ANALYSIS AND DESIGN OF CACHE MEMORY CELL FOR LEAKAGE POWER REDUCTION

A thesis report submitted in partial fulfilment of the requirements for the award of the degree of

MASTER OF TECHNOLOGY

in

VLSI Design & CAD

Submitted by

Rohit Sachdeva

Roll No. 60761017

Under the Guidance of

Mr. Arun Kumar Chatterjee

Lecturer, ECED

Department of Electronics and Communication Engineering

Thapar University, Patiala-147004, India

July, 2009

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"Analysis and Design of Cache Memory Cell for Leakage Power Reduction"** in partial fulfilment of the requirements for the award of the degree of Master of Technology in VLSI Design and CAD at the Electronics and Communication Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Mr. Arun Kumar Chatterjee, Lecturer, ECED.

The matter presented in this thesis has not been submitted in any other University/Institute for the award of any degree.

Rohit

Date: 15-07-09

Rohit Sachdeva
Roll No. 60761017

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

Arun Kumar Chatterjee

Mr. Arun Kumar Chatterjee
Lecturer, ECED
Thapar University

Counter Signed by:

22.7.09.

Dr. A. K. Chatterjee
Head, ECED
Thapar University, Patiala

22/7/09

Dr. R. K. Sharma
Dean of Academic Affairs
Thapar University, Patiala

ACKNOWLEDGEMENT

Firstly I would like to express my gratitude to my supervisor, Mr. Arun Kumar Chatterjee, Lecturer, ECED, Thapar University, Patiala, for the opportunity to work on my masters thesis under his guidance. He has provided an invaluable help with ideas and discussions throughout my entire time working on this thesis. It was both an honour and a privilege to work with him. He also provided help in technical writing and presentation style and I found this guidance to be extremely valuable.

I would like to thank the entire Electronics and Communication Engineering Department, and specifically, Dr A.K. Chatterjee, Head, and Ms. Alpana Aggarwal, PG Coordinator, for massive support with tools and ideas.

I am also thankful Ms Madhu Kushwaha, Lecturer, ECED for her regular guidance and support throughout this thesis work.

Last but not least I would like to thank my friends, Rashmi Singh, Mahendra Kumar Soni, Lokesh Kumar Srivastav and Sudhir Kumar Sharma, 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.

Rohit Sachdeva

ABSTRACT

On chip cache memories contributes a large fraction to the total power consumption of microprocessor. As technology scales down into deep-submicron, leakage power is becoming a dominant source of power consumption. As cache memory is an array structure leakage reduction in just one memory cell can on the whole reduce a large amount of leakage power.

In this thesis leakage power of conventional 6T cell at 180nm technology has been evaluated and is found to be 2.03nW. Also various circuit level leakage reduction techniques such as Autobackgate Controlled Multi-threshold CMOS (ABC-MTCMOS), Gated V_{DD} and Dynamic Voltage Scaling (DVS) has been discussed and applied on conventional 6T cache memory cell. Leakage reduction of 28.5%, 88.7% and 98.3% has been achieved by using ABC-MTCMOS, Gated-V_{DD} and DVS respectively as compared to conventional 6T cell.

Further a different architecture for memory cell, that is, 5T SRAM cell has also been adequately designed and analyzed to diminish leakage power consumption. About11.8% reduction in leakage has been achieved by conventional 5T cell than the conventional 6T cell. Using 5Tcell with ABC-MTCMOS, Gated-V_{DD} and DVS, further leakage reduction of 3.3%, 26% and 12.1% respectively has been achieved as compared to the particular case with 6T cell.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 OVERVIEW

In recent years, power consumption has become a critical design concern for many VLSI systems. Historically, one of the advantages of complementary metal-oxide-semiconductor (CMOS) over competing technologies, such as transistor-transistor logic (TTL) and emitter coupled logic (ECL), has been its lower power dissipation. When not switching, CMOS transistors have, in the past, dissipated negligible amounts of power. But now the scenario has been changed, current leaks in the transistors even when they are not switching.

In next generation technologies leakage power is expected to be more significant as threshold voltage decreases in conjunction with supply voltage as technology scales. Even in current-generation technology, subthreshold leakage power dissipation is comparable to the dynamic power dissipation, and the fraction of the leakage power will increase significantly in the near future [1]. Subthreshold leakage is a problem for all transistors, but it is a particularly important problem in memory structures, because a major part of any electronic system is the memory subsystem [2-3]. Today's microprocessor designs devote a large fraction of the chip area to the memory structures. High-performance on-chip caches are a crucial component in the memory hierarchy of modern computing systems.

Reduction in the leakage power of even a single cell of cache can on the whole reduce a large fraction of the total power dissipation of microprocessor due to large sizes of on-chip caches. Thus, cache memory is a good candidate for optimizing leakage energy consumption and reduction in leakage of cache memory can significantly improve the system power-efficiency.

1.2 OBJECTIVE

The purpose of this thesis is to analyze circuit level techniques to reduce the leakage power dissipation in cache memory cell. Various circuit level techniques have been applied to six-transistor SRAM cell (being used for caches) and their efficiency is

compared. To further increase their efficiency, five-transistor SRAM cell (already proposed) has been designed and its leakage is compared with the six-transistor cell.

1.3 OUTLINE OF THE THESIS

In **chapter 2** some basic memory concepts and overview of cache memory has been described. Various cell architectures like 6T SRAM have been discussed in detail. It also describes the need of low power and its applications.

Chapter 3 describes leakage in SRAM cell and various circuit level techniques for its reduction.

In **chapter 4** designing of 5T SRAM cell has been discussed thoroughly.

Chapter 5 summarises all simulation results and concludes the thesis report. It also discusses about the future aspects of the work done.

1.4 TERMINOLOGY USED

The following is a listing of terms and abbreviations used in this thesis and their explanations:

- **λ –** Process-technology parameter with the dimensions of V^{-1} and for a given process it is inversely to the length selected for the channel.
- **BL –** Bitline, the wire connecting the drain (source) of the memory cells' pass-transistors to the sense amplifiers.
- **Cache –** A memory used to store data or instructions likely to be used soon by the CPU. Its purpose is to speed up operation by bridging the performance gap between the CPU and the main memory.
- **CMOS –** Complementary MOS, circuits containing both NMOS and PMOS devices.
- **CPU –** Central Processing Unit, the heart of a microprocessor. It carries out the execution of instructions.
- **DRAM –** Dynamic RAM, a RAM where the value is stored dynamically on a capacitor.
- **gnd –** Ground, reference for the low potential power supply (0V).
- **I_D –** Drain current through a transistor.
- **MOSFET –** Metal-Oxide Semiconductor Field-Effect Transistor, a transistor utilizing a metal-oxide to insulate the gate from the semiconductor, and an electric field to create an inversion layer as channel.
- **NMOS –** N-channel MOSFET, a transistor utilizing an n-type inversion layer as channel for conducting current.
- **nT –** n-transistor, memory cell made up of n number of transistors. For example 6T, a cell made up of six transistors.
- **PMOS –** P-channel MOSFET, a transistor utilizing a p-type inversion layer as channel for conducting current.
- **RAM –** Random Access Memory, a memory where information can be stored and retrieved in non-sequential order.
- **Sense Amplifier –** A circuit used to amplify the differences of the bitlines during read. It is used to speed up reading and restore full-swing values.
- **SRAM –** Static RAM, a RAM where the value is stored statically in a latch.

- $\mathbf{V_{CC}}$ **–** Reference for the high potential power supply (1.8V in this thesis).
- $\mathbf{V_{DS}}$ **–** Drain-Source potential, the difference between the potential at the drain and the source of a transistor.
- $\mathbf{V_{GS}}$ **–** Gate-Source potential, the difference between the potential at the gate and the source of a transistor.
- $\mathbf{V_{PC}}$ **–** Bitline precharge voltage for the 5T SRAM.
- $\mathbf{V_{SB}}$ **–** Source-Bulk potential, the difference between the potential at the source and the bulk of a transistor.
- $\mathbf{V_{T}}$ **–** Threshold voltage, the gate-source potential required for the transistor to start conducting.
- **WL –** Wordline, the wire connected to the gate of the pass-transistors of the memory cells.

CHAPTER 2 MEMORY FUNDAMENTALS

2.1 INTRODUCTION

A memory in terms of computer hardware is a storage unit. There are many different types of hardware used for storage, such as magnetic hard drives and tapes, optical discs such as CDs and DVDs, and electronic memory in form of integrated memory or stand-alone chips. In this thesis only the electronic memory, and more specifically, random access memories (RAM) has been discussed.

An electronic memory is used to store data or programs, and is a key component in all computers today. It is built up of small units called bits which can hold one binary symbol of data (referred to as a ’1’ or a ’0’). These bits are then grouped together into bytes (8 bits) or words (usually in the range of 16-64 bits). In a normal PC several layers of abstraction are then applied to make up the memory architecture, all the way from the processor’s registers to, for example, a file on the hard drive. Within these abstract layers of memory, several physical layers (e.g. RAM, hard drive) also exist.

The main focus of this thesis is the RAM. There are four basic operations that have to be supported by a RAM. These are the writing and reading of ’0’ and ’1’ respectively.

2.2 RANDOM ACCESS MEMORIES (RAMs)

RAMs are read/write memories in which data can be written into or read from any selected address in any sequence. When a data unit is written into a given address in the RAM, the data unit previously stored at that address is replaced by the new data unit. When a data unit is read from a given address in the RAM, the data unit remains stored and is not erased by the read operation. This non-destructive read operation can be viewed as copying the content of an address while leaving the content intact. A RAM is typically used for short-term data storage because it cannot retain stored data when power is turned off.

The two categories of RAM are the static RAM (SRAM) and the dynamic RAM (DRAM). Static RAMs generally use latches as storage elements and can therefore store data indefinitely as long as the power is applied. Dynamic RAMs use capacitors as

storage elements and cannot retain data very long without the capacitors being recharged by a process called refreshing. Both SRAMs and DRAMs will lose stored data when dc power is removed and, therefore, are classified as volatile memories.

Data can be read much faster from SRAMs than from DRAMs. However, DRAMs can store much more data than SRAMs for a given physical size and cost because the DRAM cell is much simpler, and more cells can be crammed into a given chip area than in the SRAM [4].

2.3 MEMORY CELL ARRAY

The memory cells in a SRAM are organized in rows and columns, as illustrated in Figure 2.1 for the case of an n x 4 array. All the cells in a row share the same Row Select line. Each set of Data in and Data out lines go to each cell in a given column and are connected to a single data line that serves as both an input and output (Data I/O) through the data input and data output buffers.

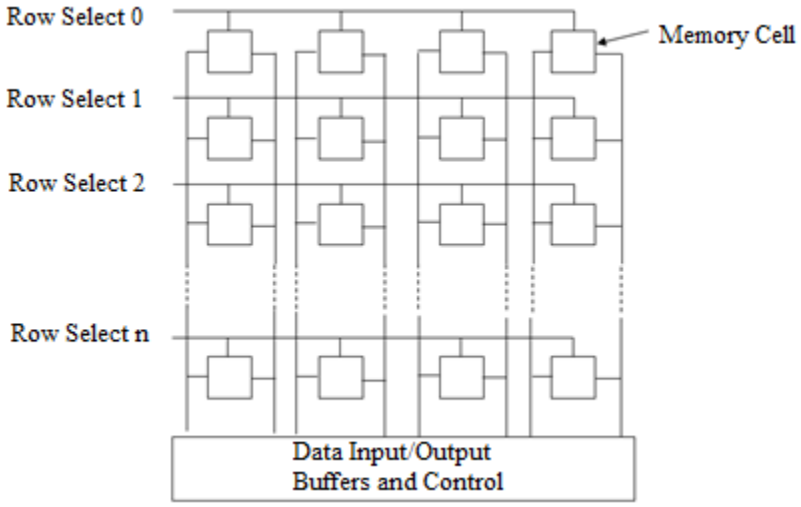

Figure 2.1: SRAM cell array [4].

To write a data unit, in this case a nibble, into a given row of cells in the memory array, the Row Select line is taken to its active state and four data bits are placed on the Data I/O lines. The Write line is then taken to its active state, which causes each data bit to be stored in a selected cell in the associated column. To read a data unit, the read line is taken to its active state, which cause the four data bits stored in the selected row to appear on the Data I/O lines.

2.4 MEMORY CELL STRUCTURES

2.4.1 6T SRAM cell

The conventional six-transistor (6T) SRAM is built up of two cross-coupled inverters and two access transistors, connecting the cell to the bitlines (figure 2.2). The inverters make up the storage element and the access transistors are used to communicate with the outside. The cell is symmetrical and has a relatively large area. No special process steps are needed and it is fully compatible with standard CMOS processes.

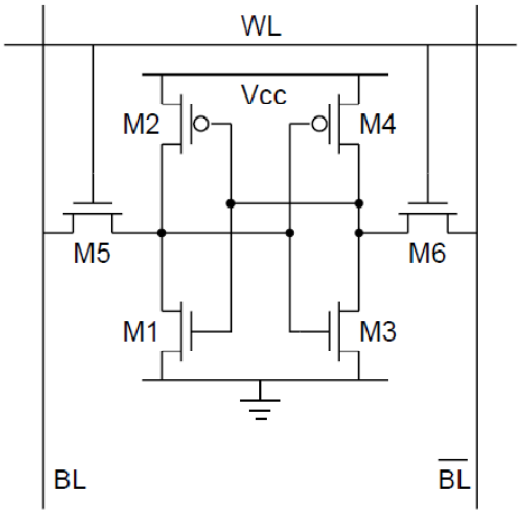

Figure 2.2: Six-Transistor (6T) SRAM Cell [5].

2.4.1.1 Read Operation

The 6T SRAM cell has a differential read operation. This means that both the stored value and its inverse are used in evaluation to determine the stored value. Before the onset of a read operation, the wordline is held low (grounded) and the two bitlines connected to the cell through transistors M5 and M6 (see figure 2.2) are precharged high (to V_{CC}). Since the gates of M5 and M6 are held low, these access transistors are off and the cross-coupled latch is isolated from the bitlines.

If a ’0’ is stored on the left storage node, the gates of the latch to the right are low. That means that transistor M3 (see figure 2.2) is initially turned off. In the same way, M2 will also be off initially since its gate is held high. This results in a simplified model, shown in figure 2.3, for reading a stored ’0’.

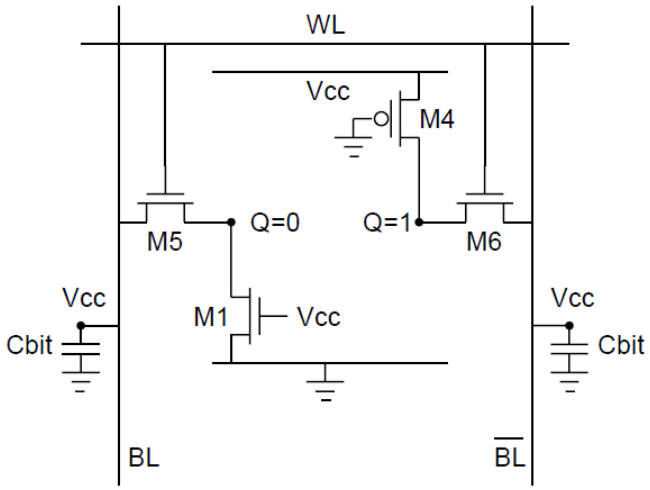

Figure 2.3: 6T SRAM Cell at the onset of read operation (reading '0') [5].

The capacitors, Cbit, (figure 2.3) represent the capacitances on the bitlines, which are several magnitudes larger than the capacitances of the cell. The cell capacitance has here been represented only through the value held by each inverter (Q=0 and Q=1 respectively). The next phase of the read operation scheme is to pull the wordline high and at the same time release the bitlines. This turns on the access transistors (M5 and M6) and connects the storage nodes to the bitlines. It is evident that the right storage node (the inverse node) has the same potential as $\overline{BL}$ and therefore no charge transfer will be take place on this side.

The left storage node, on the other hand, is charged to '0' (low) while BL is precharged to V_{CC}. Since transistor M5 now has been turned on, a current is going from Cbit to the storage node. This current discharges BL while charging the left storage node. As mentioned earlier, the capacitance of BL (Cbit) is far greater than that of the storage node. This means that the charge sharing alone would lead to a rapid charging of the storage node, potentially destroying the stored value, while the bitline would remain virtually unchanged. However, M1 is also turned on which leads to a discharge current from the storage node down to ground. By making M1 stronger (wider) than M5, the current flowing from the storage node will be large enough to prevent the node from being charged high.

After some time of discharging the bitline, a specialized detection circuit called Sense Amplifier (see figure 2.4) is turned on.

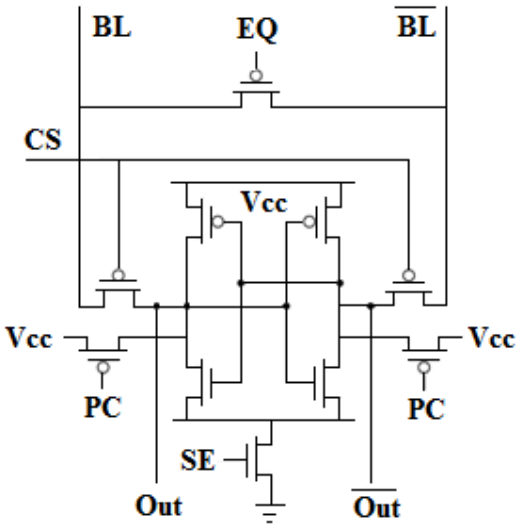

Figure 2.4: Sense amplifier for a 6T SRAM [5].

It detects the difference between the potentials of BL and $\overline{BL}$ and gives the resulting output. Initially the sense amplifier is turned off (sense enable, SE, is low). At the same time as the bitlines of the 6T cell are being precharged high, so are the cross-coupled inverters of the sense amplifier. The bitlines are also equalized (EQ is low) so that any mismatch between the precharges of BL and $\overline{BL}$ is evened out.

When the wordline of the memory cell is asserted EQ and PC are lifted and the precharge of the sense amplifier is discontinued. The column selector CS is then lowered to connect the bitlines to the latch of the sense amplifier. In figure 2.4, for purpose of clarity, only one column selector transistor for each side of the sense amplifier is present. However, normally several bitlines are connected to the same sense amplifier, each one with its own column selector transistor. In this way, several bitlines can be connected to the same sense amplifier, and the column selectors are then used to determine which bitlines should be read.

After some time, when a voltage difference of about 50-100mV (for a 0.18µm process) has developed between the two inverters of the sense amplifier, the sensing is turned on. This is done by raising SE, and thereby connecting the sources of the NMOS transistors in the latch to gnd. Since the internal nodes were precharged high the NMOS transistors are open and current is being drawn from the nodes. The side with the highest initial voltage will make the opposite NMOS (since it is connected to its gate) draw

current faster. This will make the lower node fall faster and in turn shut of the NMOS drawing current from the higher node. An increased voltage difference will develop and eventually the nodes will flip to a stable state.

The Out node in figure 2.4 is then connected to a buffer to restore the flank of the signal and to facilitate driving of larger loads. Also the $\overline{\text{Out}}$ node is usually connected to an inverter. This inverter is of the same size as the first inverter in the buffer. This is to make sure that the two sense amplifier nodes have the same load, and therefore will be totally symmetric.

Note that it is essentially the ’0’ that is detected for the standard 6T SRAM, since the side with the stored ’1’ is left unchanged by the cell. The output is determined by which side the ’0’ is on; ’0’ on the normal storage node results in a ’0’ output while ’0’ on the inverse storage ode results in a ’1’ output. Therefore the performance is mainly dependent on the constellation M1-M5 (see figure 2.2) or M3-M6 and their ability to draw current from the bitline.

2.4.1.2 Write Operation

For a standard 6T SRAM cell, writing is done by lowering one of the bitlines to ground while asserting the wordline. To write a ’0’ BL is lowered, while writing a ’1’ requires $\overline{\text{BL}}$ to be lowered. Why is this? Let’s take a closer look at the cell when writing a ’1’ (figure 2.5).

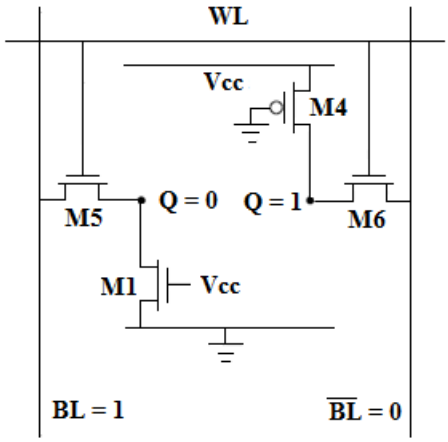

Figure 2.5: 6T SRAM Cell at the onset of writing operation (writing ‘0’ to ‘1’) [5].

As in the previous example of a read, the cell has a ’0’ stored and for simplicity the schematic has been reduced in the same way as before. The main difference now is that the bitlines no longer are released. Instead they are held at V_{CC} and gnd respectively. It can be seen from the left side of the memory cell (M1-M5) that it is virtually identical to the read operation (figure 2.3). Since both bitlines are now held at their respective value, the bitline capacitances have been omitted.

During the discussion of read operation, it was concluded that transistor M1 had to be stronger than transistor M5 to prevent accidental writing. Now in the write case, this feature actually prevents a wanted write operation. Even when transistor M5 is turned on and current is flowing from BL to the storage node, the state of the node will not change.

As soon as the node is raised transistor M1 will sink current to ground, and the node is prevented from reaching even close to the switching point. So instead of writing a ’1’ to the node, a ’0’ will be written to the inverse node. Looking at the right side of the cell we have the constellation M4-M6. In this case $\overline{BL}$ is held at gnd. When the wordline is raised M6 is turned on and current is drawn from the inverse storage node to $\overline{BL}$. At the same time, however, M4 is turned on and, as soon as the potential at the inverse storage node starts to decrease, current will flow from V_{CC} to the node. In this case M6 has to be stronger than M4 for the inverse node to change its state. The transistor M4 is a PMOS transistor and inherently weaker than the NMOS transistor M6 (the mobility is lower in PMOS than in NMOS).

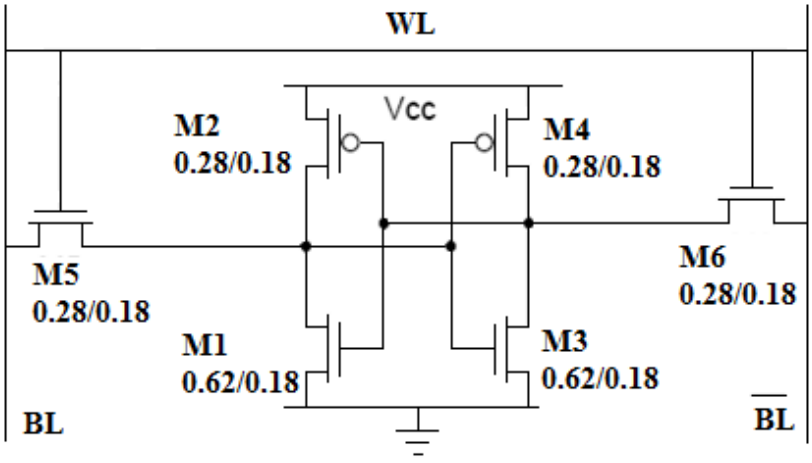

Figure 2.6: Standard 6T SRAM Cell with sizes [6].

Therefore, making both of them minimum size, according to the process design rules, will assure that M6 is stronger and that writing is possible. When the inverse node has been pulled low enough, the transistor M1 will no longer be open and the normal storage node will also flip, leaving the cell in a new stable state.

Figure 2.6 shows the sizing for the 6T SRAM cell used for comparisons in this thesis.

2.4.2 Resistive Load SRAM

2.4.2.1 Cell Structure

The resistive load SRAM is very closely related to the 6T SRAM. The only difference is that the PMOS transistors of the latch have been exchanged for highly resistive resistor elements (figure 2.7). The resistors sole purpose is to maintain the state of the cell by compensating for the leakage current. To reduce static power dissipation the resistor values must be very high. Un-doped polysilicon with a resistance of several TΩ/square is used. In terms of area this exchange is fairly good (about 30-50%), but it leads to a higher static power and a lower Static Noise Margin (SNM). Also special process steps are needed which increases the cost. The resistive load SRAM is therefore not used in sensitive applications, such as microprocessor cache.

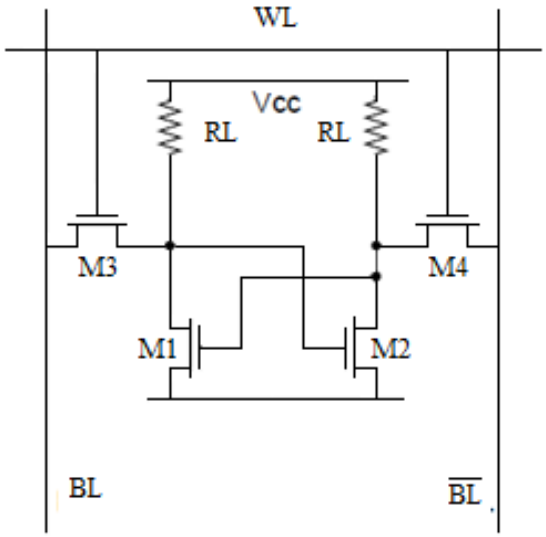

Figure 2.7: Resistive load SRAM cell.

2.4.2.2 Read Operation

The read operation is identical to the 6T SRAM case (see section 2.4.1.1).

2.4.2.3 Write Operation

The write operation is identical to the 6T SRAM case (see section 2.4.1.2).

2.4.3 3T-DRAM

The three-transistor dynamic RAM (3T DRAM) cell consists of three transistors and one capacitor and is used in DRAMs. For its detailed description refer Appendix B.

2.4.4 1T-DRAM

The one-transistor dynamic RAM (1T DRAM) cell consists of one transistors and one capacitor and is used in DRAMs. For its detailed description refer Appendix B.

2.5 CACHE

2.5.1 Introduction

One of the major applications of SRAMs is in cache memories in computers. Cache memory is a relatively small, high-speed memory that stores the most recently used instructions or data from the larger but slower main memory. Cache memory can also use dynamic RAM (DRAM). Typically, SRAM is several times faster than DRAM. Overall, a cache memory gets stored information to the microprocessor much faster than if only high-capacity DRAM is used. Cache memory is basically a cost-effective method of improving system performance without having to resort to the expense of making all of the memory faster.

The concept of cache memory is based on the idea that computer programs tend to get instructions or data from one area of main memory before moving to another area. Basically, the cache controller "guesses" which area of the slow dynamic memory the CPU (central-processing unit) will need next and moves it to the cache memory so that it is ready when needed. If the cache controller guesses tight, the data are immediately available to the microprocessor. If the cache controller guesses wrong, the CPU must go

to the main memory and wait much longer for the correct instructions or data. Fortunately, the cache controller is right most of the time.

Cache, when talking about a microprocessor, is the general term for memory that is embedded on a processor chip (however the term secondary cache is sometimes used for memory off chip). The purpose of the cache memory is to store instructions and data that are likely to be used soon by the processor. Since this memory is embedded on the chip, latency is much shorter than for the off-chip main memory. Also it can usually run at higher clock a frequency since there are much shorter interconnects and no packaging bonds, which deteriorate the signals, to pass through. With good prediction schemes and a large cache, the performance of the system can be increased enormously.

2.5.2 Cache Levels

A first-level cache (L1 cache) is usually integrated into the processor chip and has a very limited storage capacity. L1 cache is also known as primary cache. This gives an extremely short access time, and therefore provides the highest performance. This cache usually runs at the same clock frequency as the CPU. A second-level cache (L2 cache) is a separate memory chip or set of chips external to the processor and usually has a larger storage capacity than an L1 cache. L2 cache is also known as secondary cache. This is connected to CPU through an internal bus. This results in a larger latency. It is also often run at a lower frequency making it possible with smaller, less performance optimized, cells. Some systems may have higher-level caches (L3. L4, etc.), but L1 and L2 are the most common. Also, some systems use a disk cache to enhance the performance of the hard disk because DRAM, although much slower than SRAM, is much faster than the hard disk drive.

2.5.3 Cache Requirements

There are some very important requirements for a memory when it is to be embedded as on-chip cache. First and foremost it has to be reliable and stable. This is of course true for all memories, but is especially important for cache due to the more extreme performance requirements and area limitations. If embedded in a microprocessor, there is little space for redundancy (extra memory blocks used if certain memory units have defects), and because of the size and complexity of the chips the costs are very high

for each chip. Faulty memories cannot be afforded and a high yield (percentage of working chips on a wafer) is therefore extremely important.

Secondly the memory has to have high performance. The sole purpose of cache is to speed up the operation of the CPU by bridging over the performance gap between main memory and the CPU. Therefore at least some of the on-chip cache is usually clocked at the same frequency as the CPU.

Another important requirement is low power consumption. Today's advanced microprocessors use a lot of power and get very hot as a result. With increasing memory sizes these contribute with more and more power loss. This is especially important in mobile applications where prolonging battery life strongly depend on minimizing power loss. Low power architectures are therefore chosen for cache memories and low leakage is taken into account when the sizing is done.

All of these reasons together have made the 6T SRAM the choice of the day for advanced microprocessor caches.

2.6 NEED OF LOW POWER

Historically, VLSI designers have used circuit speed as the "performance" metric. Large gains, in terms of performance and silicon area, have been made for digital processors, microprocessors, DSPs (Digital Signal Processors), ASICs (Application Specific ICs), etc. In general, "small area" and "high performance" are two conflicting constraints. The IC designers' activities have been involved in trading off these constraints. Power dissipation issue was not a design criterion but an afterthought. In fact, power considerations have been the ultimate design criteria in special portable applications such as wristwatches and pacemakers for a long time. The objective in these applications was minimum power for maximum battery life time.

Recently, power dissipation is becoming an important constraint in a design. There are several reasons for emergence of this issue. Few of which are following:

- Battery-powered systems such as laptop/notebook computer, electronic organisers, etc. The need for these systems arises from the need to extend battery life. Many portable electronics use the rechargeable Nickel Cadmium (NiCd) batteries. Although the battery industry has been making efforts to develop

batteries with higher energy capacity than that of NiCd, a strident increase does not seem imminent. The expected improvement of the energy density is 40% by the turn of the century. Even with the advanced battery technologies such as Nickel-Metal Hydride (Ni-MH) which provide large energy density characteristics the life time of the battery is still low. Since battery technology has offered a limited improvement, Low-power design techniques are essential for portable devices.

- Low-power design is not only needed for portable applications but also to reduce the power of high-performance systems. With large integration density and improved speed of operation, systems with high clock frequencies are emerging. These systems are using high-speed products such as microprocessors. The cost associated with packaging, cooling and fans, required by these systems to remove the heat, is increasing significantly. The power dissipation increases with increase in frequencies.
- Another issue related to high power dissipation is reliability. With the generation of on-chip high temperature, failure mechanisms are provoked, such as silicon interconnects fatigue, package related failure, electrical parameter shift, electromigration, junction fatigue, etc.

2.7 LOW POWER APPLICATIONS

Low-power design is becoming a new era in VLSI technology, as it impacts many applications such as:

- Battery-powered portable systems, for example notebooks, palmtops, CDs, language translators, etc, represent an important growing market in the computer industry. High-performance capabilities, comparable to those of desktops, are demanded. Several low-power microprocessors have been designed for these computers.
- Electronic pocket communication products such as; cordless and cellular telephones, PDAs (Personal Digital Assistants), pagers, etc.
- Sub-GHz processors for high-performance workstations and computers. Since the power consumed is increasing with the trend of frequency increase then

processors with new architectures and circuits optimized for low-power are crucial.

- Other applications such as WLANs (Wireless Local Area Network) and electronic goads (calculators, hearing aids, watches, etc.).

CHAPTER 3 LEAKAGE IN SRAM CELL AND ITS REDUCTION TECHNIQUES

As technology scales down, the supply voltage must be reduced such that dynamic power can be kept at reasonable levels. In order to prevent the negative effect on performance, the threshold voltage (V_T) must be reduced proportionally with the supply voltage so that a sufficient gate overdrive is maintained. This reduction in the threshold voltage causes increase in leakage current every generation, which in turn can increase the static power of the device to unacceptable levels. In this chapter leakage in SRAM cell and its reduction techniques has been discussed.

3.1 LEAKAGE IN SRAM CELL

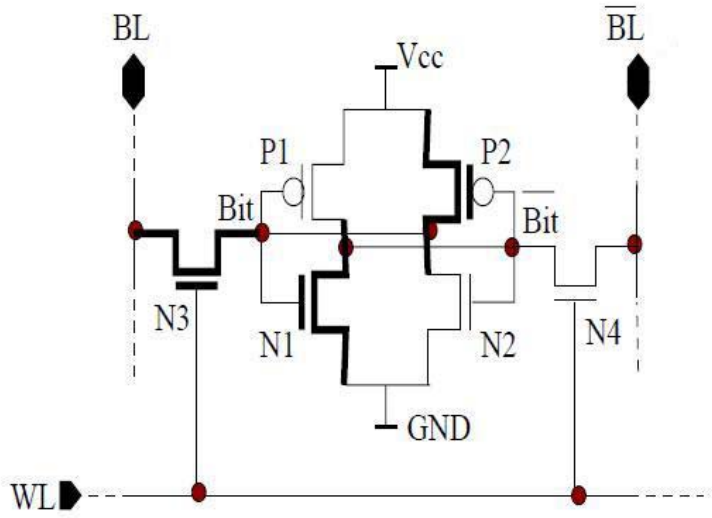

Figure 3.1: Leaking transistors (bold) in 6T cell when Bit ='0' is stored [7].

The memory core is composed of memory cells that are arranged in rows and columns. 6T memory cell architecture is used to explain the leakage in the memory cell. During an idle phase, the wordlines are deselected (WL = '0') and the bitlines are precharged (BL = '1' and $\overline{BL}$ = '1'). Depending upon the memory cell data, either transistors N4, P1, N2 (for bit = '1') or N3, P2, N1 (for bit = '0') will be leaking.

Figure 3.1 shows the transistors in the off state in bold for bit = '0'. In this case N3, N1 and P2 are off and will be leaking. The leakage current in the memory cell would be as shown in equation 3.1 [7].

$$I_{memcellIdle} = I_{Dsub}(N1) + I_{Dsub}(N3) + I_{Dsub}(P2) \quad (3.1)$$

where, I_{Dsub} is the subthreshold leakage current of the MOSFET, which is given by the equation 3.2 [7] and the subthreshold leakage in the whole memory core is given by equation 3.3 [7].

$$I_{Dsub} = I_S e^{\frac{V_{GS}}{nkT/q}} \left[1 - e^{\frac{V_{DS}}{kT/q}}\right] \quad (3.2)$$

where, I_S and n are imperical parameters with $n \geq 1$.

$$I_{memcoreIdle} = N_{rows} \cdot N_{cols} \cdot I_{memcellIdle} \quad (3.3)$$

where, N_{rows} and N_{cols} are the number of rows and columns respectively in the memory core.

Thus to reduce the leakage of a memory cell we have to concentrate on two components of leakage one is the leakage inside the cell and the other is leakage to bitlines. There are various techniques proposed to reduce these leakage components. Some most popular circuit level leakage reduction techniques have been discussed in the next section.

3.2 LEAKAGE REDUCTION TECHNIQUES

Among the emerging leakage reduction techniques [8-18], some require modification of the process technology, achieving leakage reduction during the fabrication/design stage, while others are based on circuit-level optimization schemes that require architectural support, and in some cases, technology support as well, but are applied at run-time (dynamically).

In this section various circuit level leakage reduction techniques have been discussed that have been proposed and are being used to reduce the static power dissipation in the SRAM cell.

3.2.1 Dual V_T

The dual-V_T technique employs transistors with higher threshold voltages in memory cells and faster, leakier transistors elsewhere within the SRAM. This technique requires no additional control circuitry and can substantially reduce the leakage current when compared to low V_T devices [8]. The amount of leakage current is engineered at design time, rather than controlled dynamically during operation. No data are discarded and no additional cache misses are incurred. However, high-V_T transistors have slower switching speeds and lower current drive.

3.2.2 Gated V_{DD}

Gated-V_{DD} enables a cache to "turn off" the supply voltage and eliminate virtually all the leakage energy dissipation in the cache's unused sections [11-12]. The key idea is to introduce an extra transistor with high V_T in the supply voltage (V_{DD}) or the ground path (gnd) of the cache's SRAM cells whereas rest all transistors of the cell have low V_T. The extra transistor is turned on in the used sections and turned off in the unused sections. Thus, the cell's supply voltage is "gated." Gated-V_{DD} maintains the performance advantages of lower supply and threshold voltages while reducing leakage and leakage energy dissipation.

3.2.2.1 SRAM Cell with Gated-V_{DD}

Cache data arrays are usually organized in banks; each bank contains SRAM cell rows, with each row containing one or more cache blocks. Figure 3.2 shows a cache SRAM cell using an NMOS gated-V_{DD} transistor. PMOS gated-V_{DD} is achieved by connecting the gated-V_{DD} transistor between V_{DD} and the SRAM PMOS transistors. The gated-V_{DD} transistor is turned on for the cell to be in "active" mode and turned off for the cell to be in "standby" mode.

Gated-V_{DD} transistor can be shared among multiple circuit blocks to reduce the overhead. To reduce the impact on SRAM cell speed and to ensure stability of the SRAM, the gated-V_{DD} transistor must be carefully sized with respect to the SRAM cell transistors it is gating. While a gated-V_{DD} transistor must be made large enough to sink the current flowing through the SRAM cells during a read/write operation in the active mode, too large a gated-V_{DD} transistor may reduce the stacking effect, thereby

diminishing the energy savings. Moreover, large transistors also increase the area overhead.

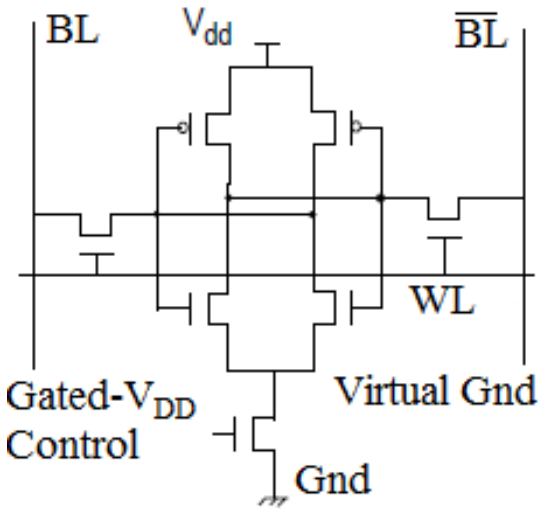

Figure 3.2: SRAM with an NMOS gated V_{DD} [11].

3.2.2.2 NMOS vs. PMOS Gated transistor

Using a PMOS or an NMOS gated-V_{DD} transistor presents a trade-off between area overhead, leakage reduction, and impact on performance. To maintain stability and high SRAM cell speed, an NMOS gated-V_{DD} transistor needs to be sufficiently wide. One estimate is to use the sum of the widths of all the transistors that could simultaneously switch in the SRAM cells.

If an entire cache block is connected to a single NMOS gated-V_{DD} transistor, the desired width of the transistor may be determined as the product of the width of one of the SRAM cell's NMOS transistors (because only one of the two is "on" during a read) and the number of cells in the cache block. Such a wide NMOS gated-V_{DD} transistor may incur a high area overhead.

Using NMOS gated-V_{DD} transistors, however, substantially reduces standby energy dissipation through the stacking effect of three NMOS transistors between the bitlines and ground. Alternatively, using a PMOS gated-V_{DD} transistor significantly reduces the required transistor width. Dual-bitline architectures typically precharge the bitlines before read operations, so the PMOS transistors simply help in holding the cell value intact and

do not contribute to read operations. It reduces the required gated-V_{DD} transistor width, resulting in a negligible area overhead.

A PMOS gated-V_{DD} transistor, however, does not create the isolation between the bitlines and the ground as does an NMOS transistor, reducing the amount of energy saving. So when PMOS gated-V_{DD} is used access transistors can also be made high-V_T to reduce bitline leakage. When an SRAM cell with an NMOS gated-V_{DD} is read, the discharging of the precharged bitlines takes longer due to the non-gnd voltage at the virtual gnd. In contrast, because the PMOS cell transistors do not contribute to read operations, a PMOS gated-V_{DD} transistor does not significantly impact the cell performance.

3.2.2.3 Limitations of Gated-V_{DD} Approach

The drawback of this approach is that the state of the cache line is lost when it is turned off because the source line and the ground line of the internal circuit become floating nodes by high-threshold transistors [16]. So, it requires some means of holding the latched data in the sleep period which increases the design complexity and the chip area, and has a significant impact on performance.

3.2.3 ABC-MTCMOS

Turning off cache lines is not the only way that leakage energy can be reduced. Significant leakage reduction can also be achieved by putting a cache line into a low-power drowsy mode. When in drowsy mode, the information in the cache line is preserved; however, the line must be reinstated to a high-power mode before its contents can be accessed. Auto-Backgate-controlled multi-threshold technique controls the backgates to reduce the leakage current when the SRAM is not activated (sleep mode) while retaining the data stored in the memory cells [14-15]. It can reduce the leakage current significantly using a simple circuit while in the sleep mode. In order to reduce undesirable leakage current in the sleep mode, the backgate bias is automatically controlled to increase the threshold voltage.

3.2.3.1 ABC-MTCMOS Circuit

Figure 3.3 shows the concept of the ABC-MT-CMOS circuit. Here Q1 and Q2, which are higher threshold transistors than those for the internal circuit, act as a switch to

cut off the leakage current. When the cell is in the active mode these transistors are turned on. The virtual source line, VVDD, becomes 1.0 V supplied by the voltage source vdd1 through Q1. Another virtual source line, VGND, is forced to ground level through Q2. The internal circuits consist of only low-threshold transistors. In the active mode, the dynamic current and static leakage current flows from vdd1 to ground, as denoted by Idd (active) in figure 3.3. If the switch transistors Q1 and Q2 turn off in the sleep mode, the current Idd (active) can be reduced, however, the data stored in the memory cell disappears.

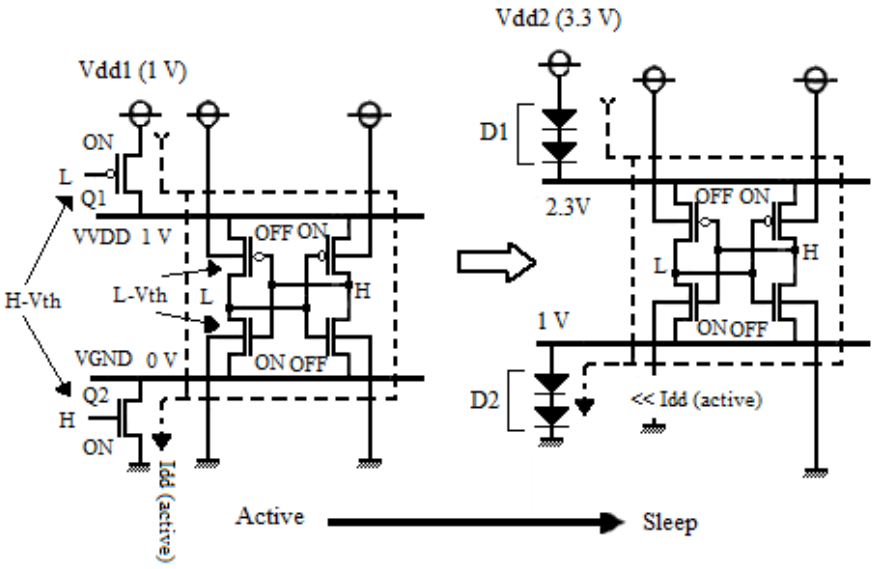

Figure 3.3: Concept of ABC-MTCMOS Circuit [14].

In the ABC-MT-CMOS, higher voltage source Vdd2 (3.3 V) and diodes D1 and D2 have been used, in order to reduce the leakage current with retaining the stored data. In the sleep mode, the VVDD is connected to Vdd2 through D1, and VGND is connected to ground through D2. Here, the diodes D1 and D2 consist of two diodes each. If it is assumed that the forward bias of one diode is 0.5 V, the forward voltages of Dl and D2 will be 1.0 V. Then, the VVDD and VGND become about 2.3 V and 1 V, respectively. The static leakage current Idd(sleep), which flows from Vdd2 to ground, decreases significantly compared with that of the active mode, because the threshold voltage of the internal transistors increase by its backgate bias effect. In the sleep mode, VVDD and VGND maintain their voltage levels owing to the weak leakage current Idd(sleep), so that the data stored in the memory cell is retained.

Figure 3.4 shows the actual configuration of the ABC-MTCMOS circuit. There are two additional high-threshold transistors Q3 and Q4. In the active mode, SL = "L" and SL (bar) ="H" is applied and Q1, Q2 and Q3 turn on and Q4 turns off. Then both VVDD and the substrate bias, BP, becomes 1.0 V. On the other hand, in the sleep mode, SL ="H" and SL (bar) ="L" is applied and Q1, Q2 and Q3 turn off and Q4 turns on. Then BP becomes 3.3 V. The static leakage current, which flows from Vdd2 to ground through D1 and D2, determines the voltages Vd1, Vd2 and Vm. Here Vd1 denotes the bias between the source and substrate of the p-MOS transistors, Vd2 denotes that of the n-MOS transistors, and Vm denotes the voltage between VVDD and VGND.

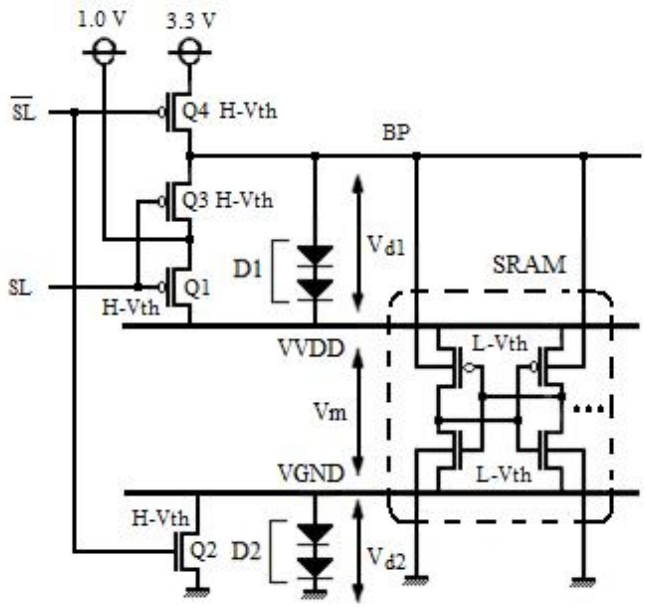

Figure 3.4: Configuration of ABC-MT-CMOS circuit [14].

3.2.3.2 Limitations of ABC-MTCMOS

It has slow switching between low power and low power modes and vice versa [16].

3.2.4 DVS (Dynamic Voltage Scaling)

In this method to reduce the leakage power of SRAM cells, in active mode, the standard supply voltage is provided to the SRAM cells. However, when cells are not intended to be accessed for a time period, they are placed in a sleep or “drowsy” mode by

supplying a low voltage (required to retain the cell information) to the SRAM cells. In drowsy mode, the leakage power is significantly reduced due to the decreases in both leakage current and supply voltage [16-17].

3.2.4.1 SRAM Leakage Power Reduction Using DVS

Basically, there are two off-state leakage current paths through the two cross-coupled inverters in the standard SRAM cell, as shown in Figure 3.5. The leakage current reduces super linearly with V_{DD}, and hence, significant reduction in leakage power can be obtained in drowsy mode. In drowsy mode, a minimum voltage must be applied to maintain state. It was found that, despite process variations, this state-preserving voltage is quite conservative, and that the state-preserving supply voltage can be even reduced further if necessary.

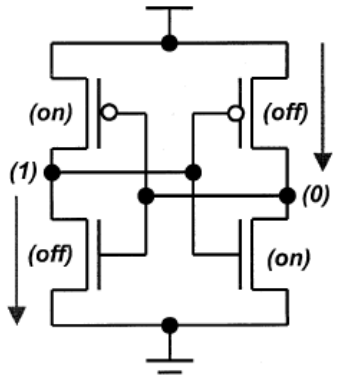

Figure 3.5: Leakage inside a SRAM cell [16].

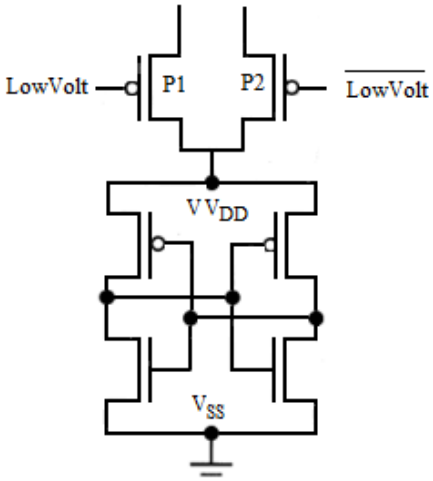

Figure 3.6: A drowsy SRAM cell with the supply voltage control mechanism [16].

Figure 3.6 shows drowsy SRAM cell with the supply voltage control mechanism. The two PMOS transistors, P1 and P2, control the supply voltage of the memory cell based on the operating mode: active or standby mode. When the cell is in the active mode, P1 supplies a standard supply voltage, and P2 supplies a standby voltage when it is in the drowsy mode. P1 and P2 are controlled by complementary supply voltage control signals.

In drowsy mode, however, SRAM cell accesses are not allowed, because the bit line voltage is higher than the storage cell core voltage, which may result in destroying the cell state. Moreover, the sense-amplifier may not operate properly at the low storage cell core voltage, because the memory cell does not have enough driving capability to sink the charge stored in the bit lines. While building a cache system for a microprocessor with a drowsy SRAM circuit technique, it can be applied to either each cache-line or on a subbank basis, depending on the microarchitectural control mechanism; each cache line or subbank shares the same virtual supply voltage node (VV_{DD}).

3.2.4.2 Limitations of DVS Approach

Dynamic Voltage scaling needs two sources and also becomes process variation dependent in the drowsy mode due to low power supply [16].

CHAPTER 4 DESIGN OF 5T SRAM CELL

One version of a 5T SRAM was presented in 1996 by Hiep Tran [19]. That cell differs fundamentally from the cell used in this thesis, in that the latch of the cell is disconnected from the gnd supply to facilitate write. This requires an additional metal wire and also destabilizes all cells on the bitline during write. The architecture used to design 5T SRAM cell in this thesis is proposed by Ingvar Carlson in 2004 [6]. The design and all simulations are carried out at 180nm technology.

4.1 CELL STRUCTURE

In a normal 6T cell both storage nodes are accessed through NMOS pass-transistors. This is necessary for the writing of the cell since none of the internal cell nodes can be pulled up from a stored ’0’ by a high on the bitline. If this was not the case an accidental write could occur when reading a stored ’0’.

However, if the bitlines are not precharged to V_{CC} this is no longer true. With an intermediate precharge voltage, V_{PC}, the cell could be constructed so that a high on the bitline would write a ’1’ into the cell, but a precharged bitline with a lower voltage would not. Also a low on the bitline could write a ’0’ into the cell, whereas the intermediate precharge voltage would not, thus giving the cell a precharge voltage window (see section 4.4.3) where correct operation is assured. This would eliminate the need for two NMOS transistors, since the cell now can be written both high and low from one side. In turn, that would also result in one less bitline. From a high density point of view this is very attractive. Figure 4.1 shows the structure of the resulting five-transistor (5T) SRAM cell. With one less bitline the 5T cell also shares a sense amplifier between two cells. This further reduces the area giving the 5T memory block an even greater advantage over the 6T SRAM.

4.2 READ OPERATION

The operation scheme when reading a 5T cell is very similar to the 6T SRAM. Before the onset of a read operation, the wordline is held low (grounded) and the bitline is

precharged. This time however, the bitline is not precharged to V_{CC}, but to another value, V_{PC}. This value is carefully chosen according to stability and performance requirements.

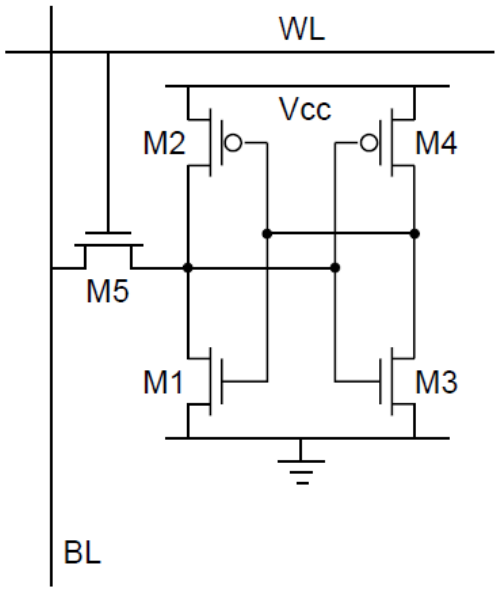

Figure 4.1: Five-Transistor (5T) SRAM cell [6].

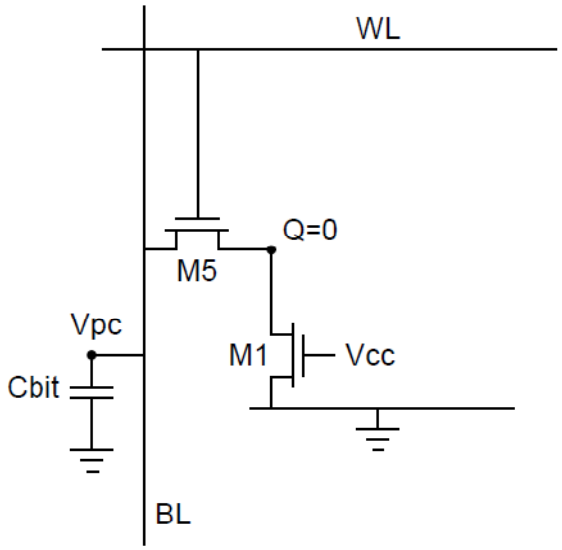

Figure 4.2: Five-Transistor SRAM cell at the onset of read operation (Reading '0') [6].

Figure 4.2 shows the simplified schematic corresponding to the onset of a read '0' operation. Note that the bitline now has been precharged to V_{PC}. One drawback of the intermediate precharge value is the apparent problem of obtaining this voltage.

One obvious way is to supply this voltage externally. The trend today is that microprocessors demand several different supply voltages, so this might in fact not be a significant drawback.

The next phase of the read operation scheme is to pull the wordline high and at the same time release the bitline. This turns on the access transistor M5 and connects the storage node to the bitline. If reading a ’0’, BL will now be pulled down through the transistor combination M5-M1. If instead a ’1’ is to be read, the situation is slightly different from the 6T case. Figure 4.3 shows the simplified schematic corresponding to the onset of a read ’1’ operation. In this case the PMOS transistor M2 is used to pull the bitline up, whereas for the 6T cell it was only used to hold the stored value internally. The implication of this is that M2 now has to be sized a bit differently since it affects the performance of the cell. These sizing issues are described more thoroughly later.

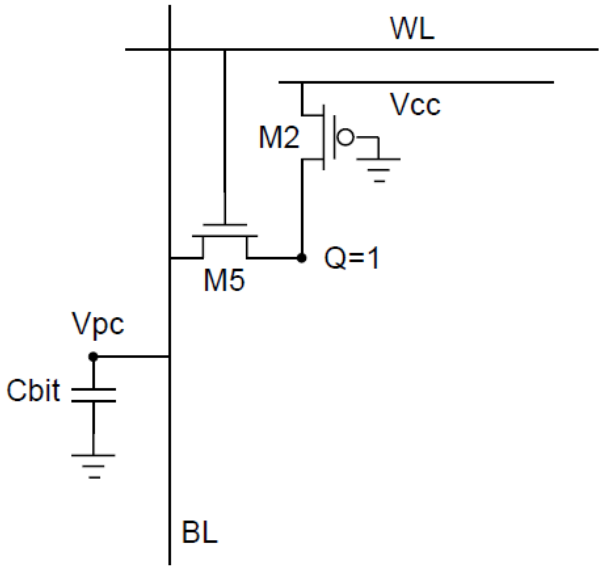

Figure 4.3: Five-Transistor SRAM cell at the onset of read operation (Reading ‘1’) [6].

Another apparent difference between the 5T SRAM and the 6T SRAM is how the sensing of the stored value is done. While the 6T cell has two bitlines and the stored value is sensed differentially, the 5T cell only has one bitline. Depending on the value stored, the 5T bitline is either raised or lowered. The challenge then becomes how to determine if the bitline voltage has increased or decreased. A few different techniques can be used for this. One idea might be to use a type of sample and hold circuit that would sample the

value before the read and then use this value as a reference in a differential sense amplifier.

The advantage of this is that the regular sense amplifier used for the 6T SRAM is quite small and fast and has been used extensively and therefore has very well known properties. The disadvantage is the extra circuit and read scheme complexity that comes with the "sample and hold" circuit [6].

Instead, another way of obtaining the needed reference can be used. If the memory is partitioned in two sections so that only one section will be accessed at any given time, the other section can be used as a reference. In other words, one bitline from section one is connected to one side of the sense amplifier, and one bitline from section two is connected to the other side of the sense amplifier. This technique is in fact often used in 1T DRAMs since the 1T cell also have only one bitline.

One implication of this scheme is that the output from the sense amplifier will either be OUT or $\overline{\text{OUT}}$ depending on which section is accessed. This is because one is connected to the $\overline{\text{BL}}$ side of the sense amplifier and a higher value on that line will result in a low output. Another thing that should be noticed is that since the bitline is not precharged to V_{CC} as in the 6T case, the column selector transistor will be more efficient if a NMOS transistor is used.

4.3 WRITE OPERATION

Writing in the 5T SRAM cell differs from the 6T cell mainly by the fact that it is done from only one bitline (see section 4.2). For the 5T cell the value to be written is held on the bitline, and the wordline is asserted. Since the 6T cell was sized so that a '1' could not be written by a high voltage on the bitline, the 5T cell has to be sized differently.

4.4 OPERATION STABILITY

4.4.1 Read Stability

The first important property when reading a static memory cell is that the cell does not flip state (accidental write) while trying to read the stored value. For the 5T SRAM cell this occurs, for a stored '0', when the voltage of the storage node (the node common

to M1, M2, and M5) exceeds the switching threshold of the M3-M4 inverter. Simplified, it can be viewed in terms of the read current drawn or supplied to the storage node (see figure 4.2).

The currents through the transistors M5 and M1 can be described as in equation 4.1 and 4.2 [5] respectively if the channel length modulation is ignored.

$$I_D = k'_n \frac{W}{L}((V_{GS}-V_T)V_{DS}-\frac{V_{DS}^2}{2}) \quad 4.1$$

$$I_D=k'_n \frac{W}{L}((V_{GS}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}) \quad 4.2$$

At the switching point, V_M, the current drawn (when reading a stored ’0’) through transistor M1 must be greater than the current supplied from the bitline through M5. Otherwise the node will rise and the cell will flip. This relation can be written as in equation 4.3. This is a simplified view not taking the changes of the feedback (output from the M3-M4 inverter) into account.

When the storage node is getting close to the switching value a significant change in the feedback will occur. However, if the formulas are used with a lower V_M than the actual switching voltage, this gives a safety margin and the changes in feedback will have little effect on the calculations. Furthermore, these formulas are used to give an understanding of the concept rather than calculating the transistor ratios.

$$k'_n W_1/L_1\,((V_{CC}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}) > k'_n W_5/L_5\,((V_{DD}-V_M-V_T)(V_{PC}-V_M)-(V_{PC}-V_M)^2/2) \quad 4.3$$

Now one thing should be pointed out. If the precharge voltage V_{PC} is chosen to be equal to the switching voltage V_M the right side of the equation equals zero and the relation is always true. This is easy to understand because if the bitline is not precharged higher than the switching point, it can never pull the storage node over that point. A more interesting situation, from a sizing point of view, is therefore when V_{PC} is chosen higher. For explanatory purposes, V_{PC} is chosen at 1.2V, and all other values are substituted as V_{CC}=1.8V, V_M=0.9V, and V_T =0.4V. The relation can then be simplified as in equation 4.4.

$$\frac{W_1}{L_1} > 0.22\frac{W_5}{L_5} \qquad 4.4$$

In other words, with V_{PC}=1.2V and both transistors at minimum length, M5 can be 4.5 times wider than M1 without destroying the stored value while reading a ’0’. This is for a sizing of the inverter M3-M4 resulting in a centred switching point (V_M=0.9V). It can also be seen from equation 4.3 that a higher V_{PC} will lower the acceptable M5/M1 ratio.

A similar discussion can be made regarding the reading of a ’1’ (see figure 4.3). The difference is that the PMOS transistor M2 is active while M1 is turned off, and that the acceptable M5/M2 ratio is lowered by a lower V_{PC}. The mobility will then also be a factor which cannot be cancelled out in the equations, but for all other purposes the discussions remain the same.

4.4.2 Write Stability

To make it possible with one sided write operations, the cell must be sized accordingly. In section 4.4.1 some guidelines were given for the sizing of transistors M1, M2 and M5 to facilitate a proper read operation. For instance it was concluded that under certain conditions M5 could be at most 4.5 times wider than M1. However, to facilitate proper write operation there is also a minimum value for the M5/M1 ratio.

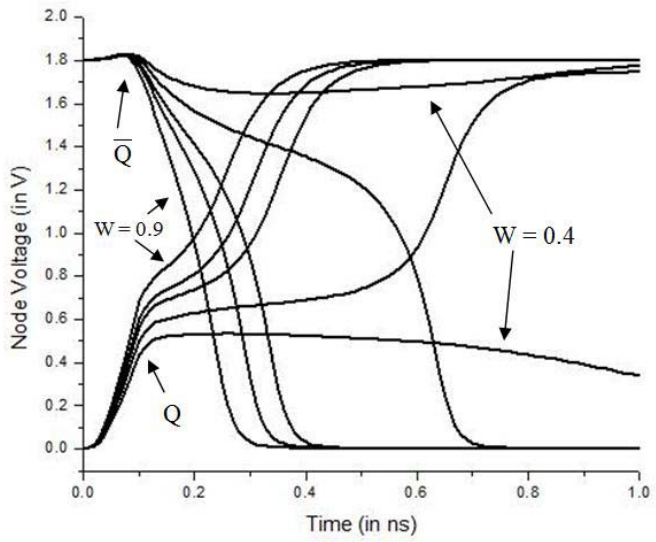

Figure 4.4: Node Voltages at different widths of access transistor.

Now to design, at first, all transistors are kept at minimum size, that is, W (=3λ) =0.28μm and L (=2λ) =0.18μm. The 5T cell with the minimum sizing will be stable for read '0' according to section 4.4.1 since M5 is not 4.5 times larger than M1. However, it also has to be instable enough so that a high voltage (V_{CC}) on BL will write the cell within a reasonable amount of time. Therefore simulations were done for different widths of transistor M5 during a write '1' operation.

Figure 4.4 shows the internal cell nodes during the operation for different values of the width of the access transistor. The width was varied from 0.4 μm to 0.9 μm. From the figure, it is evident that the width of M5 must exceed 0.4 μm or the cell will not be written at all.

4.4.3 Precharge Voltage Window

In the above discussions of read stability (section 4.4.1) it was concluded that the bitline precharge voltage (V_{PC}) was one factor in determining the stability. In fact the possible sizes of the transistors depend on the value of V_{PC}. In section 4.4.2 a preliminary sizing was proposed, which allow both reading and writing of the cell if a proper precharge voltage is used.

Now the question is within what limits BL can be precharged and still allow for proper read operation (note: V_{PC} does not affect the write operation since the bitline is held at either V_{CC} or gnd during write). In other words, what is the Precharge Voltage Window for the above sizing? To answer this question, a few simulations were made with different values of V_{PC} to show when the read operations would fail. One study was made of the read '0' case, giving the upper boundary of the window, and another one of the read '1' case, giving the lower boundary. The results can be seen in figure 4.5 and figure 4.6 respectively.

For read '0' case (figure 4.5) V_{PC} has been varied from 0.80V to 0.95V in steps of 0.05V and for read '1' case (figure 4.6) V_{PC} has been varied from 0.45V to 0.60V in steps of 0.05V. It can be seen that for the read '0' case (figure 4.5) internal nodes flip if BL is precharged to 0.95V but not for VPC = 0.90V. Also for read '1' case (figure 4.6) internal nodes flip if BL is precharged to 0.45V but not for V_{PC} = 0.50V. So, precharge window for this configuration is 0.50V to 0.90V. Thus V_{PC} can be kept within this range for proper read operation.

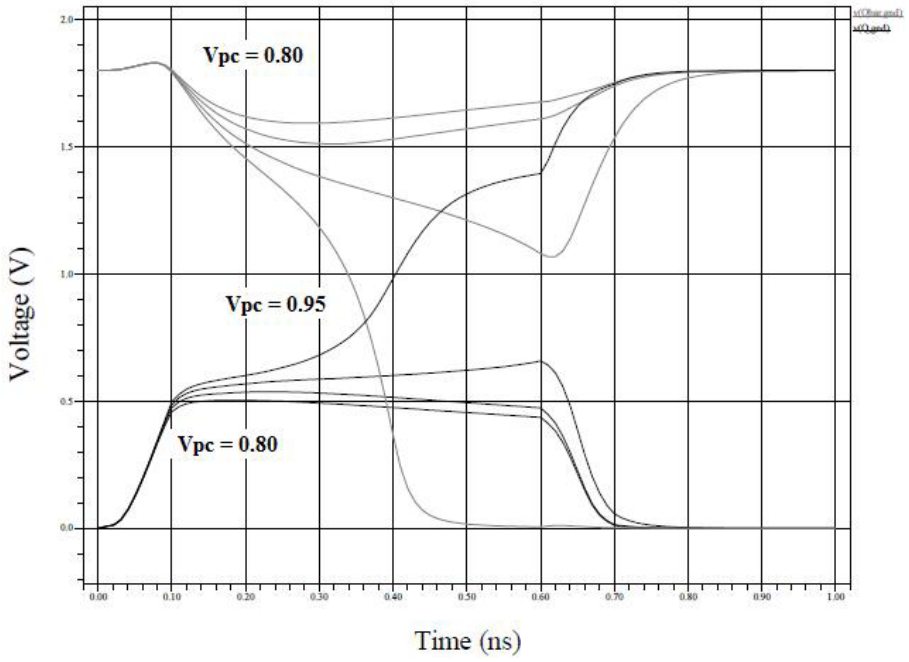

Figure 4.5: Internal cell nodes Q and $\overline{Q}$ for 5T cell reading '0' with V_{PC} varying from 0.80 V to 0.95 V.

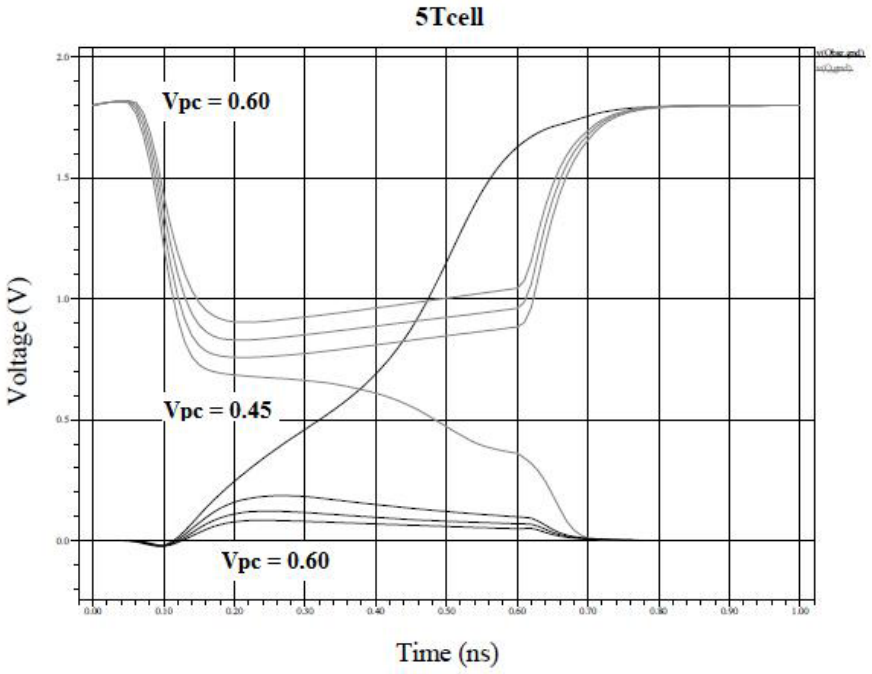

Figure 4.6: Internal cell nodes Q and $\overline{Q}$ for 5T cell reading '1' with V_{PC} varying from 0.45 V to 0.60 V.

4.5 FINAL DESIGN OF 5T CELL

Keeping all stability factors (discussed in section 4.4) in consideration simulations have been done to design 5T SRAM cell having comparable performance with 6T SRAM cell. The designed 5T SRAM cell with sizes has been shown in figure 4.7.

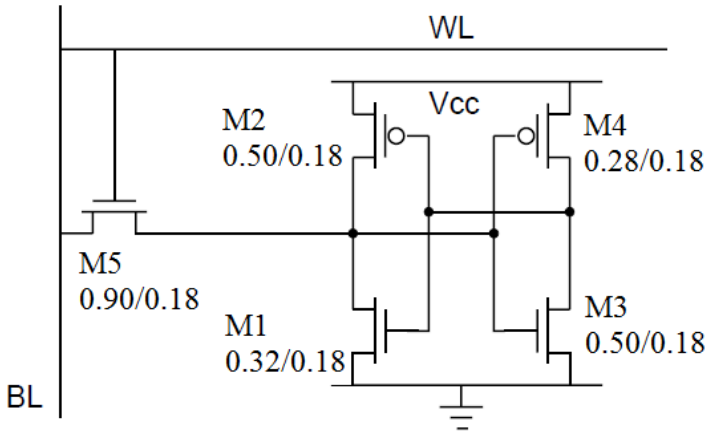

Figure 4.7: Design of 5T SRAM Cell with sizes.

CHAPTER 5 RESULTS AND CONCLUSION

5.1 RESULTS AND DISCUSSIONS

The standard 6T SRAM cell at 180 nm technology (shown in figure 2.6) is analyzed in terms of leakage power dissipation and performance. After that different circuit level leakage reduction techniques (discussed in Chapter 3) have been applied and analyzed. Leakage power in the cell has been evaluated by keeping wordline low to disconnect cell from the bitlines. Two components of the leakage have been considered, one the leakage inside the cell and other leakage to bitlines. For performance evaluation write and read access times are measured from the midpoint of the transition of wordline from '0' to '1', up to the point where both nodes (Q and $\overline{Q}$) flips for write access time, and to the point when a difference of 100mV is generated between both bitlines connected to the cell for the read access time. It is assumed that sense amplifier used for reading can read a difference of 100mV between both the bitlines.

Table 5.1 shows the leakage power dissipation and the read, write access times in the 6T SRAM cell shown in figure 2.6. Table 5.2 shows the comparison of leakage power dissipation of the conventional 6T cell and the 6T cell with various leakage reduction techniques.

Table 5.1: Leakage power and performance of 6T cell

Metrics	**Standard 6T cell**
Read time (WL high up to 100mV difference in bitlines)	336ps
Write time (WL high up to node flips)	76ps
Leakage Power/cell	2.03nW

It can be seen that Gated-V_{DD} technique can reduce leakage very efficiently but it has a disadvantage that it does not retain the cell state in the low power mode. DVS can

also reduce leakage to great extent but it needs two power supplies and in the low power mode it becomes process variation dependent. ABC-MTCMOS can retain cell state in the low power mode but it is not much efficient in reducing leakage compared to Gated-V_{DD} and DVS also it is slow in switching from low power to normal mode and vice versa.

Table 5.2: Comparison of leakage power reduction techniques

Leakage Reduction Technique	**Leakage Power Dissipation/Cell (in nW)**	**Percentage Reduction**
Conventional	2.030	-
ABC-MTCMOS	1.452	28.5
DVS	0.230	88.7
Gated-V_{DD}	0.033	98.3

Table 5.3 shows the leakage power dissipation and the read, write access times in the 5T SRAM cell shown in figure 4.7. It can be seen that 5T cell offers lesser leakage as compared to 6T cell and also gives comparable performance while consuming lesser area.

Table 5.3: Leakage power and performance of 5T cell

Metrics	**Standard 6T cell**
Read time (WL high up to 100mV difference in bitlines)	365ps
Write time (WL high up to node flips)	102ps
Leakage Power/cell	1.79nW

Table 5.4 shows the comparison of leakage in 6T and 5T SRAM cell with various leakage reduction techniques. It can be seen that leakage in 5T SRAM cell can be further reduced after applying various leakage reduction techniques.

Table 5.4: Comparison of leakage power dissipation in 6T and 5T cell

Leakage Reduction Technique	**Leakage Power Dissipation/Cell (in nW)**		**Percentage Reduction**
	6T	**5T**	
Conventional	2.030	1.790	11.8
ABC-MTCMOS	1.452	1.404	03.3
DVS	0.230	0.170	26.0
Gated-V_{DD}	0.033	0.029	12.1

5.2 CONCLUSION

Various circuit level techniques have been applied to 6T and designed 5T SRAM cell for leakage power reduction and compared. Out of all the techniques discussed DVS has found to be the best as it reduces leakage comparable to Gated V_{DD} as well as retain the cell information.

It has been found that in conventional 6T SRAM cell up to 98% reduction in leakage power can be achieved using these techniques. With conventional 5T cell about 11.8% leakage power reduction has been achieved than conventional 6T cell. Further applying the leakage reduction techniques to the 5T cell has shown 26% more reduction in leakage than in the case of 6T cell.

One can choose any method to reduce leakage depending upon requirements. DVS leakage reduction technique shows large reduction in leakage while retaining the cell information. DVS technique can be combined with the 5T SRAM cell to enhance the leakage reduction significantly.

5.3 FUTURE SCOPE

In this thesis various circuit level leakage power reduction techniques have been analyzed with 6T and 5T SRAM cell at 180nm technology. A large reduction in leakage has been observed. As memory cells being discussed have to be used in cache memory

their stability is also very important. So stability analysis of both 6T and 5T cells after applying leakage reduction techniques can be analyzed.

Further architecture level techniques such as bitline segmentation and sub-banking of memory array can also be combined with discussed circuit level techniques to enhance the reduction [20-22].

Device level techniques such as Retrograde well, Halo-doping and LDD (Light Doped Drain) implantation can be employed for leakage reduction in individual MOSFETs which eventually will reduce in large reduction. As leakage will be more significant beyond 100nm technology so this work should be extended to higher technologies such as 90nm, 70nm or beyond.

REFERENCES

[1]. V. De and S. Borkar, “Technology and design challenges for low power and high performance”, International Symposium Low Power Electronics and Design, pp.163-168, 1999.

[2]. S. Borkar, “Technology trends and design challenges for microprocessor design”, ESSIRC, pp. 7-8, Sep. 1998.

[3]. S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. Chandrakasan, “Scaling of stack effect and its application for leakage reduction,” in Proc. IEEE/ACM International Symposium on Low Power Electronics and Design, pp. 195–200, Aug. 2001.

[4]. T. Floyd, “Digital Fundamentals”, Prentice Hall, ninth edition, 2006.

[5]. J. M. Rabaey, A. Chandrakasan, and B. Nikolic, “Digital Integrated Circuits: A Design Perspective”, Prentice Hall series in electronics and VLSI, Prentice Hall, second edition, 2003.

[6]. I. Carlson, S. Anderson, S. Natarajan and A. Alvandpour, “ A high density, low leakage, 5T SRAM for embedded caches”, Proceedings of the 30th Solid State Circuits Conference, ESSCIRC, pp. 215-218, September 2004.

[7]. M. Mamidipaka, K. Khouri, N.Dutt, and M. Abadir, “Analytical models for leakage power estimation of memory array structures”, International Conference on Hardware/Software and Co-design and System Synthesis (CODES±ISSS), pp. 146-151, 2004.

[8]. J. T. Koa and A. P. Chandrakasan, “Dual threshold voltage techniques for low-power digital circuits”, in IEEE Journal of solid state Circuits, Vol. 35, No.7, pp. 1009-1018, Jul. 2000.

[9]. B. Amelifard, F. Fallah, M. Pedram, “Reducing the sub-threshold and gate-tunneling leakage of SRAM cells using dual-v_t and dual-t_{ox} assessment”, in IEEE Proceedings of Design, Automation and Test, Vol. 1, pp. 1-6, 2006.

[10]. C. H. Kim and K. Roy, “A leakage tolerant cache memory for low voltage microprocessors,” Proceedings of the 2002 International Symposium on Low-Power Electronics and Design, pp. 251-254, 2002.

[11]. M. Powell, S. Yang, B. Falsafi, K. Roy, and T. Vijaykumar, "Gated-V_{DD}: A circuit technique to reduce leakage in deep-submicron cache memories", Proceedings IEEE/ACM International Symposium on Low Power Electronics and Design, 2000, pp. 90–95.

[12]. S. Yang, M. Powell, B. Falsafi, K. Roy, and T. Vijaykumar, "An integrated circuit/architecture approach to reducing leakage in deep-submicron high-performance I-caches", in Proc. IEEE/ACM International Symposium on High-Performance Computer Architecture, 2001, pp. 147–157.

[13]. S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V power supply high-speed digital circuit technology with Multi-threshold-voltage CMOS," IEEE Journal Solid-State Circuits, vol. 30, pp. 847–854, Aug. 1995.

[14]. K. Nii, H. Makino, Y. Tujihashi, C. Morishima, Y. Hayakawa, H. Nunogami, T. Arakawa, and H. Hamano, "A low power SRAM using Auto-Backgate-Controlled MT-CMOS", in Proceedings IEEE/ACM International Symposium on Low Power Electronic Devices, 1998, pp. 293–298.

[15]. H. Makino et al., "An Auto-Backgate-Controlled MTCMOS Circuit", submitted to Symposium on VLSI Circuits, June 1998.

[16]. N. S. Kim, K. Flautner, D. Blaauw and T. Mudge, "Circuit and Micro-architectural techniques for reducing cache leakage power", IEEE Transaction on VLSI systems Vol. 12, No. 2, pp. 167-184, Feb. 2004.

[17]. K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy caches: Simple techniques for reducing leakage power", in Proc. IEEE/ACM International Symposium on Computer Architecture, 2002, pp. 148–157.

[18]. X. Chen and H. Bajwa, "Energy-efficient dual-port cache architecture with improved performances," IEE Journal of Electronic Letters, Vol. 43, No. 1, pp. 12-14, Jan. 2007.

[19]. H. Tran, "Demonstration of 5T SRAM and 6T Dual-Port RAM Cell Arrays," Symposium on VLSI Circuits, pp. 68-69, Jun. 1996.

[20]. S. Kim, N. Vijaykrishnan, M. Kandemir and M. J. Irwin, "Optimizing leakage energy consumption in cache bitlines" Journal of Design Automation for Embedded Systems, Vol. 9, No 1, pp. 5-18(14), Mar. 2004.

[21]. A. Karandikar and K. K. Parhi, "Low power SRAM design using hierarchical divided bitline approach", in Proceedings International Conference on Computer Design: VLSI in computers and Processors, pp. 82-88, 1998.

[22]. B.D. Yong and L.-S. Kim, “A low power SRAM using hierarchical bitline and local sense amplifier”, in IEEE Journal of Solid State Circuits, Vol. 40, No. 6, pp. 1366-1376, Jun. 2005.

APPENDIX A SPICE PARAMETERS USED FOR SIMULATION

A.1 Normal Spice Parameters at 180nm

.model NMOS NMOS

+Level = 49

+Lint = 4.e-08 Tox = 4.e-09

+Vth0 = 0.3999 Rdsw = 250

+lmin=1.8e-7 lmax=1.8e-7 wmin=1.8e-7 wmax=1.0e-4 Tref=27.0 version =3.1

+Xj= 6.0000000E-08 Nch= 5.9500000E+17

+lln= 1.0000000 lwn= 1.0000000 wln= 0.00

+wwn= 0.00 ll= 0.00

+lw= 0.00 lwl= 0.00 wint= 0.00

+wl= 0.00 ww= 0.00 wwl= 0.00

+Mobmod= 1 binunit= 2 xl= 0

+xw= 0 binflag= 0

+Dwg= 0.00 Dwb= 0.00

+K1= 0.5613000 K2= 1.0000000E-02

+K3= 0.00 Dvt0= 8.0000000 Dvt1= 0.7500000

+Dvt2= 8.0000000E-03 Dvt0w= 0.00 Dvt1w= 0.00

+Dvt2w= 0.00 Nlx= 1.6500000E-07 W0= 0.00

+K3b= 0.00 Ngate= 5.0000000E+20

+Vsat= 1.3800000E+05 Ua= -7.0000000E-10 Ub= 3.5000000E-18

+Uc= -5.2500000E-11 Prwb= 0.00

+Prwg= 0.00 Wr= 1.0000000 U0= 3.5000000E-02

+A0= 1.1000000 Keta= 4.0000000E-02 A1= 0.00

+A2= 1.0000000 Ags= -1.0000000E-02 B0= 0.00

+B1= 0.00

+Voff= -0.12350000 NFactor= 0.9000000 Cit= 0.00

+Cdsc= 0.00 Cdscb= 0.00 Cdscd= 0.00

+Eta0= 0.2200000 Etab= 0.00 Dsub= 0.8000000

+Pclm= 5.0000000E-02 Pdiblc1= 1.2000000E-02 Pdiblc2= 7.5000000E-03

+Pdiblcb= -1.3500000E-02 Drout= 1.7999999E-02 Pscbe1= 8.6600000E+08

+Pscbe2= 1.0000000E-20 Pvag= -0.2800000 Delta= 1.0000000E-02

+Alpha0= 0.00 Beta0= 30.0000000

+kt1= -0.3700000 kt2= -4.0000000E-02 At= 5.5000000E+04

+Ute= -1.4800000 Ua1= 9.5829000E-10 Ub1= -3.3473000E-19

+Uc1= 0.00 Kt1l= 4.0000000E-09 Prt= 0.00

+Cj= 0.00365 Mj= 0.54 Pb= 0.982

+Cjsw= 7.9E-10 Mjsw= 0.31 Php= 0.841

+Cta= 0 Ctp= 0 Pta= 0

+Ptp= 0 JS=1.50E-08 JSW=2.50E-13

+N=1.0 Xti=3.0 Cgdo=2.786E-10

+Cgso=2.786E-10 Cgbo=0.0E+00 Capmod= 2

+NQSMOD= 0 Elm= 5 Xpart= 1

+Cgsl= 1.6E-10 Cgdl= 1.6E-10 Ckappa= 2.886

+Cf= 1.069e-10 Clc= 0.0000001 Cle= 0.6

+Dlc= 4E-08 Dwc= 0 Vfbcv= -1

.model PMOS PMOS

+Level = 49

+Lint = 3.e-08 Tox = 4.2e-09

+Vth0 = -0.42 Rdsw = 450

+lmin=1.8e-7 lmax=1.8e-7 wmin=1.8e-7 wmax=1.0e-4 Tref=27.0 version =3.1

+Xj= 7.0000000E-08 Nch= 5.9200000E+17

+lln= 1.0000000 lwn= 1.0000000 wln= 0.00

+wwn= 0.00 ll= 0.00

+lw= 0.00 lwl= 0.00 wint= 0.00

+wl= 0.00 ww= 0.00 wwl= 0.00

+Mobmod= 1 binunit= 2 xl= 0.00

+xw= 0.00

+binflag= 0 Dwg= 0.00 Dwb= 0.00

+ACM= 0 ldif=0.00 hdif=0.00

+rsh= 0 rd= 0 rs= 0

+rsc= 0 rdc= 0

+K1= 0.5560000 K2= 0.00

+K3= 0.00 Dvt0= 11.2000000 Dvt1= 0.7200000

+Dvt2= -1.0000000E-02 Dvt0w= 0.00 Dvt1w= 0.00

+Dvt2w= 0.00 Nlx= 9.5000000E-08 W0= 0.00

+K3b= 0.00 Ngate= 5.0000000E+20

+Vsat= 1.0500000E+05 Ua= -1.2000000E-10 Ub= 1.0000000E-18

+Uc= -2.9999999E-11 Prwb= 0.00

+Prwg= 0.00 Wr= 1.0000000 U0= 8.0000000E-03

+A0= 2.1199999 Keta= 2.9999999E-02 A1= 0.00

+A2= 0.4000000 Ags= -0.1000000 B0= 0.00

+B1= 0.00

+Voff= -6.40000000E-02 NFactor= 1.4000000 Cit= 0.00

+Cdsc= 0.00 Cdscb= 0.00 Cdscd= 0.00

+Eta0= 8.5000000	Etab= 0.00	Dsub= 2.8000000
+Pclm= 2.0000000	Pdiblc1= 0.1200000	Pdiblc2= 8.0000000E-05
+Pdiblcb= 0.1450000	Drout= 5.0000000E-02	Pscbe1= 1.0000000E-20
+Pscbe2= 1.0000000E-20	Pvag= -6.0000000E-02	Delta= 1.0000000E-02
+Alpha0= 0.00	Beta0= 30.0000000	
+kt1= -0.3700000	kt2= -4.0000000E-02	At= 5.5000000E+04
+Ute= -1.4800000	Ua1= 9.5829000E-10	Ub1= -3.3473000E-19
+Uc1= 0.00	Kt1l= 4.0000000E-09	Prt= 0.00
+Cj= 0.00138	Mj= 1.05	Pb= 1.24
+Cjsw= 1.44E-09	Mjsw= 0.43	Php= 0.841
+Cta= 0.00093	Ctp= 0	Pta= 0.00153
+Ptp= 0	JS=1.50E-08	JSW=2.50E-13
+N=1.0	Xti=3.0	Cgdo=2.786E-10
+Cgso=2.786E-10	Cgbo=0.0E+00	Capmod= 2
+NQSMOD= 0	Elm= 5	Xpart= 1
+Cgsl= 1.6E-10	Cgdl= 1.6E-10	Ckappa= 2.886
+Cf= 1.058e-10	Clc= 0.0000001	Cle= 0.6
+Dlc= 3E-08	Dwc= 0	Vfbcv= -1

A.2 Spice Parameters at 180nm for Low Leakage

.model NMOSLL NMOS

+Level = 49					
+Lint = 4.e-08	Tox = 4.e-09				
+Vth0 = 0.5499	Rdsw = 250				
+lmin=1.8e-7	lmax=1.8e-7	wmin=1.8e-7	wmax=1.0e-4	Tref=27.0	version =3.1
+Xj= 6.0000000E-08	Nch= 5.9500000E+17				
+lln= 1.0000000	lwn= 1.0000000	wln= 0.00			

+wwn= 0.00	ll= 0.00	
+lw= 0.00	lwl= 0.00	wint= 0.00
+wl= 0.00	ww= 0.00	wwl= 0.00
+Mobmod= 1	binunit= 2	xl= 0
+xw= 0	binflag= 0	
+Dwg= 0.00	Dwb= 0.00	
+K1= 0.5613000	K2= 1.0000000E-02	
+K3= 0.00	Dvt0= 8.0000000	Dvt1= 0.7500000
+Dvt2= 8.0000000E-03	Dvt0w= 0.00	Dvt1w= 0.00
+Dvt2w= 0.00	Nlx= 1.6500000E-07	W0= 0.00
+K3b= 0.00	Ngate= 5.0000000E+20	
+Vsat= 1.3800000E+05	Ua= -7.0000000E-10	Ub= 3.5000000E-18
+Uc= -5.2500000E-11	Prwb= 0.00	
+Prwg= 0.00	Wr= 1.0000000	U0= 3.5000000E-02
+A0= 1.1000000	Keta= 4.0000000E-02	A1= 0.00
+A2= 1.0000000	Ags= -1.0000000E-02	B0= 0.00
+B1= 0.00		
+Voff= -0.12350000	NFactor= 0.9000000	Cit= 0.00
+Cdsc= 0.00	Cdscb= 0.00	Cdscd= 0.00
+Eta0= 0.2200000	Etab= 0.00	Dsub= 0.8000000
+Pclm= 5.0000000E-02	Pdiblc1= 1.2000000E-02	Pdiblc2= 7.5000000E-03
+Pdiblcb= -1.3500000E-02	Drout= 1.7999999E-02	Pscbe1= 8.6600000E+08
+Pscbe2= 1.0000000E-20	Pvag= -0.2800000	Delta= 1.0000000E-02
+Alpha0= 0.00	Beta0= 30.0000000	
+kt1= -0.3700000	kt2= -4.0000000E-02	At= 5.5000000E+04
+Ute= -1.4800000	Ua1= 9.5829000E-10	Ub1= -3.3473000E-19

+Uc1= 0.00 Kt1l= 4.0000000E-09 Prt= 0.00

+Cj= 0.00365 Mj= 0.54 Pb= 0.982

+Cjsw= 7.9E-10 Mjsw= 0.31 Php= 0.841

+Cta= 0 Ctp= 0 Pta= 0

+Ptp= 0 JS=1.50E-08 JSW=2.50E-13

+N=1.0 Xti=3.0 Cgdo=2.786E-10

+Cgso=2.786E-10 Cgbo=0.0E+00 Capmod= 2

+NQSMOD= 0 Elm= 5 Xpart= 1

+Cgsl= 1.6E-10 Cgdl= 1.6E-10 Ckappa= 2.886

+Cf= 1.069e-10 Clc= 0.0000001 Cle= 0.6

+Dlc= 4E-08 Dwc= 0 Vfbcv= -1

.model PMOSLL PMOS

+Level = 49

+Lint = 3.e-08 Tox = 4.2e-09

+Vth0 = -0.57 Rdsw = 450

+lmin=1.8e-7 lmax=1.8e-7 wmin=1.8e-7 wmax=1.0e-4 Tref=27.0 version =3.1

+Xj= 7.0000000E-08 Nch= 5.9200000E+17

+lln= 1.0000000 lwn= 1.0000000 wln= 0.00

+wwn= 0.00 ll= 0.00

+lw= 0.00 lwl= 0.00 wint= 0.00

+wl= 0.00 ww= 0.00 wwl= 0.00

+Mobmod= 1 binunit= 2 xl= 0.00

+xw= 0.00

+binflag= 0 Dwg= 0.00 Dwb= 0.00

+ACM= 0 ldif=0.00 hdif=0.00

+rsh= 0 rd= 0 rs= 0

+rsc= 0	rdc= 0	
+K1= 0.5560000	K2= 0.00	
+K3= 0.00	Dvt0= 11.2000000	Dvt1= 0.7200000
+Dvt2= -1.0000000E-02	Dvt0w= 0.00	Dvt1w= 0.00
+Dvt2w= 0.00	Nlx= 9.5000000E-08	W0= 0.00
+K3b= 0.00	Ngate= 5.0000000E+20	
+Vsat= 1.0500000E+05	Ua= -1.2000000E-10	Ub= 1.0000000E-18
+Uc= -2.9999999E-11	Prwb= 0.00	
+Prwg= 0.00	Wr= 1.0000000	U0= 8.0000000E-03
+A0= 2.1199999	Keta= 2.9999999E-02	A1= 0.00
+A2= 0.4000000	Ags= -0.1000000	B0= 0.00
+B1= 0.00		
+Voff= -6.40000000E-02	NFactor= 1.4000000	Cit= 0.00
+Cdsc= 0.00	Cdscb= 0.00	Cdscd= 0.00
+Eta0= 8.5000000	Etab= 0.00	Dsub= 2.8000000
+Pclm= 2.0000000	Pdiblc1= 0.1200000	Pdiblc2= 8.0000000E-05
+Pdiblcb= 0.1450000	Drout= 5.0000000E-02	Pscbe1= 1.0000000E-20
+Pscbe2= 1.0000000E-20	Pvag= -6.0000000E-02	Delta= 1.0000000E-02
+Alpha0= 0.00	Beta0= 30.0000000	
+kt1= -0.3700000	kt2= -4.0000000E-02	At= 5.5000000E+04
+Ute= -1.4800000	Ua1= 9.5829000E-10	Ub1= -3.3473000E-19
+Uc1= 0.00	Kt1l= 4.0000000E-09	Prt= 0.00
+Cj= 0.00138	Mj= 1.05	Pb= 1.24
+Cjsw= 1.44E-09	Mjsw= 0.43	Php= 0.841
+Cta= 0.00093	Ctp= 0	Pta= 0.00153
+Ptp= 0	JS=1.50E-08	JSW=2.50E-13

+N=1.0	Xti=3.0	Cgdo=2.786E-10
+Cgso=2.786E-10	Cgbo=0.0E+00	Capmod= 2
+NQSMOD= 0	Elm= 5	Xpart= 1
+Cgsl= 1.6E-10	Cgdl= 1.6E-10	Ckappa= 2.886
+Cf= 1.058e-10	Clc= 0.0000001	Cle= 0.6
+Dlc= 3E-08	Dwc= 0	Vfbcv= -1

APPENDIX B DRAM CELL ARCHITECTURES

B.1 3T-DRAM

Cell Structure

The three-transistor dynamic RAM (3T DRAM) is fundamentally different from both the 6T and the resistive load SRAMs. In the SRAM cells the data is stored in a latch which holds the data statically, whereas in the 3T DRAM the data is held dynamically on a capacitor. This means that, if left unused, the cell will eventually lose its information since the charge stored on the capacitor disappears through leakage. To solve this problem occasional refresh is needed. That is, each cell has to be read and then written back periodically.

In figure B.1 the schematic for a 3T DRAM cell is shown. The capacitor, CS, can either consist of the internal gate and source capacitances alone or a separate capacitor can be added. The latter is to ensure a higher capacitance, which increases stability. Due to the cells small area and relative simplicity, 3T DRAM is still used in many application specific integrated circuits.

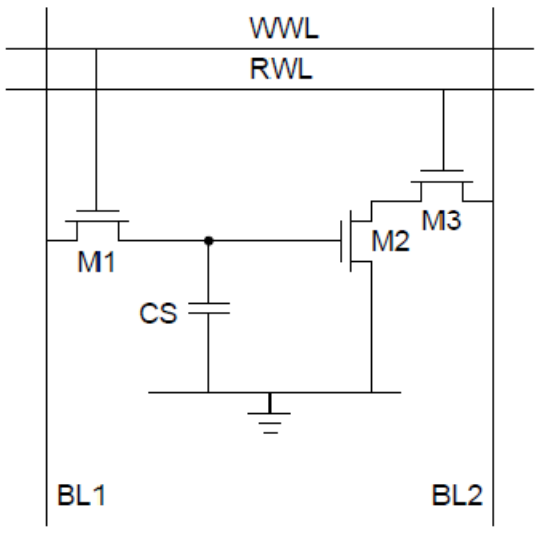

Figure B.1: Three-Transistor (3T) DRAM cell.

Read Operation

As opposed to the 6T SRAM, the 3T DRAM has a single ended read operation. This means that only one bitline is used for detecting the stored value (in this case BL2). The read operation is started by the precharge of BL2 (normally to V_{CC}). After that the read wordline (RWL) is asserted which results in M3 turning on. Now, depending on the value stored on CS, transistor M2 will either draw current from BL2 through M3 (when a ’1’ is stored), or it will be turned off (’0’ stored). Note that if a ’1’ is stored, the voltage on BL2 is lowered, and if a ’0’ is stored the bitline is left high! Consequently 3T DRAM cell is an inverting cell. Another thing that distinguishes the 3T DRAM from most other DRAMs is the fact that the read operation is non-destructive, i.e. the stored value is not affected by a read operation.

Write Operation

For write operation there is a separate bitline (BL1) and wordline (WWL). Initially BL1 is either held high (writing ’1’) or low (writing ’0’). WWL is then asserted and transistor M1 is turned on. The potential on the bitline is thereby transferred to CS before the lowering of WWL completes the write cycle. Note that, while a ’0’ can be transferred well by the NMOS transistor (M1), a ’1’ cannot. A threshold voltage is lost over the NMOS transistor and the resulting potential across CS is reduced to V_{WWL}-V_{TN}. This will reduce the current flowing through M2 during a read operation and thereby degrade performance. To overcome this problem many designs use a technique called bootstrapping to increase the voltage on the write wordline to V_{CC}+V_{TN}. This will ensure that V_{CC} will be stored when writing a ’1’.

Another thing worth noting is that there are no constraints on the transistor ratios for the 3T DRAM, as opposed to the 6T SRAM. The sizing is instead solely based on area, performance and stability considerations.

B.2 1T-DRAM

Cell Structure

In terms of cell area, the one-transistor (1T) DRAM is by far the smallest of the memories discussed here. It consists of one storage element, capacitor CS, and one pass-transistor, M1 (figure B.2).

As for the 3T case, the 1T DRAM is dynamic. It holds the stored value on a capacitor and therefore occasional refresh is needed, the same way as for the 3T in section B.1. To achieve a satisfying stability CS has to be fairly large. If the capacitor is made planar using metal layers, much of the area gain is lost. Therefore a specialized process with trenched capacitors is mainly used. While this makes the cell extremely small, it adds a large extra cost and is therefore usually not used in embedded cache. Planar capacitors made up from MOS devices can instead be used, which gives fairly large capacitance for a small area. These have, however, not yet been proved to be viable for high-yield, high-volume microprocessors.

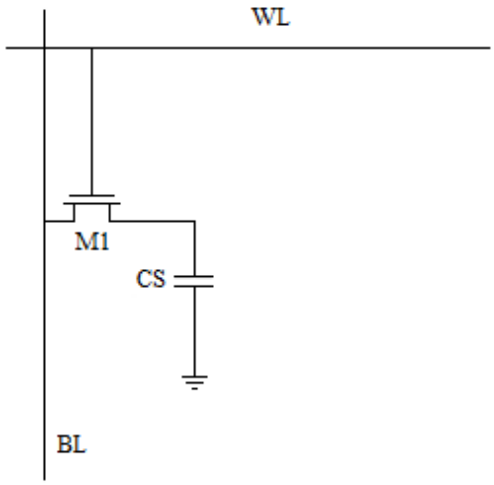

Figure B.2: One-Transistor (1T) DRAM cell.

Another term that sometimes is used for a variation of the 1T DRAM is 1T SRAM. This is a bit misleading since it is dynamic, not static. The reason why it is called SRAM is that the refresh is transparent, meaning it is hidden and in most cases will not be affecting the access time at all. This can be achieved by making separate memory banks, where all the banks not currently being accessed instead go through the refresh cycle.

Read Operation

The read operation of the 1T DRAM is very simple from the cell point of view. The bitline, BL, is precharged to some value, V_{REF}, usually around $V_{CC}/2$. Then the wordline, WL, is asserted and a charge transfer between CS and BL takes place. If a ’0’ was stored in the cell the bitline voltage will decrease, and if a ’1’ was stored it will increase. The

capacitance of the bitline is much larger than that of the cell so the voltage change on the bitline will be small.

In the 3T DRAM and the 6T SRAM one of the bitlines is continuously pulled down which means that a longer wait before trying to detect the value, results in a larger difference on the bitline. For the 1T this is not the case. Once the charges have been equalized nothing more will happen on the bitline. This means that the change must be detected and amplified using a sense amplifier, where as in the 3T and 6T, the sense amplifier only is used to speed up the read-out.

One obvious problem with the read operation in the 1T cell is that when the charge transfer occurs, the value stored in the cell is destroyed. This is called destructive read, and complicates the reading scheme further. It is now vital that a read always is followed by a write-back procedure. This can be done by having the output from the sense amplifier imposed onto the bitline, so that when the amplifier switches to full swing the BL is pulled up or down and the cell is written.

This type of operation, read followed by write-back, is exactly what should be done during the refresh. For the 1T DRAM the refresh therefore constitutes of reading all cells.

Write Operation

Writing in the 1T DRAM cell is a very simple process. The value to be written is held on the bitline, BL, and the wordline, WL, is raised. The cell storage capacitance, CS, is thereby either charged or discharged depending on the data value. When the value has been transferred the wordline is lowered again and the value is locked on the capacitor. As with the 3T DRAM, this cell will not store a ’1’ very well since one threshold voltage is lost over the pass transistor, M1. To overcome this problem, the same technique of bootstrapping is used for the 1T DRAM.