Low Power Single-Bit Cache Memory Architecture

Reeya Agrawal, Neetu Faujdar, Aditi Saxena
agrawalreeya98@gmail.com, neetu.faujdar@gmail.com, aditeesaxenaa29@gmail.com
Department of Electronics and Communication, GLA University, Mathura

Abstract- A quantitative and yield analysis of single bit cache memory architecture has been analyzed. A single bit cache memory architecture is made up of a write driver circuit, SRAM cell, and sense amplifier. Apart from it, the power reduction technique has been applied over different blocks of single bit cache memory architecture such as sense amplifier and SRAM cell, to optimize the power consumption of the circuit. To check the robustness of the circuit monte Carlo simulation and process corner simulation also have been done. The conclusion arises that Single bit cache memory architecture having VMSA with forced stack technique over SRAM in an architecture consumes the lowest power (9.18 µW).

Keywords: Single Bit SRAM VMSA Architecture (SBSVMSA), write driver circuit (WDC), Voltage Mode Sense Amplifier (VMSA), sense amplifier (SA), static random access memory (SRAM).

1. Introduction

In recent years, the exponential development of battery-operated equipment has made low power memory architecture a market impetus. As the transistor count rises, the leakage current has made the SRAM unit a power-hungry block from both static and dynamic perspectives. In SOC design today, the SRAM block is also a significant component. The main variables in the design of the memory here are power dissipation and area. Memory density scaling in memory design must continue to track logic trends in scaling. In the 90nm to 45nm CMOS technology here the experimental results are implemented in the proposed SRAM architecture, confirming the SRAM design principle.

1.1. Power Reduction Techniques

Power reduction techniques are applied over circuits to reduce the power consumption of circuits with no effect on performance, speed, and other parameters.

1.1.1 Sleep Transistor Technique

The state-destructive technique disrupts both P_MOS and N_MOS transistors from sleep transistors to supply voltage or ground. These methods are called VDD and gated-GND. These are technological varieties. When the logic circuits are in standby mode, the sleep transistor is disabled. By uninflecting sleep transistor operation from the logical networks, the technique of sleep semiconductor significantly reduces sleep power as shown in fig: 1(a) [3].
1.1.2 Forced Stack Technique

Fig: 1(b) demonstrates a forced stack technique. This second technique decreases the power by stacking transistors. When two or more transistors are uniformly switched OFF, the effect of stacking the semiconductor device reduces the sub-threshold leakage current [4].

1.1.3 Dual Sleep Technique

Both types of transistors are used in this technique: two PMOS (PM0 and PM1) and two NMOS (NM0 and NM1). Both PMOS and NMOS transistors are used in the header and footer. One transistor is ON in active mode, and another transistor is switched ON in OFF state mode. PMOS and NMOS are both used in standby mode to reduce power, as shown in fig: 1(c) [5].

Fig: 1 a) Sleep Transistor Technique, b) Forced Stack Technique, c) Dual Sleep Technique

Section 1 provides a brief presentation on the semiconductor industry and a brief introduction to power reduction techniques. Each block circuit and the operating process have been discussed in Section 2, single bit cache memory architecture. The proposed design has been discussed in Section 3 with output waveforms and a comparison table. The conclusions are present in Section 4. The goal of this paper is to design low power SBVMA in customary gpdk 45nm technology using Cadence Tool.

2. Single-Bit Memory Architecture

WDC, SRAMC, and VMSA as shown in fig: 2 [12, 13] are the single bit cache memory architectural blocks.

Fig: 2 Schematic of Single Bit Cache Memory Architecture having VMSA
The description is divided into three different parts: a) the WDC with two input pins (word enable (WE) and Bit) and two output pins (BL and BL\textsubscript{BAR}), b) the SRAMC, which is attached to the WDC via bit lines (i.e. BL and BL\textsubscript{BAR}), and an input pin (word line (WL)) and two output pins (V\textsubscript{1} and V\textsubscript{2}) and connected through bit lines having capacitance and resistance as a connector between them, c) VMSA which has five input pins (Y\textsubscript{sel}, BL, BL\textsubscript{BAR}, PCH, and SA\textsubscript{EN}) and two output pins (V\textsubscript{3} and V\textsubscript{4}).

2.1 Write Driver Circuit

The write driver is responsible for discharging the bit lines quickly to a level below the write margin cells before or when the word lines of the selected cell are involved. Two regular writing drivers are shown in Fig.3. The input of the data chooses which bit line is discharged. The WE signal is turned on just when the writing procedure is expected. The WE, otherwise, isolates the bit lines from the printed drivers. It is faster since it has less stacked transistors in the direction of discharge at the expense of complexity. Usually, the write operation is not a speed limiting transaction, so the write driver is selected for simpler configurations that relax the layout requirements. It consists of CMOS inverter and NMOS pass transistors and can be scaled because it does not depend on the number of transistors to form a column. [14].

![Fig: 3 Write Driver Circuit Schematic](image)

2.2 Conventional SRAM

It is used for operations at low power, low voltage. Here each bit is stored using bistable latching circuitry. In Fig.4, the pulls up transistors are PM6 and PM7 PMOS, while the driver transistors are NM6 and NM7 NMOS. Word line connected to NM8 and NM9 pass transistors, which include BL and BL\textsubscript{B} lines that are used for both reading and writing purposes, allows entry. These bit lines enhance the margin of noise. The value of measurable output voltage swings is given by differential circuitry. Logic 0 or 1 is stored as long as the power is on, but unlike DRAM cells, it does not need to be refreshed [4]. In SRAM design, the size of the transistors is most important for the correct operation of the transistors. The NMOS width of 0.09um is here, starting with inverters, and because the NMOS is rough twice as conductive as the PMOS. The gateway transistors, on the other hand, must be 2 or 3 times larger than the inverter NMOS transistors, so that they are sufficiently conductive to adjust the amount of logic stored by these transistors. They should not be much larger, however, so that they won’t take much space.

![Fig: 4 SRAM Cell Schematic](image)
2.3 Sense Amplifier
A small analog differential voltage produced on the bit lines in read access is amplified by the sense amplifier. The amplification results in a single-ended digital production with a full swing. SA usage decreases the size of the SRAM cell as the bit lines do not need to be completely discharged by the drive transistors. The slowest operation that gives the cell delay is usually read operations. Bit lines have greater capacitance due to the length of the metal and the number of transistors that take more time to discharge the bit lines. The difficult choice for sense amplifiers is the timing control and load capacitor selection here. The SA signal is used for that. Here, two cross-coupled inverters are used in the latch type sense amplifier, which provides the differential output that amplifies the signal and provides complete swings and two capacitors that store the bit line stage.

2.3.1 Voltage Mode Sense Amplifier
All elements needed for differential sensing are contained in the basic MOS differential voltage sense amplifier circuit. A differential amplifier requires a single-ended, small-signal output. The efficacy of a differential amplifier is characterized by its ability to amplify the true difference between the signals by ignoring common noise. A simple differential voltage amplifier is not used in memories because of the very slow operating speed provided at substantial power dissipation and inherently high offset, as shown in fig: 5 [19, 20].

![Fig: 5 Voltage Mode Sense Amplifier Schematic](image)

3. Analysis of Result
In this section output of all the circuits has been shown and described. Process corner simulation and Monte Carlo simulation are also shown and total power consumption of single bit cache memory is compared with changes in the value of resistance (R) and sees the effect of R on the power consumption of the circuit.

![Fig: 6 output waveform of WDC](image)

Fig: 6 describe the output waveform of WDC, for cases arise: a) when Bit=0V and WE=0V BL=V_DD and BL_BAR=V_DD, b) Bit=0V WE=V_DD so, BL=0V and BL_BAR= V_DD/2, c) Bit= V_DD WE=0V so, BL=0V and BL_BAR= V_DD/2 and d) Bit= V_DD WE= V_DD so, BL= V_DD and BL_BAR= 0V.
Fig: 7 describes the both write operation and hold operation of the SRAM cell. There is a pull of the network (PM6 and PM7), pull-down network (NM6 and NM7), and access transistor (NM8 and NM9) which allows data to store and sense amplifier to read the data.

![Fig: 7 Output Waveform of SRAM Cell](image)

Fig: 8 describes the read operation of VMSA, when both SA\textsubscript{EN} and WL are pulled high, during that time only the sense amplifier senses the data from the SRAM cell at bit lines and gives output at V\textsubscript{3} and V\textsubscript{4}.

![Fig: 8 Output Waveform of VMSA](image)

![Fig: 9 Process Corner Simulation](image)
Simulation has been done for six corners i.e.

a) \( V_3 \) (Corner=C4) is SS i.e. both \( N_{MOS} \) and \( P_{MOS} \) are slow.

b) \( V_3 \) (Corner=C3) is SF i.e. \( N_{MOS} \) is slow and \( P_{MOS} \) is fast.

c) \( V_3 \) (Corner=C2) is NN i.e. both \( N_{MOS} \) and \( P_{MOS} \) have normal speed.

d) \( V_3 \) (Corner=C1) is FS i.e. \( N_{MOS} \) is fast and \( P_{MOS} \) is slow.

e) \( V_3 \) (Corner=C0) is FF i.e. both \( N_{MOS} \) and \( P_{MOS} \) are fast.

f) \( V_3 \) (Corner=nom) is stat i.e. both \( N_{MOS} \) and \( P_{MOS} \) are in stat condition.

From fig:9 the conclusion arises that this circuit is not suitable for corner \( V_3 \) (Corner=C0) is FF i.e. both \( N_{MOS} \) and \( P_{MOS} \) are fast.

Fig: 10 shows the Monte Carlo simulation for the VTH_SAEN to observe the robustness of the circuit.

![Monte Carlo Simulation](image)

**Fig: 10 Monte Carlo Simulation**

Table: 1 describes that as increasing in value of resistance power consumption decreases as because resistance as a path stopper for current in a circuit and no effect on the area, performance, and speed.

**Table: 1 Different Parameter of SBSVMSA**

| S.No. | Parameters | Power Consumption | No. of Transistors | Sensing Delay |
|-------|------------|-------------------|--------------------|---------------|
| 1.    | \( R=42.3\Omega \) | 13.16 \( \mu W \) | 30                 | 13.51 \( \eta s \) |
| 2.    | \( R=42.3K\Omega \) | 11.34 \( \mu W \) | 30                 | 13.51 \( \eta s \) |

Table: 2 describes that as increasing in value of resistance power consumption decreases as because resistance as a path stopper for current in a circuit and no effect on the area, performance, and speed.

**Table: 2 Different Parameters of SBSVMSA on applying different Power Reduction Techniques over VMSA**

| S.No. | Architecture | SBSVMSA | No. of Transistors |
|-------|--------------|---------|--------------------|
|       | Techniques   | Power Consumption | Sensing Delay |
| 1.    | Sleep Transistor Technique | 11.29 \( \mu W \) | 13.51 \( \eta s \) | 32 |
| 2.    | Forced Stack Technique | 11.29 \( \mu W \) | 13.70 \( \eta s \) | 32 |
| 3.    | Dual Sleep Technique | 11.03 \( \mu W \) | 13.75 \( \eta s \) | 34 |

Table: 3 describes that applying the power reduction technique over VMSA dual sleep technique reduced power consumption i.e. 11.03 \( \mu W \) with 34 number of the transistor. There is a tradeoff between power consumption and area as on decreasing power consumption, area increases (in terms of transistors).
Table: 3 Power Consumption of SBSVMSA on applying different Power Reduction Techniques over SRAM with VMSA

| S.No. | Techniques          | Architecture         | Power Consumption | Sensing Delay | No. of Transistors |
|-------|---------------------|----------------------|-------------------|---------------|-------------------|
| 1.    | Sleep Transistor Technique |                     | 9.18 µW          | 13.61 ns      | 32                |
| 2.    | Forced Stack Technique |                     | 9.108 µW         | 13.67 ns      | 32                |
| 3.    | Dual Sleep Technique  |                     | 10.13 µW         | 13.02 ns      | 34                |

4. CONCLUSION

The conclusion arises that on applying power reduction techniques such as sleep transistor technique, footer stack technique and dual sleep technique over different blocks of single bit cache memory architecture such as SRAM and Sense Amplifier, an architecture having SRAM with Footer Stack Technique with VMSA consumes the lowest power (9.18 µW). To check the robustness of the circuit monte Carlo simulation and process corner simulation also have been done.

References

[1] Y. He, J. Zhang, X. Wu, X. Si, S. Zhen, and B. Zhang, “A Half-Select Disturb-Free 1T SRAM Cell With Built-In Write/Read-Assist Scheme for Ultralow-Voltage Operations,” in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 10, pp. 2344-2353, Oct. 2019, doi: 10.1109/TVLSI.2019.2919104.

[2] R. Fragasse et al., “Analysis of SRAM Enhancements Through Sense Amplifier Capacitive Offset Correction and Replica Self-Timing,” in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 6, pp. 2037-2050, June 2019, doi: 10.1109/TCSI.2019.2902102.

[3] Tripathi Tripti, Chauhan D. S., Singh S. K., and Singh S. V. “Implementation of Low-Power 6T SRAM Cell Using MTCMOS Technique”, In Advances in Computer and Computational Sciences, Springer, Singapore, 2017.

[4] M.Geetha Priya, Dr.K.Baskaran, D.Krishnaveni. “Leakage Power Reduction Techniques in Deep Submicron Technologies for VLSI Applications.” ELSEVIER, International Conference on Communication Technology and System Design 2011.

[5] K Sridhara, G S Biradar, Raju Ynamshetti, "Subthreshold leakage power reduction in VLSI circuits: A survey”, Communication and Signal Processing (ICCSP) 2016 International Conference on, pp. 1120-1124, 2016.

[6] S. Gupta, K. Gupta, B. H. Calhoun, and N. Pandey, "Low-Power Near-Threshold 10F SRAM Bit Cells With Enhanced Data-Independent Read Port Leakage for Array Augmentation in 32-nm CMOS," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 3, pp. 978-988, March 2019, doi: 10.1109/TCSI.2018.2876785.

[7] H. Dounavi, Y. Sifkas, and Y. Tsiatouhas, "Periodic Aging Monitoring in SRAM Sense Amplifiers,” 2018 IEEE 24th International Symposium on On-Line Testing And Robust System Design (IOLTS), Platja d’Aro, 2018, pp. 12-16, doi: 10.1109/IOLTS.2018.8474169

[8] S. Ahmad, B. Iqbal, N. Alam, and M. Hasan, "Low Leakage Fully Half-Select-Free Robust SRAM Cells With BTI Reliability Analysis," in IEEE Transactions on Device and Materials Reliability, vol. 18, no. 3, pp. 337-349, Sept. 2018, doi: 10.1109/TDMR.2018.2839612

[9] B. N. K. Reddy, K. Sarangam, T. Veeraiah, and R. Cherukuv, “SRAM cell with better read and write stability with Minimum area,” TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 2164-2167, doi: 10.1109/TENCON.2019.8929593.

[10] A. Surkak and V. Agarwal, "Delay and Power Analysis of Current and Voltage Sense Amplifiers for SRAM at 180nm Technology," 2019 3rd International Conference on Electronics, Communication, and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 1371-1376, doi: 10.1109/IECTA.2019.8822122.

[11] Gomes Iuri A.C., Meinhardt Cristina, Butzen Paulo F. “Design of 16nm SRAM Architecture” South Symposium on Microelectronics, 2012.

[12] Chakka Sri Harsha Kaushik, Rajiv Reddy Vanjarlapati, Varada Murali Krishna, Tadavarthi Gautam, V Elamaran, “VLSI design of low power SRAM architectures for FPGAs”, Green Computing Communication and Electrical Engineering (ICGCCCE) 2014 International Conference on, pp. 1-4, 2014.

[13] Richa Choudhary, Srinivasa Padhy, Nirmlal Kumar Rout, “Enhanced Robust Architecture of Single Bit SRAM Cell using Drowsy Cache and Super cut-off CMOS Concept”, International Journal of Industrial Electronics and Electrical Engineering, Volume-3, PP.63-68, July 2011.

[14] Jesal P. Gajjar, Aesha S. Zala, Sandeep K. Aggarwal, “Design and Analysis of 32 bit SRAM architecture in 90nm CMOS Technology” Volume: 03, Issue: 04, Apr-2016, pp:2729-2733.

[15] Shikha Saun, Hemant Kumar, “Design and performance analysis of 6T SRAM cell on different CMOS technologies with stability characterization”, OP Conf. Series: Materials Science and Engineering561 (2019) 012093.

[16] A. Bhaskar, “Design and analysis of low power SRAM cells,” 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, 2017, pp. 1-5, doi: 10.1109/IPACT.2017.8244888.
[17] Yong-peng Tao, Wei-ping Hu, “Design of Sense Amplifier in the High-Speed SRAM”, International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 384-387, 2015.

[18] M. Jefremow et al., "Time-differential sense amplifier for sub-80mV bit line voltage embedded STT-MRAM in 40nm CMOS," 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, 2013, pp. 216-217, doi: 10.1109/ISSCC.2013.6487706.

[19] Y. Tao and W. Hu, "Design of Sense Amplifier in the High-Speed SRAM," 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Xi’an, 2015, pp. 384-387, doi: 10.1109/CyberC.2015.32.

[20] Manoj Sinha, Steven Hsu, Atila Alvandpour, Wayne Burleson, Ram Krishnamurthy, Shekhar Borhr. “High-Performance and Low-Voltage Sense-Amplifier Techniques for sub-90nm SRAM .” SOC Conference, 2003. Proceedings. IEEE International [Systems-on-Chip].