KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

Active Supply Control in Static Random Access Memories

Promotor:
Prof. dr. ir. W. Dehaene

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Peter Geens

7th April 2009

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

Active Supply Control in Static Random Access Memories

Jury:
Prof. em. dr. ir. Y. Willems, voorzitter
Prof. dr. ir. W. Dehaene, promotor
Prof. dr. ir. M. Steyaert
Prof. dr. ir. F. Catthoor
Prof. dr. ir. P. Reynaert
Prof. dr. ir. E. Van Tuijl

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Peter Geens

UDC: 621.3.049.77

7th April 2009

D/2009/7515/29

978-94-6018-046-0

to my parents for their unconditional and unwavering support

Voorwoord

Bij dezen zal ik een poging wagen om alle mensen die al dan niet rechtstreeks een bijdrage geleverd hebben aan dit werk te bedanken. Indien ik iemand vergeten mocht zijn, onthoud dan dat een van de sleutels tot geluk een slecht geheugen is, om Rita Mae Brown te parafraseren. Ik mag dan wel gelukkig zijn, een slecht geheugen maakt geen onderscheid tussen goed en slechte dingen.

- Vooreerst mijn promotor prof. Wim Dehaene. Op een zonnige zomerdag in 2002 heeft hij mij de kans gegeven om aan het onderzoek te beginnen dat uiteindelijk tot dit werk geleid heeft. Op onbekend terrein brachten onze open discussies, zijn gedrevenheid en ervaring mij vaak weer op een mogelijk spoor. Dit avontuur met ups en downs zou nooit mogelijk geweest zijn zonder zijn onvoorwaardelijke steun.
- Prof. Francky Catthoor voor de constructieve feedback en de discussies die geheugens altijd in een grotere context plaatsten.
- de leden van mijn jury en leescomité prof. Michiel Steyaert, prof. Ed Van Tuijl, prof Patrick Reynaert en prof. em. Yves Willems als voorzitter; voor hun tijd en waardevolle opmerkingen op deze tekst en werk.
- de vele MICAS collega's voor de soms productieve en soms surrealistische discussies tijdens de koffiepauzes; Mijn bureaugenoten over de jaren in het bijzonder: Stefan, Jorg, Anselme, Junfeng, Vibhu, Kristof, Tom, Erwin, Ewout, David, Bart,…;
- Mijn collega PhD studenten binnen het IMEC TAD programma: Hua, Mandeep, Evelyn die door onze vergaderingen andere aspecten van geheugens belichtten. Ook gaat mijn dank uit naar Stefaan, Josine en Bertrand die mij veel technologische aspecten bijgebracht hebben.
- De technische staf zonder wie het onmogelijk geweest zou zijn om mijn chips op te meten, in willekeurige volgorde: Albert, Noëlla, Tony, Frederik, Jan, Ludo, Luc, Jan;
- de systeemgroep die er voor zorgde dat de servers dag en nacht bleven draaien: Frank, Stef, Rik, Marc, Piet;
- Ben, die ervoor zorgde dat we de beschikking hadden over de laatste nieuwe software updates en samen met Frederik voor de soms Bourgondische middagen;
- De secretariaten van MICAS en ESAT om al het papierwerk rond te krijgen: Chris, Daniëlle, Lut, Evelyne, Elliane;
- de laatstejaarsstudenten burgerlijk ingenieur micro-elektronica van de promotiejaren 2003-2008, voor het al dan niet vrijwillig debuggen van mijn

software tools; er is niets zo efficiënt als een “bende” ingenieursstudenten om software dingen te laten doen waar het niet voor ontworpen was.

- De middagse kaartersclub, *you know who you are*, voor de ontspanning over de middag, hoewel er soms met meer dan één “schuppe-zot” gespeeld werd;
- mijn neef Erik, voor de vele verre reizen, discussies over de elektronica industrie en troubleshooter als ik weer eens vast zat met een software probleem;
- en als laatste maar zeker niet de minste mijn ouders, die het wel en wee van dichtbij meegemaakt hebben. Hun oneindig geduld kan ik alleen maar bewonderen.

Peter Geens

Leuven, 7th April 2009

Abstract

The evolution of the cell phone from a "simple" wireless phone to a portable multimedia station is a prime example of the paradigm shift in modern day electronics. Applications evolve towards more mobility and more multimedia. This requires an increase in power efficiency of the electronic systems as battery power is limited. Battery life-time also has become one of the key sales arguments in the mobile market. As such the number of operations per watt must increase.

Modern day electronic systems are unthinkable without memory circuits. In high end single die processor systems already more than 50% of the die area is used by memory cache. This value is expected to still increase as market driving applications, such as games and multimedia, are still growing in market penetration.

As such SRAM has been chosen as the prime subject for this thesis. This introduction will start by illustrating the significance of SRAM for current technology development. This will be followed by a short section on the basic working principles of SRAMs. The key to reducing the leakage in SRAMs will be introduced in section 1.3. Section 1.4 will then briefly illustrate the consequences of a reduced supply voltage on the data retention capability of the cells. A chapter-by-chapter overview will be provided in the following section and lastly the contributions of this work will be listed.

This thesis contributes in three main domains to achieve the desired results presented in the previous sections.

To minimise the leakage currents in the matrix of an SRAM, a second, lower, supply is introduced for the non-accessed parts of the matrix. This is implemented with the finest granularity feasible: a single word. This granularity has the benefit of minimising wake-up delays while keeping the maximum possible amount of cells in a drowsy state. To achieve this the last stage of the decoder was distributed into the matrix to control the supply switches and wordline activation. This combines simplicity of control, limited area overhead and maximal reductions. The research was published in [Gee05].

A further contribution in this thesis is the development and design of a monitor and regulation circuit to guarantee the data retention. Lowering the supply voltage on the drowsy cells not only reduces the leakage currents, it also reduces the retention capability of the cells. Traditionally, the increase of inter and intra die variations in recent technologies would require margins on the drowsy supply voltage to be taken to compensate for the reduction in retention capability. These margins in turn would deteriorate the possible leakage current savings. As such it is important to be able to guarantee the retention of the data without sacrificing the benefit of leakage reduction. By measuring the actual retention capability of the cells in the live and operational environment, a regulation circuit can be designed to maximise leakage reduction on a die-to-die basis while guaranteeing the data retention. The solution published in [Gee07] fulfils this requirement.

The final contribution of this thesis is in the domain of multi-port memories. The prototype, which is the focus of the research in [Gee08], combines the leakage reduction techniques with an asymmetric width dual port SRAM. The asymmetry in the wordlength enables the reduction of active energy consumption per bit by allowing a wide, 256 bit, and narrow, 32 bit, access. It is proven that the single wide access consumes less energy than several consecutive narrow accesses. This prototype also includes the means to monitor and regulate the secondary drowsy supply.

The access time of 2ns is attained in its nominal settings a measured in section 4.3.3. In low matrix leakage settings the access time is increased to $2,5$ns with an extra leakage reduction of 33% compared to the nominal settings. The system can be put in a pure retention state by turning of periphery circuits, this reduced the leakage power consumption with 65% compared to the nominal case. The active power consumption results can also be found in table 4.5. For the narrow 32bit port the average active energy per access is 16pJ or 0.5pJ/bit. For the wide 256bit port the average active energy per access is 24pJ or 0.09pJ/bit. The difference can be mainly found in the precharge energy. These numbers confirm the possibility to save energy on the system level by accessing the memory on the wide port when more than one 32bit word of data is needed.

This thesis presents a system and the necessary background to create an SRAM where leakage currents can be minimised while guaranteeing data retention. The presented DPDSSRAM is also the first published [Gee08] dual supply SRAM that incorporates the measurement of the data retention parameter SNMh and the generation of the secondary sleep voltage on chip.

Contents

List of Figures

Symbols and Abbreviations

Symbols

C	Symbol used to describe a capacitor
V_T	transistor threshold voltage
γ	body effect parameter
η	DIBL effect parameter
μ	mobility parameter
κ	relative dielectric constant
V_{gs}	gate-source voltage
V_{ds}	drain-source voltage
V_{gb}	gate-bulk voltage

Abbrevations

DIBL	Drain Induced Barrier Lowering
GIDL	Gate induced barrier lowering
SRAM	Static Random Access Memory
ITRS	International Technology Roadmap for Semiconductors
BIST	Build-in Self Test
SNM	Static Noise Margin
SNMh	Static Noise Margin under hold
SVNM	Static Voltage Noise Margin
SINM	Static Current Noise Margin
SPNM	Static Power Noise Margin
DRV	Data Retention Voltage
DDA	Differential Difference Amplifier
SRAM	Static Random Access Memory
DSSRAM	Dual Supply SRAM

Chapter 1

Introduction

The evolution of the cell phone from a “simple” wireless phone to a portable multimedia station is a prime example of the paradigm shift in modern day electronics. Applications evolve towards more mobility and more multimedia. This requires an increase in power efficiency of the electronic systems as battery power is limited. Battery life-time also has become one of the key sales arguments in the mobile market. As such the number of operations per watt must increase.

The key technology driver for this evolution has always been scaling, as it allowed higher functionality and density at a lower production cost. Moore’s law [Moo69] modelled this evolution as a doubling of processor performance roughly every two years. With the advent of deep deep submicron technologies, new challenges have appeared for the system designers. The scaling of the feature length also increased the importance of the leakage currents on the power budget. Reducing this leakage currents has become an important facet of system-on-chip design, especially for mobile applications. Smaller feature lengths also mean fluctuations in processing have a larger impact on the transistor performance. As such variability has started to play a more important role in the design of digital circuits and systems.

Modern day electronic systems are unthinkable without memory circuits. In high end single die processor systems already more than 50% of the die area is used by memory cache. This value is expected to still increase as market driving applications, such as games and multimedia, are still growing in market penetration.

A generic memory organisation is shown in figure 1. The central data path interacts with the level 1 data and instruction caches and the loopbuffer. It has to be noted that these functions can be shared by the same physical memory block. These caches are typical small memories with a storage of less than 256 kbit and a single cycle access time. The level 2 and higher level caches only come into play should the lower level chaches not contain the necessary data for an operation in the data-path. These memories are typically larger than the lower level caches and also slower. The highest level in the memory hierarchy are the storage memories.

The slow but non-volatile memories such as flash serve as data or program storage, while the fast, but volatile memories such as DRAM or SRAM serve as work space for the processors. While the density of DRAM is higher than of SRAM, SRAM has traditionally been the first line of memory in system. This, in part, thanks to a feasible delay equal to the datapath and low power consumption compared to other memory

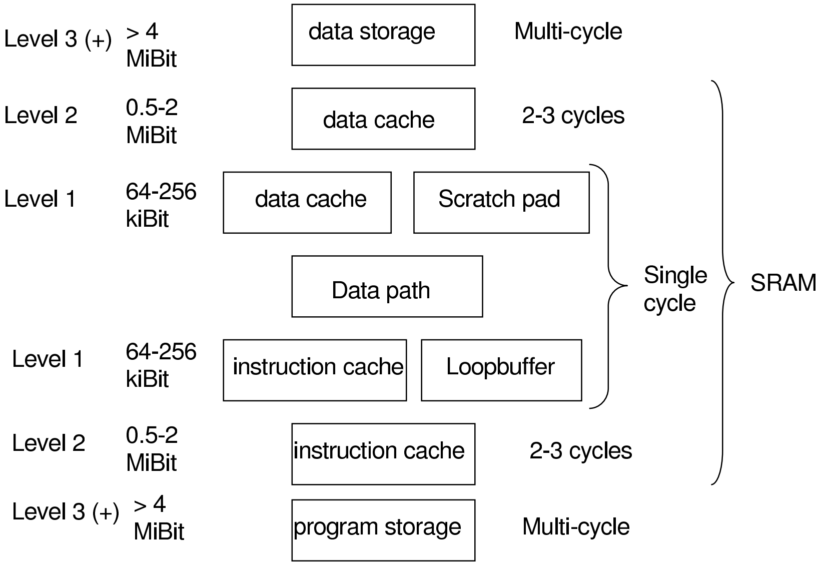


Figure 1.1: Generic memory organisation

architectures for the required sizes. This thesis will focus on these level 1 SRAM memories.

One of the first blocks to suffer from the ill effects of the technology scaling is the SRAM. The high transistor count with a low activation factor causes leakage current to become dominant in the total power consumption of the block. As the area is also kept as small as possible the variability effects also manifest here first as device lengths and width are kept near the minimal sizes. To reduce the leakage consumption of the SRAM, the circuit designer can play on only a limited set of parameters. The main design parameters that can be used are: the supply voltage, the choice of threshold voltage and the sizing of the transistors.

As such SRAM has been chosen as the prime subject for this thesis. This introduction will start by illustrating the significance of SRAM for current technology development. This will be followed by a short section on the basic working principles of SRAMs. The key to reducing the leakage in SRAMs will be introduced in section 1.3. Section 1.4 will then briefly illustrate the consequences of a reduced supply voltage on the data retention capability of the cells. A chapter-by-chapter overview will be provided in the following section and lastly the contributions of this work will be listed.

1.1 Embedded SRAM as benchmark

The main technology driver for embedded systems is profit, preferably by having more functionality for a lower production cost. One of the main contributors to the production cost is the chip area used by the system. Embedded SRAM already accounts for more than 50% of the chip area. According to the International Technology Roadmap for Semiconductors (*ITRS*) roadmap [ITR] the amount of embedded SRAM (*eSRAM*) in systems will continue to increase. To save area this means eSRAM cells are kept as small as possible by using minimum feature sizes.

As SRAM is embedded with the remainder of the system, its production must be compatible with standard CMOS processing. Combined with the need for the smallest area feasible, SRAM has become the showcase for digital system oriented technology development. eSRAM systems however work within the premises of the analog domain. The communication on the long bitlines consists out of small swing signals. This results in SRAM circuits to be among the first to suffer from the deep-deep submicron effects, such as high subthreshold and gate leakage. In combination with the high transistor count and low activity of the transistors in the eSRAM matrix, this leads to leakage power being a significant portion of the total power consumption.

In the quest for lower power circuits the previous discussion indicates that not only the active part of the power consumption has to be taken into account but also the leakage or static component. This thesis will focus on the reduction of the leakage currents in the embedded SRAM.

1.2 Basic SRAM concepts

1.2.1 SRAM system overview

The functional description of an SRAM could be written as : a circuit that can store and return data in a place identified by an address code, that can be randomly accessed. To achieve this functional description an SRAM consists of three main parts. Firstly, the cell matrix that stores the actual data in words of a predefined length. The cells store the data based on a positive feedback loop of two inverters. This loop provides the static retention of the data without the need to refresh that data. These cells will be accessed through the bitlines to the sense-amplifiers. The sense-amplifiers form the second part. They are used to translate the low swing signals generated on the bitlines by the cells to full level digital signals. The third main part consists of the decoders to decode the address into a physical position in the cell matrix. Of course additional peripheral circuits are needed, such as the timing and control section, the precharge and write system.

The timing of the activation of all subcomponents in an SRAM is critical. It plays a large part in the functionality, delays and power-consumption of the SRAM. A wordline that is turned on too soon will cause a wrong address to be accessed. If it is turned on with an excessive margin, it will be detrimental to the delay of the whole SRAM. The timing of the sense-amplifier activation, in particular, is crucial. An activation that comes to soon, can cause the sense-amplifiers to generate a false output. The cells need a sufficient amount of time to develop a voltage difference on the bitlines high enough to compensate for the mismatch in the input stage of the sense-amplifiers. An activation that comes too late, will cost in power consumption and delay. The difference on the bitlines will at that point be larger than needed to generate a correct

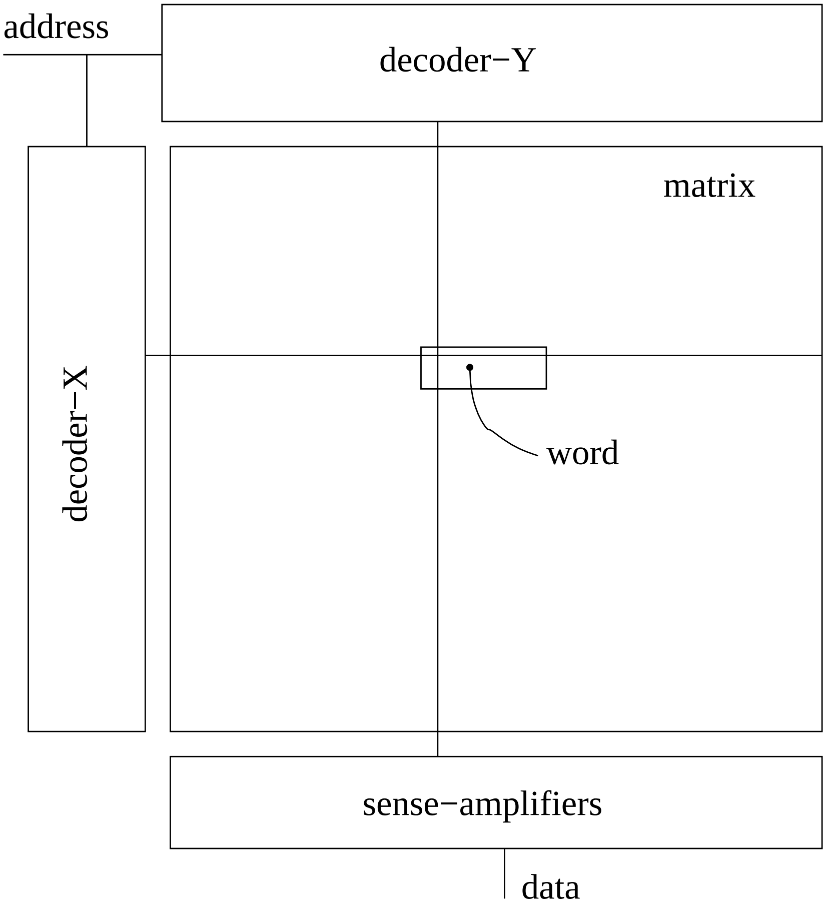


Figure 1.2: high level overview of the parts making up an SRAM

readout and require extra charge to be dumped into the bitline to be recharged to the correct precharge voltage.

1.2.2 The traditional 6T-cell

The core matrix cell traditionally consists of the six transistor circuit also shown in figure 1.3. The back to back connected inverters formed by the transistors $M1_{a,b}$ and $M2_{a,b}$ form the core storage loop. The passtransistors, also referred to as access transistors, $M3_{a,b}$ connect the inverter to the bitlines. An access operation consists of enabling the passtransistors by bringing the wordline (WL) to a high level. For a read operation the cell will now discharge one side of the bitline pair to create the needed differential voltage. For a write operation the bitlines will be brought to the correct levels before enabling the passtransistors, for the traditional cell that implies one of the bitlines is full discharged while the complementary bitline is kept or left high.

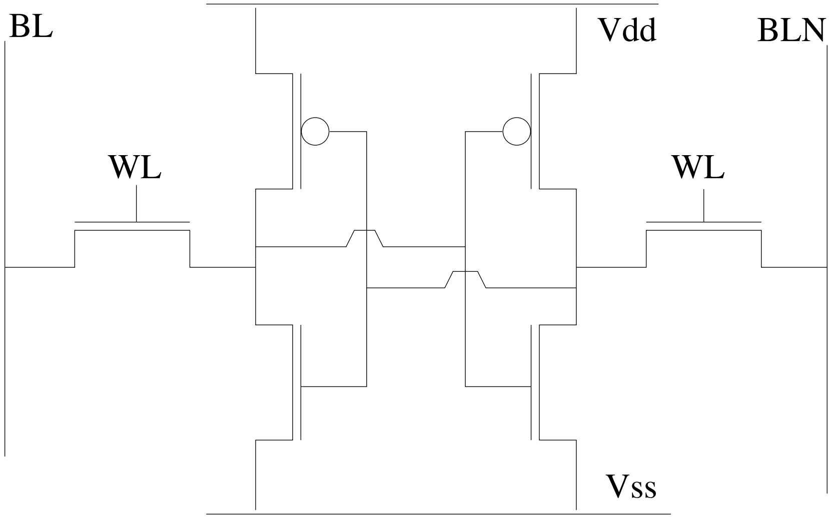


Figure 1.3: Schematic of the traditional 6T SRAM cell

As the cell is connected to the bitlines for the read and write operation through the same transistors, the risk of provoking a destructive read exists. In this scenario the inner nodes are disturbed beyond the write threshold. To avoid an unwanted write during a read operation the pull-down transistor in the cell must be strong enough to keep the low node of the cell well below the write trip point, and preferably even below the V_T of the pull-down NMOS, with the bitline at the precharge voltage. The voltage generated on the internal low node is determined by the current ratio of the pull-down transistor versus the passtransistor. The W/L ratio of the pull-down transistor to the passtransistor is referred to as the cell ratio (*CR*) and can be written as equation 1.1. [Rab03]

$$CR = \frac{r_{\text{pull-down}}}{r_{\text{pass}}} = \frac{(W/L)_{\text{pull-down}}}{(W/L)_{\text{pass}}} \tag{1.1}$$

To be able to write the cell, the passtransistors must be able to pull down the high internal node under the cell write trip point. The success and speed of execution of this operation are determined by the current ratio of the passtransistor versus the PMOS pull-up transistor. The ratio of the pull-up transistor to the passtransistor is referred to as the pull-up ratio (*PR*) and can be written as equation 1.2. [Rab03]

$$PR = \frac{r_{\text{pull-up}}}{r_{\text{pass}}} = \frac{(W/L)_{\text{pull-up}}}{(W/L)_{\text{pass}}} \tag{1.2}$$

For this thesis the traditional 6T-cell will serve as basis for the derivation of the methodologies, more specifically the four transistor that constitute the cross-couple inverters for the data storage. The formulas derived in this thesis will be specific for this situation, but could be derived for any similar architecture using the methodologies presented in this thesis.

1.3 Leakage current reduction

As power consumption has become a key specification in the mobile application market, not only active power has become the focus of research. With scaling, the leakage currents gained in importance to the level where the ITRS [ITR] predicts it to be the main contributor to the total power consumption of a system. This evolution is even more noticeable in SRAMs. The large fraction of transistors not active in the matrix tilts the power balance further towards leakage currents, consisting of subthreshold and gate leakage.

As will be shown in chapter 2 the supply voltage plays a crucial role in both the gate and subthreshold leakage. In combination with the retention capabilities of the SRAM cell even at a lower supply voltage, this creates a way of freedom to reduce the leakage currents. By lowering the supply on non-accessed cells, while ensuring the data remains intact, the leakage components can be reduced significantly. By keeping the accessed parts of the memory system on the nominal supply during an operation cycle, the delay impact can be mitigated. This technique has lead to the development of dual supply SRAMs and will be further discussed in section 2.4.

1.4 Guaranteeing data retention

While lowering the supply voltage on the SRAM cell during the non-accessed time, lowers the leakage currents, it also lowers the retention capability of the cells. To guarantee the data integrity sufficient precautions need to be taken. With the increase in variability in the state-of-the-art technologies, this is not trivial.

Traditionally worst case design methodologies would be used to ensure the data retention under worst case conditions. For the majority of dies however this would create an unnecessary overhead in power as the voltage margins taken would be higher than necessary to be able to deal with not just the intra-die variations, but also environmental changes and process corners.

The first step in reducing this power overhead due to margins, would be to calibrate on a die-to-die basis. With the help of the build in self test (*BIST*) a minimum sleep voltage can be obtained during test time. While this approach allows a die-to-die compensation of process corners, it requires costly test time. Moreover this methodology can not efficiently compensate for time dependent variations such as voltage and temperature variations or ageing effects.

On die feedback systems can compensate for both die-to-die variations and time dependent variations. Also here care must be taken to ensure the measurement and feedback system mimic the actual SRAM cells as close as possible to minimise power consuming margins. The best results will be attained by the monitor system that suffers from and reacts identically to the variations in process, voltage, temperature and time dependent changes. This thesis proposes such a monitor system based on using multiple monitor cells in parallel. The background information will be provided and the concept will be proven in a commercial 90nm technology.

Such a monitor would be impossible without a measure for the data retention capability of the cells. The noise margins introduced by Seevinck [See87] and Wann [Wan05] for read stability will be discussed and evaluated as will their extensions to measures for retention.

The table 1.1 gives a summary overview of the most important state-of-the-art dual supply SRAMs that have been published in open literature. The numbers in the table are the claims made in the concerned papers. In case the papers reported leakage reduction under the form of leakage power, the numbers have been corrected to only show the current reduction. This is done as the papers from J. Wang [Wan07a], Y. Wang [Wan08], Kim [Kim06] nor Saliba [Sal05] include a means to reduce the sleep supply voltage on chip. The system presented in Nii [Nii04], does include a programmable clamping diode, but no mention is made of the decision algorithm. The system presented in this work will deal with all these aspects. It has to be noted that the 50% leakage current reduction reported for this work in the table is versus an extrapolated current at the nominal supply of 1V to accord with the rest of the papers. The maximum sleep supply in the actual system is limited to 600mV.

This work achieves one of the highest supply voltage reductions with a factor of 5. This reduction is only possible with a reliable data retention due to the implementation of the SNMh monitor circuits. The finest granularity used, can also be found in this work. This limits the power and delay overhead for accesses. In [Wan07a, Wan08] it is not even clear whether the access can occur in a single cycle. The resulting leakage reduction can look pale in comparison with the reported numbers in the other works. However, the technology influence can not be discarded. For instance, in [Nii04] the leakage currents are dominated by the gate-leakage. The reduction of this gate-leakage current is the main contributor to the reduction reported. In this work [Gee08], the gate-leakage current is less than 1% of the total leakage current. This is not uncommon for a low V_T process. The sensitivity of the subthreshold leakage towards the supply voltage is also less than for the gate-leakage. The methodologies used and presented in this work are compatible with other technologies. These methodologies are such in nature they would outperform the other published results while guaranteeing reliable data retention in the same technology.

1.5 Chapter-by-chapter overview

This thesis will consist of 3 main parts, which are reflected in the chapter structure.

Chapter 2 will give an overview of the leakage currents and reduction techniques used in SRAMs. The power consumption in SRAM suffers most from the leakage currents, due the large number of transistors in the matrix. This will lead to the concept of dual

parameter	Nii [Nii04]	Saliba [Sal05]	Kim [Kim06]	J. Wang [Wan07a]	Y. Wang [Wan08]	this work [Gee08]
size	256kiBit	16kiBit	128kiBit	128kiBit	1MiBit	64kiBit
nominal supply	1.2V	1V	1.8V	0.5V	1.2V	1V
sleep supply	0.6V	0.3V	0.9V	0.07V best 0.15V typ	0.5V	0.2V
leakage current reduction	88%	86%	94.2%	83% best 50% typ	90%	50%
delay	2.8ns	3ns	1.02ns	N/A	0.9ns	2.5ns
overhead delay	0	9%	2%	N/A	N/A	25%
overhead active power	0	N/A	“high”	N/A	N/A	≤1%
overhead area	13.2%	3.5%	6%	0.6%	N/A	12.5%
granularity	block	row	matrix	N/A	128kiBit block	word
technology	90nm	150nm FDSOI	180nm	90nm	65nm ULP	90nm HP
data integrity	N/A	worst case	N/A	canary	N/A	SNMh

Table 1.1: Overview of the dual supply SRAMs published in the open literature. N/A notations mean the data was not available from the published material.

supply SRAMs where leakage power saving are made through the use of a secondary lower supply in the matrix. The non-active part of the SRAM matrix will be kept on this lower secondary supply while the active part will be functioning on the nominal supply.

Chapter 3 deals with the consequences of lowering the supply voltage on the matrix. Lower supply voltages mean lower noise-margins. To quantify the retention capability of the SRAM core cells, the concept of Static Noise Margin under hold (*SNMh*) will be introduced. A short discussion on possible alternatives for the bit integrity measure will also be discussed. Using this measure theoretical solutions will be presented to guarantee the data retention of the SRAMs. This leads to the concept of measuring the actual average die SNMh on a die-to-die basis with the developed monitor cells and algorithm. From this discussion it will be shown that active supply control in SRAMs can achieve the greatest savings in leakage power without jeopardising the retention capability.

Chapter 4 will detail the prototype that was designed and fabricated in a commercial 90nm technology. This prototype is a dual port dual supply SRAM. The two ports have different widths to allow power savings on the system level. For this special architecture a new cell was developed. The cell works single ended and consists of 10 transistors. Due to the read-buffer and the separation of the read and write access, the core transistors of the cell could be designed to reduced leakage currents and spread on SNMh. The latter also allows to reduce the margins needed to be taken on the secondary supply voltage to compensate for intra-die variability. The measurement results of the fabricated chip will be discussed.

Chapter 5 will give the general conclusions of this thesis and some reflection on future work.

1.6 Contributions of this work

This thesis contributes in three main domains to achieve the desired results presented in the previous sections.

To minimise the leakage currents in the matrix of an SRAM, a second, lower, supply is introduced for the non-accessed parts of the matrix. This is implemented with the finest granularity feasible: a single word. This granularity has the benefit of minimising wake-up delays while keeping the maximum possible amount of cells in a drowsy state. To achieve this the last stage of the decoder was distributed into the matrix to control the supply switches and wordline activation. This combines simplicity of control, limited area overhead and maximal reductions. The research was published in [Gee05].

A further contribution in this thesis is the development and design of a monitor and regulation circuit to guarantee the data retention. Lowering the supply voltage on the drowsy cells not only reduces the leakage currents, it also reduces the retention capability of the cells. Traditionally, the increase of inter and intra die variations in recent technologies would require margins on the drowsy supply voltage to be taken to compensate for the reduction in retention capability. These margins in turn would deteriorate the possible leakage current savings. As such it is important to be able to guarantee the retention of the data without sacrificing the benefit of leakage reduction. By measuring the actual retention capability of the cells in the live and operational environment, a regulation circuit can be designed to maximise leakage reduction on a die-to-die basis while guaranteeing the data retention. The solution published in [Gee07] fulfils this requirement.

The final contribution of this thesis is in the domain of multi-port memories. The prototype, which is the focus of the research in [Gee08], combines the leakage reduction techniques with an asymmetric width dual port SRAM. The asymmetry in the wordlength enables the reduction of active energy consumption per bit by allowing a wide, 256 bit, and narrow, 32 bit, access. It is proven that the single wide access consumes less energy than several consecutive narrow accesses. This prototype also includes the means to monitor and regulate the secondary drowsy supply.

Chapter 2

SRAM Leakage Reduction

2.1 Introduction

The rise of mobile applications, in which battery life has become a major sales argument, combined with an increasing demand in more performance, created a paradigm shift in mainstream electronics from high speed to low power design. Low power design has become the main challenge for many electronics engineers.

For a long time power reduction in digital CMOS circuits came down to reducing the dynamic power. Power was dominated by the gate capacitances, adding another incentive to scaling. However, with the advent of CMOS technologies with transistor gate lengths under $0.18\mu m$, leakage currents no longer are negligible. Without intervention leakage power would be the dominant power consumption contributor. [Kim03]

The core of most, if not all, digital systems consists of a processor and memories. Where the power consumption of processors consists mainly of dynamic power, memories have a large leakage power component. This is due to the memories consisting out of an array of cells from which only a very small amount are accessed at any one time. The other cells just need to retain the stored data. For volatile, but fast, memories as SRAMs this means there are easily a thousand times more cells “idle” than active. All of those idle cells generate leakage power consumption. Reducing this component has become of the utmost importance to reduce the total system power consumption.

This chapter will first focus on the effects that influence leakage current and the modelling of the currents. In the next sections general techniques to reduce the leakage currents will be discussed and their application within SRAMS. This will lead to the discussion of dual supply memories, where the leakage currents will be reduced by introducing a lower secondary supply voltage.

2.2 Leakage current contributors

2.2.1 Subthreshold leakage

The subthreshold or weak inversion current is the current flowing between source and drain when the gate-source voltage (V_{gs}) is below the threshold voltage (V_T). According to [Roy03] weak inversion currents are dominating off-state subthreshold leakage due to the low threshold voltages used. From the semi-log plots of the transistor current as function of V_{gs} the exponential dependence of the current on V_{gs} is clear in the subthreshold region. This can be modelled by the formula 2.1. Several deep submicron effects influence the subthreshold current and will be discussed in the next sections.

$$I_{\text{sub}} = I_0 \frac{W}{L} \exp \frac{V_{gs} - V_T}{nkT/q} \tag{2.1}$$

where

I_0 : technology dependent constant
W : width of the transistor
L : lenght of the transistor
V_{gs} : gate-source voltage
V_T : threshold voltage
n : subthreshold slope
k : Bolzmann constant
T : temperature in K
q : unity electron charge

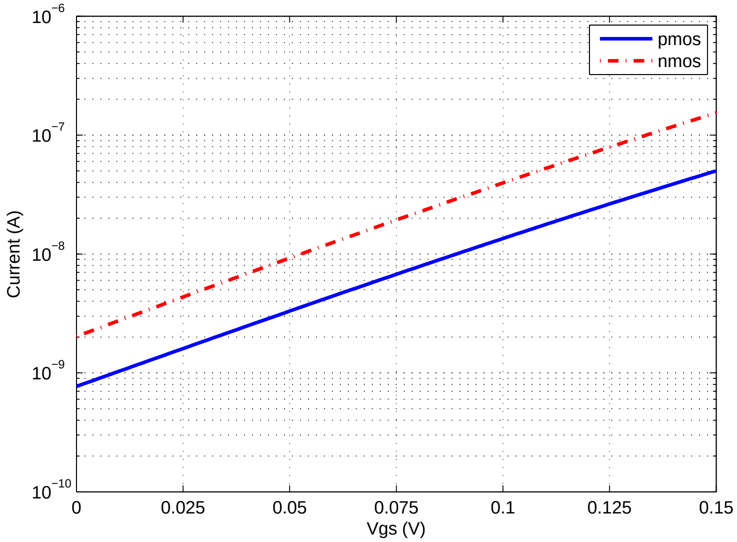


Figure 2.1: Subthreshold current as function of the gate-source voltage (V_{gs})

2.2.1.1 Drain Induced Barrier lowering

DIBL occurs when depletion regions of the source and drain interact near the channel surface. Effectively reducing the source potential barrier and as such reducing the

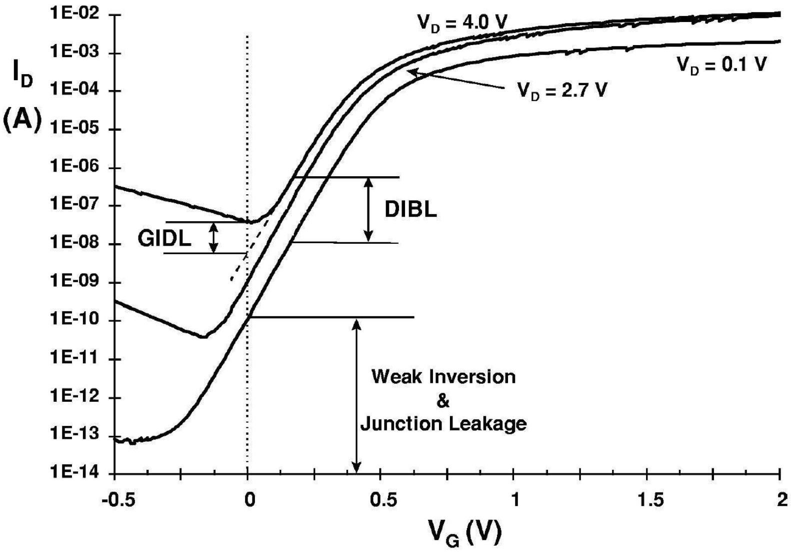


Figure 2.2: Current curves with DIBL [Roy03]

threshold voltage. This process is bound to happen in short-channel devices, where the source-drain voltage has a strong influence on the bandbending over the device.[Roy03]

As shown in figure 2.2 DIBL ideally has no influence on the subthreshold slope, but does reduce the effective threshold voltage. According to [Tay98] and [Der74], the DIBL effect can be reduced by higher surface and channel doping, or by shallow junction depths.

The influence of DIBL on the subthreshold current is illustrated in figure 2.3. The effect is exponential in nature and will be modelled in the current equation 2.2 through the coefficient η.

$$I_{\text{sub}} = I_0 \frac{W}{L} \exp \frac{V_{gs} - V_T + \eta V_{ds}}{nkT/q} \tag{2.2}$$

where

I_0	: technology dependent constant
W	: width of the transistor
L	: lenght of the transistor
V_{gs}	: gate-source voltage
V_{ds}	: drain-source voltage
V_T	: threshold voltage
n	: subthreshold slope
k	: Bolzmann constant
T	: temperature in K
q	: unity electron charge
η	: DIBL coefficient

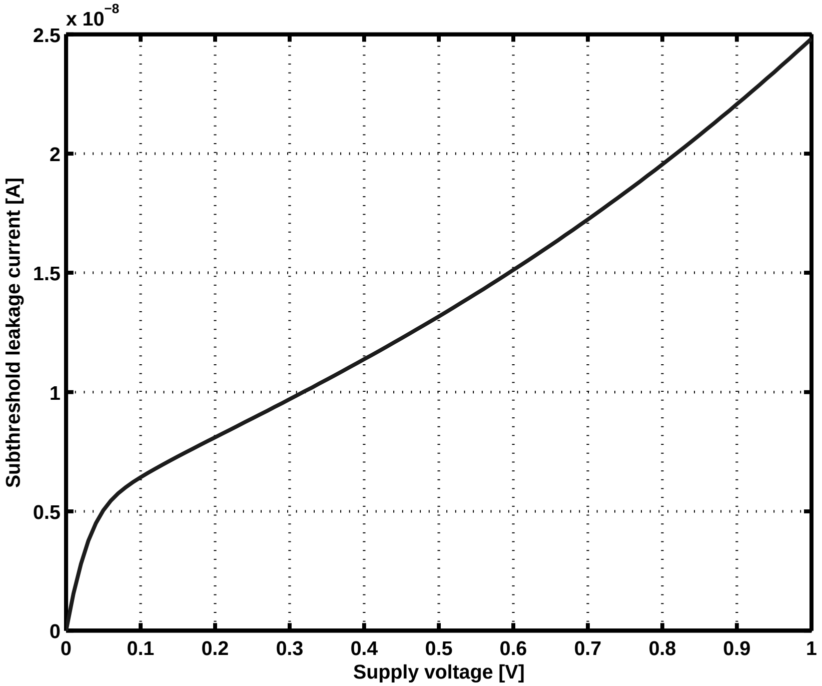


Figure 2.3: Influence of DIBL through V_{ds} on the subthreshold current of a 90nm minimal NMOS

The importance of DIBL increase with smaller technologies. Not only does the smaller feature length bring the drain and source closer together, but also the scaling of voltages is stagnating. Both these effects increase the electrical field in between the source and drain terminals of a transistor.[Roy03, Man03]

2.2.1.2 Short and Narrow Channel Effects

A mechanism similar to DIBL is caused by the geometry of short or narrow channels. When the depletion regions at the source and drain junctions are close enough together they will start to interact with the channel. Thus causing an already partly depleted region in the channel, effectively reducing the potential needed to turn on the device. In other words the short channel effects lower the effective V_T. This phenomena is commonly know as V_T roll-off.[Roy03]

A narrow width of the transistor also influences the effective V_T of the device. Fringe effects of the electrical field cause the gate-induced depletion region to spread outside the defined channel width and under the isolation implants. Additionally higher doping concentrations along the width dimension encroach under the channel stop implants, reducing the effectiveness. Hence a higher voltage is needed to turn the transistors on. In devices using trench isolation, extra depletion regions are formed under the influence of 2-D electrical fringing field. This lowers the effective V_T of the device. [Roy03]

2.2.1.3 Body Effect

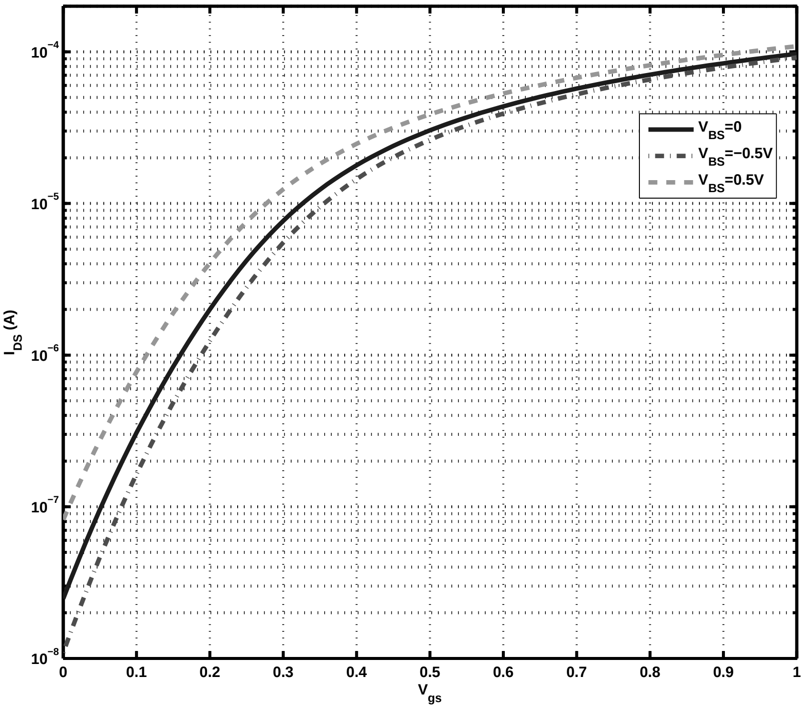


Figure 2.4: NMOS drain-source current curves under different bulk-source bias voltages

Biasing the well-to-source junction modulates the depletion layer width of the transistor and as such influences the effective threshold voltage. Following the analysis in [Roy03] the sensitivity of the threshold voltage to the biasing increases with increasing bulk doping concentration, but decreases with applied bias.

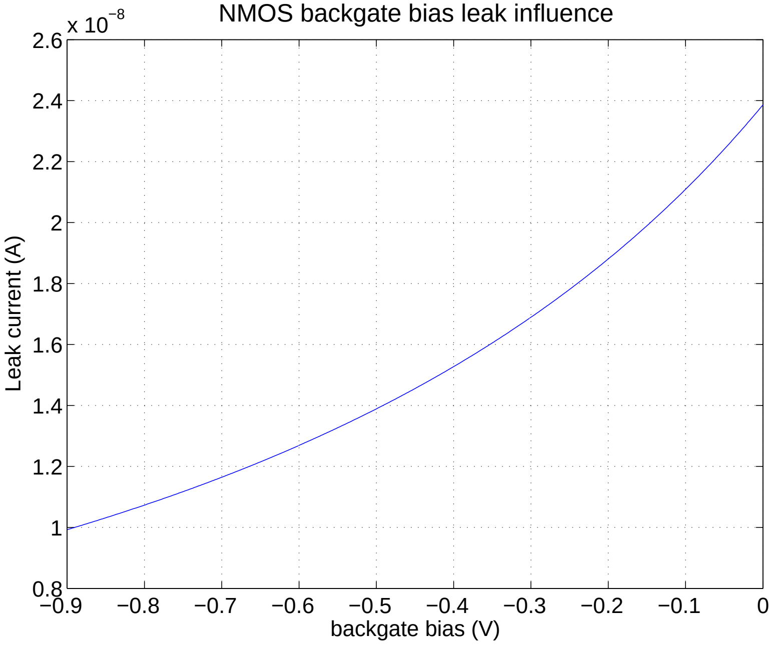


Figure 2.5: Leakage current ($V_{gs} = 0$) values as function of applied backgate bias

As shown in the figures 2.4 and 2.5 the biasing has nearly no effect on the subthreshold slope, but has an exponential influence on the subthreshold current. The influence will be modelled by the body effect parameter γ in the subthreshold current formula 2.3

$$I_{\text{sub}} = I_0 \frac{W}{L} \exp \frac{V_{gs} - V_T - \gamma V_{bs}}{nkT/q} \tag{2.3}$$

where

I_0 : technology dependent constant
W : width of the transistor
L : lenght of the transistor
V_{gs} : gate-source voltage
V_{bs} : bulk-source voltage
V_T : threshold voltage
n : subthreshold slope
k : Bolzmann constant
T : temperature in K
q : unity electron charge

2.2.1.4 Temperature

Any circuit will consume power and as such create an amount of heat. This heat will increase the temperature on die and influence the threshold voltages of the transistors through the thermal voltage (kT/q). The subthreshold current will increase with increasing temperature, creating a positive feedback loop. This effect is know as thermal runaway. By means of various cooling techniques it must be kept under control [Vas06].

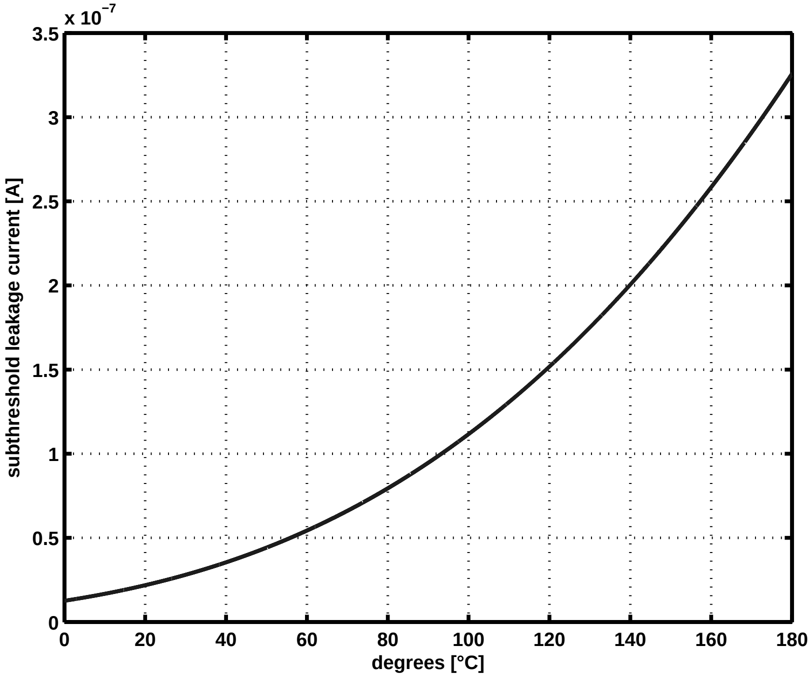


Figure 2.6: Temperature sensitivity of the subthreshold current

2.2.1.5 Full model

To reflect these influences on the subthreshold current with a zero gate-source bias the following formula has been derived [Roy03] :

$$I_{leak} = I_0 \cdot exp(\frac{-V_{T0} - \gamma \cdot V_{BS} + \eta \cdot V_{DS}}{n \cdot V_{th}}) \cdot (1 - exp(\frac{-V_{DS}}{V_{th}})) \tag{2.4}$$

$$I_0 = \mu_0 C_{ox} \frac{W_{\text{eff}}}{L_{\text{eff}}} (V_{th})^2 e^{1.8} \tag{2.5}$$

where

μ_0	: mobility
C_{ox}	: gate oxide capacitance
W_{eff}	: effective width
L_{eff}	: effective length
I_{leak}	: total subthreshold leakage
V_{T0}	: transistor threshold voltage
γ	: linearised body coefficient
V_{BS}	: Bulk-source voltage
η	: DIBL coefficient
V_{DS}	: drain source voltage
V_{th}	: thermal voltage

From formula (2.4) the dependency of the subthreshold current on bulk bias and drain induced barrier lowering (DIBL) is clearly shown. The other effects can be taken into account by the effective width (W_{eff}) and length (L_{eff}) of the device. This formula also shows which degrees of freedom a designer has and their respective impact on the current. From a circuit design perspective, the width and length of the device are determined by the operational constraints of the circuit. The technological parameters are considered to be fixed, although channel engineering to reduce leakage in transistors is a viable option [Man03]. It is clear that temperature plays an important role on the subthreshold current. As such it can be used on a system level to reduce the overall subthreshold leakage of the system by avoiding hotspots and reallocating resources. As this can only be accomplished on the operating system level of an application, this is can be done complimentary to the techniques and methodologies presented in this work. A thorough discussion on this subject falls outside the scope of this work.

If the circuit is not fixed, a circuit designer has two degrees of freedom left to reduce the subthreshold leakage current. The first one is a circuit adaptation to reduce the leakage of the circuit by introducing stacked transistors. While this can accomplish order of magnitude reductions in the leakage current due to the negative gate-source voltage, it comes with a penalty in area or delay. The second degree of freedom is the choice of threshold voltages. This can be accomplished by using differently processed transistors, by modulating it by varying bulk bias or the drain-source voltage. Given the importance of area in SRAM cell design, this thesis focuses on the threshold voltage modulation.

2.2.2 Gate Leakage

With the development of smaller feature lengths for devices also came the reduction of the gate-oxide thickness from 100nm down to 1.2nm . The voltage across the oxide did not scale with the same factor. This leads to a higher electric field across the oxide. As a result the tunnelling of electrons through the gate oxide into the channel and from the channel to the gate becomes possible. This tunnelling current is referred to as gate leakage [Roy03].

Two different modes in gate tunnelling leakage can be distinguished, accumulation and inversion. This is illustrated in figure 2.7 for a PMOS and NMOS transistor. The accumulation currents flow when a transistor is turned off, the inversion currents when the transistor is on. The accumulation currents only make up one tenth or less of the total gate leakage current [Nii04], and as such can be neglected.

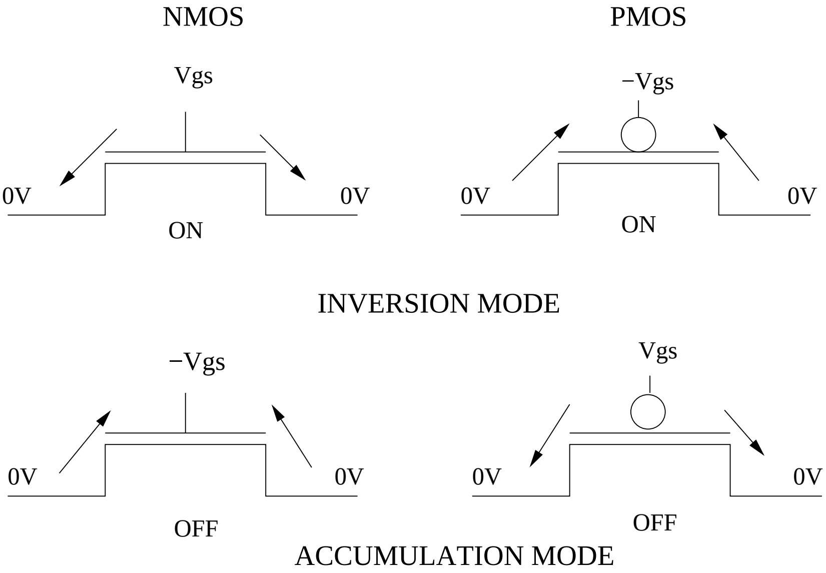


Figure 2.7: Illustration of the gate leakage current in inversion and accumulation of an NMOS and PMOS transistor

Being quantum-mechanical in nature the gate leakage current is virtually temperature-independent. For a given voltage the current can be empirically formulated as [Man03]

$$I_{\text{gate}}(t_{\text{ox}}) = A_0 \cdot exp(-B_0 \cdot t_{\text{ox}}) \tag{2.6}$$

where

I_{gate}	: total gate oxide leakage
t_{ox}	: oxide thickness
A_0 and B_0	: fit parameters

To include the effect of the gate-bulk voltage however, formula (2.6) has to be extended. Simplified equations from [Cha01] yield the formula 2.7. Figure 2.9 plots the gate leakage current as a function of the gate voltage from the simulation setup shown in figure2.8. The current *A* is measured as function of the gate-source voltage *V* on a transistor under nominal supply conditions, e.g. Vdd 1V and Vss as ground node.

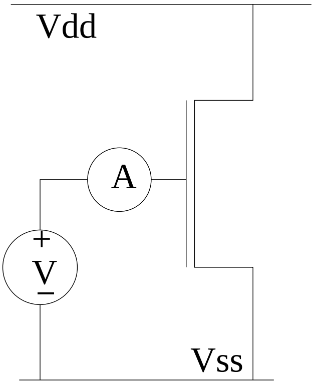


Figure 2.8: Simulation setup to measure gate leakage current

$$I_{\text{gate}} = K \cdot W \cdot (\frac{V}{t_{\text{ox}}})^2 \cdot exp(-\alpha t_{\text{ox}}/V) \qquad (2.7)$$

where

I_{gate}	: total gate oxide leakage
t_{ox}	: oxide thickness
α and K	: fit parameters
W	: gate width
V	: gate-bulk voltage

The ITRS [ITR] does not consider this leakage component to be of any great importance in future technologies, see also figure2.10. The development of gate insulator materials with a higher relative dielectric constant κ, know as high-κ materials, promises a gate leakage reduction. According to the roadmap the development of high-κ materials will allow to compensate for the reduced gate-oxide thickness. The results published in [Mis07] seem to confirm this vision as the high-κ materials used reduce the gate-leakage thousandfold. High-κ materials are expected to become available in the sub-65nm mainstream technologies.

2.2.3 Gate Induced Drain Leakage (*GIDL*)

Gate induced drain leakage is a leakage mechanism due to high electrical field effects in the drain junction of a MOS transistor. When the gate is biased to form an accumulation

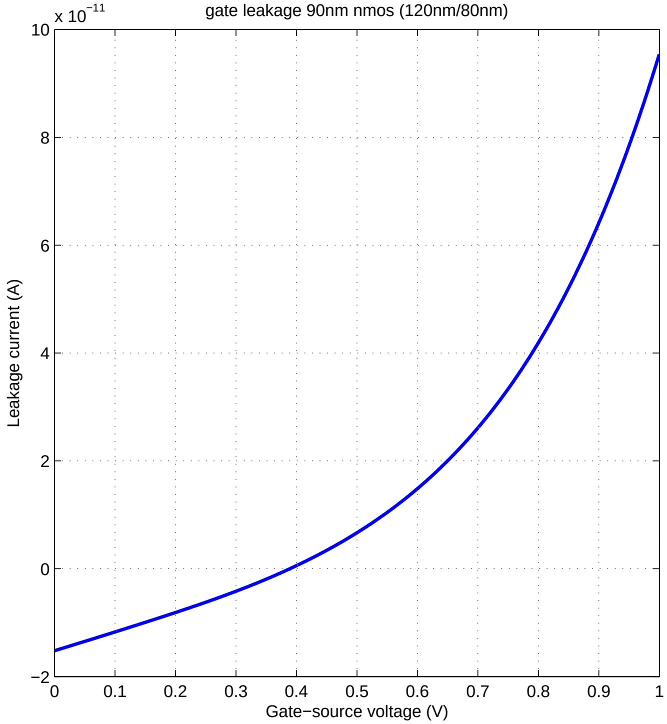


Figure 2.9: Gate leakage of a $120nm/80nm$ 90nm transistor as function of the gate voltage

layer at the silicon surface, the silicon surface under the gate has almost the same potential as the p-type substrate. Due to the presence of accumulated holes at the surface, the surface behaves like a p region more heavily doped than the substrate. This causes the depletion layer at the surface to be much more narrow than elsewhere. This depletion layer reduction results in a higher electrical field near that region, giving rise to effects like avalanche multiplication and band-to-band tunnelling. At low drain doping the field is not big enough to allow tunnelling. At very high drain doping levels the depletion width, equivalent to the tunnelling volume, is very narrow, resulting in a limited tunnelling current [Roy03]. This last statement is partly contradicted by [Man03] where GIDL was still found to be high despite having highly doped drains. In [Man03] several experiments were set up to study the influence of several kinds of doping on GIDL. The results show that great care must be taken in choosing implant materials and doping levels. GIDL is also influenced by gate-oxide thickness, gate bias, degree of abruptness of the doping and the device width [Man03].

The influence of GIDL on the total leakage current of the transistor is however mini-

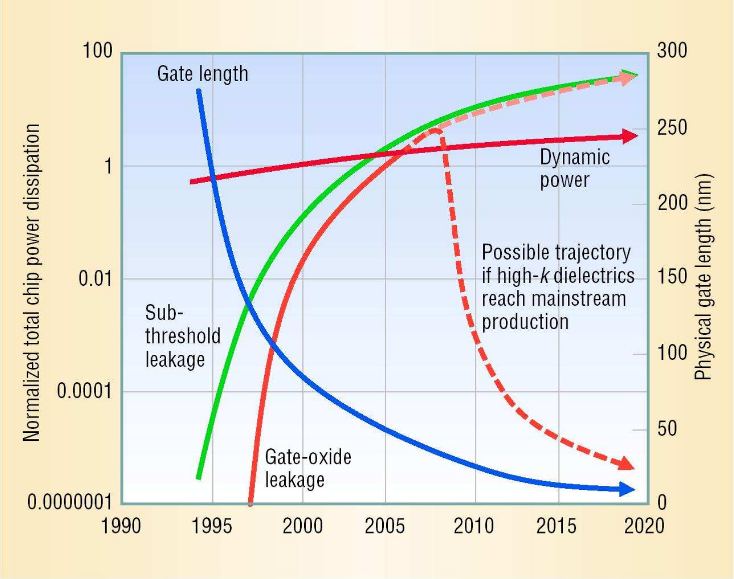


Figure 2.10: ITRS leakage current prediction [ITR, Kim03]

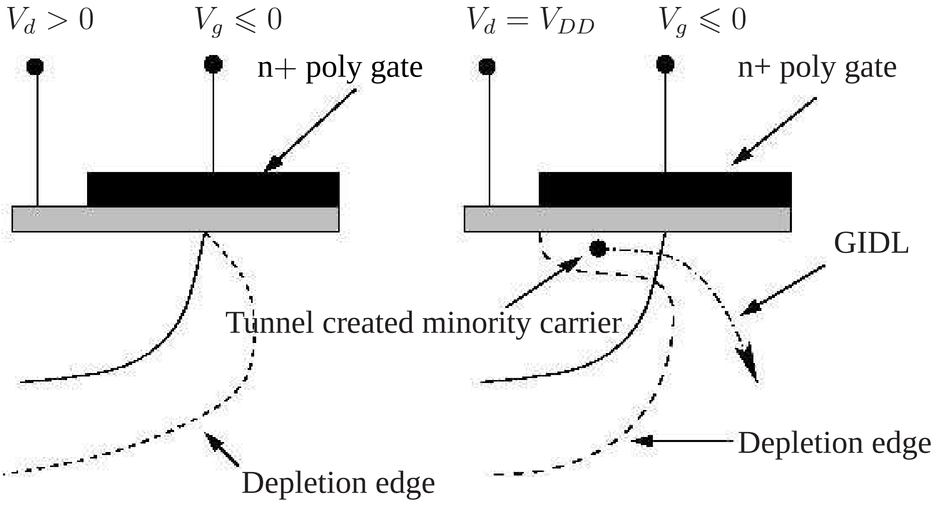


Figure 2.11: GIDL illustrated [Roy03]

mal [Man03]. For technologies with a power supply voltage lower than 1V GIDL is even further reduced as supply voltage comes near the bandgap voltage of silicium, effectively hampering the GIDL mechanism. [Nar06] As such it will be considered negligible in this thesis.

2.2.4 Leakage currents in a SRAM cell

the most prominent source of leakage power is the cell matrix. In any event only a small fraction of the cells will be accessed by the system. The rest of matrix just needs to retain its stored data. Figure 2.12 show the major leakage paths in a traditional 6T SRAM cell, where the access transistors are considered to be off.

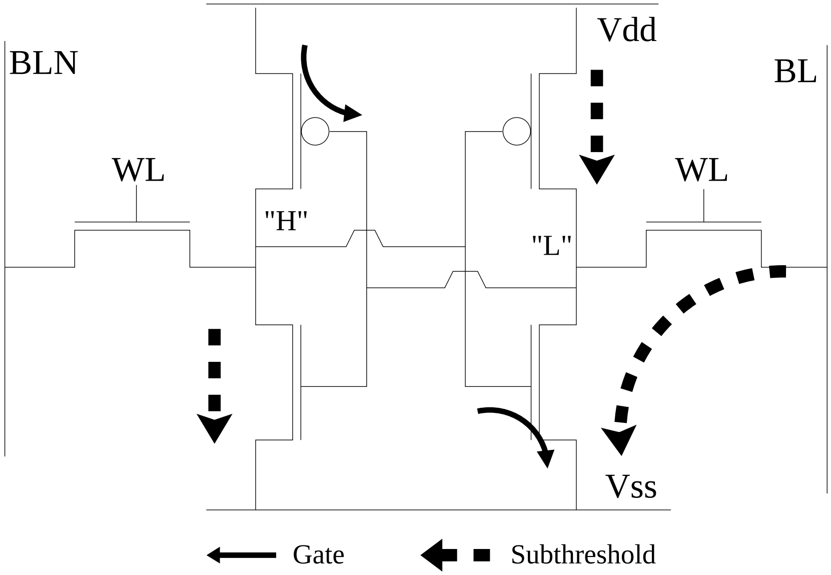


Figure 2.12: Leakage currents in a traditional 6T SRAM cell

In an SRAM cell there will always be an NMOS and a PMOS in the off-state. These transistors will draw a subthreshold leakage current from the supply. The transistors in saturation have another leakage mechanism at work, gate leakage, as they have a voltage difference between gate and drain. One of the pass transistors will also have a large drain-source voltage acrros it, causing a leakage current from the bitline to the grounded node. All these currents have one common defining factor: the supply voltage. To be able to reduce the leakage of the SRAM matrix this dependency will become crucial.

2.3 Circuit level cell leakage reduction techniques

2.3.1 Underlying Principles

To reduce the leakage currents, formulas (2.4) and (2.7) provide the usable parameters. For the subthreshold current it comes down to increasing the effective V_T. This can be achieved by using higher V_T transistors, or by influencing the V_T through the bulk and DIBL coefficient. As these coefficients are only technology dependent this leaves the back-gate bias and the voltage across the drain-source of the transistors as degrees of freedom. For the gate leakage only the voltages across the gate oxide can be used, besides technological measures on the gate dielectric.

2.3.2 Back-gate Biasing

The body effect in CMOS transistors can be used to change the V_T value. By biasing the well-source junction the width of the depletion layer can be modulated. A smaller width will lead to a lower V_T. Biasing can be done in two different ways. Firstly, by reverse biasing the transistor so its effective V_T increases. Secondly, by forward biasing the transistor, the effective V_T value will be decreased.

The effect of body bias is reflected in formula 2.4 by the parameter γ. According to [Roy03], the sensitivity of the threshold voltage increases with increased doping of the channel but decreases with applied bias.

Reverse biasing is a familiar concept in electronics design. By negatively biasing the source-bulk voltage the electric field in the channel gets changed so that the threshold voltage increases. As such the current in the subthreshold region can be dynamically reduced.

Figure 2.13 shows the relative influence of reverse body biasing on the subthreshold current of a minimal NMOS transistor in several commercial technologies and a 65nm experimental technology. The reduced influence of the back-biasing technique is clearly illustrated across the technology evolution. Where the leakage current in a .18μm could be reduced by more than a factor 10 with half the supply voltage as back bias, for 65nm only 40% is expected.

Where reverse biasing is used to increase the threshold voltage of a transistor, forward biasing has the opposite goal. Forward biasing the bulk of a transistor will lower its V_T and is normally used on transistors with a high V_T in an unbiased state. This technique can be used to reduce the V_T on the critical path, allowing the V_T to return to a high value when the transistor is not active. While the body effect still has influence on the threshold voltage in current mainstream technologies, its influence is diminishing with every new CMOS technology generation [vA04].

Formula 2.8 quantifies the backgate bias parameter as function of the oxide capacitance and substrate doping level [Lak94].

$$\gamma = \frac{t_{\mathrm{ox}} \cdot \sqrt{2N_{\mathrm{SUB}}q\epsilon_{\mathrm{Si}}}}{\epsilon_{\mathrm{ox}}} \tag{2.8}$$

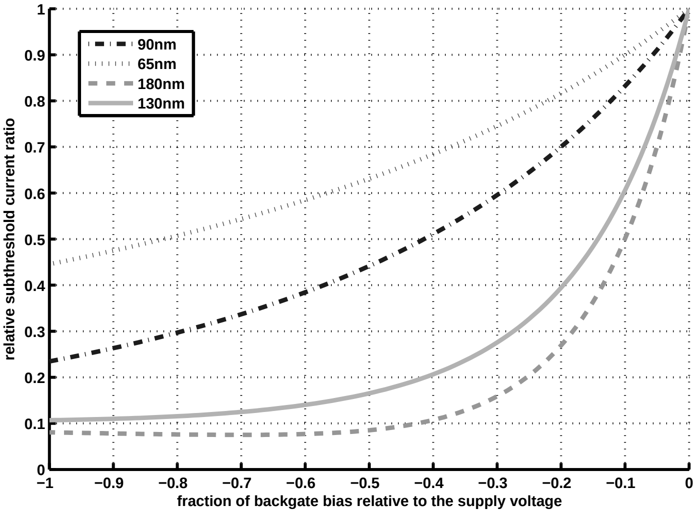


Figure 2.13: Relative influence of reverse body bias on subthreshold current. -1 is the full supply voltage as reverse bias

where

t_{ox}	: gate oxide thickness
N_{SUB}	: substrate doping level
q	: unity electron charge
ϵ_{ox}	: gate oxide permittivity
ϵ_{Si}	: Silicium permittivity

Lowering the oxide thickness as is customary with lower feature lenghts will decrease the backgate bias influence. Reverse biasing will increase the voltage across the bulk-drain and source bulk diode. Forward biasing will increase the currents flowing through the source-bulk and bulk-drain diode. This will further limit the effectiveness of body biasing because of diode breakdown. Increasing the bulk-gate voltage will also increase the gate leakage component.

To be able to implement body bias selectively triple well has to be feasible in the technology as the NMOS bulk contact needs to be separated from the wafer p-doped substrate. This will increase the area a design will occupy.

Applying body bias requires no static currents, and as such allows power efficient DC-DC conversion.

In summary, body bias still has an influence on leakage currents. However, the influence is diminishing with further CMOS scaling [vA04]. To further reduce leakage currents complimentary techniques have to be used.

2.3.3 Multi-threshold

The easiest method of reducing static power components, is the introduction of transistors with different V_T values within the same circuit.

First introduced for digital circuits in [Mut95], higher V_T transistors are used outside of the critical path to reduce static power consumption without influencing performance or more specifically operational speed. This technique combined with cutting the supply to circuits is know as Multi-Threshold CMOS (*MTCMOS*). Since then numerous variations on this method have been published such as Super Cut-off CMOS (*SCCMOS*) [Kaw00]. In SCCMOS also the power to the circuit is cut by using a low V_T transistor, which is biased outside the range of the supply voltage to further reduce the leakage when turned off.

In SRAMs, this technique is harder to apply as all transistors belong to the critical path at some point in time. The delays incurred by the decoders, cells, precharge and write circuits or sense-amplifiers, all contribute to the total delay the SRAM needs to finish its single cycle clock operation. Despite this multi-threshold techniques have been applied in the matrix, more specifically in the SRAM cells.

Solutions where high-V_T and low-V_T type transistors are used in the same cell have been published [Ame08, Ye03]. These solutions try to reduced the leakage in the matrix without compromising on the access delay. To reduce the bitline leakage when Low-V_T passtransistors are used other special techniques, like reducing the WL voltage, are required [Ye03].

2.3.4 Multi-supply

Another way of reducing the leakage currents is through the application of different supply voltages. In digital systems the drain-source voltage across a transistor is closely linked to the supply voltage. For the simplest case, namely an invertor, they are even identical. In equation 2.4 its influence is clearly shown. The second exponential factor in this equation is actually negligible when the V_{DS} stays higher than a few times the thermal voltage (25mV@300K). Due to the Drain Induced Barrier Lowering effect (*DIBL*), the effective V_T of a transistor is influenced by the drain-source voltage across the transistor, as discussed in section 2.2.

The most extreme version of this technique is also the oldest, namely power gating. In this method the supply to a non-active part of the circuit is simply cut, effectively reducing the subthreshold leakage current to near nil. An example of such a technique is SCCMOS as mentioned in the previous section.

Both MTCMOS and SCCMOS suffer from long wake up delays and current spikes on activation. Zigzag Super Cut-off CMOS (*ZSCCMOS*), proposed in [Dra04], remedies by eliminating the series connection between the power switches. This is accomplished by structuring the circuit in such a way that the logic gates are alternately switched off by an NMOS or PMOS. Further methodologies have been developed to incorporate reduction techniques for gate leakage currents, like Gate leakage Suppression CMOS (*GSCMOS*) [Dra04]. In GSCMOS a third virtual supply rail is introduced with a separate power switch. By connecting all gates on the virtual ground rail with a powerswitch on the PMOS side to this third virtual rail, all leakage paths can be eliminated. However, this reintroduces the problem of series connections on the power rails that leads to a larger wake up delay than ZSCCMOS.

While pure combinatorial circuits suffer nothing but a wake up delay from cutting power, state based systems or data storing systems will lose their state. Preventive action can be undertaken to avoid this. Registers can be excluded from the power switching for instance. For the matrix of an SRAM it is clear this offers no solution. For memories where the main goal is to maintain the stored data even under “turned-off” conditions, simply cutting the supply to memory cells is unworkable. This has given rise to the development of drowsy or sleepy memories [Fla02]. Here a secondary reduced voltage across the cells is introduced into the matrix to profit from the exponential decrease of the leakage components with lowered supply voltage, while keeping this supply high enough to maintain the stored data. The supply rail of the cells will be switched between a sleep voltage and an active voltage. As the cell supply rail is not considered to be a real supply rail, it is referred to as a virtual supply rail. The reduction of the leakage currents can be achieved by lowering the voltage on a virtual supply rail [Nii04, Kua05], by increasing the voltage on a virtual ground rail [Zha05] or by a combination of both.

Increasing the voltage on the virtual ground rail has the added benefit of back-biasing the pulldown NMOS transistors to further reduce the leakage current. However, it also introduces an extra NMOS transistor in the pulldown path, which can reduce the read current of the cells if special care is not taken. To avoid this influence extra scaling for this NMOS power switch will have be taken into account, increasing the area penalty. The use of PMOS power switches on the virtual supply rail does not influence the read current once the virtual supply is at nominal level.

The granularity of these drowsy sections goes from entire banks [Fla02] to separate words we introduced in [Gee05] and various grades in between.

Multi-supply techniques require DC-DC conversion to be available. With the goal of reducing power consumption as much as possible, this conversion should be as efficient as possible to benefit from both the current and supply reduction. High efficiency DC-DC converters integrated with SRAMs and micro-processor have been reported in open literature [Kwo08]. These implementations however come at a high cost in complexity and area and as such are unsuitable to be used solely for the generation of a second supply in an SRAM.

2.4 Dual Supply SRAMs

2.4.1 Introduction

In SRAMs two functional features are important, a low access time and retention of the stored data. To this end using two supplies is optimal to allow the SRAM to run in different modes while satisfying these requirements. A nominal supply voltage is used to power the peripheral circuits such as decoders, sense-amplifiers and control circuitry. These circuits can be powergated with techniques from the previous section 2.3. A secondary lowered supply is introduced into the matrix to lower the leakage currents in non-activated cells.

2.4.2 Granularity of control

The basic idea can be implemented in several ways. In [Fla02] whole banks are put in a drowsy mode and require one or more clock cycles to be accessed. Resulting power saving are predicted in the order of 50% to 75% , although no information is given concerning the reliable retention of the data.

The granularity of the drowsy section was further refined in [Zha05] to the level of rows in the SRAM matrix. Instead of lowering the supply voltage however the virtual ground level was raised with similar effect. This gives the added benefit of further reducing the leakage through the NMOS transistors in the cells through the body effect. No wake-up delays or wake-up power numbers were mentioned in the article. Leakage power savings were claimed to be near 90%.

The finest granularity of control was first introduced in the course of our research for this thesis and subsequently published in [Gee05]. Here only a single word is switched between the drowsy and active supply. This has the benefit of creating the lowest wake up delay combined with the highest power savings. This is due to the small parasitic capacitance of the virtual supply rail that has to be recharged to the nominal level as only the smallest needed section of the matrix is woken up. Another benefit of this fine granularity is the limited current peak it creates on the global supply when switching to the nominal level. The schematic used is shown in figure 2.14.

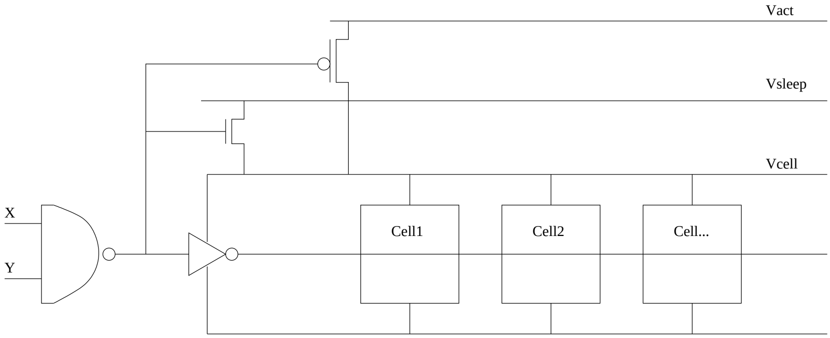


Figure 2.14: Word level schematic of the finest granular system [Gee05]

To accomplish the fine granularity of control, the decoder signals from the X and Y direction have to be recombined at word level. By distributing the final section of the decoder into the matrix and recombining right in front of the cells forming the word, the control structure for the fine granularity does not require a huge area overhead. The area overhead estimated for the solution presented in figure 2.14 is 16% for a 16bit wide word.

The resulting subdivided wordline not only reduces the global power consumption needed to drive the global wordline or the number of activated bitlines [Hir90]. The last driver for the local wordline is connected to the same virtual supply rail as the cells. This guarantees that the access transistor will not be overdriven and jeopardise the integrity of the stored data. It solves the timing problems of the previously published work elegantly.

2.4.3 Power Savings

Figure 2.15 shows the power saving possible in the matrix with the presented systems including the dynamic power overhead for the 65nm IMEC technology in function of the matrix size. A higher number of cells reduces the dynamic power influence relatively as more cells will be kept in a sleepy state. The non-continuous behaviour of the curve is due to limitations in feasible aspect ratios for the matrix.

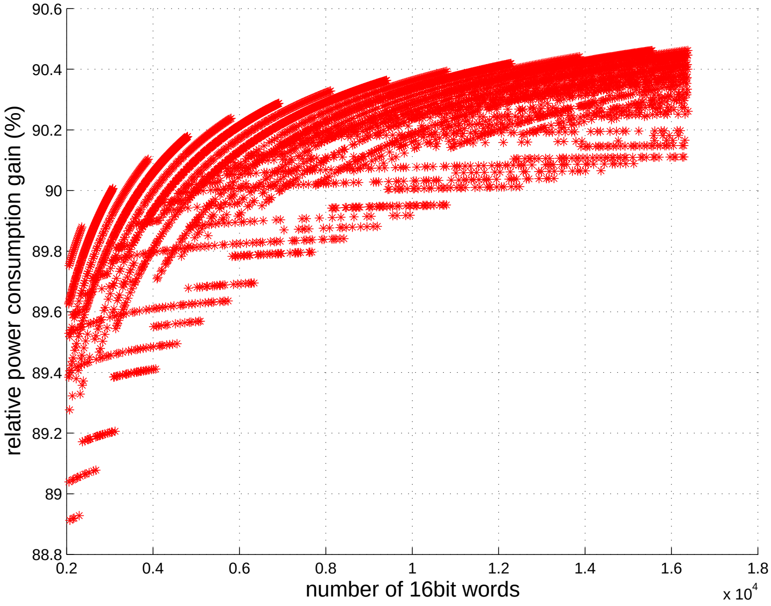


Figure 2.15: Relative modelled gain in power reduction in the matrix as function of the matrix size. Taking into account a 100MHz access rate and a sleep voltage of 400mV in IMEC 65nm technology. [Gee05]

The leakage power of an SRAM matrix can be easily modelled as formula 2.9. The leakage power is linearly depended on the number of cells in the matrix. The currents in the formula are modelled as described in section 2.2.

$$\begin{aligned} P_{\mathrm{leak}} = & V_{\mathrm{sleep}} \cdot \left(In_{\mathrm{off}} + Ip_{\mathrm{off}}\right) \cdot \left(N - W\right) + V_{\mathrm{prech}} \cdot \left(2I_{\mathrm{pass}_{\mathrm{off}}}\right) \cdot \left(N - W\right) \\ & + V_{\mathrm{act}} \cdot \left(In_{\mathrm{off}} + Ip_{\mathrm{off}}\right) \cdot W \end{aligned} \tag{2.9}$$

where

V_{act} : the active supply voltage
V_{sleep} : the sleep voltage of the cells
In_{off} : Cut-off current of the SRAM cell pulldown NMOS
Ip_{off} : Cut-off current of the SRAM cell pullup PMOS
N : the number of SRAM cells in the matrix
W : the number of SRAM cells woken up to be accessed
V_{prech} : the precharge voltage of the bitlines
$I_{\mathrm{pass}_{\mathrm{off}}}$: cut-off current of the passtransistors of the SRAM cell

In case conversion is needed to acquire the necessary voltages, the voltages in the previous formula can be easily amended to include the conversion efficiencies. This can be done by multiplying them with $1/\eta$ where η is the conversion efficiency.

From formula 2.9 it is clear that the highest reduction of leakage power will be achieved with W as small as possible. The minimal for W being the word width of the stored data [Gee05]. The fine granularity also benefits the wake-up power needed and the wake-up delay of the word as the virtual supply rail capacitance is minimised.

The dynamic power consumption in the matrix can be modelled with the familiar formula 2.10. The signal swing voltages and load capacitances depend on the matrix and decoder structure. If the matrix aspect ratio in terms of cells is kept square-like the wordline and bitline capacitance would scale $O\left(\sqrt{N}\right)$, the decoder would scale $O\left(\log_2\left(N\right)\right)$. Where N is the number of cells in the matrix.

$$P_{\mathrm{dyn}} = C_{\mathrm{load}} \cdot V_{\mathrm{supply}} \cdot V_{\mathrm{swing}} \cdot f/2 \tag{2.10}$$

where

P_{dyn} : dynamic power
C_{load} : Load capacitance
V_{supply} : supply voltage
V_{swing} : voltage swing
f : frequency

Formula 2.10 is used to estimated the dynamic power of all active components. This allows the estimation of the power savings possible in the system as illustrated in figure 2.15. The full matlab implemented model can be found in appendix A.

Formulas 2.9 and 2.10 also allow to make the evaluation of the different granularities that can be used to control the sleep section of the matrix. A granularity with a large

W compared to N will reduce the active leakage power savings and increase the active enerygy overhead of waking up the cells. The first effect is due to more cells unnecessarily contributing to the leakage with a high leakage current. The second effect comes into play in the supply capacitance. More cells to wake up means more supply capacitance to charge to the active supply voltage. If W remains small compared to N, both effects will be negligible in the total power consumption. In the context of the prototype further discussed in chapter 4, 8 words make up a single row. With a leakage reduction of 33% between active and sleeping cells, the active leakage contribution of this row would be reduced with 30% versus a row based approach. Compared with larger granularities this advantage will only increase in importance. As discussed before in section 2.4.2, the fine wordsized control granularity has additional benefits in reducing the active energy as only the bitlines that need to be accessed will be discharged and the effective wordline capacitance is reduced [Hir90].

Where figure 2.15 illustrates the possible power savings if the second supply can be generated with high efficiency, such DC-DC converters are not always feasible to be integrated with the SRAM. The secondary supply can also be generated by a series regulator, in which case only the current saving will contribute to the overall leakage power savings. Figure 2.16 shows the power savings possible when using a series regulator. The lower boundary of the supply voltage is considered to be 200mV as with lower voltage the data in the SRAM would be lost.

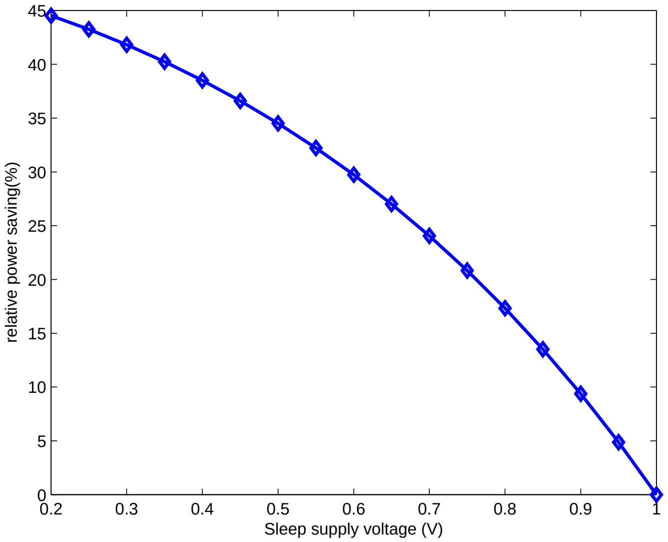


Figure 2.16: Possible estimated power savings using a series regulator in 90nm

2.4.4 Dual supply SRAMs summary

To reduce the leakage power in SRAMs, the dual supply approach is very effective. By keeping non-accessed cells in a sleepy or drowsy state where they retain the stored data and waking the needed cells up to the nominal supply, two key features of SRAM can be achieved. Access times will not be degraded significantly and the data will be retained, while both subthreshold and gate leakage currents are minimised.

The finest granularity of control has the advantage on both leakage reduction and minimised delay degradation. Only waking up the minimal amount of needed cells, the word width, the highest amount of cells can be kept in a sleepy leakage reduced state. This has the added benefit of minimising the capacitance on the virtual supply rails. As this capacitance determines the wake-up delay and power,including the current peaks on the global supply line, the overhead of both is kept minimal.

The leakage power of the system can even be further lowered by power gating the peripheral system when possible.

2.5 Chapter Conclusion

Leakage power has become the most dominant factor in the power consumption of SRAMs. To achieve low power operation SRAM leakage power is a major concern for almost any system.

The analysis of effects contributing to leakage is made in section 2.2. The resulting formulas for subthreshold and gate leakage show the parameters that can be used to reduce the leakage currents. The supply voltage plays an important role in both subthreshold and gate leakage as it defines both the drain-source as the gate voltage.

The techniques making use of one or more of the possible parameters are discussed in section 2.3. For subthreshold leakage the techniques focus on increasing the threshold voltage V_T, permanently of temporarily. The permanent solution consists of using transistor with high V_T where possible without influencing the speed performance of the SRAM. This can be combined with power gating peripheral circuits in the SRAM to further reduce the leakage in a non active state.

Techniques temporarily changing the V_T of a transistor can be divided into two categories. One technique consists of back-biasing the transistors so their effective V_T changes. This can be done in both forward and backward directions. However, the back-gate bias sensitivity of the V_T lowers with each smaller technology. As such this technique is losing effectiveness. The other category of temporarily changing the V_T value is based on the DIBL effect. The drain-source voltage across a transistor also influences the effective V_T of the transistor. This effect becomes more important with each shrinking technology node as the drain and source area come closer together.

Section 2.4 discusses the possible implementation of the reduction techniques in the specific environment of an SRAM. This leads to the introduction of dual supply SRAMs (*DSSRAM*). In a DSSRAM a secondary supply voltage is introduced, either through means of a virtual supply rail, a virtual ground rail or a combination of both. This lower supply voltage is applied to cells that are not active but just retaining data. As the cells in this drowsy mode are isolated from the bitlines, disturbing the stored data becomes less likely. The effectiveness of this approach is highly dependent on the technological parameters, especially the DIBL effect. Due to the reduced leakage current and the lowered voltage, leakage power savings up to 95% are possible in the SRAM cell matrix when a highly efficient DC-DC conversion is present.

The dual supply system can be applied in several granularities, the largest being the whole SRAM, the smallest consisting of just a single data word. This granularity has an influence on power savings, delay and control overhead. Crude granular architectures such as whole memories or banks of a memory have the simplicity of control, but suffer from drawbacks in wake-up delay and wake-up power. Waking up a whole memory to retrieve a single word has a large dynamic and passive energy cost. While at first sight the finest granular structure might be the hardest to control, we solved this issue by distributing the last stage of the decoder into the matrix [Gee05]. Combining the X and Y-signals locally before the word achieves several benefits with a small area overhead. Firstly, only the word that is needed is woken up, with a small power and delay penalty. Secondly, it creates a hierarchical wordline structure that reduces the capacitance on the global wordline and as such the dynamic power needed to drive this full-swing line. As the buffer strength can be scaled down also the leakage components originating from the buffers is diminished. Lastly, it solves one of the timing problems associated with drowsy bits. As their supply voltage is reduced, the cells become more susceptible to external influences, especially in the read case. An access transistor that would turned on fully before the cells has reached its nominal supply, compromises the stored data. By having the last stage of the decoder distributed and controlling the power switches, the last stage of the wordline buffer can also be localised and connected to the virtual supply rails. This ensures the driving voltage of access transistors scales the same way as the supply voltage of the cells waking up.

Using this fine granular system makes it possible to maximise power savings and minimise wake up delays. As such it will be used for remainder of the thesis.

Chapter 3

SRAM Data Retention

3.1 Introduction

The functional definition of a Static Random Access Memory can be described as a functional block that is able to retain the data written to it and output that data when required to do so. While this is a simple enough concept, filling it in brings numerous design challenges with it.

On the system design level the quest to reduce power consumption drives supply voltages down as far as possible. Both active and leakage power consumption are dependent on the supply voltage. Active power has a quadratic relationship , while leakage currents depend exponentially. SRAMs, where the leakage is a large portion of the total power consumption, form no exception to this trend. However, special care must be taken not to compromise the stored data integrity as the noise margins are reduced by lower supply voltages.

Another factor that has to be taken into account is the high variability in deep submicron technologies. The effect of variability is mostly felt as an increase in mismatch between transistors. The coupled transistors in the core SRAM cell will effectively have asymmetric behaviour, disturbing the bistable positive feedback loop. This results in reduced reliability for all operations.

In this chapter the current and proposed approach to solving the reliable retention of the data will be discussed. First the noise margin definition will be discussed, followed by a section detailing the use of those definitions in current design methodologies. Thereafter the more novel ways of dealing with the variability of the new processes will be discussed with respect to design. This will start with the introduction of real-time monitors, the need thereof and the proposed solution. The final section of this chapter will discuss the possible implementations of the proposed solution.

3.2 Noise Margin Definition

3.2.1 Introduction

To be able to evaluate the capability of an SRAM to retain data, a standard measure is needed. The Static Noise Margin (*SNM*) introduced by Seevinck [See87] is the oldest of such measures. The more recent N-curve measure, which can be used in combination with SNM, was introduced by Wann [Wan05]. Where SNM is fully focused on voltages, the N-curve approach brings current characteristics into the measure to give a more complete insight into the cell stability.

While traditionally the read situation is the worst case scenario for SRAM cell stability, the bit integrity can also be degenerated by lower supply voltages. This has to be weighted against the beneficial effect of lowering the supply voltage as it results in an exponential decrease of the leakage currents. To maximise the leakage power savings the supply voltage should be lowered as far as possible without jeopardising the stored data. To this end the SNM and N-curve measures have to be extended to provide bit integrity information.

This section will deal with the afore mentioned measures, SNM and N-curve. In a first step they will be briefly introduced and defined. The second step will discuss the extension of the measures to the characterisation of stand-by retention.

3.2.2 Static Noise Margin

3.2.2.1 Traditional definition

The concept of Static Noise Margin (*SNM*) with its mathematical background was first published by Seevinck in [See87]. SNM is defined to be the voltage margin the circuits has before it changes state in the worst case DC situation.

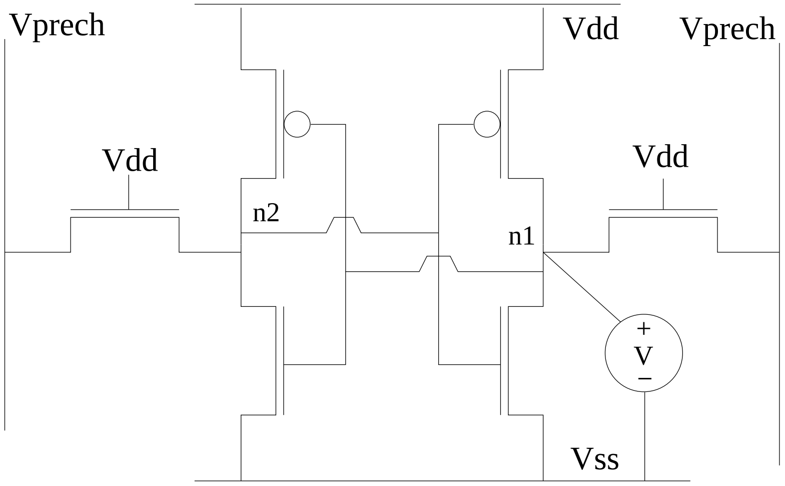


Figure 3.1: SNM measurement setup [See87]

This definition is still used today to determine the stability of SRAM memory cells. The worst case situation for data integrity is during read access for six or five transistor cells. According to the definition the cells are analysed statically under access conditions, meaning the passtransistors of the cell are turned on and the bitlines kept to their precharge value (*Vprech*). The internal storage nodes will be swept by a voltage source

(*V*) as shown in figure 3.1. This is a valid approximation of the read access situation in case the bitline capacitances are large compared to the read current of the SRAM cells. In case the bitlines capacitances were to be small, this would be an overly pessimistic view. [Cos07].

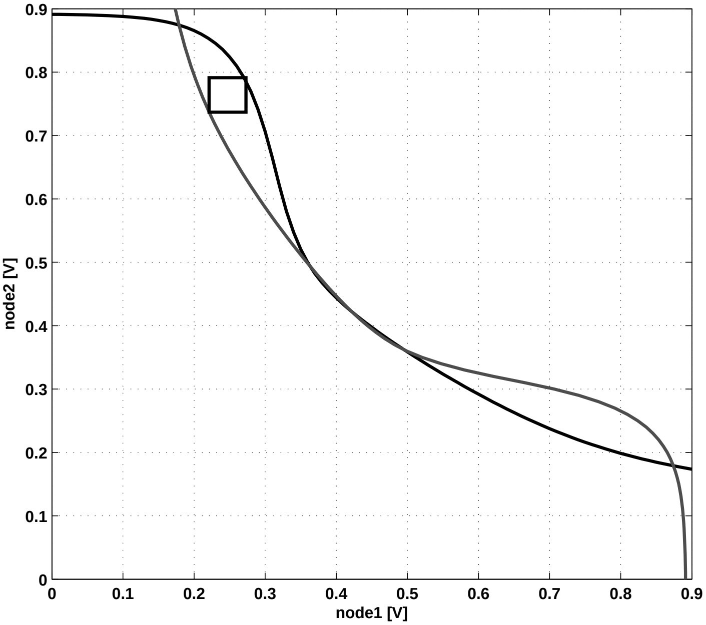


Figure 3.2: SNM graphically represented [See87]

SNM can be graphically measured on a butterfly curve. Butterfly curves are generated by plotting the DC transfer characteristics of the cell halves on the same axes. SNM is defined as the side of the minimum of the biggest squares that fit in the butterfly diagram as illustrated in figure 3.2. This is similar to the noise margin definition used in communication systems, the eye opening.

Mathematically the SNM curve can be extracted from the inverter DC transfer curves by transposing them to a 45-degree rotated axes system. The extrema of the opening can then be easily derived. SNM is then extracted on the bases of a minimax criterion from the extrema. [See87] This approach will be used in section 3.3.4.3.

3.2.2.2 Expansion to stand-by Retention

In SRAMs without matrix supply reduction, the worst possible case for data loss is during access of the cell. This is were the traditional definition of SNM focussed. When the cells have a reduced supply voltage in a drowsy or sleepy state, the worst case scenario can shift. Although the access transistors in this state are in cut-off, the reduction in supply voltage brings a reduction in noise margin for the cell as illustrated in figure3.4. Using the same methodology as described in the original definition of SNM, a similar noise margin can be measured on the sleepy cell. This margin, further referred to as static noise margin under hold (*SNMh*), is a measure for the cell's ability to keep the data under reduced voltage.

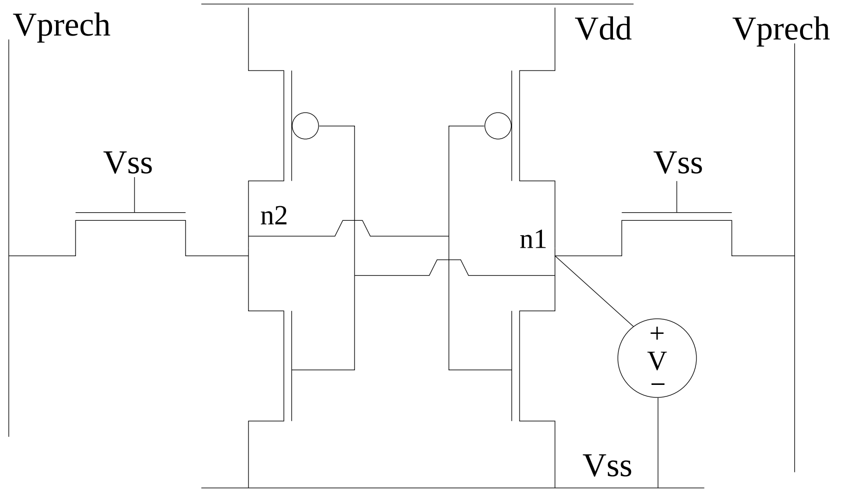


Figure 3.3: SNMh measurement setup

The SNMh evolution as function of the supply voltage is shown in figure 3.5. For the region between subthreshold and saturation the evolution of SNMh has an approximate linear relationship with the supply voltage. Although [Wan07b] and [Wan07a] imply this relationship to be valid over the whole supply, in the case of figure 3.5 the approximation is only valid between 100mV and 600mV. This can be expressed as formula 3.1.

$$\Delta SNMh \sim \frac{\Delta Vdd}{2} \tag{3.1}$$

The probability distribution of the SNMh of an individual cell under influence of transistor process variations is shown in figure 3.6. As SNMh is actually defined as a minimax criterion, a Gaussian distribution is not a good approximation.

The probability distribution of SNMh can however be calculated from the distribution of the two extrema. The distribution of the minimum of two distributions can be written in term of the probability density function (*pdf*) and the cumulative density function (*cdf*). Let f_i be the pdf of i,and F_i the cdf. The pdf of the minimum of two distributions can then be written as 3.2, under the assumption of independence.

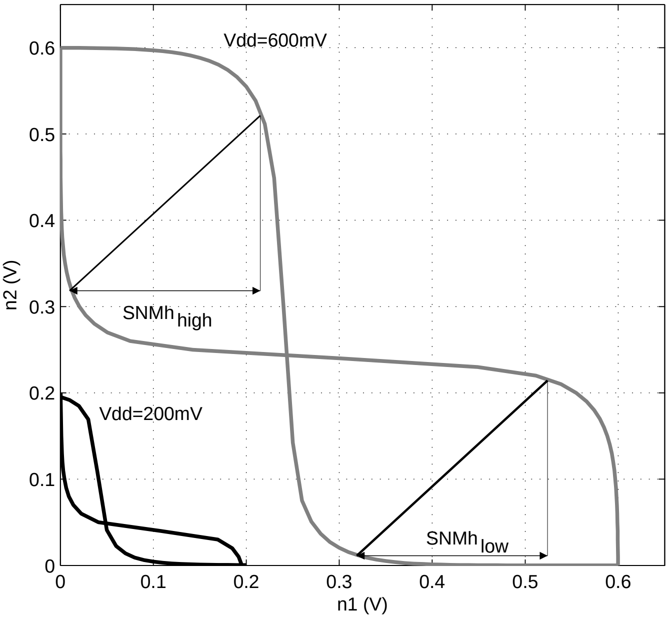


Figure 3.4: Retention butterfly curve evolution under varying supply

$$f(\min(x_1, x_2)) = f_1 \cdot (1 - F_2) + f_2 \cdot (1 - F_1) \tag{3.2}$$

The cumulative density function $F(x)$ can be written as 3.3:

$$\begin{aligned} F(x) = &\int_{-\infty}^{x} f(y)dy \\ = &\int_{-\infty}^{x} f_1(y) \cdot (1 - F_2(y))dy + \int_{-\infty}^{x} f_2(y) \cdot (1 - F_1(y))dy \\ =&F_1(x) + F_2(x) - F_1(x) \cdot F_2(x) \\ &+ \int_{-\infty}^{x} F_1(y) \cdot f_2(y)dy - \int_{-\infty}^{x} F_1(y) \cdot f_2(y)dy \\ =&F_1(x) + F_2(x) - F_1(x) \cdot F_2(x) \end{aligned} \tag{3.3}$$

The two extrema of the SNMh function, marked as $\text{SNMh}_{\text{high}}$ and SNMh_{low} in figure 3.4, can be modelled as normally distributed [Cal06]. This can be seen on the semilog

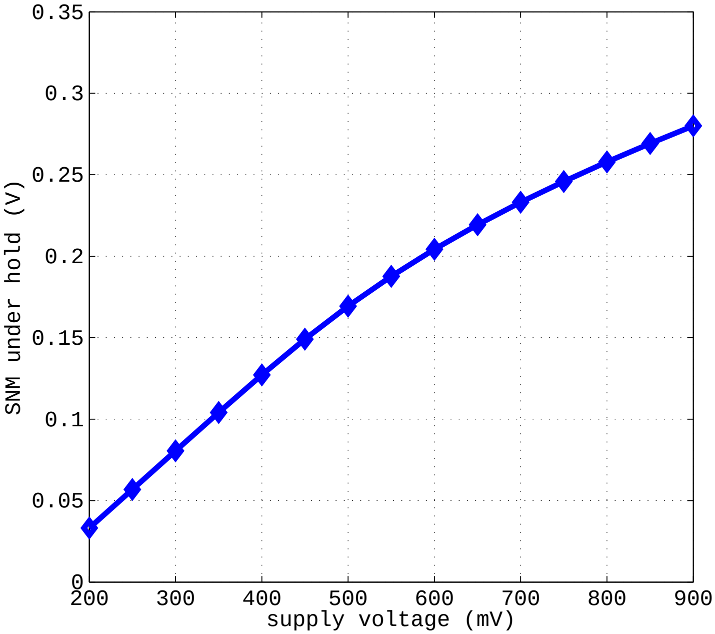


Figure 3.5: SNMh evolution under varying supply

plots 3.7 and 3.8, where the distribution is plotted versus a Gaussian fit of the data. They are however correlated and not identically distributed. To simplify further calculations the extrema will be considered to be uncorrelated. The inaccuracy introduced due to this approximation hampers the ability of the model to describe the SNMh distribution correctly over the entire range. However the tail of the distribution is modelled accurately as illustrated in figure 3.9. [Wan07b] It will be this tail that characterises the loss of data as this will represent the worst cells. As such it will have the most influence on the final yield of the matrix. The formulas 3.2 and 3.3 can then be rewritten as 3.4 and 3.5

$$\begin{aligned} f_{\text{SNMh}} = {} & \frac{1}{2\sqrt{2\pi}} \text{erfc}\left(\frac{x-\mu_h}{\sqrt{2}\sigma_h}\right) \cdot \exp\left(-\frac{(x-\mu_l)^2}{2\sigma_l^2}\right) \\ & + \frac{1}{2\sqrt{2\pi}} \text{erfc}\left(\frac{x-\mu_l}{\sqrt{2}\sigma_l}\right) \cdot \exp\left(-\frac{(x-\mu_h)^2}{2\sigma_h^2}\right) \end{aligned} \tag{3.4}$$

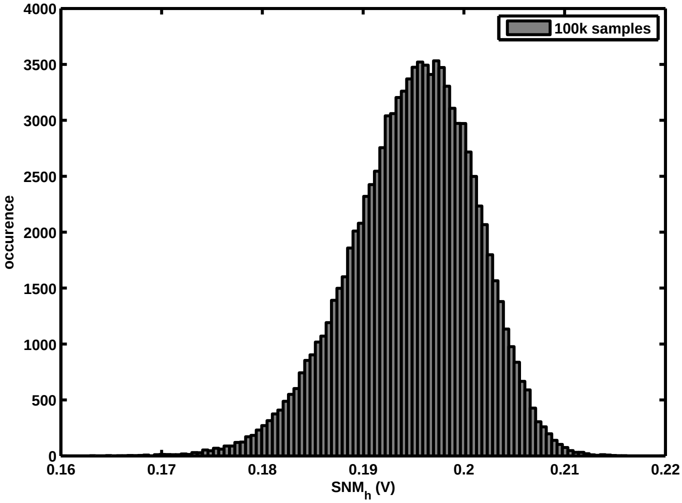


Figure 3.6: example SNMh distribution for an SRAM cell in 90nm at 600mV supply

$$F_{\text{SNMh}} = \frac{3}{4} + \frac{1}{4}\left(\text{erf}\left(\frac{x-\mu_l}{\sqrt{2}\sigma_l}\right) + \text{erf}\left(\frac{x-\mu_h}{\sqrt{2}\sigma_h}\right) - \text{erf}\left(\frac{x-\mu_l}{\sqrt{2}\sigma_l}\right) \cdot \text{erf}\left(\frac{x-\mu_h}{\sqrt{2}\sigma_h}\right)\right) \tag{3.5}$$

where

μ_l : mean of the single cell SNMh$_{\text{low}}$ distribution
μ_h : mean of the single cell SNMh$_{\text{high}}$ distribution
σ_l : standard deviation of the single cell SNMh$_{\text{low}}$ distribution
σ_h : standard deviation of the single cell SNMh$_{\text{high}}$ distribution

3.2.3 N-Curve definition

The resilience of SRAM cell towards data disturbance is not only a function of the voltage as defined by Seevinck [See87]. Under that presumption the noise margin of a cell would be almost insensitive to sizing [Gro06]. Doubling all transistor sizes in an SRAM cell would barely change the SNM, while the current needed to disturb the data would have to be twice as high. Wann [Wan05] defined three other measures that do include the influence of the current.

The measurement setup for the N-curve [Wan05, Gro06] depicted in 3.10 is similar to the setup used for the SNM as defined by Seevinck [See87]. The voltage source V is swept from ground to the supply level. The drive and sink current of this source is

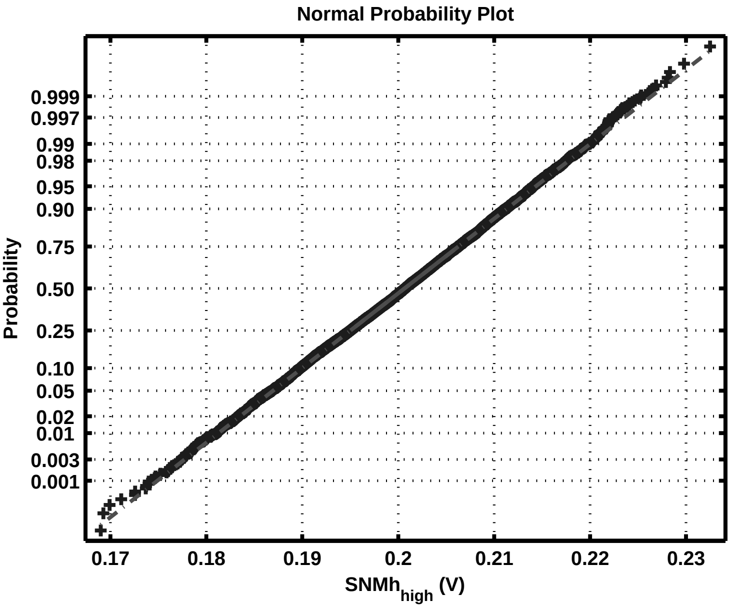


Figure 3.7: SNMh$_{\text{high}}$ distribution versus Gaussian fit

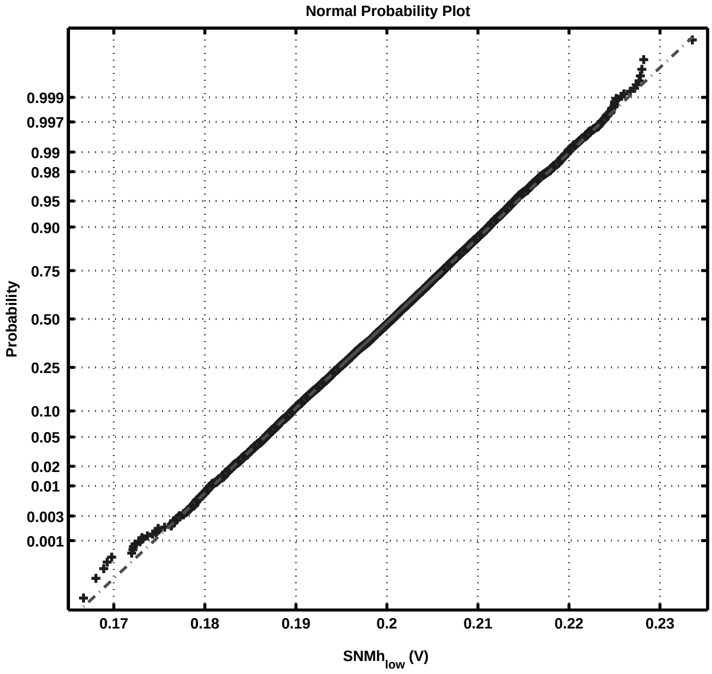


Figure 3.8: SNMh$_{\text{low}}$ distribution versus Gaussian fit

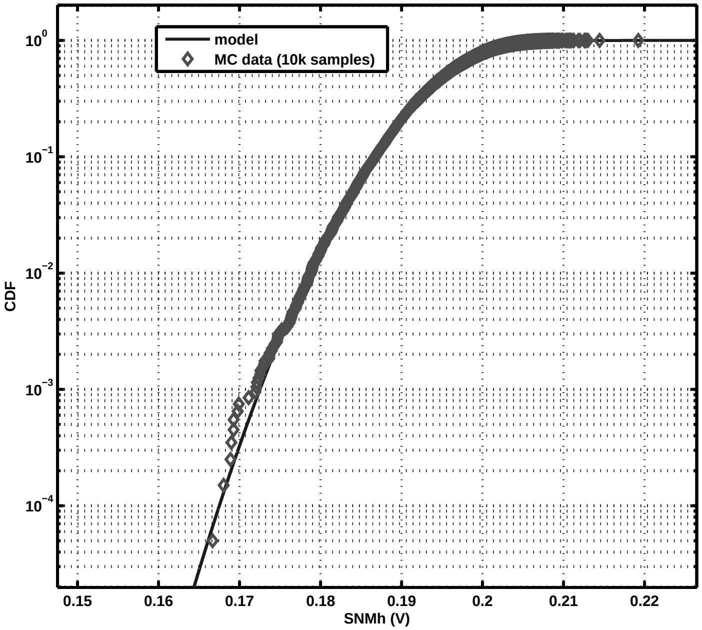


Figure 3.9: Fit of the statistical model to SNMh

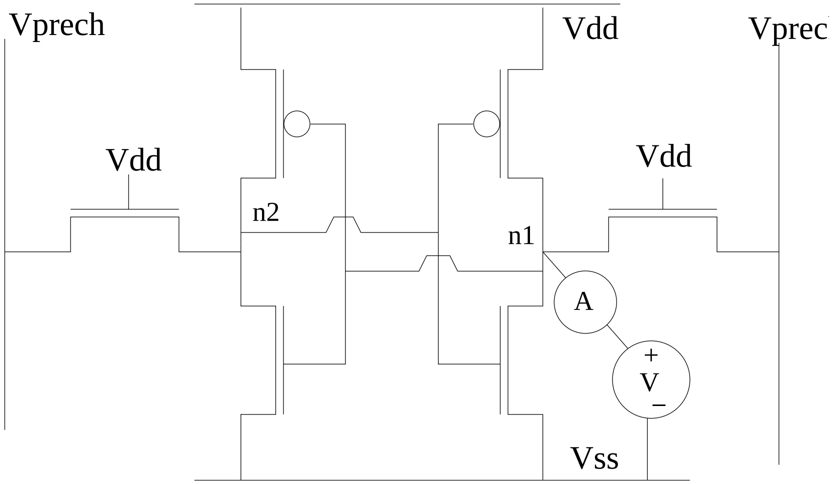


Figure 3.10: N-curve measurement setup

plotted as function of the applied voltage. This results in an N-shaped curve as can be seen in figure 3.11

3.2.3.1 Read Stability

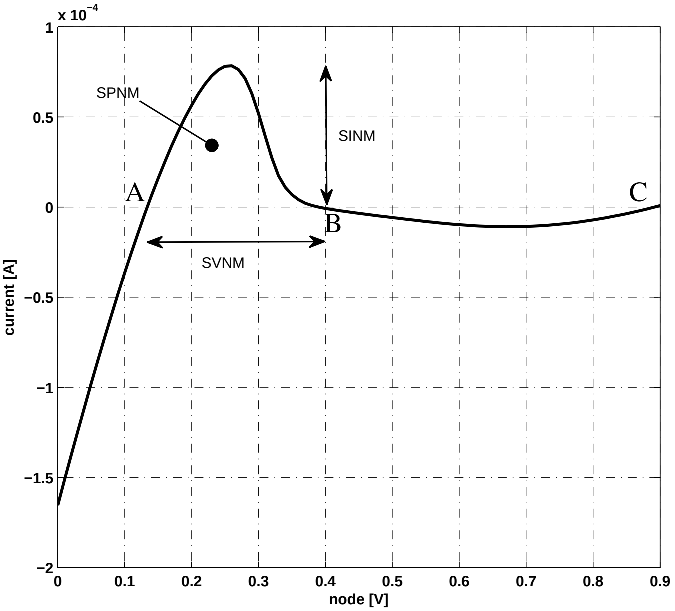


Figure 3.11: Alternative stability measures as defined on the N-curve [Wan05]

As shown on figure 3.11 three different stability measures can be defined on the N-curve. The Static Voltage Noise Margin (*SVNM*) is the voltage difference between the points A and B where the current is equal to 0. The points A and C denote the stable points of the cell, point B is the metastable point. The Static Current Noise Margin (*SINM*) is the maximum value of the current between the points A and B. It gives a figure of merit on how hard it is to create the needed voltage to make the cell flip. The Static Power Noise Margin (*SPNM*) is the integral of the current over the voltage range between points A and B. While this is the most complete metric for the SRAM stability and provides the tools to make the trade-off between SVNM and SINM, it is the hardest to measure in real time on chip.

3.2.3.2 Write-ability

The same N-curve also allows to evaluate the write-ability of the SRAM cells. The distance between the points B and C in figure 3.11 gives a measure for the voltage write margin. The current extremum between B and C gives a measure for the current write-ability. A thorough discussion on this measure is however outside the scope of this thesis but can be found in [Gro06]

3.2.3.3 Expansion to Stand-by Retention

The SVNM in the N-Curve approach would give the voltage difference between the supply and meta-stable point under hold conditions and as such is not very suited to give a measure for the retention in that case.

The current margin SINM under hold (*SINMh*) conditions provides a better measure for the retention. This is due to the exponential relationship of the leakage currents with the supply voltage. The balance of the leakage currents will define the retention capability of the SRAM cell.

The power margin under hold (*SPNMh*) combines both SINMh and SVNMh measures. It gives the most complete measure. However, due to the minimal information that comes out of the SVNMh measure, its improvement on SINMh is marginal. As SPNMh is also the most difficult to measure in real-time on chip, it is not chosen as a usuable measure for a real-time monitoring and regulation circuit.

Figure 3.12 shows the retention measures that can be defined on the N-curve in hold conditions.

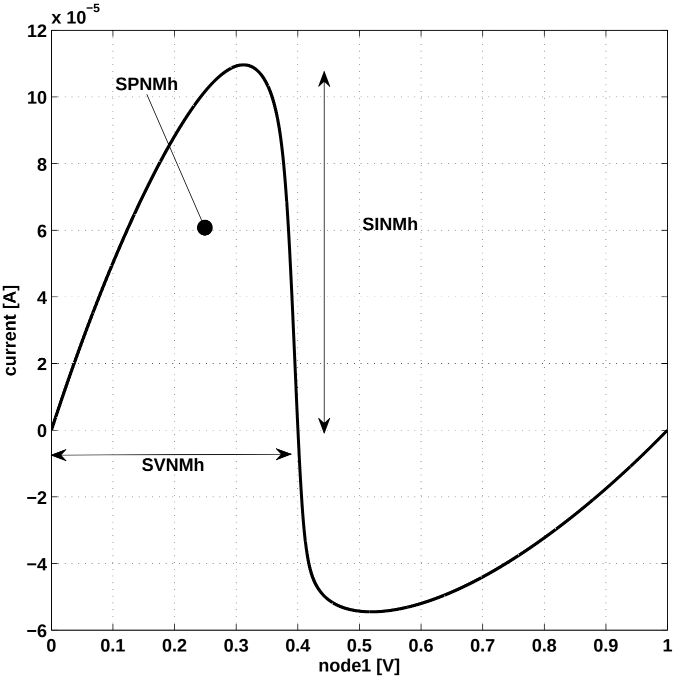


Figure 3.12: Retention measures as defined on the N-curve

The evolution of the three N-curve based measures for hold conditions as function of the supply voltage is show in figures 3.13, 3.14 and 3.15.

The SVNMh behaviour shown in figure 3.13, is as expected linear with the supply voltage. The crossing point of the butterfly curve under hold conditions is solely determined by the current balance of the pull-up pmos versus the pull-down nmos. As long as both transistors are in the saturation region, solving this current equation yields a linear dependence on the supply voltage. Once the transistors reach the subthreshold operation region the SVNmh rapidly collapses, and the cell data is easily lost.

The evolution of the SINMh measure shown in figure 3.14, follows a quadratic function as long as the core transistors operate in the saturation region. Once the transistors reach the subthreshold regime SINMh deteriorates exponentially.

SPNMh as the integral of the current over the input voltage across the range between the two leftmost combines both evolutions. The range over which to integrate reduces linearly. The current values drop quadratically. Hence the SPMNh displays a third order polynomial behaviour as long as the transistors are in saturation.

These relationships can be used to optimise the regulation algorithm. By using these relationships, the evolution of the chosen bit integrity parameter can be better predicted. This allows in turn to tune the supply voltage to the desired margin faster. This will be demonstrated in section 3.4.2.

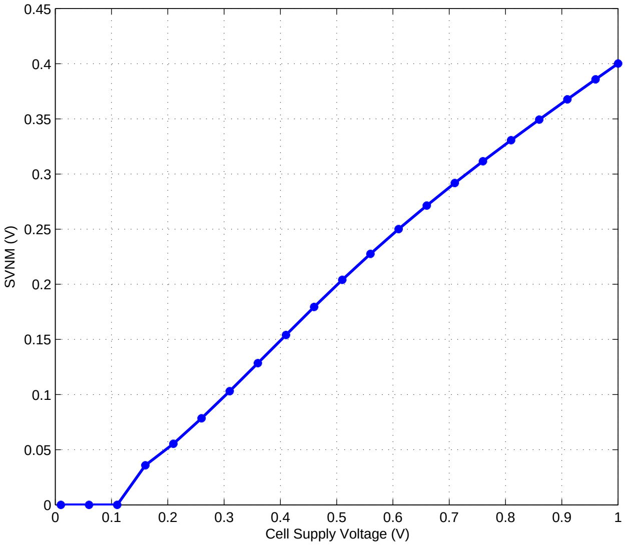


Figure 3.13: SVNMh as function of the cell supply voltage

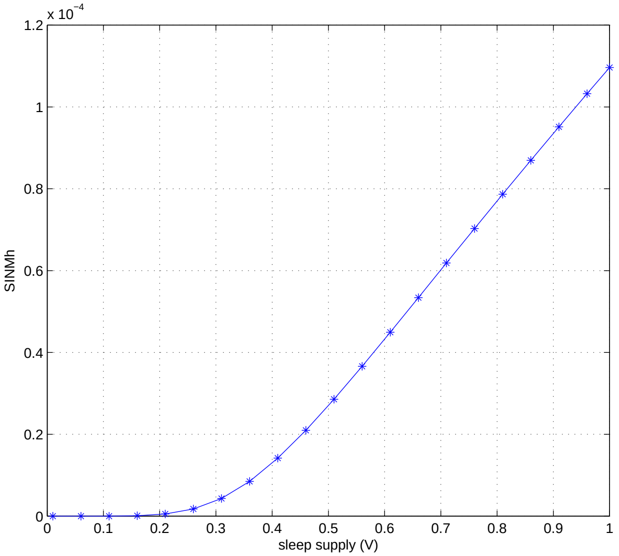


Figure 3.14: SINMh as function of the cell supply voltage

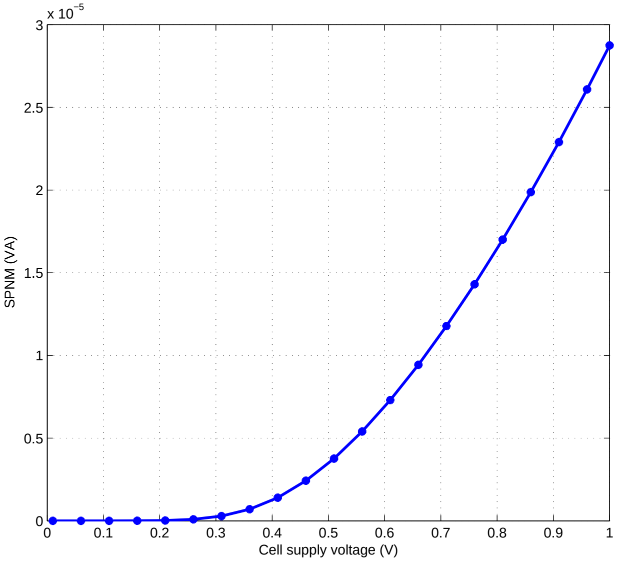


Figure 3.15: SPNMh as function of the cell supply voltage

3.2.4 Conclusion

Several measures exist to quantify the retention capability of an SRAM cell. The quantifier closest to the traditional way of measuring SRAM stability is SNMh as described in section 3.2.2.2. SNMh gives the voltage disturbance needed to flip the content of an SRAM cell under hold conditions. The probability density function for SNMh has been modelled in formula 3.4. The cumulative density function can be described by formula 3.5.

Complementary to SNMh are the measures based on the N-curve approach, as described in section 3.2.3.3, as they include the influence of the currents needed to reach the flip point of the SRAM cell. Although all three N-curve measures can be used, SINMh contains the most useful complementary measure. The additional data from SVNM, and by extension SPNM, can be derived from the combination of SNMh and SINMh.

SNMh alone already provides a consistent measure under varying supply voltages and for ease of use will be used as the main quantifier in the rest of the thesis. All measures and circuits can be extended to also use any other measure defined in the above section as methodologically the measures are near identical.

The SNMh will be used in the next section to measure and guarantee the data integrity in the cells.

3.3 Data retention solutions

3.3.1 Introduction

To guarantee retention of the data in an SRAM array, several paths have been described in open literature. They can be divided in two groups, online and offline solutions. The first path takes into account information acquired from the systems on a die to die basis and over time while the system functionality remains undisturbed. The latter path is the more traditionally followed [Wan05]. Here design time analysis is done to find the worst case scenario and its results are incorporated into design margins or costly test time die to die calibration is done.

This section will give an overview of the currently used practices in SRAM design to guarantee data retention under lower supplies. The section will start with an analysis of the off-line design solutions. To allow for solutions that need feedback over time a monitoring solution is needed. That will be the subject of the next part. Finally the monitoring solution will be linked to the supply regulation and the conclusions will be drawn.

3.3.2 Offline Design Solutions

3.3.2.1 Worst case design

In function of the transistor parameters the minimum retention voltage can be calculated as the theoretical lower boundary. By solving the current equilibrium equations in the two internal nodes of an SRAM cell, the data retention voltage (*DRV*) can be derived. In case the drain-source leakage currents are dominant and the gate leakage currents are negligible, the currents can be expressed as 3.6

$$I_i = \frac{W_i}{L_i} I_0 \cdot exp\left(\frac{V_{\mathrm{GS}} - V_T}{n_i kT/q}\right) \cdot \left(1 - e^{\left(\frac{-V_{\mathrm{DS}}}{kT/q}\right)}\right) \tag{3.6}$$

where

W_i	: Width of transistor i
L_i	: Length of transistor i
V_{GS}	: Gate Source voltage
V_{DS}	: Drain Source voltage
V_T	: thresholdvoltage of transistor i including the bulk effect and DIBL
I_0	: process specific current for transistor $W/L = 1$ and $V_{\mathrm{GS}} = V_{th}$
n_i	: subthreshold factor
k	: Bolzmann constant
T	: temperature in K
q	: electron charge

For a traditional six transistor SRAM cell as shown in figure 3.16 the current equations resolve to formulae 3.7 and 3.8:

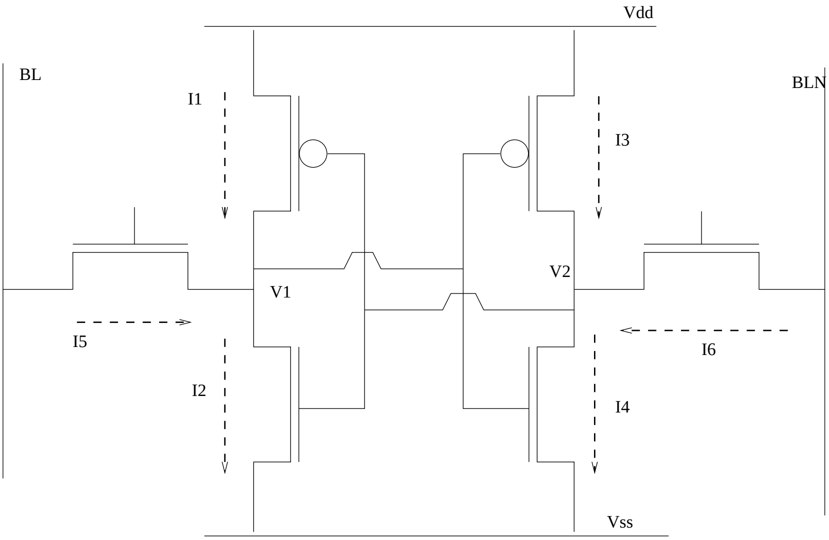


Figure 3.16: six transistor SRAM cell as used in the DRV calculation with the currents illustrated [Qin04]

$$\text{node}V_1 : I_1 + I_5 = I_2 \tag{3.7}$$

$$\text{node}V_2 : I_3 + I_6 = I_4 \tag{3.8}$$

The condition for DRV is the supply voltage is at such a level that the SNMh becomes 0. This happens when the two inverter curves touch and can be written as equation 3.9

$$\frac{df(x)}{dx} = \frac{dg(x)}{dx} \tag{3.9}$$

with f and g the function describing the DC transfer curves of the core inverters. The mathematical derivation can be found in section 3.4.2.

By substituting equation 3.6 into the equations 3.7, 3.8 and solving the system together with equation 3.9, the DRV can be found as an iterative solution. It can be expressed as formula 3.10 [Qin04]

$$DRV = DRV_1 + \left[\frac{V_1}{2} + \frac{(DRV_1 - V_2) \cdot n_2}{2}\right] \tag{3.10}$$

with

$$DRV_1 = \frac{kT/q}{n_2^{-1} + n_3^{-1}} \cdot \ln\left[\left(n_3^{-1} + n_4^{-1}\right)\frac{A_4}{A_2A_3}\left(\frac{A_5}{n_2} + \frac{A_1}{\left(n_1^{-1} + n_2^{-1}\right)^{-1}}\right)\right] \tag{3.11}$$

$$A_i = \frac{W_i}{L_i} \cdot I_0 \cdot exp\left(\frac{-V_\mathrm{T}}{n_i kT/q}\right)$$

$$V_1 = \frac{kT}{q} \cdot \frac{A_1 + A_5}{A_2} \cdot \exp\left(\frac{-DRV_1}{n_2 kT/q}\right)$$

$$V_2 = DRV_1 - \frac{kT}{q} \cdot \frac{A_4}{A_3} \cdot \exp\left(\frac{-DRV_1}{n_3 kT/q}\right)$$

where

DRV	: minimum data retention voltage
W_i	: width of transistor i
L_i	: length of transistor i
V_T	: threshold voltage of transistor i
n_i	: subtreshold slope
I_0	: technology constant
kT/q	: thermal voltage (25mv@300K)

The starting point DRV$_1$ can be obtained from the starting approximation where $V_1 = 0$ and $V_2 = Vdd$.

Formula 3.10 provides the minimal DRV in a given set of environmental parameters. To have a working memory with enough yield the parameters used should be those from the worst case corner in terms of process and temperature. Extensive simulation will need to be used to cover the effects of variability. From these stochastic simulations a minimal standby voltage can be derived. [Qin04]

While the worst case approach definitively finds a safe voltage for the stand-by voltage it also has a large margin built in to allow all dies to function with a predetermined minimal retention noise margin. This margin will in many cases be a “waste” of energy as a great deal of dies will not be in the state of statistical extremities. With a 6 σ approach only 1 in 10^9 will not be covered.

3.3.2.2 At Test Time

By lowering the sleep voltage on those dies where it is deemed possible, it is possible to create more optimal savings. To this end die to die calibration can be used. As all current commercial memories include a Build-in Self-Test (*BIST*) module, this calibration can be done at test time. By putting the BIST into a loop where the sleep voltage is also controlled the minimal stand-by supply can be found on a die to die basis. However this approach takes up valuable test time.

The “calibration at test time”-approach has the benefit of optimising the leakage current reduction further as the minimal sleep voltage can be found for each die independently. However it does not compensate for any time varying properties of the dies such as temperature or voltage variations unless a margin is build in. This margin again is the source of this solution being suboptimal. Although slow degrading effects, like ageing, can be compensated by rerunning the BIST loop over the lifetime of the system.

3.3.2.3 Conclusion

Both the design time and test time approaches can guarantee the retention of the data in an SRAM cell with a high probability. They are however also the source of a sub-optimal supply reduction and therefore sub-optimal power reduction due to the margins that have to be taken into account to be able to compensate for time-varying die parameters. In the design time case additional margins have to be taken to compensate for global process variations. The results published in [Ham08] lead to the same conclusion. To reduce the leakage as for as possible, active online regulation is needed.

3.3.3 The Need for a Monitor

3.3.3.1 Introduction

As pointed out in the conclusion of the previous section not determining the minimum stand-by voltage at run time results in sub-optimal savings. Different environmental parameters such as temperature, voltage and process variations, require different supply voltages in order to preserve the data while maximising power savings. As SNMh reduces proportionally with the supply voltage of the cells, further optimising the power savings requires feedback from the system over time. Where in section 3.3.2.1 the minimum sleep voltage was largely based on theoretical knowledge of the process, here the value will be based on actual feedback from the system. In other words a monitor of some kind has to be found and implemented.

A monitor that has the same behaviour as the core SRAM cells with regard to PVT variations will allow the maximum reduction of leakage current while keeping the stored data intact. To this end the monitor circuit needs to be in a similar, if not identical, environment as the core cells. It has to behave the same under nominal conditions and react to changes in environment as the core cells would.

This section will discuss the boundaries in which this monitor must function. These boundaries will be defined in terms of yield for the entire memory matrix. The yield will be analysed when a monitor is present in the system to allow tuning of the sleep voltage.

3.3.3.2 Yield analysis

The basic defining feature of an SRAM is the retention of its stored data. Yield in this context will be defined as the chance a full memory matrix can retain the data with a sufficient bit integrity margin. The bit integrity parameter of choice here will be SNMh as defined earlier in 3.2.2.2. Similar calculations can be done with other bit integrity parameters.

The yield of the entire matrix can be linked to the SNMh probability distribution of an individual cell. Yield is in this case the probability all cells in the matrix satisfy the SNMh constraints as illustrated in formula 3.12.

$$P_{\text{matrixWorks}} = P\left[\forall i \in \text{matrix} : \text{SNMh}_i > \text{SNMh}_{\text{min}}\right] \tag{3.12}$$

Where i represents an instance of a cell, SNMh$_i$ is the actual SNMh of cell i and SNMh$_{min}$ is the minimum SNMh to be respected.

Under the reasonable assumption that the probability of cell SNMh is independent for each cell, the right hand part of formula 3.12 can be rewritten as formula 3.13. Under the further assumption that the probability distribution for each cell is also identical, these can be further simplified to formula 3.14.

$$\begin{aligned} P[\forall i \in \text{matrix} : \text{SNMh}_i > \text{SNMh}_{\text{min}}] =& P[\text{SNMh}_1 > \text{SNMh}_{min}] \\ & \cdot P[\text{SNMh}_2 > \text{SNMh}_{min}] \cdot \ldots \\ & \cdot P[\text{SNMh}_N > \text{SNMh}_{min}] \qquad (3.13) \\ =& \left(1 - CDF\left(\text{SNMh}_{min}\right)\right)^N \qquad (3.14) \end{aligned}$$

Where N is the number of cells in the memory matrix and CDF(x) is the cummulative density function of the SNMh probability distribution of an individual cell. As earlier illustrated in section 3.2.2.2, the probability distribution of SNMh for a single cell can be modelled by the formulas 3.4 and 3.5.

Using these approximation the yield can be written as function of the error function *erf*. This leads to formula 3.15.

$$Yield = \left[\frac{1}{4}\left(1 - \text{erf}\left(\frac{\text{SNMh}_{\text{min}} - \mu_l}{\sqrt{2\sigma_l}}\right) - \text{erf}\left(\frac{\text{SNMh}_{\text{min}} - \mu_h}{\sqrt{2\sigma_h}}\right) + \text{erf}\left(\frac{\text{SNMh}_{\text{min}} - \mu_l}{\sqrt{2\sigma_l}}\right) \cdot \text{erf}\left(\frac{\text{SNMh}_{\text{min}} - \mu_h}{\sqrt{2\sigma_h}}\right)\right)\right]^N \quad (3.15)$$

where

SNMh_{min} : minimal required SNMh
μ_l : mean of the single cell SNMh_{low} distribution
μ_h : mean of the single cell $\text{SNMh}_{\text{high}}$ distribution
σ_l : standard deviation of the single cell SNMh_{low} distribution
σ_h : standard deviation of the single cell $\text{SNMh}_{\text{high}}$ distribution
N : number of cells in the matrix

In the absence of tuning, the SNMh distribution taken into account must be the worst case environmental corner to achieve the desired yield under varying supply voltages. This leads to the creation of extra margins not needed in the majority of the cases and as such to less power reduction.

With tuning it is possible to adjust the supply reduction in such a way that the power reductions can be maximised on a die-to-die basis while retaining the data with enough yield. This tuning should adjust the supply voltage such that the targeted yield is reached while reducing the supply voltage as far as possible.

The basic principle consists of a monitor that can estimate the mean of the SNMh distribution of a single cell. This value can than be used to base the supply adjustments on so that the monitor value reaches a predefined reference value. If the monitor represents the cells correctly the entire matrix will have an identical shift, guaranteeing the retention in all cells.

However, the monitor is not perfect due to variability. An additional margin has to be taken into account for the externally applied reference value to compensate for these imperfections.

In figure 3.17 the cumulative density function of the retention failure rate is plotted. Using the 6σ approach of 1 in 10^9 cells with less than the required retention noise margin, the minimum retention voltage level can be obtained under mismatch through the use of Monte-Carlo based simulations. For the commercial 90nm technology used in figure 3.17 and the cell design of section 4.2.1, this value would be 200 mV. As it is impossible for a monitor to equalise between individual cells, a monitor has to quantify the influence of global process and environmental parameters with sufficient accuracy. This means the mismatch on the monitor circuit itself should be low enough not to be of any significant influence on the supply regulation system. Failing to do so would deteriorate the savings possible as extra margins would have to be taken to compensate for the uncertainty of the monitor. The monitor would have to be considered to represent the best cell in the matrix while it could actually be worse than the worst cell in the matrix, its true position unknown as illustrated in figure 3.18.

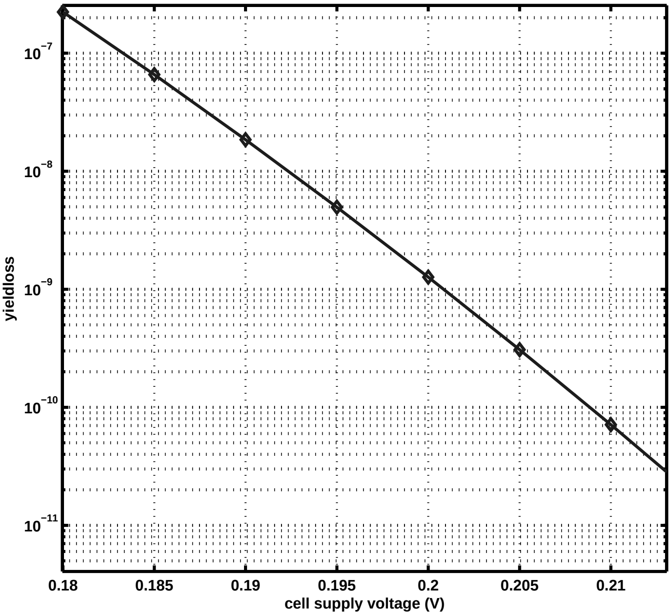


Figure 3.17: Yield analysis of SNMh for an SRAM cell in 90nm

The effective yield of the memory is determined by the chance to have a cell in the matrix under the required bit integrity parameter given a measure from the monitor under a certain supply. Formula 3.12 can be adapted to reflect this into formula 3.16.

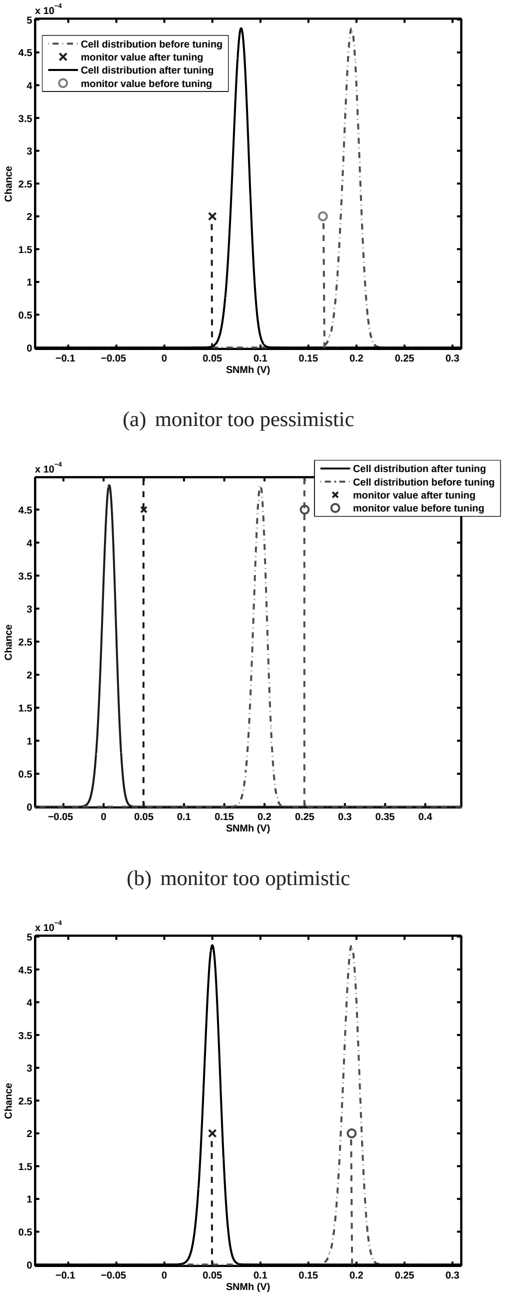
$$\begin{aligned} P_{\text{matrixWorks}} &= P[\forall i \in \text{matrix} : \text{SNMh}_i > \text{SNMh}_{\text{min}} | \text{SNMh}_{\text{mon}}] \\ &= P[\forall i \in \text{matrix} : \text{SNMh}_i > \text{SNMh}_{\text{min}}] \cdot P[\text{SNMh}_{\text{mon}}] \end{aligned} \tag{3.16}$$

With the same assumptions as used for formula 3.14, formula 3.16 can be written in the form of formula 3.17.

$$P_{\text{matrixWorks}} = \int_{-\infty}^{\infty} \left(1 - CDF\left(\text{SNMh}_{min}\right)\right)^N \cdot P_{\text{mon}}(x) dx \tag{3.17}$$

3.3.3.3 The power trade-off

Having margins on the supply voltage above the theoretical minimum DRV, see also equation 3.10, reduces possible power savings. The following paragraphs will be dedicated to the analysis of power versus margins.



(a) monitor too pessimistic

(b) monitor too optimistic

(c) monitor correct

Figure 3.18: The influence of the accuracy of the monitor illustrated. In (a) the monitor estimation will cost power. In (b) the tuned cells will cause yield problems. In (c) the monitor give the correct estimate and leads to a maximal reduction with guaranteed retention

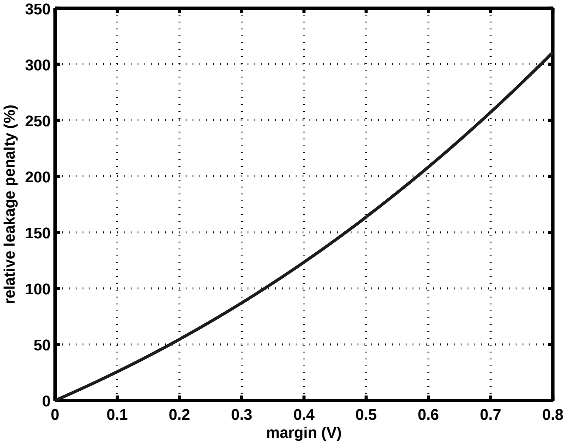


Figure 3.19: Relative power penalty due to margins compared to the DRV

In figure 3.19 the power penalty in function of the the margin size compared to the DRV is plotted. As both subthreshold and gate leakage currents are exponentially dependent on the supply voltage, see also section 2.2, so is the penalty for power.

Reducing the margins however requires the use of extra peripheral circuitry, and as such extra power. The optimum solution will then be the one that reduces the power more, by reducing the margin, than it consumes. Monitored solutions will allow for a smaller margin to be taken into account, as they provide environmental information on a die-to-die basis. However, a close eye should be kept on their own power requirements.

3.3.3.4 Conclusion

To allow for the best possible leakage power savings, the margins needed to guarantee the data retention should be minimised. To do so feedback concerning the PVT environment in which the SRAM operates is needed on a die-to-die basis. This can be accomplished by using a monitor system that observes the bit integrity and is able to adjust the voltage supply.

The influence of the monitor characteristics on the yield has been derived in formula 3.15. To achieve a high yield the monitor has to be reliable. Foremost this means the monitor has to mimic the cells as close as possible in behaviour, both in the nominal case and under influence of environmental changes.

By implementing a monitor with a sufficiently low spread and representative for the mean of the cell SNMh distribution, a regulation system can by implemented with a minimal margin overhead.

3.3.4 Online Monitored Solutions

3.3.4.1 Introduction

As illustrated in the previous section a monitor solution that compensates for time dependent variations is necessary. This monitoring solution also has to provide a good estimate for the behaviour of the bit integrity parameter under such variations and the voltage range needed to reduce leakage currents.

This section will give a short overview of the currently used monitor systems that can compensate for time dependent variations and will describe the solution proposed in this thesis.

3.3.4.2 The Canary Solution

Chandrakasan proposed in [Cal04] a solution to the similar problem in logic data path design. In systems where multiple supplies are used to scale down power, the data in the data paths has to be maintained. To that end [Cal04] proposed using banks of skewed flip-flops to predict failure of the flip-flop cells under varying supply voltage regimes.

This approach has been adapted to do the same for memories. In [Wan07a] a seperate bank of skewed memories cells is put next to the cell matrix.

The monitor cells have been adapted in layout and circuit design to fail at voltages higher than the core cells. Detection of failure is done by observing the stored data in a cell. Once the stored data flips, a cell is said to have failed. By having different cells fail at different voltages, a continuous spectrum of failure voltages can be obtained. This allows to have a die based trade-off between reliable storage and power reduction.

To reduce the spread of the DRV on the monitor cells, the constituting transistors are scaled up to profit from Pelgromm's Law [Pel89]. To further reduce the influence of rogue cells, outliers are screened out based on a majority vote.

Furthermore [Wan07a] mentions three possible means to determine the external threshold for the canary cell bank. The most obvious is the use of Monte-Carlo simulations to determine the failure rate for every possible supply. For memories where more than 6σ of margin is needed to have a yielding system, this can become time consuming. The second way to calibrate the cell banks is by doing so a test time with the help of the BIST modules, an approach already mentioned in 3.3.2.2. The final mentioned way is to base the external threshold value on the statistical model proposed in [Cal06] for the DRV.

However deliberately skewing SRAM cells also means the layout of the monitor cells and core cells will be different by design. As a consequence the influence of variations will be different between the core cells and the skewed monitor cell. Also as the cells are placed in banks, next to the matrix and each other, the environment will be different for the environment of the core cells. Again this results in a different behaviour towards process and temperature variations. These factors force extra margins to be taken into account to compensate possible misrepresentations, even if they are not explicit. These margins will limit further possible power reductions.

3.3.4.3 Proposed Solution

To have the best possible indicator of the actual behaviour the monitor has to mimic the behaviour of the SRAM cells as close as possible. This will decrease the margins that have to be taken to compensate for deviations in behaviour. To this end the monitor cells need the same electrical and geometrical properties as the core cells. The environment of the cells should also be identical, or close as possible to the environment of the core cells.

From the previous section 3.3.3.2, it is also clear that the spread on the monitor value has a direct influence on the efficiency of the power savings. A high spread will result in increased margins on the secondary sleep supply as it will give a less accurate estimate for the mean of the bit integrity parameter compared to a monitor with a lower spread.

The monitor consists of a single SRAM cell, or multiple SRAM cells in parallel, where the internal nodes are accessible. As such the conditions of similar, if not identical, behaviour to the core cells can be guaranteed. In accordance with the law of large numbers [Ber51] the mismatch influence of the transistors on the bit integrity parameter can be reduced with the square-root of the number of monitor cells, see also formula 3.18.

$$\sigma \sim 1/sqrt(N) \tag{3.18}$$

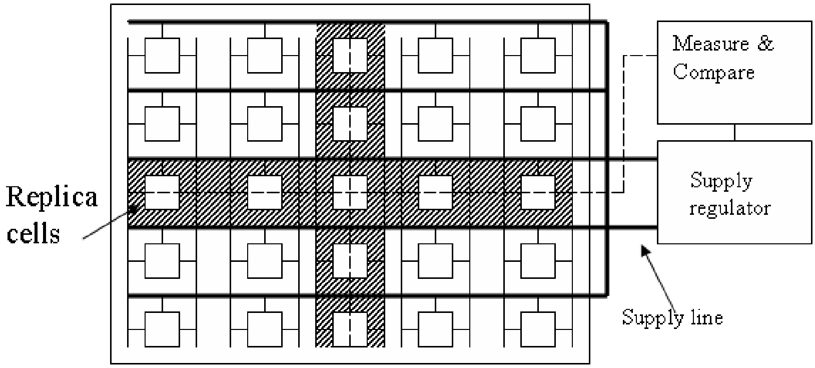


Figure 3.20: monitor organisation

Figure 3.20 gives a schematic overview of the proposed monitor system. The monitor cells are placed into the memory matrix itself to allow for the closest matching possible to the core cell environment. The access wires to the internal nodes of the cells should not interfere with the patterns used by the core cells for the word and bitlines. As such they will be put on higher metal layers. The measurement and supply regulation system can be place outside of the immediate matrix area as to not interfere with the matrix.

Another advantage of this approach is that the spread of the placement of the monitor cells will eliminate the linear dependency of location on the mismatch even further reducing the spread σ of the monitor SNM distribution.

The increased internal capacitance of the cells is not a hampering factor if the bit integrity parameter measurements are done in a static way. Both SNM and the N-Curve approaches are static bit integrity parameters and are suitable for this approach.

Due to the minimax criterion nature of SNMh, the expected value μ of the monitor will shift to a higher value if the spread σ is decreased. However, From the yield figures such as figure 3.21, it is clear this is not a hampering factor when high yields are required.

Reducing the spread on the monitor parameters also comes at a cost of power. By putting N cells in parallel the current needed to drive the internal nodes to a pre-set voltage will require N times as much current compared to a single cell. The current needed can be derived from the N-curve. In the worst case this will be the maximum current of the cell. This active power usage can be partly mitigated by a low activation rate. The needed current will also diminish with lower supply voltages, exactly those voltages where the measurement becomes critical to the yield of the system.

Nevertheless, a trade off between measurement power and leakage power reduction has to be made with yield serving as arbiter. This trade-off depends heavily on technology parameters. A high DIBL coefficient will favour a higher number of monitor cells as it will allow smaller margins and the exponential gain in power savings that entails. A process with low mismatch parameters, will favour a low number of monitor cells.

From figure 3.21 the benefit of more cells in parallel can be clearly seen. For the same yield the higher number of monitor cells compensate better for the variability. The safety margin, which needs taken to attain the required minimum SNMh after calibration, will also reduce with the increase in monitor cells to guarantee the same yield. This advantage grows further in importance with the requirement for high yields as the margin gap with a single monitorcell increases. In case the monitor would be perfect, the safety margin would be identical to the safety margin needed from the cell distribution alone.

Based on this proposed parallelised monitor cell approach the following section 3.4 will focus on the basis and implementation of the measurement algorithms.

3.3.5 Conclusion

The offline methods of worst case design, or test time calibration, can guarantee the data retention of the cells under a reduced supply voltage. However to do so requires an overhead on power as margins need to be taken to compensate for inter-die variability or time dependent environmental variations. In case of worst case design these margin will be the largest as it has to compensate for all variability mechanism by taking margins. For the test time calibration, die-to-die calibration is possible and as such will have reduced power penalties compared to worst case design. Time dependent variations, such as voltage shifts or temperature can not be compensated without taking margins. Slow effects over time, such as degrading, could be compensated by rerunning

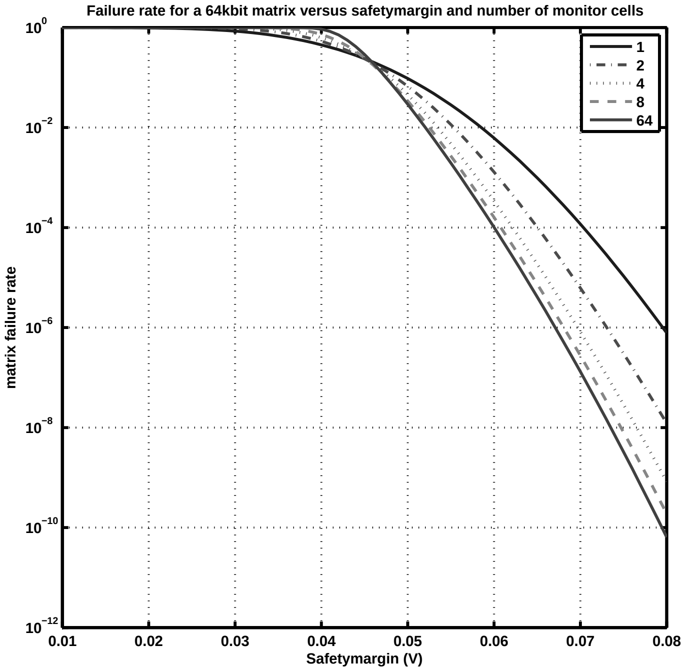


Figure 3.21: yieldloss in function of the number monitor cells and the applied sleep voltage

the BIST routines on regular intervals.

The online solutions can provide both a compensation for the inter-die variability and the time dependent variability. The canary solution however has the inherent attribute of having skewed cells. By definition this means those cells will not react identical to the core cells under environmental or process variations. As such they will need to take into account extra measures to compensate for these inaccuracies. The solution proposed in this thesis reduces those margins even further. By inserting monitor cells that are by design as close as possible to the actual core cells, a monitor system is created. This monitoring system will react in the same way as the core cells to variations and as such give the most accurate measure for the actual condition of the core cells. By using several monitor cells in parallel, the spread on the measured value can be drastically lowered. Combined with the model for the SNMh distribution, this allows the finest accuracy in determining the retention voltage in real-life conditions.

3.4 Closing the loop

3.4.1 Introduction

With access to a reliable monitor as described above in 3.3.4.3 the question remains how to interact with the rest of the circuit and have a reliable bit integrity parameter. This section will deal with the algorithm needed and its implementation to measure SNMh on the monitor. Both an analog and digital implementation will be proposed.

3.4.2 The Algorithm

The mathematical basis for the algorithm can be found in Seevinck's definition of SNM [See87] . The SNM is the minimum of maximum of the eye-openings in a butterfly plot.

To calculate this SNM the two inverter curves can be rotated over 45 degrees counterclockwise and subtracted. The extrema of this difference function are the values of which the absolute value has to be compared to know the SNM as shown in figure 3.22. The same approach can be used for the calculation of SNMh, where the access transistors are turned off.

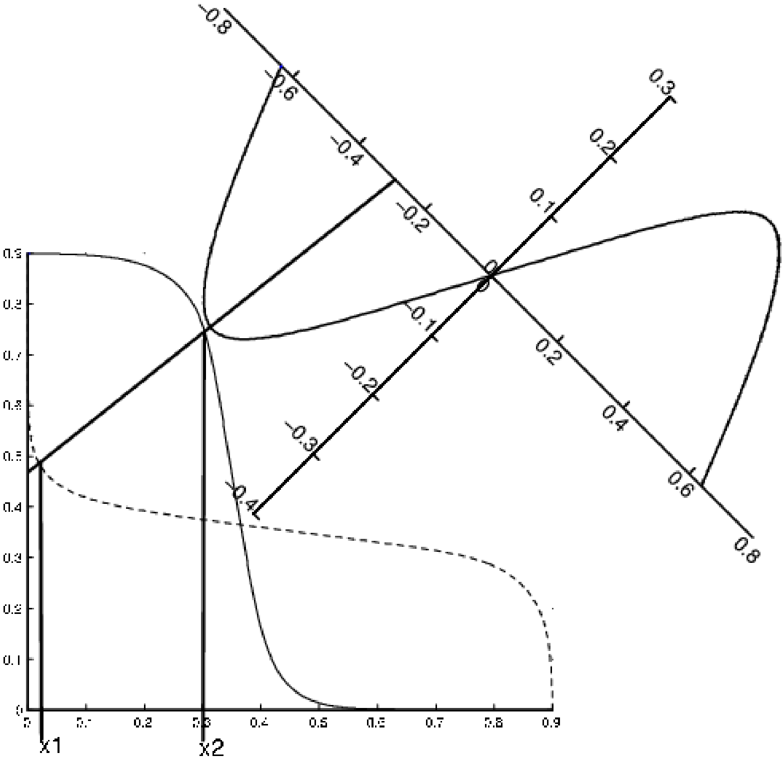


Figure 3.22: Graphical representation of the SNM calculation

If abstraction is made of the inverter curves as being the functions $f(V)$ and $g(V)$, the calculation yields the condition that needs to be satisfied to be in one of the extrema. This can then be used to implement an algorithm that measures the curves until the condition is satisfied and the comparison can take place.

Consider $f_{45}(x)$ and $g_{45}(x)$ to represent the 45 degrees counterclockwise rotated functions of $f(V)$ and $g(V)$ respectively. The difference function $h(x)$ 3.19 then becomes the function of which the extrema have to be found.

$$h(x) = f_{45}(x) - g_{45}(x) \tag{3.19}$$

The condition for extrema consist of having a derivative in the observed point equal to 0. This condition can be directly transfered to the condition that the difference of the derivatives of $f_{45}(x)$ and $g_{45}(x)$ are equal.

$$\frac{dh(x)}{dx} = 0 \tag{3.20}$$

$$\Downarrow$$

$$\frac{df_{45}(x)}{dx} - \frac{dg_{45}(x)}{dx} = 0 \tag{3.21}$$

$$\Downarrow$$

$$\frac{df_{45}(x)}{dx} = \frac{dg_{45}(x)}{dx} \tag{3.22}$$

This formula 3.22 remains valid under rotation if it is taken into account that the evaluation points have to be transformed in the same way. For this situation that implies the extra condition that the measurement points should comply with $f(V_1)$ and $g(V_2)$ to be on the same 45 degree line. In other words the difference between $f(V_1)$ and $g(V_2)$ should be equal to the difference between V_1 and V_2. This is also illustrated in figure 3.22. In practice the curves that can be measured are f and g^{-1}. So the measurement condition can be rewritten as formula 3.23

$$g^{-1}(f(x) - \Delta) = x - \Delta \tag{3.23}$$

The Δs that fulfil this condition are the measures to be evaluated as SNM candidates as they describe the SNM function in the 45 degree rotated space. It is reasonable to assume that there will be one extremum where x will be smaller than half the supply voltage and one where x will be bigger.

The measured values will than have to be compared with each other and an external applied minimum reference to regulate the secondary supply in t he system. Implementations of the measurement algorithm will have to satisfy the above conditions. Both an analog and digital implementation are presented in the following sections.

3.4.3 Analog Implementation

An analog implementation of the measurement algorithm will be based on the principle that the operating point of the measurement system can be maintained in the SNMh-extremum points. This will lead to a loop-within-loop system in which guaranteeing stability is not trivial.

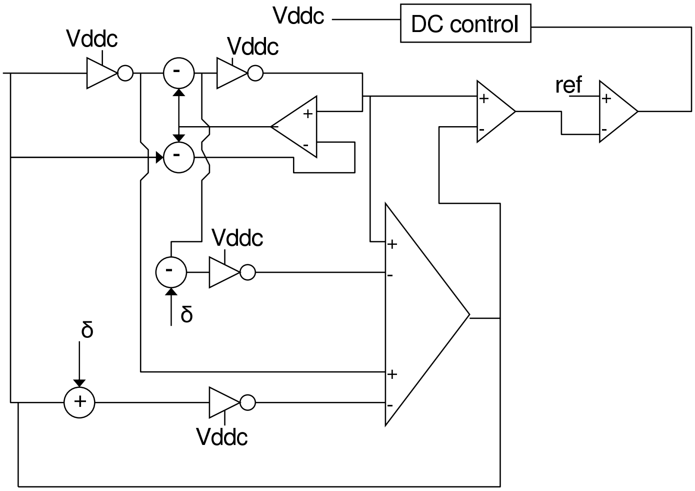


Figure 3.23: Analog example implementation

An analog implementation of such a system is depicted in figure 3.23. There are 5 main functions present in it. The inverters in the circuit repres ent half of an SRAM monitor cell.

The first function present in the schematic is the loop to find the second crossing point with the butterfly curve with the 45 degree line through x. The feedback system will force the inputs of the operation amplifier to be equal in its operating condition.

The derivatives of these two points must then be calculated using the fixed small offset δ. This is the second function in the schematic.

The third functional loop has as operating point the point where the evaluated derivatives are equal. This can be accomplished by using a differential difference amplifier (*DDA*). The DDA will feed its output back to the first function and forward to the SNMh evaluation function.

This last function evaluates the output from the DDA to the output of the first function to calculate the SNMh and compares it to the externally applied reference. The output of this last function can then be used to bias the control for the supply voltages of the SRAM monitor and data cells.

A full system that would use this schematic will need two instances. One instance will have to be biased to find the solution above half the actual supply voltage, the other to find the solution under half the actual supply voltage.

The schematic in figure 3.23 has four instantiations of the monitor circuit. The mismatch between these instantiations has a direct impact on the accuracy and reliability of the control and regulation circuit. The operational amplifiers also will have stringent requirements on offsets to implement the mathematical functions correctly. The same is also true for the wanted offsets to calculate the derivatives.

The conditions and requirements on the components forming such a system can be reduced by using a digital system.

3.4.4 Digital Implementation

3.4.4.1 Introduction

The same algorithm can also be implemented in a digital way, either by hardware or by software running instruction on a neighbouring processor. To reduce area overhead and given SRAM memories are seldom stand-alone applications, this software could run on the spare cycles of the processor. However, to accomplish this an interface is needed between the essentially analog monitor cells and the digital control and algorithmic implementation in the form of digital-to-analog and analog-to-digital converters. This section will deal with the implementation of the measurement algorithms and its possible optimisation first. Secondly the specifications for the needed interface converters will be derived.

3.4.4.2 Optimisations

Where the analog implementation will converge to the stable point in a time based on the time-constants of the intertwined loops, the digital implementation has possibilities to speed up that process without endangering the stability of the system.

The first and obvious optimisation that can be made, is not waiting until the actual SNMh has been calculated, but already reduce the supply voltage if the measured value exceeds the externally applied reference. Several refinements can be taken into account as the SNMh has a linear dependency on the supply voltage as discussed in section 3.2.2.2.

The second optimisation that is possible due to the memory capability of a digital implementation, is the possibility to bypass the implementation of the equal derivative condition 3.23. If the system is limited to evaluating points that satisfy the condition of being on the intersection of the two DC-transfer characteristics and a 45 degree line, the location where the distance between the projected x-values is maximal will also be the location for equal derivatives. Doing this is only possible if the values can be compared to values from a previous measurement. This also allows the algorithm to be implemented as a gradient steered search for that maximum.

The third and final optimisation that is made is the use of an adaptive step size to search for SNMh points. This requires the possibility to backtrack and reduce the step size in case the extremum is missed. This is only possible due to the second optimisation where the gradient is determined and the form of the SNMh function.

3.4.4.3 Flow chart

Figure 3.24 represents the flowchart of the digitally implemented algorithm. It can be divided into two main blocks. The first block will search for the points on both curves that fulfil the 45-degree line constraint. These are the steps 1 through 9. The first step will initialise the variables. The x-value is the voltage to be applied first to the monitor. It will get a small increment inc. The resulting value from applying the voltage to the monitor is stored in y in step 4. A small δ is subtracted from both x and y in step 5. $y - \delta$ is the applied to the other side of the monitor cell and stored into x'. In case x' is not equal to $x - \delta$, δ will be increased and the algorithms returns to step 5. If x' equals $x - \delta$ we have found two couples (x, y) and $(x - \delta, y - \delta)$ that belong to the respective curves on a 45-degree line. The value of the possible SNMh candidate is then the value of δ, and is stored for comparison with future candidates. This implements condition 3.23

To steer this search for candidates to find the actual SNMh within certain limits, a binary search algorithm is implemented in the section to find the next x. Due to the nature of the SNMh curve, also illustrated in figure 3.22, obtaining a lower value for the SNMh candidate indicates the actual SNMh has been passed. In this case the increment inc for x is halved and reversed to backtrack.

Steps 17 through 23 illustrated the decision making process for the supply voltage. If the measured SNMh is smaller than the external reference the supply should be increased. If it is higher the supply voltage can be decreased. The way in which to adjust the supply voltage can be optimised by using the relationship formulated in equation 3.1.

The clock-cycle accurate implementation of the algorithm in MATLAB can be found in appendix B.

3.4.4.4 Requirements towards peripheral circuitry

Regardless of the implementation used the measurement system has to be able to deliver enough current to the monitor cells as they consist out of back-to-back inverters. To be able to a apply a voltage on one side, the measurement system will have to be able to deliver current proportional to the current from the opposite half of the monitor cell. This “obstructing” current is the same as the current measured in the N-curve stability criterion [Wan05].

To accomplish this a unity-gain amplifier will have to be inserted between the monitor cells and the measurement system. For the analog implementation the applied voltage is directly available in the measurement system. For the digital implementation a digital-to-analog converter is needed. This D/A-converter can be seen as an extra overhead so its energy consumption should kept as low as possible. As all measurements are done in a quasi-static way, there is no high speed requirement for the converters, which aids the power requirement.

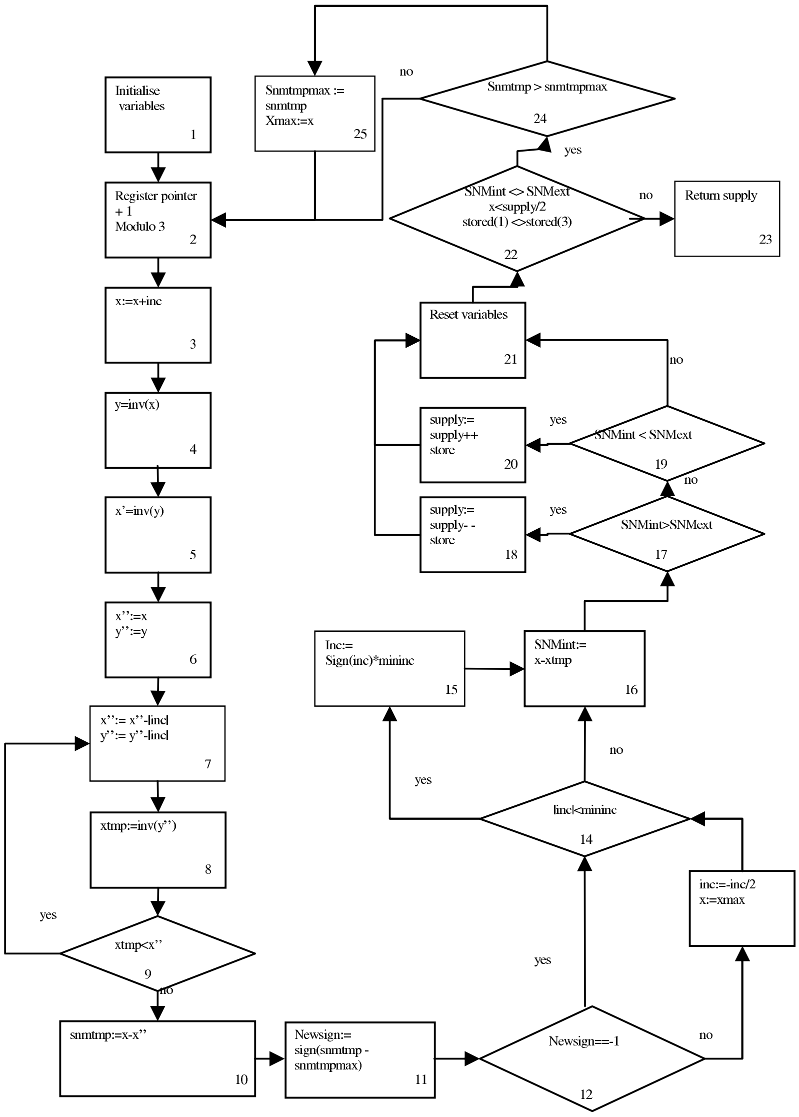


Figure 3.24: Digital implementation flowchart

3.4.5 Summary

Two possible ways to implement the measurement and feedback system have been presented in the previous sections. The first implementation proposal is based on fully analog measurement and feedback. The second implementation is digital in nature.

The analog implementation has to rely on separate monitor cells, with the possible mismatch, and on analog amplifiers. The concept results into a threefold loop-within-loop system, illustrated in figure 3.23. While the variations over time are not fast, having a convergent and stable system is not trivial. More over offsets in the amplifiers and between monitor instances will degrade the monitor performance.

The proposed digital implementation of the measurement and feedback system can be run on a separate digital processor. Alternatively the algorithm can during the spare cycles of a neighbouring processor, as memories are seldom stand alone applications. The digital part of the solution will also have a better scalability towards smaller technology nodes. Combined with the possibility to further simplify and optimise the search algorithm as discussed in section 3.4.4, the digital system has clear advantages over the analog implementation. The digital implementation also requires some periphery to interact with the,in essence analog, monitor cell. As discussed in section 3.4.4.4 the constraints for this periphery are rather relaxed and well within the boundaries of the current state of the art. Therefore the choice has been made in favour of the digital algorithm implementation in the test-chip implementation discussed in chapter 4.

3.5 Chapter Conclusion

The key defining feature of SRAMs is retaining the stored data correctly. Dual supply systems that lower the voltage across the cells not being activated, compromise this ability. To be able to evaluate the ability of an SRAM to retain its data, two bit integrity parameters, SNM and the N-curve, where introduced, respectively by [See87] and [Wan05] . Section 3.2 discusses these parameters and their extension to hold conditions. SNMh, the hold extension of SNM, is retained for the rest of the discussion as it performs adequately.

With SNMh it is possible to qualify the ability of an SRAM to retain data under lowered supply voltage. At design time, see also section 3.3.2.1, this can be used to find the minimal supply voltage at which a cell is able to retain its data, the DRV. The theoretical value for DRV can be written as formula 3.11. To compensate for PVT variations however extra margins have to taken on this value to have sufficient yield. When no feedback from the system is used this means the worst case corner has to be used to calculate the DRV. For most dies this will by definition not be the most power efficient option as only a very small minority of dies will be in those conditions.

Alternatively, every die can be screened at test time using the BIST. This approach has the benefit that the sleep supply voltage can be tuned on a die to die basis to compensate for both inter and intra die process variations. However, time dependent variations such as temperature or voltage can not be compensated. This again result in margins that have to be taken into account to allow the reliable retention of data. Changes in transistor parameters due to ageing can be compensated by rerunning the BIST loops with certain intervals, such as system reboots.

To be able to compensate for process and time dependent variations feedback from the system is needed on a die-to-die basis in real-time. To this end a monitor circuits is needed to be able to evaluate the bit integrity. This monitor can then be used in a regulation system that adjust the sleep supply as needed to guarantee reliable retention of the data. The canary approach presented in section 3.3.4.2 and first published by [Cal04] and [Wan07a], is such a monitor system. By skewing SRAM cells so they fail before the core matrix cells, failures to retain data can be detected before the data is lost. By organising such cells in a banks with different DRVs, a continuous spectrum of failure voltages can be obtained that allow a trade-off between reliable storage and power.

However, skewing cells inherently changes the behaviour of the cells towards environmental changes. To be able to qualify the effects of PVT better on the core cells, monitor cells are needed that characterise the core cells accurately. The approach presented in section 3.3.4.3 accomplished this. Cells identical in almost all aspects, be they electrical or layout, are rigged to allow measurement of the actual SNMh value. The only differences with the core cells being that the internal nodes are connect to the outside world and to other cells. The parallelism created reduces the influence of mismatch on the monitor cells while the DC characteristics barely change.

Measurement of the SNMh on the monitor setup can be accomplished in several ways. Section 3.4 discusses 2 approaches. An analog example implementation and a digital based solution are presented. The analog implementation is abandoned in favour of the digital implementation due to the stringent requirements on the analog building blocks that would be needed. The digital implementation with its algorithmic flow in figure 3.24, has the benefit of relaxing the requirements on the analog side (3.4.4.4) and to be able to run on spare cycles of a nearby processor.

By using the monitor setup described in section 3.3.4.3 and the measurement algorithm of section 3.4, the optimal point of the sleep supply voltage can be found and the retention guaranteed.

Chapter 4

Integration

4.1 Specification

To illustrate the concepts introduced in the previous chapters a 65kiBit dual port dual supply SRAM has been designed and implemented. A dual port system allows to temporarily increase the available memory bandwidth to deal with bursts of data. This is not uncommon in versatile multimedia applications where the excuted code becomes too dynamic to allow for compilation or design time optimisations on bandwidth usage over time.

The design is targeted to be used in high performance mobile application and as such power consumption is of major importance. Access frequencies in such application are typically 500MHz or below and data transfer is expected to be completed within one cycle. Leakage power consumption should be minimised while respecting this access frequency.

To limit the overhead in active energy on system level an SRAM architecture is used with two different port widths. A 32-bit and 256-bit access is implemented. This enables possible active energy savings by reducing the decoder energy overhead when reading large blocks of data [Rag07] or enables the selection of the most energy efficient for the required operation. Both ports can operate independently for both read and write operations. It has to be noted that no hardware safeguards are put into place to prevent a write or read operation on the same cells as those being accessed by the other port. The concurrent access on cells will have to be regulated at a higher level, such as the compiler.

4.2 Building Blocks

4.2.1 Dual port 10T cell

Traditionally cells used in dual port memories are based on the 6T-cell design as illustrated in figure 4.1. A pair (*BL,BLN*) of bitlines per port is used in combination with the needed access transistors. Two wordlines (*WL1,WL2*) are routed across the matrix to access the cells.

To fulfil the SNM requirements, it has to be taken into account that both pairs of access transistors can be turned on at the same time. This leads to an increased sizing of the core transistors to accommodate for the extra current. This sizing also has to take into account that the cell still has to be writeable by a single access transistor pair.

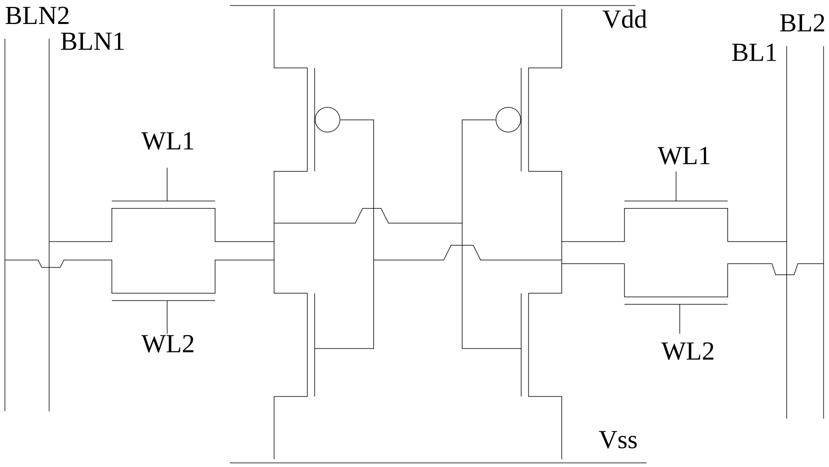


Figure 4.1: Traditional dual port SRAM cell based on the 6T-cell design

To reduce the bitline overhead of the cells, single ended cells will be used in this design. In single port SRAMs 5T designs have been used before as published in [Car04]. These single ended cells have the benefit of reducing the number of bitlines. However, the single sided access creates problems for the read and write access [Deh07].

The sizing of the transistors in the single sided cell has to be such that the single access transistor is strong enough to write both a high and low value into the cell, yet does not jeopardise the stored data during read. This leads to asymmetric sized transistors in the core transistor and a reduced precharge voltage on the bitline [Car04] or shortend bitlines [Cos07].

Alternatively, every cell is buffered for the read operation by inserting two extra transitors for the read path. The sizing of the core transistors can then be optimised for standby retention and the access transistor for write-ability without jeopardising the data integrity during the read operation [Ver07]. The internal storage nodes remain isolated from the bitlines during the read operation, effectivily making the cell SNM free. The sizing of the core cell transistors can then be optimised for retention. This will allow a further reduction in sleep voltage and hence a further reduction in leakage current. The schematic of this dual port 10T cell is shown in figure 4.2. The cell access happens through the BLW bitline for the wide access and the BLN bitline for the narrow access. For a writing operation the respective WLxW wordline will be actived. For a read operation the readbuffer will be enable by the respective WLxR wordlines.

It has to be noted that two extra wordlines have to be routed across the matrix as the read and write operation are now accomplished by different transistors in the SRAM cell. This does not create an extra power overhead as only one wordline can be active per port. Every wordline also has a reduced load as the total gate capacitance from the connected access transistors is halved compared to the dual ended access. The wordline connects to either the read transistor or the write transistor. The extra wordlines do

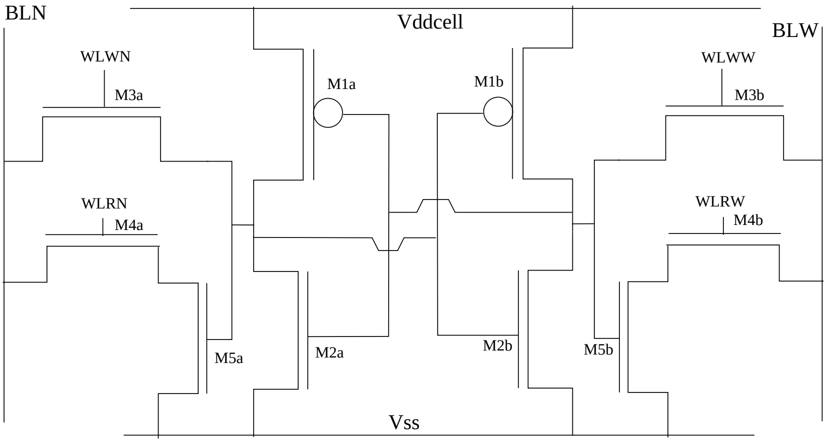


Figure 4.2: Schematic of the 10T dual port SRAM cell

generate an extra area overhead, especially in the decoders. Each wordline will need its own bufferchain, effectively doubling the area used by the buffers. In the matrix itself the area overhead is coupled with the area overhead created by the extra transistors that are inserted into the cell. Routing the extra wordlines across the matrix does not bring any extra area overhead as the lines can be placed with minimal distance and on higher metal layers.

Table 4.1 lists the used sizes for the 10T DPSRAM cell. Sizing for the cell can be done under DC conditions and is fully symmetrical for both ports. The core transistors $M1$ and $M2$ can be sized to optimise the retention at low voltages, taking into account the possible area penalty and variation on the SNMh parameter.

The main contributor to cell to cell intra-die SNMh variation is the mismatch of the core cell transistors. In accordance with Pelgrom's Law [Pel89] the mismatch reduces with the squareroot of the transistor area. As such the area occupied by the transistors should be high. This contradicts the SRAM paradigm of minimal area use. As the purpose of this thesis was to pursue leakage reduction a compromise solution was used. As there was no access to the specific SRAM rulesets for processing, the use of dogbone structures for minimal width transistors was enforced by the design and manufacturing rules. These structures create an area overhead for minimal transistors. In the same area as occupied by the dogbone minimal width transistors, transistors with increased sizing could be placed. This has the benefits of reducing the mismatch on the transistors, and as the length is also increased to satisfy the ratio constraints, reduced the leakage currents.

SNMh can be maximised by designing a cell where the eye-openings for SNMh_{low} and $\text{SNMh}_{\text{high}}$ are maximised. This can be achieved by sizing the transistor ratios of the core inverter as such that the cross-point in the butterfly curve is half the supply voltage and that the small-signal gain in that point is maximised. Hence, the theoretical and ideal upper boundary for the SNMh is half the supply voltage.

The sizing conditions to obtain the crossing point at half the supply voltage can be derived from current equilibrium. The current through the nmos pull-down has to be equal to the current through the pull-up pmos at the crossing point 4.1. Under the assumption that the transistors operate in the saturation region, the current can be written as equations 4.2 and 4.3.

$$I_{\mathrm{mp}} = I_{\mathrm{mn}} \tag{4.1}$$

where

$$I_{\mathrm{mp}} = \frac{\mu_p C_{\mathrm{ox}}}{2n} \frac{W_p}{L_p} \left(Vdd/2 - V_{T_p}\right)^2 (1 + \lambda_p Vdd/2) \tag{4.2}$$

$$I_{\mathrm{mn}} = \frac{\mu_n C_{\mathrm{ox}}}{2n} \frac{W_n}{L_n} (Vdd/2 - V_{T_n})^2 (1 + \lambda_n Vdd/2) \tag{4.3}$$

$$\tag{4.4}$$

Solving this towards the ratio of W/L this condition can be written as equation 4.5

$$\frac{W_p/L_p}{W_n/L_n} = \frac{\mu_p}{\mu_n} \cdot \frac{\left(Vdd/2 - V_{T_p}\right)^2}{(Vdd/2 - V_{T_n})^2} \cdot \frac{(1 + \lambda_p Vdd/2)}{(1 + \lambda_n Vdd/2)} \tag{4.5}$$

$$\tag{4.6}$$

The small-signal gain around the crossing point can be written as equation 4.7. The absolute value This gain should be maximised to obtain the best possible SNMh.

$$A = -\frac{gm_n + gm_p}{g_{DS_n} + g_{DS_p}} \tag{4.7}$$

In this formula gm_i and g_{DS_i} can be respectively rewritten as equations 4.8 and 4.9 [Lak94].

$$gm_i = \frac{2I_{DS_i}}{V_{gs_i} - V_{T_i}} \tag{4.8}$$

$$g_{DS_i} = \frac{I_{DS_i}}{V_{E_i} L_i} \tag{4.9}$$

where

gm_i	: transconductance of transistor i
g_{ds_i}	: drain-source transconductance of transistor i
V_{gs_i}	: gate-source voltage of the transistor i
V_{T_i}	: threshold voltage of transistor i
I_{DS_i}	: drain-source current of transistor i in saturation
V_{E_i}	: early voltage of transistor i
L_i	: lenght of transistor i

Transistor	Length (nm)	Width (nm)
$M1_{a,b}$	120	360
$M2_{a,b}$	160	240
$M3_{a,b}$	80	360
$M4_{a,b}$	80	240
$M5_{a,b}$	80	240

Table 4.1: Transistor sizes of the 10T DPSRAM cell

By filling in equations 4.2, 4.3, 4.8 and 4.9 into equation 4.7, the gain at the crossing point can be rewritten in function of the transistor lengths. The resulting equation is 4.10.

$$A = -\frac{2(Vdd/2 - V_{T_n}) + 2(Vdd/2 - V_{T_p})}{(Vdd/2 - V_{T_n})(Vdd/2 - V_{T_p})} \cdot \frac{V_{E_p}L_p \cdot V_{E_n}L_n}{V_{E_n}L_n + V_{E_p}L_p} \tag{4.10}$$

From equation 4.10 it is clear that a larger than minimal length is preferable to maximise the gain. As a trade-off between the used area, design rules and the constraints from equations 4.10 and 4.5 the values of table 4.1 were chosen.

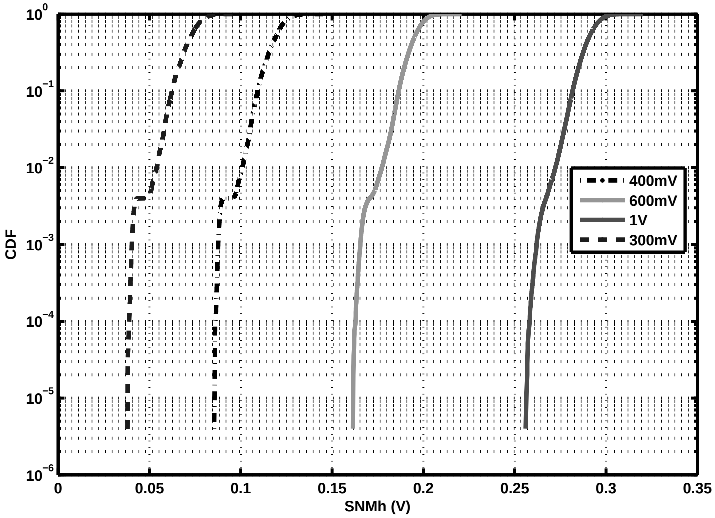


Figure 4.3: CDF of the SNM$_\mathrm{h}$ for the 10T cell for various supply voltages

Figure 4.3 shows the cumulative density functions(*CDF*) of the 10T cell under various supply voltages. The failure rate of the cell to attain a predetermined SNMh increases with lowering supply voltages. The spread on the failure rate will be taken into account when determining the reference voltage for the retention guaranteeing regulation circuit.

4.2.2 Secondary supply

The dual supply SRAM architecture requires the generation of a lower secondary supply on chip. This secondary supply voltage can be generated from the higher supply voltage by DC-DC conversion. As this constitutes a power overhead high efficient DC-DC conversion techniques are favoured. Whether these convertors use capacitors or inductors, they consume a large area [Kwo08].

Alternatively a series regulator could be used to create a lower supply at the cost of a low efficiency. The current reduction component from the power reduction will still be present, the voltage component will not. The figures 4.4 and 4.5 already mentioned in section 2.4.3 show the possible power saving factor for both the full DC-DC conversion and the series regulator solution.

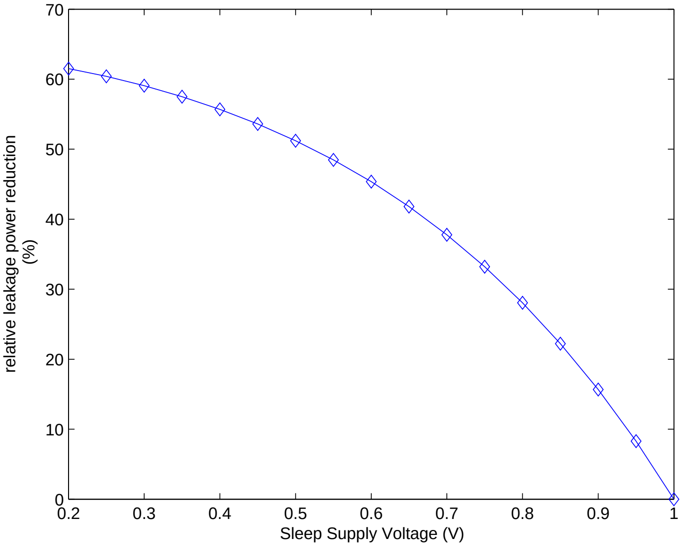


Figure 4.4: Possible power saving with a 80% efficiency DC-DC convertor as function of the sleep voltage

For DC-DC converter based approaches the fine granular solution as discussed in section 2.4.2 and published in [Gee05] can be used without further adaptation. To use a series regulator, the fine granular architecture can be adapted to include the series regulator. Instead of an independent secondary sleep supply line routed across the matrix, the series regulator now connects the local cell supply line with the global nominal

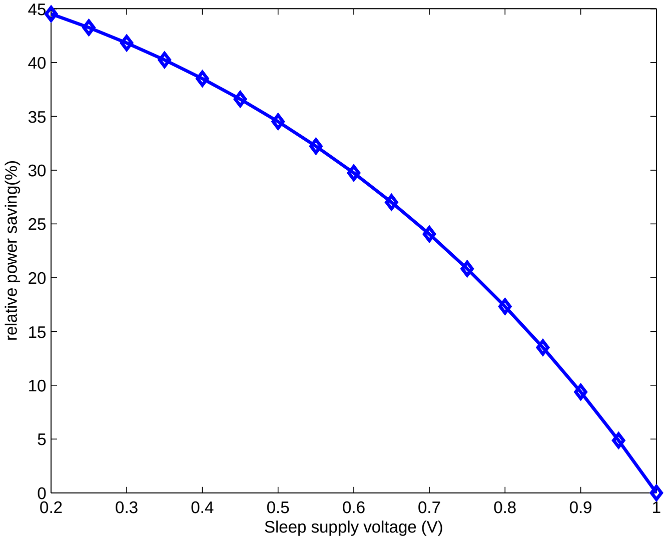


Figure 4.5: Possible power saving with a series regulator as function of the sleep voltage

matrix supply line.

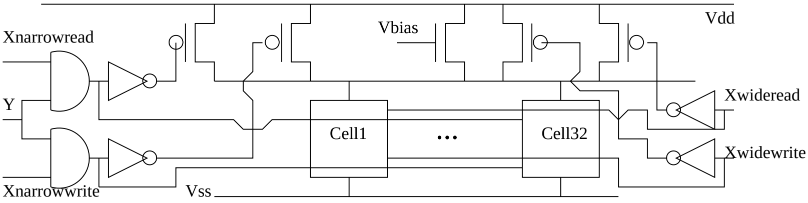


Figure 4.6: Fine granular word architecture with integrated series regulator

The series regulator is sized and biased so that it can keep the cell supply voltage on the predetermined sleep voltage. Power switches are added to allow the cell to regain nominal supply during read and sufficient current can be provided to the cells during the write operation. This leads to the schematic of the implemented circuit as show in figure 4.6.

4.2.3 Noise-margin measurement

As previously stated in section 3.3.3, minimising the leakage currents in the matrix requires the sleep voltage to be as low as possible without endangering the integrity of the stored data. This requires the implementation of a monitor and regulation circuit.

By using cells that are as close as possible to the actual matrix cells, the extra supply voltage margins to ensure data retention can be minimised. To this end the 10T DP-SRAM cell presented in section 4.2.1 is adapted to allow access to the internal storage nodes. As the SNMh value is a DC parameter, it will not be influenced by the extra capacitance that is introduced by this adaptation.

As the spread on the measured monitor value has a direct influence on the voltage of the sleep supply and as such on the total leakage power, it has to be minimised, see also section 3.3.4.3. In accordance with the law of large numbers [Ber51], the monitor cells will be connected in parallel to average out the SNMh value. This measure value is then a good estimate for the mean value of the actual cell distribution.

Five rows of 256 monitor cells are inserted into the cell matrix. These rows are spread across the matrix to also lower the dependency of the monitor SNMh on local variations in environment. The measurement and regulation circuit for the testchip are kept offchip.

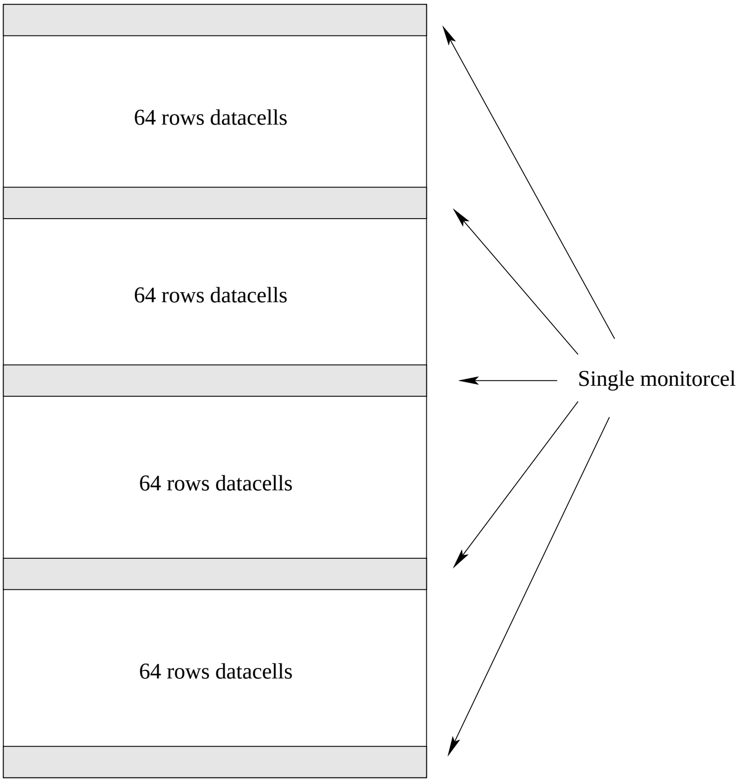


Figure 4.7: Overview of the monitor cells organisation in the SRAM core matrix

The measured estimated SNMh mean for the SRAM matrix cells combined with the predetermined failure boundary as per figure 3.17, will allow to set the reference value for the regulation circuit. This value is to be set so the retention of the data can be guaranteed with sufficient margin to guarantee the functional yield of the matrix. This value can be derived from the Monte-Carlo generated statistics as have been presented in figure 3.21 in section 3.3.4.3. The combined statistics of the cell and monitor allow to make an estimate of the yield of the matrix with the correction based on the monitor measurements included.

As a secondary function this reference value can also be used to modify the operational speed of the SRAM at the cost of leakage power. The power-speed trade off will also be clear from the measurements in section 4.3.3, more specifically table 4.4 which is also also published in [Gee08].

4.2.4 Decoders

Two decoders need to be implemented to select the correct data word for a read or write operation on the DPDSSRAM. One decoder combination is needed for the narrow 32bit access, while another one is needed for the 256bit wide access. This is shown in figure 4.8.

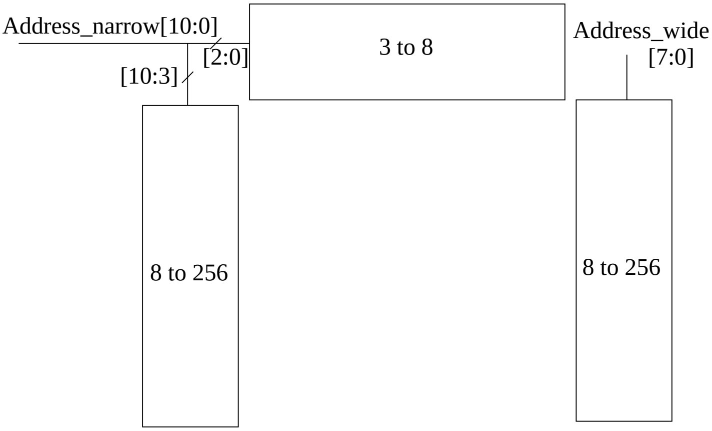


Figure 4.8: Decoder organisation of the DPDSSRAM

As the matrix is organised in rows of 256 cells, selecting a single wide word is identical to selecting a full row. Therefore the row selection decoders for both wide and narrow access will be implemented identically as 7-to-256bit decoders. Per decoded address also the difference between read and write needs to be made. This is implemented by adding a multiplexer at the last stage of the decoder to select the correct buffer tree.

The row decoder is implemented in two main stages. A first static combinatorial stage will decode the 8bit input address to 4x4 bits. The second stage recombines these four groups to create all 256 possible outputs. Figure 4.9 shows this decoder organisation.

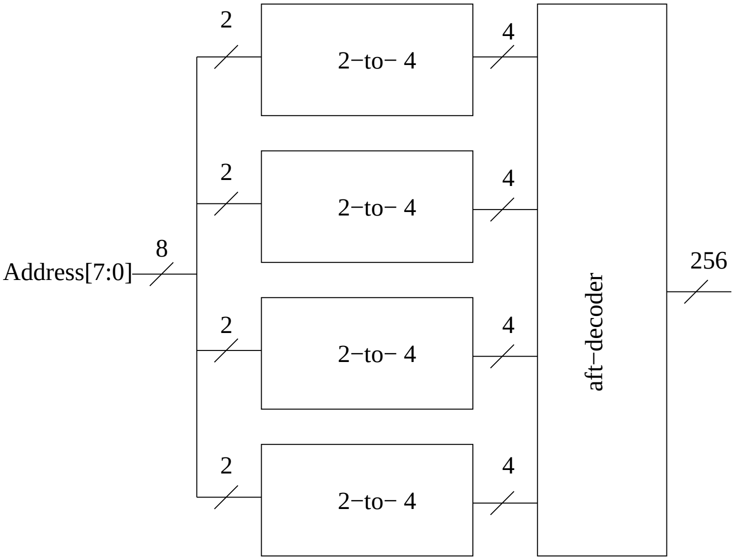


Figure 4.9: row decoder internal organisation

The four 2-to-4 decoders are built out of standard CMOS NAND gates.

The recombination stage of the decoder consists of dynamic gates as illustrated in figure 4.10. This pseudo-nor decoder has the advantage of limiting switching on the internal nodes towards the buffers [Nam98]. During the precharge phase the input on this stage can change without causing any significant switching internally. Once the input can be assumed to be stable the decoder can be put in evaluation mode by lowering the precharge signal. All but one of the decoder output stages will discharge its internal node n_1, but not cause any external switching on n_{out}.

The word column decoder decodes a 3bit address into a one hot 8bit selection signal. This Y-wordline is used in the distributed last stage of the decoder and to enable the correct set of sense-amplifiers. The Y-decoder is implemented as a single stage dynamic decoder similar to the last stage of the row decoder. The only difference is the use of 3 pull-down transistors instead of 4.

The distributed last stage of the decoder, see also figure 4.6, has a passgate based implementation of an AND-gate. As such the buffer chain for the wordlines needs to have the correct polarity. The row decoders have an active high output. The column decoder, which is only used for the narrow access, has an active low output.

4.2.5 Sense-amplifiers

To limit the delay and power consumption of a read operation, the bitlines are not fully discharged when the cell is read. This reduced swing signal has to be converted back to digital levels. This is accomplished by the sense-amplifier (*SA*).

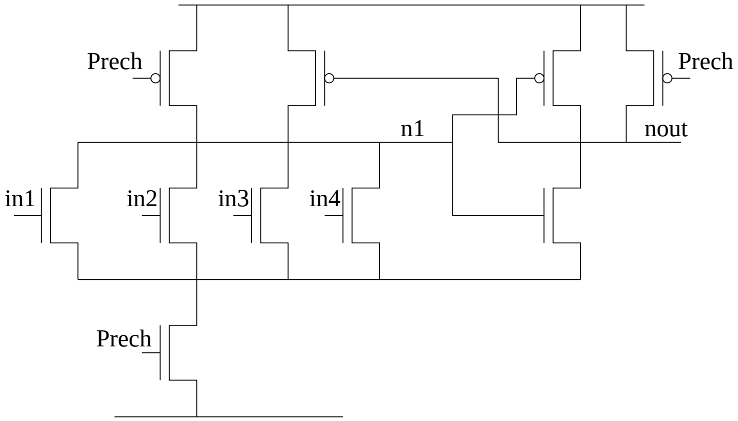


Figure 4.10: Final decoder stage [Nam98]

The SA is in essence a comparator. It amplifies the difference between a bitline (*BL*) and a reference (*Ref*) to a digital full swing signal (*Sa,San*). This reference can be an external applied reference voltage, or in the fully differential case the complementary bitline. As in this design the cell are single ended in nature, the SA will compare the bitline with an externally applied reference voltage. When the bitline drops below this predefined threshold before the SA is enabled, it is considered to be intentionally pulled down by the cell.

To reduce active biasing currents, the SA architecture used is build on the same latch architecture as the cell. Its activation is controlled by the SA_ACT signal, and as such consumes no static power when deactivated. The bitline to SA interface consist of a differential pair. This reduces the current load on the SRAM cells as no extra current is required to toggle the SA state [Kob93]. The schematic of the used SA is shown in figure 4.11

As 256 sense-amplifiers are needed to be able to read the 256bit wide word and to limit the area overhead, the SA needs to fit into the bitcolumn pitch. This limits the sizing that can be used to minimise input referenced offset on the input pair. This offset will effectively form the lower boundary for the bitline swing that needs to be developed to be able to sense the data correctly. The variance on the offset voltage can be obtained through Monte-Carlo simulations and sensitivity analysis. Taking into account that the high gain from the input stage to the output through the positive feedback stage, reduces the impact of the top transistors on the result leads to the approximation of the input referred offset as shown in formula 4.11.

$$\sigma^2_{\text{offset}} = \sigma^2_{\text{m1}} \tag{4.11}$$

with according to Pelgrom’s Law the variances on the mismatch to be function of the

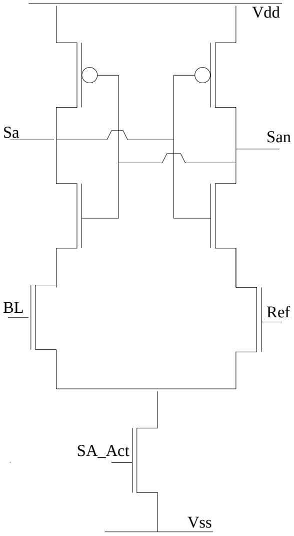


Figure 4.11: Schematic of the differential pair latch type sense-amplifier [Kob93]

transistor area [Pel89]

$$\sigma_{\text{mi}}^2 = \frac{(A_{V_T})^2}{\sqrt{W_i L_i}} + \frac{(V_{\text{GS}_i} - V_{T_i})^2}{4} \cdot \frac{A_\beta^2}{\sqrt{W_i L_i}} \tag{4.12}$$

where

W_i	: width of transistor m$_i$
L_i	: lenght of transistor m$_i$
A_{V_T}	: threshold voltage mismatch factor
A_β	: current mismatch factor

For the presented circuit a bitline swing of 100mV is sufficient to allow reliable detection. This illustrated the importance of the sense-amplifier activation timing. Letting the cells develop a higher swing on the bitlines increases the power consumption for the precharge but increases reliable read out. A lower swing would reduce the precharge power consumption but jeopardise the correct sensing of the data.

4.2.6 Timing and Control

Timing and control are an essential part of the overall DPDSSRAM design. All read or write operations on either of the ports are expected to be completed within a single clock cycle. As a result the enabling of the precharge circuits, decoders and sense-amplifiers has to be controlled in an asynchronous manner.

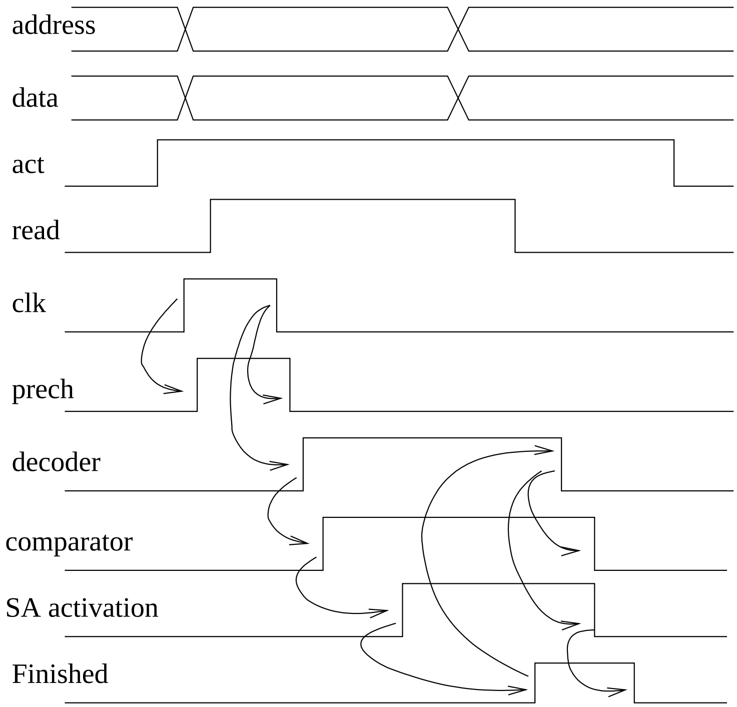


Figure 4.12: Control timing diagram

Figure 4.12 shows the correct sequence and causality , marked by arrows, of the control signals. When SRAM is activated (*act*), the clock (*clk*) will start the sequence. The rising edge of clk will enable the precharge (*prech*). The bitlines need to be precharged before an operation can take place. The falling edge of clk will disable prech. After the precharge is completed the decoder can be turned on to activate the wordline decoders. This in turn also starts the sensing on the timing bitlines with a continuous time comparator. When the bitlines are discharge sufficiently, the SA amplifiers should be turned on and the data stored for output. When all SAs have sensed their data, the finished signal is generated to reset the asynchronous state machine.

4.2.6.1 Timing

To create the timing of the control signals three main methods have been published in literature. These will be discussed briefly in the following paragraphs. While the timing of the signals is crucial to the performance and correct operation of an SRAM, the papers published on SRAM do not quantify the influence of the circuit with sufficient detail to be able to single out its effects. To be able to quantify the effects two designs only differing in timing generation should be made and evaluated.

While the delays needed between the consecutive control signals can be made by simple delay lines, their design will depend on worst case situations. The delays will have to be designed based on extensive reliable simulation of the whole SRAM. This reliable simulation depends on the correct extraction of parasitic effects based on the layout of the SRAM. Extra margins also have to be taken into account as there is a system level mismatch between the delay circuit creating the control signal and the actual needed delays in the SRAM. For instance, an inverter chain used to time the sense-amplifier activation will react differently to environmental changes than a bitcolumn. The timing still is expected to function properly under those changes, this creates an iterative design process. To compensate for the non-similar behaviour margins need to be taken into account to compensate both for the timing generation variations as the variations of the controlled elements [Nam98].

The margins taken on the timing delays takes its toll on the SRAM speed and power performance. Signals will be delayed more than necessary in most cases, creating a penalty in speed. The power penalty comes from bitlines that are being discharged further than needed and added leakage in activated parts of the memory. Using tunable delay lines can partly alleviate these effects, but requires extensive testing after fabrication [Cos07, Cos08].

The timing and control system used in this design, is based on a self-tuning approach. The most crucial part of the timing control is the activation of the sense-amplifiers. By using the actual signals generated the system a closer match between the activation signal and the required delay can be obtained than by depending on inverter generated delays [Amr98]. Dummy timing wordcolums are added to the matrix. This dummy column has cells with fixed data, allowing the bitline levels to be measured by a reliable continuous time comparator. This comparator will activate the other sense-amplifiers when the threshold level on the bitlines has been reached for reliable sense-amplifier operation. A continuous time comparator is needed to accomplish this. As the function of the SRAM has to complete in a single clockcycle, clocked comparators would require the generation of several high speed clock signal triggers to determine the right moment to activate the sense-amplifiers.

The continuous time comparator used for the detection of the bitline threshold level is shown in figure 4.13. It consists of three stages: a pre-amplifier to do a voltage to current conversion, a decision circuit and a post-amplifier to regenerate digital levels at the output [All82]. Table 4.2 shows the sizing of the transistors.

Although the input offset of this comparator can not be compensated inside the system [All82], it can be tuned out by adapting the external reference voltage.

The location of the dummy wordcolumn at the end of the wordline guarantees it will get activated as the last one on the wordline, creating an extra safety-margin. At the same time the position in the Y-direction will compensate for the cell to SA distance, which varies from row to row. Completion detection at the level of the sense-amplifiers will allow the generation of a stop signal for the decoders to turn off the activated wordline. Figure 4.14 shows an overview of this system.

The precharge phase duration for the bitlines is be controlled by a tunable delay line. This precharge phase should be long enough for the bitlines to be recharged to their

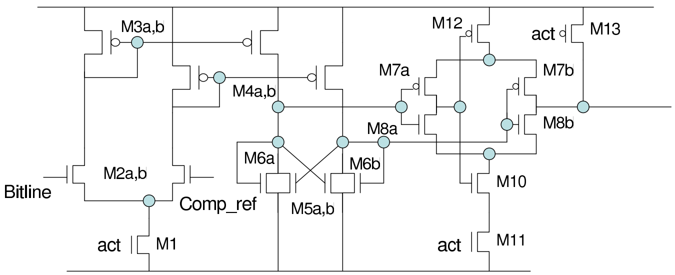


Figure 4.13: Continuous time comparator in three stages

transistor	width (nm)	lenght(nm)
M1	600	400
M2a,b	12000	400
M3a,b	600	400
M4a,b	600	80
M5a,b	1200	400
M6a,b	900	400
M7a,b	720	80
M8a,b	720	80
M10	720	80
M11	720	80
M12	720	80

Table 4.2: Transistor sizes used in the comparator

precharge voltage, and for the first stage of the decoder to be come stable for its evaluation phase.

4.2.6.2 Control

The control system is build as an asynchronous finite state machine as everything has to happen within a single clock cycle. This control circuit is build up using SR-latches and edge-detectors as shown in figure 4.15.

The edge-detectors seen in figure 4.16, consist of a small delay line made out of invertors and a NAND-gate. The pulsed activation of the SR-latches is needed to avoid race conditions in the latch. The delay in the edge-detector is build using NAND-gates instead of the normal invertors to better match with the internal delay of the SR-latches under variation.

The SR-latches are build up out of NANDs and include an external reset signal to initialise in the correct state. These latches have an active low activation.

The control circuit generates in this way all the signals illustrated in 4.12 for both the

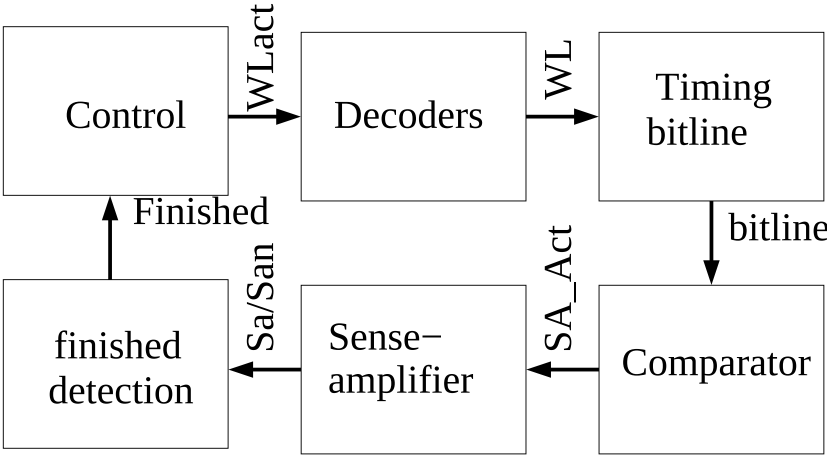


Figure 4.14: Overview of the self-tuning timing for the SA and decoders

narrow and wide cases. It has to be noted that the activation of the column-decoder (Y) in the narrow case has to precede the row-decoder(X). In the read scenario this just enables the CT-comparator before the bitlines start discharging. In the write scenario, the Y-decoder also selects the column where the data is written. The delay between the Y-decoder and X-decoder activation must ensure the bitlines are charged or discharged to the correct writing voltage before the cells are accessed.

4.2.7 Precharge and Write Periphery

Interaction of the SRAM cell with the rest of the SRAM happens through the bitlines. The cell can discharge the bitline to read or the bitline can be precharged to the level needed to write the data into the cell. The cell however is not designed to pull the bitline high. To accomplish this the precharge circuit is needed. The circuit will precharge the bitlines to the precharge voltage, typically near the supply voltage.

To reduce the leakage on the bitlines by the access-transistors, the preharge voltage is lowered to benefit from the DIBL effect[Gee05]. This leakage not only contributes to the total leakage power but also limits the number of cells that can be connected to a bitline [Gro06]. The precharge voltage is set to 600mV.

The precharge circuit is a single transistor per bitline connect to the precharge voltage rail and controlled by the *precharge* signal. A buffer tree from the control circuits to the precharge signals is included to allow driving the accumulated capacitance.

As the cells are designed to be single ended, it has to be possible to force a 0 or 1 state through the access transistor. As seen in section 4.2.1, the access transistor is sized to allow the internal node to be raised above the trip point of the cell. The voltage and current needed for this have to be provided by the write circuit. The write circuit will have to be able to pull the bitline high or low while the cell is connected to it.

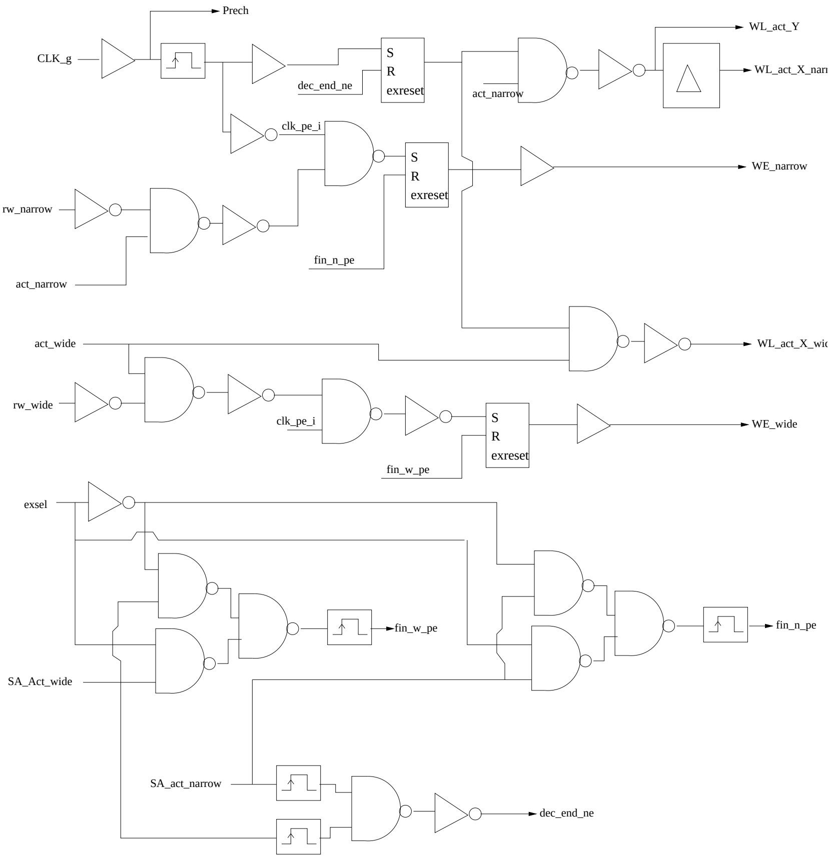


Figure 4.15: Overview of the control system

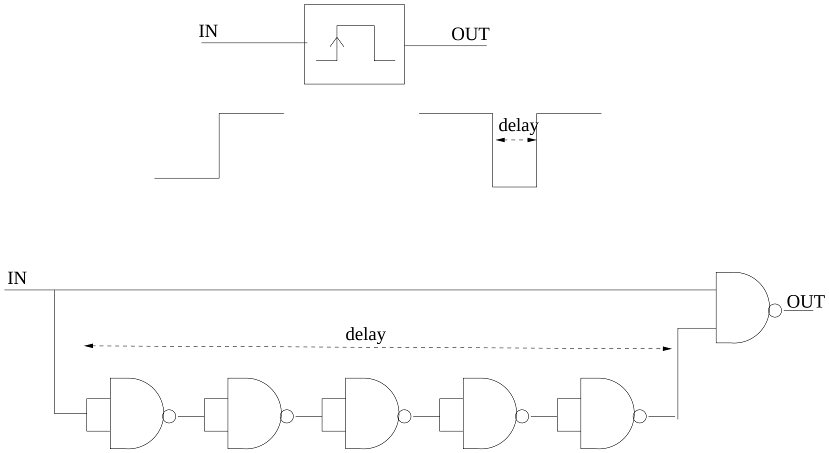


Figure 4.16: Edge-detector as used in the control system

To be able to achieve its function the write circuit requires the input data to be distributed to the transistors together with the write enable signal. To reduce area and energy overhead the pull-up and pull-down transistors are adjacent in the layout. The write enable signal for the narrow port will be combined with the Y-decoder wordcolumn selection as to only charge or discharge the needed bitlines. This eliminates any overhead on the non-access bitlines. As the wide access will need a whole row, the write signals can be applied directly.

4.2.8 Measurement Periphery

To be able to measure the the performance of the SRAM with a fine precision, off-chip measurements are nigh impossible. The added inductance from the bondwires and capacitance from the package and PCB, create an lowpass LC-filter that effectively blocks the high frequency signals we want to measure. To solve this, an on-chip solution to accurately measure the timing of the signals has been developed.

The timing signals that need to be measured will be fed into a transparent latch line. This line will be long enough to have a wide capture window. These latches will then be frozen by making them non-transparent after enough time has passed for the measured operation to be completed. This is illustrated in figure 4.17. This frozen latch line can then be loaded into a shiftregister to export the measurement data at a lower frequency offchip, where it can be interpreted by software.

The attainable measurement resolution is then the transition delay of a single latch cell. This is measured to be around 100ps. Due to process variations however this delay can change from die-to-die. To this end the latch lines delay will be calibrated by measuring the oscillation frequency of a ring oscillator built out of identical latches. This oscillation will be brought off-chip divided to lower the frequency and measured

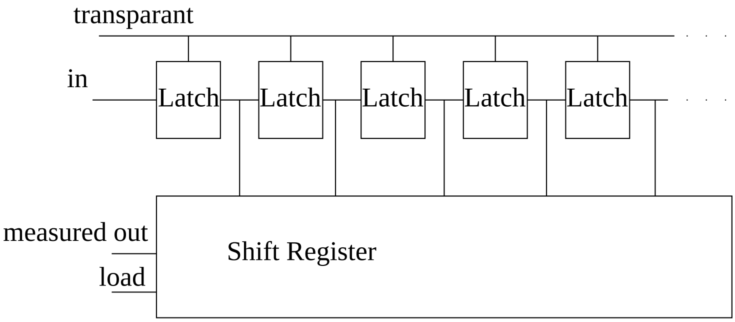


Figure 4.17: Measurement freezing latch line

as reference.

The measured signals are wordline activation, sense-amplifier enabling (sa_act) and the end of the sense operation (fin_narrow, fin_wide). This will allow the duration measurement of the important phases in an access. As seen in section 4.2.6, the wordlines will be turned off after receiving the finished signal from the sense-amplifiers. As such the time the wordline is active in the measurement is a good measure for the access delay of the memory.

To be able to generate a pulsed precharge signal with a tunable duty cycle, a tunable delay line as shown in schematic 4.18 is added on chip. By setting the correct selection bit through a shift-register the pulse signal with a variable duty cycle can be generated. The pulse width can be tuned from 200ps to 1.5ns with a stepsize of 150ps in the 450ps to 1ns range, and 250ps in the outer ranges.

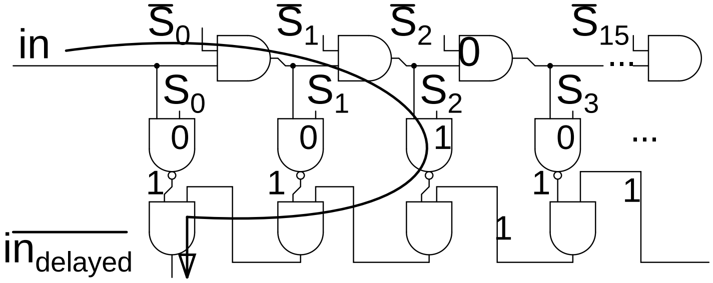


Figure 4.18: Schematic of the tunable delay line as used to generate the internal precharge signal [Cos08]

4.3 Implementation in 90nm CMOS

4.3.1 Layout

All building blocks of section 4.2 are brought together in one working SRAM system. This dual port dual supply SRAM has been processed in mainstream 90nm technology. It has to be noted that we did not have access to the special SRAM design rules and had to make cell compliant with the standard logic design rules. This leads to an area overhead compared to cells made in a dedicated process.

The layout of the dual port 10T-cell, see also section 4.2.1, is drawn in figure 4.19. To avoid the use of dog-bone structures and the associated area overhead, the width of the transistors is slightly increased. These changes are already reflected in the sizing table 4.1. This also has the added benefit of lowering the sensitivity to variability as the area slightly increases. Which in turn is beneficial to the leakage current reduction as it allows the sleep supply voltage to drop more. For the same reasons, the cell is also kept as symmetrical as possible.

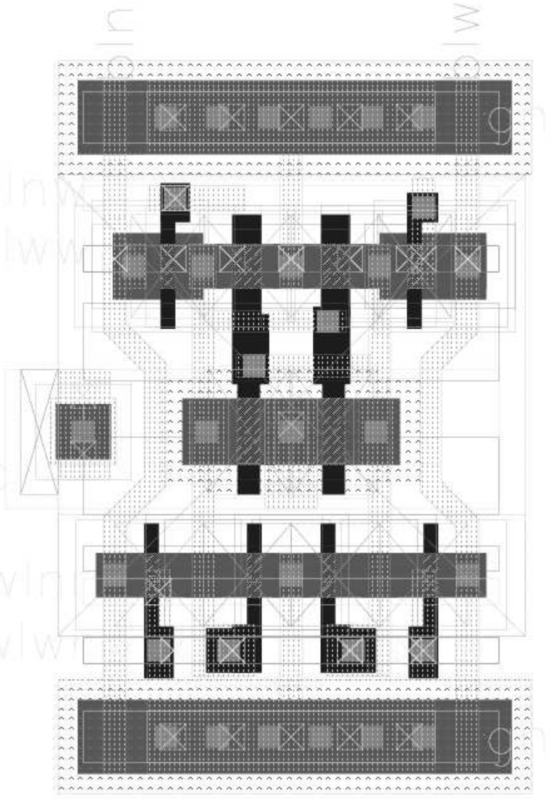


Figure 4.19: Layout of the dual port 10T-cell

The monitor cell is by design almost identical to the normal cell. It only differs in the extra contacts and metal wiring to access the internal nodes for measurement of the data integrity.

The layout of a single word with the additional headers is shown in figures 4.20 and 4.21. On the left hand side the extra word level periphery for the narrow access is situated. The wide access periphery can be seen on the right hand side. The power switching transistors are in between the access logic and the actual cell matrix. Also the transistor acting as a local series regulator is marked.

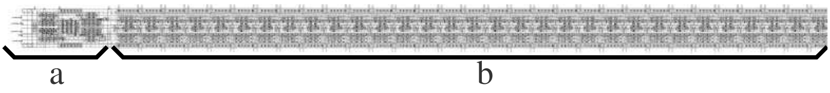


Figure 4.20: Layout of the single 32bit word with a) the periphery and b) the data cells

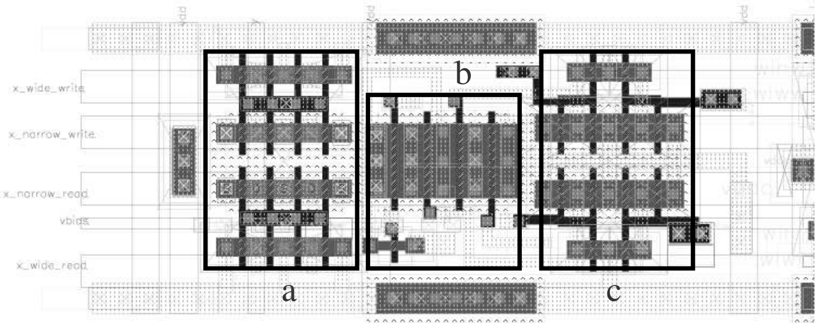


Figure 4.21: Layout of the word periphery with a) the passgate-nand access periphery, b) the power switches and series regulator, c) the local control inverters

The predecoder and control logic are based on standard logic ports and as such not reprinted in layout here.

For the analog components special care was taken to reduce systematic offset by having symmetrical layouts and similar environments. The pitch of the SAs is made identical to the cell pitch as 256 SAs are needed for the wide access. The input transistors of the continuous time comparator have also been sized and layouted to reduce mismatch influence.

The full combined layout is show in figure 4.22. The SRAM core containing the matrix, sense-amplifiers, decoders and control circuitry takes an area of 700μm x 700μm. The total chip has an area of 1200μm x 1800μm . The peripheral circuits are also annotated on the figure. These include the shiftregisters for the SPI programming interface and the on-chip measurement circuits. In figure 4.23 a micrograph of the fabricated die is shown.

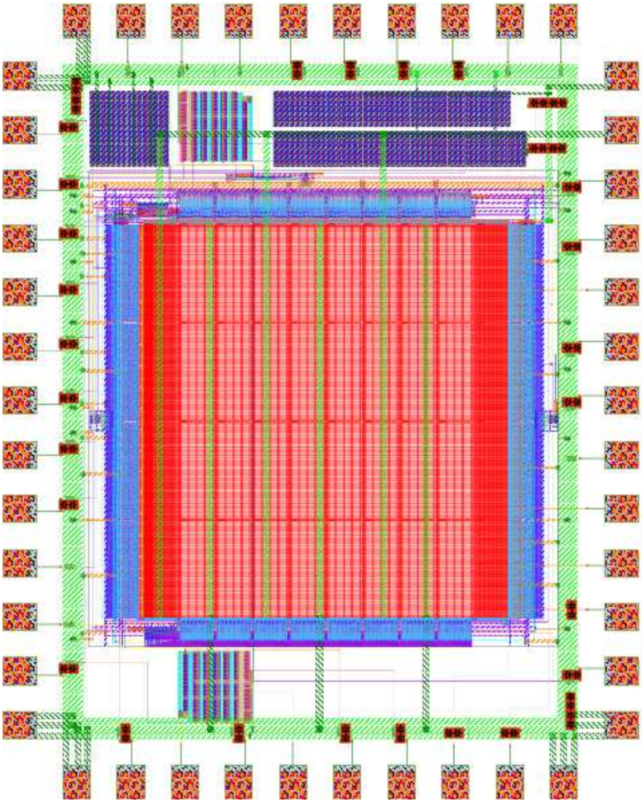


Figure 4.22: Combined layout of the full DPDSSRAM

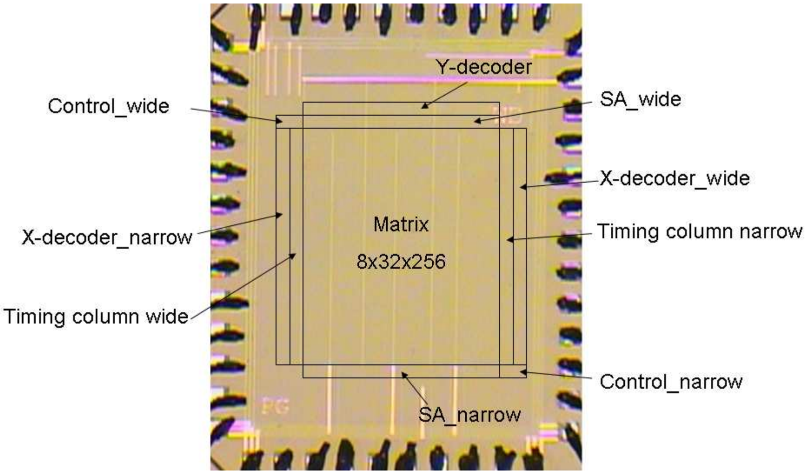


Figure 4.23: Chip micrograph of the bonded die

The digital circuits and conversion interfaces for running the monitoring algorithm were not implemented on chip. The interface to the monitorcells is brought off-chip and the algorithm runs in software on a PC for the purpose of this prototype.

4.3.2 Measurement and test setup

To limit the number of pins on the die, the inputs and outputs of the SRAM have been interfaced to the outside world with shiftregisters. The measurement delay lines are also loaded into shiftregisters to allow read out. To allow easy interpretation of the acquired data an interface with a PC had to be built.

Figure 4.24 shows a photograph of the PCB used during the measurement and test of the DPDSSRAM dies. The die is placed in a central zero insertion force socket. Level-shifters to translate the signals from the nominal chip voltage of 1V to the PCB 3.3V are inserted on the paths to and from the socket. The interface between PC and PCB uses an FTD 2231 USB-to-serial interface chip. This chip has the ability to run an SPI-like protocol to the die while having the capability to address the muxes on the PCB correctly.

This interface allows easy access to the DPDSSRAM and the measurement data, which in turn can be interpreted by the custom written software.

4.3.3 Measurement results

The leakage currents of the matrix were measured in function of the matrix sleep supply voltage. The results are plotted in figure 4.25. The highest attainable supply voltage was 606mV due to the V_T loss in the NMOS series regulator. Below 180mV all data was lost and as such does not consist a valid operational region.The evolution shows

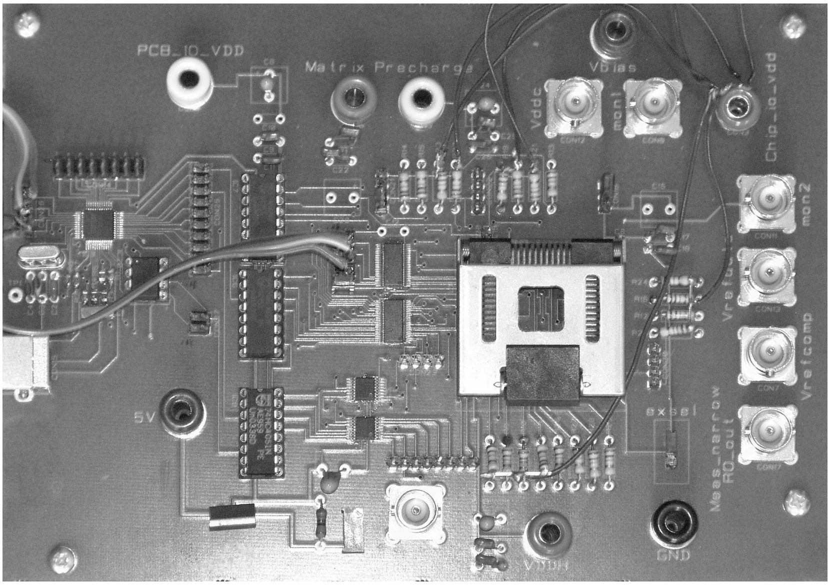


Figure 4.24: Photograph of the PCB used during measurement and test of the processed dies

a weak exponential behaviour as expected from the theoretical model in section 2.4.3. A reduction from the sleep supply from the nominal 1V to 200mV leads to a current reduction from an estimated 1100μA to 530μA, or a factor 2.

The SRAM operates in three modes. The nominal mode, where the matrix sleep supply is kept at 600mV. The low leakage mode, where the matrix sleep supply is reduced to 200mV, the minimal retention voltage that allows for the guarantee to maintain the stored data. The last mode of operation would be the pure retention mode, where the periphery of the matrix such as the decoders would be switched off. This would reduce the contribution of the periphery to near zero in the total leakage current of the SRAM system. Table 4.3 gives an overview of the leakage current contributions for the different operational modi.

The cost of the leakage power reduction comes in the form of an increased delay and active power consumption due to recharging the word supply rails for an access. The measured access delay in the two single cycle operational modes is shown in table 4.4. As the pure retention mode does not allow to access the data within one clockcycle, it is not added to the table.

The extra delay in access with a lowered sleep voltage can be attributed to two factors. It takes longer for the power switches to bring the word supply rail to the active level. Secondly, as the gate voltage on the pull down read transistor starts lower, the read current generated on the bitlines is lower. No measurable difference was found between the narrow and wide access delays. From the total access delay 450ps consists of the

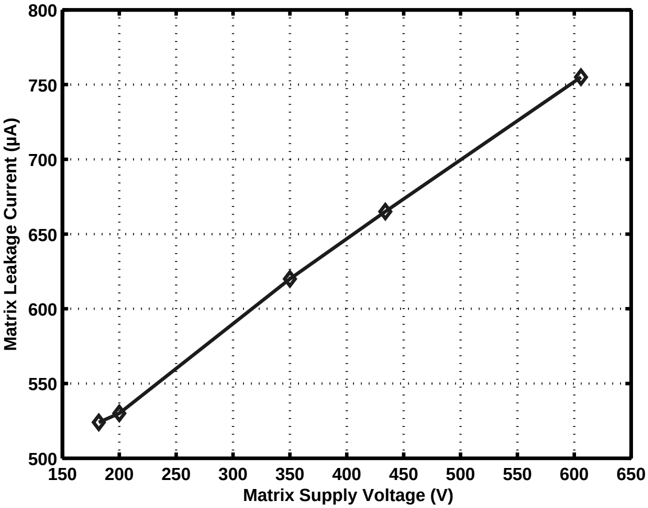


Figure 4.25: matrix leakage current as function of the sleep supply voltage

mode	matrix leakage	matrix sleep supply	periphery leakage	periphery voltage
	μA	mV	μA	V
nominal	755	606	750	1
low leakage	530	200	750	1
retention	530	200	0	0

Table 4.3: Overview of the different voltages and leakage current contributions in the different operational modi for the SRAM system

mode	delay	matrix leakage current
nominal (600mV)	2ns	750μA
low power (200mV)	2.5ns	530μA

Table 4.4: Summary of the matrix leakage and speed numbers of the DPDSSRAM in high-speed and low-leakage mode

access port	periphery (pJ)	precharge (pJ)	total (pJ)	total/bit (pJ)
narrow (32 bit)	13	3	16	0.5
wide(256 bit)	15	9	24	0.09

Table 4.5: average active energy per access

precharge phase needed to bring the bitlines to the precharge voltage.

The dual width dual port approach has the goal to reduce the energy overhead associated with reading several narrow words compared to reading a single wide word. To read one single wide word of 256 bit, eight accesses of different 32bit words would be needed without the dual width dual port approach. The gain will be in the decoder overhead power. The number of SA activations, word activations and to be precharged bitline capacitance are equivalent for the single 256bit wide access, compared to the eight 32bit accesses. It has to be taken into consideration that access energy is also data dependent. Data read-outs that do not discharge the bitlines will require less energy to be charged to the correct precharge voltage again. The energy data shown in table 4.5 is based on checkerboard data of alternating ones and zeroes (101010101...).

The energy numbers of table 4.5 are barely influenced by the sleep supply voltage. The energy difference between a 600mV and 200mV sleep supply consists only of the virtual rail capacitance that needs charging. This parasitic capacitance can be neglected compared to the other switched capacitance. The measured differences were less than 5% of the values in table 4.5 and as such fall within the error boundary of the measurement.

The high precharge energy values for the narrow access are however not in concordance with the expected values. The explanation for this anomaly lies in the limitations of the test setup. As the access times are measured on chip, the interface towards the chip was not build for high speeds. As a consequence the time between consecutive measurement cycles is unrealistically high and the bitlines are discharged by the cells through the leakage currents before a new access cycle starts. This generates an overhead in precharge energy as the system precharges all bitlines before the correct word column is fully decoded. To fundamentally solve this requires an adaptation of the precharge control to precharge only the bitlines that will be accessed. This would create extra delay as the precharge phase can only start when the Y-decoder can produce a correct evaluation of the address bits, where in the current solution the decoders can start working in parallel with the precharge phase.

$$C_{\mathrm{WLpB}} = \frac{C_{\mathrm{WL}}}{N} \cdot \sum_{k=1}^{\log_4(C_{\mathrm{WL}}/C_{\mathrm{inv}})} \left(\frac{1}{4}\right)^k \tag{4.13}$$

where

C_{WLpB} : estimated switched wordline buffer capacitance per bit
C_{WL} : total wordline capacitance
N : number of bits in a single word
C_{inv} : capacitance of a single unity size inverter

The active energy numbers confirm the thesis that it is possible to save energy on system level by accessing a single wide word instead of several narrow word operations. Under the assumptions that the total number of bits in a matrix remains constant, a similar architecture and without a restriction on the feasible aspect ratio on the SRAM, an increase in wordlenght will reduce the active energy per accessed bit. A doubling of the wordlenght would not increase the wordline-capacitance per bit. The total wordline-capacitance would increase, which in turn will increase the wordline buffers. Although this increase in buffer size will have a only a very small repercusion on the total switched capacitance per bit as can conclude from formula 4.13 where a rule of thumb fanout of 4 is taken into account. The overhead of the sense-amplifiers would also remain the same on a per bit basis. The bitline-capacitance would halve on a per bit basis. And as a doubling of the wordlenght halves the number of words in the matrix, the decoders would also reduce in size and effective switched capacitance per bit. The wordlenght and effective energy per operation per bit are hence limited by the ability of the full system to cope with wider words and the feasibility of large aspect ratios.

4.4 Comparison with State-of-the-Art

Table 4.6 retakes the information from table 1.1. The differences in the used technologies make it hard to establish a valid comparison. The ratio of the sleep versus the nominal supply would indicate how aggressive the sleep voltage could be scaled using the reported methodologies. In that regard this work realises one of the highest supply reductions while guaranteeing data retention. The work reported in [Wan07a] can claim a higher ratio under the best environmental conditions. It also claims a lower absolute sleep voltage even under typical conditions. It however gives no information on any dynamic, area or cell design aspects.

The resulting leakage reduction ratios are heavily influenced by the technology parameters and the ratio of gate versus subthreshold leakage. Gate-leakage is reported to be dominant in [Nii04] with 80% of the leakage and contributes the highest fraction to the established savings. In contrast for the technology used in this work [Gee08] gate-leakage contributes less than 1% to the cell leakage. This is not uncommon for low-V_T processes. As subthreshold current is less sensitive to the supply voltage compared to gate-leakage, the resulting reduction will also be less. It also has to be noted the reduction is listed as a current reduction as not all published work include a means to determine or generate the sleep voltage.

For the work reported in [Kim06, Wan07a, Wan08] it is unclear whether the wake-up and access to the data can happen in the same clock cycle. As waking up blocks of 128kiBit, as reported in [Wan08, Kim06], represent a significant load capacitance and hence a high delay and power overhead, it is more likely that the blocks are woken up for several accesses and prior to being accessed or that access is pipelined.

The lack of monitoring systems in [Nii04, Sal05, Kim06] suggests the use of a worst case design methodology to guarantee data retention if it was taken into account. From this work and [Wan07a] it is clear they could have benefited from a further supply voltage reduction if they would have had implemented such a monitor system. The differences between this work and [Wan07a] have already been discussed in section 3.3.4. The use of the SNMh monitor allows a finer tuning on a die-to-die basis as it has to take less margins into account for variability and environmental effects.

Hence the methodologies in this work will outperform the methodologies presented in the current state of the art.

parameter	Nii [Nii04]	Saliba [Sal05]	Kim [Kim06]	J. Wang [Wan07a]	Y. Wang [Wan08]	this work [Gee08]
size	256kiBit	16kiBit	128kiBit	128kiBit	1MiBit	64kiBit
nominal supply	1.2V	1V	1.8V	0.5V	1.2V	1V
sleep supply	0.6V	0.3V	0.9V	0.07V best 0.15V typ	0.5V	0.2V
leakage current reduction	88%	86%	94.2%	83% best 50% typ	90%	50%
delay	2.8ns	3ns	1.02ns	N/A	0.9ns	2.5ns
overhead delay	0	9%	2%	N/A	N/A	25%
overhead active power	0	N/A	“high”	N/A	N/A	≤1%
overhead area	13.2%	3.5%	6%	0.6%	N/A	12.5%
granularity	block	row	matrix	N/A	128kiBit block	word
technology	90nm	150nm FDSOI	180nm	90nm	65nm ULP	90nm HP
data integrity	N/A	worst case	N/A	canary	N/A	SNMh

Table 4.6: Overview of the dual supply SRAMs published in the open literature. N/A notations mean the data was not available from the published material.

4.5 Chapter Conclusion

A dual port dual supply SRAM was designed and fabricated in a commercial 90nm technology and measured.

By using a 10T-dual port single ended cell, energy savings where enabled. The precharge voltage on the bitlines could be lowered to reduce bitline leakage. The four cell core transistors could be designed to reduce leakage further and to have less variation in SNMh. This in turn enabled the sleep supply to be further lowered without compromising on data integrity. By adding the readbuffer the remaining access transistor could be optimised to successfully write the cell, while not risking any disturbance to the core storage nodes during read.

The noise-margin monitor as described in section 3.3.4.3 and implemented as in 4.2.3 allows the sleep supply voltage to be adjust such that the required SNMh can be guaranteed on a die-to-die basis. By connecting several cell together the spread on the measured monitor SNMh could be reduced to have an accurate measure for the most expected SNMh value of the core cells. To guarantee data retention a small extra margin must be taken into account to compensate for the spread of the SNMh on the core cells.

The timing and control circuitry as described in section 4.2.6 using dummy word columns and a continuous time comparator enable to track and adjust the control signals in function of the access address and variations. The dummy column matches best with the actual access word, creating a better timing solution than standard worst case design. This reduces the energy consumption on the bitlines as the sense-amplifiers will be triggered once the comparator senses a sufficient difference has been generated on the dummy bitlines. This minimises the margins on the sense-amplifier activation and the effective swing on the bitlines.

The on-chip measurement setup based on “freezing” latches eliminated the need to have a high speed interface to the outside world.

The dies have been measured in the laboratory. The access time of 2ns is attained in its nominal settings a measured in section 4.3.3. In low matrix leakage settings with a 200mV sleepy supply, the access time is increased to 2.5ns with an extra leakage current reduction of 33% compared to the nominal settings of 600mV as sleepy supply voltage. The system can be put in a pure retention state by turning of periphery circuits, this reduced the leakage power consumption with 65% compared to the nominal case.

Chapter 5

Conclusions

5.1 General conclusions

The paradigm shift in electronics towards more mobile and more multimedia devices has generated a number of design challenges. To accomplish the required functionality in mobile devices the energy efficiency of operations must increase as battery power is limited. Battery life-time is also defined by the power of the device used in “standby”.

The evolution of more functionality for less area and power is driven by the technology scaling describe by Moore’s Law [Moo69]. However, with the advent of deep deep submicron technology a new set of problems has appeared. The leakage currents have increased with shrinking device length and variations have a larger impact on the transistor behaviour. The leakage current has grown to such an extend that leakage power has become the most dominant factor in the power consumption of SRAMs. To achieve low power operation for almost any systems that means SRAM leakage power is a major concern.

The analysis of effects contributing to leakage is made in section 2.2. The resulting formulas for subthreshold and gate leakage show the parameters that can be used to reduce the leakage currents. One of the key parameters to reduce the leakage current is the supply voltage, as it influences both the subthreshold and gate leakage. The V_T of the transistors and gate length are other parameter that define the subthreshold leakage current.

Section 2.4 discusses the possible implementation of the reduction techniques in the specific environment of an SRAM. This leads to the introduction of dual supply SRAMs (*DSSRAM*). In a DSSRAM a secondary supply voltage is introduced, either through means of a virtual supply rail, a virtual ground rail or a combination of both. Through the DIBL effect the leakage currents will be reduced with lower drain-source voltages. This lower supply voltage is applied to cells that are not active but just retaining data. As the cells in this drowsy mode are isolated from the bitlines, disturbing the stored data becomes less likely. Due to the reduced leakage current and the lowered supply voltage, power savings up to 95% are possible in the SRAM cell matrix for technologies with a high DIBL coefficient. This thesis followed the path of reducing the supply voltage on the non-accessed cells.

The dual supply system can be applied in several granularities, the largest being the whole SRAM, the smallest consisting of just a single data word. This granularity has an influence on power savings, delay and control overhead. Crude granular architectures such as whole memories or banks of a memory have the simplicity of control, but suffer from drawbacks in wake-up delay and wake-up power. Waking up a whole memory to retrieve a single word has a large dynamic and passive energy cost. While at first sight the finest granular structure might be the hardest to control, we solved this issue by distributing the last stage of the decoder into the matrix [Gee05]. Combining the X and Y-signals locally before the word achieves several benefits with a small area overhead. Firstly, only the word that is needed is woken up, with a small power and delay penalty. Secondly, it creates a hierarchical wordline structure that reduces the capacitance on the global wordline. This results in a reduction of the dynamic power needed to drive this full-swing line, and the leakage from the wordline buffers, which can be scaled down. Lastly, it solves one of the timing problems associated with drowsy bits. As their supply voltage is reduced, the cells become more susceptible to external influences, especially in the read case. An access transistor turned on fully before the cells have reached their nominal supply, compromises the stored data. By having the last stage of the decoder distributed and controlling the power switches, the last stage of the wordline buffer can also be localised and connected to the virtual supply rails. This ensures the driving voltage of access transistors scales the same way as the supply voltage of the cells waking up.

Using this fine granular system makes it possible to maximise power savings and minimise wake up delays. However, reducing the supply voltage on the cell in a retention state comes at a cost.

The key defining feature of SRAMs is retaining the stored data correctly. Dual supply systems that lower the voltage across the cells not being activated, compromise this ability. To be able to evaluate the ability of an SRAM to retain its data, two bit integrity parameters, SNM and the N-curve, where introduced, respectively by [See87] and [Wan05] . Section 3.2 discussed these parameters and their extension to hold conditions. SNMh is the extension of SNM under hold conditions and as such purely based on voltages. The alternative parameters SINMh, SVNMh and SPNMh are derived from the N-curve. SINMh is the most complementary to SNMh as it provides information on the current, where SVNMh only provides another measure based on the crossing voltage of the butterfly curve. For ease of use SNMh has been retained for the rest of the thesis as it performs adequately.

With SNMh it is possible to qualify the ability of an SRAM to retain data under lowered supply voltage. At design time, see also section 3.3.2.1, this can be used to find the minimal supply voltage at which a cell is able to retain its data, the DRV. To compensate for PVT variations however extra margins have to be taken on this value to have sufficient yield. When no feedback from the system is used this means the worst case corner has to be used to calculate the DRV. For most dies this will by definition not be the most power efficient option as only a very small minority of dies will be in those conditions.

Alternatively, every die can be screened at test time using the BIST. This approach has the benefit that the sleep supply voltage can be tuned on a die to die basis to compensate for both inter and intra die process variations. However, time dependent variations such as temperature or voltage can not be compensated. This again result in margins that have to be taken into account to allow the reliable retention of data. Changes in transistor parameters due to ageing can be compensated by rerunning the BIST loops with certain intervals, such as system reboots.

To be able to compensate for process and time dependent variations feedback from the system is needed on a die-to-die basis in real-time. To this end a monitor circuit is needed to be able to evaluate the bit integrity. This monitor can then be used in a regulation system that adjust the sleep supply as needed to guarantee reliable retention of the data. The canary approach presented in section 3.3.4.2 and first published by [Cal04] and [Wan07a], is such a monitor system. By skewing SRAM cells so they fail before the core matrix cells, failures to retain data can be detected before the data is lost. By organising such cells in a banks with different DRVs, a continuous spectrum of failure voltages can be obtained that allow a trade-off between reliable storage and power.

However, skewing cells inherently changes the behaviour of the cells towards environmental changes. To be able to qualify the effects of PVT better on the core cells, monitor cells are needed that characterise the core cells accurately. The approach presented in section 3.3.4.3 accomplished this. Cells identical in almost all aspects, be they electrical or layout, are rigged to allow measurement of the actual SNMh value. The only differences with the core cells being that the internal nodes are connect to the outside world and to other cells. The parallelism created reduces the influence of mismatch on the monitor cells while the DC characteristics barely change.

Measurement of the SNMh on the monitor setup can be accomplished in several ways. Section 3.4 discusses two approaches. An analog example implementation and a digital solution are presented. The analog implementation is abandoned in favour of the digital implementation due to the stringent requirements on the analog building blocks that would be needed. The digital implementation with its algorithmic flow in figure 3.24, has the benefit of relaxing the requirements on the analog side (3.4.4.4) and to be able to run on spare cycles of a nearby processor.

By using the monitor setup described in section 3.3.4.3 and the measurement algorithm of section 3.4, the optimal point of the sleep supply voltage can be found and the retention guaranteed.

To prove this concept a dual port dual supply SRAM was designed and fabricated in a commercial 90nm technology and measured.

By using a 10T-dual port single ended cell, energy savings where enabled. The precharge voltage on the bitlines could be lowered to reduce bitline leakage. The four cell core transistors could be designed to reduce leakage further and to have less variation in SNMh. This in turn enabled the sleep supply to be further lowered without compromising on data integrity. By adding the readbuffer the remaining access transistor could be optimised to successfully write the cell, while not risking any disturbance to the core storage nodes during read.

The noise-margin monitor as described in section 3.3.4.3 and implemented as in 4.2.3 allows the sleep supply voltage to be adjust such that the required SNMh can be guaranteed on a die-to-die basis. By connecting several cell together the spread on the measured monitor SNMh could be reduced to have an accurate measure for the most expected SNMh value of the core cells. To guarantee data retention a small extra margin must be taken into account to compensate for the spread of the SNMh on the core cells.

The timing and control circuitry as described in section 4.2.6 using dummy word columns and a continuous time comparator enable to track and adjust the control signals in function of the access address and variations. The dummy column matches best with the actual access word, creating a better timing solution than standard worst case design. This reduces the energy consumption on the bitlines as the sense-amplifiers will be triggered once the comparator senses a sufficient difference has been generated on the dummy bitlines. This minimises the margins on the sense-amplifier activation and the effective swing on the bitlines.

The on-chip measurement setup based on “freezing” latches eliminated the need to have a high speed interface to the outside world. The measurement resolution attained by the measurement system is 150ps.

The dies have been measured in the laboratory. The access time of 2ns is attained in its nominal settings a measured in section 4.3.3. In low matrix leakage settings the access time is increased to $2,5$ns with an extra leakage reduction of 33% compared to the nominal settings. The system can be put in a pure retention state by turning of periphery circuits, this reduced the leakage power consumption with 65% compared to the nominal case. The active power consumption results can also be found in table 4.5. For the narrow 32bit port the average active energy per access is 16pJ or 0.5pJ/bit. For the wide 256bit port the average active energy per access is 24pJ or 0.09pJ/bit. The difference can be mainly found in the precharge energy. These numbers confirm the possibility to save energy on the system level by accessing the memory on the wide port when more than one 32bit word of data is needed.

The monitor and regulation system was also successfully measured, and predicted the minimum retention voltage of 200mV correctly.

As a general conclusion this thesis presents a system and the necessary background to create an SRAM where leakage currents can be minimised while guaranteeing data retention. The presented DPDSSRAM is also the first published [Gee08] dual supply SRAM that incorporates the measurement of the data retention parameter SNMh and the generation of the secondary sleep voltage on chip.

5.2 Future work

As noted before the power savings from the reduction in supply voltage for the sleepy part of the SRAM are highly dependent on the availability of high efficient DC-DC conversion. The research to create such a converter was however not the subject of this thesis. The challenge in the design of a highly efficient converter for a dual supply SRAM is twofold. The first obstacle is the full integration of the converter on chip without a huge area overhead. The second obstacle would be the conversion efficiency, the power consumption of the DC-DC converter should be negligible compared to the SRAM. If the comparator could be used not just for the SRAM but also for other IP blocks, this efficiency requirement could be mitigated. The design of such a system-on-chip would be the next logical step in research with this thesis.

The developed prototype already touched on the asymmetric access of memories as a way to reduce the energy consumption of applications. The wide access provides a significant reduction in active energy per bit compared to several narrow accesses. The optimisation of the different widths towards applications to further reduce the energy consumption would be a research subject on a higher abstraction level. The regulation system for the sleepy supply in this thesis can also serve the secondary function of a knob to tune the execution speed versus power and reliability. This illustrated the need to again design with awareness across the abstraction levels. With the projected increase in variability this way of working will only gain in importance as circuit designers need to be aware of technological and architectural developments. This will require the research and development of new design methodologies.

The reduction of leakage currents will only continue to grow in importance as mobile applications get more and more proliferated. The reduction of the supply voltage for non-accessed cells in the matrix as presented in this thesis can be further extended. The active supply can also be lowered to further reduce both active and leakage power. The development of weak-inversion logic and memories promises to cut power figures dramatically. However variability and susceptibility to noise currently make these systems to be unreliable in operation. Developing the methodologies, architectures and circuits to reliably operate in the weak-inversion region will provide a challenge and, possibly, a great reward in power reduction. Reducing the power consumption to such low level will be the key enabler to the true ubiquitous presence of sensor networks and mobile applications.

Appendix A

Matlab Implementation of the SRAM Power Model

MATLAB code

```
cellstruct=[[32 1];[32 1]];
wordstruct=couplediv(divider(nbw));

 for ind=1:length(Vsleep2)
  clear goodresults;
  goodresults{1}=[inf inf inf];
  clear nominalgood;
  nominalgood{1}=[inf inf inf];
  Vsleepa=Vsleep2(ind);
  %Vact=Vsleep2(ind); %added for pol
  %Vsleepa=Vact;
  k1=1;
  for k=1:length(cellstruct)
   for lbis=1:length(wordstruct)
     x1=cellstruct(k,1);
     x2=wordstruct(lbis,1);
     y1=cellstruct(k,2);
     y2=wordstruct(lbis,2);
     %control section;
     nbcontrolled_rows=nbcontrolled_rows1;

     if((x1*x2*xdimcell/(y1*y2*ydimcell)>maxaspect)...
       ||((y1*y2*ydimcell/(x1*xdimcell*x2)>maxaspect)))
       cost=inf;
       nominalcost=inf;
     else
       %put line or column activation control where
       %there are most.
       %This only depends on wordmatrix organisation
       %(all bits of a word must be activated at the
       % same time)
       if (x2 >y2)
         nbcontrolled= x2;
         nbuncontrol=y2;
         Csupply=nbbits*Csupplycell;
       else
         nbcontrolled=y2;
         nbuncontrol=x2;
         Csupply=nbbits*Csupplycell;
       end;
  %suppose control per word
      nbsleepcells=nbbits*(nbw-1);
      nbactleakcells=0;
      nbactcells=nbbits;
      %taking wordlines over shortest distance on
      %matrix.
      %Presuming interconnect cap to be 3/4 of total
      %average cell load
      % switching the supplies lines uses act power
      % -> Csupply
       if (x1*x2*xdimcell >y1*ydimcell*y2)
        Ceff=Cload*y2*(1+(y1-1)*3/4);
        % Csupply=Csupplycell*y1*y2;
       else
        Ceff=Cload*x2*(1+(x1-1)*3/4);
        % Csupply=Csupplycell*x1*x2;
      end;
        % worst case supply switching :
       % every access different wordline
       nomactpow=nbactcells*Pact(Ceff,Vact,Vswing,freq)
        + decodercost([[x1 y1];[x2 y2]],Vact,freq);
       nomleakpow=nbactleakcells...
       *Pleaka(Vact,Vact,Vprech,eta,gamma,n,Vt,tech)
          +nbsleepcells...
          *Pleaks(Vact,Vact,Vprech,eta,gamma,n,Vt,tech);

      actpow=nbactcells*Pact(Ceff,Vact,Vswing,freq)...
              +readpct/100*Pact(Csupply,Vact,...
              Vact-Vsleepa,freq)...
             +decodercost([[x1 y1];[x2 y2]],Vact,freq);...
              % +writepct/100*Pact(Ceff,Vsleepa,Vsleepa,...
              % freq)

      leakpow=readpct/100*nbactleakcells...
        *Pleaka(Vact,Vsleepa,Vprech,eta, ...
                      gamma,n,Vt,tech)+ ...
              (nbsleepcells+nbactleakcells*writepct/100)
                *Pleaks(Vact,Vsleepa,Vprech,eta,...
               gamma,n,Vt,tech);
               cost=actpow+leakpow;
      nominalcost=nbactcells*Pact(Ceff,Vact,Vswing,freq)
          +nbactleakcells...
           *Pleaka(Vact,Vact,Vprech,eta,gamma,n,Vt,tech)...
          +nbsleepcells...
           *Pleaks(Vact,Vact,Vprech,eta,gamma,n,Vt,tech)...
          +decodercost([[x1 y1];[x2 y2]],Vact,freq);
      dec=decodercost([[x1 y1];[x2 y2]],Vact,freq);
    end;
    if (not(cost==inf))
      nominalgood{k1}=[k lbis nominalcost];
      nomactgood{k1}=[k lbis nomactpow];
      nomleakgood{k1}=[k lbis nomleakpow];
      actgood{k1}=[k lbis actpow];
      leakgood{k1}=[k lbis leakpow];
      goodresults{k1}=[k lbis cost];
      decoder{k1}=[k lbis dec];
      k1=k1+1;
   end;
  end;
 end;
 poweruse=inf;
 nompow=inf;
 optindex=1;
 optindex2=1;
 for k2=1:length(goodresults)
   if (goodresults{k2}(1,3) < poweruse)
     poweruse=goodresults{k2}(1,3);
     optindex=k2;
   end;
   if (nominalgood{k2}(1,3) < nompow)
     nompow=nominalgood{k2}(1,3);
     optindex2=k2;
   end;
 end;
 xopt1(ind,nbwloop)=
  cellstruct(goodresults{optindex}(1,1),1);
 xopt2(ind,nbwloop)=
  wordstruct(goodresults{optindex}(1,2),1);
 yopt1(ind,nbwloop)=
  cellstruct(goodresults{optindex}(1,1),2);
 yopt2(ind,nbwloop)=
  wordstruct(goodresults{optindex}(1,2),2);
 results(ind,nbwloop)=
  goodresults{optindex}(1,3);
 nomresults(ind,nbwloop)=
  nominalgood{optindex2}(1,3);
 nomactres(ind,nbwloop)=
  nomactgood{optindex2}(1,3);
 nomleakres(ind,nbwloop)=
  nomleakgood{optindex2}(1,3);
 actres(ind,nbwloop)=
  actgood{optindex}(1,3);
 leakres(ind,nbwloop)=
  leakgood{optindex}(1,3);
 decres(ind,nbwloop)=
  decoder{optindex}(1,3);
end;

newindex=(nbw-startnbwords)/nbcontrolled_rows1+1;
possibleSaving(newindex)=
 (min(nomresults)-min(results))/ ...
   min(nomresults)*100;
tmpact=
 nomactres(find(nomresults==min(nomresults(:,nbwloop))));
plotnomact(newindex)=tmpact(1,1);
clear tmpact;
tmpleak=
 nomleakres(find(nomresults==min(nomresults(:,nbwloop))));
plotnomleak(newindex)=tmpleak(1,1);
clear tmpleak;
tmpact=
 actres(find(results==min(results(:,nbwloop))));
plotadact(newindex)=tmpact(1,1);
tmpleak=
 leakres(find(results==min(results(:,nbwloop))));
plotadleak(newindex)=tmpleak(1,1);
p2(newindex)=min(results(:,nbwloop));
nom(newindex)=min(nomresults(:,nbwloop));
tmp=find(results==min(results(:,nbwloop)));
xoptb1(newindex)=xopt1(tmp(1,1));
xoptb2(newindex)=xopt2(tmp(1,1));
yoptb1(newindex)=yopt1(tmp(1,1));
yoptb2(newindex)=yopt2(tmp(1,1));
nbw=nbw+nbcontrolled_rows1;
```

Appendix B

Clock cycle accurate monitor model in MATLAB

MATLAB code

```
while((not(eq(SNMintern, SNMextern)))&&(x<(supply/2))...
&&(not(eq(supplymem(1),supplymem(3)))))
  loopcount=mod(loopcount+1,3)+1;
   while((abs(snmtmp-snmtmpmax)>stopcrit)&&(x<supply/2)...
    &&(snmtmpmax<SNMextern)||eq(loop,0));
     if (snmtmp>snmtmpmax)
       snmtmpmax=snmtmp;
       xmax=x;
     end
     if (eq(breakloop,1))
         sprintf(’loop break’)
         break;
     end
     loop=1;
     x=x+inc;
     inc;
     y=interp2(invy,invx,invz,supply,x);
     figure(4);title(’X evolution’);hold on; grid on;
     plot(time,x,’bd’);
     plot(time,y,’r*’);
     time=time+1;
     xtempbis=interp2(invy,invx,invz,supply,y);
     time=time+1;
     newx=x;
     newy=y;
     while ((xtempbis<newx))
       if (xtempbis > supply/2)
         error(’xtempbis overflow’);
       elseif(xtempbis > newx);
         error(’stop crit failure’);
       else
         newx=newx-abs(inc);
         if(newx<0)
            newx=0;
```

```
      end;
      newy=newy-abs(inc);
      if(newy<0)
          newy=0;
      end;
      time=time+1;
      xtempbis=interp2(invy,invx,invz,supply,newy);
      time=time+1;
    end
  end
  snmtmpold=snmtmp;
  snmtmp=x-newx;
  newsign=sign(snmtmp-snmtmpmax);
  time=time+1;

  if(not(eq(1,newsign))&&not(eq(newsign,0)))
    sprintf(’switching sign inc after diff crossing’)
    time
    inc=inc*-1/2;
    trackback=1;
    x=xmax;
  end
  if(eq(increset,1))
     sprintf(’minimum inc used’)
     breakloop=1;
  elseif (abs(inc)<mininc)
      inc=sign(inc)*mininc;
      increset=1;
  end
  time=time+1;
end
SNMintern=snmtmpmax;
if(true)
  sprintf(’supply regulation signif statement’)
   if(SNMintern > 2*SNMextern)
    sprintf(’halving supply’)
    time
    supply=supply/2;
    inc=abs(incorig);
    x=x/4;
    supplymem(loopcount)=supply;
   elseif(SNMintern > SNMextern)
    sprintf(’minus sign inc SNM %g vs %g’,...
     SNMintern, SNMextern)
    time
    supply=supply - 0.01;
```

```
       inc=abs(incorig);
       supplymem(loopcount)=supply;
     elseif(SNMintern < SNMextern)
       sprintf('pos sign inc SNM')
       supply=supply+0.01;
       supplymem(loopcount)=supply;
       inc=abs(incorig);
     end
     breakloop=0;
     trackback=0;
     increset=0;
   end
   olddiff=0;
   diff1=0;
   diff2=-1;
   snmtmpmax=0.001;
   snmtmp=0.001;
   loop=0;
   inc=0.01;

end
```

Appendix C

Nederlandse Samenvatting

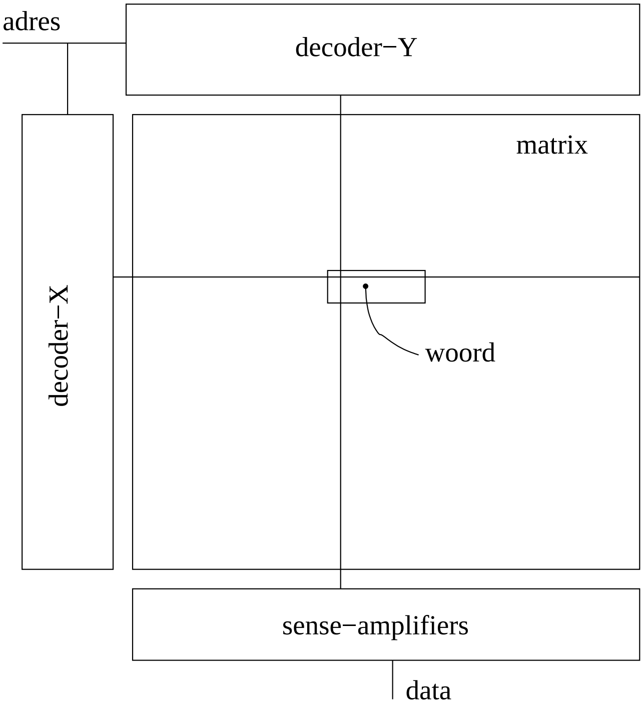
Inleiding

De evolutie van de mobiele telefoon van een “eenvoudige” draadloze telefoon tot een draagbaar multimedia platform is een uitstekend voorbeeld van de paradigma verandering in de huidige elektronica. Toepassingen evolueren alsmaar meer naar een hogere mobiliteit en grotere multimedia inhoud. Om dit te bereiken is een verbetering nodig van de energie efficiëntie, daar batterijen maar een beperkt vermogen kunnen leveren. Daarenboven is de operationele gebruikstijd van een toestel ook een belangrijk verkoopargument geworden. Het aantal bewerkingen op chip dat kan gedaan worden per watt, moet dus stijgen.

Het belangrijkste mechanisme om deze evolutie te kunnen is altijd technologische schaling geweest. Dit liet immers toe meer functionaliteit en een hogere densiteit voor chips te krijgen met een lagere productiekost. De wet van Moore [Moo69] modelleerde deze evolutie als een verdubbeling van de processor performantie ongeveer elke twee jaar. Met de komst van de deep-submicron technologieën, zijn er ook nieuwe uitdagingen voor de systeemontwerpers opgedaagd. Het verkleinen van de minimale transistorlengte vergroot ook het belang van de lekstromen op het totale vermogenbudget. Deze lekstromen reduceren in een belangrijk facet geworden van het systeemontwerp, zeker in het geval van mobiele toepassingen. Kleinere transistorlengtes betekenen ook dat de invloed van productievariaties aan belang toeneemt voor de performantie van systemen. Variabiliteit begint dus ook een belangrijke rol te spelen in digitale circuits en systemen.

Static Random Access Memories (*SRAM*) spelen een belangrijke rol in zowat alle moderne elektronische systemen. In de meest recente en meest performante processorsystemen is reeds meer dan 50% van de chipoppervlakte ingenomen door SRAMs. SRAM is echter ook een van de eerste bouwblokken om te lijden onder de negatieve effecten van schaling. Het grote aantal transistoren met een kleine activiteit maakt dat het vermogenverbruik veroorzaakt door lekstromen groter is dan het actieve vermogen. Om de gebruikte oppervlakte te beperken worden SRAM transistoren ook zo klein mogelijk gehouden. Dit leidt er toe dat variabilitietseffecten hier het eerste de kop zullen opsteken. SRAMs zijn ook gekozen als onderwerp van deze thesis, vanwege de invloed op het totale systeemvermogen en de verwachte moeilijkheden voor alle digitaal georiënteerde systemen.

De functionele omschrijving van een SRAM zou kunnen geschreven worden als: een circuit dat data kan opslaan en weergeven op een door een adres bepaalde maar willekeurige plaats. Om aan deze functionele beschrijving te kunnen voldoen bestaat een SRAM uit drie grote delen. De celmatrix, of kortweg matrix, om de data op te slaan onder vorm van woorden met een welbepaalde lengte. Deze matrix bestaat voor een SRAM uit cellen die de data bijhouden op basis van een positieve terugkoppellus van twee invertoren. Deze lus zorgt voor het statisch bijhouden van de data zonder de nood om deze te hernieuwen. De sense-amplifiers (*SA*) vormen het tweede belangrijke onderdeel. De SA's hebben als functie om de kleine signalen die op de bitlijnen door de cellen worden gezet te versterken tot volwaardige digitale signalen. Het derde belangrijke onderdeel wordt gevormd door de decoders. Deze zullen het aangelegde adres vertalen naar een fysieke locatie in de matrix. De onderdelen van een SRAM worden schematisch voorgesteld in figuur C.1



Figuur C.1: hoog niveau overzicht van de belangrijke onderdelen in een SRAM

Uiteraard zijn er nog andere onderdelen aan een SRAM: de timing controle, het schrijfcircuit en het circuit om de bitlijnen op te laden.

Het doel van deze thesis is om een besparing op het lekvermogen van SRAMs te realiseren. Zoals zal blijken uit de volgende sectie C, speelt de voedingspanning een belangrijke rol in de lekstroom. Dit verband zal dan ook gebruikt worden om de lekstroom te minimaliseren. Om de performantie van de geadresseerde cellen op pijl te houden, zullen deze op een hoge spanning gebracht worden. De niet-geadresseerde cellen zullen op een lagere spanning gebracht worden om er als het ware te sluimeren tot ze gewekt worden.

Het verlagen van de voedingsspanning voor de slapende cellen brengt echter ook een nieuw probleem mee. Een lager voedingsspanning betekent een lagere ruismarge. De vraag rijst dus: hoever mag de voedingsspanning zakken voordat de date verloren gaat. Deze vraag zal beantwoord worden in sectie C.

Voorgestelde oplossing werden geëvalueerd in een ontwerp dat besproken zal worden in sectie C

Lekstroomreductie in SRAMs

De lekstroom in SRAMs kan teruggebracht worden tot twee bijdrages: de subthreshold-lekstroom en de gate-lekstroom. De subthreshold-lekstroom is de stroom die door de transistor van drain naar source vloeit als de transistor af staat. De gate-lekstroom is de stroom die door de gate loopt omwille van kwantummechanische tunneleffecten. De recente ontwikkeling van high-κ [Mis07] reduceert echter de invloed van de gate-lek.

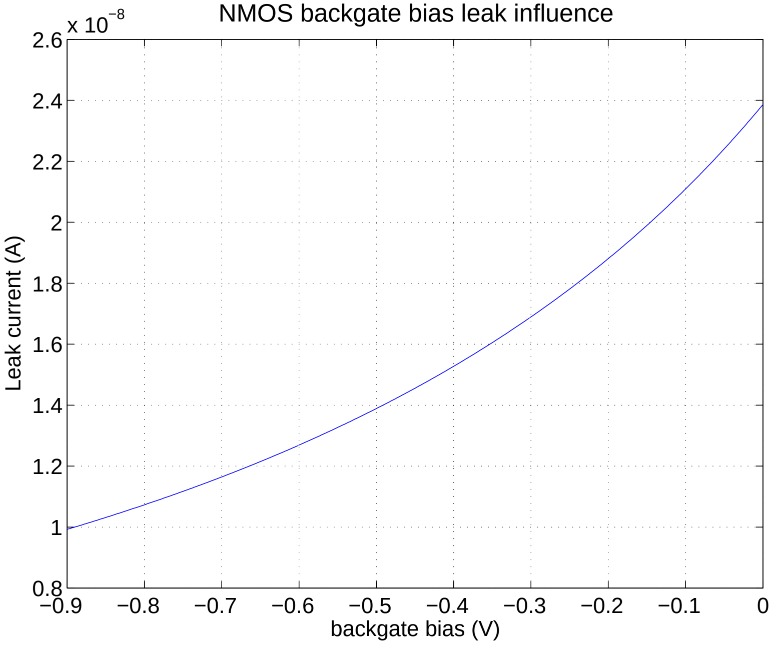
De subthreshold-lekstroom wordt benvloed door een aantal effecten waarvan enkel de belangrijkste hier worden weergegeven. Vooreerst is er de invloed van de bulk-source spanning. Door hier een negatieve spanning aan te leggen wordt de effectieve thresholdspanning V_T verhoogd. Dit zal de lek reduceren en wordt gemodelleerd met de parameter γ. Dit effect neemt echter af met elke nieuwe technologie generatie zoals gellustreerd in figuur C.2. De subthreshold-lekstroom wordt ook benvloed door de spanning over de drain-source van de transistor. Dit effect is gekend onder de naam drain induced barrier lowering (*DIBL*) en wordt gemodelleerd met de parameter η. Dit effect neemt toe met elke kleinere technologie generatie daar de afstand tussen drain en source afneemt. De derde grote invloed op de subthreshold-lekstroom is de temperatuur.

De formule C.1 verweeft al deze effecten op de subthreshold-lekstroom.

$$I_{lek} = I_0 \cdot exp(\frac{-V_{T0} - \gamma \cdot V_{BS} + \eta \cdot V_{DS}}{n \cdot V_{th}}) \cdot (1 - exp(\frac{-V_{DS}}{V_{th}})) \tag{C.1}$$

$$I_0 = \mu_0 C_{ox} \frac{W_{\text{eff}}}{L_{\text{eff}}} (V_{th})^2 e^{1.8} \tag{C.2}$$

met

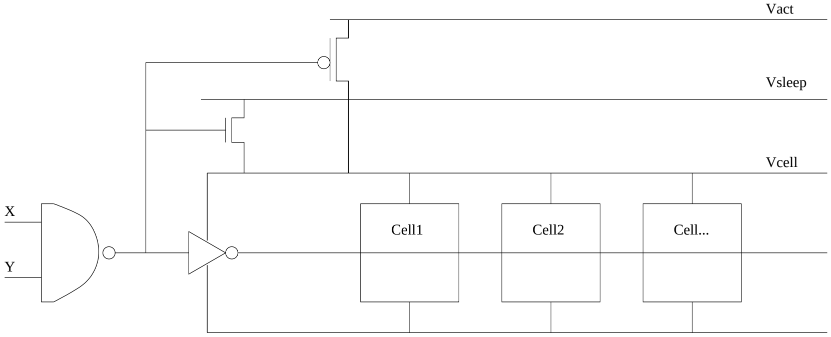


Figuur C.2: Subthreshold lekstroom als functie van de bulk-source spanning in verschillende technologieën voor een NMOS met minimale afmetingen

μ_0	: mobiliteit
C_{ox}	: gate oxide capaciteit
W_{eff}	: effectieve breedte
L_{eff}	: effectieve lengte
I_{lek}	: total subthreshold lekstroom
V_{T0}	: transistor threshold spanning
γ	: gelineariseerde bulk coëfficiënt
V_{BS}	: Bulk-source spanning
η	: DIBL coëfficiënt
V_{DS}	: drain source spanning
V_{th}	: thermische spanning

Zowel de subthreshold-lekstroom als de gate-lekstroom hangen exponentieel af van de voedingsspanning in een SRAM, daar deze zowel de spanning op de gate als de spanning over de transistoren bepaald. De voedingsspanning is dus een waardevolle ontwerpparameter om de lekstroom te benvloeden. Dit heeft de aanleiding gegeven tot het ontwikkelen van “drowsy caches” [Fla02].

In deze drowsy caches of dubbele-voedingsgeheugens worden de niet geactiveerde cellen op een lage voedingsspanning gehouden zodanig dat de lekstroom geminimaliseerd wordt maar de data bewaard blijft. De mogelijke vermogenreductie die zo kan bekomen worden is afhankelijk van een aantal factoren: de technologie factor η, de granulariteit van de gecontroleerde sectie. De hoogste reductie kan behaalt worden met de fijnste granulariteit. In het kader van dit werk werd de oplossing met de fijnste granulariteit, namelijk een enkel woord, gepubliceerd in [Gee05]. Dit zorgt voor de minste overhead in reactivatie, daar de capaciteit van de voedingslijn die geschakeld wordt minimaal is. Ook is de reductie maximaal daar enkel de benodigde cellen uit hun sluimertoestand gewekt worden. Het schema in figuur C.3 toont de controle overhead die lokaal per woord is gentegreerd. De recombinatie van de X en Y signalen uit de decoder lokaal voor het woord, levert een eenvoudige manier om de controle uit te voeren zonder een hoge kost in oppervlakte. Daarenboven laat deze fijne granulariteit toe om enkel de benodigde bitlijnen te laten ontladen, wat een actieve vermogenbesparing kan opleveren.



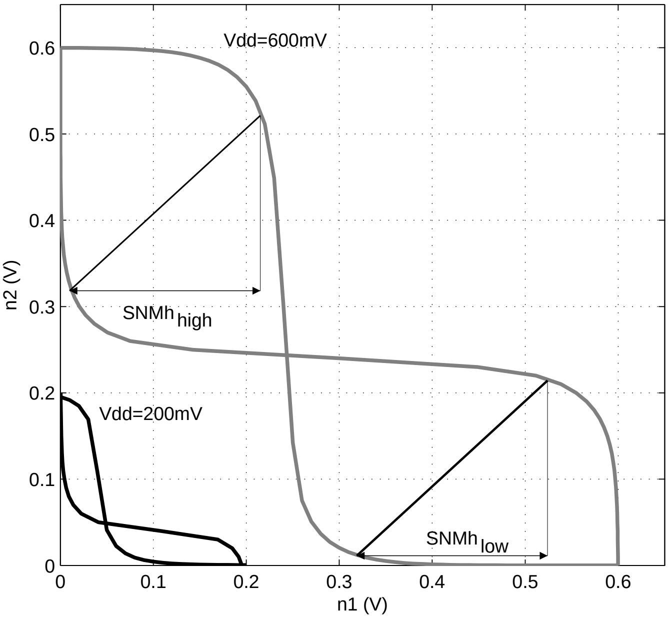
Figuur C.3: Schema op woordniveau van de oplossing met de fijnste granulariteit [Gee05]

Een reductie van de voedingsspanning creëert een situatie waarin de ruismarges voor de bewaring van de data onder druk komen te staan. Dit zal uiteindelijk de ondergrens vormen van hoever de voedingsspanning kan zakken. De volgende sectie zal hierop een antwoord formuleren.

Dataretentie in SRAMs

Om de dataretentie te kunnnen kwantificeren is een maat nodig. De maat die gebruikt wordt in dit werk is genspireerd op de door Seevinck [See87] gentroduceerde Static Noise Margin (*SNM*). De Static Noise Margin under hold (*SNMh*) wordt op dezelfde manier afgeleid maar met de cellen in niet-geadresseerde toestand. De SNM en SNMh criteria zijn minimax criteria. SNMh wordt gedefinieerd als het minimum van de maximale vierkanten die in een vlindercurve kunnen geplaatst worden. Figuur C.4 geeft een vlindercurve weer van een SRAM cell met een voeddingsspanning van 200mV en 600mV.

In het kader van de variabiliteit is niet alleen de waarde op zich nodig maar ook een modellering voor het statisch gedrag om het aantal falende cellen te kunnen bepalen.



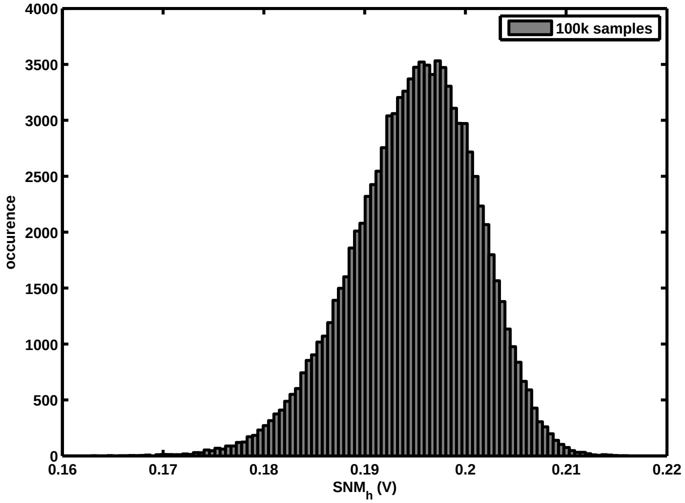
Figuur C.4: vlindercurves van de cel in bewaartoestand met aanduiding van de twee lokale extrema voor een voedingsspanning van 200mV en 600mV

Figuur C.5 geeft een voorbeeld verdeling weer voor een cel in 90nm. Daar SNMh een minimax-criterium is, kan het model afgeleid worden in functie van de verdeling voor SNMhigh en SNMlow. Deze twee verdelingen kunnen getrouw beschreven worden door gaussische normaalverdelingen. De resulterende formule is weergegeven in formule C.3. Dit model zal toelaten om de veiligheidsmarges te bepalen en als dus danig de ondergrens van de lage matrixvoeding.

$$\begin{aligned} f_{\text{SNMh}} = & \frac{1}{2\sqrt{2\pi}} \text{erfc}\left(\frac{x-\mu_h}{\sqrt{2}\sigma_h}\right) \cdot \exp\left(-\frac{(x-\mu_l)^2}{2\sigma_l^2}\right) \\ & + \frac{1}{2\sqrt{2\pi}} \text{erfc}\left(\frac{x-\mu_l}{\sqrt{2}\sigma_l}\right) \cdot \exp\left(-\frac{(x-\mu_h)^2}{2\sigma_h^2}\right) \end{aligned}$$

$$F_{\text{SNMh}} = \frac{3}{4} + \frac{1}{4}\left(\text{erf}\left(\frac{x-\mu_l}{\sqrt{2\sigma_l}}\right) + \text{erf}\left(\frac{x-\mu_h}{\sqrt{2\sigma_h}}\right) - \text{erf}\left(\frac{x-\mu_l}{\sqrt{2\sigma_l}}\right) \cdot \text{erf}\left(\frac{x-\mu_h}{\sqrt{2\sigma_h}}\right)\right) \tag{C.3}$$

met



Figuur C.5: voorbeeld van een SNMh distributie van een SRAM cell in 90nm op 600mV voedingsspanning

μ_l : verwachte waarde uit de SNMh$_{\text{low}}$ distributie van een enkele cell
μ_h : verwachte waarde uit de SNMh$_{\text{high}}$ distributie van een enkele cell
σ_l : standard deviatie van de SNMh$_{\text{low}}$ distributie van een enkele cell
σ_h : standard deviatie van de SNMh$_{\text{high}}$ distributie van een enkele cell

Om de dataretentie te kunnen garanderen onder een verlaagde voeding zijn er verschillende opties. Deze kunnen opgedeeld worden in twee grote groepen: de offline en online technieken.

Het ontwerp voor het slechtst mogelijke geval en de kalibratie tijdens de testfase zijn offline technieken. Vanuit het ontwerp voor het slechtst mogelijke ontwerp is het mogelijk de ondergrens voor de dataretentiespanning te berekenen door het oplossen van de stroomvergelijkingen in de cel. De resulterende oplossing kan geschreven worden als een iteratieve oplossing zoals weergegeven in formules C.4 en C.5.

Het startpunt DRV_1 kan bekomen worden door de startbenadering $V_1 = 0$ en $V_2 = Vdd$. Om een voedingsspanning te bekomen die voldoende garantie geeft voor het succesvol bijhouden van de data moeten de procesparameters gebruikt worden die overeenstemmen met het slechtst mogelijke geval. Aangezien deze spanning voor alle dies dezelfde is, houdt dit een verspilling van energie in. De bekomen voedingsspanning moet namelijk voldoende marge bevatten. Deze marge is echter voor het merendeel van de dies een overschatting van de effectief aanwezige degradatie van de cel-karakteristieken.

$$DRV = DRV_1 + \left[\frac{V_1}{2} + \frac{(DRV_1 - V_2) \cdot n_2}{2} \right] \tag{C.4}$$

with

$$DRV_1 = \frac{kT/q}{n_2^{-1} + n_3^{-1}} \cdot \ln \left[\left(n_3^{-1} + n_4^{-1} \right) \frac{A_4}{A_2 A_3} \left(\frac{A_5}{n_2} + \frac{A_1}{\left(n_1^{-1} + n_2^{-1} \right)^{-1}} \right) \right] \tag{C.5}$$

$$A_i = \frac{W_i}{L_i} \cdot I_0 \cdot exp \left(\frac{-V_{\mathrm{T}}}{n_i kT/q} \right)$$

$$V_1 = \frac{kT}{q} \cdot \frac{A_1 + A_5}{A_2} \cdot \exp \left(\frac{-DRV_1}{n_2 kT/q} \right)$$

$$V_2 = DRV_1 - \frac{kT}{q} \cdot \frac{A_4}{A_3} \cdot \exp \left(\frac{-DRV_1}{n_3 kT/q} \right)$$

met

DRV	: minimum data retentie spanning
W_i	: breedte van transistor i
L_i	: lengte van transistor i
V_{T}	: threshold spanning van transistor i
n_i	: subtreshold helling
I_0	: technologie constante
kT/q	: thermische spanning (25mv@300K)

Een meer optimale besparing kan bekomen worden door op die dies waar het mogelijk is de voedingsspanning verder te verlagen. Door karakterisering van de dies tijdens de testfase kan de exacte minimale dataretentiespanning bekomen worden. Om echter tijdsafhankelijke variaties op te vangen zoals temperatuur of veroudering moet een extra marge in acht genomen worden die weerom een bron is van onnodig energieverbruik. Zeer trage variaties zoals veroudering zouden eventueel nog opgevangen kunnen worden door het regelmatig laten lopen van de Built-In-Self-Test module en te herkalibreren.

Om echter de variaties die tijdsafhankelijk en verschillen van die tot die op te vangen is een online monitoring nodig van de die en het SRAM systeem. Deze monitor moet identiek reageren aan de cellen die in de matrix gebruikt worden ten overstaan van variaties in proces, temperatuur, voedingsspanning en andere tijdsafhankelijke invloeden.

Een monitor is echter niet perfect en als dusdanig ook onderhevig aan variaties. Dit heeft invloed op de effectieve faalkans van de SRAM. De kans dat alle cellen voldoen aan een minimale SNMh (SNMh$_{\mathrm{min}}$) na kalibratie aan de hand van een monitor kan geschreven worden als formule C.6. Onder de veronderstelling dat alle cellen een identieke maar statistisch onafhankelijke verdeling voor SNMh hebben, kan de opbrengst voor de matrix geschreven worden als formule C.7. De variatie op de monitor is dus van cruciaal belang voor de opbrengst van de SRAM.

$$\begin{aligned} P_{\text{matrixWerkt}} =& P[\forall i \in \text{matrix} : \text{SNMh}_i > \text{SNMh}_{\text{min}} | \text{SNMh}_{\text{mon}}] \\ =& P[\forall i \in \text{matrix} : \text{SNMh}_i > \text{SNMh}_{\text{min}}] \cdot P[\text{SNMh}_{\text{mon}}] \end{aligned} \tag{C.6}$$

$$P_{\text{matrixWerkt}} = \int_{-\infty}^{\infty} \left(1 - CDF\left(\text{SNMh}_{min}\right)\right)^N \cdot P_{\text{mon}}(x)dx \tag{C.7}$$

met

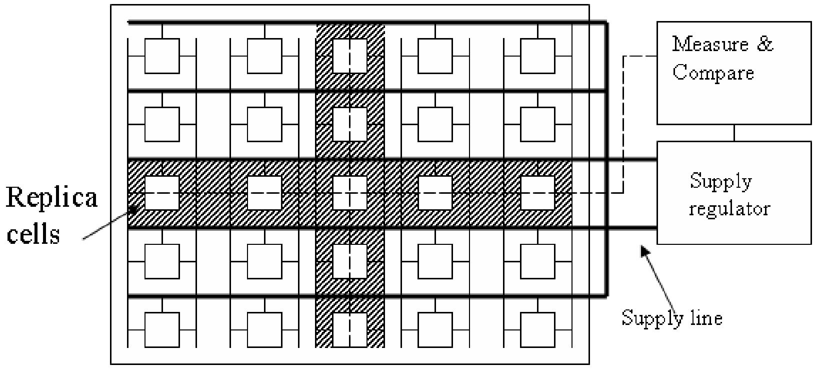
SNMh_i	: SNMh van de cel i
SNMh_{min}	: minimale vereiste SNMh
SNMh_{mon}	: SNMh gemeten op de monitor
$CDF(x)$	: cumulatieve dichtheidsfunctie voor een enkelvoudige cel
N	: aantal cellen in de matrix

De oplossing op basis van “kanarie” cellen zoals gepubliceerd in [Wan07a, Cal04] is een mogelijke implementatie van een online monitoring circuit. Hierin worden een aantal banken van cellen geplaatst die met opzet een andere, “slechtere”, lay-out en circuitontwerp hebben. Het opzet is om deze cellen te laten falen bij een voedingsspanningsreductie. Dit leidt echter tot een suboptimale besparing in lekstroom. De kanarie cellen hebben een andere omgeving en zijn anders ontworpen. Zij zullen dus ook anders reageren op veranderingen in de omgeving dan de cellen in de matrix. Dit leidt weerom tot het introduceren van een extra marge om de variabiliteit van de monitor ten overstaan van de matrix cellen te compenseren

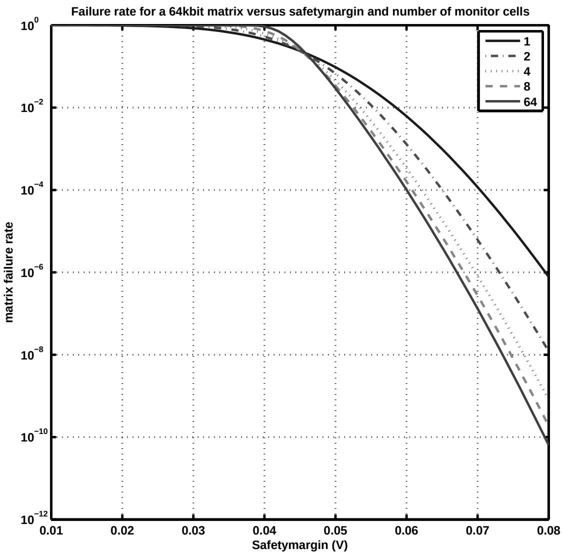
De oplossing die in dit werk wordt voorgesteld, omzeilt deze beperkingen. Door monitorcellen te gebruiken die identiek zijn in elektrisch en nagenoeg in lay-out aspecten aan de effectieve cellen van de matrix, kunnen deze in de matrix geplaatst worden. Dit heeft het voordeel dat de monitorcellen en de datacellen dezelfde omgeving hebben en identiek reageren op veranderingen. Door de monitorcellen in parallel te schakelen zoals gellustreerd in figuur C.6, kan de variatie op de opgemeten SNMh waarde beperkt worden. De variatie zal verminderen met de vierkantswortel van het aantal cellen dat in parallel geschakeld wordt [Ber51].

Deze monitor zal toestaan om enkel nog die marge in acht te nemen die nodig is voor de variatie van de SNMh waarde van de cellen binnen de matrix. Deze marge kan afgeleid worden via het opgestelde model en wordt gellustreerd in figuur C.7.

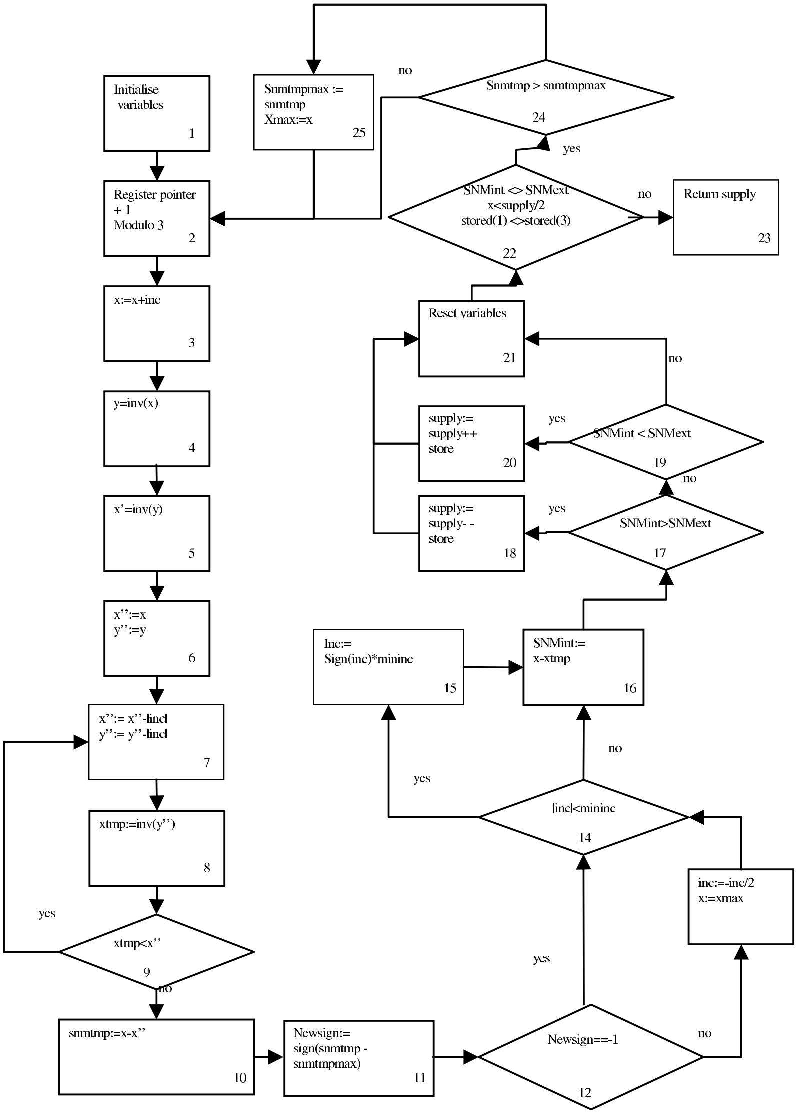
Om de SNMh van de monitor op te meten is er gekozen voor een digitale implementatie van het meetalgoritme. Het algoritme zoals voorgesteld in figuur C.8 bestaat uit drie grote delen. In een eerste deel zal een waarde op de vlinderkarakteristiek bepaald worden en het tegenoverliggende punt op de 45 lijn. De afstand tussen deze twee punten is een eerste kandidaat voor SNMh waarde. Daarna zal een volgend paar punten bepaald worden totdat via een binair zoekalgoritme het de maximale afstand en dus de SNMh bepaalt is. Deze SNMh kan dan vergelijken worden met de opgelegde minimale waarde en zal toelaten om de voedingsspanning aan te passen.



Figuur C.6: monitor organisatie



Figuur C.7: Faalkans in functie van de veiligheidsmarge



Figuur C.8: Digitale implementatie flowchart

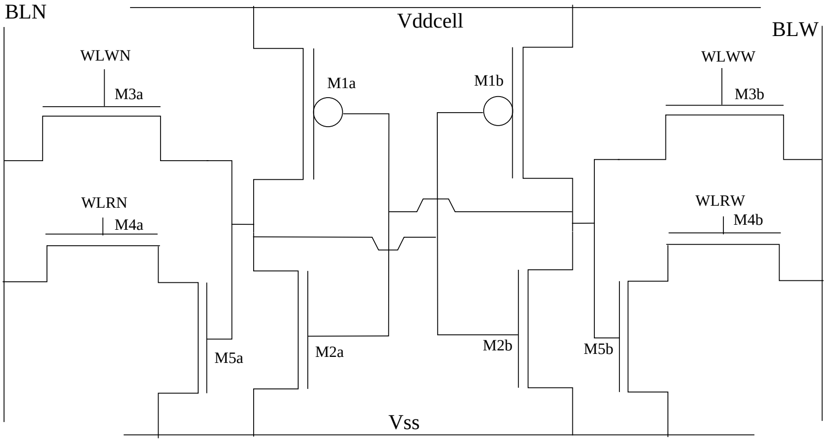
Integratie

specificatie

Om de in dit werk voorgesteld oplossingen te valideren werd een SRAM chip ontworpen in een commerciële 90nm technologie. De SRAM heeft een grootte van 64Kibit met twee toegangspoorten met een asymmetrische woordbreedte van 32bit en 256 bit. De toegang tot de data moet verzorgd worden in een enkele klokcyclus. De toegangsfrequentie voor het laag vermogen domein ligt typisch rond 500MHz of lager.

Om de overhead aan actief vermogen te verkleinen indien grote datahoeveelheden opgehaald moeten worden, wordt gebruikt gemaakt van de 256bit brede toegangspoort. Dit zal een besparing opbrengen ten overstaan van meerdere 32bit woorden op te halen. Dit vereist dat de datacel toegankelijk is via twee onafhankelijke poorten.

bouwblokken



Figuur C.9: Circuitschema van de 10T cel met dubbele toegang

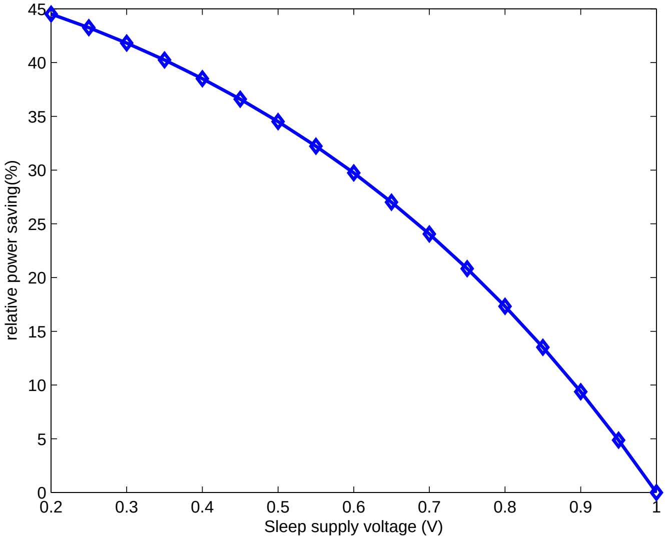
De cel die in dit werk gebruikt wordt, is schematisch weergegeven in figuur C.9. De cel bestaat uit tien transistoren die twee enkelvoudige poorten implementeren. De transistoren M1 en M2 vormen de kern van de cel om de data op te slaan. De transistoren M3 laten toe de data in de cel te schrijven, hetzij via de smalle (WLWN) of via de brede kant (WLWW). De transistoren M4, M5 vormen een leesbuffer. Door de implementatie van deze leesbuffer kan de cel gevrijwaard worden van de degenererende invloed van de bitlijnen op de opgeslagen data. Dit laat toe om de kerntransistoren M1 en M2 te schalen voor een optimale SNMh. De resulterende transistor afmetingen worden gegeven in tabel C.1

Om de tweede en lage voedingsspanning te genereren wordt gebruikt gemaakt van een lokale serie regulator op het niveau van een woord. De transistor zal geregeld worden

Transistor	Lengte (nm)	Breedte (nm)
$M1_{a,b}$	120	360
$M2_{a,b}$	160	240
$M3_{a,b}$	80	360
$M4_{a,b}$	80	240
$M5_{a,b}$	80	240

Tabel C.1: Transistor afmetingen voor de 10T DPSRAM cel

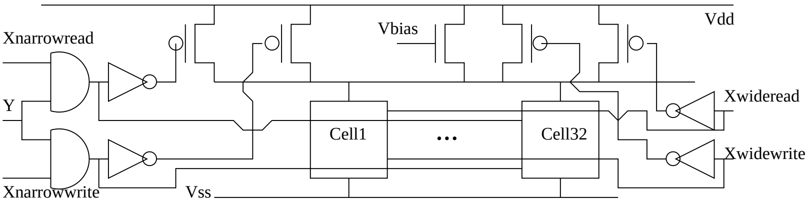
via de biasspanning om de juiste voedingsspanning te kunnen leveren. Figuur C.10 geeft de theoretisch mogelijke besparing weer die zo kan verwezenlijkt worden.



Figuur C.10: Mogelijke vermogenbesparing met een serie-regulator als functie van de voedingsspanning voor de cel

Figuur C.11 geeft schematisch de organisatie aan op het niveau van een woord. Dit niveau is teven de fijnste granulariteit die kan gebruikt worden om de toegang en sluimerende status te controleren. De vermogensschakelaars om de cellen op de nominale voeding te brengen, worden lokaal gestuurd door de decoder signalen voor een lees- of schrijfoperatie.

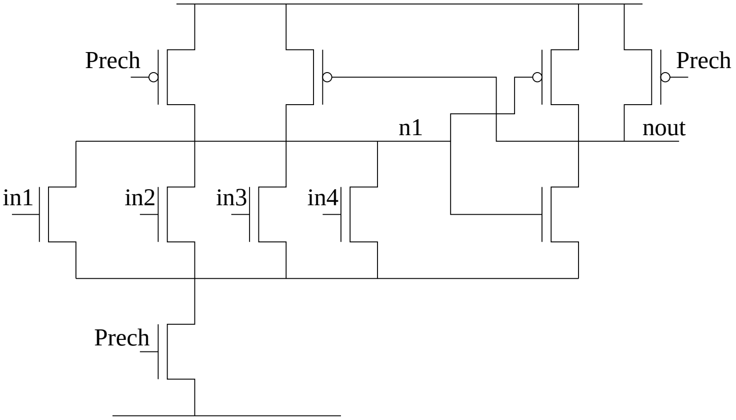
Om de laagst mogelijke voedingsspanning zonder dataverlies aan te kunnen leggen werd ook de monitorcellen gentegreerd. Hierbij moet wel vermeld worden dat het meetalgoritme niet gemplementeerd is op chip maar off-chip softwarematig draait. De



Figuur C.11: Fijnste granulaire woordstructuur met gentegreerde serie regulator

monitorcellen werden georganiseerd in 5 banken van 256bit die gelijkmatig verspreid werden over de matrix.

De decoders werden opgebouwd uit een statische standaard CMOS predecoder en een dynamische postdecoder. De dynamische postdecoder zoals voorgesteld in figuur C.12, heeft het voordeel enkel de interne nodes te ontladen bij een niet-activatie en als dusdanig een besparing op te leveren in actief vermogen voor de woordlijnbuffers [Nam98].

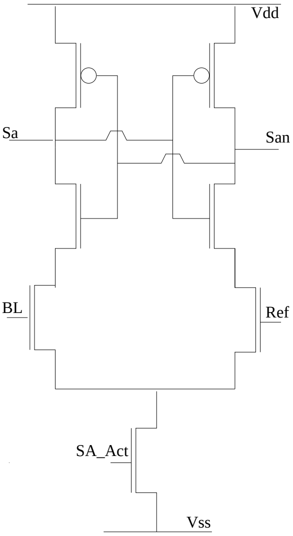


Figuur C.12: dynamisch decodercircuit zoals gebruikt in de postdecoder [Nam98]

De sense-amplifiers bestaan uit een invertor gebaseerde latch met een differentieel paar zoals voorgesteld in figuur C.13. De offsetspanning van deze SA wordt gedomineerd door het ingangspaar en bepaalt de benodigde spanningsval op de bitlijnen om een correcte uitlezing van de data te kunnen garanderen.

De timing en controle van de signalen spelen een belangrijke rol. Het overzicht van de timing van de signalen is gegeven in figuur C.14. De pijltjes in deze figuur geven de oorzakelijke verbanden aan tussen de verschillende signalen zoals gemplementeerd.

Om het activeren van de sense-amplifiers correct te laten gebeuren is een continue-

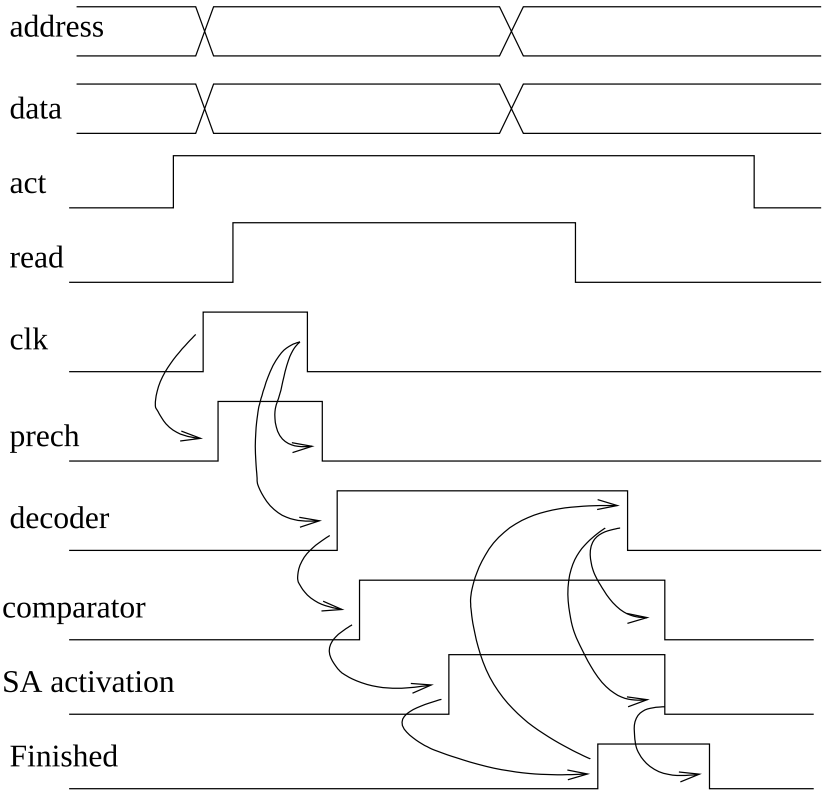


Figuur C.13: Schema van de latch type sense-amplifier met differentieel ingangspaar

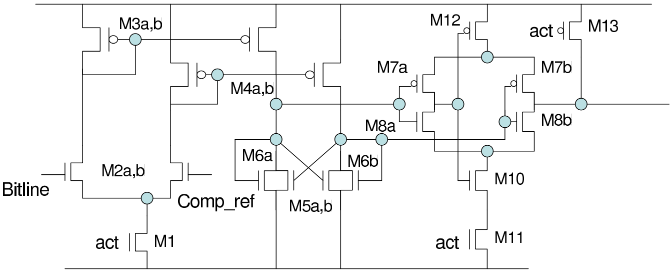
tijdscomparator gemplementeerd [All82]. Deze comparator zal op een dummy bitlijn waarvan de cellen een gekende inhoud hebben de spanningsval opmeten. Wanneer de spanning op de bitlijn voldoende gezakt is om een betrouwbare uitlezing op de sense-amplifiers te geven, zal de comparator omslaan. Dit is het signaal om de sense-amplifiers te activeren. De comparator wordt weergegeven in figuur C.15 met de bijhorende afmetingen in tabel C.2

Om de timing en de toegangstijden voor de SRAM te kunnen opmeten werd gebruik gemaakt van een gentegreerd meetsysteem. Dit systeem werkt op basis van “bevriezende” latches. Een keten van deze latches wordt voldoende lang gemaakt om een voldoende groot tijdsinterval te kunnen opmeten. Deze latches staan transparant tot het eindsignaal ze bevriest. De positie van de signalen in de latches laat dan toe om een tijdsmeting te doen, daar elke latch een vertraging heeft van ongeveer 150ps.

Figuur C.16 toont een chipfoto van de resulterende afgebonden die. De volledige core oppervlakte bedraagt 700μm op 700 μm.



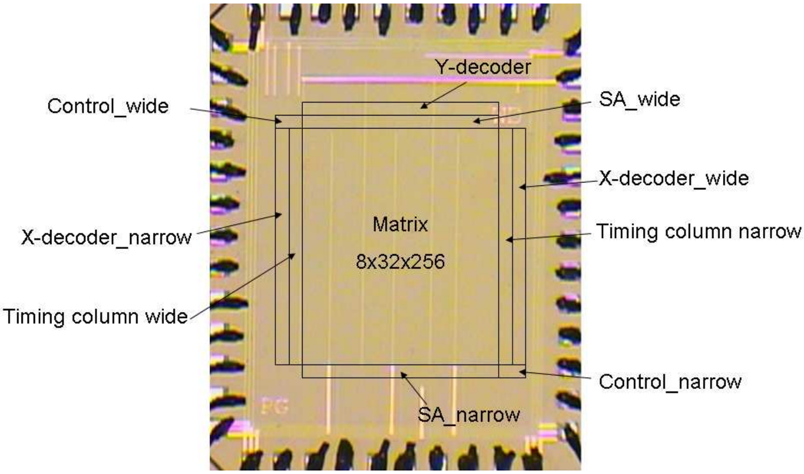
Figuur C.14: overzicht van de controlesequentie



Figuur C.15: Continue-tijdscomparator

transistor	breedte (nm)	lengte (nm)
M1	600	400
M2a,b	12000	400
M3a,b	600	400
M4a,b	600	80
M5a,b	1200	400
M6a,b	900	400
M7a,b	720	80
M8a,b	720	80
M10	720	80
M11	720	80
M12	720	80

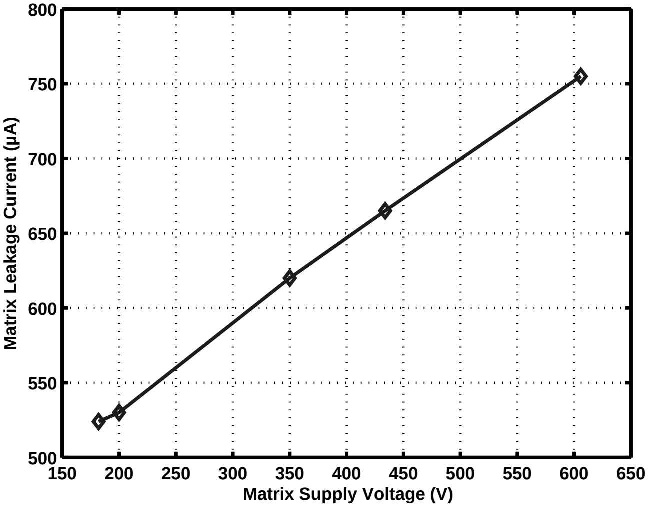
Tabel C.2: Transistor afmetingen zoals gebruikt in de comparator



Figuur C.16: chipfoto van de resulterende afgebonden die

meetresultaten

De lekstroom reductie voor de SRAM werd opgemeten voor verschillende gereduceerde voedingsspanningen zoals weergegeven in figuur C.17. De ondergrens van deze spanning werd bepaald door het falen van de dataretentie. Een voedingsspanning hoger dan 606mV is omwille van het spanningsverlies in de serie-regulator niet haalbaar. Tabel C.3 geeft ook de vastgestelde toegangstijd weer in functie van de voedingsspanning op de sluimerende cellen.



Figuur C.17: matrix lekstroom als functie van de voedingsspanning van de sluimerende cellen

mode	matrix sluimerspanning	toegangstijd	matrix lekstroom
nominaal	606mV	2ns	750μA
lage lekstroom	200 mv	2.5ns	500μA

Tabel C.3: De toegangstijd en lekstroom van de SRAM in de twee operationele modi

De meetresultaten voor de actieve energie worden weergegeven in tabel C.4. Deze data werd bekomen door het gebruik van een checkerboard (10101010...) patroon voor de data. Hieruit blijkt duidelijk dat een enkele toegang naar een 256bit woord minder kost per bit dan het uitlezen van meerdere 32bit woorden.

Het geheugen kan dus werkzaam zijn in 3 modi. De nominale modus, waarin de data bewaard wordt op een spanning van 600mV met een toegangstijd van 2ns. Een lage

toegangspoort	periferie (pJ)	precharge (pJ)	totaal (pJ)	totaal/bit (pJ)
narrow (32 bit)	13	3	16	0.5
wide(256 bit)	15	9	24	0.09

Tabel C.4: gemiddelde actieve energie per operatie

lekstroom modus, waarin de data bewaard wordt op de minimale veilige voedingsspanning 200mV met een toegangstijd van 2,5ns. Als laatste kan ook nog de zuivere retentie modus beschouwd worden. Hierin wordt de periferie van de matrix afgeschakeld, waardoor deze geen bijdrage meer kan leveren aan het totale vermogen. De cellen worden dan ook alle op de laagste veilige voedingsspanning gebracht. In deze modus wordt de data wel bewaard maar kan er geen toegang plaatsvinden zonder eerst de periferie te heractiveren. Deze heractivatie neemt echter meer dan een klokcyclus in beslag.

Tussen de nominale modus en de lage lekstroom modus is een besparing van 33% in lekstroom bereikt met een kost in toegangstijd van 25%. De retentie modus reduceert de totale lekstroom nog verder maar heeft als nadeel het niet bereikbaar zijn van de data in een enkele cyclus.

algemeen besluit

conclusie

Om het vermogenverbruik van een volledig systeem te reduceren is het nodig om het vermogen van SRAMs te reduceren daar deze een belangrijk deel uitmaken van de hedendaagse systemen. De bijdrage van de lekstromen aan het globale verbruik neemt toe met elke nieuwe technologie generatie. Het is dan ook cruciaal om deze lekstromen te minimaliseren.

De techniek die in dit werk wordt voorgesteld om dit te bereiken werkt op basis van het DIBL-effect. Door de voedingsspanning te verlagen op de cellen die niet geactiveerd zijn kan de lekstroombijdrage van de matrix gereduceerd worden. De fijne granulariteit [Gee05] waarvan in dit werk gebruikt gemaakt wordt laat toe om de controle over de activatie van cellen te reduceren tot de kleinste relevante eenheid: een enkel woord. Dit zorgt voor een minimalisatie aan lekstroom, daar een maximaal aantal cellen op een gereduceerde voedingsspanning blijft staan. Het extra energieverbruik om de cellen uit hun sluimertoestand te halen wordt op deze manier ook geminimaliseerd.

om de data integriteit te kunnen garanderen is in dit werk ook een monitoring oplossing uitgewerkt. Deze monitor in samenwerking met het ontwikkelde algoritme [Gee07] staat toe om de integriteit van de opgeslagen data te evalueren. Dit laat toe om de voedingsspanning voor de sluimerende cellen maximaal te reduceren en tevens de data integriteit te waarborgen. De ondergrens voor een betrouwbare opslag werd vastgesteld op 200mV met behulp van dit algoritme voor het ontwerp gemplementeerd in een commerciële 90nm technologie.

Deze implementatie toont ook aan dat een grote woordlengte per bit een lagere energie kost heeft dan het uitlezen van meerdere kleine woorden. In het geval van de implementatie in dit werk heeft de 256bit toegang een kost van 0.09pJ/bit, terwijl dit voor de 32bit toegang 0.5pJ/bit bedraagt. De bemerking dient hier evenwel gemaakt te worden dat de mogelijke woordbreedte beperkt wordt door een aantal fysische implementatie beperkingen, zoals de maximale aspectratio en de mogelijkheid om brede woorden op systeemniveau te verwerken [Gee08].

In dit werk werd een systeem en de nodige achtergrond gepresenteerd om een SRAM te creëren waarin de lekstromen geminimaliseerd kunnen worden en tevens de data integriteit kan gegarandeerd worden. De implementatie van de SRAM met asymmetrische breedte in toegangspoorten illustreert ook de besparing in actief energieverbruik per bit die kan gerealiseerd worden door bredere datawoorden te gebruiken.

uitbreidingen naar de toekomst

Het gebruik van een regelaar in serie zorgt voor een beperking in energiereductie. Een DC-DC convertor met een hoge efficiëntie kan hier een extra vermogenbesparing betekenen. Het ontwikkelen van deze DC-DC convertor is de eerste logische uitbreiding van dit werk. De uitdaging bestaat hier uit twee grote delen. Ten eerste, de oppervlakte gebruikt door een volledig gentegreerde DC-DC convertor mag geen te grote overhead betekenen voor het volledige systeem. Ten tweede, het verbruikte schakelvermogen van deze DC-DC convertor mag ook het vermogen van de SRAM niet overheersen. Indien een dergelijk DC-DC convertor gebruikt kan worden voor andere bouwblokken van het systeem kan een deel van deze specificatie verlicht worden.

Een tweede uitbreiding voor dit werk, is de verdere studie van de toegang met asymmetrische woordbreedte voor SRAMs. De interactie van de SRAM met de rest van het systeem is hierbij cruciaal. Hiervoor is het nodig bewust te zijn van de invloed van bouwblokken doorheen de abstractie niveaus. Dit inzicht zal leiden tot nieuwe methodologieën voor het energie-efficiënte ontwerp van elektronische systemen.

De reductie van de voedingsspanning, niet alleen omwille van de reductie in lekstroom, kan nog verder onderzocht worden. Het belang van mobiele applicaties neemt enkel maar toe. Een reductie van de voedingsspanning ook voor de actieve delen van een systeem kan het vermogenverbruik nog verder terugdringen. Het ontwerp van methodologieën, architecturen en systemen die erin slagen om dit op een betrouwbare manier te realiseren biedt een grote uitdaging en mogelijk grote opbrengsten.

Publications

International reviewed journals

1. Geens P. , Dehaene W., “A small granular controlled leakage current reduction system for SRAMs”, in *Journal on Solid-State Electronics*, vol. 49, pp. 1776-1782, November, 2005

International Conferences

1. Chen T., Geens P., Van der Plas G., Dehaene W., Gielen G., “A 14-bit 130MHz CMOS Current-Steering DAC with Adjustable INL”, in *proceedings of European Solid-State Circuits Conference*, pp. 167-170, September 21-23, 2004
2. Geens P. , Dehaene W., “A small granular controlled leakage current reduction system for SRAM”, in *proceedings of International Conference on Memory Technology and Design (ICMTD)*, May 21-25, 2005
3. Dehaene W., Cosemans S., Vignon A., Catthoor F., Geens P. , “Embedded SRAM design in deep deep submicron technologies”, in *proceedings of 33rd European Solid-State Circuits Conference (ESSCIRC)*, 2007
4. Geens P. , Dehaene W., “A Noise-Margin Monitor for SRAMs”, in *proceedings of International Conference on Memory Technology and Design*, May 7-10, 2007
5. Hua Wang, Miranda M., Geens P., Dehaene W., Catthoor, F., “A Variability Tolerant Embedded SRAM Offering Runtime Selectable Energy/Delay Figures”, in *proceedings of International Conference on Memory Technology and Design (ICMTD)*, May 7-10, 2007
6. Geens P. , Dehaene W., “A Dual Port Dual Width 90nm SRAM with Guaranteed Data Retention at Minimal Standby Supply Voltage”, in *proceedings of 34th European Solid-State Circuits Conference (ESSCIRC)*, pp. 290-293, September 15-19, 2008

Patents

1. EP1953762, “Memory device with reduced stand-by power and method for operating same”, inventors: P. Geens, W. Dehaene.
2. US 12/019,699 , “Memory device with reduced stand-by power and method for operating same”, inventors: P. Geens, W. Dehaene.

Bibliography

[All82] D.J. Allstot, "A precision variable-supply cmos comparator", *Solid-State Circuits, IEEE Journal of*, 17(6):1080–1087, Dec 1982.

[Ame08] B. Amelifard, F. Fallah, and M. Pedram, "Leakage minimization of sram cells in a dual-v_t and dual-t_{rmox} technology", *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 16(7):851–860, July 2008.

[Amr98] 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):1208–1219, Aug 1998.

[Ber51] J. Bernouilli, *Ars Conjuctandi*, 1751.

[Cal04] B.H. Calhoun and A.P. Chandrakasan, "Standby power reduction using dynamic voltage scaling and canary flip-flop structures", *Solid-State Circuits, IEEE Journal of*, 39(9):1504–1511, Sept. 2004.

[Cal06] B. H. Calhoun and A. Chandrakasan, "Static Noise Margin Variation for Subthreshold SRAM in 65-nm CMOS", *IEEE Journal of Solid-State Circuits*, 41(7):1673–1679, July 2006.

[Car04] I. Carlson, S. Andersson, S. Natarajan, and A. Alvandpour, "A high density, low leakage, 5t sram for embedded caches", *Solid-State Circuits Conference, 2004. ESSCIRC 2004. Proceeding of the 30th European*, pages 215–218, Sept. 2004.

[Cha01] A. Chandrakasan, W. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE press, 2001.

[Cos07] S. Cosemans, W. Dehaene, and F. Catthoor, "A low-power embedded sram for wireless applications", *Solid-State Circuits, IEEE Journal of*, 42(7):1607–1617, July 2007.

[Cos08] Stefan Cosemans, Wim Dehaene, and Francky Catthoor, "A 3.6pj/access 480mhz, 128kbit on-chip sram with 850mhz boost mode in 90nm cmos with tunable sense amplifiers to cope with variability", *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, pages 278–281, Sept. 2008.

[Deh07] W. Dehaene, S. Cosemans, A. Vignon, F. Catthoor, and P. Geens, "Embedded sram design in deep deep submicron technologies", *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, pages 384–391, Sept. 2007.

[Der74] R. S. Dernard et al., "Design of ion-implanted MOSFETs with very small physical dimensions", *IEEE Journal of Solid-State Circuits*, SC-9, 1974.

[Dra04] M. Drazdziulis, P. Larson-Edefors, D. Eckerbert, and H Eriksson, "A power Cut-Off Technique for Gate Leakage Suppression", *Proceedings IEEE European Solid-State Circuits Conference*, pages 171–174, 2004.

[Fla02] K. Flautner, Nam Sung Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy caches: simple techniques for reducing leakage power", *Computer Architecture, 2002. Proceedings. 29th Annual International Symposium on*, pages 148–157, 2002.

[Gee05] P. Geens and W. Dehaene, "A Small Granular Controlled Leakage Reduction System for SRAMs", *Journal of Solid State Electronics*, 49:1776–1782, November 2005.

[Gee07] P. Geens and W. Dehaene, "A Noise Margin Monitor for SRAMs", *Proceedings International Conference on Memory Technology and Design*, pages 169–172, 2007.

[Gee08] P. Geens and W. Dehaene, "A Dual Supply Dual Port SRAM in 90nm with guaranteed data retention", *Proceedings IEEE European Solid-State Circuits Conference*, pages 290–293, 2008.

[Gro06] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, "Read stability and writeability analysis of sram cells for nanometer technologies", *Solid-State Circuits, IEEE Journal of*, 41(11):2577–2588, Nov. 2006.

[Ham08] F. Hamzaoglu, Kevin Zhang, Yin Wang, H.J. Ann, U. Bhattacharya, Zhanping Chen, Yong-Gee Ng, A. Pavlov, K. Smits, and M. Bohr, "A 153mb-sram design with dynamic stability enhancement and leakage reduction in 45nm high-κ metal-gate cmos technology", *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, pages 376–621, Feb. 2008.

[Hir90] T. Hirose, H. Kuriyama, S. Murakami, K. Yuzuriha, T. Mukai, K. Tsutsumi, Y. Nishimura, Y. Kohno, and K. Anami, "A 20-ns 4-mb cmos sram with hierarchical word decoding architecture", *Solid-State Circuits, IEEE Journal of*, 25(5):1068–1074, Oct 1990.

[ITR] ITRS, http://www.itrs.net, ITRS roadmap.

[Kaw00] H. Kawaguchi, K. Nose, and T. Sakurai, "A super cut-off cmos (sccmos) scheme for 0.5-v supply voltage with picoampere stand-by current", *Solid-State Circuits, IEEE Journal of*, 35(10):1498–1501, Oct 2000.

[Kim03] N.S. Kim, T. Austin, D. Blaauw, T. Mudge, K. Flautner, J.S. Hu, M.J. Irwin, M. Kandemir, and V. Narayanan, "Leakage current: Moore's law meets static power", *Computer*, 36(12):68–75, Dec. 2003.

[Kim06] C.H. Kim, Jae-Joon Kim, Ik-Joon Chang, and K. Roy, “Pvt-aware leakage reduction for on-die caches with improved read stability”, *Solid-State Circuits, IEEE Journal of*, 41(1):170–178, Jan. 2006.

[Kob93] T. Kobayashi, K. Nogami, T. Shirotori, and Y. Fujimoto, “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):523–527, Apr 1993.

[Kua05] J.B. Kuang, H.C. Ngo, K.J. Nowka, J.C. Law, and R.V. Joshi, “A low-overhead virtual rail technique for sram leakage power reduction”, *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, pages 574–579, Oct. 2005.

[Kwo08] J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann, and A. Chandrakasan, “A 65nm sub-vt microcontroller with integrated sram and switched-capacitor dc-dc converter”, *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, pages 318–616, Feb. 2008.

[Lak94] Laker,K. and Sansen,W. , *Design of Analog Circuits and Systems*, McGraw-Hill, New York, 1994.

[Man03] R. W. Mann et al., “Ultralow-power SRAM technology”, *IBM Journal of Research and Development*, 47(5/6), September/November 2003.

[Mis07] K. Mistry et al., “A 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 cu interconnect layers, 193nm dry patterning, and 100pb-free packaging”, *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pages 247–250, Dec. 2007.

[Moo69] G.E. Moore, “Trends in silicon device technology”, *Electron Devices, IEEE Transactions on*, 16(2):234–234, Feb 1969.

[Mut95] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, “1-v power supply high-speed digital circuit technology with multithreshold-voltage cmos”, *Solid-State Circuits, IEEE Journal of*, 30(8):847–854, Aug 1995.

[Nam98] H. Nambu, K. Kanetani, K. Yamasaki, K. Higeta, M. Usami, T. Kusunoki, K. Yamaguchi, and N. Homma, “A 1.8 ns access, 550 mhz 4.5 mb cmos sram”, *Solid-State Circuits Conference, 1998. Digest of Technical Papers. 1998 IEEE International*, pages 360–361, 464, Feb 1998.

[Nar06] Narendra, Siva G. and Chandrakasan, A., *Leakage in Nanometer CMOS Technologies*, Springer, New York, 2006.

[Nii04] Nii, K. et al., “A 90-nm low-power 32-KB embedded SRAM with gate leakage suppression circuit for mobile applications”, *IEEE Journal of Solid-State Circuits*, 39(4):684–693, April 2004.

[Pel89] M.J.M. Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers, “Matching properties of mos transistors”, *Solid-State Circuits, IEEE Journal of*, 24(5):1433–1439, Oct 1989.

[Qin04] Hulfang Qin, Yu Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, “Sram leakage suppression by minimizing standby supply voltage”, *Quality Electronic Design, 2004. Proceedings. 5th International Symposium on*, pages 55–60, 2004.

[Rab03] Rabaey, J., Chandrakasan, A., and Nikolic, B., *Digital Intergate Circuits, 2nd Edition*, Prentice Hall, New Jersey, 2003.

[Rag07] P. Raghavan, A. Lambrechts, M. Jayapala, F. Catthoor, D. Verkest, and H. Corporaal, “Very wide register: An asymmetric register file organization for low power embedded processors”, *Design, Automation & Test in Europe Conference & Exhibition, 2007. DATE ’07*, pages 1–6, April 2007.

[Roy03] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, “Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits”, *Proceedings of the IEEE*, 91(2), February 2003.

[Sal05] F.R. Saliba, H. Kawaguchi, and T. Sakurai, “Experimental verification of row-by-row variable vdd scheme reducing 95 *VLSI Circuits, 2005. Digest of Technical Papers. 2005 Symposium on*, pages 162–165, June 2005.

[See87] Seevinck, E., List, F.J., and Lohstroh, J., “Static-noise margin analysis of MOS SRAM cells”, *IEEE Journal of Solid-State Circuits*, 22(5):748–754, October 1987.

[Tay98] Y. Tayr and N. H. Sing, *Fundamentals of Modern VLSI devices*, Cambridge Univ. Press, New York, 1998.

[vA04] K. von Armin, Borinski E., P. Seegebrecht, H. Fiedler, R. Brederlow, R. Thewes, J. Berthold, and C. Pacha, “Efficiency of Body Biasing in 90 nm CMOS for Low Power Digital Circuits”, *Proceedings IEEE European Solid-State Circuits Conference*, pages 175–178, 2004.

[Vas06] A. Vassighi and M. Sachdev, “Thermal runaway in integrated circuits”, *Device and Materials Reliability, IEEE Transactions on*, 6(2):300–305, June 2006.

[Ver07] Naveen Verma and A.P. Chandrakasan, “A 65nm 8t sub-vt sram employing sense-amplifier redundancy”, *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pages 328–606, Feb. 2007.

[Wan05] C. Wann, “SRAM cell design for Stability Methodology”, *IEEE VLSI-TSA International Symposium on VLSI Technology*, pages 21–22, April 2005.

[Wan07a] J. Wang and B. H. Calhoun, "Canary Replica Feedback for Near-DRV Standby VDD Scaling in a 90nm SRAM", *Proceedings Custom Integrated Circuits Conference*, 2007.

[Wan07b] Jiajing Wang, A. Singhee, R.A. Rutenbar, and B.H. Calhoun, "Statistical modeling for the minimum standby supply voltage of a full sram array", *33rd European Solid State Circuits Conference, 2007. ESSCIRC*, pages 400–403, Sept. 2007.

[Wan08] Yih Wang, Hong Jo Ahn, U. Bhattacharya, Zhanping Chen, T. Coan, F. Hamzaoglu, W.M. Hafez, Chia-Hong Jan, P. Kolar, S.H. Kulkarni, Jie-Feng Lin, Yong-Gee Ng, I. Post, Liqiong Wei, Ying Zhang, K. Zhang, and M. Bohr, "A 1.1 ghz 12 μa/mb-leakage sram design in 65 nm ultra-low-power cmos technology with integrated leakage reduction for mobile applications", *Solid-State Circuits, IEEE Journal of*, 43(1):172–179, Jan. 2008.

[Ye03] Y. Ye, M. Khellah, D. Somasekhar, A. Farhang, and V. De, "A 6-ghz 16-kb l1 cache in a 100-nm dual-vt technology using a bitline leakage reduction (blr) technique", *Solid-State Circuits, IEEE Journal of*, 38(5):839–842, May 2003.

[Zha05] K. Zhang et al., "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction", *IEEE Journal of Solid-State Circuits*, 40(4):895–901, April 2005.

biography

Peter Geens was born in Leuven, Belgium, in 1977. He received the M. Sc. degree of Electrical Engineering (Burgelijk Ingenieur), specialisation ICT:micro-electronics in 2002 from the Katholieke Universiteit Leuven. His masterthesis subject was on the design of a data recovery path for a SOPA line driver.

In 2002 he joined ESAT-MICAS as a research assistant. His research and field of interest went out to low power SRAM design.