Design and Evaluation of High Density 5T SRAM Cache for Advanced Microprocessors

Master's thesis
performed in **Electronic Devices**

by
Ingvar Carlson

Reg nr: LiTH-ISY-EX-3481-2004

23rd March 2004

Design and Evaluation of High Density 5T SRAM Cache for Advanced Microprocessors

Master's thesis

performed in **Electronic Devices**,
Dept. of Electrical Engineering
at **Linköpings universitet**

by **Ingvar Carlson**

Reg nr: LiTH-ISY-EX-3481-2004

Supervisor: **Professor Atila Alvandpour**
Linköpings universitet

Examiner: **Professor Atila Alvandpour**
Linköpings universitet

Linköping, 23rd March 2004

Avdelning, Institution Division, Department	**Datum** Date
Electronic Devices, Dept. of Electrical Engineering 581 83 Linköping	23rd March 2004

Språk Language	**Rapporttyp** Report category	**ISBN**
☐ Svenska/Swedish	☐ Licentiatavhandling	—
☒ Engelska/English	☒ Examensarbete	**ISRN**
	☐ C-uppsats	LITH-ISY-EX-3481-2004
	☐ D-uppsats	**Serietitel och serienummer** Title of series, numbering
☐ ______	☐ Övrig rapport	**ISSN** —
	☐ ______	

URL för elektronisk version

http://www.ep.liu.se/exjobb/isy/2004/3481/

Titel Konstruktion och utvärdering av kompakta 5T SRAM cache för avancerade mikroprocessorer

Title Design and Evaluation of High Density 5T SRAM Cache for Advanced Microprocessors

Författare Author

Ingvar Carlson

Sammanfattning
Abstract

This thesis presents a five-transistor SRAM intended for the advanced microprocessor cache market. The goal is to reduce the area of the cache memory array while maintaining competitive performance. Various existing technologies are briefly discussed with their strengths and weaknesses. The design metrics for the five-transistor cell are discussed in detail and performance and stability are evaluated. Finally a comparison is done between a 128Kb memory of an existing six-transistor technology and the proposed technology. The comparisons include area, performance and stability of the memories. It is shown that the area of the memory array can be reduced by 23% while maintaining comparable performance. The new cell also has 43% lower total leakage current. As a trade-off for these advantages some of the stability margin is lost but the cell is still stable in all process corners. The performance and stability has been validated through post-layout simulations using Cadence Spectre.

Nyckelord
Keywords

SRAM, high-density, cache, five-transistor, 5T, memory, microprocessor

Abstract

This thesis presents a five-transistor SRAM intended for the advanced microprocessor cache market. The goal is to reduce the area of the cache memory array while maintaining competitive performance. Various existing technologies are briefly discussed with their strengths and weaknesses. The design metrics for the five-transistor cell are discussed in detail and performance and stability are evaluated. Finally a comparison is done between a 128Kb memory of an existing six-transistor technology and the proposed technology. The comparisons include area, performance and stability of the memories. It is shown that the area of the memory array can be reduced by 23% while maintaining comparable performance. The new cell also has 43% lower total leakage current. As a trade-off for these advantages some of the stability margin is lost but the cell is still stable in all process corners. The performance and stability has been validated through post-layout simulations using Cadence Spectre.

Keywords: SRAM, high-density, cache, five-transistor, 5T, memory, microprocessor

Acknowledgment

Firstly I would like to thank my supervisor, Professor Atila Alvandpour, for the opportunity to write my masters thesis in the Electronic Devices division at Linköping University. He has been an invaluable help with ideas and discussions throughout my entire time working on this thesis. Secondly, I would like to thank the entire Electronic Devices division, and specifically Stefan Andersson for massive support with tools and ideas. He has been someone to bounce ideas with at all times no matter what the issue.

Furthermore I would like to thank my room mate Aditya Medury for the layout of the address decoder and a lot of fun discussions regarding everything from work to life in general.

Last but not least I would like to thank Sreedhar Natarajan from ATMOS Corporation, Dr. Dinesh Somasekhar, Dr. Ram Krishnamurthy, Dr. Vivek De, Director Shekhar Borkar from Intel Corporation and Professor Christer Svensson from Linköping University for invaluable input regarding my thesis and submission of a conference paper to the 2004 Symposium on VLSI Circuits. For support and input regarding the English language, I would like to thank my wife, Katherine.

Contents

Chapter 1

Introduction

1.1 Background

The need for area effective cache is ever increasing. The first Intel® Pentium® M processors, with around 77 million transistors, have 1MB of L2 cache. That constitutes about 2/3 of the total transistor count, or 50 million transistors.

Many different types of memories exist today and new technologies and circuits are developed every year. But for cache applications in advanced microprocessors, few have been proven worthy. The DRAM with its small, one-transistor cell, would be a strong candidate. It has however not been used due to the specialized process steps needed, and associated prominent increase in manufacturing cost. Planar DRAMs using a standard CMOS process, on the other hand, has not been proven to be viable for high-yield, high-volume production. This, together with the high performance demands of microprocessors, has resulted in the conventional six-transistor (6T) SRAMs being the main choice for today's cache applications.

The purpose of this thesis is to show a new approach to a fully static SRAM that can be used to replace the standard 6T. To do that, it has to have smaller area while still maintaining the high performance of the 6T.

1.2 Outline of Thesis

- **Chapter 1 - Introduction:** An introduction and overview of the entire thesis. Also includes a listing of some of the terms used in the rest of the thesis.

- **Chapter 2 - Memory Fundamentals:** Some basic memory concepts and ideas.
- **Chapter 3 - Existing Technologies:** An overview of some of the existing technologies today. These include, six-transistor SRAM, resistive load SRAM, three-transistor DRAM and one-transistor DRAM. The cell structures are discussed as well as the read and write operations. Emphasis is put on the six-transistor SRAM.
- **Chapter 4 - Proposed Technology: 5T SRAM:** The new five-transistor cell is proposed. Thorough investigations are made regarding read and write operation, stability and sizing. This chapter is the main focus of the thesis.
- **Chapter 5 - 5T-6T 128Kb Comparison:** Comparisons of area, performance, stability and leakage between two memory arrays, one using the proposed five-transistor cell and the other using a conventional six-transistor cell. These post-layout simulation comparisons determine the viability of the proposed technology.
- **Chapter 6 - Conclusion:** A summary of the results as well as a reflection of the work carried out.
- **Chapter 7 - Future Work:** A discussion of future work that could be done to build on to, and validate results from this thesis.
- **References:** A listing of the references used in the thesis.
- **Appendix A - Current Through a Transistor:** Analytical derivation of currents through the transistors during a read operation. The results are used in determining proper sizing of the memory cell.
- **Appendix B - Mismatch Simulations:** Simulations made to determine the impact of transistor mismatch within the five-transistor cell. Important results are derived regarding the sizing of the cell.

1.3 Terminology

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

λ **-** Channel length modulation, parameter describing the change of transistor gate length for different V_{DS}.

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. 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).

g$_m$ **-** Transconductance, the current amplification with respect to V_{GS}. It is defined as $g_m = \frac{\partial I_D}{\partial V_{GS}}$.

I$_D$ **-** Drain current through a transistor.

MOS(FET) - 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 a n-type inversion layer as channel for conducting current.

nT - n-transistor, memory cell made up of n 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.

Process Corner - Parameters supplied by the manufacturer delimiting the process variations for a specific transistor type. For instance the Slow N corner specifies the parameters for the NMOS transistors that result in the slowest transistors that can occur during fabrication (within a given probability).

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.

SNM - Static Noise Margin, measure of a cells stability in regard to static mismatch.

SRAM - Static RAM, a RAM where the value is stored statically in a latch.

V_{CC} - Reference for the high potential power supply (1.8V in this thesis).

V_{DS} - Drain-Source potential, the difference between the potential at the drain and the source of a transistor.

V_{GS} - Gate-Source potential, the difference between the potential at the gate and the source of a transistor.

V_{PC} - Bitline precharge voltage for the 5T SRAM.

V_{SB} - Source-Bulk potential, the difference between the potential at the source and the bulk of a transistor.

V_T - Threshold voltage, the gate-source potential required for the transistor to start conducting.

Wafer - A round disc of semiconductor material most commonly Silicon. Many chips are simultaneously fabricated on a wafer during the fabrication process.

WL - Wordline, the wire connected to the gate of the pass-transistors of the memory cells.

Yield - Measure that describes the percentage of working chips on a wafer.

Chapter 2

Memory Fundamentals

2.1 What is a Memory

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 I will only discuss the electronic memory, and more specifically, random access memories (RAM).

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, not necessarily corresponding one to one.

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. This is in contrast to read only memories (ROM) which only support the reading of ’0’ or ’1’.

2.2 Cache

2.2.1 What is Cache

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 frequencies 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.

Following the previous discussion of different abstraction layers, most advanced microprocessors today have several *physical* layers of memory, making up the cache memory, embedded on the chip. The closest one is called layer 1 (L1) cache and usually has direct contact with the Central Processing Unit's (CPU) pipeline. This gives an extremely short access time, and therefore provides the highest performance. This cache is usually run at the same clock frequency as the CPU. The strict requirements of this L1 cache, and the fact that it has to access the CPU pipeline means that it is very expensive in terms of area. The relative large area per cell of this high performance memory makes it very difficult to place large L1 caches close to the CPU. As an example the Intel® Pentium® 4 processors have only 8KB of L1 cache.

To make up for the relatively small L1 cache, a larger level 2 (L2) cache is often put on-chip. This cache is placed slightly further away from the CPU, and is connected to it 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. As a comparison most Intel® Pentium® 4 processors have 512KB of L2 cache.

Sometimes a third cache on chip is used. It is referred to as level 3 (L3) cache and is, following the same convention as above, the furthest from the CPU. It is in most cases quite comparable in performance to L2 cache or slightly slower.

2.2.2 Requirements

There are a 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 can not 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. This means clock frequencies above 2GHz.

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 with the strive for simple operation in already complex circuits have made the 6T SRAM the choice of the day for advanced microprocessor caches. The 6T SRAM is further described, along with some other existing memory techniques, in chapter 3.

Chapter 3

Existing Technologies

3.1 6T SRAM

3.1.1 Cell Structure

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 3.1). 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.

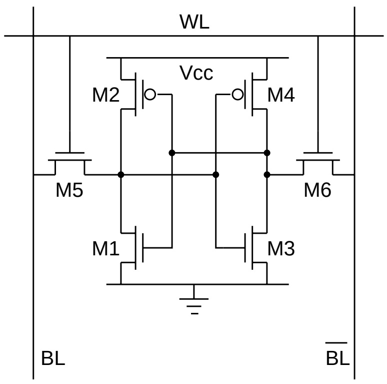

Figure 3.1: Six-transistor SRAM cell.

3.1.2 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 3.1) 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 3.1) 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 3.2, for reading a stored ’0’.

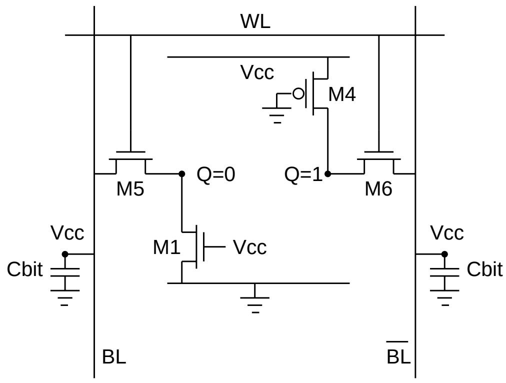

Figure 3.2: Six-transistor SRAM cell at the onset of read operation (reading ’0’).

In the figure the capacitors *Cbit* represents 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 3.3) is turned on.

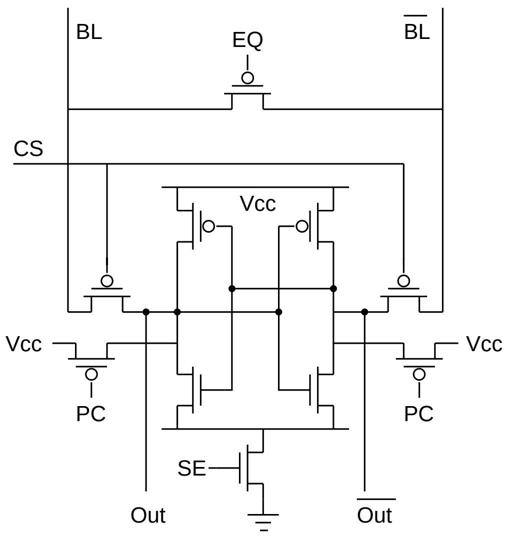

Figure 3.3: Sense Amplifier for a six-transistor SRAM.

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 3.3, 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 3.3 is then connected to a buffer to restore the flank of the signal and to facilitate driving of larger loads. Also the $\overline{Out}$ 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 node results in a '1' output. Therefore the performance is mainly dependent on the constellation *M1-M5* (see figure 3.1) or *M3-M6* and their ability to draw current from the bitline.

3.1.3 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 $\overline{BL}$ to be lowered. Why is this? Let's take a closer look at the cell when writing a '1' (figure 3.4).

As in the previous example of a read, the cell has a '0' stored and

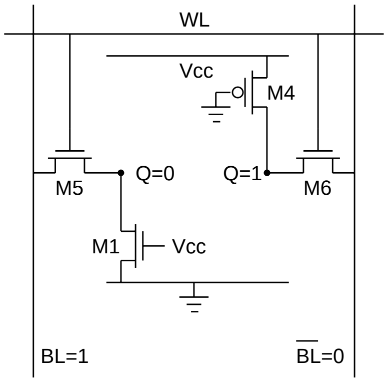

Figure 3.4: Six-transistor SRAM cell at the onset of write operation (writing ’0’→’1’).

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. If we look at the left side of the memory cell (*M1*-*M5*) it is virtually identical to the read operation (figure 3.2). 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, we are forced to write a ’0’ 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). 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 3.5 shows the sizing for the 6T SRAM cell used for comparisons in this thesis.

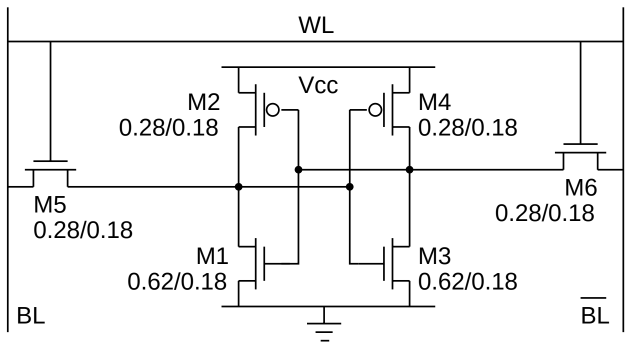

Figure 3.5: Standard six-transistor SRAM cell with sizes.

3.2 Resistive Load SRAM

3.2.1 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 3.6). 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/□ is used [6]. 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) [7]. 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.

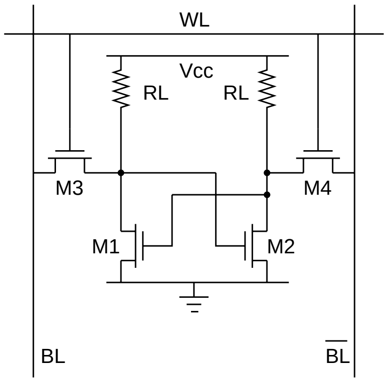

Figure 3.6: Resistive Load SRAM cell.

3.2.2 Read Operation

The read operation is identical to the 6T SRAM case (see section 3.1.2).

3.2.3 Write Operation

The write operation is identical to the 6T SRAM case (see section 3.1.3).

3.3 3T-DRAM

3.3.1 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, typically every 1 to 4ms [6].

In figure 3.7 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 [6].

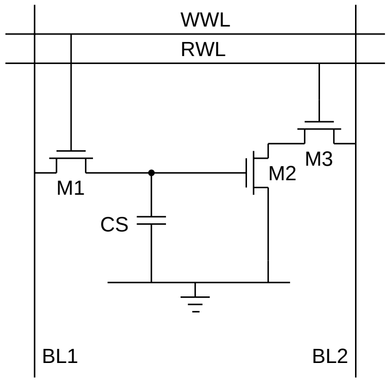

Figure 3.7: Three-transistor DRAM cell.

3.3.2 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.

3.3.3 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 transfered to *CS* before the lowering of *WWL* completes the write cycle. Note that, while a ’0’ can be transfered well by the

NMOS transistor (*M1*), a '1' can not. 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 (see sections 3.1.2 and 3.1.3). The sizing is instead solely based on area, performance and stability considerations.

3.4 1T-DRAM

3.4.1 Cell Structure

In terms of cell area, the one-transistor (1T) DRAM is by far the smallest of the memories discussed here. The cell structure is extremely simple and the most straight forward of all the memories in this thesis. It consists of one storage element, capacitor *CS*, and one pass-transistor, *M1* (figure 3.8).

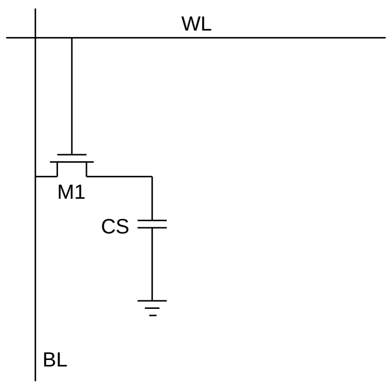

Figure 3.8: One-transistor DRAM cell.

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 3.3.1. To achieve a satisfying stability *CS* has to be fairly large (at least 30fF) [6]. If the capacitor is made planar using metal layers, much of the area gain is lost. Therefore a specialized process with *trenched* capacitors are 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.

Another term that sometimes is used for a variation of the 1T DRAM is 1T SRAM [3]. 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.

3.4.2 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.

3.4.3 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.

Chapter 4

Proposed Technology: 5T SRAM

4.1 Background

As shown in chapter 3 many different memory technologies have been proposed over the years. With their different properties, different structures have been proven useful for different applications. In this thesis the focus has been put on high-performance cache, a market that has been totally dominated by the standard 6T SRAM (see section 3.1).

The key to the microprocessor cache market is high performance, high stability and small area. With the excellent performance and stability of the 6T SRAM, it has been dominating even though the area has been comparatively very large. In this thesis it is shown that with some modifications of the cell, a reduction of area while still maintaining comparable performance can be achieved.

4.2 Cell Structure

In a normal 6T cell both storage nodes are accessed through NMOS pass-transistors (see section 3.1.1). 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 4.5.3) 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 proposed, resulting five-transistor (5T) SRAM cell.

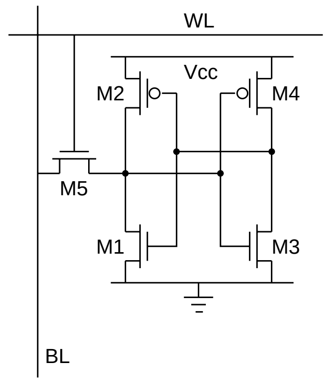

Figure 4.1: Five-transistor SRAM cell.

With one less bitline the 5T cell also shares a sense amplifier between two cells (see section 4.3). This further reduces the area giving the 5T memory block an even greater advantage over the 6T SRAM. One version of a 5T SRAM was presented in 1996 by Hiep Tran [9]. That cell differ fundamentally from the cell proposed 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 approach in this thesis is instead to mimic the behavior of the well proven 6T cell as closely as possible, while still reducing the area.

4.3 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 (see sections 4.5.1 and 4.5.3). 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}.

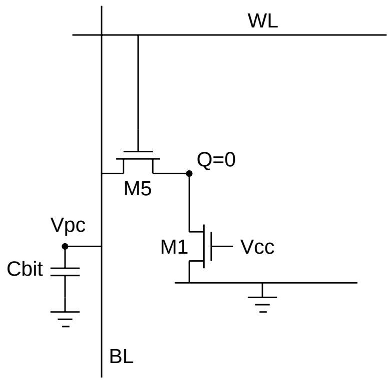

Figure 4.2: Five-transistor SRAM cell at the onset of read operation (reading ’0’).

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. However, an internal scheme shown in Fig. 4.3 is also proposed.

During precharge, one bitline is grounded and the other bitline, connected to the same sense amplifier, is charged to a value V_{CC}-V_{TN} obtained from a diode-coupled NMOS. These bitlines are then equalized during the address decoding through a NMOS transistor and the correct precharge voltage is obtained. If the diode-coupled NMOS is not desired, a similar scheme using six bitlines where two of them are charged to V_{CC} and four are grounded can be used. In this thesis the external V_{PC} approach has been used. This is to facilitate easy evaluation of the chip using different precharge voltages. The above mentioned scheme has, however, also been validated through post-layout simulations.

The next phase of the read operation scheme is to pull the wordline

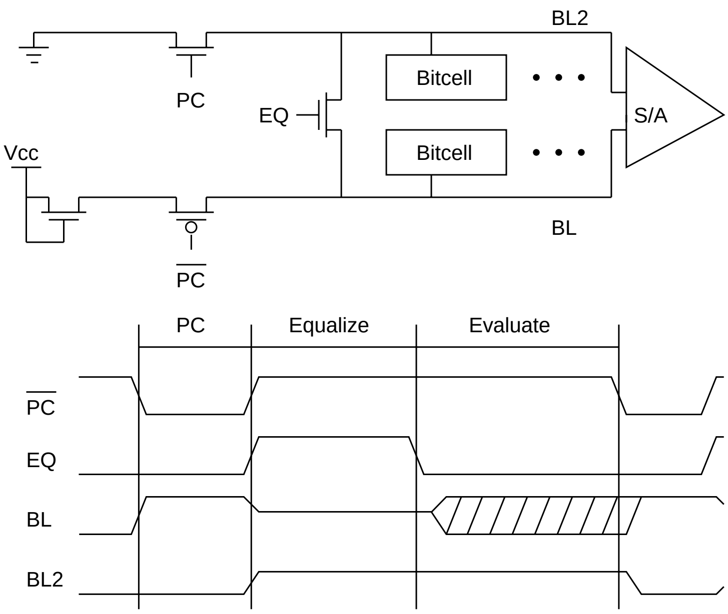

Figure 4.3: Internal bitline precharge scheme utilizing charge splitting.

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 4.4 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 in section 4.6.

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 (see section 3.1.2), 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

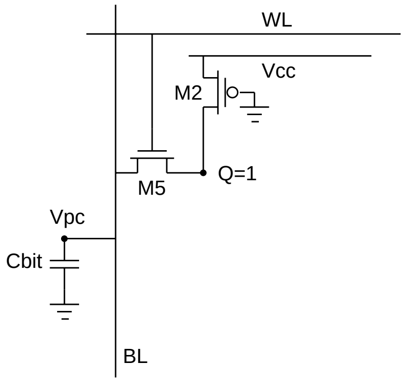

Figure 4.4: Five-transistor SRAM cell at the onset of read operation (reading ’1’).

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 (see figure 3.3) 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.

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 [4] since the 1T cell also only have one bitline (see section 3.4).

One implication of this scheme is that the output from the sense amplifier will either be OUT or $\overline{OUT}$ depending on which section is accessed. This is because one is connected to the $\overline{BL}$ side of the sense amplifier and a higher value on that line will result in a low output (see figure 3.3). 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.

4.4 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 (see section 3.1.3), the 5T cell has to be sized differently. An in-depth discussion regarding the sizing and write-ability of the 5T cell is given in section 4.5.2.

4.5 Operation Stability

4.5.1 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 respectively if the channel length modulation is ignored (see appendix A).

$$I_D = k_n' \frac{W}{L} \left((V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2} \right) \tag{4.1}$$

$$I_D = k_n' \frac{W}{L} \left((V_{GS} - V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2} \right) \tag{4.2}$$

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.

$$k_n' \frac{W_1}{L_1}\left((V_{cc}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}\right) > k_n' \frac{W_5}{L_5}\left((V_M-V_T)(V_{PC}-V_M)-\frac{(V_{PC}-V_M)^2}{2}\right) \tag{4.3}$$

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 (*M1* and *M5* are still operating in the same regions, see appendix A), and all other values are substituted as V_{CC}=1.8V, V_M=0.9V, andV_T=0.4V. The relation can then be simplified as in equation 4.4.

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

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 centered 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 4.4). 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 can not be canceled out in the equations, but for all other purposes the discussions remain the same.

4.5.2 Write Stability

To make it possible with one sided write operations, the cell must be sized accordingly. In section 4.5.1 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.

Starting out from the sizing of a standard 6T SRAM cell in section 3.1.3, the important differences in sizing considerations can be highlighted. The 5T cell with the same sizing (only without *M6* and $\overline{BL}$) will be stable for read '0' according to section 4.5.1 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 4.5 shows the internal cell nodes during the operation for different values of the width. The width was varied in steps of 0.1μm.

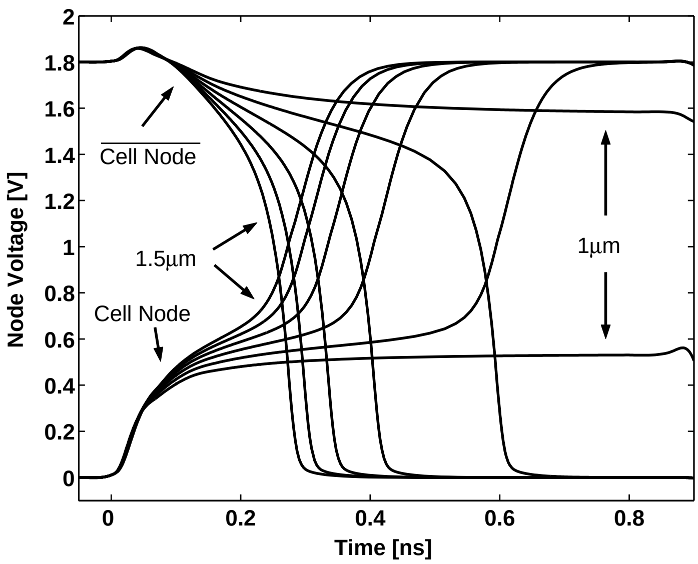

Figure 4.5: Internal cell nodes for five-transistor SRAM cell writing '1' with width of *M5* varied between 1μm and 1.5μm (typical process corner).

From the figure, it is evident that the width of *M5* must exceed 1μm or the cell will not be written at all. Given that the pass-transistors of the 6T SRAM are only 0.28μm wide and that the main objective of the 5T cell is to reduce the area, a different approach is needed. The cross-coupled inverters must be resized and the switching point adjusted. As a first step the two NMOS transistors of the inverters (*M1* and *M3*) are made minimum size (0.28/0.18μm) and the pass-transistor (*M5*) is made 0.72/0.18μm. Note that cross-coupled inverters are still symmetrical at this point. Further sizing issues are discussed in sections 4.5.4, 4.5.6, 4.6 and appendix B.

4.5.3 Precharge Voltage Window

In the above discussions of read stability (section 4.5.1) 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 4.5.2 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 effect 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 4.6 and figure 4.7 respectively.

For the read '0' case the internal nodes flip if *BL* is precharged to 1.75V but not if it is precharged to 1.65V. Also, for the read '1' case, the nodes flip if *BL* is precharged to 0.45V but not is it is precharged to 0.5V. The V_{PC} *window* for this configuration is therefore 0.5-1.65V. This, however is for typical process parameters. To get a better understanding of the stability, in regard to precharge voltage, the same procedure as above was conducted for four more process corners. The magnitudes of the resulting voltage windows are compiled in figure 4.8.

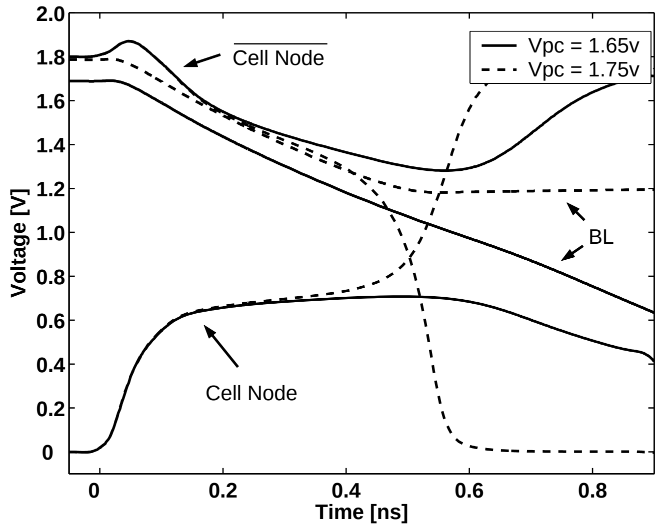

Figure 4.6: Internal cell nodes and bitline for a five-transistor SRAM cell reading '0' with V_{PC} at 1.65V and 1.75V respectively (typical process corner).

4.5.4 Sensitivity to Process Variations and Mismatch

As seen in figure 4.8, process variations can have a large impact on stability. These process variations however only take into account differences between types of transistors, not differences between transistors of the same type. For example, the corner N Fast/P Slow means that all NMOS transistors are in the fast process corner and all PMOS transistors are in the slow corner. Today another type of variation is becoming increasingly important. It is the so called *mismatch* between transistors. This means that two transistors of the same type can have different properties. For instance the lengths can vary slightly, or the size of the drain area. To simulate the effects of mismatch a *Monte Carlo* simulation is usually done.

Monte Carlo simulation is a way of using given process variations and applying statistical spread. The same simulation is performed very many times and for every time, slightly different parameters are used. The statistical spread between transistors is supplied by the manufacturers, and the values used for each simulation are determined according to Gaussian distribution.

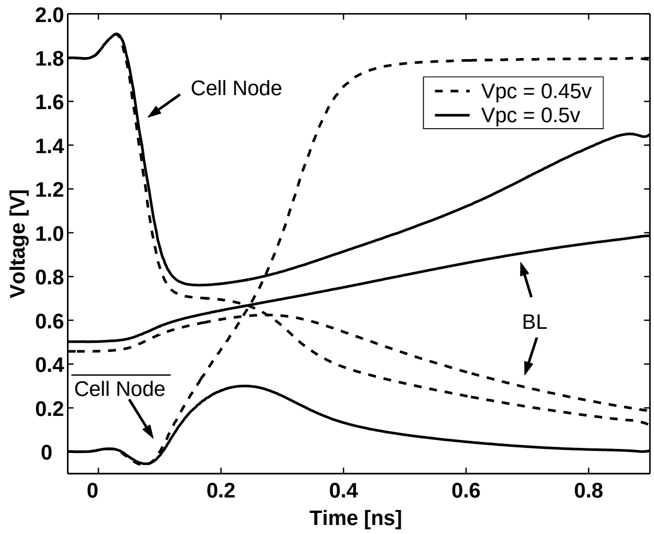

Figure 4.7: Internal cell nodes and bitline for a five-transistor SRAM cell reading '1' with V_{PC} at 0.5V and 0.45V respectively (typical process corner).

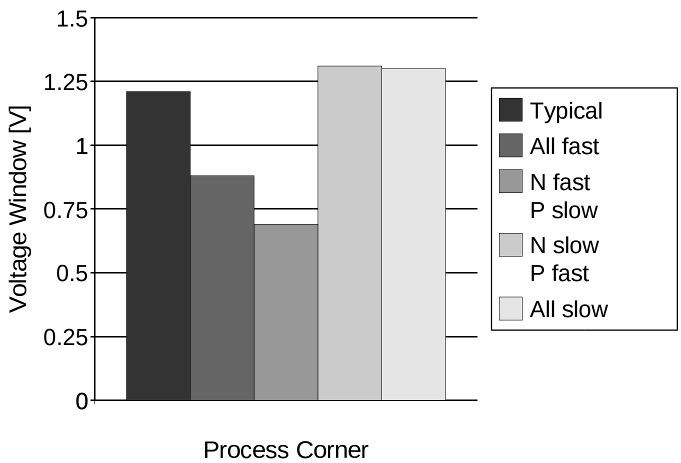

Figure 4.8: Magnitudes of precharge voltage window for different process corners.

The CMOS process used for this thesis however, did not come with Monte Carlo parameters. Instead variations between transistors of the same type were done by manually changing the V_T of the transistors and thereby making them faster or slower. Appendix B shows the different simulations and the impact given on read and write performance and the precharge voltage window. These results also provide very valuable information when sizing the cell. Changing V_T of a transistor is, in effect, making it stronger or weaker. Since this can also be done by changing sizes of the transistors, the results regarding performance and stability will also be a good indication of how a change in size will affect the cell.

4.5.5 Static Noise Margin

The static noise margin (*SNM*) of a 6T SRAM cell is defined as the minimum DC noise voltage necessary at both of the two cell storage nodes to flip the state of the cell [7]. Static noise is due to static errors in the chip that, for one reason or another, gives the cell an voltage offset or other mismatch. To measure the sensitivity to this type of mismatch a quasi-static transient simulation, where the noise sources are introduced and slowly increased (compared to the cells switching speed), is often used [5]. The most critical point in a SRAM cell is during a read [1], and SNM is therefore typically measured while holding the bitlines at the precharge value and the wordline asserted.

The same method has been used for evaluation of SNM in the 5T SRAM cell. Figure 4.9 shows the schematic used for the SNM simulations. Two noise sources have been introduced in the cross-coupling of the inverters. These are put so that their respective noise is working the same direction. If one source is increasing the gate voltage of one inverter, that inverter is giving a lower output. That output is therefore further lowered by the second noise source before being coupled back to the other inverter.

The noise sources in the simulation were swept from 0 to 500mV in 2μs which can be considered slow, since the switching speed of the cell is far less than 1ns. Figure 4.10 shows the graph of such a measurement. The two nodes are initially stable but as the noise increases the margin between the nodes diminish. At some point (about 120mV in this case) the storage nodes flip and the cell settles in this new stable state.

One difference between the 6T and 5T cells in these simulations

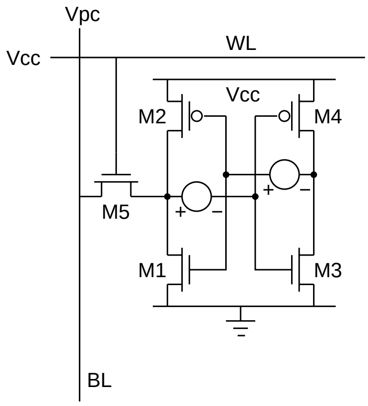

Figure 4.9: Static noise margin simulation setup for five-transistor SRAM cell

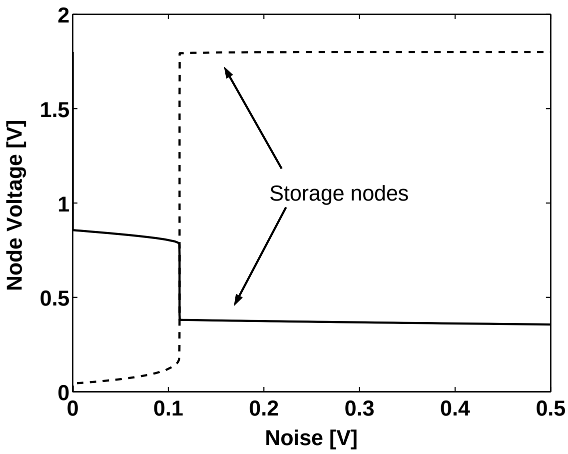

Figure 4.10: Measurement of static noise margin for a five-transistor cell.

should be pointed out. The 6T cell has two connections to bitlines, both of which are open. One bitline will try to pull up the node that has a ’0’ stored, thereby lowering the noise margin. The other one will try to hold the node with the ’1’ high, thereby keeping the cell from flipping and increasing the noise margin. For the 5T cell there is only one bitline. This bitline will pull the node towards the switching point regardless of what value is stored, and since the cell is not entirely symmetric, both values should be simulated for worst case analysis. In both cases the static noise margin will be lowered by the connection to the bitline. Since there is no other bitline to help the cell retain its state, the static noise margin can be expected to be substantially lower for the 5T cell compared to the 6T cell. On the other hand the introduction of two noise sources for the 5T cell might be a little pessimistic. Perhaps the node not connected to a bitline would experience less static noise statistically. This however, is beyond the scope of this thesis and the same SNM simulation setup has therefore been used for both the 6T and the 5T cell. The results from these simulations can be seen in section 5.3.

4.5.6 Voltage Scaling

Another important issue in today's IC design is *voltage scaling*. How well will a particular design work if the supply voltage is lowered? This will in some way be an indication of how well the design can be used in a smaller technology since the supply voltage generally decreases with every new generation of processes. But it is also important from another perspective. Power dissipation is a large problem in today's microprocessors. It limits the battery life of mobile applications and it also makes the chips so hot that malfunction can occur. To deal with these issues the supply voltage is often lowered.

To evaluate the sensitivity to voltage scaling, performance simulations were conducted, using the 5T cell sizing described in section 4.5.2. Also, to get some idea of how well the 5T cell handles the scaling compared to other cache memories, simulations were conducted for the 6T cell with sizing according to section 3.1.3. The results were then normalized for each type of operation and memory individually, and summarized in figure 4.11. The write time was measured from the assertion of the wordline to the flipping of the internal nodes. For the read time, only the time from the assertion of the wordline to a 100mV bitline development was measured. This allows for the sense amplifier to be detached from the measurements and the impact on the cell can be better evaluated. On the other hand, the relative difference between the measurements will be much larger since the total evaluation time is not taken into account.

It can be seen from the figure that the 5T SRAM compares well with the 6T SRAM for the write and read of '0'. Both memories only suffer a 10% penalty when the supply voltage is decreased by about 10%. For the read '1' case the 5T memory is slightly worse but still

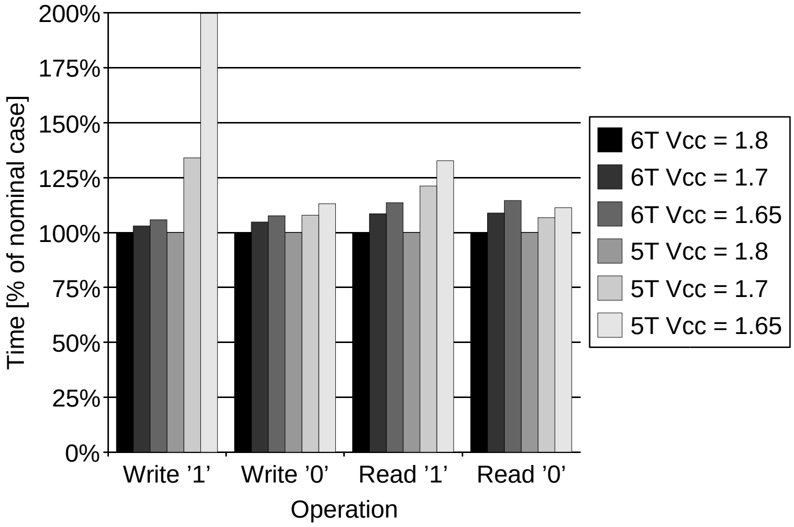

Figure 4.11: Impact of voltage scaling on internal operation times. Precharge voltages at 750mV for five-transistor memory and V_{CC} for six-transistor memory.

within reasonable limits. However, for the write ’1’ case, the 5T cell has a very dramatic increase of operation time. When the supply voltage has been decreased by 10% the write time doubles compared to the nominal case. This indicates that the 5T cell has been sized too close to the limit of when writing a ’1’ is possible. To avoid this kind of problem, which might also occur for normal supply voltage but in a worst case process corner, resizing is needed.

4.6 Sizing and Layout

The stability of the 5T SRAM cell has been thoroughly investigated in section 4.5. Several different metrics have been used to show what is important in the sizing of the 5T cell. One thing has, however, not been discussed in detail so far. That is the inherent asymmetry of the cell.

A 6T SRAM cell is absolutely symmetric in the layout. Since the cell is read differentially (see section 3.1.2) this is very important. Both storage nodes must have the same capacitances and the same sizing of the transistors connected to them. For the 5T SRAM cell this is not important. The cell does not use the inverse storage node as a reference, but instead another 5T cell (see section 4.3). Therefore it is only important that the different cells are the same, not the storage nodes within each cell. Also, the cell is by nature asymmetric since only one of the storage nodes is connected to a pass-transistor. This increases the capacitance in that node making it different from the inverse node.

From appendix B it is apparent what impact changes to individual transistors have on the performance and stability. This, together with the results from the preliminary sizing used in section 4.5, gives enough information for proper sizing of the 5T cell. Now one more factor should be considered. Will the acquired sizing be suitable for layout and therefore result in a smaller cell?

To answer this question a suitable transistor sizing must first be determined. First, to give the cell a balanced read (where read of ’0’ and ’1’ take about the same time) the inverter *M1-M2* must be resized. Since a NMOS transistor is stronger (due to better mobility) the PMOS must be made larger so that they can draw the same current from the bitline.

Next, the write time and stability issues must be addressed. As seen in section 4.5.6 the write time gets very large with a little lower supply voltage. It is actually the writing of ’1’ that is the problem not the writing in general. The cell must therefore be made a little more unstable in regard to writing ’1’. There are two ways to do this. Either the pass-transistor *M5* is made larger which will make the cell more unstable for writing in general, or the switching point of the cell could be lowered, making it easier to write a ’1’. Since the writing of ’0’ is already fast, the second option is chosen.

To lower the switching point of the cell, the switching point of the inverter *M3-M4* should be lowered. It is the gate of this inverter that is connected to the bitline through *M5*, so when that node reaches the switching point the cell changes state. Now two things can be done to lower the switching point of the inverter. Either transistor *M3* is made stronger or transistor *M4* is made weaker. None of these transistors affect the reading performance, so the the only aspects of this sizing are area, stability and write performance.

To get a sufficiently good write performance in all corners, the switching point has to be lowered considerably. If this was done only by increasing the width of *M3* it would have to be very large, which would make it very difficult to get a small cell after layout. Therefore transistor *M4* should also be made weaker. However, this transistor is already minimum width, which prevents it from being scaled down. In equations 4.1 and 4.2 it can be seen that the current through a transistor is dependent on $\frac{W}{L}$. So instead of making it thinner it can be made longer.

Now all of these parameters can be adjusted together until a satisfactory result is achieved. For instance *M5* can be made a little smaller to reduce the area and increase the stability, while the write-ability can be assured through keeping the switching point low. Figure 4.12 shows the resulting sizes.

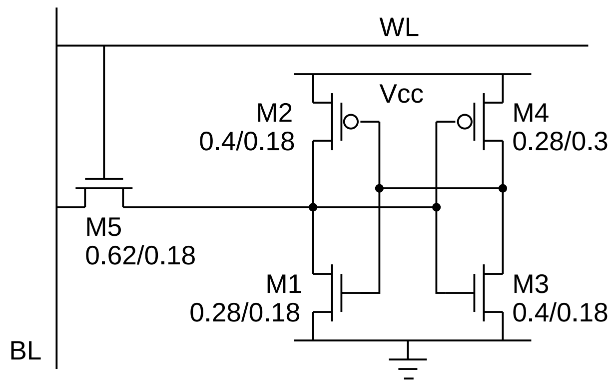

Figure 4.12: Five-transistor SRAM cell with final sizes.

When making a layout of a 6T cell it is, as mentioned above, important to match the cell so it is as symmetric as possible (figure 4.13 shows a typical layout).

This is not the case with the proposed 5T cell. To make an area efficient layout it is important to minimize the empty space. If a similar layout to the 6T layout is attempted for the 5T, there will be a hole where the second pass-transistor was (at the bottom of the cell layout). To be able to reduce the size when the cells are connected together that hole has to be filled. One thought would be to turn one cell 180° and then fit it so that the new cell's pass-transistor fits where the old cell's second pass-transistor was. The problem with that is that the new pass-transistor is much wider and would therefore not fit in that configuration without making the whole cell wider. Also, the area gain would be equivalent to one half of the total length of the pass-

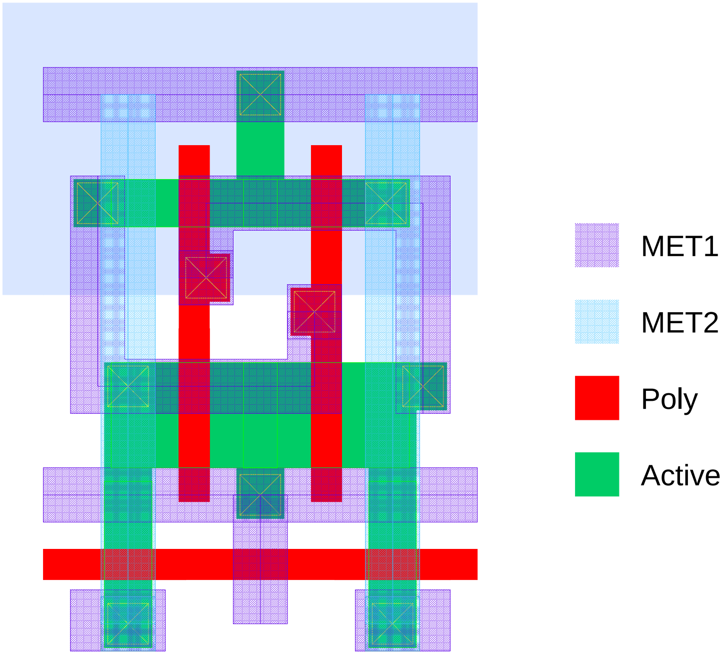

Figure 4.13: Layout of standard six-transistor SRAM cell.

transistors (including the active area) at best.

Instead the remaining pass-transistor *M5* could be moved to the side of the cell, which results in a area gain from the height of the cell equivalent to the total length of the pass-transistors. That, on the other hand, would mean that the cell would be three transistors wide instead of two. So was anything gained? The answer is yes! The 6T cell shared the top and bottom contacts with neighboring cells, therefore limiting their area contribution to half. The new 5T cell also shares top and bottom contacts with the neighboring cells, but it now also shares the left hand contact. Figure 4.14 shows the layout of the 5T cell with the pass transistor on the side and sizing according to figure 4.12.

The 6T cell was limited by metal1 on the sides (the distance to the neighboring cells was determined by the minimum space allowed between metal1). This means that the three NMOS transistors of the 5T cell (with the left contact shared) actually can fit within the same width as the metal1 took up for the 6T cell. This reduces the total cell area by 21.2%. Further comparisons, of both the cell area and the

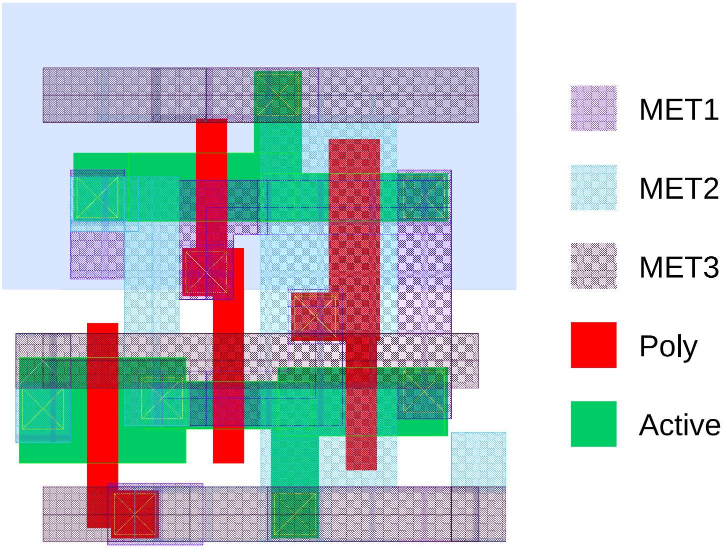

Figure 4.14: Layout of proposed five-transistor SRAM cell.

memory array area, can be found in section 5.1.

Chapter 5

5T-6T 128Kb Comparison

In this chapter all the properties of the 5T and 6T SRAM cell are compared. To be able to make this comparison two memory blocks were designed. One with a standard 6T cell and one with the proposed 5T cell. In order to show all the different aspects of the memories relatively large blocks were made. Each block is 128Kb and consists of a total of 1024 bitlines (and inverse bitlines for the 6T case), where each bitline is connected to 128 cells. These bitlines are then merged into a 64 bit wide word through the sense amplifiers. Figure 5.1 shows the organization for the 5T SRAM.

The difference for the 6T case is that the column selectors are PMOS and that *BL2* is replaced by $\overline{BL}$. Consequently, double the amount of sense amplifiers are needed, resulting in 128 outputs from the whole memory, which have to be merged into 64.

For the performance simulations a 600mV bitline precharge voltage (V_{PC}) was used for the 5T array and 1.8V (V_{CC}) for the 6T array.

5.1 Area

The purpose of this thesis it to show that a SRAM cache memory can be designed using a 5T cell with smaller area than the 6T, while not sacrificing too much performance. Assuming that the 5T is functional, the area comparison is therefore the most important one.

Figures 4.13 and 4.14 show the layouts of the two cells to scale. The most apparent difference is that the 5T is asymmetric and almost square, whereas the 6T is symmetric and oblong. With the shared left-side contact the 5T has the same width as the 6T, but the height of

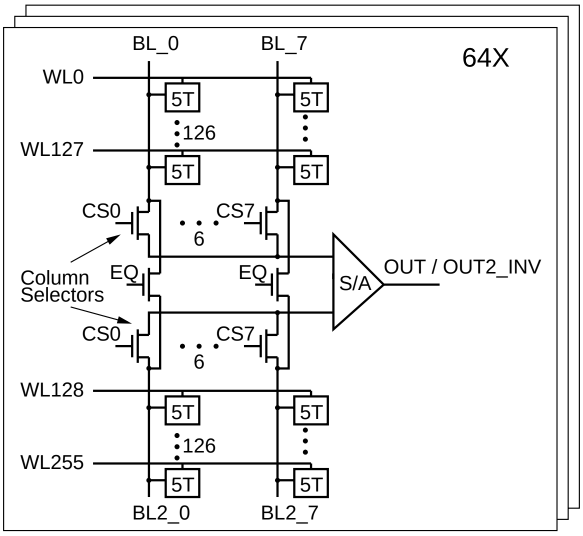

Figure 5.1: Proposed organization of 128Kb five-transistor SRAM.

the cell has been reduced by one total pass-transistor length. With the design rules available for the standard CMOS process that was used, the 6T cell is 3.12μm high and 2.56μm wide, which results in a total cell area of 7.99μm^2. The 5T cell has a height of 2.46μm and a width of 2.56μm using the same design rules. This results in a total cell area of 6.30μm^2 which is 21.2% less than for the 6T.

When making memory however, the usual CMOS design rules are not used. Instead special memory cells are developed, following much tighter rules and then evaluated for yield in a foundry. These rules are not available for fabrication (or development) for customers, which is why they have not been used here. In studying different, tighter (SRAM-like) rules, a 15-21% area improvement has still been confirmed for the cell.

The area comparison did not stop at the cell however. As discussed in the beginning of this chapter the 6T cell demands double the amount of sense amplifiers. This will increase the area slightly and further strengthen the case for the 5T cell.

Another thing that can be seen from the layouts of the cells is that the 6T cell has a poly wordline, whereas the 5T cell has a wide metal2 wordline. While not significant to the cell, it has a major impact when many cells are connected together. Figure 5.2 shows the difference of signal propagation in a poly wordline versus a thin metal2 wordline.

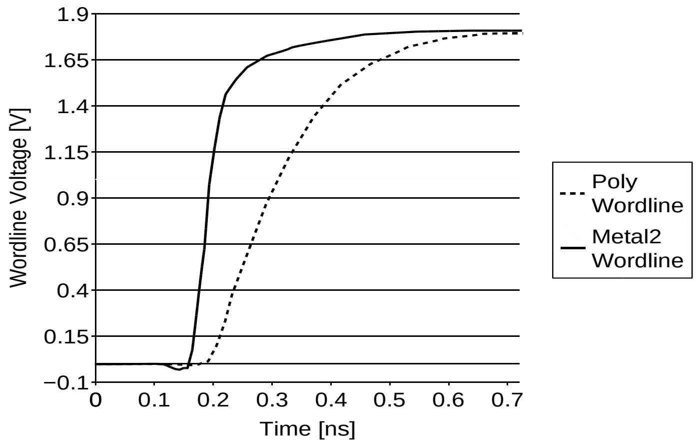

Figure 5.2: Wordline development for six-transistor memory with poly and metal wordline (64 cell/wordline).

With a difference of 100ps (more than 100%) until the midpoint (0.9V) is reached, it is clear that something other than poly is needed. The problem with the 6T cell layout is that there is no room for a poly/metal1 contact to connect the gates of the pass-transistors to a metal wordline. Instead of making every cell larger in an attempt to fit in the contact, a technique called *stitching* can be used. The main idea of stitching is to run a metal wordline on top of the poly wordline and tap down to it between cells every so often (see figure 5.3).

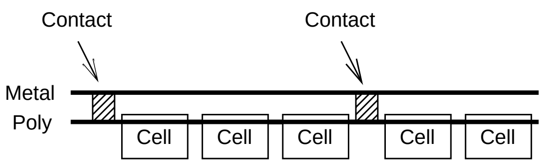

Figure 5.3: Cross-section showing stitching between metal and poly wordline every third cell.

This means that the cell can be kept small, but a little extra space is needed between cells every time a stitch occurs. In the 128Kb memory stitching has been used every 16th cell.

All this together results in a 128Kb memory array that is 1671μm wide and 686μm high for the 6T SRAM. For the 5T SRAM the corresponding measurements are 1395μm wide and 630μm high. Figure 5.4 shows a micrograph of the 5T array and the 6T array next to each other. The total area for the 6T memory array is 1.15mm^2 and for the 5T it is 0.88mm^2, a reduction of 23%.

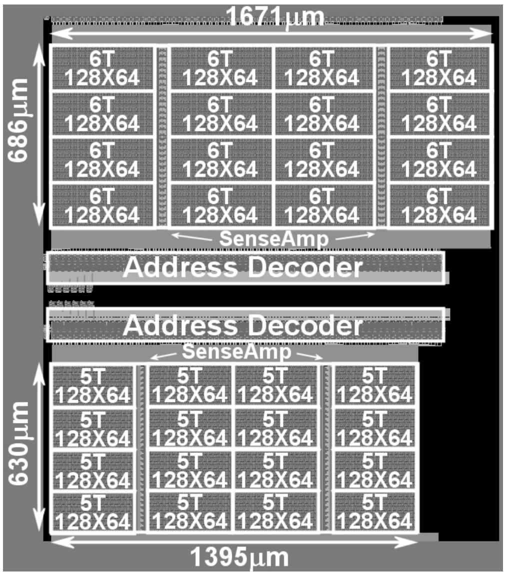

Figure 5.4: Chip micrograph of 5T and 6T memory chip.

5.2 Performance

In the performance comparisons, the extra merging of the 6T cell has been left out. Although this would make the 6T performance a little slower, the 5T cell also might need a converter after the sense amplifier. This converter should invert the output from the sense amplifier if a cell from the inverse side is read. Then again, if the cells on the inverse side instead were written with the inverse value, this value would be inverted back by the sense amplifier and therefore the converter would not be needed. Instead the write operation would be slightly slower. For a synchronous SRAM using the same cycle time for read and write this is not a problem. The read operation is generally slower and is the factor that determines the operation frequency, leaving a little idle time after a write operation.

To avoid this organizational discussion, the performance has strictly

Table 5.1: Read performance (WL high → OUT high) [ps]

Process Corner	6T	5T
Typical 110 °C	499	421
Fast N / Slow P, 3σ 110 °C	495	415
Slow N / Fast P, 3σ 110 °C	529	452
Fast, 3σ 110 °C	390	349
Slow, 3σ 110 °C	597	513

been measured from the midpoint of the wordline transition to the output of the sense amplifier buffer (read case) or flipping of the internal nodes (write case).

Table 5.1 shows the worst case read performance of the memory arrays, in different process corners.

The corners have been evaluated at deviations equivalent to 3σ Monte Carlo analysis. This means that the deviations exceed the statistical Gaussian spread of three standard deviations. In other words, less than 0.14% of transistors would exceed these values [2]. Tests have also been conducted showing a stable read with similar results when 4σ was used. Only 0.003% of the transistors would exceed this deviation [2].

From the table it is apparent that the 5T SRAM is 10-15% faster than the 6T SRAM in all corners. On closer inspection this appears to mainly be because of faster sense amplifier operation. This is due to the higher voltage difference between the bitline and the precharged latch nodes of the sense amplifier for the 5T memory. The latch nodes are therefore pulled down faster, which results in a faster switching of the sense amplifier.

For the write case the performance is compared in table 5.2.

Additional tests also verified a stable write operation when 4σ deviation was used. From table 5.2 it can be seen that the 6T array is quite a bit faster than the 5T array. Especially in the Slow N/Fast P corner. However, on comparison with table 5.1, it is apparent that the read time is larger than the write time in all corners. For a synchronous memory where the read and write cycle have the same timing, this means that the operation frequency would depend on the read time alone. Thus the 5T SRAM can operate at a slightly higher frequency than the 6T

Table 5.2: Write timing (WL high → internal switching) [ps]

Process Corner	6T	5T
Typical 110 °C	135	191
Fast N / Slow P, 3σ 110 °C	120	133
Slow N / Fast P, 3σ 110 °C	183	354
Fast, 3σ 110 °C	97	128
Slow, 3σ 110 °C	183	277

SRAM.

5.3 Stability

The stability of the 5T SRAM was evaluated in several different ways as described in section 4.5. For the comparison to the 6T SRAM, the SNM (section 4.5.5) was studied especially closely. Table 5.3 shows the results of that study.

Table 5.3: Static Noise Margin [mV/noise source]

Process Corner	6T	5T
Typical 110 °C	255	117
Fast N / Slow P, 3σ 110 °C	180	85
Slow N / Fast P, 3σ 110 °C	315	161
Fast, 3σ 110 °C	210	105
Slow, 3σ 110 °C	289	128

It is evident from the table that the 6T SRAM has a significantly higher static noise margin. As discussed in section 4.5.5 however, this comparison might be a little pessimistic for the 5T case. To answer the question of whether the 5T cell has a large enough SNM or not, the cell has to be evaluated for yield. This involves fabrication and evaluation of several chips and is discussed further in chapter 7.

5.4 Leakage

The leakage of a memory cell can be divided into two separate discussions: leakage to the bitline(s) and total leakage through the latch. The total leakage is interesting from a power dissipation point of view. The total power dissipation should be held down to stop the chip from overheating and increase battery life of mobile applications.

The bitline leakage is interesting from two different perspectives. One is that it contributes to the total power dissipation. The other is a performance issue. As described in sections 3.1.2 and 4.3 a stored value is read by amplifying a difference between one bitline (*BL*) and a reference ($\overline{BL}$ or *BL2*). To ensure a correct evaluation, the bitline and the reference must have a significant voltage difference. If we consider the case when all cells on a particular bitline (6T SRAM), except for the cell being read, has a '0' stored, an interesting situation occurs. After the precharge both bitlines are high (V_{CC}), but as soon as the precharge is lifted the sub-threshold leakage of the cells start affecting the bitline. If all the cells are storing a '0' all the leakage currents will discharge *BL* whereas the reference $\overline{BL}$ will remain high (a '1' is stored on the inverse node). When the wordline is asserted the cell to be read (storing a '1') starts to discharge $\overline{BL}$ and charge *BL*, while the other cells are still leaking. It now takes a bit of time before the voltages of *BL* and $\overline{BL}$ pass each other; time that has essentially been wasted since the same voltage difference as before must be developed for reliable read. This is a worst case scenario, but since there is no way of knowing what values are stored in the memory, the timing has to be designed for the worst case.

Another implication of this is that more cells on a bitline increases this problem. Therefore, if there is a lower leakage current from each cell, more cells can be put on the same bitline. This increases the area efficiency since less overhead per cell is needed. A lower leakage to the bitline can in other words mean smaller area and/or higher performance.

Table 5.4 shows the leakage comparison between the 5T and the 6T cells. It should be pointed out that the total leakage of a 6T cell will be equal no matter what value is stored since one of the storage nodes always store a '0' and the other a '1'.

From the table it can be seen that the 5T cell has a significantly lower leakage current, both to the bitline and as a total. The reason

Table 5.4: Leakage/Cell (standby, Typical 110 °C)

Setup	6T	5T	Diff.
To BL, low node	3.16nA	0.794nA	75%
To BL, high node	0	0.484pA	N/A
Whole Cell, ’1’ stored	7.08nA	3.36nA	53%
Whole Cell, ’0’ stored	7.08nA	4.78nA	32%

for this is the lower bitline precharge voltage. When a ’0’ is stored the V_{DS} of the pass-transistor is only one third in the 5T case compared to the 6T. This results in a leakage current reduction of 75%. For the case when a ’1’ is stored, there is instead a higher V_{DS}. The NMOS pass-transistor, however, transfers high voltages poorly. It can only pass a voltage equal to V_{CC}-V_T, and V_T is also increased in this case due to a higher source-bulk voltage (V_{SB}) [6]. The result is a very small bitline leakage current, less than 1/1000 of the current in the stored ’0’ case, and therefore practically negligible.

For the 128Kb memory arrays the total leakage currents, assuming equal amounts of ’0’s and ’1’s, are 0.93mA and 0.53mA respectively. This is a reduction of 43% for the 5T case compared to the 6T.

Chapter 6

Conclusion

In this thesis a new approach to a fully static five-transistor (5T) SRAM cell was proposed. It resulted in a smaller, yet still stable cell, and with comparable performance to the conventional 6T cell. It also has a considerably smaller leakage current.

A 128Kb memory array was implemented, using this cell, in a standard 1.8V, 0.18μm CMOS process. A separate 128Kb memory array was implemented using a standard 6T cell in the same technology. The results from the comparison between the two memory arrays are summarized in table 6.1

Table 6.1: 5T and 6T Comparison (Typical at 110°C)

Metrics	6T	5T
Read Time (WL high $\rightarrow$ OUT high)	499ps	421ps
Write Time (WL high $\rightarrow$ state flip)	135ps	191ps
Static Noise Margin	255mV	117mV
Bitline Leakage/Cell	3.16nA	0.794nA
Cell Area (Logic Design Rule)	7.99μm^2	6.30μm^2
Area (128Kb)	1.15mm^2	0.88mm^2

Further, more in-depth, studies were conducted for the 5T array detailing the memory's sensitivity to process variations. These are summarized in figures 6.1 and 6.2.

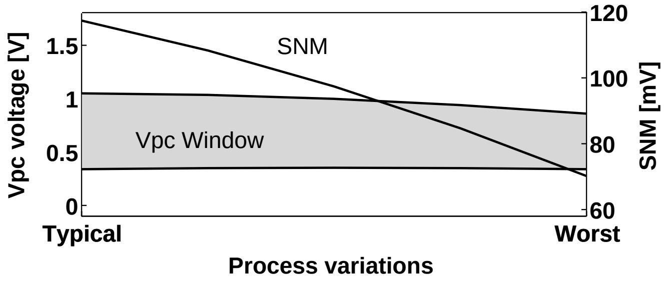

Figure 6.1: Bitline precharge voltage, V_{PC}, ensuring correct operation and static noise margin, *SNM*, vs. process variations at 110°C.

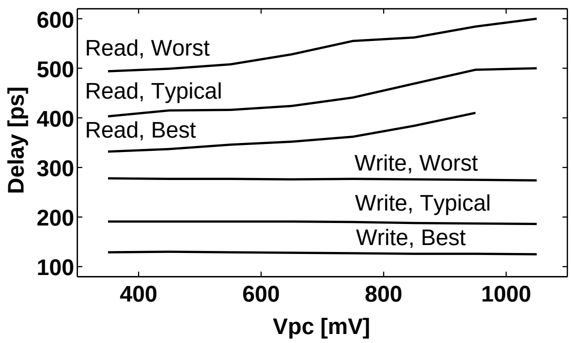

Figure 6.2: 5T read and write delay vs. bitline precharge level, V_{PC}, for different process corners at 110°C.

Chapter 7

Future Work

There are some obvious ways to continue the above described work. The first step is to fabricate a chip using the memory array layout created throughout the work with this thesis. This would be a way to validate all the simulation work, and show that the memory really works in its physical form. At present this work has already started and is expected to finish before the end of April 2004 when the chip is scheduled to go to fabrication.

Depending on the results from these physical measurements there is further interesting work in the future. For instance a porting to a 0.13μm technology could be done to evaluate the impact of scaling on the 5T SRAM cell. The alternative internal precharge scheme discussed in section 4.3 could be evaluated in a physical chip.

The layout of the cells used in this thesis have been made using regular *logic* design rules. When a SRAM memory is fabricated these rules result in much too large cells. Therefore special SRAM rules are used by the foundries when memories are fabricated. These rules are the result of manufacturing many chips and then settling for rules that give a high enough yield. One interesting continuation of this work would be a yield evaluation, using tighter rules, and compare the results to the regular 6T SRAM. This would in fact be one of the most important factors regarding the viability of the 5T SRAM in real microprocessors. Unfortunately these investigations require access to a foundry, and are completely out of the scope for a university, and even for most companies.

References

[1] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl. The impact of intrinsic device fluctuations on cmos sram cell stability. *IEEE Journal of Solid-State Circuits*, 36(4):658–665, April 2001.

[2] Gunnar Blom. *Sannolikhetsteori och statistikteori med tillämpningar*. Studentlitteratur, fourth edition, 1989.

[3] Wingyu Leung, Fu-Chieh Hsu, and M. E Jones. The ideal soc memory: 1t-sram™. In *Proceedings of the 13th Annual IEEE International ASIC/SOC Conference*, pages 32–36, September 2000.

[4] Jyi-Tsong Lin and Cheng-Chih Hsu. An initial overdriven sense amplifier for low voltage drams. *Devices, Circuits and Systems, 2000. Proceedings of the 2000 Third IEEE International Caracas Conference on*, pages C31/1–C31/4, March 2000.

[5] J. Lohstroh, E. Seevinck, and J. de Groot. Worst-case static noise margin criteria for logic circuits and their mathematical equivalence. *IEEE Journal of Solid-State Circuits*, SC-18(6):803–807, December 1983.

[6] Jan M. Rabaey, Anantha Chandrakasan, and Borivoje Nikolic. *Digital Integrated Circuits: A Design Perspective*. Prentice Hall series in electronics and VLSI. Prentice Hall, second edition, 2003.

[7] E. Seevinck, F. J. List, and J. Lohstroh. Static-noise margin analysis of mos sram cells. *IEEE Journal of Solid-State Circuits*, SC-22(5):748–754, October 1987.

[8] Ben G. Streetman and Sanjay Banerjee. *Solid State Electronic Devices*. Prentice Hall series in Solid State Physical Electronics. Prentice Hall, fifth edition, 2000.

[9] H. Tran. Demonstration of 5t sram and 6t dual-port ram cell arrays. *1996 Symposium on VLSI Circuits, Digest of Technical Papers.*, pages 68–69, June 1996.

Appendix A

Current Through a Transistor

The unified model for deep submicron MOS can be described as follows [6]:

$I_D \approx 0 \quad \text{for} \quad V_{GT} \leq 0 \quad (V_{GT} \geq 0 \text{ for PMOS})$

$I_D = k' \frac{W}{L} \left(V_{GT} V_{min} - \frac{V_{min}^2}{2} \right) (1 + \lambda V_{DS})$

with $\quad V_{min} = min(|V_{GT}|, |V_{DS}|, |V_{DSAT}|)$

and $\quad V_{GT} = V_{GS} - V_T$

Studying the 5T cell in different situations we can derive expressions for the current through transistors *M1*, *M2* and *M5* (see figure 4.1).

To determine in which region each transistor is operating, V_{min} has to be evaluated. V_{min}=V_{GT} means that the transistor is saturated, V_{min}=V_{DS} means that it is in the triode region and V_{min}=V_{DSAT} means that it is velocity saturated. The latter happens in short-channel devices where the electrical field in the channel is so high that an increase in V_{DS} will not increase the current through the transistor (if channel modulation is ignored). The current is limited by the velocity of the carriers, and the velocity is constant due to scattering effects (collisions suffered by the carriers) [6]. The value of V_{DS} where this occurs is called V_{DSAT}.

Before V_{min} can be evaluated, V_{DSAT} must be determined. Following the above discussion it is clear that it could be determined by studying the current through the transistor. However, normal saturation of a transistor behaves similarly. In both cases the current through the transistor remains constant even if V_{DS} is increased. One observation that can be made is that a velocity saturated transistor has a linear current dependence on V_{GS}, whereas a normally saturated transistor has a square dependence. Therefore the transconductance (g_m) can be used.

A transistor's transconductance is defined as $g_m = \frac{\partial I_D}{\partial V_{GS}}$ [8]. In examining the current equations it is clear that g_m is linear and increasing in the saturation region, but constant in the velocity saturation region. Therefore V_{DSAT} can be found by plotting g_m, for a transistor with gate and drain shorted (the transistor vill be saturated as soon as it is turned on). Figure A.1 shows g_m for a NMOS transistor, in a standard CMOS process. In the graph, lines have been added showing the approximate linearity and the off, saturation and velocity saturation regions respectively.

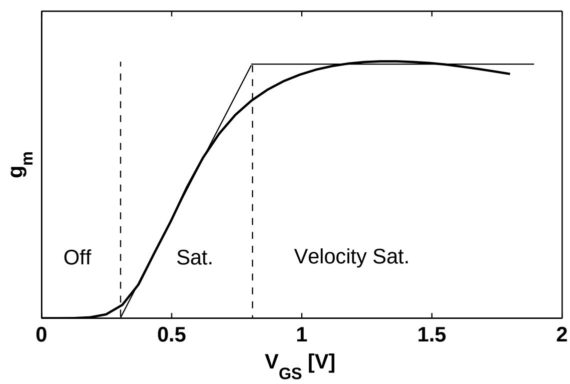

Figure A.1: Transconductance (g_m) for a NMOS transistor, in a standard CMOS process.

From the figure it can be seen that velocity saturation occurs for V_{GS_1}≈0.8V. Also, for this process V_T≈0.4V. It is furthermore stated in the above equations that the linear dependence of V_{GS} (and therefore constant g_m) occur when V_{DSAT}<V_{GS}-V_T. All this leads to the conclusion thatV_{DSAT}=V_{GS_1}-V_T≈0.4V. From this, V_{min} can be determined for different situations.

A.1 Read ’0’

When reading a ’0’ in the 5T SRAM the schematic can be simplified as in figure 4.2. To find out in which region the transistors *M1* and *M5* are operating, we study the situation around the switching point, i.e. when Q is equal to the switching voltage V_M. For transistor *M5* the gate-source voltage (V_{GS}) is V_{CC}-V_M and the drain-source voltage (V_{DS}) is V_{PC}-V_M. The value of V_{min} will therefore be the minimum of V_{CC}-V_T-V_M, V_{PC}-V_M and V_{DSAT}. With the values from figure A.1 this gives the values 1.4V – V_M, V_{PC}-V_M and 0.4V. This means that as long as the precharge voltage is lower than 1.4V and the difference between the precharge voltage and the switching point of the cell is less than 0.4V, *M5* will be in the triode region.

For transistor *M1*, the same discussion can be applied. In this case the gate-source voltage (V_{GS}) is V_{CC} and the drain-source voltage (V_{DS}) is V_M. Therefore V_{min} will be equal to V_{DSAT} (0.4V) and *M1* in the velocity saturated region as long as V_M is higher than 0.4V.

A.2 Read ’1’

When reading a ’1’, the simplified schematic is instead as in figure 4.4. Again we study the situation around the switching point V_M. For transistor *M5* the situation is identical to the situation in section A.1. For transistor *M2*, however, V_{min} has to be evaluated.

The gate-source voltage (V_{GS}) is in this case -V_{CC} and the drain-source voltage (V_{DS}) is V_M-V_{CC}. Also, V_T is approximately -0.45V and V_{DSAT} is -0.5V for PMOS in this process. All this together results in V_{min} evaluating to V_{DSAT} as long as V_M is below 1.3V, and *M2* is therefore in the velocity saturation region.

Appendix B

Mismatch Simulations

The impact of *mismatch* on the five-transistor SRAM cell was evaluated by manually changing the threshold voltage, V_T, of the transistors and thereby making them faster or slower. One at a time all the transistors were adjusted and the impact evaluated. Each transistor's threshold voltage was adjusted in the range +/- 40mV which is equivalent to about a 20% spread.

The simulations were conducted, on the schematic level, by adding an ideal voltage source in series with the gate of each transistor. In turn, these were then set at plus and minus 40 mV and the read and write performance was evaluated. In the following graphs, V_{Tn} is the threshold voltage for transistor *n* according to figure 4.1. The first graph (figure B.1) shows the performance for typical process parameters. To get a good measure of the read performance, and make it independent of sense amplifier, three different measurements have been made, measuring the time taken for a relative bitline development. The times are measured from the wordline going high until there has been a bitline development of 20mV, 50mV and 100mV respectively. The write time is measured as the time taken from a high on the wordline to the state of the cell is flipped. All values given are worst case out of read/write '0' and '1'.

The write time has been plotted in the whole range of the precharge voltage window, whereas the read times have not. The reason for this is that the pass-transistor of the 5T cell is a NMOS and can therefore not pull the bitline up more than V_{CC}-V_T. This means that a voltage development on the bitline cannot occur for too high precharge voltages. This, in effect puts another limitation on the precharge voltage

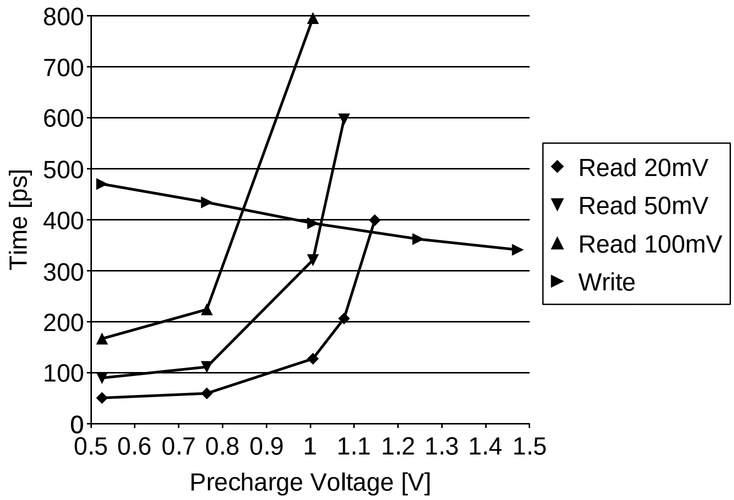

Figure B.1: Write performance and bitline development during read for typical process parameters

other than the one concerning cell stability.

Following are the graphs for the cases where V_T has been individually changed for the transistors. Figure B.2 and figure B.3 shows the graphs when V_{T1} has been changed +40mV and -40mV respectively. From these it can be seen that a higher V_T for transistor *M1* results in a faster write, but no change for read. It also decreases the possible V_{PC} window but the *actual* window (limited by bitline development during read) stays the same.

Figure B.4 and figure B.5 shows the graphs when V_{T2} has been changed +40mV and -40mV respectively. From these it can be seen that a higher V_T for transistor *M2* does not change the write time, and the read time is only increased very slightly. The possible V_{PC} window stays unchanged and the *actual* window is decreased very slightly.

Figure B.6 and figure B.7 shows the graphs when V_{T3} has been changed +40mV and -40mV respectively. From these it can be seen that a higher V_T for transistor *M3* drastically increases the write time but leaves the read time virtually unchanged. The possible V_{PC} window becomes larger, but the *actual* window is decreased slightly. This latter is due to a shifted precharge voltage window where the lower limit goes up slightly, whereas the upper limit is still limited by the read bitline development.

Figure B.8 and figure B.9 shows the graphs when V_{T4} has been changed +40mV and -40mV respectively. From these it can be seen that a higher V_T for transistor *M4* decreases the write time, but leaves the read time virtually unchanged. The possible V_{PC} window becomes smaller, and the *actual* window also decreases slightly.

Figure B.10 and figure B.11 shows the graphs when V_{T5} has been changed +40mV and -40mV respectively. From these it can be seen that a higher V_T for transistor *M5* increases the write time and the read time. The possible V_{PC} window becomes larger, but the *actual* window decreases.

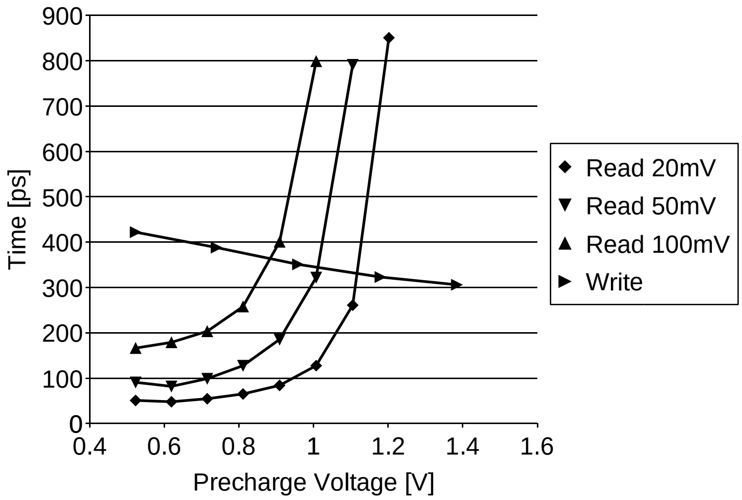

Figure B.2: Write performance and bitline development during read for a 40mV increase of V_{T1}

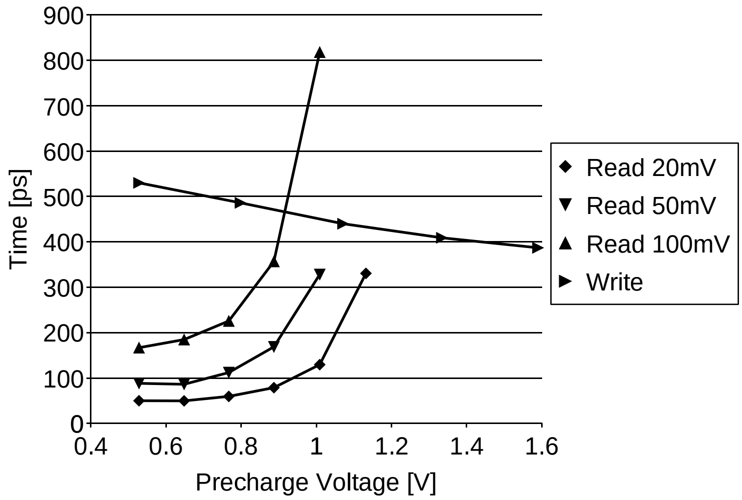

Figure B.3: Write performance and bitline development during read for a 40mV decrease of V_{T1}

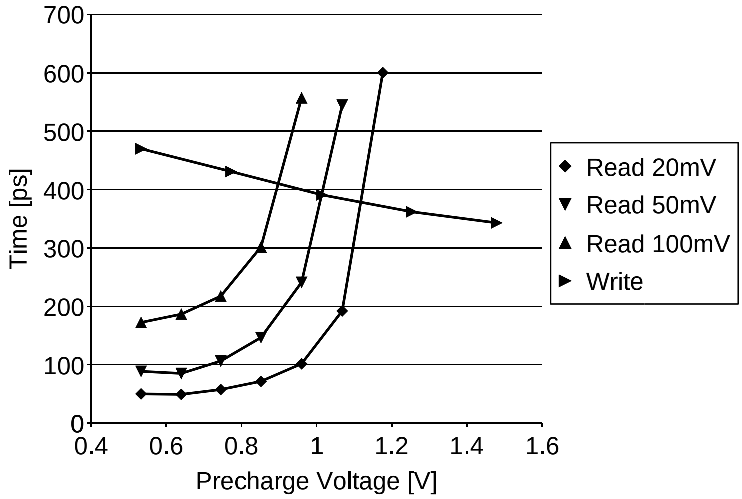

Figure B.4: Write performance and bitline development during read for a 40mV increase of V_{T2}

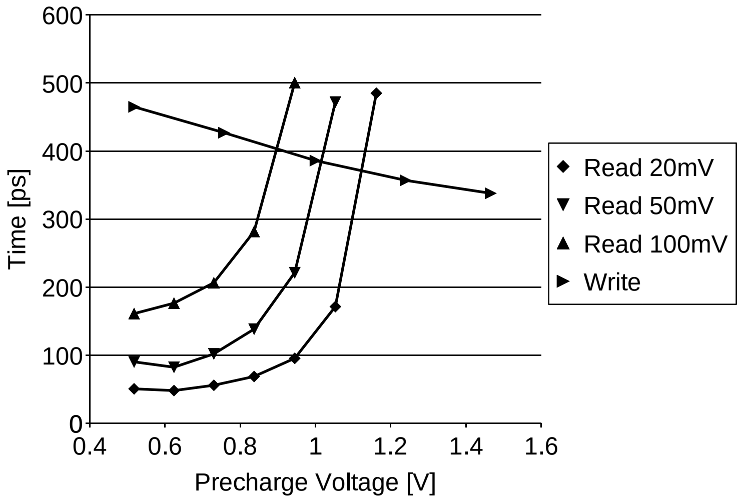

Figure B.5: Write performance and bitline development during read for a 40mV decrease of V_{T2}

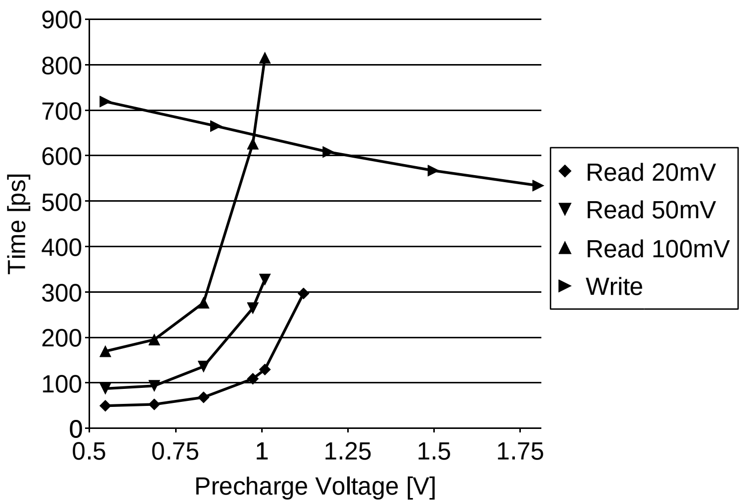

Figure B.6: Write performance and bitline development during read for a 40mV increase of V_{T3}

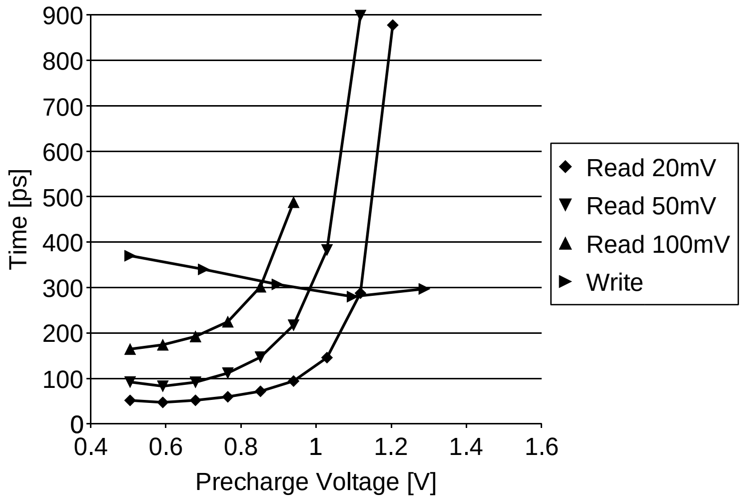

Figure B.7: Write performance and bitline development during read for a 40mV decrease of V_{T3}

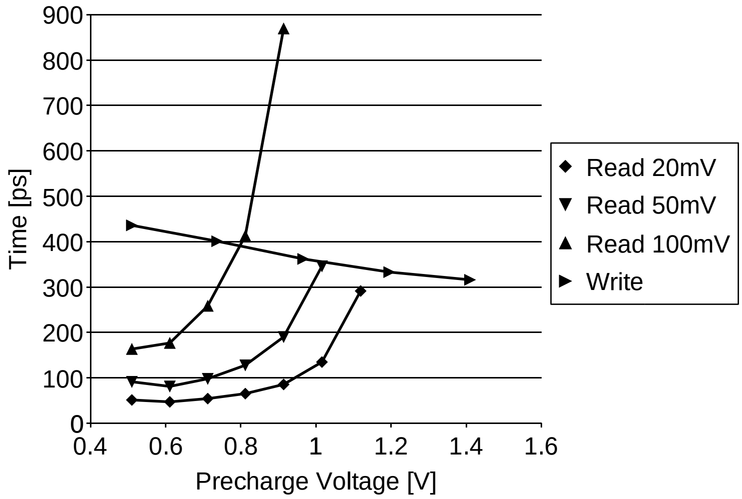

Figure B.8: Write performance and bitline development during read for a 40mV increase of V_{T4}

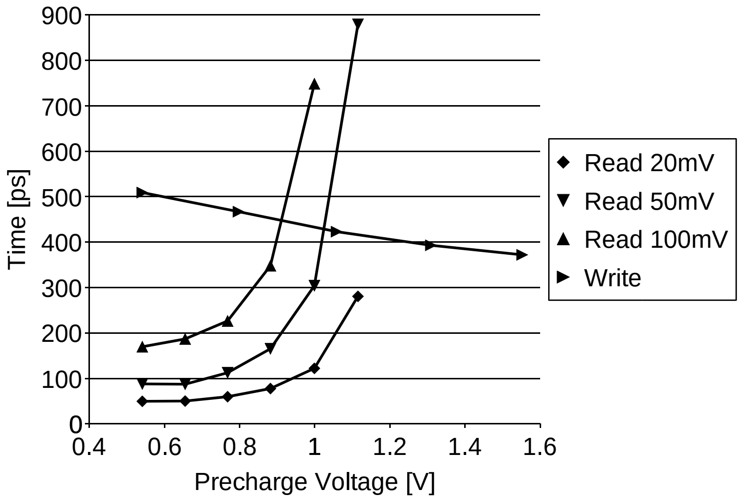

Figure B.9: Write performance and bitline development during read for a 40mV decrease of V_{T4}

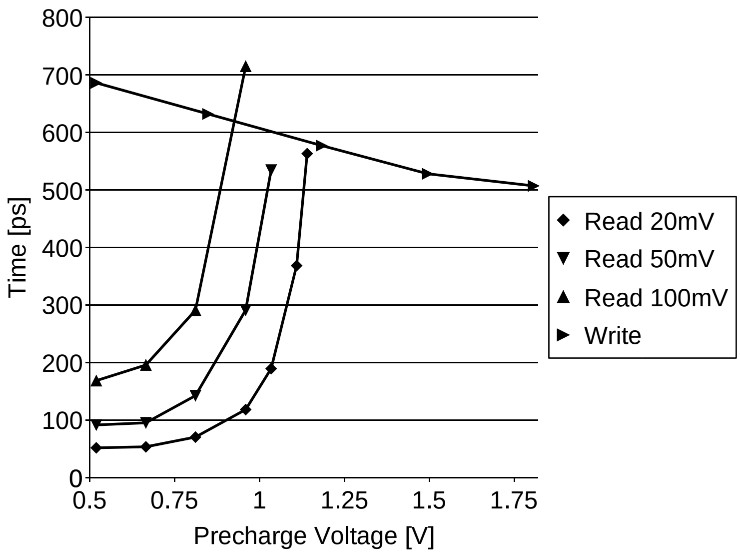

Figure B.10: Write performance and bitline development during read for a 40mV increase of V_{T5}

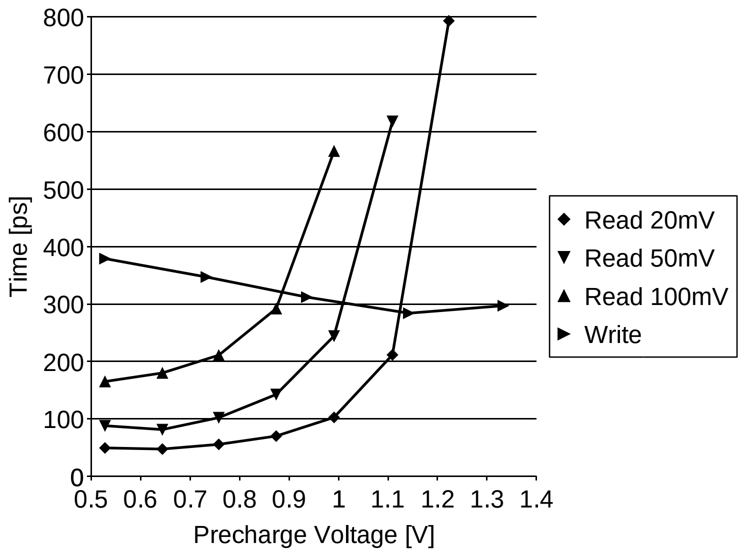

Figure B.11: Write performance and bitline development during read for a 40mV decrease of V_{T5}