Proactive Recovery for BTI in High-κ SRAM Cells

Lin Li*, Youtao Zhang†, Jun Yang*
*Dept of ECE, University of Pittsburgh; †Dept of CS, University of Pittsburgh;
*lil53@pitt.edu, †zhangyt@cs.pitt.edu, *juy9@pitt.edu

***Abstract*—Recent studies of BTI behavior in SRAM cells showed that for high-κ metal gate stack technology, PBTI induced V_{th} shift in NMOS is as significant as NBTI induced V_{th} shift in PMOS. Previous techniques of mitigating NBTI in SRAM focus mainly on PMOS and thus lack the ability to mitigate PBTI of NMOS transistors. In this paper, we propose a novel design to recover 4 internal gates within a SRAM cell simultaneously to mitigate both NBTI and PBTI effects. In the evaluated L2 cache, our technique effectively slows down the cell failure probability increase, and achieves 4.64/2.86× (best/worst case) lifetime improvement over normal design.**

***Index Terms*—high-κ, NBTI, PBTI, recovery, SRAM**

I. Introduction

With technology scaling, transistor reliability has become a major challenge in modern circuit designs. Since smaller, faster and low-power transistors tend to have thinner gate oxide, stronger electric field, and higher temperature during operation, transistors degrade over time due to BTI (Bias Temperature Instability) and other failure mechanisms. In this paper we focus on BTI which includes NBTI (Negative BTI) on PMOS transistors and PBTI (Positive BTI) on NMOS transistors.

For traditional SiO_2 dielectric technology, NBTI has been identified as the dominant effect while PBTI is negligible. Efforts have been taken to study NBTI intensively. These include NBTI modeling [17], studying its impact on circuits [10], and designing mitigation methods for PMOS gates.

Recently, high-κ metal-gate stack (high-κ) technology has been widely adopted in deep sub-micron technologies to gain high speed and low gate leakage. Unlike traditional SiO_2 dielectric technology that has dominating NBTI effect, high-κ technology exhibits comparable stress effect from both NBTI and PBTI [19]. While there are works on PBTI measurements and modeling [8], [19], only limited attempts have been given to studying the impact of PBTI on circuits and its mitigation methods for high-κ material. With the presence of both NBTI and PBTI, the degradation in high-κ based circuits is different from that in SiO_2-based circuits. For example [2], [3] showed that PBTI has a larger impact than NBTI on the degradation of a SRAM cell under the worst case BTI stress. This has made previously designed NBTI-oriented mitigating techniques inadequate, and calls for the development of new schemes for high-κ based circuits.

In this paper, we propose a novel design that can mitigate both NBTI and PBTI at the same time. We first identify that for a SRAM cell, while NBTI on PMOS shows degradations to read static noise margin and read's statistical stability [11], PBTI on NMOS is more critical to the cell's access failure and/or read flip failure. Existing NBTI-only approaches [16] recover some but not all transistors. We then devise a proactive recovery technique that recovers all four transistors in an SRAM cell simultaneously. We introduce a state *4PR* in which the inverse bias is created for all four MOS gates and the voltage level of the internal gate inputs is approximately $0.5V_{dd}$. The technique helps to recover all four MOS gates in a proactive recovery mode, at the same time, with a medium strength but 100% effective recovery time during the recovery period. Our scheme achieves less V_{th} shift, less failure probability and longer MTTF under the same BTI stress. Through testing the SPEC CPU2006 benchmarks, we find that the V_{th} shift is reduced by 4-9.5mV after 10 years, and the MTTF is extended by as much as 4.64 times on average. Our technique is orthogonal to the signal probability balancing technique [1] such that the two can be combined to achieve further reliability improvement.

978-3-9810801-7-9/DATE11/©2011 EDAA

The rest of the paper is organized as: Section 2 discusses the high-κ material and its BTI effects; Section 3 discusses the proactive recovery mode of SRAM circuits; Section 4 discusses the architecture design to utilize the proactive recovery mode; Section 5 shows the experimental results; Section 6 concludes the paper.

II. Motivation

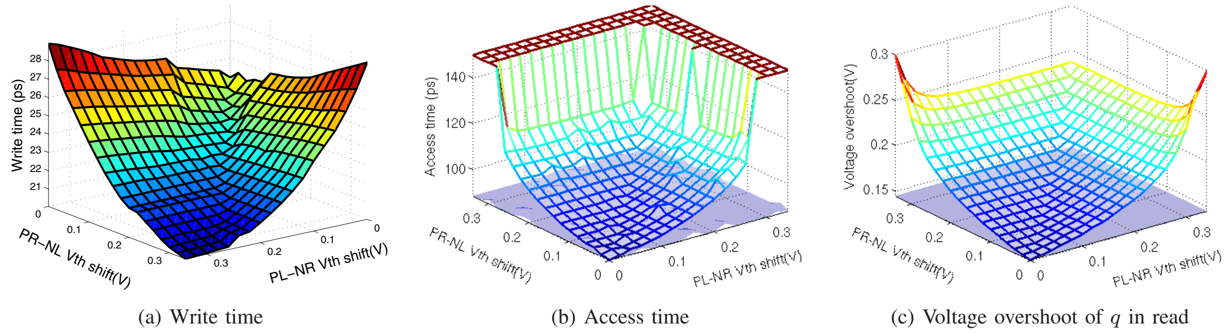
A. Emergence of high-κ material

In sub-45nm technologies, high-κ metal-gate stack is widely used to gain high speed and low leakage. Instead of using silicon dioxide (SiO_2) as the dielectric and poly-silicon gate, the new stack is composed of high-κ material dielectric and the metal gate. The high-κ material has higher relative dielectric constant than that of SiO_2. The metal gates are used on top of the high-κ dielectric. With the new stack structure, scaled gates are less leaky while keeping the high performance.

However, the reliability of the high-κ metal gate stack is yet to be solved. For SiO_2 dielectric, NBTI in PMOS has been identified as the limiting reliability issue, while the effect of PBTI in NMOS is negligible. For high-κ metal-gate stack, recent measurements [19] on V_{th} shift show that NBTI in PMOS is as severe as in SiO_2/poly-silicon. Moreover, PBTI in NMOS is now comparable to NBTI in PMOS. PBTI generates most bulk traps in the added high-κ dielectrics, which is comparable the traps generated to at the interface of SiO_2 and Si in NBTI.

Fortunately, the recovery mechanisms of NBTI used to mitigate the degradation are also observed for PBTI in high-κ dielectrics in experimental conditions. For example, V_{th} shift under PBTI is naturally healed after removing the stress and making V_{gate}, or $V_g = 0V$ [8], [9]. We term this phenomenon as "Natural Recovery". Stronger and accelerated recovery is achieved when an inverse bias (discharge bias in [8], [13]) is applied to V_g. For NBTI, the acceleration of recovery is also observed. We term this method as a "Proactive Recovery" scheme.

Both NBTI and PBTI are caused by high temperature, strong electric field, and most importantly, high stress duty cycles. Typical recovery schemes for NBTI exploit the opportunity to maximize the recovery cycles for PMOS gates [6]. However, in the presence of both NBTI and PBTI, the duty cycles for NMOS and PMOS are complementary: a gate input of "1" incurs stress on an NMOS and recovery on PMOS, while a gate input of "0" incurs recovery on NMOS and stress on PMOS. Therefore, maximizing (natural) recovery cycles for one type of gate could stress the other gate even more. The purpose of this work is therefore to redesign recovery schemes that mitigate both NBTI and PBTI at the same time.



(a) Write time (b) Access time (c) Voltage overshoot of q in read

Fig. 1. Cell stability under 1. NBTI only (surface plot) 2. Both NBTI and PBTI (mesh plot)

B. Effect of NBTI

The direct consequence of NBTI stress is the PMOS V_{th} increase due to the generated interface traps. The increased V_{th} then causes the performance and the reliability of both combinational circuits and SRAM cells to degrade. For combinational circuits, the V_{th} increase in the PMOS slows down the low-to-high transition due to a weaker drive strength. If V_{th} increases too high, the transition may fail to meet the timing requirement of the circuit. Previous study has shown that although the increased delay caused by NBTI is significant for a single gate, its impact on a combinational circuit is relatively low [10]. On the other hand, NBTI in combination with process variation can severely degrade the reliability of memory cells [10]. Such degradation can greatly reduce the statistical read stability and read static noise margin of the cell [11]. Therefore, we target the on-chip SRAM cells and mitigate both their NBTI and PBTI degradations in this work.

C. Effect of both NBTI and PBTI

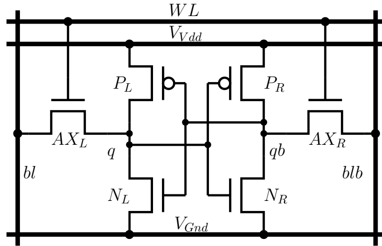


Fig. 2. Standard SRAM Cell.

Fig. 2 shows a conventional 6T SRAM cell. When "1" is stored ($q = 1, qb = 0$), P_R and N_L are under natural recovery, and P_L and N_R are under stress. When "0" is stored, P_R and N_L are under stress, and P_L and N_R are under natural recovery. As we can see, the signal probability of the cell determines the V_{th} shift of the P_R-N_L pair and P_L-N_R pair. We ignore the stress on the access transistor due to its average low duty/busy ratio. We will assume a cell storing value 0 with $q = 0, qb = 1$ for the ease of discussion in the remainder of this section.

V_{th} variation on gates can lead to the following 3 failure mechanisms in a SRAM cell [14]: write failure, access failure, and read flip failure. The write failure happens when writing an inverse value to the SRAM cell cannot finish before the word-line closes and the transaction rewinds. The access failure happens when the bit-line does not discharge below the sense amplifier's trigger threshold such that it fails to sense the value. The read flip failure happens when an inverse value is written in the cell during a normal read operation, since the voltage at q (storing a "0") is driven too high by the pre-charged bit-line and a flipping process is initiated. The SRAM cell fails if any of the above 3 failures occurs. We do not include the hold failure analysis since it is for special operations. Note that process variations (PV) during fabrication can also lead to V_{th} shifts which are additive to BTI degradation induced shifts. We will first discuss BTI effects without PV for clarity, and later present experimental evaluations with PV for completeness.

To evaluate failures caused by V_{th} shifts, we measured the write time, access time and voltage overshoot magnitude at q during a read (which can cause read flip failure) for an SRAM cell while V_{th} increases, shown in Figure 1. The dynamic simulation-based failure criteria is used instead of static criteria such as SNM and WSNM, since RSNM tends to overestimate the dynamic read failure and WSNM underestimates the dynamic write failure [12]. In simulating one cell, we used equivalent resistance and capacitance load extracted from a 128_bit×256_word SRAM bank, to obtain reasonably accurate timing and expedite the simulation. We will also compare our observations with [2], [3].

In Figure 1, the mesh plot shows results when only NBTI induced V_{th} shift is considered, and the surface plot shows the results when both NBTI and PBTI are considered. We do not put any timing constraints on read or write to show clearer trends with extreme V_{th} ranges ($0{\sim}300mV$) in the figure and no variations on access transistors) for a clear comparison. Note that in reality, V_{th} is in a much smaller region, e.g. $0{\sim}100mV$, but its impact on those failures still complies with the observations we make.

Fig. 1(a) shows write time variations with V_{th} increases. The lower the better. The mesh and surface plots almost overlap since with or without PBTI, the write times only differ by 2ps at most, indicating that PBTI does not have significant impact on write time. In other words, the NBTI effect on PMOS dominates the write time. Also, the weakened PMOS due to NBTI further weakens the pull-up strength, which makes the pull-down relatively easier. This is because the PMOS in an SRAM cell is designed weaker than the NMOS, and is more sensitive to V_{th} shifts. As a result, the write time even improves when V_{th} shifts becomes larger. This is also observed in [3]: writability is improved under symmetric degradation; only in worst case write time degrades marginally with PBTI.

Fig.1(b) shows the access time changes. Similar to Fig. 1(a), the lower the value the better the reliability. The plot clearly shows that when both NBTI and PBTI are considered (the mesh plot), access time increases quickly. However, if only the NBTI stress is considered (the surface plot), the changes in access time is not noticeable. This is because the access time is mainly determined by the driving capability of the pull-down path of the cell which does not involve the two PMOS. Hence, the degradation in the PMOS has little effect on the access process. In [2], the read access time is more sensitive to PBTI. As we can see, when considering the PBTI stress on NMOS gates,

the access failure becomes a serious reliability problem, which was not surfaced when NBTI was considered only.

Fig.1(c) shows the voltage overshoot of q in a read operation with respect to V_{th}. The higher the worse. Again, we see that the mesh plot rises above the surface plot, meaning that the read flip failure is more likely to happen when PBTI is counted. Besides PBTI affecting read stability by raising the voltage overshoot, the NBTI also degrades the read stability by reducing the flip trip point. In [3], read SNM is more sensitive to PBTI compared to NBTI. Hence, PBTI worsens the read flip failure that is already threatened by NBTI.

To summarize, the NBTI plays an important role in write time, while the PBTI plays an important role in access time. The read stability is both affected by NBTI and PBTI. Among all the failures, the access time failure dominates other SRAM cell failures [14] (in Monte-Carlo simulation). Without considering PBTI, the access reliability would be treated overly optimistically. Hence, considering both NBTI and PBTI are mandatory to evaluate the reliability of high-κ metal-gate stack technology.

III. Recovery Circuit

In this section, we discuss existing approaches [16] of mitigating the degradation caused by the BTI. Then we propose our method to improve the recovery strength for the SRAM cell circuits.

A. Existing approach

Under normal condition of a SRAM cell, one PMOS-NMOS pair is stressed ($source = drain = 1$, $gate = 0$ for PMOS) while the other is under natural recovery ($source = gate = 1$ for PMOS). The proactive recovery technique used in [16] turns the natural recovery into a strong recovery ($source = 0$, $gate = 1$) to suppress PMOS V_{th} shift. This is implemented through forcing the virtual power (VV_{dd}) to ground for the PMOS being recovered. Since the other PMOS is still under stress, the proactive recovery must be applied alternatively on both sides, and two virtual power lines are necessary to separate the two sides.

A simple extension to it is to apply proactive recovery to one PMOS-NMOS pair, instead of just one PMOS. Take P_L-N_R pair as an example: besides driving the virtual power of P_L to ground, the virtual ground of N_R is also driven to V_{dd}. Two virtual ground lines are necessary to separate N_R and N_L. Hence, the gate is positively/negatively biased to the source for P_L/N_R, resulting proactive recovery on P_L-N_R pair. We refer this technique as "Single Pair PR" ("SP PR" for short). "SP PR" will be used in rest discussion, when compared with our design since PBTI stress on NMOS cannot be neglected with high-κ technology.

The limitations of "SP PR" are two fold. First, in the recovery mode, only one PMOS-NMOS pair is being recovered while the other is still being stressed. Thus, the net improvement comes from the difference between proactive recovery and natural recovery. Second, this approach requires sources of 4 MOS connect to 4 individual virtual source/ground wire, resulting increased wiring overhead. Hence, we need to design an effective and efficient recovery scheme to mitigate both NBTI and PBTI effect.

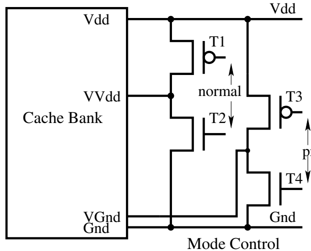


Fig. 3. Mode control design.

B. 4T proactive recovery for SRAM cells

The technique we propose in this paper targets to recover all 4 transistors in the SRAM cell simultaneously (not counting the 2 access transistors), as opposed to only 1 transistor in the previous work [16]. In a regular SRAM cell storing a "0", P_L and N_R are under natural recovery while P_R and N_L are under NBTI and PBTI stress respectively. Storing a "1" incurs the mirror effect of recovery and stress. Our proposed technique will turn all four transistors into a stronger recovery mode. Consequently, we tackle both NBTI and PBTI degradation in those 4 transistors, achieving much higher reliability improvement.

Our recovery scheme can be very easily implemented. The idea is that on a proactive recovery, the source terminals of the 2 PMOS gates are pulled to ground, and the ground terminals of the 2 NMOS gates are pulled up to V_{dd}. The circuit design is shown in Fig. 3. Assume that the power and ground terminals of all SRAM cells in a cache bank are connected to virtual power (VV_{dd}) and virtual ground (VG_{nd}). We add a mode control circuit including T_3-T_4, for VG_{nd}, and T_1-T_2, for VV_{dd}. We also add drivers to drive T_1-T_4. In reality, these drivers will be large since there is one such control circuit per cache bank. By setting different inputs to T_1-T_4, the SRAM cell can switch between *Normal mode* and *4T Proactive Recovery (4PR) mode*. The evaluation of the added drivers and virtual power/ground wires shows negligible impact on write time (0.56%), access time (0.33%) and read stability (0.36%) with the V_{th} degrading from 0 to $100mV$. Note that we don't count PV of drivers since they are less sensitive to PV due to the large size.

- *Normal mode*: T_1 and T_4 are on, T_2 and T_3 are off, virtual power/ground connects to physical power/ground.
- *4PR mode*: T_2 and T_3 are on, T_1 and T_4 are off, virtual power/ground connects to physical ground/power.

The *4PR* mode is intuitively errant since in SRAM cell the pull-up PMOS is connected to ground and pull-down NMOS is connected to power. Fig.4 (left figure) shows the transient analysis when entering from the *Normal* mode to the *4PR* mode. The *Normal* mode control signal "normal" is connected to T_1 and T_4, and *4PR* mode control signal "pr" is connected to T_2 and T_3. Initially, the SRAM cell is in *Normal* mode, storing a "1". When "normal" and "pr" commands flip, the cell enters the *4PR* mode. With the falling of "normal", both VV_{dd} and q drop. With the rising of "pr", both VG_{nd} and qb rise. When q falls and qb rises such that $q - qb < V_{th}$ is approached, P_L and N_R are off, and the cell enters the *4PR* mode. Note that N_L and P_R are always off during this process. After all 4 transistors are turned off, the q and qb slowly approach to approximate $0.5V_{dd}$, due to the sub-threshold leakage discharge of q by P_L, and charge of qb by N_R. Thus, in the *4PR* mode, all the 4 gates are off, and the internal q and qb reach about $0.5V_{dd}$. Note that we used 45nm high performance model [4] (lower V_{th} for MOS) in this study. If the low power model (higher V_{th} for MOS) is used, there is a voltage gap between q and qb when they stabilize.

Note that we have been using "off" to indicate that the gates are not in the saturation mode. However, they are not entirely off. In

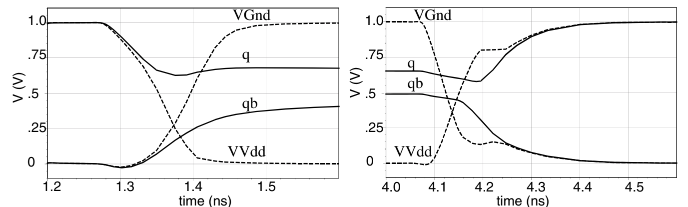


Fig. 4. Entering (left) and exiting (right) 4PR mode from the Normal mode.

fact, all 4 gates are in the *proactive recovery* mode. The “S” and “G” terminals of the two PMOS are 0 and $\sim 0.5V_{dd}$, forming a positive bias (inverse) of V_{gs}. The “S” and “G” terminals of the two NMOS are 1 and $\sim 0.5V_{dd}$, forming the negative bias (inverse) of V_{gs}. Positive V_{gs} on PMOS and negative bias V_{gs} on NMOS (all are inverse bias) result faster and stronger recovery than natural recovery ($V_{gs} = 0$) [8], [9], [13]. Although this recovery strength are weaker than using full inverse bias ($V_{gs} = V_{dd}$ in [16] and $V_{gs} = -1V$ in [13]), the measured results in [13] show that we can still achieve 89% of V_{th} shift recovery compared to using full positive bias, while the natural recovery can only achieve 50%. Moreover, since there are no transistors under stress in our *4PR* mode, we can double the recovery time that was achieved in [16]. Also, the concurrent recovery for mirror MOS's eliminates the requirement for more complex circuit and alternating recovery for both sides.

C. Further discussions

To evaluate the effectiveness of our *4PR*, we need a closer examination of recovery characteristics under different magnitudes of discharge bias in real SRAM cell. As mentioned earlier, in multiple experiments [8], [9], [13], [15] the recovery is accelerated under inverse bias of V_g with source, drain and substrate terminals grounded. Although the settings (temperature, stress time, stress voltage, measurement method) of these experiments vary, it is clear that the inverse bias accelerates the PBTI recovery significantly. In [8], at 10^4 seconds, applying V_g of $-1.2V$ results in 43% recovered V_{th} shift, comparing to 8% for grounded gate. In [15], the final degradation is reduced by 31%, compared to that after same cycles of alternating stress and natural recovery. In [13], applying $-1.5V$ bias achieves full recovery, while applying $-0.5V$ achieves 87% recovery and natural recovery only achieves 48%.

However, in real SRAM cell circuits both the natural recovery and proactive recovery under *4PR* mode differ from that in experiments. In natural recovery, inverse bias occurs at the gate-drain overlap (for NMOS, $V_g = V_s = 0, V_d = V_{dd}$), contrary to the setting of natural recovery in experiments without any bias (for NMOS, $V_g = V_s = V_d = 0$). In *4PR*, a mild inverse bias is forced at the original gate-source overlap (for NMOS, $V_g = V_d = 0.5V_{dd}, V_s = V_{dd}$), which is different from the mild/full inverse bias on both gate-drain and gate-source ends.

For natural recovery, in [15], the inverse bias at the gate-drain end shows negligible impact on recovery strength: even raising the V_{ds} from $1.2V$ to $1.6V$ only results in limited improvement of recovered V_{th} from 3% to 10%. The great diminish is due to the reduced electrical field in the channel where there is no inverse bias for V_{gs} or V_{gb} and consequent reduced electron tunneling current for charge neutralisation. Although inverse V_{gd} marginally accelerates recovery, it might be hasty to assume the inverse V_{gs} bias in *4PR* has more impact on recovery strength than V_{gd}. Since there are no experiments performed under varying gate-source bias for authors' best knowledge, it is hard to quantify the individual impact of 2 inverse biases (V_{gs}, V_{gd}) on the detrapping process and the recovery strength. Since *4PR* puts gates into a new recovery state, we conservatively studied two extreme cases to include possible scenarios where V_{gd} might play a role.

1) *4PR*-best. The first case is that V_{gd} has little impact on the gate recovery, so only V_{gs} dominates the recovery strength, as with all previous researches. The *4PR* can achieve strongest recovery in this case.
2) *4PR*-worst. The second case is that V_{gd} has the same impact on gate recovery, so combined V_{gs} and V_{gd} should be used to determine the recovery strength. The *4PR* can achieve weakest recovery in this case.

TABLE I
RECOVERY COMPARISON: THE MAGNITUDE OF V_{gs}, V_{gd}, RECOVERY STRENGTH AND TIME UNDER DIFFERENT MODES. P REPRESENTS PMOS. N REPRESENTS NMOS.

condition	V_{gs} (P/N)	V_{gd} (P/N)	strength	time
Stress	-/+V_{dd}	-/+V_{dd}	N/A	1-x
Natural	0	+/-V_{dd}	0.5	min(x, 1-x)
SP PR	+/-V_{dd}	+/-V_{dd}	1	50%
4PR-best	+/-0.5V_{dd}	0	0.89	100%
4PR-worst	+/-0.5V_{dd}	0	0.52	100%

However, even under the worst case, it is expected that more V_{th} shift is recovered when the inverse V_{gs} bias is applied in experiment of [15] after multiple cycles of stress and natural recovery. Since in real circuits the gate-drain has the inverse bias while the gate-source has only 0 bias, the recovery on gate-drain is always stronger than gate-source. Then, the accumulated trapped charges in dielectrics at the gate-source end is denser than at the gate-drain end. As a result, periodically applying inverse V_{gs} effectively removes the charges, compared to the saturated recovery at the gate-drain end.

We list V_{gs}, V_{gd}, recovery strength and recovery time for different modes of P/NMOS in Table I. Take the NMOS as an example: in the stress mode, $V_{gs} = V_{gd} = V_{dd}$. In the natural recovery mode, $V_{gs} = 0$ and $V_{gd} = -V_{dd}$. In the “SP PR” mode, $V_{gs} = V_{gd} = -V_{dd}$. Previous study [13] has shown that a full inverse bias to the gate results in full recovery of V_{th} while a zero bias results a 50% of recovery. Hence, the recovery strength ratio of Natural vs. SP PR is 1:2. Regarding recovery time, the SP PR mode dedicates certain percentage of time for recovery on SRAM cache bank. Within this recovery time, only one PMOS is recovered, so the total effective recovery time for PMOS gates is 50%. Outside the recovery time, the PMOS still enjoys the natural recovery opportunities. For natural recovery, the recovery time of any gate in a cell depends on the cell's signal probability x. Since the recovery time is limited by the weakest gate in the cell, the effective recovery time is always less than 50%.

The recovery strengths of *4PR*-best and *4PR*-worst are shown in Table I. The full recovery under SP PR mode is normalized as 1. In *4PR* mode of an NMOS, $V_g = V_d = 0.5V_{dd}$ and $V_s = V_{dd}$. Hence, $V_{gs} = -0.5V_{dd}$ and $V_{gd} = 0$. In *4PR*-best, V_{gd} can be ignored and V_{gs} is half of that in SP PR mode. Hence, *4PR* can achieve 89% of the recovery strength of SP PR mode, for all 4 gates. The *4PR*-worst mode might be weaker than the Natural mode but due to the effective recovery on gate-source end, the recovery score is 52%.

When it comes to recovery time, the *4PR* mode has a clear advantage over all other modes. As we will explain later, both *4PR* and SP PR modes have the same dedicated time for SRAM bank recovery. During this time, *4PR* ensures recovery for all 4 gates while SP PR provides recovery for only one PMOS gate. Outside this recovery time, both schemes enjoy natural recovery opportunities.

The last issue is the potential impact of body terminal on the gate recovery. Although this impact is likely to be small, it can still be easily solved by applying a proper body bias to cancel its effect on the gate.

IV. RECOVERY ARCHITECTURE

A. Using a spare bank for recovery

We apply our proposed recovery technique in L2 cache, similar to the one in [16]. Since L2 cache is naturally sub-banked, the basic recovery scheme is to recover each bank in rotation. However, since the bank loses data once entering the recovery mode, we

adopt similar strategy as in [16] to leverage the spare bank for data backup and restore. At any time, there are 64/32/16 cache banks (3 configurations) in the normal mode and 1 cache bank, either existing or spare cache bank, in the recovery mode. The recovery cycles varies from 10K to 5000K cycles.

With a spare cache bank, the overheads in our design include area, latency and energy. The latency and energy overheads come from copy data to/from the spare bank. Alternatively, we can also simply discard the bank's contents (with dirty lines flush) for recovery and use the spare to make up for the cache capacity. No data copying is necessary. The latency and energy overheads come from the cold start of the empty spare bank, which generates more memory accesses. We will consider the former design in this paper.

B. Design overhead

We built a 128_bit×256_word bit bank (without considering the peripheral circuitry) using 45nm technology to measure the design overhead, including the power, latency and area. The power includes mode transition power, data read/write power for backup and restore, and leakage power. The breakdown of each category is listed in Table II.

TABLE II
ENERGY FOR TEST 128_BIT×256 SRAM BANK.

Dynamic Energy		Leakage Power	
operation	energy(pJ/bit)	mode	power (mW)
read	0.043	Normal	1.906
write	0.315 (max)	SP PR [16]	1.424
pre-charge	0.78 (read preceded)	*4PR*	1.409
	4.37 (write preceded)		
mode switch	100.35 per bank		

The mode switch power is incurred on each mode switch: the mode control circuit (with 5-stage drivers for VV_{dd} and VG_{nd}) consumes $100.35pJ \times 2 = 0.201nJ$ for the whole bank. Additional cache bank copy and restore consists of 2×256 reads and writes which amount to 0.421 and 2.399nJ respectively. In summary, the dynamic energy overhead in each recovery transaction is a little over 512 normal reads and 512 normal writes. Consider that this extra energy is charged on every $10^4 \sim 10^6$ cycles, the overall dynamic energy overhead is very small.

For the leakage power, the test bank under *SP PR* and *4PR* modes consumes 1.424mW and 1.409mW, respectively. The reduction over the *Normal* mode (1.906mW) is due to the reduced V_{DS} of the gates in cut-off mode. For *SP PR*, the V_{DS} of the MOS under proactive recovery is 0; for *4PR*, the V_{DS} of 4 transistors are $\sim 0.5V_{dd}$.

We also measure the latency of mode switches between *Normal* and *4PR* . Based on our test SRAM bit bank, with a properly designed 5-stage driver, it takes less than 5ns to finish the transition. Compared to the recovery cycles of 10K to 1000K, the transition overhead can be omitted.

The area overhead mainly comes from the spare cache bank, virtual power/ground wires and virtual power/ground drivers. We use the same spare cache bank and same number of wires for virtual power ground as in [16]. The only difference is that our design uses two wires for virtual power and ground, and the previous design uses two wires for left/right side virtual powers. As shown in [16], the virtual power wires already exist in modern memory designs, so our wiring requirement does not incur too much overhead. The spare bank's area overhead is 1/17 at most since there are altogether 17 banks with 1 being the spare. The area of drivers is roughly proportional to bank size, and is found to be negligible.

V. RESULT ANALYSIS

A. Experiment setup

We build the aforementioned SRAM bit bank circuit based on the PTM 45nm high-performance model [4] for high-κ metal-gate stack. To measure the failure events discussed in Section II-C, we use the metric of dynamic stability through transient analysis in Synopsis Hspice [12] due to its high accuracy. The normal circuit is designed with certain guard-band according to the specification, then the circuits with V_{th} shift are simulated. The write time, access time, voltage overshoot during a read are all measured and displayed in Fig. 1(no PV). The same framework is used in evaluating cell failure stability (Section V-C with PV).

Since the degree of stress in the 4 gates of a SRAM cell depends on the stored value, we extract the signal probability of sample bits in a 256KB 16-way L2 cache by running the SPEC CPU2006 benchmarks on the PTLsim [18] micro-architecture simulator using X86 ISA. We fast-forward the initialization phase, and run 1 billion instructions of all 20 benchmarks. The extracted signal probability is used to calculate the V_{th} degradation by Eqn (1) [11]:

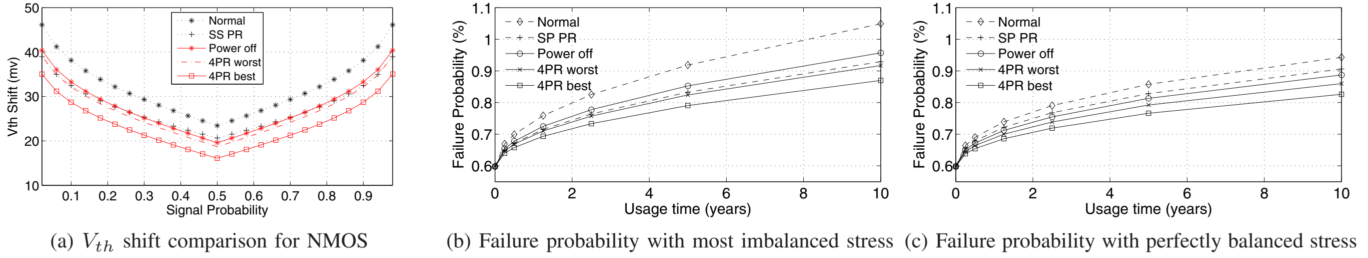
$$\Delta V_{t,AC} = K_{AC} \cdot t^n = \alpha(S) \cdot K_{DC} \cdot t^n \qquad (1)$$

Here K_{DC} is the technology dependent constant. $\alpha(S)$ is the AC factor determined by the signal probability of the cell S, independent of operating frequency f. $\alpha(S)$ is calculated using the multiple cycle stress and recovery model: the stress phase is modeled by the popular accepted power-law model with the exponent $n = 0.16$ for P/NMOS [7], [8]; the recovery phase is modeled by the accurate dispersive model [7]. The parameters (B and ξ) of dispersive model for P/NBTI are individually calibrated to match the experiment data in [15]. For stronger recovery, both in the single-pair proactive recovery mechanism derived from [16], denoted as "SP PR", and our *4PR* with stronger and faster recovery, the parameter B is calibrated by the published data in [9], [13].

B. V_{th} shift comparison

We first compare the effectiveness of the aforementioned schemes in suppressing V_{th} shifts: *Normal* mode, "SP PR" which is the single-pair proactive recovery derived from [16], *4PR-worst* and *4PR-best* are two extremes of our proposed method. In addition, we also experimented for the "Power Off" scheme which turns off the cache bank and all voltages are 0 such that all 4 gates are put in a weak-natural recovery mode (Normal recovery still has a bias in V_{gd}). We include this case because it is also a practical recovery scheme and natural to apply in reality. Fig. 5(a) shows the larger V_{th} shift (in 10 years) between two NMOS (similar trend for PMOS) in SRAM cell versus different signal probabilities in 16 normal with 1 spare bank configuration. All curves are symmetric to signal probability of 0.5, because under signal probabilities of α and $1 - \alpha$, the V_{th} shift of two pairs are symmetric. And the best result happens when the cell stores 0 and 1 in a 50%-50% split in time where stress is perfectly balanced to the two pairs.

The largest V_{th} shift is always seen in the *Normal* mode, as no strong recovery mechanism is performed, and natural recovery alone is inadequate. Then comes the "SP PR", which are nearly equally effective with "SP PR" being slightly better for balanced stress and "Power Off" being slightly better for unbalanced stress. This is not surprising as both techniques intentionally recover internal MOS gates and "SP PR" uses strong recovery for 1 PMOS-NMOS pair but "Power Off" uses weaker recovery for all 4 MOS's. Finally, our proposed *4PR* wins over the previous ones in all data-points even in the worst case since we use stronger recovery (than "Power Off") for



(a) V_{th} shift comparison for NMOS (b) Failure probability with most imbalanced stress (c) Failure probability with perfectly balanced stress

Fig. 5. V_{th} shifts comparison and cell failure probabilities considering PV.

all 4 MOS's and longer recovery time than SP PR. On average, 4PR reduce the V_{th} shift by 9.5mV and 4mV compared to *Normal* and "SP PR" in best case. For configurations of 64/32 banks + 1 spare bank, the reductions are 6.2mV/4.9mV and 3.3mV/2.6mV compared to *Normal* and "SP PR". Considering the improvement of lifetime, 1/17 area overhead is tolerable.

C. Cell failure probability analysis

In addition to BTI stress induced V_{th} shifts, process variations (PV) during fabrication also produces initial V_{th} shifts. The combined shifts will create cell failure statistically depending on whether they cancel or enhance each other. We perform Monte Carlo simulation to evaluate the probability of cell failure after a sustained amount of time (10 years) accounting for both PV and BTI induced V_{th} shifts.

We first use R [5] to generate initial V_{th} shifts for cells in all banks under normal distribution with $\sigma(V_{th}) = 30mV$ due to PV. These initial shifts may cause initial cell failures. Using the access time as the metric of measurement, we found this initial cell failure probability is 0.598%. Some of these defects can be mitigated by redundant rows in each cache banks or other failure protection mechanisms. Those that cannot be fixed will cause yield loss.

Next, we add V_{th} shifts introduced by BTI stress to the initial shifts, and run Monte Carlo simulation to evaluate probability of failure over time. We assume BTI induced V_{th} to be $50mV$ in the worst case after 10 years as we collected in Fig. 5(a). We experimented two extreme signal probabilities: (1) most imbalanced case that stresses only 1 pair of internal gates; and (2) perfectly balanced case that equally stresses both pairs of gates. The results are shown in Fig. 5(b) for (1) and Fig. 5(c) for (2) respectively.

Our proposed *4PR* technique slows down the increase of the cell failure probability, since more V_{th} shift is recovered. The cell failure probability after 10 years under most imbalanced stress reduces from 1.05%/0.93% for Normal/SP PR to 0.87%/0.92% for *4PR*-best and *4PR*-worst. Under perfectly balanced stress condition, the cell failure probability is reduced from 0.94%/0.91% for Normal/SP PR to 0.83%/0.86% for *4PR*-best and *4PR*-worst.

D. MTTF improvement

Last, we calculate the mean-time-to-failure (MTTF) using the V_{th} shifts and the collected signal probabilities from 20 SPEC2006 benchmarks. The lifetime is determined when the 3rd bit in a cache line failed because typically, there are failure protection mechanisms such as ECC for on-chip L2 cache lines. Notice that there are no large MTTF variations among different benchmarks. This is because cells that tend to fail sooner have extremely imbalanced signal probabilities close to either 0 or 1. Hence, those cells do not change very much. For any type of applications, it is intuitive to understand that there are always some cells that do not change that much (e.g. the most significant bits of integers).

On average, "SP PR" extends SRAM cell's lifetime to 2.43 times that of "Normal", with "Power Off" falling slightly behind: 2.0 times better. Our "4PR" consistently outperforms those two techniques and achieves 4.64/2.86 times lifetime extension over "Normal", under *4PR*-best/worst. These are considerable improvements compared with cell failure probabilities and V_{th} shift amount in Fig. 5. This is because V_{th} shift follows power law of time. In the long run, it take longer and longer time to achieve the same amount of V_{th} shift. Hence, a slight recovery of shift can increase the lifetime significantly. This is where a better recovery scheme becomes most effective.

VI. Conclusion

We have shown that PBTI in NMOS transistors has a large impact to SRAM cell's reliability using high-κ metal gate stacking technology. We developed a technique to recover all MOS transistors in a SRAM cell to mitigate both NBTI and PBTI effect. Our technique achieves 3× lifetime extension over the non-protected baseline cell.

References

[1] J. Abella, X. Vera, and A. Gonzalez. Penelope: The NBTI-Aware processor. In *Proc of Micro-40*, pages 85–96, 2007.

[2] A. Bansal, R. Rao, J. Kim, S. Zafar, J. Stathis, and C. Chuang. Impact of NBTI and PBTI in SRAM bit-cells: Relative sensitivities and guidelines for application-specific target stability/performance. In *IRPS*, pages 745–749, 2009.

[3] A. Bansal, R. Rao, J. Kim, S. Zafar, J. H. Stathis, and C. Chuang. Impacts of NBTI and PBTI on SRAM static/dynamic noise margins and cell failure probability. *Microelectronics Reliability*, 49(6):642–649, June 2009.

[4] Y. Cao. Predictive technology model. http://www.eas.asu.edu/~ptm.

[5] R. foundation. The r project for statistical computing. http://www.r-project.org.

[6] X. Fu, T. Li, and J. Fortes. NBTI tolerant microarchitecture design in the presence of process variation. In *Proc. of MICRO-41*, pages 399–410, 2008.

[7] T. Grasser, W. Gos, V. Sverdlov, and B. Kaczer. The universality of nbti relaxation and its implications for modeling and characterization. In *Proc. of 45th IRPS*, pages 268–280, Phoenix, 2007. IEEE.

[8] D. P. Ioannou, S. Mittl, and G. L. Rosa. Positive bias temperature instability effects in nmosfets with hfo2/tin gate stacks. *IEEE TDMR*, 9(2), June 2009.

[9] B. Kaczer, T. Grasser, P. Roussel, J. Martin-Martinez, R. O'Connor, B. O'Sullivan, and G. Groeseneken. Ubiquitous relaxation in bti stressing–new evaluation and insights. In *Proc. of 46th IRPS*, pages 20–27, Phoenix, 2008. IEEE.

[10] K. Kang, S. Gangwal, S. P. Park, and K. Roy. NBTI induced performance degradation in logic and memory circuits: how effectively can we approach a reliability solution? In *ASPDAC*, pages 726–731, Seoul, Korea, 2008.

[11] K. Kang, H. Kufluoglu, K. Roy, and M. A. Alam. Impact of Negative-Bias temperature instability in nanoscale SRAM array: Modeling and analysis. *IEEE TCAD*, 26(10):1770–1781, 2007.

[12] D. Khalil, M. Khellah, N. Kim, Y. Ismail, T. Karnik, and V. De. Accurate estimation of SRAM dynamic stability. *IEEE TVLSI*, 16(12):1639–1647, 2008.

[13] J. Mitard, X. Garros, L. Nguyen, C. Leroux, G. Ghibaudo, F. Martin, and G. Reimbold. Large-Scale time characterization and analysis of PBTI in HFO2/Metal gate stacks. In *44th IRPS*, pages 174–178, 2006.

[14] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS. *IEEE TCAD*, 24(12):1859–1880, 2005.

[15] S. Ramey, C. Prasad, M. Agostinelli, S. Pae, S. Walstra, S. Gupta, and J. Hicks. Frequency and recovery effects in high-k bti degradation. In *Proc. of 47th IRPS*, pages 1023–1027, Montreal, 2009.

[16] J. Shin, V. Zyuban, P. Bose, and T. M. Pinkston. A proactive wearout recovery approach for exploiting microarchitectural redundancy to extend cache SRAM lifetime. In *Proc. of 35th ISCA*, pages 353–362, 2008.

[17] R. Vattikonda, W. Wang, and Y. Cao. Modeling and minimization of PMOS NBTI effect for robust nanometer design. In *Proc. of 43rd DAC*, pages 1047–1052, San Francisco, CA, USA, 2006.

[18] M. Yourst. PTLsim: a cycle accurate full system x86-64 microarchitectural simulator. In *ISPASS*, pages 23–34, 2007.

[19] S. Zafar, Y. Kim, V. Narayanan, C. Cabral, V. Paruchuri, B. Doris, J. Stathis, A. Callegari, and M. Chudzik. A comparative study of NBTI and PBTI (Charge trapping) in SiO2/HfO2 stacks with FUSI, TiN, re gates. In *VLSI Technology, Digest of Technical Papers.*, pages 23–25, 2006.