Reliable and Self-Repairing SRAM in Nano-scale Technologies using Leakage and Delay Monitoring

Saibal Mukhopadhyay, Kunhyuk Kang, Hamid Mahmoodi, and Kaushik Roy

Dept of ECE, Purdue University, West Lafayette, IN-47907, USA
<sm, kang18, mahmoodi, kaushik>@ecn.purdue.edu

Abstract

The inter-die and intra-die variations in process parameters result in large number of failures in an SRAM array degrading the design yield. In this paper, we propose an adaptive repairing technique for SRAM based on leakage and delay monitoring. Leakage and delay monitoring is used to effectively separate dies with different inter-die Vts from each other. Using the leakage (or delay) monitoring and adaptive body bias, we propose a reliable and self-repairing SRAM which has reduced number of parametric failures under high inter-die and intra-die Vt variations. The proposed self-repairing SRAM improves the design yield by 5%-40% in predictive 70nm technology from BPTM.

1. Introduction

Die-to-die and within-die variations in process parameters result in mismatch in the strengths of different transistors in an SRAM cell (Fig. 1), resulting in functional failures (read, write, access and hold failures) [1-2]. The functional failures due to parametric variations (hereafter, referred to as parametric failures) degrades the memory yield (i.e. the number of non-faulty chips) [2]. The principal reason for parametric failures is the intra-die variation in threshold voltage of the cell transistors due to random dopant fluctuations [1-2]. The die-to-die variation in process parameters (say, *Vt*) also has a strong impact on the failure probability of a cell. In particular, low-Vt dies has a higher probability of read and hold failures while high-Vt dies suffer mostly from access and write failures. Thus die-to-die variations significantly increase the yield degradation. Hence, a self-repairing technique in SRAM that reduces the read/hold failures in low-Vt dies and access/write failures in high-Vt dies can considerably improve yield. This can be achieved by using adaptive repairing technique such as application of Adaptive Body Bias (ABB) [3-6]. Application of Reverse-Body-Bias (RBB) in low-Vt dies increases their Vt thereby reducing possible read/hold failures in SRAM cells. Similarly, application of Forward Body Bias in high-Vt dies decreases their Vt, which reduces the access and write failures in the SRAM cells. However, major obstacle in the application of the adaptive repair techniques in memory is the presence of large intra-die variation. Due to high intra-die variation it becomes difficult to distinguish between a die from low Vt (inter-die) process corner and a die from high Vt (inter-die) process corner. Hence, separation of the dies in different inter-die Vt corners (hereafter, referred to as Vt-binning) is very important for the application of adaptive and self-repair techniques. In this paper, we propose a self-repairing SRAM that successfully detects the inter-die Vt corners and apply a proper body-bias to improve yield. In particular,

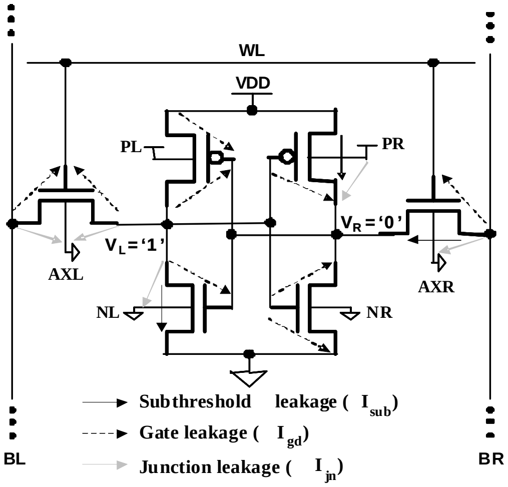


Fig. 1: SRAM Cell storing "0" at node R.

- We show the application of body-bias in reducing memory failures.
- We propose an efficient technique for Vt-binning by monitoring the leakage of a memory array or delay of a ring-oscillator. We show that even under a large intra-die variation monitoring the total memory leakage (or delay of a long inverter-chain) is an effective and reliable technique for separating high-Vt dies from low-Vt ones (i.e. Vt binning).
- Finally, using delay and leakage monitoring we propose a reliable and self-repairing SRAM array.

In the proposed design using on-chip delay and leakage

0-7803-9039-3/$20.00 © 2005 IEEE

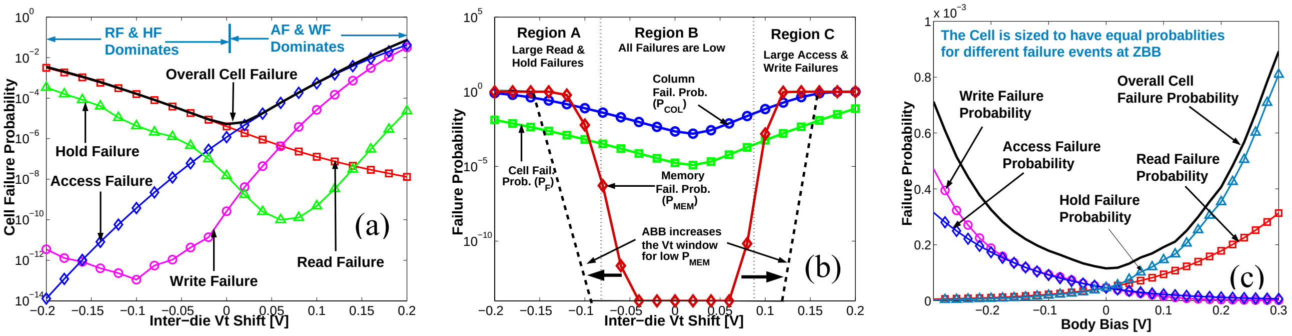


Fig. 2: Effect of inter-die Vt shift and body-bias on the failure probabilities: (a) cell failure probability with inter-die Vt shift, (b) memory failure probability with inter-die Vt shift, and (c) effect of body-bias on cell failure.

monitors, forward or reverse body bias is applied adaptively in an SRAM die depending on its inter-die Vt corner. The proposed design is implemented in BPTM 70nm technology [12] and simulated in HSPICE. Our analysis shows that the self-repair technique in the SRAM improves the yield by 5%-40% depending on the inter-die and intra-die Vt variations.

2. Background: Parametric Failures in SRAM and Effect of Body-Bias

2.1 Parametric Failures in SRAM Cell and Array

The intra-die Vt variation ((Vt) due to random dopant fluctuations (RDF) results in failures in SRAM cell. The Vt shifts of the cell transistors due to RDF, can be considered as independent Gaussian random variables with standard deviation given by [1, 7, 8]:

$$\sigma_{\delta Vt_i} = \frac{qT_{ox}}{\varepsilon_{ox}}\sqrt{\frac{N_{SUB}W_{dm}}{3LW}} \qquad (1)$$

where, T_{ox} is the oxide thickness, W_{dm} is the width of the depletion region, and N_{SUB} is the doping concentration in substrate. The parametric failures in an SRAM cell are principally due to [2, 8]:

Read Failure - Flipping of the SRAM cell data while reading. The read failure can be reduced by increasing the difference between the voltage rise at the node storing "0" while reading (say, V_{READ}) and the trip-point of the inverter (V_{TRIPRD}) associated with the node storing "1".

Write Failure – Unsuccessful write to the SRAM cell. Write failure occurs if the node storing "1" cannot be discharged through the access transistors during the word-line turn on time.

Access Failure –Access failure occurs if the voltage difference between the two-bitlines (bit-differential) at the time of sense amplifier firing reduces below the offset voltage of the sense-amplifier [15]. Access failure occurs due to the reduction of the bit-line discharging current through the access and pull-down NMOS transistors.

Hold Failure - The destruction of the cell data in the standby mode with the application of a lower supply voltage. The hold failure occurs due to high-leakage of the NMOS transistors connected to the node storing "1". At a lower V_{DD}, due to the leakage of the NMOS, the node storing "1" reduces from V_{DD} (which is enhanced by a weak PMOS). If that voltage becomes lower than the trip-point of the inverter storing "0" the cell flips in the standby mode.

If a cell in a column fails (*column failure*), the column is replaced by an available redundant column. If the number of faulty columns is larger than the number of redundant columns, the SRAM array fails (*memory failure*). The column (P_{COL}) and the memory failure prob abilities (P_{MEM}) are estimated as:

$$P_{COL} = 1-(1-P_F)^N;\ P_{MEM} = \sum_{i=N_{RC}+1}^{N_{COL}} \binom{N_{COL}}{i} P_{COL}^{i}\left(1-P_{COL}\right)^{N_{COL}-i} \qquad (2)$$

2.2 Effect of Inter-die Vt shift on Cell Failures

A negative shift in the threshold voltage, due to inter-die variation, (i.e. for the SRAM arrays shifted to the low-Vt process corners) increases the read and the hold failures (Fig. 2a). This is because of the fact that, lowering the Vt of the cell transistors increases V_{READ} and and reduces V_{TRIPRD}, thereby increasing read failures. The negative Vt shift increases the leakage through the transistor N_L, thereby, increasing the hold failures. In case of the SRAM arrays in the high-Vt process corners, the access failures and the write failures are high (Fig. 2a). This is principally due to the reduction in the current drive of the access transistors. The hold failure also increases at the high Vt corners, as the trip-point of the inverter PR-NR increases with positive Vt shift. Hence, the overall cell failure increases both at low and high-Vt corners and is minimum for arrays in the nominal corner (Fig. 2a). Consequently, the probability of memory failure is high at both low-Vt and high-Vt inter-die process corners (Fig.2b).

2.3 Effect of Body-bias on Cell Failures

Let us now discuss the effect of the body-bias (applied only to NMOS) on different types of failures. Application of reverse body-bias increases the Vt of the transistors which reduces V_{READ} and increases V_{TRIPRD}, resulting in a reduction in the read failure (Fig. 2c). The Vt increase due to RBB also reduces the leakage through the NMOS thereby reducing hold failures (Fig. 2c). However, increase in the Vt of the access transistors due to RBB increases the access and the write failures. On the other hand, application of FBB reduces the Vt of the access transistor which reduces both access and write failures. However, it increases the read (V_{READ} increase and V_{TRIPRD} reduces) and hold (leakage through NMOS increases) failures (Fig. 2c).

2.4. Application of Adaptive Body Bias to Enhance Yield

From Fig. 2b, it can be observed that, above a certain Vt-shift (~100mV) small changes in inter-die Vt results in a large memory failure probability (~1) (regions A and C). However, for chips with Vt in the window of -100mV to 100mV (region B) the memory failure probability (P_{MEM}) is negligible (~0). Using the memory failure probability the yield of the memory can be defined as [2, 8]:

$$Yield = 1 - \left(\sum_{INTER=1}^{N_{INTER}} P_{MEM}\left(Vt_{INTER}\right) \Big/ N_{INTER} \right) \qquad (3)$$

where, N_{INTER} is the total number of dies. Let us now assume that, due to inter-die distribution of Vt, the number of dies in region A, B, and C are N_A, N_B and N_C, respectively. Hence, yield can be obtained as:

$$Yield = 1 - \left(\frac{P_A N_A + P_B N_B + P_C N_C}{N_A + N_B + N_C} \right) = \frac{N_B}{N_A + N_B + N_C} \qquad (4)$$

where, P_A(~1), P_B(~0), and P_C(~1) are the memory failure probabilities in the region A, B and C. Hence, to improve yield, N_A and N_C have to be reduced (in other words N_B needs to be increased). This can achieved by applying RBB to the dies in region A thereby reducing their read and hold failure probability. Similarly, application of FBB to the chips in region C reduces their write and access failure probability. This effectively, increases the ΔVt(inter) window for region B resulting in a higher value for N_B and lower values for N_A and N_C. Hence, adaptive application of the body bias (Adpatibe Body Bias, ABB) based on the inter-die process corner of a die, can effectively improve the memory yield.

3. Leakage and Delay Monitoring

The effective identification of the inter-die process corner of a memory die is the key element in the application of ABB. Since the responses a circuit (such, as delay or leakage) in a die depends on the Vt of values of the transistors in that die, such responses can be used to identify the inter-die process corner of the die. However, the random within-die variation in Vt tends to mask the difference in the response of the circuits with different inter-die Vt shift. Hence, to determine the inter-die Vt corner, the effect of intra-die variation needs to be cancelled. In this section, we describe a low-cost method to determine inter-die process corners even under a large random within-die variation. The proposed method is essentially based on the application of the Central Limit Theorem [11]. Using Central Limit Theorem, the distribution of a random variable (say, Y) which is the summation of a large number of independent random variables (say, X_1, ..., X_n) can be assumed to be Normal with mean and the standard deviation given by:

$$\mu_Y = \sum_{i=1}^{n} \mu_{Xi} \text{ and } \sigma_Y^2 = \sum_{i=1}^{n} \sigma_{Xi}^2 \qquad (5)$$

If all the variables are identically distributed (i.e. all with equal mean μ_X and standard deviation σ_X) we further obtain:

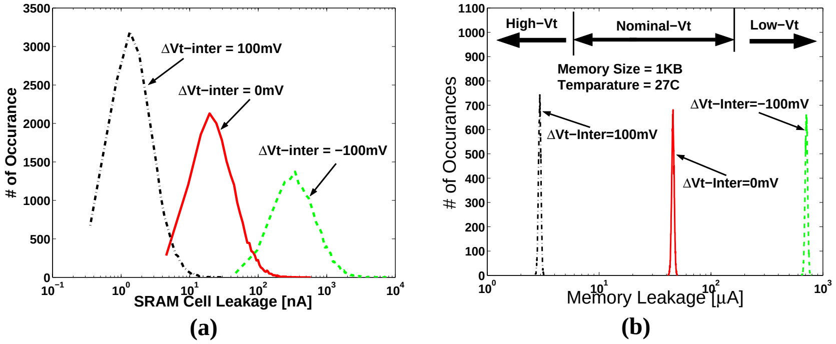
$$\mu_Y = N\mu_X \text{ and } \sigma_Y = \sqrt{N}\sigma_X \Rightarrow \frac{\sigma_Y}{\mu_X} = \frac{1}{\sqrt{N}} \frac{\sigma_X}{\mu_X} \qquad (6)$$

From (6), it can be observed that, the spread (standard deviation/mean) of the variable Y is less than the spread in the variable X and the spread of Y reduces as more number of variables are added together. Using the above theory we develop a method for canceling the effect of intra-die Vt variation on a circuit response and determine the inter-die corner of a die.

3.1. Vt-binning with Leakage Monitoring

The leakage of an SRAM cell is composed of the subthreshold, the gate and the junction tunneling leakage as shown in Fig, 1 [2]. The random intra-die variation in threshold voltage results in significant variation in cell leakage, particularly, the subthreshold leakage (as it exponentially depends on Vt [2], [7]). The leakage of an SRAM array is obtained by adding the leakage of all the cells (say, N_{CELL}) in the array. As explained earlier the major source of intra-die Vt variation in SRAM array is the RDF. Since RDF induced Vt variation is completely random, the leakage of different cells can be considered as independent random variables. Hence, the Central Limit Theorem can be applied to estimate the overall memory leakage.

Due to the exponential dependence of the leakage on Vt, the inter-die Vt shift results in a large change in the leakage of an SRAM cell. However, due to the large leakage spread caused by intra-die Vt variation, the leakage distribution of memory cells from different inter-die Vt corners overlap with each other. To illustrate this we have simulated the leakage of memory cells and SRAM array designed in BPTM 70nm technology node, with different inter-die Vt shift. Random intra-die Vt variation is applied to the different transistors in the cell.



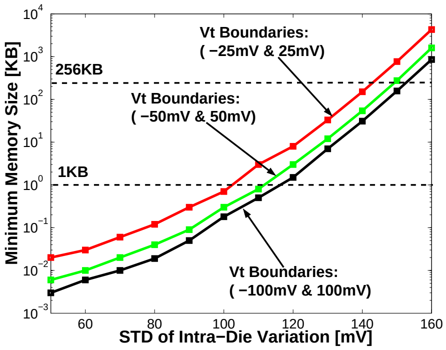


Fig. 3: Effect of random intra-die Vt variation at different inter-die Vt corners of SRAM: (a) leakage distribution (due to intra-die variation) of an SRAM cell, (b) leakage distribution (due to intra-die variation) of the 1KB SRAM array.

Fig. 4: Minimum memory size vs Intra-die distribution.

The intra-die Vt variations applied to different cells of the SRAM array are also independent. Fig. 3a shows that the distributions of leakage (due to within-die Vt variation) of memory cells at low, normal, and high inter-die Vt corners are overlapping, making it impossible to discriminate between the cells from different inter die process corners. However, the leakage distribution (due to within-die Vt variation) of the SRAM array (composed of large number of cells) are well separated as predicted by (6) (Fig. 3b). Hence, by monitoring the leakage of an SRAM array we can determine the inter-die corner of a die. From (6) it can be observed that increasing the memory size increases the separation between the leakage distribution (due to within-die Vt variation) of an SRAM array from low-Vt and high-Vt inter-die corners. Hence, for an intra-die Vt variation, there exists a minimum memory size that is required to effectively discriminate between a memory chip from a high-Vt corner and low-Vt corner of the inter-die distribution. To estimate the minimum size we use the following process:

$$
\begin{aligned}
&Let, Vt_{bnd} = boundary\ of\ the\ different\ Vt\ region\\
&\sigma_{Intra} = std.\ of\ intra\text{-}die\ Vt\ variation(\sigma_{Vt0})\\
&High\text{-}Vt\ bin: \Delta Vt_{iner} > Vt_{bnd}\\
&Low\text{-}Vt\ bin: \Delta Vt_{iner} < -Vt_{bnd}\\
&Nominal\text{-}Vt\ bin: -Vt_{bnd} < \Delta Vt_{iner} < Vt_{bnd}\\
&Let, N_{min} = minimum\ number\ of\ cells\ required\ for\\
&effetive\ binning\ of\ inter-die\ Vt\\
&N_{min}\ is\ given\ by\ the\ minimum\ N_{Cell}\ for\ which:\\
&1:\ P\left[I_{MEM}(\Delta Vt_{iner} = Vt_{bnd}) > I_{MEM}(\Delta Vt_{iner} = 0)\right] < 10^{-12*}\\
&2:\ P(I_{MEM}(\Delta Vt_{iner} = 0) > I_{MEM}(\Delta Vt_{iner} = -Vt_{bnd})) < 10^{-12}\\
&{*10^{-12}}\ represents\ 7\sigma\ point\ of\ a\ Standard\ Normal\ Dist.
\end{aligned}
\qquad (7)
$$

It is observed, that for reasonable values of intra-die variations the effective separation of the low-Vt and high-Vt process corners can be obtained for memory sizes equal to and higher than 1 KB (Fig. 4). Hence, we can conclude that, leakage monitoring can effectively be used to differentiate between memory chips from different inter-die process corners (inter-die Vt-binning).

3.2. Vt-binning with Delay Monitoring

The delay of an inverter reduces with a reduction in the Vt of the NMOS and PMOS transistors. Hence, the delay of an inverter form a low-Vt die will be lower than that of a inverter from a high-Vt die. Thus, monitoring the delay of an inverter can be used to detect the inter-die process corner of an SRAM array. However, the random intra-die variation (due to RDF) in the Vt of the transistors also results in a significant spread in the delay of inverter (can be modeled as a Normal variable). Hence, if the delay of inverter chain of a small length (say, 3) is observed, the inter-die shift in the delay can be masked by the intra-die delay variation. Fig. 5a and 5b shows the, delay distribution due to random within-die Vt variation of a 3-stage and 300-stage inverter chain (designed with minimum size NMOS transistors) with different inter-die Vt shift obtained through Monte-Carlo simulations in SPICE using BPTM 70nm technology. The intra-die Vt variation applied to the different inverters are independent of each other. As observed from Fig. 5a, the delay distributions (due to intra-die variation) for 3-stage inverter chain with different inter-die Vt shift are overlapping. However, the delay distributions of the 300-stage inverter are well separated and provide good discrimination between the dies from different regions (Fig. 3b). The discrimination increases with an increase in the inverter length and there exists a minimum number of stages for which successful inter-die Vt detection is possible (under a certain intra-die variation). We estimated the minimum number of stages for different intra-die variation using (7). It was observed that for reasonable intra-die Vt variations (< 90mV) delay of a 300 stage inverter chain can be used for Vt binning (Fig. 5c). As the intra-die variation increases and/or the Vt boundaries for the different Vt corners become closer to

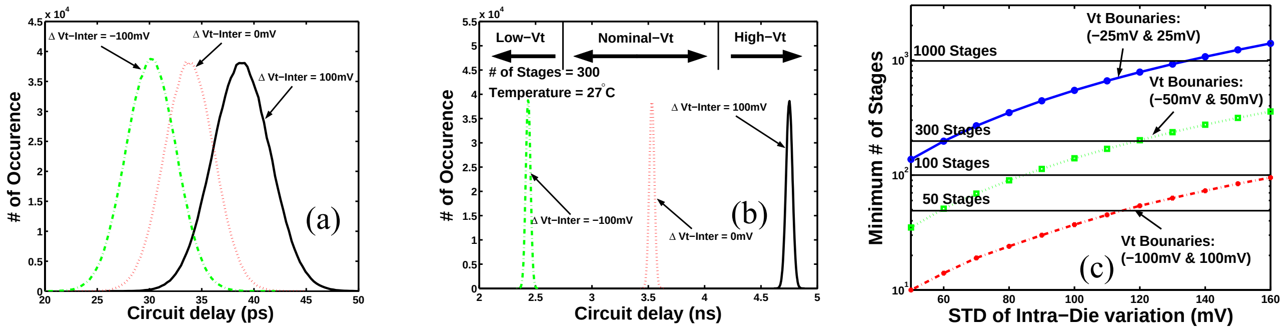


Fig. 5: Vt-binning with monitoring the delay of an inverter-chain: (a) 3-stage inverter-chain, (b) 300-stage inverter chain, and (c) minimum number of stages for effective separation.

each other, the minimum number of stages required for effective separation also increases (Fig. 5c).

In this section, we discussed that monitoring of the leakage of the SRAM array or the delay of a long inverter chain can be efficiently used to detect the inter-die process corners of SRAM dies. In next section we will discuss the design of self-repairing SRAM using ABB and delay/leakage monitor

4. Self-Repairing SRAM using ABB

Self-repair is a useful technique in order to improve the design yield. In this method, an on-chip system detects the inter-die process corner of the chip and accordingly applies adaptive repair technique (in this case, proper body bias) to fix the parametric failures in that process corner. Hence, the first step is on-chip detection of the inter-die process corner. Based on the discussions in the previous sections, the process corner can be estimated by monitoring the standby leakage of the SRAM or monitoring the delay of an on-chip inverter chain. In this section we discuss the two self-repair strategies for SRAM based on the leakage and delay monitors.

4.1. Self-Repairing SRAM using Leakage Monitoring

In a Self-repairing SRAM using "Leakage Monitoring", the leakage (memory leakage) of the SRAM die is monitored using an on-chip leakage monitor. The measured leakage is then compared with the reference currents to identify the inter-die process corner of the chip. Based on this measurement, the right body bias is applied to the chip. The schematic of a self-repairing SRAM array with self-adjustable body-bias generator is shown in Fig. 6. A simple current mirror circuit is used for leakage monitoring (Fig. 7). The transistors in the current mirror are designed to be large to reduce the effect of variations in the monitor circuit. The monitor generates an output voltage (Vout) that is proportional to the leakage of the SRAM array (Fig. 7). The output of the leakage monitor is compared with the reference voltages corresponding to the different inter-die process corners. The reference voltages can be generated using band-gap voltage sources [13]. Based on the results of this comparison, the body bias generator applies the right body bias to the SRAM array (Fig. 8). If an SRAM die is in the low inter-die Vt corner, the output of the leakage monitor (Vout) will be greater than both the reference voltages (V_{REF1} and V_{REF2}) and both comparators generate zero, resulting in application of a reverse body bias (RBB). If the SRAM chip is in the high Vt corner, Vout of the leakage monitor will be less than both V_{REF1} and V_{REF2}. Hence, the outputs of the comparator will be at logic one. This results in application of a forward body bias (FBB). For SRAM dies that are in the nominal Vt corner, the leakage monitor output will be between V_{REF1} and V_{REF2} (V_{REF1}>V_{REF2}). Hence, the output of one the comparators will be at zero and the other one at one. This results in application of a zero body bias (ZBB=0). To avoid any performance loss due to the voltage drop across the leakage monitor (which is in the supply path) the leakage monitor is bypassed in the regular mode of operation. This bypassing is implemented by the PMOS bypass switch. Since the leakage monitor has to be bypassed when the memory is accessed, the outputs of the comparators are sampled to flip-flops (FF) by the "calibrate" signal. The repairing circuit starts operating when the calibrate signal is turned "on".

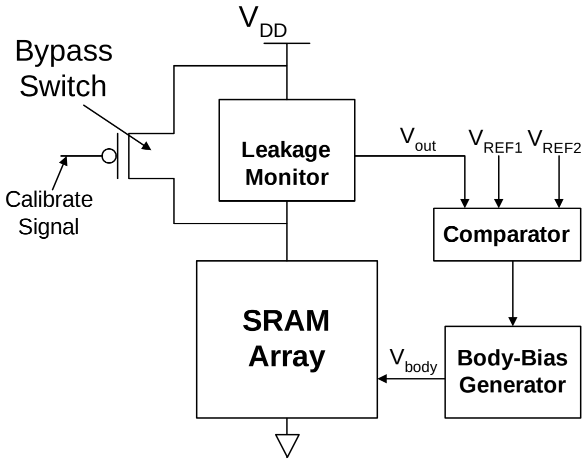


Fig. 6: Block-diagram of self-repairing SRAM using ABB

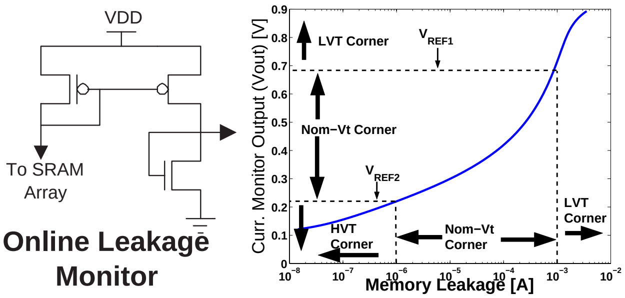


Fig. 7: Current mirror circuit and characteristics

The self-repairing circuit shown in Fig. 6 is implemented in BPTM 70nm technology to evaluate its effectiveness. We have used the comparator from [14]. However, to reduce the offset voltage, a sense-amplifier based comparator with offset compensation technique will improve the design [15]. A PVT tolerant current monitoring circuit can also be used to improve the design [16]. A large number (10000) of Monte-Carlo simulations was performed to generate inter-die Vt shifts in the SRAM array (Fig. 9a). The inter-die distributions of the memory chips results in the inter-die distributions of the memory leakage (Fig. 9b). The variation in the memory leakage results in different Vout voltage in the dies in different inter-die process corners (Fig. 9c). Finally, based on the comparator results in each die, the correct body-bias is generated (Fig. 9d) and the dies get grouped based on the applied body-bias voltages (Fig. 9d). The reference voltage levels are selected based on the pre-calibrated values of the memory leakage at different inter-die process corner. However, it should be noted that, an increase in the intra-die variation increases the mean of the leakage of dies shifted to different inter-die corner (Fig. 10a). Hence, the pre-calibration of the reference voltage has to consider the intra-die Vt variation. Moreover, for a die with a certain inter-die Vt-shift, the leakage spread due to within-die variation results in a spread in the generated Vout voltage for that die (Fig. 10b). In our experiment we observed that, the Vout distribution corresponding to different Vt corners are well separated (Fig. 10b for a 1KB cache with Vt boundaries at ±80mV). The separation increases with an increase in the Vt boundaries and/or the memory size.

Let us now investigate the different sources of error in the application of the proper body bias in the SRAM array. To simplify the analysis let us only consider the errors that can occur for dies shifted to low inter-die Vt corners.

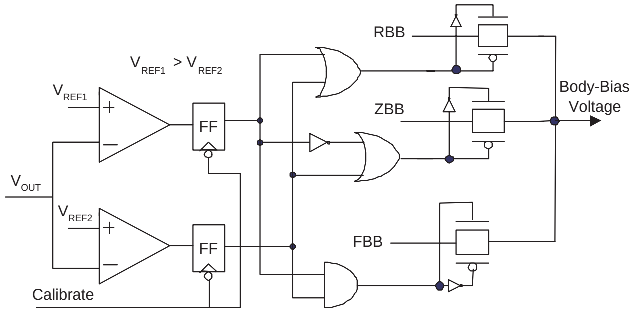


Fig. 8: Body-Bias generation circuit for self-repairing SRAM.

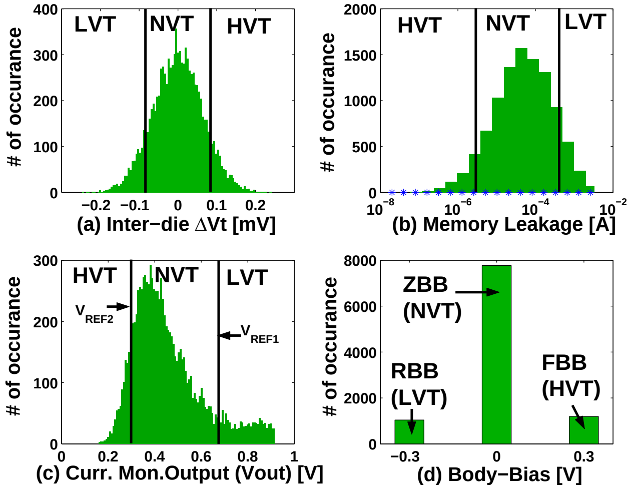


Fig. 9: Operation of the self-repair strategy: (a) Inter-die Vt distribution, (b) inter-die leakage distribution, (c) distribution of Vout, and (d) generation of ABB. (LVT-lowVt, NVT-nominal Vt, HVT- high Vt .

- First, due to intra-die variation the Vout for a low-Vt die with Δ*Vt-inter* = (-Vt_{bnd}-Δ), (where, Δ is a small voltage ~5-10mV) may become lower than V_{REF1} (Fig. 11a) which results in the application of ZBB (instead of RBB). This results in a miss-prediction in the low-Vt corner (P_{MISLVT}) and such a miss-prediction reduces the

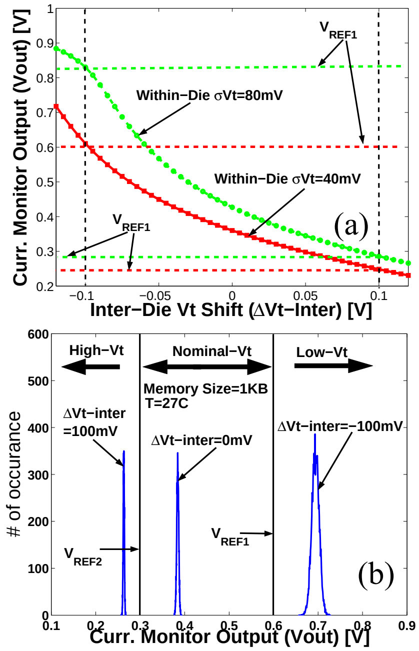


Fig. 10: Effect of intra-die variation on Vout and the selection of reference voltage: (a) selection of reference voltage with intra-die variation, and (b) variation of Vout due to intra-die variation at different inter-die process corner.

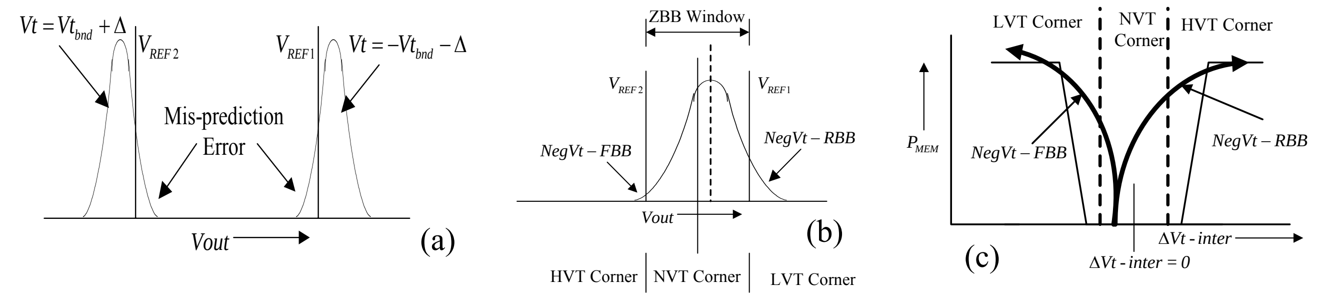


Fig. 11: Errors due to intra-die variations in Vout: (a) Mis-prediction error, (b) application of incorrect body-bias resulting in new failures- NEGVT-FBB and NEGVT-RBB case, (c) increase in the number of faulty dies due to NEGVT-RBB and NEGVT-FBB

yield improvement using ABB. Similar, miss-prediction can occur for high-Vt dies with $\Delta Vt\text{-}inter = (Vt_{bnd} + \Delta)$. The probability of miss-prediction can be formally defined as:

$$\begin{aligned} P_{MISLVT} &= P\left\{V_{out}\left(-Vt_{bnd} - \Delta\right) < V_{REF1}\right\} \\ P_{MISHVT} &= P\left\{V_{out}\left(Vt_{bnd} + \Delta\right) > V_{REF2}\right\} \end{aligned} \qquad (8)$$

The miss-prediction error reduces with reduction in the intra-die Vt variation or increase in the memory size, which reduce the Vout spread (Table-I). Our simulation shows that the error is less than 7% for 1KB memories and large within-die variation (Table-I). The error becomes negligible for memories larger than 2KB. However, it should be noted that this type of mis-prediction does not increase the number of faulty dies compared to the no-body-bias case. The mis-prediction error can be reduced by designing the Vt boundaries little inside the ZBB window (instead of exactly at the region boundaries) as shown in Fig. 17 in section 5.

- Let us consider a die in ZBB window (i.e. non-faulty one) and ΔVt-inter < 0. Due to presence of the intra-die variation in Vout and the "offset" voltage of the comparator (V_{off}), this die may be either detected as a HVT die or LVT die (Fig. 11b). If it is detected as a HVT die, FBB will be applied which will result in further reduction of its Vt, making the die a faulty one (Fig. 11c). We will refer to this event as *NegVt-FBB* and the probability of this error ($Y_{\text{NegVt-FBB}}$) is defined as:

$$P_{NegVt-FBB}(-\Delta Vt) = P\left(Vout(-\Delta Vt) < V_{REF2} + V_{off}\right)$$
$$\times P_{MEM}\left(-\Delta Vt, V_b = FBB\right)$$

$P_{MEM}\left(-\Delta Vt, V_b = FBB\right) =$ Memory Failure Probability for a die with inter-die Vt= $-\Delta Vt$ and FBB

$$Y_{NegVt-FBB} = \sum_{\Delta Vt=0}^{Vt_{bnd}} P_{NegVt-FBB}(-\Delta Vt) \Big/ \#\text{ of dies within 0 and } \text{-}Vt_{bnd}$$

Table-I: Miss-prediction Error (%) (with Δ=5mV)

Miss-prediction Type	Vt Boundary	STD. of Intra-Die Variation		
		60mV	70mV	80mV
P_{MISLVT}	25mV	0.006%	0.62%	7.%
	50mV	0.014%	0.61%	7.5%
	100mV	0.002%	0.72%	7.98%
P_{MISHVT}	25mV	0.008%	0.52%	6.46%
	50mV	0.008%	0.51%	6.02%
	100mV	0	0.42%	5.7%

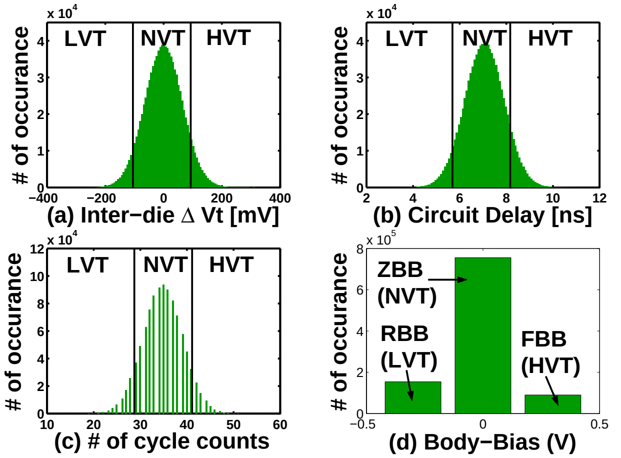
On the other hand, if it is detected as a LVT die, a RBB will be applied. Application of RBB to a die in the ZBB region may shift it to the HVT corner resulting in a faulty die. We will refer to this event as *NegVt-RBB* and the probability of this error ($Y_{\text{NEGVT-RBB}}$) is defined as:

$$P_{NegVt-RBB}(-\Delta Vt) = P\left(Vout(-\Delta Vt) > V_{REF1} - V_{off}\right)$$
$$\times P_{MEM}\left(-\Delta Vt, V_b = RBB\right)$$
$$Y_{NegVt-RBB} = \sum_{\Delta Vt=0}^{Vt_{bnd}} P_{NegVt-RBB}(-\Delta Vt) \Big/ \#\text{ of dies within 0 and } \text{-}Vt_{bnd}$$

Similar arguments hold for the dies in the ZBB region with ΔVt-inter > 0. For those dies, the Vout variation and comparator offset may result in the application of either RBB (resulting in a higher Vt, referred to as *PosVt-RBB*) or FBB (resulting in a lower Vt, referred to as *PosVt-FBB*). The probability of occurrence of *NegVt-FBB* and *NegVt-RBB* increases with an increase in the intra-die variation and comparator offset voltage, and reduction in the separation between the Vt boundaries. However, our simulation result shows that, even with a comparator offset of 40mV and standard deviation of intra-die variation of 100mV this error is negligible (<0.01%). The low error due to *NegVt-FB* is due to the fact that $P\left(Vout(-\Delta Vt) < V_{REF2} + V_{off}\right)$ is negligible. This probability increases as ΔVt approached 0. However, as ΔVt→0, the $P_{MEM}\left(-\Delta Vt, V_b = FBB\right)$ reduces since the die shifts further away from low-Vt corner. Hence, the product of the above two probabilities remain negligible in the entire negative ΔVt region of the ZBB window. Due to similar reason, errors due to *NegVt-RBB, PosVt-RBB* and *PosVt-FB* are also negligible. The probabilities of *NegVt-FBB* and *NegVt-RBB* (or, *PosVt-RBB* and *PosVt-FBB*) can be further reduced using a multi-cycle repair strategy.

4.2. Self-Repairing SRAM using Delay Monitoring

The self-repair using leakage monitoring is an effective and low-overhead method. However, the detection of the inter-die Vt corner with leakage is difficult in gate-leakage dominant technologies, (particularly, at room temperature) since, gate leakage is a weak function of Vt. This problem is avoided in the delay monitor based self-repair circuit shown in Fig. 12. In this technique, we first design a long inverter chain (600 stages). Since the delay of a 600-stage ring-oscillator is significantly higher than the clock-period, we can use the clock and a counter-based detection technique. The counter is first initialized to zero state and at the rising edge of the "calibrate" signal (which needs to be synchronized with the rising edge of the clock) counting begins. The counter is disabled at the rising edge of the signal from the output of the final inverter. The total delay of the path is determined by the state of the counter. Finally, the final state of the counter is compared with pre-calibrated state values representing low-Vt and high-Vt corners (similar to VREF1 and VREF2 in case of the leakage monitor circuit). Depending on the comparison result, a proper body-bias is applied. Since the delay is measured in the quanta of the clock cycle, the number of stages should large enough to minimize the quantization error. In our design we used a 5GHz clock and 600 stages (~64 clock periods) which require a 6-bit counter. The quantization error can be further minimized using larger inverter chain and using dual-edge triggered counters [17]. Moreover, as the "comparator" in this case is a set of digital logic gates, there is no error due to the offset of the comparator. As delay is a weaker function of temperature (compared to leakage), the temperature sensitivity of the monitor is also lower. The major drawback of the delay monitoring technique is the higher area overhead associated with the repair circuit (ring-oscillator, counter and the comparator logic). However, it should be noted that the area of the repair circuit is independent of the memory size. In our case, the 600 stage ring-oscillator has area similar to ~250 SRAM cells and can be insignificant for large memories. This is in contrary to the leakage monitoring case, where, size of the current mirror increases (proportionally) with the memory size as the leakage current drawn by the SRAM array increases with the size of the array. Hence, the delay monitoring technique can be very effective for large SRAM array.



(a) Inter−die Δ Vt [mV]

(b) Circuit Delay [ns]

(c) # of cycle counts

(d) Body−Bias (V)

Fig. 13: Operation of the self-repair strategy: (a) Inter-die Vt distribution, (b) inter-die delay distribution, (c) distribution of output state, and (d) generation of ABB. (LVT-lowVt, NVT-nominal Vt, HVT- high Vt .

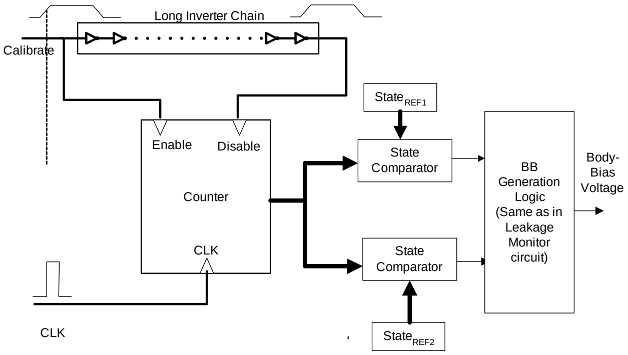


Fig. 12: Delay monitor based self-repair circuit.

The proposed self-repairing SRAM with delay monitor is implemented and simulated in BPTM 70nm technology. The delay of the inverter chain is counted using the counter, which converts the delay distribution to a distribution of the final state of the counter (Fig. 13). The final counter state is compared (digital comparator, can de designed using decoding logic) to the reference states and proper body-bias is generated based on the comparison output.

The major advantage of the delay monitor based design is

<table>
<tr><th colspan="2">Table-II: Comparison of Leakage and Delay Monitors</th></tr>
<tr><th>Leakage Monitoring</th><th>Delay Monitoring</th></tr>
<tr><td>+ Simple circuit
+ Low-overhead
- Ineffective with high gate leakage.
- larger intra-die spread in Vout. Reference voltage selection depends on the intra-die variation.
- Offset of the comparator
- Area overhead depends on the memory size.</td><td>+Digital comparator, and hence no "offset" issue
+Negligible variation in output state-virtually no mis-prediction error.
+Low process and temperature sensitivity.
+ Area overhead is independent of memory size
- Complex circuitry
- Large area overhead for small array size</td></tr>
<tr><td colspan="2"><b><i>Leakage monitor is good for small SRAM in subthreshold leakage dominant technologies.</i></b>
<b><i>Delay Monitor is good for large SRAM and unavoidable in gate leakage dominant technologies.</i></b></td></tr>
</table>

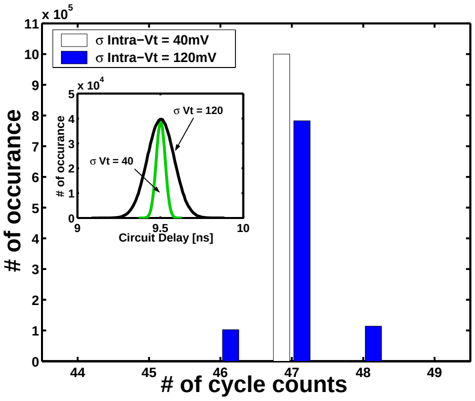


Fig. 14: Effect of intra-die distribution on the reference state generation.

the elimination of the analog components (such as current mirror, comparators). Due to the use of digital comparison, the issue of comparator offset in leakage monitors is resolved. Moreover, simulation result shows that intra-die variation has a very weak impact on the distribution of the output state (at any inter-die Vt corner) since the mean of the delay distribution is weakly sensitive to intra-die Vt variation (Fig. 14, which shows the distribution of the output state considering 100000 Monte-Carlo simulations). This has two major significances: first, the reference state generation can be done without considering intra-die variation. Second, due to the very low spread in the generated output state, mis-prediction errors are negligible (Fig. 14). Table-II summarizes the positive and negative aspects of the delay and leakage monitor based self-repair SRAM. We would like to mention that, the proposed ABB based yield enhancement technique can also be implemented using off-chip selection and application (using programmable fuses) of body-bias voltages.

5. Results and Discussion

The proposed self-repairing SRAM is designed using BPTM 70nm technology and simulated in SPICE to evaluate its effectiveness in improving yield. The size of the transistors in the SRAM cell is first optimized to minimize the cell failure probability at ΔVt-inter =0mV and FBB and RBB of ±0.3V. The self-repairing technique is applied on a 64KB and 256 KB SRAM array. It is observed that without self-repair the designed SRAM array has large number of failures in low-Vt and high-Vt inter-die corners resulting in a low yield, particularly, for large inter-die variations (Fig. 15, 16, 17). The proposed circuit successfully applies the proper body bias depending on the inter-die Vt corner of the circuit. A large reduction in number of failures is observed in both low and high inter-die Vt corners (Fig. 15). The application of ABB widens window of the low-memory failure probability (i.e. region B in Fig. 2b) as shown in Fig. 16. The application of the self-repair technique results in 8%-40% improvement in yield over the SRAM array designed using the optimized cell (Fig. 17). The effectiveness of the technique improves with an increase in the intra-die and inter-die variations.

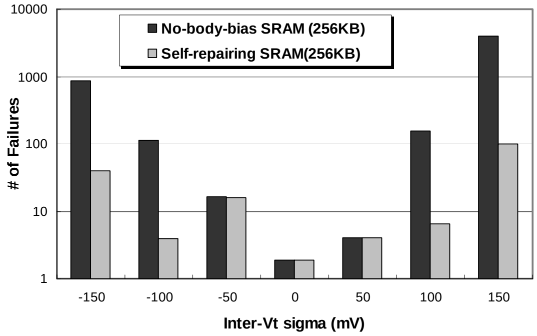


Fig. 15: Reduction in number of failures in 256KB memory array using self-repairing SRAM.

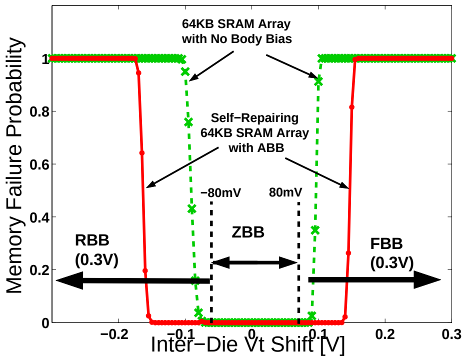


Fig. 16: Effect of ABB on memory failure probability.

One of the major design parameter in the proposed design is the applied FBB and RBB voltage levels. The voltage levels are pre-calibrated based on their impact on yield, leakage and Vt. With an increase in the FBB and RBB levels, the yield improvement increases (Fig. 18a). However, application of too high FBB (and RBB) may degraded the yield by shifting the dies from the boundaries of nominal Vt window to low-Vt or high-Vt corners (i.e. from region B to A or C), particularly, if mis-prediction occurs. A maximum applicable FBB is also bounded by the maximum allowable subthreshold leakage

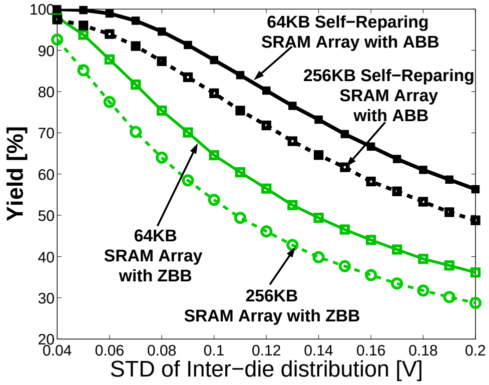


Fig. 17: Yield Enhancement using Self-Repairing SRAM

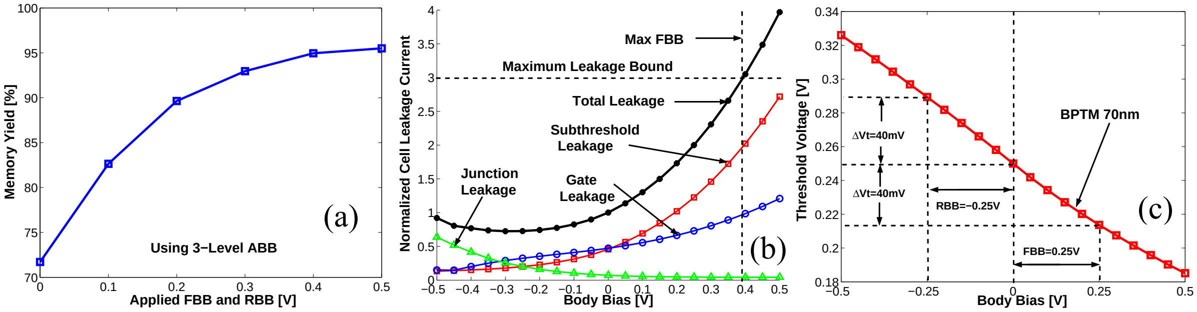


Fig. 18: Impact of adaptive body bias on (a) memory yield, (b) cell leakage, and (c) threshold voltage.

(Fig. 18b). The leakage bound on RBB comes from the increasing junction leakage with RBB. However, for 70nm node the junction leakage is not very significant. Finally, the optimal value of the applied body bias is also decided by the amount of Vt change required to shift dies from region A (or C) to B. Based on the required Vt shift and the pre-calibrated Vt versus body-bias curve, the body-bias voltage levels is decided. For example, in the 70nm devices, 250mV body bias results in approximately 40mV shift in Vt (Fig. 18c). This shift is sufficient to move memory chips from region A (or C) to B in Fig. 16.

In the proposed approach ABB is used to compensate for the global variation in Vt and not the local one (i.e. due to RDF). However, as shown in Fig. 2, the global variation directly impacts the failure probability due to local random variations. For example, an equal amount of local Vt variation (due to RDF) results in a larger number of read/hold failures in a low-Vt die compared to a high-Vt die (Fig. 2). Application of RBB in low-Vt dies increases Vt of all the transistors, which reduces the number of cells that were failing in read/hold mode in that die. Hence, by compensating for the global variation, the proposed approach reduces the *impact of local variability* on cell failures.

6. Conclusions

In this paper we propose a self-repairing SRAM using on-chip current and delay monitoring technique. In the proposed SRAM array, array-leakage and/or delay of an inverter chain are monitored and used to separate different SRAM dies in appropriate inter-die Vt corners. Adaptive body bias is applied to the dies in the different Vt corners resulting in a significant reduction in failures. The onchip monitoring of leakage and delay is observed to be efficient in identifying the inter-die Vt corners of the SRAM dies. While leakage monitoring has a lower cost and is simpler to implement, the delay monitoring is observed to be more robust and scalable. It is observed that for SRAM array of small size and in subthreshold leakage dominant technologies, leakage monitoring is effective. However, for large SRAM array and in scaled technologies (where gate leakage is high), delay monitoring is more effective. Since parametric failures in SRAM's are becoming an increasing problem, the proposed self-repairing SRAM can be very effective in achieving high yield in nano-meter technologies.

7. Acknowledgements

This work was supported in part by Gigascale System Research Center, Semiconductor Research Corporation (under contract #1078.001), Intel and IBM PhD Fellowship Program.

8. References

[1] A. Bhavnagarwala, et. al., "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," IEEE J. Solid State Circuits, vol. 36, no. 4, pp. 658-665, April 2001.

[2] S. Mukhopadhyay, et. al, "Statistical design and optimization of SRAM coll for yield enhancement," Int. Conference on Computer Aided Design, 2004, pp. 10-13, Nov. 2004.

[3] J.W. Tschanz, et. al, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," IEEE JSSC., vol. 37, no. 11, Nov, 2002, pp. 1396-1402.

[4] J.T. Kao, et. al, "A 175-MV multiply-accumulate unit using an adaptive supply voltage and body bias architecture," IEEE J. Solid State Circuits, vol. 37, no. 11, Nov, 2002, pp. 1545-1554.

[5] C. H. Kim et. al., "On-die CMOS leakage current sensor for measuring process variation in sub-90nm generations," Symposium on VLSI Circuits, June 2004, pp. 250 – 251.

[6] K. Itoh, VLSI Memory Chip, Springer, 2001

[7] Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, New York: Cambridge Univ. Press, 1998.

[8] S. Mukhopadhyay, et. al, "Modeling and estimation of failure probability due to parameter variations in nano-scale SRAMs for yield enhancement," Symp. on VLSI Circuits, 2004, pp. 64 – 67, June 2004

[9] R. Rao, et. al, "Parametric yield estimation considering leakage variability," DAC, 2004, pp. 442 - 447 June, 2004.

[10] N. C. Beaulieu, et. al, "Estimating the distribution of a sum of independent lognormal random variables," IEEE Trans. on Comm., Vol. 43, No. 12, pp. 2869-2873, Dec. 1995.

[11] A. Papoulis, Probability, Random Variables and Stochastic Process

[12] BPTM 70nm: Berkley Predictive Technology Model,

[13] Banba et.al, "A CMOS bandgap reference circuit with sub-I-V operation," IEEE JSSC, pp. 670-674, 1999.

[14] Source: http://ecircuitcenter.com/OpModels/V_Limit/

[15] B. Wicht, et. al, "Yield and speed optimization of a latch type voltage sense amplifier", IEEE Journal of Solid-State Circuits, vol. 39, July, 2004, pp. 1148-1158.

[16] C.H. Kim, et.al, "On-die CMOS leakage current sensor for measuring process variation in sub-90nm generations," Symposium on VLSI Circuits, 2004, pp. 250 – 251

[17] H.Mahmoodi, et. al, "Dual-edge triggered level converting flip-flops", ISCAS 2004. Proceedings of the 2004 International pp. II - 661-4,Vol.2.