A

Thesis Report

On

Design and Analysis of Low Power SRAM

Submitted in the partial fulfilment for the

Degree of

Master of Technology

In

VLSI Design & CAD

By

Nitin Gupta

Reg. No. 601061030

Under the supervision of

Dr. Ravi Kumar

Assistant Professor

Department of Electronics & Communication Engineering

Thapar University, Patiala-147004, India

Jun 2012

DECLARATION

I hereby declare that the Thesis report entitled **"Design and Analysis of Low Power SRAMs"** is an authentic record of my own work carried out as requirement for the award of Master of Technology in VLSI Design & CAD at Thapar University, Patiala under the guidance of **Dr. Ravi Kumar**, Assistant Professor, (ECED).

nitin gupta

Date: 26-06-12

Nitin Gupta
601061030

It is certified that the above statement is correct to the best of my knowledge and belief.

Ravi kumar

Date 26/06/12

Dr.Ravi Kumar
Assistant Professor
Thapar University, Patiala

Countersigned by:

Dr. Rajesh Khanna
Professor & Head, ECED
Thapar University, Patiala

Dr. S.K. Mohapatra
Dean, Acadmic Affairs
Thapar University, Patiala

ACKNOWLEDGEMENT

First of all, I would like to express my gratitude to **Dr. Ravi Kumar, Assistant Professor,** Electronics and Communication Engineering Department, Thapar University, Patiala for her patient guidance and support throughout this report. I am truly very fortunate to have the opportunity to work with her. I found this guidance to be extremely valuable.

I am also thankful to our Head of Department**, Dr. Rajesh Khanna,** Electronics and Communication Engineering Department, entire faculty, staff of Electronics and Communication Engineering Department, and then friends who devoted their valuable time and helped me in all possible ways towards successful completion of this work. I thank all those who have contributed directly or indirectly to this work.

Lastly, I would like to thank my parents for their years of unyielding love. They have always wanted the best for me and I admire their determination and sacrifice.

Nitin Gupta

601061030

ABSTRACT

Power has been a major issue in SoC designs with the contemporary sub-micron technologies. . It has thus become very important to control the power and address the power dissipation throughout the design cycle right from the architectural level. For 180nm and below technologies, leakage is the main factor which dominates over the dynamic power and contributes to almost or more than 50% of total power dissipation. In many new high performance designs, the leakage component of power consumption is comparable to the switching component. According to some authenticated reports, 40% or more of the total power consumption is due to the leakage of transistors. This percentage is likely to increase with technology scaling unless effective techniques are introduced to bring leakage under control. This report presents several topology based leakage reduction mechanism applicable to a standard 6-T SRAM cell. Substantial reduction in the leakage current in standby mode has been obtained. The leakage current components considered in present work are gate leakage and subthreshold leakage. This thesis has been organized into 6 sections. The first chapter presents an introductory overview of semiconductor memories in general and SRAM in particular with a compressive review of contemporary literature. The second chapter reviews the leakage current components inside an SRAM cell and popular techniques to tackle the issue. Third and Fourth chapter summarizes the proposed topologies when yield substantial reduction in off state leakage. Finally chapter five and six describe simulation based performance comparison and conclusion derived from this work respectively.

TABLE OF CONTENTS

LIST OF FIGURES

Page

Table

Abbreviations

RAM	Random Access Memory
DRAM	Dynamic Random Access Memory
EEPROM	Electrically Erasable Programmable Read-Only Memory
SDRAM	Synchronous Dynamic Random Access Memory
MRAM	Magneto resistive Random Access Memory SRAM
nMOS	n-type MOSFET
pMOS	p-type MOSFET
USRS	Upper Supply Rejection Scheme
LPRS	Lower Potential Raising Scheme
WL	Word Line
BL	Bit line
BLB	Bit line bar
Sense_en	Sense enable
Read_en	Read enable

CHAPTER 1

Introduction and Literature Review

1.1 Motivation

High power consumption in portable electronics devices is an issue of serious concern. Shortening of battery life and additional packaging and cooling requirement are associated with high power consumption. Static power dissipation due to standby leakage currents is an important component of total power dissipation. Ubiquitous electronics devices contain different types of component of which many remain idle during a particular operation. Static power dissipation occurring in these idle components accounts for a huge percentage of total power dissipation in the system. Therefore, minimization of this leakage component becomes crucial for effective power management. As a result of continued scaling of MOS devices, a dramatic enhancement in the performance of MOS devices has been achieved. This has led to increase power dissipation due to leakage currents. Till now, the drain to source sub-threshold current has been the dominant leakage component.

The second driving force behind the low power design phenomenon is a growing class of personal computing devices, such as portable desktops, digital pens, audio and video-based multimedia products, and wireless communications and imaging systems, such as personal digital assistants, personal communicators and smart cards. These devices and systems demand high-speed, high-throughput computations, complex functionalities and often real time processing capabilities. The performance of these devices is limited by the size, weight and lifetime of batteries. Serious reliability problems, increased design costs and battery-operated applications prompted the IC design community to look more aggressively for new approaches and methodologies that produce more power-efficient designs, which means significant reductions in power consumption for the same level of performance. Memory circuits form an integral part of every system design as Dynamic RAMs, Static RAMs, Ferroelectric RAMs, ROMs or Flash Memories, significantly contributing to the system level power consumption. Reducing the power dissipation in memories can significantly improve the system power-efficiency, performance, reliability and overall costs. RAMs have experienced a very rapid development of low-power low-voltage memory design during

recent years due to an increased demand for notebooks, laptops, hand-held communication devices and IC memory cards.

1.2 Literature review:

Chen and Peh, August 2003 [1]: Power will be the key limiter to system scalability as inter connection networks take up an increasingly significant portion of system power. In this paper, the authors propose an architectural leakage power modeling methodology that achieves 95- 98% accuracy against HSPICE estimates. When applied to interconnection networks, combined with previous proposed dynamic power models, The authors provide valuable insights on total network power consumption. Their modeling shows router buffers to be a prime candidate for leakage power optimization. They thus investigate the design space of power-aware buffer policies, propose a suite of policies, and explore the impact of various circuits mechanisms on these policies. Simulations show power-aware buffers saving up to 96:6% of total buffer leakage power.

Andrei et. al., 2004 [2]: In contemporary and future embedded as well as high performance microprocessors, power consumption is one of the most important design considerations. Since, in current technologies the dynamic power consumption dominates the static power consumption, voltage scaling is an effective technique to reduce the power consumption. The most common way to reduce the power consumption of multi-processor systems is to schedule a program to run on as many processors as possible and apply voltage scaling afterwards. As technology scales to increasingly smaller feature sizes, however, the static power consumption is expected to grow exponentially. In this paper, the authors first show for which combinations of leakage current, supply voltage, and clock frequency the static power consumption dominates the dynamic power dissipation. Based on these results, it is at a certain point no longer advantageous to use as many processors as possible. They then present a heuristic to schedule task graphs on a number of processors that is sufficient to meet the deadline, but at the same time minimizes the power consumption.. Their results show that their scheduling algorithm reduces the total energy consumption by up to 65%, compared to the strategy that schedules the tasks on the maximum number of processors and then exploits the remaining slack to lower the supply voltage.

Mutoh et. al. ,1995 [3]: For the most recent CMOS feature sizes (e.g., 90nm and 65nm), leakage power dissipation has become an overriding concern for VLSI circuit designers. ITRS reports that leakage power dissipation may come to dominate total power consumption. The authors propose a novel approach, named "sleepy keeper," which reduces leakage current while saving exact logic state. Sleepy keeper uses traditional sleep transistors plus two additional transistors – driven by a gate's already calculated output to save state during sleep mode. Dual Vth values can be applied to sleepy keeper in order to dramatically reduce sub threshold leakage current. In short, like the sleepy stack approach, sleepy keeper achieves leakage power reduction equivalent to the sleep and zigzag approaches but with the advantage of maintaining exact logic state (instead of destroying the logic state when sleep mode is entered). Based on experiments with a 4-bit adder circuit, sleepy keeper approach achieves up to 49% less delay and 49% less area than the sleepy stack approach. Unfortunately, sleepy keeper causes additional dynamic power consumption, approximately 15% more than the base case (no sleep transistors used at all). However, for applications spending the vast majority of time in sleep or standby mode while also requiring low area, high performance and maintenance of exact logic state, the sleepy keeper approach provides a new weapon in a VLSI designer's arsenal.

Karimi and Alimoradi, 2011[4]: Rapid growth in semiconductor technology has led to shrinking of feature sizes of transistors using deep submicron (DSM) process. As MOS transistors enter deep submicron sizes, undesirable consequences regarding power consumption arise. Until recently, dynamic or switching power component dominated the total power dissipated by an IC. Voltage scaling is perhaps the most effective method to decrease dynamic power due to the square law dependency of digital circuit active power on the supply voltage. As a result, this demands a reduction of threshold voltage to maintain performance. Low threshold voltage results in an exponential increase in the sub-threshold leakage current. On the other hand as technology scales down, shorter channel lengths result in increased sub-threshold leakage current through an off transistor. Therefore, in DSM process static or leakage power becomes a considerable proportion of the total power dissipation. For these reasons, static power consumption, i.e. leakage power dissipation, has become a significant portion of total power consumption for current and future silicon technologies.

There are several different approaches tackling leakage. Each technique provides an efficient way to reduce leakage power, but disadvantages of each technique limit the application of each technique. Previously proposed work can be divided into following techniques:

(1) state-saving techniques: where circuit state (present value) is retained.

(2) state-destructive techniques:

where the current Boolean output value of the circuit might be lost. MTCMOS power gating is a well-known way to reduce leakage and it continues to be applied to very-deep submicron CMOS technologies. This can be done by using one PMOS transistor and one NMOS transistor in series with the transistors of each logic block to create a virtual ground and a virtual power supply. Notice that in practice only one transistor is necessary, because of their lower on-resistance, NMOS transistors are usually used.

Amrutur et.al.,2000 [5]: They propose a novel approach that leverages circuit and architecture level techniques to drastically reduce leakage power dissipation in high-performance caches even when most of the cache cells are actively used. They observe that the cache resident memory values of ordinary programs exhibit a strong bias towards zero or one at the bit level. They introduce a family of high-speed dual-Vt SRAM cell designs that exploit this bit-level bias to reduce leakage power while maintaining low access latency. The key characteristic of this cell family is asymmetry: Leakage power dissipation depends on the actual bit value stored. In the preferred state, the leakage power is smaller by as much as approximately 10x and comparable to that of an high-Vt cell. Asymmetry is also key to maintaining high performance reads (the main disadvantage of an high-Vt cell). They propose two asymmetric-cell cache (ACC) designs. The first is statically biased towards the zero bit value. The other uses run-time selective inversion to increase the number of zero-holding bits. They evaluate their designs using the SPEC2000 benchmarks and for a commercial 0.13m, 1.2V CMOS technology. They find that for most programs the majority of memory bits are zero with the actual fraction varying from 52% to 88% for the level one data cache. They also find that this bias is less evident in the instruction stream (around 60% on the average). Using selective inversion, it is possible to further increase the fraction of zero holding bits by another 6% and 11% for the level one data and instruction caches respectively. Overall, for one cell design, leakage power is reduced by 96% and 94% for the

level one data and instruction caches compared to conventional caches. Finally, the novelty of their proposed method has been established.

Roy et.al., 2003 [6]: High leakage current in deep-submicrometer regimes is becoming a significant contributor to power dissipation of CMOS circuits as threshold voltage, channel length, and gate oxide thickness are reduced. Consequently, the identification and modeling of different leakage components is very important for estimation and reduction of leakage power, especially for low-power applications. This paper reviews various transistor intrinsic leakage mechanisms, including weak inversion, drain-induced barrier lowering, gate-induced drain leakage, and gate oxide tunneling. Channel engineering techniques including retrograde well and halo doping are explained as means to manage short-channel effects for continuous scaling of CMOS devices. Finally, the paper explores different circuit techniques to reduce the leakage power consumption.

To achieve higher density and performance and lower power consumption, CMOS devices have been scaled for more than 30 years. Transistor delay times decrease by more than 30% per technology generation, resulting in doubling of microprocessor performance every two years. Supply voltage has been scaled down in order to keep the power consumption under control. Hence, the transistor threshold voltage has to be commensurately scaled to maintain a high drive current and achieve performance improvement. However, the threshold voltage scaling results in the substantial increase of the sub threshold leakage current.

In this paper, all leakage mechanisms contributing to the off-state current (not just the current from the drain terminal) have been explored. Other leakage mechanisms are peculiar to the small geometries themselves. As the drain voltage increases, the drain to channel depletion region widens, resulting in a significant increase in the drain current. This increase in drain current is due to channel surface current caused by drain-induced barrier lowering (DIBL) or due to deep channel punch through currents. Moreover, as the channel width decreases, the threshold voltage and the off current both get modulated by the width of the transistor, giving rise to significant narrow-width effect. All these adverse effects which cause threshold voltage reduction (leakage current increase) in scaled devices are called short-channel effects (SCE). To maintain a reasonable SCE immunity while scaling down the channel length, oxide thickness has to be reduced nearly in proportion to the channel length. Decrease in

oxide thickness results in increase in the electric field across the gate oxide. The high electric field and low oxide thickness result in considerable current flowing through the gate of a transistor. This current destroys the classical infinite input impedance assumption of MOS transistors and thus affects the circuit performance severely.

Powell et.al., 2000 [7]: Deep-submicron CMOS designs have resulted in large leakage energy dissipation in microprocessors. While SRAM cells in on chip cache memories always contribute to this leakage, there is a large variability in active cell usage both within and across applications. This paper explores an integrated architectural and circuit level approach to reducing leakage energy dissipation in instruction caches. They propose, gated-Vdd, a circuit-level technique to gate the supply voltage and reduce leakage in unused SRAM cells. Their results indicate that gated-Vdd together with a novel resizable cache architecture reduces energy-delay by 62% with minimal impact on performance.

The ever-increasing levels of on-chip integration in the recent decade have enabled phenomenal increases in computer system performance. Unfortunately, the performance improvement has been also accompanied by an increase in a chip's power and energy dissipation. Higher power and energy dissipation require more expensive packaging and cooling technology, increase cost, decrease product reliability in all segments of computing market, and significantly reduce battery life in portable systems. Historically, chip designers have relied on scaling down the transistor supply voltage in subsequent generations to reduce the dynamic energy dissipation due to a much larger number of on chip transistors. Maintaining high transistor switching speeds, however, requires a commensurate down-scaling of the transistor threshold voltage giving rise to a significant amount of leakage energy dissipation even when the transistor is not switching.

Sumita et. al, 2005 [8]: There remains a need to improve sub-1-V CMOS VLSIs with respect to variation in transistor behavior. In this paper, to minimize variation in delay and the noise margin of the circuits in processors, They propose several mixed body bias techniques using body bias generation circuits. In these circuits, either the saturation region of the current between source and drain or the threshold voltage of PMOS/NMOS is permanently fixed, regardless of temperature range or variation in process. A test chip that featured these body bias generation circuits was fabricated using a 130-nm CMOS process

with a triple-well structure. The mixed body bias techniques which keep the current (drain to source) of the MOS in the decoder and I/O circuits of a register file fixed and maintain the threshold voltage of MOS in both the memory cell and domino circuits of the register file fixed resulted in positive temperature dependence of delay from 40 C to 125 C, 85% reduction of the delay variation compared with normal body bias (NBB) at = 0 8 V. In addition, the results using these techniques show a 100-mV improvement in lower operating voltage compared with NBB at -40 degree C on a 4-kb SRAM.

Wei et. al., 1999 [9]: Reduction in leakage power has become an important concern in low-voltage, low-power, and high-performance applications. In this paper the dual-threshold technique was used to reduce leakage power by assigning a high-threshold voltage to some transistors in noncritical paths, and using low-threshold transistors in critical path(s). In order to achieve the best leakage power saving under target performance constraints, an algorithm is presented for selecting and assigning an optimal high-threshold voltage. A general leakage current model which has been verified by HSPICE simulations is used to estimate leakage power. Results show that the dual-threshold technique is good for leakage power reduction during both standby and active modes. For some ISCAS benchmark circuits, the leakage power can be reduced by more than 80%. The total active power saving can be around 50% and 20% at low- and high-switching activities, respectively.

Okuyama et. al., 1988 [10]: A 256K CMOS static RAM (SRAM) which achieves an access time of 7.5 ns and 50-mA active current at 50 MHZ operation is described. A 32-block architecture is used to achieve high-speed access and low power dissipation. To achieve faster access time, a double activated- pulse circuit which generates the word-line-enable pulse and the sense-amplifier-enable pulse has been developed. The data-output reset circuit reduces the transition time and the noise generated by the output buffer. A new self-aligned contact technology reduces the diffused region capacitance. This RAM has been fabricated in a twin-tub CMOS 0.8 um technology with double-level poly silicon (the first level is polycide) and double-level metal. The memory cell size is 6* 11 um^2 and the chip size is 4.38X9.47 mm^2.

Mohammad, 2011 [11]-Leakage power becomes big percentage of total active power especially for small geometry CMOS technology. It is estimated that 20-50% of total average

power during normal operation lost to leakage power. Leakage power is even more important for mobile devices where ideal time is long and battery life is important. This paper presents a low leakage SRAM cell and array architecture targeting high performance, low power embedded memory. The proposed novel 7-Transistor (7T) based memory provides 50% lower leakage power compare to 8T cell and 30% faster access time than traditional 6-Transistor (6T) SRAM cell with increased area of20% compared to the compact 6T cell. All comparisons are based on 28nm foundry low power process technology.

Do et.al , 2011 [12]: A new current-mode sense amplifier is presented. It extensively utilizes the cross-coupled inverters for both local and global sensing stages, hence achieving ultra low-power and ultra high-speed properties simultaneously. Its sensing delay and power consumption are almost independent of the bit- and data-line capacitances. Extensive post-layout simulations, based on an industry standard 1 V/65-nm CMOS technology, have verified that the new design outperforms other designs in comparison by at least 27% in terms of speed and 30% in terms of power consumption. Sensitivity analysis has proven that the new design offers the best reliability with the smallest standard deviation and bit-error-rate (BER). Four 32*32-bit SRAM macros have been used to validate the proposed design, in comparison with three other circuit topologies. The new design proposed by author can operate at a maximum frequency of 1.25 GHz at 1 V supply voltage and a minimum supply voltage of 0.2 V. These attributes of the proposed circuit make it a wise choice for contemporary high-complexity systems where reliability and power consumption are of major concerns.

E. Aly et. al.,2005 [13]: In this paper, a low-power SRAM design by reducing the accessed bit line capacitance has been proposed. Each column consist of several divisions. Each division consists of equal number of cells connected together with a local bit line and one local sense amplifier. The divisions are connected together with a global bit line that is also connected to the write, precharge, and reading sense amplifier circuits. During write operation, the global bit lines is slightly discharged to develop a small voltage difference on the accessed local bit lines enough for the local sense amplifier to amplify it to full swing on the local bit lines pair only. The experimental results show reduction in the power

consumption that can reach more than 88% with no performance degradation. The architecture is flexible to be used according to the power, and area budget.

Cheng and Huang, 2005 [14]: This paper presents a low-power SRAM design with quiet-bitline architecture by incorporating two major techniques. Firstly, the authors use a one-side driving scheme for the write operation to prevent the excessive full-swing charging on the bitlines. Secondly, they use a precharge free pulling scheme for the read operation so as to keep all bitlines at low voltages at all times. SPICE simulation on a 2K-bit SRAM macro shows that such architecture can lead to a significant 84.4% power reduction over a self-designed baseline low-power SRAM macro.

W.Mann et. al., 2010 [15]: Large scale 6T SRAM beyond 65 nm will increasingly rely on assist methods to overcome the functional limitations associated with scaling and the inherent read stability/write margin trade off. The primary focus of the circuit assist methods has been improved read or write margin with less attention given to the implications for performance. In this work, they introduce margin sensitivity and margin/delay analysis tools for assessing the functional effectiveness of the bias based assist methods and show the direct implications on voltage sensitive yield. A margin/delay analysis of bias based circuit assist methods is presented, highlighting the assist impact on the functional metrics, margin and performance. A means of categorizing the assist methods is developed to provide a first order understanding of the underlying mechanisms. The analysis spans four generations of low power technologies to show the trends and long term effectiveness of the circuit assist techniques in future low power bulk technologies.

Ming, 2005 [16]: This paper describes a low-power write scheme by adopting charge sharing technique. By reducing the bitlines voltage swing, the bitlines dynamic power is reduced. The memory cell's static noise margin (SNM) is discussed to prove it is a feasible scheme. Simulation results show compare to conventional SRAM, in write cycle this SRAM saves more than 20% dynamic power.

Itoh, 1995 [17]: Trends in low-power circuit technologies of CMOS RAM chips are reviewed in terms of three key issues: charging capacitance, operating voltage, and dc current. The discussion includes a general description of power sources in a RAM chip, and covers both DRAM’s and SRAM’s. In DRAM’s, successive circuit advancements have

produced a power reduction equivalent to two to three orders of magnitude over the last decade for a fixed memory capacity chip. Coupled with the low-power advantage of CMOS circuits, two technologies have been the major contributors to power reduction: lower charging capacitance due to partial activation of multi-divided arrays that use multi-divisions of data and word lines and lower operating voltage resulting from external power supply reduction, half-VDD pre-charging, and on-chip voltage down converting scheme. In SRAM's, partial activation of a multi-divided word line drastically reduces the dc current from the data-line load to the selected cell. In addition to advances in the sense amplifier circuit, an auto power down scheme that uses address transition detection for word driver and column circuitry further reduces the dc current. It is also shown that to design ultralow voltage DRAM's and SRAM's, the application of sub threshold current reduction circuits (such as source-gate back biasing) to cell and iterative circuit blocks will be indispensable in the future.

1.3 Introduction to Static RAM

1.3.1 SRAM Memory

SRAM or Static random Access memory is a form of semiconductor memory widely used in electronics, microprocessor and general computing applications. This form of semiconductor memory gains its name from the fact that data is held in there in a static fashion, and does not need to be dynamically updated as in the case of DRAM memory. While the data in the SRAM memory does not need to be refreshed dynamically, it is still volatile, meaning that when the power is removed from the memory device, the data is not held, and will disappear.

1.3.1.1 SRAM memory basics

There are two key features to SRAM - Static random Access Memory, and these set it out against other types of memory that are available:

1. The data is held statically: This means that the data is held in the semiconductor memory without the need to be refreshed as long as the power is applied to the memory.

2.SRAM is a form of random access memory: A random access memory is one in which the locations in the semiconductor memory can be written to or read from in any order, regardless of the last memory location that was accessed.

The circuit for an individual SRAM memory cell comprises typically four transistors configured as two cross coupled inverters. In this format the circuit has two stable states, and these equate to the logical "0" and "1" states. In addition to the four transistors in the basic memory cell, and additional two transistors are required to control the access to the memory cell during the read and write operations. This makes a total of six transistors, making what is termed a 6T memory cell. Sometimes further transistors are used to give either 8T or 10T memory cells. These additional transistors are used for functions such as implementing additional ports in a register file, etc for the SRAM memory.

Although any three terminal switch device can be used in an SRAM, MOSFETs and in particular CMOS technology is normally used to ensure that very low levels of power consumption are achieved. With semiconductor memories extending to very large dimensions, each cell must achieve very low levels of power consumption to ensure that the overall chip does not dissipate too much power.

1.3.1.2 SRAM memory cell operation

The operation of the SRAM memory cell is relatively straightforward. When the cell is selected, the value to be written is stored in the cross-coupled flip-flops. The cells are arranged in a matrix, with each cell individually addressable. Most SRAM memories select an entire row of cells at a time, and read out the contents of all the cells in the row along the column lines. While it is not necessary to have two bit lines, using the signal and its inverse, this is normal practice which improves the noise margins and improves the data integrity. The two bit lines are passed to two input ports on a comparator to enable the advantages of the differential data mode to be accessed, and the small voltage swings that are present can be more accurately detected [18].

Access to the SRAM memory cell is enabled by the Word Line. This controls the two access control transistors which control whether the cell should be connected to the bit lines. These two lines are used to transfer data for both read and write operations.

1.3.1.3 SRAM memory applications

There are many different types of semiconductor memory that are available these days. Choices need to be made regarding the correct memory type for a given application. Possibly two of the most widely used types are DRAM and SRAM memory, both of which are used in processor and computer scenarios. Of these two SRAM is a little more expensive than DRAM. However SRAM is faster and consumes less power especially when idle. In addition to this SRAM memory is easier to control than DRAM as the refresh cycles do not need to be taken into account, and in addition to this the way SRAM can be accessed is more exactly random access. A further advantage if SRAM is that it is more dense than DRAM.

As a result of these parameters, SRAM memory is used where speed or low power are considerations. Its higher density and less complicated structure also lend it to use in semiconductor memory scenarios where high capacity memory is used, as in the case of the working memory within computers.

1.4 SRAM Architecture: Different types of SRAM cells are based on the type of load used in the elementary inverter of the flip-flop cell. There are currently three types of SRAM memory cells:

• The 4T cell (four NMOS transistors plus two poly load resistors)

• The 6T cell (six transistors—four NMOS transistors plus two PMOS transistors)

• The TFT cell (four NMOS transistors plus two loads called TFTs)

1.4.1 4-Transistor (4 T) Cell

The most common SRAM cell consists of four NMOS transistors plus two poly-load resistors (Figure 1.1 [19]). This design is called the 4T cell SRAM. Two NMOS transistors are pass-transistors. These transistors have their gates tied to the word line and connect the cell to the columns. The two other NMOS transistors are the pull-downs of the flip-flop

inverters. The loads of the inverters consist of a very high polysilicon resistor. This design is the most popular because of its size compared to a 6T cell. The cell needs room only for the four NMOS transistors. The poly loads are stacked above these transistors. Although the 4T SRAM cell may be smaller than the 6T cell, it is still about four times as large as the cell of a comparable generation DRAM cell.

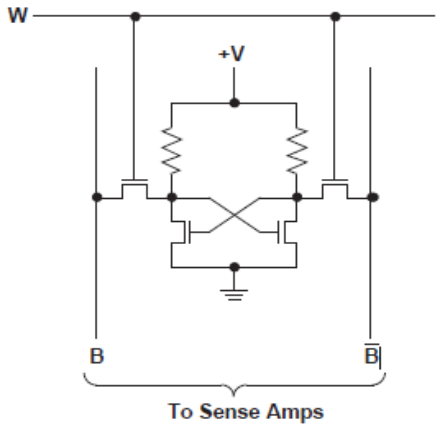

Figure 1.1 4-T SRAM cell

The complexity of the 4T cell is to make a resistor load high enough (in the range of giga ohms) to minimize the current. However, this resistor must not be too high to guarantee good functionality. Despite its size advantage, the 4T cells have several limitations. These include the fact that each cell has current flowing in one resistor (i.e., the SRAM has a high standby current), the cell is sensitive to noise and soft error because the resistance is so high, and the cell is not as fast as the 6T cell.

1.4.2 6 Transistor (6T) Cell

A different cell design that eliminates the above limitations is the use of a CMOS flip-flop. In this case, the load is replaced by a PMOS transistor. This SRAM cell is composed of six transistors, one NMOS transistor and one PMOS transistor for each inverter, plus two NMOS transistors connected to the row line. This configuration is called a 6T Cell. Figure 1.2 [20] shows this structure. This cell offers better electrical performances (speed, noise immunity, standby current) than a 4T structure. The main disadvantage of this cell is its large size. Until recently, the 6T cell architecture was reserved for niche markets such as military or space that needed high immunity components. However, with commercial applications needing

faster SRAMs, the 6T cell may be implemented into more widespread applications in the future. Much process development has been done to reduce the size of the 6T cell. At the 1997 ISSCC conference, all papers presented on fast SRAMs described the 6T cell architecture.

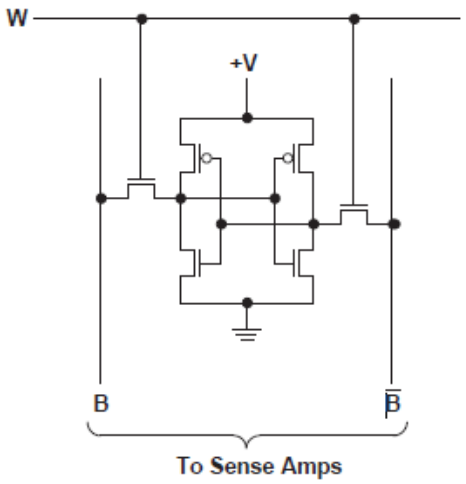

Figure 1.2: 6-T SRAM cell

1.4.3 TFT (Thin Film Transistor) Cell

Manufacturers have tried to reduce the current flowing in the resistor load of a 4T cell. As a result designers developed a structure to change, during operating, the electrical characteristics of the resistor load by controlling the channel of a transistor. This resistor is configured as a PMOS transistor and is called a thin film transistor (TFT) as shown in figure 1.3 [21]. It is formed by depositing several layers of polysilicon above the silicon surface. The source/channel/drain is formed in the polysilicon load. The gate of this TFT is polysilicon and is tied to the gate of the opposite inverter as in the 6T cell architecture. The oxide between this control gate and the TFT polysilicon channel must be thin enough to ensure the effectiveness of the transistor.

The performance of the TFT PMOS transistor is not as good as a standard PMOS silicon transistor used in a 6T cell. It should be more realistically compared to the linear polysilicon resistor characteristics.

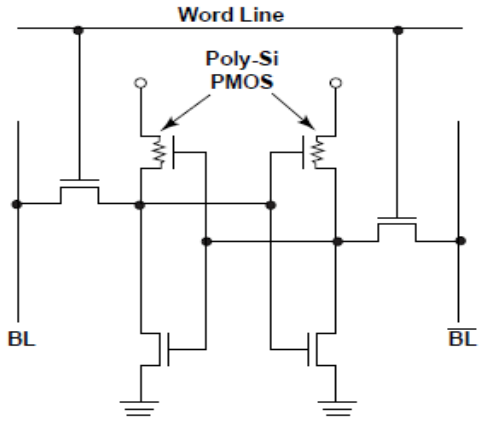

Figure 1.3: TFT cell

1.5 Design challenges: Achieving low-voltage operation in SRAM faces a many of challenges, originating from process variation, and related to bit cell stability, sensing, architecture, and efficient CAD methodologies. The trend toward increased quantity of embedded SRAM in scaled technology compounds the specific need of SRAM in low-power systems. Integrating more memory on chip provides an effective means to use silicon because of memory's lower power density, layout regularity, and performance and power benefits from reduced off-chip bandwidth. As a result, the ever increasing integration of embedded SRAM continues.

The workhorse of embedded memory is SRAM based on the 6T (six-transistor) cell. Each memory cell is associated with one or more WLs and one or more BLs. Generally, the bit cell supplies and well biases are globally connected to static voltage sources. During read, the WL voltage WL is raised, and the memory cell discharges either BL (bitline) or BLBAR (bitline complement), depending on the stored data on nodes port1 and port2. A sense amplifier converts the differential signal to a logic-level output. Then, at the end of the read cycle, the bls return to the positive supply rail. During write, WL is raised and the BLs are forced to either VDD or VSS (depending on the data), over powering the contents of the memory cell. During hold, WL is held low and the BLs are left floating or driven to VDD. The stability of the 6T memory cell can be verified by examining its butterfly curve, which contains the voltage transfer characteristics (VTCs) of the two inverters. The input output relations from PORT1 to PORT2 and from PORT2 to PORT1 are plotted on the same set of

axes, assuming the BLs are driven by DC voltage. During read or hold, three roots of intersection are desired, indicating bistability. During write, only one root of intersection is desired, so that the cell will deterministically flip to one of the two data states, as set by the BL polarity. The severity of local variation among devices in the memory cell, primarily from threshold voltage fluctuation in transistors, threatens the capability to preserve stability across 106 to 109 memory cells within one die.

Single-event-upsets from radiation also corrupt data in SRAM. When alpha particles from packaging materials or neutrons from space penetrate a silicon wafer, they can generate a charge that perturbs the state nodes of a memory element, causing it to flip. This failure rate increases with a reduction of supply voltage because of the decrease in stored charge on internal nodes. To address these soft errors, an SRAM can be protected with an error correcting code (ECC), requiring extra memory bits in each word and adding latency to both write access (for encoding) and read access (for detection and correction). If more than 1 bit must be corrected, ECC complexity increases significantly. Therefore, multi bit errors from soft-error phenomena are avoided by interleaving multiple words onto the same physical row. For example, a row of 128 adjacent bits comes from eight 16-bit words interleaved. Therefore, physical multi bit errors show up as multiple single-bit errors in different words.

When applying the dynamic voltage control to SRAM, the difficulty of SRAMs operating at low supply voltage comes from the degradation of access speed due to a lack of driving current. This is based on the fact that the transistor delay increases exponentially as the lowered supply voltage. The other obstacles to low voltage operation include instability of SRAM cells such as an increase in soft-error rate (SER) and reductions of static noise margin (SNM). Although the conventional approach boosting the supply voltage in memory cells improves the deteriorated SNM it cannot avoid the issue of exponential increase in the write/read access time in low voltage operation.

To summarize, the primary design challenge before the designer is to devise a topology based mechanism for reduction of leakage current when the device is in standby mode. Many such techniques have the potential to be applied however, the area requirements pose a serious constraint to the application of any such techniques where the number of components

in a particular device are likely to be increased. This situation demands careful sizing of individual components and application of effective power and area management techniques.

With a special emphasis on SRAM cells the specific aspects of this thesis keeping in view the above mentioned design challenges can be enlisted as follows.

1.6 Specific Aspect of the thesis:

- The present work puts forth a topology based solution for leakage current minimization problem. This is in contrast to various fabrication based methodologies which require a greater amount of time, effort and investment.
- Substantial reduction in standby leakage current has been obtained using a couple of simple techniques with only a few extra transistors for a given cell.
- The present work also leaves a lot of scope for future improvements related to sizing and area optimization of sleep transistors so that the power area constraints could be taken care of.

CHAPTER 2

Leakage Current Components Inside an SRAM Cell

2.1 Dominant leakage mechanism in CMOS Transistor:

There are four main source of leakage current in a CMOS transistor as shown in figure:

1. Reverse biased junction leakage(Irev)
2. Gate induced drain leakage(I_{GIDL})
3. Gate direct tunneling leakage(I_G)
4. Subthreshold leakage(I_{sub})

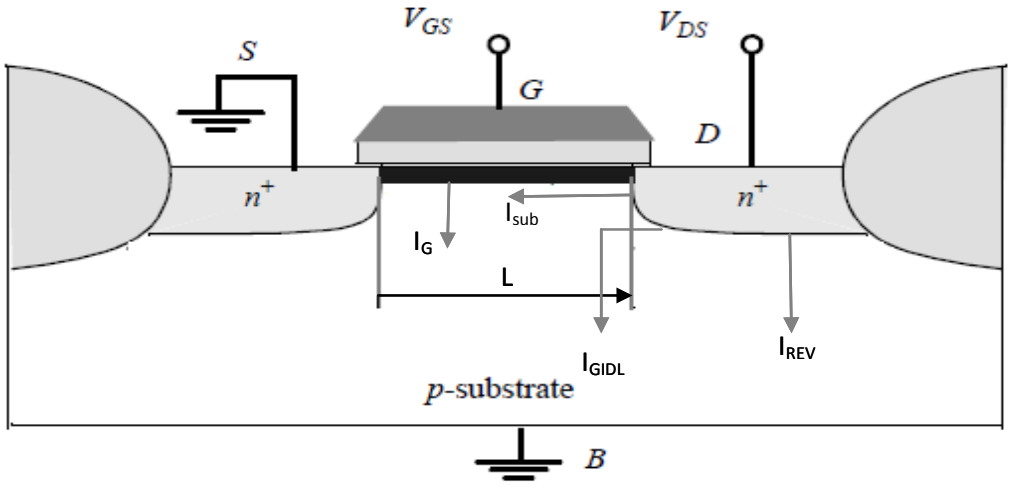

Figure 2.1 : Main source of leakage current

2.1.1 Junction leakage: The junction leakage occurs from the source or drain to the substrate through the reverse biased diodes when a transistor is OFF. A reverse-biased P-N junction

leakage has two main components: one in minority carrier diffusion/drift near the edge of the depletion region, the other is due the electron-hole pair generation in the depletion region of the reverse biased junction. For example in the case of inverter with a low input voltage, the NMOS is OFF, PMOS is ON and the output voltage is high. Subsequently the drain to substrate voltage of the OFF NMOS transistor is equal to the supply voltage. This result in a leakage current from the drain to the substrate through the reverse-biased diode. The magnitude of the diode leakage current depends on the area of the drain diffusion and the leakage current density which is in turn determined by the doping concentration.

If both the n and p regions are heavily doped, band to band tunneling (BTBT) dominates the p-n junction leakage. Junction leakage has a very high dependency on the temperature. However junction reverse leakage components from the both source drain diodes and the well diodes are generally negligible with respect to the other three leakage components.

2.1.2 Gate Induced Drain Leakage: The gate induced drain leakage (GIDL) is caused by high field effect in the drain junction of MOS transistors. For a NMOS transistor with grounded gate and drain potential at VDD, a significant band bending in the drain allow electron-hole pair generation through avalanche multiplication and the band to band tunneling. A deep depletion condition is created since the holes are rapidly swept out to the substrate. At the same time electrons are collected by the drain, resulting in GIDL current. This leakage mechanism is made work by high drain to body voltage and high drain to gate voltage.

Transistor scaling has led to the increasingly steep halo implants, where the substrate doping at the junction interface is increased, while the channel doping is low. This is done mainly to control punch-through and drain-induced barrier lowering while having a low impact on the carrier mobility in the channel. The resulting steep doping profile at the drain edge increases band to band tunneling current there, particularly a V_{DB} is increased. Thinner oxide and higher supply voltage increases GIDL current.

2.1.3 Gate Direct Tunneling Leakage: With scaling of the channel length, maintaining good transistor aspect ratio by the comparable scaling of the gate oxide thickness, junction depth and depletion depth are important for ideal MOS transistor behavior. Unfortunately

with the technology scaling, maintaining good transistor aspect ratio has been a challenge. In other words, reduction of the vertical dimensions has been harder than that of horizontal dimensions with the silicon oxide gate thickness approaching scaling limits there is now a rapid increase in gate direct tunneling leakage current.

Due to quantum mechanical and poly silicon gate depletion effects, both the gate charge and inversion layer charge will be located at the finite distance from the oxide channel interface with the charge location being a strong function of bias applied to the gate. The location of the inversion layer in the silicon substrate for a transistor with a typical bias when quantum mechanical effects are taken into account is 1 nm from the oxide channel interface. This increase the effective oxide thickness by 0.3 nm. Taking charge spread on the both sides of the interface along with poly depletion charge the 1nm oxide tunneling limit into an effective oxide thickness of 1.7 nm.

To compact this limit researchers have been exploring several alternatives including the use high permittivity gate dielectric, metal gate, novel transistor structure and circuit based techniques. The use of high permittivity gate dielectric will result in thicker and easier to fabricate dielectric for iso- gate oxide capacitance with potential for significant reduction in gate leakage. Identification of a proper high permittivity dielectric material that has good interface state with silicon along with limited gate leakage is in progress. However it has also been shown that use of high permittivity gate dielectric has limited value. In addition novel transistor structure such as self aligned double gate. FinFET and tri-gate MOS transistors that promise better aspect ratio are being explored.

2.1.4 Subthreshold Leakage: The subthreshold leakage is current flowing from drain to source when a transistor operated in weak inversion region. Unlike the strong inversion region in which the drift-current dominates, the subthreshold conduction is due to the diffusion current of the minority carriers in the channel for a MOS device. For instance, in case of an inverter with a low input voltage, the NMOS is turned off and the output voltage is high. In this case, although V_{GS} is 0V, there is still a current flowing in the channel of the OFF NMOS transistor due to the V_{DD} potential of the $V_{DS.}$ The magnitude of the sub threshold current is a function of temperature, supply voltage, device size and the process parameter out of which the threshold voltage plays a dominant role.

For the current CMOS technologies, the sub threshold leakage current, I_{SUB} is much larger than the other leakage current components. This is much larger than the other leakage current components. This is mainly because V_T is lower in modern device.

$$I = \left(\frac{W}{L}\right) \mu V_{th}^2 C e^{\left(V_{GS} - V_{th} + \frac{\eta VDS}{Vth}\right)} \left(1 - e^{-\frac{V_{DS}}{V_{th}}}\right) \quad \text{(1.1)}$$

Equation describes sub threshold current in terms of other device parameters

Here W and L denotes the transistor width and length, μ denotes the carrier mobility

$$V_{th} = \frac{KT}{q} \quad \text{(1.2)}$$

Vth is the thermal voltage

η is given by:-

$$\eta = 1 + \frac{C}{C_{ox}} \quad \text{(1.3)}$$

Here Cox denotes the gate input capacitance per unit area,C is the sum of depletion region capacitance and interface trap capacitance.

2.2 Leakage mechanism inside an SRAM

An SRAM cell is in the inactive state, when the word line is held low and bit line is charged to VDD. These inactive states come in between read and write operations. In the inactive state, different transistors dissipate leakage power depending on the value stored in the cell. This leakage current primarily owes its origin to two dominant leakage mechanism viz., sub-threshold leakage and gate leakage. Major contributors to the gate leakage current are gate oxide tunneling and injection of hot carrier from substrate to the gate oxide. Gate-induced drain leakage (GIDL) is another significant leakage mechanism, resulting due to the depletion at the drain surface below the gate-drain overlap region. Due to the substantial

increase in the leakage current, the static power consumption is expected to exceed the switching component of the power consumption unless effective measures are taken to reduce the leakage power. Due to adverse SCEs, the channel length cannot be arbitrarily reduced even if allowed by lithography. For digital applications, the most undesirable SCE is the reduced gate threshold voltage at which the device turns on, especially at high drain voltages. Therefore, to take the best advantage of the new high-resolution lithographic techniques, new device designs, structures, and technologies should be developed to keep SCEs under control at very small dimensions. In addition to gate oxide thickness and junction scaling, another technique to improve short-channel characteristics is well engineering. By changing the doping profile in the channel region, the distribution of the electric field and potential contours can be changed. The goal is to optimize the channel profile to minimize the off-state leakage while maximizing the linear and saturated drive currents. Super steep retrograde wells and halo implants have been used as a means to scale the channel length and increase the transistor drive current without causing an increase in the off-state leakage current.

Till recently, the drain source sub-threshold current had been thought to be the dominant leakage mechanism. A number of techniques have been proposed in the particular literature for reducing drain to source sub-threshold leakage when the SRAM is in the inactive state.

2.2.1 Sub threshold Leakage

Sub threshold leakage is the drain-source current of a transistor when the gate-source voltage is less than the threshold voltage (Figure 2.2 [22]). More precisely, sub threshold leakage happens when the transistor is operating in the weak inversion region. The sub threshold current depends exponentially on threshold voltage, which results in large sub threshold current in short channel devices. To reduce the sub threshold leakage of an SRAM cell, one can increase the threshold voltage of all or some of the transistors in the cell. The drawback of this technique is an increase in read/write delay of the cell. If the threshold voltage of the pull up PMOS transistors is increased, the write delay increases whereas 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 whereas the effect on the write delay would be marginal. By increasing the threshold voltage of the pass transistors both read and write delays increase. Due to the delay of sense amplifiers and output buffers in a read path, the write delay of an SRAM cell tends to be smaller than its read delay. Therefore, one can think of reducing the subthreshold leakage by increasing the threshold voltage of the PMOS transistors as long as the write delay is less than the read delay.

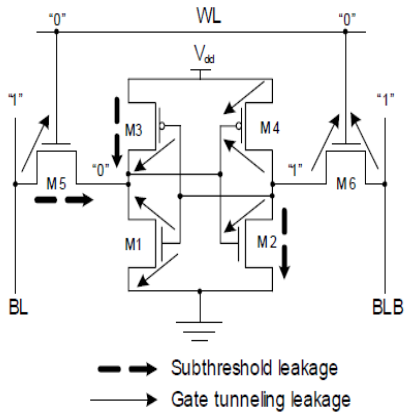

Figure 2.2 Subthresold Leakage

2.2.2 Gate Tunneling Leakage

Electrons (holes) tunneling from the bulk silicon through the gate oxide into the gate results in gate tunneling current in an NMOS (PMOS) transistor. Gate tunneling current is composed of three major components:

(1) Gate to source and gate to drain overlap current

(2) Gate to channel current, part of which goes to source and the rest goes to drain

(3) Gate to substrate current.

In bulk CMOS technology, the gate to substrate leakage current is several orders of magnitude lower than the overlap tunneling current and gate to channel current [22]. 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 condition. Since the gate to

source and gate to drain overlap regions are much smaller than the channel region, the gate tunneling current in the OFF state is much smaller than gate tunneling in the ON state [22].
If SiO2 is used for the gate oxide, PMOS transistors will have about one order of magnitude smaller gate leakage than NMOS transistors [22, 23]. Therefore, in an SRAM cell, the power saving achieved by increasing the oxide thickness of the PMOS transistors is marginal.
The subthreshold and gate tunneling leakage currents of an SRAM cell storing "0" are shown in Figure 2.3[22].

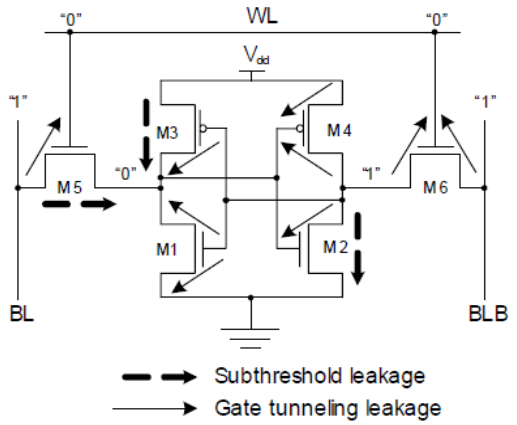

Figure 2.3: Gate Tunneling Leakage

2.3 Power consideration in SRAM cell: To reduce the power consumption in SRAMs all contributors to the total power must be targeted. The most efficient techniques used in recent memories are:-

1. Capacitance reduction of word-lines and the number of cells connected to them, data lines,1/0 lines and decoders.
2. DC current reduction by using new pulse operation techniques for word-lines, periphery circuits and sense amplifiers.
3. AC current reduction by using new decoding techniques (i.e. multi-stage static CMOS decoding)
4. Operating voltage reduction.
5. Leakage current reduction (in active and standby mode) by utilizing multiple threshold voltage (MT-CMOS) or variable threshold voltage technologies (VT-CMOS).

2.3.1 Capacitance reduction

The largest capacitive elements in a memory are word-line, bit lines and dateline each with a number of cells connected to them. Therefore, reducing the size of these lines can have a significant impact on power consumption reduction. A common technique used often in large memories is called Divided Word Line (DWL) which adopts a two-stage hierarchical row decoder structure as shown in Figure 2.4[24]. The number of sub-word lines connected to one main word line in the data line direction is generally four, substituting the area of a main row decoder with the area of a local row decoder. DWL features two-step decoding for selecting one word-line, greatly reducing the capacitance of the address lines to a row decoder and the word- line RC delay.

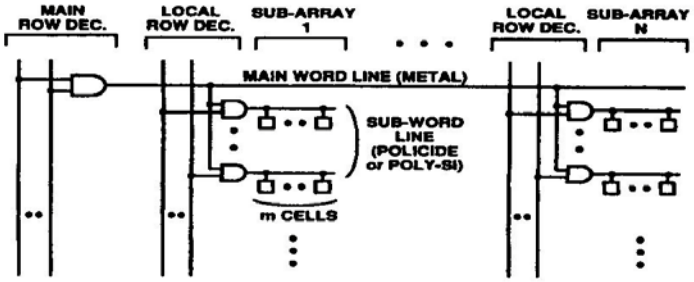

Figure 2.4: Divide word line structure

A single bit line cross-point cell activation (SCPA) architecture reduces the power further by improving the DWL technique .The architecture enables the smallest column current possible without increasing the block division of the cell array thus reducing the decoder area and the memory core area. The cell architecture is shown in Figure 2.5[25]. The Y-address controls the access transistors and the X-address. Since only one memory cell at the cross-point of X- and Y- is activated, a column current is drawn only by the accessed cell. As a result, the column current is minimized. In addition, SCPA allows the number of blocks to be reduced because the column current is independent of the number of block divisions in the SCP A. The disadvantage of this configuration is that during the write "high" cycle, both X- and Y-lines have to be boosted using a word-line boost circuit.

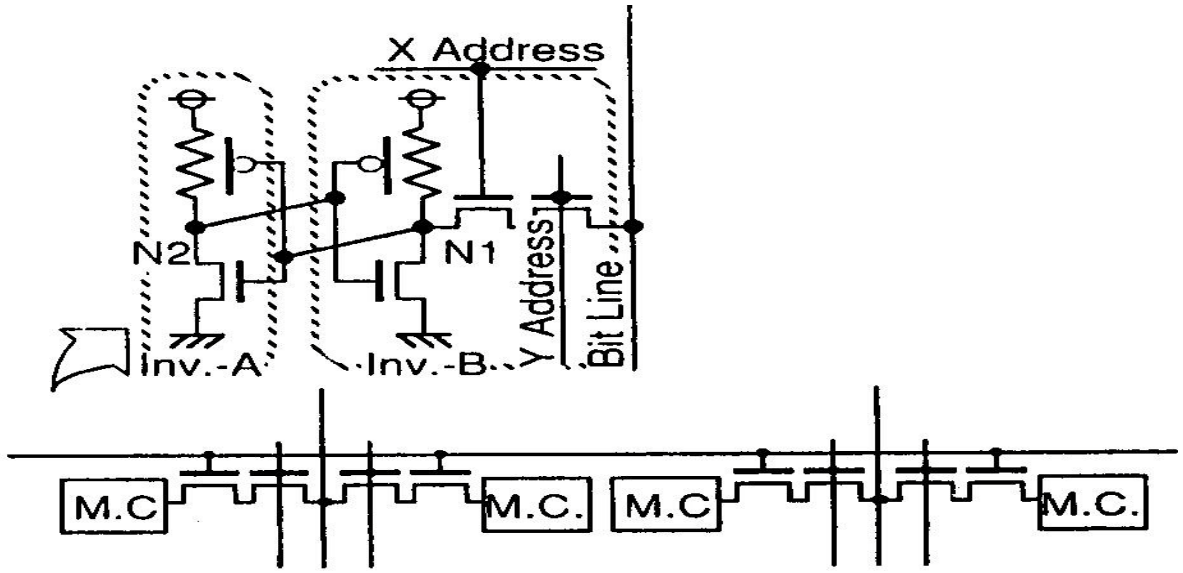

Figure 2.5: Memory cell used for SCPA Architecture

2.3.2 Pulse operation techniques

Pulsing the word-lines, equalization and sense lines can shorten the active duty cycle and thus reduce the power dissipation. In order to generate different pulse signals, an on-chip address transition detection (ATD) pulse generator is used. This circuit, shown in Figure 2.6 [26], is a key element for the active power reduction in memories.

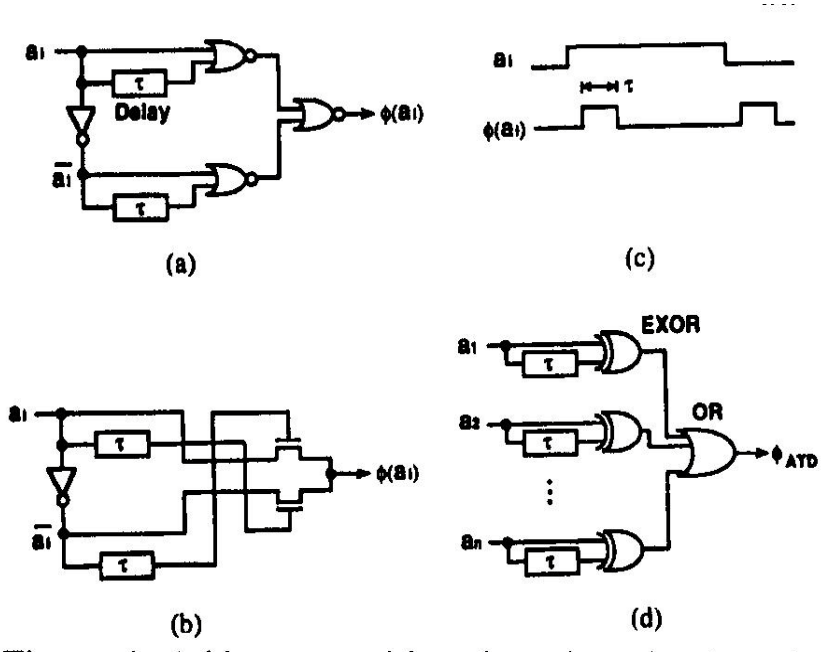

Figure 2.6: Address transition detection circuit

The principle of the operation is in a merger of a voltage-level converter with a logical AND. A positive half-swing (transitions from a rest state $V_{dd}/2$ to V_{dd} and back to $V_{dd}/2$) and a negative half-swing (transitions from a rest state $V_{dd}/2$ to Gnd and back to $V_{dd}/2$) combined with the receiver-gate logic style result in a full gate overdrive with negligible effects of the low-swing inputs on the performance of the receiver. This structure is combined with a self-resetting circuitry and a PMOS leaker to improve the noise margin and the speed of the output reset transition (Figure 2.7 [27]).

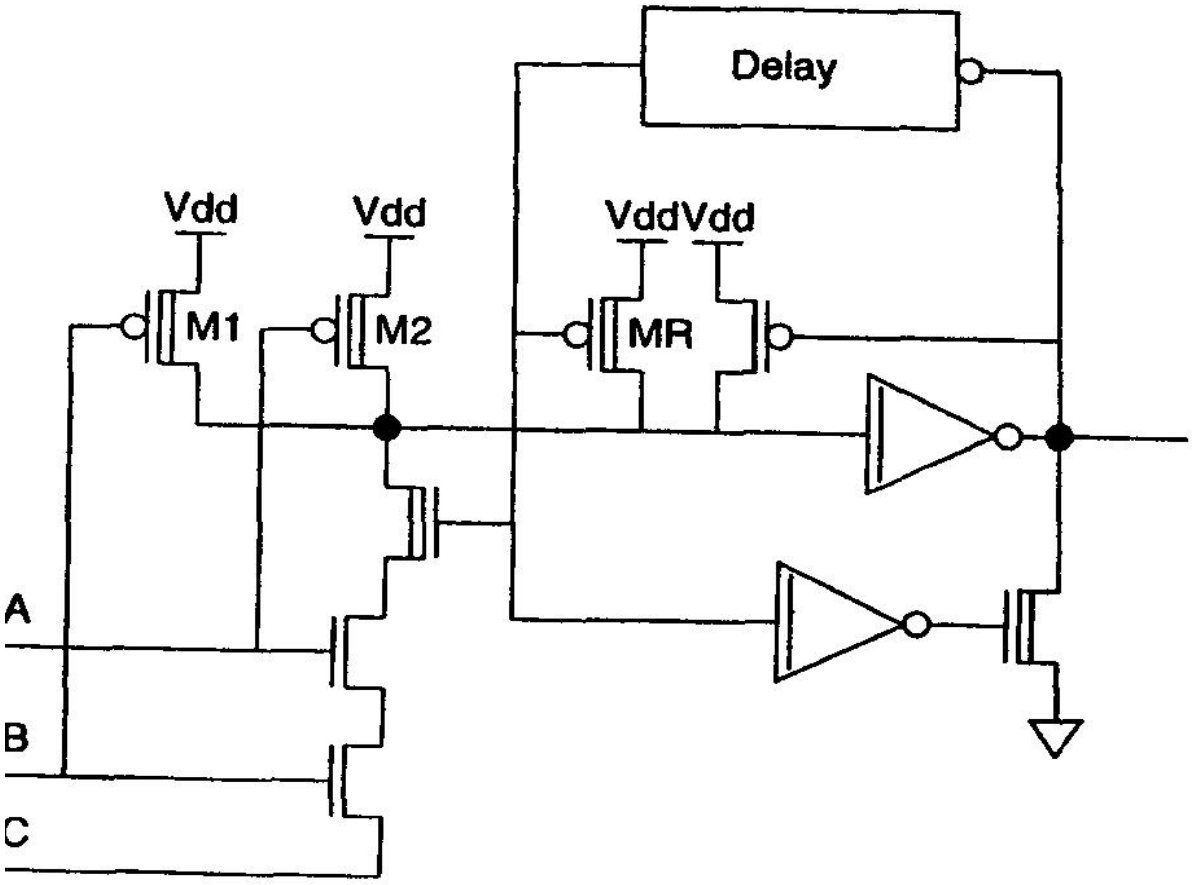

Figure 2.7: Self setting half swing pulse mode gate with a PMOS leaker

Both negative and positive half-swing pulses can reduce the power consumption further by using a charge recycling. The charge used to produce the assert transition of a positive pulse can also be used to produce the reset transition of a negative pulse. If the capacitances of positive and negative pulses match, then no current would be drawn from the power supply.

Combining the half swing pulse mode logic with the charge recycling techniques, 75% of the power on high-capacitance lines can be saved.

2.3.3 AC current reduction

One of the circuit techniques that reduce AC current in memories is a multi-stage decoding. It is a common fact that as fast static CMOS decoders are based on decode architecture, the number of transistors, fan-in and the loading on the address input buffers is reduced (see Figure 2.8 [27]). As a result, both speed and power are optimized. The signal, generated by the ATD pulse generator, enables the decoder and secures pulse activated word-line. OR/NOR and AND/NAND architectures.

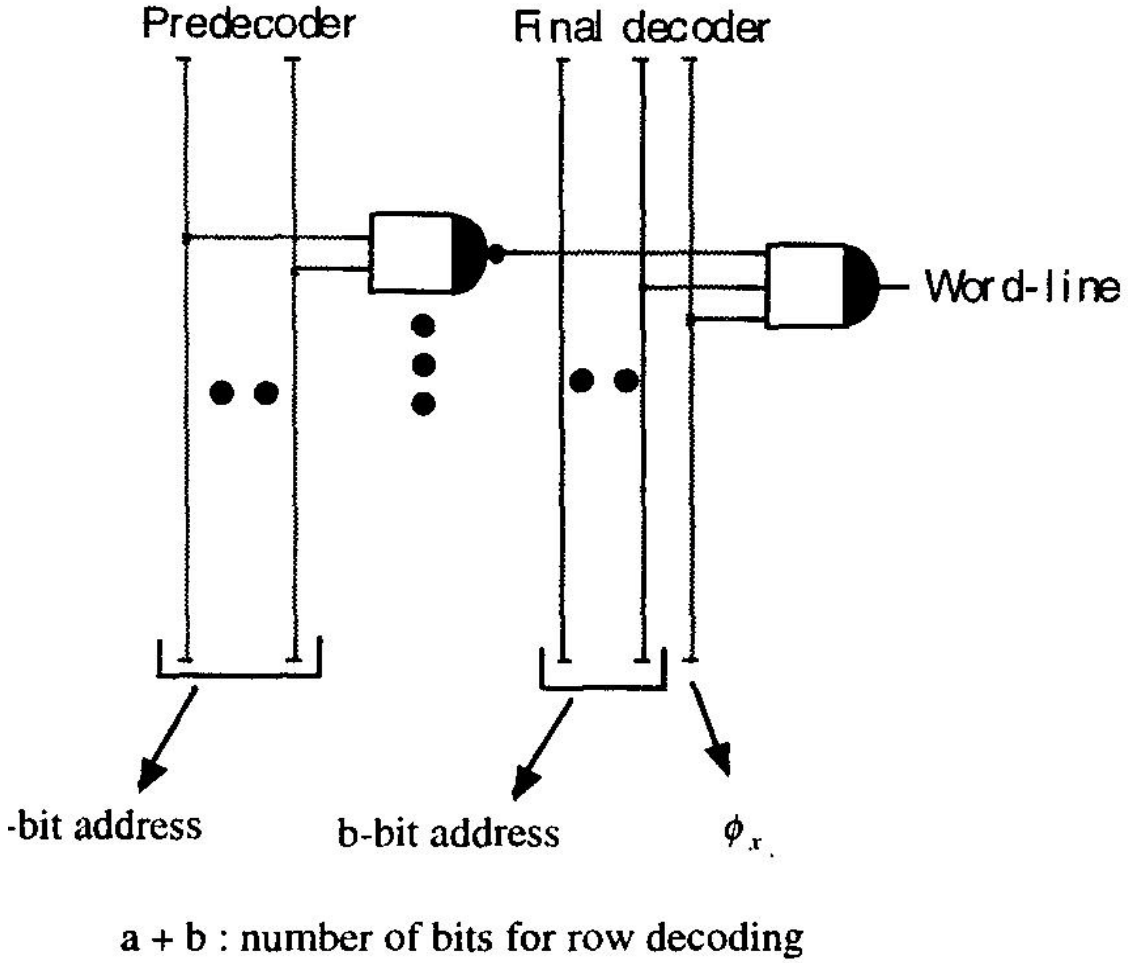

Figure 2.8 : A two stage decoder architecture

2.3.4 Operating voltage reduction:

A new current-mode sense amplifier for 1.5V power supply was proposed by Wang and Lee. The new circuit overcomes the problems of a conventional sense amplifier with pattern dependency by implementing a modified current conveyor. Pattern dependency problem limits the scaling of the operating voltage. Also, the circuit does not consume any DC power because it is constructed as a complementary device. As a result, the power consumption is reduced by 61-94% compared with a conventional design. The circuit structure of the modified current conveyor is similar to a conventional current conveyor design. However, an extra PMOS transistor MP7, as seen in Figure 2.9 [28], is used. The transistor is controlled by RX signal (a complement of CS). After every read cycle, transistor MP7 is turned on and equalizes nodes RXP and RXN which eliminates any residual differential voltage between these two nodes (limitation in conventional designs).

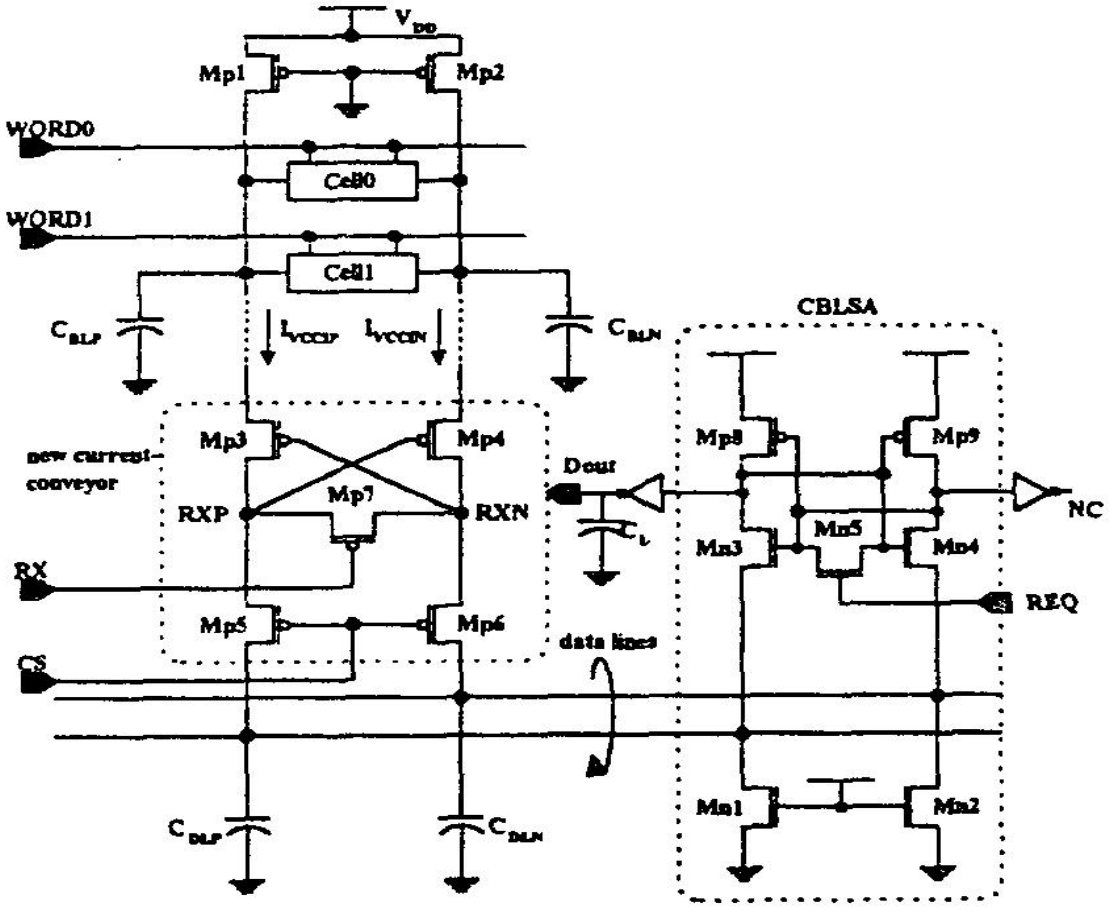

Figure 2.9 : SRAM read circuitry with a new current mode sense amplifier

2.3.5 Leakage current reduction

Power consumption has two aspects:

Dynamic power – The power that is consumed by a device when it is actively switching from one state to another. Dynamic power consists of switching power, consumed while charging and discharging the loads on a device, and internal power (also referred to as short circuit power), consumed internal to the device while it is changing state

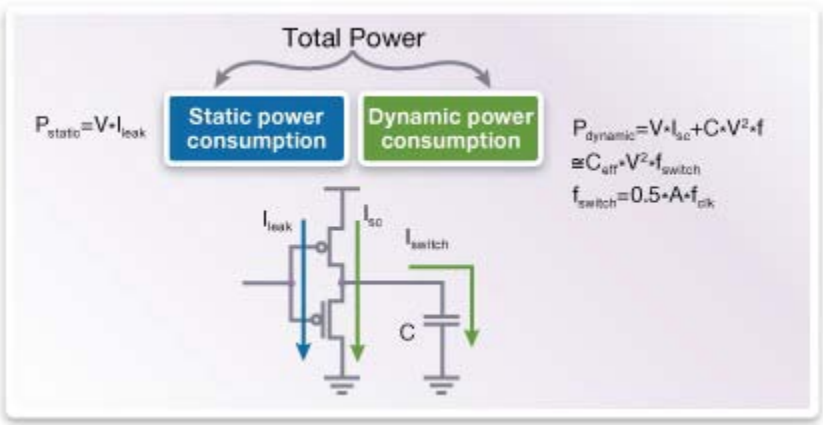

Figure 2.10: Total power

Leakage power – The power consumed by a device not related to state changes (also referred to as static power). Leakage power is actually consumed when a device is both static and switching, but generally the main concern with leakage power is when the device is in its inactive state, as all the power consumed in this state is considered “wasted” power.

Various techniques have been developed to reduce both dynamic and leakage power. The two most common traditional, mainstream techniques are:

2.3.5.1 Clock gating – The disconnecting of the clock from a device it drives when the data going into the device is not changing. This technique is used to minimize dynamic power [29].

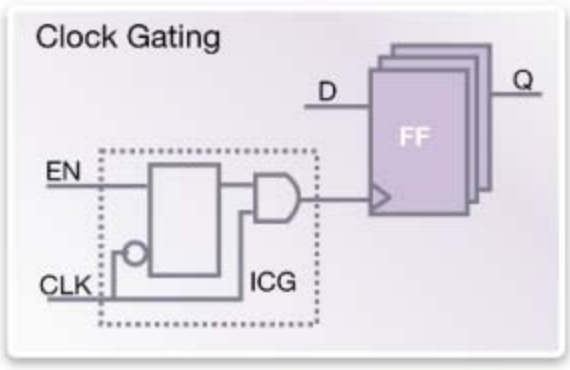

Figure: 2.11 Clock gating

2.3.5.2 Multi-Vth optimization – The replacement of faster Low-Vth cells, which consume more leakage power, with slower High-Vth cells, which consume less leakage power. Since the High-Vth cells are slower, this swapping only occurs on timing paths that have positive slack and thus can be allowed to slow down(Fig. 2.12 [29]).

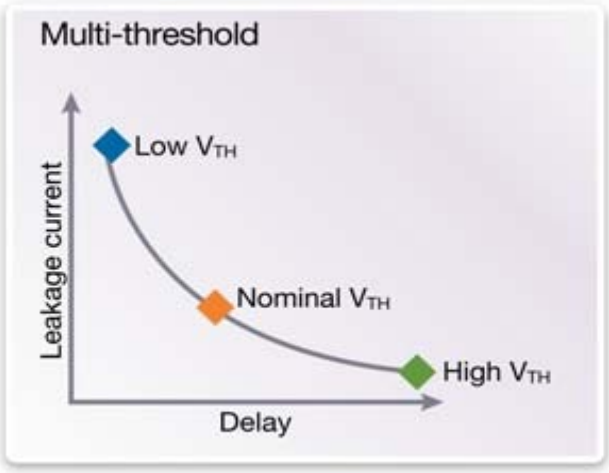

Figure: 2.12 Multi V_t optimization

As technologies have shrunk, leakage power consumption has grown exponentially, thus requiring more aggressive power reduction techniques to be used. Similarly, clock frequency increases have caused dynamic power consumption of the devices to outstrip the capacity of the power networks that supply them, and this becomes especially acute when high power consumption occurs in very small geometries, as this is a power density issue as well as a power consumption issue.

2.3.5.3Multi-Voltage

Multiple voltage rails (multi-Vdd) can be supplied to a design to impact power and performance. A higher voltage yields a faster the circuit, but with higher the dynamic power. In many designs, only discrete portions of the design need to run at high speed. Other portions may only operate at lower speeds, and thus require lower voltages (and therefore consume less power) [30].

2.3.5.4 Dynamic Voltage and Frequency Scaling (DVFS)

Modifying the operating voltage and/or frequency at which a device operates, while it is operational, such that the minimum voltage and/or frequency needed for proper operation of a particular mode is used.

2.3.5.5 Well Biasing

Separate voltage supplies can be used to connect to the NMOS and PMOS bulk regions in triple well CMOS technologies. Modification of these voltages with respect to the primary power and ground supplies is called well-biasing. These supplies can be modulated to provide a back-bias voltage which causes an increase in device Vth, reducing the sub-threshold leakage. These supplies can also be modulated in the reverse direction to provide a forward-bias voltage which causes a decrease in device Vth that increases the speed at which the transistors switch, at a cost of increased sub-threshold leakage. Thus, well-biasing can be used to directly adjust between high performance and low power consumption.

Chapter 3

Upper Supply Reduction Scheme (USRS) for Reduction in Leakage

3.1 Topology and current reduction: In the present work I propose three leakage reduction schemes with the use of self controlled switches .The first scheme involves placing of switch at the upper end of the cell to reduce supply voltage and will now be termed as Upper supply reduction scheme (USRS). Whereas in the second scheme the switch is placed at the lower end of the cell to raise the potential of the ground node and will be termed as Lower potential raising scheme(LPRS). In addition to these two techniques, a modified version of the USR scheme has also been proposed in which the bit lines are made floating during the standby mode. The impact of USRS and its variant on leakage currents is described in the next two subsections.

3.2 Leakage control using USRS: An SRAM cell incorporating a USRS scheme is shown in figure 3.1, along-with its impact on leakage currents flowing through different transistors. When the SRAM is in active mode, a full supply voltage is applied. However the supply voltage level is applied to SRAM is reduced to voltage level V_d in inactive mode. Since transistor M4 is in on state, voltage at the drains of M2 and M4 is also reduced. As before let us consider first the impact on gate leakage current. Since the gate voltage of transistor M1 is decreased, gate leakage current through it is also reduced. A decrease in drain voltage of transistor M2 results in lower gate to drain voltage across it and thus gate leakage current through it is also reduced .A decrease in source voltage of M6 results in a decrease in one component of EDT leakage across it while leaving the other unchanged. Gate leakage across transistor M5 remains unchanged. Transistor PU1 being a PMOS transistor does not result in any significant added leakage current as a result of transistors used in USPS circuit. One can thus see that USRS scheme has a better impact on gate leakage current reduction than LPRS scheme. However this scheme is inferior with respect to sub-threshold leakage current. While, sub-threshold leakage through transistors M3 and M2 is reduced, leakage across

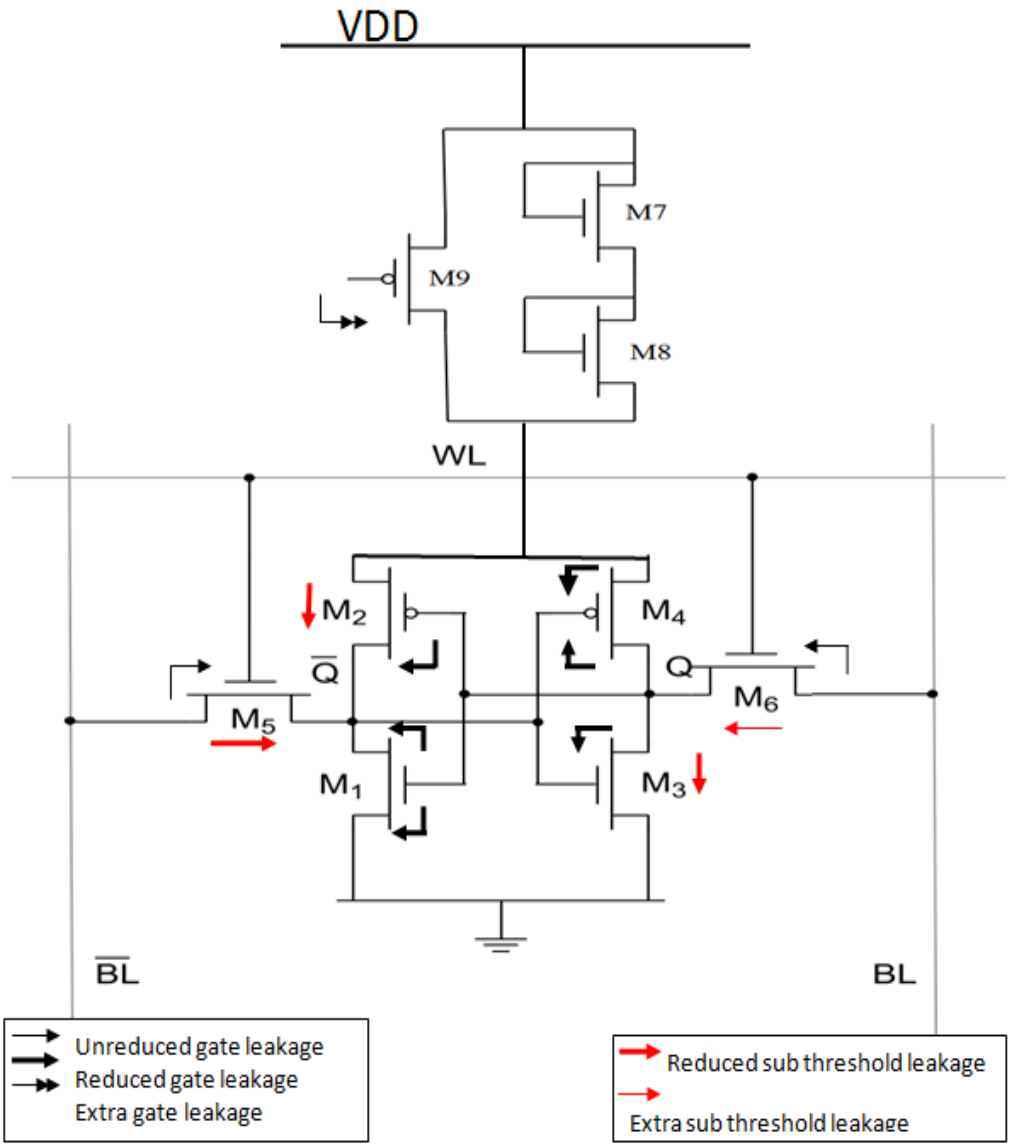

Figure 3.1: USR Scheme

transistor M5 remains unaltered. Also, a new sub-threshold leakage current appears in transistor M6 as a result of reduction in its source voltage.

3.2.1 Effect on subthreshold leakage: As we know that there are two techniques that we apply here to reduce leakage currents in SRAM cell. Out of these two techniques first we discuss USR scheme.USRS stand for Upper Supply Rejection Scheme. As we discussed earlier in USR scheme we place three transistors above the bit cell in which one transistor is

operated through clock. When SRAM cell is in off mode then we provide a alternative path from the bit cell to power supply. This path is provided by off the clock when Cell is neither in ON state nor in off mode or we can say that transistor is in standby mode. During the idle mode, different transistors in the cell dissipate leakage power depending on the value stored in the cell. Let us consider different individual cases:

Case 1: when '0' is stored. As shown in figure there is a significant gate leakage current through N type transistor M2, M5 and M6 and the mechanism is primarily is edge direct tunneling (EDT).Although a similar mechanism operates in transistors M3 as well, the gate leakage here is negligible because of its P-type nature. Gate leakage is maximum in transistor M1 and the mechanism here is EDT. Similarly, leakage current in M4 is relatively insignificant due to its P-type nature.

Sub-threshold current originates in transistors that are in OFF state and that include transistors M2 & M3 in the cross-coupled inverter pair and access transistors M5and M6.Except for the access transistor M6 whose drain source voltage is zero, all three transistors have significant sub-threshold leakage currents.

To summarize there are three sub-threshold leakage current components through M2, M3 and M5. A leakage current strategy should thus address all these leakage current components.

To summarize, the USRS approach, while more successful in reducing gate leakage current, still leaves two gate leakage component in access transistors unaltered. It also leaves one sub-threshold component in access transistor unchanged and results in an additional sub-threshold leakage current across the other access transistor. However there is an overall improvement in performance while using USRS scheme as compared to LPRS scheme.

3.2.2 Effect on gate leakage: As we discuss earlier, during the idle mode, different transistors in the cell dissipate leakage power depending on the value stored in the cell. Let us consider different individual cases:

As we know that in USRS topology we add three transistors to reduce the supply voltage during idle or standby mode. In write or read operation there is no reduction in supply voltage. Due to this reason we get the correct bit value stored on ports. As we see from

simulation diagram the gate leakage in access transistor is unchanged. In transistors M1, M2, M3 and M4 there is a significant reduction in gate leakage.

3.3 USR scheme with floating bit lines:

The USR scheme though considerably successful in reducing gate leakage current, still leaves two gate leakage current components in access transistors unaltered. It also leaves one Sub threshold current component in access transistor unchanged and results in an additional subthreshold leakage current across the other access transistor. One thus notes that leakage currents in access transistors are not adequately addressed through the USRS approach. According to authenticated reports, leakage current through access transistors contributes about 20% of the total OFF state leakage in an SRAM. Furthermore, this estimated 20% contribution to the OFF state leakage is obtained when only subthreshold leakage is considered to be important. The importance of leakage current through access transistor becomes even greater when gate leakage current is considered along with subthreshold leakage current. As a result, a leakage current reduction scheme must address currents through access transistors as well. These currents can be significantly reduced if along with reduction in supply voltage to the cross coupled inverter pair, the bit lines are made floating by putting the pre-charge transistors in cut-off. A mechanism for implementing this idea has been shown in fig.3.2

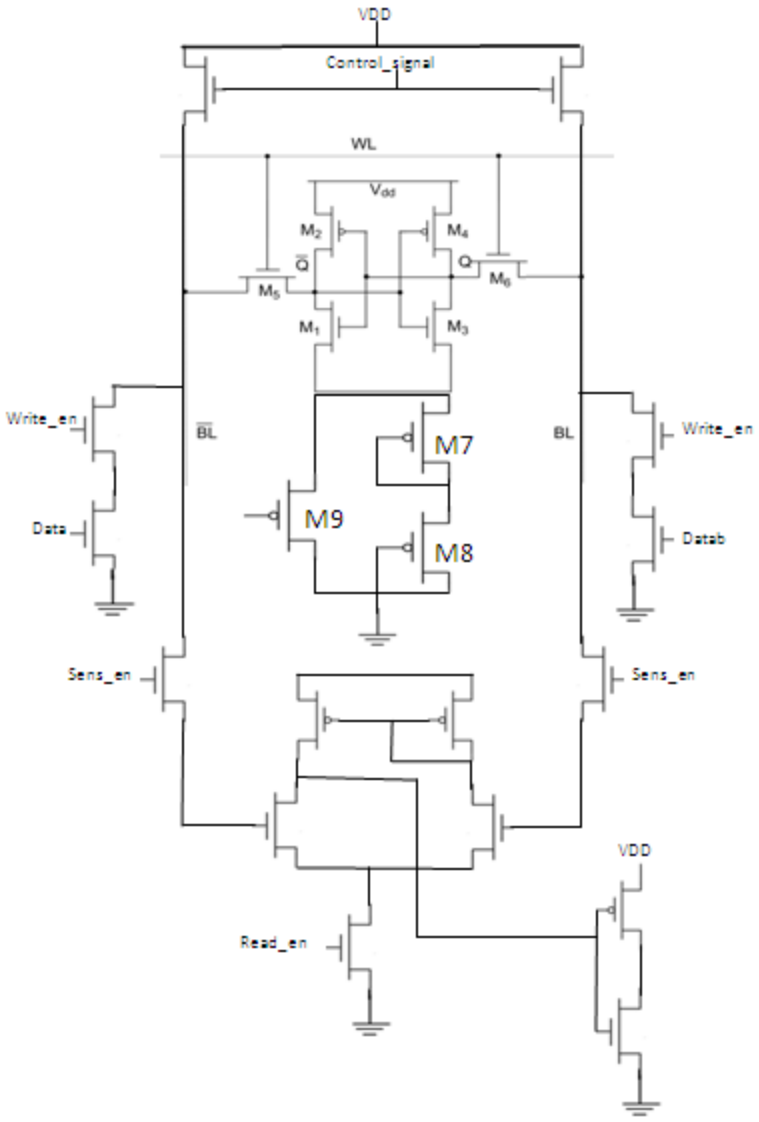

Figure 3.2 USR-A scheme

Let us call this scheme USRS-A, to differentiate it from the USRS discussed at the beginning of this chapter. This scheme can be used to lower the subthreshold leakage in access transistors along with that in SRAM latch, by gating the pre-charge signal with the wakeup signal. In the active mode, the pre charge transistors will work on full Vdd supply voltage.

The bit line voltages are made floating by putting the pre-charge transistors in the OFF state in the inactive mode. In this way, the leakage current is reduced. Individual reduction in each transistor has been depicted in Fig 3.2.The simulation results are shown in chapter 5.

CHAPTER 4

Lower Potential Raising Scheme (LPRS) for Reduction of Leakage

4.1 Topology and current reduction mechanism: As discussed earlier there are three methods that have been used in the present work to reduce leakage current in standby mode. Out of the three techniques one has been discussed in the previous chapter. Here we discuss the second techniques called LPRS scheme. In this technique we raise the ground potential to a some a extent in standby mode. Due to rise in ground potential overall swing will be reduced so leakage current reduced. As we also know that leakage current also depends upon the value stored in cell. As we know that we are using 6-T SRAM cell. Every transistor has different leakage current in standby mode. Some transistor dissipates so much leakage power that will affect the circuit performance. So our aim to minimize that leakage power. In extra topology we add three transistors. Out of three transistors one transistor is driven by clock, which is used as a control switch. In read or write mode we will enable the clock so that there is a direct connection between ground supply and SRAM bit cell. However, if we disable the clock when SRAM is in active mode then there will be reduction in power supply due to which our stored value will be affected and we will not get the proper output. In standby mode we disable the clock so that there is an alternate path and all the current flows through that path. Due to this there is a reduction in input voltage swing and due to this alternate path there is a reduction in leakage current.

Figure 4.1shows a schematic of an SRAM cell in which LPRS is applied. The switch provides 0 volt at ground node during the active mode and raises ground level (virtual ground) during the inactive mode. This scheme is similar to the diode footed cache design scheme proposed to control gate and sub-threshold leakage in SRAM cell, in which a diode designed with high Vt MOS transistor was used to raise the ground level in inactive mode [Agrawal and Roy, 2003].

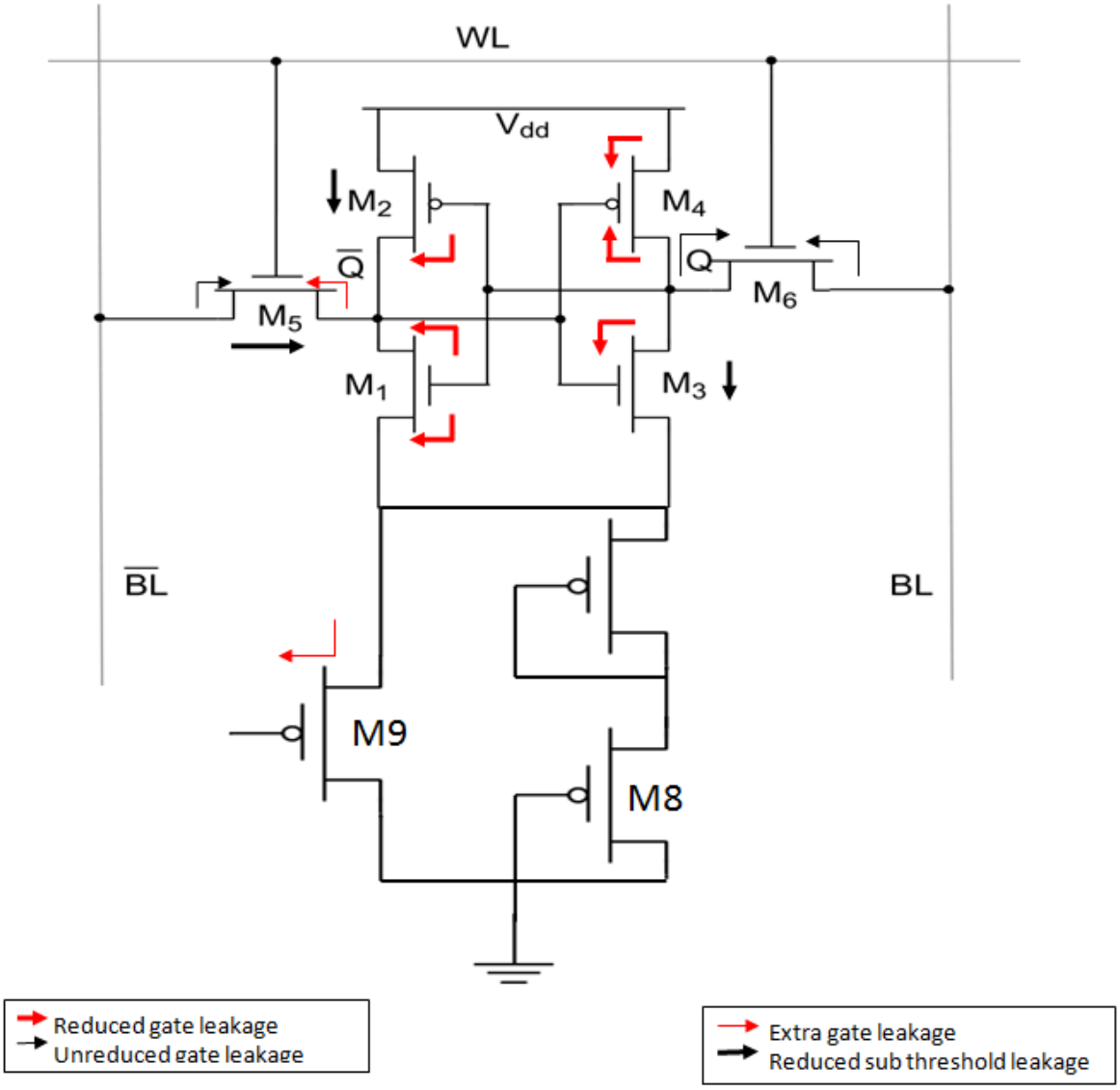

Figure 4.1: LPR scheme

4.2 Effect on gate leakage:

Let us consider the impact of this approach on gate leakage first. As we know that in this approach we increase the ground potential to increase to reduce the leakage current. An increase in the virtual ground voltage (V_S in fig.4.1), result in decrease of gate-source voltage and gate-drain voltage of transistor M1 and gate-drain voltage of transistor M2 and result in sharp reduction in gate leakage currents of two transistors. However there is no improvement

in gate leakage currents for transistors M5 and M6. We know that M5 and M6 are the access transistor. In fact as a result of increase in drain voltage of M1, a new gate leakage current appear in transistor M5 as indicated in Figure 4.1.In corporation of LPRS results in another new gate leakage current through NMOS transistor in MOS switch.

Although only one transistor is normally used for one bank of SRAM cells, leakage current through it is not necessarily negligible because its size has to be much larger than NMOS transistors within the SRAM cell to avoid performance degradation in the active state.

4.3 Effect on subthreshold leakage:

Sub threshold leakage is also an important component in leakage current. This leakage comes in picture when transistor is in off state and operated in sub threshold mode. If this leakage increases from its maximum value then it will affect the whole circuit performance and we will not be able to get a proper output. As we also know that this leakage current is also depend on the value stored in the SRAM bit cell. As for as sub-threshold leakage current are concerned, LPRS approach is successful in reducing currents through M3, M2 and M5 as well. To summarize it should be added that while all sub-threshold currents are reduced using LPRS approach, it is the only partially successful in reducing gate leakage current.

CHAPTER 5

Simulation Results &Performance Comparison of Proposed Topologies

5.1 Performance comparison

This chapter summarizes the simulation results verified for 10 cycles of 'Read' and 'Write' operations with the conventional 6 transistor SRAM cell and SRAM cells incorporating the proposed leakage reduction schemes. The simulations have been carried out on Cadence Design environment platform. Figs. 5.1, 5.7 and 5.14 are the optimized schematics for conventional 6 T SRAM cell, and cells incorporating USRS and LPRS respectively. Furthermore, the rest of the figures are tool generated waveforms for currents in the individual transistors and bit line and word line signals.

The leakage currents in the conventional and the schemes suggested in previous chapters are shown in Table 5.1. Although from the waveforms it is evident that current in an individual transistor fluctuate during a 'Read' 'Write' cycle, in this case, only average value of current in standby mode has been measured and reported. Compared to the conventional 6 T SRAM cell, LPRS suppresses the total leakage by 59.8%. While USR scheme without changing the bit line voltages provides a leakage reduction of 69.7%. However, USR-A scheme does not give any significant advantage over the conventional SRAM cell. One of the possible reasons for this is the large current swing obtained when the pre-charge transistors are cut off from the supply. The efforts to minimize this large current swing would comprise the task for the future workers in the field. However, it can be safely presumed that both USRS and LPRS have served to reduce both gate and sub threshold leakages substantially.

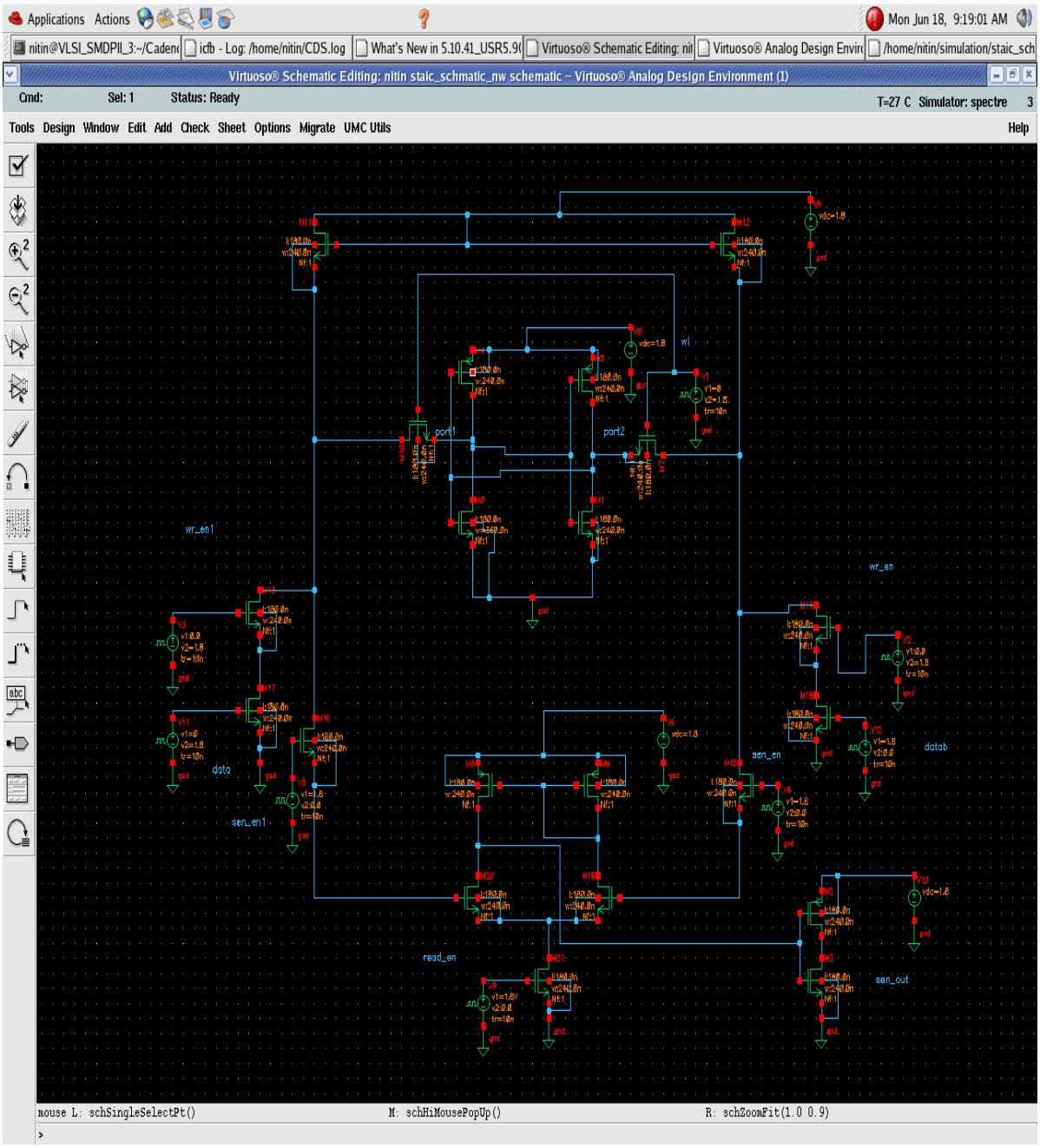

Figure 5.1: 6-T Static RAM cell schematic with sense amplifier

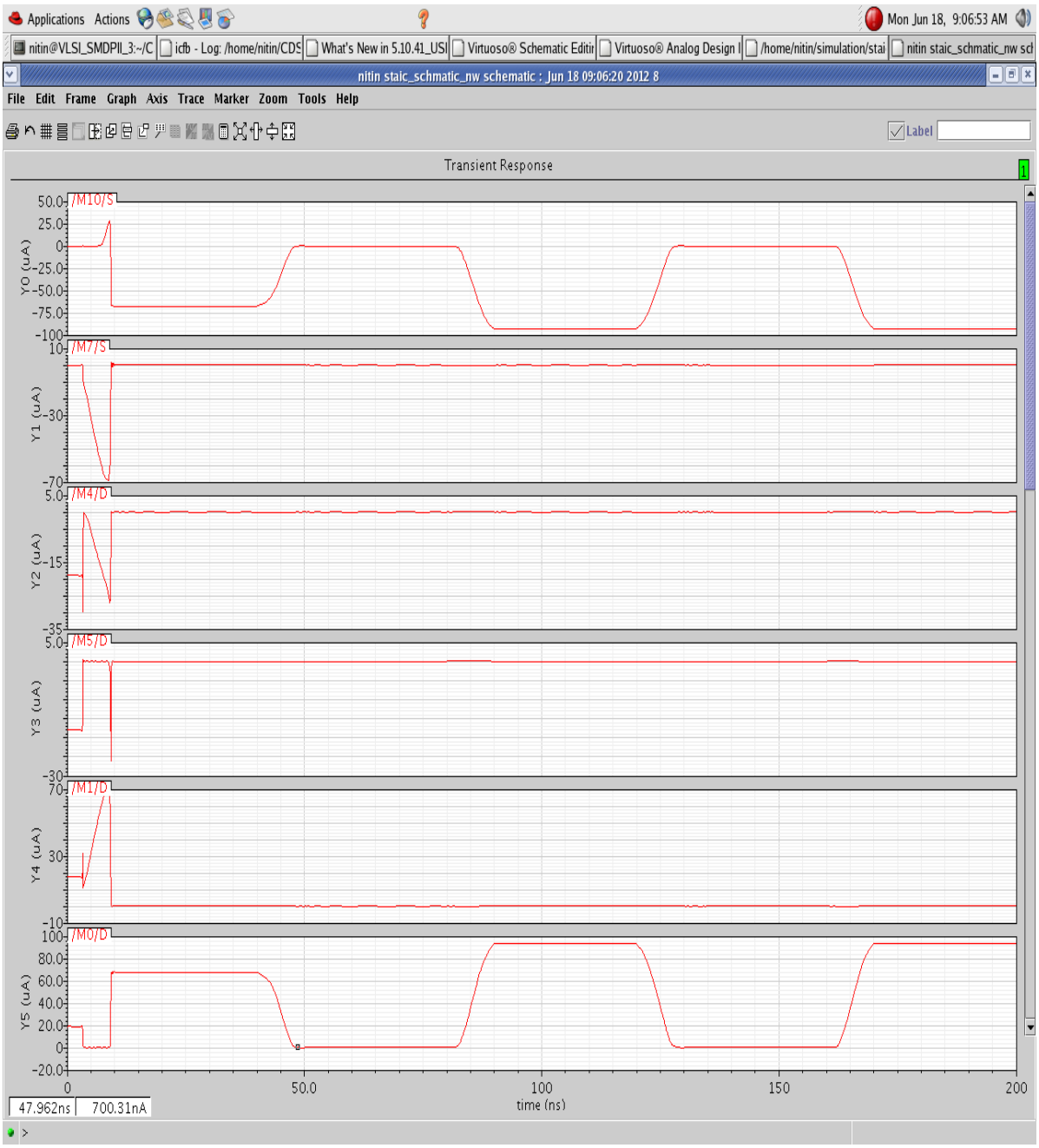

Figure 5.2: Current waveforms of Read '0' operation (for conventional 6-T cell)

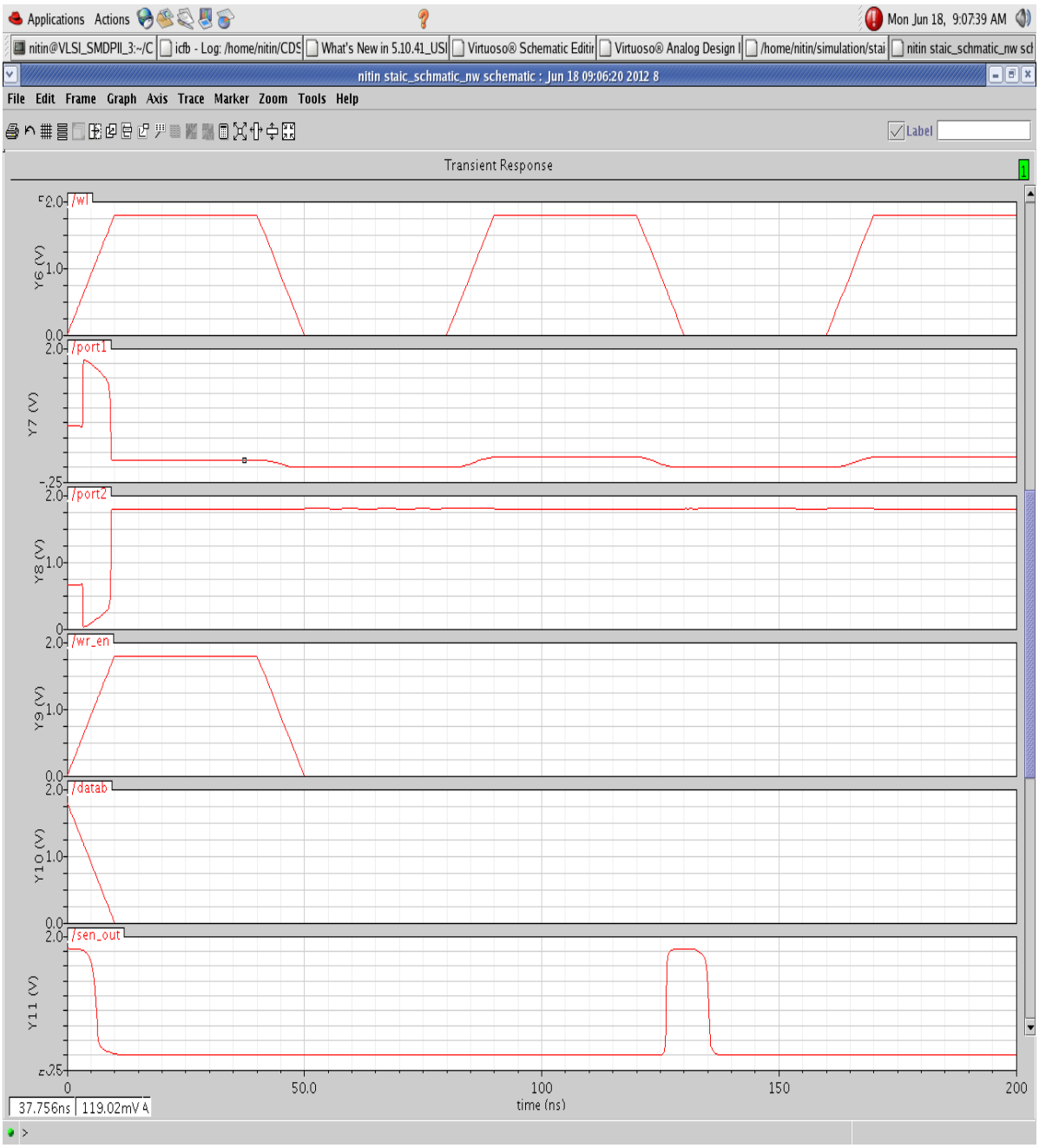

Figure 5.3:Word line and bit line waveform for Read ‘0’ operation

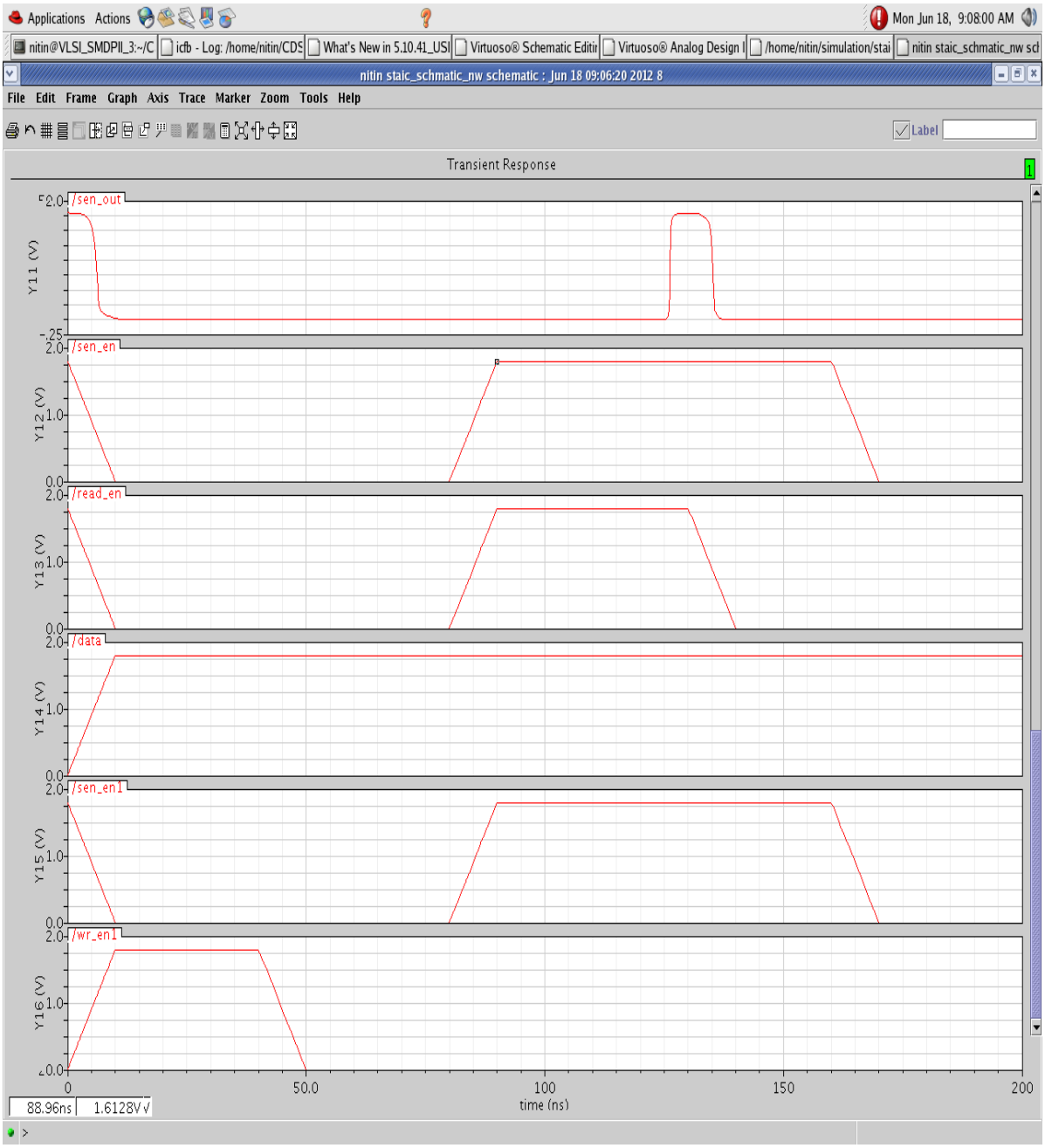

Figure 5.4 control signal and data waveforms for Read ‘0’ operation

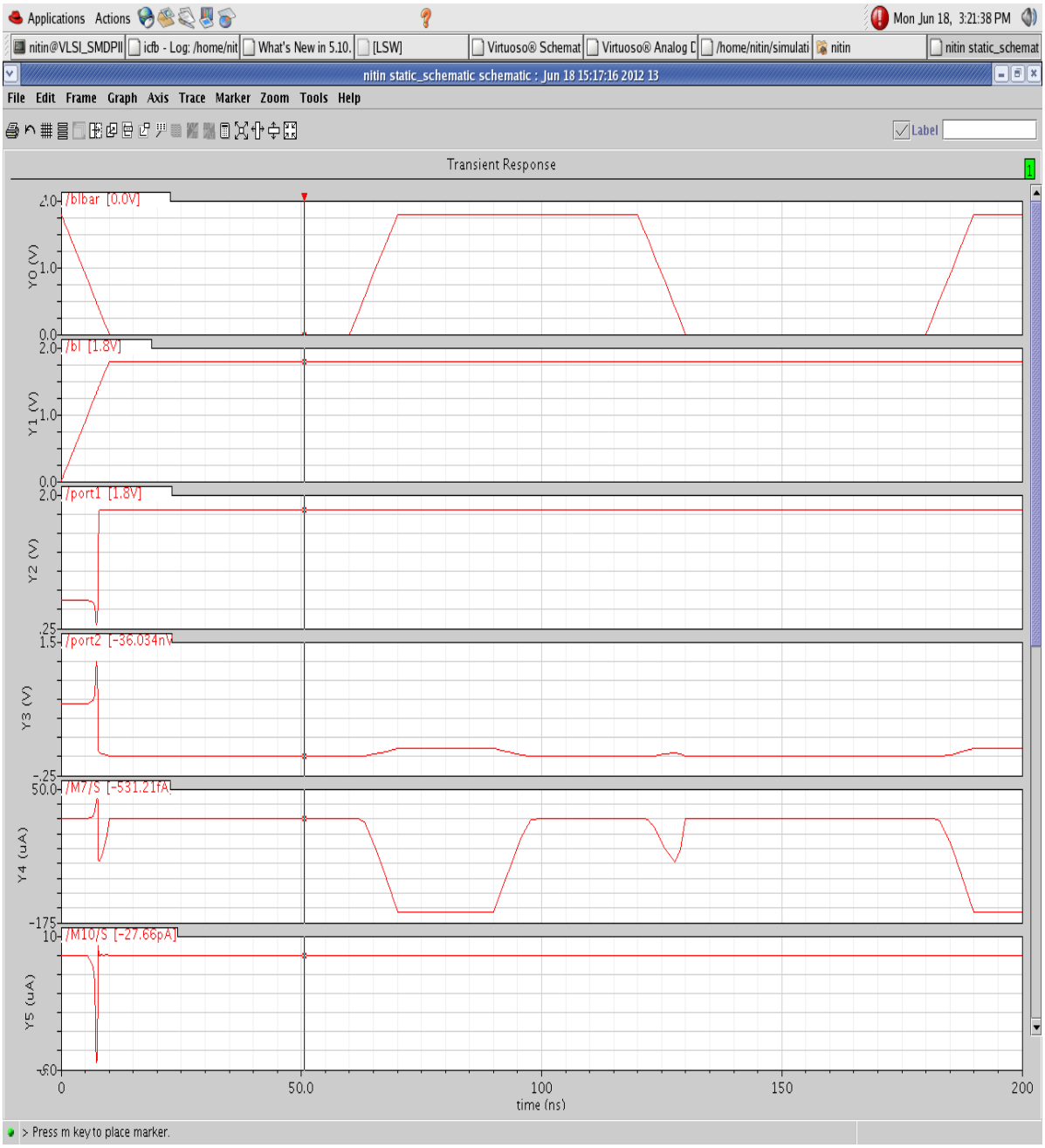

Figure 5.5 Bit line and current waveforms for Read '1' operation

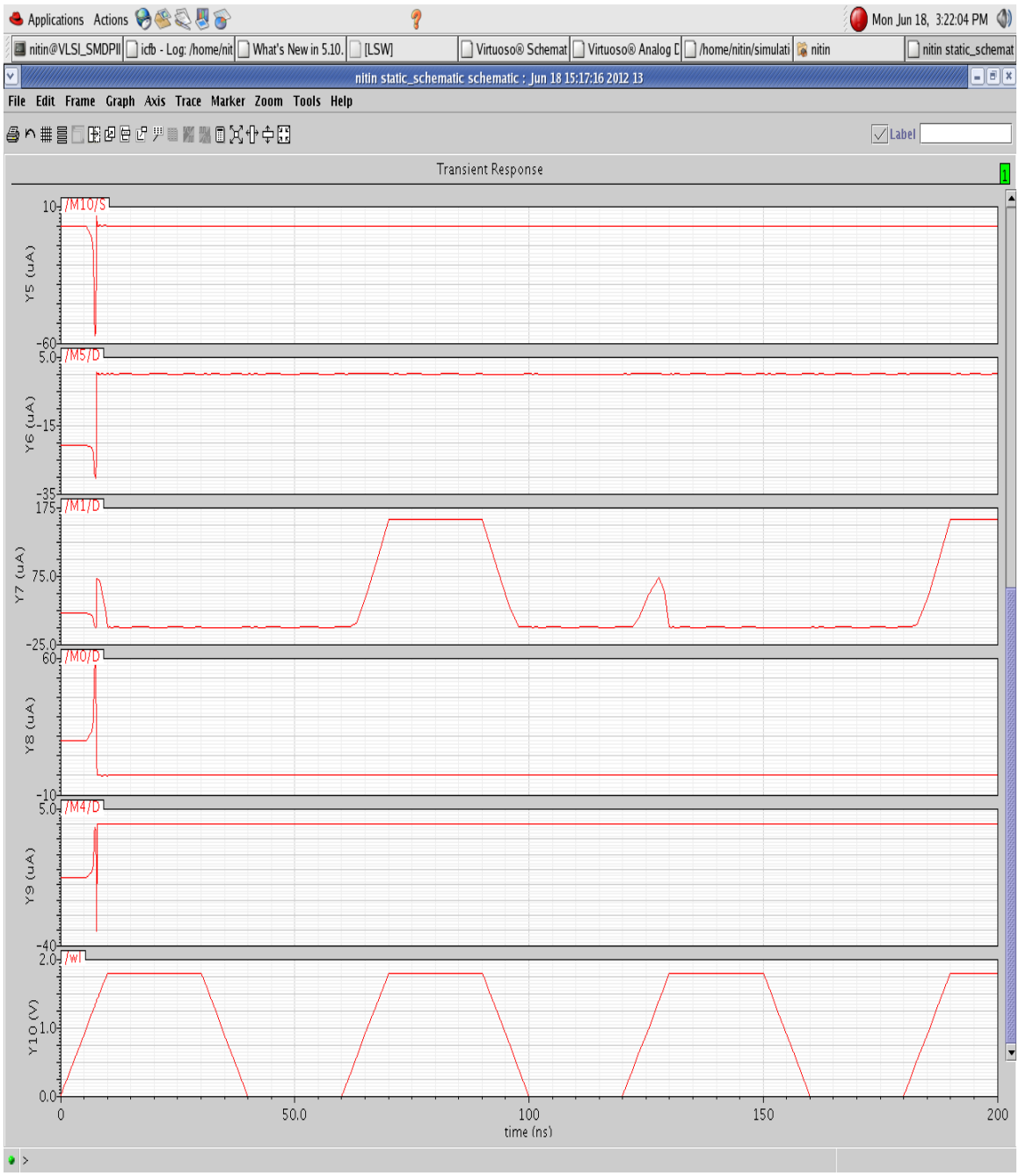

Figure 5.6 Current waveforms for Read '1' operation

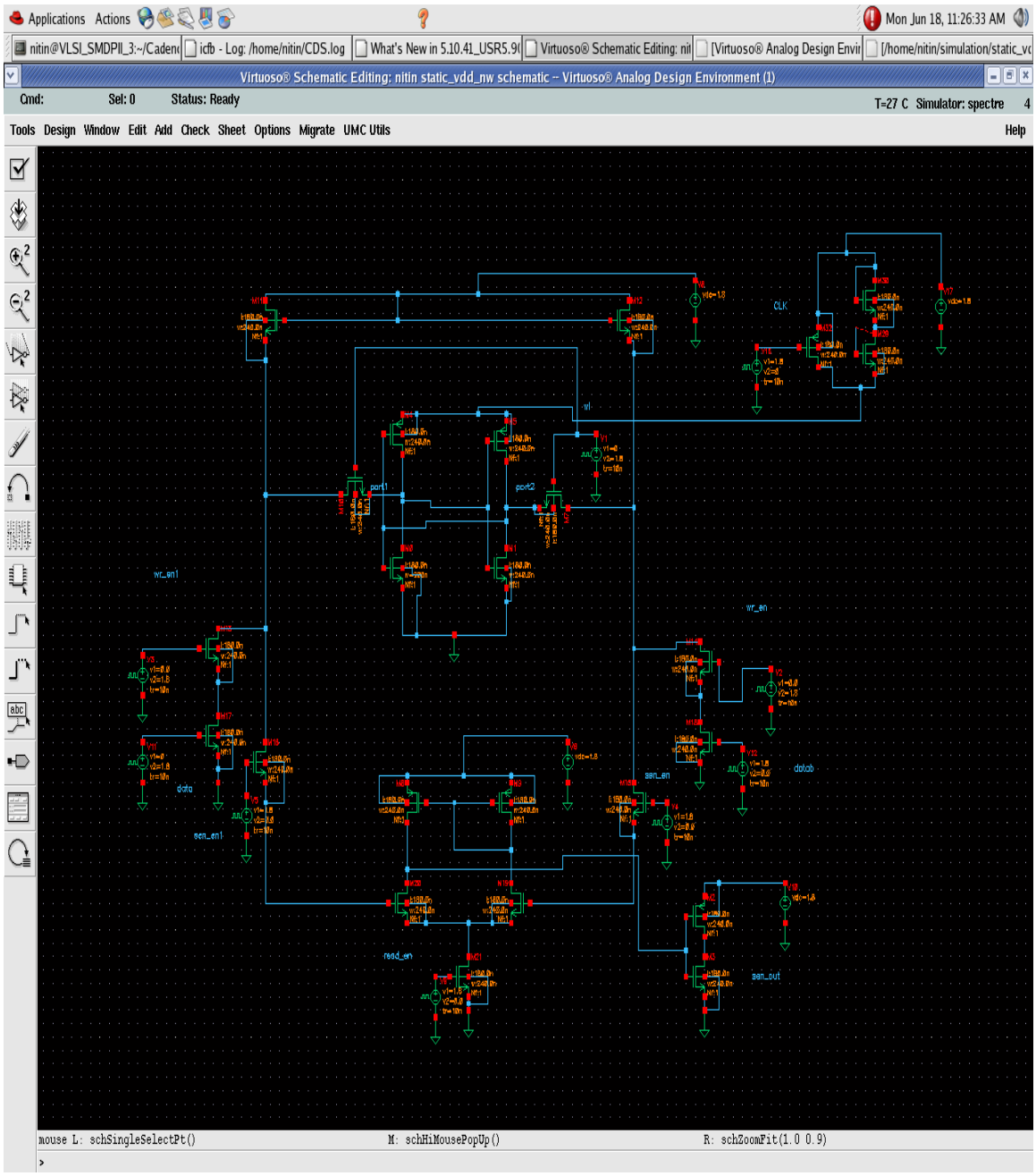

Figure 5.7: Schematic for USR scheme

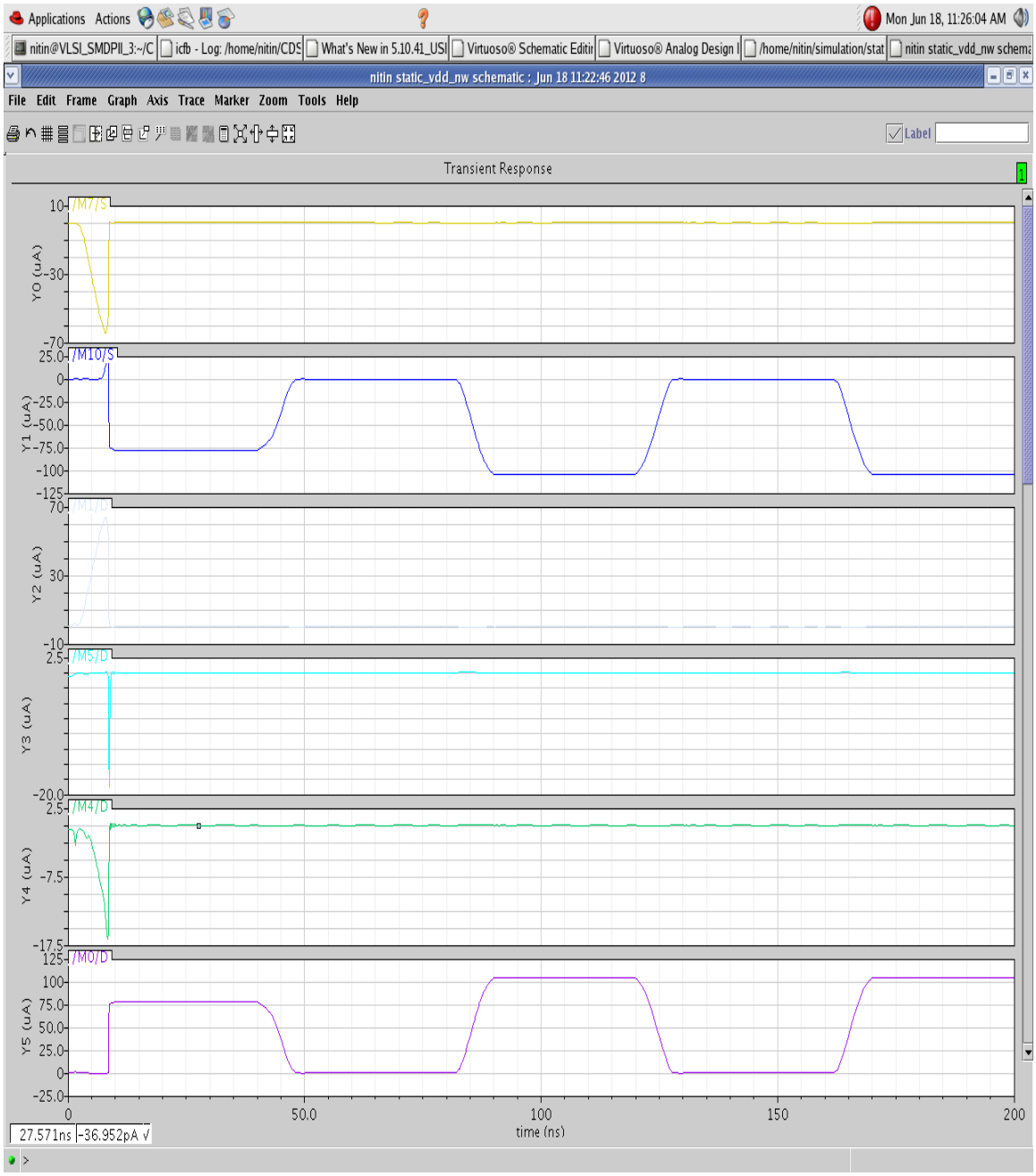

Figure 5.8 : Current waveform for Read ‘0’ operation

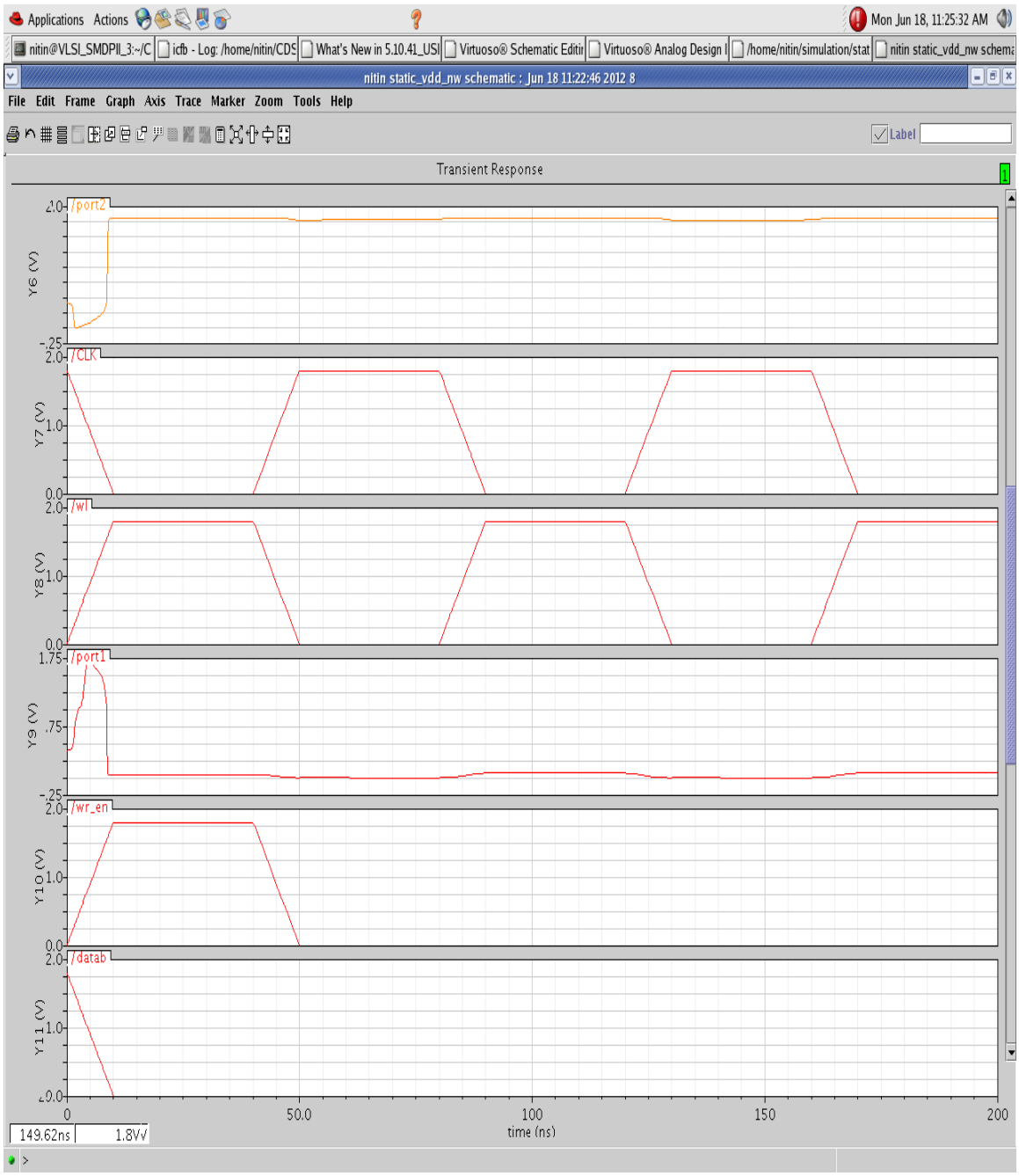

Figure 5.9: Data and word line waveforms for Read ‘0’ operation

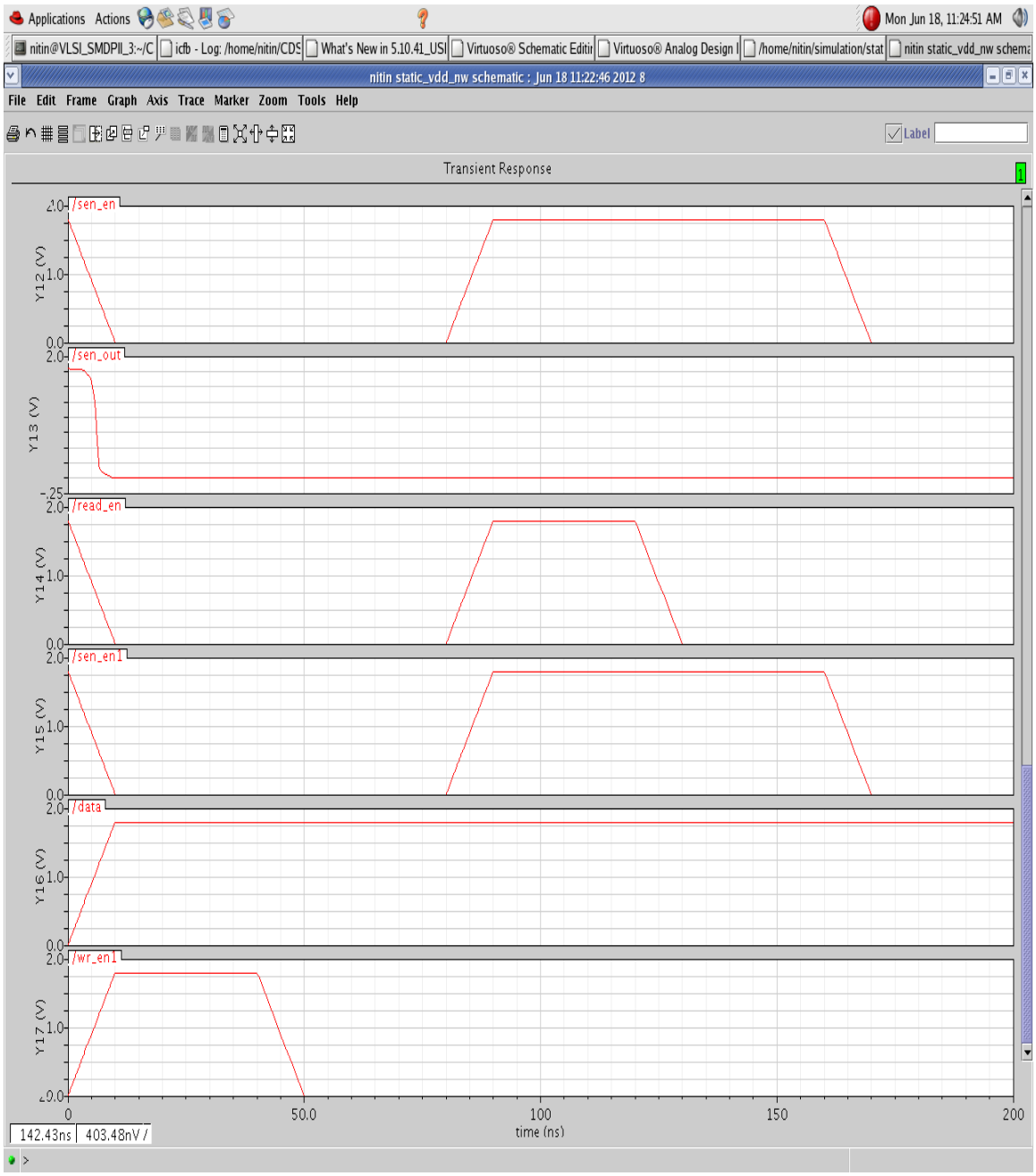

Figure 5.10 Control signals waveforms for Read ‘0’ operation

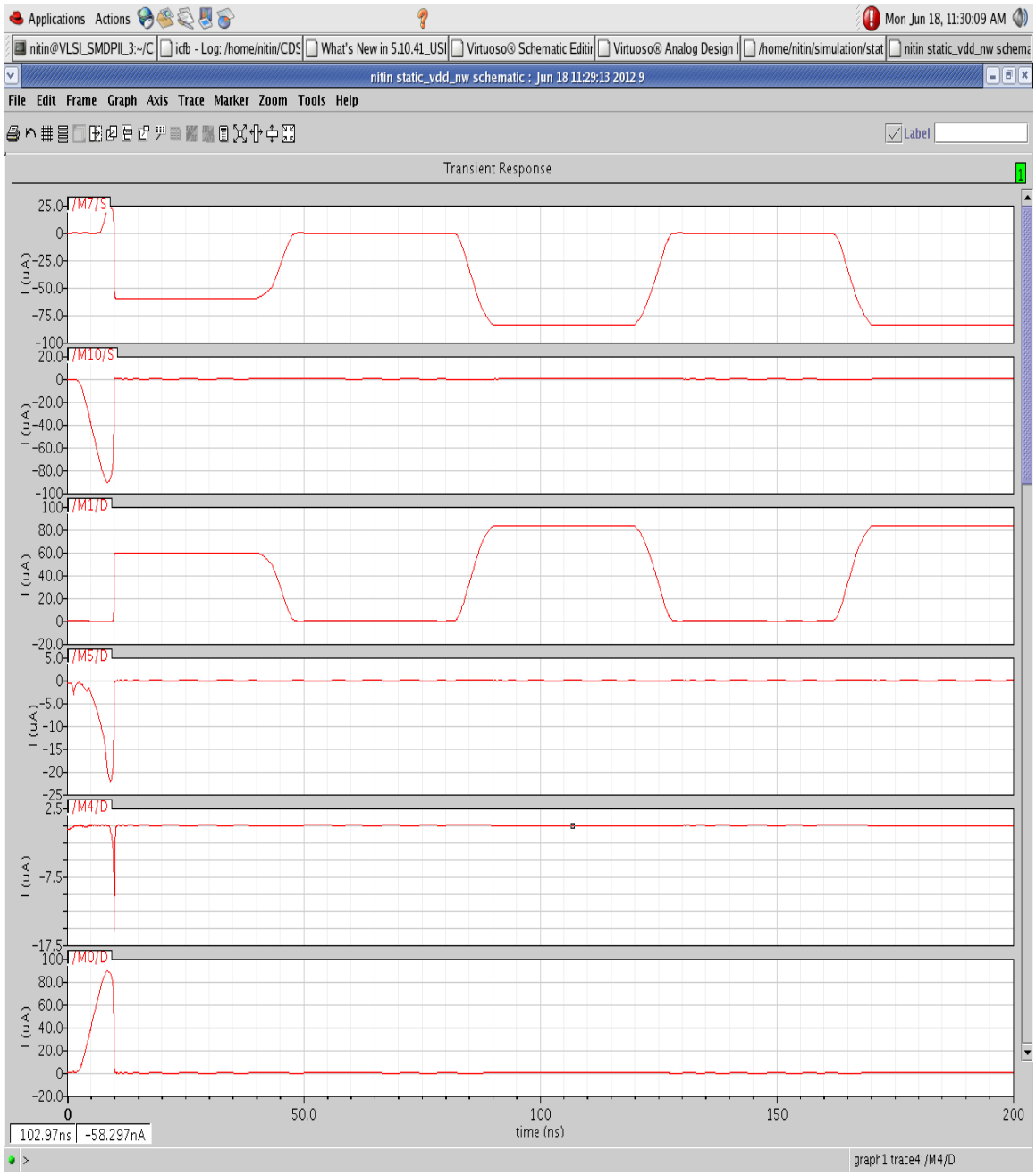

Figure 5.11: Current waveforms for Read ‘1’ operation

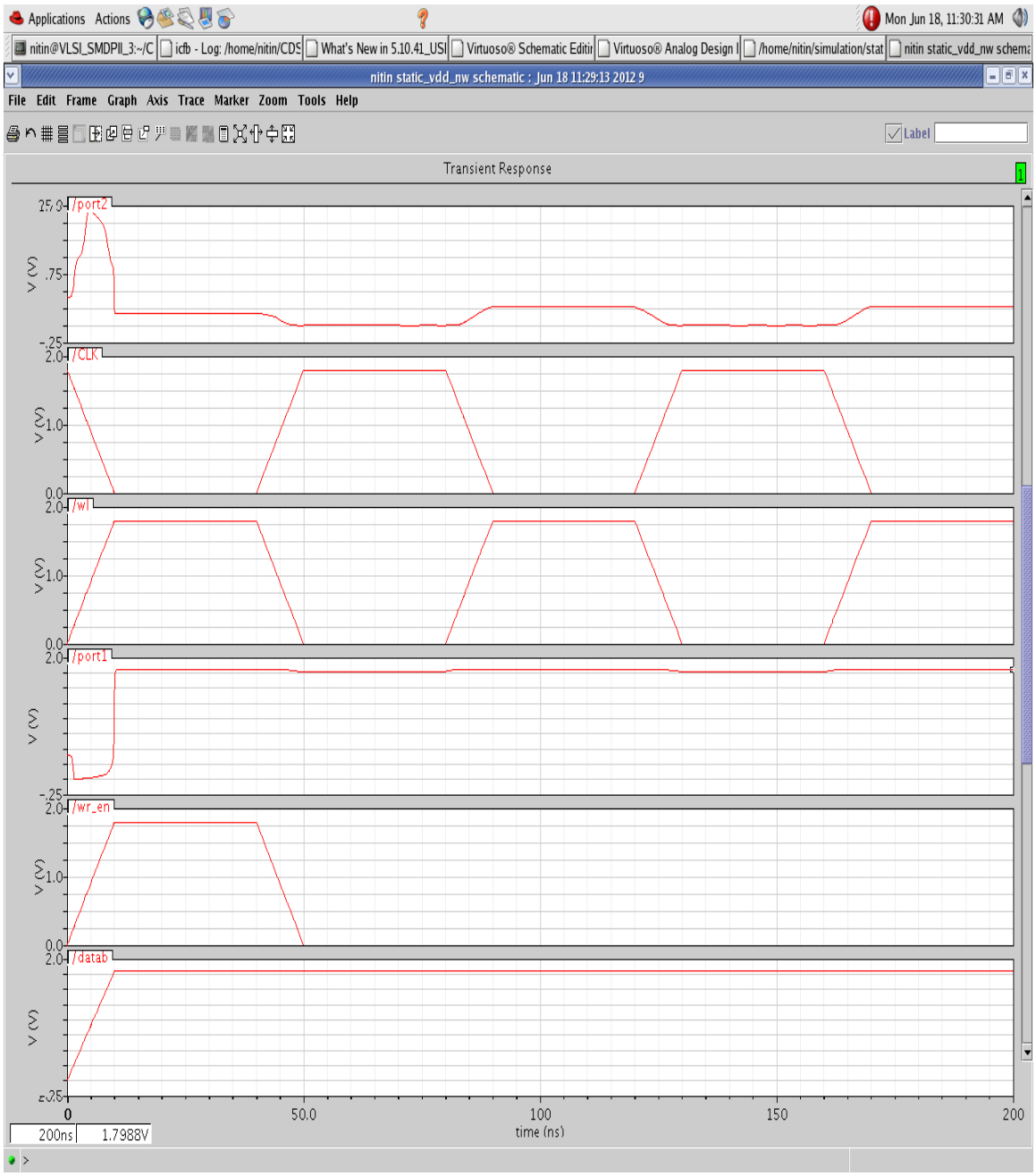

Figure 5.12: Data and word line waveforms for Read ‘1’ operation

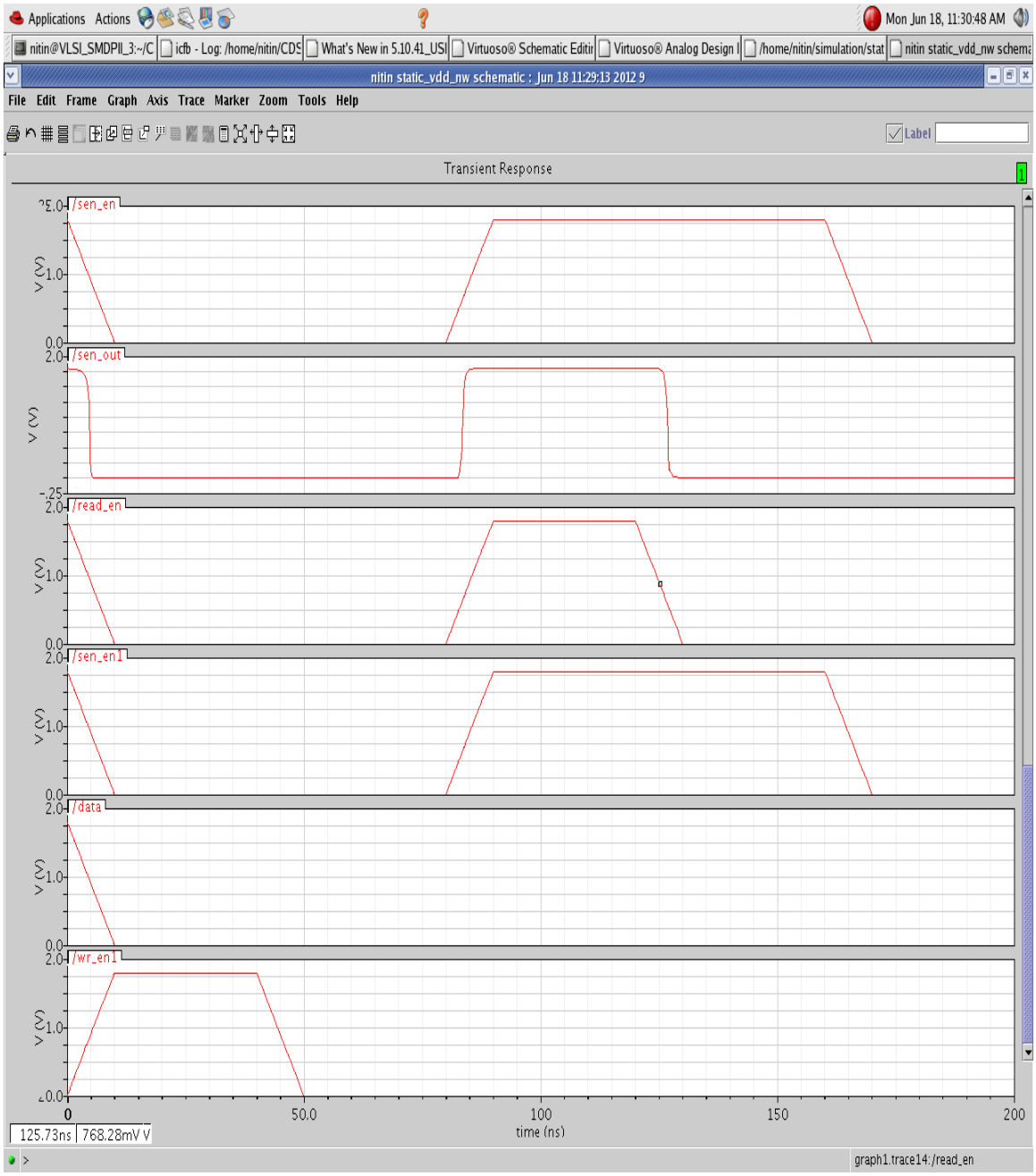

Figure 5.13: Control signals waveforms for Read ‘1’ operation

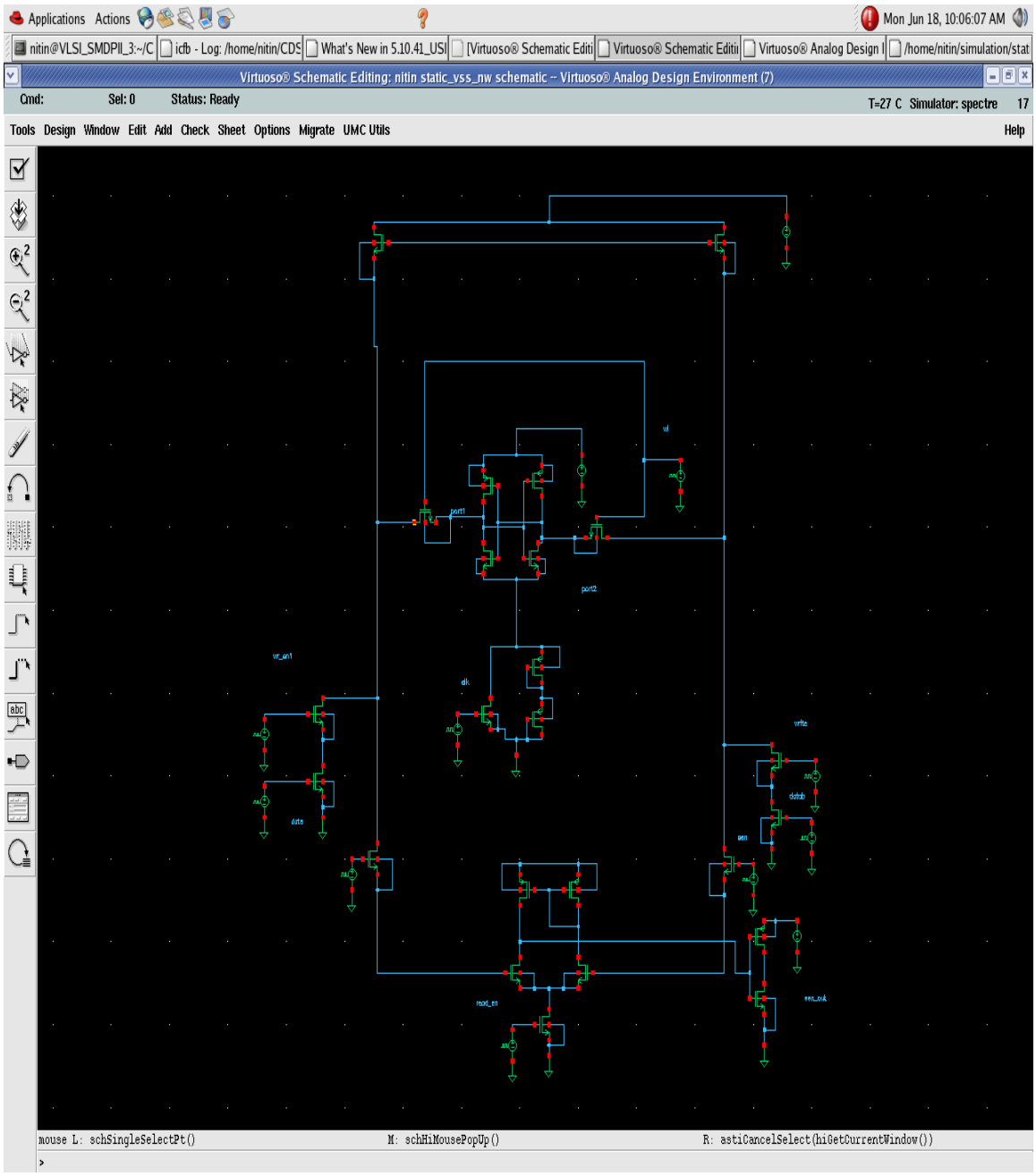

Figure 5.14: Schmatic for LPR scheme

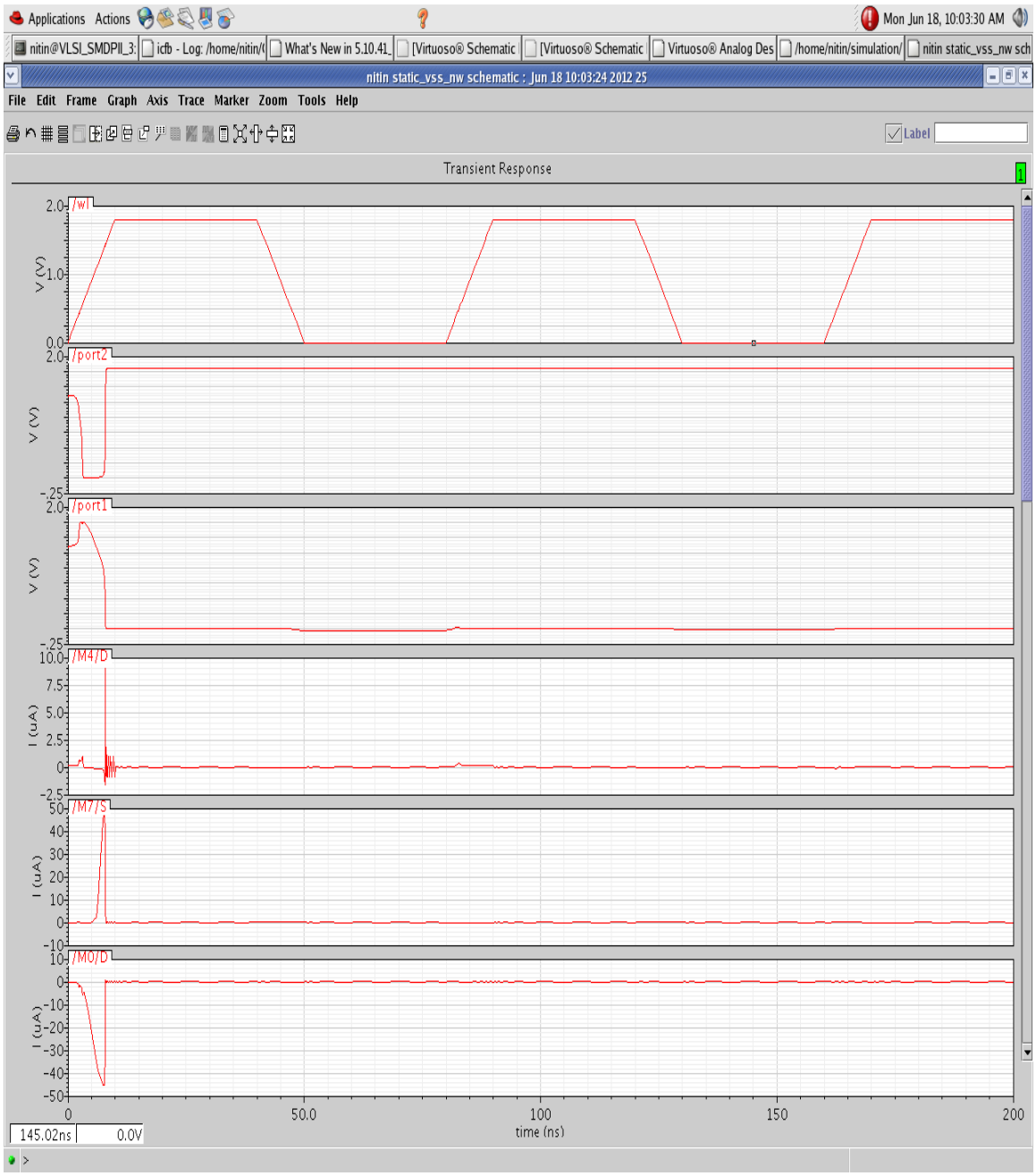

Figure 5.15: Current and word line waveforms for Read ‘0’ operation

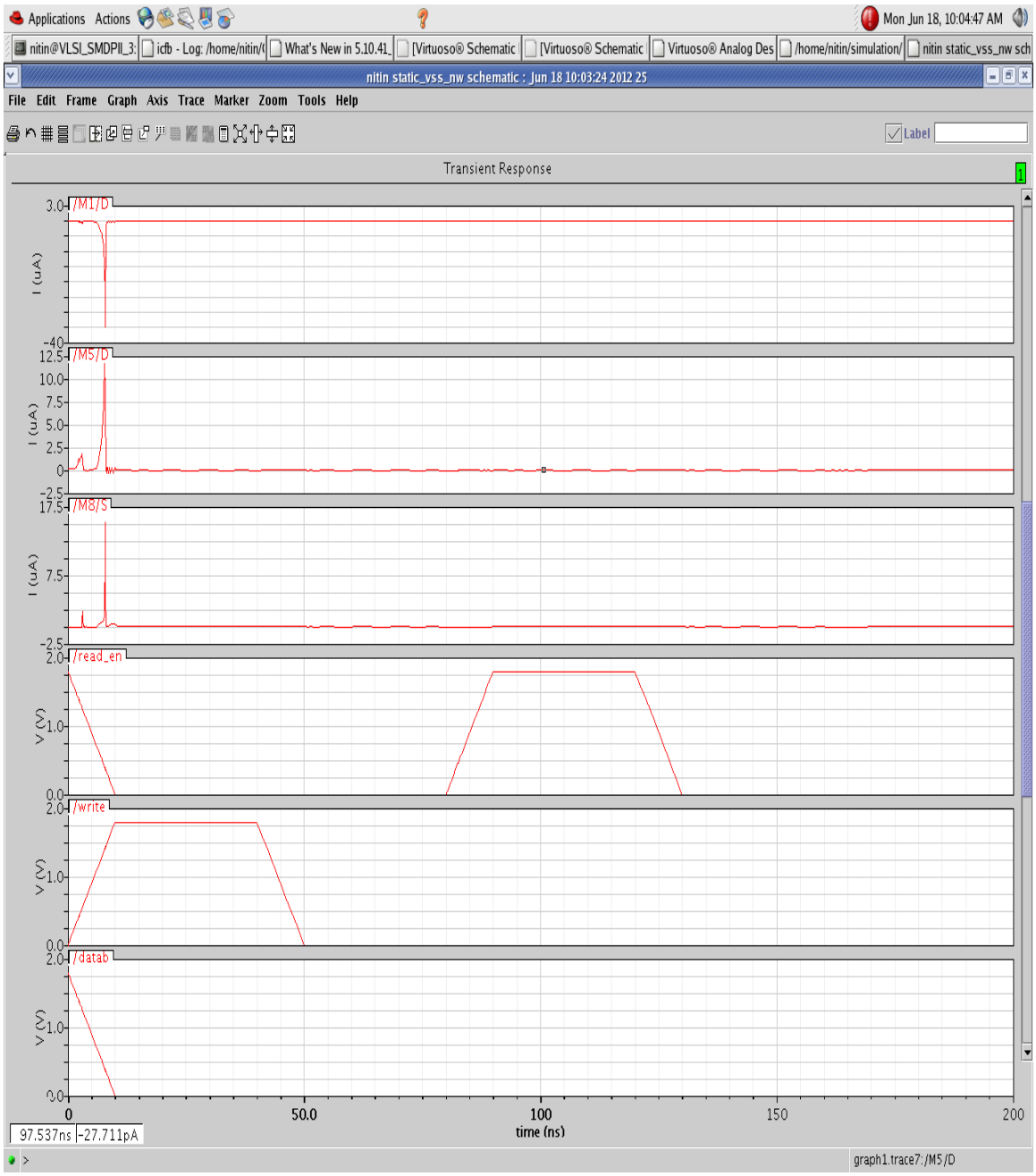

Figure 5.16: Control signals and current waveforms for Read '0' operation

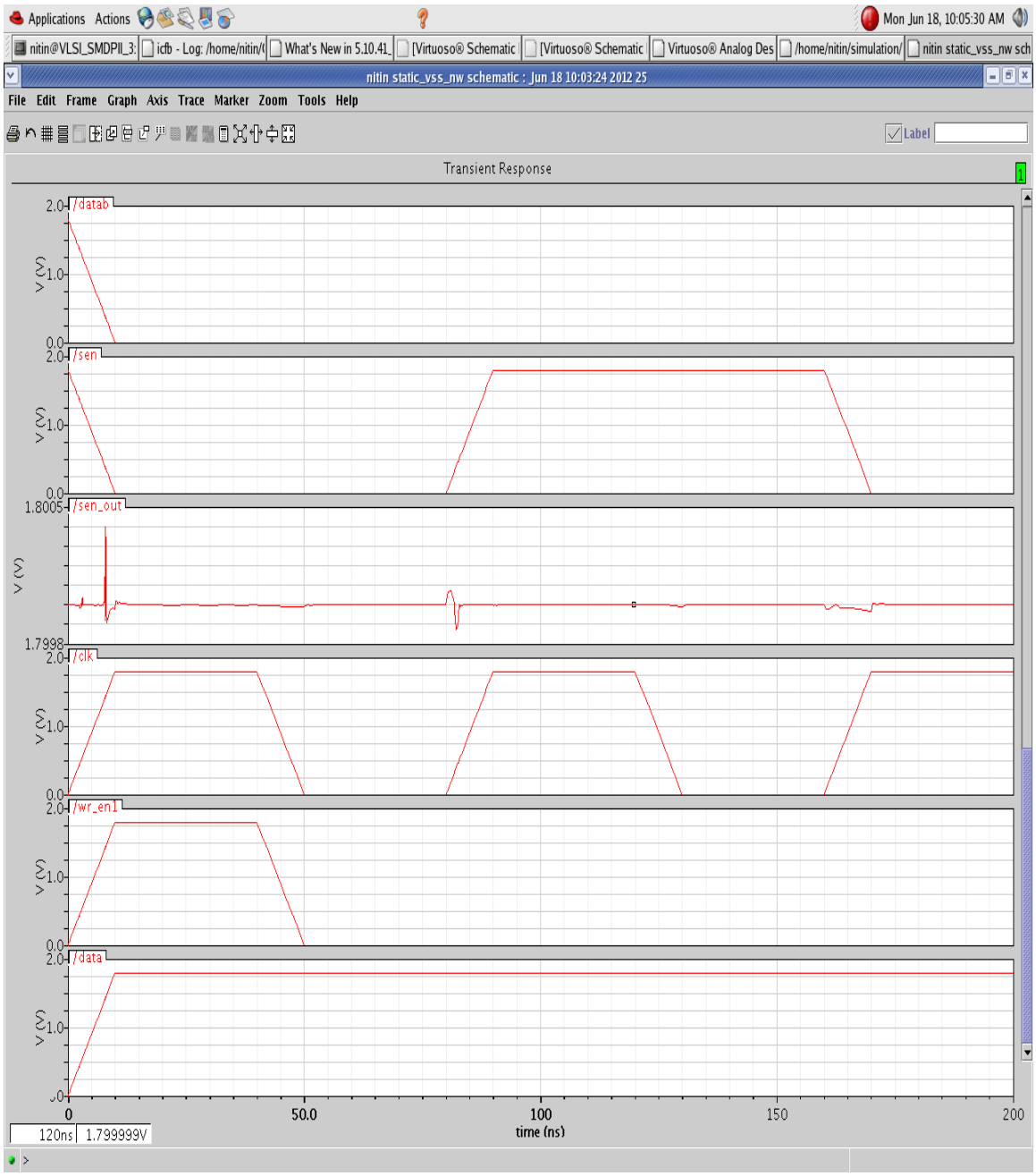

Figure 5.17 Control and data waveforms for Read ‘0’ operation

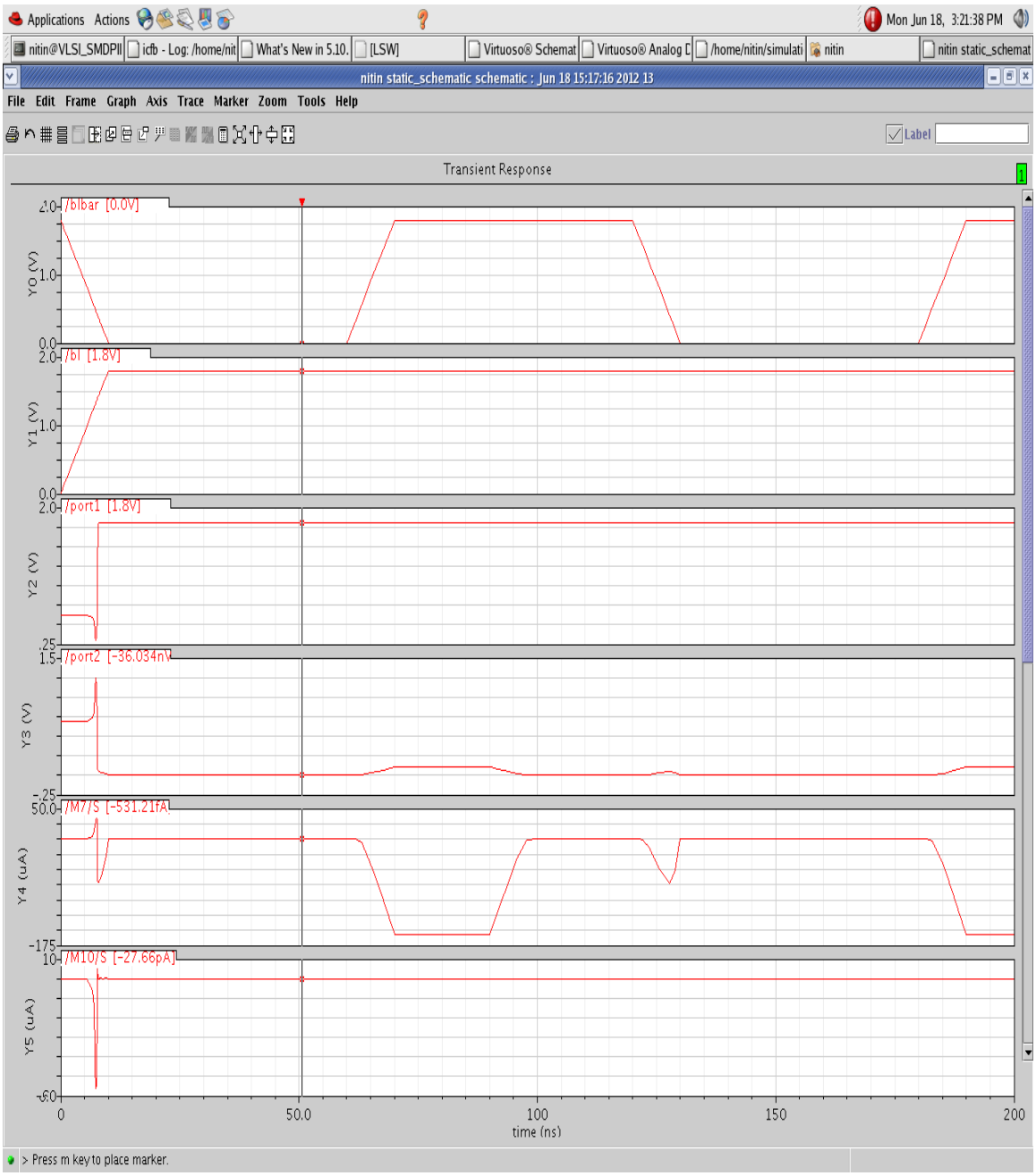

Figure 5.18: Current and stored value waveform for read '1' operation

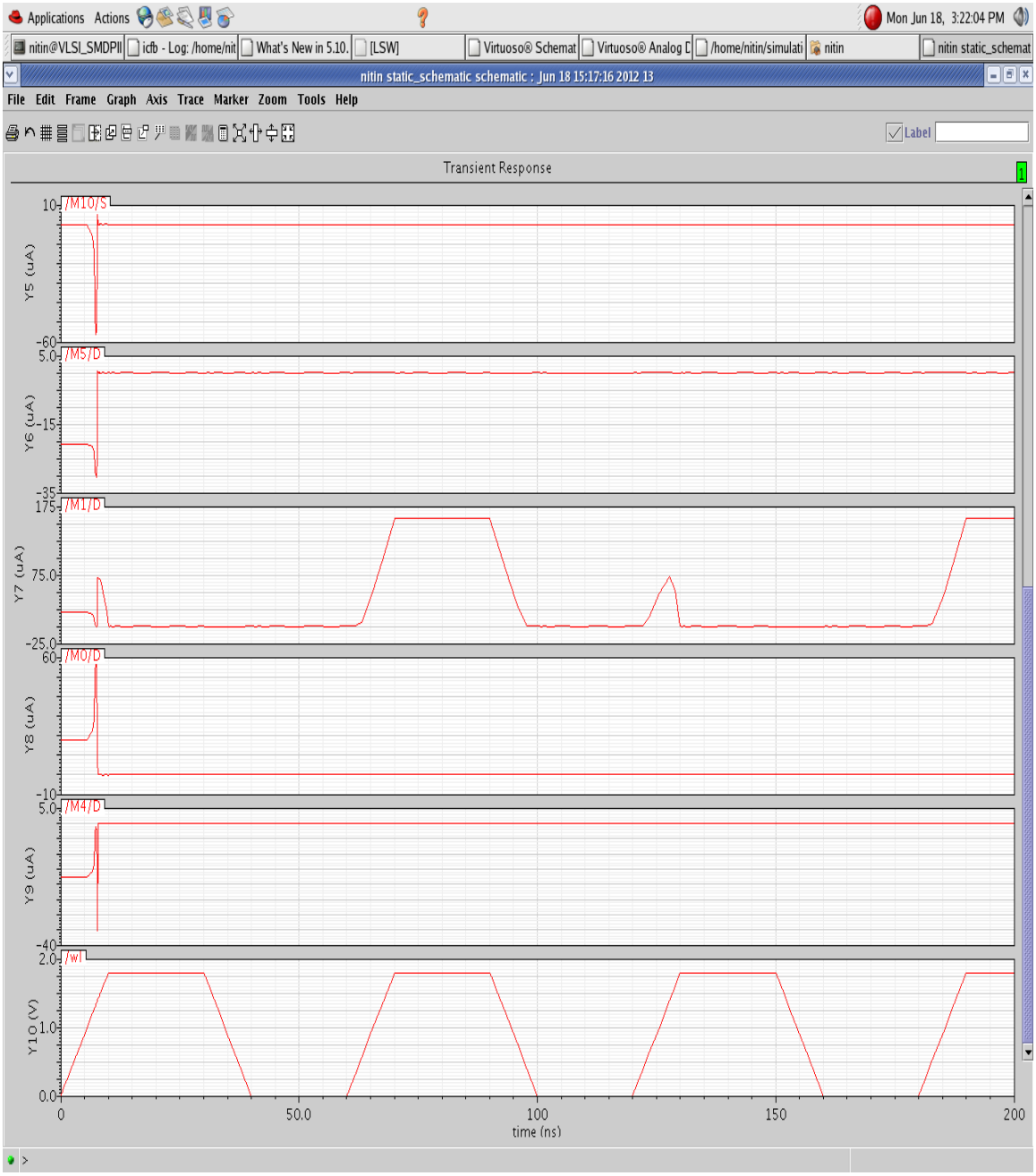

Figure 5.19: Current and WL waveform for read '1' operation

Table 5.1

	M1	M2	M3	M4	M5		M6	I cell	% REDUCTION		
	I_{gate} (pA)	I_{gate} (pA)	I_{sub} (pA)	I_{sub} (pA)	I_{gate} (pA)	I_{sub} (pA)	I_{gate} (pA)		Ig	Is	Io
Conventional cell	37.1	13.9	0.86	0.56	9.84	0.6	19.5	83.6	--	--	--
USRS	2.38	1.14	0.58	--	9.84	0.6	10.6	25.3	70.2	41.6	69.7
LPRS	2.35	1.08	0.17	0.29	10.7	--	19.4	33.6	58.3	77.2	59.8
USRS-A	40.6	1.23	0.4	0.56	11.6	0.7	5.4	64.7	30.7	20.5	23.7

CHAPTER 6

Conclusion and Future Scope

6.1 Conclusion

The present work analyzes the latest developments in low-power circuit techniques and methods with an emphasis on SRAMs. Appropriate methods for reduction of power consumption were studied such as capacitance reduction, very low operating voltages, DC and AC current reduction and suppression of leakage currents to name a few.. Many of reviewed techniques are applicable to other applications such as ASICs, DSPs, etc. Battery and solar-cell operation requires an operating voltage environment in low voltage area. These conditions demand new design approaches and more sophisticated concepts to retain high device reliability. The proposed techniques (USRS and LPRS) are topology based and hence easier to implement. The substantial reduction in leakage currents obtained proves them to be as effective as any fabrication level technique would be.

6.2 Future Scope

To perform the write operation in the SRAM cell to flip the data value, nearly full voltage swings is required on the bit line. This full voltage swing on the highly capacitive bit lines will consume a greater amount of power according to law of CV^2f. Thus voltage swing reduction is an effective way to decrease the power dissipation. The current mode sensing technique is also proposed to give the small voltage swing on the bit lines during write operation. In the proposed method the layout and simulation is done for the one bit line pair for three different methodologies. The bit line interference of selected cell with adjacent selected and non selected cell is also checked out. The proposed current conveyor method has shown an improvement in terms power dissipation over the voltage write and current read (VWCR) and current write and current read (CWCR) method without comprising the performance. In the past, power dissipation was not a constraining factor because device density and operating frequencies were low enough. But nowadays due to increased integration and operating frequency of integrated circuits, power consumption has become an important factor. Battery operated portable devices which performing the high performance

processing task also consume lots of power. Various methodologies are used to reduce the power dissipation by optimizing the parameters that are related to power consumption of circuit. Also the Static RAM is used as a cache memory in the processor and also has an application in the embedded system. Due to continuous advances in the integrated circuit technology, the density of SRAMs in embedded application has grown substantially in recent years. The SRAM block is becoming indispensable block in the system-on-chips (SoCs). The larger density SRAM block has a highly capacitive bit lines and data lines. The dynamic power of SRAM is mainly due to charging and discharging of highly capacitive lines. In view of all these, the future course of action involves effective reduction of leakage in an SRAM cell. It is proposed here that appropriate leakage reduction techniques would be developed with an emphasis on the reduction of gate leakage. Leakage reduction in SRAM is also possible using self controllable switch either at the upper end of the cell to reduce supply voltage (USR scheme) or at the lower end of the cell to raise the potential of the ground node (LPR scheme). This method would also be tested for its efficacy when this work is advanced.

References

[1] Xuning Chen and LiShiuan Peh “Leakage Power Modeling and Optimization in Interconnection Networks” in proceeding ACM Islped, koria, 2003

[2] Alexandru Andrei, Marcus Schmitz, Petru Eles, Zebo Peng, and Bashir M. Al-Hashimi. Overhead-Conscious Voltage Selection for Dynamic and Leakage Energy Reduction of Time Constrained Systems. In Proc. Conf. on Design, Automation and Test in Europe, page 10518, Washington, DC, USA, 2004. IEEE Computer Society.

[3] S. Mutoh, “1-V Power Supply High-speed Digital Circuit Technology with Multithreshold-Voltage CMOS,” IEEE Journal of Solis-State Circuits, Vol. 30, No. 8, pp. 847-854, August 1995.

[4] Gholamreza Karimi1 and Adel Alimoradi “Multi-Purpose Technique to Decrease Leakage Power in VLSI Circuits” Canadian Journal on Electrical and Electronics Engineering Vol. 2, No. 3, March 2011.

[5] B. S. Amrutur and M. A. Horowitz, Speed and power scaling of SRAM's, IEEE Journal of Solid-State Circuits, vol. 35, February 2000.

[6] Kaushik Roy “Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits Proceedings of the IEEE, Vol. 91, NO. 2, Feb.2003.

[7] Michael Powell Se-Hyun Yang, Babak Falsafi, Kaushik Roy, and T. N. Vijay Kumar “Gated-Vdd: A Circuit Technique to Reduce Leakage in Deep-Submicron Cache Memories, in proceeding ACM, ISLPED, koria 2000.

[8] Masaya Sumita, Shiro Sakiyama, Masayoshi Kinoshita, Yuta Araki, Yuichiro Ikeda, and Kohei Fukuoka., “Mixed body bias techniques with Fixed Vt and Ids Generation Circuits, IEEE Journal of solid-state circuits, VOL. 40, NO. 1, Jan 2005.

[9] Liqiong Wei., Zhanping Chen, Kaushik Roy, Mark C. Johnson, Yibin Ye and Vivek K. De. “Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications” IEEE Transactions on very large scale integration system, vol. 7, no. 1, March 1999.

[10] Hiroaki Okuyama, Takeshi Nakano, Shuichi Nishida, Etsuro Aono. Hisahiro Satoh and Shigeru Akita. “A7.5-ns 32Kx8 CMOS SRAM” IEEE Journal of solid-state circuits. vol. 23. no. 5. Oct, 1988.

[11] Baker Mohammad, “Low Leakage Power SRAM Cell for Embedded Memory” International Conference on Innovations in Information Technology, 2011.

[12] Anh-Tuan Do,Zhi-Hui Kong, Kiat-Seng Yeo, and Jeremy Yung Shern Low “Design and Sensitivity Analysis of a New Current-Mode Sense Amplifier for Low-Power SRAM” IEEE Transaction on very large scale integration systems vol. 19, no. 2, Feb. 2011.

[13] Ramy E. Aly, A. Bayoumi and Mohamed Elgamel “Dual Sense Amplified Bit Lines (DSABL) Architecture for Low-Power SRAM Design”, IEEE, 2005.

[14] Shin-Pao Cheng and Shi-Yu Huang “A Low-Power SRAM Design Using Quiet-Bitline Architecture” Proceedings of the 2005 IEEE International Workshop on Memory Technology, Design, and Testing, 2005.

[15] Randy W. Mann, Jiajing Wanga, Satyanand Nalam, Sudhanshu Khanna, Geordie Braceras , Harold Pilo, Benton H. Calhoun a “Impact of circuit assist methods on margin and performance in 6T SRAM” Solid State Electronics Elsevier, 2010.

[16] Gu Ming Yang Jun, Xue Jun. “Low Power SRAM Design Using Charge Sharing Technique”, IEEE, 2005.

[17] Kiyoo Itoh. “Trends in Low-Power RAM Circuit Technologies” IEEE, 1995.

[18] Zhang Zheng-xuan, Zhang En-xia “Practical considerations in the design of SRAM cells on SOI” Microelectronics Journal 39 (2008) 1829– 1833.

[19]K. Takada**,** “A 16Mb 400MHz Loadless CMOS Four- Transistor SRAM Macro,” ISSCC 2000.

[20]G. Razavipour, A. Afzali-Kusha, and M. Pedram “Design and Analysis of Two Low-Power SRAM Cell Structures” IEEE Transaction on VLSI system, vol. 17, no. 10, Oct. 2009.

[21] Leland Chang, Robert K. Montoye, Yutaka “An TFT-SRAM for variability tolerance and low-voltage operation in high-performance caches,” IEEE J. Solid- State Circuits, vol. 43, no. 4, April 2008.

[22] G. Razavipour, A. Afzali-Kusha “Design and Analysis of Two Low-Power SRAM Cell Structures” IEEE Transaction on very large scale integration systems, VOL. 17, NO. 10, Oct. 2009.

[23] A. Agarwal, C. Kim, S. Mukhopadhyay, and K. Roy, “Leakage in nano-scale technologies: mechanisms, impact and design considerations,“ in Proc. of Design Automation Conf.,2004.

[24]K. Itoh and K. Sasaki, Trends in low power circuit design, Proc. IEEE, Vol. 83, pp 524-543, 1993.

[25] M. Ukita, "A Single Bitline Cross-Point Cell Activation (SCPA) Architecture for Ultra Low Power SRAMs," in ISSCC Digest of Technical Papers, February 1994.

[26] Kenneth W. Mai, Toshihiko”MoriLow-Power SRAM Design Using Half-Swing Pulse Mode Techniques” IEEE Journal of solid-state circuits, Nov. 1998.

[27] K. Itoh “Trends in Low-Power RAM Circuit Technologies", Proceedings of the IEEE, April 1995.

[28] Anh-Tuan Do, Zhi-Hui Kong “Design and Sensitivity Analysis of a New Current-Mode Sense Amplifier for Low-Power SRAM” IEEE transaction on very large scale integration systems, VOL. 19, NO. 2, FEBRUARY 2011.

[29] L. Carro, D. Matos “A New Reconfigurable Clock-Gating Technique for Low Power SRAM-based FPGAs” 978-3-9810801-7-9/DATE11/© EDA, 2011.

[30] Nobutaro Shibata, Mayumi Watanabe “A High-Speed Low-Power Multi-VDD CMOS/SIMOX SRAM with LV-TTL Level Input/output Pins—Write/Read Assist Techniques for 1-V Operated Memory Cells” IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 45, NO. 9, SEPTEMBER 2010.