NTNU – Trondheim
Norwegian University of
Science and Technology

Ultra-Low Voltage SRAM in 130nm CMOS Process

Ole Samstad Kjøbli

Electronics System Design and Innovation
Submission date: June 2014
Supervisor: Snorre Aunet, IET
Co-supervisor: Jan Rune Herheim, Atmel Norway AS

Norwegian University of Science and Technology
Department of Electronics and Telecommunications

Preface

This thesis is the finishing project for the degree *Master of Science* in *electronics, design of digital systems* at the Department of Electronics and Telecommunication, Faculty of Information Technology, Mathematics and Electrical Engineering (IME) at the Norwegian University of Science and Technology (NTNU). The thesis is a continuation of a specialization project completed in the autumn-term of 2013 for Atmel Norway AS.

Working on this thesis has been both challenging and interesting. It has provided me with a lot of insight in the process flow of integrated circuits, general circuit design, low-power design and CAD tools used in the industry.

I would like to thank my supervisors Snorre Aunet (Proffessor at NTNU) and Jan Rune Herheim (Analog Team leader at Atmel Norway AS) for their help and support throughout this last semester. Getting to grips with all the intricacies of subthreshold circuits, CAD-tools and more would have been a much more challenging task without their assistance. I would also like to thank fellow student Glenn Andrè Johnsen for his insightful discussions and ideas during the course of this last semester.

Lastly I would also like to thank my family for their help and support.

Trondheim, June 2014

Ole Samstad Kjøbli

Abstract

Energy harvesting systems typically contain a low-power embedded processor in order to collect and interpret sensory data and such a processor will need memory to store that data. The most effective method of reducing power consumption in an electronic circuit is to decrease the supply voltage and this thesis explores the viability of implementing an ultra-low voltage SRAM architecture in a 130nm CMOS process for Atmel Norway AS. The architecture supports voltage scaling between 400mV and a regular supply voltage of 1.2V.

The architecture was implemented with conventional 6T SRAM cells and 10T SRAM cells designed for low-voltage operations using state of the art design techniques and literature. The SRAM architecture is asynchronous and self-timed to more easily cope with the effects of process and temperature variations. To realize the architecture a small set of logic gates were also designed for ultra-low voltage operation and used in the SRAM read and write control circuitry. All building blocks in the architecture were simulated with extracted parasitics to get more realistic simulation results. Corner and Monte Carlo simulations were used to show how temperature and process variations statistically affected the building blocks and their performance.

Simulation results showed that the 10T SRAM cell is more robust with a 60-100% larger static noise margin compared to the conventional 6T cell, but draws 1.2-1.6 times more leakage power and is physically 64% larger. The differential nature of the 6T cell makes its read operations faster compared to the 10T cell, but the offset voltage in the sense amplifiers used for reading reduces the potential speed gain somewhat. The 6T cell also experience a disturb voltage during read operations and the nature of this disturbance is different at subthreshold and superthreshold voltages, making it difficult to assess yield in a system supporting voltage scaling. the 10T cell does not experience this problem which makes it the more predictable and safe choice for future implementations. Reducing the voltage from 1.2V to 400mV gives a power saving in the range 4-18 depending on process variations and temperature. At low temperatures the supply voltage must be increased either permanently or by using dynamic voltage compensation to perform a read operation within a 32kHz clock cycle.

This thesis has showed that it is viable to implement a subthreshold SRAM architecture in the Atmel 130nm CMOS process and some important effects of applying voltage scaling have been explored. Reducing the power supply to such an extent reduces performance and will need some form of voltage compensation to increase performance at low temperatures.

Sammendrag

Energisankende systemer har typisk en laveffekts mikroprosessor som samler og tolker sensordata og en slik prosessor trenger minne for å lagre disse dataene. Den mest effektive metoden for å redusere effektforbruk i en elektrisk krets er å redusere forsyningsspenningen og denne oppgaven utforsker mulighetene for å implementere en SRAM-arkitektur for ultralave spenninger i en 130nm CMOS prosess for Atmel Norway AS. Arkitekturen støtter spenningsskalering mellom 400mV og en standard forsyningsspenning på 1.2V.

Arkitekturen ble implementert med konvensjonelle 6T SRAM-celler og 10T SRAM-celler designet for operasjon ved lave spenninger ved hjelp av designteknikker og litteratur på aktuelt teknisk nivå. Arkitekturen er asynkron og selvklokket får å håndtere effektene av prosess og temperaturvariasjoner bedre. Et lite sett av logiske porter ble også designet for operasjon ved ultralave spenninger for å realisere periferikretser for lese og skrivekontroll. Alle byggeblokker ble simulert med ekstraherte parasitter fra layout for å få realistiske simuleringsresultater. Hjørne og Monte Carlo simuleringer ble brukt for å vise hvordan prosessvariasjoner og temperatur statistisk påvirket byggeblokkene og deres ytelse.

Simuleringsresultatene viser at 10T-cellen er mer robust og har en 60 til 100% større statisk støymargin sammenlignet med den konvensjonelle 6T-cellen, men den drar 1.2 til 1.6 ganger større lekkasjestrømmer og er fysisk 64 % større. Den differensielle lesemetoden til 6T-cellen gjør leseoperasjoner raskere sammenlignet med 10T-cellen, men offsetspenningen i leseforsterkeren gjør at vinningen i ytelse blir noe redusert ved lave spenninger. Leseoperasjonen med 6T celler skaper også en forstyrrelsesspenning i cellene og magnituden av denne påvirkes forskjellig ved subterskel og superterskelspenninger, noe som gjør det vanskelig å forutse produksjonsutbytte med spenningsskalering. 10T-cellene har ikke dette problemet og er derfor et mer forutsigbart og sikkert valg for framtidige implementasjoner. Resultatene viser også at å redusere forsyningsspenningen fra 1.2V til 400mV kan gi en effektsparing på 4 til 18 ganger avhengig av prosessvariasjoner og temperatur. Ved lave temperaturer reduseres ytelsen såpass at forsyningsspenningen må økes enten permanent eller ved hjelp av dynamisk spenningskompensering for å øke ytelsen slik at en operasjon kan fulføres i løpet av en 32kHz klokkesykel.

Oppgaven har vist at det er mulig å implementere en SRAM-arkitektur for ultralave spenninger i Atmels 130nm CMOS prosess og noen viktige effekter av spenningsskalering har også blitt vurdert. Ved å redusere forsyningsspenningen såpass mye trengs det en form for spenningskompensering for å øke ytelsen ved lave temperaturer.

Table of Contents

Acronyms

ULP	Ultra-Low Power
ULV	Ultra-Low Voltage
MCU	Microcontroller Unit
SRAM	Static Random-Access Memory
MOSFET	Metal-Oxide Semiconductor Field-Effect Transistor
CMOS	Complementary Metal-Oxide Semiconductor
PVT	Process, Voltage and Temperature
IMRAD	Introduction, Methods, Results and Discussion
PUN	Pull-up Network
PDN	Pull-down Network
DRAM	Dynamic Random-Access Memory
SA	Sense Amplifier
NM	Noise Margin
SNM	Static Noise Margin
VTC	Voltage Transfer Characteristic
DIBL	Drain-Indiced Barrier Lowering
FBB	Forward Body Biasing
RBB	Reverse Body Biasing
IF	Imbalance Factor
ADE	Analog Design Environment
DRC	Design Rule Check
LVS	Layout vs. Schematic
PEX	Parasitic Extraction
$\boldsymbol{VV_{DD}}$	Virtual-V_{DD}
ATD	Address Transition Detector
TD	Transition Detector
DWL	Divided Wordline
IO	Input/Output
RNCE	Reverse Narrow Channel Effect

PDP	Power-Delay Product
FOM	Figure of Merit
PRCP	Power-per-Read Cycle Product
HVT	High-V_{th}
RVT	Regular-V_{th}
LVT	Low-V_{th}

List of Figures

List of Tables

1. Introduction

The trends of the consumer electronics market indicates that future products will become even more mobile and connected. The concept of *"the internet of things"* predicts that even the simplest gadget will be connected to the internet, and since no person wants to run around every day and change batteries in every household appliance the battery life of said appliances must be increased. Modern mobile phones have processing capabilities that surpasses the most powerful computers of earlier decades while being battery powered. Unfortunately there have been relatively few breakthroughs in consumer battery technology so optimizations for increased battery life have been achieved through ingenious optimizations in hardware and/or software, but even with these optimizations it is often necessary to recharge a mobile phone after a day of constant use.

Lowering the supply voltage is a very efficient, if not the most efficient way of reducing the power consumption of an electronic circuit. One field that is gaining interest and traction among both corporate and scientific institutions is the concept of ultra-low power (ULP) design, which is equivalent to ultra-low voltage (ULV) design because of the relationship between power consumption and voltage. This design methodology is also known as subthreshold or near-threshold design because the supply voltage is often lowered to values below or near the absolute value of the transistors threshold voltage.

Atmel Corporation designs and produces microcontrollers units (MCU) used in many different markets, and as such it is in their interest to explore the possibilities of ULV design. In order to make a ULV microcontroller certain elements must be in place; A standard cell library for various logic gates, flip-flops, and memory to store data. This thesis focuses mainly on a memory architecture consisting of static random-access memory (SRAM) cells for ULV applications.

1.1 Previous Work

Circuits operating at subthreshold voltages is by no means a new concept. Low power circuits were experimented with as early as the late 1960s[1] and models describing subthreshold operation of the metal-oxide semiconductor field-effect transistor (MOSFET) was described in the early 1970s. The work of E. Vittoz showed that subthreshold circuits could be a viable option in applications where low power consumption was the most important metric and one example that was implemented in a physical product was a low power quartz oscillator used in wrist watches [2].

Very few commercial products on the market today utilizes components operating at subthreshold voltages, but one recent example is a real time counter developed by the Texas-based Ambiq Micro [3]. The research of subthreshold operation has been limited mostly to academic and scientific institutions. Many academic papers have been written about the subject and many prototypes have been produced, but a small quantity of prototype chips do not translate well to a large-scale commercial production line. Because of low yield and difficulties of large scale production of subthreshold circuits the commercial sector has been reluctant to adapt subthreshold design, but this might change with the increasing needs for low power consumption.

This thesis is a continuation of a previous specialization project from 2013 also conducted for Atmel Norway AS. The purpose of the specialization project was to compare various SRAM cell designs for subthreshold operation[4].

1.2 Problem Description

The project description given by Atmel Norway AS is as follows:

> ***Ultra-Low-Voltage scalable SRAM topology including sense amplifier***
>
> *Energy harvesting systems typically contain an embedded processor to collect, process, and interpret sensory input data. The system typically includes a CPU, memories, buses, and peripherals. As memory in such a system an ultra low voltage SRAM could be used.*
>
> *This assignment involves specifying and designing a sense amplifier for an already existing ULV SRAM cell. These building blocks will be used to build a configurable SRAM topology. The design should be done using state of the art design techniques and literature.*
>
> *The required software, hardware and a working place at Atmels office could be provided.*

1.3 Thesis Purpose and Scope

The purpose of this thesis is to use state of the art design techniques and literature to design a memory architecture for subthreshold operation in a 130nm complementary metal-oxide semiconductor (CMOS) process provided by Atmel Norway AS. The memory architecture will consists of SRAM cells that must maintain robustness at low voltages and peripheral circuits needed to interface with the SRAM cells like decoders, driving circuits and sense amplifiers. The peripheral circuitry will need a digital logic library for ULV operation. This is not available so a small set of logic gates must be designed for ULV operation as well.

The specifications given by Atmel Norway AS is as follows:

- The architecture should support voltage scaling from a decided minimum V_{DD} up to the regular supply voltage of 1,2V.
- Speed is not an issue, but the SRAM architecure must function at frequencies set by a standard internal oscillator. Atmel internal oscillators operate at 32kHz.
- The memory architecture must function in the standard temperature range for consumer electronics (-40^{o}C $\rightarrow$ 85^{o}C).
- It can be assumed that a voltage regulator that compensates for temperature and process variations are present in a potential system.
- The size of SRAM arrays is not important, but it should be in the range of 1K to 8K to be of any use in commercial products.

In this thesis the entire architecture will be explained in detail and the benefits and challenges of moving from subthreshold to superthreshold voltages will be explored.

The scope of this thesis is limited to the SRAM array with peripheral devices. In order to make these peripheral devices a few standard cells will be created. The SRAM architecture is asynchronous and self-timed in order to combat the effects of process, voltage and temperature (PVT) variations and ease timing constraints. Two variants of the architecture will be implemented; one with the conventional 6T SRAM cell and one with a 10T SRAM cell designed for increased robustness at low voltages. The positives and negatives of both implementations are compared based on metrics like speed, power consumption, robustness and area.

1.4 Thesis Structure Overview

This thesis folows the introduction, methods, results and discussion (IMRAD) structure and the contents of each chapter is as follows:

- **Chapter 1** presents the motivation for the thesis, introduces some previous work done in the field of ULV design. The verbatim problem description given by Atmel Norway AS and the purpose and scope of this thesis is also presented.
- **Chapter 2** gives an introduction to power consumption, SRAM operation, subthreshold design challenges and some of the tools used in the thesis.
- **Chapter 3** presents the proposed memory architecture and how it was designed.
- **Chapter 4** presents the building blocks used in the memory architecture.
- **Chapter 5** present the simulation methods used on each individual component and the memory architecture.
- **Chapter 6** contains the results from the methods described in chapter 4.
- **Chapter 7** contains the physical layout of some components used in the memory architecture.
- **Chapter 8** discusses some of the main aspects of this thesis.
- **Chapter 9** concludes the work done in the thesis based on results and discussion.
- **Chapter 10** provides some ideas for further work with the concepts presented in the thesis.
- **Appendix A** Contains simulations results for all logic gates.
- **Appendix B** contains layouts for all logic gates.
- **Appendix C** contains Verilog-A source code used in simulations.

2. General Background

This chapter contains the necessary background theory of power consumption and its sources, basic SRAM operation, subthreshold design challenges and the tools used in simulations.

2.1 Power Consumption

The power consumption of a logic gate can be divided into three sources as shown in Fig. 2.1. The total power consumed by the logic gate can be expressed as the sum of these three sources as shown by equation[5].

$$P_{total} = P_{dynamic} + P_{leakage} + P_{short-circuit} \tag{2.1}$$

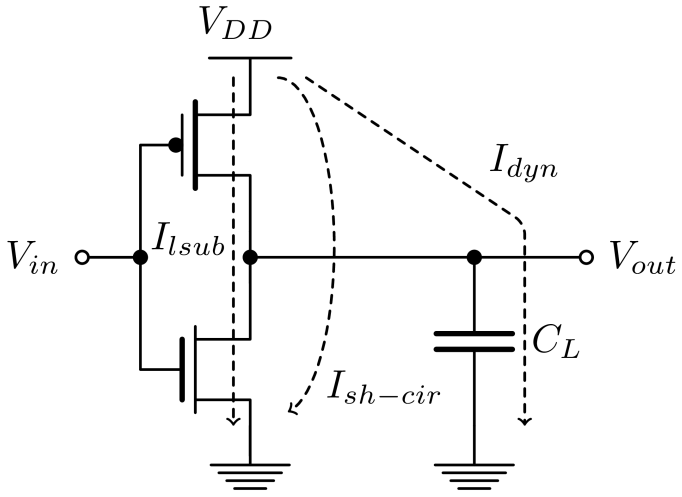


Figure 2.1: Power contributions in an inverter

2.1.1 Dynamic Power Consumption

Dynamic power consumption occurs when a logic gate changes/switches from one state to another. During this change a current is drawn from the power supply into the load of the logic gate. Assuming the load is mostly capacitive the dynamic power consumption can be estimated using equation 2.2[6].

$$P_{dynamic} = \alpha f_{switch} V_{DD}^2 C_L \tag{2.2}$$

where f_{switch} is the switching frequency, C_L is the capacitive load, V_{DD} is the supply voltage and α is the switching activity factor. Reducing any of these factors will lower the total power consumption of the circuit.

2.1.2 Leakage Power Consumption

Leakage power consumption is caused by an inherent subthreshold leakage current that is always drawn by transistors and is given by equation 2.3[6].

$$P_{leakage} = V_{DD} I_{lsub} \tag{2.3}$$

where I_{lsub} is the subthreshold leakage current. The value of I_{lsub} is dependent on transistor design parameters like width, length and the biasing of the transistor. The nature the I_{lsub} will be explained in more detail in section 2.3.

2.1.3 Short-Circuit Power Consumption

The short-circuit power consumption can be lumped with the dynamic power consumption as it only occurs when the gate is switching. In digital logic gates there is always a time interval when there exists a short-circuit path between V_{DD} and the ground potential through the pull-up networks (PUN) and pull-down networks (PDN) of the logic gate as shown in Fig. 2.2.

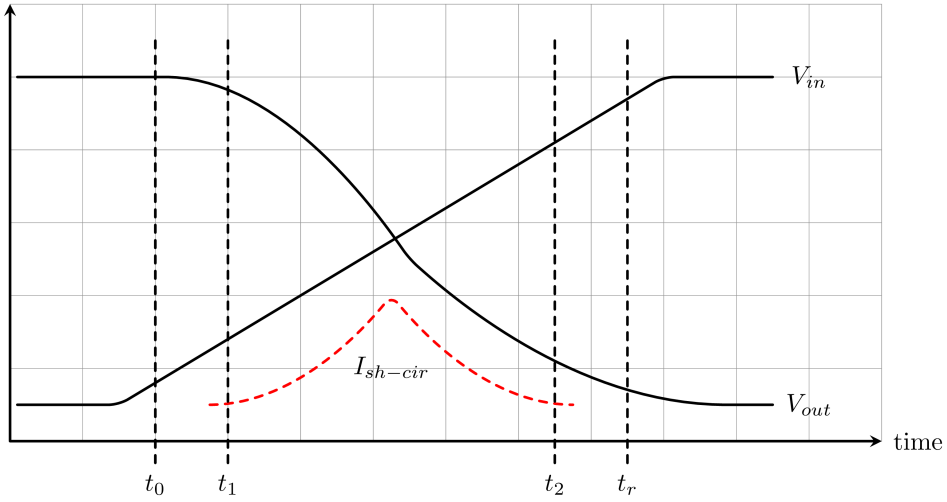


Figure 2.2: Short-circuit power

The short-circuit current drawn increases and decreases in the time interval between t_1 and t_2 when $V_{tn} < V_{in} < V_{tp}$. Peak short-circuit power is drawn when the PUN and PDN draws equal amounts of current. In a perfectly balanced logic gate this should ideally occur when $V_{in} = V_{out} = V_{DD}/2$.

2.2 Introduction to SRAM

Memories based on MOSFETs emerged in the late 1960s with the dynamic random-access memory (DRAM) cells. These memory cells are simply a capacitor storing one bit of data with a MOSFET switch controlling read and write access. The performance of DRAM has not been able to follow the performance increase of modern processors because of access times and power consumption. DRAM cells must also be periodically refreshed to prevent the capacitor from discharging [7]. SRAM provides faster access times and robustness, and is often used as on-chip memory/cache in modern processors. The improvements provided by SRAM comes at the cost of more area on chip, as a conventional SRAM cell consists of 6 transistors compared to the DRAMs single capacitor and transistor.

2.2.1 The 6T SRAM Cell

The conventional 6T SRAM cell is shown in Fig. 2.3. The cell consists of a bistable latch made from two cross-coupled inverters ($M1 - M4$) and two access transistors ($M5$ and $M6$) that provides read and write access to the latch. Data is stored as two complementary logic values in the internal nodes Q and $\overline{Q}$. The SRAM cell must be designed with contradicting design requirements so that a read operation does not disturb the data stored in the cell and that a write operation is able to force new values to the internal nodes of the cell.

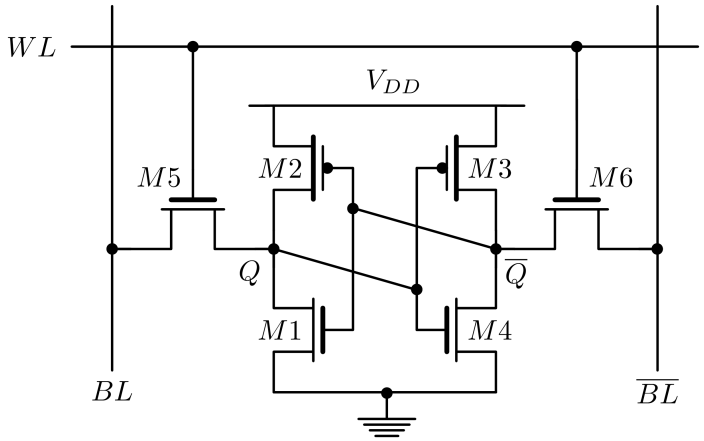


Figure 2.3: conventional 6T SRAM cell

The SRAM cell has three modes of operation: read, write and hold.

Read Operation

Before the read operation the bitlines (BL and $\overline{BL}$) are precharged to V_{DD}. The read operation is initialized by enabling the wordline (WL) and turning on the access transistors. One of the precharged bitlines will be pulled down by the corresponding access transistor because one of the internal nodes hold a logic "0". To speed up the read operation a sense amplifier (SA) is often used to perform differential sensing between BL and $\overline{BL}$. Once a sufficient differential voltage is established between BL and $\overline{BL}$ the SA is able to read the data long before the bitline has discharged completely. A complete discharge of the bitline takes more time and consumes more

dynamic power compared to a partial bitline discharge, so SAs are beneficial in many ways. A read "0" operation is shown in Fig. 2.4.

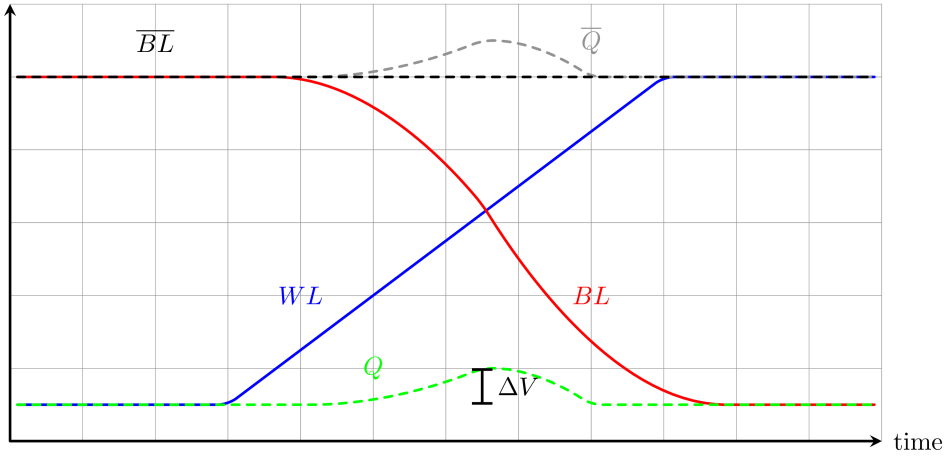


Figure 2.4: SRAM read "0" operation

When the access transistors are turned on a voltage divider is formed between one of the access transistors and the adjacent NMOS transistor of the cross-coupled inverter generating a logic "0". In the case when Q = "0" $M6$ and $M4$ form a voltage divider and the voltage at Q is $0V + \Delta V$. If ΔV is too large it might cause the other inverter to flip its output and cause a destructive read operation. To prevent destructive read operations from happening the cross-coupled inverters must be sized to withstand the added ΔV.

The design requirements for achieving non-destructive read operations with the conventional 6T cell are given by equation 2.4[7].

$$\left(\frac{W_1}{L_1}\right) > \left(\frac{W_5}{L_5}\right), \left(\frac{W_4}{L_4}\right) > \left(\frac{W_6}{L_6}\right) \tag{2.4}$$

Write Operation

The bitlines are precharged to V_{DD} before the write operation as well. The write operation is initialized by driving BL the value that will be written to Q and $\overline{BL}$ to the complementary value. When the wordline is enabled the access transistors forces the value of the bitlines into the internal nodes of the SRAM cell. A write "1" operation is shown in Fig. 2.5. The cross-coupled inverters must be strong enough to withstand the added ΔV during read operations, but they must be weak enough to allow new data to be written to the cell.

The design requirements for achieving successful write operations with the conventional 6T cell are given by equation 2.5[7].

$$\left(\frac{W_5}{L_5}\right) > \left(\frac{W_2}{L_2}\right), \left(\frac{W_6}{L_6}\right) > \left(\frac{W_3}{L_3}\right) \tag{2.5}$$

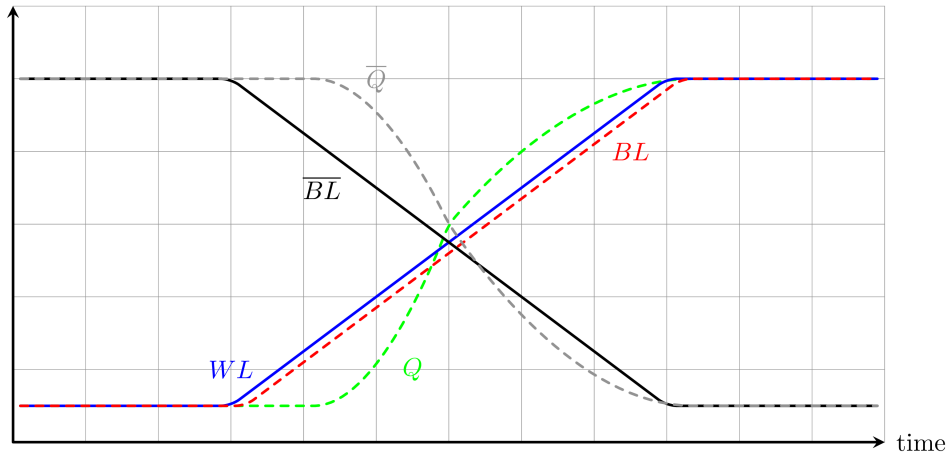


Figure 2.5: SRAM write "1" operation

Hold Operation

When the SRAM cell is not accessed for reading or writing the cell is holding the stored data because of the positive feedback provided by the cross-coupled inverters. SRAM is a volatile type of memory, meaning that if the power supply is turned off the stored data will be lost.

2.2.2 Static Noise Margins

In the previous section it was mentioned that the cross-coupled inverters must tolerate a voltage change ΔV over ground potential in order to maintain a non-destructive read operation. The maximum noise voltage the cross-coupled inverters can handle is often called the *noise margin* (NM) and it is a common way of measuring the robustness of SRAM cells. The *static noise margin* (SNM) can be extracted for all three modes of operation and are called read SNM, hold SNM and Write SNM. The read SNM is always the smallest in the conventional 6T SRAM cell and is usually considered the worst-case SNM.

Fig. 2.6 shows how the SNM can be simulated with a simple DC analysis where the noise voltages are modeled as DC voltge sources[8]. The noise voltages are swept from 0V to V_{DD} and the voltage transfer characteristics (VTC) of the cross coupled inverters are used to construct an aptly named butterfly plot. The different SNMs can be extracted by setting the wordline and bitline voltages according to table 2.1.

SNM	$\boldsymbol{V_{WL}}$	$\boldsymbol{V_{BL}}$	$\boldsymbol{V_{\overline{BL}}}$
HSNM	0V	Dont care	Dont care
RSNM	V_{DD}	V_{DD}	V_{DD}
WSNM	V_{DD}	0V	V_{DD}

Table 2.1: SNM test cases

Fig. 2.7a, 2.7b and 2.7c shows butterfly plots for all three operations. During a hold operation the VTCs of both inverters are ideally equal because no voltage dividers are

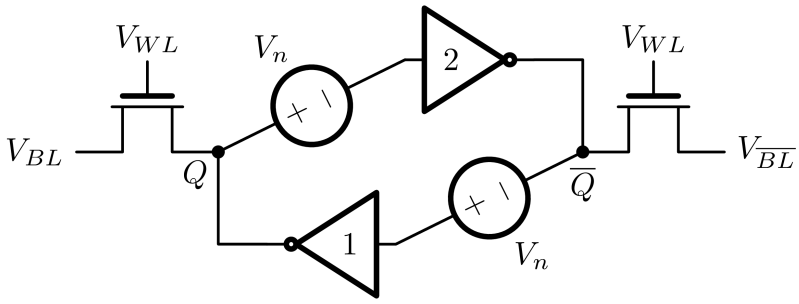
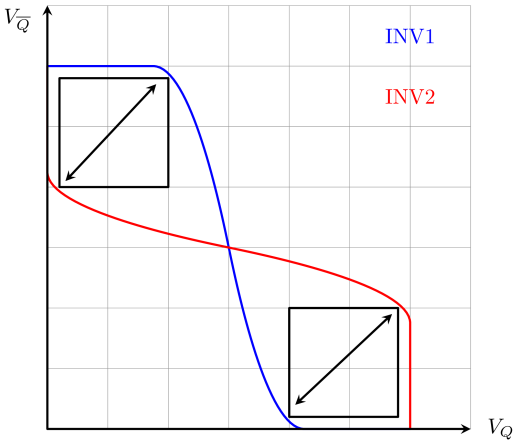
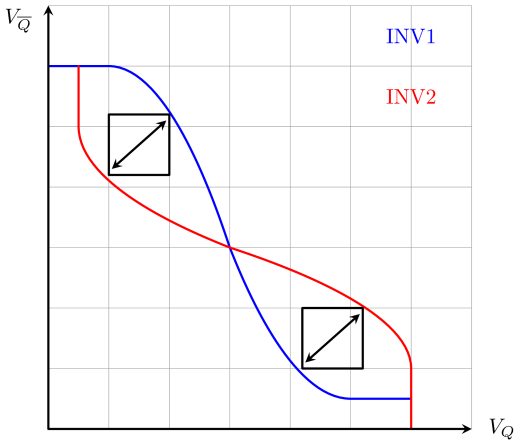


Figure 2.6: SNM extraction

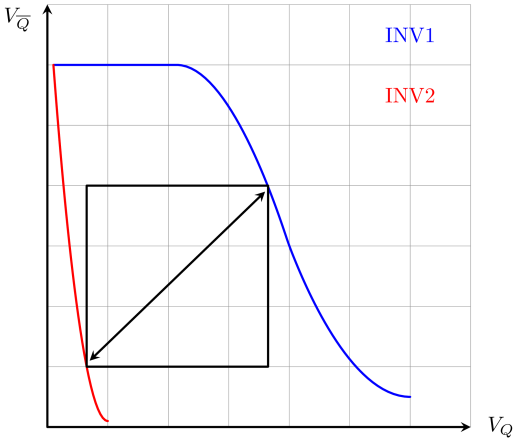
formed, but mismatch between individual transistors will make the VTCs different. Because of the voltage divider phenomenon the symmetry is degraded when reading and completely broken when writing. The SNM is equal to the length of the diagonal of the largest square that fits into the lobes of the butterfly plot. During write operations the VTCs share only one crossing point on the butterfly plot, indicating that the SRAM cell is monostable and that the write operation is successful[9].



(a) Butterfly plot for HSNM extraction



(b) Butterfly plot for RSNM extraction



(c) Butterfly plot for WSNM extraction

2.3 Subthreshold Design Challenges

A MOSFET operates in the subthreshold domain when the supply voltage is set below the absolute value of the transistors threshold voltage V_{th}. Operating at these voltages can offer large savings in power consumption, but introduces other new design challenges not present at superthreshold voltages.

2.3.1 Subthreshold Currents

In superthreshold designs a MOSFET operates in three modes of operation: cutoff, triode and saturation. When the transistor is in cutoff there is still some current flowing through it, but these off-currents are so small compared to the on-current that they are often ignored for simplicity. When transistors are operate at subthreshold voltages these leakage currents becomes the main contributing currents. The three largest leakage current contributions at subtreshold voltages are: the gate currents, caused by tunnelling through the dielectric layer; the junction currents, caused by band-to-band tunneling across the depletion layer; and the subthreshold current, caused by carrier diffusion [10]. Fig. 2.8 shows these currents in a NMOS transistor.

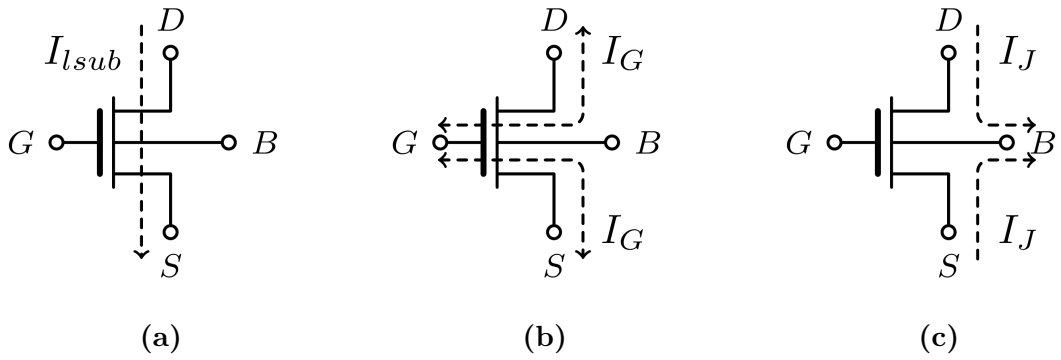


Figure 2.8: (a) Subthreshold current (b) Gate currents (c) junction currents

The impact of each current depends on the circuit and its design. The channel length is often increased in SRAM circuits to combat the effects of random dopant fluctuations during wafer production, and this increases the impact of the gate current I_G [10] to a certain extent. The subthreshold current I_{lsub} usually dominates and can be estimated using equation 2.6[11].

$$I_{lsub} = \beta e^{V_{GS}/n \cdot V_t}[e^{\lambda_{DS} V_{DS}/n \cdot V_t}](1 - e^{-V_{DS}/V_t}) \tag{2.6}$$

β is the MOSFET driving strength given by equation 2.7.

$$\beta = I_0 \frac{W}{L} e^{-(V_{TH0} - \lambda_{BS} V_{BS})/n \cdot V_t} \tag{2.7}$$

where I_0 is the residual current when $V_{GS} = 0$, V_t is the thermal voltage, V_{TH0} is the threshold voltage with zero substrate bias, W/L is the transistor size, V_{BS} and V_{DS} are the body-to-source and drain-to-source voltage respectively and n is the subthreshold slope

factor given by equation 2.8[12]. λ_{BS} and λ_{DS} are the body effect and drain-induced barrier lowering (DIBL) coefficients respectively.

$$n = 1 + \frac{C_d}{C_{ox}} \tag{2.8}$$

The driving strength β can be changed by adjusting the transistors size, selecting a different V_{TH0} or by applying body biasing. When a positive voltage is applied to the bulk node of the transistor forward body biasing (FBB) is applied while a negative voltage results in reverse body biasing (RBB). FBB increases β while RBB decreases it. The usefulness of body biasing depends on the process, as single-well processes only allows body biasing of either the NMOS or PMOS transistors.

2.3.2 On/Off-Current Ratio

From a digital point of view a transistor is either on or off. The currents that flow trough the transistor for both these states can be found by inserting $V_{GS} = V_{DD}$ and $V_{GS} = 0V$ into equation 2.6. Assuming no body biasing and no DIBL effect the two currents are given by equations 2.9 and 2.10.

$$I_{on} \approx \beta e^{V_{DD}/n \cdot v_t} \tag{2.9}$$

$$I_{off} \approx \beta \tag{2.10}$$

At superthreshold voltages the ratio between these two currents is very large, but at very low voltages the off-current can become significant in magnitude compared to the on-current and the number of transistors sharing a common node in a circuit becomes limited. The problem is illustrated in Fig. 2.9.

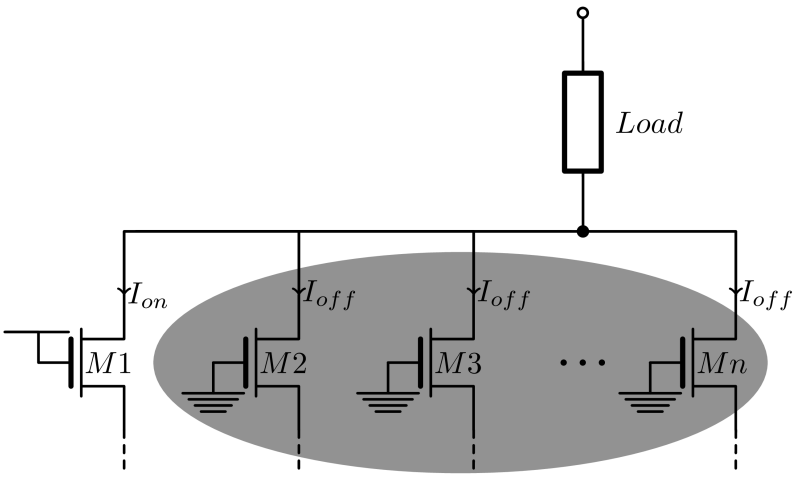


Figure 2.9: On and off-currents with a shared node

If we assume the shared node is a bitline in a SRAM column and the transistors are access transistors, the on-current of the selected cell must be larger than the combined off-currents of all the other cells[11]. If the on and off-currents are too similar the voltage levels of a logic ”1” and logic ”0” can become indistinguishable.

2.3.3 Delay

Using a perfectly balanced inverter as the basis, the delay can be estimated as a CV/I-metric using equation 2.11[12].

$$\tau_D = \frac{C_G V_{DD}}{I_{on}} = \frac{C_G V_{DD}}{\beta e^{V_{DD}/nv_t}} \tag{2.11}$$

Where C_G is the inverter gate capacitance. When operating at subthreshold voltages the on-current is exponentially dependent on V_{DD} and as a result the delay increases. This suggest that subthreshold circuits should be used for in systems and modules where reduced power consumption is more important than speed like always-on modules. Subthreshold and/or near-threshold circuits could also be used to combat the increased number of unused components/dark silicon in modern CMOS processes if the delays are within acceptable margins[13].

2.3.4 NMOS/PMOS Imbalance

At superthreshold voltages a NMOS transistor usually has a driving strength approximately 2-3 times that of a PMOS transistor. At subthreshold voltages this imbalance is typically higher, and depending on the process either the PMOS or NMOS transistor can be the stronger device. The imbalance factor IF is used to evaluate the driving strength ratio between PMOS and NMOS transistors [11].

$$IF = max\left[\frac{\beta_p}{\beta_n}, \frac{\beta_n}{\beta_p}\right] \geq 1 \tag{2.12}$$

A low IF is desired to achieve good VTC balance in logic gates. Lower balance means that the output logic levels of the logic gate is degraded and the leakage currents of subsequent logic gates are increased. As a result the IF determines the maximum logic depth for a given V_{DD} and vice versa.

2.3.5 Parallel Transistors

The XOR gate in Fig. 2.10 implemented in a 65nm CMOS process fails to drive the output to logic "1" for the input AB = "10" when V_{DD} = 100mV[12]. The parallel-connected $M1$, $M3$ and $M4$ are turned off and their combined leakage current is too significant to the on-current of $M2$.

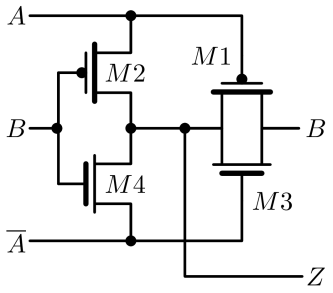


Figure 2.10: Problematic subthreshold XOR implementaion

The alternative XOR implementation in Fig. 2.11 has equal amounts of on/off transistors in the PDN and PUN for all input combinations resulting in approximately equal on/off-current ratio for all input combinations.

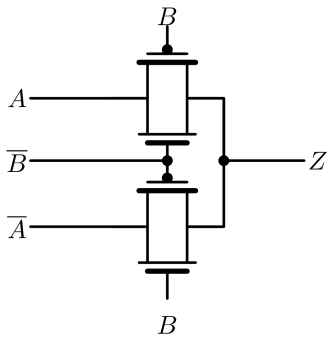


Figure 2.11: Good subthreshold XOR implementaion

2.3.6 The Stack Effect

Stacked transistors reduce both the on-current and leakage current so the stack effect can be exploited to reduce power consumption, but transistor sizes must be increased to compensate for the decreased driving on-current if the logic gate is driving large capacitive loads.

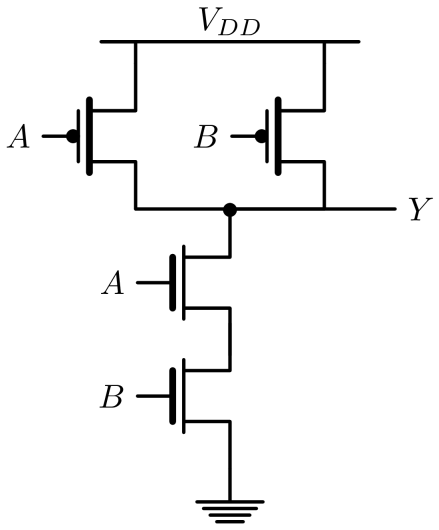


Figure 2.12: A conventional 2-input NAND gate

The conventional boolean 2-input NAND gate in Fig. 2.12 might not be able to drive the output to logic ”0” at very low voltages since the stack reduces the pull-down current and the leakage of the PMOS transistors in the PUN might becomes too significant. As shown in Table 2.2 the stack decreases the effective NMOS driving strength β_n by a factor of two when AB = ”11”. The strength is further reduced for every new boolean input that is added, so logic gates with fan-in higher than 2-3 should be avoided at subthreshold voltages

A	B	Pull-up strength	Pull-down strength
0	0	$2\beta_p$	-
0	1	β_p	-
1	0	β_p	-
1	1	-	$0.5\beta_n$

Table 2.2: NAND2 input vectors and effective drive strength

2.3.7 Process, Voltage and Temperature Variations

The exponential dependencies between transistor parameters, environment conditions and current makes subthreshold circuits more sensitive to PVT variations[11]. Equation 2.13 and 2.14 shows that higher temperatures decreases the mobility factor μ and lowers the threshold voltage V_{th} of a transistor. Lower mobility increases the delay while a lower threshold voltage decreases the delay [14].

$$\mu(T) = \mu(T_0)\left(\frac{T}{T_0}\right)^{-M} \tag{2.13}$$

$$V_{th}(T) = V_{th}(T_0) - KT \tag{2.14}$$

Where T_0 is 300K^{o}C, M is the mobility-temperature exponent and K is the threshold voltage-temperature coefficient. Superthreshold circuits experience increased delays for high temperatures because the decrease in mobility dominates while subthreshold circuits experience higher delays for low temperatures because the decrease in threshold voltage dominates. At higher temperatures the delay of subthreshold circuits decrease, but leakage current increase with for low V_{th}.

Process variations are caused by inaccuracies in the manufacturing process and can be divided into global and local process variations.

Global Process Variations

Global process variations are equal all over the produced die like wafer-to-wafer misalignment or temperature resulting in variations in the threshold voltage. Ideally these process variations affect every transistor in the system by an equal amount.

Local Process Variations

Local process variation affect different parts of a die or circuit. The cause of these effects can be systematic or random in nature. Some sources are aberrations in the processing equipment or random dopant fluctuations where the placement and concentration of doping atoms have an inherently random deviation.

2.4 Tools, Simulation and Analysis

This section will explain some of the tools and simulation methods used in this thesis.

2.4.1 Parametric Analysis

When designing a circuit it is beneficial to see the effect of adjusting various design parameters affect the performance of the circuit. Sweeping over a parameter is called parametric sweep and can be performed in many circuit simulators like Cadence Analog Design Environment (ADE). An example could be to evaluate the VTC of an inverter by sweeping over the sizes of the NMOS and PMOS transistor. The simulation time increases for every swept variable and number of steps. If the simulations are complex the parametric analysis can become bery computationally intensive if care is not taken when setting up the analysis.

2.4.2 Corner Simulations

Process corners are simulation files that applies a specific statistical case of process variations to the transistors in a circuit. Corner models are usually divided into three even corners and two uneven corners. Typical-typical (TT), fast-fast (FF) and slow-slow (SS) are the even corners where both the NMOS and PMOS transistor are equally affected by variation. In the uneven fast-slow (FS) and slow-fast (SF) corner one transistor is affected more than the other. The first letter refers to the NMOS transistor and the second letter refers to the PMOS transistor. In the FS corner the NMOS exhibits fast behavior and the PMOS exhibit slow behavior. Corner simulations does not take mismatch between individual transistors into account, but corner simulations are useful for to show the effect of process variations in computationally intensive simulations on large and complex systems.

2.4.3 Monte Carlo Analysis

To verify the robustness and performance of a circuit it is necessary to simulate it with statistical models to check if the circuits performance with process variations and transistor mismatch. When simulating a circuit with Monte Carlo analysis a set of statistical files for the process can be applied to verify the circuit in various process corners and with both random process variations and mismatch. The variations comes in the form of a variation on the transistor parameters that represent the errors in manufacturing and wafer production. The yield target is usually 3-6σ, meaning that if $\mu \pm 3 - 6\sigma$ meets the design specification the yield of the circuit 99.86-99.99%. Ideally every simulated parameter should fit into a Gaussian distribution like in Fig. 2.13, but depending on the simulated parameter the amount of Monte Carlo iterations can vary from several thousand to achieve the ideal distribution. More iterations results in more accurate results.

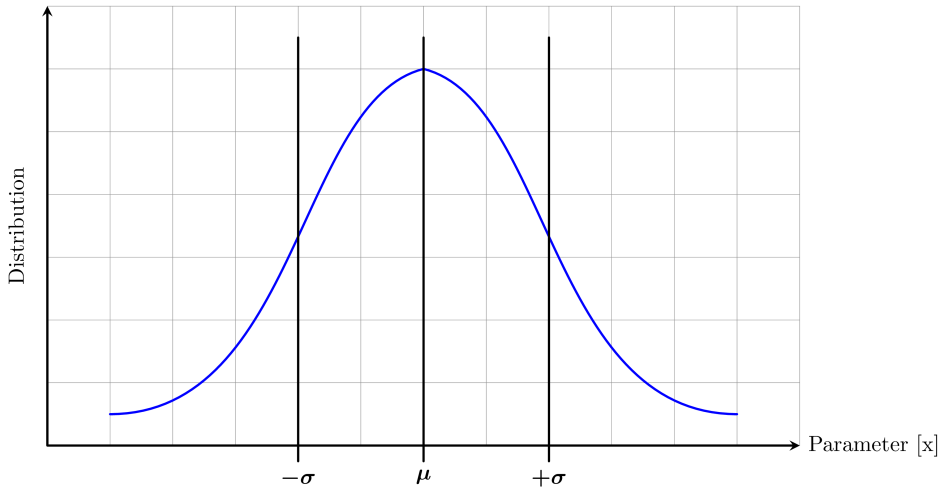


Figure 2.13: Gaussian distribution

A common method of reducing simulation time is to run the simulation until the changes in the standard deviation σ becomes minuscule for each new iteration and set the number of Monte Carlo iterations to that number where σ starts settling. In this thesis this number was found to be ≈ 100, but to add some precision the amount of Monte Carlo iterations was set to 200. Monte Carlo simulations in this thesis were performed in Cadence ADE XL with statistical process data from Atmel Norway AS.

2.4.4 Parasitic Extraction

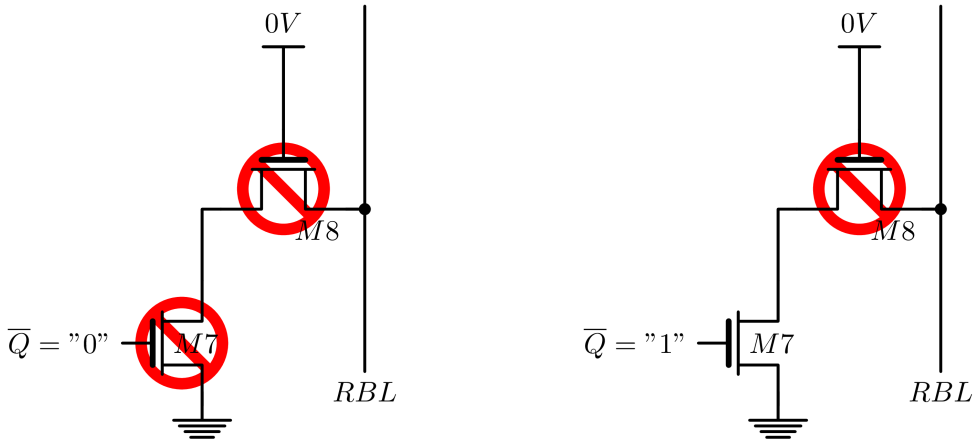
Simulating directly on the schematic ignores the added parasitic resistance and capacitance that is introduced by the physical layout. When a physical layout has been made in Cadence Layout XL it must be verified by Mentor Graphics Calibre design rule check (DRC) and layout versus schematic (LVS) check to confirm that the layout follows the specified design rules and that there are no discrepancies between the layout and the schematic. Once the layout passes both DRC and LVS the parasitics can be extracted with Mentor Graphics Calibre parasitic extraction (PEX) and included in previously defined testbenches. Parasitic extraction can be performed with xRC for 2D extraction and XACT3D for 3D extraction. The latter is more accurate but require tremendous computational power to use in Monte Carlo simulations so xRC were used.

3. Memory Architecture

This chapter will present some of the desicions made when deciding the general parameters of the SRAM architecture.

3.1 Choice of SRAM cells

In an earlier project for Atmel Norway AS[4] a litterature study was performed in order to evaluate SRAM cells at subthreshold voltages and a 8T SRAM cell with a gated-read buffer was evaluated at $V_{DD} = 350$mV. The 8T cell was found to be more robust at low voltages because the gated-read buffer decouples the read and write operation and the disturb voltage caused by the voltage divider in conventional 6T cells does not occur. The conflicting sizing requirements for the SRAM cell becomes a non-issue, but the writeability requirements of equation 2.5 still applies. The 8T cell had an average read SNM of 156mV which was a huge improvement over 50mV with a similarly sized conventional 6T cell. The amount of SRAM cells connected to a single bitline was heavily reduced because of the nature of the gated-read buffer as shown in Fig. 3.1. The leakage current of the buffer varies with the data stored in the cell and the worst-case leakage occurs when storing a logic ”0” as the NMOS transistor in the buffer is turned on.

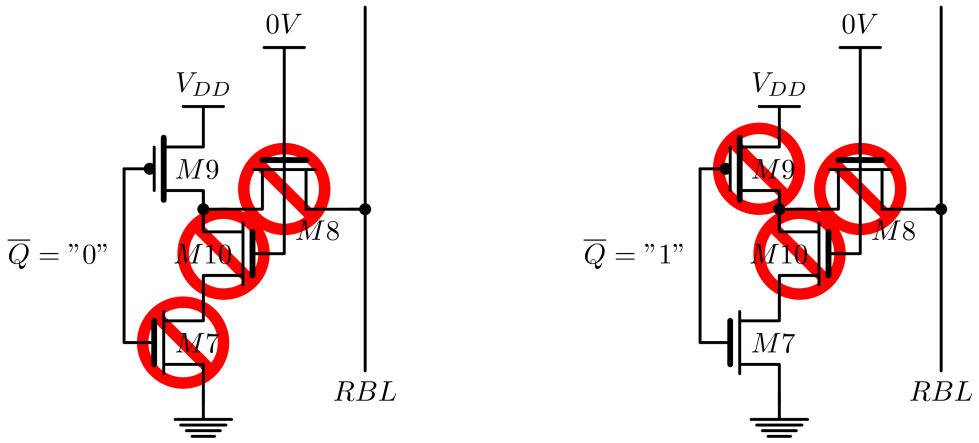


(a) Holding logic ”1” **(b)** Holding logic ”0”

Figure 3.1: 8T read buffer holding both logic levels

The leakage currents from the 8T cells deteriorated the on/off-current ratio and the bitline length became limited to 10-100 cells depending on the temperature and contents of the cells. To increase the on/off-current ratio a 10T SRAM cell with a 4T gated read buffer was chosen instead[9]. The 10T read buffer operates similarly to the 8T read buffer in Fig. 3.1, but the addition of $M9$ and $M10$ exploits the stack effect to reduce the leakage

current and ensures approximately the same leakage whether the cell stores a logic ”1” or logic ”0” as shown in Fig. 3.2.



(a) Holding logic ”1” **(b)** Holding logic ”0”

Figure 3.2: 10T read buffer holding both logic levels

The disadvantage of this read buffer is that the stack effect reduces the drive current as well (see Table 2.2), but the benefit of consistent on/off-current ratio was deemed more important than speed. The single-ended nature of the read operation also means that sense amplifiers can not be used with the 10T cell unless one of the SA inputs is connected to a stable and carefully chosen reference voltage[15]. The simplest reading method for single ended SRAM cells is to use an inverter for sensing, but this is slower because the bitline must discharge for a longer period of time to correctly sense data.

Simulation results of the 6T and 10T cells showed that the read SNM of the 6T cell is ≈ 67% of the 10T read SNM at V_{DD} = 400mV. Considering the differential nature of the 6T read operation the potential speed increase when using 6T cells makes them a lucrative candidate if the reduced SNM is within acceptable margins. The 6T cell is also 39% smaller than the 10T cell and consumes ≈ 32% less leakage power.

3.2 Write-Assist

PVT variations can causes write failures to occur at subthreshold voltages because the access transistors of the cell becomes too weak to overpower the cross-coupled inverters. Increasing the size of he access transistors improves writeability according to equation 2.5, but considering the area constraints on memory cells other write-assist methods are usually preffered. Boosting the voltage of the wordline signal increases the strength of the access transistors by increasing its V_{GS} (see equation 2.6)[16]. Another solution is to decreases the supply voltage of the cross-coupled inverters during write operations [15]. The downside of the latter approach is that a separate V_{DD}-connection is required for every SRAM row, but no complex circuitry like charge pumps or additional voltage references are needed to generate the boosted voltages.

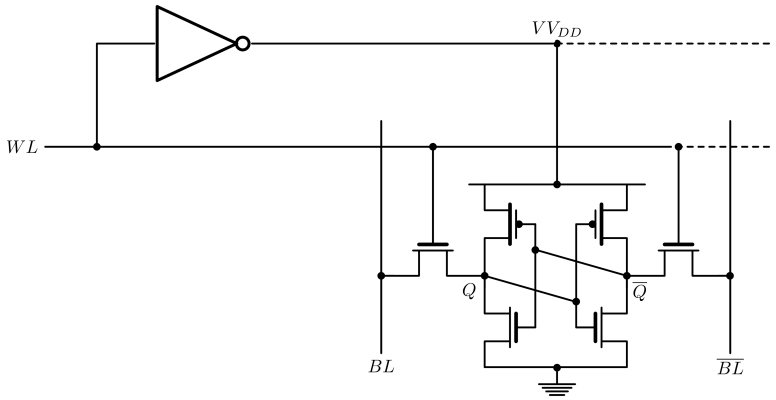


Figure 3.3: VV_{DD} write-assist method

The virtual-V_{DD} (VV_{DD}) write-assist method shown in Fig. 3.3 was chosen because of its simplicity. Using charge pumps or voltage references to generate voltages $\geq V_{DD}$ introduces a potential problem with voltage scaling as the circuit generating the boosted voltage levels must be turned off when increasing the supply voltage to superthreshold levels or the transistors might be destroyed. The write path is equal in 10T and 6T cell so the same write-assist is used with both cells.

When the VV_{DD} output is lowered towards ground potential the voltage in the internal SRAM nodes will be lowered, but positive feedback will ensure the node holding a logic "1" will settle somewhere between ground and V_{DD} and new data can be written more easily. When a write operation is complete the VV_{DD} output is reset to V_{DD} and the positive feedback in the SRAM cell pulls the logic "1" node back to proper logic levels as Fig. 3.4 illustrates.

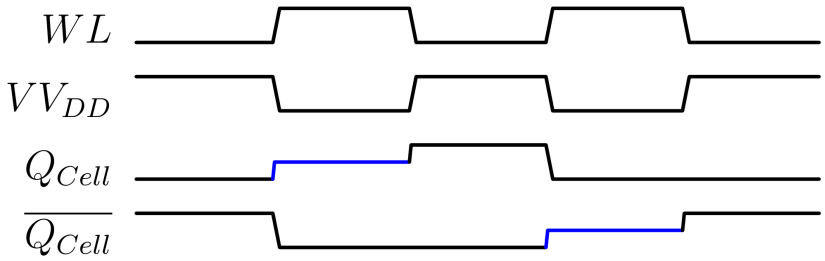


Figure 3.4: Wordline driver write operation

3.3 Asynchronous SRAM

Asynchronous SRAMs respond to changes on the address or other inputs to generate a self-timed internal "clock"-signal that controls the read and write circuitry. Synchronous SRAMs use a periodic clock to start a operation and outputs data on a subsequent clock edge. Synchronous SRAMs can be pipelined to increase throughput and many support burst mode read and write operations to increase performance[17]. Asynchronous SRAMs are generally slower compared to synchronous SRAMs, but since no clocks are used there are no problems associated with clock distribution and clock skew and the

switching power of clocked latches or registers are reduced[18].

Synchronous SRAMs require that data is available on the outputs of the SRAM circuitry after one or more clock cycles. If data fails to arrive because the bitlines were not able to discharge fast enough or if some other component causes slowdown a read failure will occur. Delays increase at subthreshold voltages and the exponential dependencies between the environment, transistor parameters and transistor drive current mean that the amount of clock cycles needed to complete the read and write operations must be set pessimistically low, especially at low temperatures. Current solutions used by Atmel Norway AS are asynchronous so the ultra-low voltage SRAM architecture was implemented as an asynchronous architecture as well.

3.4 Replica-Based Self-Timing

The read circuitry must be enabled when the bitlines have discharged to an acceptable level so that the read circuitry can sense correct data. The sensing operation should be as short as possible to increase overall throughput and to reduce power consumption. A common solution in the past have been to use a chain of inverters to match the bitline discharge delay[18] , but the major drawback of this method is that matching the delay of an inverter chain to the bitline discharge of SRAM cells is not trivial and is most likely more difficult at subthreshold voltages because of PVT variations. Most modern Asynchronous SRAMs use a replica bitline for self-timing as two bitlines within the same array should ideally be equally affected by PVT variations[19].

A redundant row and column are added to the SRAM array and a set number of SRAM cells in the replica column stores a logic ”0” and discharges together with the actual bitline during a read operation. Because of the redundant row the delay of the dummy wordline should ideally track the delay of the actual wordline. An inverter is used to sense the replica bitline voltage and activates the read circuitry when the replica bitline passes the trip voltage of the inverter. Fig. 3.5 illustrates the replica bitline concept.

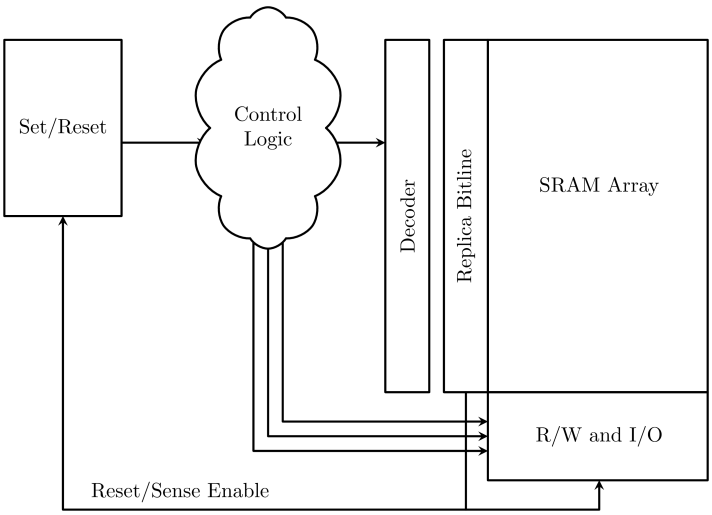


Figure 3.5: Replica bitline loop

The amount of dummy cells discharging the replica bitline must be set to achieve the desired discharge delay. The discharging current of the replica bitline must be set according to equtaion 3.1[18].

$$I_{replica} = \left(\frac{V_{inv-trip}}{V_{SA-diff}} \right) I_{bitline} \tag{3.1}$$

If the trip voltage of the inverter is $V_{DD}/2$ and the differential trip voltage of a sense amplifier is $V_{DD}/10$ the discharging current of the replica bitline must be five times that of the actual bitline. The unaccessed cells connected to the replica bitline and redundant row are static and have their internal nodes fixed to a logic value.

The replica timing of the 10T SRAM architecture is simple as the replica bitline only need to discharge equally with the actual bitlines. The 6T architecture is more difficult because of the crucial aspect of the SA signal timing. Mismatch between the replica bitline and actual bitline can cause the SA to activate prematurely and cause a erroneous read operation. To reduce variability in the 6T bitline matching the single replica bitline was replaced with a dual replica bitline[20]. Both configurations are shown in Fig. 3.6.

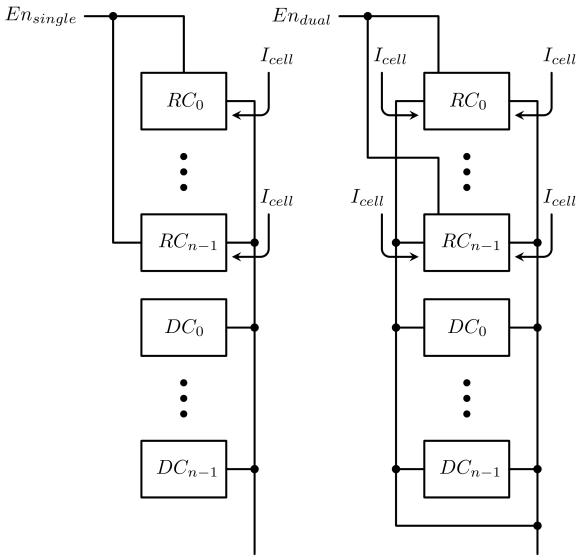


Figure 3.6: Single and dual replica bitline configurations

Where RC are the discharging cells and DC are locked cells. The replica cells discharge both bitlines towards ground potential and both bitlines are connected to a sensing inverter. The configuration effectively doubles the bitline capacitance, but also doubles the discharging current resulting in the same delay as the single bitline configuration as shown by equation 3.2.

$$\tau_{replica} \approx \frac{2C_{BL}}{2nI_{cell}} = \frac{C_{BL}}{nI_{cell}} \tag{3.2}$$

Simulation results for this configuration indicate that the read cycle time deviation in a 65nm process at V_{DD} = 500mV is reduced by 50% compared to a single bitnline configuration[20].

3.5 SRAM Control

The asynchronous SRAM control must detect a change on one or more of the inputs to the SRAM, store the information that an operation is currently in progress and wait for a reset condition to return the entire system to an idle state. The conventional address transition detector (ATD) in Fig. 3.7 detects changes on the address inputs and generates a pulse on the output that is used to initiate an operation. The pulse width is set by the delay element before the XOR gates which can be a chain of inverters or a large capacitive load.

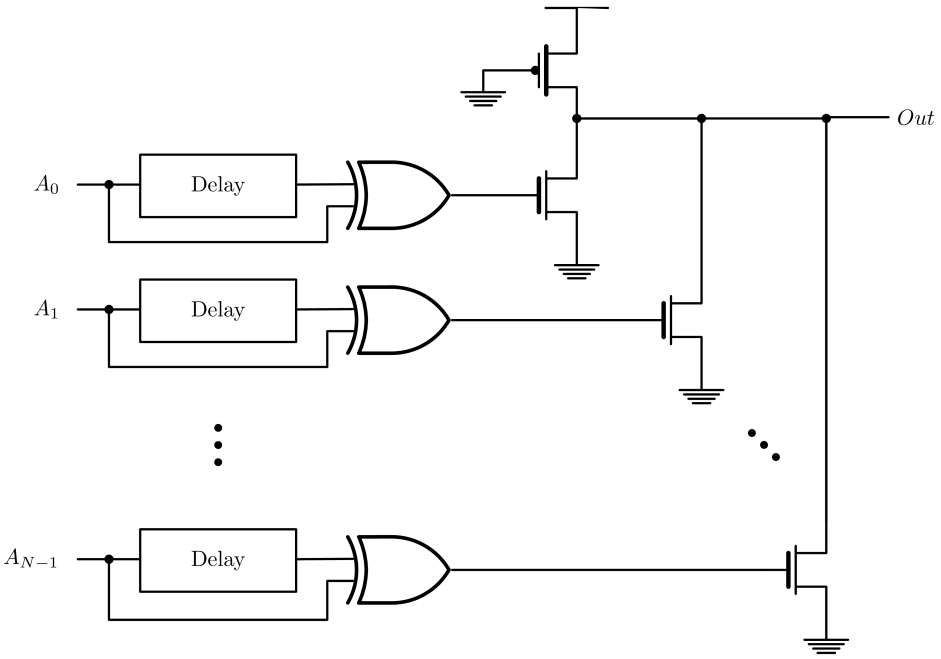


Figure 3.7: Address Transistion Detector

The ATD is in essence a pseudo-NMOS NOR gate where a PMOS transistor keeps the output high as long as none of the NMOS transistors have a logic ”1” on their inputs. During simulations it was discovered that this circuit encountered a similar problem as the XOR gate in Fig. 2.10 where the leakage current of the NMOS transistors prevented the PMOS transistor to pull the output to logic ”1” after a address transition. To reduce the fan-in the simpler transition detector (TD) in Fig. 3.8 was designed to detect a positive edge on the input of a single enable signal.

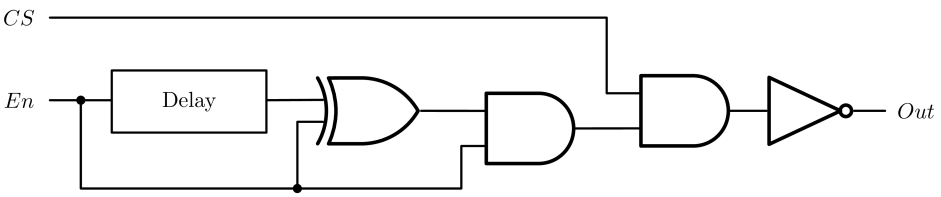


Figure 3.8: Transistion Detector

The output of the transition detector is applied to the clock inputs of a latch with an asynchronous reset input that serves as a storage register for a flag that indicates an operation is in progress. The schematic of the latch is shown in Fig. 3.9.

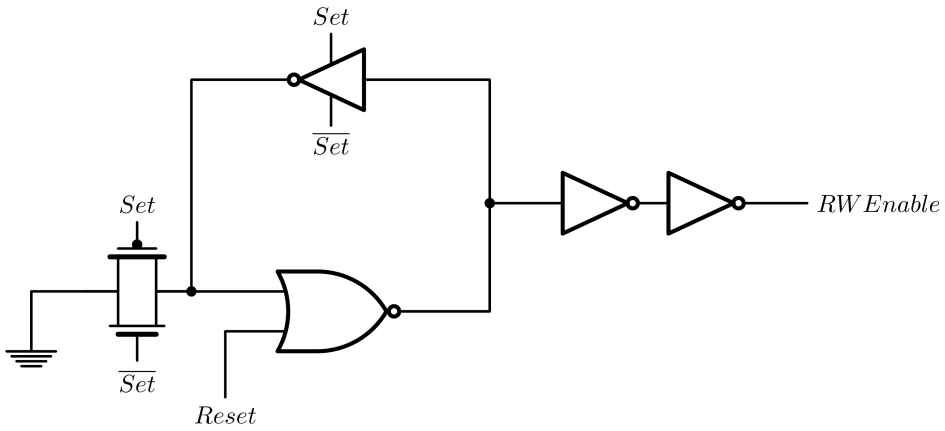


Figure 3.9: Enable Register

When the pulse generated by the TD is applied to the *Set*-inputs the feedback inverter is turned off and the NOR gate propagates a logic "1" to the output. When the read or write operation is complete the reset input is set and the NOR gate propagates a logic "0" to the output. Since the input is tied to ground potential the setup and hold time is not an issue compared to a latch with a variable data input.

Fig. 3.10 shows the entire SRAM control circuit. A multiplexer (MUX) after the TD output has its select signal connected to the *RWEnable* signal so that any pulses generated during an operation is ignored.

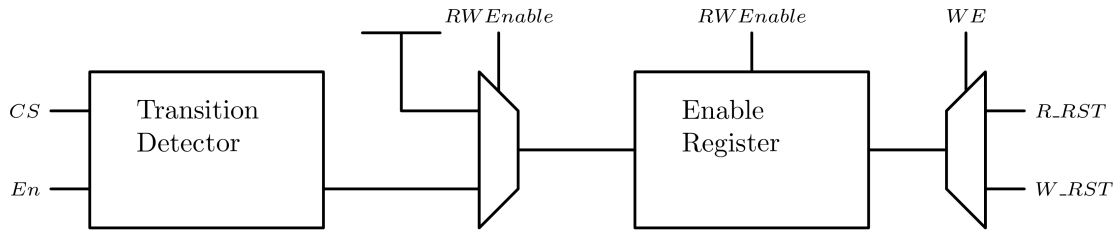


Figure 3.10: SRAM Control state machine

Depending on the write enabe (WE) input the second MUX propagates either the read or write reset signal to the latch.

3.6 Decoder

The SRAM architecture is divided into SRAM blocks with a set number of rows consisting of of 32-bit words. Each block has its own block select signal and a word is only selected when both the block select and it's decoded address are asserted. This is called the divided wordline (DWL) architecture[21]. Only one block is selected at a time so the delay and power consumption of the word-select operation is reduced because the capacitive load seen from the decoder is smaller compared to a unified wordline architecture. The concept is shown in Fig. 3.11. The decoding is performed by a multi-stage decoder. In these decoders the address input are pre-decoded in the first decoder stages and shared by the

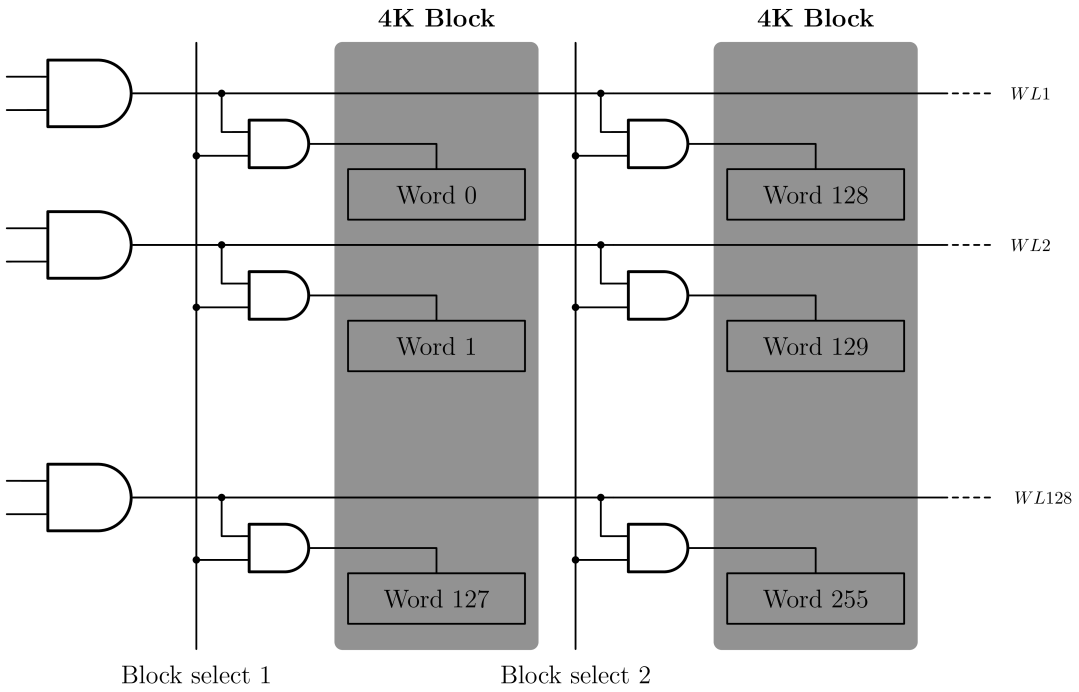


Figure 3.11: Divided Wordline Architecture

subsequent decoder stages. The amount of address inputs needed to decode all words is calculated with equation 3.3.

$$N_{decoder} = log_2 \left(\frac{\text{Array size in bits}}{\text{Word length in bits}} \right) \tag{3.3}$$

Using a 8k memory consisting of of two 4K SRAM as an example: In order to decode all 256 memory locations 2 block select signals and $log_2\ 128 = 7$ address inputs are needed. Without the decoder 256 inputs would be needed and the amount would reach several thousand or even million for larger megabyte-sized memories.

Table 3.1 shows the decoding scheme for the 8K SRAM decoder. The decoder uses three predecoders to decode 7 address inputs. Two 2-to-4 predecoders and one 3-to-8 predecoder are used to create a larger 8-to-16 predecoder. The 7-to-128 decoder is shown in Fig. 3.12.

Block Select	Address	Word	SRAM Block
01	0000000	0	1
01	0000001	1	1
01	...	...	
01	1111111	127	1
10	0000000	128	2
10	...	..	
10	1111111	255	2

Table 3.1: SRAM architecure decoding scheme

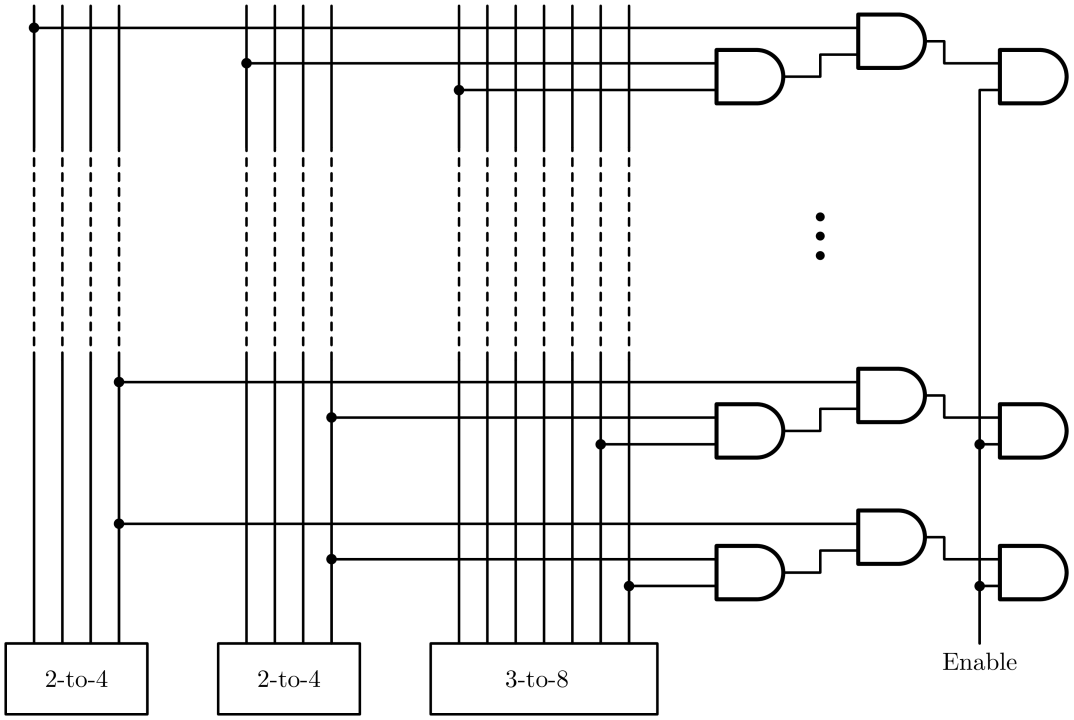


Figure 3.12: 7-to-128 decoder

.

The outputs of the predecoders are connected together with AND gates to form the total 7-to-128 decoder. Evey output is paired with the *RWEnable*-signal from the enable register with another set of AND gates.

3.7 Architecture Operation

This section will explains how the read and write operations function in detail.

3.7.1 10T Read Operation

Fig. 3.13 shows a simplified overview of the 10T SRAM architecture. Some driving buffers and logic gate chains have been omitted to simplify the figure.

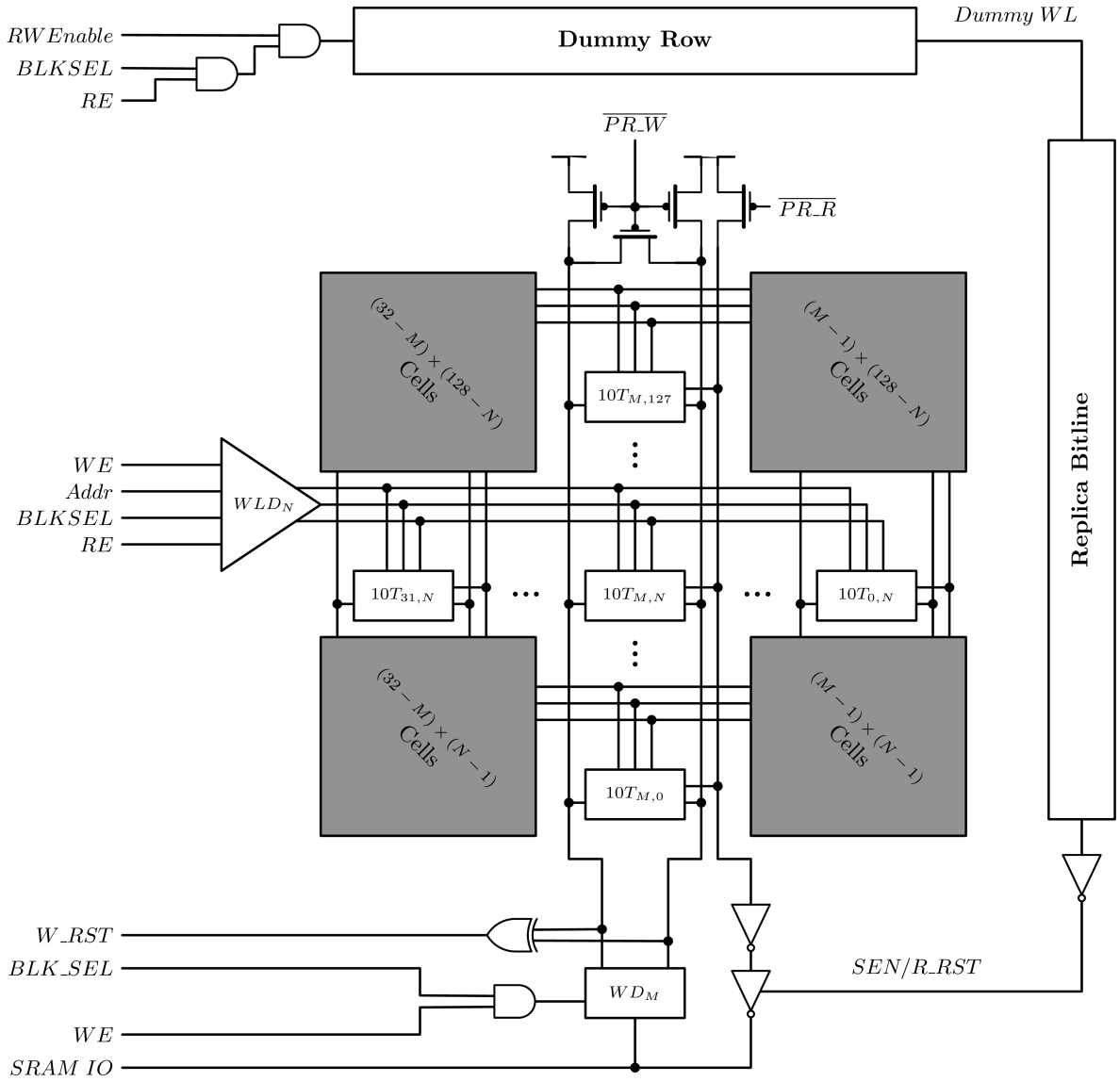


Figure 3.13: 10T architecture overview

The read operation starts when a positive edge is applied to the input of the TD. It is assumed that the address inputs, block select and RE/WE-inputs are set prior to the arrival of the positive edge. When the output pulse of the TD is logic "0" the latch will propagate a logic "1" to the output and activates the *RWEnable*-signal which remains at logic "1" until the replica biltine activates the reset signal. Fig. 3.14 shows a simplified timing diagram for the SRAM control.

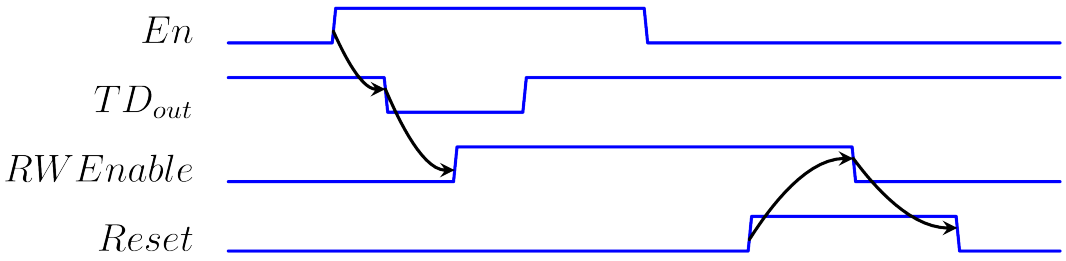


Figure 3.14: SRAM control timing diagram

The *RWEnable*-signal propagates the output of the address decoder to the wordline drivers and activates the read wordline RWL of the selected memory address at the same time as the dummy wordline. The precharge circuits are controlled by the following boolean functions:

$$\overline{PR\_R} = \overline{(RWenable \vee SEN)}$$
$$\overline{PR\_W} = \overline{(RWenable \vee WE)}$$

When the precharge circuits turn off the bitlines starts discharging if a logic "0" is stored in the accessed cell or the bitline remains at V_{DD} if a logic "1" is stored. When the replica bitline has discharged past the tripping point of the sensing inverter the SEN/R_RST-signal is activated and the SRAM control circuit then resets the system to an idle state. Fig. 3.15 shows a simplified timing diagram for the 10T read operation.

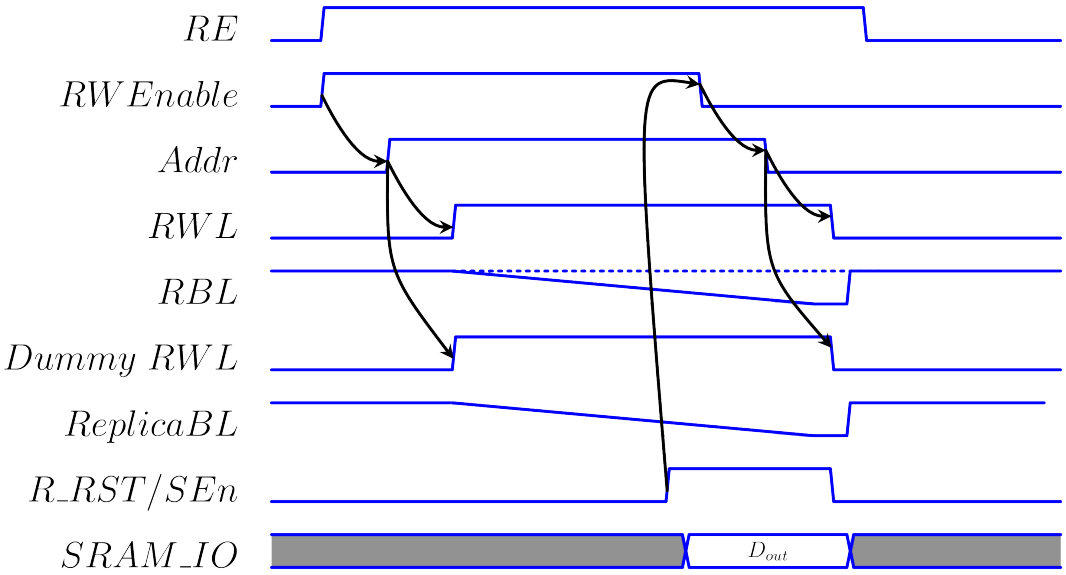


Figure 3.15: 10T read timing diagram

Latches connected to the outputs of the IO bus activates with the SEN-signal and stores the data on the output until new data is read.

3.7.2 6T Read operation

Fig. 3.16 shows a simplified overview of the 6T SRAM architecture. Some driving buffers and logic gate chains have been omitted to simplify the figure.

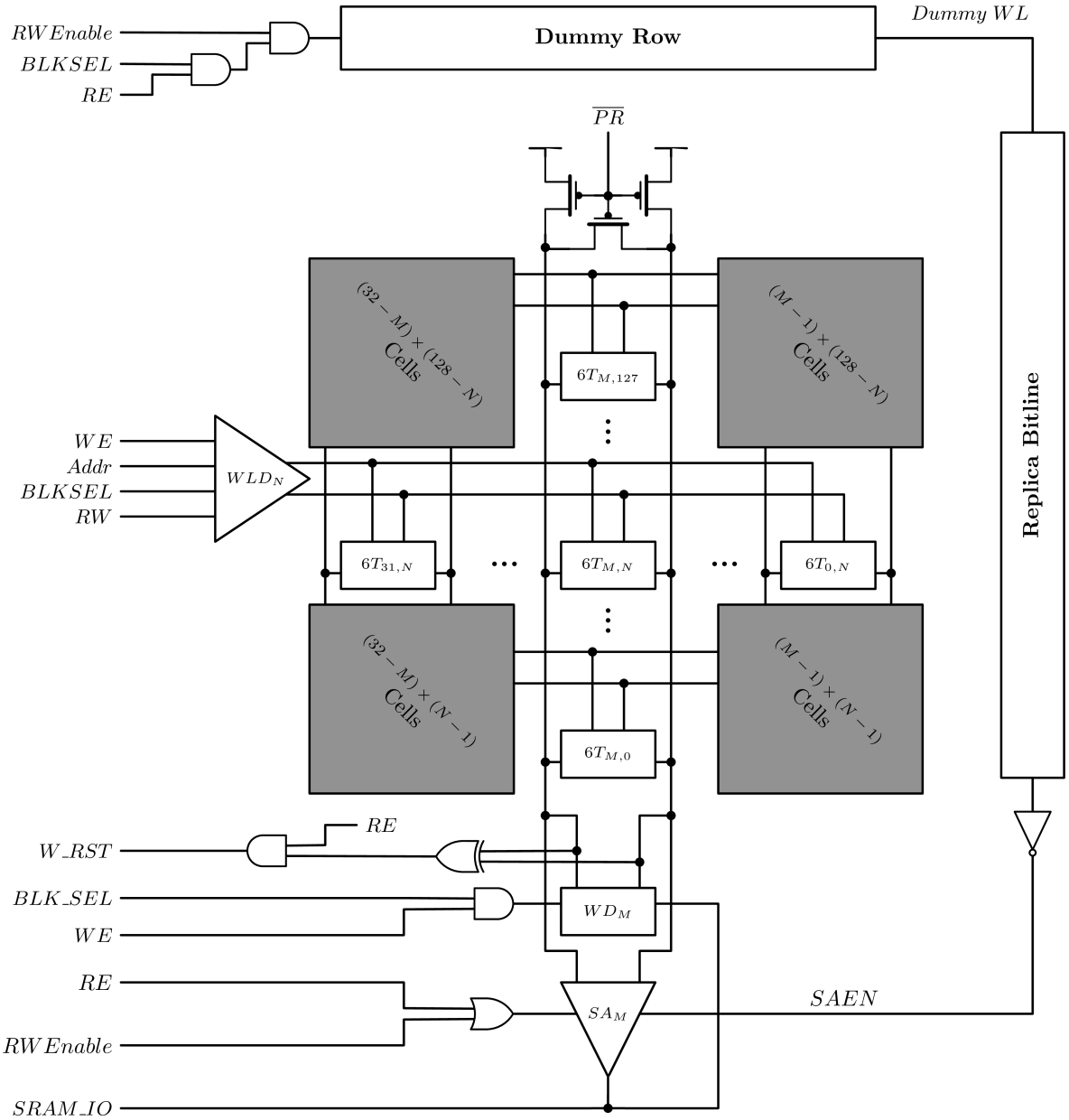


Figure 3.16: 6T architecture overview

The read operation starts in the same way as in the 10T architecture with the TD and enable register generating the *RWEnabl*-signal. The *RWEnable*-signal propagates the output of the address decoder to the wordline drivers and activates the wordline WL of the selected memory address at the same time as the dummy wordline. The precharge circuits are controlled by the following boolean function:

$$\overline{PR} = \overline{[(RWenable \vee W\_RST) \vee RWEnable]}$$

The signal RW is controlled by the following boolean function:

$$RW = WE \vee RE$$

The RW-signal is used to activate the wordline drivers for both read and write operations. Once the replica bitline has discharged past the tripping point of the sensing inverter the $SAEN$-signal activates tha SA by turning on its footer transistor and turning off the reseting PMOS transistors. The differential bitline voltage ΔV_{BL} is sensed by the SA and the SRAM control circuit is reset to an idle state. Fig. 3.17 shows a simplified timing diagram for the 6T read operation.

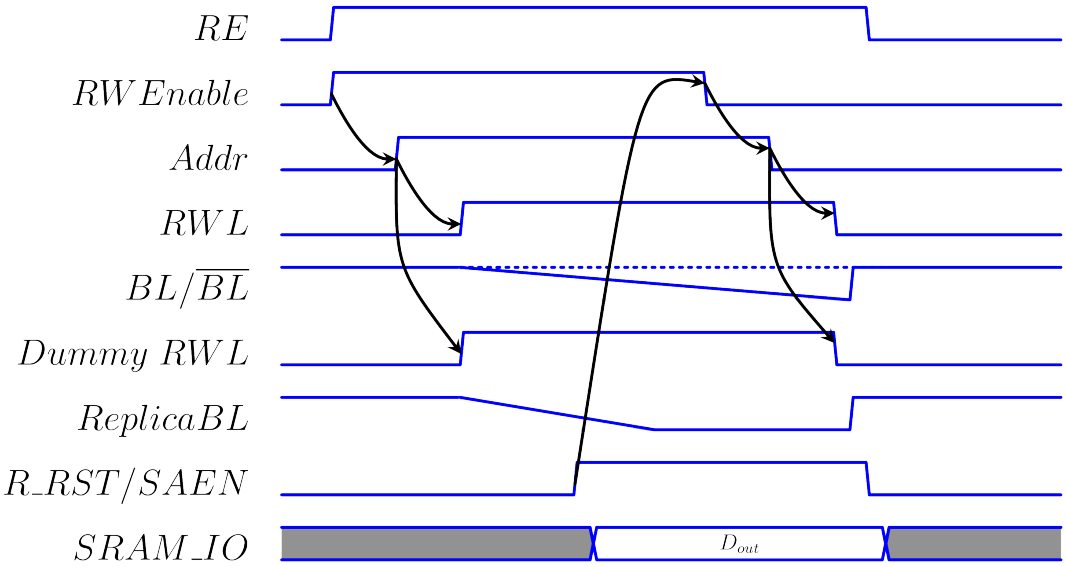


Figure 3.17: 6T read timing diagram

Latches connected to the outputs of the IO bus activates with the $SAEN$-signal and stores the data on the output until new data is read

3.7.3 Write Operation

The Write operation is almost identical for 10T and 6T cells, but because of more complex precharge control and the generation of the RW-signal the propagation delay in the 6T architecture will be slightly longer. Data is applied to the IO bus and the write operation is started the same way as the read operation in Fig. 3.14, but with the WE-signal set to V_{DD} instead of the RE-signal. The wordline/write wordline is raised and the VV_{DD}-output of the selected wordline driver is driven towards ground potential. The write driver forces the bitlines to complementary logic values and new data is written to the internal nodes of the selected SRAM cells. Fig. 3.18 shows a simplified timing diagram for the 10T and 6T write operation.

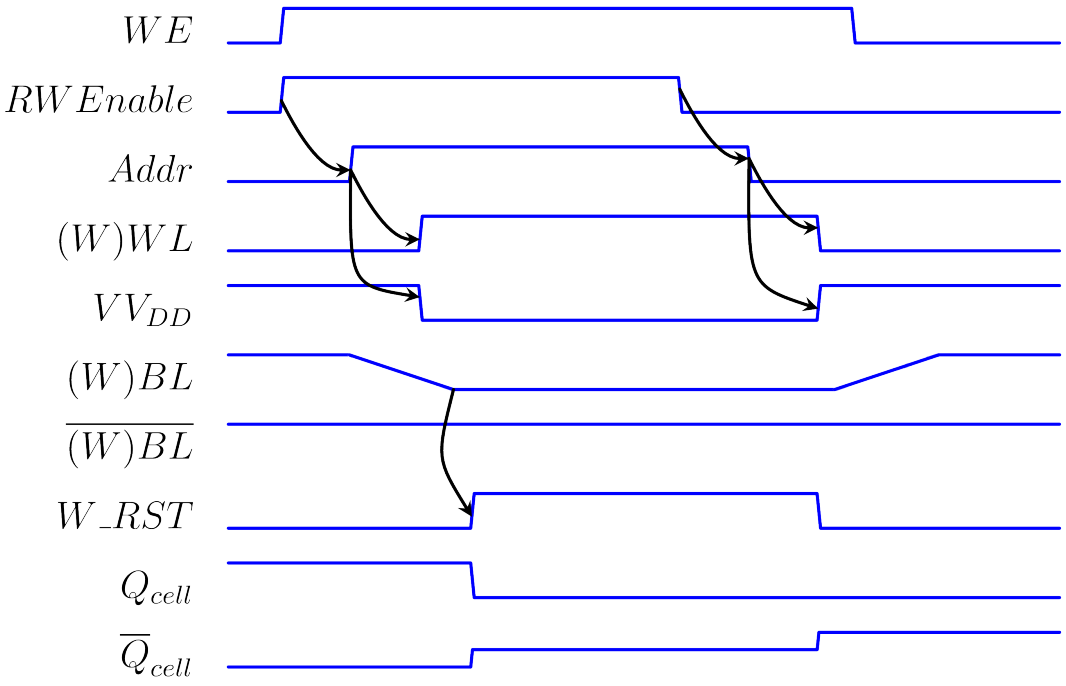


Figure 3.18: Write ”0” timing diagram

When the bitlines have reached complementary logic values a XOR gate will activate and propagate the reset signal to the SRAM control which resets the system to an idle state.

4. Architecture Components

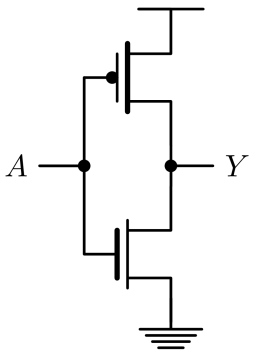
This chapter presents the components used in the memory architecure.

4.1 Logic Gates

No logic gate library for low-voltage operation exists at Atmel Norway AS so a small subset of logic gates had to be designed for the SRAM architecture. The transistors sizes were chosen by evaluating the balance, performance and area of the logic gate in a process flow that is further explained in chapter 5.

4.1.1 Inverter

The conventional 2T CMOS inverter is used to generate AND, OR and XOR logic when connected in series with other logic gates. 2T inverters are also used for driving large capacitive loads because the inverter has no stacked or parallel devices degrading the logic gate balance or reducing the driving strength.



A	Y
0	1
1	0

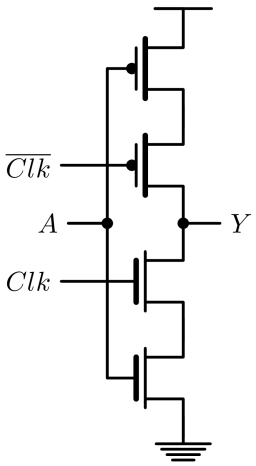
Figure 4.1: 2T CMOS inverter

Transistor sizes:
$(W/L)_n = 160nm/520nm$, $(W/L)_p = 160nm/260nm$

Simulations showed that to increase driving strength it was more area-effective to increase length and m-factor of the transistors instead of increasing the width. Increasing the m-factor means connecting m transistors in parallel. The reverse narrow-channel effect (RNCE) in the transistor caused a reduction in drive strength for $W = W_{min} \rightarrow 10W_{min}$ which is consistent with previous research[11]. Driving inverters with m-factor 2 and 4 were constructed to drive wordlines, bitlines and other signals with large capacitive loads.

4.1.2 Gated Inverter

The 4T CMOS gated inverter only inverts the input signal when the clock/enable signal is set. Stacked transistors results in reduced driving strength and speed compared to the 2T non-gated inverter, but variability caused by PVT variations are reduced [22].



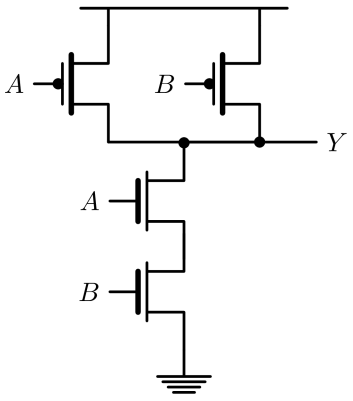
Clk	A	Y
0	X	Z
1	0	1
1	1	0

Figure 4.2: 4T CMOS gated inverter

Transistor Sizes:
$(W/L)_n = 160nm/520nm$, $(W/L)_p = 160nm/260nm$,

4.1.3 NAND Gate

The conventional 4T NAND gate will exhibit worse balance and sped at subthreshold voltages because of parallel and stacked transistors in the PUN and PDN. 8T NAND gates have been shown to be more robust at low voltages [22], but the NAND gate is usually the most utilized logic gate after the inverter (because of it's functional completeness) so the use of 4T NAND gates can greatly reduce the area of a system.



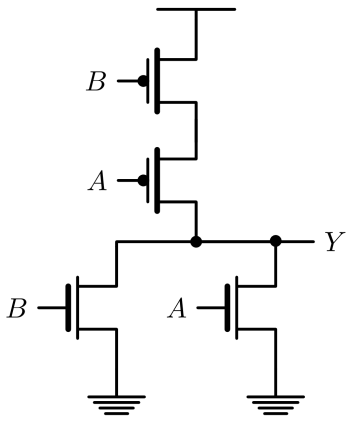
A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

Figure 4.3: 4T CMOS NAND gate

Transistor Sizes:
$(W/L)_n = 160nm/650nm$, $(W/L)_p = 160nm/650nm$,

4.1.4 NOR Gate

The conventional 4T NOR gate also faces the same problems as the conventional 4T NAND gate in terms of stacked and parallel devices. Depending on whether the PMOS or NMOS transistor is the stronger transistor in a given process the conventional NAND or conventional NOR gate will exhibit the worst-case balance.



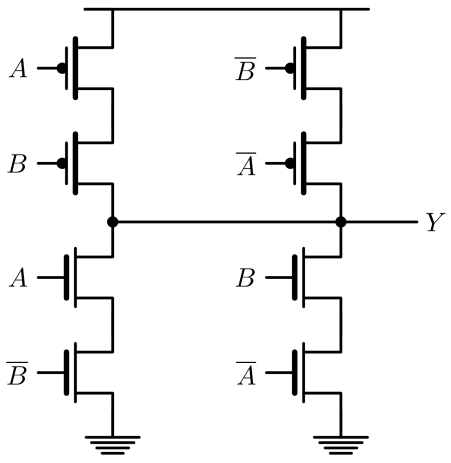
A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0

Figure 4.4: 4T CMOS NOR gate

Transistor Sizes:
$(W/L)_n = 160nm/520nm$, $(W/L)_p = 160nm/260nm$.

4.1.5 XNOR Gate

All branches in the 8T XNOR gate has equal amounts of stacked and parallel transistors meaning which results in less variability at subtreshold voltages compared to the NAND and NOR gate. Since transistors are stacked in both the PUN and PDN of the gate driving strength is reduced for all input combinations.



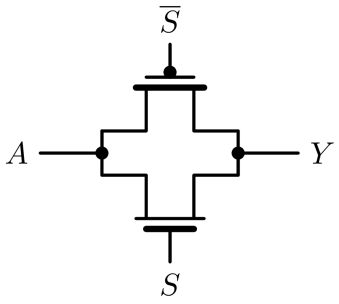
A	B	Y
0	0	1
0	1	0
1	0	0
1	1	1

Figure 4.5: 8T CMOS XNOR gate

Transistor Sizes:
$(W/L)_n = 160nm/520nm$, $(W/L)_p = 160nm/260nm$.

4.1.6 Transmission Gate

The transmission gate is a bidirectional switch that conducts a logic value when the S-signal is activated. Transmission gates are preferred over single-transistor pass gates at subthreshold voltages because the transmission gate conducts equal amounts of current independently of the input and whether the PMOS or NMOS transistor is the stronger transistor in a given process[12]. The transmission gate is also used to create 2-to-1 multiplexers (MUX).



S	Y
0	Z
1	A

Figure 4.6: 2T CMOS transmission gate

Transistor Sizes:
$(W/L)_n = 160nm/520nm$, $(W/L)_p = 160nm/260nm$.

4.2 SRAM Cells

The 10T SRAM cell in Fig. 4.7 is a 6T SRAM cell with a gated-read buffer connected to the $\overline{Q}$ node. The cross-coupled inverters in both SRAM cells use the same sizes as the 2T inverter because the SNM of the cells is directly tied to the VTC balance. The read buffer in the 10T cell also uses these sizes because the stack effect reduces the on and off-currents by a factor of the smallest transistor in the stack. The access transistors are slightly smaller than the other NMOS transistors to fulfill the writeability requirements set by equation 2.5.

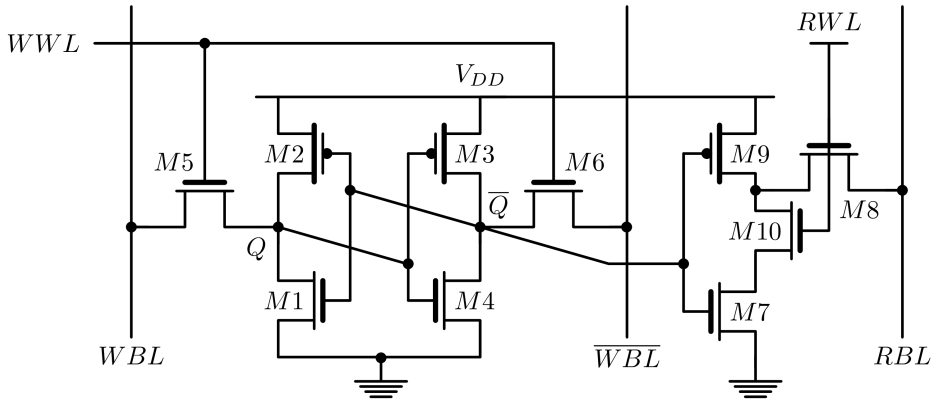


Figure 4.7: 10T SRAM cell with gated-read buffer

Transistor Sizes:
$(W/L)_n = 160nm/520nm$, $(W/L)_p = 160nm/260nm$, $(W/L)_{5,6} = 160nm/390nm$

4.3 Sense Amplifier

Fig. 4.8 shows a latch-based SA used with 6T cells [23]. The SA is similar to a 6T SRAM cell in that data is stored as complementary logic values by exploiting the positive feedback of cross-coupled inverters.

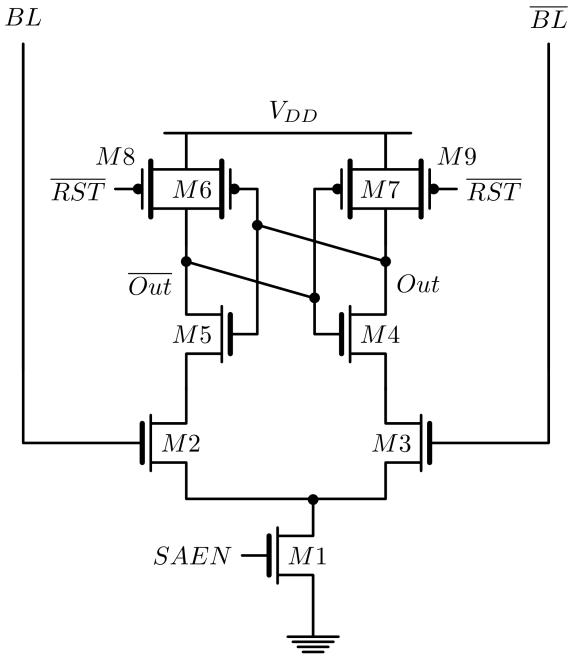


Figure 4.8: Sense Amplifier

When a differential input voltage is applied to the inputs the current drawn by the transistor connected to the discharging bitline ($M2$ or $M3$) will conduct less current than the other and when the difference in current becomes large enough the latch will set the internal nodes to the corresponding logic value. Ideally an infinitesimal differential voltage should be enough to engage the latch, but because of V_{th}-mismatch between the input and inverter transistors the differential bitline voltage must be larger than an offset voltage caused by the mismatch to ensure the read operation succeeds. At subthreshold voltages the offset voltage can be estimated using equation 4.1[24].

$$V_{os} = (\Delta V_{th2} - \Delta V_{th3}) + (\Delta V_{th4} - \Delta V_{th5}) \cdot 10^{-(V_{DD} - V_{INDC})/n} \tag{4.1}$$

Where V_{INDC} is the input DC voltage without ΔV_{BL} and n is the subthreshold slope factor. The read operation will not initiate until the footer transistor $M1$ is enabled. The timing of the $SAEN$-signal is crucial and it must not be activated before the differential bitline voltage has surpassed the offset voltage. $M1$ also reduces the leakage power consumption as the current in both branches of the SA flows through $M1$, which is turned off when not reading. To prevent the SA to act like a memory cell two PMOS transistors are activated with the $\overline{RST}$-signal after a completed read operation, forcing both outputs to V_{DD}.

Transistor Sizes:
$(W/L)_n = 520nm/130nm$, $(W/L)_p = 260nm/130nm$, $(W/L)_{8,9} = 160nm/130nm$

The resetting PMOS transistors are kept as small as possible to prevent extensive capacitive coupling on the output. The cross coupled inverters are sized the same as the 2T inverter because more balanced VTCs reduces the offset voltage. The footer transistor is also sized the same as the other NMOS transistors to minimize leakage[24]. Fig. 4.9 shows a sensing operation where a logic ”0” is read from a SRAM cell.

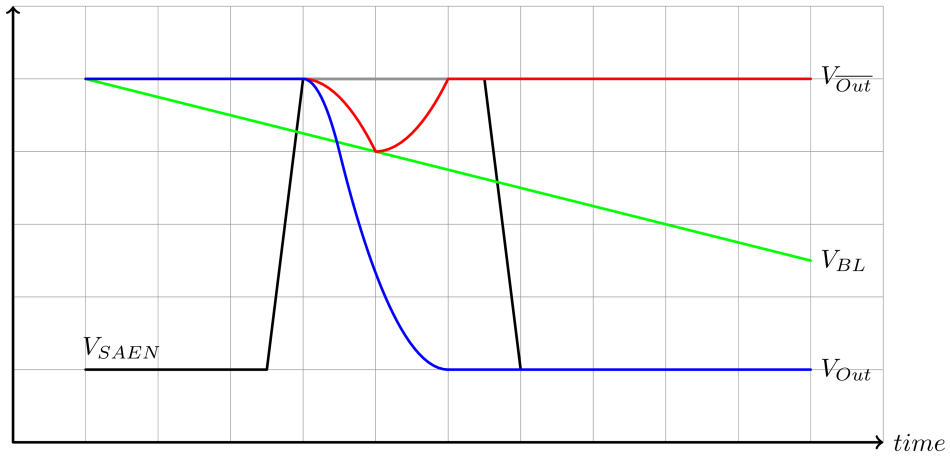


Figure 4.9: Sense amplifier read ”0” operation

4.4 Bitline Precharge

The bitlines are discharged by a single access transistor when reading, and it is easier to discharge the bitline with an NMOS transistor than to charge it. Precharging the bitlines to V_{DD} prior to, or between operations ensure that all bitlines are in a known state and reduces power consumption since the bitlines will not float. Fig. 4.10 shows a common circuit used to precharge bitlines. When the precharge signal goes low the two precharge transistors will pull the bitlines to V_{DD}. The middle PMOS transistor ensures equalization of the bitlines[18].

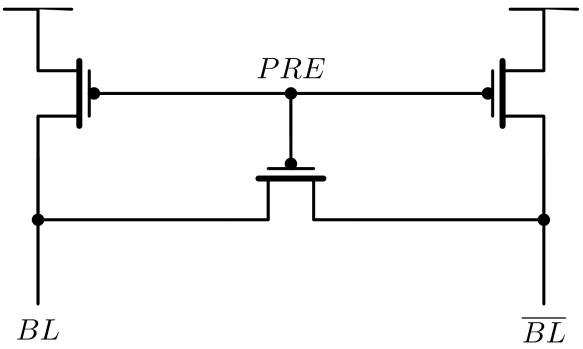
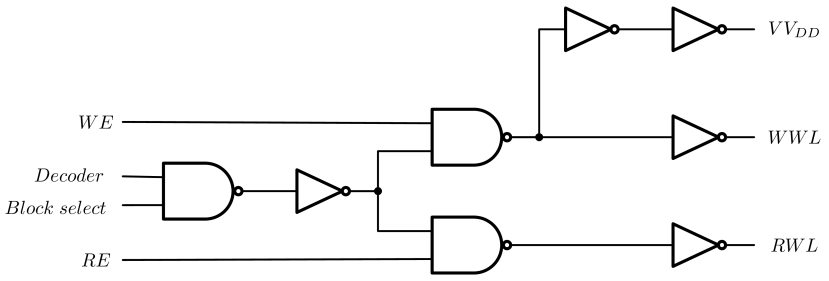


Figure 4.10: Precharge circuit

A single PMOS transistor is used for read bitline precharge with 10T cells as there is no need for equalization.

4.5 Wordline Drivers

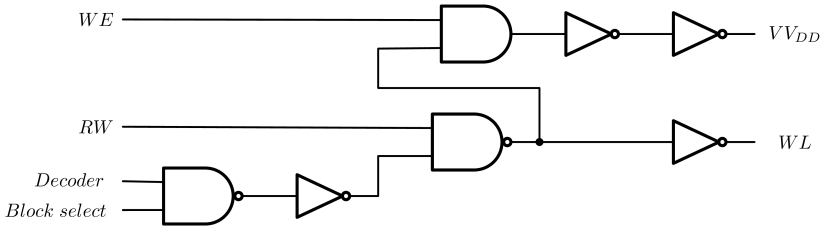
Fig. 4.11 shows the wordline driver for 10T SRAM cells and its truth table.



Block select	Decoder	WE	RE	VV_{DD}	WWL	RWL
0	X	X	X	1	0	0
1	0	X	X	1	0	0
1	1	0	1	1	0	1
1	1	1	0	0	1	0
1	1	1	1	Not Allowed		
1	1	1	1	Not Allowed		

Figure 4.11: Wordline driver for 10T SRAM

The read enable (RE) and write enable (WE) inputs decide if the read or write worldline is raised. During write operations the VV_{DD} output is also lowered. The output from the address decoder starts the operation after it has been activated with the $RWEnable$-signal. The block select signal prevents triggering of the wordline drivers in unaccessed SRAM blocks. The wordline driver for 6T SRAM cells shown in Fig. 4.12 is similar to the 10T driver except that it only drives one wordline. The RW-input is activated for both read and write operations.

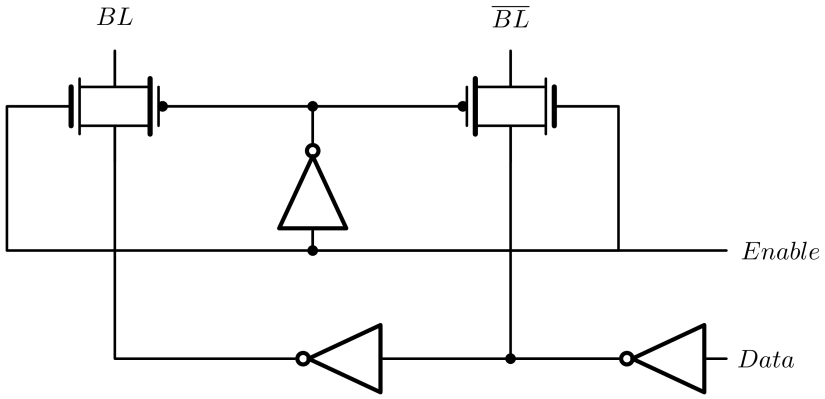


Block select	Decoder	WE	RW	VV_{DD}	WL
0	X	X	X	1	0
1	0	X	X	1	0
1	1	0	1	1	1
1	1	1	0	0	1

Figure 4.12: Wordline driver for 6T SRAM

4.6 Write Driver

The write driver forces the bitlines to complementary logic values during write operations. The driving portion of the circuit consists of two inverters and two transmission gates. A third inverter is used to generate the inverted enable signal for the transmission gates. Fig. 4.13 shows the write driver and its truth table[18].



Enable	Data	BL	$\overline{BL}$
0	X	Z	Z
1	0	0	1
1	1	1	0

Figure 4.13: Write driver

To prevent unnecessary switching the *Enable* signal is only asserted when the block select signal for the selected block is active. The bitlines amounts to a significant capacitive load so the output inverters must be sized accordingly. The 10T write operation is faster compared to the 10T read operation because the write operation is differential in nature while the read operation is single-ended and an inverter can discharge the write bitline faster then a 10T cells read buffer can discharge the read bitline. With 6T cells the read and write operations are more equal in terms of speed as both operations are differential.

5. Simulation Methodologies

This chapter presents the testing and simulation methods used to decide and verify device/architecture parameters. The circuit schematics and simulations were performed in Cadence Virtuoso. Statistical data and corner/mismatch models from Atmel Norway AS were used in Monte Carlo simulations and corner simulations in Cadence ADE XL.

5.1 Optimal Supply Voltage

The SRAM is arguably the most difficult component to implement for subthreshold voltages and will most likely set the minimum supply voltage V_{DD} for a system. The SRAM can become very slow at low voltages due to the large amount of capacitance on wordlines/bitlines and will most likely not work reliably at supply voltages below 300mV unless extensive read and write assist mechanisms are introduced. The optimal V_{DD} for logic gates is usually found by connecting the logic gate in a ring oscillator and extracting the power-delay product (PDP) which is a common figure of merit (FOM) for low power operation [25]. The PDP is useful for finding the minimum-energy per gate switching operation in a CMOS process, but it is less useful considering the SRAM due to its relatively low duty-cycle. A new FOM named the power-per-read-cycle product (PRCP) was introduced and it gives the minimum power-delay product for a SRAM read "0" -cycle with a fixed frequency and is calculated with equation 5.1.

$$PRCP = P_{average} \cdot t_{f-avg\,read\,0} \tag{5.1}$$

where P_{avg} is the average power consumption of a the SRAM cell when a sequence of logic "0" are read from the cell. $t_{f-read\,0}$ is the average fall time on the read bitline. The bitline does not discharge when reading a logic "1", so the fall time is the worst-case scenario in terms of delay. The simulation for the PRCP was performed a minimum sized 10T SRAM cell in the TT process corner at 25^oC with a 32kHz toggling frequency on the read wordline.

32 SRAM cells were connected to the wordline and 128 cells were connected to the bitlines to get a result of a 4K array like the system in Fig. 3.13. The wordline signal was applied to the input of a buffer consisting of two minimum sized inverters to get more realistic and non-ideal rise and fall times.

5.2 Logic Gates

Fig. 5.1 shows the process flow used to designing the logic gates.

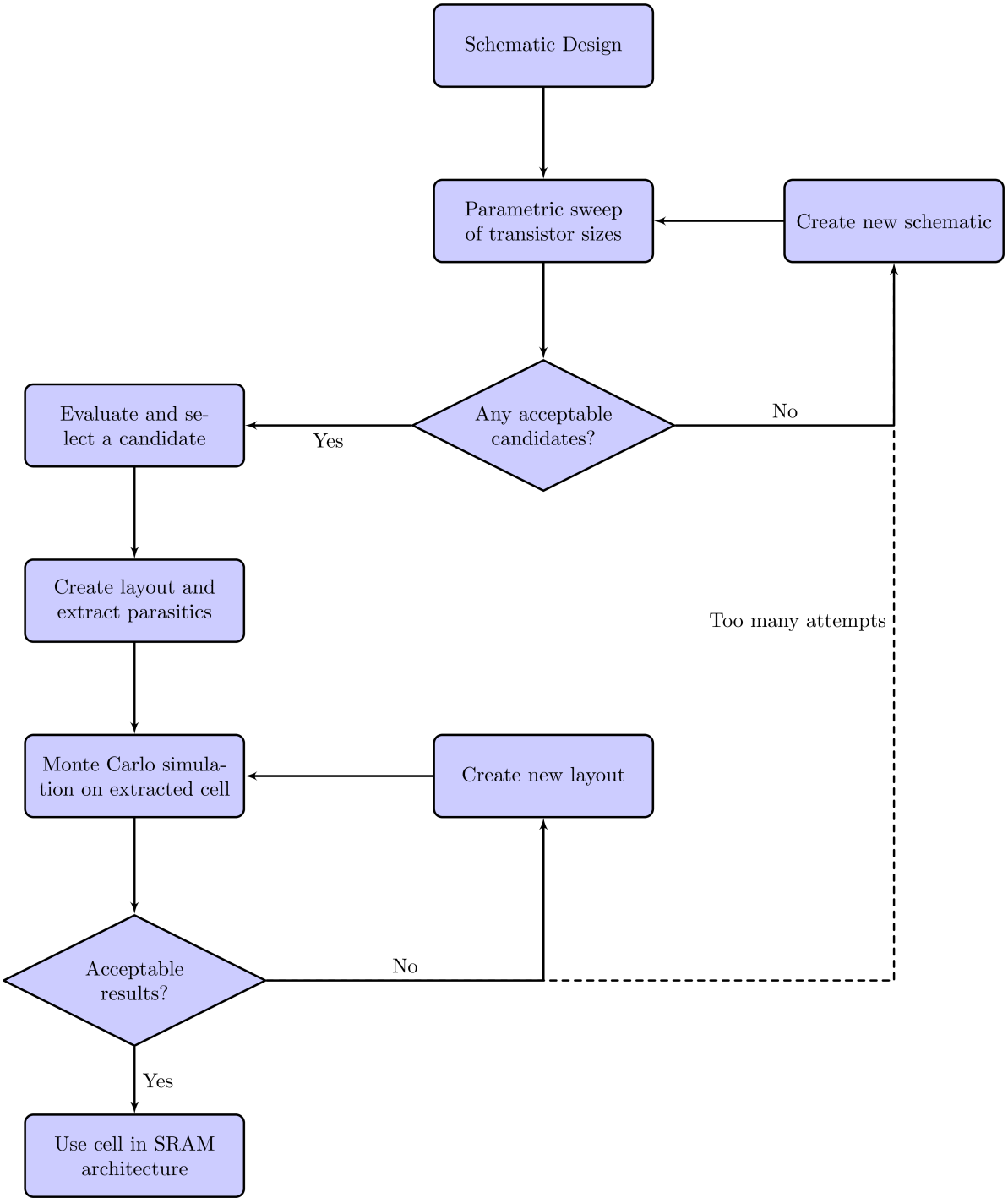


Figure 5.1: Gate sizing process flow

The logic gates were evaluated by their VTC balance, propagation delay, rise and fall times. The area and performance at superthreshold voltages was also taken into account when choosing transistor sizes. To reduce simulation time the transistor sizes were adjusted as multiples of the minimum transistor width and length defined by the 130nm CMOS process ($W_{min} = 160nm$, $L_{min} = 130nm$).

5.2.1 VTC Balance

The balance of a logic gate was characterized as its deviation from a perfectly balanced logic gate where $V_{in} = V_{out} = V_{DD}/2$ as shown in Fig. 5.2. Imbalance in the logic gates causes power consumption to go up and increases the optimal V_{DD} for a system [11].

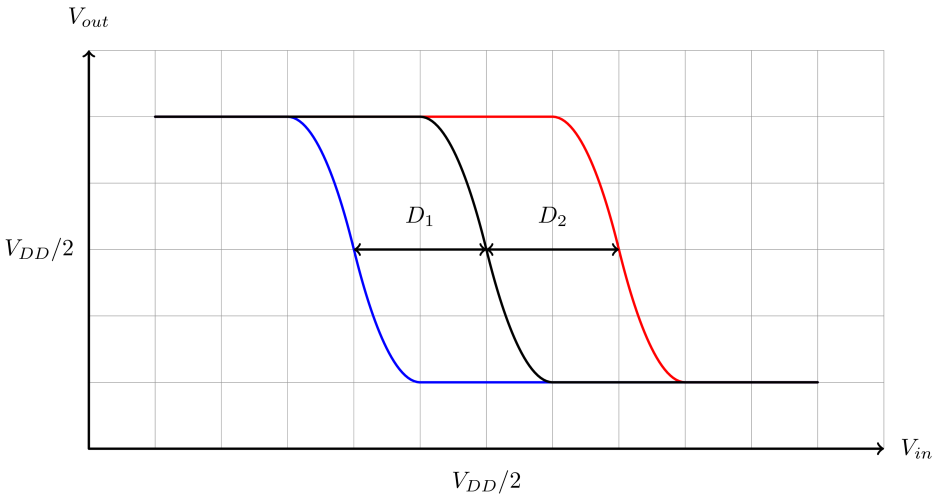


Figure 5.2: VTC balance deviation

The deviation was calculated using equation 5.2.

$$D = \left(\frac{V_{out}}{V_{DD}/2} - 1\right) \cdot 100\% \tag{5.2}$$

Since the inverters have one input they have one VTC, but multi-input gates like the NAND gate have three VTCs depending on the input combination. The VTC for a given input combination was found by sweeping a DC voltage on the selected inputs from 0V to V_{DD} while keeping the other inputs tied to ground. Every logic gate was connected as shown in Fig. 5.3 for a NAND gate.

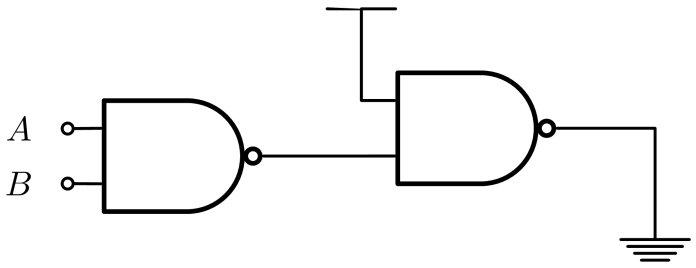


Figure 5.3: Testbench for VTC deviation

The gate balance was simulated using Monte Carlo analysis with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both both supply voltages.

5.2.2 Dynamic Performance

The propagation delay is the time it takes for a change in the input to be observed on the output of a circuit. The propagation delay is calculated using equation 5.3[26].

$$\tau_D = \frac{t_{pLH} + t_{pHL}}{2} \tag{5.3}$$

Where t_{pLH} and t_{pHL} is the low-to-high and high-to low transition time from $V_{DD}/2$-to-V_{DD} and V_{DD}-to-$V_{DD}/2$ respectively. The delay was measured by constructing a 5-stage ring oscillator of the logic gate under test as shown in Fig. 5.4 for the NAND gate. To initialize oscillation the B-input was set to V_{DD} by a voltage controlled switch (*switch* from the cadence *AnalogLib* library) for $1\mu s$ before the delay measurements. The propagation delay was measured only for the first logic gate in the chain.

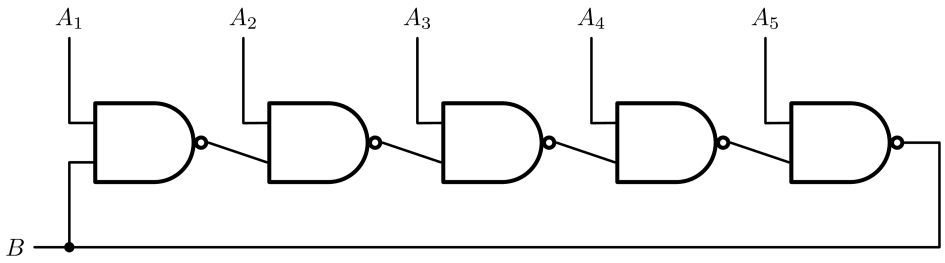


Figure 5.4: Testbench for propagation delay

The A-inputs are set depending on the logic gate under test. The inverters will always oscillate, but the second input for the other logic gates must was set according to Table 5.1 to maintain oscillations.

Gate	A_1	A_2	A_3	A_4	A_5
Inverters	No second input				
NAND	V_{DD}	V_{DD}	V_{DD}	V_{DD}	V_{DD}
NOR	0	0	0	0	0
XNOR	V_{DD}	0	V_{DD}	0	V_{DD}

Table 5.1: Delay testbench input setup

Rise and fall times also tell something about the dynamic performance of a circuit. A weak PUN and/or PDN will increase the rise and fall time due to imbalance. It is diffucult to achieve completely equal rise and fall times, but it is desirable to have them as close as possible to decrease power consumption. The rise and fall times were extracted from the same testbench as the propagation delay.

The propagation delay, rise and fall times were simulated using Monte Carlo analysis with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both both supply voltages.

5.2.3 Propagation Delay and Fan-Out

The propagation delay of a logic gate increases with the output load. To examine the delay and fan-out relationship the ring oscillator used for simulating the propagation delay was modified by adding multiple inverters to the output of the first logic gate in the oscillator as shown in Fig. 5.5 where $N-1$ is the inverter fan-out of the logic gate. N was increased in powers of two from 2 to 128 to show the effect of large fan-outs which are present in multiple signal paths in the SRAM architecture.

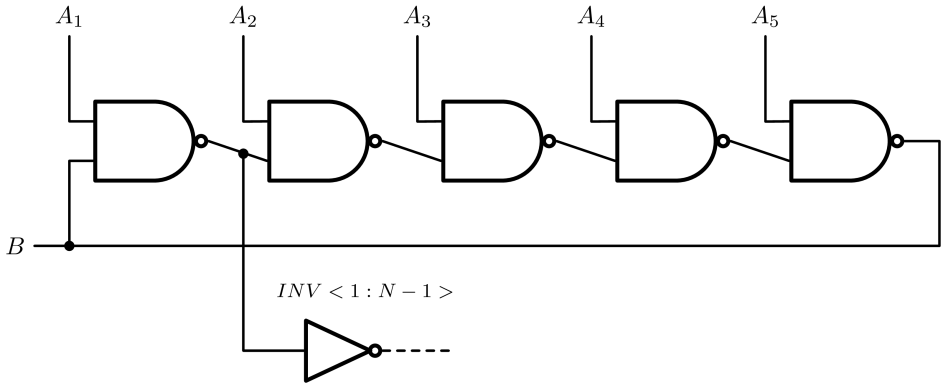


Figure 5.5: Testbench fot fan-out/delay

Simulation time increases for large fan-outs because the simulator needs a lot of time to set initial voltages in every node in all subcircuits ans schematics generated after parasitic extraction will add several resistors and capacitors to the original schematic so the number of nodes can reach several hundred even for a simple logic gate. To show the general trends the 2 to 128 fan-out was simulated in the FF, FS, TT, SF and SS process corners at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV and 1.2V.

5.2.4 Power Consumption

The power consumption was measured with the average current drawn by the power supply over 10 32kHz clock cycles using equation 5.4.

$$P = I_{DD-avg} V_{DD} \tag{5.4}$$

The leakage power mas measured for all input combinations and the total power consumption (leakage + dynamic power) was measured with a 32kHz clock toggling the worst case input combinations in Table 5.2 for 10 clock cycles. The dynamic power consumption is proportional to the frequency (see equation 2.2) so the total power consumption for a given frequency f can be estimated using equation 5.5.

$$P_{tot-f} = P_{leak} + \frac{f}{32KHz} P_{32kHz} \tag{5.5}$$

The toggling pulse was applied to the input of a buffer consisting of two inverters to get more realistic and non-ideal rise and fall times. The power consumption of these inverters were connected to another power supply through cadence inherited connections to not affect the result.

Gate	*Input*1	*Input*2
Inverter	$A \uparrow\downarrow$	-
Clocked Inverter	$Clk \uparrow\downarrow$	$A \uparrow\downarrow$
NAND	$A \uparrow\downarrow$	$B \uparrow\downarrow$
NOR	$A \uparrow\downarrow$	$B \uparrow\downarrow$
XNOR	$A \uparrow\downarrow$	$B-$

Table 5.2: Total power consumption input configurations

The power consumption was simulated using Monte Carlo analysis with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both supply voltages.

5.3 Decoder Critical Path Delay

The propagation delay of the decoder adds to the total delay of the read and write operations. To simulate the decoder a critical path with the largest fan-out was used to find the worst-case delay. The critical path of the decoder in Fig. 3.12 is from the input of a 2-to-4 decoder to the output of the total decoder. The simulation setup is shown in Fig. 5.6 where the critical path is indicated by the dotted lines.

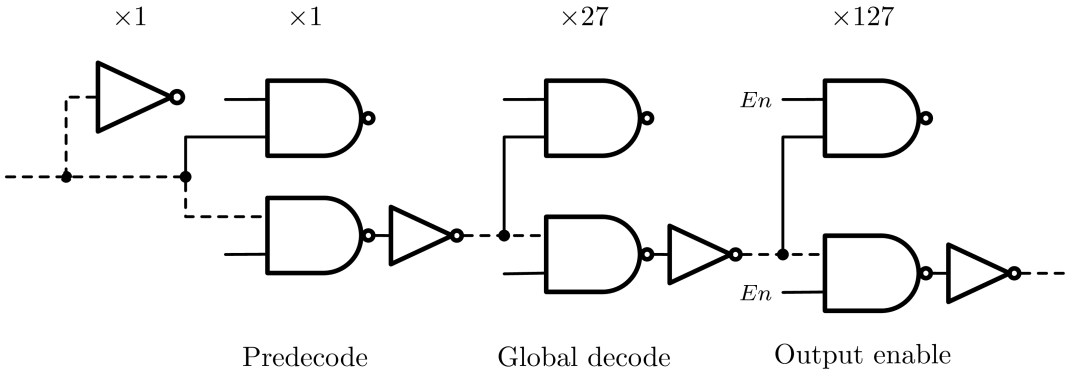


Figure 5.6: Critical Path of two-stage decoder

The decoder propagation delay was found by applying a 32kHz pulse to the address input and calculating the delay with equation 5.3. The input pulse was applied to the input of a buffer consisting of two inverters to get more realistic and non-ideal rise and fall times.

The delay simulation was performed using Monte Carlo analysis with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both supply voltages.

5.4 SRAM Cells

The SRAM cells were evaluated based on their noise margins and leakage power consumption. The 6T cell was also evaluated based on the severity of disturb voltage ΔV during read operations caused by the voltage divider between the access transistor and the adjacent NMOS transistor in the node storing a logic ”0”.

5.4.1 Static Noise Margins

Extracting the SNM graphically as described in chapter 2 is inefficient with Monte Carlo simulations due to the amount of VTCs that has to be evaluated and implementing a method that extracts the largest diagonal between the VTC curves is also not trivial. The testbench in Fig. 5.7 allows direct extraction of the SNM in any simulator[8].

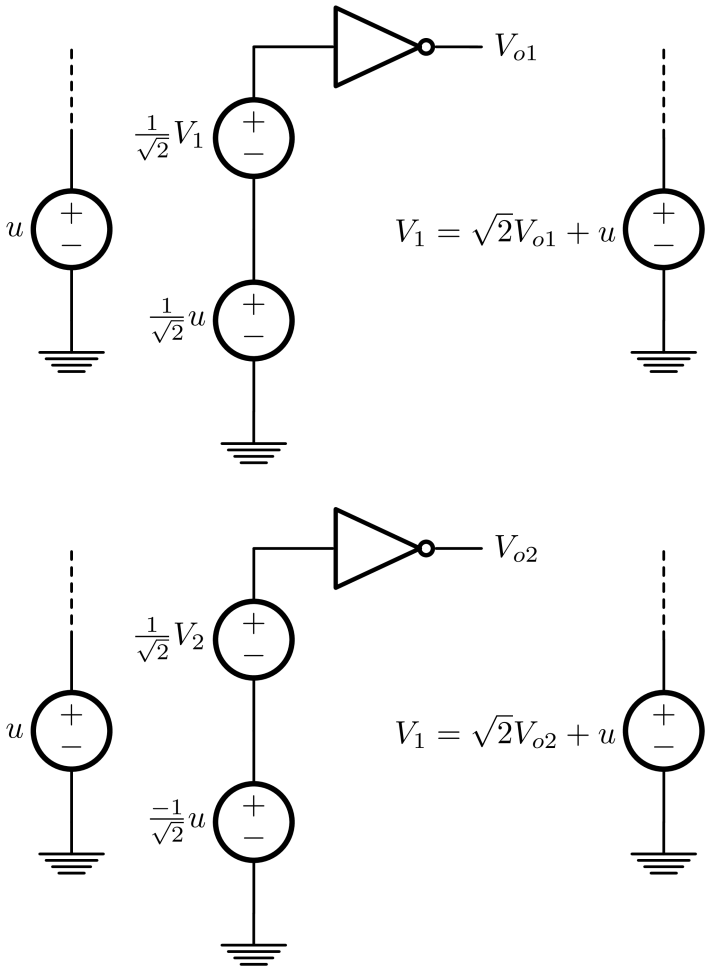


Figure 5.7: SNM simulation method

The inverters represent the inverters in the SRAM cell (access transistors are not shown to simplify the figure). The variable u is increased from $-\sqrt{2}V_{DD}$ to $\sqrt{2}V_{DD}$. The outputs $V1$ and $V2$ combines into the butterfly plot, but in a new coordinate system (u, v) where the VTCs are tilted 45^o. The lesser of V_{o1} and V_{o2} is divided with $\sqrt{2}$ to obtain the SNM as shown in Fig. 5.8. Because voltage-controlled voltage sources (VCVS) in Cadence have limited functionality the tesbench was modeled in a Verilog-A module. The source code is available in Appendix C.

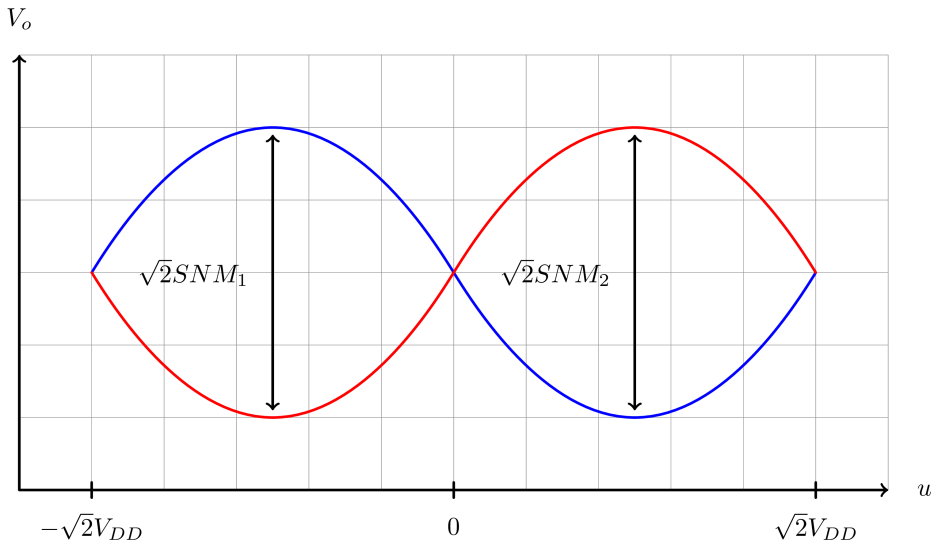


Figure 5.8: Butterfly plot in (u,v) coordinate system

The SNM was simulated using Monte Carlo analysis with random process and mismatch variations for temperatures -40^{o}C, 25^{o}C and 85^{o}C at V_{DD} = 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both supply voltages.

5.4.2 On/off-current Ratio

The operation of the SRAM at subthreshold voltages can be limited by the on/off-current ratio as explained in section 2.3.2. If the ratio is too low the difference between a logic ”1” and logic ”0” can become indistinguishable and the length of a bitline must be reduced. The on/off-current ratio was simulated by applying a pulse to the gate of a minimum sized NMOS transistor as shown in Fig. 5.9. The off-current was measured for logic ”0” input and the on-current was measured for logic ”1” input.

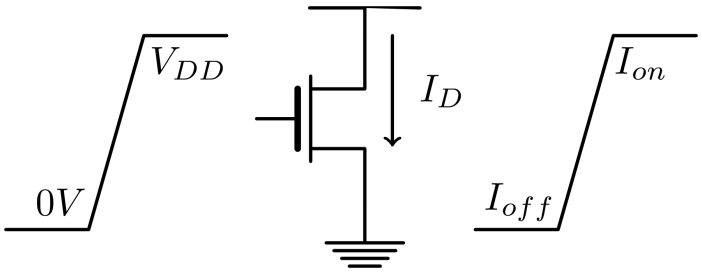


Figure 5.9: On/off-current ratio simulation with NMOS transistor

The on/off-current ratio was simulated in the FF, FS, TT, ST and SS corner with a fixed temperature of 25^C where the supply voltage was increased from 10mV to 1.2V to show the effects of increased supply voltage. The simulation was repeated in the same corners with V_{DD} = 400mV in the temperature range -40^{o}C to 85^{o}C to show the effects of temperature.

5.4.3 10T Bitline Length and Read Delay

For V_{DD} = 400mV the bitline discharge delay would pose a larger limitation before the on/off-current ratio would become an issue. To decide the number of SRAM cells connected to a bitline the number of 10T SRAM cells was increased in powers of two from 16 to 512. The delay was measured as the time from the read wordline was set to V_{DD} to the output of the sensing inverters was set to logic "0". The simulation was only performed for 10T SRAM cells because the read buffer stack decreases both the on and leakage currents resulting in a slower bitline discharge. The delay was aslo only measured for a read "0" operation as the read bitline does not discharge for a read "1" operation.

Simulation time increases tremendously for large fan-outs because the simulator needs a lot of time to set initial voltages in every node in all subcircuits and schematics generated after parasitic extraction will add several resistors and capacitors to the original schematic so the number of nodes can reach several hundred even for a simple logic gate. To show the general trends in increasing the number of SRAM calls from 16 to 512 the read time was simulated FF, FS, TT, SF and SS process corners at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV and

5.4.4 6T SRAM Cell Read Disturb

As shown in Fig. 2.4 the 6T SRAM cell experiences a noise voltage ΔV in the node holding a logic "0" because of the voltage divider formed by the access transistor and it's adjacent NMOS transistor. The peak value of this spike is dependent on supply voltage, the amount of SRAM cells connected to the bitlines and the pulse width of the wordline signal. The noise voltage is tied to the SNM, but even if a single cell shows satisfactory static behavior it might still fail dynamically in a SRAM array.

The read disturb was modeled as the peak voltage present on the internal nodes in the 6T SRAM cell during read operations. Temperature and frequency dependencies were simulated using Monte Carlo analysis with random process and mismatch variations for temperatures -40^{o}C, 25^{o}C and 85^{o}C at supply voltages 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for each supply voltage value. Temperature and frequency dependencies were also simulated with 32 SRAM cells connected to the bitlines. Because the simulation setup time increases tremendously with many component instances the bitline length simulations were only performed in the FF, FS, TT, ST and SS corner at temperatures -40^{o}C, 25^oC and 85^{o}C for V_{DD} = 400mV and 1.2V.

5.4.5 Leakage Power Consumption

The leakage power consumption of the SRAM cells can amount to a large portion of the overall power consumption of a system due to the large amounts of cells and as CMOS processes shrinks in size the leakage power can easily become the dominating power consumption. The leakage power consumption was measured for a single SRAM cell by measuring the average current drawn from the power supply and calculating the power consumption using equation 5.4.

The leakage power consumption was simulated using Monte Carlo analysis with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both supply voltages.

5.5 Sense Amplifier

At subthreshold voltage the speed and accuracy of a circuit is traded for lower power consumption. The differential nature of the latch-based SA results in a faster read operation compared to single-ended sensing, the most important metric at subthreshold voltages is the SA offset voltage. The SA was evaluated by its offset voltage and power consumption.

5.5.1 Read Acces Yield

Equation 4.1 gives the offset voltage of a SA at subthreshold voltages. The threshold voltage differences ΔV_{th} are not directly available, however it can be estimated using the standard deviation of σ_{th} of a minimum sized NMOS transistor[23]. To simplify the simulation the offset voltage is defined as the minimum differential bitline voltage ΔV_{BL} needed to perform a correct sensing operation. When a ΔV_{BL} is applied to the SA inputs both the Out and $\overline{Out}$ nodes will begin to discharge, but after some time after the $SAEN$-signal is set the latching operation will cause one node to settle at ground potential and the other will return to V_{DD}. If ΔV_{BL} is too small the wrong node might discharge to ground potential and the sensing operation fails. To characterize the offset voltage of the SA the read yield for a given ΔV_{BL} was simulated. The read yield is given by equation 5.6[27]

$$Y_{read} = \frac{Succesfull\ read\ iterations}{Total\ iterations} \times 100\% \tag{5.6}$$

Y_{read} was simulated with ΔV_{BL} increased from 0V until 100% read yield was achieved. A successful sensing operation was defined as when the internal node Out of the SA reached 0V. For iterations where the sensing operation failed Out would be equal to VDD. The $SAEN$-signal and the $\overline{RST}$ was applied with a 32kHz pulse through two inverters to achieve realistic rise and fall times.

The read yield was simulated using Monte Carlo analysis with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both supply voltages.

5.5.2 Leakage Power Consumption

The leakage power consumption was measured the SA cell by measuring the average current drawn from the power supply over 10 32kHz clock cycles and calculating the power consumption using equation 5.4. The leakage power was measured with $SAEN = 0$, $\overline{RST} = 0$ and $\Delta V_{BL} = 0$ (both bitlines at V_{DD} because of precharge).

The leakage power consumption was simulated using Monte Carlo analysis with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C at 85^{o}C for V_{DD} = 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both supply voltages.

5.6 Enable Register Delay

The enable register is in essence a latch with asynchronous reset, but since the input is tied to V_{DD} the setup and hold time is not of much interest compared to an ordinary latch. The propagation delay from the enable input to $RWEnable = V_{DD}$ adds to the overall delay of the read and write operations and was simulated the delay from $V_{DD}/2$ on the enable input to $V_{DD}/2$ on the $RWEnable$ output.

The propagation delay was simulated for both the set and reset operations and they were simulated using Monte Carlo simulations with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C, 85^{o}C for V_{DD} 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both supply voltages.

5.7 Transition Detector Pulse Width

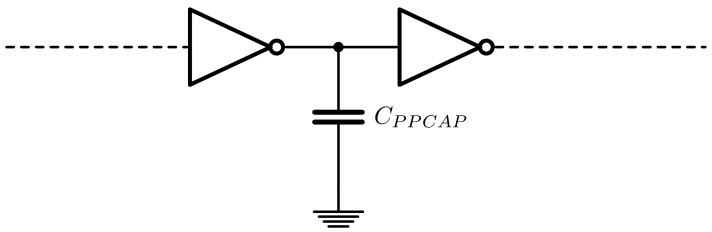


Figure 5.10: Transition detector delay element

The value of the PPCAP in the transition detectors delay element shown in Fig. 5.10 decides the pulse width of the output signal and adds to the overall delay of read and write operations. The pulse generated by the transition detector must be long enough so the enable flag register is able to set, but should also be as small as possible to keep the overall delay short. 35fF was chosen

The pulse width was simulated using Monte Carlo simulations with random process and mismatch variations at temperatures -40^{o}C, 25^{o}C, 85^{o}C for V_{DD} 400mV and 1.2V. Each Monte Carlo simulation ran for 200 iterations resulting in a total of 600 iterations for both supply voltages.

5.8 SRAM Architecture

The individual components of the SRAM architecture have been simulated with Monte Carlo analysis to show the impact of PVT variations and mismatch. Testing of the entire architecture using Monte Carlo analysis is not practical due to the long simulation times so the SRAM architectures were simulated using corners. To further reduce simulation time the testbench was simplified to the replica wordline and bitline, control logic, the critical path of the decoder and a single row and column of SRAM cells. 32 SAs, write drivers and precharge circuits were connected to apply realsistic loads to all control signal paths as shown in Fig. 5.11 and 5.12 for the 10T and 6T testbench respectively.

5.8.1 Read and Write Cycle Times

The read and write operations starts when a positive edge is applied to the input of the transition detector. It is assumed the address and read/write enable inputs are set before the pulse arrives. The definition of the operation cycle times are as follows:

- **Write cycle time:** The time from a positive edge on the transition detector to the VV_{DD}-output of the wordline driver is pulled back to V_{DD}.
- **10T read cycle time:** The time from a positive edge on the transition detector to the read bitline is precharged back to V_{DD}.
- **6T read cycle time:** The time from a positive edge on the transition detector to the $SAEN$-signal is reset to ground potential.

The cycle times were simulated in the FF, FS, TT, SF and SS corner for temperatures -40^{o}C, 25^{o}C, 85^{o}C for V_{DD} at 400mV and 1.2V.

5.8.2 Power Consumption

The total power consumption (leakage + dynamic power consumption) was measured for both the 6T and 10T testbenches by measuring the average current drawn from the power supply and calculating the power consumption using equation 5.4. A complete 4K SRAM array will have higher consumption because of the 3937 remaining unaccessed SRAM cells omitted from the testbench. The logic gates in the decoder that is not on the critical path will also contribute to the total leakage currents so the total power for a complete 4K array can be estimated by adding the mean leakage found from previous simulations of the SRAM cells and logic gates.

The power consumptions were simulated in the FF, FS, TT, SF and SS corner at temperatures -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV and 1.2V.

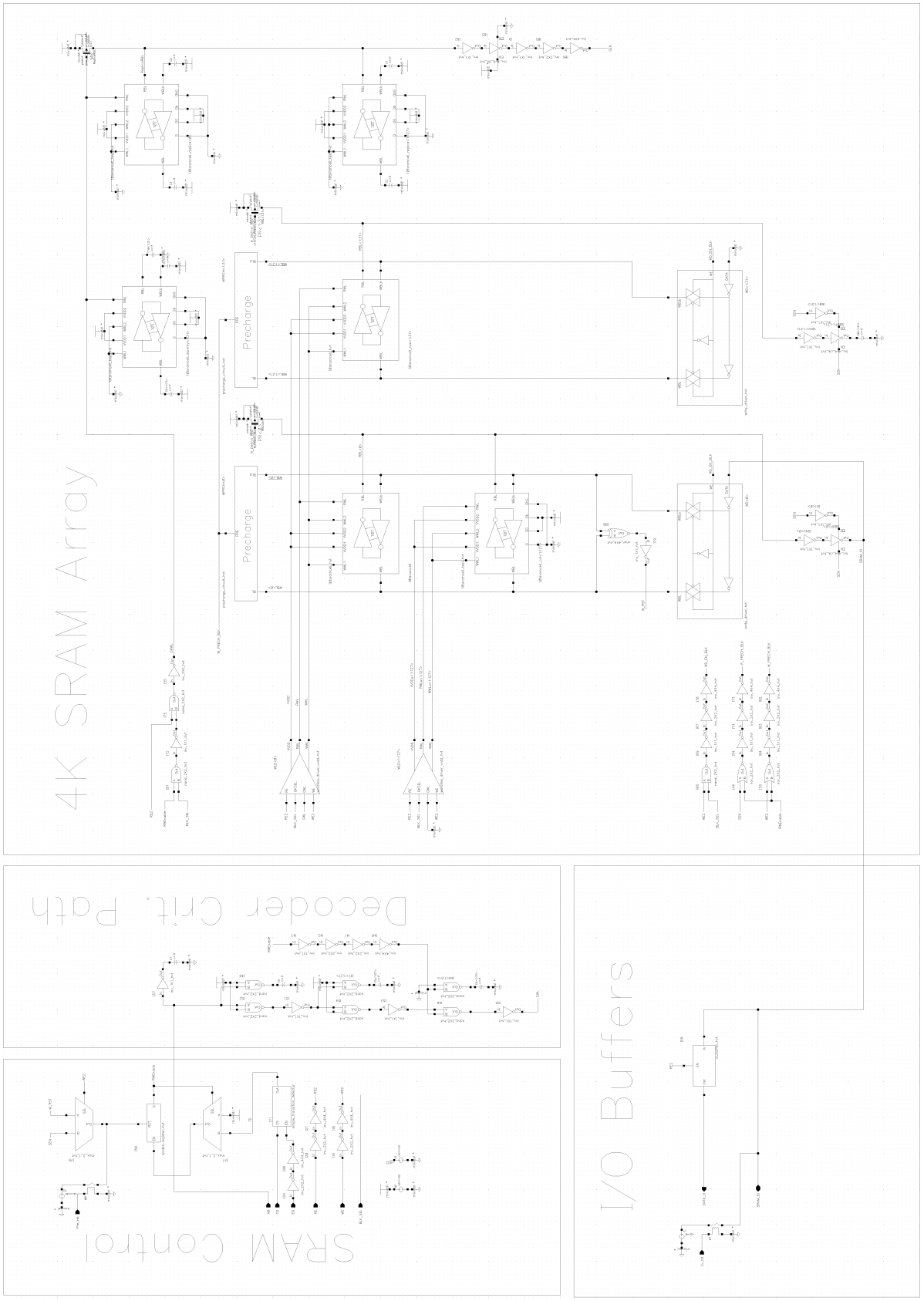


Figure 5.11: 10T SRAM architecure testbench

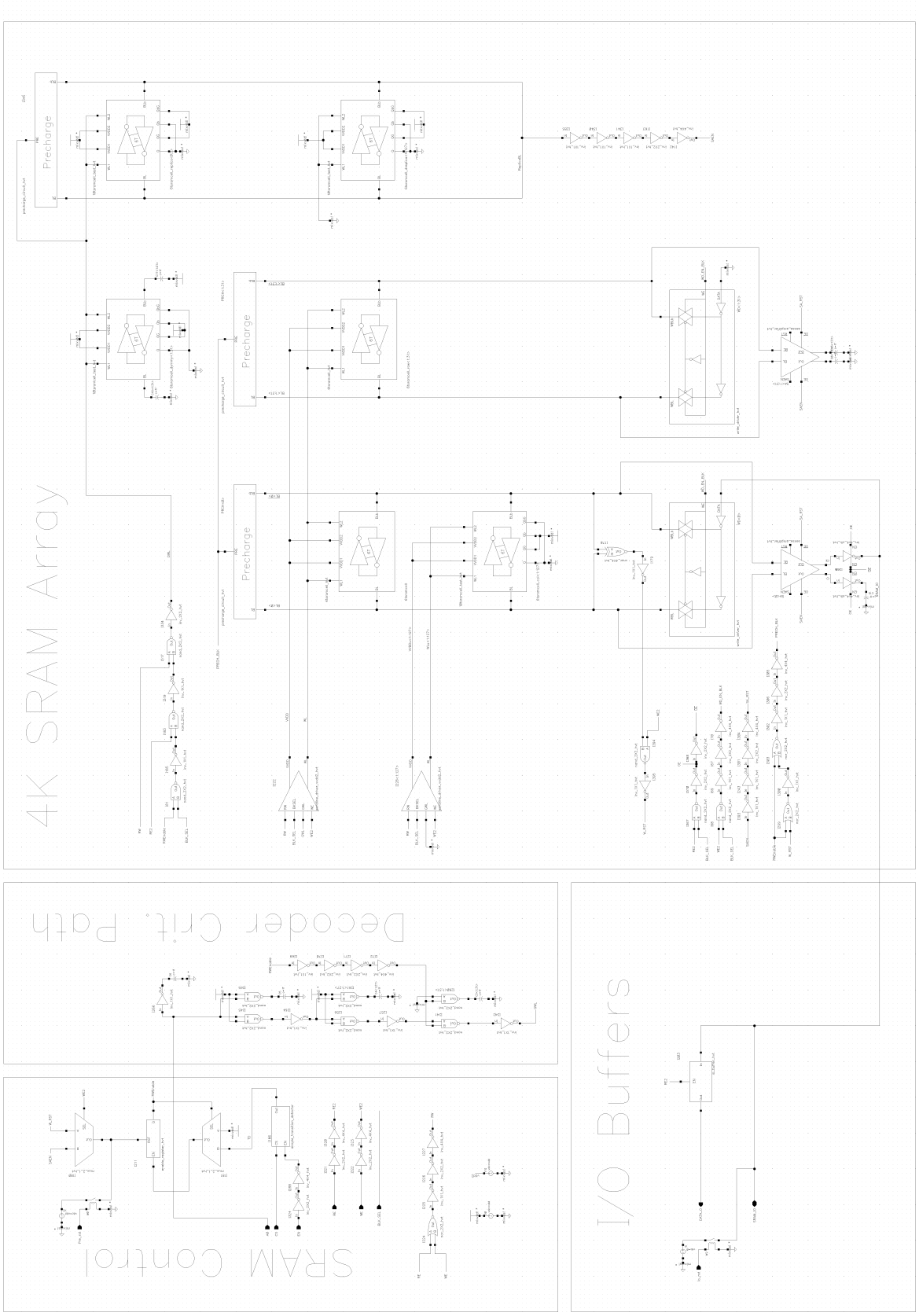


Figure 5.12: 6T SRAM architecure testbench

6. Results

The results from the simulations described in chapter 5 are presented in this chapter. The results are extracted from Cadence Virtuoso ADE and ADE XL and post-processed in MathWorks MATLAB.

6.1 Power-Read Cycle Product

Fig. 6.1 shows the PRCP in the nominal process corner at 25°C. The global minimum is located approximately at the absolute value of the NMOS threshold voltage, but exhibits little devation in the voltage range 400mV to 550mV.

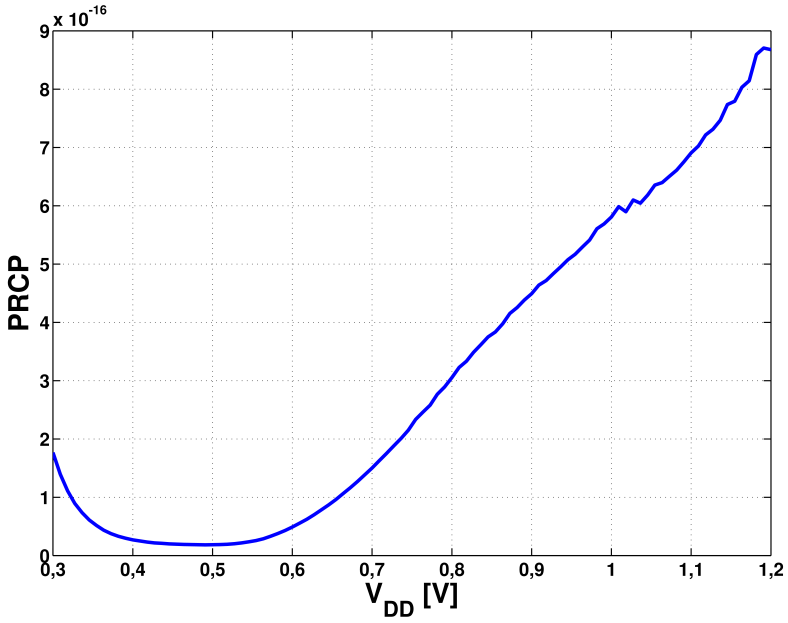


Figure 6.1: Power-read cycle product

To minimize power consumption while maintaining a low PRCP the supply voltage for low voltage operation was set to 400mV.

6.2 Logic Gates

This section will present the simulation results from the tests of the NAND gate only. The NAND gate is the most used logic gate in the SRAM architecture except from the inverter. While the inverter is the most used logic gate it is also the least interesting because it is the most robust gate because of its simplicity and symmetry. Results for the other logic gates are available in Appendix A.

6.2.1 Balance

Fig. 6.2 shows the two worst case VTCs that occurred for the NAND2 gate after 600 Monte Carlo iterations at -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV.

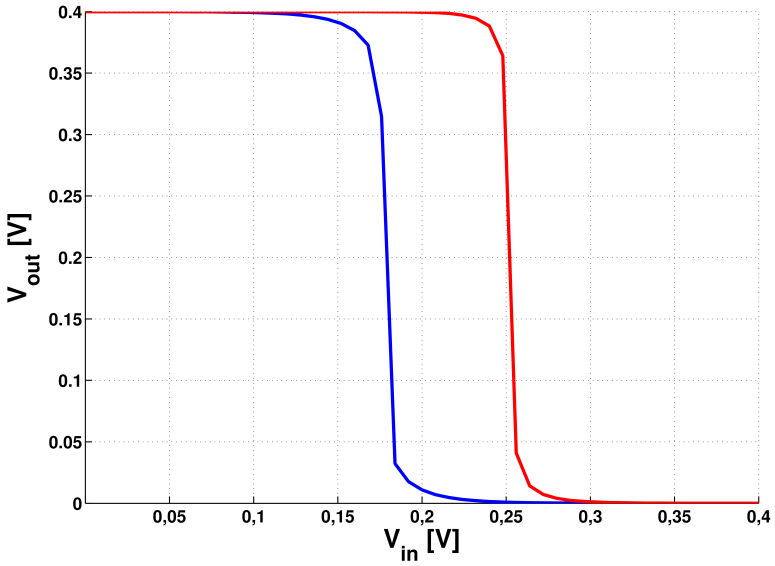


Figure 6.2: Worst-case VTCs for NAND2

The VTC to the left occurs for the input combinations AB = 01 and AB = 10 at 85^{o}C and the VTC to the right occurs for the input combination AB = 11 at -40^{o}C. Table 6.1 shows the results for all input combinations at both supply voltages.

	-40^oC				25^oC				85^oC			
0.4V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D_{AB}[%]	5.36	15.65	29.03	4.33	2.52	14.87	25.73	4.32	2.14	14.57	25.26	4.3
D_{A0}[%]	-5.6	6.75	18.3	5.19	-6.88	3.84	14.64	5.15	-10.09	1.429	13.86	5.3
D_{0B}[%]	-5.85	5.78	18.42	4.9	-9.62	2.815	14.62	4.99	-10.37	0.1	13.84	4.94
1.2V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D_{AB}[%]	-2	2.24	5.98	1.56	-1.93	3.25	6.07	1.71	-1.77	4.28	6.35	1.85
D_{A0}[%]	-10.06	-5.55	-1.96	2.06	-10.78	-7.18	-2.13	2.04	-13.96	-8.23	-2.44	2.2
D_{0B}[%]	-10.12	-6.53	-2.01	1.98	-14.03	-8.95	-5.16	2.1	-15.24	-10.94	-6.08	1.99

Table 6.1: NAND gate balance results

6.2.2 Dynamic Performance

Fig. 6.3 shows the mean (μ) and maximum propagation delay for the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

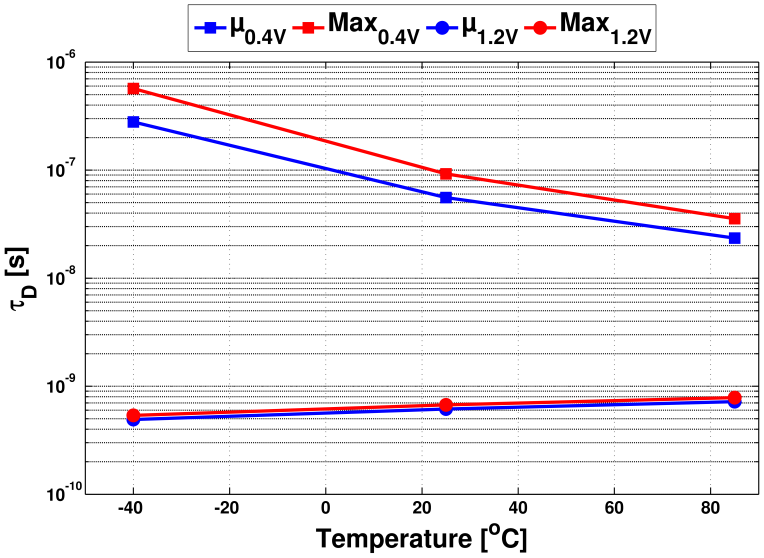


Figure 6.3: Propagation delay for NAND2 gate

Fig. 6.4 shows the delay standard deviation (σ) relative to the mean (μ) for the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

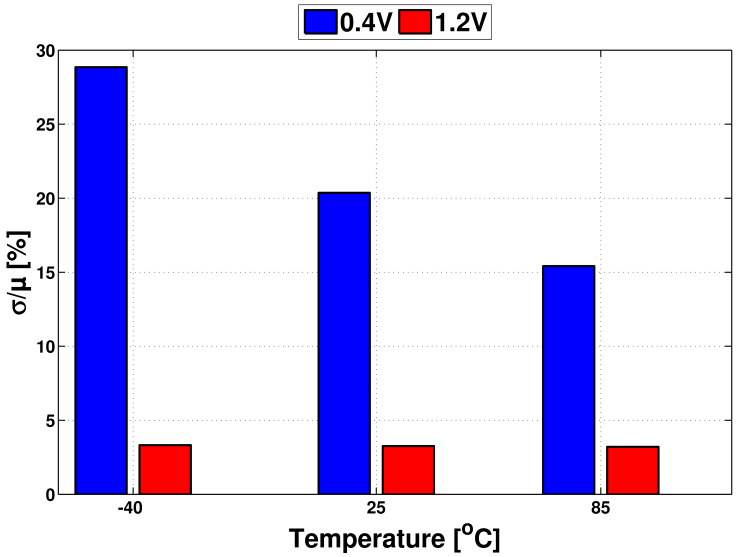


Figure 6.4: NAND2 propagation delay relative σ

Fig. 6.5 shows the mean (μ) and maximum rise time of the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

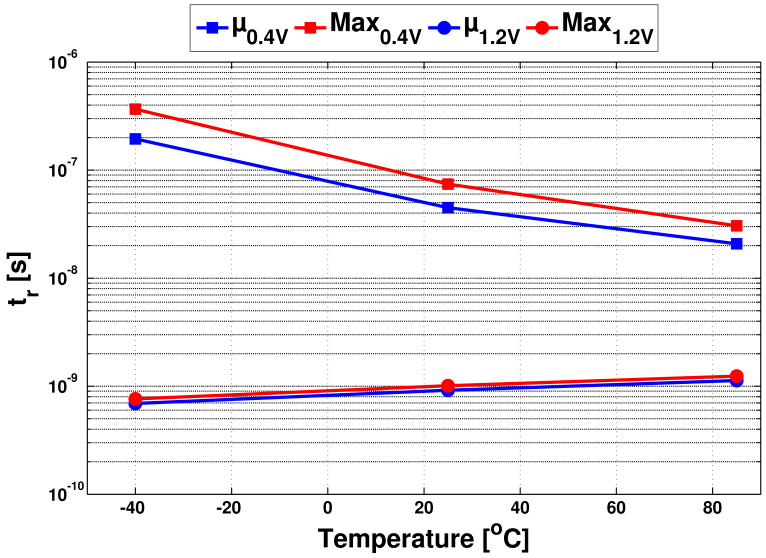


Figure 6.5: Rise time for NAND2 gate

Fig. 6.6 shows the mean (μ) and maximum fall time of the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

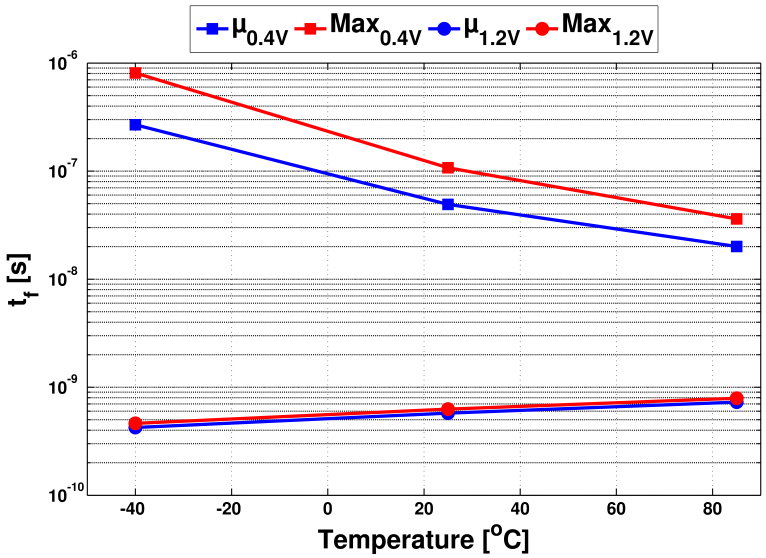


Figure 6.6: Fall time for NAND2 gate

Fig. 6.7 shows the rise and fall time standard deviation (σ) relative to the mean (μ) for the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

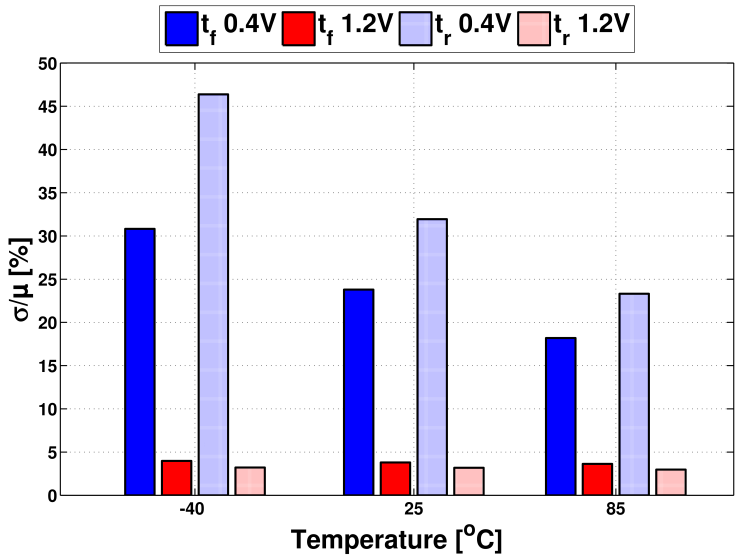


Figure 6.7: NAND2 rise and fall times relative σ

6.2.3 Power Consumption

Fig. 6.8 shows the worst-case leakage power consumption of the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V. The worst-case leakage occurs for the input AB = 11.

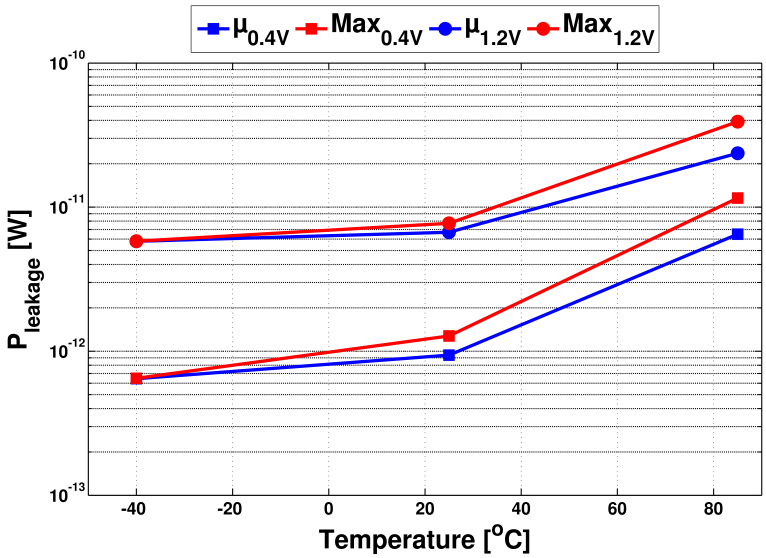


Figure 6.8: Leakage power consumption for NAND2 gate

Fig. 6.9 shows the leakage power standard deviation (σ) relative to the mean (μ) for the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

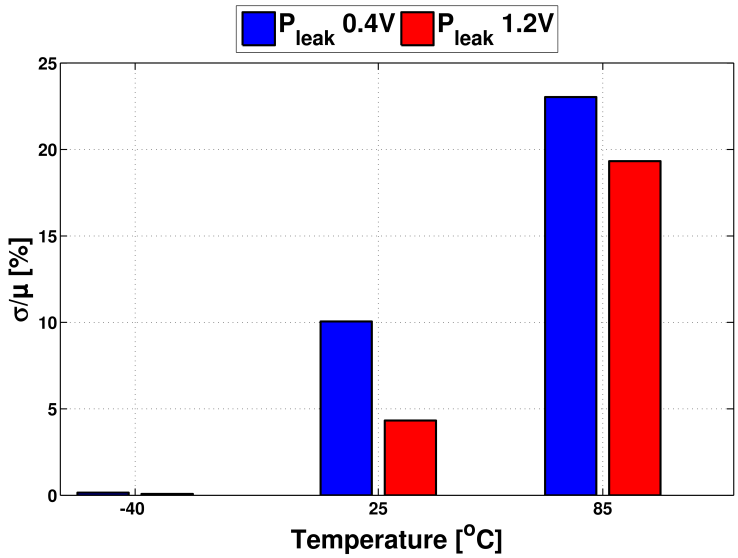


Figure 6.9: NAND2 leakage power relative σ

Fig. 6.10 shows the worst-case total power consumption of the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V. The worst-case power consumption occurs when both inputs toggle simultaneously.

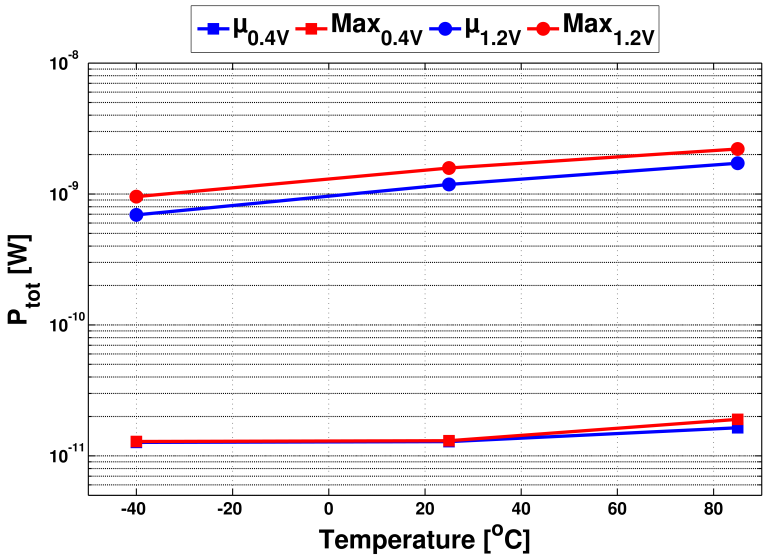


Figure 6.10: Total power consumption for NAND2 gate

Fig. 6.11 shows the total power standard deviation (σ) relative to the mean (μ) for the NAND2 gate after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

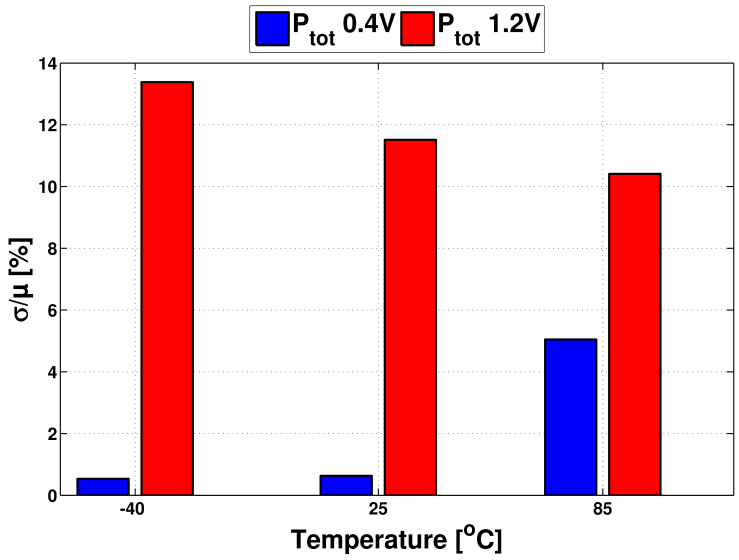


Figure 6.11: NAND2 total power relative σ

6.2.4 Propagation Delay and Fan-Out

Fig. 6.12 shows the delay and fan-out relationship of the NAND2 gate for V_{DD} = 400mV in the TT corner at 25^{o}C and the two most extreme corner and temperature cases.

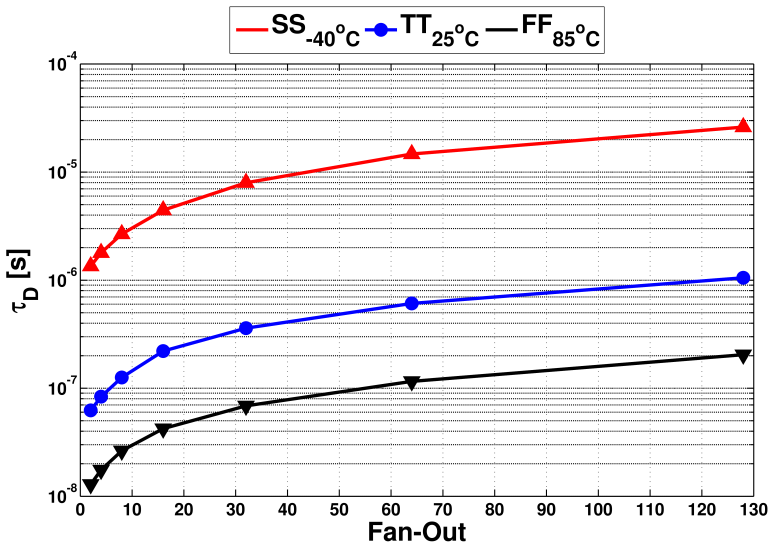


Figure 6.12: Delay-fan-out relatinship NAND2 gate

6.3 Decoder

Fig. 6.13 shows the propagation delay from the input of the decoder critical path to the output of the decoder after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

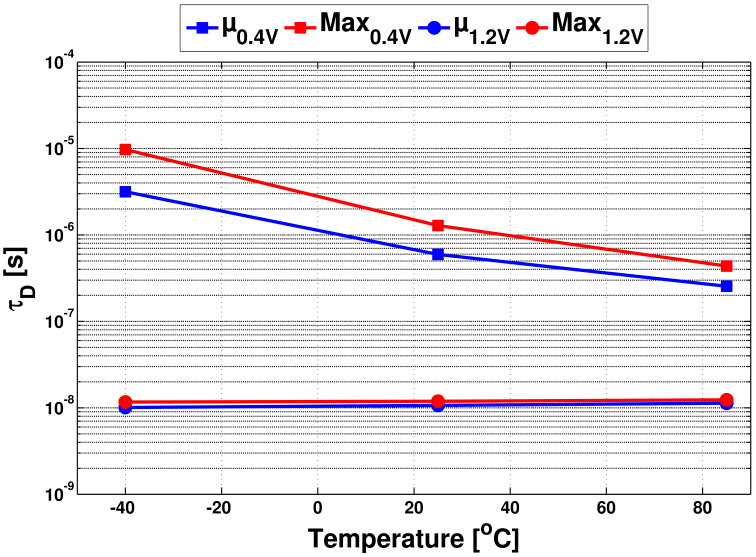


Figure 6.13: Decoder propagation delay

Fig. 6.14 shows the decoder propagation delay standard deviation (σ) relative to the means (μ) for the decoder after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

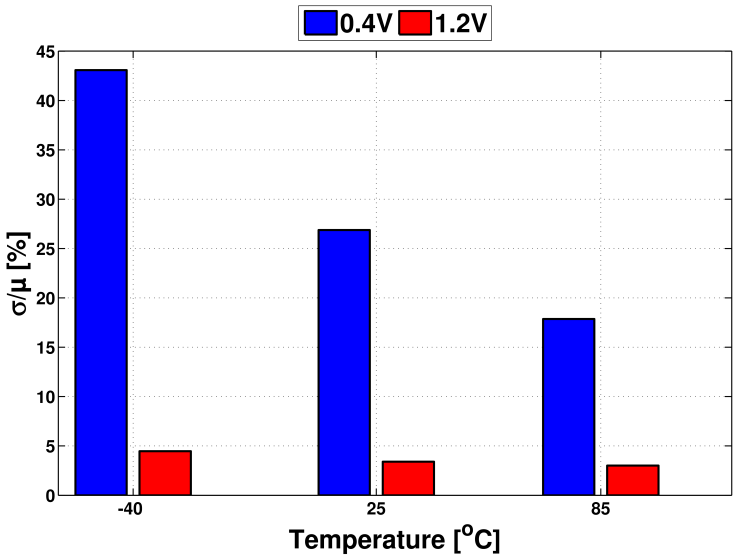


Figure 6.14: Decoder propagation delay relative σ

6.4 SRAM Cells

This section presents the simulation results from the tests of the SRAM cells.

6.4.1 Read Static Noise Margins

Fig. 6.15 and 6.16 shows the RSNM distribution and butterfly plot of the 10T SRAM cell for V_{DD} = 400mV and temperatures -40^{o}C, 25^{o}C and 85^{o}C.

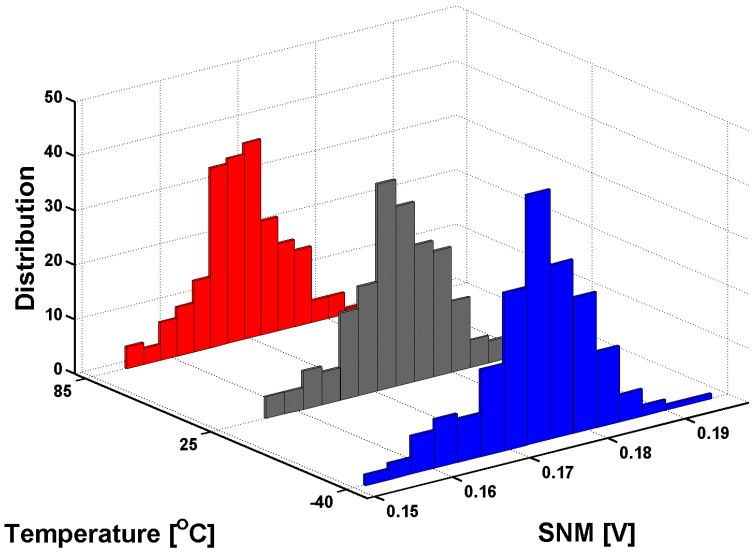


Figure 6.15: RSNM distributiins for 10T SRAM cell at 400mV

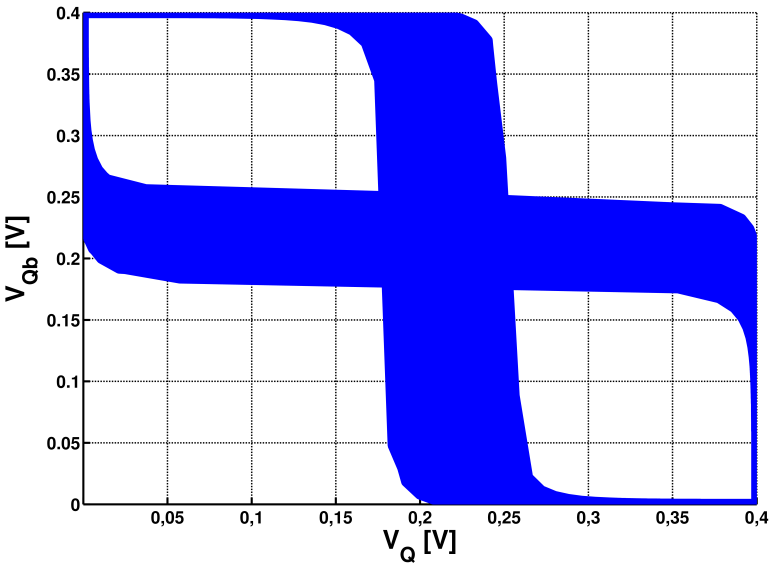


Figure 6.16: RSNM butterfly plot for 10T SRAM cell at 400mV

Fig. 6.17 and 6.18 shows the RSNM distribution and butterfly plot of the 10T SRAM cell for V_{DD} = 1.2V and temperatures -40^{o}C, 25^{o}C and 85^{o}C.

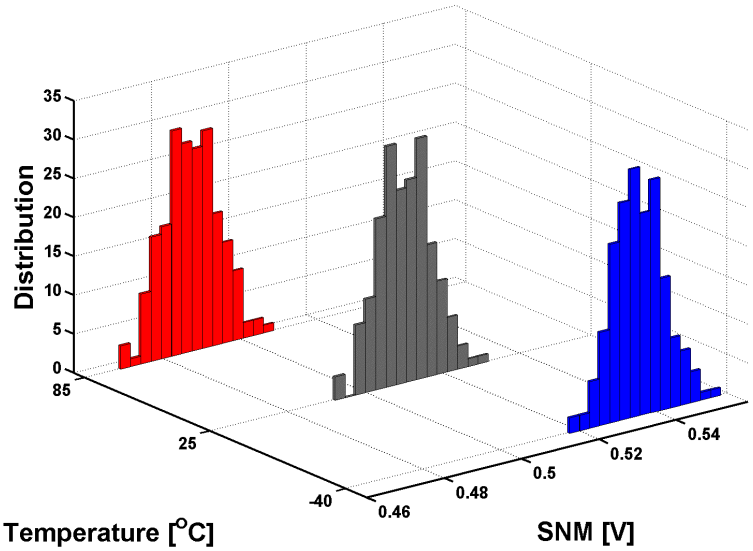


Figure 6.17: RSNM distributiins for 10T SRAM cell at 1.2V

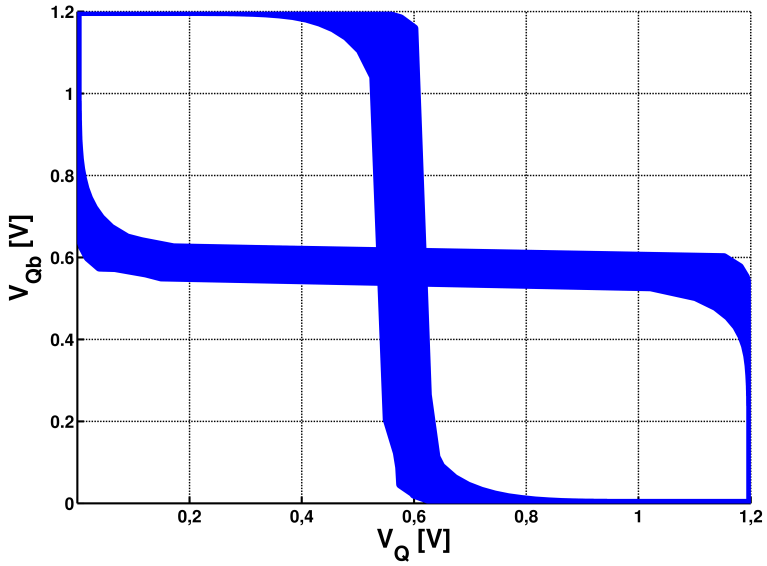


Figure 6.18: RSNM butterfly plot for 10T SRAM cell at 1.2V

Fig. 6.19 and 6.20 shows the RSNM distribution and butterfly plot of the 6T SRAM cell for V_{DD} = 400mV and temperatures -40^{o}C, 25^{o}C and 85^{o}C.

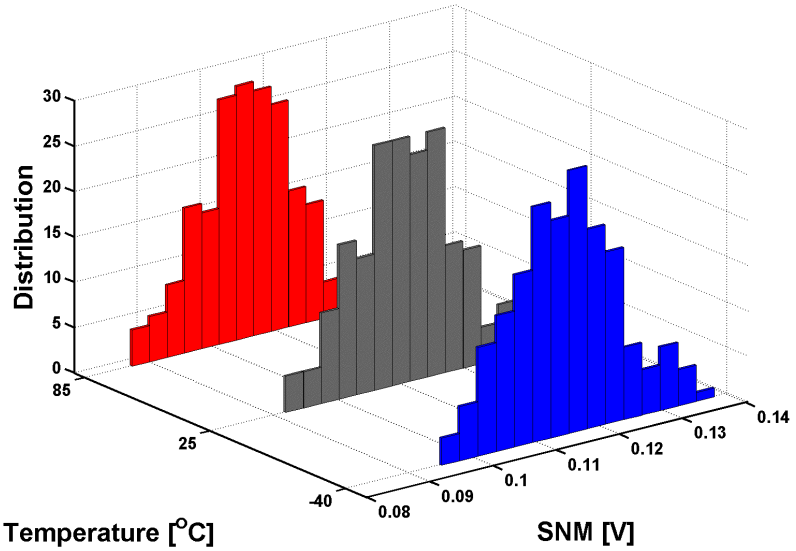


Figure 6.19: RSNM distributiins for 6T SRAM cell at 400mV

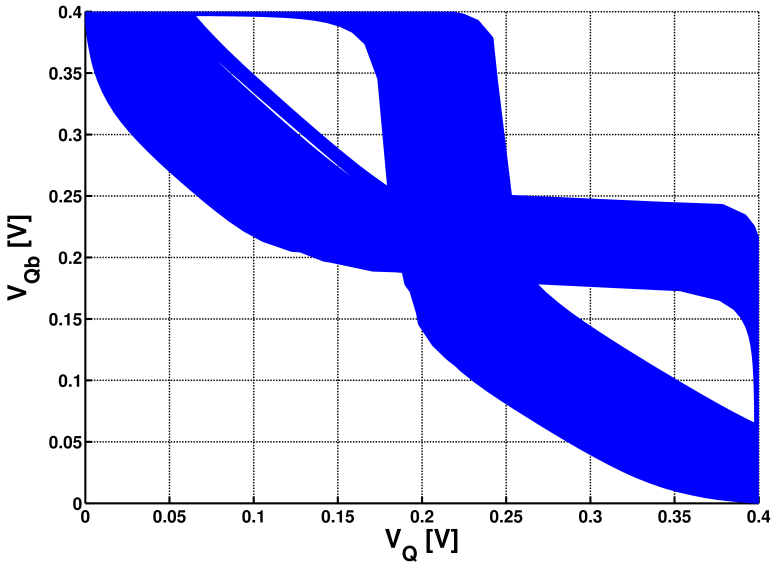


Figure 6.20: RSNM butterfly plot for 6T SRAM cell at 400mV

Fig. 6.21 and 6.22 shows the RSNM distribution and butterfly plot of the 6T SRAM cell for V_{DD} = 1.2V and temperatures -40^{o}C, 25^{o}C and 85^{o}C.

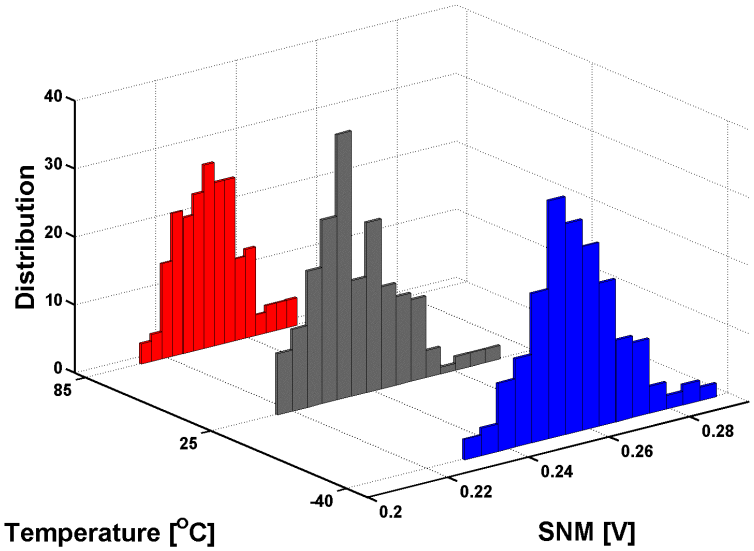


Figure 6.21: RSNM distributiins for 6T SRAM cell at 1.2V

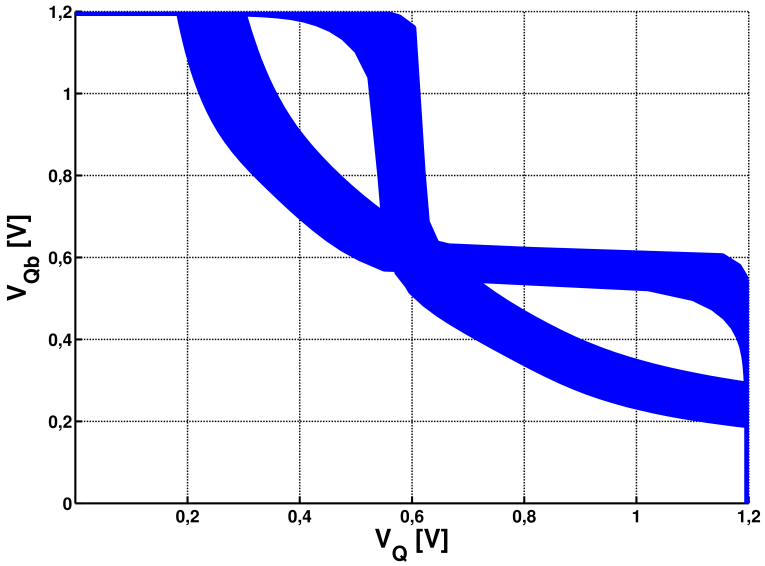


Figure 6.22: RSNM butterfly plot for 6T SRAM cell at 1.2V

6.4.2 Write Static Noise Margins

Fig 6.23 and 6.24 shows the WSNM distribution and butterflyplot for V_{DD} = 400mV and temperatures -40^{o}C, 25^{o}C and 85^{o}C with the VV_{DD} write assist method.

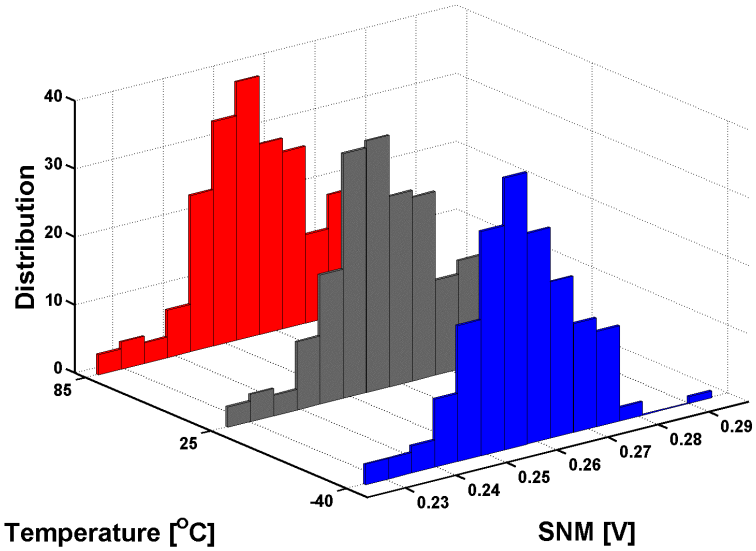


Figure 6.23: WSNM distributions for 10T and 6T SRAM cells at 400mV

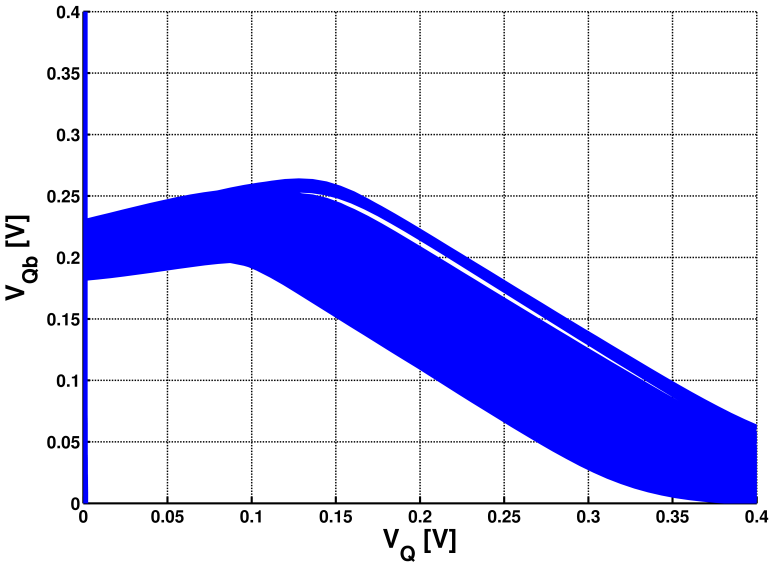


Figure 6.24: WSNM butterfly plot for 10T and 6T SRAM cells at 400mV

Fig 6.25 and 6.26 shows the WSNM distribution and butterflyplot for V_{DD} = 1.2V and temperatures -40^{o}C, 25^{o}C and 85^{o}C with the VV_{DD} write assist method.

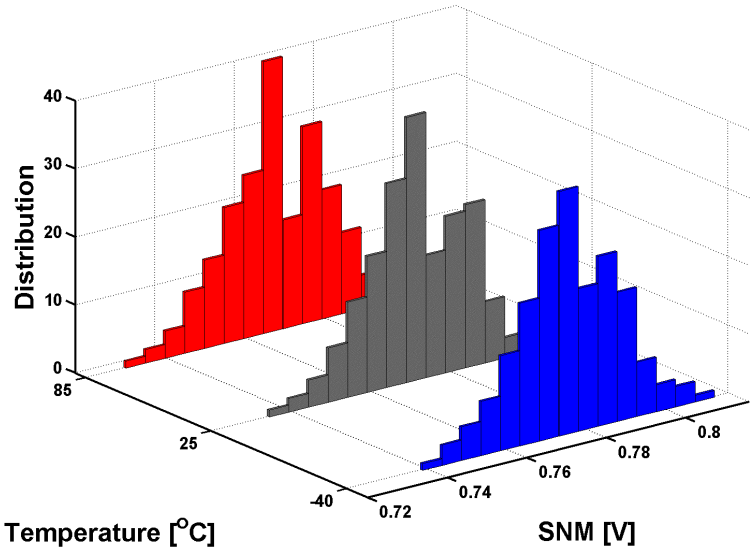


Figure 6.25: WSNM distributions for 10T and 6T SRAM cells at 1.2V

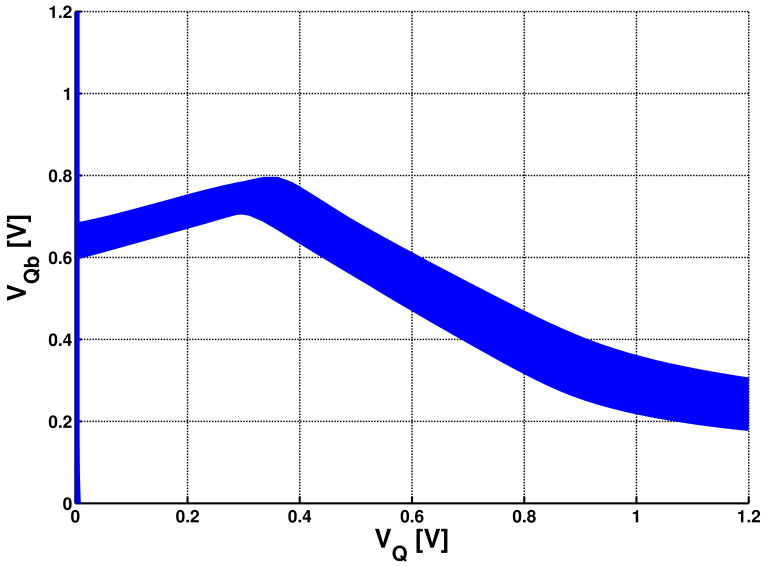


Figure 6.26: WSNM butterfly plot for 10T and 6T SRAM cells at 1.2V

The *VVDD* write assist method lowers the maximum WSNM in return for increased writability at subthreshold voltages. Fig. 6.27 shows the WSNM butterfly plot for temperatures -40^{o}C, 25^{o}C and 85^{o}C at V_{DD} = 400mV without the VV_{DD} write assist mechanism. The red circle indicates an area of the plot where the two VTCs might cross. Since the VTCs can cross more than once the circuit might not be monostable during write operations, resulting in a failed write operation.

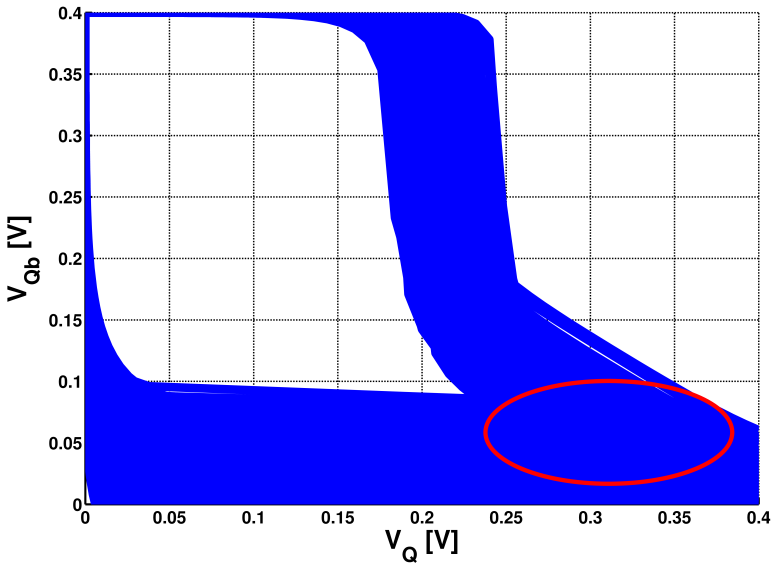


Figure 6.27: WSNM Butterfly plot at 400mV without VV_{DD} assist

At 1.2V the circuit was found to be monostable for all butterfly plots without the VV_{DD} write assis method after a the same Monte Carlo simulation.

6.4.3 On/off Current Ratio

Fig. 6.28 show the on/off-current ratio in all five process corners when the voltage supply is increased from 10mV to 1.2V at 25^{o}C and Fig. 6.29 shows how the temperature range -40^{o}C - 85^{o}C affects the on/off-current ratio at 400mV.

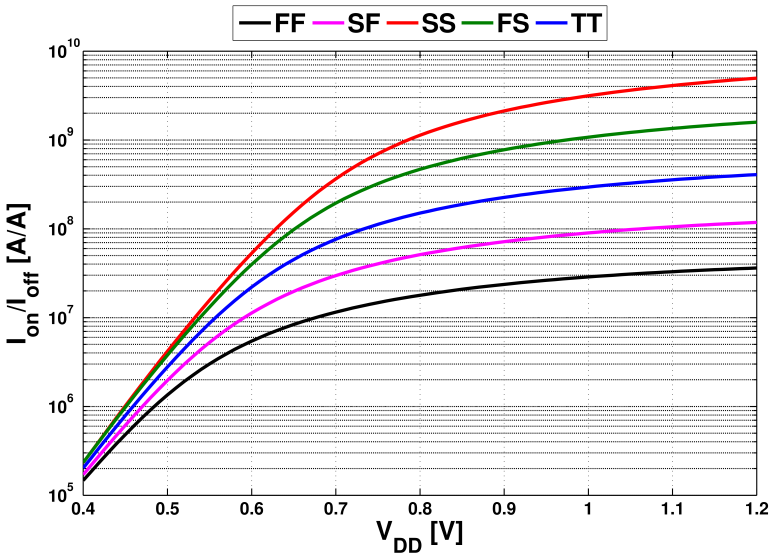


Figure 6.28: On/off-current ratio with increasing supply voltage

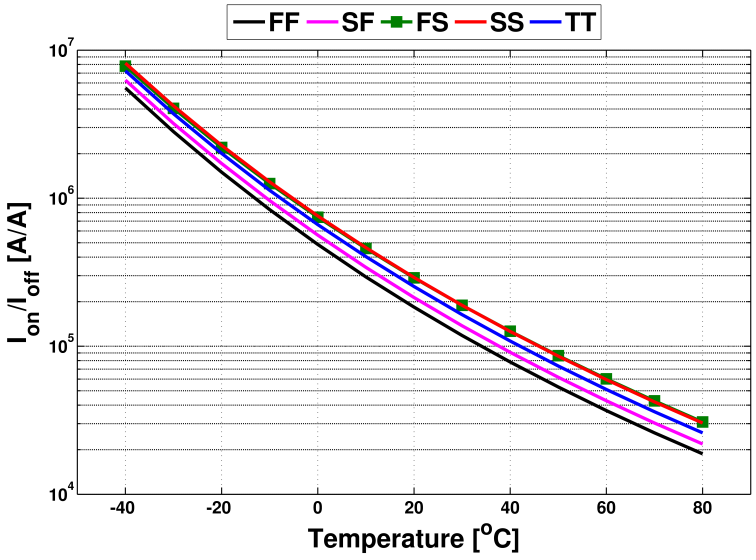


Figure 6.29: On/off-current ratio with increasing temperature at 400mV

6.4.4 SRAM Cell Leakage Power

Fig. 6.30 shows the leakage power consumption of a 10T SRAM cell after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V

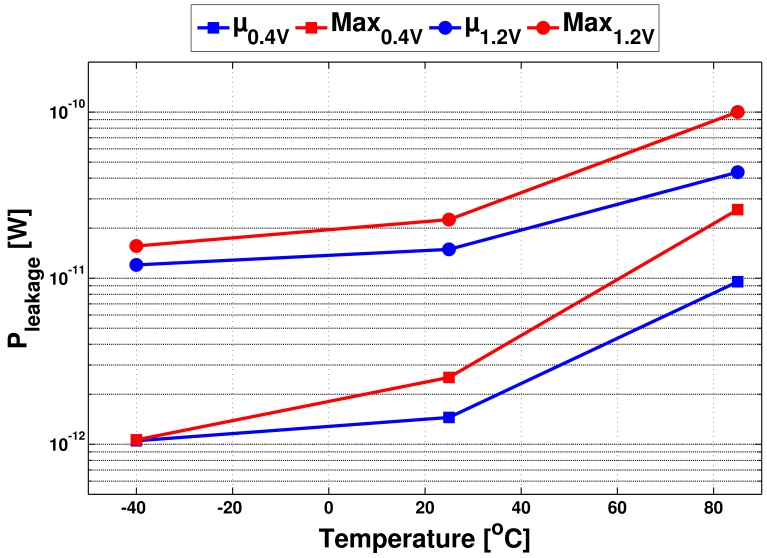


Figure 6.30: Leakage power of 10T SRAM cell

Fig. 6.31 shows the leakage power consumption of a 6T SRAM cell after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V

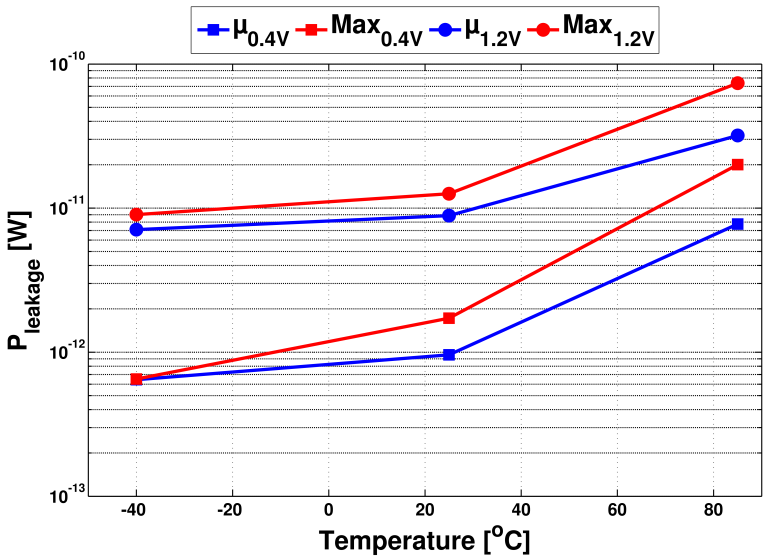


Figure 6.31: Leakage power of 6T SRAM cell

Fig. 6.32 shows theS RAM cell leakage power standard deviation (σ) relative to the means (μ) for the 10T and 6T SRAM cells after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

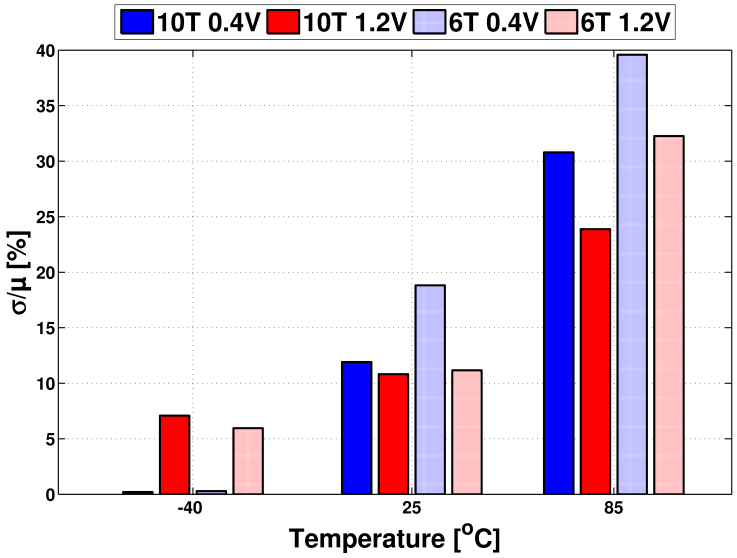


Figure 6.32: SRAM leakage power relative σ

6.4.5 Bitline Length and Read Delay

Table 6.33 shows the maximum frequency for different numbers of 10T SRAM cells in the SS, TT and FF corners for V_{DD} = 400mV and 1.2V.

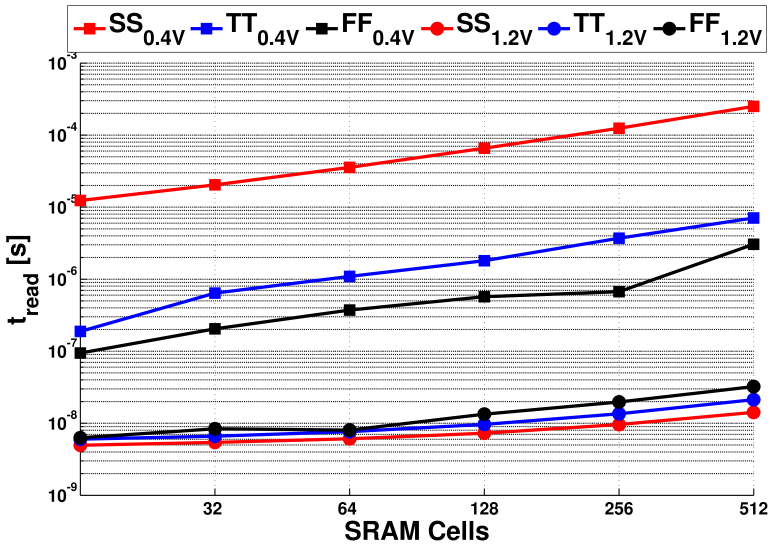


Figure 6.33: Read delay for different number of 10T SRAM cells

6.4.6 6T SRAM Read Disturb

Fig. 6.34 shows the internal node Q of a 6T SRAM cell in different corners when a logic "0" is read at 25^{o}C and V_{DD} = 400mV with a 32 kHz frequency and a bitline lelght of 32 SRAM cells.

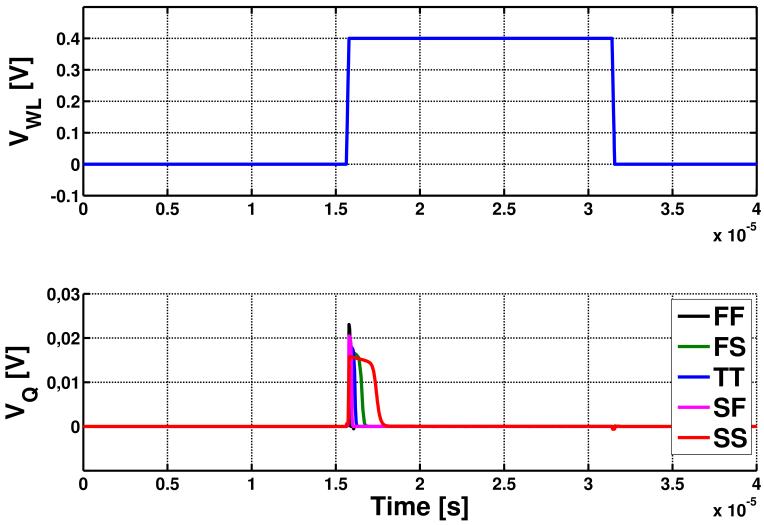


Figure 6.34: 6T read disturb

Fig. 6.35 shows the read disturb voltage in the 6T SRAM cell when reading a logic "0" for different temperatures and for V_{DD} = 400mV and 1.2V.

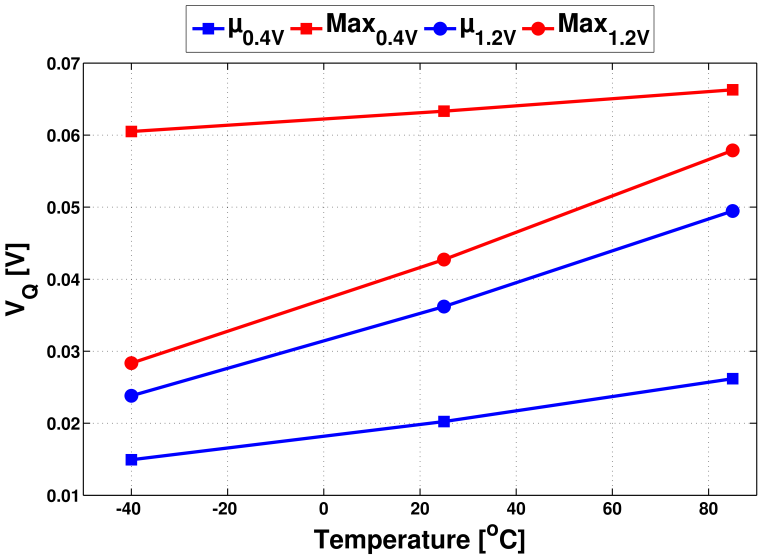


Figure 6.35: 6T read disturb voltage for different temperatures

Fig. 6.35 shows the read disturb voltage standard deviation (σ) relative to the mean (μ) in the 6T SRAM cell when reading a logic "0" for different temperatures and for V_{DD} = 400mV and 1.2V.

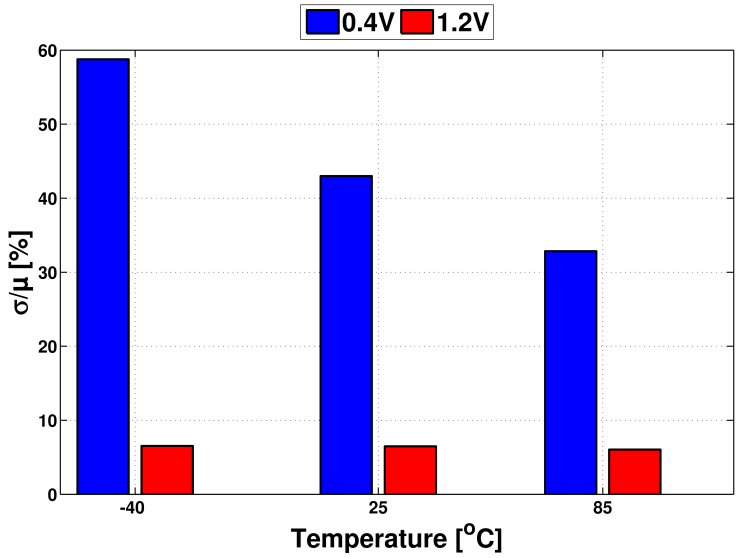


Figure 6.36: 6T read disturb voltage relative σ at different temperatures

Fig. 6.37 shows the read disturb voltage in the 6T SRAM cell when reading a logic "0" at 25°C in different corners with an increasing number of SRAM cells connected to the bitlines for V_{DD} = 400mV and 1.2v .

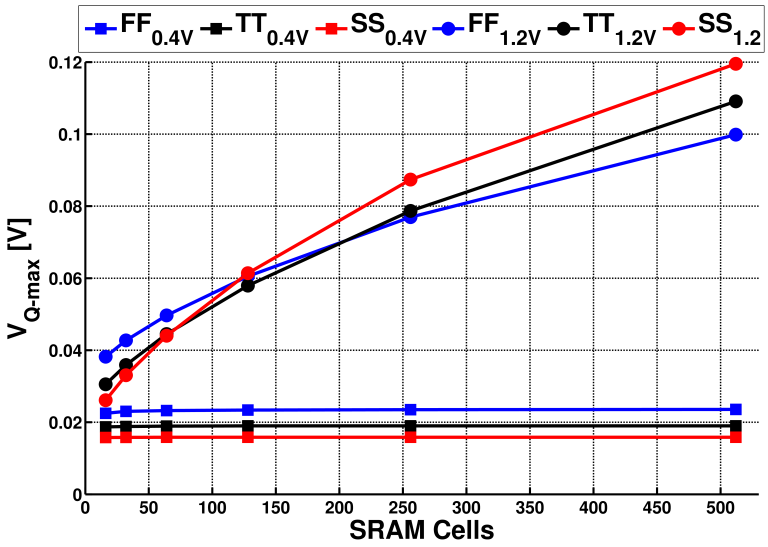


Figure 6.37: 6T read disturb voltage for different number of SRAM cells

Fig. 6.38 shows the read disturb voltage in the 6T SRAM cell when reading a logic ”0” at 25^{o}C where the pulse width of the WL-signal is decreased for V_{DD} = 400mV and 1.2V.

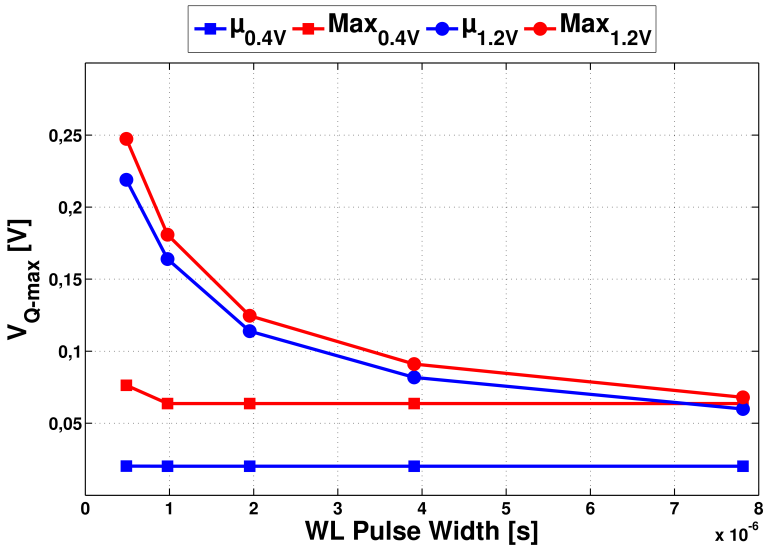


Figure 6.38: 6T read disturb voltage for different read frequenciess

Fig. 6.39 shows the read disturb voltage standard deviation (σ) relative to the mean (μ) in the 6T SRAM cell when reading a logic ”0” for different WL pulse widths at V_{DD} = 400mV and 1.2V.

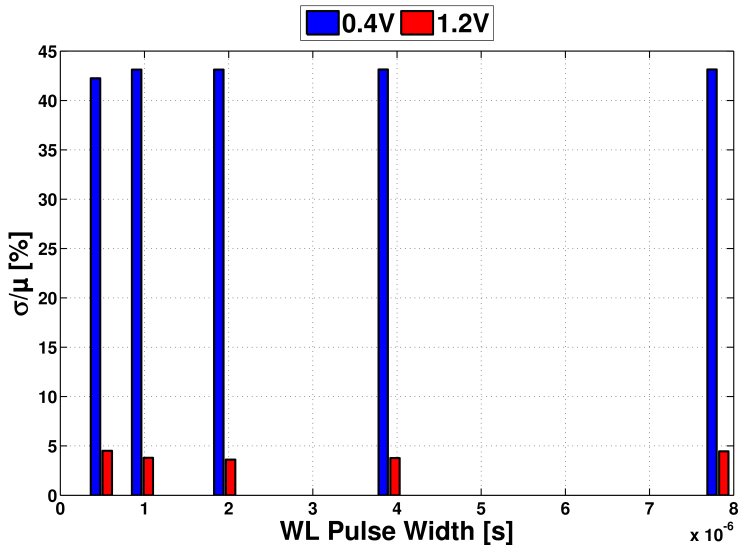


Figure 6.39: 6T read disturb voltage relative σ at differentpulse widths

6.5 Sense Amplifier

This section presents the simulation results from the test of the SA.

6.5.1 Read Access Yield

Fig. 6.40 and Fig. 6.41 shows the read access yield of the sense amplifier at -40^{o}C, 25^{o}C and 85^{o}C for V_{DD} = 400mV and 1.2V respectively .

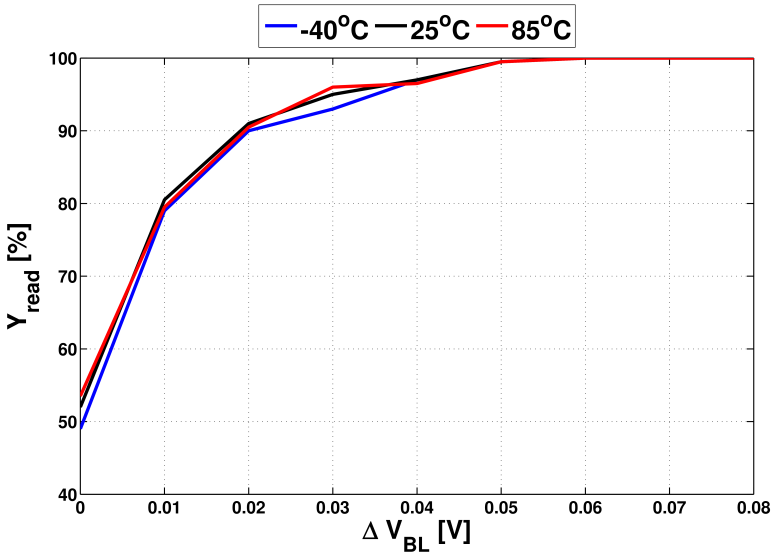


Figure 6.40: Sense amplifier read access yield distribution at 400mV

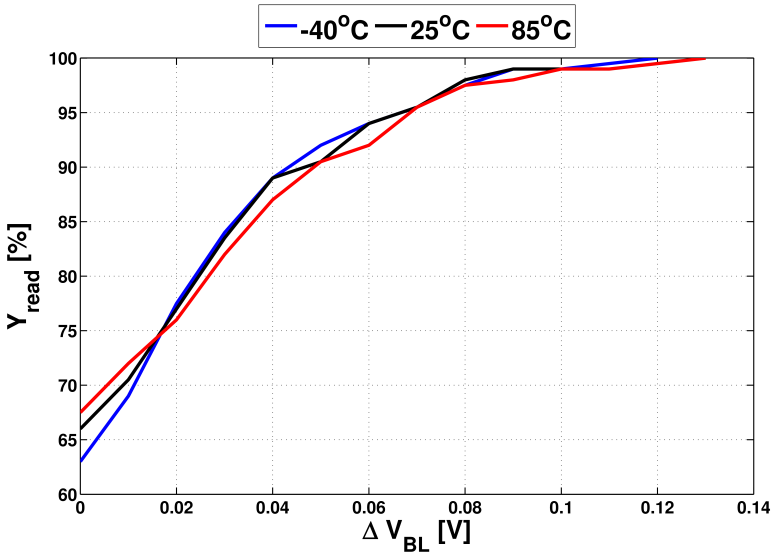


Figure 6.41: Sense amplifier read access yield distribution at 1.2V

6.5.2 Leakage Power Consumption

Fig. 6.42 shows the leakage power consumption of the SA after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

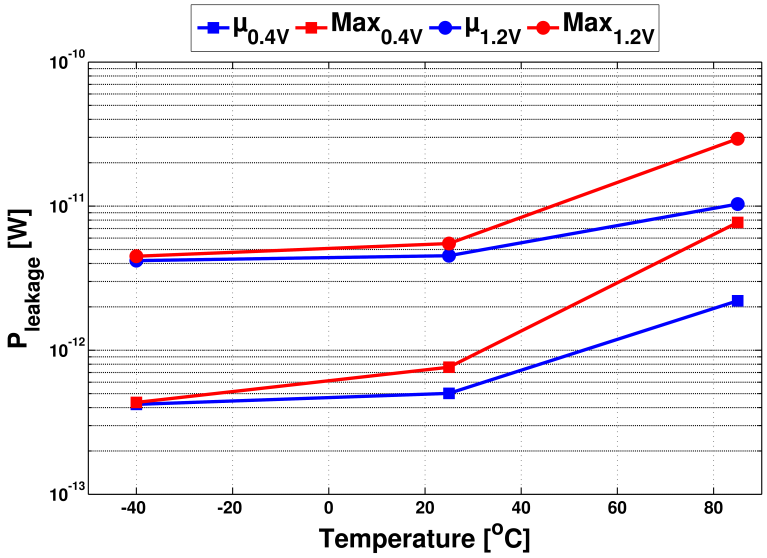


Figure 6.42: Leakage power consumption for SA at 400mV and 1.2V

Fig. 6.43 shows the leakage power standard deviation (σ) relative to the mean (μ) for the SA after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

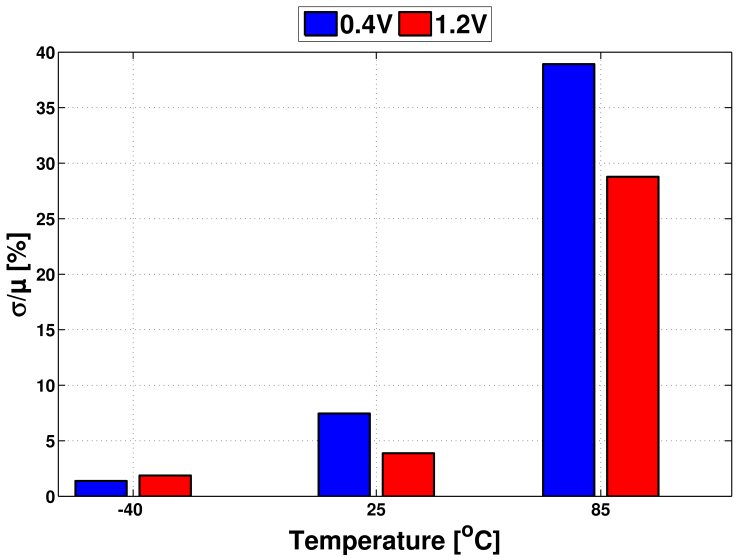


Figure 6.43: Relative σ for SA leakage power at 400mV and 1.2V

6.6 Enable Register Delay

Fig. 6.44 shows the mean (μ) and maximum propagation delay for the enable register after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

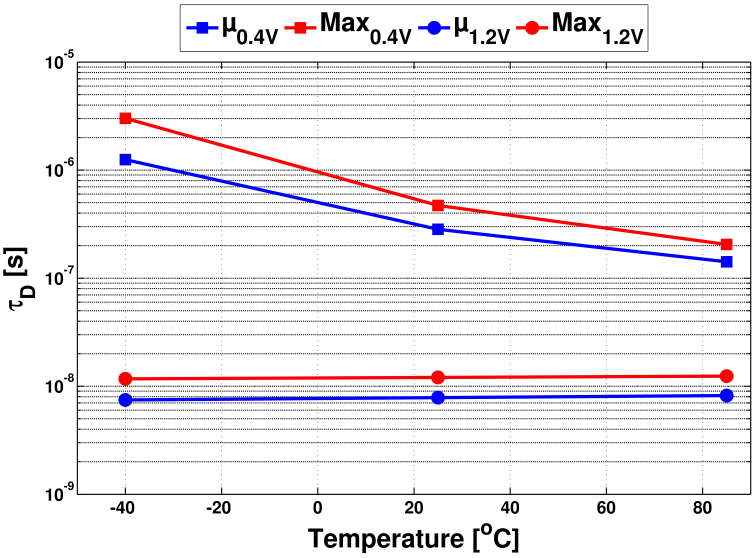


Figure 6.44: Propagation delay for enable register

Fig. 6.45 shows the delay standard deviation (σ) relative to the mean (μ) for the enable register after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

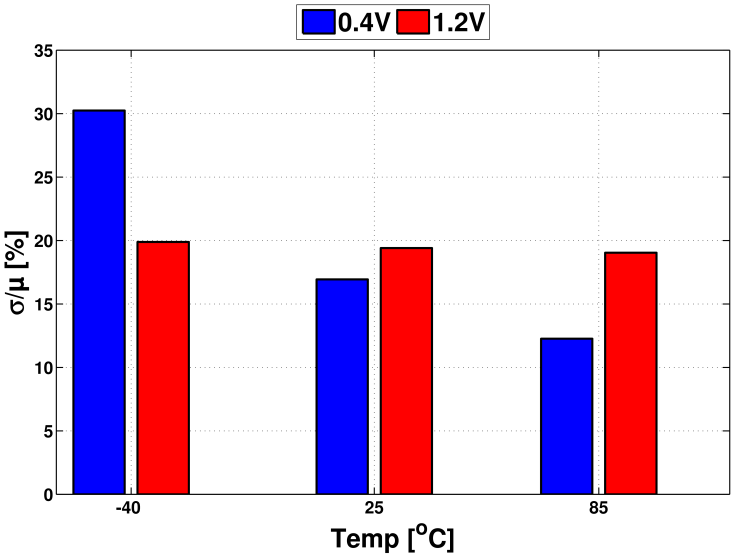


Figure 6.45: Enable register propagation delay relative σ

6.7 Transition Detector Pulse Width

Fig. 6.46 shows the mean (μ) and maximum pulse width generated by the transistion detector after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

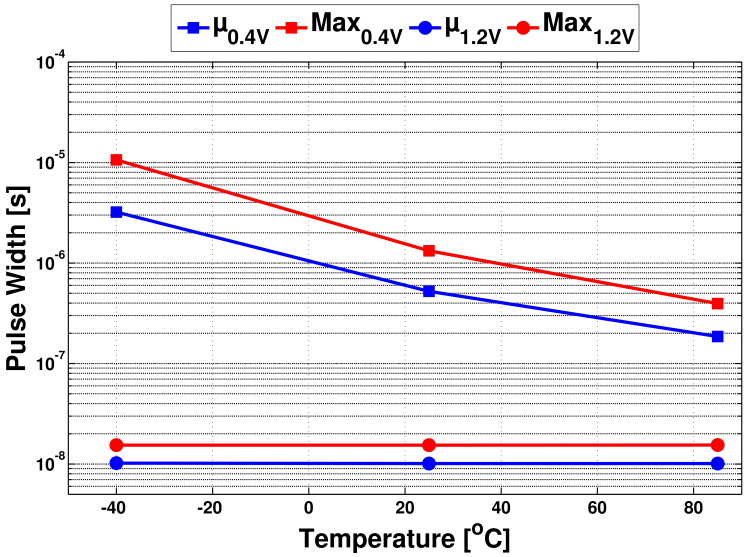


Figure 6.46: Pulse width from transistion detector

Fig. 6.47 shows the pulse width standard deviation (σ) relative to the mean (μ) for the transistion detector after 600 Monte Carlo iterations at V_{DD} = 400mV and 1.2V.

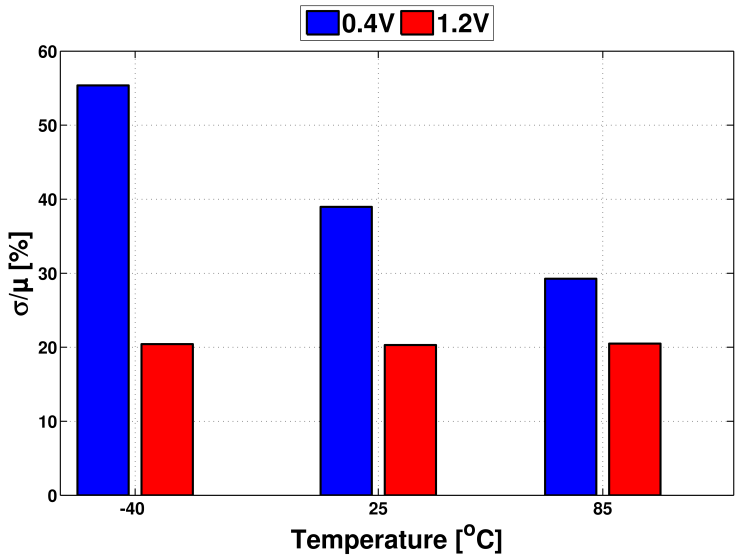


Figure 6.47: Transistion detector pulse width Relative σ

6.8 SRAM Architecture

Fig. 6.48 shows a transient plot of the 10T SRAM read ”0” cycle in the TT corner at 25^{o}C and V_{DD} = 400mV.

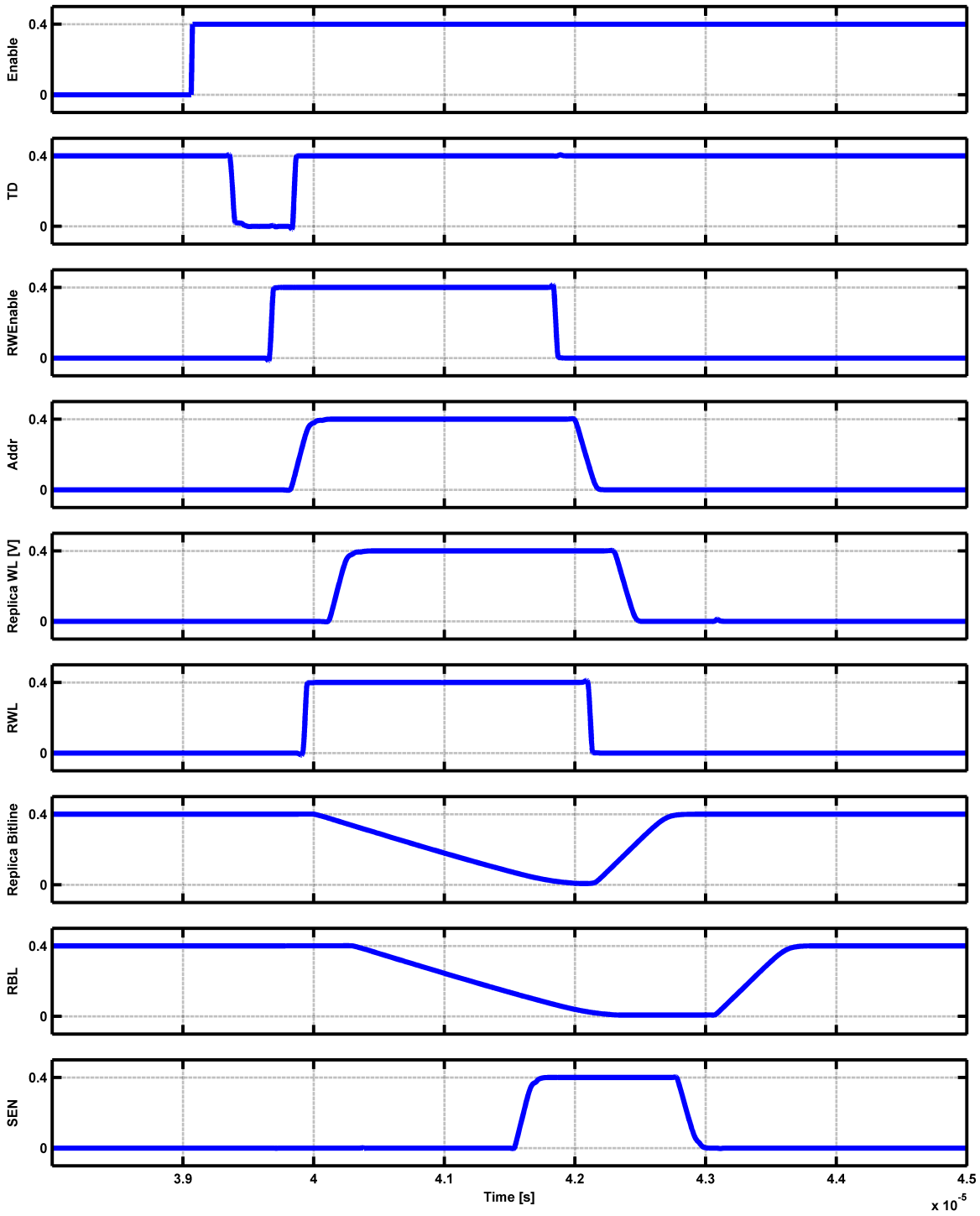


Figure 6.48: 10T Read ”0”

Fig. 6.49 shows a transient plot of the 6T SRAM read ”0” cycle in the TT corner at 25°C and V_{DD} = 400mV.

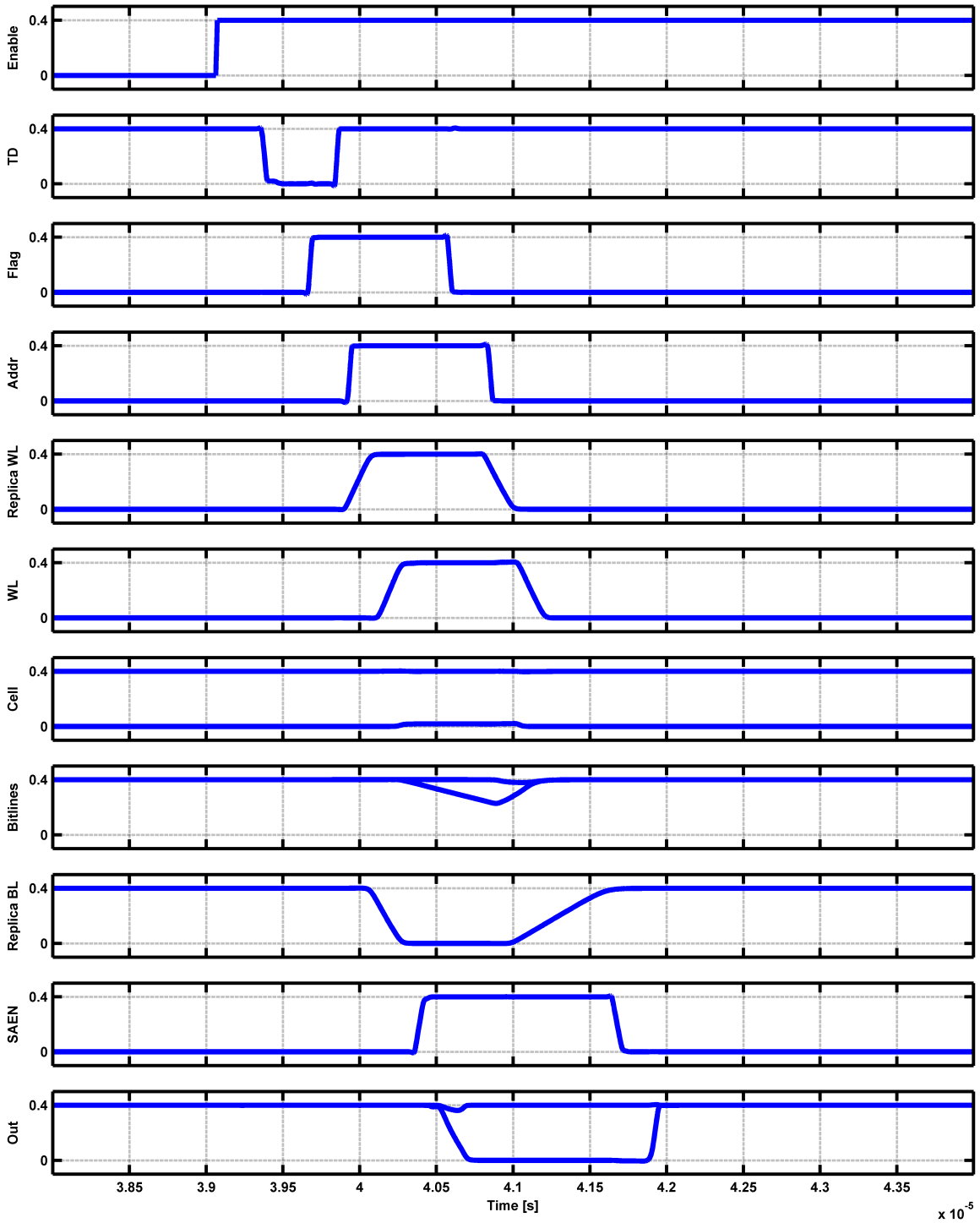


Figure 6.49: 6T Read ”0”

Fig. 6.50 shows a transient plot of the 10T and 6T SRAM write ”1” cycle in the TT corner at 25^{o}C and V_{DD} = 400mV.

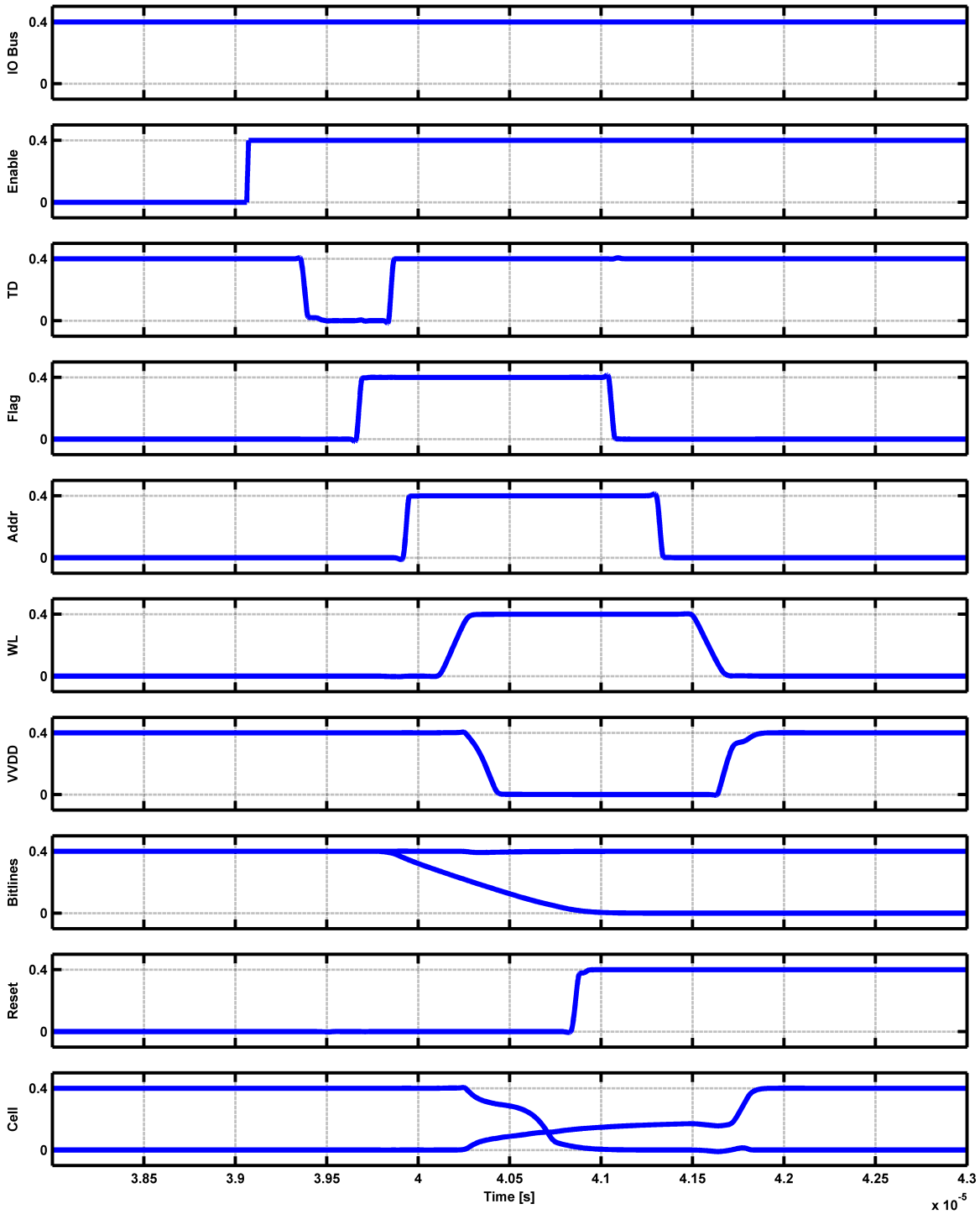


Figure 6.50: 10T and 6T Write ”1”

Table 6.2 shows the read "0" and write "1" cycle times for 10T SRAM architecture in the FF, FS, TT, SF and SS corners at V_{DD} = 400mV and 1.2V.

	-40^{o}C		**25^{o}C**		**-85^{o}C**	
0.4V	**Read** [μs]	**Write** [μs]	**Read** [μs]	**Write** [μs]	**Read** [μs]	**Write** [μs]
FF	5.54	3.28	1.44	0.87	0.70	0.43
FS	40.5	20.22	5.72	3.04	1.93	1.06
TT	26.19	14.59	4.64	2.67	1.74	1.02
SF	27.82	16.82	4.93	3.04	1.83	1.14
SS	144.3	76.68	16.68	9.24	4.83	2.75
1.2V	**Read** [ns]	**Write** [ns]	**Read** [ns]	**Write** [ns]	**Read** [ns]	**Write** [ns]
FF	27.11	20.15	32.12	23.24	36.47	25.86
FS	32.63	23.25	38.58	26.77	43.67	29.75
TT	32.31	23.47	38.12	27.0	43.07	29.94
SF	32.13	23.79	37.7	27.33	42.44	30.29
SS	40.49	28.59	47.52	32.8	53.43	36.29

Table 6.2: Cycle times for 10T SRAM architecture

Table 6.3 shows the read "0" and write cycle times for 6T SRAM architecture in the FF, FS, TT, SF and SS corners at V_{DD} = 400mV and 1.2V.

	-40^{o}C		**25^{o}C**		**-85^{o}C**	
0.4V	**Read** [μs]	**Write** [μs]	**Read** [μs]	**Write** [μs]	**Read** [μs]	**Write** [μs]
FF	3.19	3.44	0.89	0.91	0.44	0.45
FS	18.85	24.37	2.97	3.54	1.086	1.20
TT	13.89	15.22	2.63	2.79	1.042	1.07
SF	17.93	17.5	3.22	3.16	1.223	1.19
SS	70.27	82.58	8.53	9.74	2.768	2.87
1.2V	**Read** [ns]	**Write** [ns]	**Read** [ns]	**Write** [ns]	**Read** [ns]	**Write** [ns]
FF	21.05	20.57	24.7	23.78	39.63	26.55
FS	24.57	23.83	28.88	27.5	27.89	30.64
TT	24.61	24.01	28.84	27.68	32.64	30.76
SF	24.72	24.33	28.9	27.99	32.5	31.08
SS	30.07	29.32	35.23	33.74	29.63	37.36

Table 6.3: Cycle times for 6T SRAM architecture

Table 6.4 shows the total power consumption for the read "0" and write cycles for the 10T SRAM architecture in the TT corners at V_{DD} = 400mV and 1.2V.

	-40^{o}C		**25^{o}C**		**-85^{o}C**	
0.4V	**Read [nW]**	**Write [nW]**	**Read [nW]**	**Write [nW]**	**Read [nW]**	**Write [nW]**
FF	0.821	0.870	7.4	8.70	41.6	41.93
FS	0.802	0.935	6.17	9.71	21.73	24.19
TT	0.801	0.860	5.86	7.29	15.61	16.85
SF	0.800	0.817	5.76	6.36	14.83	15.66
SS	0.796	0.874	5.50	7.25	8.76	10.24
1.2V	**Read [nW]**	**Write [nW]**	**Read [nW]**	**Write [nW]**	**Read [nW]**	**Write [nW]**
FF	14.01	14.61	80.96	99.33	198.6	216.9
FS	9.916	10.52	68.75	87.39	122.3	140.9
TT	9.903	10.43	67.41	83.66	102.3	120
SF	9.901	10.37	66.99	81.76	99.9	117.3
SS	8.395	8.88	62.36	77.52	75.02	91.22

Table 6.4: Total power consumption of 10T SRAM architecture

Table 6.5 shows the total power consumption for the read "0" and write "1" cycles for the 6T SRAM architecture in the TT corners at V_{DD} = 400mV and 1.2V.

	-40^{o}C		**25^{o}C**		**-85^{o}C**	
0.4V	**Read [nW]**	**Write [nW]**	**Read [nW]**	**Write [nW]**	**Read [nW]**	**Write [nW]**
FF	0.783	0.736	10.59	8.40	42.17	40.37
FS	0.724	1.0	8.04	13.11	22.86	26.37
TT	0.741	0.722	8.38	7.04	17.89	16.42
SF	0.789	0.717	10.15	6.87	19.12	15.64
SS	0.723	0.781	7.59	7.55	11.0	10.36
1.2V	**Read [nW]**	**Write [nW]**	**Read [nW]**	**Write [nW]**	**Read [nW]**	**Write [nW]**
FF	13.64	12.26	139.7	94.46	247	209.4
FS	9.957	8.85	119.5	82.26	169.1	136.6
TT	10.15	8.81	124.3	79.89	156.1	116.4
SF	10.35	8.80	130	79.23	160.6	114.5
SS	8.765	7.53	115.6	74.25	127.4	88.69

Table 6.5: Total power consumption of 6T SRAM architecture

7. Layout

This chapter presents the physical layouts of the components in the SRAM array that were constructed for use in simulations. Parasitics from these layouts were extracted to add wire and via capacitance + resistance to get more realistic simulation results. The layout of remaining logic gates can be found in Appendix B.

7.1 Logic Gates

Fig. 7.1 shows the layout of the smallest inverter.

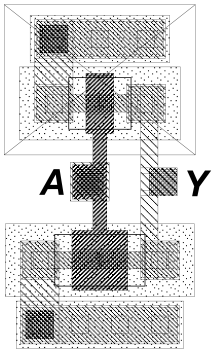


Figure 7.1: Layout of inverter

Fig. 7.2 shows the layout of the 2-input NAND gate.

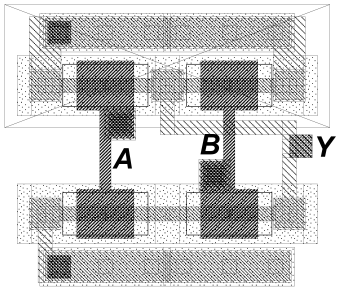


Figure 7.2: Layout of NAND gate

The physical area of the inverter is $1.78\mu m \times 3.2\mu m = 5.67\mu m^2$ and the physical area of the NAND gate is $3.7\mu m \times 3.2\mu m = 11.84\mu m^2$.

7.2 SRAM Cells

Fig. 7.3 shows the layout of the 6T SRAM cell.

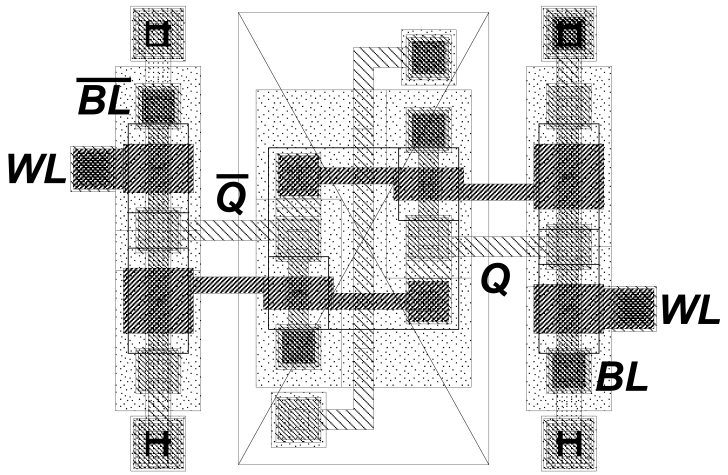


Figure 7.3: Layout of 6T SRAM cell

Fig. 7.4 shows the layout of the 10T SRAM cell.

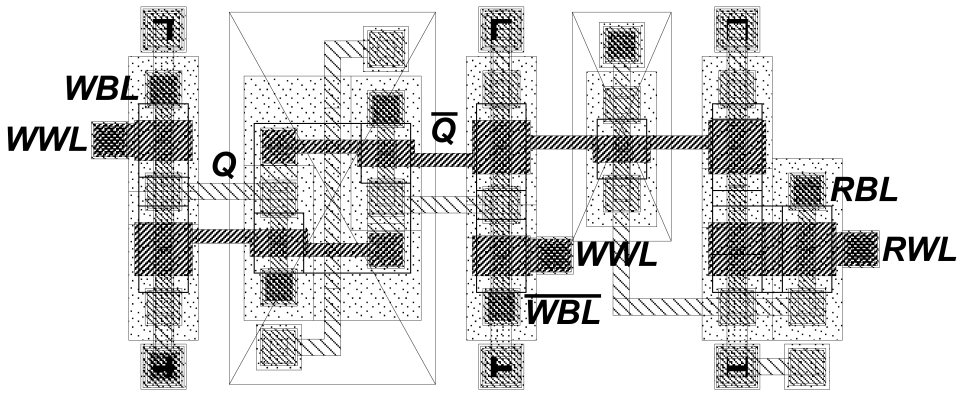


Figure 7.4: Layout of 10T SRAM cell

Both cells are designed as "thin cells" where all transistors are oriented in the same direction in order to reduce the effects of mismatch[28].

The physical area of the 6T SRAM cell is $4.68\mu m \times 3.71\mu m = 17.36\mu m^2$ and the physical area of the 10T SRAM cell is $7.65\mu m \times 3.71\mu m = 28.38\mu m^2$.

7.3 Sense Amplifier

Fig. 7.5 shows the layout of the SA used with 6T SRAM cells.

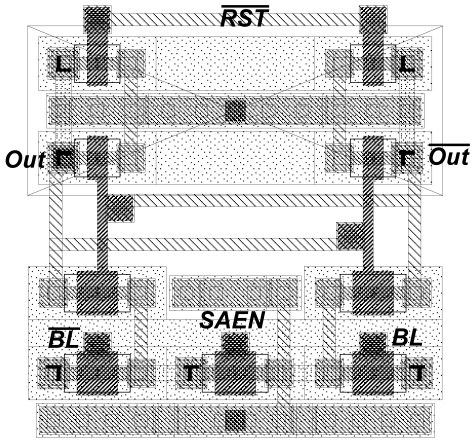


Figure 7.5: Layout of Sense Amplifier for 6T SRAM cells

The physical area of the SA is $5.3\mu m \times 5.5\mu m = 30.25\mu m^2$. The width is slightly wider than the 6T SRAM cell in order to orient all transistors in the same direction, which leads to some wasted space between SRAM columns.

The outputs of the SA must be loaded with clocked inverters or transmission gates (not shown) to isolate the internal nodes from the IO bus.

7.4 Write Driver

Fig. 7.6 shows the layout of the write driver. All transistors are oriented in the same direction to reduce the effect of mismatch.

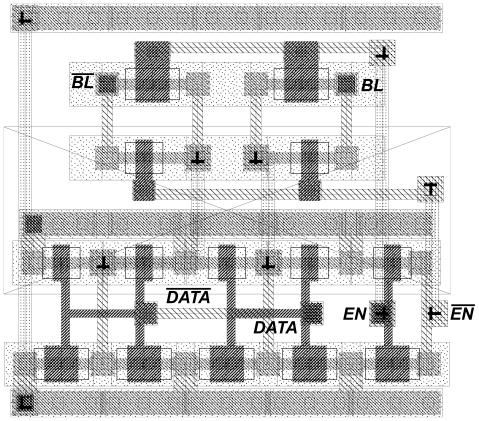


Figure 7.6: Layout of write driver

The physical area of the write driver is $6.88\mu m \times 6.38\mu m = 43.89\mu m^2$. The width is slightly wider than the 6T SRAM cell in order to orient all transistors in the same direction, which leads to some wasted space between 6T SRAM columns.

7.5 Wordline Drivers

Fig. 7.7 shows the layout of the 6T wordline driver.

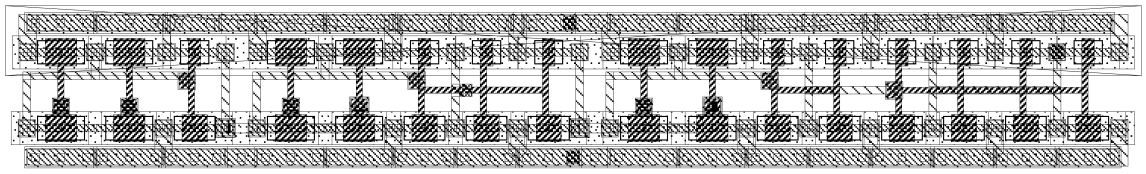


Figure 7.7: Layout of 6T wordline driver

Fig. 7.8 shows the layout of the 10T wordline driver.

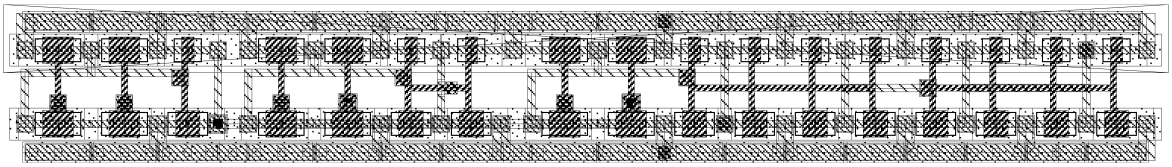


Figure 7.8: Layout of 10T wordline driver

The physical area of the 6T wordline driver is $23.42\mu m \times 3.2\mu m = 74.95\mu m^2$ and the physical area of the 10T wordline driver is $24.74\mu m \times 3.2\mu m = 79.17\mu m^2$. This layout applies a single wordline driver for every SRAM row.

7.6 Precharge Circuit

Fig. 7.9 shows the layout of the 1precharge circuit.

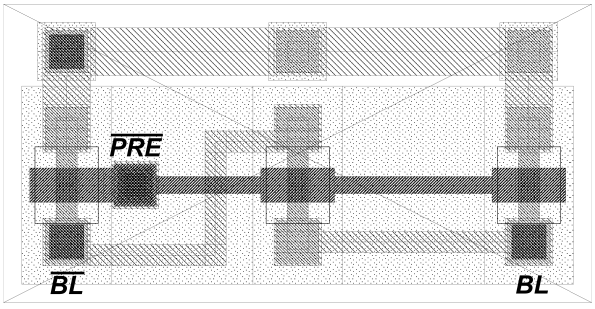


Figure 7.9: Layout of precharge ciruit

The physical area of the precharge circuit is $4.47\mu m \times 2.26\mu m = 10.1\mu m^2$.

8. Discussion

This chapter contains discussions and reflections of the simulation results and design of the SRAM architecture.

8.1 Transistor Type

At the start of the project high-V_{th} (HVT) transistors were chosen because they have lower leakage currents compared to reguar-V_{th} (RVT) and low-V_{th} (LVT) transistors. The trade-off is that the driving on-current is lower as well. In a 65nm process the differences in driving strength β has been shown to be as much as 18 times between LVT and HVT transistors[11]. Since most of the SRAM is in an idle state at any given time HVT transistors were the best choice for leakage reduction, but performance could be increased by utilizing RVT transistors as they would be able to drive capacitive loads more effectively. Another choice could be to use a combination of HVT and LVT transistors to increase the performance of driving circuits while retaining the leakage reduction in the SRAM cells themselves. The latter option provides some difficulties in terms of area as a combination of LVT and HVT transistors require the insertion of a guard-ring around HVT transistors to protect them from latch-up. Using two types of transistors would also increase the monetary cost of producing a chip.

8.2 Supply Voltage

The PRCP results in Fig. 6.1 indicates that the PRCP global minimum lies very close to the threshold voltage of the transistor. $V_{DD} = 400$mV was chosen as the PRCP is not that much higher for that value. At room temperatures and above the circuit does perform well withing the constraints of a 32kHz clock cycle, but as shown by the cycle times in Table 6.2 and 6.3 the read cycle time speed increases by a factor of 5-8 when moving from 25^{o}C to -40^{o}C. If process variations are severe the SRAM might need up to 5 32kHz clock cycles to complete the read operation with 10T cells and 3 clock cycles with 6T cells. This indicates that temperature is a very important factor when it comes to reliable operation. Increasing the supply voltage to 500mV or above might allow both the 10T and 6T cells to complete read and write operations within a 32kHz clock period in the SS corner, but overall power consumption would also increase. Another approach could be to utilize some kind of dynamic voltage regulation to compensate for the decreased temperatures and process variations.

The specification from Atmel Norway AS states that a PVT compensating regulator would be used with any potential product. Fig. 8.1 shows a simplified model of such a regulator suggested by Atmel Norway AS. A stable and reliable current generator

provides a bias current tuned for 400mV at 25^{o}C into the drain terminals of two diode-connected transistor. Both PMOS and NMOS transistors are used so the regulator will compensate for the transistor type that is most affected by PVT variations. When the temperature and process variations are applied to the circuit the threshold voltage of the diode-connected transistors will increase or decrease and the amplifier generates a new supply voltage.

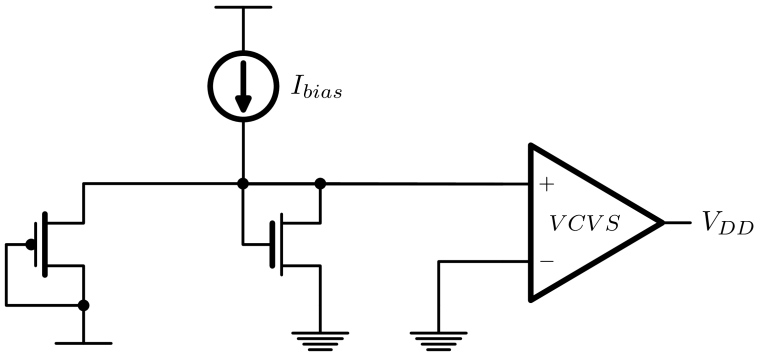


Figure 8.1: Simplified PVT compensating voltage regulator

Since the amplifier will also have imperfections it will have an offset error on the output voltage. The specification from Atmel Norway AS estimated the error would be $\approx \pm 10\%$. This type of regulation may not be the best option for subthreshold voltages because of the output error. With $V_{DD} = 400$mV a 10% error can result in a supply voltage from 360mV to 440mV which can decrease performance even further.

8.3 Logic Gates

The results from the logic gate sizing process showed that increasing transistor lengths was a more effective method of balancing logic gates compared to increasing the widths which is the usual way of balancing logic gates at superthreshold voltages[26]. According to equation 2.7 increasing the width and length should be equally effective methods of increasing the driving strength of a transistor, but during simulations it was discovered that increasing the widths made the balance statistically worse unless the width was larger or equal to $10W_{min} = 1.6\mu m$. The cause of this phenomenon is the RNCE and is consistent with previous research[11]. Because of the RNCE it was also deemed more area-effective to increase the m-factor for driving inverters instead of increasing the widths.

The results for the NAND gate shows that the VTC deviation varies between -10% and 25% at $V_{DD} = 400$mV depending on the input combination and temperature and the results in Appendix A shows that this range is consistent for all logic gates. The logic gates exhibited valid VTCs for all Monte Carlo iterations meaning none of the logic gates failed to operate correctly.

The delay of the NAND gate increases for lower temperatures. At 400mV there is a 120% mean delay increase when the temperature is lowered from 25^{o}C to -40^{o}C while the delay decrease for $V_{DD} = 1.2$V for the same temperatures is only 20%. The relative-σ is also 25-12% higher at subthreshold voltages and changes with temperature,

while at V_{DD} = 1.2V the relative-σ is relatively unchanged across temperatures. The rise and fall times exhibit similar characteristics, and the fall-time is the shows the worst behavior because of the NMOS stack in the PDN.

The NAND gates also consumes $\approx$ 3-10 times less mean leakage power at V_{DD} = 400mV compared to 1.2V and 46-106 times less mean total power with a 32kHz toggling frequency over the temperature range -40^{o}C to 85^{o}C. As expected reducing the supply voltage is a very effective method of reducing the power consumption.

The propagation delay naturally increase for larger fan-outs and increasing the fan-out from 2 to 128 increases the delay by several orders of magnitude and PVT variations can cause the delay to approach 0.1ms for very large fan-outs. This indicates that the decoder delay will be limited by the output AND gates of the decoder. 128 AND gates must toggle in order to start a read and write operation so to reduce the delay cause by the large fan-out the decoder outputs was divided into partitions of 16 AND gates, each partition driven by its own driving chain connected to the *RWEnable*-signal.

8.4 SRAM Cells

As expected the gated-read buffer improves the read SNM of the 10T cell. At 400mV the mean 10T read SNM is $\approx$ 170mV which is a 60-70% increase over the 6T cell. At 1.2V the increase is even higher at 60-100%. The Write SNM is the same for both SRAM cells and the VV_{DD} write-assist ensures the butterfly plots are monostable in all iterations of the Monte Carlo analysis. As shown in Fig. 6.27 the write SNM is higher without the write-assist, but the potential of unwriteable SRAM cells becomes a problem. If a voltage-boosting write-assist method had been implemented the WSNM could probably be increased with guaranteed writeability, but as mentioned in chapter 3 this would require the design of a charge pump or voltage reference and level-shifters which would increase the complexity of the design and introduce potential dangers in terms of voltage scaling.

The bitline length becomes a limiting factor of the SRAM design long before the on/off-current ratio at V_{DD} = 400mV. The read delay of the 10T cell increases for linger bitline lengths, but PVT variations cause the variation to be higher compared to at V_{DD} = 1.2V. A bitline length of 128 cells was chosen because the delay in the TT corner was $2\mu s$ and the specification from Atmel Norway AS stated compensation would be applied to counteract the effects of low temperature. As shown by the operation cycle times in Table 6.2 and 6.3 the read times at low temperatures are very long. To reduce them the bitlines could be divided into partitions and selected using MUXes in a hierarchical bitline structure[12] as shown in Fig. 8.2. This approach does not require any complex circuitry, but will have a significant impact on the area. Splitting the read bitline into two partitions would require one multiplexer per bit in a word and precharge circuits for both partitions. Assuming the smallest 2-to1 multiplexer available this would amount to $32 \cdot 7 = 224$ extra transistors for a 4K SRAM array with 32 bits per word. Assuming the length of each bitline partition can be expressed using powers of two the amount of multiplexers needed for the hierarchical read operation can be calculated with equation 8.1.

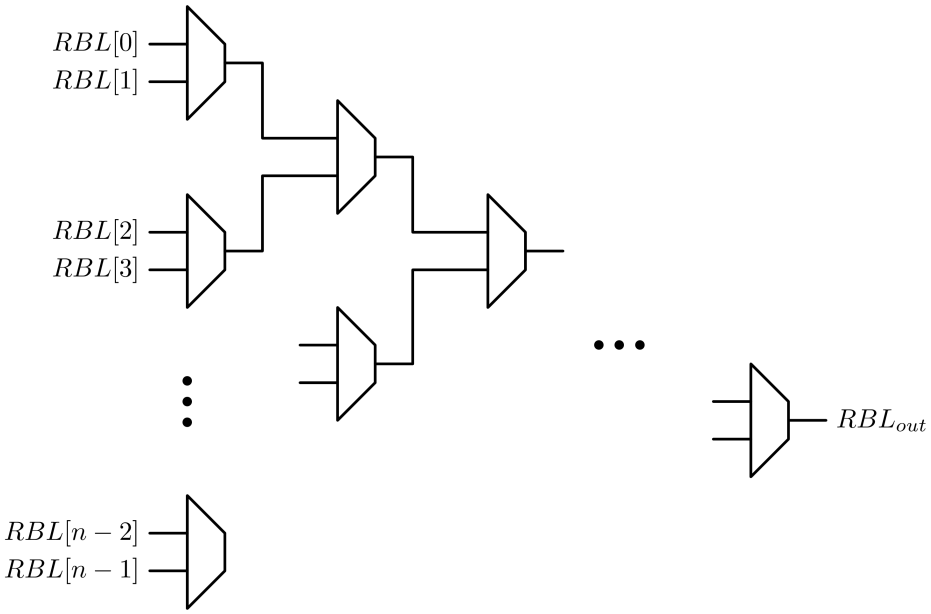


Figure 8.2: Hierarchical-read access with 2-to-1 multiplexers

$$\frac{N_{par}^2 - N_{par}}{2} \tag{8.1}$$

Where N_{par} is the number of bitline partitions. This approach adds a lot of area overhead to the design if the bitlines are split into many small partitions and design-wise it will also become difficult to fit many multiplexers within the width constraint of a single SRAM cell.

The 10T cell consumes 1.2 to 1.6 times more mean leakage power compared to the 6T cell depending on the temperature. Leakage power increases with temperature and the leakage difference between 10T and 6T cell is smallest at 85^oC. Both cells experience a mean leakage power reduction of one order of magnitude when reducing V_{DD} from 1.2 to 400mV, but there is only a reduction of 4 times at 85^oC because of the increased leakage at 400mV.

The disturb voltage experienced by the 6T cell is dependent on several factors. The peak value increases by 10-15mV over the temperature range -40^oC to 85 ^{o}C, but variations at $V_{DD} = 400$mV is a lot larger compared to 1.2V. The bitline length and WL pulse width has little to no effect on the disturb voltage at subthreshold voltages, but at superthreshold voltages the peak disturb voltage increase with the bitline length and with shorter WL-pulses. As these results indicate the 6T SRAM cell is less prone to dynamic failures and is mostly affected by temperature at subthreshold voltages while at superthreshold the opposite relationship is true. This indicates that the 6T cell is more difficult to implement in a system utilizing voltage scaling as more sources for dynamic read failures are introduced when the supply voltage is increased.

8.5 Sense Amplifier

Monte Carlo simulations with 200 iterations for each temperature indicates that the read access yield reaches 100% at 400mV when the differential bitline voltage ΔV_{BL} is 60mV or larger. Inserting this into equation 3.1 the number of discharging replica bitline cells must be set to 2 or 3 assuming the trip voltage of the sensing inverter is $V_{DD}/2 \pm 10\%$. For simulations of the SRAM architecture the number was set to 3 as the delay of driving the $SAEN$-signal will allow the bitline to discharge longer before the SA is activated. At $V_{DD} = 1.2$V a differential bitline voltage of ΔV_{BL} is 130mV or greater is needed to reach 100% read access yield and the number of discharging replica bitlines must be set to 4-5. This indicates that the number of discharging replica cells should be programmable depending on the supply voltage. The SRAM architecture should have a programmable register controlling the number of replica cells that can be opened by the replica wordline. The number of cells increase for high supply voltage to increase speed.

The leakage power consumption of the SA is lower than the leakage of the 6T SRAM cell due to the added footer transistor and increased stacking of transistors. When the $SAEN$-signal is not active both "branches" in the SA consists of a transistor stack of three transistors and the combined current of both enters the drain of the footer transistor which is turned off.

8.6 SRAM Architecure

Fig. 6.48, 6.49 and 6.50 shows the 10T read "0" operation, 6T read "0" operation and write "1" operation respectively. As the cycle times show the 6T architecture provide faster read operations, but both architectures suffer from severe increases in delay at -40^oC and variable cycle times caused by process variations at $V_{DD} = 400$mV. This further emphasizes the need for supply voltage compensation to combat the effect of PVT variations in subthreshold circuits. The PVT variations also increase the delays of the TD pulse width and enble register propagation delay and together with the delay of the decoder they can amount to a large portion of the overall delay of the SRAM operations.

The power consumption of both circuits increase with temperature which indicates that leakage current become a more dominating contribution at higher temperatures. Moving from $V_{DD} = 1.2$V to 400mV provides a 4-18 times reduction in power for both architectures. Since both architectures use an almost identical write method the write cycle times and power consumption is close between architectures but not entirely due to some differences in control logic. The total power of the 6T architecture is on average higher than the 10T architecture during read operations because of the faster discharge of the dual replica bitline and the dynamic power consumption of the SAs.

9. Conclusion

The simulation results show that the SRAM topology implemented in the Atmel 130nm CMOS process is viable for subthreshold voltages. The 10T cell is more robust with a 60-100% larger static noise margin compared to the conventional 6T cell, but the increased robustness comes at a cost of increased leakage power consumption and increased area. The 10T cell is physically 64% larger than the 6T cell and also requires more time to complete a read ”0” operation due to the single-ended nature of the read buffer. A full bitline swing is required to complete the 10T read ”0” operation while the 6T cell only require a lower differential bitline voltage. The offset voltage of the sense amplifier at 400mV is relatively large so the speed gains of using 6T cells is somewhat diminished, but still faster compared to 10T cells. The read operation of the 6T cell also creates a disturb voltage in one of the internal nodes of the SRAM cell and the its magnitude is dependent on several factors. The amount of SRAM cells connected to the bitline, the width of the wordline signal and the severity of process and temperature variations all affect the disturb voltage and their impact are grater at high supply voltages, making it difficult to asses the yield in systems with voltage scaling. The 10T cell uses a read buffer to decouple the read and write operation and do not encounter this problem and this makes the 10T cell more predictable and the safest choice for future implementations with voltage scaling. If voltage scaling is not used the 6T cells becomes a more lucrative option because the fixed supply voltage gives a more predictable yield and the area and leakage power consumption are reduced.

Simulation of both architectures shows that the power savings of moving from 1.2V to 400mV are within the range of 4-18 times depending on the severity of process variations and temperature, but these saving comes at the price of increased delays in both implementations. The active power consumption of the 6T implementation is greater because of the dynamic power consumption of the sense amplifier, but the read ”0” speed is approximately 2-3 times faster compared to the 10T implementation. The lowest power savings occur at high temperatures due to increased leakage currents. The largest savings occurs at low temperatures, but the performance is degraded to such a degree that the 10T implementation requires 5 32kHz clock cycles to complete a read ”0” operation while the 6T implementation requires 3 at -40^{o}C in the SS process corner. To combat the extreme degradation in speed the supply voltage must be raised either permanently or through some kind of dynamic supply voltage compensation to perform a read operation within a 32kHz clock cycle.

This thesis has shown that it is viable to implement a subthreshold SRAM architecture with 10T and 6T cells in the Atmel 130nm CMOS process and that 10T SRAM cells are more robust and more predictable with voltage scaling in terms of yield. Some important effects of applying voltage scaling have also been explored. 1Reducing the power supply to such an extent reduces speed of the SRAM, but the self-timed architecture approach ensures that read and write operations will finish after 3-5 32kHz clock cycles with the worst-case process variations and temperature. Some form of voltage compensation must be applied to increase performance at low temperatures.

10. Future Work

This chapter present some ideas for further work on the SRAM architecure.

10.1 Prototype Chip

Both the 10T and 6T SRAM architectures should be taped out and physically tested on a chip. While the simulation results indicate that both architectures are viable options there might be some aspect of physical production that will affect the SRAM architecures more than the simulations results have accounted for.

10.2 Output Level Shifters

During an earlier specialization product a subthreshold ring oscillator using inverters was made, but when the supply voltage was dropped below a voltage of 600mV the level shiftes which converted signals from the low-power domain to output logic levels failed, meaning there was no way of knowing if the oscillator worked for the voltages it was designed for.

10.3 Self-Testing Voltage Compensation

The proposed PVT compensating regulator in Fig. 8.1 might not be a good solution for subthreshold circuits because of the output offset error. A self-testing voltage compensation scheme would be a better option. Self-testing should be performed on a minimum sized circuit replicating the critical path of the system. Small devices are affected more by process variations and will increase margins for the actual circuit. As an example: a programmable regulator is able to output supply voltages from 300mV to 600mV with a step size of 50mV. When the system boots the supply voltage is set to 600mV and a state machine gradually reduces V_{DD} until a read ”0” operation on a minimum sized replica SRAM bitline fails. When it fails the state machine will revert to the previous successful voltage level. A test must be performed periodically to adapt to the environment.

The interval of the test should be programmable by the developer as he/she will most likely know more about the environment in which the system will be used. A ultra-low voltage system measuring temperature will most likely not require frequent re-tests as temperature rarely changes rapidly, but component in a smartphone can experience frequent temperature changes when the user takes the smartphone out from his/her pocket into the winter cold.

References

[1] Fritz Leuenberger and Eric Vittoz. Complementary-MOS Low-Power Low-Voltage Integrated Binary Counter. *Proc. IEEE*, 57(9), September 1969. 2

[2] E. Vittoz and J. Fellrath. New Analog CMOS IC'S Based on Weak Inversion Operation. In *Solid State Circuits Conference, 1976. ESSCIRC 76. 2nd European*, 1976. 2

[3] Ambiq Micro. `http://ambiqmicro.com/`, January 2014. 2

[4] Ole Samstad Kjøbli. Ultra-Low Voltage SRAM Cell. Specialization project, Norwegian University of Science and Technology (NTNU), December 2013. 2, 19

[5] Preeti Ranjan Panda et. al. *Power-efficient System Design.* Springer Science+Buisness Media, 2010. 5

[6] Kristian Granhaug and Snorre Aunet. Six Subthreshold Full Adder Cells Characterized in 90 nm CMOS Technology. In *Design and Diagnostics of Electronic Circuits and systems, 2006 IEEE*, 2006. 5, 6

[7] Jawar Singh et. al. *Robust SRAM Design and Analysis.* Springer Science+Buisness Media New York, 2013. 7, 8

[8] Evert Seevinck et. al. Static-Noise Margin Analysis of MOS SRAM Cells. *IEEE J. Solid-State Circuits*, 22(5), 1987. 9, 47

[9] Benton Highsmith Calhoun and Anantha P. Chandrakasan. A 256kb Sub-threshold SRAM in 65nm CMOS. In *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, 2006. 10, 19

[10] Siva G. Narendra and Anantha Chandrakasan. *Leakage in Nanometer CMOS Technologies.* Springer Science+Buisness Media, 2006. 11

[11] Massimo Alioto. Ultra-Low Power VLSI Circuit Design Demystified and Explained: A Tutorial. *IEEE Trans. Circuits Syst. I, Reg. Papers*, 59(1), 2012. 11, 12, 13, 15, 33, 43, 89, 90

[12] Alice Wang et. al. *Sub-Threshold Design for Ultra Low-Power Systems.* Springer Science+Buisness Media, 2006. 12, 13, 36, 91

[13] M.B. Taylor. A landscape of the new dark silicon design regime. *Micro, IEEE*, 33(5), Sept 2013. 13

[14] A. Bellaouar et. al. Supply Voltage Scaling for Temperature Insensitive CMOS Circuit Operation. *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, 45(3), 1998. 15

[15] Naveen Verma and Anantha P. Chandrakasan. A 256kb 65nm 8t subthreshold sram employing sense-amplifier redundancy. *IEEE Journal of Solid-State Circuits*, 43(1), 2008. 20

[16] Ik-Joon Chang et. al. A 32 kb 10t sub-threshold sram array with bit-interleaving and differential read scheme in 90 nm cmos. *Solid-State Circuits, IEEE Journal of*, 44(2), Feb 2009. 20

[17] IBM. Applications note: Understanding static sram operation, 1997. 21

[18] Andrei Pavlov and Manoj Sachdev. *CMOS SRAM Circuit Design and Parametric Test in Nano-Scale Technologies.* Springer Science+Buisness Media, 2008. 22, 23, 38, 40

[19] B.S. Amrutur and M.A. Horowitz. A replica technique for wordline and sense control in low-power sram's. *Solid-State Circuits, IEEE Journal of*, 33(8), Aug 1998. 22

[20] Yi Li et. al. An area-efficient dual replica-bitline delay technique for process-variation-tolerant low voltage sram sense amplifier timing. *IEICE Electronics Express*, 11(3), 2014. 23

[21] Yoshimoto M. et. al. A divided word-line structure in the static ram and its application to a 64k full cmos ram. *Solid-State Circuits, IEEE Journal of*, 18(5), Oct 1983. 25

[22] Jonathan Edvard Bjerkedok. Subthreshold real-time counter., 2013. 34

[23] Kobayashi T. et. al. A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture. *Solid-State Circuits, IEEE Journal of*, 28(4), Apr 1993. 37, 50

[24] Joseph F. Ryan. and Benton. H Calhoun. Minimizing Offset for Latching Voltage-Mode Sense Amplifiers for Sub-Threshold Operation. In *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, 2008. 37, 38

[25] Dipanja Sengupta and Resve Saleh. Power-delay metrics revisited for 90 nm cmos technology. In *Quality of Electronic Design, 2005. ISQED 2005. Sixth International Symposium on*, March 2005. 41

[26] John P. Uyemura. *Introduction to VLSI Circuits and Systems.* John Wiley & Sons, 2002. 44, 90

[27] Mohamed H. Abu-Rahma and Mohab Anis. *Nanometer Variation-Tolerant SRAM: Circuits and Statistical Design for Yield.* Springer Science+Buisness, 2013. 50

[28] Benton H. Calhoun Randy W. Mann. New category of ultra thin notchless 6t sram cell layout topologies for sub-22nm. *Quality Electronic Design (ISQED),2011 12th International Symposium*, March 2011. 86

Appendix A: Additional Results

A.1 Inverter

	-40^oC				25^oC				85^oC			
0.4V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D[%]	-9.86	8.47	25.76	6.7	-10.17	7.52	25.6	6.75	-10.97	6.6	24.32	5.75
τ_D[ns]	20.42	80.39	273.8	35.03	4.87	13.55	32.46	4.03	2.29	5.08	10	1.14
t_r[ns]	15.02	52.35	137.4	23.05	3.92	10.08	21.39	3.39	1.981	4.08	7.39	1.08
t_f[ns]	14.45	74.61	350.5	56.1	3.76	12.03	37.9	6.07	1.96	4.46	10.78	1.64
P_0[fW]	480	480.4	481.6	0.26	490.8	535.1	758.9	40.82	989.9	2.154	6875	944.2
P_1[fW]	480.1	481.4	485.8	1.11	528.6	712.4	1191	132.8	1942	5486	13200	2294
P_T[pW]	10.16	10.36	10.65	0.069	10.3	10.48	10.69	0.065	11.58	14.03	19.66	1.451
1.2V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D[%]	-9.76	-2.34	2.86	2.5	-9.99	-2.94	2.25	2.53	-10.1	-3.42	2.15	2.58
τ_D[ps]	77.2	88.85	99.82	3.715	94.88	108.9	122.1	4.47	110.6	126.3	141.2	5.057
t_r[ps]	94.23	110.9	126.6	6.25	120.9	142.1	161.4	7.85	146.2	171.2	194.4	9.34
t_f[ps]	79.16	88.42	97.38	3.59	105.6	117.8	129.9	4.71	132.7	147.7	163.5	5.77
P_0[pW]	4.397	4.397	4.487	0.027	4.41	4.624	5.42	0.16	6.17	10.26	26.67	3.32
P_1[pW]	4.32	4.32	4.34	0.038	4.48	5.09	6.67	0.44	9.04	20.57	45.65	7.457
P_T[nW]	0.428	1.02	1.72	0.223	0.942	1.74	2.91	0344	1.55	2.63	4.13	0.452

Table A.1: Inverter simulation results

	-40^oC					25^oC					85^oC				
0.4V	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]
τ_2	17.4	142	80.2	78.3	430	4.8	21.5	14.8	14.3	52.6	2.5	7.7	5.9	5.6	16
τ_4	23.1	212	109	90.9	601	6.4	31.4	19.9	17.1	71.9	3.4	11.2	8	7	21.7
τ_8	34.5	351	167	115	942	9.5	51.5	30.2	22.3	111	5	18.2	12.1	9.3	33.2
τ_{16}	57	632	282	158	1626	14.8	91	50.1	29.9	188	7.2	32.1	19	11.7	55.2
τ_{32}	92.6	1196	502	209	2990	21.3	173	81.2	37	332	9.4	57	27.8	13.6	90.5
τ_{64}	138	2338	826	278	5428	28.9	322	120	44.6	535	11.4	98	37.1	14.5	133
τ_{128}	204	4479	1288	380	9060	37.5	580	170	50.4	817	13.1	207	47.5	130	184

Table A.2: Inverter fan-out/delay simulation results

A.2 Gated Inverter

	-40^oC				25^oC				85^oC			
0.4V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D[%]	-10	7.86	21.98	6.67	-10.13	6.95	21.82	6.52	-10.44	6.05	21.15	6.54
τ_D[ns]	163	362.9	845.2	118.5	35.32	62.4	108.2	14.49	15.38	23.61	36.08	4.24
t_r[ns]	58.41	237.5	682.1	99.32	16.08	45.65	103.7	14.38	8.56	18.71	36.19	4.58
t_f[ns]	87.19	331.8	1751	220.1	21	52.73	182.4	23.37	10.12	19.66	51.77	6.32
P_{00}[fW]	372.9	374	375.7	0.638	397.3	421.1	472.1	13.88	751.5	1311	2689	307.5
P_{01}[fW]	372.6	373.4	375.4	0.560	406.8	458.2	559.2	28.34	1044	2065	4064	565.2
P_{10}[fW]	408.2	416.7	427.5	4.03	452.5	499.6	605.2	30.07	918.3	2152	4686	714.9
P_{11}[fW]	408.3	419.9	430.2	4.25	494	672.9	1170	115.5	2234	5528	13730	2034
P_T[pW]	11.83	12	12.14	0.048	11.98	12.22	12.52	0.091	13.1	14.1	15.65	0.479
1.2V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D [%]	-9.821	-2.41	1.98	2.5	-9.99	-3.07	1.89	2.45	-10.13	-3.67	1.81	2.55
τ_D[ps]	320.6	357.2	394.1	14.02	394.7	441.5	486.3	17.07	455.2	510.3	562.1	19.74
t_r[ps]	403.8	470.2	530.4	23.04	523.5	604.6	683.3	28.41	636	729.2	824.2	32.79
t_f[ps]	292.8	318.7	353	12.38	394.9	428.4	473.2	15.87	489.9	539.8	593.6	19.09
P_{00}[pW]	3.39	3.44	3.54	0.028	3.52	3.71	3.96	0.089	4.57	6.57	11	1.02
P_{01}[pW]	3.35	3.36	3.37	0.004	3.46	3.63	3.99	0.102	6.19	9.34	16.26	1.73
P_{10}[pW]	4.02	4.15	4.37	0.062	4.24	4.5	4.96	0.143	5.86	10.2	18.88	2.48
P_{11}[pW]	3.87	3.93	3.98	0.019	4.18	4.78	6.42	0.381	9.77	20.37	46.92	6.56
P_T[pW]	283.3	470.2	692.3	78.69	562.1	845.2	1185	114.9	880.8	1278	1715	159

Table A.3: Gated inverter simulation results

	-40^oC					25^oC					85^oC				
0.4V	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [μs]	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]
τ_2	73.1	657.3	350.9	325	1.98	19.9	98	64.3	58.3	235.6	10.3	34.9	25.4	22.6	71.2
τ_4	93.7	920.3	459.7	373.1	2.7	25.6	134.1	82.9	68.5	308	13.3	47.1	32.7	27.3	92
τ_8	134.9	1446	675.2	463.7	3.98	36.7	206	119.9	88.2	451.1	19.1	71.4	47.1	36.1	133.2
τ_{16}	216.2	2492	1103	637.3	6.6	57.8	349.7	192.7	122.9	735.8	29.1	120	74.6	49.7	214.5
τ_{32}	366.2	4518	1942	890.8	11.9	90	638.4	326.8	166	1290	41.9	214.8	117.8	63.1	363.6
τ_{64}	588.8	8776	3392	1228	21.9	132.5	1205	520.7	212.5	2232	55.5	377.6	170.6	74.3	577.7
τ_{128}	911.1	17081	5561	1747	38.7	186.6	2189	777.9	268.5	3549	70.9	62.9	234.3	82.4	842.9

Table A.4: Gated inverter fan-out/delay simulation results

A.3 NAND Gate

	-40^oC				25^oC				85^oC			
0.4V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D_{AB}[%]	5.36	15.65	29.03	4.33	2.52	14.87	25.73	4.32	2.14	14.57	25.26	4.3
D_{A0}[%]	-5.6	6.75	18.3	5.19	-6.88	3.84	14.64	5.15	-10.09	1.429	13.86	5.3
D_{0B}[%]	-5.85	5.78	18.42	4.9	-9.62	2.815	14.62	4.99	-10.37	0.1	13.84	4.94
τ_D[ns]	118.3	279.7	569.4	80.49	30.94	55.77	92.31	11.36	15.39	23.51	35.55	3.63
t_r[ns]	96.78	194.3	366.7	59.89	26.2	44.92	74.17	10.69	13.71	20.82	30.49	3.79
t_f[ns]	67.46	268.2	810.8	124.4	19.01	49.21	107.4	15.72	10.05	20.08	36.13	4.68
P_{00}[fW]	403.2	404.5	406.8	0.667	411.1	435.7	503.2	16.84	894.6	1428	2831	325
P_{01}[fW]	483	484.1	485.6	0.541	493.5	546.5	661.6	32.48	1020	2356	4837	742
P_{10}[fW]	410	421	429.8	3.758	463	515.2	666.6	37.12	1136	2395	5682	822.5
P_{11}[fW]	643	644.8	648.6	1.002	736.2	940.2	1275	94.54	2998	6495	11570	1496
P_T[pW]	12.49	12.68	12.87	0.068	12.66	12.85	13.07	0.081	14.66	16.38	19.03	0.826
1.2V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D_{AB}[%]	-2	2.24	5.98	1.56	-1.93	3.25	6.07	1.71	-1.77	4.28	6.35	1.85
D_{A0}[%]	-10.06	-5.55	-1.96	2.06	-10.78	-7.18	-2.13	2.04	-13.96	-8.23	-2.44	2.2
D_{0B}[%]	-10.12	-6.53	-2.01	1.98	-14.03	-8.95	-5.16	2.1	-15.24	-10.94	-6.08	1.99
τ_D[ps]	447.6	491.6	538.1	16.32	561.6	615.7	672	20.04	655	720.4	783.9	23.1
t_r[ps]	631.6	693.2	763.7	27.59	841.4	917.9	1011	34.81	1039	1129	1241	41.05
t_f[ps]	392.3	424.5	463.7	13.65	531.1	576.4	627.7	18.36	627.7	729	790.1	21.67
P_{00}[pW]	3.79	4	4.45	0.122	3.89	4.24	4.99	0.192	5.58	7.47	12.12	1.14
P_{01}[pW]	4.34	4.53	4.76	0.061	4.54	4.82	5.35	0.152	6.32	10.95	19.3	2.53
P_{10}[pW]	4.01	4.2	4.44	0.067	4.19	4.6	5.24	0.176	6.7	10.9	22.06	2.79
P_{11}[pW]	5.78	5.79	5.8	0.004	6.07	6.7	7.72	0.289	12.97	23.65	39.17	4.57
P_T[nW]	0.457	0.693	0.956	0.093	0.829	1.18	1.58	0.136	1.249	1.719	2.205	0.179

Table A.5: NAND gate simulation results

	-40^oC					25^oC					85^oC				
0.4V	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [μs]	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]
τ_2	78.3	447.5	294.9	320.1	1.35	23.5	78.6	62.5	68	194.8	12.9	31.2	27.3	29.1	66.3
τ_4	105.5	602.2	393.6	427.7	1.79	31.8	105.5	83.6	91	258.2	17.5	42.1	36.6	39.1	88
τ_8	159.9	908.6	592.5	641.7	2.67	48.1	158.7	125.6	136.4	385.1	26.4	63.5	55	58.4	131.4
τ_{16}	267.3	1518	987.4	1061	4.44	78.8	265	220.9	222.2	638.4	42.3	105.7	89.8	93.5	216
τ_{32}	463.3	2741	1760	1834	7.97	131.7	477	359.4	377	1131	68.6	184.2	148.7	156.1	371.2
τ_{64}	805.2	5188	3107	3314	14.7	224.4	866.2	609.7	674.1	1969	115.8	311.3	246.6	276	618.5
τ_{128}	1442	9735	5454	6213	26.1	397.7	1490	1052	1259	3345	204.2	515.4	424	511	1035

Table A.6: NAND gate fan-out/delay simulation results

A.4 NOR Gate

	-40^oC				25^oC				85^oC			
0.4V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D_{AB}[%]	-7.25	8.91	27.92	6.202	-6.7	8.33	26.33	6.25	-9.32	7.28	25.99	6.29
D_{A0}[%]	-9.12	8.72	26.04	6.31	-9.94	8.2	25.91	6.39	-10.27	7.06	25.63	6.42
D_{0B}[%]	-9.7	8.7	29.02	5.91	-9.95	8.16	26.49	5.81	-10.14	7.36	26.06	5.89
τ_D[ns]	119.1	242.7	566.7	76.03	25.09	42.25	75.62	9.7	10.32	15.8	24.52	2.85
t_r[ps]	73.5	222.1	538	83.17	18.85	42.89	86.26	12.28	9.36	17.42	30.74	3.94
t_f[ps]	47.08	181.9	493.8	91.83	12.48	30.99	63.7	10.69	6.46	11.85	20.25	2.93
P_{00}[fW]	541.2	624	643.7	0.409	669.3	748.8	1012	53.37	1831	3906	9737	1269
P_{01}[fW]	410.7	420.8	433.4	4.426	487.5	681.1	1381	128.5	1917	5661	16750	2245
P_{10}[fW]	480.6	482.1	487.4	1.241	533.3	725.9	1296	141.6	2116	5770	14840	2429
P_{11}[fW]	400.3	401	404.5	0.545	421.8	508.7	694.8	50.14	1409	2987	6404	891
P_T[pW]	12.3	12.54	12.75	0.075	12.42	12.66	13.02	0.101	13.96	15.75	19.59	0.902
1.2V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D_{AB}[%]	-7.57	-3.07	2.1	2.24	-10	-4.74	1.87	2.3	-10.33	-6.15	0.78	2.24
D_{A0}[%]	-10.01	-3.3	2.04	2.3	-10.57	-4.7	1.82	2.35	-13.94	-6.2	0.37	2.42
D_{0B}[%]	-7.7	-2.2	4.69	2.18	-9.83	-2.91	2.19	2.2	-10.01	-3.46	1.97	2.19
τ_D[ps]	214.8	237.3	257.9	8.9	259.9	286.8	311.3	10.44	297	327.7	355.2	11.59
t_r[ps]	368.5	431.2	494.3	24.98	477.7	53.2	623.9	29.62	577.5	663	741.7	32.93
t_f[ps]	181.9	198.3	216.7	6.68	236.2	258.9	281.3	8.65	290.3	317.6	346.2	10.27
P_{00}[pW]	5.84	5.93	6.13	0.054	5.98	6.38	7.53	0.224	10	17.39	38.07	4.46
P_{01}[pW]	3.9	3.94	3.99	0.018	4.17	4.81	7.11	0.424	8.73	20.8	56.49	7.238
P_{10}[pW]	4.37	4.41	4.52	0.029	4.59	5.25	7.12	0.482	9.73	21.63	51.11	7.922
P_{11}[pW]	3.605	3.605	3.616	0.002	3.67	3.95	4.58	0.170	6.51	11.63	22.1	2.782
P_T[nW]	0.475	0.809	1.32	0.146	0.949	1.49	2.14	0.207	1.44	2.147	3.07	0.265

Table A.7: NOR gate simulation results

	-40^oC					25^oC					85^oC				
0.4V	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [μs]	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]	**FF** [ns]	**FS** [ns]	**TT** [ns]	**SF** [ns]	**SS** [ns]
τ_2	56.34	290.2	241.7	353.2	1.18	15.3	47.1	45.4	59.8	151.1	7.5	17.3	17.7	21.8	46.4
τ_4	73.4	377.4	312.1	459.7	1.51	19.7	61.3	58.7	78.3	193.2	9.9	22.7	23.1	28.8	59.7
τ_8	107.3	551.1	453.4	669	2.18	28.8	89.5	85.5	114.2	279.7	14.4	33.4	33.5	41.9	86.5
τ_{16}	174.2	898.4	732.2	1075	3.52	45.9	146.7	137.4	181.9	451.7	22.4	54.2	53	66.1	138.2
τ_{32}	296	1595	1262	1852	6.14	76.2	257.8	230.1	310.4	767.4	36.7	91.3	86.7	112.5	228.6
τ_{64}	519.5	2956	2204	3394	10.79	132.2	449.8	394.9	566.3	1309	63.6	154.3	148.2	205.2	384.1
τ_{128}	947.4	5294	3962	6469	19.12	240.2	782.9	705.2	1082	2293	116.4	283.7	264.8	396.4	673.5

Table A.8: NOR gate fan-out/delay simulation results

A.5 XNOR Gate

	-40^oC				25^oC				85^oC			
0.4V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D_{AB}[%]	-6.76	8.18	25.96	6.13	-9.56	7.34	25.77	6.11	-9.89	6.44	25.02	6.21
D_{A0}[%]	-9.64	8.43	22.25	6.19	-9.98	7.62	21.94	6.2	-10.18	6.73	21.72	6.12
D_{0B}[%]	-9.92	8.8	25.82	6.23	-10.16	7.93	25.64	6.18	-10.71	6.93	25.25	6.28
τ_D[ns]	317.3	625.8	1313	197.6	67.68	111	192.6	24.97	28.91	42.24	64.97	7.37
t_r[ns]	134.9	449	1247	188	34.93	86.76	187	26.96	17.96	36.38	68.23	8.79
t_f[ns]	155.8	588.6	2460	362.1	38.73	95.41	263.5	39.09	19.04	36.67	80.23	11.19
P_{00}[fW]	835.6	878.2	909.8	24.15	962.2	1123	1776	99.09	3366	7514	21400	2300
P_{01}[fW]	907.5	945	973	23.45	1184	1722	2983	306.7	8460	18160	39810	5273
P_{10}[fW]	773.3	822.3	859.2	25.17	1137	1651	2919	305.7	9156	18250	38740	5259
P_{11}[fW]	836.2	878.3	909.5	23.65	1094	1511	2381	267.3	6523	14810	28440	4604
P_T[pW]	32.17	33.38	34.47	0.452	36.51	37.97	39.82	0.596	48.91	55.42	67.35	3.41
1.2V	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ	**Min**	μ	**Max**	σ
D_{AB}[%]	-6.43	-2.51	2.2	2.2	-9.56	-3.14	2.06	2.37	-9.87	-3.67	1.99	2.43
D_{A0}[%]	-9.84	-2.37	2.01	2.38	-10	-3.03	1.93	2.32	-10.13	-3.57	1.83	2.35
D_{0B}[%]	-9.78	-2.56	2.2	2.24	-9-98	-3.36	2.06	2.32	-10.07	-3.9	1.99	2.34
τ_D[ps]	560.6	606	579.4	22.8	690	746.8	835.6	28.37	790.0	859.9	965.4	32.94
t_r[ps]	811.9	926.8	1046	43.09	1049	1187	1340	56.63	1270	1427	1607	62.44
t_f[ps]	549.2	616.3	685.6	24.66	740.7	831.2	917.9	31.94	935.8	1046	1148	38.38
P_{00}[pW]	8.59	9.71	12.07	0.578	9,7	11.41	15.84	0.982	18.62	34.25	83.45	8.36
P_{01}[pW]	8.91	9.96	12.16	0.547	10.25	13.35	20.52	1.43	34.66	67.78	142.1	17.49
P_{10}[pW]	8.01	9.04	11.05	0.507	10.05	12.55	18.17	1.42	37.35	67.1	134.6	17.42
P_{11}[pW]	7.89	8.28	8.73	0.224	8.78	10.47	13.47	0.934	27	54.28	98.65	15.07
P_T[nW]	1.13	1.84	2.78	0.321	2.25	3.43	4.85	0.496	3.57	5.15	6.97	0.646

Table A.9: XNOR gate simulation results

	-40^oC					25^oC					85^oC				
0.4V	**FF**	**FS**	**TT**	**SF**	**SS**	**FF**	**FS**	**TT**	**SF**	**SS**	**FF**	**FS**	**TT**	**SF**	**SS**
	[μs]	[μs]	[μs]	[μs]	[μs]	[ns]	[ns]	[ns]	[ns]	[ns]	[ns]	[ns]	[ns]	[ns]	[ns]
τ_2	0.146	1	0.660	0.794	3.5	39.4	152.2	123.5	143.1	431.8	19.7	54.5	48.1	54.1	130.8
τ_4	0.190	1.33	0.857	1.04	4.51	51.2	201.2	158.8	185.6	550.8	25.7	71.8	62.1	69.9	167.4
τ_8	0.281	1.98	1.26	1.52	6.62	75.7	298	232.9	268.8	804.2	38	106.2	91	101	243.8
τ_{16}	0.465	3.26	2.08	2.48	10.38	124.4	490.9	381.3	434.7	1309	62	175.9	148.5	162.1	3.96
τ_{32}	0.829	5.84	3.7	4.34	19.25	217.5	886.6	675.9	745.5	2321	105	321	258.2	271.2	697.8
τ_{64}	1.509	11.08	6.88	7.86	36.08	313.7	1720	1212	1322	4282	177.8	623.4	444.7	473.3	1244
τ_{128}	2.713	21.93	12.61	14.47	68.41	665.8	2515	2126	2441	7659	312.6	1135	769.6	868.5	2155

Table A.10: XNOR gate fan-out/delay simulation results

Appendix B: Additional Layouts

B.1 Driving Inverters

Fig. B.1 shows the layout of the inverter with m-factor 2.

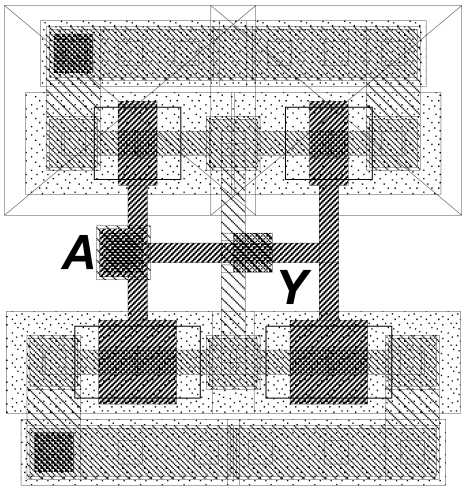


Figure B.1: Layout of 2x2 inverter

Fig. B.2 shows the layout of the inverter with m-factor 4.

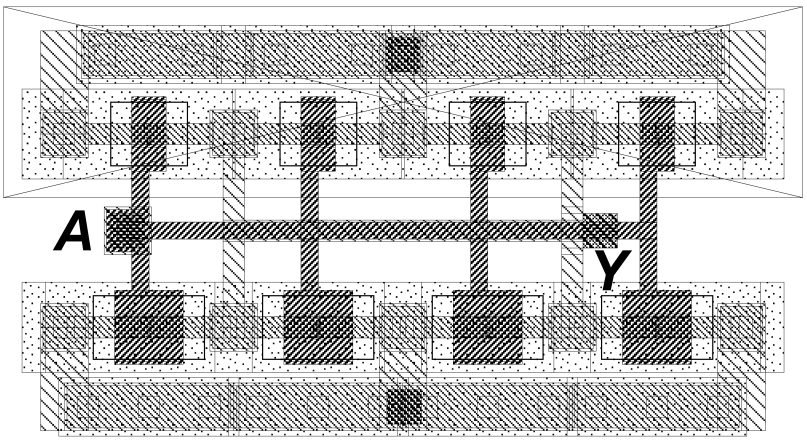


Figure B.2: Layout of 4x4 inverter

B.2 Gated Inverter

Fig. B.3 shows the layout of the gated inverter.

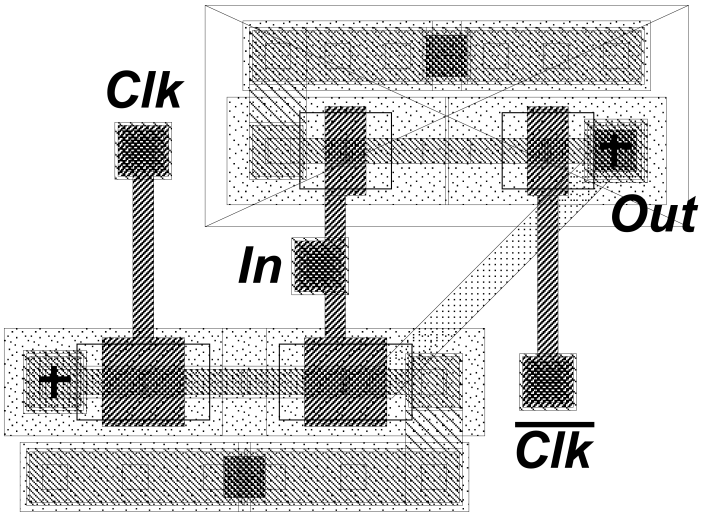


Figure B.3: Layout of gated inverter

B.3 NOR Gate

Fig. B.4 shows the layout of the NOR gate.

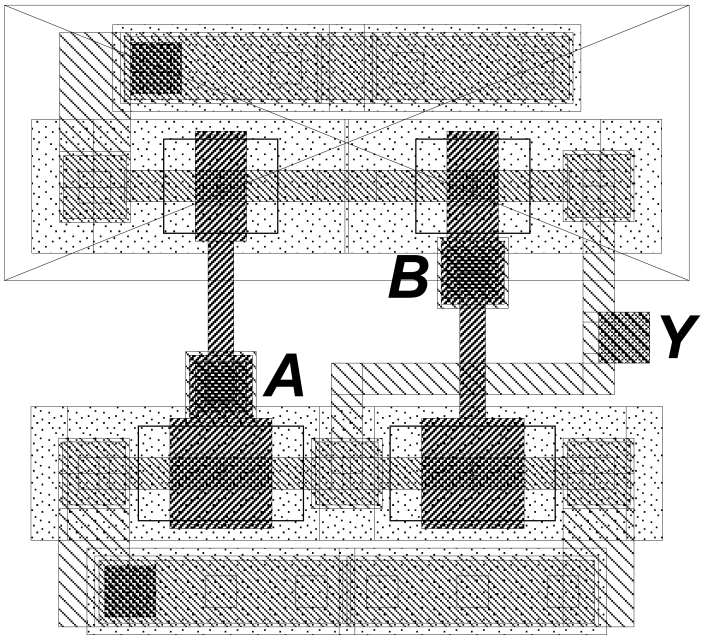


Figure B.4: Layout of NOR gate

B.4 XNOR Gate

Fig. B.5 shows the layout of the XNOR gate.

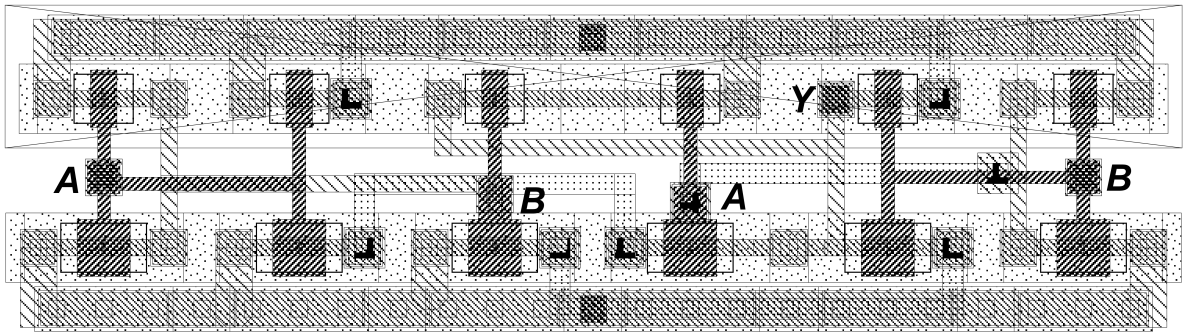


Figure B.5: Layout of XNOR gate

Fig. B.6 shows the layout of the transmission gate.

B.5 Transmission Gate

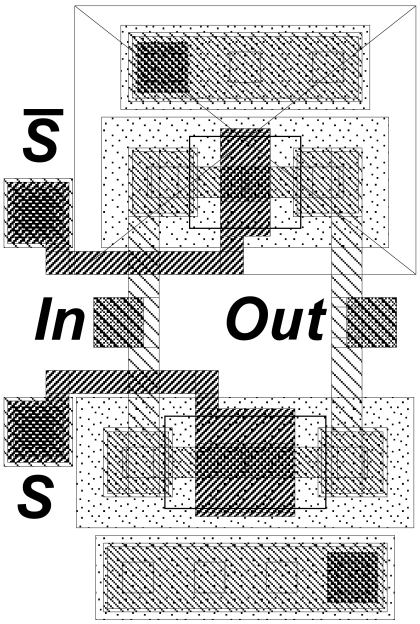


Figure B.6: Layout of transmission gate

B.6 2-to-1 Multiplexer

Fig. B.7 shows the layout of the 2-to-1 multiplexer.

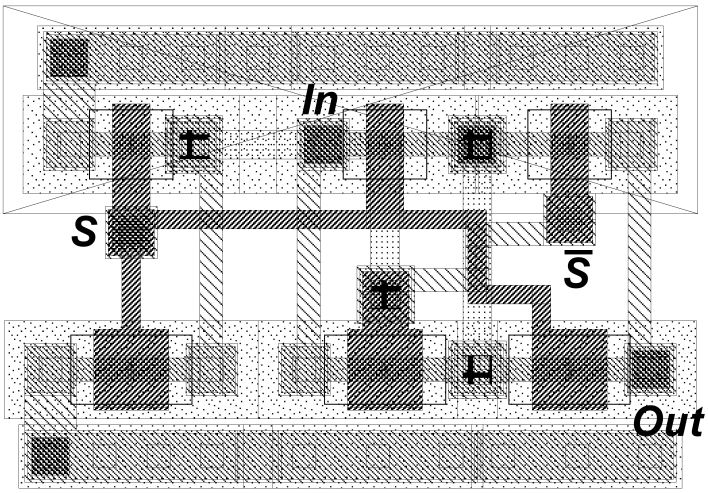


Figure B.7: Layout of 2-to-1 multiplexer

Appendix C: Source Code

C.1 SNM Extraction Module

```
`include "constants.vams"
`include "disciplines.vams"

module read_snm_gen(vu, vq, vqg, vqb, vqbg, v1, v2, vsnm1, vsnm2);
  // Connection for sweeping vatiable "u"
  input vu;
  // Connection for SRAM node "Q" and "Qb"
  input vq, vqb;
  // Conection for cross-coupled inverters
  output vqg, vqbg;
  // Connection for "V1" and "V2"
  output v1, v2;
  // Tilted butterfly plot output
  output vsnm1, vsnm2;

  electrical vu, vq,vqg, vqb, vqbg, v1, v2, vsnm1, vsnm2;

  analog begin
    V(vqg)  <+ ((1/sqrt(2))*V(vu))+((1/sqrt(2))*V(v1));
    V(vqbg) <+ ((-1/sqrt(2))*V(vu))+((1/sqrt(2))*V(v2));
    V(v1)   <+  V(vu)+(sqrt(2)*V(vq));
    V(v2)   <+ -V(vu)+(sqrt(2)*V(vqb));

    V(vsnm1) <+ abs(V(v1)-V(v2));
    V(vsnm2) <+ abs(V(v2)-V(v1));
    end
endmodule
```