LUND UNIVERSITY

DEPARTMENT OF ELECTRICAL AND INFORMATION TECHNOLOGY

MASTER OF SCIENCE THESIS

Custom-Cell Design for Sub-Vt Memories

Supervisor:
Joachim Rodrigues

Author:
Babak Mohammadi

Advisers:
Pascal Meinerzhagen
Yasser Sherazi
Oskar Andersson

Lund 2012

©

The Department of Electrical and Information Technology
Lund University
Box 118, S-221 00 LUND
SWEDEN

This thesis is set in Computer Modern 10pt,
with the LaTeX Documentation System

Abstract

Supply voltage scaling is a very popular technique to reduce energy dissipation, and leads to near-threshold or even subthreshold circuit operation when applied aggressively. Digital designs may be conveniently synthesized with commercially available standard-cell libraries which unfortunately are not optimized for scaled voltages. Moreover hard macros like RAM may not work at all at aggressively scaled supply voltages. Therefore, it is desirable to a small custom designed standard-cell library allowing for the automated synthesis of memories. In this project, various leakage minimization techniques like transistor stacking, transistor channel stretching, and hardware minimization are evaluated. The target technology is 65 nm CMOS.

Acknowledgement

This project took place within the Department of Electrical and Information Technology (EIT) part of Faculty of Engineering, LTH (Lunds Tekniska Högskola) at Lund University.

I would like to take this opportunity to thank my supervisor Joachim Rodrigues who supported me in all steps during this project. He guided and encouraged me and provided a perfect working environment. His valuable comments and advices improved my report.

Special thanks to my advisers Pascal Meinerzhagen, Yasser Sherazi and Oskar Andersson who were always available for help with countless problems that I faced. Pascal's exact and detailed clarifications were really useful and helped me a lot.

Also my thanks and appreciations go to Mattias Andersson who shared his valuable experience and knowledge with me so many times. Without his help I would not be able to do certain parts of project.

I should also thank Stefan Molund, Reza Meraji, Carl Bryant and Jonas Lindstrand for their technical support and advice.

Babak Mohammadi
Lund, Jan, 2012

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Today's modern devices combine multiple tasks of yesterday's super computers in a better, faster and more efficient way. Increasing functionality in smaller hand-held devices demand for batteries with higher power and longer service time. For example, in today's mobile-phones, the energy consumption varies between 10 mW(jpeg encoding in a cell phone) to 10 W (Peak power in a mobile device). But improvements in battery technology is slow and the gap between the increasing energy demands and available energy storages is getting bigger and bigger. Thus It puts additional loads on circuit designers' shoulders to optimize and employ low energy methods more intensively while adding more and newer functionalities to their systems. To have an energy efficient design, these methods should be applied in all design levels from physical transistor implementation up to energy reduction algorithms and techniques used in software and operating systems.

In SoC[1] design, moving to newer technologies has always been an efficient option to reduce energy consumption, shrink design size and add more functionality on smaller chip area. But in sub-micron devices, it was demonstrated that static power turns to be an intolerable issue. A study in [1] shows that in a 15 mm die, when the total transistor width on the die increases by 50 %, the total leakage current increases by 7.5 X, which results 5 X increase in leakage power. Since active power remains constant (per scale theory) for constant die size, the leakage power will dominate. Also the ITRS[2] made following predictions [2]:

[1] System on Chip
[2] International Technology Roadmap for Silicon

Table 1.1: Power comparison between generations.

Power per cm^2	90nm	60nm	45nm
Dynamic	1X	1.4X	2X
Static	1X	2.5X	6.5X
Total	1X	2X	4X

As it can be seen in Table 1.1, the leakage power is the main source of power consumption in deep sub-micron technologies. To reduce this unacceptable range of power consumption and make it more tolerable, different techniques have been developed and illustrated (chapter 3).

In many digital circuits, memories are the main area and power consumers. Their leakage can be dominant static energy dissipation source and switching their capacitive lines would cost lots of dynamic energy [3]. Thus By decreasing energy dissipation in memory cells, a significant reduction in system's total energy consumption could be achieved.

For a given activity factor and frequency, it is possible to find an optimum supply voltage (V_{DD}) and threshold voltage (V_{TH}) pair which gives minimum energy consumption. To do this, a good work is done by A. Wang and A.P. Chandrakasan [5] [6]. By using BSIM3 model for Ring-Oscillator composed of cascaded NAND gates, they drew constant performance and constant energy contours for different activity factors. It is shown that for clock rates less than 10 MHz, minimum energy is achieved in sub-threshold region. The drawback with sub-threshold design when comparing with strong-inversion region is its long delays and lower performance since in sub-threshold regime the ON-current is reduced by orders of magnitudes compared to strong-inversion region and as a consequence it takes much longer time to charge and discharge node capacitances when switching. Because of weak performance, sub-threshold region normally is used in applications that timing is not critical and having ultra-low-power is the main concern. To improve timing properties threshold voltage could be reduced, but it will cost with increased leakage and static energy, since as threshold voltage decreases, leakage current increases exponentially [4].

In this master thesis project, different latch architectures and leakage reduction techniques in sub-threshold region were studied. A detailed comparison over selected latch topologies is performed and the best latch which fulfils the project constraints is selected. Further customization on latch to minimize its leakage current is done, while keeping an eye on performance to assure that the minimum speed requirement of project is submitted. The layout for final topology is designed, in a way that it is completely compatible with SCL (Standard Cell Library). In this project, the optimal point for V_{DD}/V_{TH} is not studied, since the supply voltage was

set by other projects to be around 300 mV. All the simulation results in this report are for V_{DD}=300 mV if is not specified. The dynamic energy reduction methods are not covered in this thesis as the activity factor is assumed to be very low. Default transistor option in this project, if not mentioned, is LP HVT (Low power, High V_T family.

Chapter 2

Leakage sources

In digital circuits, MOS transistors are used as switches which their ON/OFF states are controlled by the Gate-Source voltage (V_{GS}). In transistor's ideal model, when $V_{GS} = 0$, the current is supposed to be blocked completely as the switch is OFF, but in practice there is always a current flowing through OFF transistors which is called *leakage current*. Lots of parameters and effects contribute in final leakage current, like threshold voltage, channel physical length and effective dimensions, channel/surface doping profile, drain/source junction depth, gate oxide thickness, V_{DD} and temperature [1].

Figure 2.1 shows the different leakage paths in a deep sub-micron NMOS transistor.

- I_1 **(Junction Reverse Bias Current):** This current is a minimal contributor to total transistor I_{OFF}, it gets even less important in sub-threshold region due to lower voltage scale. It has two main components: Minority carrier diffusion/drift near the edge of the depletion region and electron-hole pair generation in the depletion region of the reverse bias junction [7].

- I_2 **(Weak Inversion):** This current occurs between source and drain when the gate voltage is below V_T. The carriers move by diffusion along the surface like the charge transport across the base of bipolar transistors. This current dominates OFF-state leakage in modern devices due to their low V_T [8] [9].

- I_3 **(Drain Induced Barrier Lowering):** DIBL occurs when a high voltage is applied to the drain where the depletion region of the drain interacts with the source near the channel surface to lower the source potential barrier. The source then injects carriers into the channel surface without the gate playing a role [8] [9].

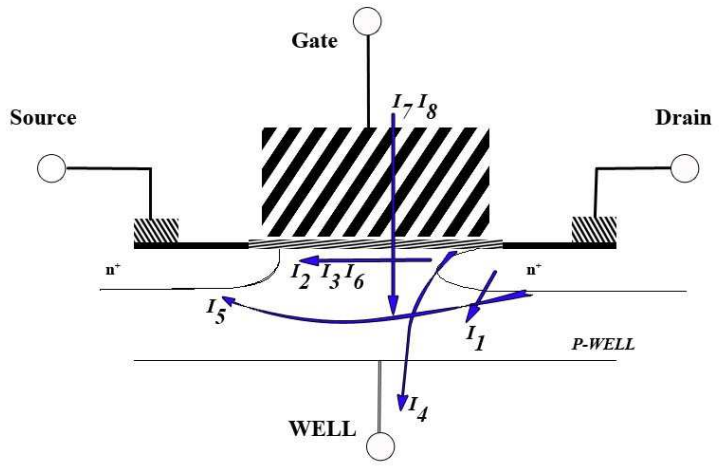


Figure 2.1: Leakage current mechanism in deep sub-micron transistors

- I_4**(Gate-Induced Drain Leakage):** GIDL currents are due to tunneling of electrons from the valance to conduction band in the transition zone of the drain-substrate junction below the gate-to-drain overlap region where a high electric field exists [8] [9].
- I_5 **(Bulk Punch-through):** This current flows from source to the drain due to lateral bipolar transistor formed by the source(emitter), the bulk(base), and the drain(collector) [10]
- I_6 **(Narrow Width Effect):** Transistors V_T in non-trench isolated technologies increases for geometric gate widths on the order of $\leq 0.5\mu m$. An opposite and more complex effect is seen for trench isolated technologies that show decrease in V_T for effective channel widths on the order of $W \leq 0.5\mu m$ [11].
- I_7 **(Gate oxide tunnelling):** This is the current across the thin gate oxide between the gate and the substrate I_G [12] [13], due to high electric field in the gate oxide. The responsible mechanism in nano-metric devices is direct tunneling through the oxide bands.
- I_8 **(Hot carrier injection):** This current is the result of injection of hot carriers (holes and electrons) in the oxide. Short-channel devices are more susceptible to this kind of current. This current increases as L_{eff} (effective length) decrease unless V_{DD} is scaled accordingly [1].

Except for sub-threshold leakage current (I_2), other leakage currents through specified paths in Figure 2.1 can be quite significant in deep sub-micron devices in

moderate and strong inversion regimes. But because of voltage scaling in sub-threshold region, they tend to be negligible [6]. Except for rare cases, sub-threshold leakage current dominates in weak inversion operation region. The simulation results approves this for 65 nm technology. It was observed that the gate leakage current components were negligible when comparing with total leakage current in sub-threshold ($\ll 10^{-5}$).
Summarized sub-threshold leakage current including weak inversion and DIBL effect is specified as following:

$$I_{subth} = A \times e^{\frac{V_{GS}-V_{TH0}-\gamma\prime V_{SB}+\eta V_{DS}}{n v_T}} \times \left(1 - e^{\frac{-V_{DS}}{v_T}}\right), \tag{2.1}$$

where

$$A = \mu_0 C_{OX} \frac{W}{L_{EFF}} (v_T)^2 e^{1.8} e^{\frac{(-\Delta V_{TH})}{\eta v_T}}, \tag{2.2}$$

V_{TH0} is the zero bias threshold voltage, $v_T = kT/q$ is the thermal voltage. The body effect for small values of source to bulk voltages is very nearly linear and is represented by the term $\gamma\prime V_{SB}$, where $\gamma\prime$ is the linearized body effect coefficient. η is the DIBL coefficient, $C_{ox}$is the gate oxide capacitance, μ_0 is the zero bias mobility and n is the sub-threshold swing coefficient for the transistor. V_{GS}, V_{DS}, and V_{SB} correspond to transistor's gate-source, drain-source and source-bulk voltages respectively.

Chapter 3

Reduction of OFF-State Current

3.1 Introduction

There are different methods and mechanisms developed for reducing the leakage current ranging from changing transistor's physical properties up to hybrid mechanisms which control system.
Employing circuits in sub-threshold operation region is a leakage reduction mechanism by itself, but by combining other general leakage reduction mechanisms used in moderate and strong inversion regions, as we will see later in this chapter, leakage in sub-threshold operation region is reduced even further.
The obvious method for leakage reduction is to remove leakage paths where possible. In the case of latches, there is at least one buffer which charges the storage node inside the cell in order to keep the hold-data's state, improving the transition properties and reducing read/write errors. So there will be a minimum of two V_{DD} to GND leakage paths which should be treated as the main leakage sources. In this chapter leakage reduction methods from these paths will be discussed.

Studying leakage reduction techniques will be useful in topology selection phase by choosing architectures that include these reduction techniques by default.

3.2 Leakage Reduction Techniques

There are several approaches for minimizing leakage which are normally used in moderate and strong inversion region:

- **Multi-V_{TH}**:
  Using Multi-V_{TH} transistor in design is one solution, where HVT (High V_T)

transistors are used wherever design goals allow, and lower threshold voltage (V_{TH}) transistors are used where it is necessary to meet timing constraints.

- **Power Gating**:
  A second approach is to shut down the power supply of a logic block when it is not active. This approach is useful for bigger and complex designs where it is possible to divide the circuit to active and inactive blocks and take care of the power of each block with an additional logic circuit.

- **Variable Threshold CMOS (VTCMOS)**:
  Another effective method which can reduce leakage by up to three orders of magnitude. In this method a reverse bias is applied to the substrate which increases V_{TH}. However VTCMOS adds complexity to the library and required additional power network. Also effectiveness of this method has been shown to be decreasing with scaling technology [2].

- **Stack effect**:
  The stack effect, or self-reverse bias, can reduce the sub-threshold leakage when more than one transistor in the stack is turned off.

- **Long Channel Devices**: From the equation of sub-threshold current (2.1), it is clear that there is reverse relation between the effective length of transistor and sub-threshold current when device is off ($V_G = 0$).

3.3 Methods Used In This Project

Except for VTCMOS and power-gating, other mentioned methods are used directly or indirectly in this project. *Stack-effect* and *Long channel devices* are discussed more in detail.

3.3.1 Stacking

The *Stack Effect* refers to the leakage reduction effect in a transistor stack when more than one transistor is turned off. Stacking effect is known as a very effective technique for leakage reduction that can reduce the leakage current at least an order of agnitude [14]. Figure 3.1 shows stack-effect with 2 transistors.
When the Gate voltages (V_G) of stacked transistors in Figure 3.1 are zero, a small sub-threshold current will follow which :

- will keep V_X positive, causing V_{GS1} to be negative. According to relation 2.1, a negative V_{GS} will reduce leakage.

- will also reduce V_{DS1} resulting minimized DIBL effect across it, which will further reduce the leakage.

- makes (V_{BS1}) negative, causing the threshold voltage to increase and ultimately reducing the sub-threshold leakage even further [14].

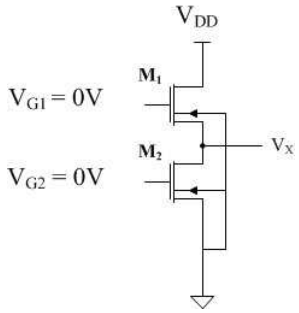


Figure 3.1: Stacking of 2 NMOS transistors in OFF-State

Again consider 2 stacked transistors in Figure 3.1 when both of them are in OFF-mode. In this mode, an OFF-state current will follow through both transistors which is almost equal in both devices (neglecting other current components). To find this current, first V_X should be calculated by equating (2.1) for both transistors by putting corresponding voltages and relating V_X. After doing simplifications V_X will be [15]:

$$V_X = \frac{v_T}{(1+2\eta+\gamma\prime)} ln(\frac{A_1}{A_2} e^{\frac{\eta V_{DD}}{v_T}} + 1), \tag{3.1}$$

V_X is equal to V_{DS2} and $V_{DD} - V_{DS1}$, so by placing (3.1) in (2.1) for M_1 or M_2, leakage current could be found. The off current for two stacked transistors in 3.1 will be:

$$I_{OFF-stack} = A_2 \times e^{\frac{-V_{TH0}+\eta V_X}{n v_T}} \times \left(1 - e^{\frac{-V_X}{v_T}}\right), \tag{3.2}$$

In (3.2), V_X is a very small voltage in sub-threshold region (in the range of 20-80 mV). If we compare it with 3.3 we can see that stacked effect is relatively big.

$$I_{OFF-single} = A \times e^{\frac{-V_{TH0}+\eta V_{DD}}{n v_T}} \times \left(1 - e^{\frac{-V_{DD}}{v_T}}\right) \approx A \times e^{\frac{-V_{TH0}+\eta V_{DD}}{n v_T}}, \tag{3.3}$$

The leakage reduction achievable with stacking can be calculated using *"stack effect factor"* which is equal to $\frac{I_{OFF-single}}{I_{OFF-stack}}$. Unfortunately *stack effect factor* becomes a complicated and difficult to analyze relation in sub-threshold region. Simulations done for sub-threshold region show that the assumption of $V_X > 3kT/q$ used in [18] for calculating a general expression for *stacking factor effect* is not valid for 65 nm

technology in sub-threshold region and interpolation in [17] has a big error in sub-threshold region with this technology.

To see the *stack effect* in sub-threshold region different simulations were done. As Figure 3.2 shows in sub-threshold regime having two stacked devices reduces leakage almost by 50%, but stacking effect diminishes for orders higher than two transistors in stack. The leakage current after passing two transistors decreases significantly and becomes comparable with substrate leakage components.

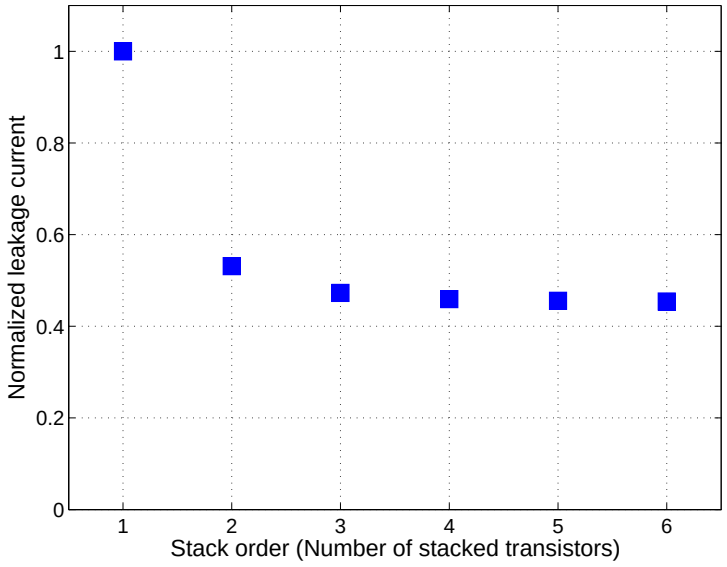


Figure 3.2: Leakage current of stacked transistors normalized to single device's leakage current

Different leakage current components flowing through substrate path, have a weak dependency on physical design parameters. To limit substrate current, a combination of *stack effect* and *floating-body-biasing* is used. In this mode, these current components have to pass a series of drain-source resistances of preceding/succeeding transistors in parallel with a reverse-biased diode (in nMOS stacked transistors) (Figure 3.3) instead of a short circuit to V_{DD} or GND. Figure 3.4 compares stacking of mentioned two body-biasing methods. As it can be seen, floating-body has much less leakage compared to normal biasing and by using 6 transistors in the stack, leakage reduces to almost 10 % of a single-transistor leakage current.

Despite of having relatively less leakage current, floating-body-biasing has two ma-

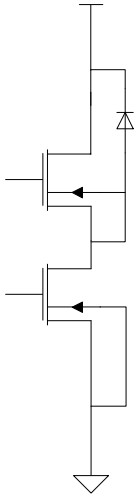


Figure 3.3: Stacked NMOS transistors with reverse diode which is formed as a result of having triple N-Well layer that isolates the body.

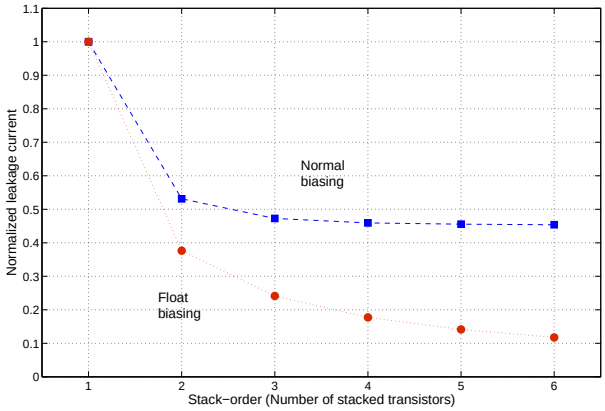


Figure 3.4: Leakage current of stacked transistors in *floating-body-biasing* and *normal-body-biasing* normalized to single device's leakage current

.

jor drawbacks. First and important one is the huge area overhead due to large n/p guard-rings and second one is parasitic bipolar transistors which would cause latch-up problems. The area overhead problem will be compared in chapter 6. Because of this problem, further analysis about the latch-up and possible solutions are not investigated in this project.

3.3.2 Stack-Effect vs. Channel Length

Another method for OFF-state leakage reduction is using longer channels. As it can be seen from relations 2.1 and 2.2, effective channel length has a reverse relation with the leakage current reduction. By increasing the channel length, threshold voltage decreases (100% increase in length, results in 7.5% V_{TH} reduction) which in turn increases leakage, but as Figure 3.5 shows, overall leakage current decreases dramatically up to 150 nm and then saturates. The main reason for threshold voltage reduction due channel-length increasing in short channel devices is non-uniform lateral doping (Halo implant).

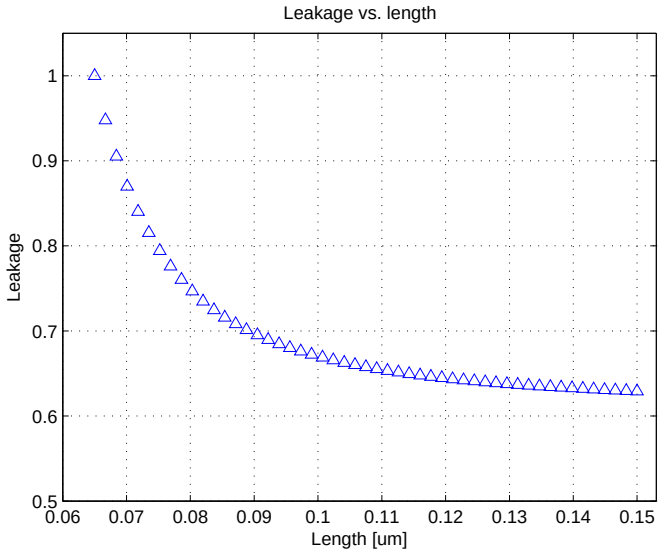


Figure 3.5: Leakage current vs. gate length normalized to minimum gate length's leakage.

3.3.3 Combining Long-Channel Transistors and Stack-Effect

As we saw previously, both methods reduce the leakage currents to lower levels and then saturate (*Floating-body-biasing* is not the case). We can push this limit even further down by combining both techniques. It should be noticed that by doing this, both ON and OFF mode currents will decrease which means slower performance.

In normal body-biasing, stacking more than 2 devices does not pay back area and

performance cost with leakage reduction, so we will concentrate on stacking order of up to two. Figure 3.6 shows the normalized current of two stacked transistors whose gate length is increased from 65 nm to 140 nm. Using this method, another 10% of leakage current will be reduced. As it can be seen, after 120 nm, leakage reduction almost saturates and there is no benefit in using longer devices in stack chain.

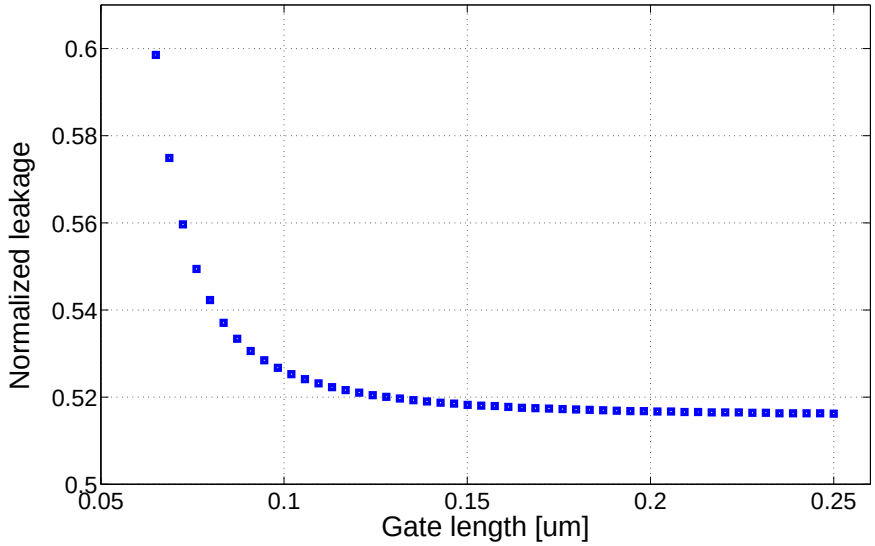


Figure 3.6: OFF-state Leakage current of two stacked transistors normalized to the leakage of single minimum sized off transistor. Length of both transistors in the stack are swept.

In the case of having multiple transistors with different gate lengths in the stack, placing the longer transistors at the nodes closer to V_{DD} and shorter transistors closer to GND will result in a lower leakage.

As an example in Figure 3.7, case A will have a lower leakage than case B. This is shown in Figure 3.8 where in curve B, gate length of M_1 (in Figure 3.7) is 65 nm and the gate length of M_2 is swept. Curve C shows the case where M_1's gate length is swept and M_2 has minimum gate length. As it can be seen, the curve C has lower leakage. Also as curve D in Figure 3.7 suggests, increasing the gate length of all stacked transistors has the minimum leakage.

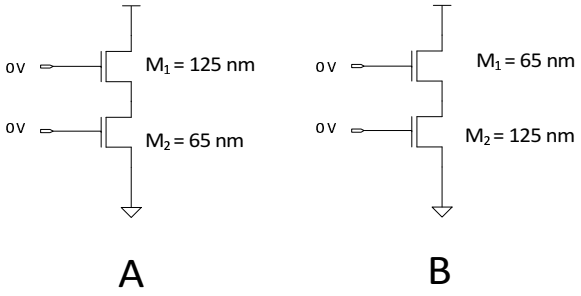


Figure 3.7: Having transistors with different lengths in the stack:
A) Longer transistor close to V_{DD}, B) shorter transistor close to V_{DD}.

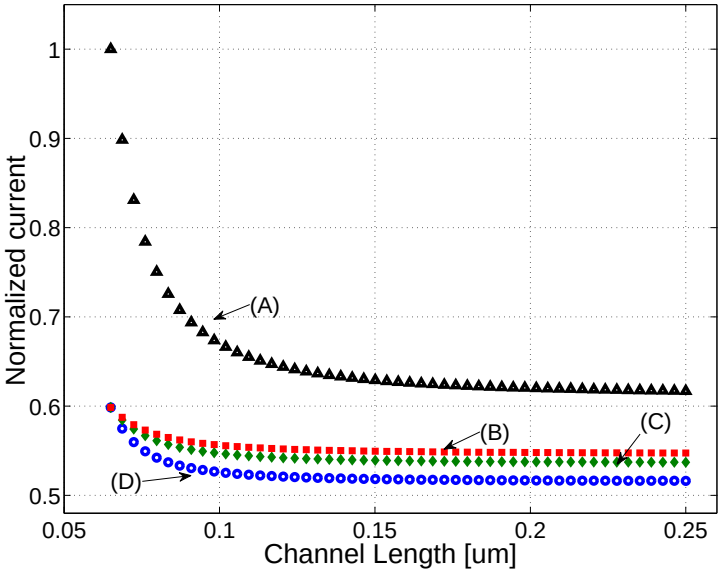


Figure 3.8: OFF-state Leakage current normalized to maximum leakage current of single transistor:
A) Single transistor, L=swept,
B) 2 stacked transistors, L_1(in Figure 3.1)=65 nm, L_2=swept,
C) 2 stacked transistors, L_1=swept, L_2=65 nm,
D) 2 stacked transistors, L_1=swept, L_2=swept

3.4 Summary

In this chapter different leakage reduction techniques in $V_{DD} - GND$ paths were studied. *Stack-effect* and the effect of increasing the gate length were analyzed and it was shown that combining these mechanisms gives better results. Increasing the devices' gate length in the stack chain up to 120 nm has a sharp leakage reduction ratio, but after 120 nm it almost saturates. Stacking more than 2 devices is not efficient, since consumes extra chip area and degrades the performance. Also having *floating body-biased* transistors in stack reduces both substrate and drain-source leakage currents. $Multi - V_{TH}$ and power gating are discussed in next chapters.

Chapter 4

Comparative Analysis of Latch Topologies

After knowing leakage sources and leakage reduction mechanisms in previous chapters, following 3 points were considered for topology selection:

- Functional in sub-threshold region
- Minimum number of $V_{DD} - GND$ paths
- Maximum number of stacked transistors in $V_{DD} - GND$ paths

After making initial selection according to mentioned 3 points, final selection is done by considering leakage current and other secondary factors like performance, layout size and reliability.

In the first stage, the functionality of different topologies was analyzed. All the topologies were assessed under same conditions. They were all designed with the same class of transistors with same size, power class and threshold voltage and same test conditions like supply voltage, timing properties and etc. After this phase, topologies that were not functional in sub-threshold operation region or had poor performance were filtered out.

4.1 Topologies

Figure 4.1 contains the outcome of first evaluation phase and shows architectures which were functional at sub-threshold operation region and had relatively lower average leakage. Selected architectures are simplified and their input/output buffers are removed where possible. The reason for this was to have a functional latch with

minimum number of leakage paths, since most of the architectures contain buffers at their inputs/outputs which adds additional leakage paths and this can cause wrong leakage assessment when comparing with architectures that do not have any buffer at their input/output.

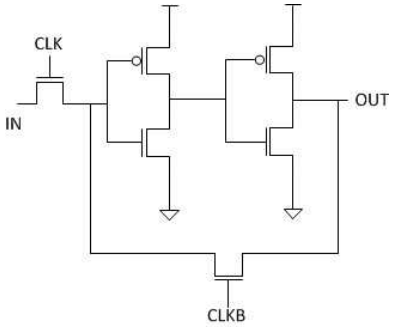
4.2 Timing Properties

There are different timing specifications for a latch, but since the latches in this project will not be used in critical paths, we leave advanced timing analyses which are used to determine the propagation delay of cascaded latches. Figure 4.2 shows main four timing properties of a latch. According to the simulations, hold-times of selected architectures are negligible compared to their setup times in sub-threshold region. The hold-time for all architectures were below 20 ns.
Table 4.1 lists the simulation results for all topologies listed in Figure 4.2. It should be noticed that these values are just for comparing topologies at the same conditions and do not represent the real and practical timing values. Different output controlling mechanisms can have different loads and therefore the presented values in Table 4.1 could be different, since the store-node is not completely isolated from output at this stage. To have more realistic values outputs of all latches are loaded with an inverter made from 2 pairs of stacked nMOS and pMOS transistors. Simplified test-bench is shown in Figure 4.3.
As Table 4.1 and Figure 4.4 show, the setup times of *PassTr* are completely asymmetric and writing high values takes much longer time than writing low values. The reason for this issue is the voltage drop over pass-transistor at the input when passing high values causes. If voltage drop on pass-transistor is ΔV and the value at the output of a symmetric latch flips at V_M, the switching threshold in this topology will be $V_M + \Delta V$. Replacing nMOS transistor with pMOS will cause a delayed swing from high to low, since pMOS transistors have a voltage drop when passing low values. Replacing pass-transistor with low-V_{TH} transistor will improve this effect and replacing it with transmission-gate will fix this problem.

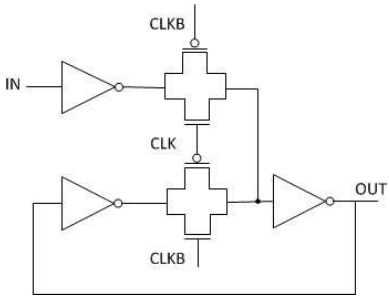
Table 4.1: Setup time for data-high and data-low for minimum size transistors(measured at output)

Topology	**Low (A)**[μs]	**High (C)**[μs]
PassTr	0.522	6.892
MUX	0.645	0.827
nRam	1.944	3.859
D-Latch-3	1.373	0.819
D-Latch-Sw	0.089	0.192
SRIS	2.017	0.918

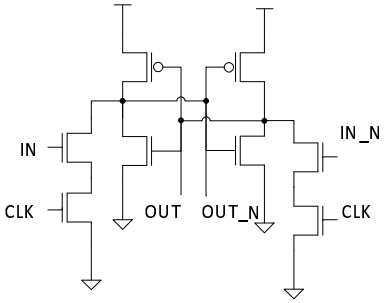
Figure 4.4 shows a complete swing of all latches. As the values in Table 4.1



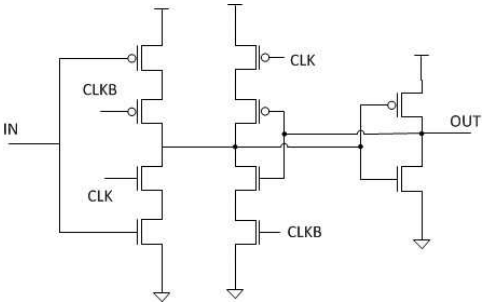
(a) Multiplexer based latch using pass-transistors /PassTr



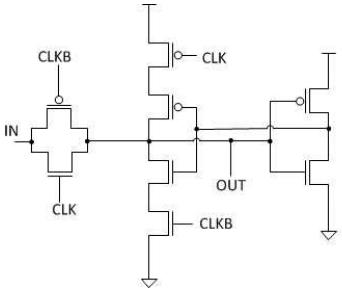
(b) Multiplexer based latch using transmission gates/MUX



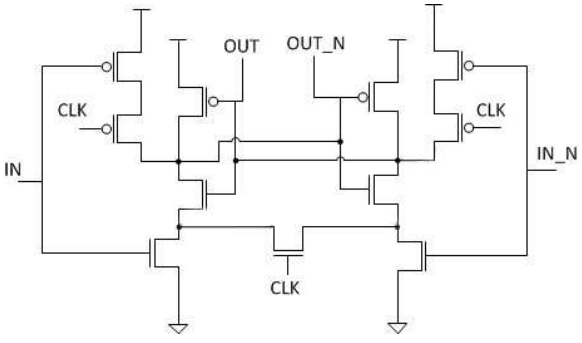
(c) RAM type n-latch/nRam



(d) D-latch using tri-state/D-Latch-3



(e) D-latch using tri-state and transmission-gate/D-Latch-Sw



(f) Static ratio-insensitive (SRIS) p-latch/SRIS

Figure 4.1: Schematic of topologies which passed functionality test phase.

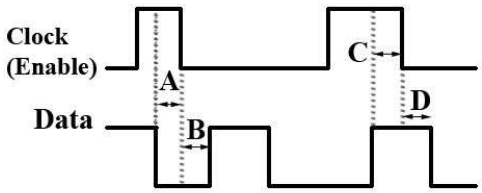


Figure 4.2: Timing specification of an active-high latch. A = Data-low setup time, B= data-low hold time, C = data-high setup time, D = data-high hold time.

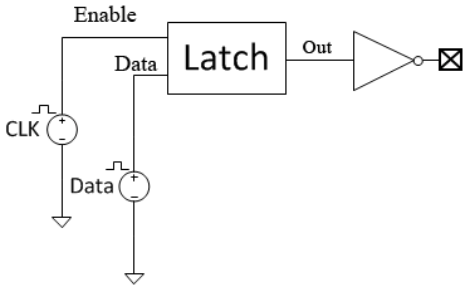


Figure 4.3: Test-bench used for timing analysis.

and plots in Figure 4.4 suggest, *D-Latch-Sw, MUX and D-Latch-3* show the best performances.

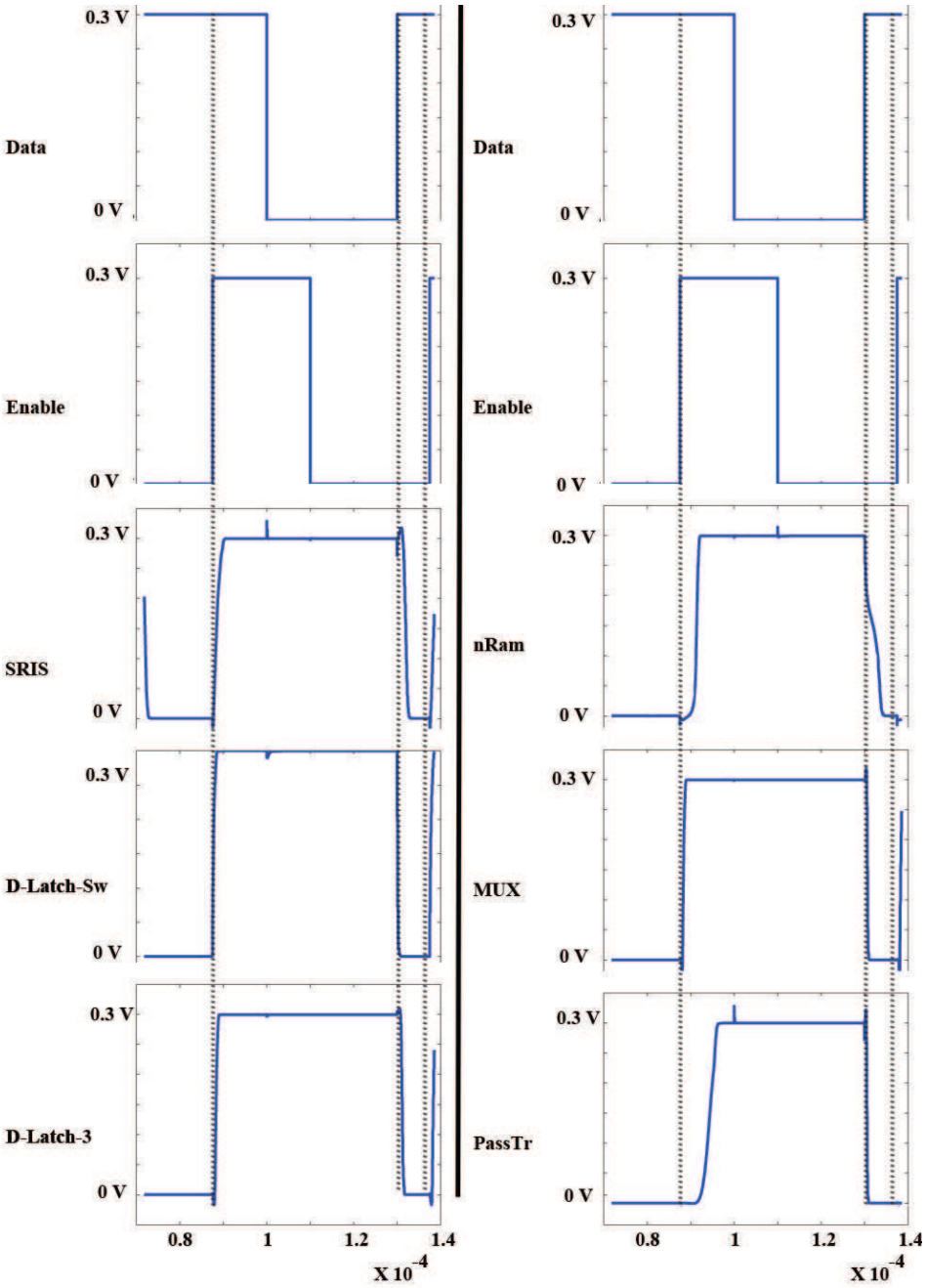


Figure 4.4: Performance assessment of 6 selected topologies from phase 1.

4.3 Leakage Current Comparison

To find the leakage current of each topology, 4 data/hold states should be considered which are the combination of input (low/high) and hold-value (low/high). To have the appropriate hold-data combination, a transient simulation is performed. At first, the desired hold-data is written to the latch and locked and then at the second stage, the leakage current can be measured after a relatively long wait-time to have all leakage currents in their steady state. This value is assumed as the leakage current of that specific state. The average leakage should be calculated for all 4 cases, since as we will see, some topologies have a high dependency to input/hold data combination.

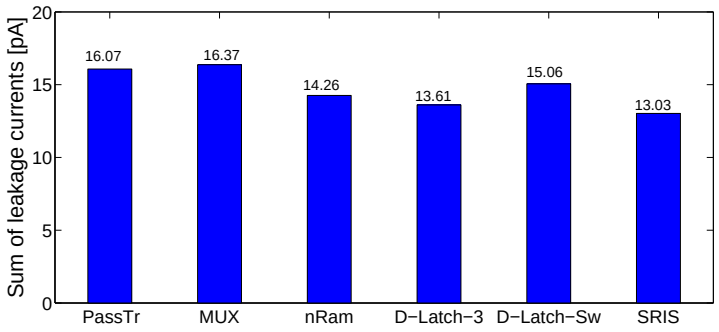


Figure 4.5: Sum of leakage current for all data/hold-value combinations.

Table 4.2: Average leakage currents of different topologies.

Topology	**Average leakage current [pA]**
PassTr	4.02
MUX	4.1
nRam	3.57
D-Latch-3	3.40
D-Latch-Sw	3.76
SRIS	3.26

As we can see in Figure 4.5 and Table 4.2, *SRIS* and *D-Latch-3* have the least total leakage among all selected topologies. But by checking the structures of *SRIS* and *nRam* it can be seen that these topologies need both data and inverted data inputs

and therefore an inverter should be used. Adding another inverter will increase the leakage and totally the leakage of these topologies will be more than *D-Latch-3*. Two mentioned topologies have an advantage that they do not need inverted *Enable* input, but these inputs could be shared between all bits in the word (Figure 4.7). For example if we have a memory word with 8 bits, an inverter could be used to feed all 8 bits and by doing this the leakage of this single inverter will be divided between 8 bits which makes it negligible when comparing with single latch's leakage. On the other hand the layout area of *SRIS* will be bigger than *D-Latch-3*, since:

- It has one extra transistor compared to *D-Latch-3*. With counting the transistors for inverted-data, it will be at least 3.
- The number of n/pMOS transistors is not equal and in layout the p/n network won't be symmetric and the area of a transistor will be wasted.

Topologies that contain transmission-gates or pass-transistors, have a leakage path between their outputs and inputs in some data combinations. This leakage path is shown by longer shadowed-line in Figure 4.6 and is a result of impedance of pass-transistors/transmission gates in OFF-mode which is shown in Table 4.3. This leakage current is sourced by the buffer that feeds inputs of bits in memory column (see Figure 4.7) which can be relatively large current when accounting the leakage current of other bits connected to data-line. Because of small I_{ON}/I_{OFF} ratio in sub-threshold region more buffers will be required to dominate leakage current at data-line while writing new data which means higher leakage and area. These buffers can be added internally to the input of each bit which will solve data-line buffer loading, but will add a $V_{DD} - GND$ leakage path per bit.

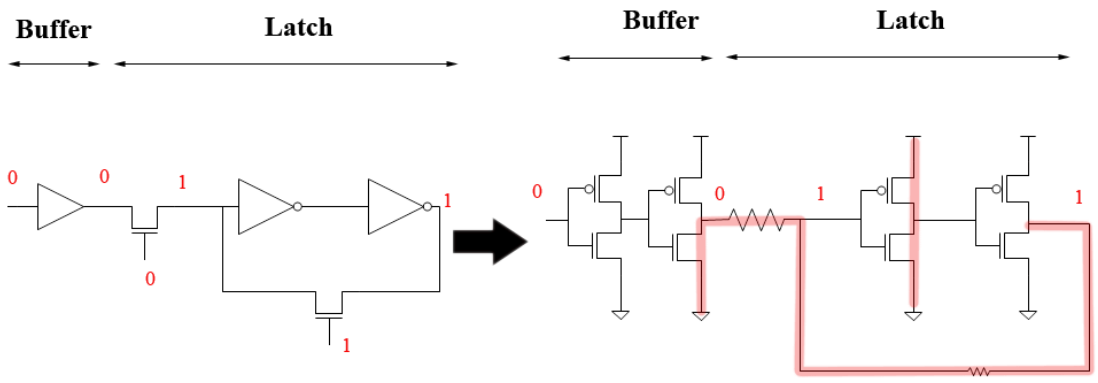


Figure 4.6: Main leakage paths in *PassTr* topology. Other topologies containing transmission-gates have similar situation.

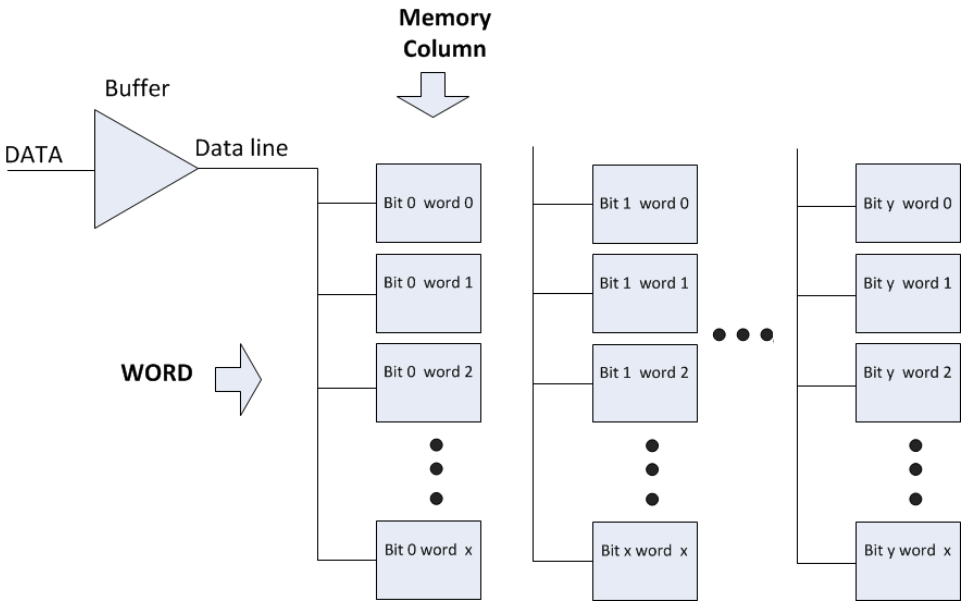


Figure 4.7: Buffer and data-line in a memory.

As mentioned in timing section, since pass-transistors cannot conduct high or low values very well, both transistors in the first internal inverter may be turned on simultaneously in the two states - depending on p or n-type pass-transistor - and therefore a high short-circuit or cross-over current will flow through this path which is shown by shorter shadowed-line in Figure 4.6. This effect can be improved by using low-threshold transistors which their voltage drop is lower, but it has other side effects and causes the leakage current through the before mentioned leakage path through pass-transistor/transmission-gate to increase significantly(refer to Table 4.3).

Table 4.3: ON and OFF mode resistance of different pass-transistors/transmission-gates and their OFF-mode leakage current.

	ON $\times 10^6\Omega$	OFF $\times 10^9\Omega$	OFF-current [pA]
trans.-gate/HVT	97.56	187.85	1.597
trans.-gate/SVT	5.43	16.21	18.51
trans.-gate/LVT	1.41	4.49	66.75
nMOS pass-tr./HVT	103.59	234.01	1.282
nMOS pass-tr./SVT	6.39	16.26	18.45
nMOS pass-tr./LVT	1.40	3.81	78.82
pMOS pass-tr./HVT	340.02	457.25	0.656
pMOS pass-tr./SVT	19.93	82.12	3.653
pMOS pass-tr./LVT	10.69	53.66	5.591

By comparing the OFF-current in Table 4.3 and the average leakage of different topologies in Table 4.2 it can be seen that the leakage of a SVT/LVT pass-transistors/transmission-gates are orders of magnitude higher than latch itself, which will result in having a higher average total leakage. So using low V_{TH} transistors as pass-transistors or transmission-gates is not advised.

Last point to mention in this section is the data dependency of leakage currents in some topologies. As expected, topologies containing transmission-gates/pass-transistors have a relatively high leakage when the hold-value is different with data available at the input. Figure 4.6 shows this case when the hold-value is logic 1 and data at the input port is logic 0. Having data dependent leakage will cause a high dependency of total memory leakage to input data which is not desirable in most cases. Table 4.4 shows the data dependency of leakage in two cases. As it can be seen, topologies without transmission-gates/pass-transistors have a much less variation for different data inputs. For example, *D-Latch-3* has 16% variation, where the same structure with transmission-gate (*D-Latch-Sw*) has 30% variation

in leakage current in these two cases.

Table 4.4: Data dependency of leakage in selected topologies. First column shows the leakage for the case that the input data and hold-value are same(0), second column shows the case that the input data(1) is different with hold value(0).

Topology	Input=0, Hold=0 [pA]	Input=1, Hold=0 [pA]
PassTr	2.243	3.279
MUX	3.444	4.739
nRam	2.924	3.780
D-Latch-3	3.178	3.780
D-Latch-Sw	3.135	4.430
SRIS	3.017	3.496

4.4 Final Topology and Conclusion

So far a few aspects of advantages and disadvantages of selected topologies are discussed. Each topology has specific advantages and disadvantages that makes it to be the target architecture for different applications. For example applications with the chip area as the first constraint, *PassTr* could be a good option. This topology could be modified a little to meet a better timing specification.

Other topologies containing transmission-gates generally have a good timing specifications. But in addition to the leakage through data-line, they add high capacitance load (considering the capacitance of internal storage node) to the data-line which will degrade overall timing properties of memory and increase the dynamic power consumption. Their layout size could be classified in average range among other architectures.

Remaining topologies have quite close specifications in both performance and leakage and the inputs are well-isolated from storage nodes. But as mentioned earlier, *SRIS* and *nRam* need an extra inverter to provide the inverted data signal and this will result higher leakage and area overhead in these topologies.

Considering the leakage and second order constraints, *D-Latch-3* topology was selected as the target topology in this project, as it has low leakage and good performance. The only drawback with this topology is the number of transistors which is 10, but since most of the transistors are stacked and they don't need any contacts at the interconnection nodes, a dense layout can be designed. This will be

discussed in detail in layout chapter.

Table 4.5 summarizes the important parameters for selected latches.

Table 4.5: Summary

Architecture	Sum of leakages $[pA]$	Leakage via data-line $[pA]$	Min no. of transistors	$t_p[\mu s]^*$
Standard cell	31.79	**0**	16	1.25
PassTr	16.07	1.04	**6**	3.71
MUX	16.37	1.59	10	0.74
nRam	14.26	**0**	8+2	2.90
D-Latch-3	13.60	**0**	10	1.06
D-Latch-Sw	15.05	1.59	8	**0.14**
SRIS	**13.02**	**0**	11+2	1.47

* Average rise-fall time, $t_p = \left(\frac{risetime+falltime}{2}\right)$.

In next chapter we will try to customize *D-Latch-3* further to minimize the leakage even more.

Chapter 5

Customization

After selecting the final topology, some optimizations are done to reduce the leakage current even further which is the main constraint in this project. The leakage reduction techniques discussed in chapter 3 will be applied to the selected topology and results will be compared to see the effect.

5.1 Applying Leakage Reduction Mechanisms

One of the most effective leakage reduction mechanisms is stacking. *D-Latch-3* has three $V_{DD} - GND$ paths, one of which is off in hold mode. The second inverter in the latch is the main source of leakage in this topology and should be stacked with two more transistors to minimize the leakage. The tri-stated feedback inverter is the same as input tri-stated inverter in write-mode and changes to a normal inverter in hold mode. This path's leakage will be reduced by increasing the length of the transistors.

As discussed earlier in Chapter 3, combining longer channel with stack-effect will give a much better result. So we will change the *D-Latch-3* as shown in Figure 5.1. Also as mentioned, leakage reduction rate reduces after 125 nm and almost saturates. Because of this 125 nm is selected as the length of transistors in the latch.

By making these changes, the average leakage current of *D-Latch-3* will be 2.075 pA which means 39% reduction compared to original structure. To have a reference for comparing new version's results, *Standard cell* in library will be used. Table 5.1 compares some properties of *Standard cell* and *Customized cell*. As it can be seen, the leakage is almost 4 times higher than custom version of *D-Latch-3*. The area cost of customizations made in this chapter will be analyzed in Chapter 6.

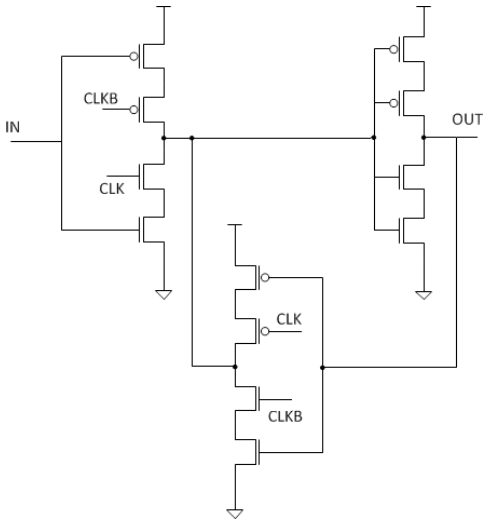


Figure 5.1: Customized *D-Latch-3*. The output inverter stacked with two extra transistors and the length of all transistors are increased to $125\mu m$.

Table 5.1: Performance and leakage comparison between *Standard-cell, D-Latch-3* and its custom version(length of transistors in custom version is 125 nm).
* Setup time

Topology	$(0to1)[\mu s]^*$	$(1to0)[\mu s]^*$	Average leakage [pA]
Standard cell	1.18	1.329	7.948
D-Latch-3/base	0.739	1.373	3.402
D-Latch-3/custom.	3.42	4.22	2.075

5.2 Output Buffer in D-Latch-3/Custom

The same as input, storage node should be isolated from output, since normally multiple bits are connected to the same output data-line and any change at this line can affect the data in storage node and reduce the latch's reliability. This can be done by inserting a pass-transistor, transmission gate, MUX or tri-state between storage node and output are some possible solutions.

- **Pass-transistor and transmission gate**
  The same as input, a latch with pass-transistor/transmission gate at the output will suffer from high leakage. Also there will be some performance issues,

since in this mode, the buffer which is used in storage node will be responsible for driving the high capacitive load on output data-line. As this buffer is optimized for leakage, will have a low I_{ON} current and this will cause a long charge/discharge time resulting a poor performance. Also backward-driving in these gates could be problematic and can cause reliability issues as these gates can drive current in both directions. Voltage drop and reduced noise margins should be considered when using these solutions as well.

- **MUX**
  Using multiplexers can be good solution for small memory sizes, but when it comes to higher memory sizes, multiplexers add huge area, leakage and delay overhead to the design.

- **Tri-state**
  Tri-states isolate storage node and output very well and do not have back-driving. Furthermore with good transistor sizing, they can act as good buffers which can drive the capacitive output-data line.

In this project, tri-state buffer is used because of the mentioned advantages. The input of tri-states should be moved to internal node, otherwise the final data value will be inverted. The final latch and output controller is shown in Figure 5.2

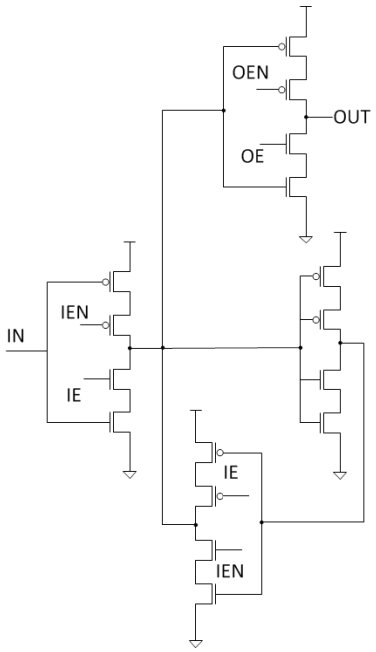


Figure 5.2: *D-Latch-3* with output controller.

As Figure 5.2 shows, there are roughly 6 gate-source and 4 drain-bulk capacitances which are connected to storage node in parallel. The total capacitance at this node will be relatively large and because of low I_{ON} current at sub-threshold region, the setup time will take longer time and performance will be degraded. This effect and solution is studied in next section.

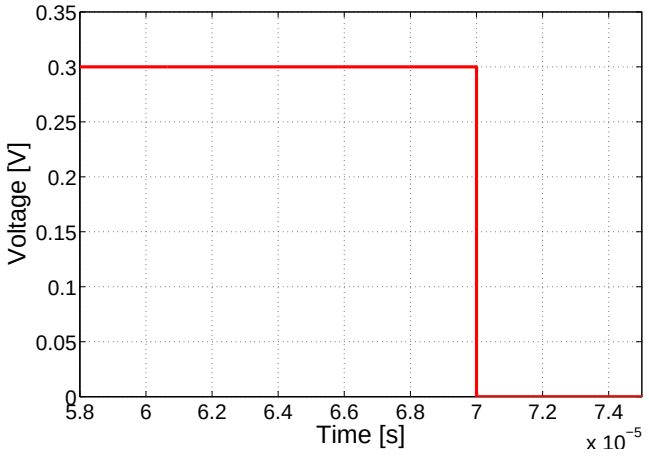
5.3 Improving Timing Specifications

As Figure 5.3 shows, the new version has a longer setup time both in low-to-high and high-to-low cases which was expected, but still the requirement of project to have a speed of a few hundred kHz is fulfilled. Table 5.2 contains timing information of customized *D-Latch-3* with connected tri-state at the output. As it can be seen, because of the weakness of pMOS transistors in driving ON-mode current compared to nMOS transistors, there is a big error between rise-time and fall-time of transitions. To solve this issue in strong inversion region, the width of pMOS transistors are increased to compensate current driving capability of pMOS transistors/network.

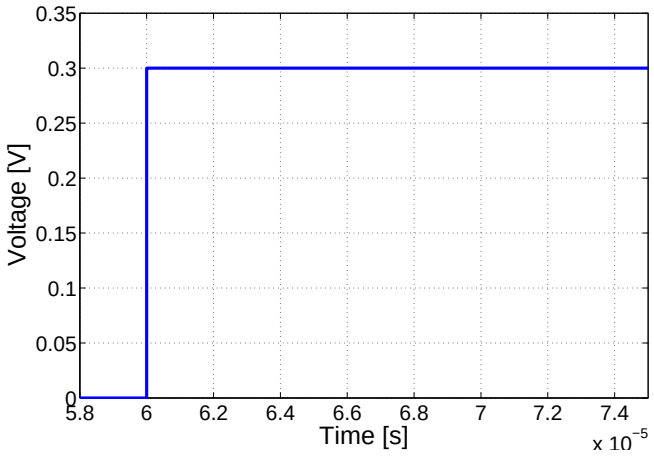
To validate this effect in sub-threshold region, a simple stack inverter is used. In rough estimation, last customized version is a combination of 4 stacked inverters, so if we can speed up the stacked inverter in Figure 5.4, final gate will speed up as well.

To do this, a parametric-analysis over the width of pMOS transistors from $0.135\,\mu m$ to $1.675\,\mu m$ with 40 points was performed. Figure 5.5(a) shows the transient pulse responses for mentioned 40 points. As it can be seen the variation of fall-times is more than variation of rise-times, as the width of nMOS transistors are constant and by increasing the width of pMOS transistors, the total capacitance connected to store node is increasing where the current driving capability of nMOS transistors is the same. This results in higher variation in falling edge. Figure 5.5(b) shows rise-time and fall-time vs. pMOS width. As 5.5(b) shows, increasing the width of pMOS transistors more than 0.6 μm does cannot decrease the rise-time and rise-time remains constant at $1.519\,\mu s$, but since capacitance at the test node increases, the fall-time continues to increase constantly and finally rise-time and fall-time become equal at $1.306\,\mu m$. Placing this value into our *customized cell* perfectly matches this result and rise-time and fall-time become equal. As it is obvious this value is not practical and efficient solution, since increasing the capacitance on inputs, outputs and internal nodes by a factor of 20-60 is not a wise idea.

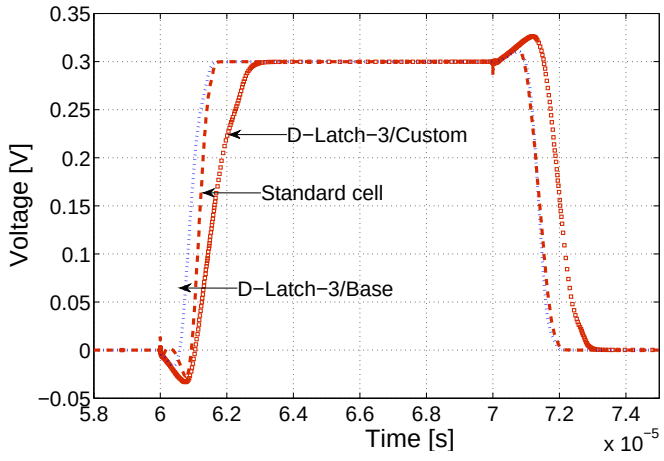
Another solution to have symmetrical rise/fall time is to increase P-network's driving by using shorter pMOS transistors which will cost increased leakage. Figure 5.6 shows the effect of length reduction of pMOS transistors on rise-time. As it can



(a) Data



(b) Clock(Enable)



(c) Outputs

Figure 5.3: Transient response simulation of custom and base versions of *D-Latch-3* and *Standard-cell* a) Data, b) Clock(Enable), c) Outputs of D-Latch-3 Base and Custom versions and Standard-cell.

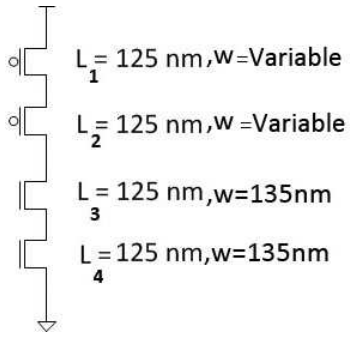


Figure 5.4: Stacked inverter for width analysis

be seen, when using minimum length stacked pMOS transistors, rise-time becomes almost equal to fall-time. The small error can be removed by increasing nMOS transistors' length or pMOS transistors' width.

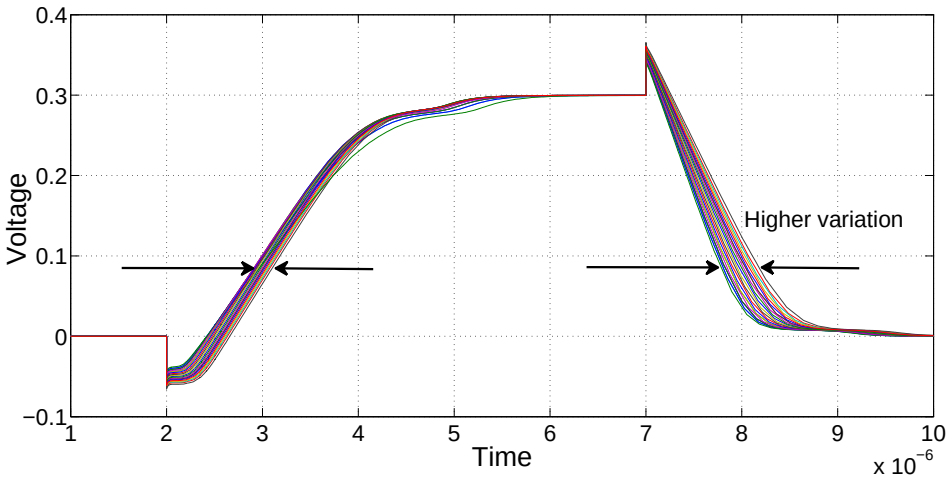
The results of simple stacked inverter is verified on custom version of *D-Latch-3* by changing the width of all pMOS transistors to 135 nm. The effect of length reduction in pMOS transistors can be seen in Table 5.2. The first column of table belongs to the case that the length of all transistors' is 125 nm and the second column belongs to the case that we had symmetrical input/outputs. As it can be seen, the result in final design matches as well and there is a great improvement in rise-time reduction.

Table 5.2: Timing specification of customized version

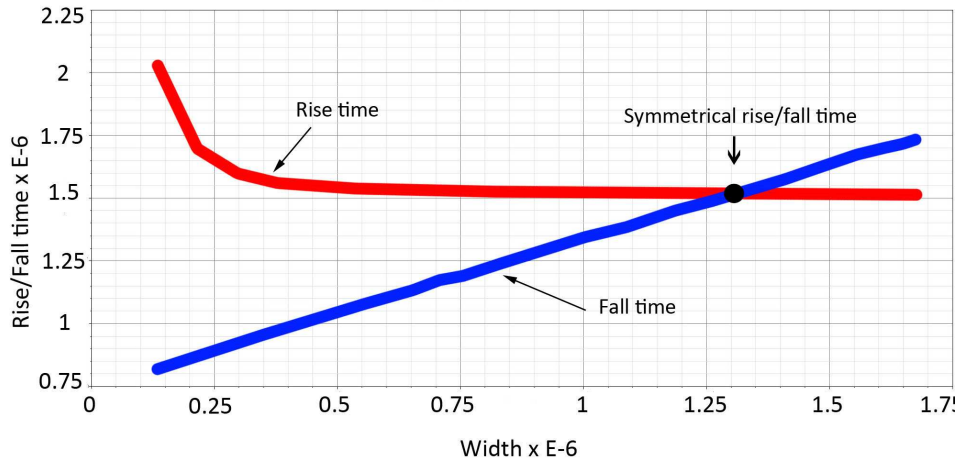
Topology	Time[μs] *	Time[μs] **
Rise time	3.57	1.86
Fall time	1.53	1.44
Setup time(low to high)	2.20	1.16
Setup time(high to low)	0.97	0.91

* Length of transistors: pMOS = 125 nm, nMOS = 125nm.
** Length of transistors: pMOS = 65 nm, nMOS = 125nm.

Both mentioned methods, increasing width and decreasing length of pMOS transistors increase the leakage current and since the leakage was the first constraint in this project, using these methods are avoided in this project. These results could be used as a reference for applications.



(a) Pulse responses of the inverter in Figure 5.4 with 40 different pMOS widths swept from 135 μmm to 1.6 μm



(b) Rise/Fall time vs. pMOS transistors' width shift

Figure 5.5: Fall/Rise time analysis with different pMOS widths for inverter in Figure 5.4.

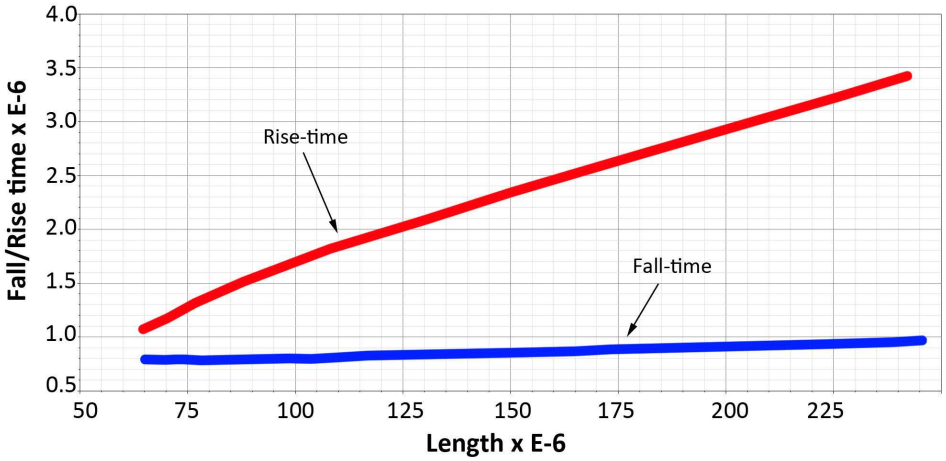


Figure 5.6: Fall/Rise vs. Length of pMOS transistors.

5.4 Sensitivity to Variation

Generally there are three sources of failure in SRAMs [16]: a) read-failures, b) write-failures and c) hold-failures.

In read and write modes, *D-Latch-3* can be considered as the generic latch represented in [16], consequently the same as generic latch, it will be immune from the same read and write failures. The remaining hold-failure is studied in the next section.

5.4.1 Hold-Failure Analysis

The hold-failure can be analyzed by using SNM (Static Noise Margin) estimation. SNM is defined as the minimum DC noise voltage needed to flip the cell state [19]. SNM is extracted as the largest square embedded in the butterfly curves (VTC curves of INV2 and INV3 which are achieved from Montecarlo simulations) of INV2 and INV3 which are measured from V_{in} and V_{out} nodes shown in Figure 5.7. These curves are based on 1000-point Monte Carlo simulations for die-to-die (process) and within-die (mismatch) variation at 25°C.

As Figure 5.8 suggests above 300 mV is reliable region for selected topology.

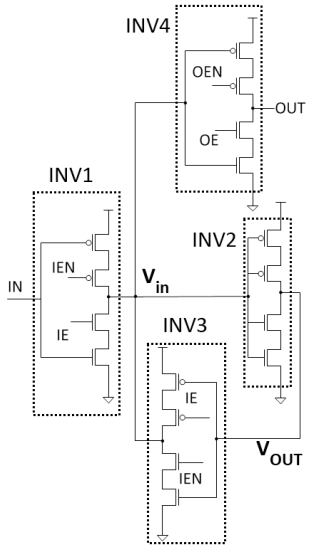
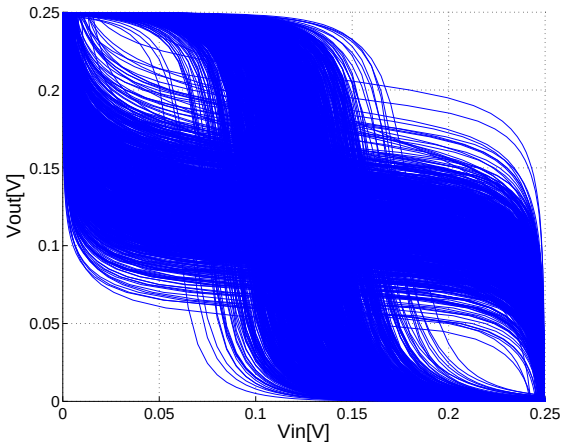
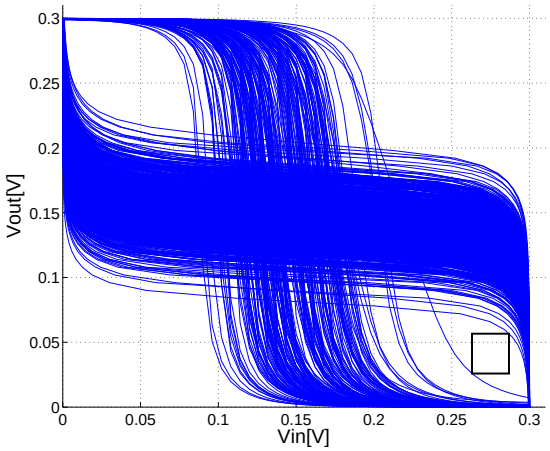


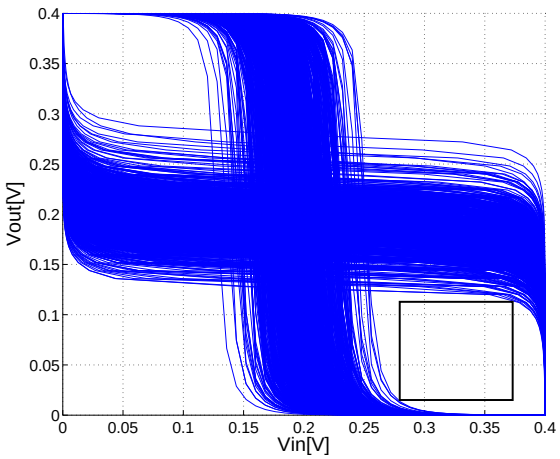
Figure 5.7: *D-Latch-3* with output controller.



(a)



(b)



(c)

Figure 5.8: Hold static noise margin (SNM) of *D-Latch-3* for (a) VDD = 250mV, (b) VDD = 300mV, and (c) VDD = 400mV.

5.4.2 Delay Variation

In addition to long delays, there is a significant variation of delay around the nominal delay in sub-threshold region. This can be problematic in time sensitive applications and add long wait margins, specially where multiple gates are cascaded. To see this variation in 65 nm technology, delays of a standard cell inverter is studied. This gives a general idea about the delay variation in 65 nm at different supply voltages. Depending on observation points in *D-Latch-3*, the results of simple inverter can be generalized to the latch. For example, in Figure 5.7, by choosing the storage node as the observation point for delay measurements, we have just INV1 between input and storage node which is a simple inverter in write-mode. Similarly, we have INV4 between storage mode and output which is a simple inverter in read mode.

The data for delay variation gathered from 1000 point Monte Carlo simulation at 1.2 V (nominal voltage) and 300 mV (sub-threshold voltage). The test-bench and transistor sizes are shown in Figure 5.9.

As it is shown in Figure 5.10a, delay variation around nominal voltage (1.2 V) has a Gaussian distribution and is minimal. The delay variation around mean value for this case is 5.3%. This value changes to 42% for 500 mV and 53.2% for 300 mV which is a huge variation. As it can be seen in Figure 5.10c and 5.10b, delay distribution is not Gaussian in these cases and delays below average are more likely to happen, however above-average delays have wider range and can be several times longer than mean delay.

Using LVT transistors is not an effective way for reducing delay variation in sub-threshold operation region. Figure 5.11 compares two versions of inverter tested in previous part. The threshold voltages of transistors used in Low-VT version is higher than 300 mV (416 mV for LVT and 697 mV for HVT family of transistors). As Figure 5.11a shows, the delay variation around mean delay of Low-VT version is 51.7% which is 1.6% lower than High-VT version.

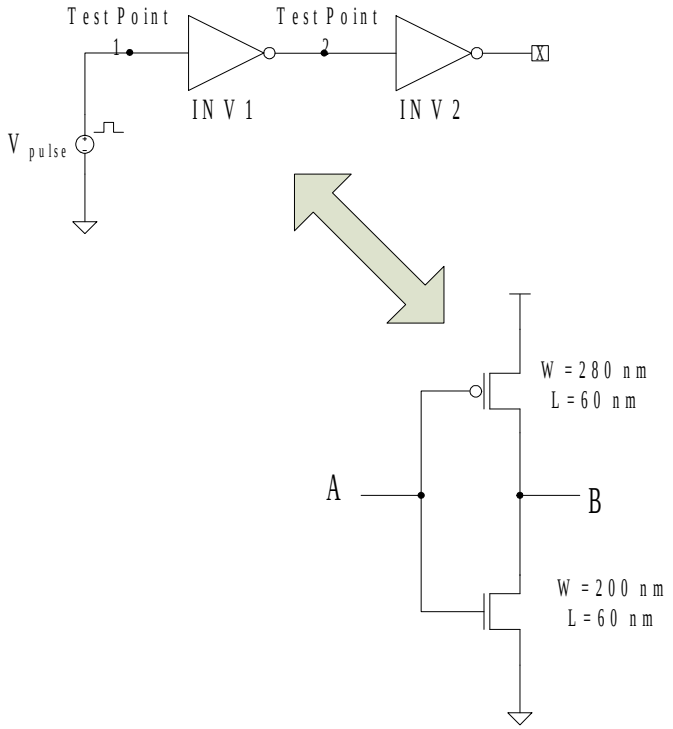
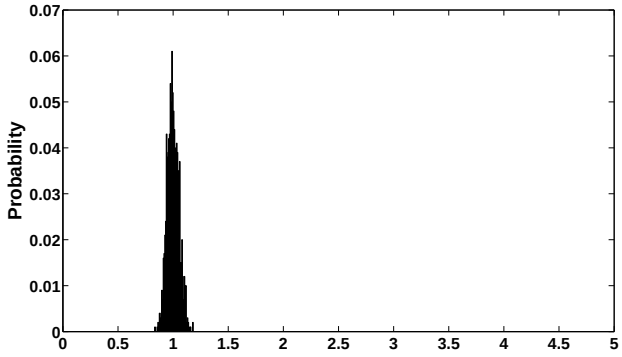
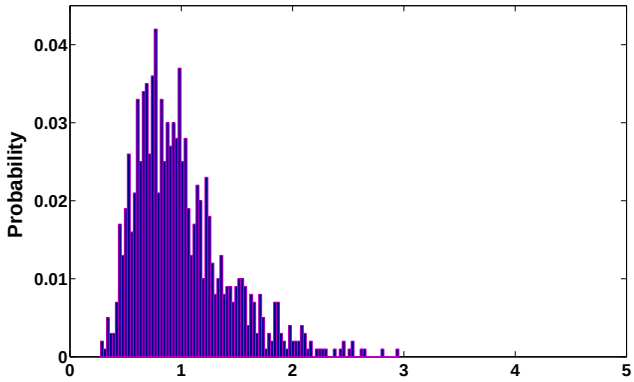


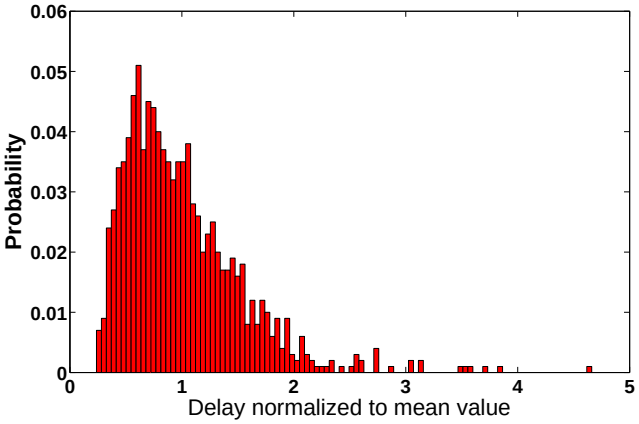
Figure 5.9: Test bench for delay variation analysis and transistor sizes which are used in standard cell inverters. Both LVT and HVT versions of inverters use the same transistor dimensions.



(a) 1.2 V

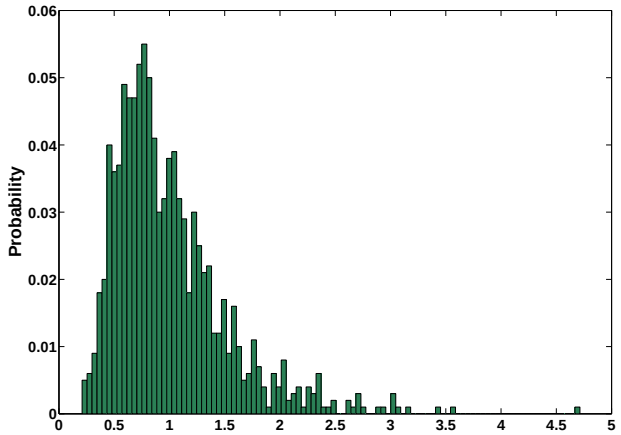


(b) 500 mV

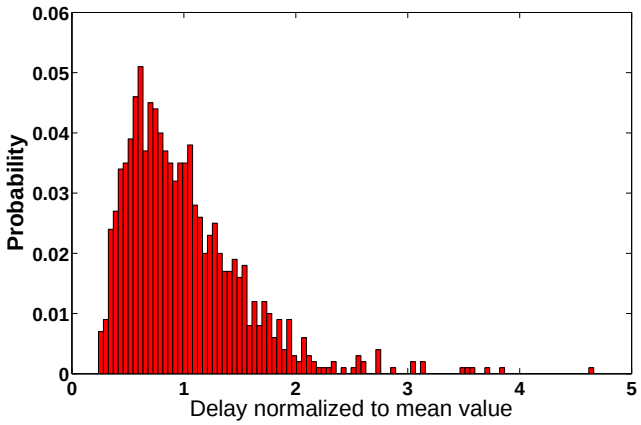


(c) 500 mV

Figure 5.10: Delay variation in nominal voltage and sub-threshold region.



(a) 300 mV, Low-VT



(b) 300 mV, Hight-VT

Figure 5.11: Delay variation of High-VT and Low-VT versions under same conditions.

5.5 Summary

In this chapter we customized selected topology according to techniques discussed in Chapter 3 which were mainly stack-effect, long-channel devices, Floating-body-biasing and combination of these. As we will see on next chapter using FBB is not possible in most of the designs and because of this it is not covered in detail in this project.

By combining stack-effect and long-channel methods, another 39% of leakage current was reduced. But at the same time performance is degraded, since the same as leakage, I_{ON} was reduced and node capacitances were increased. Also, customized topology showed a very good reliability at the range of 300-400 mV.

Delay variation in sub-threshold has a very wide range and a designer should consider the variation of 50%. Shorter delays have higher probability to happen, but longer delays could be several times longer than average delay.

Chapter 6

Layout

Chip area is one of the most important design factors and needs a special effort to design an area efficient layout. In digital-custom-cell design, in addition to usual noise, isolation and design rule considerations, there are special integration rules that need to be considered, since these cells will be automatically placed and routed and should not violate with other cells design rules. In most of the cases, especially when the design dimensions are small, these constraints add a huge area overhead to design's area. Also usually in digital designs wide power rails on Metal-1 are used which takes large area. Since Metal-1 is widely used in devices for interconnection of lower layers, a large area around these power rails remain unused. Consequently it is very essential to try to use the area as efficient as possible.

With keeping leakage in mind as the first design constraint, different methods for area minimization were used that the important points will be explained briefly.

6.1 Standard-cells

As mentioned earlier, standard-cells have a defined format that assures automatic violation-free placement of cells and their interconnection. The main rules could be listed as following:

- Height of the cell: As shown in Figure 6.1 the height of all cells need to be a multiple of $2.6\,\mu m$.
- Width: The width of the cell must be multiple of $0.1\ \mu m$.
- Power interconnections: A certain part of the cell will overlap with power rails which are on Metal-1. The overlapped area should not contain any Metal-1 connection from other nets.

- Implants and N-Well size: The size of p/n-type implants and N-Well at the borders should have their specific values.

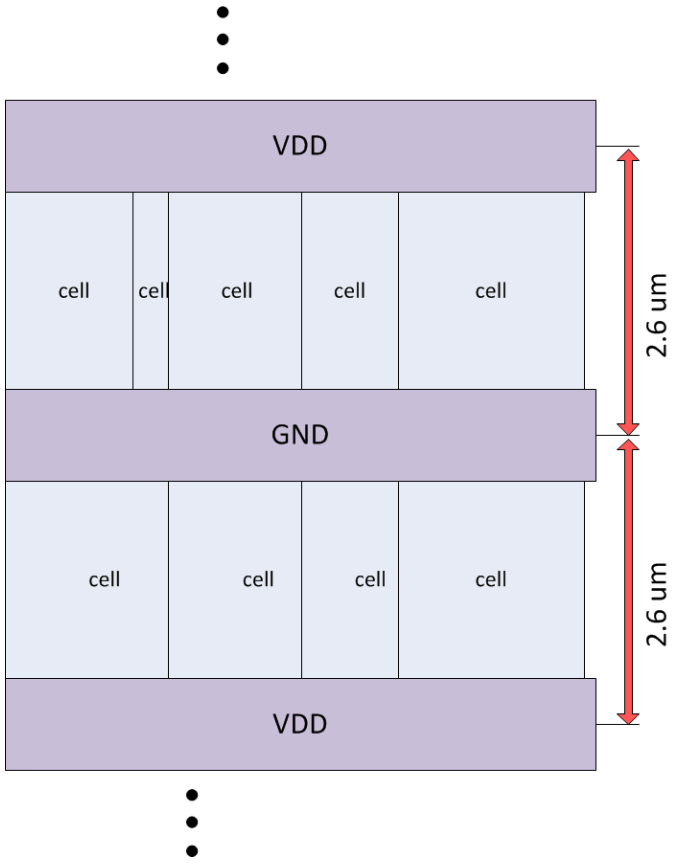


Figure 6.1: Power rails and standard cells placements.

6.2 Area vs. Leakage

To have different selection options for area/leakage optimized designs, different versions of the *D-Latch-3* is implemented. Extra to transistor sizes that were chosen for leakage reduction in Chapter 5, two smaller versions of *D-Latch-3* were designed. In new versions top and bottom transistors in stack changed to be 65 nm (Figure 6.2). By doing this, in addition to area reduction, the capacitance load in store-node decreases as leads to a better timing specifications. As expected decreasing area costs with increased leakage. Table 6.1 lists area-leakage profile for 3 cases.

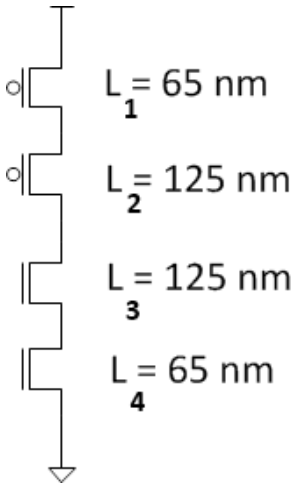


Figure 6.2: Length of transistors for having a dense layout.

Table 6.1: Area and leakage comparison of implemented layouts(in *Standard-cell format*).
* Compared to *Standard-cell*

$L_1, L_4[\mu m]$	$L_2, L_3[\mu m]$	Width$[\mu m]$	Area$[\mu m^2]$	Leakage reduction [%]*	Area reduction [%] *
125	125	2.3	5.98	77.2	10.74
65	125	2.1	5.46	70	18.50
65	90	1.9	4.94	67.7	26.26
65	65	-	-	61.5	-

Values listed in Table 6.1 are for the case that cells are being used with a*Commercial Standard-cell* library. For custom libraries or other applications, there is no need to follow the *Standard-cell* rules which increases the height of the cell. The pure area of designed layout is listed in Table 6.2.

6.3 Other Area Minimization Techniques

It is possible to combine the layout of multi-bit cells together to have a denser layout, since as mentioned before, there are some unused areas between cells both in horizontal and vertical directions. By using the vertical area gaps and placing

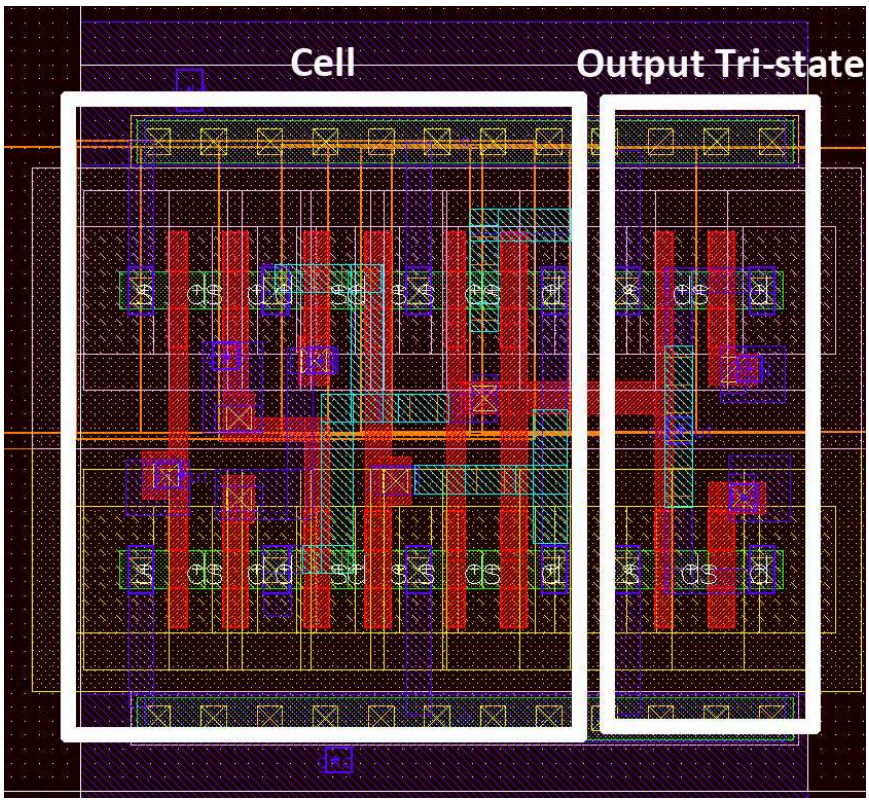


Figure 6.3: Layout of *D-Latch-3* with 65 and 90 nm transistors and connected output tri-state.

Table 6.2: Area comparison of implemented layouts with different cell-heights (The height in these layout designs do not follow the cell-height (2.6 μm) in SCL which is used in this project, .
* Compared to *Commercial Standard-cell*

$L_1, L_4[\mu m]$	$L_2, L_3[\mu m]$	Width$[\mu m]$	Area$[\mu m^2]$	Area reduction [%] *
125	125	2.1	4.2	37.9
65	125	2.0	3.20	52.7
65	90	1.8	2.88	57.1
65	65	-	-	-

another bit-cell (this method is tested in another project which was about near-threshold memory cells), 50% of area reduction was achieved. Also by using the

horizontal gaps between 2-bit cells and combining 4 bits, another 14% area was saved. So by employing this method, 64% area reduction is possible. It should be noticed that mentioned results are possible for narrow and near to minimum width transistors with the SCL and technology used in this project and combining wider transistors in different SCLs and technologies is not tested yet.

6.4 Post-Layout Simulation

After designing the layout, a post-layout simulation was performed. Figure 6.4 shows the applied signals and outputs. The simulation is done for *D-Latch-3* with 65 and 90 nm transistors. As it can be seen, there is a relatively small error between the results of schematic and layout. *Calibre* is used for layout extraction. The extraction type is R+C+CC and format is *Calibreview*.

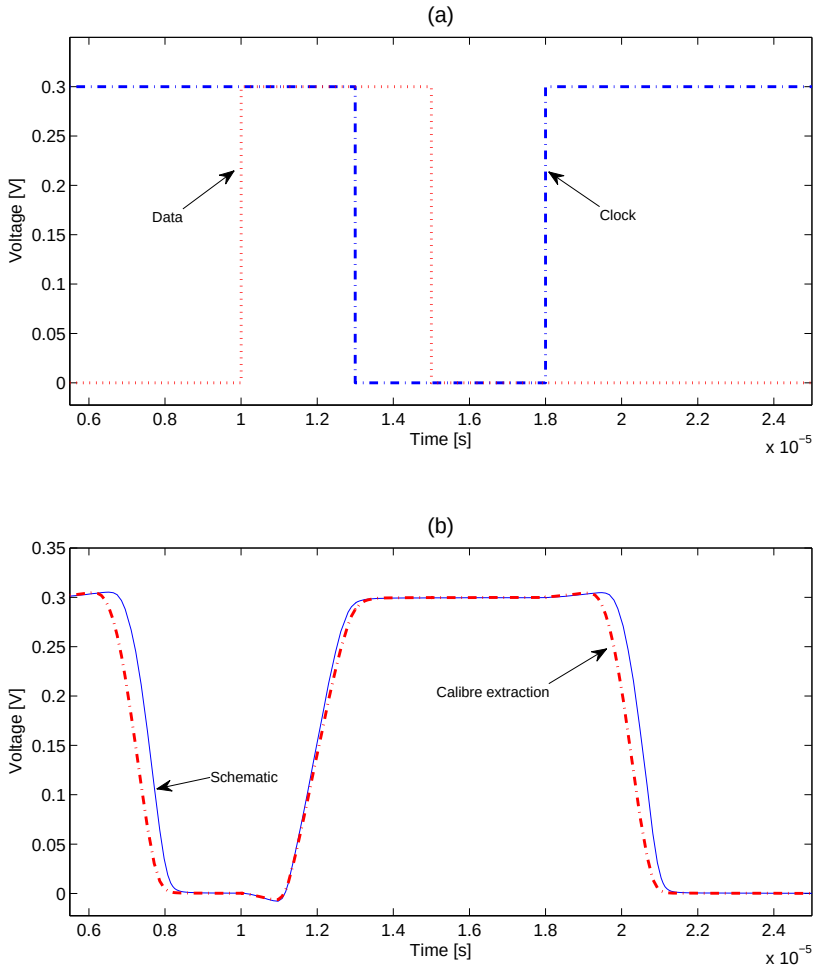


Figure 6.4: Post-Layout simulation. a) Data and Input-Enable signals, b) Schematic and CalibreView extraction simulations.

6.5 Summary

In this chapter the implemented layouts were explained and it was seen that despite of having long and extra stacked transistors, the layout is smaller than *Standard-cell's* dimensions. By using the unused gaps between power-rails and neighbouring cells and also combining multiple cells in single layout, it is possible to reduce the area by 50% for two bits and 64% for 4 bits. Using this techniques a layout was designed for a similar cell which confirmed these results.

Chapter 7

Conclusion

Moving to new technologies with smaller design rules has always been a good solution to minimizing the size and power and integrating more functionality. Beginning from 180 nm, static power consumption has changed to a major issue which consumes higher percentage of total energy consumption in each new technology compared to older technologies.

To have a more power efficient design, different design levels from physical transistor implementation methods up to software algorithms should be optimized. Static memories are widely used components in today's devices and even in some designs memories are dominant power consumers. So in this project different leakage reduction methods in gate-level memory design were analysed.

For low-activity designs, working in sub-threshold regime gives the minimum energy consumption. Depending on activity factor and working-frequency it is possible to find the optimal supply voltage and threshold voltage to have minimum energy dissipation. The supply voltage in this project was 300 mV which was determined by other parameters and projects and all leakage reduction techniques were studied for this supply voltage.

With leakage as the first design constraint, the latch with the least leakage which was functional in sub-threshold region was selected and by using transistor level leakage reduction techniques customized to reduce leakage even more. From leakage reduction techniques, **stacking-effect**, using **long-channel** devices and their combination found to be very effective and were used in customization process.

Timing properties of customized cell was analyzed and different methods for decreasing delay and rise/fall time using transistor sizing was introduced. It was

seen that using two times longer nMOS transistors compared to pMOS transistors produced almost symmetric timing properties both on data low-to-high and high-to-low transitions which could be important for some applications.

Monte carlo simulations for variation sensitivity of customized cell showed a good static noise margin for 300 mV and above.

Multiple layouts with different area, leakage and delay profiles were designed. Despite of using extra transistors in the stack chain and using almost two times longer transistors, the designed layout for customized cell consumed almost 11% less area. Other versions with shorter transistors consume even less area. Combining multiple bits in layout found to be a good method to win chip area. In a similar project for near-threshold memories, doing this gained 64% area reduction over standard cell.

Leakage current of custom cell was much less than *Standard cell*'s leakage(77% with 125 nm transistors and 62% with 65 nm transistors).

7.1 Future Work

As results in this project confirm, by customizing gates for sub-threshold region and leakage both area and energy consumption could be optimized. With increasing static energy consumption in newer technology as well as increasing demand for sub-threshold design, having a special *sub-threshold* library with customized cells would gain a considerable energy and chip area in a given design. Doing this can improve the performance and timing properties of system as well, since most of available standard cells are not time optimized in sub-threshold region and techniques used for performance improvements in strong-inversion region may have a reverse effect in sub-threshold region and make performance worse.
Also in this project mostly gate-level and basic improvements were studied. A complementary work could be done for optimization at higher design levels, like memory blocks. Because of low and limited driving capability of transistors in sub-threshold region, the optimal, minimum and maximum values for fan-in/out, cascaded devices and so on for a given timing constraint could be studied.

Bibliography

[1] Vivek De, Yibin Ye, Ali Keshavarzi, Siva Narendra, James Kao, Dinesh Somasekhar, Raj Nair, Shekhar Borkar "Design of high performance microprocessor circuits", Intel Corporation, Chapter 3

[2] Michael Keating, David Flynn, Robert Aitken, Alan Gibbons, Kaijian Shi "Low Power Methodology Manual For System-on-Chip Design", Springer publications, 2-3

[3] Benton Highsmith Calhoun, Member, IEEE, and Anantha P. Chandrakasan,"A 256-kb 65nm sub-threshold SRM design for ultra-low-voltage operation" IEEE (2007)

[4] Alice Wang and Anantha P. Chandrakasan "Optimal Supply and Threshold Scaling for Subthreshold CMOS Circuits", IEEE Computer Society (2002)

[5] 1 A. Chandrakasan, R. Brodersen, Low Power Digital CMOS Design, Kluwer Academic Publishers, 1995.

[6] Anantha Chandrakasan, Joyce Kwong, "Sub-threshold Design for Ultra Low-Power Systems" Massachusetts Institute of Technology Cambridge, Massachusetts, USA, 94-95

[7] Antoni Ferr, Joan Figueras "Low-power Electronic Design, Ch.3, Leakage in CMOS Nanometric Technologies", CRC Press 2005, 3-5 - 3-8

[8] M-J. Chen et al. Back-Gate Bias Enhanced Band-to-Band Tunneling Leakage in Scaled MOSFETs.IEEE Electron. Device Lett., Vol. 19, April 1998.

[9] 19 K-F. You and C-Y. Wu. A new quasi-2-D model for hot-carrier band-to-band tunneling current.IEEE Trans. Electron. Devices, Vol. 46, June 1999.

[10] S. M. Sze, Ed. High-Speed Semiconductor Devices. John Wiley & Sons, New York, 1990.

[11] R. F. Pierret, Semiconductor Device Fundamentals. Addison-Wesley, Reading, MA, 1996.

[12] H-S. P. Wong et al. Nanoscale CMOS. Proc. IEEE, Vol. 87, April 1999.

[13] C.T. Liu. Circuit requirement and integration challenges of thin gate dielectrics for ultra small MOSFETs. In IEDM Tech. Dig., pp. 747750, 1998.

[14] S.Narebdra,S.Borkar,V.De,D.Antoniadis,A.Chandrakasan,Scaling of stack effect and its application for leakage reduction",ACM/IEEE International Symposium on low Power Electronics and Design,67,August2001, pp. 195200.

[15] Zhanping Chen, Mark Johnson, Liqiong Wei, and Kaushik Roy "Estimation of standby leakage power in CMOS circuits considering accurate modeling of transistor stacks", IEEE, 1998

[16] Pascal Meinerzhagen, Student Member, IEEE, S.M. Yasser Sherazi, Andreas Burg,Joachim Rodrigues "Benchmarking of Standard-Cell-Based Memories in the Sub-VT Domain in 65-nm CMOS Technology", IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS (JETCAS), VOL. 1, ISSUE 2, JUNE 2011

[17] UDSM subthreshold leakage model for NMOS transistor stacks.

[18] Scaling Of Stack Effect and its Application for Leakage Reduction, Siva Narendran, Shekhar Borkar, Vivek De, Dimitri Antoniadisn, and Anantha Chandrakasann, 2001

[19] 26 Seevinck, E., List, F.J., Lohstroh, J. 1987. Static-Noise Margin Analysis of MOS SRAM Cells. IEEE Journal of Solid State Circuits, SC-22, 5 (Oct. 1987), 748-754.