Design and Simulation of Low Power Stacked SRAM Cell

Balwinder Raj, Member IEEE,
Asst. Prof., Department of ECE, National Institute of Technology Jalandhar
(NIT Jalandhar) P.B-144011, India

Abstract – Low power SRAM has been under its renovation stage for high density and low power applications. This paper proposes 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 has significant power reduction in the standby mode. In addition to these, the proposed cell has a superior Static Noise Margin (SNM) at applied supply voltage. The significant improvements in the results obtained validate our approach for the proposed stacked SRAM cell design for low power memories design.

I. INTRODUCTION

In nanoscale memories design low power design techniques are required for high performance SRAM cell design. 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 [1, 2]. The cell stability is further degraded by scaling of supply voltage. The focus of the SRAM cell design in nanoscale domain 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 [2, 7].

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. In addition to the standby power consumption, access time and the cell stability are important parameters of consideration during the design of SRAM cell [5, 8]. 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 [4, 10].

The stacked SRAM design has been proposed on 90nm technology 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 [11-12]. However, very few applications of stacking effect on SRAM have 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 [16].

II. STACKING EFFECT OF TRANSISTOR

International Technology Road for Semiconductor (ITRS) in 2007 predicted that, memories 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 subthreshold 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 maximizing the number of off transistors by stacking and applying proper input vectors can reduce the standby leakage of a functional block [6].

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.

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 SRAM CELL DESIGN

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 are ‘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.

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.

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.
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 writes 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. RESULTS AND DISCUSSION

The power consumed during the read, write and the standby mode for the Half-stacked, Full-stacked and the conventional memory cell with and without a sleep transistor has been analysis. The various schematics used for the analysis are shown in the Figures 2, 3, 4. We explore the impact of variation in transistor width of the driver (NMOS) and load (PMOS) transistors on the power consumed at 100MHz. The widths mentioned in the figures are that for the transistors in the conventional cell which is twice the width of stacked transistor. For example a width of 200nm in the figures corresponds to 100nm width of each transistor in the corresponding transistor stack. The supply voltage was kept at 1.1V.

The variation of power consumed during the idle period is shown in Figure 5. Fig. 5(a) shows the variation of power with width of load transistors and Fig.5 (b) with width of driver transistors. The power was calculated after the current had reached a steady state value. The figures show that the power consumed is least in case of Full Stacked cell. Half Stacked cell also shows reduced power consumption as compared to that of conventional cell with and without sleep transistor. In addition the power dissipation is less susceptible to process variation in case of both the proposed designs.

![Write power with nmos](image)

![Idle with NMOS](image)

The variation of active power is shown in the Figures 6, 7. Although the stacking was done primarily to reduce the leakage power the figures show a reduction in the active power too.

These figures also show that the variation in power consumed is reduced for the proposed designs. The drop across the lower transistor in the stack, with a zero input vector is very small [14]. It was observed that with increase in the width of the driver transistor, the drop further decreases. As subthreshold current exponentially decreases with decrease in V_DS and increases linearly with increase in W/L ratio, the corresponding increase in the subthreshold current is fractional. Thus the relative variation of power is also less due to more or less fixed value of the current through the stack.
Fig 7: Variation of read power with width of load transistor

CONCLUSION

Stacked SRAM cells design has been proposed for ultra-low power applications. The proposed designs have enhanced performance in terms of power. The full stacked design shows a reduction of 30% in the standby power. 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


[16] T-SPICE 9 User Guide and Reference