Architecture for Isolating Defective Two Port SRAM Memories

A. Padma Sravani and M. Satyam

***Abstract*—This paper presents a method of implementing fault isolation for word-oriented SRAM memories. It uses Built in Self Test (BIST) technique to locate the faults. Isolation circuit is used to isolate the faults by using switches. The detection and isolation of faulty rows is done at power on. Based on defect injection in SPICE simulation, faults are generated and tested.**

***Index Terms*—Fault isolation, BIST, process variations, Two-port SRAM memory.**

I. Introduction

The density and complexity of the memory is high and is increasing rapidly as technology scales down. Due to this scaling, process variations increase significantly causing parametric variations in transistor feature sizes and threshold voltages due to random dopant fluctuations, line edge roughness, sub-wavelength lithography[1] [2]. The threshold voltage mismatch between the neighboring transistors in a cell, results in the failure of a cell. Since these failures are caused by the variation in the device parameters, these are known as the parametric failures. A failure in any of the cells in a column (or row) of the memory will make that column (or row) faulty. Even though the failure rate of today's memory chips is relatively low, it still remains the most unreliable part due to high density of hardware components used in it. Therefore, investing in memory failure analysis, fault modeling and simulation, test algorithm development and evaluation, DFT, BIST, diagnosis, etc., has been considered one of the key factors in producing successful memory. Memory fault diagnosis thus is important so far as yield improvement is concerned. It is used in memory development to find design and/or process errors and inconsistency. So, methods to improve reliability by providing some internal fault diagnosis and hence provide tolerance are gaining its importance.

Many techniques have been proposed for fault tolerance in SRAM memories. Fault masking methods like Triple Modular Redundancy (TMR) [3] may be used for small size memories. But such techniques are not feasible for high capacity memories due to large additional hardware overhead. Error correcting codes can be implemented on memories [4] [5] for transient faults which can also be extended to PVT faults. ECC can be implemented for correction of 1-bit errors in memory words. Here a Single error-correcting and double error detecting codes (SECDED) can be implemented. However to detect multiple bit errors, the coding circuit complexity increases excessively and may not be cost effective implementation. In [6] a dedicated Content addressable memory (CAM) structure is used for storage and remapping of faulty addresses. But the use of CAM is largest part of hardware overhead. The memory unit used in the CAM is SRAM which may be prone to PVT variations and may have read and write failures.

In this paper a test pattern generator to generate the input patterns during test operation for word-oriented two-port SRAM array is used. Also a controller to generate signals necessary to switch between normal and test operations is used. Detection of faults is done during the test operation. In the SRAM array, a fault in at least one SRAM cell will mark the entire array as faulty. The fault status is stored in 1-bit D Flip-flop. This information is used to isolate the faulty rows.

Manuscript received October 14, 2012; revised November 25, 2012.

Padma Sravani Annam is with IIIT Hyderabad, India (e-mail: padmasravani.a@research.iiit.ac.in)

II. Two Port SRAM Cell

A. Circuit Description

Multi-port SRAMs are often implemented using static random access memory (SRAM) due to its fast operation and the ability to support multiple read and write operations simultaneously, thus increasing data throughput in embedded systems and meeting the expected demands of parallel or pipelined microprocessors. With the continuous scaling of transistor feature size, designing low power robust memories and investigating their failure characteristics become critical. A multi-port memory, such as two-port memory is actually a single memory array with two entirely independent sets of data, address and control lines as shown in Fig 1. Such a memory can be written to or read from through two different paths. By using two-port SRAMs, the efficiency of the memory accesses can be doubled. In the 2-port memories, the read and write ports are completely independent and can access any location simultaneously (but simultaneous access to same cell by both ports cannot happen).

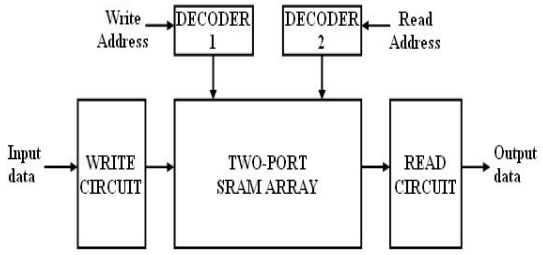

Fig. 1. Two-port SRAM memory array

In the present design one port is dedicated to write and another port is dedicated to read. An 8–T SRAM cell is shown in Fig 2. The write and read operations are carried by two wordlines and two sets of complimentary bitlines.

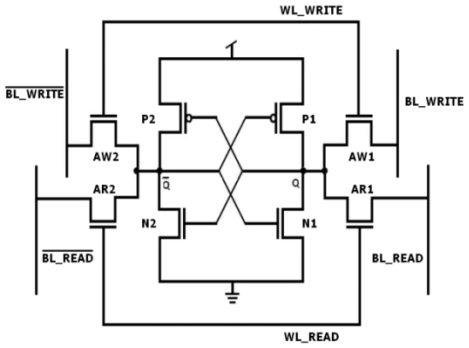

Fig. 2. Two port SRAM cell

In this paper, we study the defects occurring in the two-port SRAM cells. The memory is modelled at the transistor level and analyzed for defects by applying a set of test patterns. The memory used is word - oriented. A failure in any of the cells in a row of the memory will make that row faulty and entire row is isolated.

B. Faults in SRAM

With the down-scaling of integrated circuits the operating voltages of SRAM cells are lowered causing the reduction in static noise margin. Reduced read and write margins may cause errors in the respective read and write operations. Process variations in SRAM cells may cause one of the above failures.

1) Access failure

The cell access time is defined as the time required to produce a pre-specified voltage difference between two bit-lines (bit-differential). An increase in V_{th} of access transistors may cause decrease in current that discharge the bitlines through access transistors during read operation. This leads to reduced bit-differential voltage during sense operation resulting in wrong evaluation of sense amplifier.

2) Read failure

During the read operation, Q (Fig 2) increases due to the voltage divider action of AR1 and N1 to a positive value. Decrease in V_{th} of access transistor may result in voltage higher than the trip point of the inverter N2-P2 causing the cell to flip resulting in a read failure.

3) Write failure

During write operation, $\overline{Q}$ should be reduced below the trip point of inverter N1-P1. Variation in strength of the access transistors and the trip point of the cross coupled inverters may result in unsuccessful write.

4) Hold-stability failure

During stand by when the V_{DD} of the cell is reduced to reduce the leakage power consumption, the inability of an SRAM cell to hold its content may cause hold failure. Excessive mismatch in inverters may cause this failure.

III. Fault Isolation Architecture

A. Built in Self Test

In addition to process technology issues, guaranteeing the performance, quality, and reliability of the memory cores in a cost effective way requires further research efforts. A promising solution to this dilemma is BIST. In this design BIST is integrated with the SRAM array as shown in Fig 3. The BIST module will perform testing, analysis upon every power-up. Faults are stored in isolation circuit and it disconnects the address line with the wordline of SRAM array.

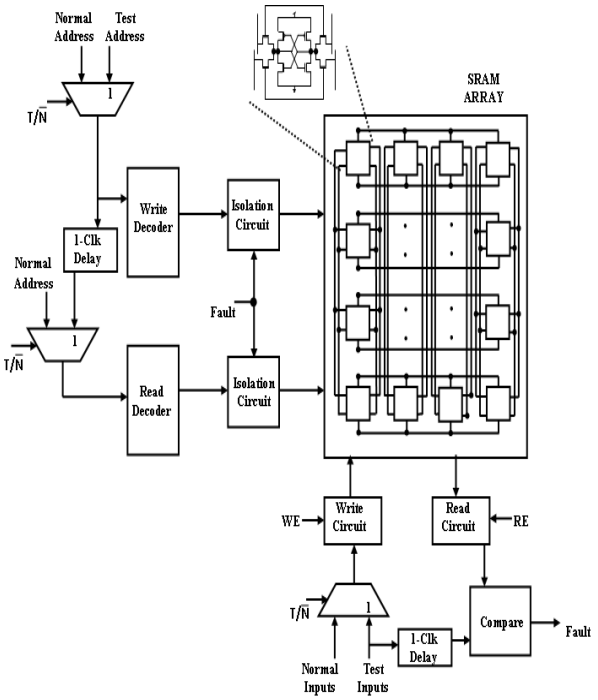

Fig. 3. SRAM array with BIST and isolation circuit

A fault in SRAM cell can occur during write or read, but it is detected only after a cell is read. During power on, the $T / \overline{N}$ goes high; test input addresses and test input are applied during this stage. Each input address is delayed for one clock period and applied to read decoder i.e., reading of the row is performed in the next clock cycle. For each read operation, the sense amplifier outputs are compared with the delayed inputs to determine the presence of fault. When *fault* = 1, the row is detected as faulty. Once a fault is detected, it informs the isolation circuit module to mark the defective row as faulty through the *fault* signal. The pointers are updated to invalidate the row. The size of the memory is reduced, as one row is removed. The system still works though performance may be affected. This approach effectively increases the product yield. The number of blocks that can be invalidated normally depends on the performance penalty that can be tolerated.

B. Fault Detection

The fault is detected by comparing the sense amplifier output with the delayed inputs. The output of compare circuit after the read operation should be considered. Exclusive – OR gates are used and bit by bit comparison is done in parallel. The fault is generated while RE (Read Enable is high).

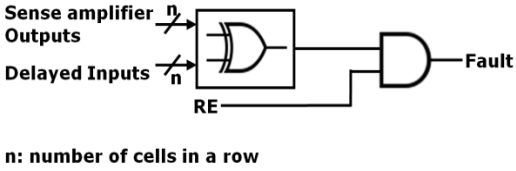

Fig. 4. Compare circuit

C. Isolation Circuit

The isolation circuit consists of d-flip-flop to store the fault status of that row. Each row has an isolation circuit which isolates the faulty rows from the array. The 1-bit fault status controls the switches which connect the address decoder outputs to the wordlines of SRAM array.

Initially at power-on all d-flip-flops are reset to '0'. Writing and subsequent reading of the row will detect the fault. The fault signal combined with the WL_read signal will detect the faulty row. The d-flip-flop is now set to '1'. The output of it controls the switch. Logic high to the gate of PMOS switch will turn it OFF. The OR gate is used to make sure the logic 1 remains during other clock cycles when fault and WL_read signals are updated. In this way the faulty rows are isolated by making respective wordlines pulled to ground. After all the rows in the memory are checked, the $T/\overline{N}$ signal goes low signalling normal operation. Here 1-bit d-flip-flop is sufficient to store the fault status of the row. During normal operation the additional delay will be due to multiplexer at input address and switch for isolation which is kept minimum.

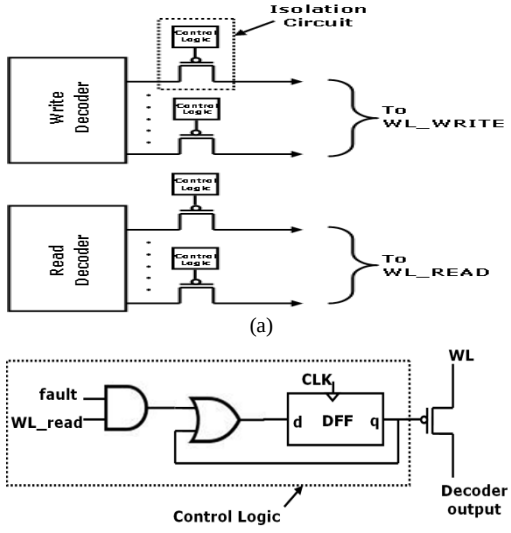

Fig 5. (a) Switches between address decoder output and wordlines (b) Control logic to the switch

IV. Simulation Results

The above circuit has been simulated in HSPICE using 65nm technology PTM model [7] at supply voltage of 1V. The circuit operates at 50MHz. During fault free condition the fault signal remains low. The decoder outputs are connected to the wordlines.

In Fig 6 the clock (Clk), write enable (WE), read enable (RE) with address decoder outputs (w_dadd_0, w_dadd_1 for write; r_dadd_0, r_dadd_1 for read of row #0 and row #1 respectively) and wordlines (WL_write_0, WL_write_1 for write; WL_read_0, WL_read_1 for read of row #0 and row #1 respectively) during fault free condition are shown. The decoder outputs are connected to the wordlines.

In Fig 7 the SRAM cell has a read failure. During read the cell data (Q) is flipped from '0' to '1'. The sense amplifier output (SA output) is obtained as '1' and the fault signal goes high, indicating the fault.

Fig 8 shows that the wordline for write and read are held low for the input address making sure the row contents are neither read nor written.

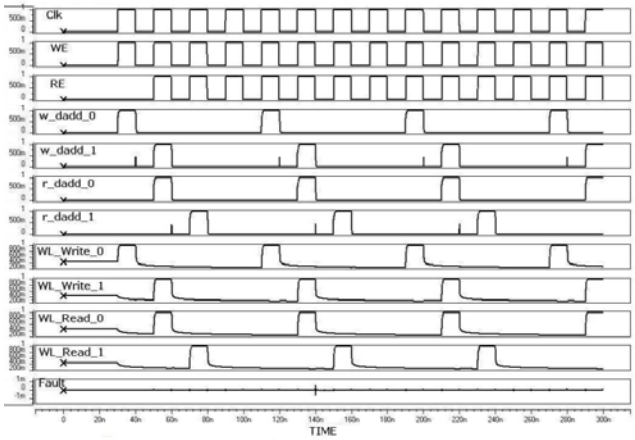

Fig. 6. Address decoder outputs and wordlines for write and read during fault free condition

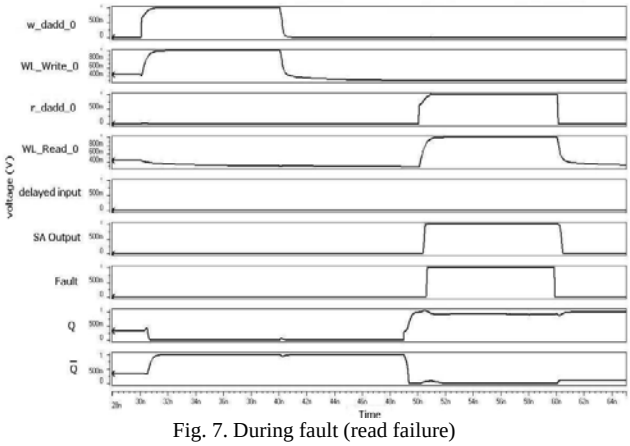

Fig. 7. During fault (read failure)

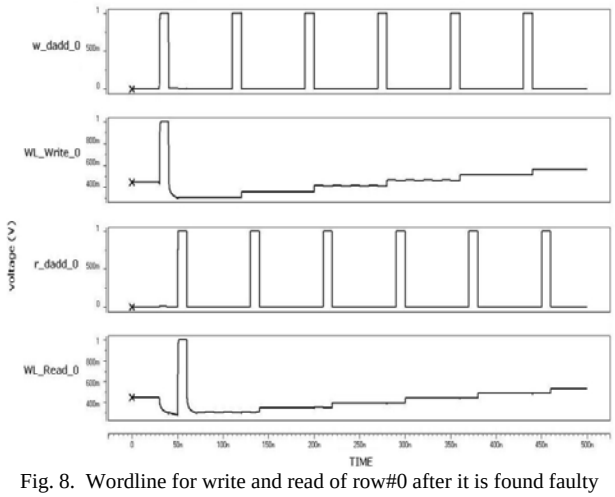

Fig. 8. Wordline for write and read of row#0 after it is found faulty

V. Conclusion

In this paper, we presented the two port SRAM cells and proposed fault isolating architecture to improve the yield in nanoscale memories under process variation. The architecture is word-oriented array integrated with built in self test. Circuit to detect and isolate faults are proposed. Any row having one or more faulty cells is considered to be faulty. The faulty row is isolated from the array by disconnecting wordline from input address by use of switches. This scheme has very less energy and area

overhead. The scheme has minimum affect the access time and performance. The circuit implementation is verified in HSPICE using 65nm technology.

REFERENCES

[1] A. Bhavanagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid State Circuits*, vol. 36, no. 4, pp. 658-665, Apr. 2001.

[2] G. Gielen *et al.*, "Analog and digital circuit design in 65 nm CMOS: end of the road," in *Proceedings of the Design*, Automation and Test in Europe, 2005

[3] D. P. Siewiorek and R. S. Swarz, The Theory and Practice of Reliable System Design, Digital Press, 1982.

[4] H. T. Vergos and D. Nikolos, "Efficient fault tolerant cache memory design," *Microprocessing and Microprogramming*, vol. 41 no. 2, pp.153-169, May 1995

[5] H. L. Kalter, C. H. Stapper, J. E. Barth, J. DiLorenzo, C. E. Drake, J. A. Fifield, G. A. Kelley, S. C. Lewis, W. B. V. D. Hoeven, and J. A. Yankosky, "A 50-ns 16-Mb DRAM with a 10-ns Data Rate and On-Chip ECC," *IEEE J Solid-State Circuits*, vol. 25, pp. 1118–1128, 1990

[6] S. Lu and C. Hsu, "Fault Tolerance Techniques for High Capacity RAM," *IEEE Transactions on Reliability*, vol. 55, pp. 293–304, June 2006

[7] Berkeley Predictive Technology Model website. [Online]. Available: http://www.eas.asu.edu/~ptm/