Design Principles of SRAM Memory in Nano-CMOS Technologies

Ezeogu Chinonso Apollos Scholar, National Information Technology Development Agency (NITDA) Nigeria

ABSTRACT

Static Random Access Memory (SRAM) is a volatile memory that is widely used in every embedded system - Silicon on Chip (SoC). Digital Signal Processing (DSP). Microcontroller, Field Programmable Gate Array (FPGA) and Video applications. It is also used in register, cache and cache-less applications due to large storage *density, reduced read-write access time, low power consumption and stability. Thus, this paper presents the design principle of SRAM at the 45nm technology node, the peripheral building blocks and functionalities, operations, transistor scalene challenges, and process variation effects of SRAM designs. A clear detail schematic diagrams using Cadence Virtuoso design tool for IC design was used for designing the peripheral circuitry and the SRAM cell.

General Terms

SRAM, CMOS, Design Principles, Nano Technology

Keywords

Memory cell, Embedded System, Read, Write, Process variation, Leakage current, Power consumption.

1. INTRODUCTION

SRAM memory is still currently the main memory block of today's embedded systems and computing devices cache and register designs. It is used in portable devices and in embedded systems due to high demand for data storage, computing speed, data stability, and low power consumption [1] and plays essential role in all Intel products in achieving power performance goals; process defect sensitivity and detection [2]. In high performance computing systems, the use of SRAM for cache design helps speed up data communication between the Central Processing Unit (CPU) and the memory block; this ensures that frequently accessed data are retrieved from the cache rather than the CPU - this technique is called the John Von Neumann stored program computation concept that is widely used. However, this multicore system computational efficiency is getting saturated in high data computation because of decreased efficiency, power consumption, and stability as CMOS scaling continues to get smaller below the 90nm technology node.

The SRAM cell is classified into different configurations which are named according to the number of transistors used in designing the memory cell. We have 4T, 5T, 6T, 7T, 8T, 9T, 10T and higher order SRAM configurations [1]. SRAM is used in main memory for cache-less embedded processors and hence must be optimized in terms of power, density, area and delay [4]. The design factors that are paramount to be considered when designing SRAM are power consumption, leakage current and stability under process variation. As CMOS scaling down is reaching its physical limit this poses other challenges in leakage power, reliability, test complexity, cost for mask and design, yield and fabrication processes [7]. Figure 1 shows 45nm 6T SRAM logic technology which was

believed to be the first fully functional SRAM using 45nm process technology manufactured on January 26, 2006 by Intel Corporation [1].



Figure 1a: Intel 45nm SRAM chip [2]



Figure 1b: Intel 45nm SRAM in CMOS technology [11]

From figure 1a, the Intel's SRAM bits are arranged in subarrays with rectangular matrix-like structure with rows and columns such that in a read or write operation from the memory subarray, a specific row and column are activated depending on the address, and a group of bits called a word are read or written [2]. The subarray is designed to be very compact since it is tiled many times, to form an on-die cache in a real product. After completion of the subarray design, then large portions of the X-chip area can be tiled with these subarrays with minimal additional effort [2].

2. SRAM DESIGN ARCHITECTURE

2.1 SRAM Block Diagram

SRAM complete block structure is shown in figure 2, with other peripheral circuitry such as sense amplifier, row and column decoders, read and write drivers, and timing control logic required for the complete implementations and simulation of the SRAM cell for read, write and hold states. The SRAM architecture is arranged in cores for a larger system then in blocks and arrays depending on the design specifications. The memory arrays are arranged in rows (word lines) and column (bit lines) of memory cells and have unique location defined by the intersection of the rows and column of the array. And each address or memory cell has its own precharged circuit, read buffer and write driver, sense amplifier, and activation word lines and bit lines for active cell selection.



Figure 2: SRAM Block Diagram [1].

2.2 SRAM Memory Cell

Static random access memory (SRAM) is a type of memory that uses bi-stable flip-flop (two cross-coupled inverters) made up of at least four transistors; the flip-flop may be in either of two stable states, 1 or 0, and also the access transistors grant access to stored data for read and write. Thus, the term static means that it holds data as long as power is applied which differentiates it from dynamic RAM which must be periodically refreshed. There are different types of SRAM cell based on their functionality, transistors number used and memory size. Based on functionality we have synchronous and asynchronous SRAMs. The synchronous SRAM uses one or more clock to time the operations of the SRAM and it is the most used design nowadays. While the asynchronous SRAM is independent of clock frequency with a sequential pattern of *read* and *write* operations and has some limitations when it comes to performance as compared to synchronous SRAMs.

Transistor number is now used to classify the operation, performance and improvement in SRAM's operations; the

number of transistors include 4T, 5T, 6T, 7T, 8T, 9T, 10T, 11T, and 12T. The cell name is given base on the number of transistors it contains, where "T" stands for "transistor". The fundamental building block of a Static Random Access Memory (SRAM) is the SRAM memory cell. The cell is activated by raising the word line and is read or written through the bit lines. There are three different states of an SRAM cell, namely: standby, reading and writing states to be discussed later in section 3.



Figure 3a: 4T SRAM Cell







Figure 3c: 7T SRAM Cell



Figure 3d: 8T SRAM Cell



Figure 3e: 9T SRAM Cell



Figure 3f: 10T SRAM Cell

2.3 Pre-charged Circuitry

The pre-charge circuitry is used for charging the bit lines to V_{dd} prior a read operation. It is necessary for both bit lines (BL and BLB) to be exactly at V_{dd} and perfectly equalized for a correct read operation [8]. Thus, the circuit is made up of three PMOS transistors, M1, M2, and M3. The M1 is the

International Journal of Computer Applications (0975 – 8887) Volume 178 – No. 11, May 2019

equalization transistor which ensure that asymmetric defect is eliminated for correct read operation than if it was not included; in order words it helps to minimize the voltage difference between the bit lines, reduce pre-charge time by making sure that the bit lines (BL and BLB) are at nearly at equal potential. A good pre-charge circuit should not have bit line voltage difference greater than 80mV for correct read operation of the SRAM cell [8]. Consequently, M2 and M3 are the load transistors that connect the bit lines to the V_{dd} for the pull up. It is also possible to use NMOS transistors which will pre-charge the bit lines to V_{dd} - V_{th}; this gives a faster single-ended bit line sensing because the bit lines do not swing as much to the V_{dd}, but the disadvantage is that it reduces the noise margins and require more pre-charge time.



Figure 4: Pre-charged Circuit [1].

2.4 Sense Amplifier

The sense amplifier detects a small differential voltage difference in the bit lines during read cycle then it amplifies it to full swing close to the V_{dd} thereby reducing the time required for read operation, in order words, the read operation is very fast. Because SRAM does not require constant refreshing as DRAM, hence sensing must be non-destructive unlike the destructive sensing of the DRAM cell [1].



Figure 5: Sense Amplifier [1].

Thus, the sense amplifier robustness is an important component of the SRAM block; the choice of sense amplifier design used affects the robustness of the bit line sensing, the read speed, reliability and power consumption of the design. Sense amplifiers reduce delay in heavily capacitive loaded logic circuits. Furthermore, the sense amplifier in figure 5 is a clocked control sensed amplifier. The clocking helps save power because it consumes power only when the clock is activated, however it requires good timing chain to be activated at the right time. If it is activated too early then the bit lines may not be pre-charged enough to operate reliably while if it activates too late then the SRAM operation will be slow.

From figure 5, when the sense clock (SAE) is low the amplifier is inactive and it is said to be in pre-charging mode and the transistors P1 and P4 are called isolation transistors, that charge the output nodes to V_{dd} . Furthermore, when it is high (SAE=1) then M4 and M8 are activated and the sense amplifier is said to be in sensing mode. Transistors M4 and M8 work as common source differential amplifier [9], turning on the cross-coupled inverter pair, which puts one of the output low and the other high through regenerative feedback. And power dissipation can be reduced during read operations by turning off the word lines and the sense amplifier once a sufficient differential voltage has been reached and end of data sensing at the output on the bit lines so as to save power. The design constraints of the sense amplifier can be summarized as follows: Input differential voltage should be small; less sensitive to process variations (environmental factors such as noise, temperature and voltage supply); Transistors sizing is critical and the amplifier should be made symmetrical; and area should be minimum.

2.5 Row and Column Decoder

The decoders make the number of interconnects to be quantified to a factor of $\log_2 N$ where N is the number of independent address locations. In order words, they help in

reducing the total number of pins such that if there are N + K address lines then the bit storage will be 2^{N+K} bits. Thus, supposing we have three address bits (input pins) A0, A1, A2 as shown in our SRAM block design in figure 2; then N+K=3 therefore memory address space is $2^3 = 8$ bits (1byte) wide. Row decoder is used for the selection of rows (word lines). While the column decoder is used to select the bit lines. In addition, there is a multiplexer used to select one or more columns for input/output of data [1].

2.6 Write Driver

The write driver is used for writing data into the SRAM cell through the access transistors. The write driver will pull one of the bit lines low and the other high depending on the data input. Supposing we have data = 0, then the bit line (BL) is pulled down to gnd while the BLB is at V_{dd}. Similarly, if data =1 then one of the bit lines (BL) is pulled to V_{dd} by transistor M3 while the BLB is pulled to gnd by the transistor M2. Thus, to write data to a memory cell, the data is applied to the data input pin. Then the cell must be selected using its corresponding row and column coordinates. This is achieved by pulling the specific address column low for that cell while enabling the write line (write enable) and also inserting the word line; when the data is established in the cell then the word line is turned off to preserve the power. In addition, the write driver transistors are designed to be stronger than the relatively weak transistors of the cell so that data can easily override the previous state data of the memory. Thus, careful sizing is necessary to ensure correct operation of the entire design.



Figure 6 Write Driver [1].

2.7 Timing Block Diagram

The timing block design as shown in figure 7, is the control circuit for pre-charge clocking, word line (WL), read wordline (RWL), sense amplifier, write enable signal activation in order to make sure that the correct timing sequence is achieved during read and write operations. Timing hazards have to be prevented so not to have unintended read/write operation. The timing block presented in this paper was designed to help improve read and write of the SRAM by using wordline boosting scheme for write - that is, using a slightly higher voltage than the V_{dd} for the wordline [1].



Figure 7 Timer Schematic Diagra[1]. 3. SRAM OPERATIONAL PRINCIPLE

3.1 Basic Operation

To understand the basic function of SRAM, the 6T SRAM will be used for the detailed analogy. The 6T SRAM is made up of six transistors, whereby two of the transistors are PMOS and four are NMOS. The configuration is such that the PMOS and NMOS form a cross-coupled inverter while two NMOS transistors are connected one each to the bit lines (see Figure 8). Thus, these NMOS bit lines connected transistors are referred to as the "access transistors" which are controlled by the word line. The 6T configuration expels most of the limitations of the 4T and 5T to a certain degree especially with the advantage of better noise immunity.

However, external noise during read operation, power consumption, and stability are still issues to be tackled. Static power dissipation in 6T SRAM is relatively small. The cell only draws current from the power supply during switching. But, during the idle state the cell's leakage current in the deep-sub-micron technology becomes an issue because of the leakage current and data retention at low operating voltages. The 6T SRAM consists of two PMOS (MP1 and MP2) known as the load transistors, two NMOS (MN1 and MN2) known as the drivers' transistors and also two NMOS (MN3 and MN4) known as the access transistors. There are three major operations of SRAM: retention/standby, read and write operations. These operations are explained below:



Figure 8: 6T SRAM Architecture [1]

3.1.1 Standby (Retention) Operation

This is the state when the SRAM cell is idle (data is held in latch) and the bit line and bit line bar (data path) are kept at *gnd* when the access transistors are disconnected because the word line is not inserted. Thus, the PMOS transistors will continue to re-enforce each other as long as they are connected to the power supply in order to keep the data stored in the latch as shown in Figure 9a. During this idle/retention mode, when "1" is stored in the cell, MP1 and MN1 are ON thus there exists a positive feedback between Q and QB nodes making Q to be pulled to Vdd. Similarly, when "0" is stored in the cell, MP1 and MN1 are OFF while QB is pulled to V_{dd}.



Figure 9(a): Retention Mode [1]

3.1.2 Read Operation

This is the state when data is requested from the memory cell. Therefore, to read data, both bit line (BL) and bit line-bar (BLB) are initially pre-charged to a logic state 1 (V_{dd}), when the word line (WL=0) is low. After the pre-charge cycle the word line (WL) is enabled (WL=1) thus the access transistors

as shown in figure 9b, MN3 and MN4, are switched ON thereby connecting them to the bit lines. See figure 9(b) for simplified schematic during read 0. For instance, if data Q = 0through the bit line (BL) then the gate of MN2 is turned off and MP1 initially held high due to the pre-charging will then go "off" - PMOS are turned ON when the gate input is low while NMOS when the gate input is high. Note that Q = 0, then QB=1 which is then fed as feedback to MP1 and MN1 hence switching MN1 ON while MP1 OFF. Thereafter the current (Icell) now moves from the bit line (BL) through MN3 to the storage node Q thereby charging node Q while discharging the bit line (BL). Since MN1 is ON then the current from the node Q is further discharged to gnd; this is possible by making the width (cell ratio) of MN1 wider than MN3. The bit line voltage, (VBL), having been discharged to $(V_{dd} - V_{th})$ the sense amplifier detects this voltage difference in the bit lines and is then triggered and speedily amplifies the small differential voltage between the bit lines to full swing close to (V_{dd}) by identifying the bit line with the higher voltage raising it V_{dd} while the lesser voltage is discharged slowly through gnd. Then the data is kept at a stable state by a sense amplifier. Conversely, if the data to be stored is "1" (see figure 9c) the potential at node Q and the bit line potential will be equal so no discharge will take place; however, at node QB = 0, the bit line-bar (BLB) potential is higher so the discharge current will move through transistor MN4 to MN2, thus discharging BLB to V_{dd} -V_{th}. The sense amplifier will pull BLB to gnd while BL remains at V_{dd} .



3.1.2.1 Read Constraint

The increase in voltage at node Q should not be too large to switch ON transistor MN2. The 6T SRAM design shown in figure 8 is symmetrical, therefore all dimension for left symmetry will be used for the right symmetry. Therefore, the condition below must be satisfied, Cell Ratio (CR): The ratio between sizes of driver transistor to the load transistor during read operation.

Cell Ratio (CR) =
$$\frac{W_{MN1}}{L_{MN1}} > \frac{W_{MN3}}{L_{MN3}}$$
.....(1)

Since $L_{MN1} = L_{MN3} = 45$ nm, then $W_{MN1} > W_{MN3}$ by a factor of at least 1.2 in order to ensure adequate noise margin and no-destructive read. And as CR increases, the speed of the SRAM cell increases [9].

Similarly, $W_{MN3} = W_{MN4}$ and $W_{MN1} = W_{MN2}$

The differential voltage developed on the bit lines depends on the cell current, I_{cell} , bit line capacitance, C_{bit} , and the length of time, δ_t , the word line is activated. The current should be large enough in order to discharge the bit line capacitance.

$$I_{cell} = C_{bit} \times \frac{\delta V}{\delta t} \dots (2)$$

Read Operation Summary:

1. Pre-charge the bit lines to $V_{\rm dd}$ while the word line and sense

amplifiers are disabled;

2. Lower the column decoder of the given memory cell to be

read;

3. Enable the word line and sense amplifier after the pre-

charge cycle;

4. Sense amplifier reads data from bit lines;

5. Read outputs using the bit lines, a drop in bitline (BL) indicates data = 0 else data =1.

3.1.3 Write Operation

This is the state when data is been written/updated in the cell (see Figure 9d). To write data into a cell, the sense amplifier and pre-charge circuits are deactivated while write enable and the word line are first activated then the input data is driven through the write driver input pin then the bit line is pulled to the value of the given data while the bitline bar (BLB) takes the complementary value. For instance, if data=0 then BL =0 while BLB = 1 (V_{dd}); conversely, if data=1 then BL =1(V_{dd}) while BLB = 0 (*gnd*). Hence, given that transistors MP1 and MN3 are correctly sized then cell will flip and the data is effectively written.



Figure 9(d): Simplified Schematic During WriteOperation (switching data $0 \rightarrow 1$)

Thus, consider when data = 1 to be written to a cell node initially storing a "0", then the transistors MP2 and MN4 will function as pseudo-NMOS (MN3 is ON) inverter then current flows through the storage node to bit line-bar (BLB) and also through MP2 to the storage node (QB) as soon as the potential at this node starts decreasing. This results in a voltage drop at

node Q then MP2 pulls down Q and cause the SRAM to switch values $% \left({{{\rm{A}}} \right)^{2}} \right)$

$$Pull up Ratio(PR) = \frac{W_{MP2}}{L_{MP2}} / \frac{W_{MN4}}{L_{MN4}} \dots (3)$$

In equation 3, the NMOS has mobility higher than the PMOS, therefore for correct write operation the NMOS can be sized to be equal to or greater than PMOS at minimum size.

Write Operation Summary:

1. Drive one of the bit lines high and the other low;

2. Load data into write driver input pin;

3. Enable the "write line" and the "word line" simultaneously;

4. Data is overwritten due to the weak SRAM cell transistors compared to the write driver.

3.2 Transistor Scaling and Challenges

Moore's law states that the shrinking in density of transistor size will double every 1.5 years. This predictive law for CMOS has always proven to be true; the technology-node scaling is driven by the need for high integration density and performance required in cache designs and microprocessors. This has led to increase in statistical variation in the process parameters which can cause increase in the total leakage current. Hence, the reduction in threshold voltage, channel length, drain/source junction depth, gate oxide thickness and V_{dd} has become a major contributor to increase in leakage current. The sub threshold leakage is the drain-source current of the transistor when the gate-source voltage is less than the threshold voltage; this is large for short channel devices. The gate current leakage is due to low oxide thickness and the high electric field resulting to current flow through gate of the transistor even during the off state because the classical infinite impedance assumption of MOS is destroyed by the energy field. Due to this increase in leakage current, the static power consumption thus exceeds switching component of the power consumption.

3.3 Effect of Process Variation in SRAM

Process variations are the critical design parameters – die to die and intra-die variation – from equipment processing in the semiconductor design technology due to inability to precisely control the fabrication process at small feature technologies at the nano-scale which in turn results in large variation in the operation and functionality of the design. This is very severe in the case of memory components as minimum sized transistors are used in their design [11]. These variations include the *film thickness*, *lateral dimensions*, *doping concentration* and *threshold voltage variation*. All these contribute to the circuit optimization for performance and power consumption. Doping concentration affects the threshold voltage, the V_{th} increases steadily as a result of more random dopant fluctuations in channel, source and drain due to increase delay distribution and delay spread. Consequently,

these random and systematic fluctuations affect the stability of the SRAM [1]. Therefore, in the 6T SRAM design, the read stability of the cell is determined by the ratio of the current produced by the access transistors MN3 and MN4. Furthermore, the impact of variation increases as the supply voltage, V_{dd} , scales down to V_{th} because the sensitivity of the circuit delay amplifies. Temperature and voltage variation are environmental variations which are primarily a function of intra-die (within die) variations, and contribute to failure rate (write ability and read stability) in SRAM cells.

4. SRAM CHALLENGES MITIGATION

SRAM is designed to reduce power loss, mitigate scaling challenges, process variation and single and multiple event upset and other challenges. Many architectures have been proposed to evade these problems, some of which include:

- (i) The use of Statistical DOE-ILP Power-Performance-Process (P3) optimisation of Nano-CMOS SRAM technique [10] is very efficient because this approach can reduce 61% power consumption and 13% SNM increase when 6T and 8T were subjected to this method at 45nm technology node.
- (ii) Block permutation scheme [3] is another proposed scheme to minimize process variation by permuting cache blocks to maximize the distance between blocks with consecutive addresses, thus if the area increases then power density is minimized as a result of increase in the working sets.
- (iii) The N-curve based power metrics (SPNM and WTP) [5] which takes into account both voltage and current for the power metric analysis to measure the stability of an SRAM cell, thus, in addition statistical model for estimating the static power noise margin (SPNM) and the write trip power (WTP) are given for process variation in threshold voltage, Vth.
- (iv) Double ended read decoupled 9T SRAM designed [4].
- (v) Read-Write Assist techniques, V_{dd} lowering [6], Vss raising, boosted word-line gate voltage and negative bit-line.
- (vi) The use of dynamic sleep design to lower SRAM power supply to effectively reduce static power consumption by reducing leakage [12].

5. CONCLUSION

In this paper, the design principles of SRAM including the peripheral circuitry, operations, challenges, mitigation techniques, how to improve the stability were clearly explained in a very simplified manner including the design schematic to show the memory cell and logic gates of the SRAM block. Meanwhile, designers and researchers are faced with issues with process variation which include: process technology, voltage, temperature, stability and leakage power as transistors sizes are scaled down, and effects under single and multiple event upset. However, SRAM is still currently the most widely used memory in embedded systems; its use is now extended in designing non-volatile memory using memristors [1]. Consequently, researchers are currently exploring its use for SRAM-based physical unclonable function architecture, SRAM-based Computation-in-Memory architecture, and field programmable gate arrays. It is worth to note that SRAM will still continue to play major role in embedded system and computing devices even in decades to come as more applications of its use are been explored.

6. REFERENCES

[1] Ezeogu, Apollos. 2013 "Process Variation Aware Non-Volatile (Memristive) 9T SRAM Memory Design in Nano-CMOS Technologies", M.Sc. Theses submitted to University of Bristol, United Kingdom.

- [2] Uddalak Bhattachara et al.,2008 "45nm SRAM Technology Development and Technology Lead Vehicle" Intel Technology Journal, Volume 12, Issue 2
- [3] Milad Zamani, Sina Hassanzadeh, Khosrow Hajsadeghi and Roghayeh Saeidi, 2013"A 32kb 90nm 9T -cell Subthreshold SRAM with Improved Read and Write SNM" 8th International Conference on Design and Technology of Integrated Systems in Nanoscale Era(DTIS).
- [4] Arvind Chakrapani 2018, "Survey on the design methods of low power SRAM cell" in International Journal of Pure and Applied Mathematics
- [5] Singh Jawar, Mathew Jimson, Pradhan Dhiraj K., Mohanty Saraju P. 2008 "Failure analysis for ultra low power nano-CMOS SRAM under process variations". Soc Conference, IEEE international, IEEE conference publications, P251 -254.
- [6] Mohammad, M. O, Saint-Laurent, P. Bassett, andAbraham J., 2008 "Cache design for low power and high yield," in Proc. 9th International Symposium on Quality Electronic Design ISQED 2008, 17–19 March 2008, pp. 103–107.
- [7] Hoang Anh Du Nguyen, Lei Xie, Mottaqiallah Taouil, Razvan Nane, Said Hamdioui, Koen Bertels, 2015 "Computation-In-Memory Based Parallel Adder" Laboratory of Computer Engineering, Faculty of EE, Mathematics and CS Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.
- [8] Luigi Dilillo, Patrick Girard, Serge Prevossoudovitch, Arnaud Virazel, "Resistive-Open Defect Influence in SRAM Pre-Charge Circuits: Analyshis and Characterization" Proceeding of European Test Symposium(ETS'05) Copyright 2005, IEEE.
- [9] Shalinin, Anand Kumar, 2013 "Design of High Speed and Low Power Sense Amplifier for SRAM Applications", International Journal of Scientific & Engineering Research Volume 4, Issue 7, ISSN 2229-5518, pp 402 -406.
- [10] Thakral Garima, Mohantu Saraju P., Ghai Dhru, Pradhan Dhiraj K. 2010 "P3 (Power- Performance-Process) Optimisation of Nano-CMOS SRAM using statistical DOE-ILP ". Quality Electronic Design (ISQED), 11th International Symposium on, p176- 183.
- [11] Mutyam M, Narayanan V. 2007 "Working with Process Variation Aware Cache". Design, Automation & Test in Europe Conference & Exhibition, p1-6.
- [12] K.-S. Min, K. Kanda, and T. Sakurai, 2003 "Row-by-row dynamic source-line voltage control (RRDSV) schemefor two orders of magnitude leakage current reductionof sub-1-V-VDD SRAM's." In *Proceedings IEEEInternational Symposium Low Power Electronics* andDesign (ISLPED), pp. 66–71.