Gate Leakage Current Analysis in READ/ WRITE/ IDLE States of a SRAM Cell

Valmiki Mukherjee Saraju P. Mohanty Elias Kougianos
Email: vm0058@unt.edu Email: smohanty@cs.unt.edu Email: eliask@unt.edu
Rahul Allawadhi Ramakrishna Velagapudi
Email: ra0139@unt.edu Email: rv0063@unt.edu
VLSI Design and CAD Laboratory (http://www.vdcl.cse.unt.edu)
P. O. Box 311366, University of North Texas, Denton, TX 76203.

Abstract

The increasing market demand for ever smaller and application packed portable electronic devices has been fueling the relentless scaling of the CMOS transistor. The ITRS roadmap envisages that high performance CMOS circuits will require ultra-low gate oxide thickness to overcome the effects of shorter channel lengths. However, such devices will be susceptible to a more profound leakage mechanism due to carrier tunneling through the gate oxide. Consequently, the gate oxide tunneling current has emerged as the major component of the leakage power consumption of nanoscale CMOS devices. In the case of an important CMOS circuit like Static RAM (SRAM) there is a high probability for the leakage currents to be manifested with more prominence. SRAMs form a vital component of the CPU cache therefore there is a critical need for analysis, explanation, and characterization of the various tunneling mechanisms SRAMs. This paper explores the gate leakage current scenarios in the READ, WRITE and IDLE states of the SRAM which can make significant contribution to modeling and reduction of gate leakage in SRAM circuits.

1 Introduction

There has been a significant increase in the demand for low power and high performance digital VLSI circuits. Designers are implementing very high-order scaling of both device dimensions and supply voltage. At this stage there are several short channel effects (SCE) such as drain induced barrier lowering (DIBL), large V_{th} roll-off, diminishing I_{ON}/I_{OFF} and band-to-band tunneling (BTBT.) As a result, there has been a drastic change in the leakage components of the device both in the inactive as well as active modes of operation. The leakage current in short channel nanometer transistors has diverse forms, such as reverse biased diode leakage, subthreshold leakage, SiO_2 tunnel current, hot carrier gate current, gate induced drain leakage, channel punch through current [1]. While biased diode leakage and SiO_2 tunnel current flow during both active and sleep mode of the circuit, the other currents flow during the sleep mode only.

In this paper we focus on a typical and extensively used CMOS circuit, the static RAM (SRAM) cell. SRAM structures are used for cache-memory and compromise a large number of the on-chip transistors in bulk-CMOS as well as System-on-Chip (SOC) systems. We have developed a thorough understanding of the phenomenon of gate leakage right from the transistor level up to the level of functional units and architectural blocks. In this research we plan to analyze and evaluate the effective gate leakage current for an SRAM cell in READ/WRITE/IDLE operation. Eventually, this analysis of the gate leakage current will help in exploring new techniques to reduce gate leakage in these structures.

Most of the works in gate leakage analysis and reduction have focussed on combinatorial circuits. Memory circuits need more attention as they are equally susceptible to the phenomenon of gate leakage. Various gate leakage reduction methodologies have been described in the literature such as analytical modeling for gate leakage in the behavioral domain presented in [2]. In the logic domain a dual dielectric technique for reduction of gate leakage was presented in [3]. Limits on scaling of SRAM have been discussed in [4, 5]. Issues in SRAM leakage suppression and corresponding techniques have been proposed in [6, 7, 8, 9, 10, 11, 12]. The authors in [13] carried out a feasibility study for subthreshold SRAM. This shows the growing concern for leakage, especially direct tunneling through the gate oxide of a MOS transistor. This work presents a comparative study of the different states of operation of the SRAM which can be further used for analysis and reduction techniques.

2 Analysis of a SRAM Operation

Static Random Access Memory (SRAM) is a type of semiconductor memory. Each bit in an SRAM is stored on four transistors that form two cross-coupled inverters. This storage cell has two stable states which are used to denote "0" and "1". Two additional access transistors help controlling the access to the cross coupled unit formed by the inverters during read and write operations. So typically it takes six transistors to store one memory bit. The design of a basic SRAM cell is shown in Fig. 1. Access to the cell is enabled by the word line (WL) which controls the two access transistors N3 and N4 which allow the access of the memory cell to the bit lines: BL and $\overline{BL}$. They are used to transfer data for both read and write operations. The presence of dual bit lines i.e. BL and $\overline{BL}$ improves noise margins over a single bit line. The symmetric circuit structure allows for accessing a memory location much faster than in a DRAM. Also the faster operation of an SRAM over DRAM can be attributed to the fact that it accepts all address bits at a time where as DRAMs typically have the address multiplexed in two halves, i.e. higher bits followed by lower bits.

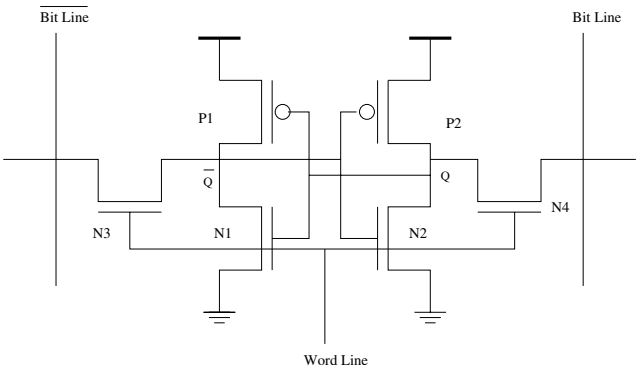


Figure 1: Basic diagram of a 6T SRAM cell.

The operation of a CMOS SRAM cell can be described in terms of three states viz. WRITE, READ and IDLE operations. The start of a write cycle begins by applying the value to be written to the bit lines. In order to write a "0", we would apply a 0 to the bit lines, i.e. setting $\overline{BL}$ to 1 and BL to 0. A "1" is written by inverting the values of the bit lines. WL is then made high and the value that is to be stored is latched in. The input-drivers of the bit lines are designed to be much stronger than the relatively weak transistors in the cell itself, so that they can easily override the previous state of the cross-coupled inverters. Proper operation of an SRAM cell however needs careful sizing of the transistors in the unit. The read cycle is started by asserting the word line WL, enabling both the access transistors N3 and N4. The second step occurs when the values stored in Q and $\overline{Q}$ are transferred to the bit lines BL and $\overline{BL}$ through N1 and N3. On the BL side, the transistors P2 and N4 pull the bit line towards V_{DD} (when a "1" is stored at Q). If the content of the memory was a 0, the reverse would happen and $\overline{BL}$ would be pulled towards 1 and BL towards 0. For the idle state, the word line is not asserted and the access transistors N3 and M4 disconnect the cell from the bit lines. The two cross coupled inverters formed by N1, N2, N3 and N4 will continue to reinforce each other as long as they are disconnected from any external circuits.

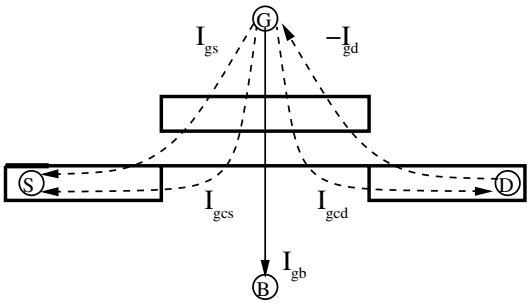
3 Gate Leakage in a SRAM cell

An analysis of the SRAM would need the basic analysis of its building blocks, NMOS and PMOS transistors. So here we review the physical mechanism of the gate leakage in a MOS transistor and present the case for gate leakage in SRAM. We identify the regions of operations of an NMOS device (which can then be extrapolated for a PMOS) distinguishing its transient and steady states. Different mechanisms contribute to the overall current during different phases of the switching cycle. The physical mechanism of the tunneling current can be studied in separate regions as steady-state region (ON or OFF) and transient state (during Low-to-High and High-to-Low transition).

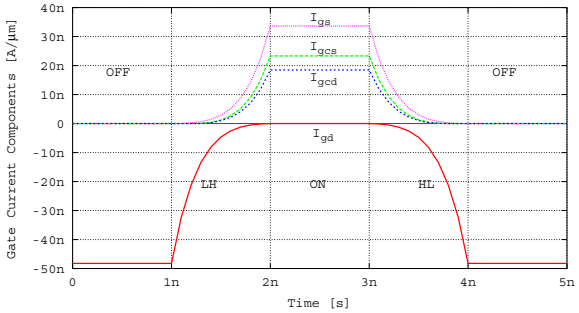
In the steady-state ON region both the gate and drain of the device are held at high with the source being grounded. In this state a well-formed channel exists and three separate components of the gate tunneling current I_{gs}, I_{gcs} and I_{gcd} are active. The component from gate to drain overlap (I_{gd}) is absent due to the almost zero electric field in that region of the oxide. The overall current flow is from gate to source and channel, opposite to the flow in the OFF state. In the steady-state OFF region both gate and source are at ground while the drain is at high (V_{DD}) voltage. Since no channel is formed in this condition, the only active component is I_{gd}.

The transient state prevails when the device changes from ON to OFF or OFF to ON state, which is not an instantaneous process. During Low-to-High (LH) and High-to-Low (HL) region all four components of the gate tunneling current become active as shown in Fig. 2(b) . In this case the source is at ground, the drain is at V_{DD} and the gate is switched from low to high or high to low. In the LH transition, the channel gradually originates at the source and extends to the drain and the components I_{gs}, I_{gcs} and I_{gcd} start becoming significant, in that order. Conversely, as the field across the oxide region over the drain is reduced, I_{gd} decreases to almost total extinction. A study of this state is important in the case of SRAM because it can show the effect that transition from one of the states to the other has on the gate leakage.

As discussed above, the SRAM can operate in three modes viz. WRITE, READ and IDLE. These modes have different states and different combinations of transistors are



(a) Gate leakage current paths in a NMOS transistor



(b) Components of I_{ox} in NMOS corresponding to a pulse input

Figure 2: Gate oxide tunneling current (I_{ox}) components in BSIM4.4.0 model. I_{gs} and I_{gd} are the components due to the overlap of gate and diffusions, I_{gcs} and I_{gcd} are the components due to tunneling from the gate to the diffusions via the channel and I_{gb} is the component due to tunneling from the gate to the bulk via the channel. We made similar observations in the case of PMOS with relatively smaller yet of comparable magnitude currents.

active during these states leading to the prevalence of different leakage components and different values of gate leakage. The test bench for analysis of gate leakage in SRAM is shown in Fig. 3 based on circuit from [14].

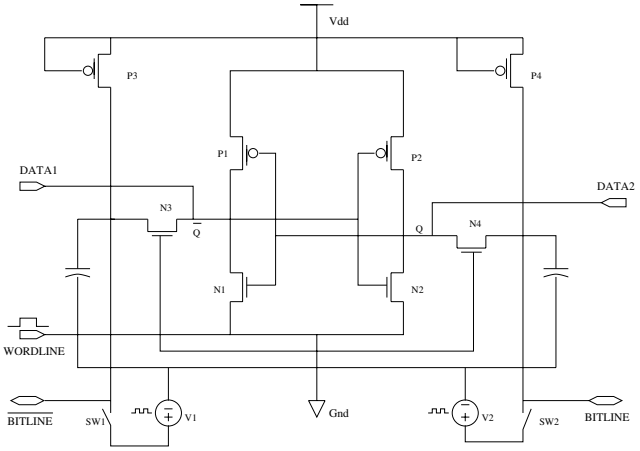


Figure 3: Testbench used for the analysis of the gate leakage in an SRAM based on circuit from [14]

3.1 WRITE Operation

The analysis is started with the WRITE operation. In case of writing a "1" the BL is held high and the $\overline{BL}$ is held low. In this state N2 and P2 leak the most as they are connected to the BL which is high. Also the access transistor on the BL side, N4 leaks significantly as both BL and the WL are high. These set of transistors are ON and provide a path for the gate leakage to flow. P1 and P3 are turned ON too and leak in this state. N1 and N3 are OFF and leakage is low in their case. The case for writing a "1" is exactly the reverse of this case where the transistors that are OFF while writing a "0" are ON during writing a "1" due to the symmetry in the SRAM cell.

3.2 READ Operation

In case of a READ operation the supplies V1 and V2 shown in Fig. 3 are not required and hence disconnected from the circuit by turning off switches SW1 and SW2. For the sake of analysis, let us assume that the content of the memory is "1" (stored at Q). The read cycle is started by making the word line WL high, enabling both the access transistors N3 and N4 which contribute to the gate leakage as during the WRITE operation. As a consequence of the READ operation BL is left at its precharged value and $\overline{BL}$ is discharged through N1 and N3. On the BL side, the transistors P2 and N4 pull the bit line towards V_{DD} and as they are both ON, contribute to leakage. If the content of the memory was a "0", the opposite would happen and $\overline{BL}$ would be pulled towards "1" and BL towards "0". So the same set of transistors that contributed to the leakage in the WRITE operation contributed to the leakage during the READ state. So the leakage profile for a READ operation is expected to the be same as in the case of WRITE and reading a "0" or "1" does not make any difference.

3.3 IDLE Operation

The SRAM goes into the IDLE state when the word line WL is maintained low, then, the access to the pass transistors N3 and N4 disconnect the cell from the bit lines. The two cross coupled inverters formed by N1-P1 and N2-P2 will continue to be active and hence leakage will take place in them even if they are disconnected from any external circuit. In this state the transistors connected to the power supply i.e. P1 and P2 will still leak and hence a considerable gate leakage is still expected even if not of the order of READ or WRITE.

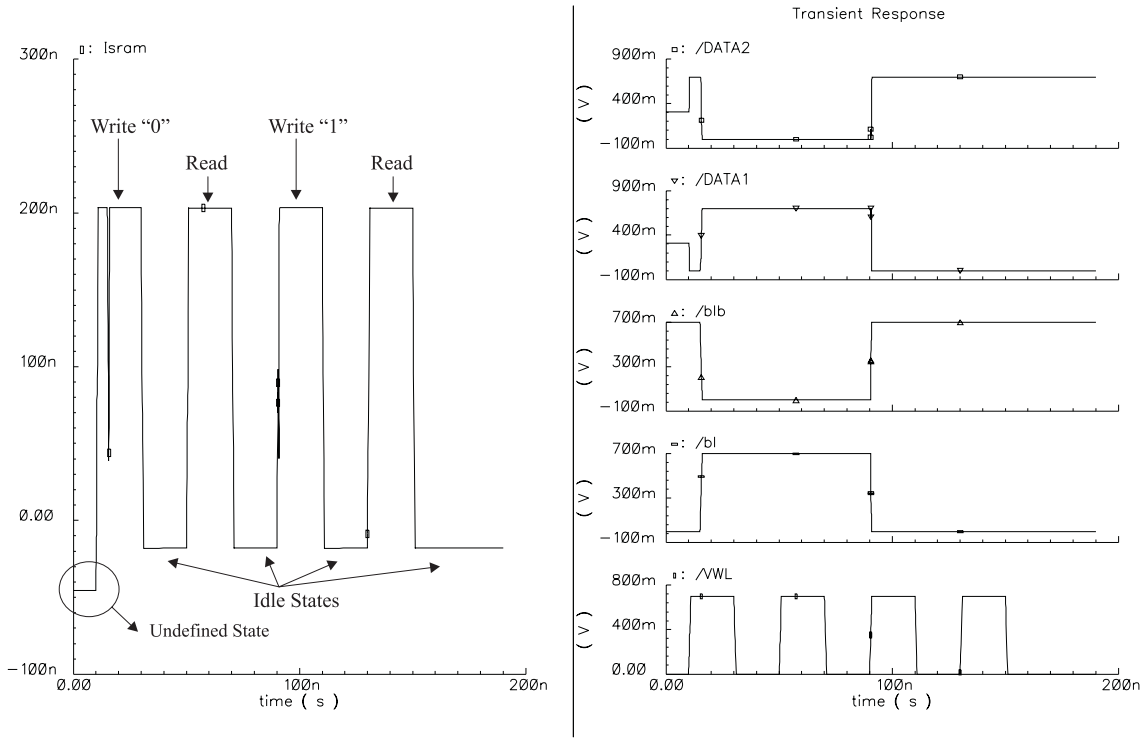


Figure 4: Plots representing simulation results from the input waveform and the corresponding gate leakage current in an SRAM cell

4 Experimental Results

At the device level we use the Berkeley Predictive Model (BPTM) for a $45nm$ device technology node with $T_{ox} = 1.4nm$, threshold voltage $V_{th} = 0.22V$, and supply voltage $V_{DD} = 0.7V$. The width of the device is chosen to be very large ($W = 1\mu m$), thus eliminating any narrow-width or width-modulation effects. We use Cadence Design Systems' Analog Design Environment and Spectre circuit simulator for the purpose of design and simulation of the circuit. We analyze gate leakage as the gate direct tunneling current by evaluating all components (source, drain and bulk) in each of the transistors of the SRAM from the BSIM4.4.0 model [15].

The various states of the SRAM were simulated using the same test bench by including a pair of switches implemented in VerilogA. The switches SW1 and SW2 connected to the BL and the $\overline{BL}$ during the WRITE operation to the inputs V1 and V2 and helped in precharging them. These were turned off during the READ and the IDLE operations when the bitlines were only sensed. So using a combination of piece wise linear inputs for the wordline and the bit lines a sequence of input and outputs states were achieved. The sequence comprised of WRITE "1" - IDLE - READ - IDLE - WRITE "0" - IDLE - READ -IDLE. This simulates all the real scenarios involved in the SRAM where there is a gap between the various operations and different values are written to and read from the SRAM. The state of the various input and outputs during the various operations are shown in Fig. 4. This corresponds to the analysis presented in the section 3 and shows that the gate leakage in READ and WRITE operations is almost similar as the same set of transistors leak while it is lower in the case of the IDLE operation.

5 Conclusions and Future Works

In this work we presented a systematic comparison of the various modes of operation of an SRAM. This work is notable in its contribution to future reduction techniques based on this analysis. It was seen that both the WRITE and READ mode dissipates the maximum amount power through gate leakage. In case of the IDLE state it was seen that the SRAM device gate leakage is not as significant as in the WRITE/READ mechanism however there is still a considerable dissipation through gate leakage which is not at all desirable. A number of future works in this regard is under consideration including sleep state assignments for IDLE state as well as for designing circuits for preventing gate leakage in SRAM.

References

[1] K. Roy, S. Mukhopadhyay, and H. M. Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, Feb 2003.

[2] S. P. Mohanty, V. Mukherjee, and R. Velagapudi, "Analytical Modeling and Reduction of Direct Tunneling Current during Behavioral Synthesis of Nanometer CMOS Circuits," in *Proceedings of the 14th ACM/IEEE International Workshop on Logic and Synthesis (IWLS)*, pp. 248–249.

[3] V. Mukherjee, S. P. Mohanty, and E. Kougianos, "A Dual Dielectric Approach for Performance Aware Gate Tunneling Reduction in Combinational Circuits," in *Proceedings of the IEEE International Conference on Computer Design (ICCD)*, pp. 431–437.

[4] A. Bhavnagarwala and et. al., "The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Scalability," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 4, pp. 658 – 665, Apr 2001.

[5] A. Bhavnagarwala, A. Kapoor, and J. Meindl, "Fluctuation Limits on Scaling of CMOS SRAMs," in *Proceeding of the 30th European Solid-State Device Research Conference*, pp. 472 – 475.

[6] N. Azizi, A. Moshovos, and F. N. Najm, "Low-Leakage Assymetric-Cell SRAM," in *International Symposium on Low Power Electronics and Design*, 2002, pp. 48–51.

[7] N. Azizi and F. N. Najm, "An Asymmetric SRAM Cell to Lower Gate Leakage," in *Proceedings of IEEE International Symposium on Quality Electronic Design*, 2004, pp. 534–539.

[8] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," in *Proceedings of IEEE International Symposium on Quality Electronic Design*, 2004, pp. 55–60.

[9] K. Nii and et. al., "A 90-nm Low-Power 32-KB Embedded SRAM with Gate Leakage Suppression Circuit for Mobile Applications," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 4, pp. 684–693, Apr 2004.

[10] B. Ameliard, M. Pedram, and F. Fallah, "Reducing The Sub-Threshold And Gate-Tunneling Leakage Of SRAM Cells Using Dual-VT And Dual-Tox Assignment," in *Proceedings of the Design Automation and Test in Europe*, 2006.

[11] A. Goel and B. Mazhari, "Gate leakage and its reduction in deep submicron SRAM," in *Proceeding of the International Conference on VLSI Design*, 2005, pp. 606–611.

[12] K. Zhang and et. al., "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 876885, Apr 2005.

[13] A. Raychowdhury, S. Mukhopadhyay, and K. Roy, "A Feasibility Study of Subthreshold SRAM Across Technology Generations," in *Proceedings of International Conference on Computer Design*, pp. 561–564.

[14] "Aries: An LSI Macro-Block for DSP Applications," `http://www.vlsi.wpi.edu/aries/5.html`.

[15] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New Paradigm of Predictive MOSFET and Interconnect Modeling for Early Circuit Design," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 201–204.