Dual-T_{ox} 7T SRAM Cell Design for Leakage Power Reduction on 45nm Technology

[1]ManiramRawat, [2]Anshul Jain ,[3]Vinod Rajput

[1,2,3] Department of Electronics & Communication Engineering, S.R.C.E.M

Email: maniramrawatmits@gmail.com[1] ,anshuljaineng@yahoo.co.in[2]

Abstract – **This paper presents techniques based on dual oxide thickness assignment to reduce the leakage power of SRAM but maintaining their performance. The proposed a new seven transistors (7T) dual oxide thickness SRAM cell is proposed in this paper for simultaneously reducing the active and standby mode power consumption while enhancing the data stability and the read speed. With the new 7T SRAM cell, the storage nodes are isolated from the bit lines during a read operation, thereby enhancing the data stability as compared to the standard six transistors (6T) SRAM circuits. The transistors of the cross-coupled inverters are not on the critical read delay path with the new technique. Minimum sized dual-oxide thickness transistors are therefore conveniently used in the cross-coupled inverters for significantly reducing the leakage power consumption without causing degradation in the read speed. With the proposed 7T SRAM circuit, the static noise margin and the read speed are enhanced by up to 83% and 15%, respectively, as compared to the conventional 6T SRAM circuits. Furthermore, the leakage and the write power consumptions of the proposed dual-Tox SRAM circuit are reduced by up to76% as compared to the conventional6T SRAM circuits in a 45nm CMOS technology.**

I.INTRODUCTION

CMOS scaling technology node requires not only very low threshold voltages to retain the device Switching speeds, but also ultra-thin gate oxides to maintain the current drive and keep threshold voltage variations under control when dealing with short-channel effects [1]. Low threshold voltage results in an exponential increase in the sub threshold leakage current, whereas ultra-thin oxide causes an exponential increase in the gate leakage current. The leakage power dissipation is roughly proportional to the area of a circuit. Since in many processors caches occupy about50% of the chip area [2], the leakage power of caches is one of the major sources of power consumption in high performance microprocessors.

While one way of reducing the sub threshold leakage is to use higher threshold voltages in some parts of a design, to reduce the gate leakage, it is necessary to use multiple oxide thickness. There are different ways to achieve a higher threshold voltage [3], among them are adjusting the channel doping concentration and applying a body bias. To achieve multiple oxide thicknesses, on the other hand, Arsenic implantation into the silicon substrat.before thermal oxidation can be used. Leakage current is a primary concern for low power, high performance digital CMOS circuits for portable applications, and industry trends show that leakage will be the dominant component of power in future technologies. New leakage mechanisms, such as tunneling across thin gate oxides, which lead to gate oxide leakage current (Igate), are coming into play from the 90nm node onwards. According to the International Technological Roadmap for Semiconductors (ITRS) [1], physical oxide thickness (Tox) values of 3–7 Å will be required for high performance CMOS circuits, and quantum effects that cause tunneling will play a dominant role in such ultra-thin oxide devices. The probability of electron tunneling is a strong function. Of the barrier height (i.e., the voltage drop across gate ox-ide) and the barrier thickness, which is simply Tox, and a small change in Tox can have a tremendous impact on Igate. For example, in MOS devices with SiO_2 gate oxides, a difference in Tox of only 2 Å can result in an order of magnitude increase in I_{gate} [2], so that reducing Tox from 3Å to7Å increases I_{gate} current. The other component of leakage, sub threshold leakage (Isub), forms a reducing fraction of the total leakage as Tox is reduced, so that I_{gate} will become the dominant leakage mechanism in the future. The most effective way to control Igate would be through the use of high-k dielectrics, but such materials are not expected before the 45nm technology. This chapter will explore the use of dual Tox values for performance optimization. Although this optimization can be exploited at a number of points in the design methodology, our solution considers Tox assignment as a step that is performed after placement and transistor sizing, at which point it is used to achieve a final performance improvement. Unlike earlier stages of design, there is less design uncertainty at this point and minor changes in layout parasitic due to Tox assignment can be dealt with an incremental update. As a result, all of the delay gains from our procedure can be guaranteed in the final design, with a low leakage power overhead. Leakage power can be broadly divided into two categories: standby leakage, which corresponds to the situation when the circuit is in a non-operating or sleep mode, and active leakage, which relates to leakage during normal operation. Numerous effective techniques for controlling standby leak-age have

been proposed in the past, including state assignment [3], the use of multiple threshold CMOS (MTCMOS) sleep transistors, body-biasing [5], and dual Tox combined with state assignment. Active leakage, however, has not been addressed very widely in the literature so far, primarily because it has not been a major issue in the present technologies. However, leakage power dissipation in the active mode has grown to over 40% in some high-end parts today. Therefore, reducing active leakage is vital for advanced technologies in current-generation circuits, and for Next-generation technologies. The range of options that are available for reducing active leakage is considerably more limited than for standby leakage, and the use of dual Tox assignments is a powerful method for this purpose.

IILEAKAGE CURRENT IN LOW TOX

Ideally an MOS transistor conducts zero current when the gate-to-source voltage is less than the threshold voltage (sub-threshold regime) as shown in Fig. 1. A closer examination of the IDS-VGS curve with a logarithmic scale shows that the drain current is not zero in the sub threshold regime. The sub-threshold drain current, however, drops exponentially with the reduction in the gate-to-source voltage as shown in Fig. 3.3. The sub-threshold drain current is caused primarily by the diffusion of the minority carriers in the channel region [1]. This sub threshold leakage current depends on the bias voltages of the transistor, the threshold voltage, the device dimensions, the doping profile of the channel, the source and the drain, and the junction temperature [1]. A derivation of the drain current in the sub-threshold regime and the parameters that affect the sub-threshold leakage current are presented in this section

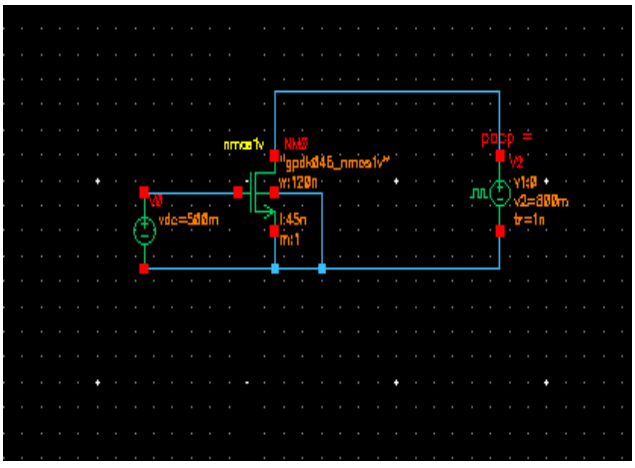


Fig.1 NMOS Transistor With Normal Case

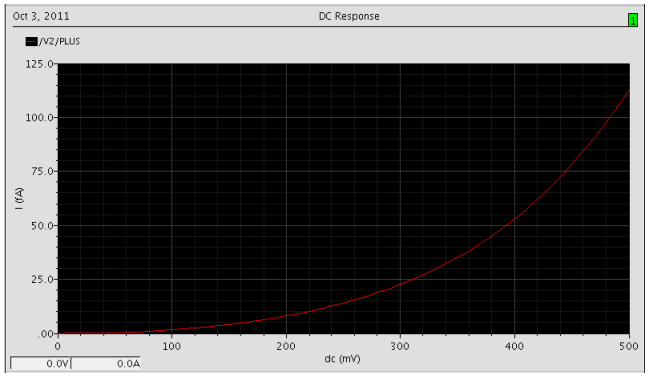


Fig.2. Current in nMOS Transistor in Normal Vt and Tox

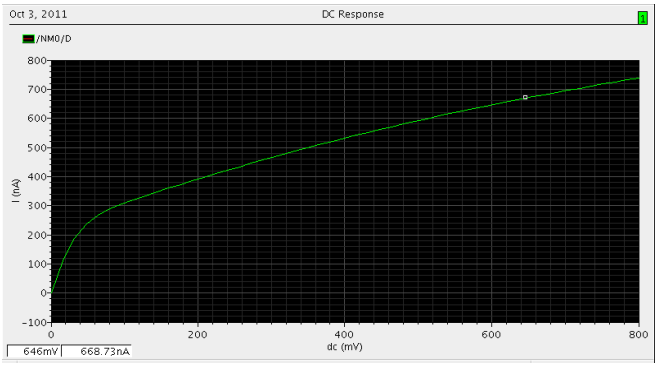


Fig 3 Current in nMOS Transistor in Normal Vt and Tox

Sub threshold leakage is the drain-source current of a transistor when the gate-source Voltage is lower than the threshold voltage. There are two dominant sub threshold leakage paths in a 7TSRAM cell:

1) To ground paths inside the SRAM cell and

2) The bit-line (or bit-bar line) to ground path through the pass transistor.

To reduce the first type of leakage, the threshold voltages of the pull-down NMOS transistors and/or pull-up PMOS transistors can be increased, whereas to lower the second type of leakage, the threshold voltages of the pull-down nMOS transistors and/or pass transistors can be increased. If the threshold voltage of the pull up PMOS transistors is increased, the write delay increases while the effect on the read delay would be negligible. On the other hand, if the threshold voltage of the pull down NMOS transistors is increased, the read delay increases while the effect on the write delay would be marginal. By increasing the threshold voltage of the pass transistors, both read and write delay increases.

2.1 Selecting Tox and Leff:

While an increased value of Tox succeeds in significantly reducing Igate, several other physical effects must be taken into consideration. Increasing the value of Tox while keeping the channel length constant may adversely impact the functionality of the transistor. Specifically, due to drain induced barrier lowering (DIBL), an increase in Tox may result in a situation where the drain terminal takes control of the channel, so that the "on" or "off" state of the transistor is no longer completely governed by the gate terminal.

This effect has been recognized during technology scaling, and scaling trends have shown that Tox reduces nearly in proportion with Leff. We maintain this proportion for each of the chosen values of Tox by setting. The term Tox,e in this equation refers to the electrical Tox, which is related to the physical value of Tox as follows:

$$\frac{Leff@ToxLo}{Tox,eLo} = \frac{Leff@ToxHi}{Tox,eHi}$$

The term T_{ox},e in this equation refers to the electrical Tox, which is related to the physical value of Tox as follows:

$$T_{ox,e} = T_{ox} + T_{oxoffset}$$

The Tox offset term is added to account for the gate depletion and channel quantization effects, and a typical value is 0.3nm. It will be implicit that as we change Tox, the value of Leff will also be scaled. Before determining reasonable values for T_{oxLo} and $T_{ox\ Hi}$,we will study the effect of varying Tox on leakage for an inverter. The gate oxide leakage, I_{gate}, and the sub threshold leakage, I_{sub}, for both the NMOS and PMOS transistors in the inverter, are graphically depicted for various values of T_{oxHi} , at T_{oxLo} = 3 Å ; The values of Isub are obtained through cadence simulations on predictive technology models, and an analytical model is used to generate Igate. The average leakage of the inverter is calculated as the sum of the average Igate and I_{sub} leakages.

A change in Tox of a transistor leaves the load capacitance presented to the previous stage of logic un-changed. As a result, the delay of a fan in logic gate does not change significantly, and hence our optimization method needs only to consider the delay change of a given logic gate when it's Tox is altered. A change in Tox of a transistor leaves the load capacitance presented to the previous stage of logic unchanged. As a result, the delay of a fan in logic gate does not change significantly, and hence our optimization method needs only to consider the delay change of a given logic gate when it's Tox is altered.

Since the capacitance is unchanged, the CVdd2f (dynamic) power remains unaffected by changes in Tox. This is extremely important since our optimization targets the active mode of operation

2.2 Leakage Models:

We will now describe the models used to calculate Isub and Igate for each transistor, and the approach for computing the average Isub and Igate values for a given logic gate. The total leakage current for a logic gate is then computed as the sum of its corresponding average Isub and Igate..

2.3. Tunneling Leakage Current:

Gate oxide leakage can be primarily attributed to electron [hole] tunneling in NMOS [PMOS] devices. Physically, this tunneling occurs in the gate-to-channel region, and in the gate-to-drain/source overlap regions. Of tunneling, referred to as edge direct tunneling (EDT) is ignored in our case for two reasons: firstly, because the gate-to-drain/source overlap region is significantly smaller than the channel region, and secondly, because the oxide thickness in this overlap region can be increased after gate patterning to further suppress. Our work focuses on gate-to-channel tunneling, and we use the following analytic tunneling current density (J_{tunnel}) model based on the electron [hole] tunneling probability through a barrier height (EB).

$$J_{tunnel} = (kT)2\left(1+\frac{\square kT}{z\sqrt{E_B}}\right)\times \exp\left(\frac{E_{F0,Si,SiO2}}{kT}\right)\exp\left(-\square\sqrt{E_B}\right) \quad \ldots(4)$$

Where E_{F0},Si/SiO2 is the Fermi level at the Si/SiO2 interface and m is 0.19Mo for electron tunneling and 0.55Mo for hole tunneling, where Mo is the electron rest mass. The terms k, hand q correspond to physical constants (respectively, Boltzmann's constant, Planck's constant and the charge on an electron, T is the operating temperature, and E_B is the barrier height. Electron tunneling from the conduction band, which is only significant in the accumulation region, results in direct tunneling gate leakage current in nMOS transistors. In pMOS transistors, on the other hand, hole tunneling from the valence band results in the tunneling gate leakage current.

The tunneling gate current is composed of three main components:

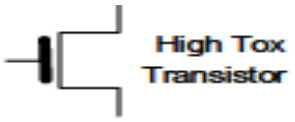
1) gate-to-source and gate-to-drain overlap current;
2) gate-to-channel current, part of which goes to the source.
3) gate-to-substrate current.

In CMOS technology, the gate-to-substrate leakage current is several orders of magnitude lower than the overlap tunneling and gate-to-channel current [6]. On the other hand, while the overlap tunneling current dominates the gate leakage in the OFF state, gate-to-channel tunneling dictates the gate current in the ON state of the transistor. Since the gate-to-source and gate-to-drain overlap regions are much smaller than the channel region, the tunneling gate current in the OFF state is much smaller than gate tunneling in the ON state.

If SiO2 is used for the gate oxide, PMOS transistors will have about one order of magnitude smaller gate leakage than NMOS transistors. Therefore, in SRAM cell, the power saving achieved by increasing the oxide thickness of the PMOS transistors is marginal.

2.4 Effect of High Tox in leakage reduction:

While an increased value of Tox succeeds in significantly reducing Igate, several other physical effects must be taken into consideration. Increasing the value of Tox while keeping the channel length constant may adversely impact the functionality of the transistor. Specifically, due to drain induced barrier lowering (DIBL), an increase in Tox may result in a situation where the drain terminal takes control of the channel, so that the "on" or "off" state of the transistor is no longer completely governed by the gate terminal of the barrier



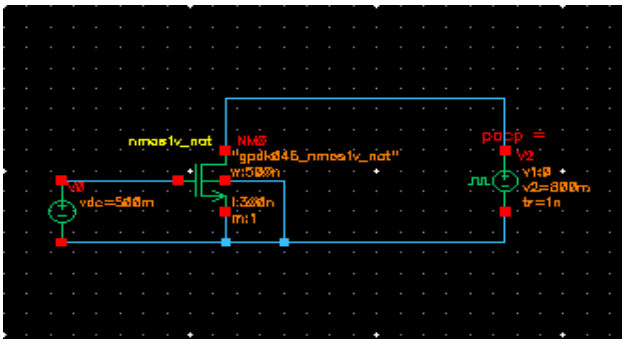


Fig 2.4 a nMOS Transistor with High-Tox

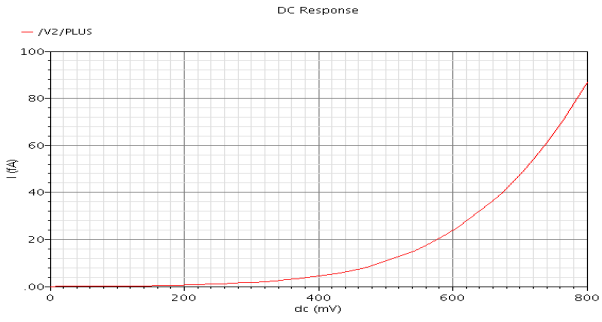


Fig 2.4b Leakage Current in nMOS Using High-Tox

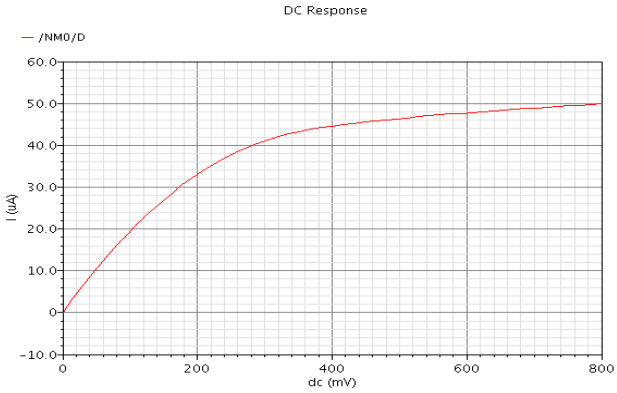


Fig 2.4 c Tunneling Leakage Current in nMOS Using High-Tox

Hight (i.e., the voltage drop across gate oxide) and the barrier thickness, which is simply T_{ox}, and a small change in T_{ox} can have a tremendous impact on Igate. For example, in MOS devices with SiO2 gate oxides, a difference in Tox of only 2Å can result in an order of magnitude increase in Igate [2], so that reducing Tox from 7 Å to3 Å increases I_{gate} by approximately 1000×. The other component of leakage, sub threshold leakage (I_{sub}), forms a reducing fraction of the total leakage as Tox is reduced, so that I_{gate} will become the dominant leakage mechanism in the future. The most effective way to control I_{gate} would be through the use of high-k dielectrics.

The use of dual Tox values for performance optimization. Although this optimization can be exploited at a number of points in the design methodology, our solution considers Tox assignment as a step that is performed after placement and transistor sizing, at which point it is used to achieve a final performance improvement. Un-like earlier stages of design, there is less design uncertainty at this point and minor changes in layout parasitic due to T_{ox} assignment can be dealt with an incremental update. As a result, all of the delay gains from our procedure can be guaranteed in the final design, with a low leakage power overhead.

While an increased value of T_{ox} succeeds in significantly reducing I_{gate}, several other physical effects must be taken into consideration. Increasing the value of T_{ox} while keeping the channel length constant may adversely impact the functionality of the transistor. Specifically, due to drain induced barrier lowering (DIBL), an increase in Tox may result in a situation where the drain terminal takes control of the channel, so that the "on" or "off" state of the transistor is no longer completely governed by the gate terminal. The transistor with the minimum (most negative) cost provides the largest delay reduction for the smallest increase in leakage power, and is selected for assignment to T_{oxLo}. The corresponding L_{eff} is also concurrently changed. If two

transistors have the same cost, ties are heuristically broken, first by selecting the transistor with the higher fan out.

III.Dual Tox 7T SRAM cell

The use of dual Tox values for performance optimization. Although this optimization can be exploited at a number of points in the design methodology, our solution considers Tox assignment as a step that is per-formed after placement and transistor sizing, at which point it is used to achieve a final performance improvement. Unlike earlier stages of design, there is less design uncertainty at this point and minor changes in layout parasitic due to Tox assignment can be dealt with an incremental up date. As a result, all of the delay gains from our procedure can be guaranteed in the final design, with a low leakage power overhead.

Leakage power can be broadly divided into two categories: standby leakage, which corresponds to the situation when the circuit is in a non-operating or sleep mode, and active leakage, which relates to leakage during normal operation. Numerous effective techniques for controlling standby leakage have been proposed in the past, including state assignment [3], the use of multiple threshold CMOS (MTCMOS) sleep transistors, body-biasing [5], and dual Tox combined with state assignment. Active leakage, however, has not been addressed very widely in the literature so far, primarily because it has not been a major issue in the present technologies. However, leakage power dissipation in the active mode has grown to over 45% in some high-end parts today [6]. Therefore, reducing active leakage is vital for advanced technologies in current-generation circuits, and for next generation technologies. The range of options that are available for reducing active leakage is considerably more limited than for standby leakage, and the use of dual T_{ox} assignments is a powerful method for this purpose.

3.1 SRAM cell configuration by using dual-Tox

The reversed biased p-n junction leakage has two main components one is minority carriers' diffusion near the edge of the depletion region and the other is due to electron-hole pair generation in the depletion region Of the reverse biased junction [12]. The junction tunneling current is an exponential function of junction doping and reverse bias voltage across the junction.

To reduce the gate tunneling leakage of an SRAM cell, only the oxide thickness of the pull down NMOS transistors and pass-transistors need to be increased. Although this is seemingly desirable from a low power point of view, it is not applicable for all cells in the cell array; thin oxide needs to be used in the cells far from the address decoder and sense amplifiers. It should be emphasized that increasing the oxide thickness. Also increases the threshold voltage, resulting in a Decrease in the sub threshold leakage. In the following, high-V_t transistors refer to those transistors whose threshold voltage have been modified by increasing the channel doping, not the ones whose threshold voltage has been boosted as a result of increasing the oxide thickness. The simulation results in this segment are obtained by using 45nm technology using cadence software, which accurately models sub threshold gate leakage current. The value of low threshold voltage is .25v and the thin oxide thicknesses 3.10 Å while the thick oxide is 3.20 Å. The supply technology is .8v.

Table 3.1 Possible configurations for high oxide thickness assignment

cell	Leakage reduction current in over (000)	% leakage reduction current in over (000)
0,0,0	20.813	–
0,0,1	20.66	0.035
0,1,0	17.82	0.69
0,1,1	17.7	0.71
1,0,0	19.45	0.31
1,0,1	19.3	0.34
1,1,0	16.76	0.94
1,1,1	16.55	0.98

3.2Leakage current reduction for each configuration 3.2.1 Configuration (0, 0, 0)

For (0, 0, 0) Corresponds to a configuration with normal pull-down transistors (M1, M2), normal pull-up transistors (M3, M4) and normal pass transistors (M5, M6).

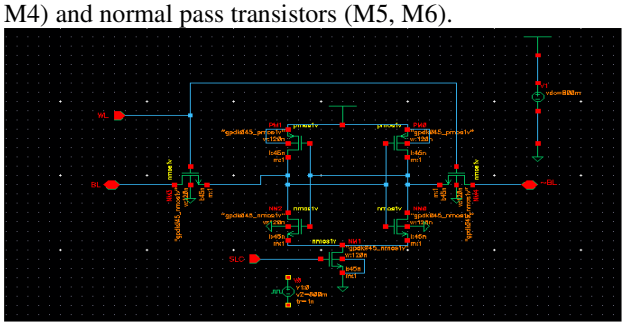


Fig.3.2.1a, Configuration (0, 0, 0) 7T SRAM Cell

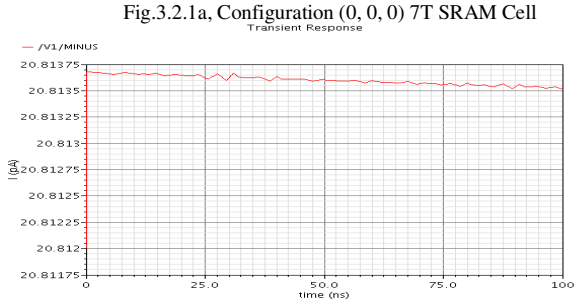


Fig 3.2.1b Leakage Current Reduction Waveform

3.2.2Configuration (0, 0, 1)

For (0, 0, 1) Corresponds to a configuration with normal pull-down transistors (M1, M2), normal pull-up transistors (M3, M4) and high-T_{ox} pass transistors (M5, M6).

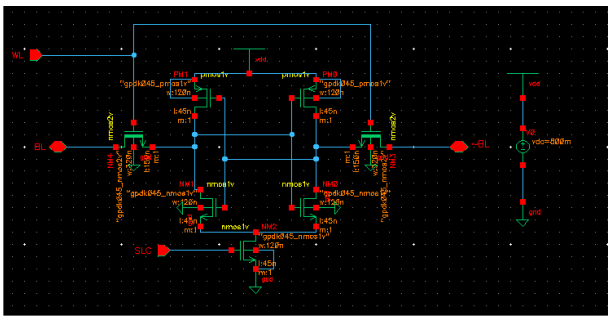


Fig.3.2.2a Configuration (0, 0, 1) 7T SRAM Cell in Dual Tox

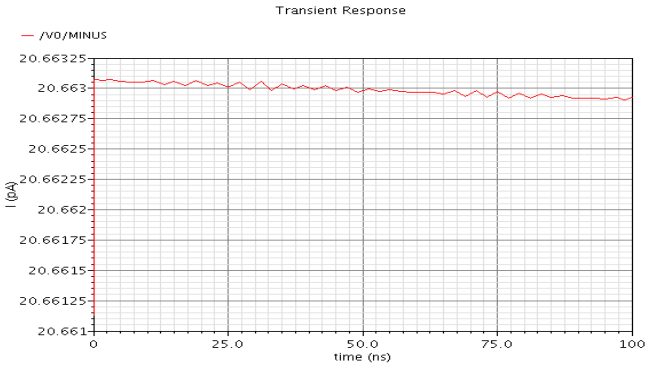


Fig 3.2.2b Leakage Current Reduction Waveform

3.2.3 Configuration (0, 1, 0)

For (0, 1, 0) Corresponds to a configuration with normal pull-down transistors (M1, M2), high-T_{ox} pull-up transistors (M3, M4) and normal pass transistors (M5, M6).

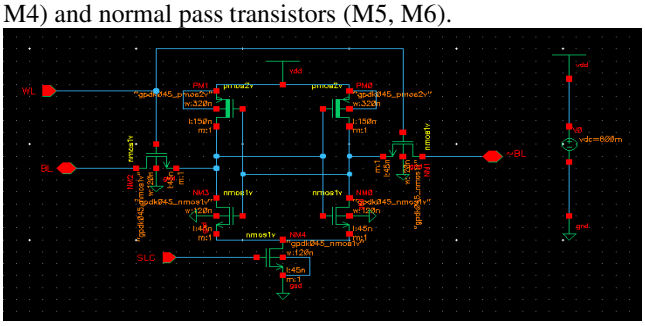


Fig.3.2.3a Configuration (0, 1, 0) 7T SRAM Cell in Dual-Tox

Fig 3.2.3b Leakage Current Reduction Waveform

3.3.4 Configuration (0, 1, 1)

For (0, 1, 1) Corresponds to a configuration with normal pull-down transistors (M1, M2), high-T_{ox} pull-up transistors (M3, M4) and high-T_{ox} pass transistors (M5, M6).

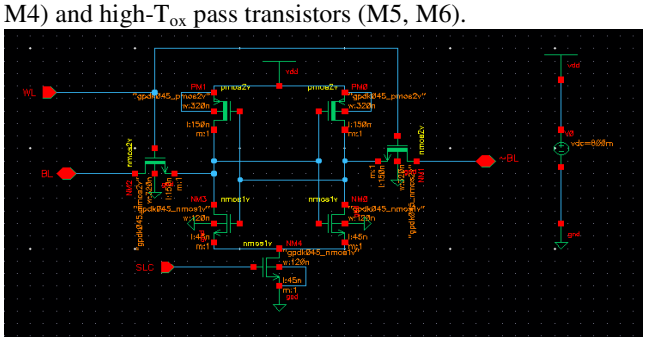


Fig. 3.3.4a Configuration (0, 1, 1) 7T SRAM cell in Dual-Tox

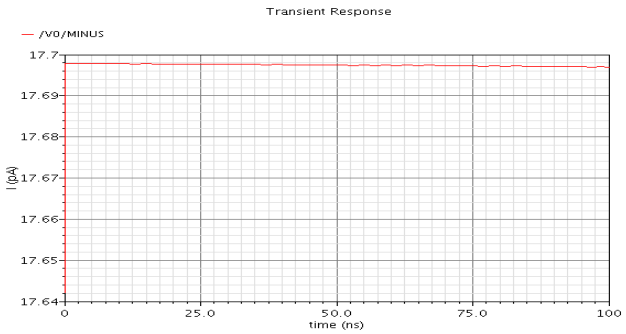


Fig. 3.3.4b Leakage Current Reduction Waveform

3.2.5 Configuration (1, 0, 0)

For (1, 0, 0) Corresponds to a configuration with high-T_{ox} pull-down transistors (M1, M2), normal pull-up transistors (M3, M4) and normal pass transistors (M5, M6).

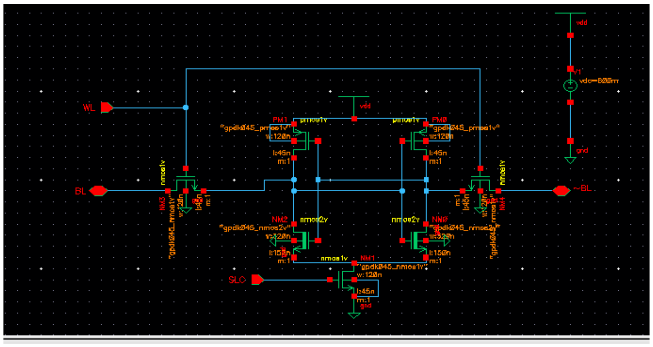


Fig. 3.2.5a Configuration (1, 0, 0) 7T SRAM cell in Dual-Tox

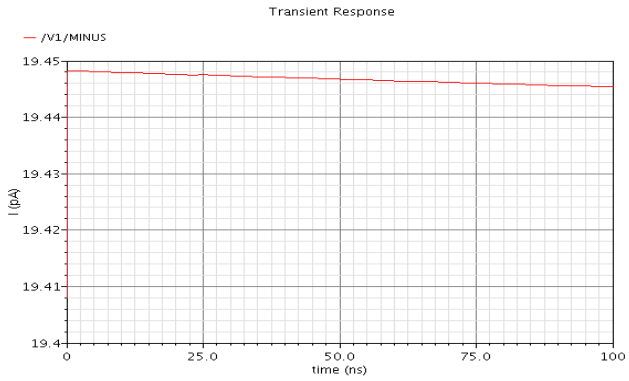


Fig 3.2.5b Leakage current reduction waveform

3.2.6 Configuration (1, 0, 1)

For (1, 0, 1) Corresponds to a configuration with high-T_{ox} pull-down transistors (M1, M2), normal pull-up transistors (M3, M4) and high-T_{ox} pass transistors (M5, M6).

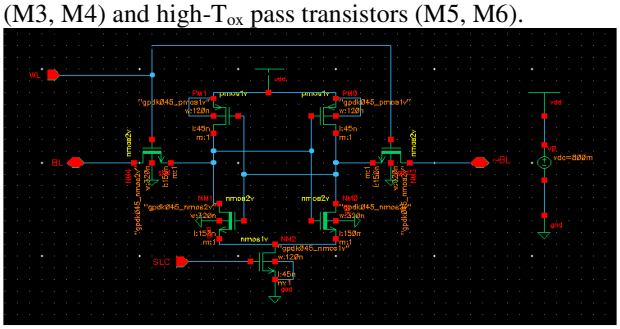


Fig. 3.2.6a Configuration (1, 0, 1) 7T SRAM cell in dual-Tox

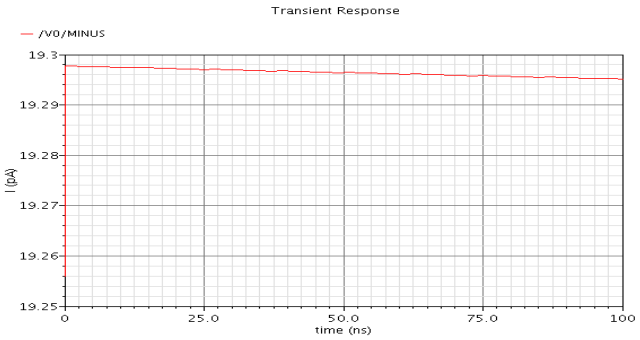


Fig 3.2.6b Leakage Current Reduction Waveform

3.2.7 Configuration (1, 1, 0)

For (1, 1, 0) Corresponds to a configuration with high-T_{ox} pull-down transistors (M1, M2), high-T_{ox} pull-up transistors (M3, M4) and normal pass transistors (M5, M6).

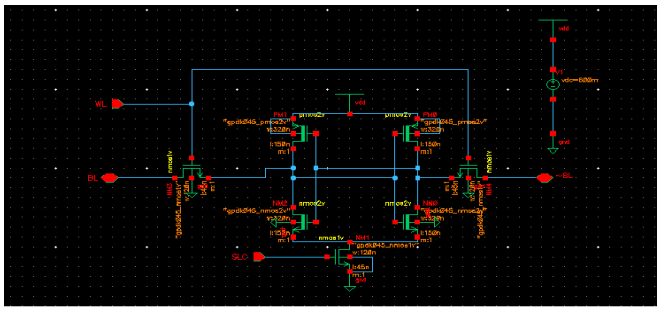


Fig.3.2.7a Configuration (1, 1, 0) 7T SRAM Cell In Dual-Tox

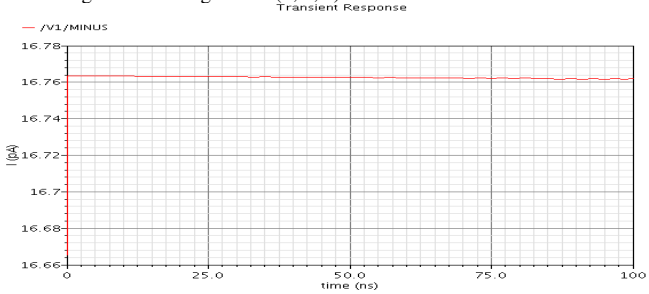


Fig.3.2.7bLeakage Current Reduction Waveform

3.2.8 Configuration (1, 1, 1)

For (1, 1, 1) Corresponds to a configuration with high-T_{ox} pull-down transistors (M1, M2), high-T_{ox} pull-up transistors (M3, M4) and high-T_{ox} pass transistors (M5, M6).

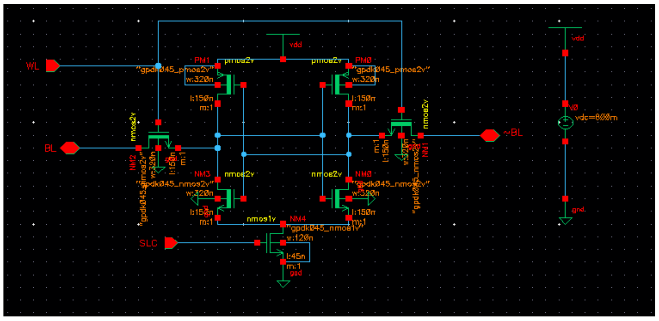


Fig.3.2.8a Configuration (1, 1, 0) 7T SRAM Cell In Case of Dual-Tox

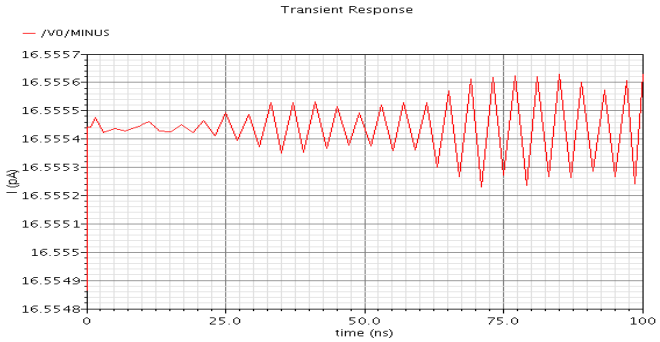


Fig 3.2.8b Leakage Current Reduction Waveform

IV.CONCLUSION

From the above study it is concluded that the proposed technique is used for reducing the total active leakage, including gate oxide leakage, by determining appropriate values of Tox, and iteratively assigning them to the individual transistor in the circuit. Our approach shows a clear tradeoff between leakage and delay, and an achievable delay reduction of 25%.

REFERENCES

[1] Semiconductor Industry Association, "International Technology Roadmap for Semiconductors," 2002. Available at http://public.itrs.net.

[2] F. Hamzaoglu and M. R. Stan, "Circuit-Level Techniques to Control Gate Leakage for Sub-100 nm CMOS," in Proc. Of ACM/IEEE ISLPED, pp. 60–63,Aug. 2002.

[3] D. Lee and D. Blaauw, "Static Leakage Reduction through Simultaneous Threshold Voltage and State Assignment," in Proc.of ACM/IEEE DAC, pp. 191–194, Jun. 2003.

[4] J. Kao et al., "Transistor Sizing Issues and Tool for Multi-Threshold CMOS Technology," in Proc. of ACM/IEEE DAC, pp. 409–414, Jun. 1997.

[5] Y. Oowaki et al., "A sub-0.1 µm Circuit Design with Substrate Over-Biasing," in IEEE ISSCC Dig. of Tech. Papers, pp. 88– 89, Feb. 1998.

[6] S. Narendra et al., "Leakage Issues in IC design: Trends, Estimation, and Avoidance." Tutorial at the IEEE/ACM ICCAD,Nov. 2003.

[7] D. Lee et al., "Analysis and Minimization Techniques for To-tal Leakage Considering Gate Oxide Leakage," in Proc. Of ACM/IEEE DAC, pp. 175– 180, Jun. 2003.

[8] C.-H. Choi et al., "Impact of Gate Direct Tunneling on CircuitPerformace: A Simulation Study," IEEE Trans. on ElectronDevices, pp. 2823–2829, Dec. 2001.

[9] N. Sirisantana et al., "High-Performace Low-Power CMOS Circuits Using Multiple Channel Length and Multiple Oxide Thickness," in Proc. of IEEE ICCD, pp. 227–232, Sept. 2000.

[10] K. Bernstein, Private Communication. IBM T. J. Watson Research Center, Yorktown Heights, NY,2003.

[11] Y. Taur, "CMOS Design Near the Limits of Scaling," IBM J.R&D, vol. 46(2/3), pp. 213–222, Mar./May 2002.

[12] K. Chen et al., "Predicting CMOS Speed with Gate Oxide and Voltage Scaling and Interconnect Loading Effects," IEEE Trans. On Electron Devices, vol. 44(11), pp. 1951–1957, Nov. 1997.

[13] Device Group at UC Berkeley, "Berkeley Predictive Technology Model," 2002. Available at http://www-device.eecs.berkeley.edu/_ptm/.

[14] S. Sirichotiyakuletal., "Duet: An Accurate Leakage Estimationand Optimization Tool for Dual-Vt Circuits," IEEE Trans. On VLSI Systems, vol. 10(2), pp. 79–90, Apr. 2002.

[15] A. Chandrakasan et al., Design of High-Performance MicroprocessorCircuits. Piscataway, NJ: IEEE Press,2001.

[16] K. A. Bowman et al., "A Circuit-Level Perspective of the Optimum Gate Oxide Thickness," IEEE Trans. on Electron Devices, vol. 48(8), pp. 1800–1810, Aug. 2001.

[17] J. Fishburn and A. Dunlop, "TILOS: A Posynomial Programming Approach to Transistor Sizing," in Proc. of ACM/IEEEICCAD, pp. 326–328, Nov. 1985.

[18] E. M. Sentovich et al., "SIS: A System for Sequential Circuit Synthesis," Tech. Rep. UCB/ERL M92/41, Electronics Research Laboratory, Dept. of EECS, University of California, Berkeley,May 1992.