voltage in this state, the chance of any cell failure occurring at this stage is remote. After the BFR has been captured and scanned out, the entire cycle is repeated. An extension of this approach can be used to track $V_{MIN.}$ (Fig. 6). Here, V_{MEAS} is stepped down until a target BFR is reached.

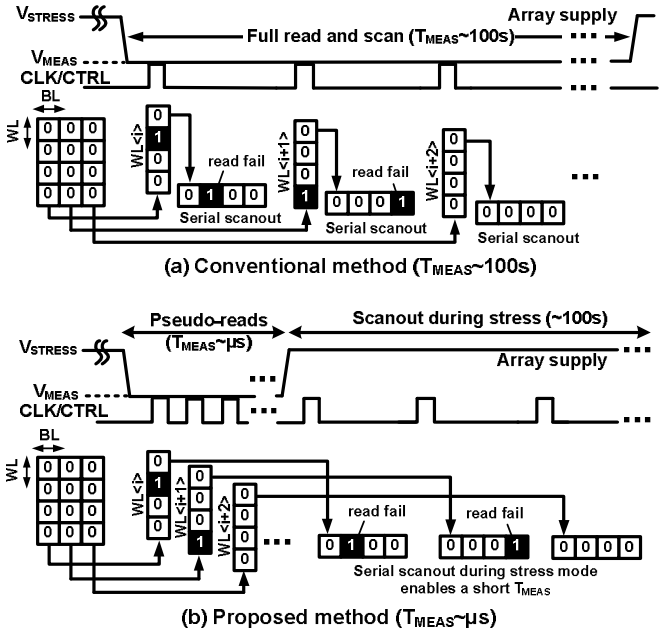


Fig. 5 Read BFR measurement sequence example for an array initialized to zero. (a) In the conventional method, supply is lowered to V_{MEAS} followed by a full read and slow scan out which results in a long T_{MEAS} (b) The proposed approach consists of a pseudo-read (=sequential WL perturbations) which stores pass/fail info in the array. The array is immediately put back into stress mode to prevent unwanted recovery followed by a full reliable read and scan out.

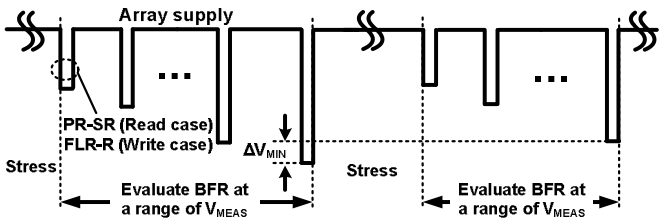


Fig. 6 Extension of the read BFR test sequence in Fig. 5 for read V_{MIN} measurements with microsecond range T_{MEAS}. Here, V_{MEAS} is stepped down until a target BFR is reached. Similar concept can be applied for tracking write V_{MIN}.

Write Timing Sequence

An approach similar to the above would not work for write case. A 'good' cell will flip easily on a write. Consequently, BTI due to the prior DC stress, would start to recover, unless an immediate second flip (or write-back) to the original state is done. Hence, the cell cannot be used as a temporary storage for BTI information, and a full readout into shift registers is needed to capture the first flip information. The ensuing timing sequence is shown in Fig. 7. The initialization step and stress resembles the read case. The T_{MEAS} window consists of the critical flip with array and peripheral supplies kept at V_{MEAS}, followed by a reliable read-latch and restore at V_{NOM}. This biasing ensures that we isolate out the first flip fails. After FLR, array supply goes to V_{STRESS} and we do a slow scan out of the data stored in the on-chip shift registers. Then, FLR-S is repeated for the next row. The main concern here is that the latter rows would observe a somewhat AC stress behavior, which could possibly induce some error. The error was minimized by inserting a programmable offset stress of $1000xT_{MEAS}$ between successive FLR-S steps using the approach claimed in [7]

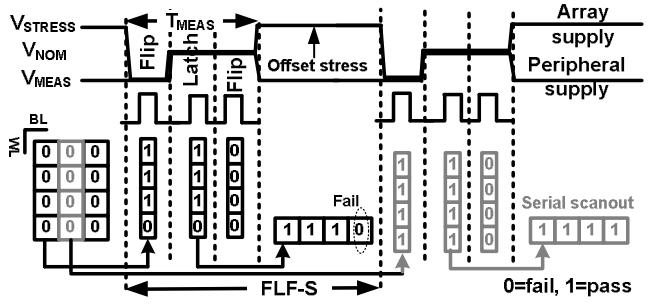


Fig. 7 Write BRF measurement sequence for an array initialized to zero. First, the opposite data is forced (or write 1) at V_{MEAS} Next, supply is raised to V_{NOM} and a reliable full read samples the cell data into a shift register. To prevent the cells from recovering, they are flipped back to initial state (data 0), and the array is put back to stress. A serial scan out is performed at this time.

Read Failure Measurements

Fig. 8 shows read BFR with stress time at different T_{MEAS} showing expected degradation trends. The upper and lower panels correspond to results at 0.52V, 85°C and 0.45V, 25°C, respectively. The right column shows BFR captured after T_{STRESS}=10s at different T_{MEAS}. Over T_{STRESS}=2000s, with T_{MEAS} kept at 3µs, the BFR rises by around 10 times. Without using the proposed PR-SR technique, T_{MEAS} is more than few milliseconds, causing errors of as much as 10-100X in terms of BFR.

Fig. 9 shows the effect of BTI on measured V_{MIN}. Over T_{STRESS}=2000s, and T_{MEAS} =3µs, V_{MIN} changes by an amount close to 25mV. Also, by ensuring an at-least three decade smaller T_{MEAS}, the proposed method alleviates 35mV error from the conventional methods. Note that measurements of V_{MIN} required external supply changes as shown in Fig. 6 leading to larger time between measurement samples. This time discrepancy was calibrated out during post-processing.

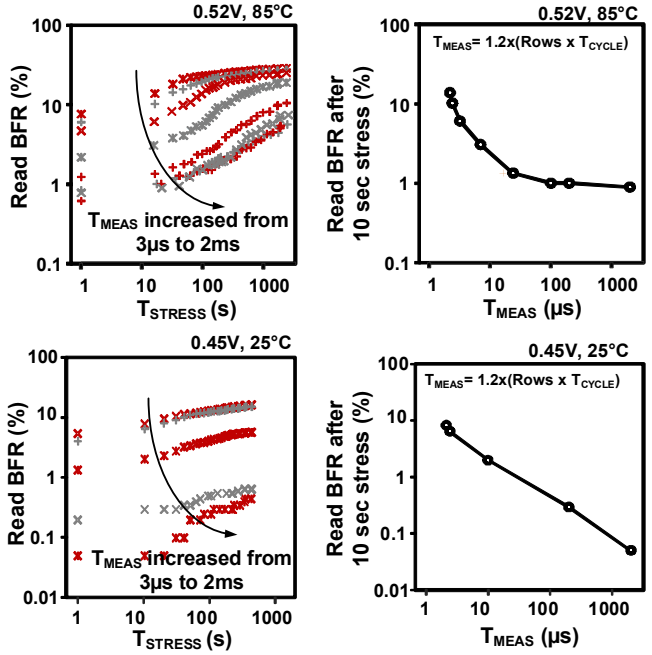


Fig. 8 Read BFR degradation with different T_{MEAS}. BFR at 0.52V, 85°C (upper panels) and 0.45V, 25°C (lower panels). The minimum T_{MEAS} possible by our test setup in order to cover the whole array at T_{CYCLE}=10ns is 3µs (20% allocated time for supply switching). A high BFR range (e.g. >0.1%) was chosen to obtain a smooth BFR curve.

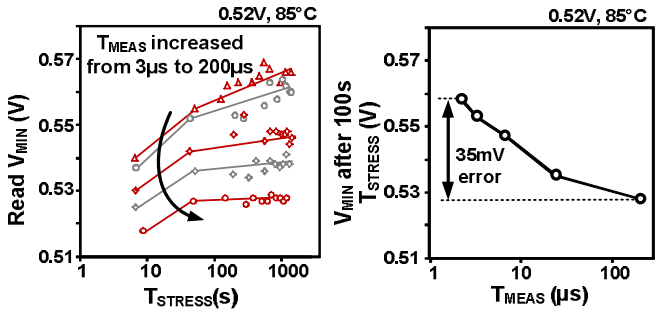


Fig. 9 (Left) Read V_{MIN} versus T_{STRESS} for different T_{MEAS}. (Right) Read V_{MIN} after a 100s stress period as a function of T_{MEAS}.

Write Failure Measurements

Fig. 10 shows the BFR evolution for write case using the test sequence in Fig. 7. As expected, there is an improvement seen in BFR. The sensitivity to T_{MEAS} was found to be much greater than the read, especially at 25°C, and BFR is seen to drop sharply below 3.6µs. At 85°C for T_{STRESS}=2000s, the BFR drops 2x, pointing to lower sensitivity overall to BTI stress, compared to read case. Overall, at least a 100X error in BFR is obtained from the conventional methods due to the smaller T_{MEAS}.

Spatial distribution of the read flips is depicted in Fig. 11 at three measurement interrupts on T_{STRESS}. The upper panel correspond to low initial BFR while the lower panel corresponds to high initial BFR. The marker cells consisting of an alternate 1-0 pattern indicates correct FSM operation. Overall, no observable spatial correlation seen.

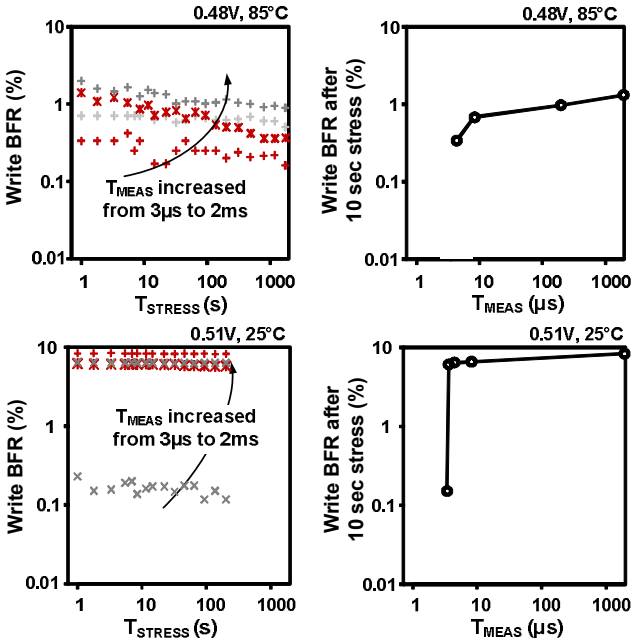


Fig. 10 Write BFR degradation at 0.48V, 85°C (upper panels) and at 0.51V, 25°C (lower panels). Compared to read case in Fig. 8, lower sensitivity seen towards T_{STRESS}, and higher towards T_{MEAS}. Actual stress voltage undisclosed due to confidentiality.

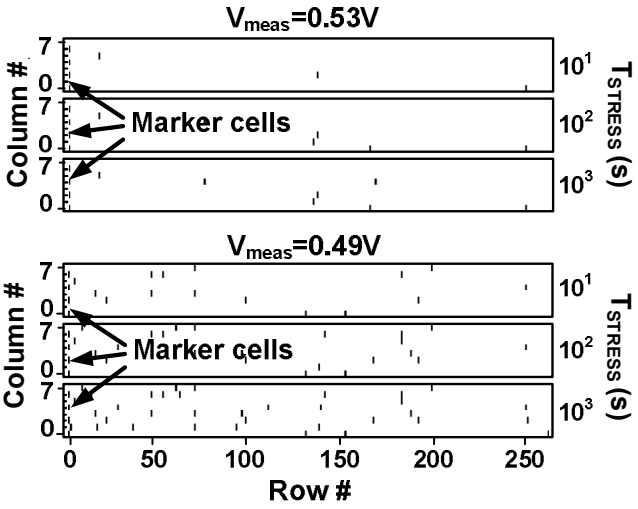
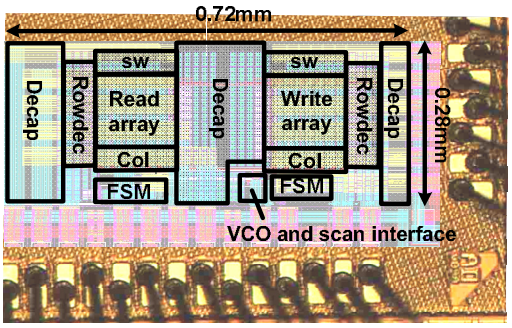


Fig. 11 Spatial distribution of read failures. The array initialized with data '0' in all cells. The black dots correspond to fail cells. No significant spatial correlation observable for fail bits.

The micro-photograph of the SRAM macro fabricated in a 32nm SOI process along with a feature table are shown in Fig. 12. Comparison with previous approaches is given in Table 1.



Process	0.9V 32nm HKMG SOI
Bitcell size	0.898 x 0.269µm^2
Ckt dim.	0.72 x 0.28mm^2
Density	2x32kb RD/WR macros
T_{MEAS}	>3µs @ 0.5V, 100MHz >1µs @ 0.7V, 500MHz
V_{MEAS}	0.45-0.9V
V_{STRESS}	0.9-2.5V

Fig. 12 Test chip micro-photograph and feature summary Measurements were automated using a Labview™ controlled data acquisition board.

Acknowledgements

The authors would like to thank the Semiconductor Research Corporation (SRC) and the Texas Analog Center of Excellence (TxACE) for financial support, DARPA Leading Edge Access Program (LEAP) for chip fabrication support, and industrial liaisons from SRC member companies for technical feedback.

References

[1] A.T. Krishnan et al., IEDM, pp. 4.4.1-4.4.4, 2006
[2] V. Huard et al., IRPS, pp. 655-664, 2010
[3] T. Kim et al., CICC, pp. 231-234, 2009
[4] J. Keane et al., IEDM, pp. 4.2.1-4.2.4, 2010
[5] T. Grasser et al., IEDM, pp. 4.4.1-4.4.4, 2010
[6] S. Drapatz et al., ESSDERC, pp. 146-149, 2010
[7] H. Reisinger et al., IRPS, pp., 448-453, 2006

	[3]	[6]	This work
Min. T_{MEAS}	Few seconds	1ms for 1st row. No macro level data	3µs for a 32kb array
Fail metric	V_{MIN}	Read failures from a single row only	Read or write failure from 32kb array
Features	Read and write mixed	Only read fails evaluated	Read and write separately evaluated
	Cell/sensing fails mixed	Cell fails only	Cell fails only
Evaluation	Progressive	End of life only	Progressive
Type of aging	NBTI only	NBTI only	NBTI and PBTI

Table 1 Comparison to previous approaches for BTI evaluation in SRAM. This work was the first work to target recovery free evaluation from a SRAM macro. Also, the scope of this work was much broader than the previous works by including both Read and Write failure modes affected by combined NBTI and PBTI