
DESIGN AND ANALYSIS OF LOW POWER SRAM

Nitin Kumari¹, Dr. Hitanshu²

¹M. Tech Scholar, Ganga Institute Of Technology & Management Kablana Jhajjar, India.

²Assistant Professor, Ganga Institute Of Technology & Management Kablana Jhajjar, India.

ABSTRACT

SRAM, or static random access memory is essential to the functionality and energy efficiency of contemporary integrated circuits. It is becoming more and more important for SRAM to minimise power consumption as the market for portable, battery-operated products grows. An extensive examination of design approaches and analysis strategies for low-power SRAM implementations is presented in this study.

The first part lists the essential variables affecting power consumption and describes the basic ideas behind SRAM functioning. It talks about the trade-offs between space, performance, and power and emphasises how important SRAM is to creating system designs that are energy-efficient.

The following sections discuss several approaches to lower the power consumption of SRAM designs. Among these are circuit-level strategies like power gating, dynamic voltage and frequency scaling (DVFS), and voltage scaling. In addition, architectural optimisations are studied, including cache coherence techniques, bank-level power management, and hierarchical bitline architectures.

Additionally, this work explores low-power optimised advanced SRAM cell designs. It investigates the effectiveness of methods for reducing both static and dynamic power dissipation, including subthreshold operation, multi-threshold voltage transistors, and decreased supply voltage. Moreover, thorough analytical techniques for assessing the trade-offs between power and performance in low-power SRAM architectures are provided. This includes methods for predicting energy consumption under different operating circumstances and workloads that are based on simulation as well as analytical models.

There is additional discussion of the difficulties and constraints posed by low-power SRAM design, such as performance deterioration, production variances, and reliability problems. Here are some ideas for overcoming these obstacles without sacrificing energy efficiency.

Key Words: SRAM (Static Random Access Memory), Low power, Design optimization, Energy-efficient SRAM, Power reduction techniques, Leakage current

1. INTRODUCTION

In current time the demand of high-speed processors which are operated in very less power and required lesser area, is increasing exponentially with time. A perfect memory is reliable, space efficient and quick with consuming minimum power. Almost all VLSI chips now include fast low power SRAM as a vital component, and it's particularly true for all the processors, where cache memory sizes are increasing on the SOCs with generation to overcome the performance gap between the memory and the processors [1, 2]. Furthermore due to greater integration, running speed and the exponential expansion of battery operated product, power dissipation has become an essential problem [3]. The design of SRAM investigated in this thesis, with an emphasis on optimizing latency and power, but scaling of supply [6, 11] and the process [4, 5] will always be most important factor. This thesis looks at certain approaches that will be utilised in certain combinations with scaling to produce fast low-power operation.

As there is a demand in battery-operated portable devices increasing, low power and compact size design becoming a crucial factor and most favourable area of research for researchers. To design a fast processor, fast sampling of data is one of the major constraints, and SRAM plays this crucial role. SRAM works as mediator between processor and main memory which are interfaced with other slower peripheral. RAM work as cache memory in a processor. SRAM or cache memory consists of instruction data temporarily, which are frequently in use by the processor. SRAM is 2nd fastest memory in memory organization. It comes after the register type memory. With passing of time as the demand of high-speed operation is increasing, size of cache memory is also increasing proportionally. Since larger size of SRAM memory provides a wider operational bandwidth, which implies larger size of data can be received or can send to destination. Bandwidth is directly related to the speed of operation of processor. i.e., as bandwidth increases speed of operation will also improve. However, as the size of SRAM on System.

on Chip increases, it becomes a greater area consumer on SOC board. On an average it consumes about 50% of the total area on a SOC board. This is a major concern. In this project, we will try to optimize our SRAM cell design at block level so that the requirement of area can be minimized. The next work will upon low power techniques at block

level and at transistor level to minimize the power dissipation. Some features of SRAM are like it does not require data refreshing after certain period of interval. These properties of SRAM cell eliminate the requirement of complex and area taking peripherals. It implies that SRAM can retain the data until unless power is not removing, so SRAM need to be connected to the power supply when it is in use. As SRAM is a mediator between processor and other peripherals and memories, it has wide application in many areas for example in wireless communication, in DSPs, in portable devices etc.

SRAM cell generally designed using only MOSFETs. 6T SRAM is a most used design in SRAM cell design. Many other designs is also proposed by the researchers, which have an edge in terms operational speed, data stability, reduced recite and inscribe cycle time, but still 6T SRAM cell is the first choice of memory manufacturing industry. Since SRAM cell already consumes about 50% area of the total SOC board which is a major constraint. With passing of time much advancement is also proposes by the researcher in 6T SRAM cell design and it is highly compatible with other SRAM design. In this project the work will upon designing of a low power 6T SRAM cell, and using this cell we design a SRAM memory using the concept of memory banking. The design should be low power and high- speed peripherals of SRAM memory.

2. OBJECTIVES

To work on this thesis various investigation done and investigate several strategies and different techniques for lowering the SRAM's leakage power design using CMOS technology. Stacked SRAM are proposed here, in order to lowering the leakage power without compromising the performance of the SRAM. It is also investigated the effect of different temperature on SRAM circuits

3. LITERATURE REVIEW

Shokoufeh Naghizadeh, Mohammad Gholami [1], explained about the importance of SRAM. In present days there is increase in demand of portable device like mobile laptops, tablets and many more. These device provide limitless functionality, to provide limitless functionality they limitless power. But the battery technology is not developed at that speed. Due to this reason the portable device runs on battery, the battery should not drain too fast. So, for lowering the leakage power consumption of the battery operated device. They show different techniques. RAM is most important device for the electronic circuit. It is used in the many integrated chip or SOCs due to its speed. It is used for the fetching the data and instruction from the main memory.

- Hemanta Kr. Mondal and Debasis Mukherjee [2], described the butterfly approach's methods for calculating, reading, and writing the static noise margin of a 6T SRAM cell. In his work, he also detailed the pull-up and cell ratios, their calculations, and how they impact the stability of memory cells. The pull-down or NMOS transistor in a normal 6T SRAM cell has to be larger than the pull-up and access transistors in order to maintain stability throughout the reading operation. To lower bit line capacitances, the access transistor size must be smaller. On the other hand an access transistor with strong current capacity and a high size should be used for write operations. There are competing design requirements for the cell's write capacity and read stability. Create distinct paths for reading and writing to address this issue.

Jan M. Rabaey, Anantha Chandrakasan B. Nikolic [3], Proposed SRAM design and also explained the design of memory using the SRAM cells, also explain about the other peripheral like row and tree decoder, pre-charge, sense amplifier and the connection between them for memory creation.

Sung M. Kang and Y. Leblebici [4], explained SRAM cell design and its reading writing and hold operation. Also explain that how a cell is connected to create memory cell, all the cell are connected together in array fashion. That the alternation of stored data is not permitted in reading process. In his book DRAM is also introduced. ITRS predicted that the SRAM take 90% area of the chip and consume massive power it's about 50% of total power. So, we require the SRAM which will take less power and give high performance. Most of time SRAM is idle and it take power to store data. In other hand the passive does not required power to store the data. But we use SRAM because of its high speed. It helps to enhance the speed of the operation of the processor.

K. Yamaguchi, H. Nambu [5], proposed the 64KB CMOS In which ECL are the word line drivers. He uses combination of ECL work line, cell array of CMOS SRAM and some write circuit. The ECL word line drivers and write circuit drive the CMOS SRAM Cell arrays without the use of an intermediate voltage level converter.

Rakesh Dayaramji Chandankhede [6], Proposed a decoupled latch sense amplifier which is a current control that reduces power consumption while improving performance. When the enable signal is logic low, the bit line logic and Bit-line Bar logic grows on the latch output, i.e. differential voltage. Whenever the pull down NMOS transistor in SRAM is off, the logic on bit - line and logic on bit-line bar does not expand, and the enable signal remains high.

When the EN signal is high, the low- voltage line will go to ground.

Sreerama Reddy G M, P Chnadra sekhara Reddy [7], Proposed an 8KB SRAM which consume less power than congenital 6T SRAM, also introduce memory banking approach which works at 800MHz. This 8KB low power SRAM was based on 180 nm technology. They also addressed the reduction in power dissipation and clock latency here.

Harekrishna Kumar, V. K. Tomar, [8], proposed the various type of leakage power in SRAM through this paper and explains it. In present use of the portable devices has increased exponentially. And it is increasing day by day. SRAM is major part of embedded memory and SOC. As technology size decrease the leakage power of devices has become the major issue. It also degrades the power supply. We need to develop the device that consume less power, and gives high performance. Here a comparison has been done and the analysis of 6T, 7T, 8T, 9T, and 10T SRAM. Then give the different idea to reduce it, introduced many schematic diagrams to reduce the leakage power and current.

4. METHODOLOGY

Over time, improvements are happened in SRAM array organisation in memory and circuit design that leads to lower the delay and power of practical SRAMs. This chapter is going to explores about both of these subjects as well as the concerns that this thesis addresses. In Section 3.1, The different strategy like partitioning strategies before highlighting the key circuit solutions which were published in the literature to make certain improvements in speed and power.

PARTITIONING OF SRAM

In huge SRAMs, partitioning arrays of cell into smaller sub units as in cell arrays, instead having a single monolithic array as depicted in Figure 3.1, can result in significant gains in delay and power. A big array is typically made into several no. of same sized sub arrays (usually known as macros), each of which holds a portion of the accessible word, termed the sub word, and all of which are triggered at the same time to access the entire word [10,12]. Low-power SRAMs normally have only one macro, whereas high presentation SRAMs can have up to 16 macros. Except for sharing sections of the decoder, the macros can be regarded of as self-regulating RAMs. The basic structure of every macro is the same as the one shown in below Figure 3.1. Word line plays an important role while access to a row, it enables all of the cells in particular row, and the column multiplexers access the requested sub word. For macros with a large set of columns, this layout has two drawbacks: The RC latency on the WL increases as the row's square and number of cells. The number of columns increases the bit-line power linearly. Both of these flaws can be addressed by subdividing into small – small pieces units of cells by using approach. The DWL approach divides a typical array's lengthy word line into k pieces, which doesn't depend to any circuit, and triggered separately, lowering the RC delay by K^2 and the word line length by k. The DWL architecture is shown in Figure It splits a 256-column matrix into four blocks, each with 64 columns. There are now two steps involved in picking a row.

The word line which is enabling first which is global, and then a block select signal is sent into the appropriate block to enable the specified local word line. The local word line has a reduced RC latency since it is smaller (just 64 columns long). Despite the fact that the global.

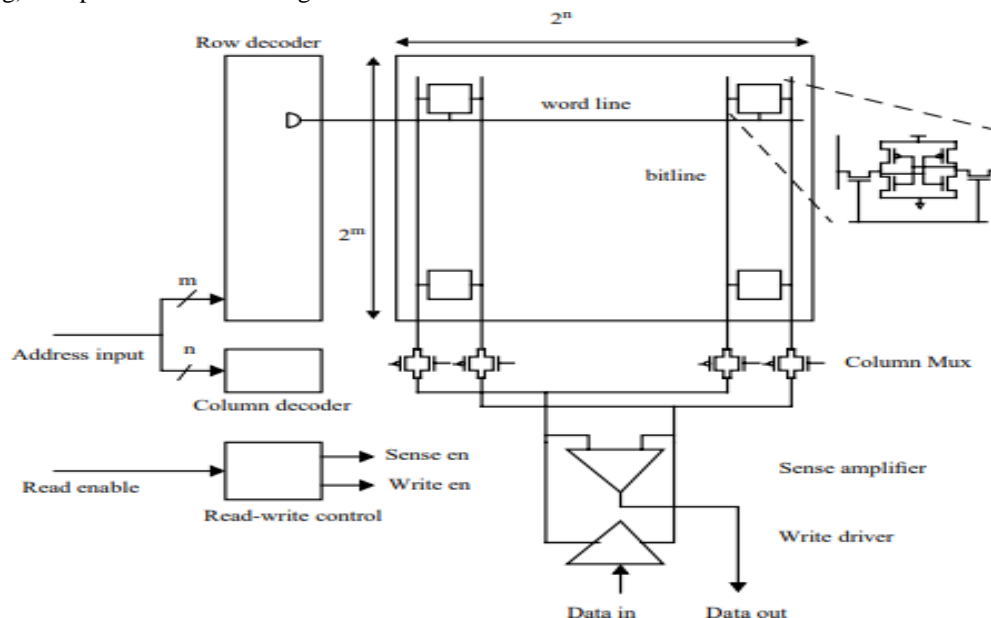


Fig.: 1 Cell design framework using SRAM

The macro's width and word line are nearly same; it will possess lesser RC latency than a whole word line due to its reduced loading capacitance. Despite the fact that the main word lines are almost similar to the macro's width, it has a very less RC latency than a whole word line due to its relatively low load capacitance. Rather than seeing the loading of all 256 cells, the four word line drivers can be seen input loading. Furthermore, because it is used bigger wires on an advanced level metal layer, its resistance may be lower. The word line RC latency is further reduced by a factor of four by keeping the word drivers in the centre of the word line segments and cutting each section's length in half. The column current is also lowered by a factor of four because among 256 cells first 64 cells are activated in the undivided array. The Hierarchical Decoding (HWD) method [14] is based on the idea of partitioning the word line recursively on the Main word line (and the blocking select line) for large RAMs. Partitioning can also be used to minimise bit-line heights, as explained in the following section.

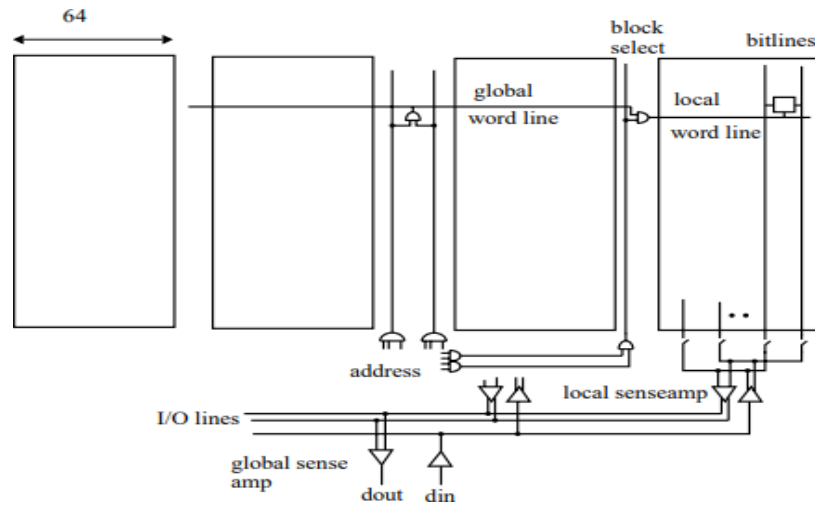


Fig.:2 Architecture of Divided Word Line (DWL)

RAM partitioning results in area overhead at the partition boundaries. A partition that deconstructs the word lines, for illustration, necessitates the deployment of word line drivers at the boundaries. Because the area of RAM defines the decoder's main wires length and the data channel, it has a direct impact on their latency and energy. Just look at the Adjustment in latency, area and energy achieved through the dividing.

CIRCUIT METHOD IN SRAM

The data path the decoder and are two components that make up the SRAM access path. The address input circuit to the word line are included in the decoder. The circuits make cells to the input and output ports and make up the data path.

The logic purpose of decoder is similar to n-input AND gates, with a level arrangement of the more fan-in AND operation. Figure 3.3 depicts the architecture of a two-level 8 to 256 converter. The pre decoder is the first level, where among the four address inputs two groups and the complement of (A0, A0, A1, A1,...) are decoded active first of its 16 pre- decoder o/p the wires, resulting in partially decoded products (A0A1A2A3, A0A1A2A3).

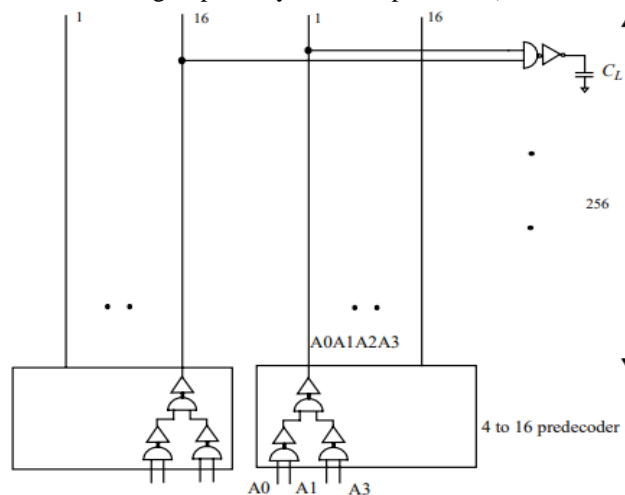


Fig.:3 Diagram of decoders

To enabling the word line, the pre-decoder o/p is merged at the other stage. The gate latency along path (crucial), as well as interconnection delays of the word line and pre- decoder and wires, makes up the decoder's latency. In large SRAMs, the latencies of wire in the structure, particularly of the word line, become critical as the RC latency generate due to the wire length. The size of the gates which are available in the decoder provides always a trade-offs in between the consumption of power and latency. A lot of researchers have looked into transistor size for both fast speed and low power. Because of the presence of intermediary link from the pre-decode wires, the decoder sizing problem is slightly difficult. We look at this issue and propose lower bounds on latency. We also look at some simple scaling strategies for achieving high speed and minimal power. By optimising the circuit style utilised to design the decoder gates, the decoder delay can be considerably reduced. Designs in old days very simple use in combinational method using static CMOS circuits to implement the decode logic function

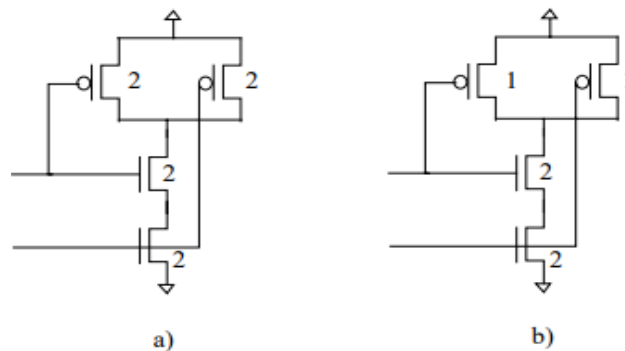


Fig. 4 (a) NAND gate (conventional) b) NAND gate by Namura.

The maximum latency to the ancient line(word) and the time to insert a new line(word) is therefore the decoder latency of gate in such a design, and its reduced when each gate available in the decoder will be designed in such a way so that it has equal timing of rising and falling. Using pulsed circuit approaches, where the word line is a pulse that stays active for a given minimum length before shutting off, then the latency of gate can will be minimal. As a result, all of the before any accessing, the word lines which are disabled, and the decoder only needs to reactivate them for the updated row address. Because the decoder logic chain only requires transmitting one type of transition, the sizes of transistor plays an important role to increase the speed and reduce the latency and delay timing. Diagram illustrates this method, in which the size of enhancement P type transistor is half in the NAND gates Structure. The size of PMOS can be lowered further by a half in the pulsed design while maintaining the same rising latency since both inputs are guaranteed to de-assert, decreasing the load of the first stage hence the entire decoder's latency.

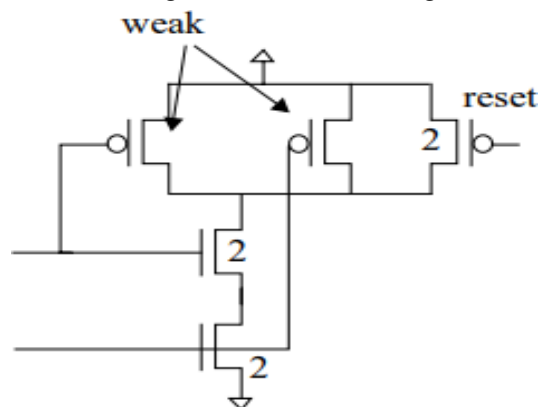


Fig.: 5 Skew NAND gate

In low-power design there are more other ways like a pulsed decoder can be used is by lowering the power of the bit-line path, which we'll go into momentarily. A multiplexor may be developed for reads from the SRAM data, and a demultiplexer for writes. In the simplest implementation, the multiplexor alone consists of two stages. At the next level, column pass transistors multiplex a restricted number of bit-lines initially. Extra metal layers may be utilised to divide a bit-line height into many levels of bit-line hierarchy when it is very high. Specifically, there are several methods to construct the multiplexor structure. Two potential designs for a 512-row block are shown in Figure 3.6. While the single-ended bit-line was chosen to save complexity in the picture, only the N-mosfet pass gates are depicted in the design for the genuine multiplexor. Complementary MOS pass gates for differential bit-lines would enable for reads and writes. A single sensing amplifier is multiplexed from two neighbouring columns of 512 cells in a single level multiplexor design, as shown in Figure.

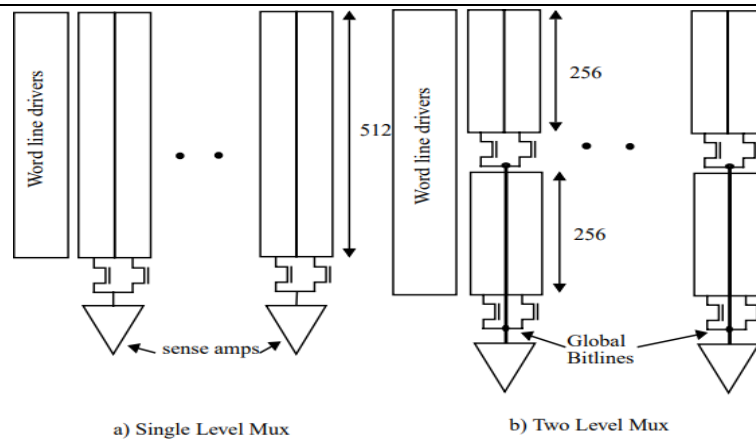


Fig:6 Bit-line MUX architecture

Figure shows a two-story building. The primary equal complexes the two 256high columns in the above two-level structure. The global bit lines that feed into the sense amplifier may be created by multiplexing the aforementioned output at the second level. Hierarchical MUX may also be implemented using the I/O lines, which are linked to the output of each other SA via the I/O ports. Because of its tiny size, a memory cell has a very low bit-line slew rate after the read is complete. Consequently, bit-line signals are amplified by SAs to enable the detection of signals as low as 100 mV. In a conventional design, after the sense amplifiers detect the bit-lines, they slew further, producing a sizable voltage differential. The bit-lines lose a significant amount of power as a consequence of their high capacitance. By limiting the word line pulse width, we can decrease pd by controlling the amount of charge that the bit-lines push down. In this theory, we suggest a method to control the word line pulse width, which must be just sufficiently broad. for the sensing amplifiers to continuously detect and prevent the bit-lines from sliding additional under a range of operational conditions.

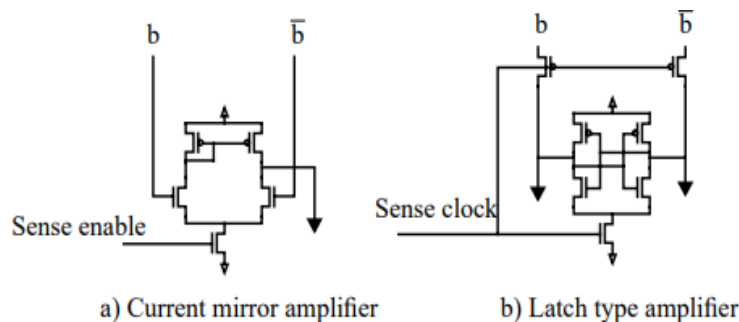


Fig.:7 Two different type of sense Amplifier a) Current mirror type b) Latch type

There have been a variety of SA circuits presented in the past, but generally they fall into two categories: linear amplifiers and latch amplifiers [10]. A simple prototype of each class is shown in Figure. The linear amplifier type requires a DC bias current is to set it up in high gain area before the bit line signal arrives. To convert the tiny swing bit line signal to a full swing Complementary MOS signal, many stages of amplification are required. High-end designs often use these amplifiers. Since low voltage and low power designs need biasing power and have a restricted supply voltage, they are not advised. In these designs, latch type designs are often used. Due to the extra timing margins, the sense clock timing is typically modified for a worst case operating as well as process conditions, which bring it down under, average situations. The outputs of SA's are connected to the I/O lines in big SRAMs, adding another level to data flow hierarchy. The signal is transported between Memory blocks and RAM I/O ports through the I/O lines. Because the power consumption of these lines can be severe in high access width SRAMs, signalling on these lines is also done via tiny swings. We'll use the low swing bit-line technique to lower the power of the I/O lines as well.

5. RESULT

In present era of microprocessor SRAM become the unavoidable part of it. But leakage current of the SRAM makes it more challenging. As size of the transistors are shrinking this leads to increase in leakage current. There is different type of leakage current are available (explained in previous chapter) which makes SRAM power hungry. Therefore, design engineer always tries to make SRAM with minimal power consumption. In this chapter simulation has been done. Different kind of SRAM circuit is simulated and the parameter like SNM and leakage current is calculated. In SRAM there are problem of power loss as discuss in previous chapter, like static, dynamic and short circuit power

loss. To reduce or minimize these losses there are different technique. Here through the simulation it is tried to minimize the static power. For that different stacked transistor method is used. In stacking technique NMOS and PMOS transistor is used. The schematic diagram of stacked transistor. For simulation purpose cadence virtuoso tool is preferred here. To analyse its stability butterfly curve is used.

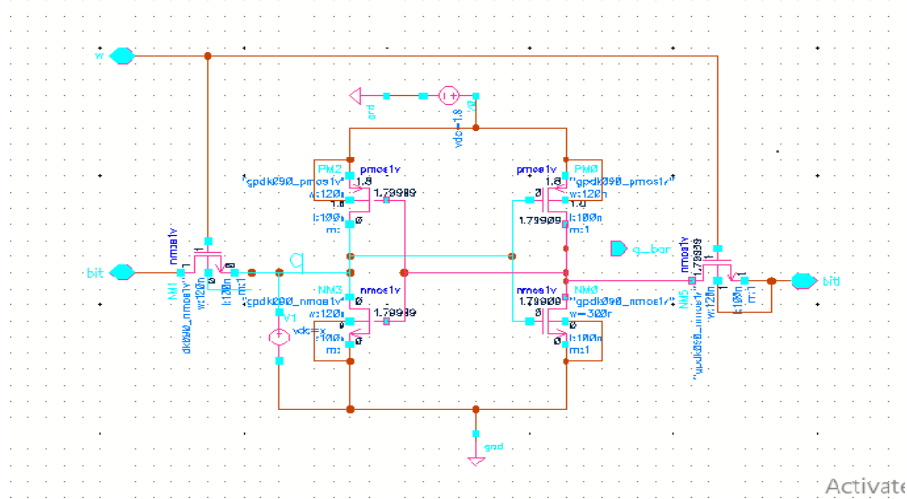


Fig.:8 Schematic of 6T SRAM

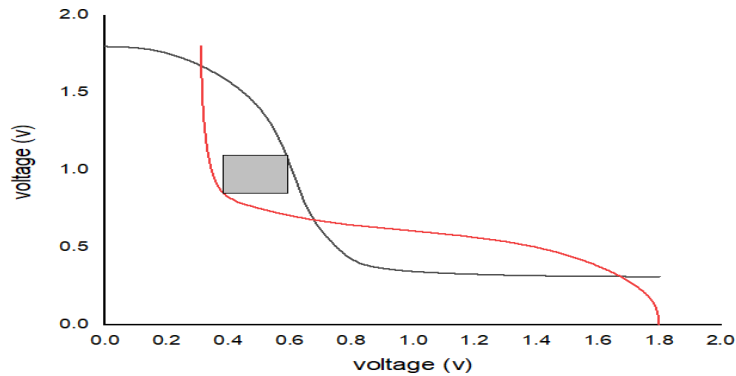


Fig.:9 SNM curve of 6T SRAM

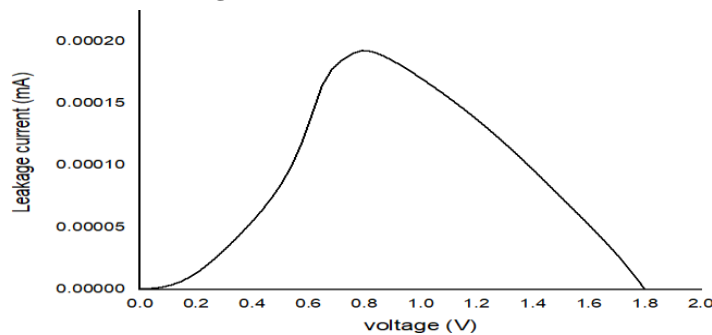


Fig.:10 leakage current of 6T SRAM

The above diagram is representing the conventional 6T SRAM. The transistor used here is based on 90nm technology.

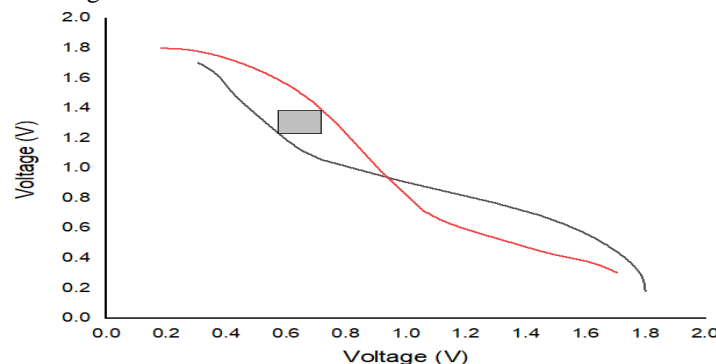


Fig.:12 Schematic of 2nd topology

6. CONCLUSION

A change has been made to the normal 6T SRAM cell in order to reduce the leakage current from the device. To lower the leakage power, more transistors are employed in this revision. The stacking approach refers to the connecting those more transistors in series. An additional transistor is employed in the stacking approach. As a result, the SRAM's total area increases. On the other hand, it lowers power loss by lowering leakage current. In the second topology, an additional transistor is used in an attempt to better understand SRAM stability and examine how designing an SRAM schematic affects the stability of a traditional 6T SRAM. Compared to the 6T SRAM, it has a higher static noise margin. There is less leakage current. Compared to a 6T SRAM cell, it lowers leakage current by 40.45%, and its performance is superior to that of a 6T SRAM with a different temperature. Again, two more NMOS transistors are added to the 6T SRAM. The storage cell and the ground are separated by this transistor. The performance of this stacking approach is likewise superior than that of the 6T SRAM. Improved SNM may be approaching. The second and third topologies' SNMs are almost identical. By reducing leakage current, it also reduces power loss. Comparing it with the 6T SRAM cell, the leakage current is reduced by 54.29%. Additionally, it performs better than a 6T SRAM cell at a different temperature. The 4th topology requires 10 transistors. It gives better noise margin as well as lesser leakage current than the 6T SRAM. It reduces the leakage current nearly 9 times than the 6T SRAM. The Noise Margin of 6th topology is far better than the 6T SRAM. So overall 6th topology is better than among all the above topologies, with one demerit it will take more area because of 10 transistors.

7. REFERENCES

- [1] Shokoufeh Naghizadeh1, · Mohammad Gholami2, “Two Novel Ultra-Low-Power SRAM Cells with Separate Read and Write Path” Springer Science Business Media, LLC, part of Springer Nature 2018.
- [2] Debasis Mukherjee, Hemanta Kr. Mondal, “Static Noise Margin Analysis of SRAM Cell For High Speed Application”, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.
- [3] Jan M. Rabaey, Anantha Chandrakasan, Borivoje Nikolic “Digital Integrated Circuits”, Pearson Education Electronics, 2003.
- [4] Sung-Mo Kang and Yusuf Leblebici “CMOS Digital Integrated Circuits”, TATA McGRW-HILL EDITION 2003.
- [5] Kunihiko Yamaguchi, Hiroaki Nambu, et. al. —A 1.5-ns Access Time, 78-pm² Memory-Cell Size, 64-kb ECL-CMOS SRAM, IEEE Journal of Solid-State Circuits, Vol. 27, No.2. February 1992.
- [6] Rakesh Dayaramji Chandankhede —Design of High Speed Sense Amplifier for SRAM, 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).
- [7] Sreerama Reddy G M, P Chnadraseskhara Reddy, “Design and implementation of 8Kbits Low Power SRAM in 180nm Technology”, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol II IMECS 2009, March 18-20, 2009.
- [8] Harekrishna Kumar1 · V. K. Tomar, “A Review on Performance Evaluation of Different Low Power SRAM Cells in Nano-Scale Era” Springer Science Business Media, LLC, part of Springer Nature 2020.
- [9] Pavankumar Bikki, Pitchai Karuppanan, “SRAM Cell Leakage Control Techniques for Ultra Low Power Application: A Survey “scientific research publication, Motilal Nehru National Institute of Technology, Allahabad, India, February 2017.
- [10] K. Sasaki, et. al., “A 15-ns 1-Mbit CMOS SRAM”, IEEE Journal of Solid State Circuits, vol. 23, no. 5, pp. 1067-1071, October 1988.
- [11] M. Matsumiya, et. al., “A 15-ns 16-Mb CMOS SRAM with interdigitated bit-line architecture”, IEEE Journal of Solid State Circuits, vol. 27, no. 11, pp. 1497-1502, November 1992.
- [12] Anantha P. Chandrakasan and Jan M. Rabaey, “Digital Integrated Circuits: A Design Perspective”.
- [13] M. Yoshimoto, et. al., “A 64kb CMOS RAM with divided word line structure”, 1983 IEEE International Solid State Circuits Conference, Digest of Technical Papers, pp. 58-59.
- [14] T. Hirose, et. al., “A 20nS 4Mb CMOS SRAM with Hierarchical Word Decoding Architecture”, 1990 IEEE International Solid State Circuits Conference, Digest of Technical Papers, pp. 132-133.