SRAM POWER REDUCTION – AN ULTRA-LOW-POWER SRAM ARCHITECTURE IN 45nm TECHNOLOGY

June 23, 2009

Author: Khawar Sarfraz

Supervisors: Dr. Nick P. van der Meijs (Delft University of Technology, Delft, the Netherlands)
Toby S. Doorn (NXP Semiconductors, Eindhoven, the Netherlands)
Roelof H. W. Salters (NXP Semiconductors, Eindhoven, the Netherlands)

SRAM POWER REDUCTION – AN ULTRA-LOW-POWER SRAM ARCHITECTURE IN 45nm TECHNOLOGY

THESIS

submitted in partial fulfilment of the
requirements for the degree of

MASTER OF SCIENCE

in

ELECTRICAL ENGINEERING
(Specialization: Microelectronics)

by

Khawar Sarfraz

Circuits & Systems Group
Faculty of Electrical Engineering, Mathematics & Computer Science
Delft University of Technology, Delft, the Netherlands

SRAM POWER REDUCTION – AN ULTRA-LOW-POWER SRAM ARCHITECTURE IN 45nm TECHNOLOGY

MSc Thesis Committee

Group: Circuits & Systems

Identifier: CAS-MS-2009-06

Chair: Prof. Dr. Edoardo Charbon

Advisor: Dr. Nick P. van der Meijs

Advisor: Toby S. Doorn

Advisor: Roelof H. W. Salters

Member: Dr. Said Hamdioui

DEDICATION

In loving memory of my father, Sarfraz Ahmad, who couldn't live to see me graduate

To my mother, Nighat Sarfraz, for her unending prayers and love

To my brother, Sohab Sarfraz, for always being there for me

And

To my wife, Hina Khawar, for her patience and understanding

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis has come into existence as a consequence of an eleven-month research carried out at NXP Semiconductors, the Netherlands. The work was undertaken in partial fulfilment of the requirements for the degree of Master of Science at the Circuits and Systems Laboratory, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, the Netherlands.

I would like to start by thanking Dr. Nick P. van der Meijs, my university advisor, for developing my interest in the discipline of Digital Integrated Circuits. I am extremely thankful to him for his guidance, moral support and for providing me with a fantastic opportunity to conduct my thesis at NXP Semiconductors. I would also like to thank Toby Doorn, my daily supervisor at NXP Semiconductors, for his fine supervision, meticulous reviews, numerous fruitful discussions and good guidance over the entire duration of the project.

In addition, I am grateful to Roelof Salters, my second supervisor at NXP Semiconductors, for his invaluable input in design reviews and precious feedback during our weekly meetings. Many thanks are also due to Patrick van de Steeg for his advice, participation and suggestions in progress meetings and propositions on design improvement.

I would also like to express my sincere thanks to Delft University of Technology and NXP Semiconductors for providing me with nearly full scholarship for my Master of Science study, without which perhaps I wouldn't have been here to start with.

Moreover, I am greatly appreciative of my family's continual support during my stay out of Pakistan. Most importantly, I am exceedingly thankful to Allah (Almighty) for providing me with the strength, determination and resolve to complete this work.

Khawar Sarfraz
Eindhoven, the Netherlands
June, 2009

ABBREVIATIONS

Chapter – 1

CMOS	Complementary Metal Oxide Semiconductor
DBL	Divided Bit Line
DRAM	Dynamic Random Access Memory
DWL	Divided Word Line
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
SA	Sense Amplifier
SNM	Static Noise Margin
SoC	System on a chip
SOI	Silicon on Insulator
SRAM	Static Random Access Memory

Chapter – 2

NCI	Non Cell Implant
TSMC	Taiwan Semiconductor Manufacturing Company

Chapter – 3

ACTIVE mode	Time during which a memory read/write operation takes place
DRV	Data retention voltage
GIDL	Gate Induced Drain Leakage
NMOS	N-channel MOSFET
PMOS	P-channel MOSFET
SRC	The source node of the pull down transistors in a memory cell
STANDBY mode	Time during which no memory read/write operation takes place (time during which DRV is maintained across the SRAM cell)
SUP	The source node of the pull up transistors in a memory cell
V_{SRC}	The voltage at the SRC node
V_{SUP}	The voltage at the SUP node

Chapter – 4

H-V_T	High threshold-voltage logic transistor
L-V_T	Low threshold-voltage logic transistor
S-V_T	Standard threshold-voltage logic transistor
SRAM-PD	The NMOS pull down transistor in the SRAM cell

Chapter – 5

SRAM-PU	The PMOS pull up transistor in the SRAM cell

Chapter – 7

DCT	Diode-connected transistor

Chapter – 8

CSBA	Cyclic Single Block Access
SQBA	Sequential Block Access

Appendix – B

BL/BLb	Bit line pair
FSM	Finite State Machine
CR	Cell ratio
PR	Pull up ratio

Appendix – C

DIBL	Drain Induced Barrier Lowering

ABSTRACT

Static Random Access Memory (SRAM) constitutes the cornerstone of data storage in most SoCs employed in hand-held devices that we find on the market today. Products that rely heavily on the use of SoCs are notebooks, digital cameras, cell phones, high-end audio-visual equipment and navigation systems to name a few. SRAMs help reduce the gap between the fast processors and the slow DRAMs. Since SRAMs comprise a large fraction of a modern SoC's total area budget, they also contribute substantially to the total power budget of the chip. The challenges associated with SRAM design have increased manifolds with the shrinking of technology. SRAMs tend to become leakier with reduction in device threshold voltage, which leads to high standby power consumption.

Numerous prominent publications have appeared over the recent years targeted at SRAM power reduction in an effort to reduce the overall power consumption of the chip. The work presented in this thesis, conducted at NXP Semiconductors, is on similar lines. A *1.1V* ultra-low-power embedded SRAM architecture is investigated in TSMC 45nm technology that employs source biasing leakage reduction technique to reduce both the active as well as standby power consumption of the memory. The proposed architecture has been shown to be robust over a temperature range of $-40°C \leq T \leq +125°C$ as well as over *10%* variation in supply voltage. The idea is to place a large fraction of the non-accessed memory cells in standby so that their leakage could be reduced. The memory instance is therefore divided into equal-sized blocks, where each block contains *4k* cells and *2* additional transistors: a switch and a diode-connected transistor. Only one block is active during any read/write cycle. The non-accessed memory blocks maintain data retention voltage across their terminals, hence they leak less (compared to maintaining V_{DD} across their terminals).

The proposed solution achieves a maximum reduction in block leakage current of *64.5%* with an area overhead of *6.6%*. The maximum memory operating frequency is *740MHz*. A memory instance of *4Mb* achieves a maximum of *30.7mW* of power savings at *500MHz*, compared to using no power saving technique.

Chapter 1

Introduction to low-power SRAM design

1.1 Relevance of low-power design

The significance of low-power design was realized when designers were first faced with challenges of portability, packaging, high data throughput [1], higher operating frequencies and temperature control, all the while maintaining focus on circuit reliability. The importance of low-power design was also brought into limelight as a consequence of mediocre advances in battery technology [2, 3].

The past two decades have seen an unprecedented surge in efforts geared towards achieving reliable circuit operation while consuming little power. The widespread deployment of wireless communication systems further fuelled research activity in the field of low-power electronic design, which resulted in numerous related patents and inventions getting filed. Even today, low-power system architectures and associated circuit design techniques are one of the most researched and well-funded areas in the field of integrated circuit (IC) design.

Notebooks, digital cameras, high-end audio-visual equipment, navigation systems and cellular telephones are some of the notable products that have emerged as a consequence of advances in the field of low-power electronic design and technology.

At the heart of these products are complex system-on-chip (SoC) architectures that rely heavily on the presence of large on-chip data storage resources (SRAM or cache memories) in order to function effectively. Cache memories help reduce the gap between the fast processors and the slow DRAMs. Since SRAMs comprise a large fraction of a modern SoC's total area budget, they also contribute substantially towards the total power budget of the chip. Numerous prominent publications have appeared over the recent years targeted at SRAM power reduction in an effort to reduce the overall power consumption of the chip. The work presented in this thesis is also on the same lines. An ultra-low-power SRAM architecture is investigated that helps reduce both the active and standby power of the SRAM.

The aim of this chapter is to highlight some of the design techniques prevalent in the industry that help achieve low-power operation of the SRAM by providing references to some of the notable works published to date. Also presented is an overview of how this thesis is organized.

1.2 Low-power SRAM design techniques

The total power budget of a typical SRAM comprises both the active and the standby power [4], and a ton of literature has been published in the past two decades on effective reduction methods of both these contributors.

1.2.1 Active power consumption

The sources of active power consumption are as follows:

i) The memory matrix
ii) The entire SRAM peripheral circuitry, comprising all building blocks identified in Appendix B

The following are some of the key active power reduction techniques in SRAMs:

i) DWL and HWD approach
ii) Pulsed word line and bit line design
iii) Divided/hierarchical bit line (DBL) approach
iv) Low-voltage memory matrix
v) Dual-V_{DD} design

1.2.1.1 DWL and HWD approach

The increasing length of the word line, leading to an increased capacitance and speed degradation [4], poses a serious concern for memory designers. Divided Word Line (DWL) [5] and Hierarchical Word Decoding (HWD) [6] considerably assist in mitigating this problem. They are two of the most common multi-stage decoding schemes, employed in large decoders. In the DWL approach (Figure 1.1), a block word line is selected when both the block select and the global word line signals are active. This reduces the power consumption on the word line due to smaller capacitance being charged and discharged, and hence the delay. For large SRAM instances (over *4Mb*), the HWD scheme is more effective [7]. The HWD (Figure 1.2) differs from the DWL scheme in that it incorporates an additional level of decoding by employing an intermediate sub-global word line. The hierarchical word line architecture is therefore able to select the minimum required number of cells in a row thus significantly reducing parasitic column currents. In addition, the total capacitance associated with each smaller word line is considerably reduced, which lowers the memory access time and power consumption.

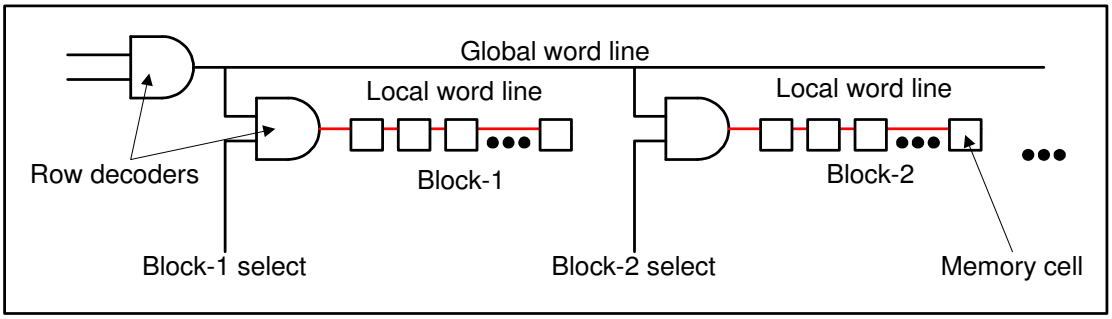


Figure 1.1 Divided Word Line (DWL) approach (re-drawn from [7])

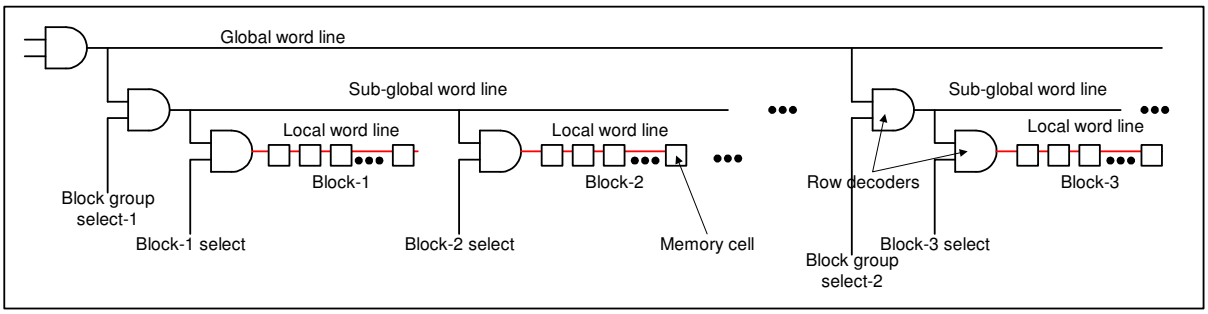


Figure 1.2 Hierarchical Word Decoding (HWD) approach (re-drawn from [7])

1.2.1.2 Pulsed word line and pulsed bit line design

Significant power savings can be achieved by reducing the active duty cycle of a word line that is by pulsing it high for a short duration during a read/write operation [8]. Pulsed operation can also be applied to bit lines with a positive swing from $V_{DD}/2$ to V_{DD} and back to $V_{DD}/2$ and a negative swing from $V_{DD}/2$ to *GND* and back to $V_{DD}/2$ [9]. This approach however degrades the static noise margin of the cell. In [10], pulsed word line and pulsed bit line schemes are used in combination with a read-modify-write approach with the local SA being used as a write driver. The scheme provides *26X* improvement in cell stability with *8%* area overhead. The expected penalty in speed for the proposed design has not been discussed in [10].

1.2.1.3 Divided bit line (DBL) approach

Bit line capacitance can be reduced via a DBL approach as proposed in [11], where two or more SRAM cells are combined together to divide the bit line into several sub bit lines (Figure 1.3). The sub bit lines can again be combined to form more levels of hierarchy.

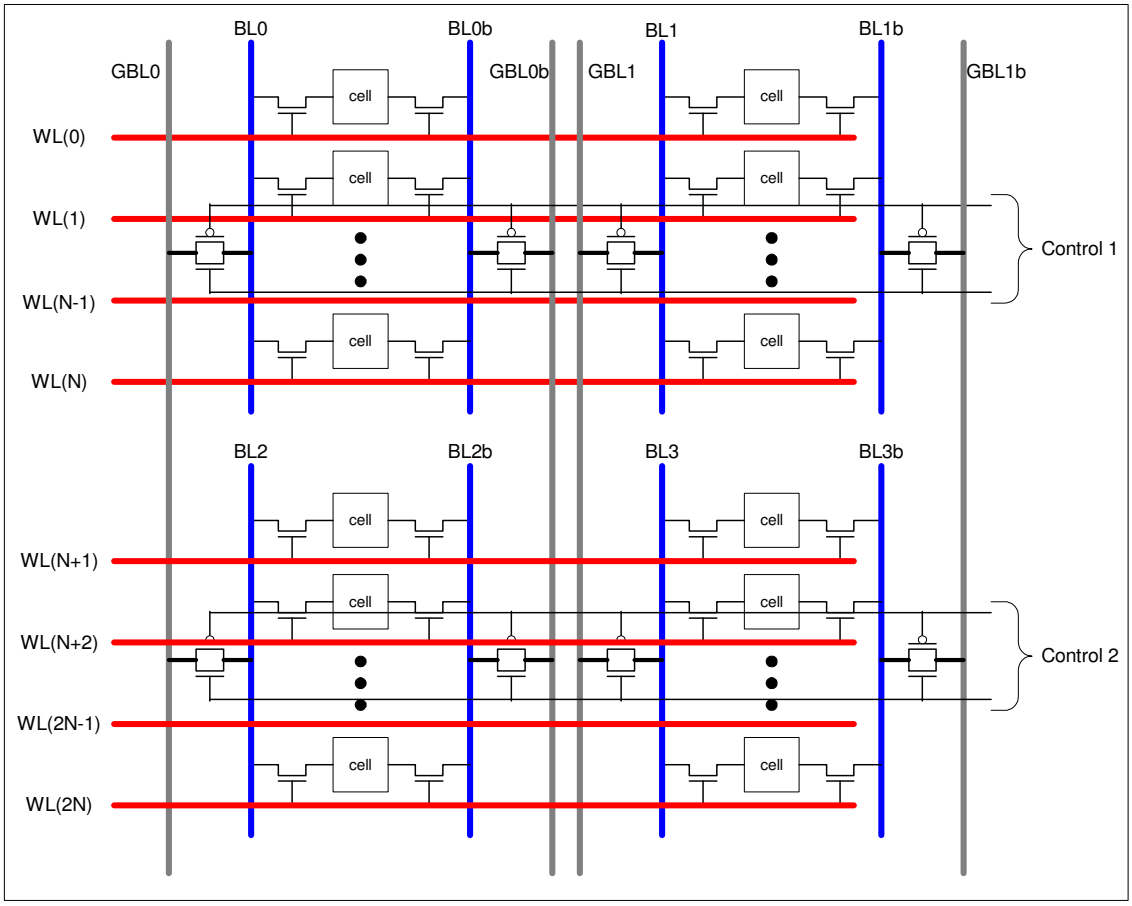


Figure 1.3 Divided Bit Line (DBL) approach

The proposed architecture results in reduction of bit line capacitance, which reduces the access time and power consumption. Active power savings of *50%-60%* have been reported in addition to access time reduction by *20%* at the cost of *5%* increase in transistor count [11]. A dual sense amplified bit line architecture is presented in [12] which achieves *88%* reduction in active power at a cost of *49%* increase in area. In the suggested architecture, a group of cells in a column share a local bit line pair as well as a local SA that works both during read and write operations. The local bit line pairs are connected to a global bit line pair via PMOS pass gates.

1.2.1.4 Low-voltage memory matrix

The contribution of the dynamic and short-circuit power of memory matrix decreases primarily due to reduction in supply voltage level and SRAM cell capacitance with the introduction of each subsequent technology generation. The leakage current of the matrix does not contribute significantly to the active power dissipation of the SRAM for high-V_T transistors.

1.2.1.5 Dual-V_{DD} design

Low-voltage peripheral circuit design promises significant power savings for embedded SRAMs. Typically dual-V_{DD} memory architectures allow ultra-low-voltage operation of the periphery that interfaces with the matrix (running on higher supply voltage) using efficient level shifting circuits. High performance can be guaranteed without significant area overhead when the difference in the two voltage rails is small. The drawbacks of this scheme are higher levels of leakage of the memory matrix and increased chip complexity.

1.2.2 Standby power consumption

The sources of standby power consumption are as follows:

i) Leakage of the memory matrix (primarily)
ii) Leakage of the periphery in single-V_{DD} designs

In standby mode, that is when no read/write operation takes place, the dense memory matrix in most modern SoCs contributes heavily to the standby power rating of a chip due to high leakage currents. For single-V_{DD} designs, the leakage of the periphery also appreciably contributes towards the total memory standby current. Most of the literature published to date that focuses on the reduction of standby power in SRAMs is targeted towards efficient cell architecture design, control of cell leakage current paths, MOSFET threshold voltage adjustment, cell supply voltage control, the use of alternate technology (e.g. silicon on insulator - SOI) and the possibility of using high-*k* gate dielectrics. References to some of the notable publications along with the gains and trade-offs associated with these techniques are presented here.

1.2.2.1 Dynamic control of V_B

An effective approach to reduce leakage current is to dynamically control MOSFET substrate bias [13]. This technique allows the critical SRAM transistors to have $V_{SB}=0$ zero during a read/write operation and $|V_{SB}|>0$ during standby mode. The reduction in leakage is achieved at the cost of increased chip area.

1.2.2.2 Re-designing the SRAM cell

A low-leakage 8-T single-ended SRAM cell is proposed in [14] that achieves *60%* reduction in total leakage at high temperatures and an improvement of *2.2X* in Static Noise Margin (SNM) during reading at the cost of *33%* increase in memory matrix area. The two added transistors help reduce gate leakage current at reduced temperatures and improve SNM during reading.

1.2.2.3 Different types of SRAM cells in the matrix

In [15], a scheme based on dual-V_T and dual-t_{ox} assignment to SRAM cells is proposed that can reduce leakage power dissipation of a *32*x*512* instance by *40%*. The idea is based on the fact that delays in a memory cell depend on the physical distance of the cell from the SA and the row decoder. The approach therefore recommends the use of two different types of SRAM cells in the matrix. The downside of the approach is the associated change to the SRAM design flow and increased fabrication costs.

1.2.2.4 Dual-V_T SRAM cell

SRAM cells containing dual-V_T transistors have also been investigated in an attempt to reduce leakage currents. A 6T cell is presented in [16], which employs minimum-length high-V_T pass gates. A modified precharge scheme is also proposed in which only those columns that are to be read from are precharged at the beginning of the read cycle. Leakage currents are reduced by a factor of *1.9* at an access time penalty of *7.4%* and an area overhead of *2.4%* at *T=110°C*.

1.2.2.5 Leakage of periphery in single-V_{DD} designs

For embedded SRAMs running on a single voltage supply, the leakage of the peripheral circuits can become an appreciable fraction of the total memory standby current. A novel approach is presented in [17], where a statistical analysis of the active SRAM column has been carried out which not only helps optimize the access time of the memory (*6%* improvement) but also reduces the size of the sense amplifier. A smaller sense amplifier would also leak less.

1.3 Summary

DWL, HWD and DBL approaches are common to most modern SRAMs, since they allow efficient memory operation above a certain minimum threshold. The pulsed word line and pulsed bit line technique sounds impressive in light of the figures quoted but an architecture based on this scheme has not appeared in the mainstream business to date, primarily due to the speed penalty involved. As the memory matrix becomes denser with the introduction of each new technology generation, leakage currents are expected to increase to unacceptable levels, more so with the reduction in device threshold voltage. The reduction in matrix leakage is therefore a challenge that designers are expected to be faced with in every technology generation to follow.

1.4 Thesis organization

Chapter 2 focuses on the motivation for tackling SRAM leakage in sub-*100nm* technologies, and provides a thorough description of the assignment and the boundary conditions.

The concept of source biasing and an overview of the proposed solution are presented in Chapter 3. Also highlighted are the two sub-sets of the proposed solution.

Chapter 4 presents a detailed analysis on the first sub-set of the proposed solution. It contains a detailed discussion on the NMOS footer switch when used in conjunction with a memory block.

Chapter 5 presents a detailed analysis on the second sub-set of the proposed solution. It contains a detailed discussion on the PMOS header switch when used in conjunction with a memory block.

In Chapter 6, a model is developed that estimates total area overhead associated with the switch as a function of the number of fingers in layout. In light of the results obtained in Chapters 4 and 5, one type of switch (in a header and footer role) is then identified based on an area/type trade-off.

Two techniques to achieve data retention voltage during periods of standby are introduced in Chapter 7. The first option (using a diode-connected transistor together with a switch and the memory block) is thoroughly investigated for the two sub-sets of the proposed solution. The second technique (using an actively clamped switch) is not fully investigated but its feasibility is determined for use with the block-switch architecture. The most promising options are then highlighted based on an area/leakage reduction trade-off.

In Chapter 8, the highlighted options are investigated for power savings at the desired design operating frequency. Power savings are investigated for two different memory access schemes: cyclic single

block access and sequential block access. The results of the thesis including complete area/speed/power trade-off for the promising options are presented at the end and the most suitable architecture is identified.

Some of the potential directions of future work in the area of low-power SRAM design are indicated in Chapter 9.

Appendix A contains an overview of sources of power dissipation in digital circuits.

Appendix B presents a functional overview of the SRAM. Also covered in this section is a basic review of the read and write operation of the 6T SRAM cell.

Appendix C addresses the mechanisms of leakage in a MOSFET. In particular, sub-threshold leakage, gate leakage and gate induced drain leakage are covered.

1.5 References

[1] A. P. Chandrakasan, S. Sheng and R. W. Brodersen, "*Low-Power CMOS Digital Design*," IEEE Journal of Solid State Circuits, Vol. 27, Issue 4, Apr 1992, pp. 473-484

[2] J. Rabaey and M. Perdam, "*Low Power Design Methodologies*," Kluwer Academic Publishers, 1996

[3] W. Nebel and J. Mermet, "*Low Power Design in Deep Submicron Electronics*," Kluwer Academic Publishers, 1997

[4] K. Itoh, "*VLSI Memory Chip Design*," Springer-Verlag Berlin Heidelberg, 2001

[5] M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, S. Nagao, S. Kayano and T. Nakano, "*A Divided Word-Line Structure in the Static RAM and its Application to a 64k Full CMOS SRAM*," IEEE Journal of Solid State Circuits, Vol. 18, Issue 5, October 1983

[6] 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*," IEEE Journal of Solid-State Circuits (JSSC), Vol. 25, Issue 5, pp. 1068–1073, October 1990

[7] A. Pavlov, M. Sachdev, "*CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies: Process-Aware SRAM Design and Test*," Springer, 2008

[8] K. Itoh, K. Sasaki and Y. Nakagome, "*Trends in Low-Power RAM Circuit Technologies*," Proceedings of the IEEE, Vol. 83, Issue 4, pp. 524-543, April 1995

[9] M. Margala, "*LowPower SRAM Circuit Design*," Proceedings of 7th IEEE International Workshop on Memory Technology, Design and Testing (MTDT '99), San Jose, August 1999

[10] M. Khellah, Y. Ye, N. S. Kim, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb and V. De, "*Wordline & Bitline Pulsing Schemes for Improving SRAM cell stability in Low-Vcc 65nm CMOS Designs*," Digest of Technical Papers 2006, Symposium on VLSI Circuits 2006, pp. 9-10

[11] A. Karandikar and K. K. Parhi, "*Low Power SRAM Design using Hierarchical Divided Bit-line Approach*," Proceedings of IEEE Int. Conference on Computer Design (ICCD), Austin, October 1998

[12] R. E. Aly, M. A. Bayoumi and M. Elgamel, "*Dual Sense Amplified Bit Lines (DSABL) Architecture for Low-Power SRAM Design*," IEEE International Symposium on Circuits and Systems 2005, ISCAS 2005, Vol. 2, pp. 1650-1653, 23-26 May, 2005

[13] H. Kawaguchi, Y. Itaka, T. Sakurai, "*Dynamic leakage cut-off scheme for low-voltage SRAMs*", Symposium of VLSI Circuits, Honolulu, HI, pp. 140-141, June 1998

[14] S. K. Jain and P. Agarwal, "*A Low leakage and SNM free SRAM cell design in Deep Sub micron CMOS Technology*," 19th International Conference on VLSI Design 2006, held jointly with 5th International Conference on Embedded Systems and Design, 3-7 Jan. 2006

[15] B. Amelifard, F. Fallah and M. Pedram, "*Leakage Minimization of SRAM cells in a Dual –V_t and dual –T_{ox} Technology*," IEEE Transactions on Very Large Scale Integration Systems, Vol. 16, Issue 7, July 2008

[16] F. Hamzaoglu, Y. Ye, A. Keshavari, K. Zhang, S. Narendra, S. Borkar, M. Stan and V. De, "*Analysis of dual-V_T SRAM cells with full-swing single-ended bit line sensing for on-chip cache*," IEEE Transactions on VLSI Systems, Vol. 10, Issue 2, April 2002

[17] T.S. Doorn, J.A. Croon, E. J. W. ter Maten and A. Di. Bucchianico, "*A yield centric statistical design method for optimization of the SRAM active column*," to appear in ESSCIRC 2009 proceedings
on of the SRAM active column," to appear in ESSCIRC 2009 proceedings

Chapter 2

Motivation and assignment description

2.1 Motivation

It is evident from the discussion presented in Chapter 1 that both the active and standby power needs to be reduced in order to minimize power consumption in SRAMs. However, as the memory matrix becomes denser with the introduction of each new technology generation, sub-threshold leakage currents are expected to increase to unacceptable levels, more so with the reduction in device threshold voltage. A classic case is that of pass-gate leakage in the SRAM cell, where leakage of non-accessed cells can approach a significant fraction of the read current of the selected cell in the same column [1]. Thin gate oxides in *90nm* technology and beyond, coupled with high electric fields across the oxide result in high gate leakage current [2].

It is therefore a logical approach to tackle the issue of increasing standby currents in embedded memories in an effort to bring down the total standby power consumption of the chip. Significant reduction in leakage current for dual-V_{DD} SRAMs, at a relatively low area overhead and performance penalty could be a valuable contribution to knowledge. The motivation behind addressing the problem of increasing leakage currents is further fueled by the average-rate advances in battery technology [3] in recent years.

2.2 Assignment description

The assignment is to conceive and design a *1.1V*, *500MHz, 1kb* x *32* Non-Cell-Implant (NCI) SRAM architecture in TSMC *45nm* technology such that power reduction is achieved both during the active as well as standby modes of operation by using source biasing leakage reduction technique. Preliminary work on leakage reduction schemes for embedded SRAMs [4] was conducted at NXP Semiconductors, Eindhoven, the Netherlands, before the start of this work and it was concluded that significant reduction in cell leakage current can be achieved using the concept of source biasing. The novelty here is the reduction of active SRAM power by employing leakage reduction techniques. The operating temperature range for the design is specified to be *–40°C* to *+125°C*. Moreover, the design is expected to be robust over supply voltage variation of $V_{DD(nom)}$*±10%*. A thorough evaluation of the achieved gain (i.e. power savings) in light of area and speed trade-offs associated with the architecture is also expected at the end of the assignment. The focus in this assignment will be on low-power design and operation of the memory matrix and not on the peripheral circuit that drives it.

2.3 Simulation environment and process corners

All simulations for this work have been carried out using Philips Pstar version *5.4.4* circuit design software with Simkit version *3.1.2* and process block A11, running under Red Hat Linux *3.1.3-6.12* environment. The simulation corners governing results presented in this work are defined in Table 2.1. Specifically in chapter 7, SNFP and FNSP processes are also employed in simulations.

Table 2.1 Definition of simulation corners

Defined simulation corner	**Process**	**Voltage / V**	**Temperature / °C**
FAST	FNFP	$1.1V_{DD(nom)}$	– 40
NOMINAL	NOMINAL	$V_{DD(nom)}$	25
SLOW	SNSP	$0.9V_{DD(nom)}$	125
MOST LEAKY	FNFP	$1.1V_{DD(nom)}$	125
LEAST LEAKY	SNSP	$0.9V_{DD(nom)}$	– 40

All plots showing temperature dependence of parameters have their temperature axis defined in degrees Celsius, unless otherwise stated.

2.4 References

[1] T. S. Doorn, R. Salters, L. E. Villagra, "*SRAM Design Challenges: A research view on current status and future work*," Technical Note NXP-R-TN-2007/00066, Issued 6/2007

[2] Y. Takeyama, H. Otake, O. Hirabayashi, K. Kushida and N. Otsuka, "*A Low Leakage SRAM Macro with Replica Cell Biasing Scheme*," IEEE Journal of Solid State Circuits, Vol. 41, Issue 4, April 2006

[3] J. Rabaey and M. Perdam, "*Low Power Design Methodologies*," Kluwer Academic Publishers, 1996

[4] T. S. Doorn, "*Leakage reduction in SRAM cells*," Technical Note NXP-R-TN-2008/00084, Issued 4/2008

Chapter 3

Concept of source biasing and approach towards a solution

3.1 Background work

Source biasing leakage reduction technique for embedded SRAMs has been investigated at NXP Semiconductors, Eindhoven, the Netherlands [1]. Source biasing allows the reduction of all three types of SRAM cell leakage currents (refer to Appendix C). It has been found that the total leakage of a TSMC $0.374\mu m^2$ Non-Cell-Implant High-Speed SRAM cell in *45nm* technology can be reduced by *77%* at elevated temperatures and nominal supply voltage when the source node of the pull down transistors of the SRAM cell (here onwards referred to as SRC node) is raised from *GND* to $V_{DD(nom)} - 0.7V$, keeping the NMOS bulk and word line voltages at *GND* [1]. *41%* reduction in leakage current can be achieved if the source node of the pull up transistors of the cell (here onwards referred to as SUP node) is lowered to *0.7V* keeping the PMOS bulk at V_{DD} and the word lines at *GND* [1]. These figures promise significant power savings for a potential SRAM architecture employing source biasing leakage reduction technique.

3.2 Leakage reduction with source biasing

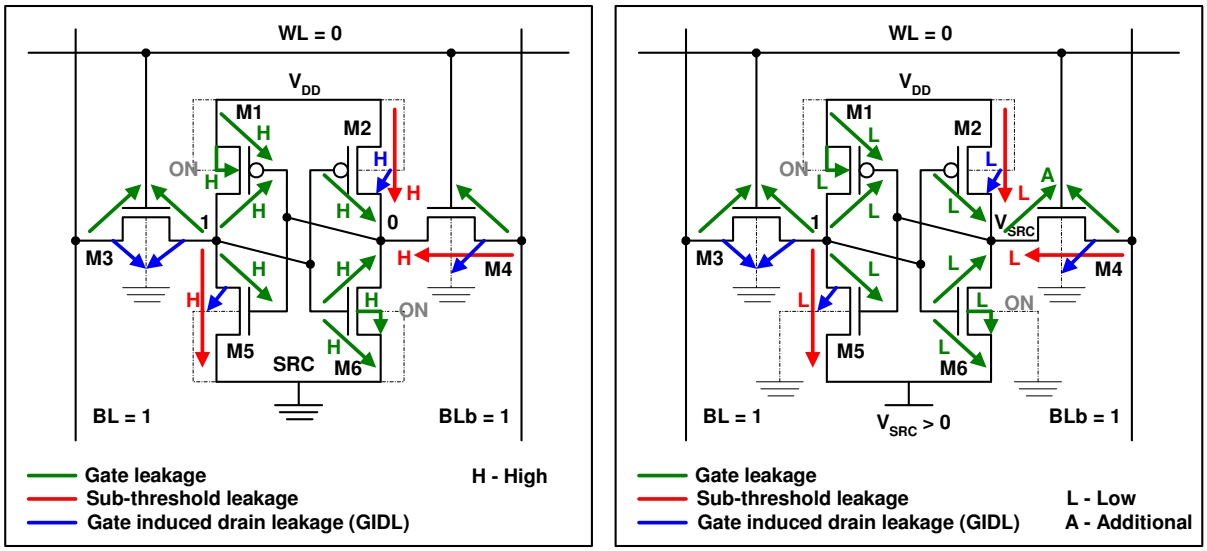
The concept of source biasing is thoroughly explained here since it forms the basis of the work presented in this thesis. The threshold voltage of a MOSFET can be adjusted by incorporating

substrate bias. When $V_{SB}=0$, inversion of the silicon surface occurs when the surface potential equals twice the Fermi potential. However if $V_{SB}>0$, then the charge stored in the depletion region increases. In order to achieve inversion with $V_{SB}>0$, the charge on the gate terminal has to increase in order to compensate for the increased charge in the depletion region. Therefore when $V_{SB}>0$, the threshold voltage of an NMOS increases. The threshold voltage of a PMOS also increases for $|V_{SB}|>0$. Equation (3.1) [2] presents the equation for device threshold voltage where V_{T0} is the device threshold voltage when $V_{SB}=0$, γ is the body-effect coefficient and ϕ_F is the Fermi potential.

$$V_T = V_{T0} + \gamma\left(\sqrt{|(-2)\phi_F + V_{SB}|} - \sqrt{|2\phi_F|}\right) \tag{3.1}$$

The increase in device threshold voltage results in the decrease of sub-threshold leakage currents, since sub-threshold leakage is exponentially dependant on threshold voltage and increases by an order of magnitude for every *100mV* decrease in threshold voltage [3].

Figure 3.1a illustrates the leakage currents in the SRAM cell under normal bias whereas Figure 3.1b illustrates the same when the SRC node voltage is raised to a potential greater than zero. The causes for the reduction in cell leakage currents under the influence of source biasing (applied to the SRC node) are provided in Table 3.1. The transistor instance names tie in with those appearing in Figure 3.1 to aid understanding.



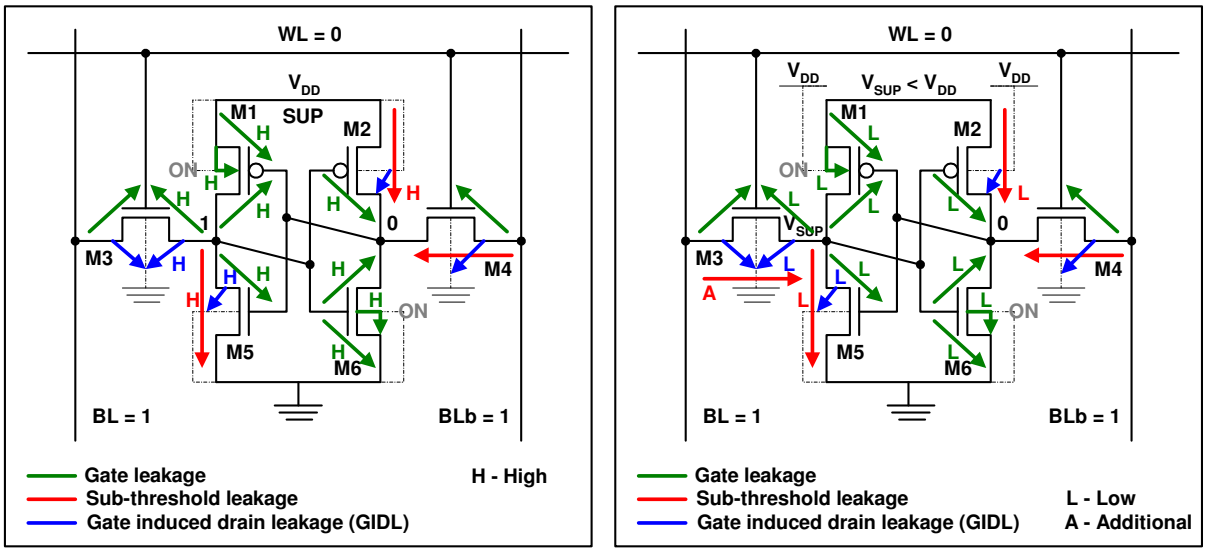
(a) (b)

Figure 3.1 High leakage with $V_{SRC}=0$ (a) and reduced leakage with $V_{SRC}>0$ (b)

Table 3.1 Causes of cell leakage reduction obtained as a consequence of source biasing the SRC node

Transistor	Type of leakage current altered	Cause
M5	Sub-threshold leakage decreases Gate leakage decreases	Reduced V_{DS}, increased V_{SB} Reduced $\|V_{GD}\|$
M4	Sub-threshold leakage decreases Additional gate leakage	Reduced V_{DS}, negative V_{GS} Increased $\|V_{GS}\|$
M2	Sub-threshold leakage decreases, GIDL decreases Gate leakage decreases	Reduced $\|V_{DS}\|$ Reduced $\|V_{DB}\|$ Reduced V_{GD}
M6	Gate leakage decreases	Reduced V_{GS}, reduced V_{GD}
M1	Gate leakage decreases	Reduced $\|V_{GS}\|$, reduced $\|V_{GD}\|$

Figure 3.2a illustrates the leakage currents in the SRAM cell under normal bias whereas Figure 3.2b illustrates the same when the SUP node voltage is lowered to a potential less than V_{DD}. The causes for the reduction in cell leakage currents under the influence of source biasing (applied to the SUP node) are provided in Table 3.2. The transistor instance names tie in with those appearing in Figure 3.2 to aid understanding.



(a) (b)

Figure 3.2 High leakage with V_{SUP}=V_{DD} (a) and reduced leakage with V_{SUP}<V_{DD} (b)

Table 3.2 Causes of cell leakage reduction obtained as a consequence of source biasing the SUP node

Transistor	Type of leakage current altered	Cause
M5	Sub-threshold leakage decreases, GIDL decreases, Gate leakage decreases	Reduced V_{DS}, reduced V_{DB}, Reduced $\|V_{GD}\|$
M3	Additional sub-threshold leakage, GIDL decreases, Gate leakage decreases	Increased V_{DS}, reduced V_{SB}, Reduced $\|V_{GS}\|$
M6	Gate leakage decreases	Reduced V_{GS}, reduced V_{GD}
M1	Gate leakage decreases	Reduced $\|V_{GS}\|$, reduced $\|V_{GD}\|$
M2	Sub-threshold leakage decreases, Gate leakage decreases	Reduced $\|V_{DS}\|$, increased $\|V_{SB}\|$, Reduced V_{GD}

Source biasing the SRC node (while maintaining word line voltage and NMOS bulk at *GND* potential and the PMOS bulk and SUP node at V_{DD}) achieves greater leakage reduction because the larger pull down transistor is back-biased and the pass gate gets a negative gate-source voltage.

Source biasing the SUP node (while maintaining word line voltage and NMOS bulk at *GND* potential and the PMOS bulk at V_{DD}) achieves lower leakage reduction because only the smaller pull up transistor is back-biased.

3.3 Approach towards a solution

In this contribution, the time during which a memory read/write operation takes place is termed as the ACTIVE mode and the time during which no read/write operation takes place is termed as STANDBY mode/period.

TSMC guarantees correct read/write operation of the memory cell when the potential across its terminals is maintained close to V_{DD} since that maintains sufficiently high SNM. The data retention voltage (DRV) of the cell is specified at *0.7V*. During the ACTIVE mode of a cell, the word line is active and either read or write current flows through the cell depending on the type of memory cycle (refer to Appendix B). During STANDBY, the word line is inactive and leakage current of the cell is dominant. Source biasing leakage reduction technique can therefore only be employed during the STANDBY period of a memory cell.

In order to achieve maximum power savings in both the ACTIVE and STANDBY modes of operation, the entire memory instance can be broken down into equal-sized blocks so that when a read/write operation takes place in a specific block, all other blocks can be placed in STANDBY such that

reduced leakage flows through the cells housed therein as a consequence of DRV maintained across the block terminals. The architecture of a block is such that all cells within a block share the SUP and the SRC nodes. All cells also have their PMOS bulks connected to the same node. The same is true for the NMOS bulk terminals. Also, all cells in a column share the same bit line pair and all cells in a row share the same word line. A graphical representation of a block containing *B* cells is presented in Figure 3.3. The abbreviations used in Figure 3.3 tie in with those used in Section 3.2 to aid understanding. The selection of the block size is discussed in Chapter 4.

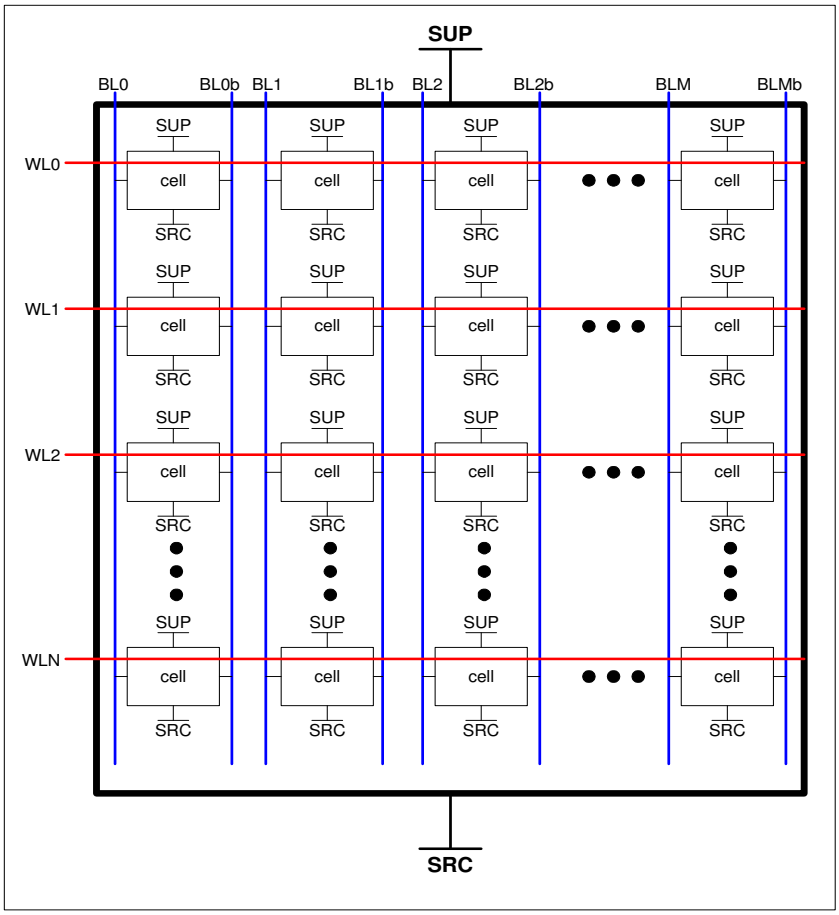


Figure 3.3 Memory block containing *B* cells

In order to reduce leakage of memory cells housed in blocks during STANDBY periods, a simple NMOS footer switch can be connected between the common SRC node of a block and *GND*. During the ACTIVE mode, the gate of the NMOS switch is driven high, which allows near-V_{DD} potential across all cells in the memory block, thus meeting TSMC's recommendation for correct cell operation. At the end of the ACTIVE mode, the gate of the switch is driven low, which places the SRC node in a floating state. Since the SRAM cell pull down transistor connecting the SRC node and the cell node storing logic *0* is on (refer to Figure 3.4), the node storing logic *0* is also placed in a floating state as a

result. Leakage currents of the memory block then begin to charge the total SRC node capacitance, as a result of which V_{SRC} begins to rise.

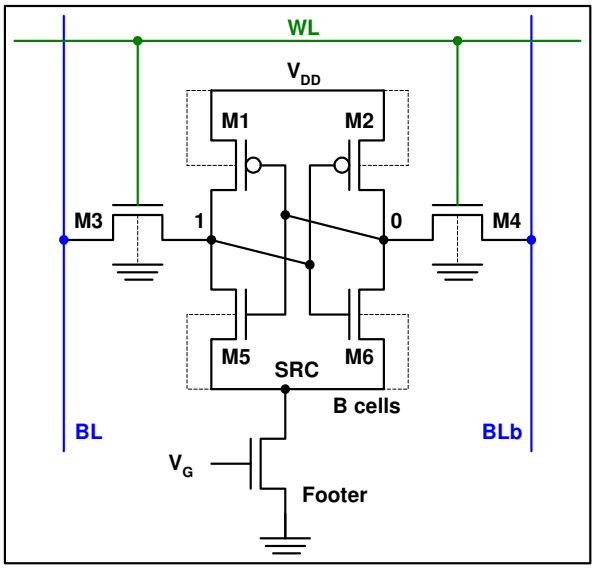


Figure 3.4 Inclusion of a footer switch to a block of *B* cells

When V_{SRC} starts to increase, memory cell leakage currents start to decrease. Sub-threshold leakage currents in particular are reduced. M5 in Figure 3.4 starts to get back-biased. A negative gate-source voltage starts to develop across the terminals of M4 and $|V_{DS}|$ across M2 starts to fall, as mentioned in Table 3.1.

If the rising SRC node voltage is not held at $V_{DD} - 0.7V$, the memory cells in the block would lose their stored data since DRV would be violated. In order to stop this from happening, two DRV-maintaining techniques are investigated in Chapter 7: a diode-connected transistor and an actively clamped footer. The DRV-maintaining circuit would hold the SRC node at $V_{DD} - 0.7V$ during STANDBY allowing reduced leakage current to flow through the memory block.

ACTIVE power would therefore be reduced since only one block would have V_{DD} across its terminals during any read/write cycle, while all other blocks would be maintaining DRV, hence leaking less. STANDBY power consumption would be reduced since all blocks would be maintaining DRV when no read/write cycle is in progress.

A top-level diagram of the proposed solution is presented in Figure 3.5. A similar approach has also been investigated using a PMOS header switch. The main difference between the footer and header options is the fact that the footer must be wide enough to sink the read current generated through an

entire row of cells in the block. This condition does not apply to a header since it does not fall in the path of read current.

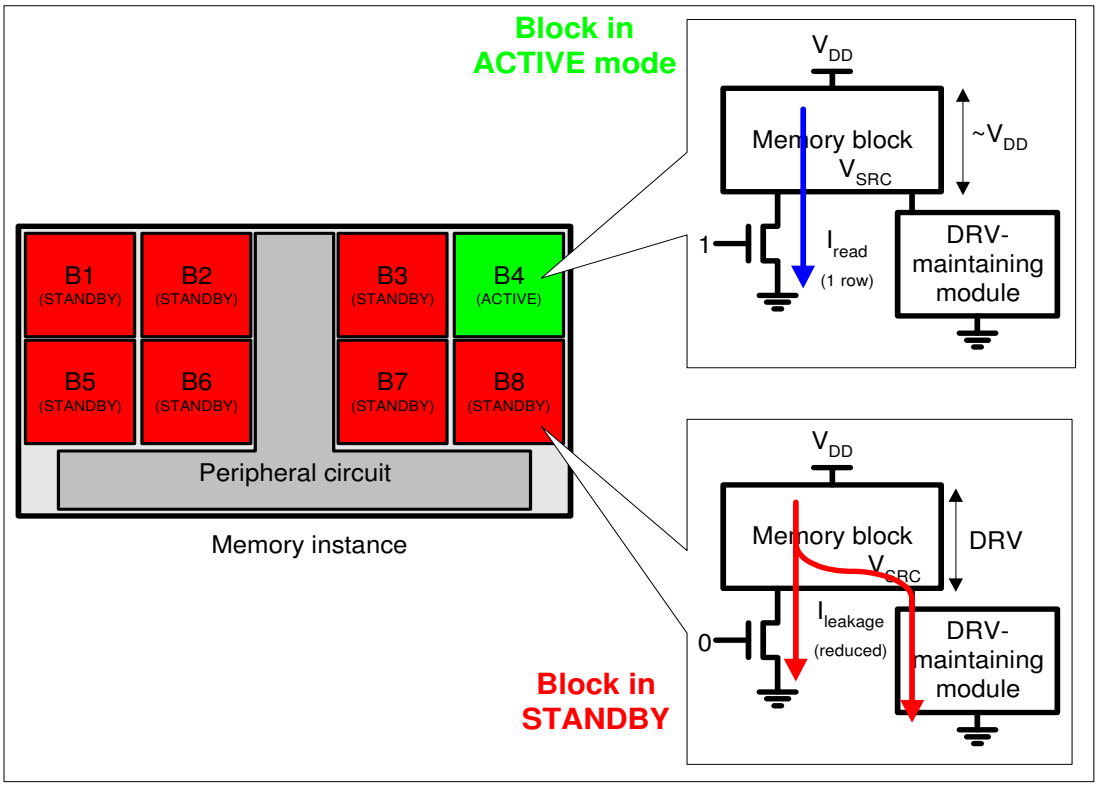


Figure 3.5 A memory instance showing *1* ACTIVE block and the rest in STANDBY

3.4 Conclusion

The concept of source biasing and its relevance to leakage reduction has been highlighted. A top-level layout of the potential solution to the assignment has also been presented. In the following chapter, the behavior of a footer switch is investigated when used in conjunction with a memory block.

3.5 References

[1] T. S. Doorn, "*Leakage reduction in SRAM cells*," Technical Note NXP-R-TN-2008/00084, Issued 4/2008

[2] J. Rabaey, A. Chandrakasan, and B. Nicoloc, "*Digital Integrated Circuits: A Design Prospective*," Second Edition. Prentice Hall, 2003

[3] Deepaksubramanyan, B.S., Nunez, A., *Analysis of subthreshold leakage reduction in CMOS digital circuits*, Circuits and Systems 2007, MWSCAS 2007, 50th Midwest Symposium on Volume, Issue, 5-8 Aug 2007, pp. 1400-1404

Chapter 4

Footer switch analysis with a memory block

4.1 Selection of memory block size

A memory block containing a certain number of rows and columns must first be identified in order to proceed with the switch analysis. The justification of the block width lies in the fact that the existing NXP layout of a sense amplifier (SA) fits within the pitch of two memory cells [1]. That means each SA is shared between two memory columns. For the desired 32-bit read/write access, a total of 32 SAs are required, which means selecting a block width of at least 64 cells/columns.

Figure 4.1 shows the percentage of total cells in STANDBY plotted as a function of number of rows in a block, when the number of columns per block is fixed at *64*. The numbers indicated on the plot itself indicate the number of blocks in a *32kb* instance. From the curve it can be seen that there is no sharp decision-aiding point on the curve that indicates what block height to use. However, what is evident is the fact that increasing the number of rows beyond *64* (*8* blocks) significantly starts to decrease the percentage of cells placed in STANDBY.

The extreme right end of the curve suggests the use of *1* block only, placing zero cells in STANDBY. No leakage reduction/power savings can be expected out of this option. The value on the left end of the curve suggests using a block as wide as a single row, i.e. *64* bits. It will become clear later in this

chapter that this option would lead to significant area overhead associated with the footer switch. So the selection of block height is based more on choice. An instance of *32kb*, when divided into *8* equal blocks would allow sufficient granularity in order to observe the effect of reduced leakage in the non-accessed blocks. Hence a block height of *64* columns is selected. Each block is therefore *64* columns wide and *64* rows high, containing a total of *4096* cells.

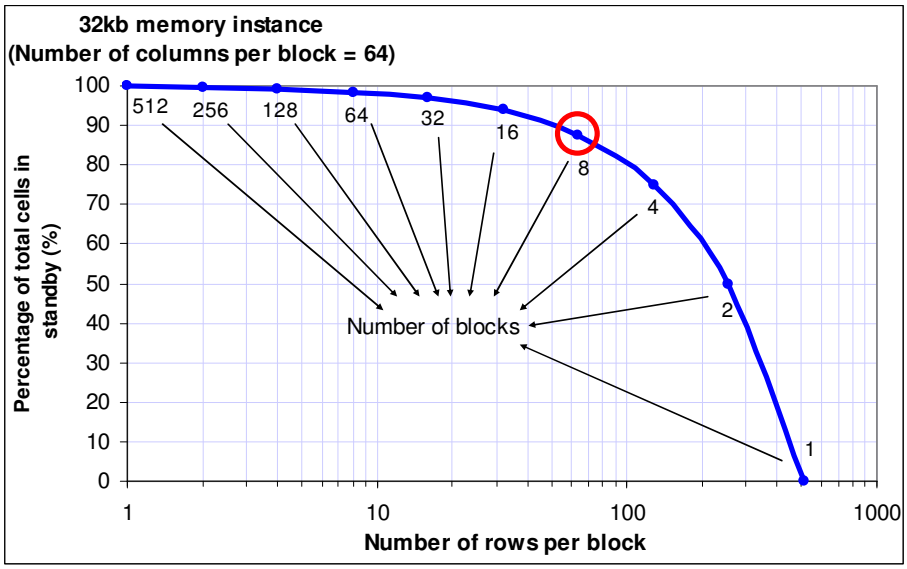


Figure 4.1 Block height selection

4.2 Factors governing footer size and type

The behavior of four different types of N-channel switches is now investigated in conjunction with a *4096*-cell memory block: high threshold voltage (H-V_T) logic transistor, standard threshold voltage (S-V_T) logic transistor, low threshold voltage (L-V_T) logic transistor and SRAM cell pull down transistor (SRAM-PD). The length of all transistors is fixed at *55nm* (length of the SRAM-PD transistor) so that process variations influencing channel length would be better correlated between the block and the switch. The minimum length of the logic transistors is however *40nm*. The SRAM-PD transistor is incorporated in the analysis since it is expected it would allow accurate tracking of process variation (particularly ΔV_T) in the memory block.

When a footer switch is used in combination with a memory block, three critical parameters influence the selection of the optimal switch in terms of type and size. These include:

iii) SRC node voltage in the ACTIVE mode

iv) The time required to pull down the SRC node from $V_{DD} - 0.7V$ to near-*GND* potential (i.e. the time needed to bring the block into ACTIVE mode from STANDBY mode)

v) Leakage current of the block with respect to that of the switch during STANDBY mode

Detailed simulations are carried out to fully understand the role each of the above-mentioned parameters. A comprehensive account of the simulation setup and a thorough discussion on the results obtained is presented in Sections 4.3 – 4.6.

4.3 SRC node voltage (V_{SRC}) in the ACTIVE mode

4.3.1 Simulation setup

During a read/write operation, the voltage at the SRC node needs to be sufficiently close to *GND* potential in order to achieve sufficient noise margin. TSMC guarantees correct read/write operation of the memory cell only when the potential difference across the memory cell is sufficiently close to V_{DD}. The maximum drain-source potential allowed across a footer switch in the current NXP memory range is *10mV*. This number is not a standard but merely a reference since it is sufficiently close to zero potential. The footer must ensure $V_{SRC} \leq 10mV$ in all process corners, over the entire temperature range and for all variations in supply voltage. This requirement on the footer governs its size.

In case of a *4096*-cell block, a single local word-line activates an entire row of cells, *64* cells to be exact. The switch therefore needs to be wide enough to sink the read current flowing through *64* cells as well as the leakage current flowing through all *4096* cells, and still maintain $V_{SRC} \leq 10mV$.

Two modified versions of an extracted Pstar netlist of *8* cells, arranged in *4* rows and *2* columns (known as a *double-quadro* model), are used as a starting point for all simulations detailed here onwards. The modifications relate to the separation of V_{DD} and SUP nodes (for block analysis with PMOS header switch) and *GND* and SRC nodes (for block analysis with NMOS footer switch). Instances of this model (with individual scaling factors "*S*") can be connected together to realize different block sizes and number of bits accessed. When a scaling factor "*S*" is used with an instance of the model, the simulator sees the resultant model as "*S*" layers of the instance piled on top of each other. All the layers however share the same port signals. Figure 4.2 shows the block diagrams of the modified *double-quadro* models.

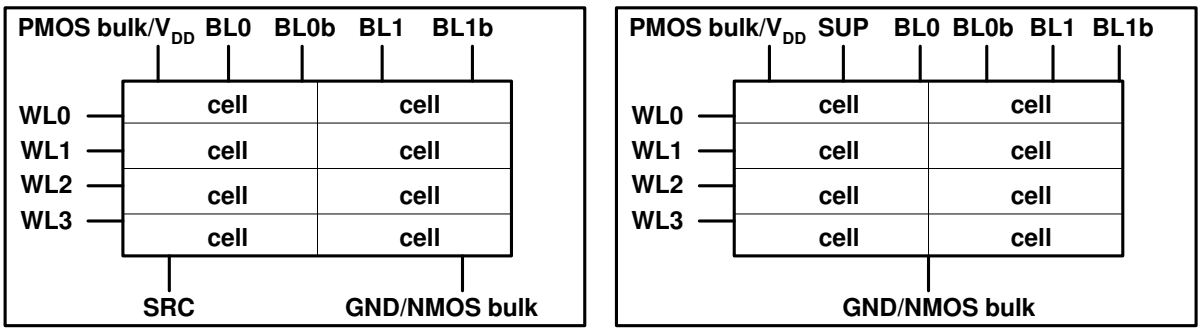


Figure 4.2 Modified double-quadro model for footer analysis (a) and header analysis (b)

For each type of N-channel switch, a DC simulation is set up in which one word line (WL0) is driven high to activate *64* cells in the memory block, all other word lines (represented by WL1 in Figure 4.3) are kept inactive. The footer gate is driven high, the bit lines and PMOS bulk are tied to V_{DD}, the NMOS bulk is tied to *GND* and the SRC node is initialized at *10mV*. The temperature is swept over the entire operating range. Also, the supply voltage is varied from $V_{DD} = 0.9V_{DD(nom)}$ to $V_{DD} = 1.1V_{DD(nom)}$. For each supply voltage level, the width of the footer is swept in order to achieve the value that guarantees $V_{SRC} \leq 10mV$ in the worst case. Figure 4.3 illustrates the simulation setup, which can also be used to obtain block read current value.

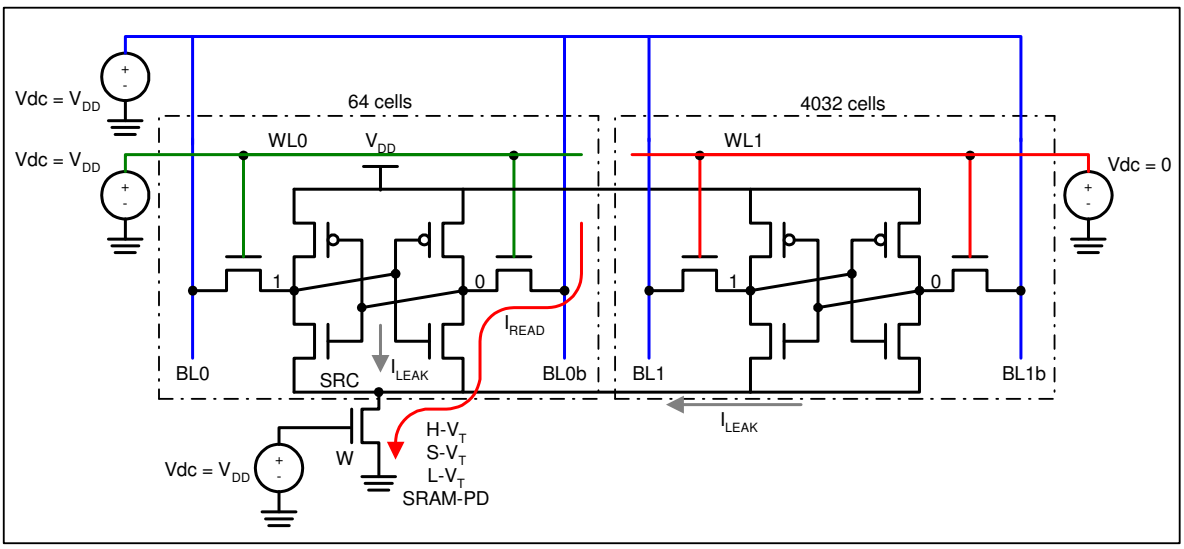


Figure 4.3 Sizing the footer

The reason of inclusion of the SRAM-PD transistor in the analysis was the fact that it would accurately track process variation in the block. We therefore use multiple SRAM-PD transistors (exact copies of the transistor used in the memory cell) to realize the total required width of the switch. The

individual transistor width is fixed at *215nm*, which is the width of the pull down transistor in the SRAM cell.

4.3.2 Results and discussion

4.3.2.1 High-V_T footer

Figure 4.4 shows SRC node voltage plotted as a function of temperature in FAST, NOMINAL and SLOW corners and with supply voltage variation, when using H-V_T footer.

The highest read current through the block-footer system flows in the FAST corner whereas the lowest read current flows in the SLOW corner. A width of *261µm* achieves $V_{SRC} \leq 10mV$ in the FAST corner. The FAST corner presents the worst case for SRC node voltage because the on-resistance of the block reduces far more than the on-resistance of the footer. V_{SRC} is therefore higher compared to NOMINAL and SLOW corners. Within each corner, V_{SRC} scales with the supply voltage, similar to how the voltage across the *lower* resistor would scale with supply voltage in a simple resistive potential divider circuit. Also, the on-resistance of the footer decreases with an increase in width, hence V_{DS} of the footer decreases with an increase in width.

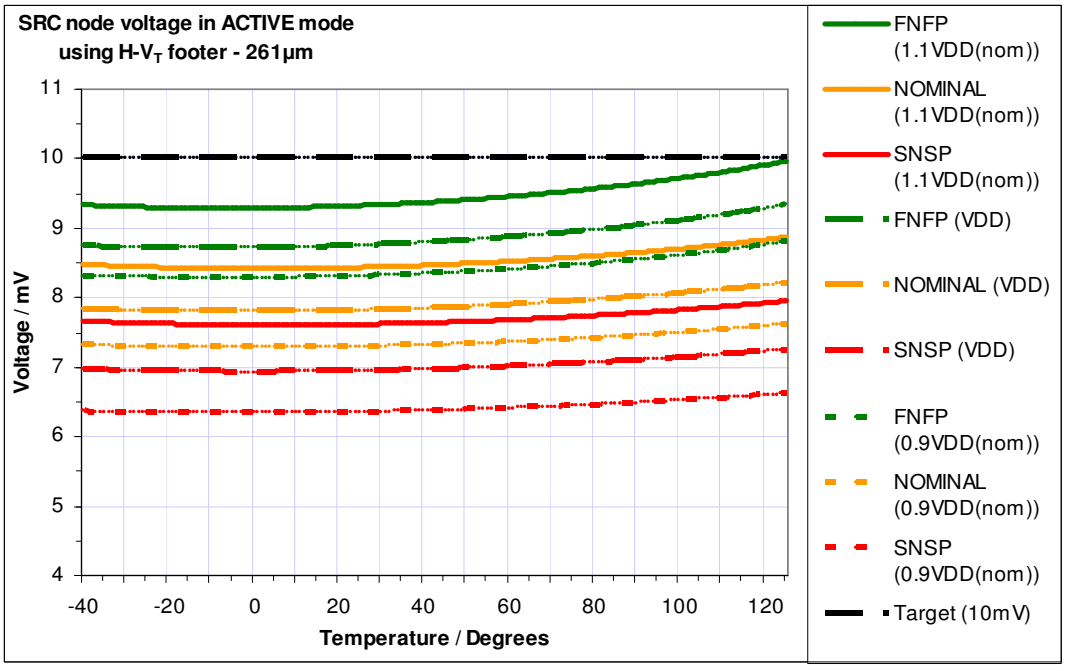


Figure 4.4 V_{SRC} in ACTIVE mode using H-V_T footer

Threshold voltage and mobility are the two key variables that vary with temperature. Both decrease with temperature. A decrease in threshold voltage causes an increase in current whereas a decrease in mobility leads to a reduction in current. The two variables vary in different proportion for the footer

and the memory bock. The shape of the curves over the full temperature scale is governed by whichever of the two parameters dominates for the block-footer system at a certain temperature value. The graphs are presented to indicate that the SRC node voltage does not surpass the *10mV* mark in any corner, over the entire operating temperature range and across full supply voltage variation. A detailed study on the behavior of mobility and threshold voltage variation with temperature of all memory cell transistors and the footer is outside the scope of this work.

4.3.2.2 Standard-V_T footer

Figure 4.5 shows SRC node voltage plotted as a function of temperature in FAST, NOMINAL and SLOW corners and with supply voltage variation, when using S-V_T footer. A width of *213µm* achieves $V_{SRC} \leq 10mV$ in the FAST corner.

A smaller S-V_T transistor (compared to H-V_T) suffices to guarantee $V_{SRC} \leq 10mV$ in the ACTIVE mode. This is because with a lower threshold voltage, the gate overdrive voltage is higher.

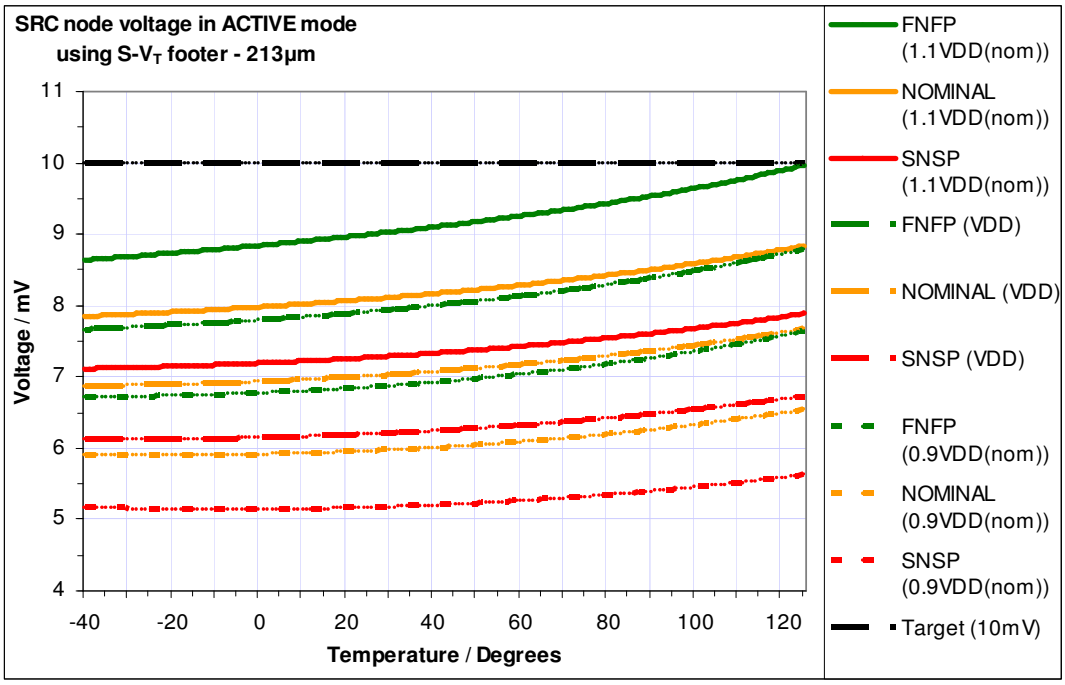


Figure 4.5 V_{SRC} in ACTIVE mode using S-V_T footer

4.3.2.3 Low-V_T footer

Figure 4.6 shows SRC node voltage plotted as a function of temperature in FAST, NOMINAL and SLOW corners and with supply voltage variation, when using L-V_T footer. A width of *176µm* achieves $V_{SRC} \leq 10mV$ in the FAST corner.

A smaller L-V_T transistor (compared to H-V_T and S-V_T) suffices to guarantee $V_{SRC} \leq 10mV$ in the ACTIVE mode. This is because with an even lower threshold voltage, the gate overdrive voltage is now much higher.

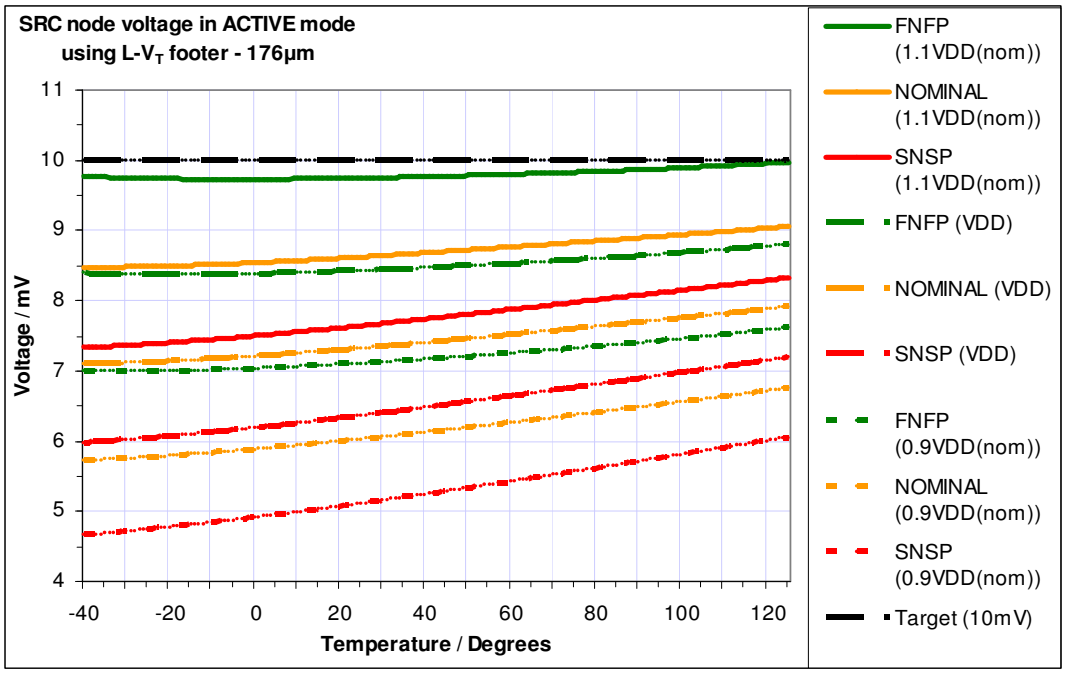


Figure 4.6 V_{SRC} in ACTIVE mode using L-V_T footer

4.3.2.4 SRAM-PD footer

Figure 4.7 shows SRC node voltage plotted as a function of temperature in FAST, NOMINAL and SLOW corners and with supply voltage variation, when using SRAM-PD transistor as the footer. A multiplication factor of *996* (*W = 996* x *215nm = 214µm*) achieves $V_{SRC} \leq 10mV$ in the FAST corner.

The size of the SRAM-PD is roughly the same as that of the S-V_T logic transistor (*213µm*). This indicates that the threshold voltage of the SRAM-PD transistor is roughly the same as that of the S-V_T logic transistor.

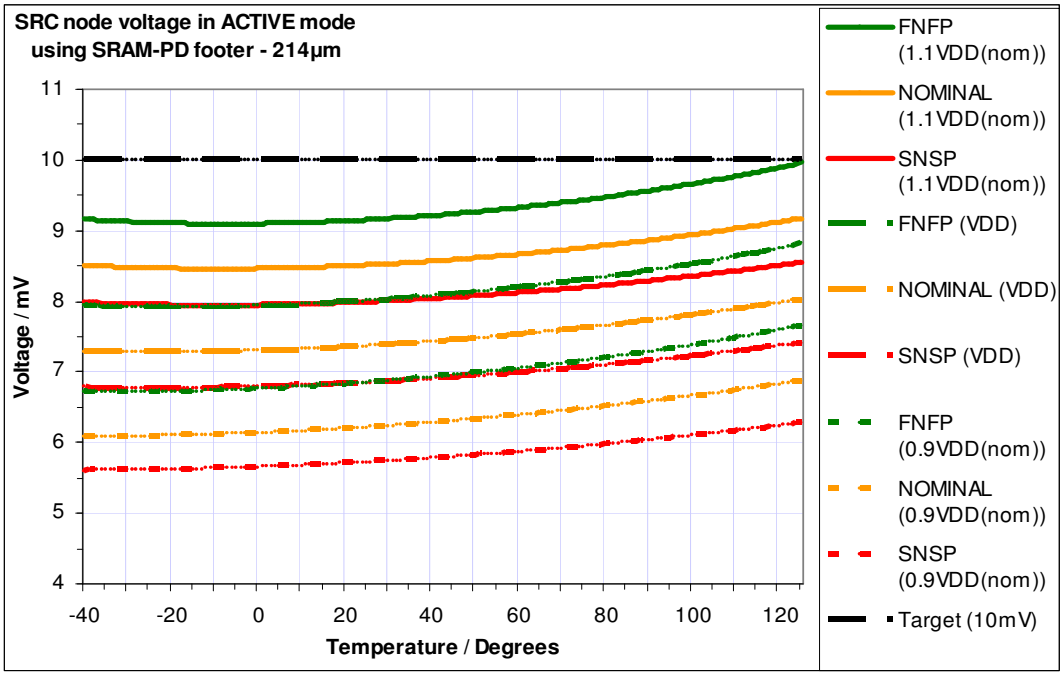


Figure 4.7 V_{SRC} in ACTIVE mode using SRAM-PD footer

4.3.2.5 Effect on read current

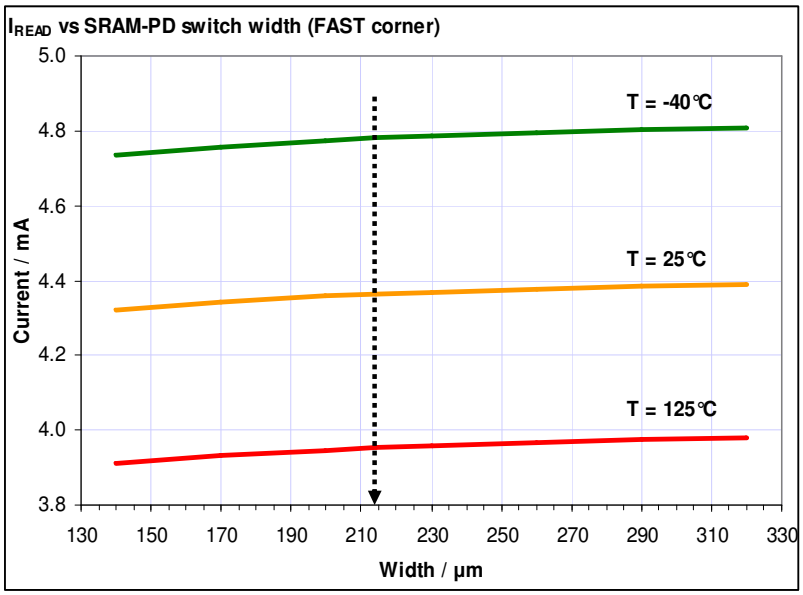


Figure 4.8 Read current of *64* cells for different switch sizes (FAST corner)

Figure 4.8 shows read current of *64* cells (*1* row) plotted as a function of width of SRAM-PD switch for different temperature values in the FAST corner. The read current increases with an increase in switch size due to reduced resistance of the switch. It is clear that further increases in the switch size (beyond *214µm*) lead to little increase in read current. The results are similar for other footer types (H-

V_T, S-V_T, and L-V_T). It can therefore be concluded that the switch sizes obtained from the simulation setup outlined in section 4.4.1 do not severely limit the read current flowing through the block-footer system.

4.4 SRC node pull down time (STANDBY to ACTIVE mode)

4.4.1 Simulation setup

The time taken by the footer to pull down the SRC node from $V_{DD} - 0.7V$ to $10mV$ contributes towards the access time of the memory. Normally the word lines are the fastest signals in the memory. They are active for approximately 60% of the total read access time [2]. In order to minimize the access time penalty associated with pulling down the SRC node, the word line and the gate of the footer are activated simultaneously. Since the cycle time of existing NXP memories in $45nm$ technology is in the range of $1ns$ and the target frequency for this work is $500MHz$, therefore a maximum access time penalty of $1ns$ can be added to the existing cycle time. The pull down time should therefore be less than or equal to $1ns$ for $500MHz$ operation (i.e. $2ns$ cycle time). This condition should be met across all process corners and over the full variation in supply voltage and temperature. A transient simulation is set up where all word lines in the block are held inactive till $t=0$, the bit lines tied to V_{DD} and the SRC node initialized to $V_{DD} - 0.7V$ in order to simulate STANDBY state. The internal storage nodes of all cells are also initialized to V_{DD} and $V_{DD} - 0.7V$.

At $t=0$, word line WL0 and the gate terminal of the footer are simultaneously driven high. The footer pulls down the SRC node voltage from $V_{DD} - 0.7V$ towards GND and the simultaneous activation of the word line simulates the start of a read operation. The simulation is carried out for the four types of N-channel footers for sizes obtained from the simulation setup outlined in section 4.4.1, and the time taken to pull down the SRC node to within $10mV$ is recorded. Figure 4.9 highlights the simulation principle.

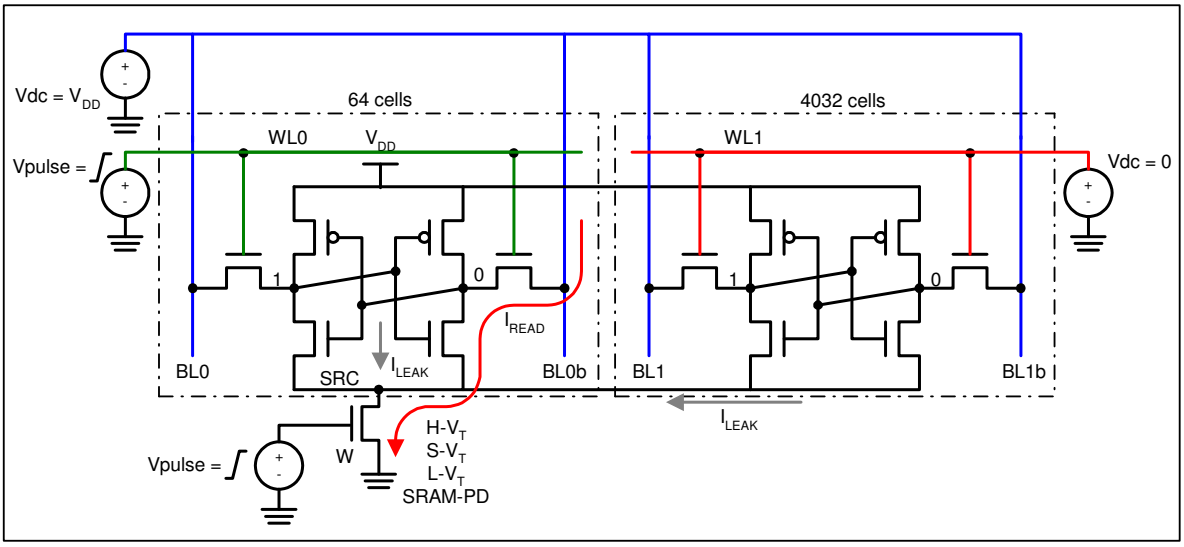


Figure 4.9 Simulation setup for the measurement of SRC node pull down time

4.4.2 Results and discussion

When data is to be read from or written to the memory, the memory block must first be brought out of STANDBY mode into ACTIVE mode. In order to switch between the modes, the footer must pull down the SRC node from $V_{DD} - 0.7V$ to $10mV$. Figure 4.10 shows the SRC node voltage and word line/footer gate signals plotted as a function of time in the FAST, NOMINAL and SLOW corners, when H-V_T footer is used.

V_{SRC} has been initialized to $0.51V$, $0.4V$ and $0.29V$ in the FAST, NOMINAL and SLOW corners respectively so as to simulate DRV condition ($0.7V$) across the memory block in STANDBY mode before the footer is turned on.

The pull down time is highest in the SLOW corner because SNSP transistors have reduced drive capability as a result of longer channel length, shorter width, thicker oxide layer (lower gate capacitance) and higher channel doping levels (higher threshold voltage). Increased temperature leads to greater scattering of carriers in the channel, hence reduced mobility. And a reduced supply voltage leads to a lower gate overdrive voltage ($V_{GS} - V_T$). The time taken to pull down the node is lowest in the FAST corner.

The SRC node temporarily rises beyond $V_{DD} - 0.7V$ mark (i.e. towards V_{DD}) before falling towards GND. This transitory increase in voltage is due to Miller effect and no more than $5mV$ in the worst case.

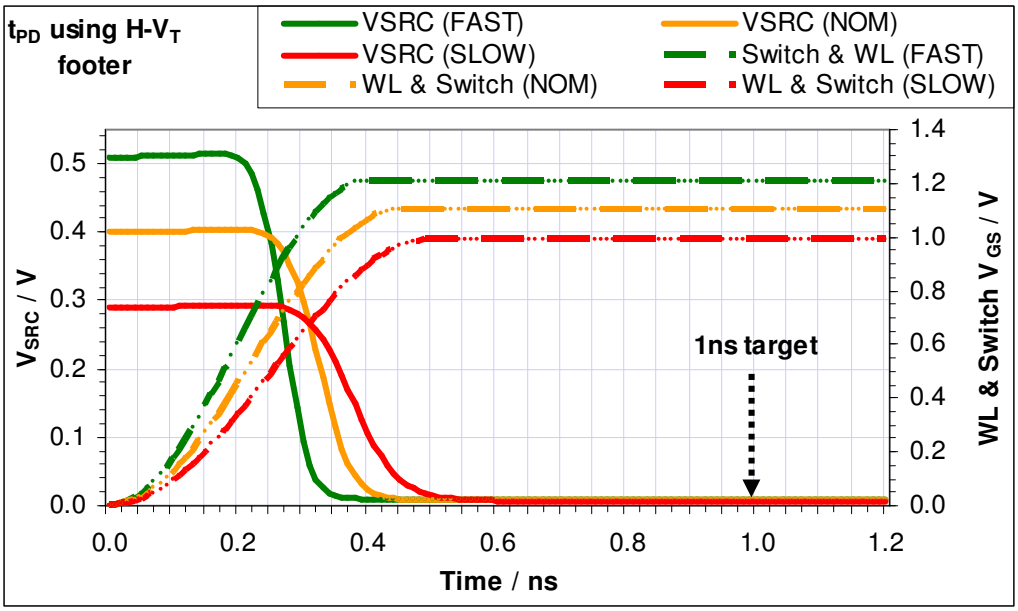


Figure 4.10 SRC node pull down time with H-V_T footer

Figure 4.11 shows the SRC node voltage and word line/footer gate signals plotted as a function of time in the FAST, NOMINAL and SLOW corners, when S-V_T footer is used. The pull down time in the three corners is lower, compared to the H-V_T footer. The lower pull down time can be attributed to increased drive capability of the S-V_T transistor due to a higher gate overdrive voltage ($V_{GS} - V_T$).

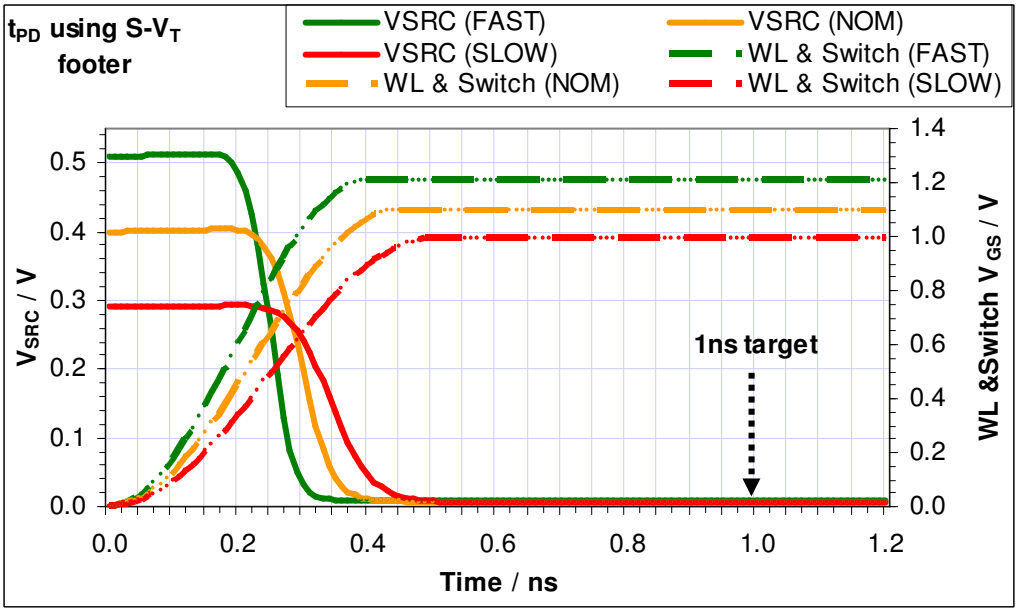


Figure 4.11 SRC node pull down time with S-V_T footer

Figure 4.12 shows the SRC node voltage and word line/footer gate signals plotted as a function of time in the FAST, NOMINAL and SLOW corners, when L-V_T footer is used. The pull down time is now even lower in the three corners for reasons provided in the above paragraph.

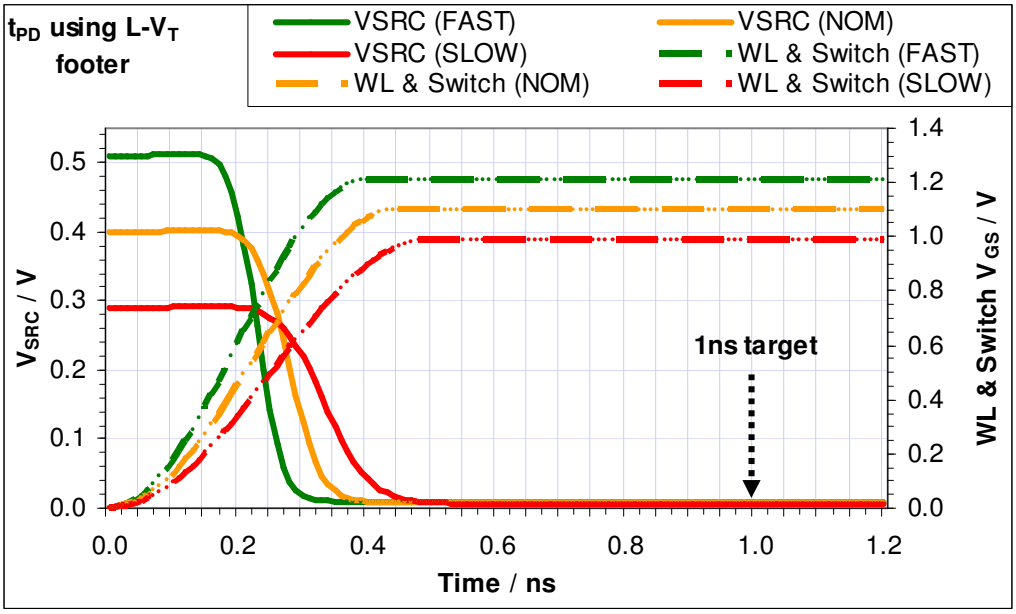


Figure 4.12 SRC node pull down time with L-V_T footer

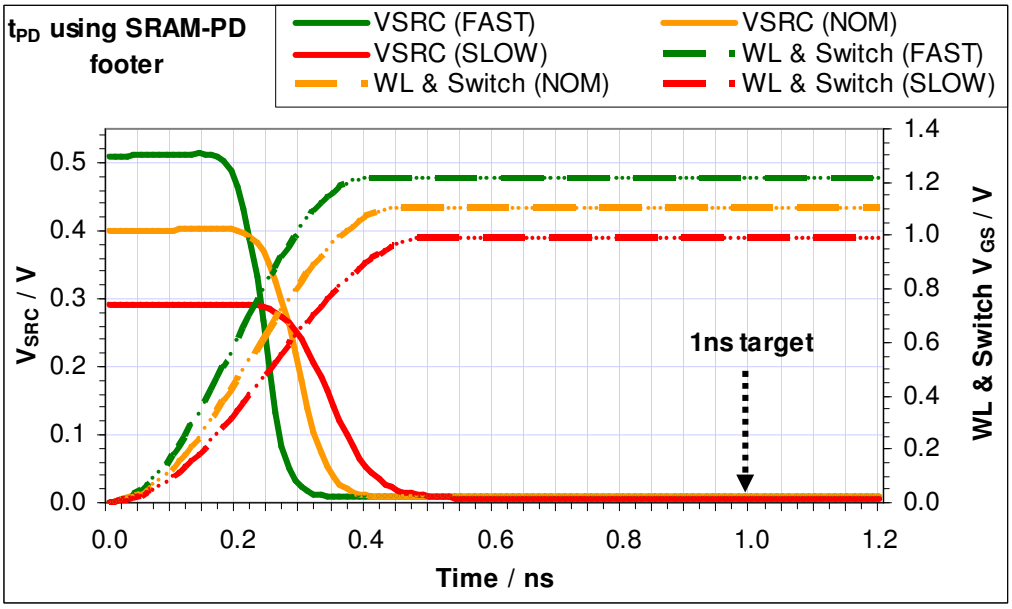


Figure 4.13 SRC node pull down time with SRAM-PD footer

Figure 4.13 shows the SRC node voltage and word line/footer gate signals plotted as a function of time in the FAST, NOMINAL and SLOW corners, when SRAM-PD footer is used. The pull down time in the three corners is similar to that of the S-V_T transistor.

The plots presented in Figure 4.10 through to Figure 4.13 clearly show the SRC node being pulled down to within *10mV* much before the *1ns* mark. Values obtained for the pull down time in different corners represent the access time penalty associated with the memory when a particular type of footer is used.

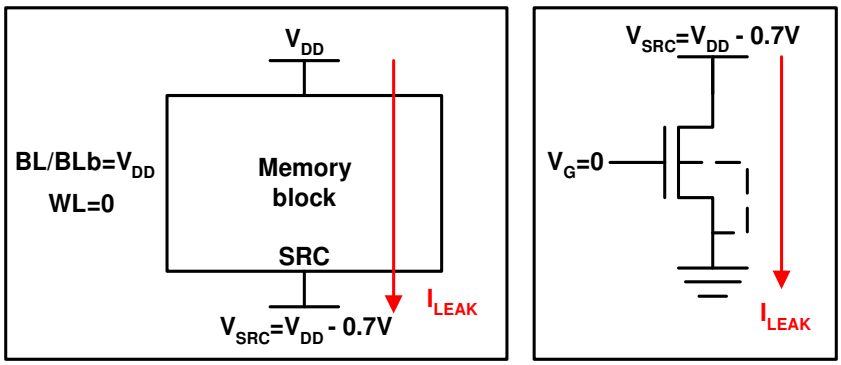
4.5 Block and switch leakage currents (STANDBY mode)

4.5.1 Simulation setup

During STANDBY mode, the intention is to be able to maintain DRV across the memory block. In order to achieve maximum benefit out of source biasing leakage reduction technique, the switch must leak less than the block when V_{SRC} is maintained at $V_{DD} - 0.7V$ during STANDBY. This condition must be met in all process corners, across supply voltage variation and over the entire temperature scale. In case the leakage of the switch exceeds that of the block at the condition specified above, the actual potential across the block would be higher than the DRV and maximum leakage reduction would not be achieved.

Two independent DC simulations are set up, one for the block and the other for the footer. In the first (Figure 4.14a), the leakage current of the block is simulated in different process corners over the full temperature scale. The supply node is maintained at V_{DD} and the SRC node is maintained at $V_{DD} - 0.7V$. Variations in supply voltage are also incorporated. All word lines are deactivated. The bit lines and the PMOS bulk are tied to V_{DD} and the NMOS bulk is tied to *GND*.

The second simulation (Figure 4.14b) is carried out for each type of footer for the sizes obtained through the simulation setup outlined in section 4.4.1. The drain of the footer is maintained at $V_{DD} - 0.7V$ and the bulk, source and gate terminals are tied to *GND*. The total leakage of the transistor over the full temperature scale is then recorded in different corners.



(a) (b)

Figure 4.14 Block (a) and footer (b) leakage current simulation principle

4.5.2 Results and discussion

Figure 4.15 shows block and switch leakage plotted as a function of temperature for each type of switch. The highest leakage, both for the block and the switch, occurs in the MOST LEAKY corner and the lowest leakage occurs in the LEAST LEAKY corner. The details of these corners can be found in chapter 3.

The leakage curves for the memory block, in Figure 4.15, tend to become horizontal towards the lower end of the temperature scale. This can be explained by the fact that at room temperature and below gate leakage and gate induced drain leakage (GIDL) currents dominate, which are much less temperature dependent. The curve for the LEAST LEAKY corner is flatter over a wider range of temperature values because sub-threshold currents decrease by the highest proportion (compared to MOST LEAKY and NOMINAL corners) as a result of increased threshold voltage, and because gate leakage and GIDL dominate heavily at low temperatures. As the temperature rises, so does the leakage current due to the fact that sub-threshold currents, which depend exponentially on temperature and threshold voltage, start to dominate.

Unlike the leakage curves for the S-V_T, L-V_T and SRAM-PD switches, the leakage current curves for the H-V_T footer (Figure 4.15a) resemble those of the block at the lower end of the temperature scale. This is because a higher threshold voltage causes maximum reduction in sub-threshold leakage at reduced temperatures. Gate induced drain leakage and gate leakage dominate as a result.

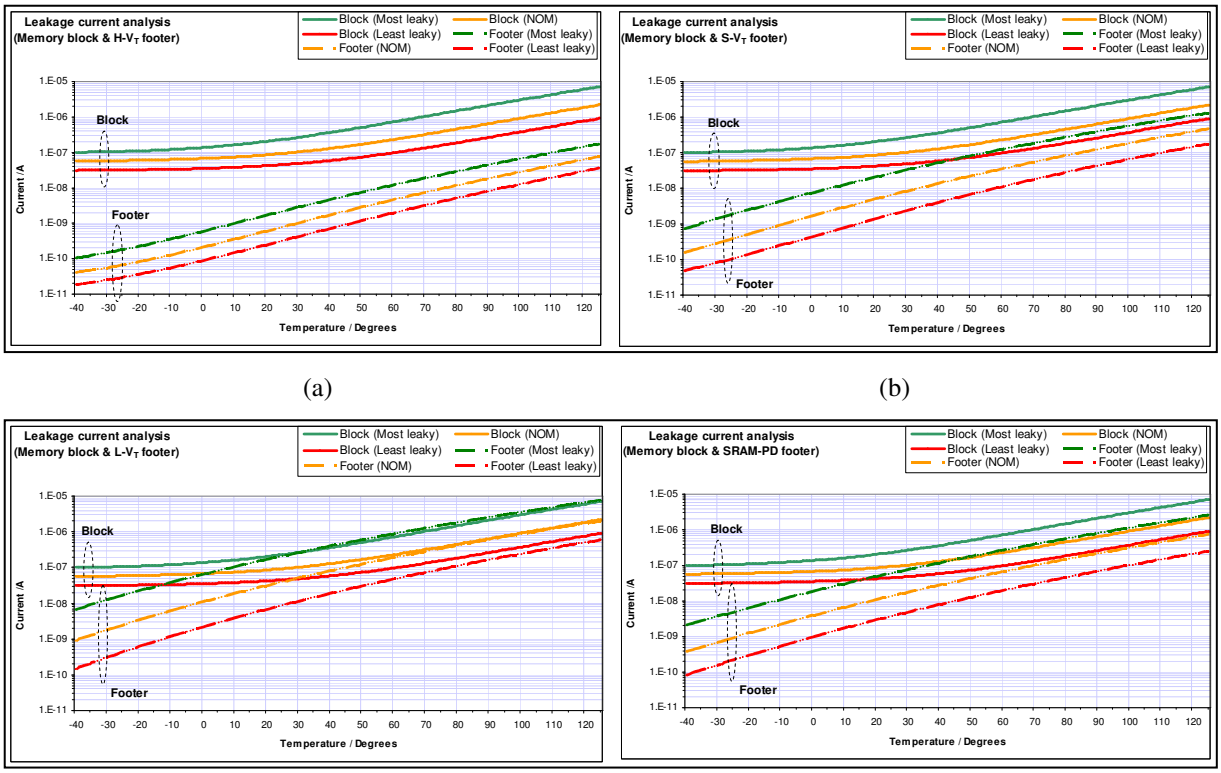


Figure 4.15 Block and H-V_T switch leakage (a) Block and S-V_T switch leakage (b) Block and L-V_T switch leakage (c) and Block and SRAM-PD switch leakage (d)

The leakage of the logic transistors (Figure 4.15a, Figure 4.15b and Figure 4.15c) increases with a decrease in transistor threshold voltage. Reduction in threshold voltage makes it harder to fully turn off the transistor, hence the level of sub-threshold current flowing in the channel at $V_{GS}=0$ increases. Sub-threshold currents increase exponentially with temperature and roughly by an order of magnitude for every *100mV* decrease in threshold voltage [3]. For S-V_T, L-V_T and SRAM-PD transistors, sub-threshold leakage is so heavily dominant even at reduced temperatures that GIDL and gate leakage can be neglected.

In Figure 4.15c, the leakage of the L-V_T footer exceeds that of the block in the MOST LEAKY corner for temperature values $T \geq 34°C$. Unlike the H-V_T, S-V_T and SRAM-PD transistors, the use of this footer will not permit V_{SRC} to rise all the way from *10mV* in the ACTIVE mode to $V_{DD} - 0.7V$ in the STANDBY mode. Instead, V_{SRC} will be maintained at a value less than $V_{DD} - 0.7V$ during STANDBY. The obtained result does not suggest not using the L-V_T footer under any circumstances, what it does suggest is that its use would not allow maximum leakage reduction to be achieved during STANDBY. The L-V_T footer therefore falls lower on the merit list compared to the other three candidates.

4.6 SRC node voltage during STANDBY

4.6.1 Simulation setup

This simulation (setup shown in Figure 4.16) has been included to prove that if the leakage of the switch exceeds that of the block at $V_{SRC} = V_{DD} - 0.7V$, the SRC node would not be able to float up to $V_{DD} - 0.7V$ during STANDBY. The voltage at the node would go on to settle at a level lower than $V_{DD} - 0.7V$. As a result, the full potential of the leakage reduction technique (source biasing) would not be realized.

The word lines are not held in the active state for the entire duration of the memory cycle time. As soon as the sense amplifiers start sensing the differential voltage on the bit lines, the word lines are switched off to save power [2]. Therefore, by the time the footer is turned off to enter the STANDBY mode in the proposed architecture the word lines would already be in the inactive state. A transient simulation is set up in which the block supply node and the bit lines are tied to V_{DD}, the word lines are tied to *GND* and the internal storage nodes of the cells initialized to V_{DD} and *GND*. The SRC node is also initialized to *GND* potential since no read current is flowing then. When the gate of the footer is deactivated at $t=0$, SRC node voltage begins to rise as a consequence of leakage charging it. Slowly the node floats up to and beyond $V_{DD} - 0.7V$.

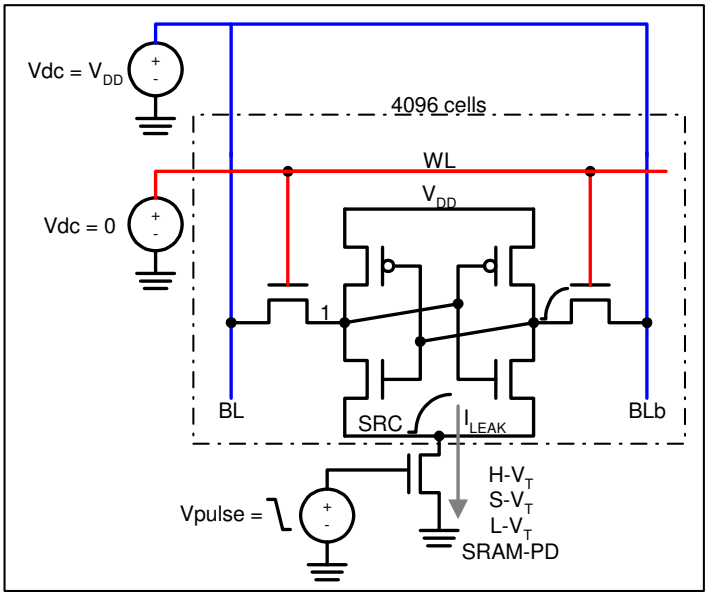


Figure 4.16 Floating the SRC node

4.6.2 Results and discussion

The moment the gate of the footer is deactivated the SRC node voltage temporarily falls below zero due to Miller effect in the footer before the block leakage starts to charge the SRC node capacitance. This effect is visible in Figure 4.17 immediately after $t=0$.

The voltage rises the fastest in the MOST LEAKY corner due to the highest magnitude of leakage current flowing through the memory block. It is clearly visible from the curves that V_{SRC}, in case of L-V_T footer, does not reach $V_{DD} - 0.7V$. This is evidence of the results observed in Figure 4.15c.

Within the remaining curves, the time taken to reach $V_{DD} - 0.7V$ is least in case of the H-V_T footer and highest for the SRAM-PD footer. This is because the H-V_T footer has the highest off-resistance, hence the lowest leakage. The SRAM-PD footer has the lowest off-resistance, hence the highest leakage.

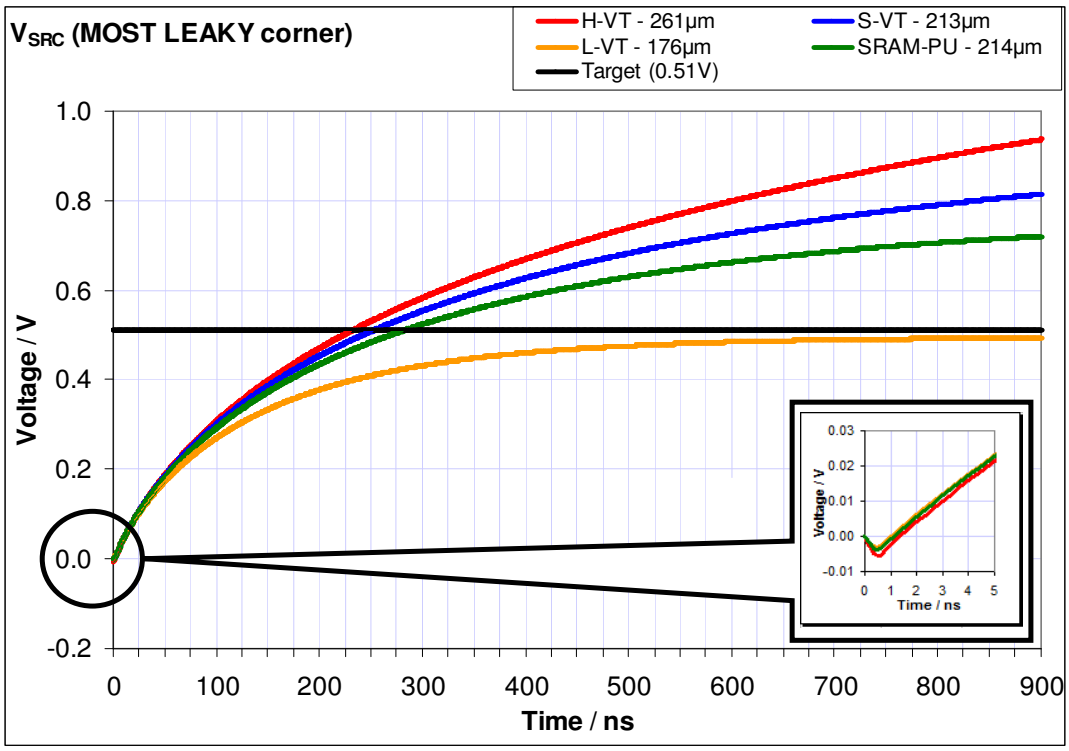


Figure 4.17 V_{SRC} during STANDBY mode (MOST LEAKY corner)

The simulation was repeated for the NOMINAL and LEAST LEAKY corners, where the trend of the curves was found to be similar to that in Figure 4.17, except that the time to reach $V_{DD} - 0.7V$ was greater. It was highest in the LEAST LEAKY corner due to lowest levels of leakage current flowing through the block.

The curves in Figure 4.17 show V_{SRC} rising beyond $V_{DD} - 0.7V$ mark, which would lead one to believe that the memory would eventually lose its data. Note that this condition will not occur in practice since a DRV maintaining circuit shall be employed to clamp the rising SRC node voltage to $V_{DD} - 0.7V$. This is discussed in chapter 7.

4.7 Summary and conclusion

As outlined earlier in section 4.3, the selection of the optimal footer is governed by the following three parameters:

i) V_{SRC} in ACTIVE mode
ii) The time required to pull down the SRC node from $V_{DD} - 0.7V$ to $10mV$
iii) Leakage current of the block with respect to that of the switch during STANDBY mode

Simulation results have shown that out of the three parameters influencing switch size, the first is the most critical and therefore governs the switch size. Three out of the investigated four footer types are suitable candidates for a footer switch that can be used in conjunction with a memory block. Since the L-V_T footer would not allow maximum leakage reduction to be achieved during STANDBY, it should therefore not be used. Out of the remaining three, the S-V_T switch is nearly the same size as the SRAM-PD switch. The advantage of using the SRAM-PD transistor is that it would accurately track process variation in the memory block, particularly ΔV_T. The H-V_T switch is simply too large without any obvious advantage over the SRAM-PD switch. A summary of the results is presented in Table 4.1. All transistors have a length of *55nm*. At this stage, the logical switch choice would be the SRAM-PD transistor.

Table 4.1 Summary of results

	Footer type			
	High-V_T	Standard-V_T	Low-V_T	SRAM-PD
Minimum Size	261µm	213µm	176µm	214µm

In the following chapter, the behavior of the memory block, during ACTIVE and STANDBY modes, is investigated with the use of a header switch.

4.8 References

[1] NXP Intranet, "*Memory estimator for 45nm Embedded Memories*" 2009

[2] A. Pavlov, M. Sachdev, "CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies: Process-Aware SRAM Design and Test," Springer, 2008

[3] Deepaksubramanyan, B.S., Nunez, A., *Analysis of subthreshold leakage reduction in CMOS digital circuits*, Circuits and Systems 2007, MWSCAS 2007, 50th Midwest Symposium on Volume, Issue , 5-8 Aug 2007, pp. 1400-1404

Chapter 5

Header switch analysis with a memory block

In this chapter the behavior of a P-channel header switch is investigated when used in conjunction with a *4096*-cell memory block. Similar to the case of footer, four different types of P-channel switches are analyzed with the memory block: high threshold voltage (H-V_T), standard threshold voltage (S-V_T), low threshold voltage (L-V_T), and SRAM cell pull up transistor (SRAM-PU). The length of all transistors is fixed at *l=55nm*, (the length of the SRAM-PU transistor), for reasons specified in chapter *4*.

5.1 Factors governing header size and type

When a header is used in combination with a memory block, three critical parameters influence the selection of the optimal switch in terms of type and size. These include:

vi) The time required to pull up the SUP node of the memory block from *0.7V* to $V_{DD} - 10mV$ (i.e. the time needed to bring the block into ACTIVE mode from STANDBY mode)

vii) SUP node voltage of the cell/block in the ACTIVE mode

viii) Leakage current of the block with respect to that of the switch during STANDBY mode

Detailed simulations are carried out to fully understand the role each of the above-mentioned parameters. A comprehensive account of the simulation setup and a thorough discussion on the results obtained is presented in Sections 5.2 – 5.5.

5.2 SUP node pull up time (STANDBY to ACTIVE mode)

5.2.1 Simulation setup

Similar to the footer the time taken by the header to pull up the SUP node from *0.7V* to V_{DD} – *10mV* contributes towards the access time of the memory. A transient simulation is set up where all word lines in the block are held inactive till *t=0*, the bit lines are tied to V_{DD} and the SUP node is initialized to *0.7V* in order to simulate STANDBY state. The internal storage nodes of all cells are also initialized to *0.7V* and *GND*. Figure 5.1 illustrates the simulation principle.

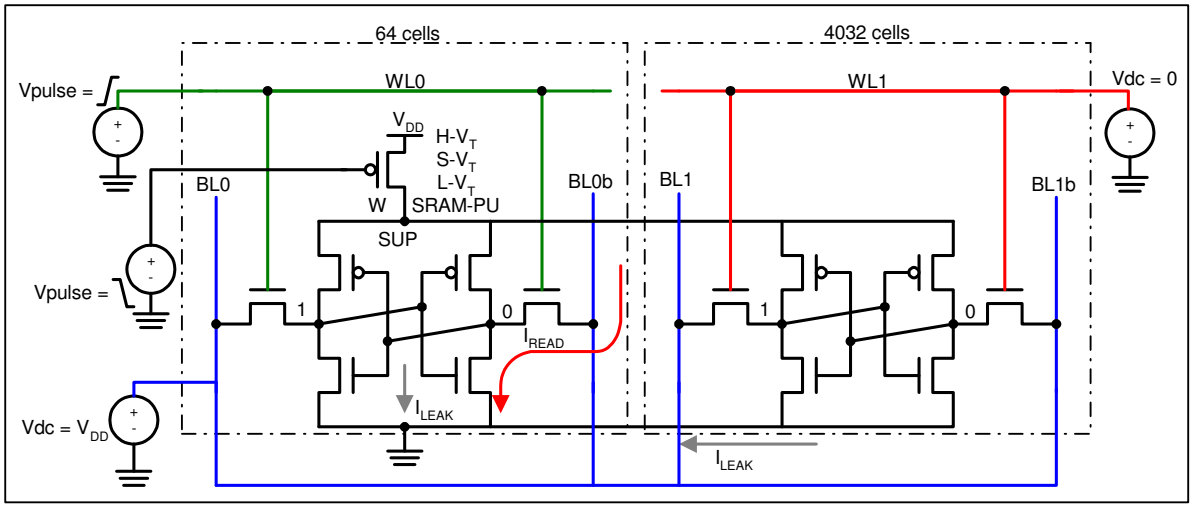


Figure 5.1 Sizing the header

At *t=0*, word line WL0 is driven high and the gate terminal of the header is driven low. The header pulls up the SUP node voltage from *0.7V* towards V_{DD} and the simultaneous activation of the word line simulates the start of a read operation. The simulation is carried out for the four types of P-channel headers. Also, the supply voltage is varied from $V_{DD} = 0.9V_{DD(nom)}$ to $V_{DD} = 1.1V_{DD(nom)}$. For each supply voltage level, the width of the header is swept in order to achieve the value that guarantees a maximum pull up time of *1ns* in the worst case (for reasons explained in section 4.5.1). The time to pull up the SUP node to within *10mV* of V_{DD} is recorded.

5.2.2 Results and discussion

5.2.2.1 High-V_T header

Figure 5.2 shows the SUP node pull up time plotted as a function of header width in different corners, when an H-V_T header is used. The pull up time is highest in the SLOW corner as a consequence of reduced drive capability of the transistor. Reduced supply voltage leads to a lower gate overdrive voltage ($|V_{GS}| - |V_T|$), and increased temperature leads to lower carrier mobility as a result of higher scattering in the channel.

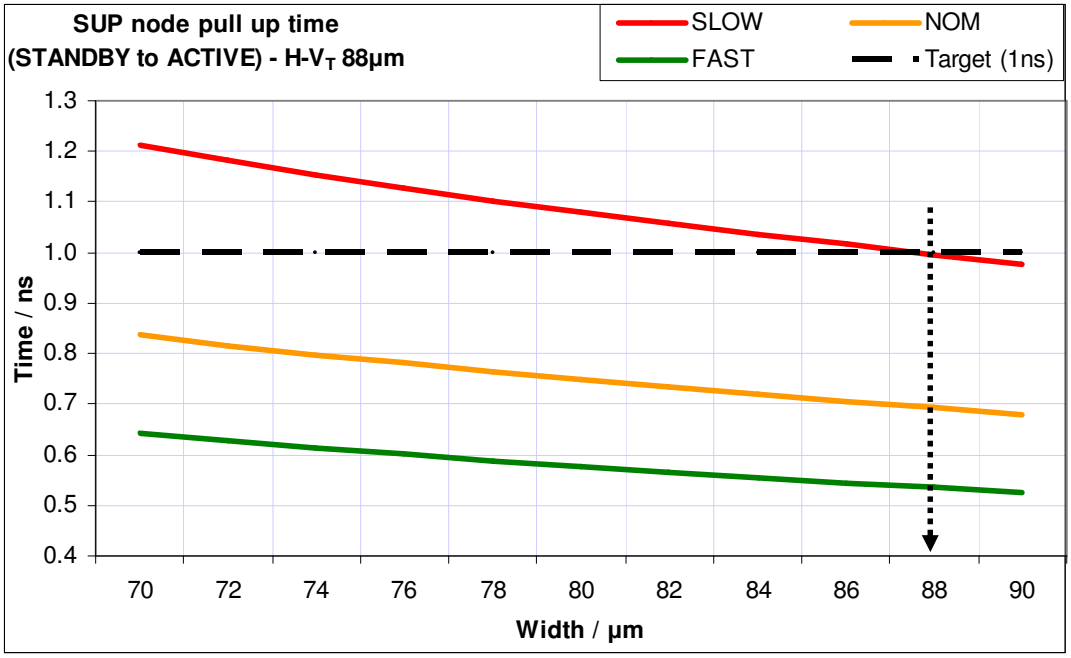


Figure 5.2 SUP node pull up time with H-V_T header

A width of *88µm* guarantees a worst case pull up time of *1ns* in the SLOW corner. The general trend of the curves can be explained by the fact that the current drive of the transistors increases with an increase in width, hence the time required to pull up the SUP node decreases with increasing transistor size.

5.2.2.2 Standard-V_T header

Figure 5.3 shows the SUP node pull up time plotted as a function of header width in different corners, when an S-V_T header is used. A width of *63.5µm* guarantees a worst case pull up time of *1ns* in the SLOW corner. A smaller S-V_T transistor compared to its H-V_T counterpart suffices because with a lower threshold voltage, the gate overdrive voltage is higher.

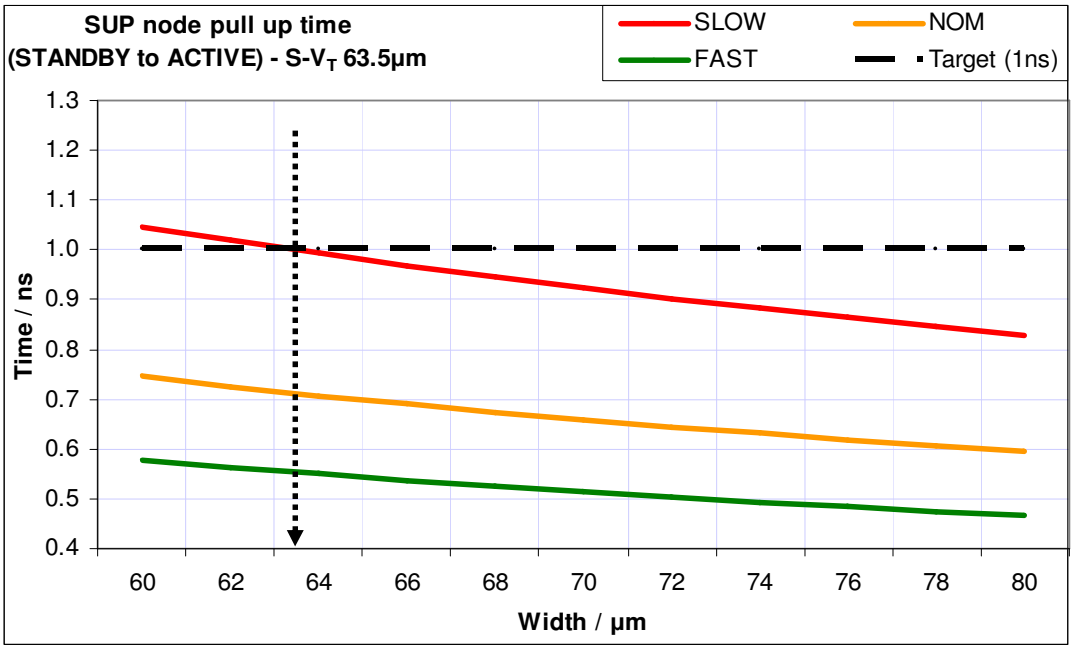


Figure 5.3 SUP node pull up time with S-V_T header

5.2.2.3 Low-V_T header

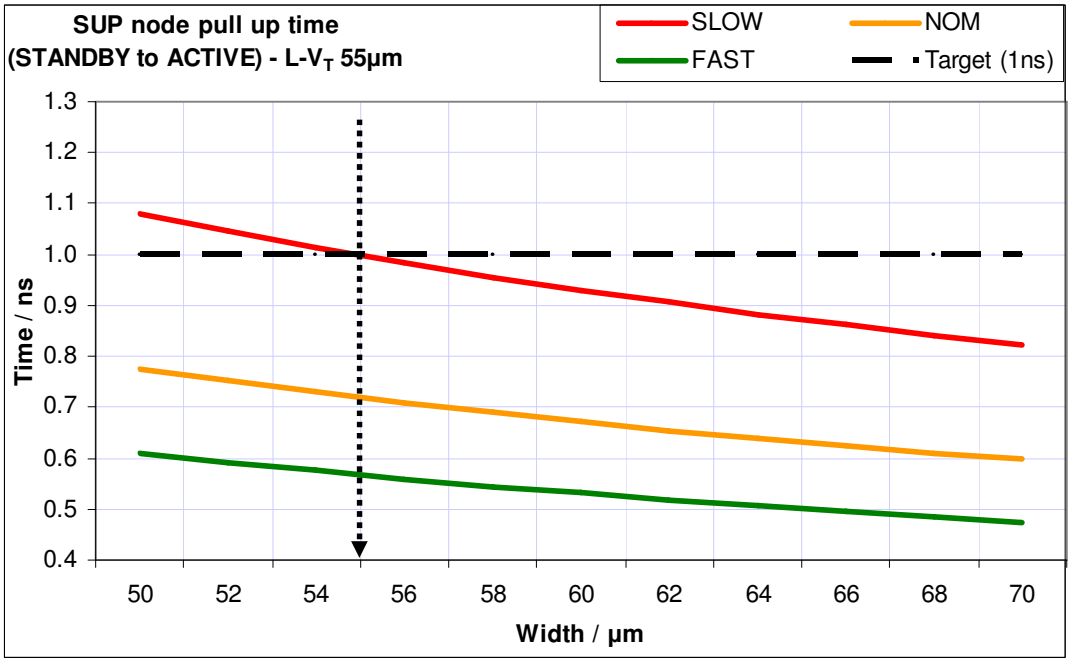


Figure 5.4 SUP node pull up time with L-V_T header

Figure 5.4 shows the SUP node pull up time plotted as a function of header width in different corners, when an L-V_T header is used. A width of *55µm* guarantees a worst case pull up time of *1ns* in the

SLOW corner. An even smaller L-V_T transistor compared to its S-V_T counterpart suffices for the reasons presented in section 5.2.2.2.

5.2.2.4 SRAM-PU header

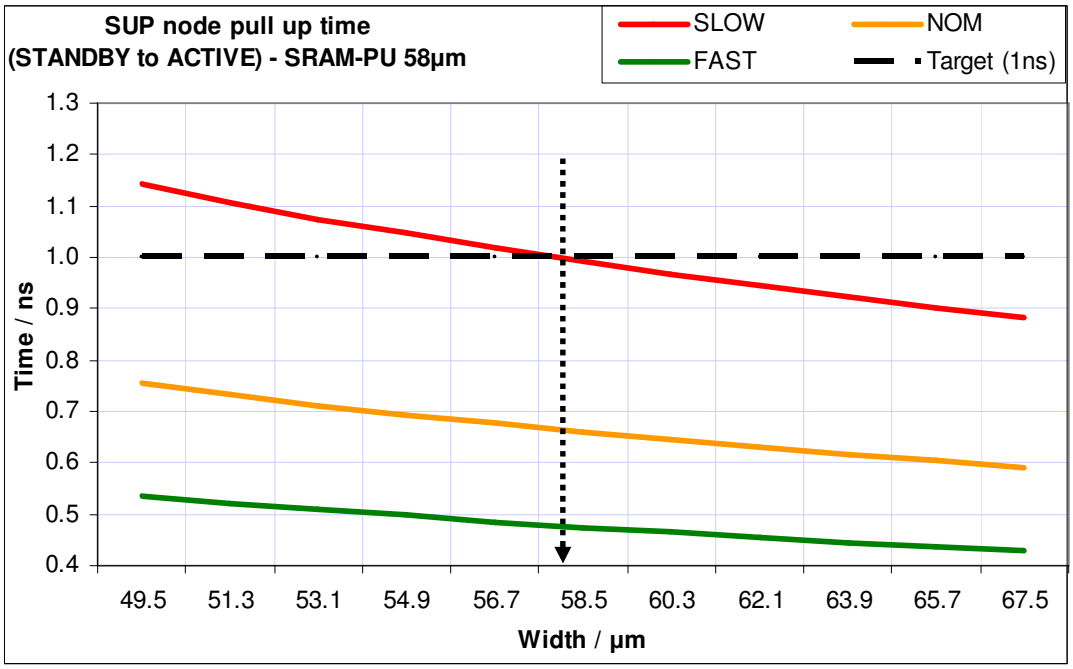


Figure 5.5 SUP node pull up time with SRAM-PU header

Figure 5.5 shows the SUP node pull up time plotted as a function of header width in different corners, when an SRAM-PU header is used. A multiplication factor of *645* (*W=58µm*) guarantees a worst case pull up time of *1ns* in the SLOW corner.

5.3 SUP node voltage (V_{SUP}) in the ACTIVE mode

5.3.1 Simulation setup

During a read/write operation, the voltage at the SUP node needs to be sufficiently close to supply rail potential, V_{DD}. Similar to the footer, the maximum drain-source potential allowed across a header switch is *10mV*. The header size (obtained from the simulation setup in section 5.2.1) must ensure the SUP node voltage remains within *10mV* of V_{DD} during a read/write operation in all corners, over the entire temperature range and across supply voltage variation.

For each type of P-channel switch, a DC simulation is set up in which one word line (WL0) is driven high to activate *64* cells (an entire row) in the memory block, all other word lines (represented by WL1) are kept inactive. Figure 5.6 illustrates the simulation setup.

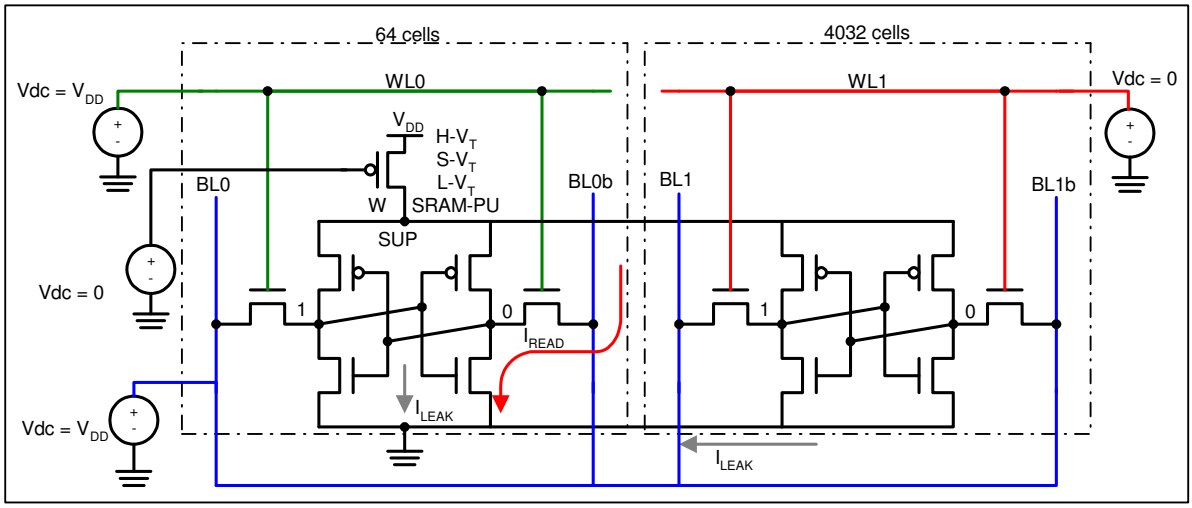


Figure 5.6 V_{SUP} in ACTIVE mode

The header gate is driven low, the bit lines are tied to V_{DD}, the NMOS bulk is tied to *GND*, the PMOS bulk is tied to V_{DD} and the SUP node is initialized to $V_{DD} - 10mV$. The temperature is swept over the entire operating range. Also, the supply voltage is varied from $V_{DD} = 0.9V_{DD(nom)}$ to $V_{DD} = 1.1V_{DD(nom)}$. For each supply voltage level, V_{SUP} is recorded as a function of temperature.

5.3.2 Results and discussion

Figure 5.7 shows SUP node voltage plotted as a function of temperature for FAST, NOMINAL and SLOW corners and with supply voltage variation, when using H-V_T header.

V_{SUP} scales with supply voltage variation, but decreases (varies) the most in the FAST corner because the on-resistance of the memory block decreases far more than the on-resistance of the header. The level to which V_{SUP} falls and the characteristic shape of the curves can be linked to the two key parameters that vary with temperature: threshold voltage and mobility. The graphs are presented to indicate that V_{SUP} stays well within $10mV$ of V_{DD} in all corners, over the entire operating temperature range and across supply voltage variation.

Figure 5.8 and Figure 5.9 show V_{SUP} plotted as a function of temperature for FAST, NOMINAL and SLOW corners and with supply voltage variation, when using S-V_T and L-V_T headers respectively.

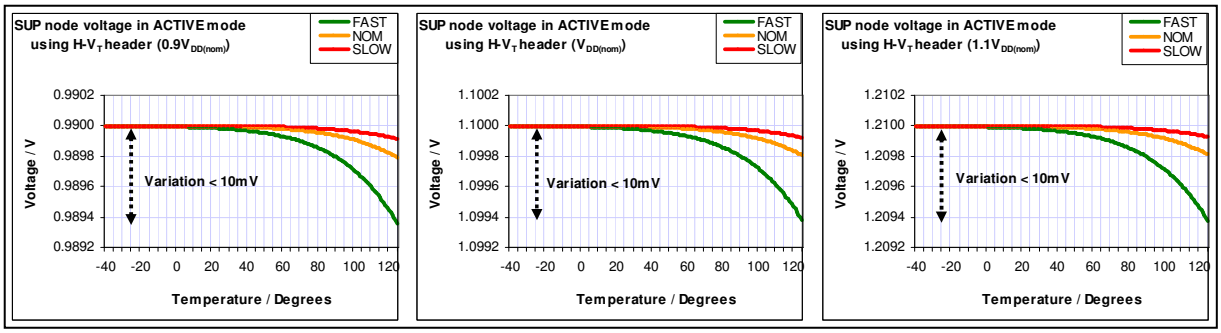


Figure 5.7 V_{SUP} in ACTIVE mode using H-V_T header at different supply voltage levels

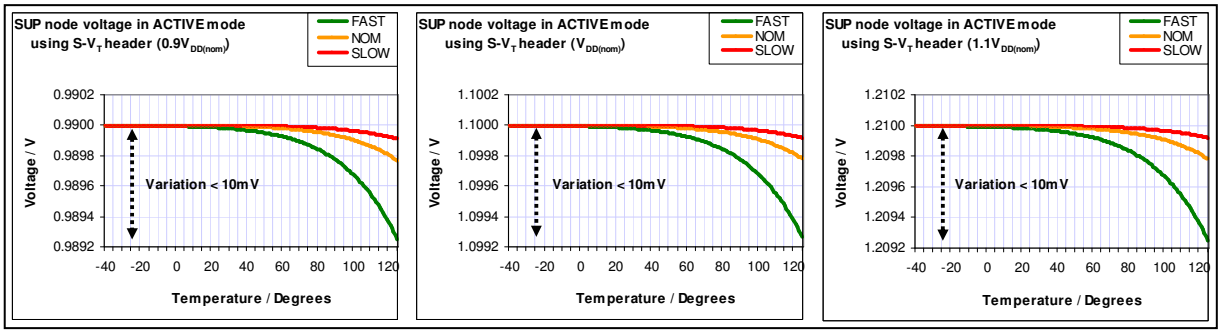


Figure 5.8 V_{SUP} in ACTIVE mode using S-V_T header at different supply voltage levels

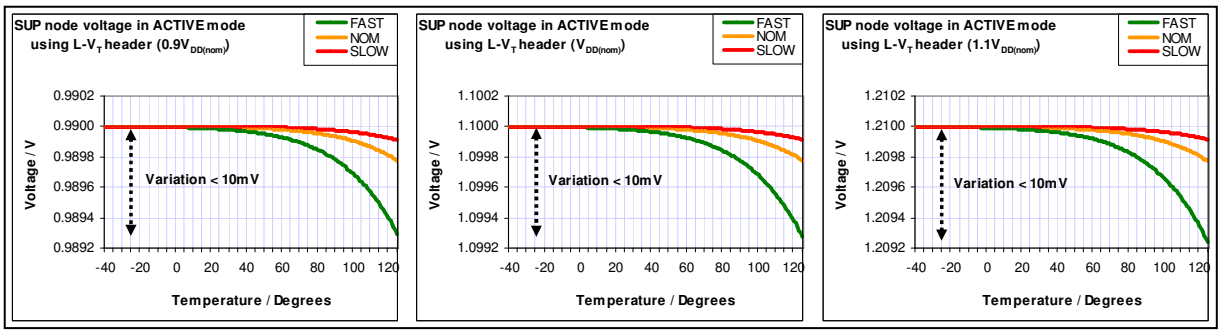


Figure 5.9 V_{SUP} in ACTIVE mode using L-V_T header at different supply voltage levels

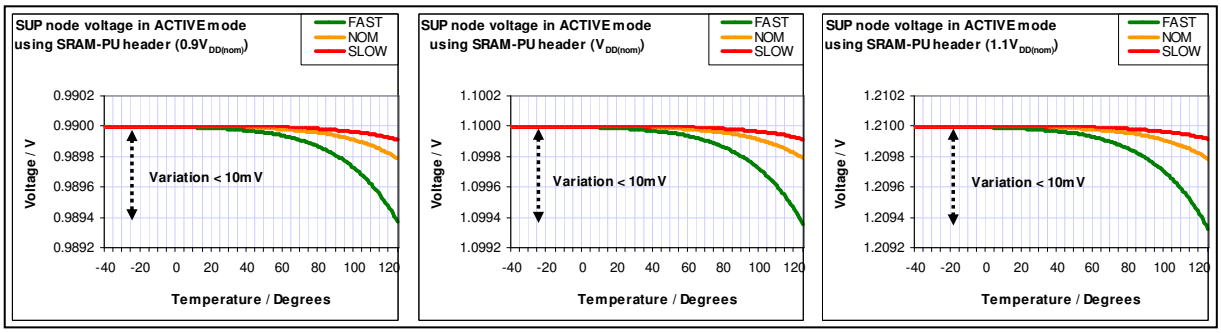


Figure 5.10 V_{SUP} in ACTIVE mode using SRAM-PU header at different supply voltage levels

Figure 5.10 shows V_{SUP} plotted as a function of temperature for FAST, NOMINAL and SLOW corners and with supply voltage variation, when using SRAM-PU header.

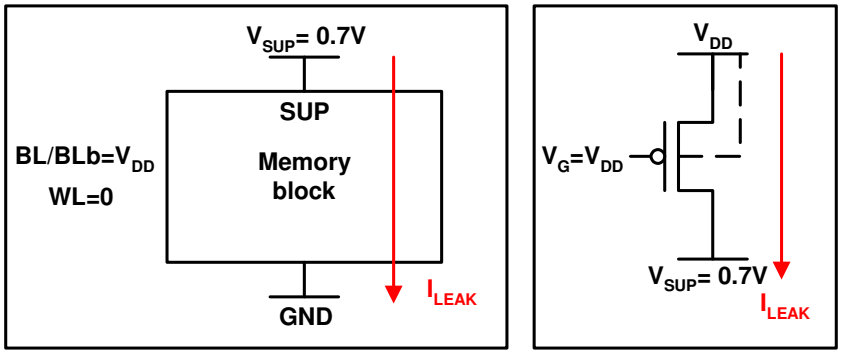
Simulation results presented in this section clearly demonstrate that the header sizes obtained through the simulation setup outlined in section 5.2.1 do not pose a limitation on V_{SUP} in the ACTIVE mode.

5.4 Block and switch leakage currents (STANDBY mode)

5.4.1 Simulation setup

Like in the case of the footer switch, the intention here is also to be able to maintain DRV across the memory block during STANDBY. In order to achieve maximum benefit out of source biasing leakage reduction technique, the switch must leak less than the block when V_{SUP} is maintained at $0.7V$ during STANDBY. This condition must be met in all corners, across supply voltage variation and over the entire temperature scale. In case the leakage of the switch exceeds that of the block at the condition specified above, then the actual potential across the block would be higher than the DRV and maximum leakage reduction would not be achieved.

Two independent DC simulations are set up, similar to those carried out for the block and footer. Figure 5.11 illustrates the leakage current simulation principle for block and header. The total leakage of the block and the switch as a function of temperature is then recorded in different corners.



(a) (b)

Figure 5.11 Block (a) and header (b) leakage current simulation principle

5.4.2 Results and discussion

Figure 5.12 shows block and switch leakage plotted as a function of temperature for each type of switch. The highest leakage, both for the block and the switch, occurs in the MOST LEAKY corner and the lowest leakage occurs in the LEAST LEAKY corner.

It is clear from Figure 5.12 that the use of all header types will permit V_{SUP} to fall all the way from V_{DD} – *10mV* in the ACTIVE mode to *0.7V* in the STANDBY mode. Therefore, they are all viable candidates to be used in a header role.

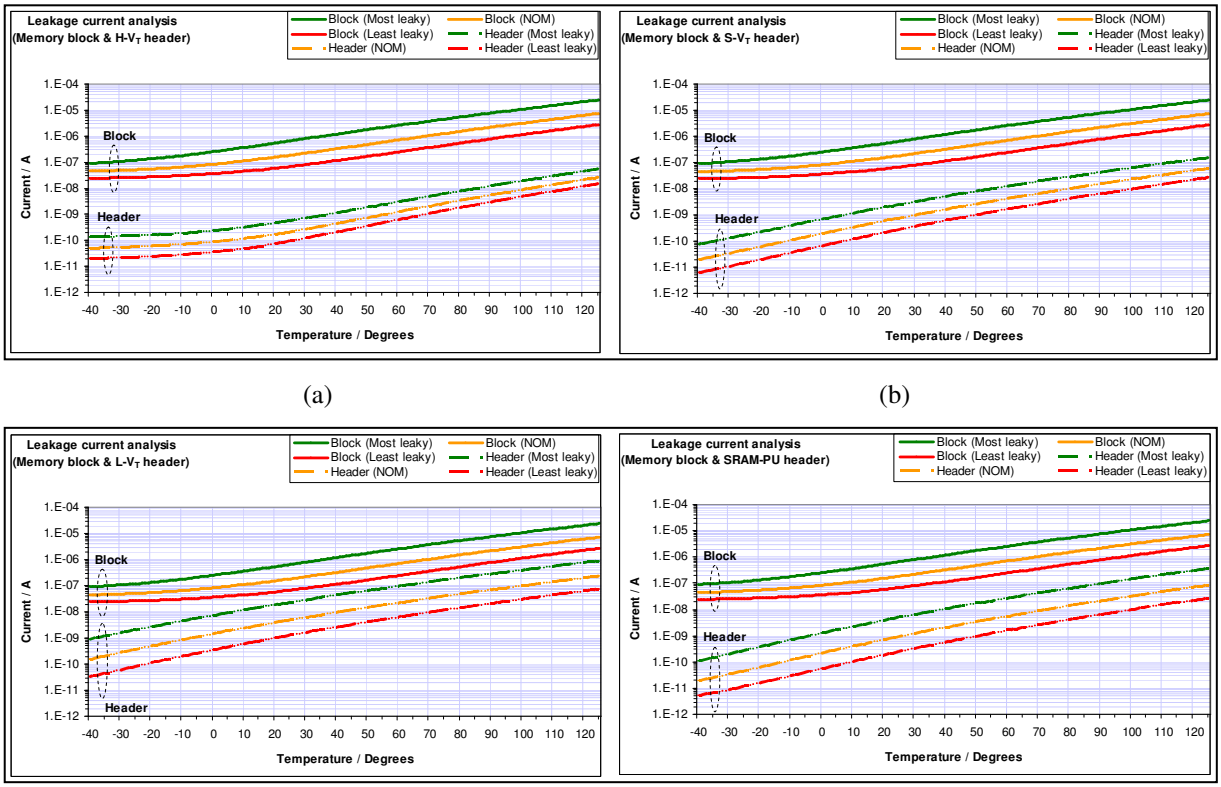


Figure 5.12 Block and H-V_T switch leakage (a) Block and S-V_T switch leakage (b) Block and L-V_T switch leakage (c) and Block and SRAM-PU switch leakage (d)

5.5 SUP node voltage during STANDBY

5.5.1 Simulation setup

This simulation (setup outlined in Figure 5.13) has been included to prove that when the leakage of the switch is less than that of the block at V_{SUP} = *0.7V*, the SUP node would always be able to float down to *0.7V* during STANDBY. This would allow maximum benefit to be achieved out of source biasing technique.

A transient simulation, comparable in setup to the one outlined in section 4.7.1, is executed for the memory block and a header switch.

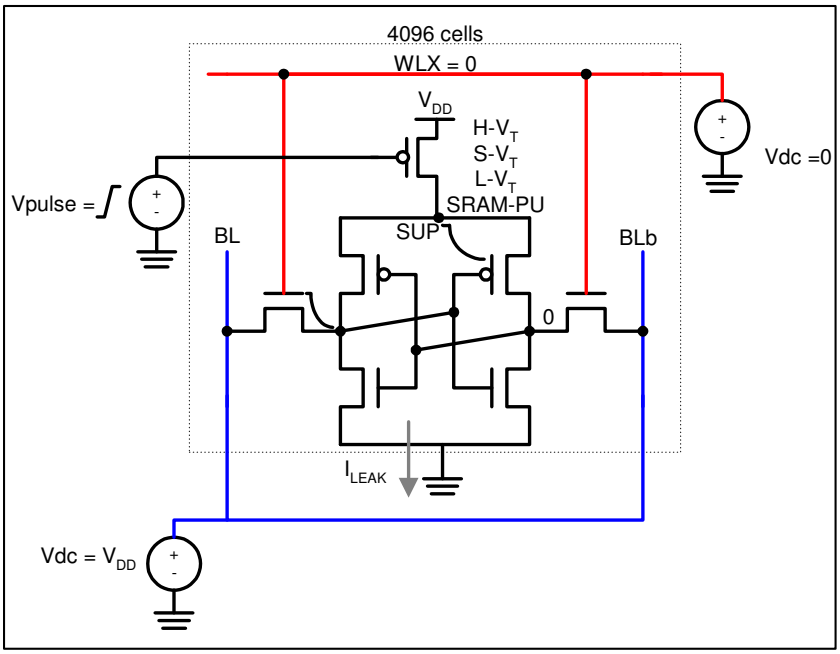


Figure 5.13 Floating the SUP node

5.6 Results and discussion

The moment the gate of the header is deactivated, the SUP node voltage temporarily rises above V_{DD} due to Miller effect before the block starts to lose charge. This effect is visible in Figure 5.14 immediately after *t=0*.

Soon after *t=0*, the block starts to lose its charge since the SUP node is placed in a floating state. As the SUP node voltage begins to fall, the header $|V_{DS}|$ gradually increases and the header starts to leak as well. The SUP node voltage decreases at approximately the same rate for all types of headers in a given corner. That is because the node voltage fall time is determined by how quickly the memory block loses its charge. The rate at which the voltage drops is independent of the type of header incorporated in the circuit. Simulations repeated for the NOMINAL and LEAST LEAKY corners showed that the SUP node voltage fall time was the highest in the LEAST LEAKY corner due to lowest levels of leakage current flowing through the block.

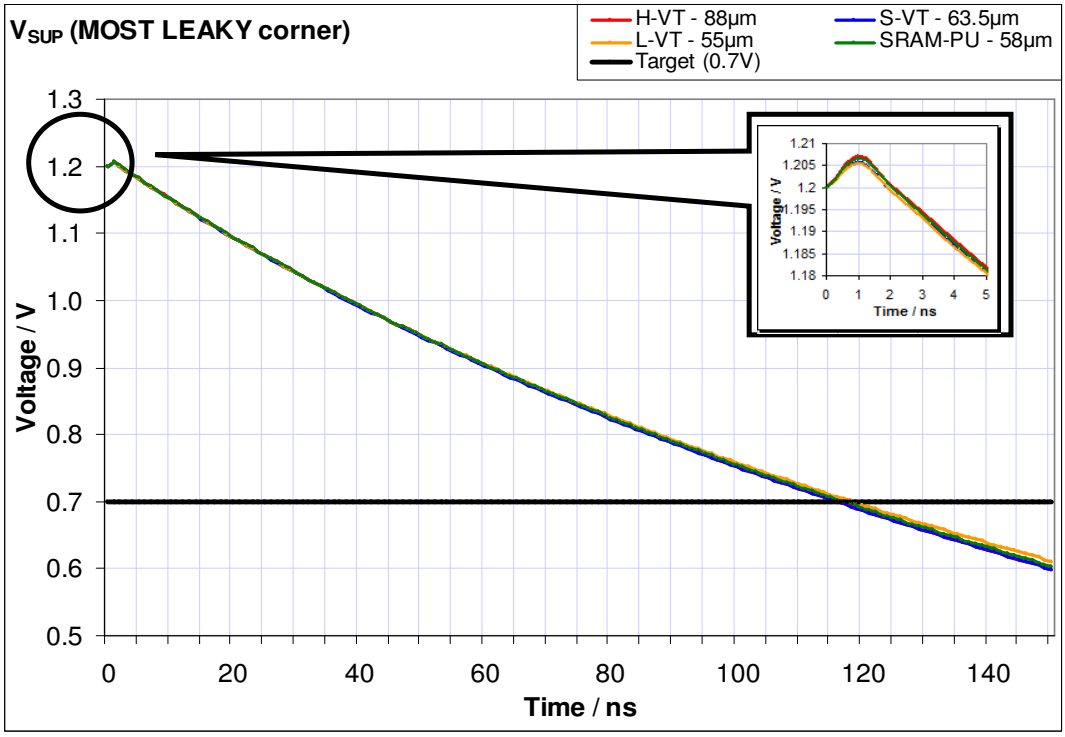


Figure 5.14 V_{SUP} during STANDBY mode (MOST LEAKY corner)

The curves in Figure 5.14 show V_{SUP} dropping below the DRV level (*0.7V*), which would lead one to believe that the memory would eventually lose its data. Note that this condition will not occur in practice since a DRV maintaining circuit shall be employed to clamp the falling SUP node voltage to the DRV level. This is discussed in chapter 7.

5.7 Summary and conclusion

As outlined earlier in section 5.1, the selection of the optimal header is governed by the following three parameters:

i) The time required to pull up the SUP node from *0.7V* to V_{DD} – *10mV*

ii) V_{SUP} in ACTIVE mode

iii) Leakage current of the block with respect to that of the switch during STANDBY mode

Simulation results have shown that out of the three parameters influencing switch size, time required to pull up SUP node when switching from STANDBY to ACTIVE mode is the most critical and therefore governs the switch size. All header types investigated are suitable candidates for a switch that can be used in conjunction with a memory block. The L-V_T switch is slightly smaller than the SRAM-PU switch. The advantage of using the SRAM-PU transistor is that it would accurately track process variation in the memory block, particularly ΔV_T. The S-V_T and H-V_T switches are large in size

without any obvious advantage over the SRAM-PU switch. A summary of the results is presented in Table 5.1. All transistors have a length of *55nm*. At this stage, the logical switch choice would be the SRAM-PU transistor, despite being a little larger than the L-V_T switch.

Table 5.1 Summary of results

	Header type			
	High-V_T	Standard-V_T	Low-V_T	SRAM-PU
Minimum Size	88µm	63.5µm	55µm	58µm

In the following chapter, a model is developed to estimate the area overhead associated with each type of switch, compared to the area of the memory block. In order to evaluate the area-speed-power trade-off associated with the proposed design, an estimate of the switch area overhead is therefore necessary.

Chapter 6

Area overhead estimation & switch selection

In chapters *4* and *5*, minimum sizes for various footer and header types were identified that were suitable to be used as a switch in conjunction with a *4096*-cell memory block. Since the total benefit of reduced leakage during STANDBY mode can only be realized at the cost of increased area (that of the switch for now), it becomes necessary to estimate the total area overhead associated with each kind of switch. In this chapter therefore, a model is developed that estimates the total switch area overhead as a function of number of fingers in layout. In order to better track process variation in the block, multiple instances of the exact SRAM cell load and driver transistors were used in the switch in chapters *4* and *5*, allowing duplication of the memory cell transistor outside the memory matrix. Since the logic transistors are fundamentally different, i.e. in terms of V_T implant, their selected sizes can be implemented using a wide range of finger widths in layout. For the logic transistors in particular, it is therefore important to know how the area overhead would vary with decreasing finger width (i.e. from the maximum allowed OD width down to the width of the SRAM transistors).

The largest device width possible from a layout perspective in *45nm* technology is a little less than *18µm* (*17.82µm* for this model allowing room for a POLY contact point on either side). This limitation is set by POLY layout rules for the stated technology [1], according to which the maximum length of POLY between two POLY contact points cannot exceed *18µm*. Increasing this distance beyond the specified limit has an adverse effect on POLY resistance value and hence the signal propagation time

along the length of POLY, which can lead to degradation of device performance. Hence for widths exceeding *17.82μm*, multiple devices are placed in parallel, with POLY running horizontally.

6.1 Overview and limitation of the model

The model exploits TSMC *45nm* technology layout rules for area estimation. As stated in chapter 3, the memory block is organized as an array of *64* columns by *64* rows. Each Non-Cell-Implant High-Speed memory cell measures *1μm* by *0.374μm* on layout, giving an area of $0.374\mu m^2$. The area of the memory block is therefore $1531.9\mu m^2$. The idea is to be able to lay the footer or header directly above (in case of header) or below (in case of footer) the memory block and preferably within the block width of *64μm* in order to reduce the effect of current crowding. For a given OD width, the proposed model therefore lays out the fingers in parallel along the width of the memory block (with POLY running horizontally) and gradually adds additional finger rows in order to realize the required switch size. Note that POLY is routed horizontally within the SRAM cell. In *45nm* technology, it is recommended that POLY be laid out in the same direction throughout the design due to better results achieved during lithographic patterning of POLY. Therefore, in order to be consistent, POLY has been routed horizontally for the switch also.

The threshold voltage of a transistor with three fingers, each of width *W* is less than the threshold voltage of a transistor with one finger of width *3W* due to narrow channel effects. A comprehensive study of threshold voltage variation with OD width is outside the scope of this work. The proposed model therefore does not take into account the variation in threshold voltage with changing OD width of individual fingers. It is therefore assumed that the total required header/footer width does not change significantly if the layout implementation is varied over different number of fingers. The above-stated limitation does not apply to the SRAM transistors since multiple instances of the same have been used to realize the switch. Derivation of the model is discussed in detail in sections to follow.

6.2 Model derivation

Figure 6.1 illustrates the proposed layout of a footer switch. The variables used in the illustration match those used in the equations presented in the following sub-sections to aid understanding of the model.

6.2.1 Number of fingers, width of OD and number of rows

The width of an individual finger for a given number of fingers N_f is given by Equation (6.1), where W_{req} is the total width of the switch.

$$W_f = \left(\frac{W_{req}}{N_f} \right) \tag{6.1}$$

The total width of active (OD) region is fixed for a specified switch size, irrespective of the number of fingers, and is given by Equation (6.2).

$$W_{T(OD)} = N_f \cdot W_f \tag{6.2}$$

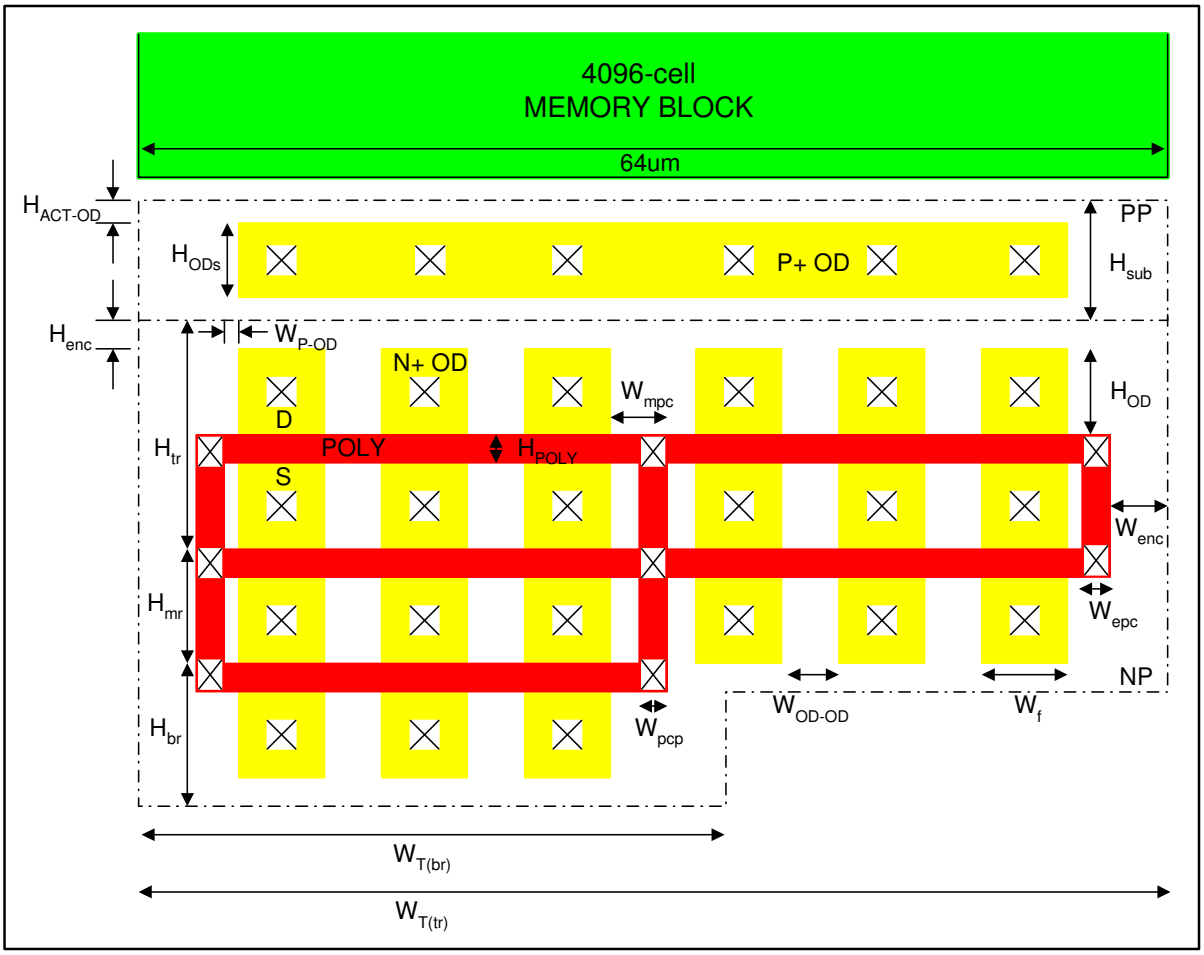


Figure 6.1 Proposed layout of a footer switch (not drawn to scale)

When two or more fingers are laid out in parallel (along the width of the memory block), OD-OD spacings ($W_{OD\text{-}OD}$ = *110nm* [1]) are introduced between the fingers. The total width of OD and spacings therefore is given by Equation (6.3).

$$W_T = N_f \cdot W_f + (N_f - 1) \cdot W_{OD-OD} \tag{6.3}$$

The number of rows required to implement W_T is given by Equation (6.4), where W_{blk} is the width of the memory block ($64\mu m$).

$$N_r = roundup\left(\frac{W_T}{W_{blk}}\right) \tag{6.4}$$

The number of middle rows is given by Look-up table (6.5), where ($N_r - 1$) represents number of rows other than the top row.

$$N_{mr} = \begin{cases} 0 & if \quad N_r - 1 = 0 \\ N_r - 1 - 1 & \quad N_r - 1 > 0 \end{cases} \tag{6.5}$$

6.2.2 Width due to POLY contact points

The addition of POLY contact points increases the width of a row. The number of POLY contact points in a row is governed by Look-up table (6.6), where W_r represents the width of any row without POLY contacts. When the number of POLY contact points is *2*, they represent the left and right end POLY contacts in a row. Any additional POLY contacts are placed between fingers within a row.

The minimum number of POLY contact points in a row of fingers is set to two (one at each end). This has been done to overcome a potential scenario where the POLY run is just less than the maximum allowed between two contact points. In such a situation, signal propagation time may increase to unacceptable levels for devices located at the far end of the contact point.

$$N_{pc} = \begin{cases} 0 & W_r = 0 \\ 2 & W_r \le 17.82\mu m \\ 3 \quad if & W_r \le 35.64\mu m \\ 4 & W_r \le 53.46\mu m \\ 5 & W_r > 53.46\mu m \end{cases} \tag{6.6}$$

The additional width added to a row due to inclusion of POLY contacts is given by Equation (6.7), where W_{epc} is the width due to POLY contact at the row ends and W_{mpc} is the width due to each middle POLY contact.

$$W_{pc} = 2 \cdot W_{epc} + (N_{pc} - 2) \cdot W_{mpc} \tag{6.7}$$

The width of the end and middle POLY contact point is given by Equation (6.8) and Equation (6.9) respectively, where $W_{P\text{-}OD}$ represents the POLY contact point to OD distance (*90nm* [1]) and W_{pcp} represents the width of a POLY contact point (*120nm* [1]).

$$W_{epc} = W_{P-OD} + W_{pcp} \tag{6.8}$$

$$W_{mpc} = \left(2 \cdot W_{P-OD}\right) + W_{pcp} - W_{OD-OD} \tag{6.9}$$

6.2.3 Top-level area estimation

The total width of top, middle and bottom row is given by Look-up tables (6.10 – 6.11) and Equation (6.12) respectively, where W_{enc} represents the extension of NP/PP source-drain implantation region over POLY (*110nm* [1]).

$$W_{T(tr)} = \begin{cases} W_T + W_{pc} + \left(2 \cdot W_{enc}\right) \\ W_{blk} + W_{pc} + \left(2 \cdot W_{enc}\right) \end{cases} \quad if \quad \begin{matrix} W_T \le W_{blk} \\ W_T > W_{blk} \end{matrix} \tag{6.10}$$

$$W_{T(mr)} = \begin{cases} 0 \\ W_{T(tr)} + W_{pc} + \left(2 \cdot W_{enc}\right) \end{cases} \quad if \quad \begin{matrix} N_{mr} = 0 \\ N_{mr} > 0 \end{matrix} \tag{6.11}$$

$$W_{T(br)} = remainder\left(\frac{W_T}{W_{blk}}\right) + W_{pc} + \left(2 \cdot W_{enc}\right) \tag{6.12}$$

The height and area of the bottom row is given by Equation (6.13) and Equation (6.14) respectively, where H_{OD} is the height of a source/drain region with the inclusion of a contact point (*140nm* [1]), H_{POLY} is the height of the POLY (*55nm* [1] = channel length) and H_{enc} is the extension of NP/PP region over OD (*80nm* [1]).

$$H_{br} = H_{OD} + H_{POLY} + H_{enc} \tag{6.13}$$

$$A_{br} = W_{T(br)} \cdot H_{br} \tag{6.14}$$

The height of each middle row and total area of middle rows is given by Equation (6.15) and Look-up table (6.16) respectively.

$$H_{mr} = H_{OD} + H_{POLY} \tag{6.15}$$

$$A_{mr} = \begin{cases} 0 \\ W_{T(mr)} \cdot H_{mr} \cdot N_{mr} \\ \left[(N_{mr} - 1) \cdot W_{T(mr)} \cdot H_{mr}\right] + \left[W_{T(br)} \cdot H_{mr}\right] + \left[\left(W_{T(mr)} - W_{T(br)}\right) \cdot \left(H_{mr} + H_{enc}\right)\right] \end{cases}$$

$$\begin{array}{ll} & N_{mr} = 0 \\ if & W_{T(br)} = W_{T(mr)} \\ & W_{T(br)} < W_{T(mr)} \end{array} \tag{6.16}$$

The height and area of the top row are given by Equation (6.17) and Look-up table (6.18) respectively.

$$H_{tr} = \left(2 \cdot H_{OD}\right) + H_{POLY} + H_{enc} \tag{6.17}$$

$$A_{tr} = \begin{cases} 0 & \\ W_{T(tr)} \cdot H_{tr} & \\ W_{T(tr)} \cdot \left(H_{tr} + H_{enc}\right) & \\ \left[W_{T(br)} \cdot H_{tr}\right] + \left[\left(W_{T(tr)} - W_{T(br)}\right) \cdot \left(H_{tr} + H_{enc}\right)\right] & \end{cases} \quad \begin{array}{ll} & W_{req} = 0 \\ if & N_{mr} > 0 \\ & N_r = 1 \\ & N_r = 2 \end{array} \tag{6.18}$$

The height and area of the substrate contact are given by Equations (6.19) and Equation (6.20) respectively, where H_{ODs} represents the minimum height of substrate OD with a contact point (*120nm* [1]) and $H_{ACT\text{-}OD}$ is the extension of NP/PP region over OD for substrate contact (*20nm* [1]).

$$H_{sub} = H_{ODs} + \left(2 \cdot H_{ACT-OD}\right) \tag{6.19}$$

$$A_{sub} = W_{T(tr)} \cdot H_{sub} \tag{6.20}$$

The total area of the switch is given by Equation (6.21).

$$A_{tot} = A_{tr} + A_{mr} + A_{br} + A_{sub} \tag{6.21}$$

6.3 Results and discussion

The results presented in this section are obtained from the model proposed in section 6.2. Figure 6.2 shows the total switch area plotted as a function of equal-sized fingers used to implement the required

width of the switch. The maximum finger width is fixed at *17.82um* in light of POLY layout rules. The minimum finger width is fixed at *215nm* for footer transistors and *90nm* for header transistors since these values correspond respectively to the width of SRAM-PD and SRAM-PU transistors used in the cell. It is important to state here that unlike the area plots of the logic transistors, the plots for the SRAM-PD and the SRAM-PU transistors are only valid at their highest x-axis value for reasons provided in section 6.1.

In order to better understand the model, important parameters have been plotted for the SRAM-PU header (*58um*). Figure 6.3a shows the decrease in W_f and corresponding decrease in (W_f + $W_{OD\text{-}OD}$) with increasing number of fingers. Figure 6.3b shows the total OD width, which is independent of the number of devices. Also shown is W_T, which increases linearly with N_f due to addition $W_{OD\text{-}OD}$ for every unit increment in N_f.

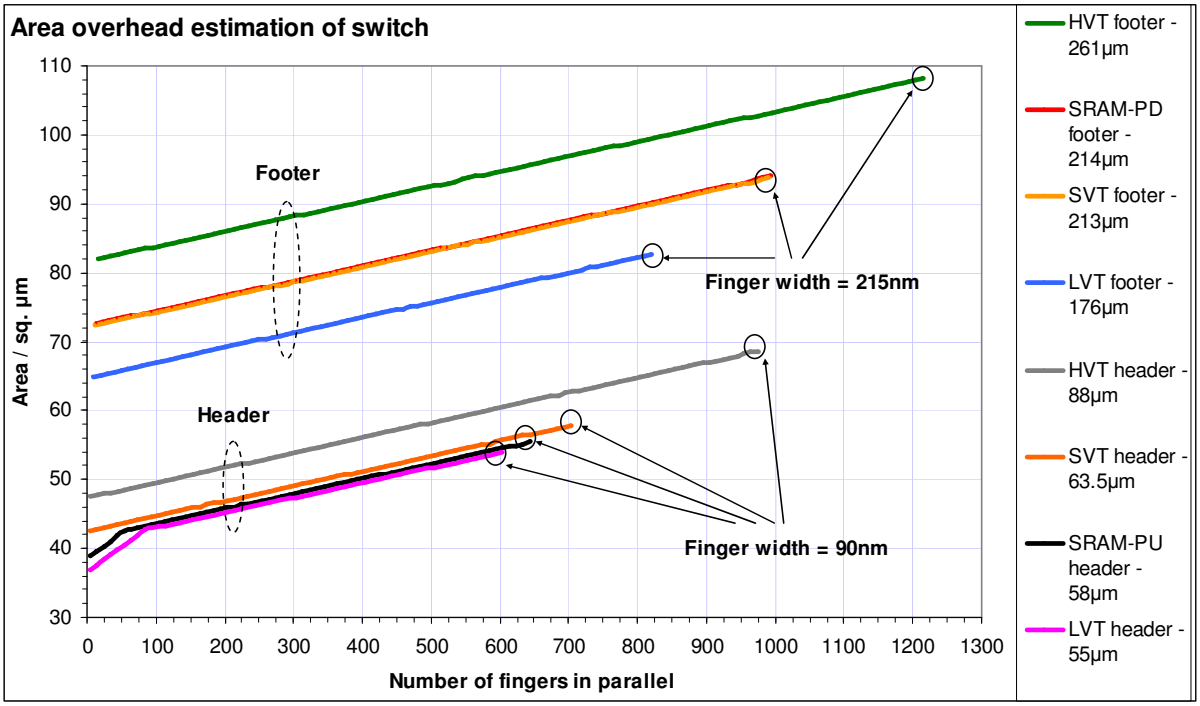
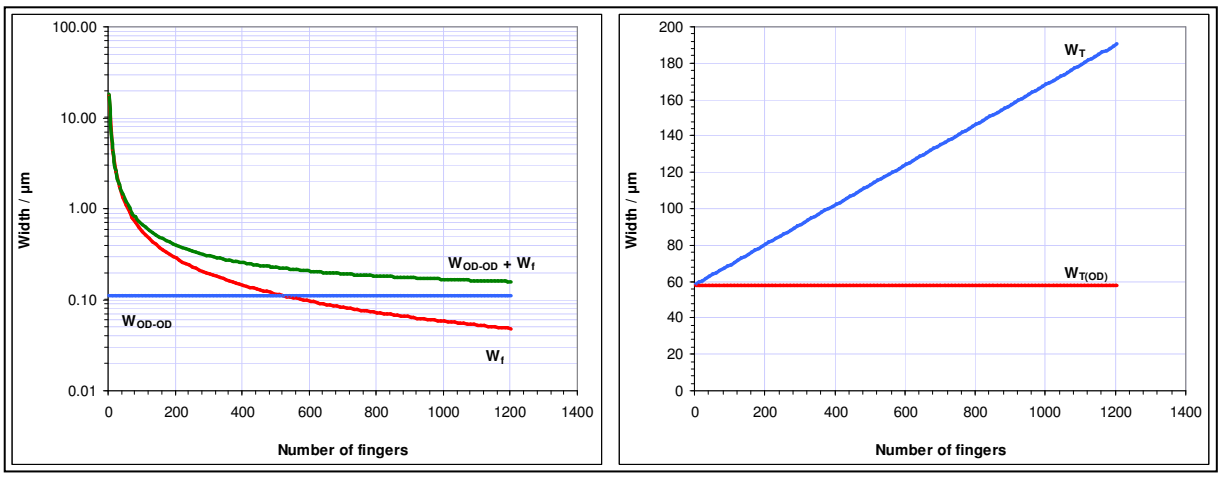


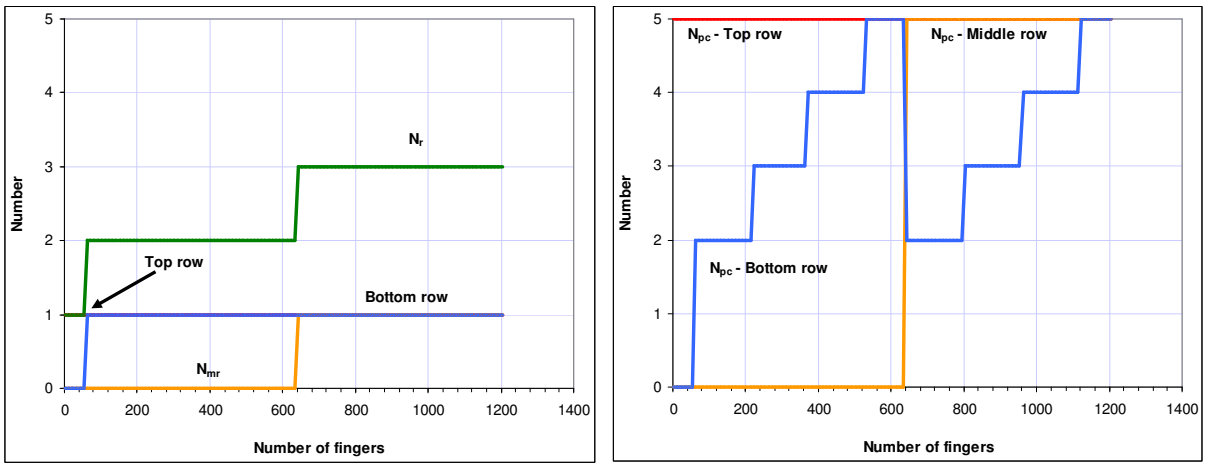
Figure 6.2 Area overhead for different header and footer types as a function of number of fingers

Figure 6.4a illustrates how the number of rows change with increasing N_f. A top row is always present if the switch size is greater than zero. A bottom row does not come into existence unless $W_{OD\text{-}OD}$ additions increase the top row width beyond the block width (*64µm*). At that point a second row is incorporated, which increases the total number of rows (N_r) to *2*. Once the second row (bottom row) hits the *64µm* mark due to decreasing finger size and continuous $W_{OD\text{-}OD}$ additions, the existing bottom row transforms into a middle row and a new bottom row comes into existence. At this point, the row count jumps to *3*.



(a) (b)

Figure 6.3 W_f, $W_{OD\text{-}OD}$ and (W_f + $W_{OD\text{-}OD}$) (a) $W_{T(OD)}$ and W_T (b)

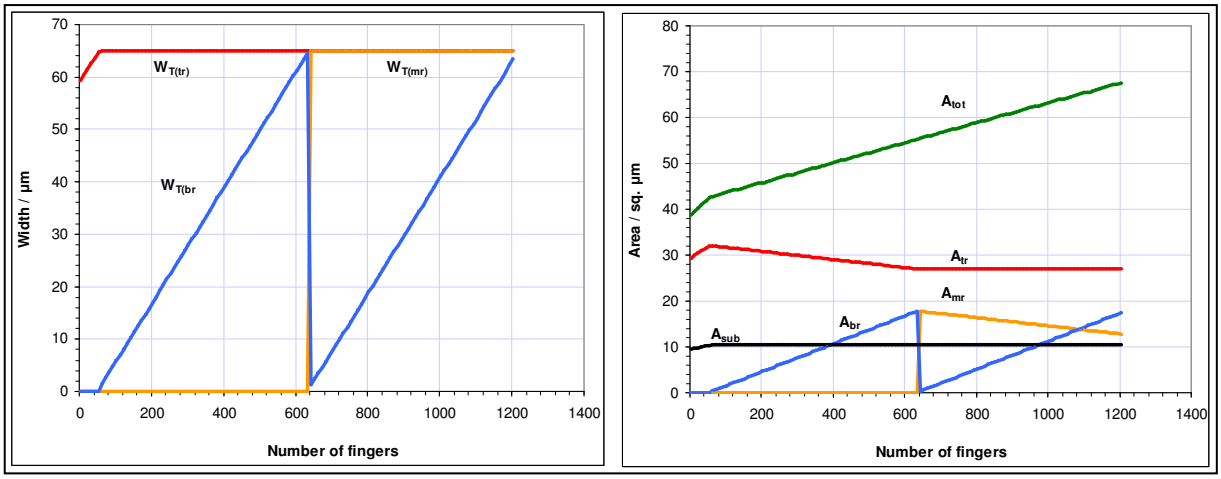


(a) (b)

Figure 6.4 Number of rows (a) number of POLY contacts (b)

Figure 6.4b shows the number of POLY contact points in each row plotted as a function of N_f. Since the width of the top row is always greater than *53.46µm* in this case, it contains a maximum of *5* POLY contact points (refer to look-up table 6.6). A bottom row does not exist initially but as the number of fingers in the top row increases, a bottom row soon comes into existence. As the width of the bottom row increases with increasing N_f, the number of POLY contact points also increase till they also hit the maximum value of *5*. There onwards, addition of a few more devices takes the width of the bottom row to the *64µm* mark, at which point the bottom row transforms into a middle row with the maximum (*5*) number of POLY contacts and a new bottom row starts to grow.

Figure 6.5a shows the total width of different rows plotted as a function of N_f. The width of the top row is initially less than the maximum (block width), but with increasing number of fingers and $W_{OD\text{-}OD}$ additions, it eventually hits the *64µm* mark and becomes constant. At that point the bottom row width begins to increase from zero and keeps increasing with increasing N_f till it also reaches *64µm*. At that point, the bottom row is transformed into a middle row and a new bottom row starts to grow with further increase in N_f.



(a) (b)

Figure 6.5 Total width of different rows (a) total area, including area of rows and area of substrate (b)

Figure 6.5b plots the total area as well as area of individual components of the switch as a function of N_f. The width of the substrate contact equals the width of the top row. Since the width of the top row increases from a minimum starting value to the maximum (Figure 6.5a), so does the substrate contact area. The area of the top row initially increases for as long as no other row is present. Once the bottom row comes into existence, the area of the top row starts to decrease due to a decrease in the total height of the top row (refer to look-up table 6.18). The lost height is actually incorporated in the height of the bottom row. The area of the top row hits a constant value when the bottom row gets converted to a middle row. This allows no more fluctuation in the height of the top row (refer to look-up table 6.18). A similar pattern then continues between the area of the middle and bottom row.

The initial higher gradient of the area curve is absent for larger sized switches in Figure 6.2 because for those, even with the minimum number of fingers, multiple rows exist that lead to a single gradient area graph.

6.4 Summary and switch selection

The maximum area overhead associated with each type of header and footer is stated in Table 6.1. Even with the highest number of fingers, the maximum area overhead of all switches does not surpass *10%* of memory block area.

In light of the results obtained in this as well as the previous two chapters, one type of switch (in header and footer role) needs to be identified that can be incorporated in subsequent analyses. Hence a decision needs to be taken considering trade-off in the results obtained thus far.

Table 6.1 Maximum percentage area overhead of footer and header switches with respect to memory block

Configuration	**Footer**				**Header**			
Switch type	H-V_T	S-V_T	L-V_T	SRAM-PD	H-V_T	S-V_T	L-V_T	SRAM-PU
Minimum switch size	261µm	213µm	176µm	214µm	88µm	63.5µm	55µm	58µm
Maximum switch area overhead (compared to memory block)	7.06%	6.12%	5.40%	6.13%	4.48%	3.78%	3.52%	3.62%

6.4.1 Footer switch selection

The H-V_T footer has the largest size (*261µm*) with an area overhead of *7.06%*. On the other hand, the L-V_T footer has the smallest size (*176µm*) with an area overhead of *5.4%*. The pull down time of SRC node is much less than the target value of *1ns* for all footer types. Also, the read current flowing through the block-footer system is not severely limited by the type of footer incorporated. However in case of leakage current comparison, unlike other footer types the L-V_T footer is found to leak more than the block at $V_{SRC} = V_{DD} - 0.7V$ (for $T \geq 34°C$ in the MOST LEAKY corner). Since the use of L-V_T footer will guarantee DRV>*0.7V*, maximum leakage reduction will not be achieved during STANDBY. The use of L-V_T footer can therefore be ruled out. Between the S-V_T and the SRAM-PD footers, the difference in size and hence the percentage area overhead (for finger width of *215nm*) is very little. However the advantage of using the SRAM-PD footer is that it would able to track process variations in the memory block better than the S-V_T footer, particularly ΔV_T.

6.4.2 Header switch selection

The H-V_T header has the largest size (*88µm*) with an area overhead of *4.48%*. On the other hand, the L-V_T header has the smallest size (*55µm*) with an area overhead of *3.52%*. V_{SUP} in active mode is within *10mV* of V_{DD} for all types of headers. In case of leakage current comparison, all header types are found to leak less than the block at V_{SUP} = *0.7V*. Since the difference in size and percentage area overhead is very little between the S-V_T and the SRAM-PU headers, the SRAM-PU header would therefore be preferred over its S-V_T counterpart due to reasons provided in section 6.4.1.

In the following chapter, the selected header and footer will be analyzed in conjunction with a diode connected transistor (of the same type) and the memory block in order to achieve DRV during STANDBY. An alternate technique to achieve DRV is also briefly investigated, in which the switch is actively driven by an op-amp during STANDBY.

6.5 References

[1] C. H. Luf, "TSMC 45nm CMOS Logic and 40nm CMOS Logic Design Rule (CLN45LP/LPG, CLN40LP/LPG)," Document No. T-N45-CL-DR-001, Version 1.1, April 2, 2008

Chapter 7

Techniques to achieve data retention voltage during STANDBY

The SRAM-PD and SRAM-PU transistors have been identified at the end of the previous chapter as the most suitable candidates in a header and footer role. Having identified the type and size of the switch, a feedback mechanism now needs to be devised in order to achieve DRV across the memory block during STANDBY. Therefore in this chapter, we investigate the use of a simple diode-connected transistor (DCT) in conjunction with a switch and the memory block. Since the SRAM transistors have been found to be the most suitable in a switch role, they will therefore also be employed in the diode-connected configuration. The H-V_T, L-V_T and S-V_T transistors are not employed in the analyses presented here onwards. The use of a DCT for achieving DRV was first investigated in [1].

Also investigated is an alternate approach to achieve DRV, which is by using an actively clamped switch. This technique, first published in [2], has been briefly investigated for feasibility of use with the proposed block-switch architecture in the sense that accurate figures for the associated area overhead and current consumption of different modules in the published work have been taken from literature in order to determine the system-level feasibility of this scheme. A brief analysis of this scheme is presented later in this chapter.

7.1 The diode-connected transistor option

In a DCT, the gate terminal is shorted to the drain and the output *I/V* characteristics resemble that of a forward biased diode. This circuit can be used to clamp the output to a specific voltage level and therefore does not allow the output voltage to increase without limit. Figure 7.1 illustrates the concept of achieving DRV using a DCT. During ACTIVE mode, M1 is on and therefore V_{SRC} is maintained at *10mV* as explained in chapter 4. When M1 is switched off to enter STANDBY mode, V_{SRC} begins to rise as a consequence of block leakage charging the SRC node. As the gate-source capacitance of the DCT (and more generally the total SRC node capacitance) is slowly charged the DCT gradually begins to conduct as a result of increasing V_{GS} and V_{DS}. Depending on the size of the DCT, the output voltage can be clamped to a maximum of one threshold voltage.

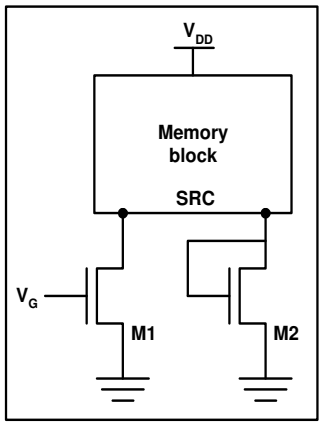


Figure 7.1 Achieving DRV using a DCT in conjunction with a footer switch and a memory block

7.2 Motivation for the DCT option

The DCT solution has a number of advantages as well a disadvantage. They are listed as under:

7.2.1 Benefits of using a DCT

i) The DCT does not consume any additional power.

ii) The technique is simple from an implementation point of view [3].

iii) Internal voltage references are not required to achieve the desired voltage at the SRC/SUP node during STANDBY.

iv) The solution does not contain multiple transistor types with different threshold voltages, which would otherwise significantly add to the fabrication cost due to inclusion of additional mask layers.

v) A reasonably low leakage current level through the system can still be maintained during STANDBY.

vi) Only a small area overhead is associated with the proposed solution.

vii) If the DCT is of the same type as the switch (which it is in this case), the DCT could be made a part of the switch itself in layout to obtain further reduction in total area overhead.

7.2.2 Disadvantage of using a DCT

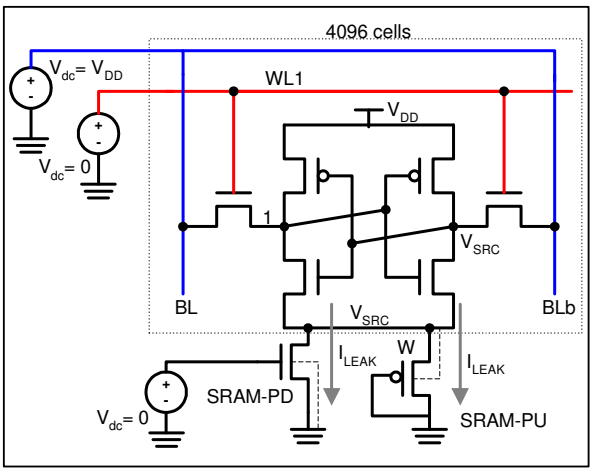
i) The DCT option does not complement the *maximum* benefits of source biasing in different process corners. In the MOST LEAKY case, threshold voltage is lowered the most and therefore high leakage in the block requires a correspondingly high V_{SRC} and lower V_{SUP}. The DCT would however also have reduced threshold voltage and therefore V_{SRC} (in case of footer) would be maintained at a value lower than $V_{DD} - 0.7V$. In the case of header, V_{SUP} would be maintained at a value higher than $0.7V$. On the other hand, in the LEAST LEAKY corner, the threshold voltage increases the most but there is not much block leakage current anyway, therefore the increased threshold voltage of the switch (leading to higher V_{SRC}, lower V_{SUP}) is not really needed. Aspects of the DCT regarding its temperature dependence and threshold voltage are in fact opposite to what is actually desired.

7.3 DCT with SRAM-PD footer switch

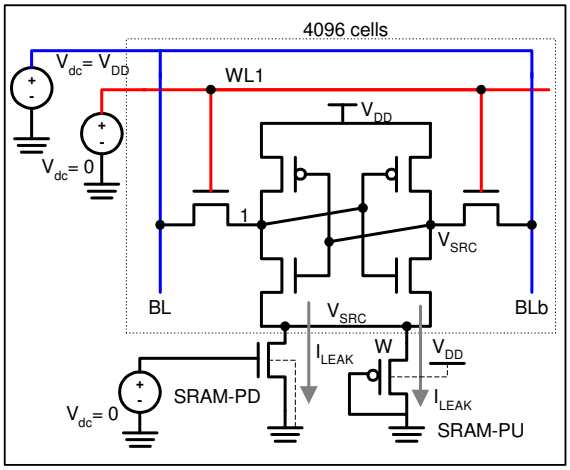
7.3.1 Simulation setup

During STANDBY, in order to retain stored data, the voltage at SRC node must not be allowed to exceed $V_{DD} - 0.7V$. In order to size the DCT, a DC simulation is set up in which the SRC node is initialized to $V_{DD} - 0.7V$, the bit lines and cell PMOS bulk are tied to V_{DD} and the cell NMOS bulk is tied to *GND*. The footer bulk is also tied to *GND*. All word lines in the block (represented by WL1 in Figure 7.2) are maintained in the inactive state. The DCT is placed between the SRC and *GND* node. The length is fixed at *55nm* and a transistor multiplication factor is used to obtain the size that would be able to maintain DRV across the memory block in all corners. The simulation is run separately for different supply voltage levels, and in each case V_{SRC} is recorded as a function of temperature in different corners. The simulation setup can also be used to get data on the total leakage current through the system.

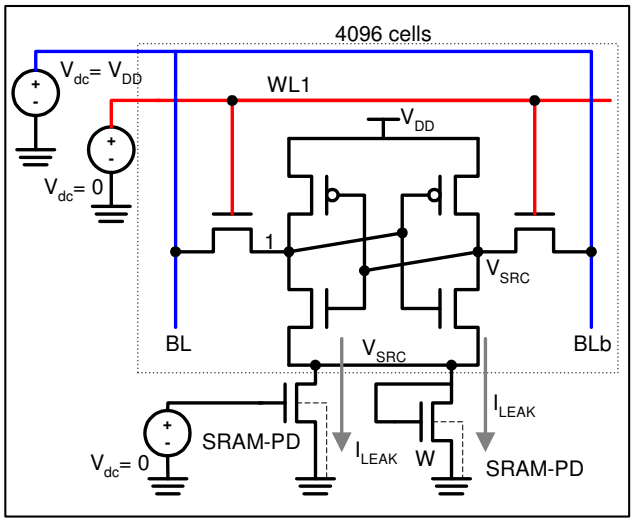
The simulation is carried out for three DCT configurations. In the first (Figure 7.2a), the SRAM-PU (PMOS) transistor is used in the DCT configuration with its bulk tied to the SRC node. The second configuration (Figure 7.2b) is similar to the first except that that PMOS bulk is tied to V_{DD} instead of the SRC node. In the third configuration (Figure 7.2c), the SRAM-PD (NMOS) transistor is used in the DCT configuration with its bulk tied to *GND*.



(a) with PMOS DCT (bulk tied to SRC node)



(b) with PMOS DCT (bulk tied to V_{DD})



(c) with NMOS DCT

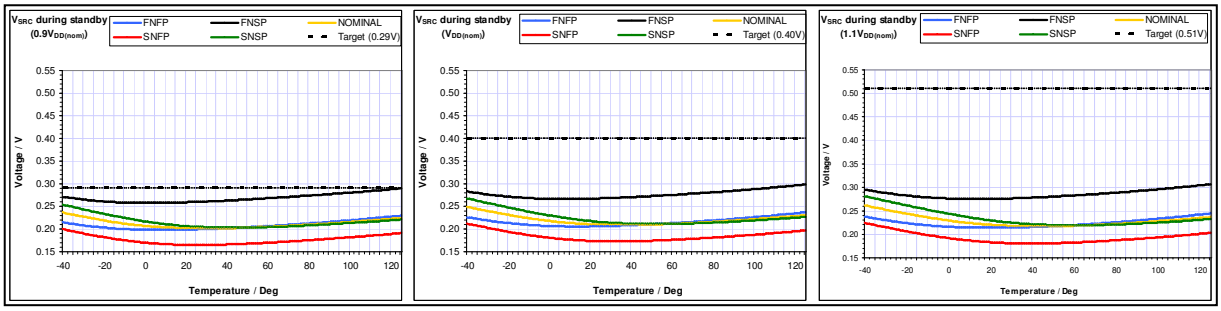
Figure 7.2 DCT configurations with a footer switch

7.3.2 Results and discussion

7.3.2.1 SRAM-PU DCT with bulk tied to SRC node

Figure 7.3 shows V_{SRC} plotted as a function of temperature during STANDBY for the configuration depicted in Figure 7.2a. The three plots show V_{SRC} variation as a function of temperature for different supply voltage levels. The minimum DCT width that guarantees DRV in all corners as well as over the full variation in supply voltage is *15.7µm*, which is achieved with the simulation conducted at V_{DD} = $0.9V_{DD(nom)}$. The FNSP process achieves the highest SRC node voltage because of the fact that the off-resistance of the block decreases the most compared to the off-resistance of the DCT. On the other hand, the SNFP process achieves the lowest SRC node voltage due to the fact that the off-resistance of

the DCT decreases the most compared to the off-resistance of the block. The SRC node voltage for other processes lies between these two extremes over the entire temperature scale.



(a) (b) (c)

Figure 7.3 V_{SRC} during STANDBY using PMOS DCT (bulk tied to SRC node) at $V_{DD} = 0.9V_{DD(nom)}$ (a) at $V_{DD} = V_{DD(nom)}$ (b) at $V_{DD} = 1.1V_{DD(nom)}$ (c)

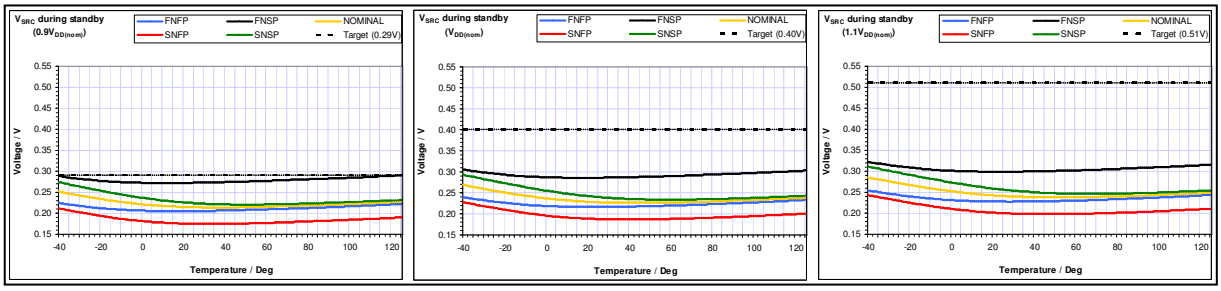
Ideally we would have liked to see V_{SRC} rise to $0.4V$ in case of $V_{DD} = V_{DD(nom)}$ and to $0.51V$ in case of $V_{DD} = 1.1V_{DD(nom)}$, but the voltage across the DCT is maintained at a much lower value in these two cases. It is important to state here that although V_{SRC} is the highest in the FNSP process, as explained above, the maximum leakage current through the entire system does not flow for this process. The maximum leakage current flows in the MOST LEAKY corner (FNFP process, $1.1V_{DD(nom)}$ and $T=125°C$) The results presented in Figure 7.3 illustrate that the DCT solution does not help achieve maximum leakage reduction when it is desired.

The two primary contributors that influence the shape of the curves in Figure 7.3 are threshold voltage and mobility, both of which are known to decrease with temperature. As previously stated, a detailed study on mobility and threshold voltage variation with temperature of all memory cell transistors, the footer and DCT was not carried out. The graphs are presented to indicate that the selected DCT width guarantees DRV across the block in all corners, over the entire operating temperature range and across full supply voltage variation.

7.3.2.2 SRAM-PU DCT with bulk tied to V_{DD} node

Figure 7.4 shows V_{SRC} plotted as a function of temperature during STANDBY for the configuration depicted in Figure 7.2b. When the bulk of the PMOS DCT is tied to V_{DD} instead of the SRC node, $|V_{SB}|$ is already greater than zero at $V_{DD} = 0.9V_{DD(nom)}$ and increases further with increase in supply voltage. This increases the threshold voltage of the DCT with increasing V_{DD}. Unlike the topology presented in Figure 7.2a, this particular circuit topology is able to achieve a higher V_{SRC} with increasing supply voltage level, as is desired. With a higher threshold voltage, the minimum width needed to achieve DRV across the block is also greater at $60.7\mu m$. The slightly higher voltage levels achieved for

different processes (compared to the results in Figure 7.3) are evidence to the discussion presented above.

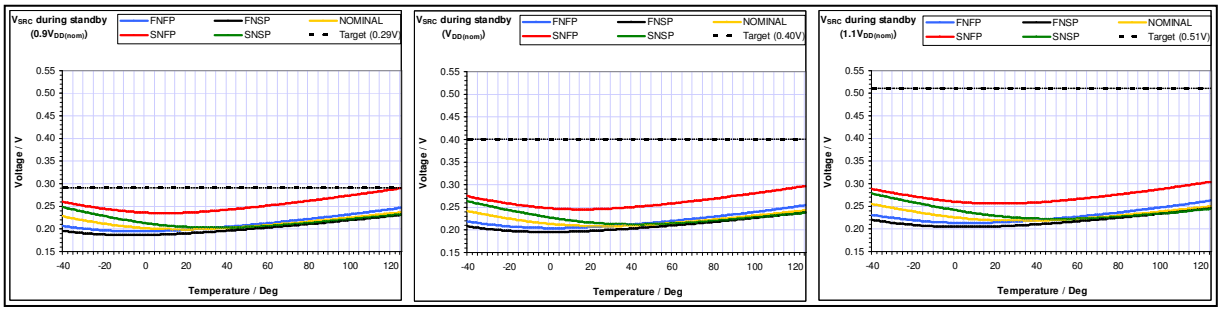


(a) (b) (c)

Figure 7.4 V_{SRC} during STANDBY using PMOS DCT (bulk tied to V_{DD}) at $V_{DD} = 0.9V_{DD(nom)}$ (a) at $V_{DD} = V_{DD(nom)}$ (b) at $V_{DD} = 1.1V_{DD(nom)}$ (c)

7.3.2.3 SRAM-PD DCT with bulk tied to GND

Figure 7.5 shows V_{SRC} plotted as a function of temperature during STANDBY for the configuration depicted in Figure 7.2c. The three plots show V_{SRC} variation as a function of temperature for different supply voltage levels. The minimum DCT width that guarantees DRV in all corners as well as over the full variation in supply voltage is *6.9µm*, which is achieved with the simulation conducted at $V_{DD} = 0.9V_{DD(nom)}$. Unlike the case of PMOS DCT, the SNFP process achieves the highest SRC node voltage for the NMOS DCT because of the fact that the off-resistance of the NMOS DCT increases the most compared to the off-resistance of the block. On the other hand, the FNSP process achieves the lowest SRC node voltage due to the fact that the off-resistance of the NMOS DCT decreases the most compared to the off resistance of the block.



(a) (b) (c)

Figure 7.5 V_{SRC} during STANDBY using NMOS DCT (bulk tied to *GND*) at $V_{DD} = 0.9V_{DD(nom)}$ (a) at $V_{DD} = V_{DD(nom)}$ (b) at $V_{DD} = 1.1V_{DD(nom)}$ (c)

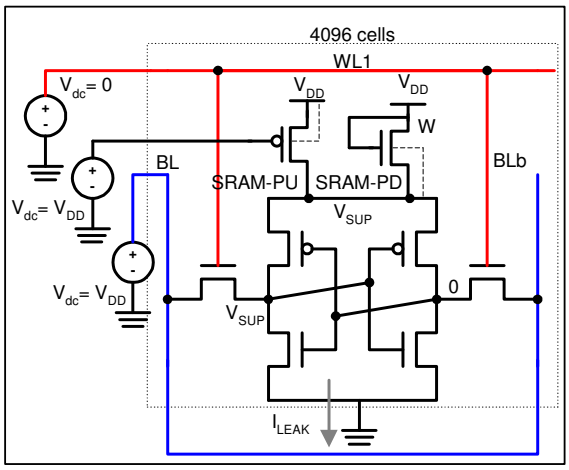
A smaller NMOS DCT suffices because of the higher carrier mobility in an N-channel device, which leads to higher conductivity of the DCT. Since source and bulk are tied to the same node (V_{SB}=*0*), the

threshold voltage does not change with variation in supply voltage, similar to the circuit topology discussed in section 7.3.2.1.

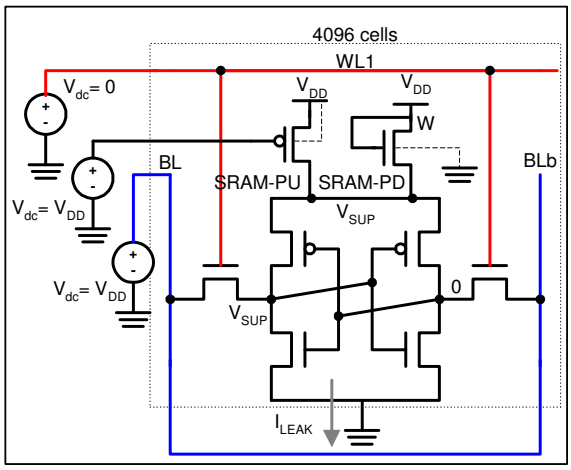
7.4 DCT with SRAM-PU header switch

7.4.1 Simulation setup

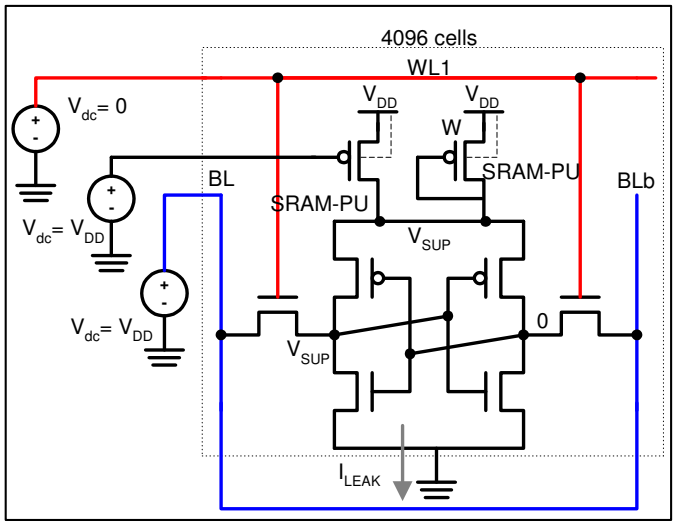
During STANDBY, in order to retain stored data, the voltage at SUP node must not be allowed to drop below *0.7V*. In order to size the DCT, a DC simulation is set up (Figure 7.6), similar to the one outlined in section 7.3.1. The DCT is placed between the V_{DD} and SUP node. The simulation is again run separately for different supply voltage levels, and in each case V_{SUP} is recorded as a function of temperature in different process corners.



(a) with NMOS DCT (bulk tied to SUP node)



(b) with NMOS DCT (bulk tied to *GND*)



(c) with PMOS DCT

Figure 7.6 DCT configurations with a header switch

The simulation is carried out for two DCT configurations. A potential option is presented in Figure 7.6a but it is not analyzed for reasons provided in section 7.4.2.1. In the configuration of Figure 7.6b, the bulk of the SRAM-PD DCT is tied to *GND*. In the configuration of Figure 7.6c, the SRAM-PU (PMOS) transistor is used in the DCT configuration with its bulk tied to V_{DD}.

7.4.2 Results and discussion

7.4.2.1 SRAM-PD DCT with bulk tied to SUP node

The topology presented in Figure 7.6a is not feasible from a fabrication point of view since the bulk of the NMOS DCT is connected to the SUP node. TSMC allows all NMOS transistor bulk connections to *GND* only. Even if this option is approved for fabrication, a separate p-well would need to be created, which would add to fabrication costs.

7.4.2.2 SRAM-PD DCT with bulk tied to GND node

Figure 7.7 shows V_{SUP} plotted as a function of temperature during STANDBY for the configuration depicted in Figure 7.6b. The three plots show V_{SUP} variation as a function of temperature for different supply voltage levels. The minimum DCT width that guarantees DRV in all process corners as well as over the full variation in supply voltage is *46.7µm*, which is achieved with the simulation conducted at $V_{DD} = 0.9V_{DD(nom)}$. When the bulk of the NMOS DCT is tied to *GND* instead of the SUP node, $|V_{SB}|$ is already greater than zero at $V_{DD} = 0.9V_{DD(nom)}$, and increases further with increase in supply voltage. This increases the threshold voltage of the DCT with increasing V_{DD}. Unlike the topology presented in Figure 7.6a, this particular circuit topology is able to achieve a lower V_{SUP} with increasing supply voltage level, as is desired. The SNFP process achieves the lowest SUP node voltage because of the fact that the off-resistance of the NMOS DCT increases the most compared to the off-resistance of the block. On the other hand, the FNSP process achieves the highest SUP node voltage due to the fact that the off-resistance of the DCT decreases the most compared to the off-resistance of the block. The SUP node voltage in other processes lies between these two extremes over the entire temperature scale.

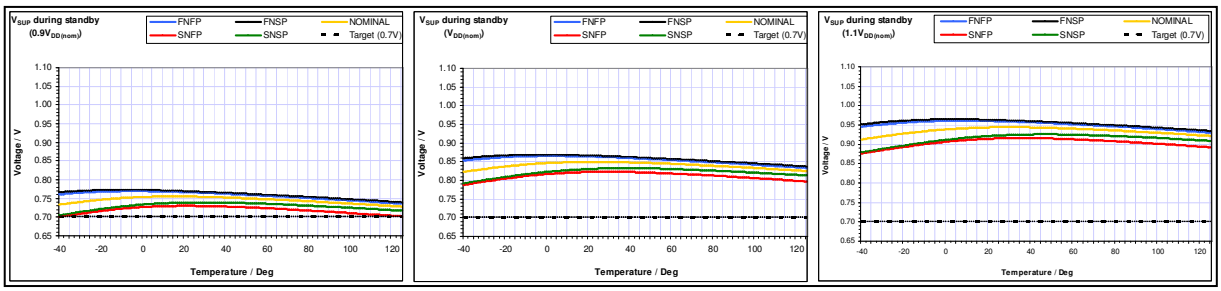


Figure 7.7 V_{SUP} during STANDBY using NMOS DCT (bulk tied to *GND*) at $V_{DD} = 0.9V_{DD(nom)}$ (a) at $V_{DD} = V_{DD(nom)}$ (b) at $V_{DD} = 1.1V_{DD(nom)}$ (c)

Variation in supply voltage scales the voltage across the block and the DCT accordingly. A higher voltage is therefore maintained across the block at $V_{DD} = 1.1V_{DD(nom)}$ when actually the lowest voltage (closest to or equal to $0.7V$) is desired in order to achieve maximum reduction in block leakage current.

7.4.2.3 SRAM-PU DCT with bulk tied to V_{DD} node

Figure 7.8 shows V_{SUP} plotted as a function of temperature during STANDBY for the configuration depicted in Figure 7.6c. The three plots show V_{SUP} variation as a function of temperature for different supply voltage levels. The minimum DCT width that guarantees DRV in all corners as well as over the full variation in supply voltage is $43.5\mu m$, which is achieved with the simulation conducted at $V_{DD} = 0.9V_{DD(nom)}$. Unlike the case of NMOS DCT, the FNSP process achieves the lowest SUP node voltage for the PMOS DCT because of the fact that the off-resistance of the DCT increases the most compared to the off-resistance of the block. On the other hand, the SNFP process achieves the highest SUP node voltage due to the fact that the off-resistance of the DCT decreases the most compared to the off-resistance of the block.

A relatively large PMOS DCT is needed because of the lower carrier mobility in a P-channel device, which leads to lower conductivity of the DCT. Since source and bulk are tied to the same node ($V_{SB}=0$), the threshold voltage does not change with variation in supply voltage.

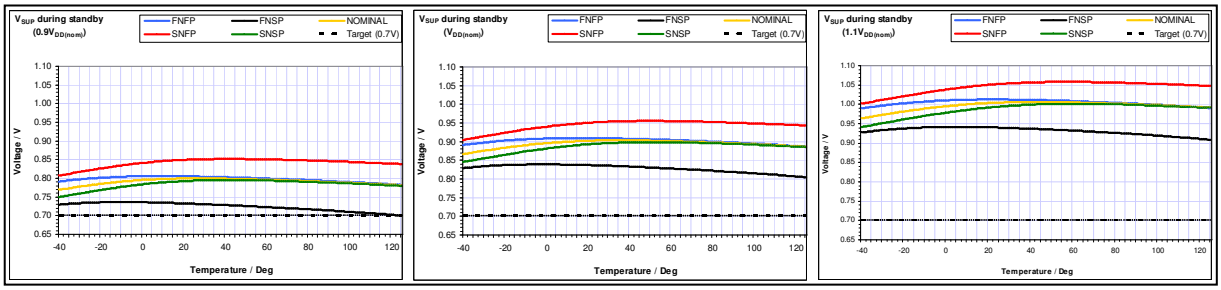


Figure 7.8 V_{SUP} during STANDBY using PMOS DCT (bulk tied to V_{DD}) at $V_{DD} = 0.9V_{DD(nom)}$ (a) at $V_{DD} = V_{DD(nom)}$ (b) at $V_{DD} = 1.1V_{DD(nom)}$ (c)

7.5 Selection of DCT

The use of the DCT helps achieve DRV across the block during STANDBY, however it does not allow maximum possible reduction in leakage currents to be achieved. Table 7.1 and Table 7.2 present the results of the analysis for the block-footer-DCT and block-header-DCT systems respectively. In addition, the area estimation model developed in chapter 6 has been employed to determine the area overhead associated with each type of DCT.

For the block-footer-DCT system (Table 7.1), the configuration shown in Figure 7.2c (NMOS DCT with bulk tied to GND) is able to achieve the highest SRC node voltage (*0.263V*) in the MOST LEAKY corner, and hence maximum leakage reduction (*64.5%*). The leakage reduction achieved with this configuration in the NOMINAL and LEAST LEAKY corners also closely matches the figures of the other two options. It is also the most efficient option from an area overhead perspective, requiring only *0.5%* area of the memory block.

Table 7.1 Comparison of results for different DCT options with a footer switch

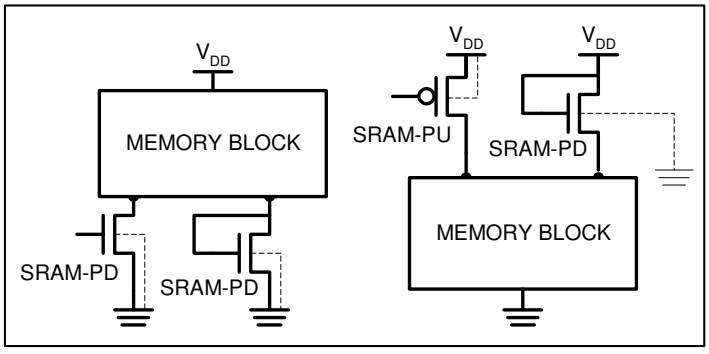
	Percentage area overhead	MOST LEAKY		NOMINAL		LEAST LEAKY	
		V_{SRC}	Max. leakage reduction	V_{SRC}	Max. leakage reduction	V_{SRC}	Max. leakage reduction
Footer + PMOS DCT with bulk tied to SRC	7.63%	0.244V	62.5%	0.210V	57.4%	0.253V	43.0%
Footer + PMOS DCT with bulk tied to V_{DD}	9.83%	0.244V	62.5%	0.228V	59.6%	0.274V	44.7%
Footer + NMOS DCT with bulk tied to GND	6.63%	0.263V	64.5%	0.208V	57.2%	0.248V	42.5%

For the block-header-DCT system (Table 7.2), the configuration shown in Figure 7.6b (NMOS DCT with bulk tied to GND) is able to achieve lower SUP node voltage in the MOST LEAKY and NOMINAL corners (*0.930V* and *0.849V*), and hence higher leakage reduction of *25.5%* and *36.5%* respectively in the two corners. It is also more efficient from an area overhead perspective requiring only *2.9%* area of the memory block.

Table 7.2 Comparison of results for different DCT options with a header switch

	Percentage area overhead	MOST LEAKY		NOMINAL		LEAST LEAKY	
		V_{SUP}	Max. leakage reduction	V_{SUP}	Max. leakage reduction	V_{SUP}	Max. leakage reduction
Header + NMOS DCT with bulk tied to SUP	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Header + NMOS DCT with bulk tied to GND	6.52%	0.930V	25.5%	0.849V	36.5%	0.704V	57.5%
Header + PMOS DCT with bulk tied to V_{DD}	6.82%	0.990V	21.0%	0.902V	31.1%	0.750V	52.9%

Figure 7.9 shows the selected block-switch-DCT configurations.



(a) (b)

Figure 7.9 Selected block-footer-DCT configuration (a) and block-header-DCT configuration (b)

Figure 7.10 shows *32kb* instance-level leakage current in STANDBY mode in different corners for the configurations shown in Figure 7.9. Source biasing the SRC node (compared to source biasing the SUP node) achieves greater reduction in leakage current in the MOST LEAKY and NOMINAL

corners but not in the LEAST LEAKY corner. That is because the block-header-DCT configuration is able to maintain V_{SUP} closer to DRV in the LEAST LEAKY corner, relative to the block-footer-DCT configuration maintaining V_{SRC} closer to $0.9V_{DD(nom)} - 0.7V$.

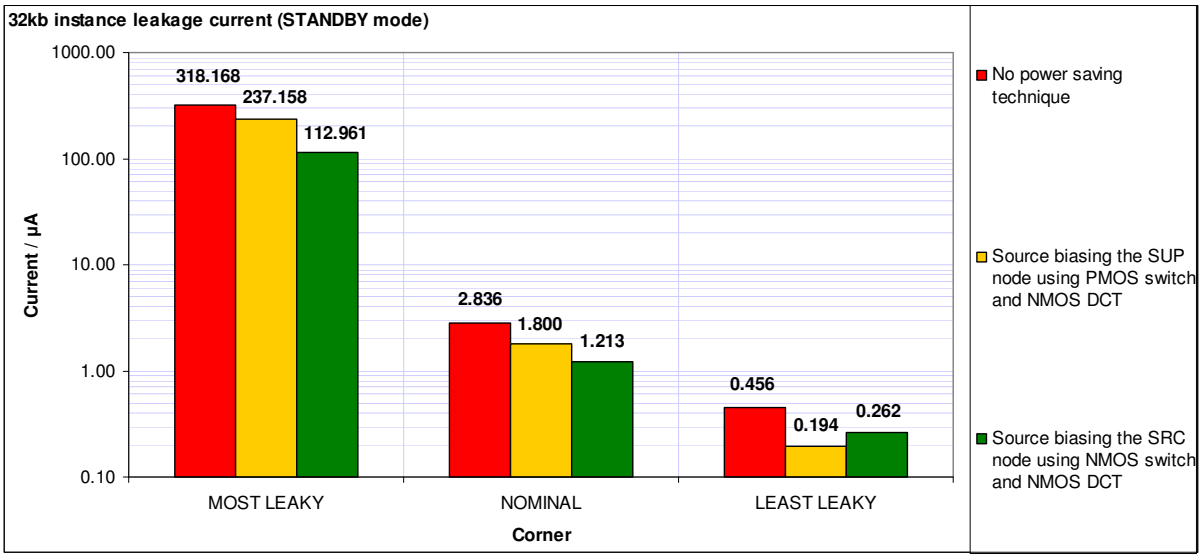


Figure 7.10 STANDBY mode block leakage current comparison for the investigated architectures

Figure 7.11 shows the percentage reduction in block/instance leakage in comparison to using no power saving technique. Maximum leakage reduction of *64.5%* and *25.5%* can be achieved in the MOST LEAKY corner for the block-footer-DCT and block-header-DCT configurations respectively.

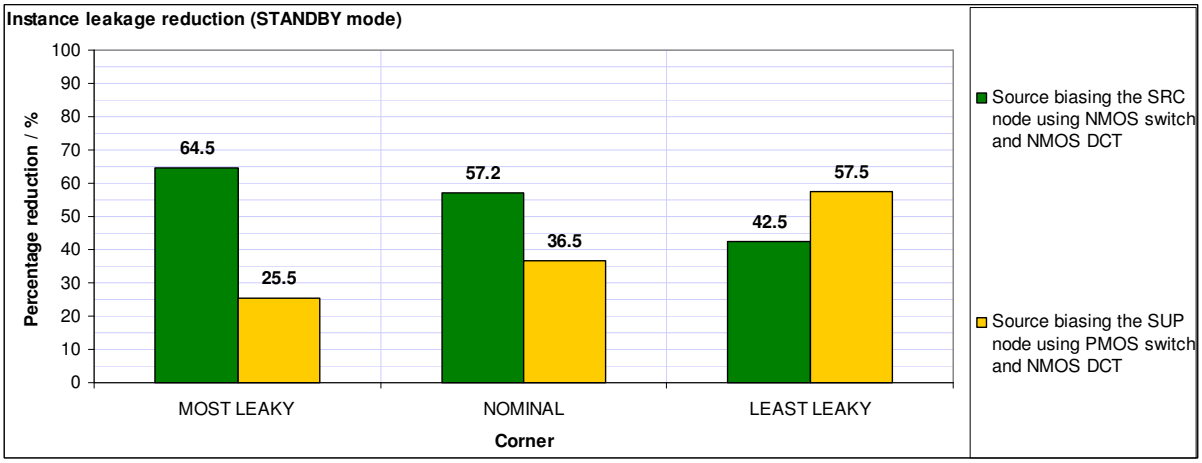


Figure 7.11 Percentage block leakage reduction (compared to using no power saving technique)

7.6 The actively clamped switch

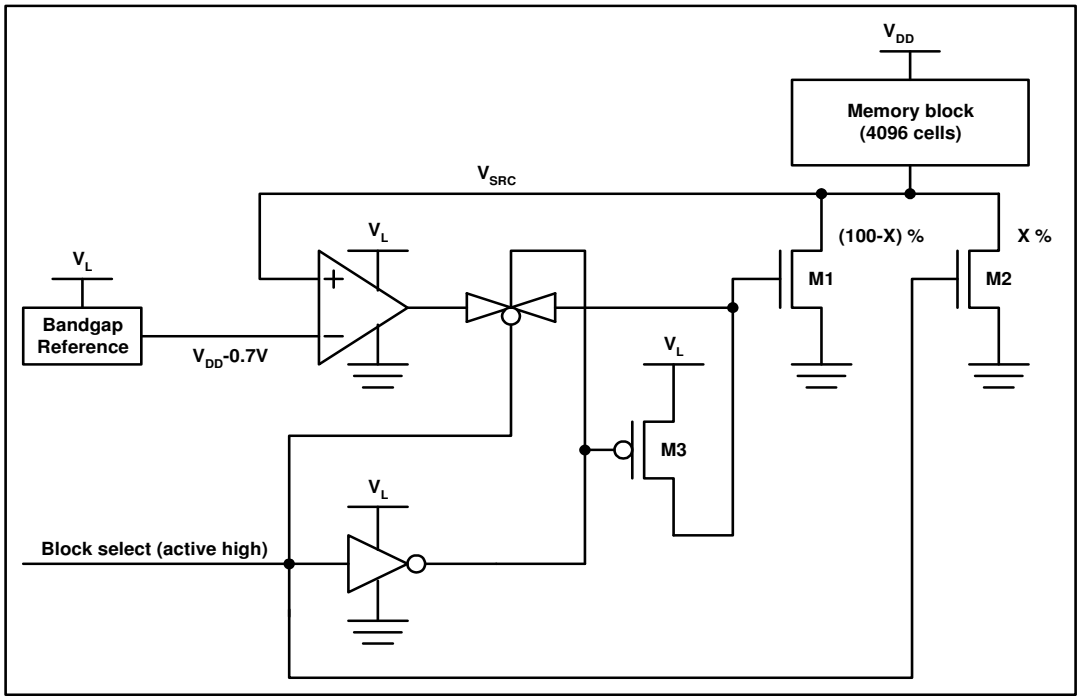


Figure 7.12 Actively clamped switch scheme (for source biasing the SRC node)

A modified version of the actively clamped sleep transistor scheme presented in [2] is employed to estimate leakage reduction achieved during STANDBY. A schematic of this approach is presented in Figure 7.12. A proposed version of this scheme to suit the header switch is presented in Figure 7.13. The area overhead, current consumption and output voltage variation figures for the bandgap reference and the op-amp are taken from NXP's core IP library. The current consumption of an existing embedded bandgap is *20µA*, *10µA* and *5µA* and that of the op-amp is *20µA*, *15µA* and *10µA* in the MOST LEAKY, NOMINAL and LEAST LEAKY corners respectively [4]. The total area overhead associated with the bandgap and the op-amp is *15210µm*^{2} and *10µm*^{2} respectively [4], which is *993.4%* and *0.7%* of the memory block area. The variation on the bandgap output voltage is *±5%* of the NOMINAL value [4]. Keeping the expected variation of the bandgap output in mind, a corresponding *5%* variation on the SRC and SUP nodes is taken into account as a result.

The circuit itself was not simulated. Note that a single bandgap is required per memory instance, but each memory block needs to have an independent op-amp.

During ACTIVE mode (block select signal is active), the output of the op-amp is cut-off and both M1 and M2 are on, thus maintaining V_{DD} across the memory block. When the block select signal is driven

low, M2 is switched off and the gate voltage of M1 is controlled by the output of the op-amp. The op-amp sets its output voltage such that the difference at its inputs is minimized. V_{SUP}/V_{SRC} is therefore maintained at the reference voltage level provided by the bandgap during STANDBY. The op-amp needs to fulfill certain criteria in order to be used in this role. The op-amp should not only have a low settling time but should also be able to guarantee minimum overshoot on the SRC node [2].

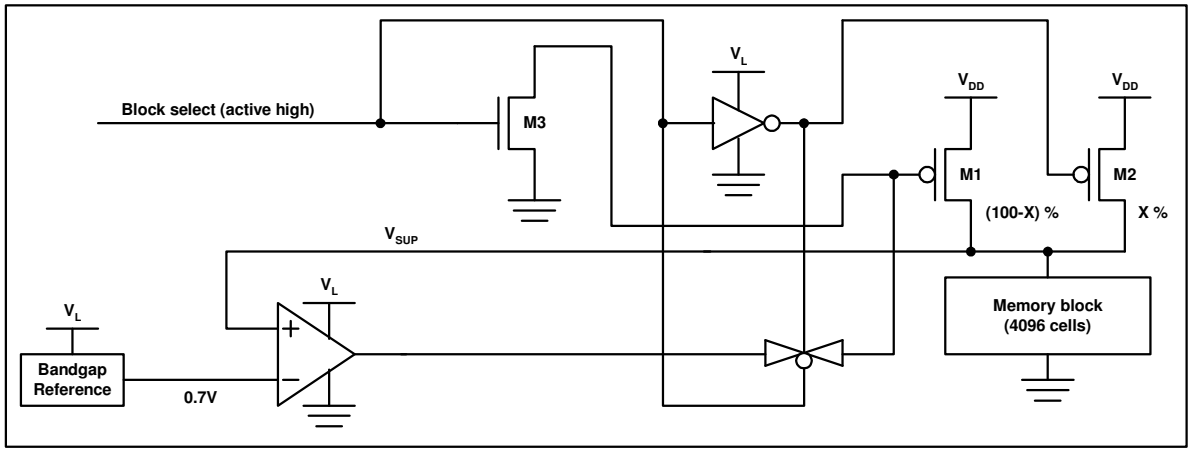


Figure 7.13 Proposed actively clamped switch scheme (for source biasing the SUP node)

7.6.1 Simulation setup

A simulation setup, similar to the one outlined in sections 7.3.1 and 7.4.1, can be used to determine the total leakage current through the block-switch system. The simulation setup shown in Figure 7.14 mimics the behavior of an op-amp driving the gate of the switch during STANDBY.

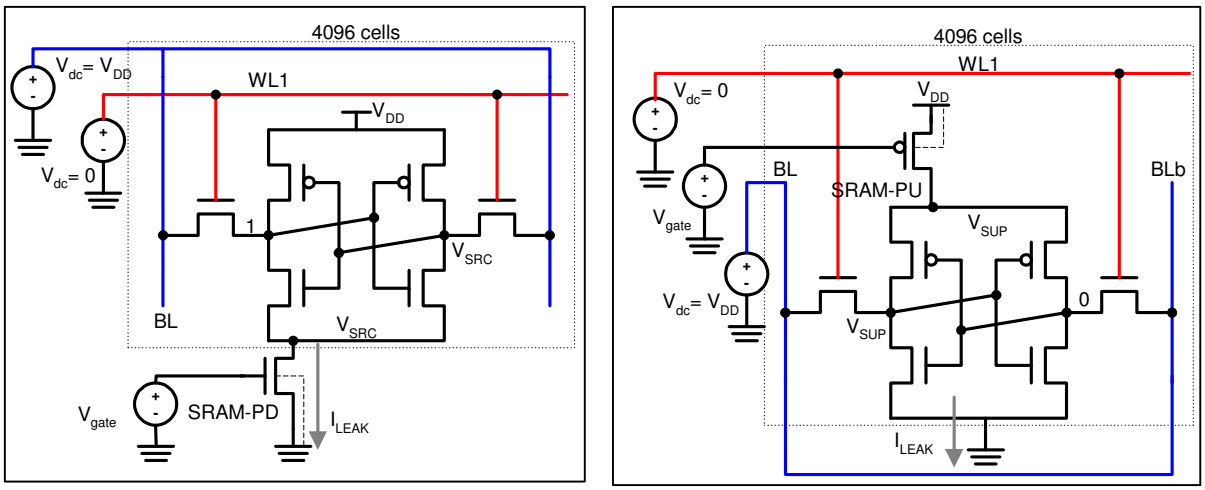


Figure 7.14 Simulation setup used to determine the magnitude of leakage current

The only difference is that the DCT is removed and a gate voltage is applied to the footer (header) to achieve the desired voltage at the SRC (SUP) node in the different corners. The voltage at the stated

nodes would be slightly different in different corners due to the *5%* expected variation at the bandgap output.

7.6.2 Results and discussion

Table 7.3 and Table 7.4 provide system-level results for the actively clamped footer and header systems respectively.

Table 7.3 System-level results for actively clamped footer switch

	Percentage area overhead	MOST LEAKY		NOMINAL		LEAST LEAKY	
		V_{SRC}	Max. leakage reduction	V_{SRC}	Max. leakage reduction	V_{SRC}	Max. leakage reduction
32kb instance + 8 actively clamped footers + 8 op-amps + bandgap	130.9%	0.32V	13.3%	0.31V	-4868.9%	0.29V	-18592.1%

Table 7.4 System level results for actively clamped header switch

	Area/ Percentage area overhead	MOST LEAKY		NOMINAL		LEAST LEAKY	
		V_{SRC}	Max. leakage reduction	V_{SRC}	Max. leakage reduction	V_{SRC}	Max. leakage reduction
32kb instance + 8 actively clamped headers + 8 op-amps + bandgap	128.5%	0.77V	-20.9%	0.74V	-4889.5%	0.7V	-18578.6%

It is clear that for a *32kb* instance, little leakage reduction is possible with the actively clamped footer configuration only, in the MOST LEAKY corner. The current consumption of the op-amps and the bandgap is far greater than the reduction in leakage achieved in the memory blocks in the NOMINAL

and LEAST LEAKY corners. For the actively clamped header scheme, more power is actually consumed in all corners. One method to achieve more benefit out of this scheme is to use larger-sized blocks.

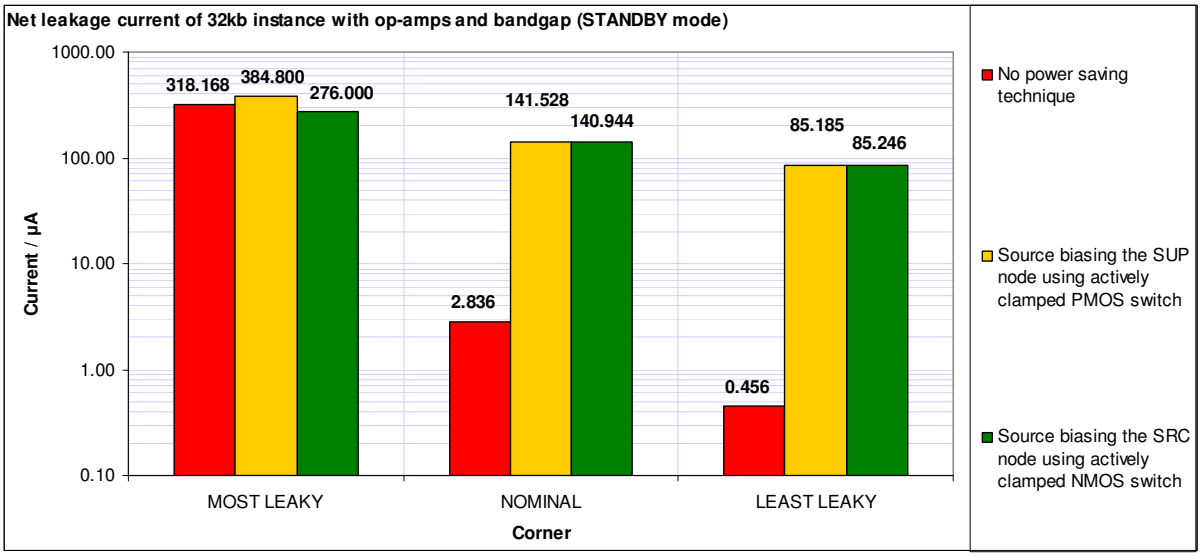


Figure 7.15 STANDBY mode system leakage current comparison for architectures using actively clamped switch

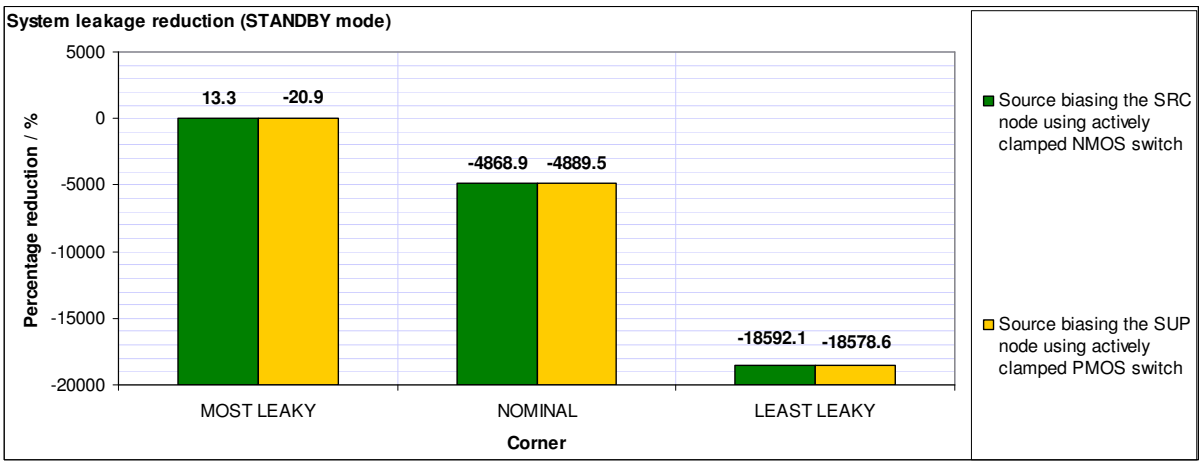


Figure 7.16 Percentage system leakage reduction (compared to using no power saving technique)

Figure 7.15 shows a graphical representation of system STANDBY leakage levels with the use of an actively clamped switch scheme. Figure 7.16 shows the percentage reduction in leakage compared to using no power saving technique. The figures tie in with the data presented in Table 7.3 and Table 7.4.

7.7 Comparison of the two schemes

Although the actively clamped switch architecture achieves greater block leakage reduction, the presence of an op-amp with each memory block significantly increases the overall system STANDBY current. For a *32kb* instance, the area overhead for the actively clamped switch scheme is also significantly higher, which is primarily due to the large area overhead of the bandgap. For larger instances, the actively clamped scheme may be more area efficient, but it can only achieve leakage reduction in the MOST LEAKY corner for the actively clamped footer configuration.

The most promising option appears to be the block-footer-DCT architecture from the results seen in Table 7.1.

7.8 Conclusion

Optimum candidates in a DCT role both for the block-footer and block-header architectures have been identified. An area/leakage current trade-off has also been presented for an alternate scheme that can be used to achieve DRV during STANDBY. The highlighted configurations in Table 7.1, Table 7.2 and Table 7.3 will be analyzed for potential power savings at the desired frequency of *500MHz* in the following chapter. A power estimation analysis is necessary so that a complete design trade-off picture in terms of area, speed and power can be sketched. In the following chapter, a model is developed that estimates power savings for the investigated architectures for two different memory addressing modes.

7.9 References

[1] M. Yamaoka, Y. Shinozaki, N. Maeda, Y. Shimazaki, K. Kato, S. Shimada, K. Yanagisawa and K. Osada, "*A 300-MHz 25-µA/Mb-Leakage On-Chip SRAM module featuring process-variation immunity and low-leakage-active mode for mobile-phone application processor*," IEEE Journal of Solid State Circuits, Vol. 40, Issue 1, January 2005

[2] M. Khellah, D. Somasekhar, Y. Ye, N. S. Kim, J. Howard, G. Ruhl, M. Sunna, J. Tschanz and V. De, "*A 256-Kb Dual-V_{cc} SRAM Building Block in 65-nm CMOS Process with Actively Clamped Sleep Transistor*," IEEE Journal of Solid State Circuits, Vol. 42, Issue 1, January 2007

[3] T. S. Doorn, "*Leakage reduction in SRAM cells*," Technical Note NXP-R-TN-2008/00084, Issued 4/2008

[4] NXP Intranet IP portfolio, "*45nm Analog/Mixed Signal Library*" 2009

Chapter 8

Power estimation

In the work presented thus far, optimum header and footer switches as well as optimum DCTs have been identified keeping in view the maximum achievable leakage reduction and associated area penalty. An alternate technique to achieve DRV during STANDBY has also been briefly investigated, and area overhead/leakage reduction figures have been calculated. In this chapter, a power estimation model is developed that estimates power savings achieved from the selected architectures as a function of memory operating frequency.

The power estimation analysis is presented for the MOST LEAKY, NOMINAL and LEAST LEAKY corners using two types of memory access schemes: cyclic single block access and sequential block access.

8.1 Power consumed by switch

The power consumed by the driving of the switch is directly proportional to the frequency of operation, f_{access}, of the memory.

Figure 8.1 shows the voltage and current profiles for the NMOS switch. The area under a power-time graph represents the energy consumed per transition, $E_{consume}$. The power consumed by the switch is therefore given by Equation (8.1).

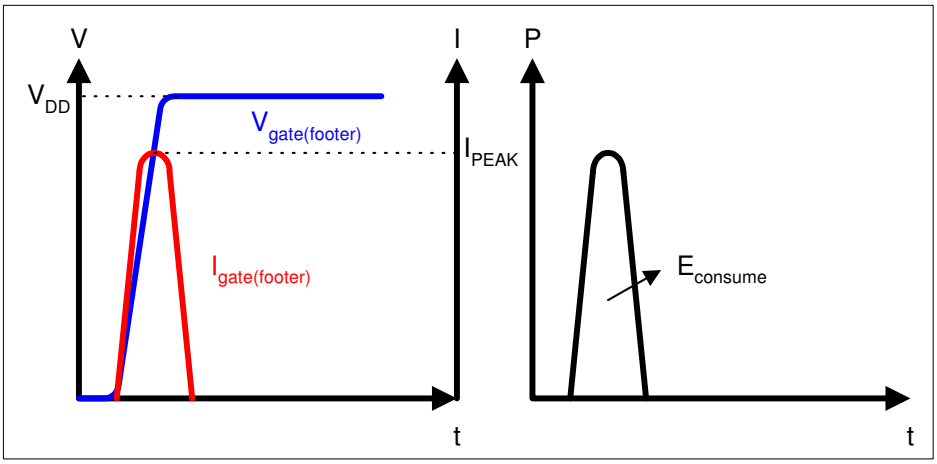


Figure 8.1 Energy consumption per transition for footer switch

$$P_{SWITCH} = E_{consume} \cdot f_{access} \tag{8.1}$$

8.2 Power savings in memory block

As soon as the switch is de-activated, the leakage of the block begins to charge the SRC node capacitance (in case of a footer switch) and the voltage on the node begins to rise. As a consequence of that happening, the leakage current through the memory block begins to decrease. The rising SRC node voltage is finally clamped to a fixed potential by the DCT. Consequently, the decreasing leakage level through the memory block also hits a constant level at time t_0. This principle is illustrated in Figure 8.2.

The leakage current savings achieved over time are expressed by Equation (8.2), where $I_{LEAK\ (MAX)}$ is the leakage of the memory block when a potential difference of V_{DD} exists across its terminals.

$$I_{SAVE} = I_{LEAK\,(MAX)} - I_{LEAK} \tag{8.2}$$

The corresponding instantaneous power savings are expressed by Equation (8.3).

$$P_{SAVE} = V_{DD} \cdot I_{SAVE} \tag{8.3}$$

The energy savings achieved at a particular instant in time can be obtained by integrating the instantaneous power savings over the STANDBY time period, as shown by Equation (8.4).

$$E_{SAVE} = \int_0^t P_{SAVE} \cdot dt \tag{8.4}$$

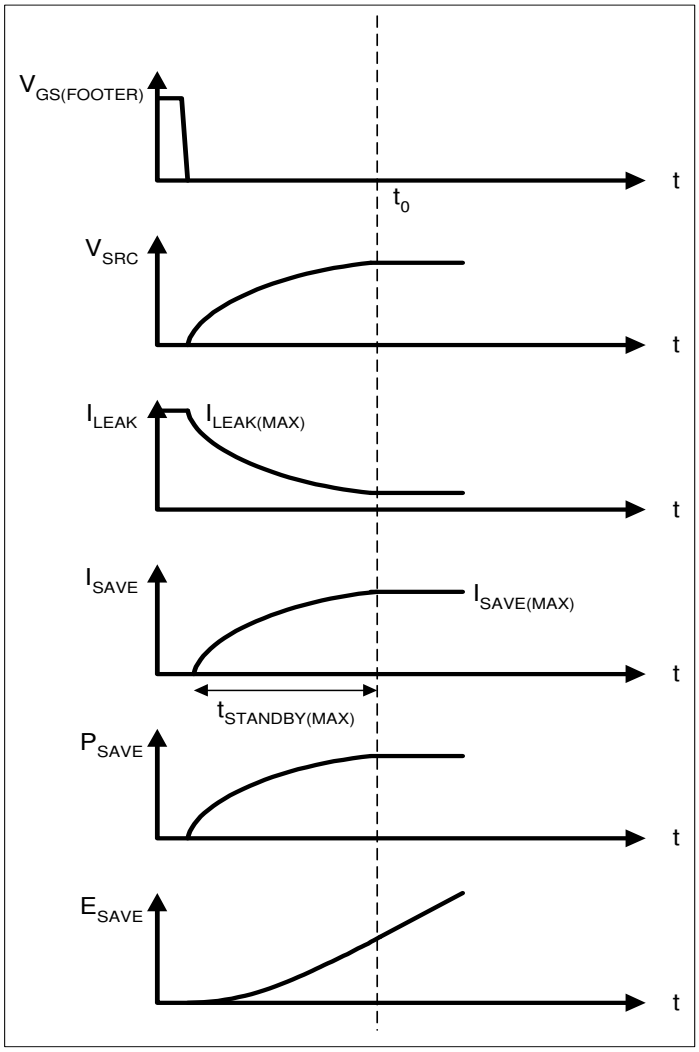


Figure 8.2 Deactivating the footer switch

8.3 Memory access schemes

8.3.1 Cyclic single block access (CSBA)

In this particular access scheme, a single memory block is accessed cyclically while all other blocks are in STANDBY, i.e. they are never accessed. The power saved by the accessed block is inversely proportional to the frequency of operation, while all non-accessed blocks save maximum power. Figure 8.3 illustrates this principle. This access scheme mimics designs using high locality of reference.

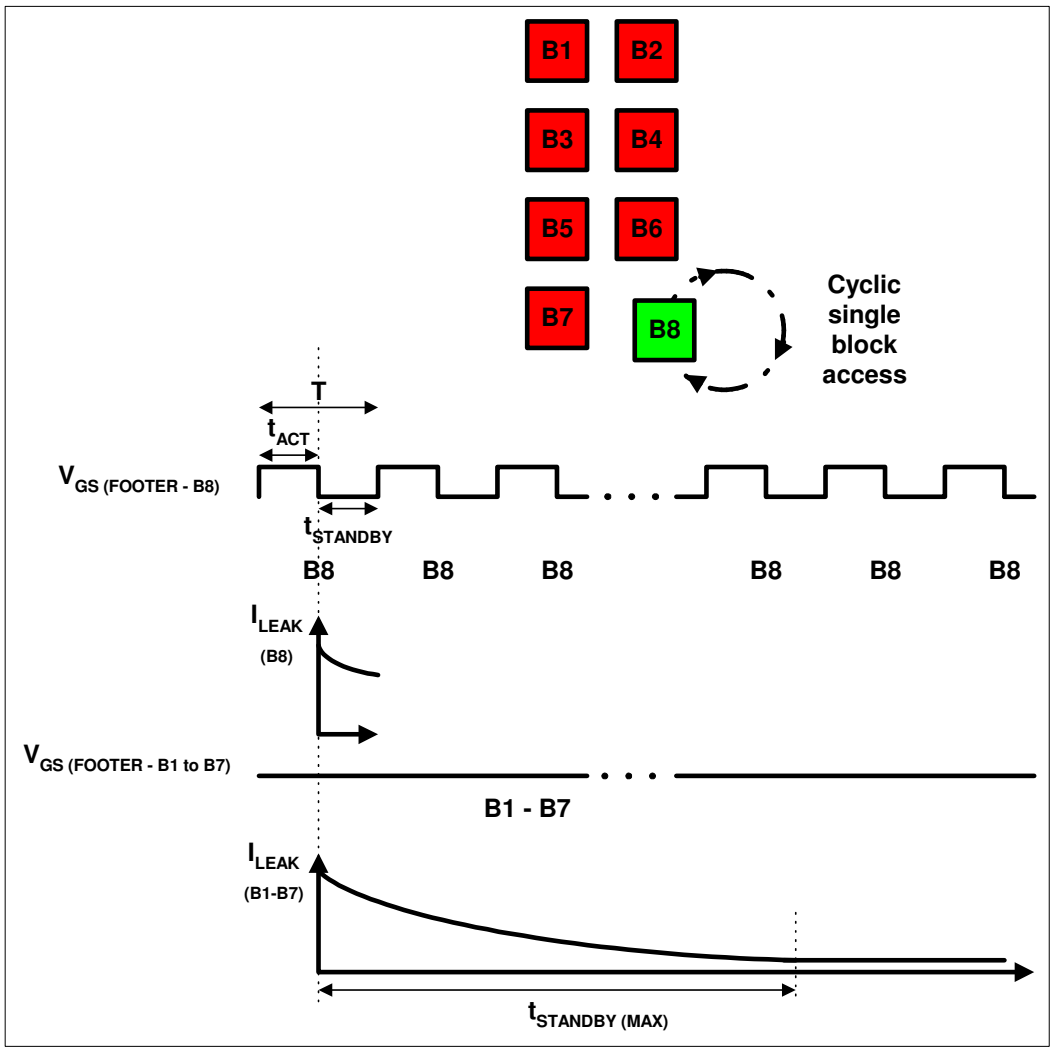


Figure 8.3 Principle of cyclic single block access scheme

The pull down time (t_{PD}) of the SRC node in the MOST LEAKY, NOMINAL and LEAST LEAKY corners is *480s*, *410ps* and *450ps* respectively for the SRAM-PD footer. The ACTIVE time (minimum cycle time) of the proposed memory architecture is therefore given by Equation (8.5).

$$t_{ACT} = t_{CYC(\min)} = t_{PD} + 1ns \tag{8.5}$$

where *1ns* is the cycle time of the existing memories and t_{PD} is the access time penalty associated with the use of a footer switch. The penalty (t_{PU}) in case of the SRAM-PU header switch is *800ps*, *670ps* and *860ps* respectively in the MOST LEAKY, NOMINAL and LEAST LEAKY corners. Therefore the ACTIVE time of the proposed memory architecture with a header switch is given by Equation (8.6).

$$t_{ACT} = t_{CYC(\min)} = t_{PU} + 1ns \tag{8.6}$$

The period T of the switch gate signal can be expressed by Equation (8.7) and the frequency of memory access by Equation (8.8).

$$T = t_{ACT} + t_{STANDBY} \tag{8.7}$$

$$f_{access} = \frac{1}{T} \tag{8.8}$$

The power savings of the accessed block is expressed by Equation (8.9).

$$P_{AB} = f_{access} \cdot E_{SAVE} \tag{8.9}$$

The frequency for all non-accessed blocks is zero, hence the power savings of each non-accessed block is given by Equation (8.10), where $I_{SAVE(MAX)}$ represents the maximum savings in leakage current (refer to Figure 8.2).

$$P_{NAB} = I_{SAVE(MAX)} \cdot V_{DD} \tag{8.10}$$

The total power savings of the memory instance is therefore given by Equation (8.11), where N_B represents the total number of blocks the instance is divided into.

$$P_I = (N_B - 1) \cdot P_{NAB} + P_{AB} \tag{8.11}$$

8.3.2 Sequential block access (SQBA)

In the SQBA scheme, all memory blocks are accessed sequentially in a cyclical pattern. In other words, all blocks save equal power for a given f_{access} value. Figure 8.4 illustrates this principle. This particular access scheme mimics situations where random block wise hopping is observed.

The common time for which each block remains in STANDBY is given by Equation (8.12).

$$t_{STANDBY(COMMON)} = (N_B - 1) \cdot T + t_{STANDBY} \tag{8.12}$$

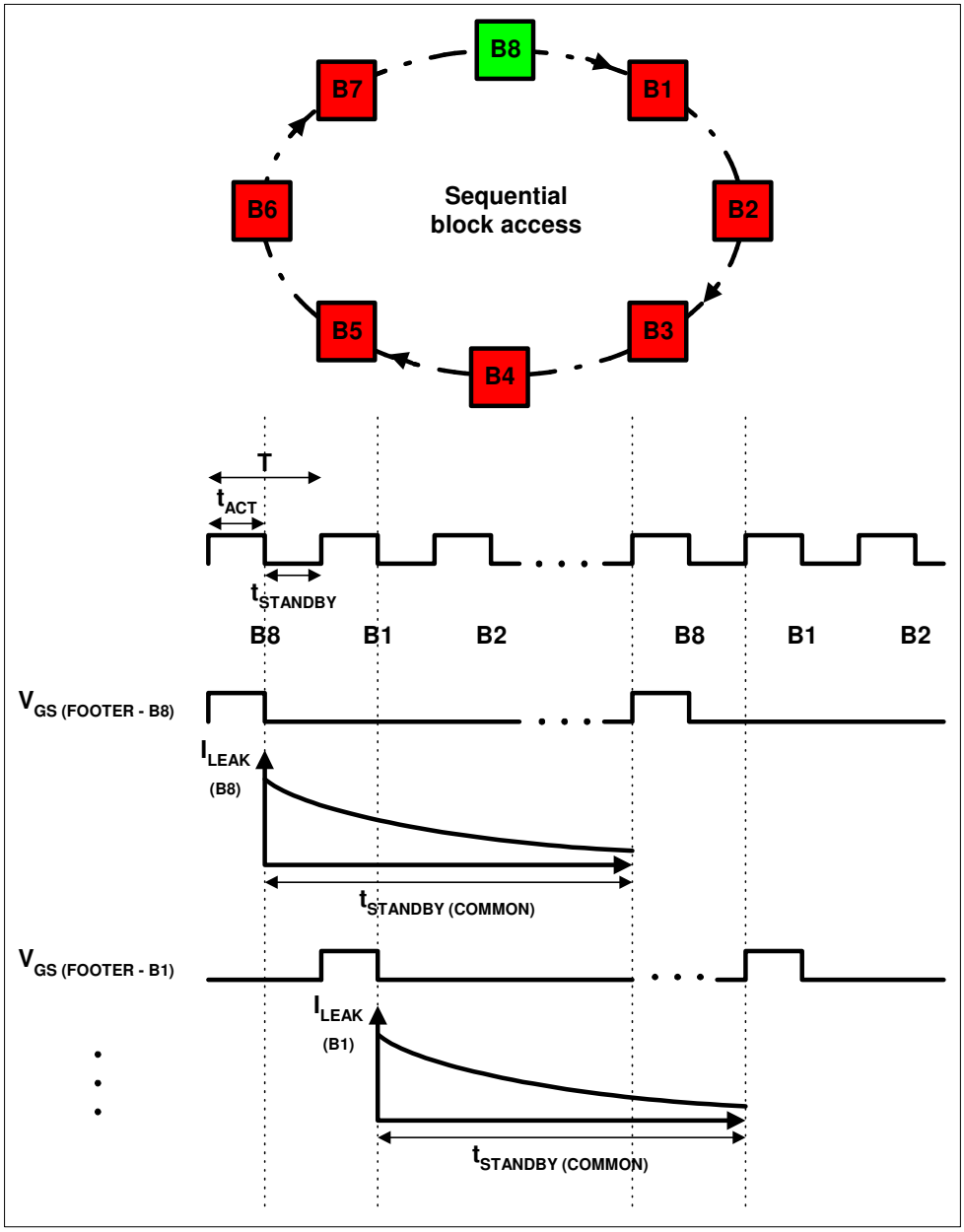


Figure 8.4 Principle of sequential block access scheme

The common internal frequency of accessed blocks is expressed by Equation (8.13).

$$f_{COMMON} = \frac{1}{t_{STANDBY(COMMON)} + t_{ACT}} \tag{8.13}$$

The power savings of each block is therefore given by Equation (8.14) and the total power savings of the instance are given by Equation (8.15).

$$P_{AB} = f_{COMMON} \cdot E_{SAVE(t_{STANDBY(COMMON)})} \tag{8.14}$$

$$P_I = N_B \cdot P_{AB} \tag{8.15}$$

8.4 Power savings with the block-switch-DCT architecture

Figure 8.5 shows power savings of a *32kb* memory instance together with the power consumption of the NMOS switch, plotted as a function of f_{access} for the two memory access schemes, using the block-footer-DCT architecture. The net power savings at a specific frequency are the power savings of the instance minus the power consumption of the switch. Figure 8.6 shows the same for the block-header-DCT architecture.

It is evident that with a *32kb* instance, the only power savings possible at the desired frequency of *500MHz* are in the MOST LEAKY corner. What is also evident from the results is that the investigated memory architecture is mostly suited to high-temperature applications. For a *32kb* instance, more power is actually consumed in the NOMINAL and LEAST LEAKY corners for frequencies above *10MHz* for the block-footer-DCT and above *40MHz* for block-header-DCT configurations respectively.

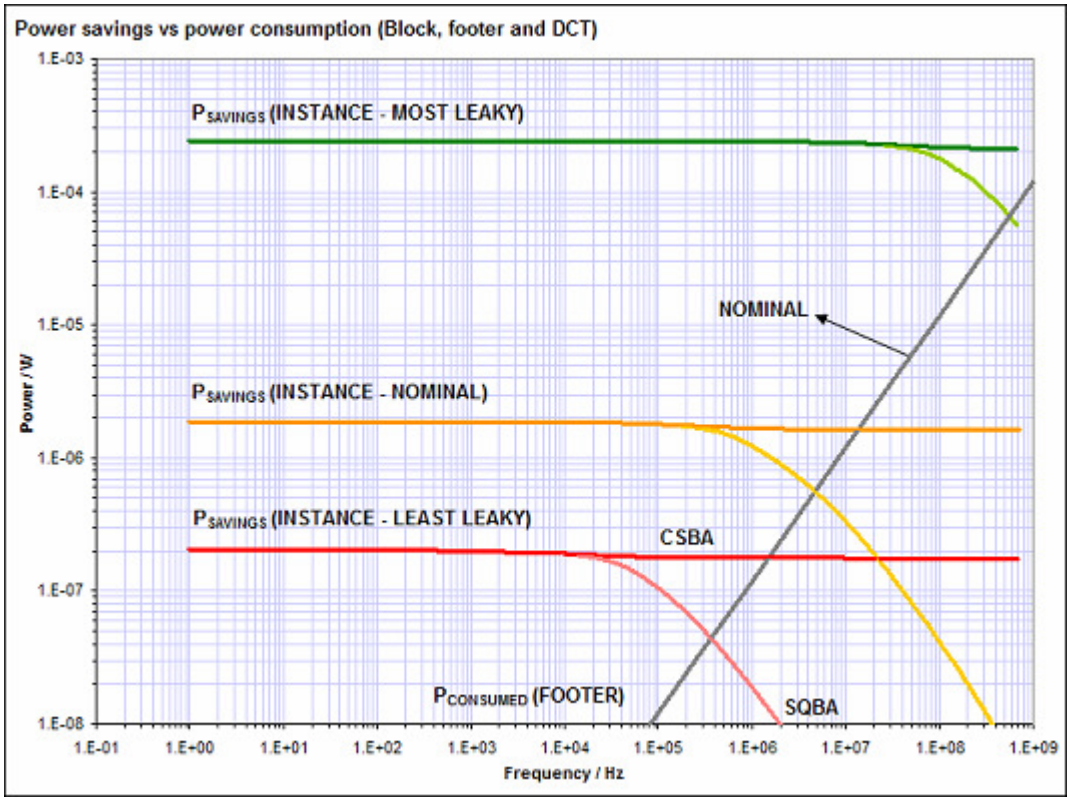


Figure 8.5 *32kb* instance power savings against power consumed by footer switch (block-footer-DCT)

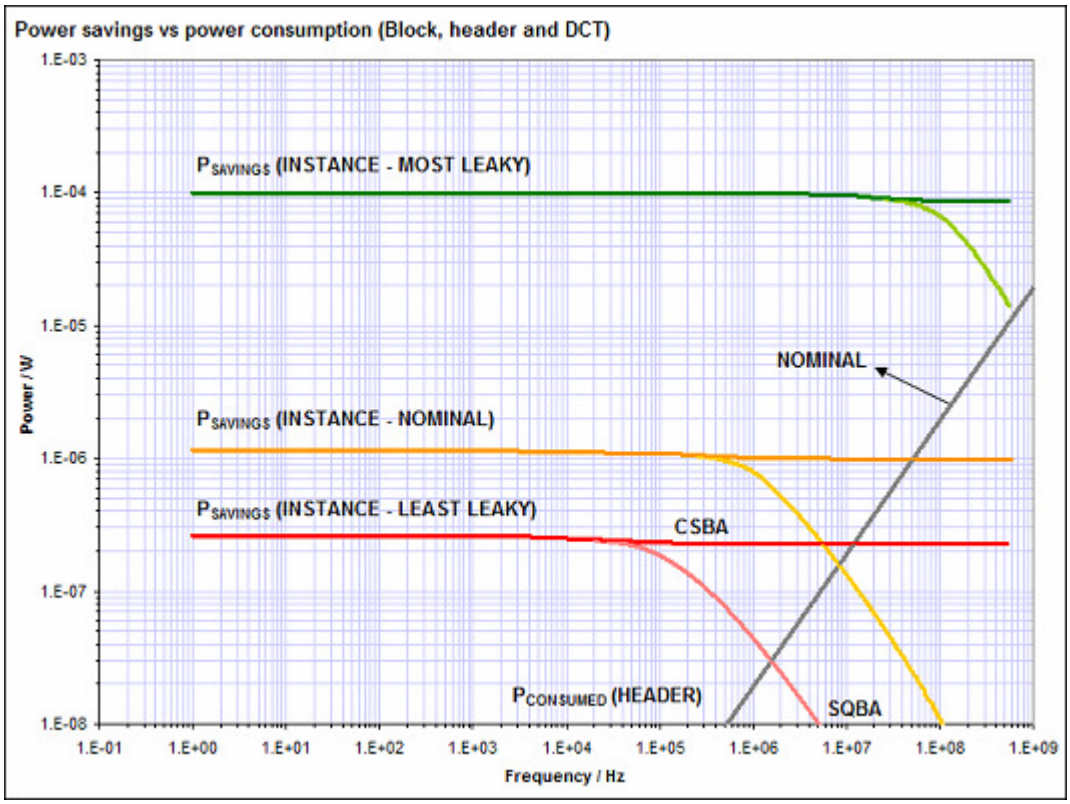


Figure 8.6 *32kb* instance power savings against power consumed by header switch (block-header-DCT)

Plots that curve downward represent the power savings of the instance using SQBA scheme and the plots that are relatively horizontal represent the power savings of the instance using the CSBA scheme. The achieved power savings are lowest at very high frequencies since sufficient time is not available to the SRC (SUP) node to float up (down) to the *designed-for* potential during STANDBY. Block leakage currents therefore do not reduce substantially after the switch has been turned off. As f_{access} decreases, more STANDBY time is available to the memory blocks hence leakage currents reduce by a greater amount leading to higher power savings. Below a certain frequency, the plot for instance power savings hits a constant value because maximum reduction in leakage currents is achieved.

In case of the CSBA scheme, the power savings of only the accessed block depend on f_{access}. Since the remaining blocks are always in STANDBY, they save maximum power. This is the reason why the plots for the CSBA scheme are relatively horizontal in Figure 8.5 and Figure 8.6.

Note that if only one *4096*-cell memory block was used in the analysis, the power curves for both memory access schemes would be identical, horizontal till a certain frequency value and curving downwards with further increase in frequency.

As the number of blocks is increased, the separation between the CSBA and SQBA curves decreases at high frequencies because a larger number of blocks increases the value of $t_{STANDBY\ (COMMON)}$ in Equation (8.12), bringing its value closer to when maximum leakage reduction can be achieved.

In Figure 8.5 and Figure 8.6, the maximum operating frequency of the instance in the three corners is different due to the difference in pull-down (pull-up) times of the SRC (SUP) nodes in the three corners, which contribute differently to the access time penalty thus giving different values for minimum cycle time.

Since it is clear that significant power savings cannot be achieved with an instance of *32kb*, the memory instance size is now increased so that a near break-even between power saved and power consumed at *500MHz* is achieved in the NOMINAL corner for the SQBA scheme. This would allow power savings to be achieved at room temperature and above for the desired frequency of operation. In order to achieve a near break-even at the stated condition a memory instance of *4Mb* (*1024* blocks) is required using the block-footer-DCT configuration (Figure 8.7), and a memory instance of *2Mb* (*512* blocks) is required using the block-header-DCT configuration (Figure 8.8).

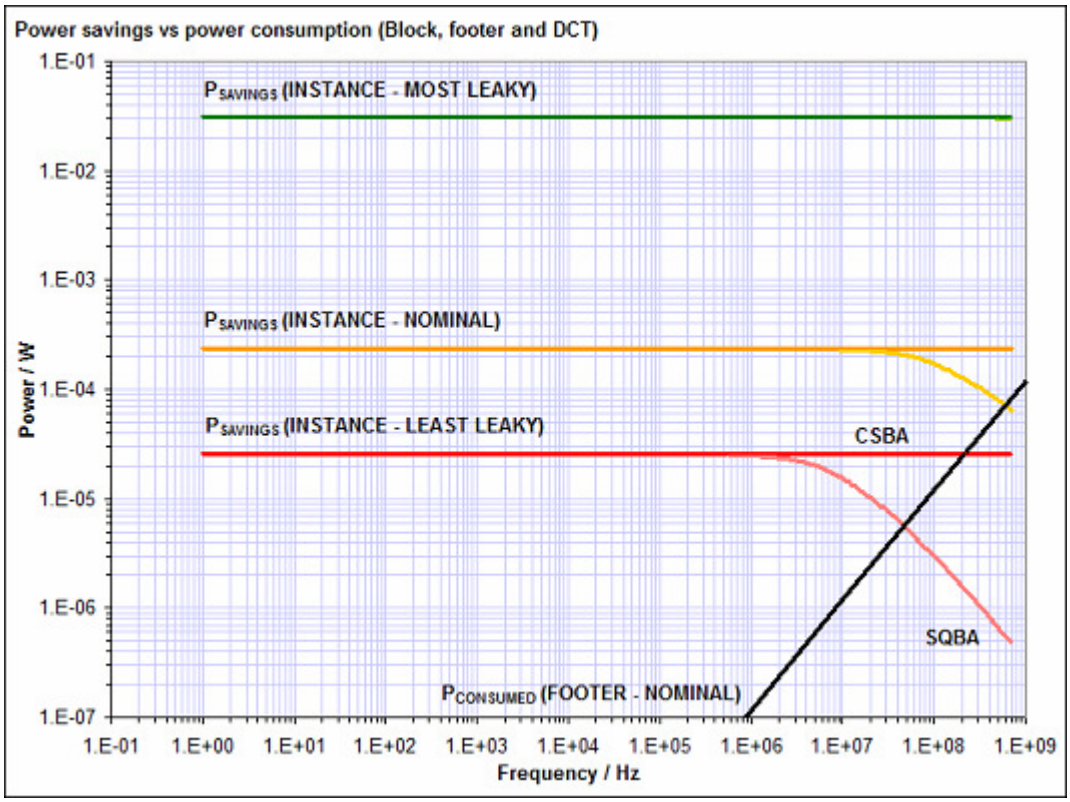


Figure 8.7 *4Mb* instance power savings against power consumed by footer switch (block-footer-DCT)

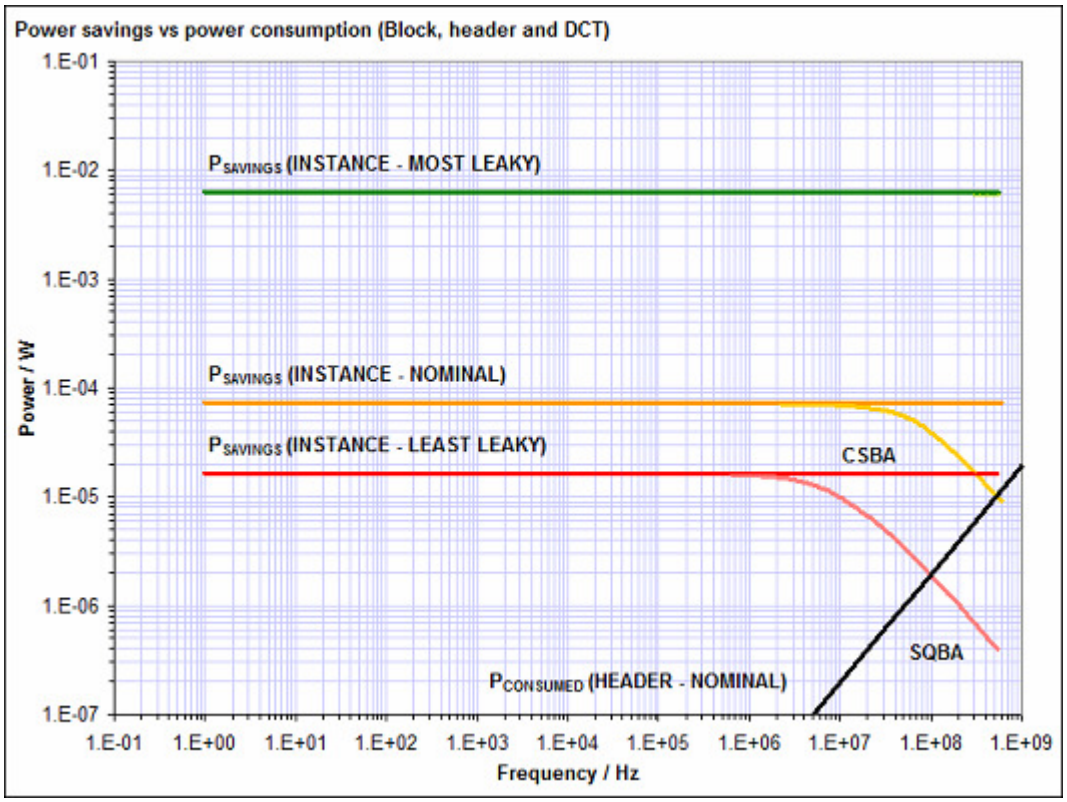


Figure 8.8 *2Mb* instance power savings against power consumed by header switch (block-header-DCT)

In connection to the data presented in Figure 8.7 and Figure 8.8, an estimate of net instance power savings at *500MHz*, in comparison to employing no power saving technique, is presented in Figure 8.9. Significant power savings can now be achieved in the MOST LEAKY corner by both memory architectures. Lesser power savings are achieved in the NOMINAL corner for both memory architectures, using the CSBA scheme. The power savings are the least for the SQBA scheme in the NOMINAL corner for both architectures as it presents the near break-even case. For the block-footer-DCT configuration, power is actually consumed in the LEAST LEAKY corner. However, for the block-header-DCT architecture, some power is still saved in the LEAST LEAKY corner using the CSBA scheme, due to the lower power consumption of the header switch at *500MHz*.

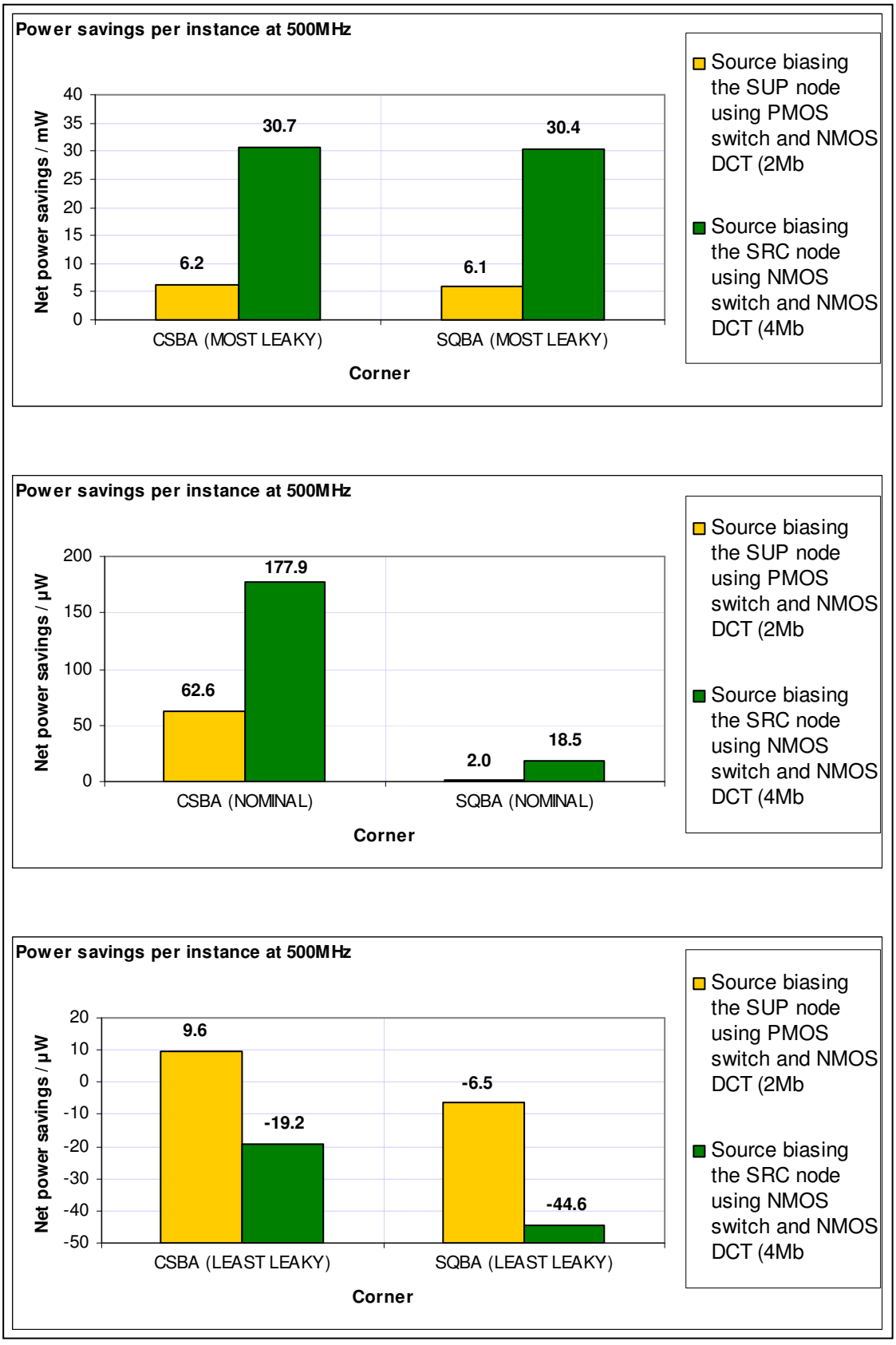


Figure 8.9 Power savings achieved at *500MHz* operation

8.5 Power savings with the actively clamped switch scheme

In chapter 7, it was shown that leakage reduction was only possible for the actively clamped footer scheme in the MOST LEAKY corner. Unlike the case of the DCT architecture, a break-even between power saved and that consumed is not possible in the NOMINAL corner since power is not saved in that corner. As the number of blocks in the instance is increased, leakage reduction improves but the power consumption of the op-amp units also increases since their power consumption is directly proportional to their number. A near break-even in the MOST LEAKY corner has been achieved at *500MHz* with an instance size of *256kb*, using the SQBA scheme. This is graphically illustrated in Figure 8.10.

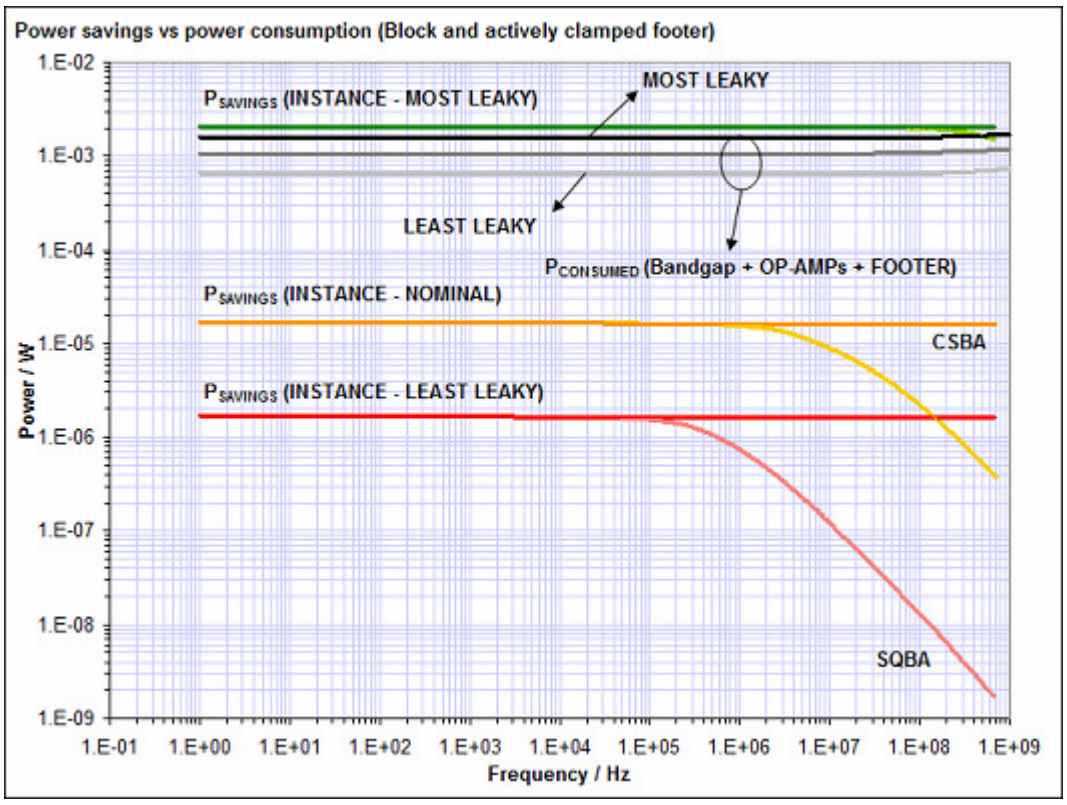


Figure 8.10 *256kb* instance power savings against power consumed by bandgap, op-amps and footer switch

The plots for power consumption curve upwards at high frequencies because the increasing power consumption of the switch is being added to the constant power consumption of the bandgap and the op-amp units.

The results indicate that this particular architecture is only suited to very high temperature applications. Also, significant power savings cannot be achieved with this architecture if the instance size is increased beyond *256kb* since the power consumption of the op-amp units increases proportionally.

In connection to the data presented in Figure 8.10 and, an estimate of instance power savings, in comparison to employing no power saving technique, at the desired frequency of operation is presented in Figure 8.11. As explained earlier, no power is saved in any corner except the MOST LEAKY.

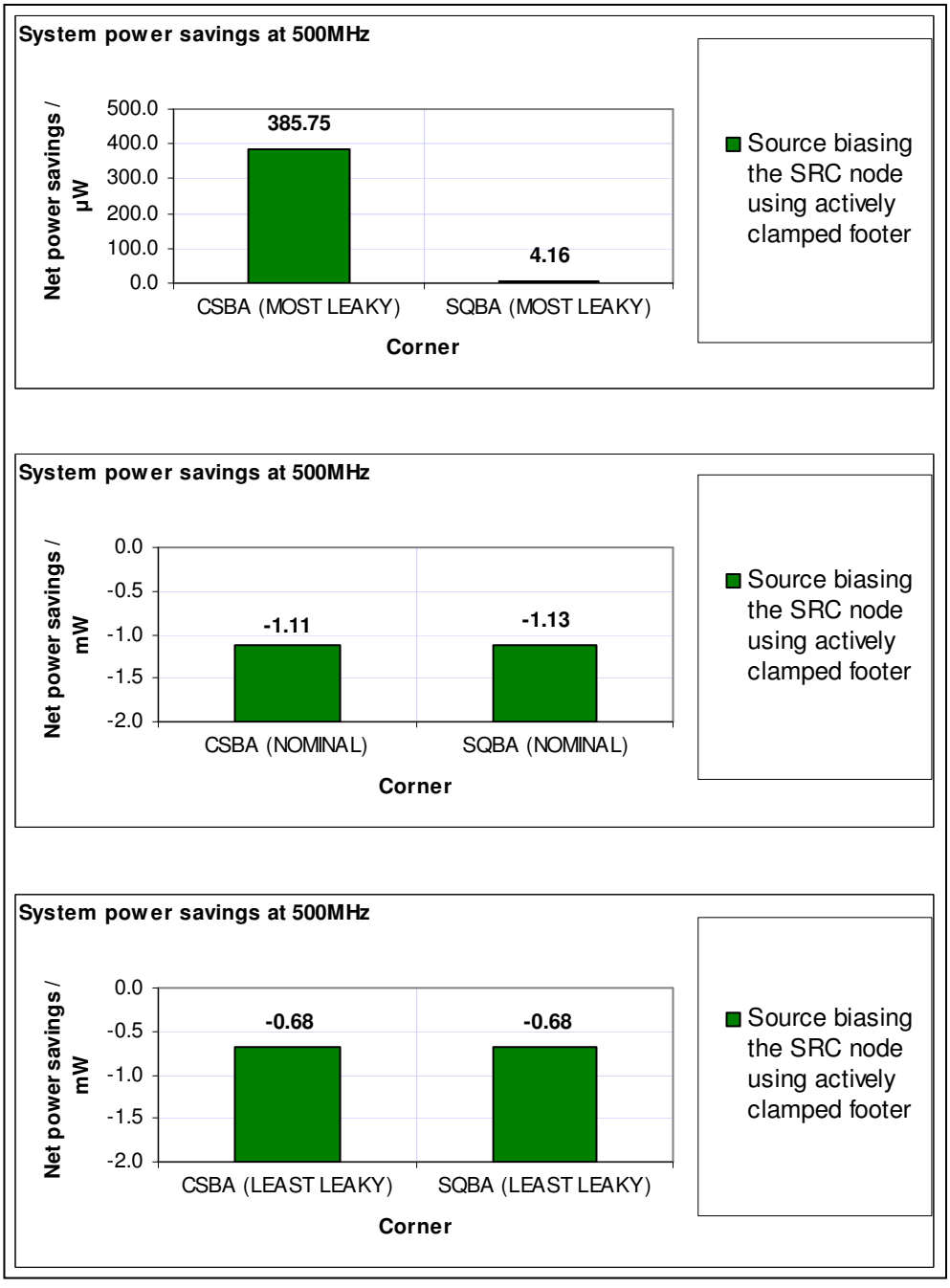


Figure 8.11 Power savings achieved at *500MHz* operation using actively clamped NMOS switch

8.6 Results

A power estimation model has been presented in this chapter to estimate power savings for the architectures selected at the end of chapter 7. Power savings have been estimated for two different memory access schemes. Significant power savings can be achieved using the DCT approach in the MOST LEAKY and NOMINAL corners, while lower power savings are possible with the actively clamped switch scheme in the MOST LEAKY corner only. The complete design trade-off picture for the investigated options is presented here. Figure 8.12 illustrates the block-footer-DCT architecture and Table 8.1 provides a summary of results for this configuration.

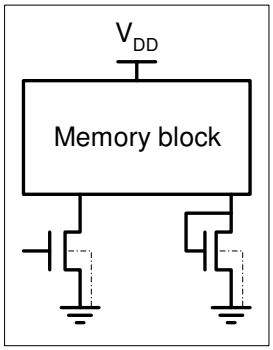


Figure 8.12 The block-footer-DCT architecture

Table 8.1 Summary of results for the block-footer-DCT architecture

	MOST LEAKY	**NOMINAL**	**LEAST LEAKY**
Block size	4096 bits		
SRAM-PD footer size	214µm (996 x 215nm)		
SRAM-PD DCT size	6.9µm (32 x 215nm)		
Total area overhead (switch and DCT)	6.63%		
Percentage reduction in leakage current for a 32kb instance	64.5%	57.2%	42.5%
t_{CYCLE} of proposed architecture	1.48ns	1.41ns	1.45ns
Frequency of proposed architecture	675MHz	709MHz	689MHz
Net power savings of a 4Mb instance at 500MHz	30.7mW (CSBA) 30.4mW (SQBA)	177.9µW (CSBA) 18.5µW (SQBA)	-19.2µW (CSBA) -44.6µW (SQBA)

The maximum operating frequency of the design is *740MHz* in the FAST corner.

Figure 8.13 illustrates the block-header-DCT architecture and Table 8.2 provides a summary of results for this configuration.

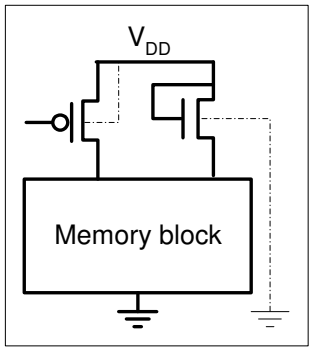


Figure 8.13 The block-header-DCT architecture

Table 8.2 Summary of results for the block-header-DCT architecture

	MOST LEAKY	**NOMINAL**	**LEAST LEAKY**
Block size	4096 bits		
SRAM-PU header size	58µm (645 x 90nm)		
SRAM-PD DCT size	46.7µm (217 x 215nm)		
Total area overhead (switch and DCT)	6.52%		
Percentage reduction in leakage current for a 32kb instance	25.5%	36.5%	57.5%
t_{CYCLE} of proposed architecture	1.80ns	1.67ns	1.86ns
Frequency of proposed architecture	555MHz	598MHz	537MHz
Net power savings of a 2Mb instance at 500MHz	6.2mW (CSBA) 6.1mW (SQBA)	62.6µW (CSBA) 2.0µW (SQBA)	9.6µW (CSBA) -6.5µW (SQBA)

The maximum operating frequency of the design is *675MHz* in the FAST corner.

Figure 8.14 illustrates the actively clamped footer architecture, a configuration that was also briefly investigated to determine leakage reduction and power savings. Table 8.3 provides a summary of results for this configuration.

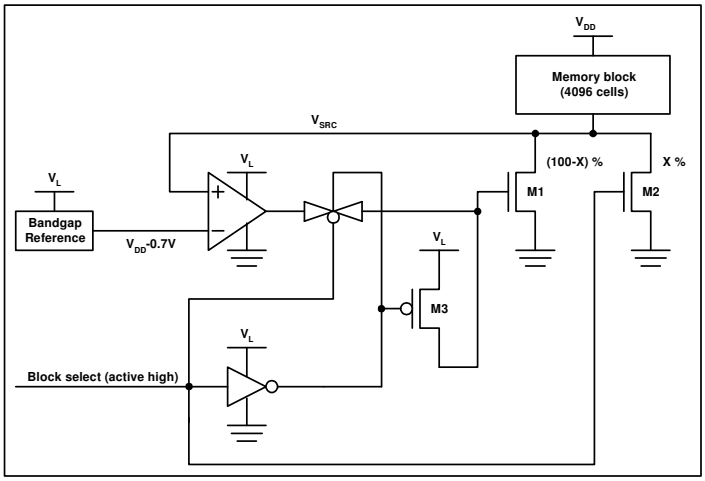


Figure 8.14 The actively clamped footer architecture

Table 8.3 Summary of results for the actively clamped footer architecture

	MOST LEAKY	NOMINAL	LEAST LEAKY
Block size	4096 bits		
SRAM-PD footer size	214µm (996 x 215nm)		
Op-amp area	10µm^2		
Bandgap area	15210µm^2		
Total area overhead for 32kb instance (switch, op-amp units and bandgap)	130.9%		
Percentage reduction in leakage current for a 32kb instance	13.3%	-4868.9%	-18592.1%
t_{CYCLE} of proposed architecture	1.48ns	1.41ns	1.45ns
Frequency of proposed architecture	675MHz	709MHz	689MHz
Net power savings of a 256kb instance at 500MHz	385.8µW (CSBA) 4.1µW (SQBA)	-1.11mW (CSBA) -1.13mW (SQBA)	-0.68mW (CSBA) -0.68mW (SQBA)

The results of the block-footer-DCT architecture are the most promising out of the three architectures presented. It achieves the maximum leakage reduction in the MOST LEAKY and NOMINAL corners for the desired *32kb* instance. It also meets the *500MHz* operating frequency requirement in all corners. It also achieves the highest power savings in the MOST LEAKY and NOMINAL corners. The proposed architecture sees feasibility of use at room temperature and above. Finally, the total area overhead associated with this scheme is only 6.63%.

In Chapter 9, some recommendations for future work in the field of low-power SRAM design are presented.

Chapter 9

Recommendations for future work

SRAMs are a vital building block of most modern SoCs, therefore low-power SRAM design is expected to be an active area of interest amongst researchers for a number of years to come. In most of the publications available to date, the emphasis has been on the reduction of either ACTIVE or STANDBY power. That is because for approximately the same performance figures, one of the two can be reduced at an acceptable penalty in area. Several broad areas of future work are suggested in the following sections.

9.1 Test chip layout

Although a test chip layout was an extended target of this assignment but due to lack of time, the layout could not be finished. It would be a worthwhile exercise to layout the selected block-footer-DCT architecture and carry out a post layout simulation on the design. A difference in pre- and post layout simulations would indicate how well the design was laid out and the effect of increased routing capacitance on the timing critical signals. Finally, a silicon measurement of leakage currents would confirm exactly how much leakage reduction is possible with the proposed architecture, compared to simulation results.

9.2 Sub-threshold SRAM design

Sub-threshold operation is the lower limit to low-power digital design. Since low-power design is driving the electronics market today, it would be a worthwhile effort to investigate the operation of the TSMC *45nm* 6T SRAM cell in sub-threshold region of operation. A complete sub-threshold SRAM

design together with a comprehensive analysis of the associated area-speed-power trade-offs would be a notable contribution to knowledge.

9.3 V_{SRC} controller design

This suggestion is indirectly related to low-power SRAM design. Work could be undertaken to investigate a potential circuit/module that can accurately control the gate-source voltage of the footer during periods of STANDBY. A block level diagram of the proposal is provided in Figure 9.1.

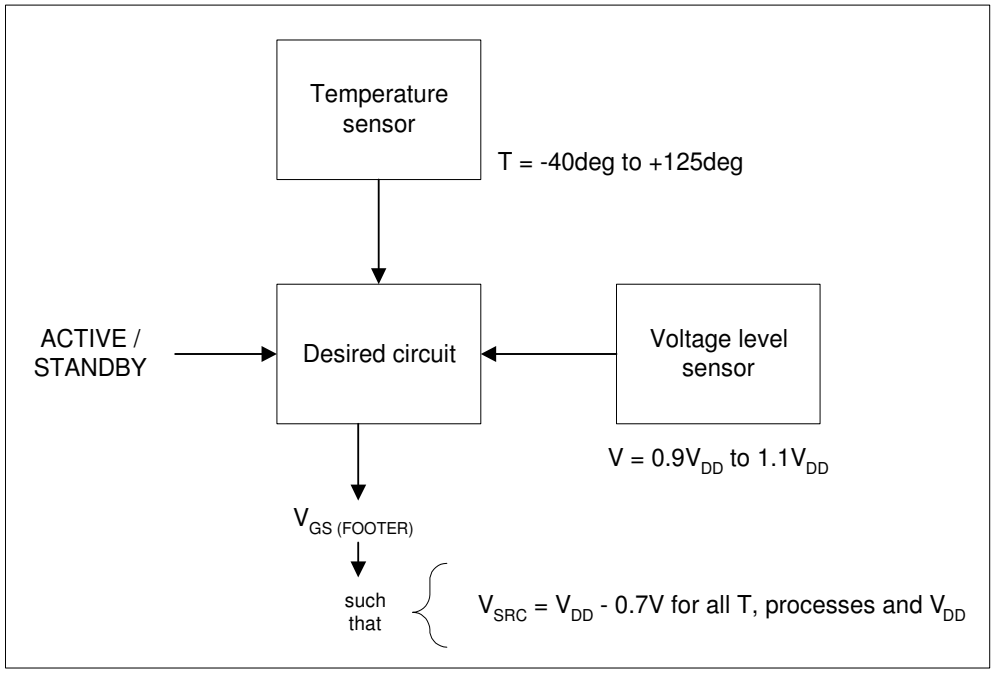


Figure 9.1 Block diagram for a potential SRC node voltage controlling circuit

The area overhead may be acceptable if the desired circuit could be shared between all memory blocks and if it is able to control the footer V_{GS} accurately over the entire operating temperature range as well as for supply voltage variation. Realization of this circuit would allow maximum possible reduction in leakage currents for large memories using source biasing, provided the highlighted modules do not consume significant power.

9.4 Block address decoder design

For SRAM designs employing a high locality of reference (that is when a lot of read/write operations take place within a single block), a block address decoder could be designed such that the software running on top of the hardware is able to keep the footer (block select signal) on for multiple clock cycles. That would significantly reduce the power consumption associated with driving the footer at every clock edge. Also, after the first read/write operation, subsequent read/writes to the same block

would not incur the additional access time penalty (associated with bringing the block from STANDBY mode into ACTIVE mode) talked about this report. For subsequent read/writes, the proposed architecture would be able to run at *1GHz* (*1ns* cycle time). The design of such a decoder would lead to higher power savings in the MOST LEAKY and NOMINAL corners and could even allow power savings in the LEAST LEAKY corner for the block-footer-DCT architecture at high frequencies.

Appendix A

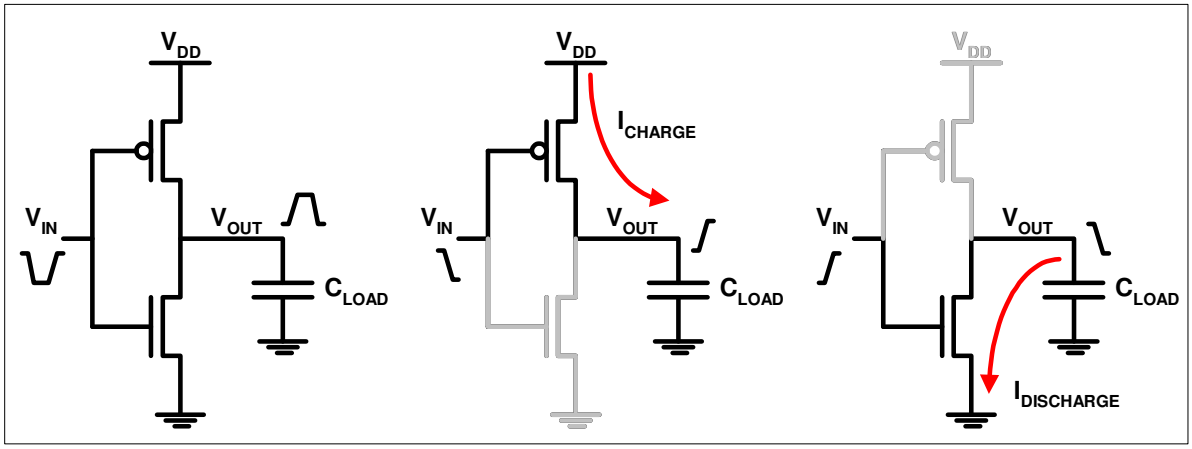
Sources of power dissipation in CMOS circuits

A brief overview of sources of power consumption in digital circuits is presented here to aid the understanding of work presented in this thesis. There are three sources of power dissipation, namely dynamic power dissipation, short-circuit power dissipation and power dissipation due to leakage currents. These contributions can be combined together in Equation (A.1) [1] to represent the total power dissipation.

$$P_{tot} = P_{dynamic} + P_{short-circuit} + P_{leakage} \quad \text{(A.1)}$$

A.1 Dynamic power dissipation

Dynamic power dissipation is attributed to the charging and discharging of capacitances in a digital circuit. In Figure A.1b, when the input signal goes through a falling transition, the PMOS starts to conduct and the NMOS is switched off. Current is drawn from the power supply that charges the load capacitor to V_{DD}. The energy drawn from the power supply during the charging phase is $C_L \cdot V_{DD}^2$ [2]. Half of this energy is stored on the load capacitor and the other half is dissipated in the PMOS device and associated interconnects during charging.



(a) (b) (c)

Figure A.1 CMOS inverter (a) inverter charging a capacitive load (b) inverter discharging a capacitive load (c)

When a rising transition takes place on the input terminal (Figure A.1c), the NMOS transistor starts to conduct and the PMOS is switched off. The pull down transistor discharges the load capacitor and the stored energy is therefore dissipated in the NMOS transistor and the associated interconnects. The dynamic power dissipation can then be expressed as Equation (A.2) [2]:

$$P_{dynamic} = C_L \cdot V_{DD}^{\ 2} \cdot f \cdot \alpha \qquad \text{(A.2)}$$

where C_L is the load capacitance, V_{DD} is the supply voltage, f is the frequency of operation and α is the activity factor. Dynamic power dissipation can be reduced by reducing any one or all of the parameters listed on the right hand side of Equation (A.2).

A.2 Short-circuit power dissipation

The short-circuit power dissipation occurs as a consequence of the availability of a direct current path between the supply node and ground during a switching operation. The direct current path comes into existence during the high-to-low as well as low-to-high input transitions, in which both the NMOS and PMOS transistors simultaneously conduct for a short duration. Figure A.2 provides a pictorial representation of the input voltage and the short circuit current in an inverter, plotted as a function of time.

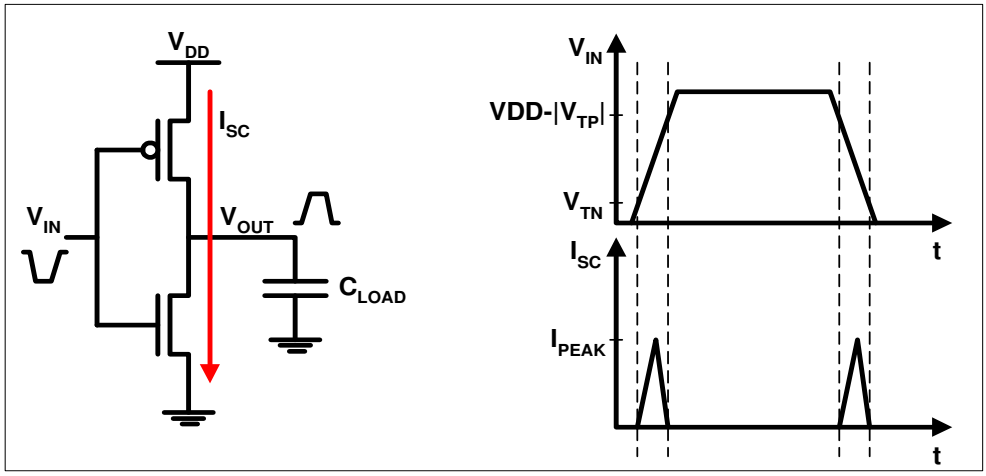


Figure A.2 Short circuit current in an inverter (re-drawn from [2])

The peak value of current is a function of transistor size and the slopes of V_{IN} and V_{OUT} [2]. If the current spikes are approximated as triangles and a symmetrical inverter drive is assumed, the energy consumed per switching can be expressed by Equation (A.3) and the power consumption can be expressed by Equation (A.4) [2].

$$E_{short-circuit} = V_{DD}\frac{I_{PEAK}\cdot t_{sc}}{2} + V_{DD}\frac{I_{PEAK}\cdot t_{sc}}{2} = V_{DD}\cdot t_{sc}\cdot I_{PEAK} \tag{A.3}$$

$$P_{short-circuit} = V_{DD}\cdot t_{sc}\cdot I_{PEAK}\cdot f \tag{A.4}$$

where I_{PEAK} is the maximum short-circuit current, V_{DD} is the supply voltage, t_{sc} represents the time for which both devices are on and f is the frequency of operation. Short-circuit power dissipation can be lowered by reducing the transistor aspect ratio and the supply voltage level, by using advanced technology generations and by reducing the input signal transition times [1].

A.3 Power dissipation due to leakage currents

In CMOS technologies beyond *90nm*, leakage currents of low-V_T transistors heavily dominate the standby power consumption of a design [3]. Leakage currents can be broadly categorized into four main types. These include sub-threshold leakage, gate leakage, gate induced drain leakage and junction leakage. A detailed account of MOSFET leakage currents in *45nm* technology is provided in Appendix C.

Gate leakage occurs for as long as a potential difference exists across the oxide layer. Sub-threshold leakage and gate-induced drain leakage occur when the device is in the off state. Junction leakage

occurs for as long as a drain potential with respect to the bulk is present. Different leakage mechanisms dominate under different bias conditions and temperatures.

Leakage currents can be reduced by using high-V_T transistors, operating at reduced temperatures, by employing high-*k* gate dielectrics and by incorporating innovative circuit design techniques including operating the circuit at reduced supply voltage levels.

A.3 References

[1] D. Soudris, C. Piguet and C. Goutis "*Designing CMOS Circuits for Low Power*," Springer, 2002

[2] J. Rabaey, A. Chandrakasan, and B. Nicolić, "*Digital Integrated Circuits: A Design Perspective*," Second Edition, Prentice Hall, 2003

[3] Y. Takeyama, H. Otake, O. Hirabayashi, K. Kushida and N. Otsuka, "*A Low Leakage SRAM Macro with Replica Cell Biasing Scheme*," IEEE Journal of Solid State Circuits, Vol. 41, Issue 4, April 2006

Appendix B

SRAM functional overview and 6T SRAM cell operation

B.1 SRAM: Functional overview

A block-level diagram of the SRAM is presented in Figure B.1. An SRAM instance comprises a memory matrix, normally arranged in a two-dimensional array of N rows and M columns, extending over one or more pages. SRAMs are usually self-timed in that an internal module generates all the necessary timing signals required for the operation of the SRAM.

B.1.1 The row-decoder

The word line driver (part of the row/X-decoder in Figure B.1) allows the selection of a specific word line in the SRAM instance using the internally generated timing signals. The row decoder can be either single or multi-stage in its architecture. In a single stage decoder, the entire decoding is realized in a single block, which makes the architecture a promising option for small-sized memory instances. Typically in a multi-stage decoding scheme, the most significant bits are pre-decoded in the first stage, which provide enable signals for a second stage decoder, which then selects a particular word line. This type of decoding is more suited to larger SRAM instances.

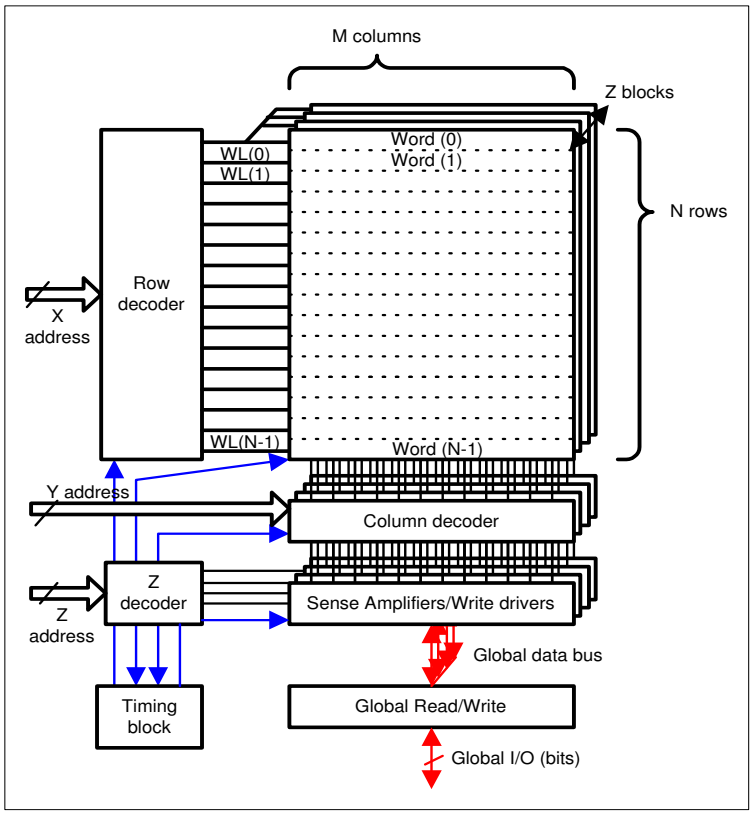


Figure B.1 SRAM block diagram (re-drawn from [1])

B.1.2 The column-decoder

The column/Y-decoder allows multiple columns in the array to share a single sense amplifier (SA), which greatly reduces the area overhead. The Y-decoder also provides isolation between bit lines and the SA inputs during reading and prevents the complete discharge of bit lines by the SA.

B.1.3 The timing block

The function of the timing block is to accurately mimic the delay associated with bit line discharge and therefore to provide a SA enable signal at the right instant. The timing module normally employs a finite state machine (FSM) that moves through a set of pre-defined states in order to generate sequentially timed signals for each block in the SRAM. Normally a chip select signal, global clock and a reset signal are inputs to the FSM [2]. The timing module is driven by a dummy-loop, which consists of a dummy column and a dummy row, similar in height and width to the actual memory matrix. This approach allows replication of load capacitance and delays associated with bit- and word lines. The dummy-loop based timing block provides accurate timing control over deactivation of the word line and activation of the SA. The SA employed in the dummy-loop is usually an inverter, which resets the FSM in the timing module when it toggles its state.

B.1.4 The write driver

The function of the write driver circuit (Figure B.2) is to quickly discharge one of the bit lines from the pre-charge level (V_{DD}) to below the write margin of the SRAM cell [1].

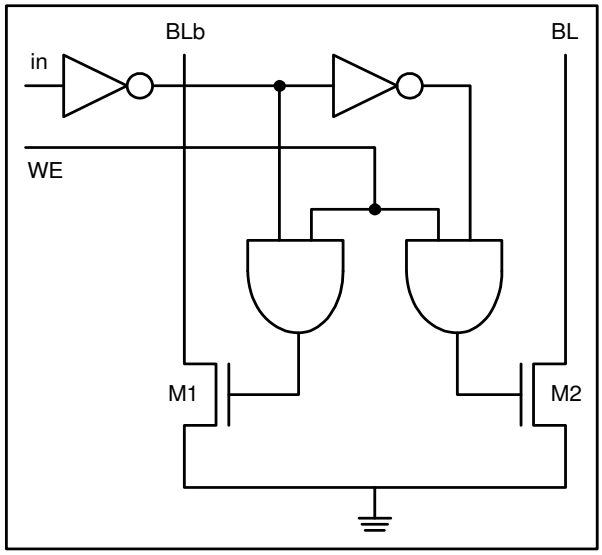


Figure B.2 A typical write driver circuit (re-drawn from [1])

Each bit line pair (or a column) in the SRAM instance requires one write driver circuit so that data may be written to each cell within the column. Depending on the data to be written to the cell, and using a write enable signal, either M1 or M2 is activated in order to discharge the desired bit line. A write operation is usually faster than a read operation primarily because a write operation does not include bit line discharge by the cell and sensing delays.

B.1.5 Precharge, equalization and sense amplifier

The role of precharge, equalization and SA circuits can be better explained with the aid of Figure B.3, which shows a classic SRAM circuit with precharge, SRAM cell, column multiplexer, SA reset and equalization and SA. Also shown are signal waveforms during a read operation.

The role of the pre-charge circuit is to charge the bit lines to V_{DD} potential in the absence of a read/write operation. The equalization transistor equalizes the voltage between the bit lines and ensures that the differential voltage present between the bit lines at the start of a read operation is very small, if not zero. At the start of a read operation, the bit lines are precharged to V_{DD} and then left floating. The precharge level is set to V_{DD} to improve noise robustness and to ensure a non-destructive read under process variations [3]. Also, since the V_{DD} pin is already available on the chip, it makes

sense to use that potential for precharge rather than using an alternate voltage level, which would normally come at an area cost, even if it provided higher design efficiency.

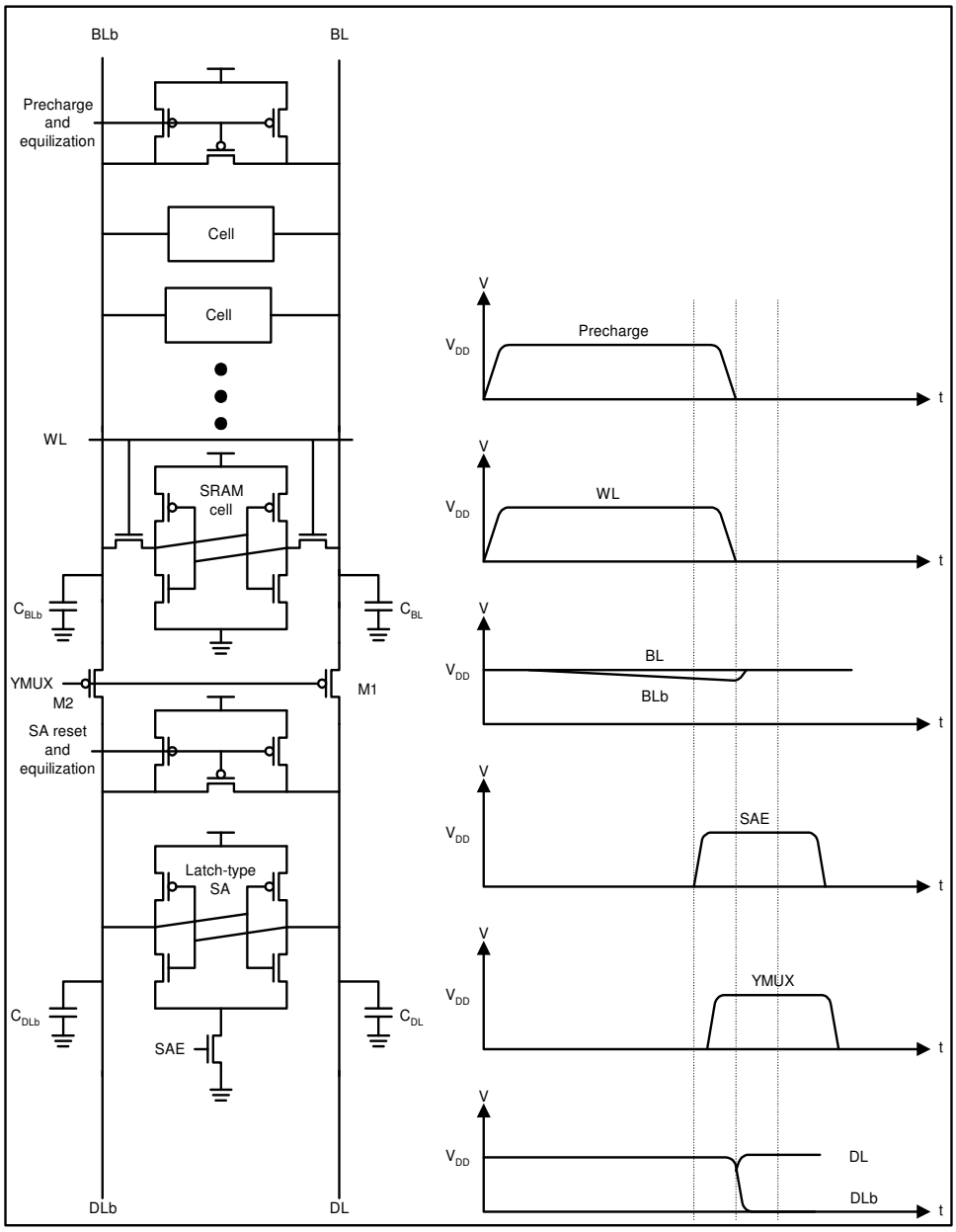


Figure B.3 SRAM circuit with pre-charge, column MUX, SRAM cell, SA and signal waveforms during a read operation (diagram is a modified version of the one appearing in [4])

When the word line is activated, the SRAM cell starts to discharge one of the bit lines (the one connected to the cell node storing logic *0*). The minimum acceptable differential voltage on the bit lines directly governs the read access time of the memory and therefore its speed. A larger bit line differential voltage is advantageous for reliable sensing but the cell takes longer to develop that differential voltage. An optimum exists between time taken by the cell to develop the bit line

differential (usually *100m*V [5]) and the SA to amplify the input data to full swing CMOS signal levels [4].

The inclusion of the SA in the SRAM not only allows faster reading but also permits the use of a smaller SRAM cell [3]. Since the inputs of the SA are not isolated from the outputs, M1 and M2 are needed to isolate the bit lines from the SA and therefore prevent full discharge of one of the bit lines, which would otherwise add to delay and power.

B.1.6 SRAM access and cycle time

The access time is the minimum amount of time required to read data from the memory, measured with respect to the initial rising edge of clock in the SRAM read operation. The typical value of this parameter is usually the figure quoted on product data sheets [6]. It comprises delay of the word line decoder, the bit line discharge time (to a sufficient differential voltage), time taken by the SA to amplify the bit line differential voltage to full swing CMOS voltage level and the propagation delay though interconnects before data appears at the pins.

The cycle time is the amount of time required to perform a single read or write operation and reset the internal circuitry before a second operation can begin [6]. This time is usually designated by the fastest clock cycle. Normally the access time is less than the cycle time. In certain cases however, the access time may be equal to or greater than the cycle time, particularly where pipelining is employed.

B.2 6T CMOS SRAM cell

The standard six-transistor CMOS SRAM cell, shown in Figure B.4, is the most widely employed architecture due to its small size, differential nature (as opposed to single ended) and high robustness [1].

The cell consists of a cross-coupled inverter pair (M1/M5 & M2/M6) that forms a latch structure, and a pair of NMOS pass gates (M3 & M4). The pass gates connect the internal storage nodes of the cross-coupled inverters to the bit lines, allowing data to be read (non-destructively) from or written to the memory cell. During STANDBY mode, they provide cell isolation.

The cell design must maintain a balance between robustness, speed, area, leakage and yield [5]. Reduced cell size normally increases the speed and lowers the power consumption due to reduced capacitances. They also help reduce the total area, which in turn lowers the bit line and word line capacitance, which in turn improves the memory access time. On the other hand, the drawback of smaller transistors is reduced cell stability [1].

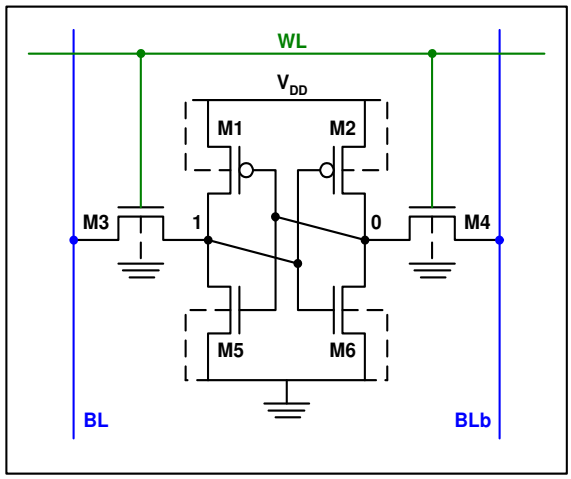


Figure B.4 Six-transistor standard CMOS SRAM cell

Normally the NMOS pull-down transistors are the largest and the PMOS pull-up transistors the smallest. The NMOS pass gates are generally bigger than the pull-up devices and smaller than the pull-down devices. They are carefully sized so that they are strong enough to allow data to be changed at the storage nodes during writing but weak enough not to flip the state of the cell during reading.

B.2.1 Read operation and static noise margin

At the start of a read cycle the bit line pair (BL/BLb) is precharged to V_{DD} and then left floating. These bit-lines then act as charged capacitors. Thereafter, the word line (WL) is asserted, which enables the two pass gates. As a result, read current (I_{READ}) begins to flow from BLb through the pass-gate M4 and the pull-down transistor M6 to *GND* (refer to Figure B.5). BLb is therefore discharged by the memory cell. A sense amplifier is used to speed up the reading process, and typically kicks in when the potential difference between the bit lines reaches approximately *100mV* [5]. Figure B.5 shows the direction of read current flow through the SRAM cell.

The amount of read current flowing through the cell directly determines how fast the bit lines can be discharged. The magnitude of the read current therefore significantly influences the maximum speed of the memory.

At the time read current flows through the cell, voltage at the node storing logic '*0*' temporarily rises. Since the pull-down device M6 is stronger than the pass gate M4, the node is securely held below the trip point of the inverter M1/M5. Data is therefore prevented from getting corrupted during a read operation. The maximum voltage that can be applied to the internal nodes of the cell before the cell flips its state is termed as the static noise margin (SNM). SNM is investigated in detail in [7].

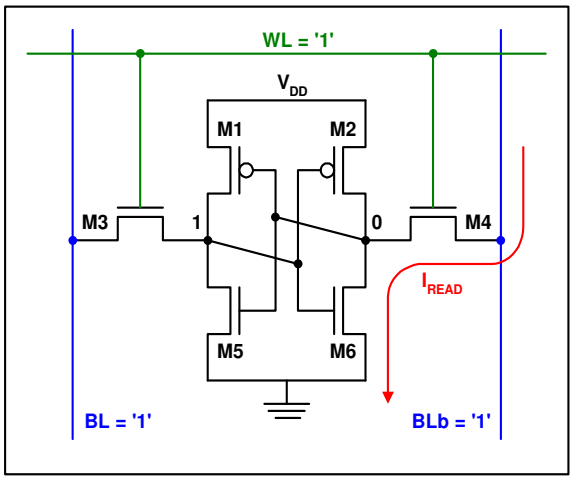


Figure B.5 Read current flow through SRAM cell

An important term associated with the SRAM cell is the cell ratio (CR), which is defined by Equation (B.1) [8]. CR is the same for M5/M3 due to symmetrical cell structure.

$$CR = \frac{W_{M6} / L_{M6}}{W_{M4} / L_{M4}} \tag{B.1}$$

Higher values of CR make it less likely for the voltage at the internal node (storing a '*0*') to go past the trip point of inverter M1/M5 during reading. Also, higher values of CR provide higher I_{READ} values and improved stability at the cost of increased cell area. Lower values of CR decrease the cell area but also compromise somewhat on speed and stability [1].

B.2.2 Write operation and write margin

During writing, a write driver discharges one of the bit lines and maintains the other at V_{DD}. Thereafter, the word line (WL) is asserted, which enables the two pass gates. As a result, write current (I_{WRITE}) begins to flow from V_{DD} through the pull-up transistor M1 and pass-gate M3 into BL. Figure B.6 shows the direction of write current (I_{WRITE}) flow through the SRAM cell.

The pull-up device M1 tries to maintain high voltage at the node storing logic '*1*' but is unable to do so due to the stronger pass gate M3. Eventually M3 pulls down the internal node originally held at logic '*1*.' The cell therefore flips it state. The minimum bit line voltage at which the cell flips it state is termed as the WM [9]. The WM value and variation is dependant on the cell design, the memory matrix size and process variation. A cell is considered not writeable if the value of WM drops below zero in the worst case [1].

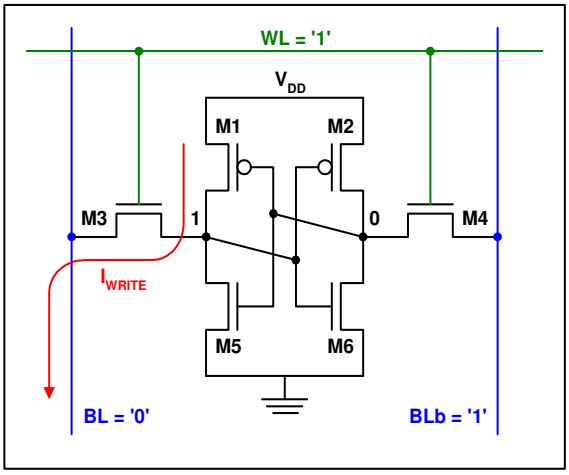


Figure B.6 Write current flow through SRAM cell

In addition to maintaining high voltage at the node storing logic '*1*,' M1 also prevents the discharge of this node that can occur as a result of leakage currents of the pull down transistor M5 during STANDBY.

Another important term associated with the SRAM cell is the pull-up ratio (PR), which is defined by Equation (B.2) [8].

$$PR = \frac{W_{M1} / L_{M1}}{W_{M3} / L_{M3}} \tag{B.2}$$

The maximum allowed PR is governed by the switching threshold of inverter M2/M6 and threshold voltage process corner of the pass gate [1]. Stronger pass gates and slightly weaker pull up transistors are normally employed to ensure a robust write operation in the worst case process corner.

Also, at the time write current I_{WRITE} flows through the accessed cells, parasitic read current may flow through all other cells enabled by the same word line but which are not written to during that particular write cycle.

B.3 References

[1] A. Pavlov, M. Sachdev, "*CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies: Process-Aware SRAM Design and Test*," Springer, 2008

[2] X. Dong, "*Low Voltage Embedded SRAM Design*," MSc Thesis, CAS Group, TU Delft, the Netherlands, 2008

[3] K. Itoh, M. Horiguchi, H. Tanaka, "*Ultra-Low Voltage Nano-Scale Memories*," Springer, 2007

[4] T.S. Doorn, J.A. Croon, E. J. W. ter Maten and A. Di. Bucchianico, "*A yield centric statistical design method for optimization of the SRAM active column*," to appear in ESSCIRC 2009 proceedings

[5] T. S. Doorn, R. Salters, L. E. Villagra, "*SRAM Design Challenges: A research view on current status and future work*," Technical Note NXP-R-TN-2007/00066, Issued 6/2007

[6] IBM, "*Understanding Static RAM Operation*," Application Note, 1997, http://www.ece.cmu.edu/~ece548/localcpy/sramop.pdf

[7] E. Seevinck, F. J. List, J. Lohstroh, "*Static-noise margin analysis of MOS SRAM cells*," IEEE Journal of Solid-State Circuits, pp. 748-754, vol. 22, Issue no. 5, 1987

[8] J. Rabaey, A. Chandrakasan, and B. Nicolić, "*Digital Integrated Circuits: A Design Perspective*," Second Edition, Prentice Hall, 2003

[9] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepali, Y. Wang, B. Zheng, and M. Bohr, "*A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply*," IEEE Journal of Solid State Circuits (JSSC), vol. 41, pp. 146–151, January 2006

Appendix C

Leakage currents in 45nm SRAM cell

The data retention current in an SRAM cell consists of a number of components. Three sub-threshold currents, eleven gate leakage (or gate-tunneling) currents and five gate induced drain leakage (GIDL) currents flow in each cell. Figure C.1 illustrates the leakage current components of the SRAM cell.

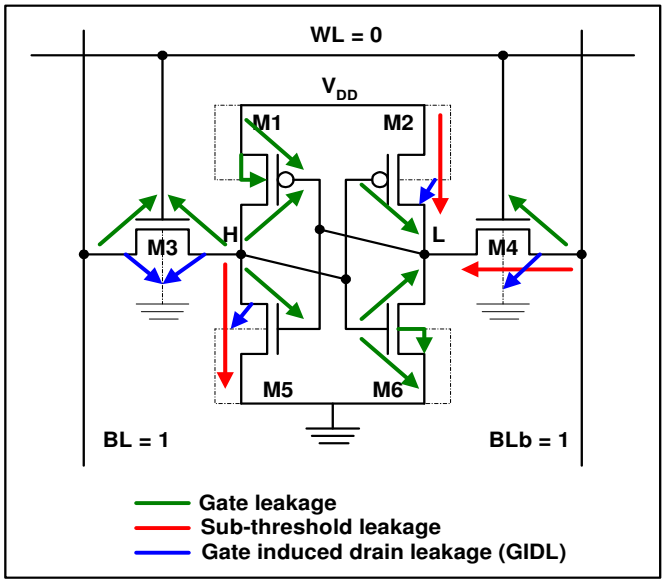


Figure C.1 Leakage currents in a standard 6-T CMOS SRAM cell

Junction leakage (comprising of diffusion current, generation current, avalanche current and zener tunneling current) [1] is a second order effect in *45nm* technology [2] and is therefore ignored.

C.1 Sub-threshold leakage current

Sub-threshold leakage current is the current that flows in the channel (between drain and source) for $V_{GS}<V_T$. This current increases exponentially with temperature and roughly by an order of magnitude for every *100mV* decrease in threshold voltage [3]. The threshold voltage itself is temperature and V_{DS} dependant. Drain induced barrier lowering (DIBL) and temperature-dependant number of carriers in the channel are the reasons for the dependencies. V_T typically varies by about *1mV/K* [2].

In short-channel devices V_T reduces due to an effect called DIBL. In normal transistor operation, carriers face a barrier close to the source, which they have to cross. An increased drain-source bias lowers the height of this barrier. This results in reduced threshold voltage both in sub-threshold and super-threshold operating regions [4]. Also, DIBL becomes stronger with back-biasing [2]. Sub-threshold leakage current is governed by Equations (C.1 – C.3) [5], where W and L are the transistor width and length, T is the temperature, μ_{eff} is the effective mobility, C_{ox} is the gate oxide capacitance, C_D is the depletion capacitance, k is the Boltzmann constant, q is the electronic charge, V_{DS} is the drain-source voltage and V_{GS} is the gate-source voltage.

$$I = I_0 \cdot \exp\left(\frac{V_{GS} - V_T}{mkT/q}\right) \cdot \left(1 - \exp\left(-\frac{V_{DS}}{kT/q}\right)\right) \tag{C.1}$$

$$I_0 = \mu_{eff} \cdot C_{ox} \cdot \left(\frac{W}{L}\right) \cdot (m-1) \cdot \left(\frac{kT}{q}\right)^2 \tag{C.2}$$

$$m = 1 + \frac{C_D}{C_{ox}} \tag{C.3}$$

Sub-threshold currents dominate the SRAM cell leakage current at and above room temperature. They can be reduced by using high-V_T transistors. Sub-threshold currents can cause read failures in cases where the pass gates are low-V_T devices. In such a case the leakage currents of all non-accessed cells in a column may equal or surpass the read current of the accessed cell in the sane column. This problem can be alleviated by using high-V_T pass gate transistors [6].

C.2 Gate induced drain leakage (GIDL or gate induced band to band tunneling) current

When gate voltage is low and drain voltage is high (in case of NMOS device), electron-hole pairs can be generated right under the gate-drain overlap region. In such a scenario, a negative gate-drain

voltage then exists since $V_G<V_D$. This condition would force electrons in the drain region to be pushed away from the surface. Since the drain is heavily doped, it is hard to significantly deplete it of electrons. Due to the high doping concentration in the drain, the depletion region formed between the gate oxide and drain as a result is very thin. The electrons therefore tunnel across this depletion region from the valence band into the conduction band of the drain [1].

As a result, additional electrons flow out the drain terminal. Also, empty states are left behind in the valence band. Those holes flow out through the substrate contact. The band to band tunneling is therefore controlled by the gate. The leakage principle is illustrated in Figure C.2.

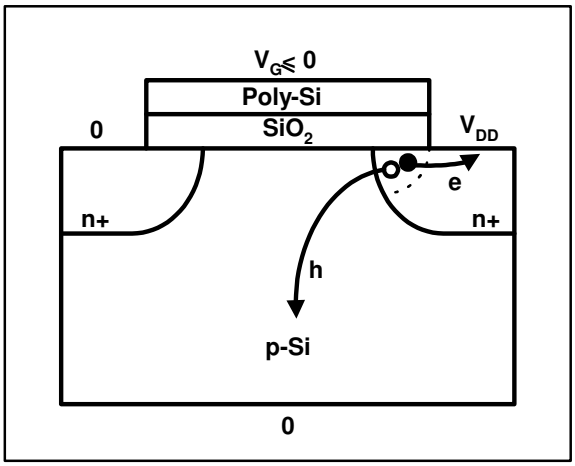


Figure C.2 GIDL in NMOS

The current depends heavily on the size of the gate-drain overlap and the field across it. With a negative V_{GB} for NMOS transistors and a positive V_{GB} for PMOS transistors, the drain potential is lowered which in turn lowers the drain barrier height, hence increasing the leakage [2]. GIDL is very weakly temperature dependent and governed by Equation (C.4) [2], where A, B and C are constants, V_{DB} is the drain-bulk junction voltage and V_{GD} is the gate-drain overlap oxide voltage.

$$I = A \cdot V_{DB} \cdot \left({V_{GDov}}^2 + \left(C \cdot V_{DB} \right)^2 \right) \cdot \exp\left(- \frac{B}{\sqrt{{V_{GDov}}^2 + \left(C \cdot V_{DB} \right)^2}} \right) \tag{C.4}$$

V_{GD} is the most important voltage for GIDL, but GIDL also increases with an increase in V_{DB} [2]. GIDL can be reduced by lowering V_{DD} or by using high-k gate insulator materials [6].

C.3 Gate leakage current

Gate tunneling current primarily flows from gate to the source for NMOS devices and from source to the gate for PMOS devices, through the channel [6]. Gate leakage also occurs due to gate-drain and gate-source overlap regions. A thorough gate-leakage model for the 6-T SRAM cell is presented in [7]. Gate leakage in a MOSFET is primarily of two types depending on the thickness of the oxide layer. If the oxide layer is thick, the tunneling current through the insulator is small and the MOS structure behaves like a typical capacitor. However, if the oxide layer is thin, the tunneling current is high.

The difference between GIDL and gate leakage is that GIDL is dependant on the drain-bulk voltage, whereas gate leakage is not. However, gate leakage, similar to GIDL, can be reduced by lowering V_{DD} or by using high-*k* gate insulator materials [6]. Gate leakage, like GIDL, is also very weakly temperature dependant.

C.3.1 Fowler-Nordheim Tunneling

When gate oxides are greater than roughly *5nm* in thickness [8], Fowler-Nordheim Tunneling (FNT) mechanism dominates. Under this mechanism, when a positive voltage is applied to the gate (in case of NMOS devices), a potential drop is established across the oxide layer, which effectively reduces the barrier height for electrons to tunnel through. As a result, the carriers (electrons) tunnel through only a part of the insulator layer, after which the rest of the insulator layer does not hinder the current flow. Therefore, the higher the electric field across the oxide, the smaller is the tunneling distance for the carriers. FNT current is a temperature independent component of the total SRAM cell leakage and is described by Equation (C.5) [8], where E is electric field across the insulator and C_1 and C_2 are constants.

$$J = C_1 \cdot E^2 \cdot \exp\left(-\frac{C_2}{E}\right) \tag{C.5}$$

C.3.2 Direct Tunneling

When gate oxides are less than *5nm* in thickness [8], Direct Tunneling (DT) mechanism dominates under which electrons (in case of NMOS devices) near the bottom of the conduction band can tunnel directly through the entire oxide layer. It is a far more efficient process in that it does not depend on the electric field inside the insulator. It depends only on the oxide thickness and the height of the barrier the electrons have to tunnel through. Readers are encouraged to refer to [9] for detailed information on DT current.

C.4 References

[1] M. Lundstrom, "*EE-612 Lecture 16 MOSFET Leakage*," Electrical and Computer Engineering, Purdue University, West Lafayette, IN USA, Fall 2008
http://nanohub.org/resources/5690/download/2008.10.28-ece612-l16.pdf

[2] T. S. Doorn, "*Leakage reduction in SRAM cells*," Technical Note NXP-R-TN-2008/00084, Issued 4/2008

[3] Deepaksubramanyan, B.S., Nunez, A., *Analysis of subthreshold leakage reduction in CMOS digital circuits*, Circuits and Systems 2007, MWSCAS 2007, 50th Midwest Symposium on Volume, Issue, 5-8 Aug 2007, pp. 1400-1404

[4] Harry Veendrick, "*Deep-Submicron CMOS ICs: From Basics to ASICs*," Kluwer Academic, 1998

[5] M. Lundstrom, "*EE-612 Lecture 12 Subthreshold Conduction*," Electrical and Computer Engineering, Purdue University, West Lafayette, IN USA, Fall 2006
http://nanohub.org/resources/1825/download/2006.09.18-ece612-l12%20subthreshold.pdf

[6] K. Itoh, M. Horiguchi, H. Tanaka, "*Ultra-Low Voltage Nano-Scale Memories*," Springer, 2007

[7] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, Y. Yamagami, T. Suzuki, A. Shibayama, H. Makino and S. Iwade, "*A 90-nm Low-Power 32-KbB Embedded SRAM with gate leakage suppression circuit for mobile applications*," IEEE Journal of Solid State Circuits, Vol. 39, Issue 4, April 2004

[8] S. M. Sze and K. K. Ng, "*Physics of Semiconductor Devices*," Third Edition, Wiley Inter Science, USA, 2007

[9] Y. Taur and T. H. Ning, "*Fundamentals of Modern VLSI Devices*," Cambridge University Press, USA, 2000