Noise Margin and Delay Analysis of Half Stacked and Full Stacked SRAM Cell Design

Balwinder Raj,
Member IEEE
Asst. Prof., Department of ECE, National Institute of Technology Jalandhar
(NIT Jalandhar) P.B-144011, India
{balwinderraj@gmail.com}

Abstract - In this paper we propose a half stacked and full stacked SRAM cell design for low power application. This based on the "Stacking Effect of Transistors" with stacking of the driver and the load transistors to reduce the total power consumed in the SRAM cell. The results obtained on basis of proposed half stack and full stack SRAM cell are compared and contrasted with the conventional SRAM cell with sleep transistor and normal mode transistor. The proposed full stack cell gives a 30% power reduction in the standby mode. In addition to these, the proposed cell has a superior Static Noise Margin (SNM) of 380mV at a supply voltage of 1.1V. The significant improvements in the results obtained validate our approach for the proposed stacked SRAM cell design for low power memories design.

1. INTRODUCTION

As memory begins to dominate the area of chip in high performance applications, SRAM has become the focus of aggressive scaling. But cell miniaturization also introduces several new challenges in Very Large Scale Integrated (VLSI) Circuits such as sensitivity to process variations and increasing transistor leakage. The cell stability is further degraded by scaling of supply voltage. The focus of the SRAM cell design in sub-90nm technologies is to achieve a balanced cell design with optimized device sizing while choosing right cell topology to ease the difficulty of fabrication process to reduce defect density. However there is no universal way to avoid tradeoffs between power, delay and area and thus designers are required to choose appropriate techniques that satisfy application and product needs.

In CMOS based devices the total power dissipation is the sum of active and standby power components. In many event driven applications, like a processor running an X-server, circuits spend most of their time in an idle state where no computation is being performed. Minimizing standby power (leakage power) consumption can be especially important in mobile devices where leakage drains the battery when the circuit is idle for a long time. Thus we focus on the reduction of the leakage power for ultra low-power applications. The need for low power in digital devices is responsible for the scaling of the supply voltage.

This requires scaling down of threshold voltage which degrades the SRAM cell stability by increasing the leakage. The leakage power is composed of subthreshold leakage, gate oxide tunneling leakage, band-to-band-tunneling leakage (BTBT) etc. The BTBT leakage is very small as compared to the subthreshold and gate leakage around the 90nm node [1]. Aggressive scaling down of the threshold voltage leads to exponential increase in subthreshold leakage due to decrease in threshold voltage [2]. Thus at sub-100nm technology node subthreshold leakage will become the major contributor to standby power consumption [3]. Various techniques like gated supply voltage (V_{DD}) technique [4], gated ground technique [5] [6], dual threshold voltage (V_T) [7] [8], multi V_T [9] etc. have been implemented for the reducing the standby leakage power.

In addition to the standby power consumption, access time and the cell stability are important parameters of consideration during the design of SRAM cell. Earlier attempts achieved low power consumption at the cost of time. However with advancement in technology the need for faster devices is on the rise. So it is very important to ensure that the cell design does not increase the access time drastically. The cell stability represented by the noise margin is another important criterion during the design of the cell. Higher noise margin ensures that the data is secure and probability of accidental changes is kept to minimum.

In this paper we propose a new SRAM design on 90nm node based on the 'stacking effect of transistors' having ultra low-power consumption, improved static noise margins with an acceptable delay. The stacking effect is based on the fact that there is a large reduction in the leakage current when more than one transistor in NMOS or PMOS stacks are turned off simultaneously. Also an alternate design of stacking of only driver transistors (half stacking) has been proposed. The half stacking reduces the area overhead of the cell without much compromise in the performance. The stacking effect has been previously used to reduce the leakage in gates where there are already transistor stacks present [3] [10] [11] [12]. However no application of stacking effect on SRAM has been reported to the best of our knowledge. The power consumption, noise margin and access time of the stacked cell have been compared with that of conventional cell and extensive simulations have been carried out on T-SPICE.

The main contributions of this paper is as follows: 1) a new architecture for the SRAM cell called stacked cell (full and half) is introduced that reduces leakage power considerably 2) comparison of various noise margins (static, read, write) of full stacked and half stacked cell with that of conventional SRAM cell 3) comparison of delay and power of half stacked and full stacked cell with that of conventional cell, with and without a sleep transistor (gated ground technique).

The remainder of the paper is organized in the following way. In Section II the stacking effect of transistors is explained. In Section III the designs of the proposed stacked memory cells are discussed. In Section IV the various simulation results are presented and analyzed. Finally conclusions are given in Section V.

II. STACKING EFFECT

According to ITRS road map in 2001, memory will occupy about 90%of the chip area in 2013 [13]. In such a system the leakage current of an embedded SRAM dominates the standby current. Thus, the reduction of standby current is the most important to achieve low power consumption. Leakage currents in NMOS or PMOS transistors depend exponentially on the voltage at four terminals of transistor.

Increasing the source voltage of NMOS transistor reduces the sub threshold leakage current exponentially due to negative V_{GS}, lowered signal rail (V_{CC}-V_S), reduced Drain Induced Barrier Lowering (DIBL) and body effect. This effect is also called self-reverse biasing of transistor. The self-reverse biasing effect can be achieved by turning off a stack of transistors. Turning off more than one transistor raises the internal voltage (source voltage) of the stack which acts as reverse biasing the source [14].Thus

© 2013. The authors - Published by Atlantis Press

maximizing the number of off transistors by stacking and applying proper input vectors can reduce the standby leakage of a functional block.

Stacking principle had been implemented to reduce the leakage power in gates and logic circuits [3] [10] [11] [12].The leakage current flowing through transistors connected in series depends upon the number of 'off' transistors in the stack. Turning off the stacked transistors raises the intermediate voltage to a positive value due to a small drain current.

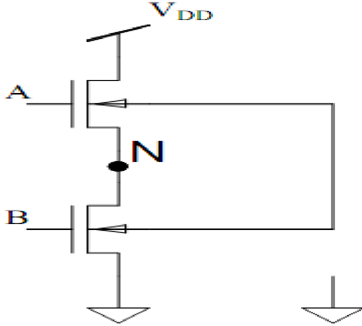

Figure 1. Fig1: Two transistor NMOS stack

Consider the Fig. 1. It consists of a stack of two NMOS transistors. When an input vector "00" is applied to the gate both the transistors are turned off. As discussed earlier a small positive potential develops at the node N. This potential developed at the intermediate node has the following effects:

1) The gate-to-source junction becomes reverse biased since V_{GS} is negative. As the subthreshold current is exponentially proportional to V_{GS}, it is also reduced.

2) There is an increased body effect in the top transistor due to a negative body- to-source potential and V_T is increased. Since the subthreshold current is exponentially proportional to V_T also, it is reduced.

3) Drain-to-source potential of top transistor decreases due to increase in source potential. This results in lesser Drain-Induced Barrier Lowering (DIBL). As a result the subthreshold leakage is further reduced.

This phenomenon is called stacking effect.

In [12] the effect of stacking on leakage current was extensively discussed. It was shown that the power consumption depends upon the input vectors applied to the gates. At the 90nm node applying "10" vector at the gate, reduces the gate leakage current and applying "00" vector input to a two transistors stack only reduces subthreshold leakage and does not change the gate leakage component. However the total standby power was found to be minimum when the input vector is "00" as the subthreshold leakage is dominating at the 90nm node[1][12]. This fact is exploited to reduce the standby power consumption of SRAM. However this reduction is achieved at the expense of delay penalty as the effective width of the transistor becomes W/N^2 (where W is width of the transistor before stacking and N is the number of transistors after stacking) after stack forcing. It is similar to replacing a low-V_T device with a high-V_T device in a multiple-V_T design. [9].

III. STACKED CELL DESIGN

A basic SRAM cell consists of two cross-coupled CMOS inverters with NMOS pass transistors to access the cell. Separate pull up and pull down circuitry are provided for READ and WRITE operations. In the stable state one of the NMOS transistors and one of the PMOS transistors is 'off' and the other two are 'on' as shown in the Fig.2. As discussed above the leakage of the 'off' transistors can be reduced by stacking the transistors.

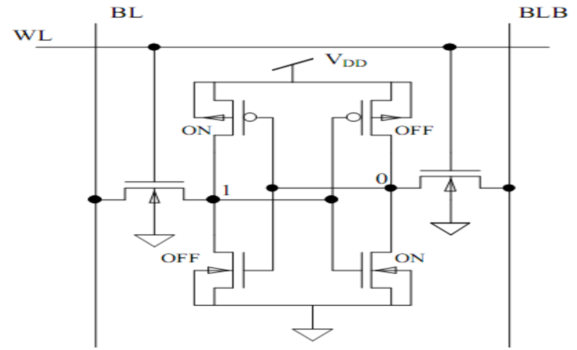

Figure 2 Conventional SRAM cell

In the proposed full stacked SRAM design each of the NMOS transistors and the PMOS transistors are replaced by a stack of two transistors as shown in the Fig.3. The width of each stacked NMOS (PMOS) transistor is half of that of the NMOS (PMOS) transistor in the conventional cell. This ensures that there would be negligible overhead in area of the cell [10]. The input vector for the stack of 'off' transistors is '00' which results in substantial reduction in total standby power of the cell. The stacked cell is more robust due to increased V_T of the transistors. However the stacked cell design has increased delay with reduced read and write noise margins.

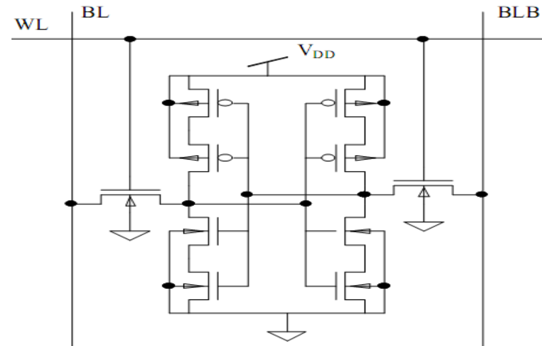

Figure 3 Full Stack SRAM cell

The increase in delay is due to addition of extra transistors in the critical path and the reduction in width of the transistors that results in higher V_T. Higher the threshold of the PMOS transistors, lesser the inverter trip point. But in case of NMOS higher the threshold, higher is the inverter trip point [15]. Also the current driving capability of NMOS transistor is more than that of the PMOS transistor. Hence stacking of NMOS transistors tends to have more impact on the trip voltage of the inverter as compared to the stacking of PMOS transistors. Hence stacking of NMOS transistors only is expected to have higher read and write noise margins along with reduced power consumption.

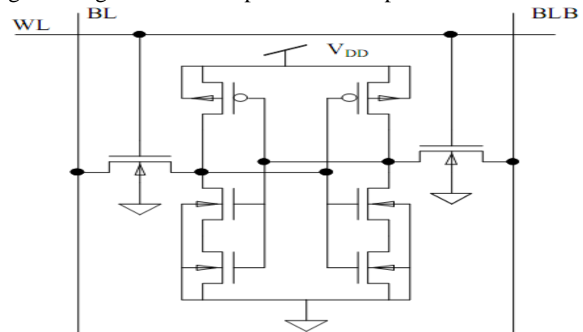

Figure 4 Half Stack SRAM cell

Motivated by this fact we also propose an alternate design called Half Stacked Cell by 'stacking only the driver transistors' as

shown in the Fig.4. This architecture has advantages of lesser area overhead, reduced delay and improved read and write noise margins as compared to the Full Stacked cell.

In the following section the proposed architectures are simulated on T-SPICE to measure power, noise margins and the delay of the sense amplifier. The obtained results are compared with that of the conventional cell with and without sleep transistor.

IV. SIMULATION RESULTS NOISE MARGIN

Noise margin is the maximum amount of voltage noise that can be introduced at the outputs of the two invertors such that the cell retains its data. In the standby mode this is referred to as Static Noise Margin (SNM), during the read operation as Read Noise Margin(RNM) and during write operation as Write Noise Margin (WNM) [16][17][18].

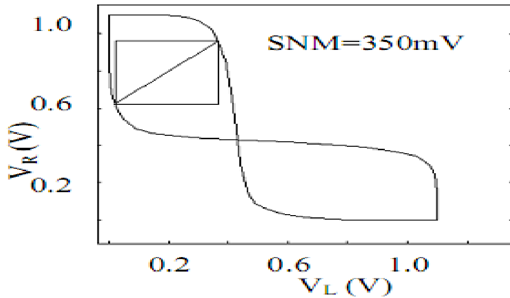

Figure 5.a SNM curves for conventional cell

Table I shows the noise margin data for different cell configurations and corresponding butterfly curves are shown in figure 8. There is an improvement of about 30 mV in SNM value of full stack (379mV) as compared to the conventional cell (350.7mV). However earlier works have reported an SNM value of 300mV at 130nm at V_{DD} of 1.5V [19], 200mV at 130nm at 1.2 V_{DD} [20], 110mV at 50nm at V_{DD} of 1V [21]. The best SNM is however observed in case of half stack (384.7mV). This is so because stacking of NMOS transistor increases the inverter trip point whereas stacking of PMOS transistor reduces it. As a result the increase in inverter trip point in case of half stack cell is more than that of full stack. This accounts for higher SNM in case of half stacked cell as compared to full stack.

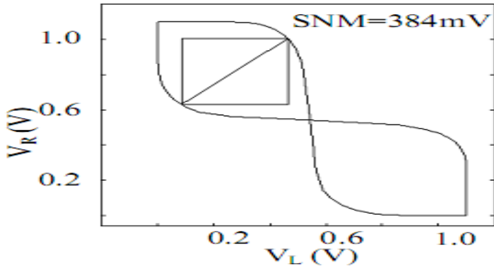

Figure 6.a SNM curves for half stacked cell

Table I Noise Margin values for different cell configurations

	Conventional cell	Half stacked cell	Full stacked cell
SNM(mV)	350.7	384.7	379
RNM(mV)	265.9	192.3	84.9
WNM(mV)	265.9	192.3	90.5

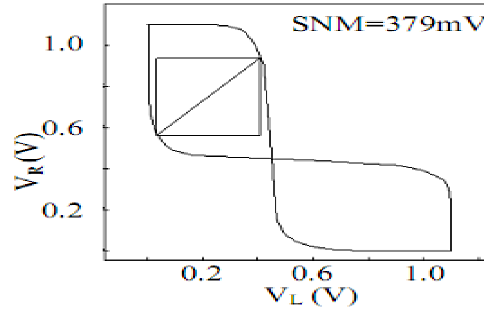

Figure 7.a SNM curves for full stacked cell

DELAY:

The access time for the SRAM consists of the sum of the delay times of the four circuit stages-address buffers, decoders, memory cell arrays and sense amplifier circuits. However the major contributors to the access time are the decoder and the sense amplifier circuits [22].In this section we discuss the variation of delay of the sense amplifier for the different cell designs.

Fig.9 shows the schematic of the differential sense amplifier circuit used. Since the amplifier is differential, it can be directly employed in SRAMs where the SRAM cell utilizes both the BL and BLB**.** The delay was measured when the voltage levels had reached 90% of their steady state values of 1.1V and 0V. Since the number of transistor in the critical path are more in case of the stacked cells there is increase in the delay of the sense amplifier as shown in the Table II. In the table, variation of delay with respect to bitline capacitance is shown. The table shows that the variation in delay with increase in capacitance in case of stacked cell is less as compared to that of conventional cell. So addition of more number of cells in the same column would incur less delay penalty as that of the conventional cell.

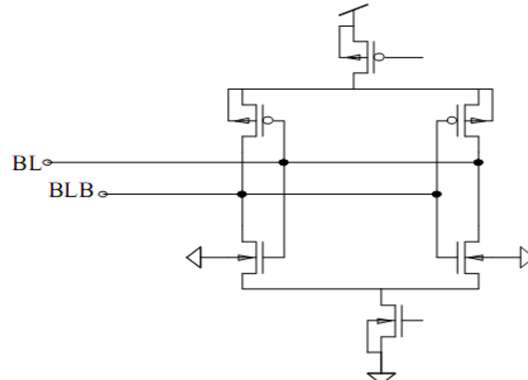

Figure 8: Schematic of Sense amplifier circuit

Table II Power consumed by different cell configurations

	Conventional cell	Conventional cell with sleep transistor	Half stacked cell	Full stacked cell
Idle power(nW)	3.491	2.949	2.662	2.46
Write power(μW)	16.761	16.344	16.3	16.21
Read power(μW)	14.509	13.881	12.5	11.764

The optimized cell was designed keeping the tradeoffs between power, delay and noise margin in mind. Table 3 shows the

variation in power consumed for different optimized cell configurations. The power consumed was found to decrease with decrease in width of pass transistors. Hence the width of pass transistors was kept at its minimum value of 100nm. The width of the driver transistor was kept at 150nm and that of the load transistors was kept at 100nm for each of the stacked transistor to achieve optimized operation in terms of power and noise margin.

With further scaling of the supply voltage there was an appreciable decrease in both active and idle power. However this was achieved at the expense of a delay penalty of about 10%. Also the noise margins were also reduced and the performance of the cell was severely degraded.

CONCLUSION

Two new SRAM cell design have been proposed for ultra low power applications. The proposed designs have enhanced performance in terms of power and SNM. The full stacked design shows a reduction of 30% in the standby power. In addition an enhanced static noise margin of 380mV was obtained at a supply voltage of 1.1V. Also the proposed design is less susceptible to process variation. This ensures that minor errors in the fabrication process do not affect the cell performance drastically. In addition, the half stacked cell offers better noise margins with lesser area overhead as compared to that of the full stacked cell. Thus the proposed cell architectures offer superior performance as compared to earlier cell designs and can be implemented for ultra low-power applications.

REFERENCES:

[1] A. Agarwal, S. Mukhopadhyay, C.H .Kim, A. Raychowdhury and K. Roy, "Leakage power analysis and reduction: models, estimation and tools", IEE Proc.-Comput. Digit. Tech., Vol. 152, No. 3, May 2005, pp. 353-368.

[2] Shengqi Yang, Wayne Wolf, Wenping Wang, N. Vijaykrishnan and Yuan Xie, "Low-leakage Robust SRAM cell design for Sub-100nm Technologies", ASP-DAC 2005, pp.539-543.

[3] Siva Narendra, Shekhar Borkar, Vivek De, Dimitri Antoniadis, and Anantha Chandrakasan, "Scaling of Stack Effect and its Application for Leakage Reduction", ISLPED '01, August 6-7, 2001, pp. 195-200.

[4] Michael Powell, Se-Hyun Yang, Babak Falsafi, Kaushik Roy, and T.N. Vijaykumar, "Gated-V_{dd}: A Circuit Technique to Reduce Leakage in Deep-Submicron Cache Memories", ISLPED '00, pp.90-95.

[5] Amit Agarwal, and Kaushik Roy, "A Noise Tolerant Cache Design to Reduce Gate and Sub-threshold Leakage in the Nanometer Regime", ISLPED '03, August 25-27, 2003, pp.18-21.

[6] Amit Agarwal, Hai Li, and Kaushik Roy, "DRG-Cache: A Data Retention Gated-Ground Cache for Low Power", DAC 2002, June 10-14, 2001, pp.473-478.

[7] Navid Azizi, Andreas Moshovos, Farid N. Najm, "Low Leakage Asymmetric-Cell SRAM", ISLPED '02, August 12-14, 2002, pp.48-51.

[8] Fatih Hamzaoglu, Yibin Ye, Ali Keshavarzi, Kevin Zhang, Siva Narendra, Shekhar Borkar, Mircea Stan, and Vivek De, "Dual-V_T SRAM Cells with Full-Swing Single-Ended Bit Line Sensing for High-Performance On-Chip Cache in 0.13 µm Technology Generation", ISLPED '00, pp. 15-19.

[9] W. Hung, Y. Xie, N. Vijaykrishnan, M. Kandemir, M.J. Irwin and Y. Tsai, "Total Power Optimization through Simultaneously Multiple-V_{DD} Multiple-V_{TH} Assignment and Device Sizing with Stack Forcing", ISLPED '04, August 9-11, 2004.

[10] Yibin Ye, Shekhar Borkar and Vivek De, "A New Technique for Standby Leakage Reduction in High-Performance Circuits", in Symp. VLSI Circuits Dig. Tech. Papers, 1998, pp. 40-41.

[11] Jun Cheol Park and Vincent J. Mooney III, "Sleepy Stack Leakage Reduction", IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 14, no. 11, pp.1250-1263.

[12] Saibal Mukhopadhyay, Cassondra Neau, Riza Tamer Cakici, Amit Agarwal, Chris H. Kim and Kaushik Roy, " Gate Leakage Reduction for Scaled Devices Using Transistor Stacking", IEEE Trans. Very Large Scale Integr. (VLSI) Systems, vol. 11, no. 4, August 2003, pp. 716-730.

[13] International Technology Roadmap for Semiconductors, 2001Update, Semiconductors Industry Assoc. and SEMATECH.

[14] Amit Agarwal, Chris H. Kim, Saibal Mukhopadhyay and Kaushik Roy, " Leakage in Nano-Scale Technologies: Mechanisms, Impact and Design Considerations", DAC 2004, June 7-11, 2004, pp. 6-11.

[15] Zheng Guo, Sriram Balasubramanian, Radu Zlatanovici, Tsu-Jae king and Borivoje Nikolić, "FinFET-Based SRAM Design", ISLPED '05, August 8-10, 2005, pp. 2-7.

[16] Kanak Agarwal and Sani Nassif," Statistical Analysis of SRAM Cell Stability", DAC 2006, July 24-28, 2006, pp. 57-62.

[17] Benton H. Calhoun and Anantha P. Chandrakasan, " Static Noise Margin Variation for Sub-threshold SRAM in 65-nm CMOS", IEEE J. Solid-State Circuits, Vol. 41, No. 7, July 2006, pp. 1673-1679.

[18] Evert Seevinck, Frans J. List and Jan Lohstroh, " Static-Noise Margin Analysis of MOS SRAM Cells", IEEE J. Solid-State Circuits, Vol. SC-22, No. 5, October 1987, pp. 748-754.

[19] W.Kong,R.Venkatraman,R.Castagnetti,F.Duan and S.Ramesh, "High-Density and High-Performance 6T-SRAM for System-on-Chip in 130nm CMOS Technology", in Symp. VLSI Technology Digest of Tech. Papers,2001, pp.105-106

[20] Ramnath Venkatraman, Ruggero Castagnetti,, Olga Kobozeva, Franklin L. Duan, Arvind Kamath, S. T. Sabbagh, Miguel A. Vilchis-Cruz, Jhon Jhy Liaw, Jyh-Cheng You and Subramanian Ramesh, "The Design, Analysis, and Development of Highly Manufacturable 6-T SRAM Bitcells for SoC Applications"IEEE Trans. on electron devices,vol. 52,No. 2, February 2005,pp. 218-226.

[21] Chris Hyung-il Kim, Jae-Joon Kim, Saibal Mukhopadhyay, , and Kaushik Roy, "A Forward Body-Biased Low-Leakage SRAM Cache:Device, Circuit and Architecture Considerations"IEEE Trans. on VLSI sys., Vol 13, No. 3, March 2005, pp. 349-357.

[22] Hiroaki Nambu, Kazuo Kanetani, Kaname Yamasaki, Keiichi higeta, Masami Usami, Yasuhiro Fujimura, Kazumasa Ando, Takeshi Kusunoki, Kunihiko Yamaguchi and Noriyuki Homma. " A 1.8-ns Access, 550-MHz, 4.5-Mb CMOS SRAM", IEEE J. Solid-State Circuits, Vol 33, No.11, November 1998, pp. 1650-1658.