Design and Development of BIST Architecture for Characterization of S-RAM Stability

M. Krishna Chaitanya and V. Ravi *
SENSE Department, VIT - Chennai, Vandalur to Kelambakkam Road, Chennai - 600127, Tamil Nadu, India; mkrishna.chaitanya2014@vit.ac.in, ravi.v@vit.ac.in

Abstract

Objectives: The objective is to find the optimum SNM of SRAM and then a BIST architecture is designed and implemented to test the SRAM cells varying the voltage of the bit lines. Methods/Analysis: In this paper we use a detection technique which is digitally programmable to detect the defective SRAM cells by using some additional set of SRAM cells to change the bitline voltage and then apply stress to the CUT. Findings: By using this programmable detection technique, cells can be tested even after the fabrication and accordingly one can find out the bit line voltages at which even weak and bad cells can be made useful. Improvement: The main advantage of programmability is that we can maintain considerable tradeoff between test yield and quality. The results at the end will justify the effectiveness of the BIST architecture

Keyword: Built in Self Test (BIST), SRAM, Programmable Detection Technique, Static Noise Margin (SNM), Stability.

1. Introduction

Now a days, the need for robust memory architectures is very high in embedded systems. As SRAM cells offer more robustness compared to DRAM cells, we use the former more in modern real life SOC applications1–4. But as the packing density is inversely proportional to the yield as SRAM cells have high packing density the yield becomes less with respect to noise margins. Issues like nano scaled technologies, non catastrophic defects5,6 dependent on environmental parameters can cause instability in SRAM cell. The stability of an SRAM cell is also dependent upon the neighboring blocks when used in embedded applications. Because of all these it’s very important to keep track of the stability of SRAM.

In nano technologies it’s a difficult task to design good robust SRAM chips even in suitable conditions. It takes lots of time to detect faults like stability fault and data retention fault. Traditional test which was used to detect such faults is the pause test, which required high temperatures, moreover it has less fault coverage and not proven effective for most of the faults. If the faults remain undetected in the production test phase, it leads to a huge loss once the chip has been manufactured. The manufacturers can also face long term reliability problems because of manufacturing defects.

The SRAM cells which poses these kind of faults are termed as weak cells in this paper. In order to check the weak cells the stress should be applied in such a way that only the weak cells below the normal threshold level gets flipped and the good cells doesn’t. Therefore many techniques have been proposed earlier to overcome all these problems. But none of them have the ability to program the fail/pass threshold of the SRAM cell which is introduced in this paper.

Cell stability tests are of two types7 as shown in Figure 1, 1) Passive and 2) Active. Passive techniques doesn’t require additional circuitry to apply the stress to the cell. Whereas it’s the opposite in active type of tests where special additional circuit is used to create required stress to the CUT. Again the active test technique is divided into two sub types, 1) Single stress test technique and 2) Multiple or programmable test stress technique. The advantage of programmable test technique is that we can choose or program the test stress which is to be applied to the cell.

*Author for correspondence
The rest of the paper is structured as follows. Section 2 discusses the 6T SRAM schematic and its working. Section 3 presents the definition of SNM and measuring it in different modes of operation of a 6T SRAM. Section 4 explains the weak cell concept and measurement of SNM of a weak cell. Section 5 describes the previous work and the concept of programmable weak cell detection threshold used in the proposed DFT. Section 6 explains the implementation of the weak cell detection technique. Obtained results are presented in Section 7.

2. 6T SRAM Cell

As shown in Figure 2, the SRAM considered in this paper has a total of 6 transistors. Among the 6 transistors, 2 of them are known as access transistors (AT1 and AT2) which helps to access the SRAM cell either to write into or to read from the cell. It also has two cross couple inverters which makes a total of four transistors. Two acts as pull up (U1 and U2) and the other two acts as pull down (D1 and D2). It has an input signal WL called as word line. Whenever the word line is logic high then the data can be written onto or read from the SRAM cell. It also has two bit lines labelled as BL and BLB. The data carried by BL is the compliment of what is carried by the bit line BLB.

2.1 Write Operation

For doing the write operation, we need to set WL as logic one and then we give the input data to the bitlines BL and BLB. This data gets stored at the outputs of each of the cross coupled inverters. The term writability or write margin is defined as the minimum voltage of the bitline needed in order to flip the value stored in the SRAM cell. If the write margin is below the ground voltage which is zero the cell is not writable. Figure 3 shows the write cycle.

2.2 Read operation

For doing the read operation we use extra circuitry like the precharge circuit which pre charges both the bit lines to $V_{dd}$ the read cycle and the sense amplifier connected to the bit lines. After the bit lines are precharged, the word line WL is set to logic high. Now the bit line voltage of BL remains at $V_{dd}$ as logic one is stored but the voltage of BLB gets discharged through the transistors AT2 and D2 connected to ground in series. Now the sense amplifier is enabled, it senses the small difference in voltage developed between the bit lines (BL, BLB) and amplifies to output to full swing. Figure 4 shows read cycle.

3. Static Noise Margin

SNM or Static Noise Margin is defined as the maximum amount of DC noise the SRAM cell can withstand without flipping its data. It’s generally calculated in three different stages SRAM cell operation hold or retention mode, read mode and write mode. SNM is obtained graphically from the butterfly curve. To obtain the butterfly curve, we isolate the cross coupled inverters, then we perform DC
analysis at the output by varying the input voltage from 0 to 1.8v\textsuperscript{15-17}. After which we toggle the axis of the obtained graph and merge both of them to obtain the required butterfly curve. The maximum square that can fit in either of the halves of the butterfly curve is the static noise margin. SNM is the measure of writeability and readability of SRAM cell. While calculating the SNM for writeability if both the curves are meeting at a point the cell is said to possess unistability. In order to obtain good SNM and less leakage current the transistors of SRAM cell are sized as shown in Table 1.

- **SNM in idle mode.**

SNM is generally high in idle mode. As it can be seen in Figure 5 butterfly curve the area is more in between the curves.

- **SNM in read mode.**

In read mode the SNM gets reduced compared to idle mode as in read mode the access transistors comes into picture. Read SNM can be seen in Figure 6.

- **SNM in write mode.**

A cell is said to be stable in write mode if both the curves intersect at a single point which is termed as unistability. SNM of SRAM in write mode can be seen in Figure 7.

![Figure 5. SNM of SRAM in idle mode.](image1)

![Figure 6. SNM of SRAM in read mode.](image2)

![Figure 7. SNM of SRAM in write mode.](image3)

Table 1. Transistor sizes for SRAM

<table>
<thead>
<tr>
<th></th>
<th>Width(um)</th>
<th>Length(nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pull up</td>
<td>4</td>
<td>180</td>
</tr>
<tr>
<td>Pull down</td>
<td>8</td>
<td>180</td>
</tr>
<tr>
<td>Pass gate</td>
<td>6</td>
<td>180</td>
</tr>
</tbody>
</table>

4. **SRAM Fault Model**

A fault in SRAM cell can happen because of stability fault or data retention fault where the former is the subset of
the latter. Stability fault may occur because of numerous reasons like defects in resistances, misalignment of mask, mismatch of transistors. All these leads to unstability of the cell as a result of which the stored value flips and leads to errors. A fault model is something which is helpful to represent the fault while simulating the circuit. Since the presence of fault degrades the SNM of the SRAM cell, the fault model used will degrade the SNM by introducing transistor mismatches and also resistive defects. The fault model is created by inserting a resistance between the inverters as shown in the Figure 8. As we vary the value of the resistance the SNM value changes accordingly and so is the weakness of the SRAM cell. For resistances larger than 200K ohm SNM is not affected. For resistances below 100K ohm till 15K ohm SNM reduces linearly and for values below 15K ohm SNM becomes zero and results in catastrophic failure. This method gives equal SNM in both halves of the butterfly curve, which means the response is symmetrical.

SNM is now calculated by introducing the above discussed fault in to the SRAM cell. As we can see from the Figure 9 (a,b,c) SNM is degraded for the resistance of 5K hence acts as the required weak cell.

In general the VTC of the normal SRAM cell is not symmetric because of the mismatches. Therefore we can say that the asymmetry is proportional to the SNM of the cell. The point at which both the curves intersect is called the metastability point, at which the value of the cell flips. Its different for weak cell compared to good one. Because of the asymmetric property the metastable point is not equidistant from the node potentials. Therefore the node which is closer to the metastability point is more prone to disturbance than the other. VM normal and VM weak are the node voltages, M normal, M weak are the metastable points of good and weak cells respectively. If we consider V test as the test voltage and by some manipulation if the voltage of the node B of the SRAM is reduced below the test voltage then the weak cell will flip its state while the good cell doesn’t. If we choose V test greater than VM weak then weak cells will pass through the tests successfully which is not preferred. And if V test is less than VM normal then even normal cells will flip. Therefore we need to select V test range as VM weak is less than V test less than VM normal.

In another method of creating a weak cell, we use independent power supply or the cell. As we reduce the supply voltage of the SRAM its SNM also gets reduced. For example to reduce the SNM of the cell to 50% the power supply has to be reduced to around 1v. In the following BIST architecture we can use any of the above two methods. In this paper the second method have been used. The way in which SNM is affected because of reduced supply can be seen in the Figure 10. The voltage have been swept from 1.8V to 1.2V and snm has been calculated for the values of 1.8V ,1.5V and 1.2V.

5. Weal Cell Detection Techniques

As discussed earlier there are two types of methods to detect the weak cell 1) Single or fixed threshold method and 2) Programmable threshold method. The disadvantage with single threshold method is that it can result in either over or under coverage. Its mostly dependent upon the pre silicon simulation data. Another disadvantage is
that in this method the yield loss is more. An example of single threshold is weak write test mode, in which a weak overwrite stress is applied. The main principle is, in the write operation the access transistors of the SRAM are underdriven by using relatively lesser word line voltage which in turn used to create weak write stress. In single threshold technique $V_{test}$ is fixed to a single value and it allows only single fail or pass threshold. In multiple threshold technique we have flexibility to vary the $V_{test}$ value of the weak cell and change the pass or fail threshold according to the requirements. This process is called digitally programmable technique.

In digitally programmable technique\textsuperscript{19–21} we use a column of SRAM cells connected to a word line decoder which has been modified in a way to simultaneously enable more than one word lines. This technique starts with writing a1/0 ratio to the group of cells in the column followed by the normal precharge of the bit lines BL and BLB for certain amount of time after which they are disabled again which leaves the bit lines floating. Then a short pulse is sent to all the word lines of the n cells which are used to form the ratio R. Now depending on the chosen ratio R, bit lines gets discharged till certain voltage. Now we enable the word line WL of the CUT which is the weak cell. As the bit lines are already discharged to some extent because of the earlier step, these partially discharged bit lines helps to check the stability of the cell. If the value gets flipped then its termed as the weak cell. Now to check all the possibilities one an try for different ratios. Detection of the weak cell is done by deciding the minimum SNM that can be accepted and the rest of the cells below that pre determined value gets flipped once they are connected to these floating bit lines. The voltage of the floating bit lines depends mainly on the ratio R and the width of the pulse used to enable certain number of SRAM cells to get the ratio R. For certain ratios the value of the cell doesn’t flip or non-destructive write happens which means that the cell is digitally programmed to withstand the stress for that ratio.

In order to prevent the good cells from flipping, i.e; to stop $V_{test}$ from dropping below the metastable point $V_{M_{normal}}$ we can either reduce the ratio R or the width of the pulse used to enable set of cells can be shortened. By controlling the extent to which the floating bit lines discharge the fail/pass threshold can be varied and thus programme the threshold to detect the weak cell.

To increase the effectiveness of the proposed architecture one can increase the number of cells forming the ratio R. The accuracy can also be increased by using larger capacitance for the bit lines BL and BLB. As the number of bits in a column increases the capacitance pertaining to that particular column increases too, so this technique is much more helpful for bit lines with higher values of capacitance.

In order to implement the whole BIST architecture we have used a modified word line decoder which has an enable_all switch in between the SRAM cells and the decoder, SRAM cells and one of which is a CUT. In order to implement enable_all switch we have used a 2:1 mux whose output is connected to the wordlines of the SRAM cells used to form the ratio R. One of the inputs

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure10}
\caption{Variation of SNM with respect to supply voltage.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure11}
\caption{Schematic of whole BIST architecture setup}
\end{figure}
of all the multiplexers is connected to a common signal enable_all, while the other input is connected to the output of decoder. The select line of all these multiplexers is connected to a common single input as we need to enable all of them at the same time with a common input. The whole architecture is shown in the Figure 11. Here we are using other SRAM cells to provide the stress.

In the Figure 12 the enable_all signal ‘net102’ is of 10ns. As we can see there is an error in the signal ‘b’ at around 30ns. Now we re-programme the threshold level with changing the time period of the enable_all signal to 500ps. In the Figure 13 no error can be seen while reading the output of the CUT content.

6. Conclusion

In this paper a digitally programmed weak cell detection technique is proposed with variable threshold level to have maximum coverage and more yield. It is based upon varying the bit line voltage with the help of other SRAM cells by writing 1’s or 0’s into them by enabling the word line at the same time with the enable_all signal. Another way to alter the bit line voltage is by changing the time period of the enable_all signal which makes the bit lines partially discharge, after which the CUT is connected to these partially discharged bit lines which applies the required stress to the CUT.

7. References