Project Number: FP7-611404

D2.2.2 - Characterization of failure mechanisms for future systems

Authors

Serkan Ozdemir, Nivard Aymerich (INTEL), Marc Riera, Ramon Canal, Antonio González (UPC), Manolis Kaliorakis, Sotiris Tselonis, Nikos Foutris, Dimitris Gizopoulos (UoA), S. Di Carlo (Polito), P. Prinetto (Polito)

Version 1.0 – 24/03/2016

Lead contractor: UPC

Contact person:

Antonio González
Dep. of Computer Architecture
Universitat Politècnica de Catalunya
Campus Nord UPC, Cr. Jordi Girona 1-3, 08034 Barcelona (ES)
Tel. +34-934016988
Fax. +34-934017011
E-mail: antonio@ac.upc.edu

Work package: WP2

Affected tasks: T2.2

Nature of deliverable[1]	R	P	D	O
Dissemination level[2]	PU	PP	RE	CO

[1] *R: Report, P: Prototype, D: Demonstrator, O: Other*

[2] ***PU**: public, **PP:** Restricted to other program participants (including the commission services), **RE** Restricted to a group specified by the consortium (including the Commission services), **CO** Confidential, only for members of the consortium (Including the Commission services)*

COPYRIGHT

© COPYRIGHT CLERECO Consortium consisting of:

- Politecnico di Torino (Italy) – Short name: POLITO
- National and Kapodistrian University of Athens (Greece) - Short name: UoA
- Centre National de la Recherche Scientifique - Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (France) - Short name: CNRS
- Intel Corporation Iberia S.A. (Spain) - Short name: INTEL
- Thales SA (France) - Short name: THALES
- Yogitech s.p.a. (Italy) - Short name: YOGITECH
- ABB (Norway and Sweden) - Short name: ABB
- Università politecnica della Catalunya (Spain) – Short name: UPC

CONFIDENTIALITY NOTE

THIS DOCUMENT MAY NOT BE COPIED, REPRODUCED, OR MODIFIED IN WHOLE OR IN PART FOR ANY PURPOSE WITHOUT WRITTEN PERMISSION FROM THE CLERECO CONSORTIUM. IN ADDITION TO SUCH WRITTEN PERMISSION TO COPY, REPRODUCE, OR MODIFY THIS DOCUMENT IN WHOLE OR PART, AN ACKNOWLEDGMENT OF THE AUTHORS OF THE DOCUMENT AND ALL APPLICABLE PORTIONS OF THE COPYRIGHT NOTICE MUST BE CLEARLY REFERENCED

ALL RIGHTS RESERVED.

INDEX

COPYRIGHT **2**

INDEX **3**

Scope of the document **5**

1. Introduction **7**

2. Target technologies, Modeling and Circuit Design **8**

2.1. Technologies Review **8**

2.2. Predictive Models **8**

2.3. Circuit Design **9**

2.3.1 SRAM Cells 10

2.3.2 Latch and Flip Flop 11

2.3.3 Logic Gates 12

3. Description of Failure mechanisms **13**

3.1. Random Dopant Fluctuations (RDF) **14**

3.2. Line Edge Roughness (LER) **15**

3.3. Random Telegraph Noise (RTN) **16**

3.4. Electromigration (EM) **17**

3.5. Metal Stress Voiding (MSV) **18**

3.6. Gate Oxide Wearout (GOW) **19**

3.7. Hot Carrier Injection (HCI) **20**

3.8. NBTI/PBTI Aging **21**

3.9. Radiation Induced Faults (RIF) **22**

3.10. SOI Self-Heating (SHE) **24**

3.11. Other Sources **25**

3.12. Sources of Failure mapped with technologies **25**

4. Characterization of Different Sources of Failure **26**

4.1. Methodology to Characterize Soft Errors **27**

4.1.1 Modeling Circuit Level Soft Error Rates (SER) 28

4.1.2 Critical Charge (Qcrit) 28

4.1.3 Mapping Qcrit to SER 30
4.1.4 Neutron Flux 32
4.1.5 Time Vulnerability Factor and Masking Effects 33
4.1.6 Evaluation Framework and Tools 34
4.2. Multi Cell Upsets (MCU) Model 37
4.3. SER and Aging Combined Effects 39
5. Analysis of Basic Components 41
5.1. Analysis of SRAM Cells 41
5.2. Analysis of a Latch 42
5.3. Analysis of Logic Gates 43
5.4. Analysis of SRAM Cells with Aging 44
6. Trends 47
6.1. Technology Trend 47
6.2. Voltage Trend 48
6.3. Temperature Trend 49
6.4. Fanout Trend 50
6.5. Location Trend 51
7. Conclusions 53
8. Acronyms 54
9. Bibliography 55

Scope of the document

This document is an outcome of task T2.2, "***Reliability failure mechanisms for future systems***", elaborated in the description of work (DoW) of the CLERECO project under the Work Package 2 (WP2).

Figure 0.1 depicts graphically the goal of this deliverable, its main results, the inputs it uses and which work packages will use its outputs.

D2.2.2 focuses on describing the most important failure mechanisms in current and future technologies, and performs a characterization of how these failure mechanisms affect the reliability of basic circuit components. The technologies considered in this deliverable are those identified in deliverable D2.1 (Report on future technologies that may be used in future computer systems) and their characterization is performed taking into account the reliability metrics identified in deliverable D2.4.1 (Report on system level reliability metrics v.1). Environmental conditions are also considered as described in deliverable D2.3 (Definition of operation modes for future systems".

This deliverable produces two main outcomes for the CLERECO project. First a detailed list of failure mechanisms that may arise in future technology. These failure mechanisms represent the main source of unreliability of complex system. Second, a characterization of the characteristic of each considered failure mechanism in order to compute vulnerability data to be exploited for the upper layers (e.g., error rates, etc.).

The outputs of this deliverable will be strongly exploited within WP3, WP4 and WP5 activities.

It has to be pointed out that CLERECO project does not deal with software bugs/errors but only with the effect of hardware faults and their propagation to software layers.

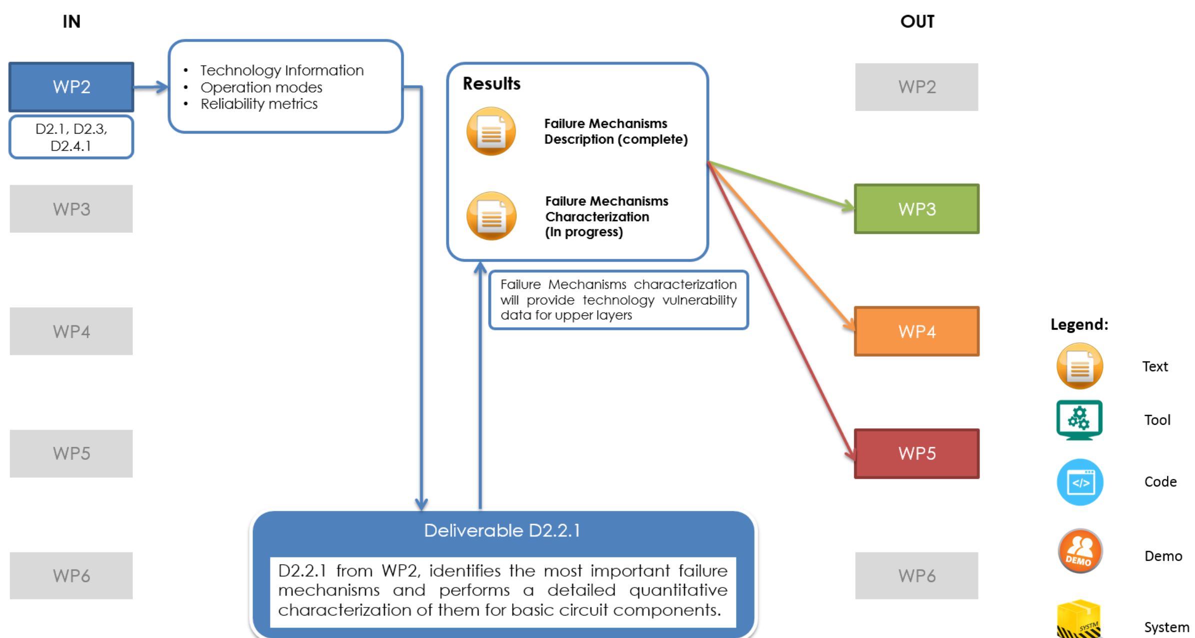


Figure 0.1: Deliverable summary

The document is organized in the following sections:

- **Introduction**. This section sets the background for the document. The objectives of the document and the investigations made for its development are included.
- **Target Technologies, Modeling and Circuit Design.** This section makes a review of the most promising future technologies and how they are modeled and designed.
- **Description of Failure Mechanisms.** This section describes the most important failure mechanisms divided in a subchapter per source of failure, and maps the sources of failure with the technologies affected.
- **Characterization of different sources of failure.** This section explains how the different sources of failure can be characterized to obtain the vulnerability factor at technology level, focusing on soft errors and its combination with aging.
- **Analysis of Basic Components.** This section shows the data obtained on soft error rates of the most basic elements of any electronic device.
- **Trends on Soft Error Rates.** This section shows the trends on soft error rates for a variety of technologies and components obtained from the previous characterization.
- **Conclusions.** This section summarizes the document and takes some conclusions on how this part of the project is going on.
- **Acronyms and Definitions.** A section containing a list of the most important acronyms used in the document and their definitions.
- **Bibliography.** A section containing a list of the references used to make this part of the project and this document.

The changes with respect to the preliminary version 2.2.1 of this document are the following:

- **Circuit Design:** New section describing how our circuits have been designed.
- **Methodology:** New section describing our own methodology to compute the SER reviewing all the elements that have been taken into account to develop our methodology.
- **MCU Model:** New section describing our Multi Cell Upsets model.
- **Aging:** Two new sections, one describing the aging model we used and another describing our results.
- **Analysis and Trends:** Two new chapters, one showing the results of each basic component and another showing different trends on soft error rates.
- **Conclusions:** Final conclusions taking into account the final results obtained.

1. Introduction

System reliability has become an important design aspect for computer systems due to the aggressive technology miniaturization, which introduces a large set of different sources of failure for hardware components [1][2][3][4][5][6][7]. Errors are strongly related to the technology used to build the hardware blocks composing the system and are caused by effects such as physical fabrication defects, aging or degradation (e.g., NBTI), environmental stress (e.g., radiations), etc.

After a raw fault manifests in a given hardware block, it can be propagated through the different hardware structures composing the full system and reach the software layer by corrupting either data or instructions composing a software application.

The reliability stack depicted in Figure 1.1 summarizes the basic idea of system reliability evaluation of CLERECO. Every system is split into three main layers: (1) technology, (2) hardware and (3) software. CLERECO's goal is to contribute with a full system reliability estimation methodology, which takes into consideration all these factors to provide an accurate estimate of the expected reliability of the system as early as possible during design.

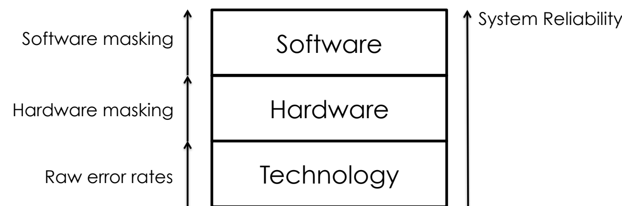


Figure 1.1: CLERECO reliability stack

Each layer included in Figure 1.1 defines an interface with the upper layer, which in turns sets how the errors can be propagated from one layer to the next one. In this deliverable we focus on errors that can cross the interface between the technology and the hardware layer. The main relevant elements required to analyze the impact of technology on the reliability of a system are shown in Figure 1.2.

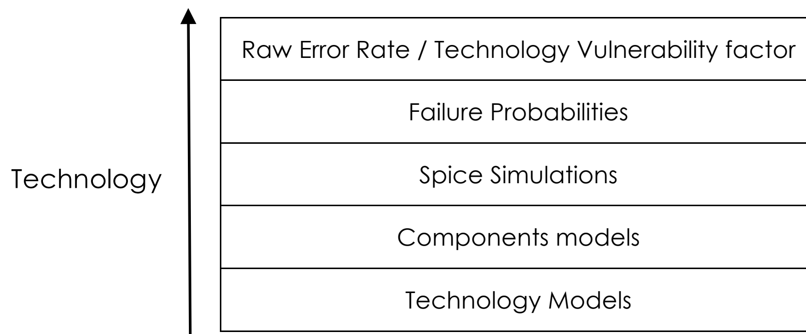


Figure 1.2: The technology layer

In this document, the most important failure mechanisms are described and in this preliminary version we will focus on the characterization of soft errors. The first step is to find predictive models for future technologies and develop models for the components that need to be analyzed. Next step is to perform Spice simulations to test the reliability of these components in the new technologies in order to compute failure probabilities and derive the Technology Vulnerability Factor (TVF) that will be required for the next layers of the stack.

2. Target technologies, Modeling and Circuit Design

This chapter is divided into three sections. In the first section, the technologies that are strong candidates to be used in a near future are briefly reviewed. A detailed list of these technologies is provided in deliverable D2.1. Second section describes the models for future technologies, and comments which are the main models and which models are used for our simulations. The second section gives the general guidelines about circuit design in SPICE using the predictive technology models, and describes the components analyzed in this project.

2.1. Technologies Review

Planar CMOS technology is still being used and will stay here for a long time. Planar CMOS has been scaled down during many generations but physical limitations and reliability problems are starting to be a serious challenge for newer technology nodes. As planar CMOS have their limitations, other technologies such as multi gate FinFET transistors are gaining interest and are analyzed in this project. FinFETs [9] have the conduction channel wrapped by a thin silicon "fin" which forms the body. The thickness of the fin is the major challenge for FinFETs fabrication as it determines the effective length of the channel. Another technology being used nowadays is silicon on insulator (SOI), which refers to the use of layered silicon-insulator-silicon substrate instead of the conventional silicon substrate to reduce parasitic capacitance and improve performance. Finally, newer technologies that are still being investigated such as III-V HEMT will be considered in this project.

2.2. Predictive Models

Transistors are simple devices with a complicated physical behavior. Transistor models are used for almost all modern electronic design work. Circuit simulators such as SPICE use models to predict the behavior of a design and ensure the reliability of the circuit. Most design work is related to integrated circuit designs, which have a very large tooling cost, and there is a large economic incentive to get the design working without any iterations. Complete and accurate models allow a large percentage of designs to work the first time. Transistors are modeled using compact models with predicted parameters [13][14]. Compact models include effects of the transistor layout such as width, length, current-voltage characteristics, parasitic capacitances, resistances, time delays and temperature effects, among other physical effects.

The models used in SPICE are a hybrid of physical and empirical models. Physical models are based on the physical phenomena within a transistor, while empirical models are based on fitting measured data. Such models are incomplete unless they include specification of how parameter values are to be extracted for a specific technology node. In SPICE, these parameters are specified in the model card of each technology. To attempt standardization of model parameters used in different simulators, an industry working group was formed, the Compact Model Council (CMC), to choose, maintain and promote the use of standard models. One of their main goals is to predict how circuits using the next generation of devices should work, to identify which direction the technology should take, and have models ready beforehand.

In the area of predictive modeling, the most important models are the BSIM (Berkeley Short-channel IGFET Model) Group [10] and the Arizona State University (ASU) PTM [11] based on BSIM, which were developed for Planar CMOS technology nodes up to 7nm. BSIM was developed by empirically extracting model parameters from early stage silicon data while ASU PTM improved the methodology by taking into account significant physical correlations among model parameters. Both groups also developed PTM models for multi-gate transistors, mainly FinFETs, for sub-20nm technology nodes. Moreover, the Berkley group has also developed some SOI models.

All the predicted models are developed based on the scaling theory of planar CMOS and multi-gate devices, physical models and the International Technology Roadmap for Semiconductors (ITRS) projections [12], which recollects data of the industry and makes projections about the future technologies.

We use the ASU PTM models for Planar and FinFET technologies since they include the model cards of the most recent technology nodes, which can be directly used to simulate in SPICE. For SOI technology, we tried the Berkley model (BSIM-SOI) but as the model cards are not included the results were not accurate. Then, we found an alternative model, the UTSOI model from the *Laboratoire d'électronique des technologies de l'information* (CEA-Leti) [40], which has a model card of planar SOI with values for 20/22nm. For the SOI FinFET technology we have used a 10nm model from the European project named TRAMS [41]. Finally, we have obtained an III-V HEMT model from [42].

2.3. Circuit Design

In Table 2.1 there is the list of all the hardware components, technologies and technology nodes to be analyzed. The technologies in red are still not available as there are no public technology models for them, and efforts are being made to find these models for future work. All the components have been modeled and analyzed with SPICE. For this purpose, we developed a description of the necessary circuits at transistor level and use the appropriate predictive technology model (PTM) of the technology node to be analyzed.

Technology (CMOS)	Technology Nodes		Circuits
Bulk Planar (ASU PTM Models)	22nm and 16nm		SRAM Cells 6T/8T/10T
Bulk FinFET (ASU PTM Models)	20nm and 14nm		Flip Flop - D
SOI Planar (UTSOI Model)	22nm	X	Latch
SOI FinFET (TRAMS Model)	10nm		Logic Gates (AND, OR, NOT…)
III-V HEMT (Offered from [42])	20nm		

Table 2.1: Hardware elements and Technologies analyzed

SPICE (Simulation Program with Integrated Circuit Emphasis) is an electronic circuit simulator used in integrated circuits design to check the integrity of the circuit and predict its behavior. To simulate in SPICE, one needs to describe the base components with transistors, then the circuit netlist and finally select which type of simulation will be performed (e.g. transient, montecarlo) [43]. We use HSPICE, a commercial version of SPICE, to make transient simulations of the components listed to compute their Qcrit under different conditions.

CMOS technology provides two types of transistors: an n-type transistor (NMOS) and a p-type transistor (PMOS). These are also defined in the circuit description and their symbols are shown in Figure 2.1. Further details can be found in [44].

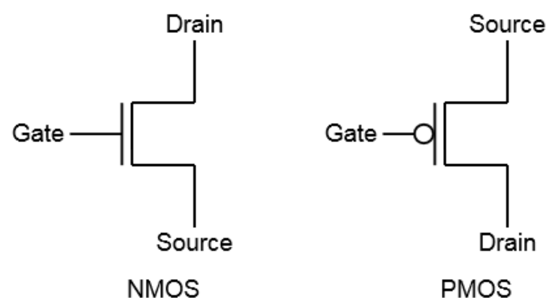


Figure 2.1: NMOS and PMOS Symbols

Transistors of each component need their size to be specified in the SPICE circuit description. Transistors sizing depends on the technology used. In the case of Planar CMOS, the sizing means to specify the length and the width of the transistor in lambdas or nanometers. Examples of most of the circuits can be found in the literature with the sizes in lambdas [44]. In a similar way, the transistors for FinFETs are sized in terms of number of fins that determines the effective width of a FinFET transistor [9].

2.3.1 SRAM Cells

SRAM is a type of memory widely used in current CPUs. For example, it is used in cache memories and register files of processors. SRAMs are made of arrays of cells, each one storing a bit of memory. There are different types of cells depending on the number of transistors used to make the cell, being 6T, 8T and 10T the most common ones. They are depicted in Figure 2.2, Figure 2.3 and Figure 2.4, respectively.

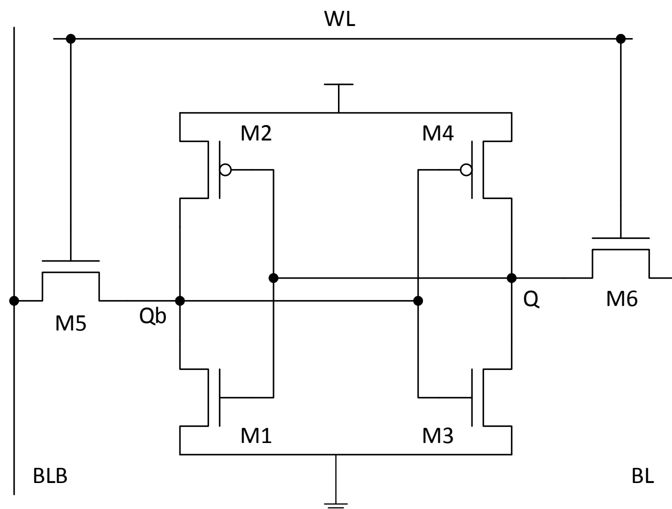


Figure 2.2: Scheme of a 6T Cell

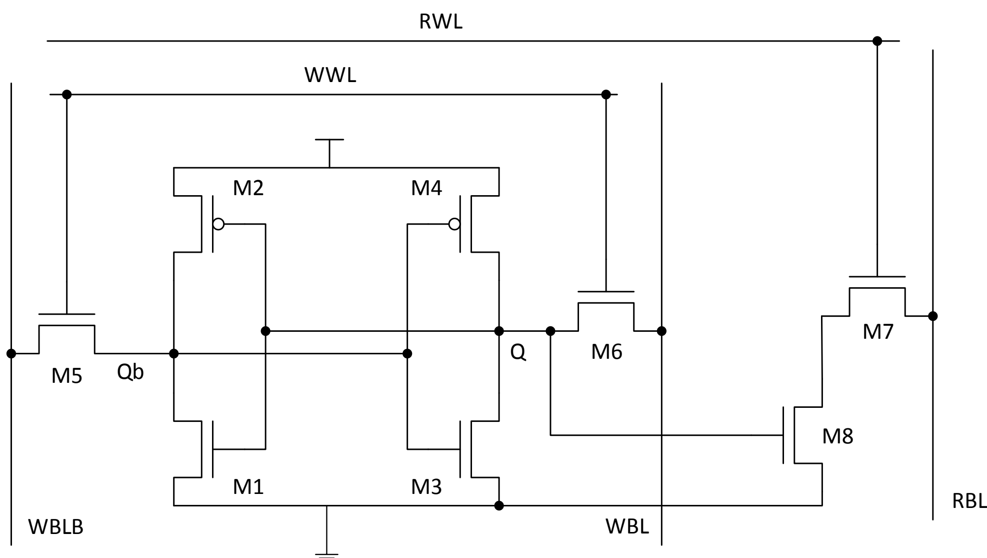


Figure 2.3: Scheme of an 8T Cell

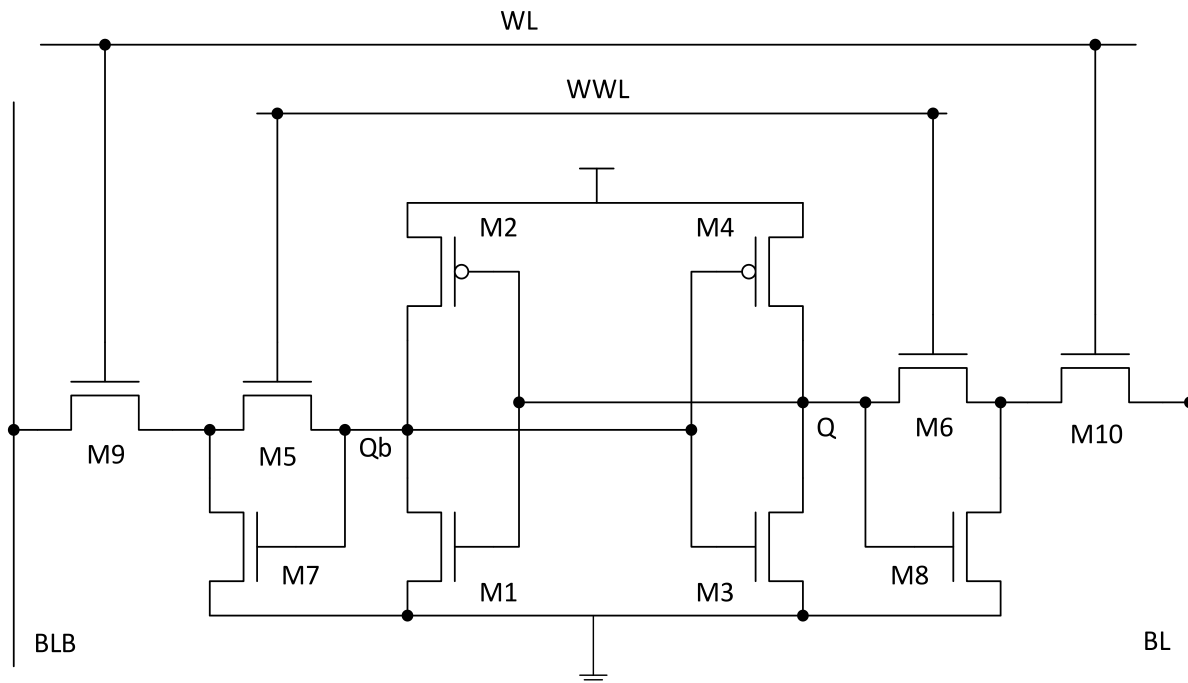


Figure 2.4: Scheme of a 10T Cell

Choosing the 6T cell as example, this cell has a pair of inverters (M1-M4) and two access transistors M5 and M6. This cell needs a careful transistor design as the strength (i.e. Width/Length ratio) of the transistors is crucial to write new values in the cell (Q and Qb) and, at the same time, perform read operations without losing the content. The nMOS transistors in the cross-coupled inverters must be the strongest. The access transistors are of intermediate strengths and the pMOS transistors must be weak. Therefore, for bulk and SOI planar, we have made the nMOS transistors in the inverters (Width/Length) 8/2 λ , access transistors 4/2 λ and pMOS transistors 3/3 λ , which are typical lambda values [44]. In the case of FinFETs, we use 2 fins for M1 and M3 and 1 for the rest of transistors, which are values obtained from the literature [45].

The 8T and 10T cells have as base the 6T cell, but in the case of the 8T, it adds more transistors to decouple the reading from the writing; and, in the case of the 10T, it adds more transistors to be more robust. Transistors of these cells are sized to conserve the same strength of the 6T cell.

2.3.2 Latch and Flip Flop

Latches are the most basic sequential logic elements. Their output values depend not only in the current inputs but also in the previous ones. Therefore, latches are used to store data like state information. Figure 2.5 shows the scheme of the latch used in our simulations being the Flip-flop composed of two of this latches.

Latches and Flip flops are sized similarly to SRAM cells. For our latch, the first logic structure is a combination of a latch and an inverter that forms a tristate buffer. To be able to transfer new data into this latch, the first tristate buffer must be stronger as compared to the feedback inverter and the second tristate buffer [46]. Therefore, we start from minimum sizes (2/2 λ or 1 fin) and then increase following this principle for each technology.

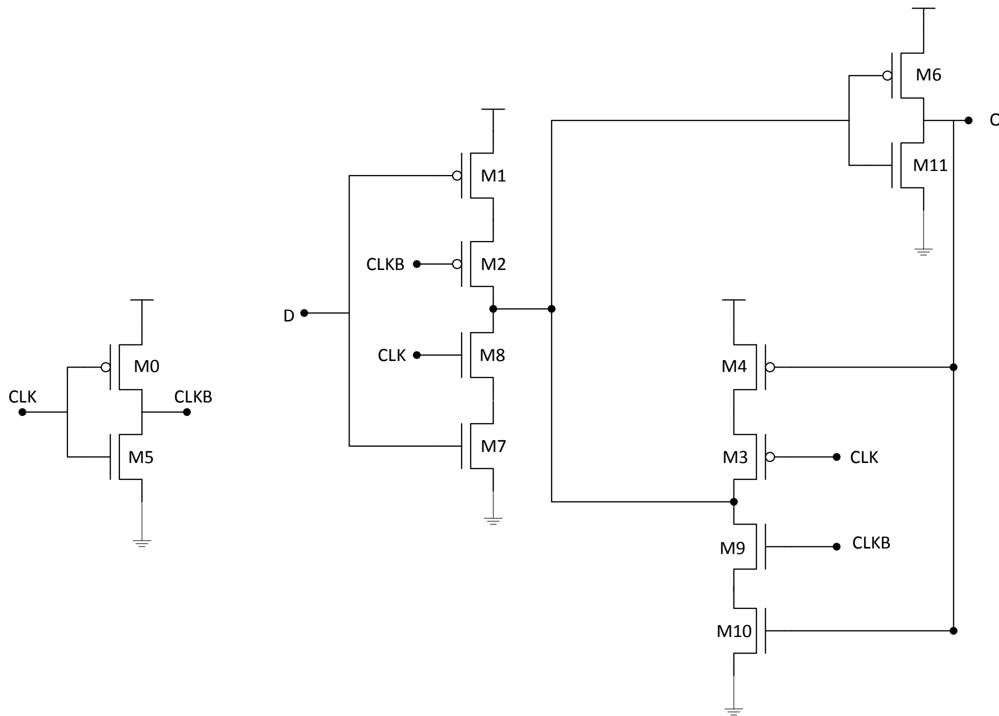


Figure 2.5: Scheme of a Latch

2.3.3 Logic Gates

Logic gates are the basis of any electronic device being the most common the NAND, NOR and NOT, which are depicted in Figure 2.6 with four inputs. Further details on how to implement other common gates and functions can be found in [44].

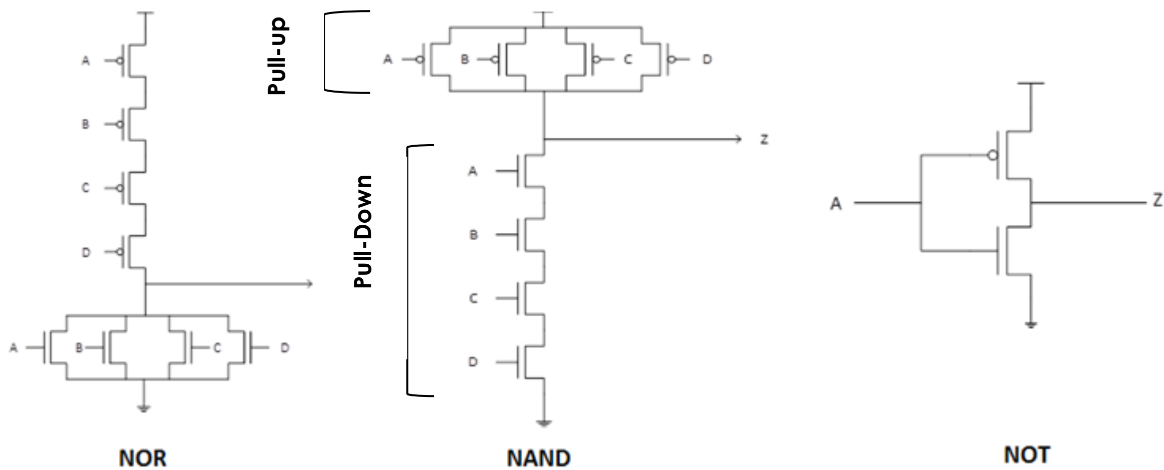


Figure 2.6: NAND, NOR and NOT Schemes

Logic gates are sized to have minimum size but being symmetric so that the delays to commute from 0 to 1 and back are equal or at least similar. This symmetry is achieved by matching the strength (Rs=Width/Length) of the pull down and pull up to 1 Rs. In the case of bulk planar, the PMOS transistors from the pull up have a lower strength (2x-3x) than the NMOS transistors from the pull down. Therefore, to make the gate symmetric, the PMOS transistors are sized with a higher width to increase their strength. PMOS and NMOS transistors of FinFET and SOI technologies have a similar strength relation, which has been tested with SPICE.

3. Description of Failure mechanisms

The first step of this project is to select the failure mechanisms that will be analyzed. In order to do that, we have made a study from the literature looking for the failure mechanisms that may have a highest impact on the vulnerability of current and future technologies. The results are described in this section.

Faults, errors and failures [16] are terms that are often confused but have different meanings. A fault is a defect that may trigger an error or stay dormant. Faults in hardware structures could arise from defects, imperfections, or interactions with the external environment. Examples of faults include manufacturing defects in silicon chip or bit flips caused by cosmic ray strikes.

Faults are usually classified into three categories: permanent, intermittent and transient. Permanent faults remain for indefinite periods till corrective action is taken. Oxide wear out leading to a transistor malfunction is an example. Intermittent faults appear, disappear, and then reappear and are often early indicators of permanent faults. Finally, transient faults are those that appear and disappear in a very short period of time (typically one cycle). Bit flips or gate malfunctions due to an alpha particle or a neutron strike are examples of transient faults. A fault in a particular system layer may not show up at the user level. This may be because the fault is being masked in an intermediate layer, a defective transistor may affect performance but not the correct operation, or because any of the layers may be designed to tolerate some faults.

Errors are manifestation of faults. Faults could cause an error, but not all faults show up as errors, as they may be masked or tolerated. Errors can be classified in the same way as faults, so a permanent fault may cause a permanent error and so on. The final term, failure, is defined as a system malfunction that causes the system not to meet its correctness, performance, or other guarantees. Figure 3.1 summarizes this terms in the way of when they can arise. As an example, Figure 3.2 shows the different types of SRAM failures, which can arise from manufacturing defects, process variations and alpha particles or neutron strikes.

Figure 3.1: Summary of fault, error and failure terms

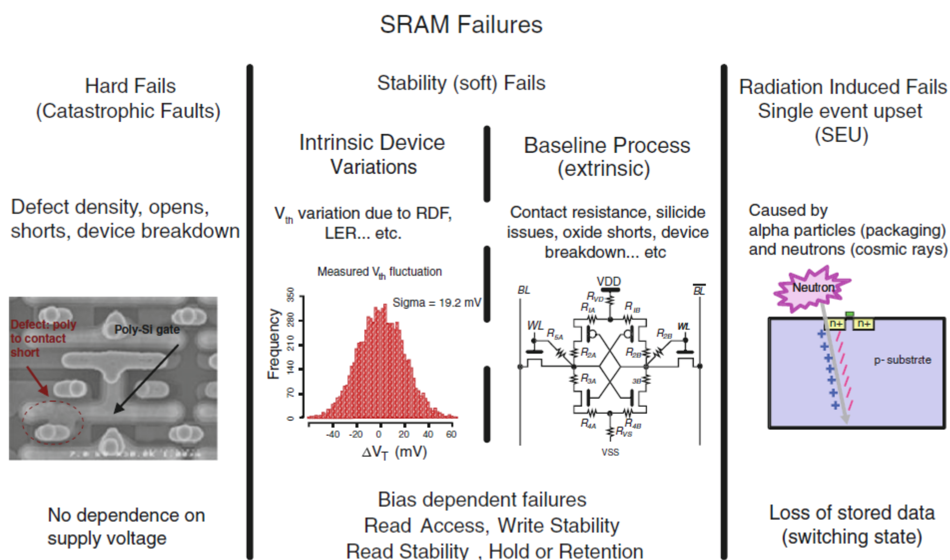


Figure 3.2: Different types of SRAM failures

3.1. Random Dopant Fluctuations (RDF)

Random Dopant Fluctuations (RDF) [15][8] is a type of process variation that may cause a failure, and are primarily caused due to the random fluctuation in the number of dopant atoms in the channel gate and their placement. The effect is more pronounced as devices are scaled down, as the total number of dopant atoms in the depletion region decreases with subsequent technology nodes. This fluctuation in the number of dopants in the transistor channel results in variations in the threshold voltage (V_{th}) for the device.

The problem of RDF has been well documented over the last three decades and it has been predicted to be a major challenge for controlling device performance. Due to the random nature of this phenomenon, the threshold voltage (V_{th}) of the transistor undergoes significant variation. This is because the intrinsic value of V_{th} is dependent on the charge of the ionized dopants in the depletion region. The standard deviation of V_{th} follows the inverse square law of the device area. In other words, with scaling of technology, σV_{th} dependent on RDF increases for transistors with smaller area. The variation in V_{th} due to RDF has been demonstrated to follow a Gaussian distribution with its standard deviation derived as:

$$\sigma_{V_{th}} = \left(\sqrt[4]{2q^3\varepsilon_{Si}N_a\phi B}\right) x \frac{T_{ox}}{\varepsilon_{ox}} x \frac{1}{\sqrt{3WL}}$$

Where q represents electron charge, ε_{Si} and ε_{ox} are permittivity of silicon and gate oxide, N_a is the channel dopant concentration, ϕB is the difference between Fermi level and intrinsic level, T_{ox} is the gate oxide thickness, W and L are the channel width and length of the transistor, respectively.

The trend to reduce the total number of dopant atoms when reducing device dimensions is shown in Figure 3.3. It is evident that reducing the total number of dopant atoms in subsequent process nodes makes σV_{th} increase significantly. Even two equal transistors with the same number of dopants can have different voltage thresholds due to their position in the channel. As RDF is inversely proportional to the device area, SRAM cells, which are usually constructed with the minimum geometry transistors available, are intrinsically the most susceptible to this type of variation.

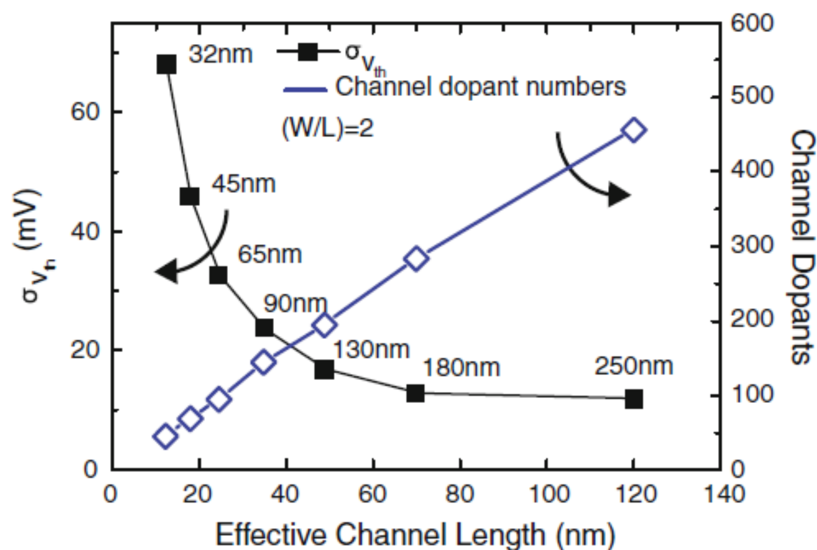


Figure 3.3: Impact of RDF on Vth variation and number of dopants of a MOSFET

3.2. Line Edge Roughness (LER)

Line-edge roughness (LER) [15][8] is caused by the change in the shape of the gate along the channel width direction as can be seen in Figure 3.4. This roughness in the edge of the gate is caused by the inherent characteristics of the materials forming the gate and additional process steps such as etching and imperfection in lithography.

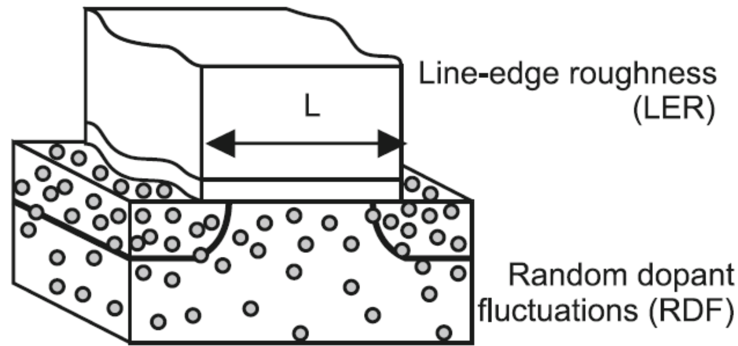


Figure 3.4: Primary sources of variation: RDF and LER

The impact of this phenomenon is more pronounced at technologies below 50nm, as process technologies use light sources with wavelengths much higher than the minimum feature size, increasing gate variation due to LER. LER impacts directly on V_{th} variation following a Gaussian distribution, and is inversely proportional to the gate width of the transistor. The impact of LER when changing the device dimension from W_1 to W_2 on σV_{th} is given by the following equation:

$$\sigma_{V_{th}\,|\,W_2} = \sqrt{W_1/W_2}\,\sigma_{V_{th}\,|\,W_1}$$

Figure 3.5 shows the impact of LER on V_{th} fluctuation while scaling transistor widths. As explained in [8], the variance of this phenomenon does not decrease with technology scaling despite improvements in the underlying manufacturing technology. As a result, the problem can become critical for devices such as memory cells that are extremely susceptible to V_{th} mismatch.

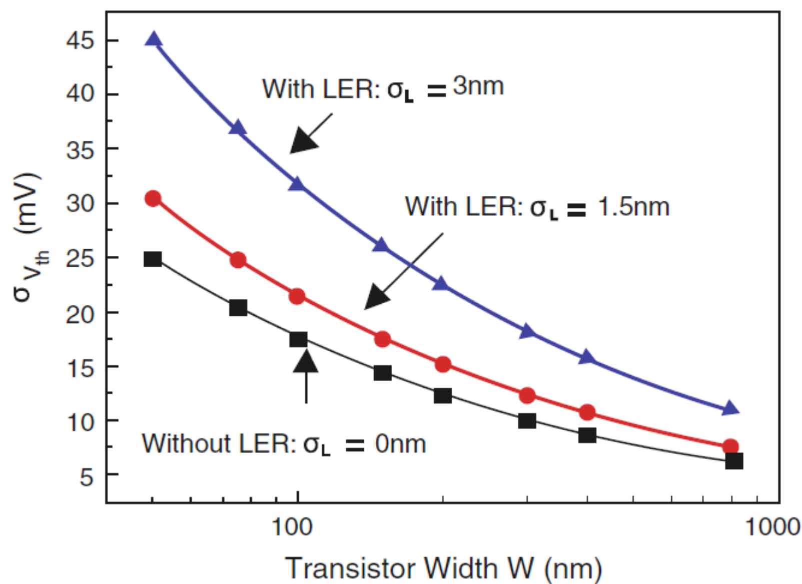


Figure 3.5: Combined effect of LER and RDF on Vth variation

3.3. Random Telegraph Noise (RTN)

Random Telegraph Noise (RTN) [8], also known as random telegraph signal (RTS), is a random fluctuation in the device drain current due to the trapping and detrapping of channel carriers in the dielectric traps at the oxide interface, as shown in Figure 3.6, which causes variation in Vth. The fluctuation in drain current is caused by the change in the number of carriers as well as the changes in surface mobility due to scattering by the trapped charges in the gate dielectric.

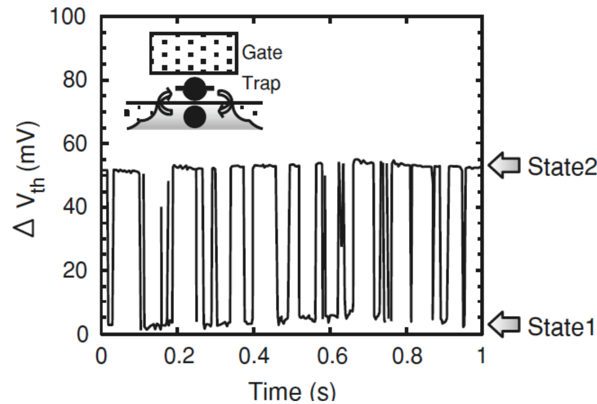


Figure 3.6: RTN Vth variation is caused by trapping and detrapping of charges in the channel

Both RTN and RDF arise due to discreteness in charges, however, RTN differs from RDF in that is time dependent, and fewer charges are involved. Technology scaling increases RTN due to reduction in the number of channel carriers caused. The impact of RTN on V_{th} variations can be estimated as follows:

$$\Delta V_{th,RTN} = \frac{q}{W_{eff} L_{eff} C_{ox}}$$

where *q* is the elementary charger, L_{eff} and W_{eff} are the effective channel length and width, respectively, and *Cox* is the gate capacitance per unit area. The equation shows that V_{th} variation is inversely proportional to device area, and it can become a serious concern for highly scaled technologies and a critical problem for SRAM cells. Figure 3.7 shows that V_{th} variation due to RTN has a non-Gaussian distribution with a long tail, which is a critical concern related to RTN, and RTN may exceed RDF in design impact.

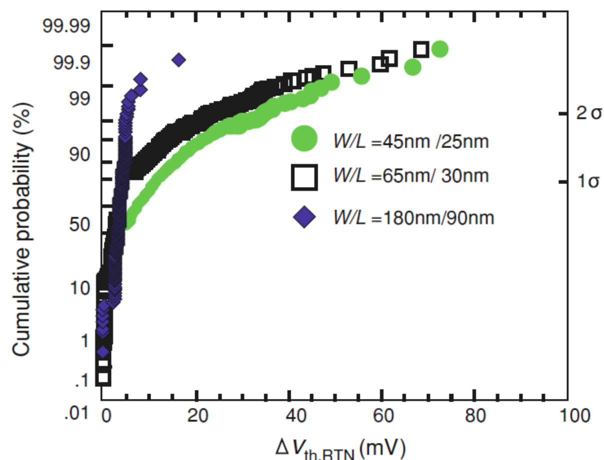


Figure 3.7: Distribution of Vth fluctuation due to RTN in 22nm technology

3.4. Electromigration (EM)

Electromigration (EM) [16] is a failure mechanism that causes voids in metal lines or interconnects in semiconductor devices. Often, these metal atoms from the voided region create an extruding bulge on the metal line itself. EM is caused by electron flow and exacerbated by rise in temperature. As electrons move through metal lines, they collide with the metal atoms. If these collisions transfer sufficient momentum to the metal atoms, these atoms may get displaced in the direction of the electron flow. The depleted region becomes the void, and the region accumulating these atoms forms the extrusion. Figure 3.8 shows the Electromigration effect and Figure 3.9 shows a real example of voids caused by these phenomena.

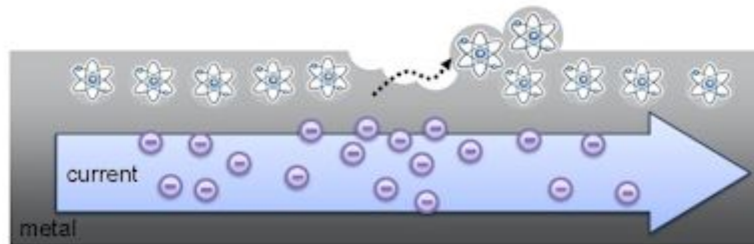


Figure 3.8: Electromigration

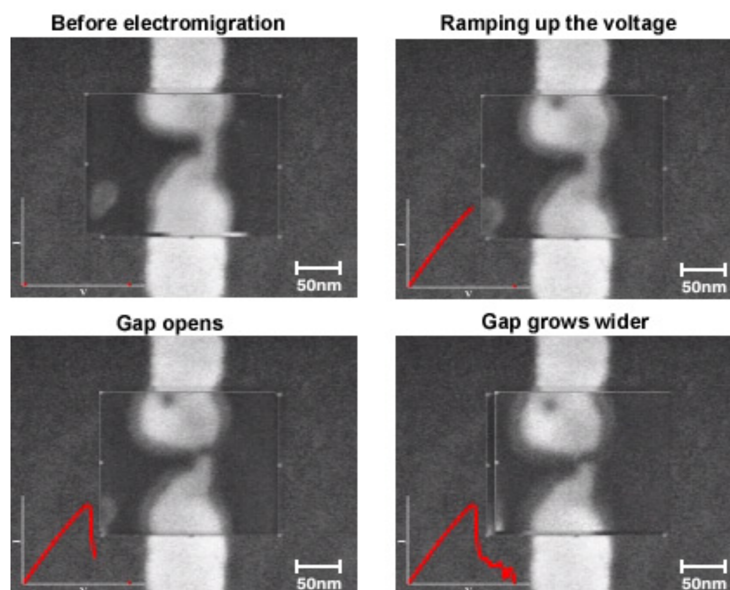


Figure 3.9: Example of a Void due EM [17]

Black's law is commonly used to predict the Median Time to Failure (MeTTF) of a group of aluminum interconnects. This law was derived empirically and applies to a group of metal interconnects, so cannot be used to predict the TTF of an individual wire. The equation is as follows:

$$MeTTF_{EM} = \frac{A_0}{j_e^2} e^{\frac{E_a}{kT}}$$

where A_0 is a constant dependent on technology, j_e is electron current density (A/cm^2), T is the temperature (K), E_a is the activation energy (eV) for EM failure and k is the Boltzmann constant.

As technology shrinks, the current density usually increases, so designers need effort to keep the current density at acceptable levels to prevent EM. Nevertheless, the exponential temperature term has a more serious effect on MeTTF than current density.

3.5. Metal Stress Voiding (MSV)

Metal stress voiding (MSV) [16], also known as Stress Migration, causes voids in metal lines due to different thermal expansion rates of metal lines and the passivation material they bond to. This can happen during the fabrication process itself, when deposited metal reaches very high temperatures (400 ^{0}C or more) for a passivation step, and the metal lines expand and tightly bond to the passivation material. However, when cooled to room temperature, enormous tensile stress appears in the material due to the differences in the thermal coefficient of expansion of the two materials. If the stress is large enough, then it can pull a line apart and the void can show up immediately or years later. Figure 3.10 shows an example of a void caused by stress migration.

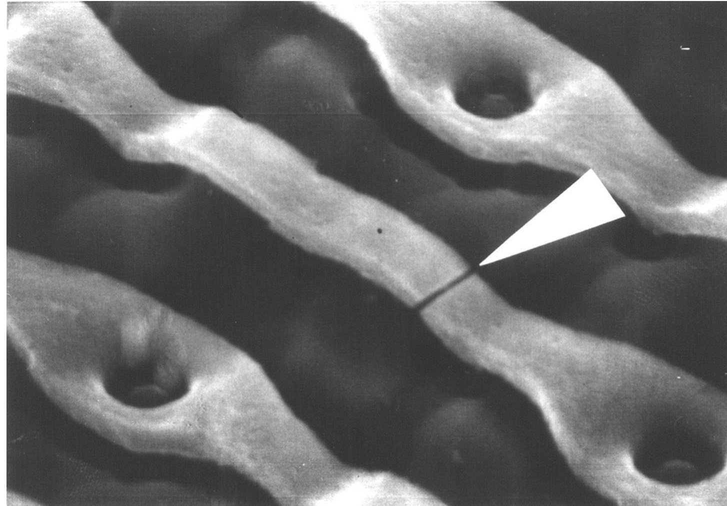


Figure 3.10: Example of a void due to Stress Migration

The Mean Time to Failure (MTTF) due to MSV is given by the following equation:

$$MTTF_{MSV} = \frac{B_0}{(T_0 - T)^n} e^{\frac{E_b}{kT}}$$

Where T is the temperature, T_0 is the temperature at which the metal was deposited, B_0, n, and E_b are material dependent constants, and k is the Boltzmann constant. For copper, n = 2.5 and E_b = 0.9. The higher the operating temperature, the lower the term $(T_0 - T)$ is and the higher the MTTF is. However, the exponential term drops rapidly with a rise in the operating temperature and usually has the more dominant effect.

In general, copper is more resistive to EM and MSV than aluminum and for this reason has replaced aluminum for metal lines in the semiconductor industry. However, copper can cause severe contamination in the fab and therefore needs a more controlled process.

3.6. Gate Oxide Wearout (GOW)

Gate oxide reliability has become an increasing concern in the design of high performance silicon chips. Gate oxide consists of thin noncrystalline and amorphous silicon dioxide (SiO_2). In a bulk CMOS transistor the gate oxide electrically isolates the polysilicon gate from the substrate or bulk of the transistor as can be seen in Figure 3.11. The switching speed of a CMOS transistor is a function of the gate oxide thickness. As technology shrinks, the supply voltage is reduce to maintain the overall power consumption, but this reduces the switching speed. To increase the switching speed, the gate oxide thickness is reduced and rapidly approaches molecular dimensions. Oxides with such a low thickness are referred to as ultrathin oxides and introduce some failure mechanisms.

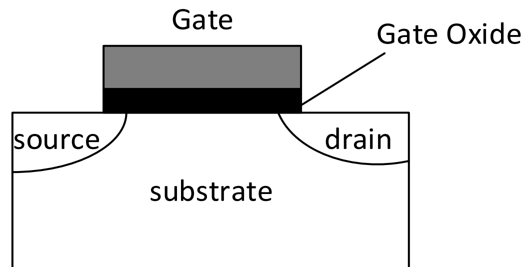


Figure 3.11: Structure of a bulk CMOS transistor

Ultrathin oxide breakdown [16] causes a sudden discontinuous increase in conductance often accompanied by an increased current noise, causing a reduction in the current of the transistor. Gradual oxide breakdown may initially lead to intermittent faults but may eventually cause a permanent fault in the device.

The breakdown is caused by gradual buildup of electron traps, which are oxide defects produced by missing oxygen atoms. The breakdown occurs when a statistical distribution of these traps is vertically aligned and allows a thermally damaging current to flow through the oxide. This is known as the *percolation* model of wearout and breakdown and the time to breakdown for a gate oxide can be expressed with the following equation:

$$T_{bd} = Ce^{\gamma(\alpha t_{ox} + \frac{E_a}{kT_j} - V_G)}$$

Where C is a constant, t_{ox} is the gate oxide thickness, Tj is the average junction temperature, Ea is the activation energy, VG is the gate voltage, and γ and α are technology dependent constants. Therefore, the time to breakdown decreases with decreasing oxide thickness but increases with decreasing V_G. This model is still an area of active research.

3.7. Hot Carrier Injection (HCI)

Hot Carrier Injection (HCI) [16] arises from impact ionization when electrons in the channel strike the silicon atoms around the drain-substrate interface. This could happen from one of several conditions, including a higher power supply or short channel lengths, among others. HCI results in a reduction of the maximum operating frequency of the chip.

The ionization produces electron-hole pairs in the drain as can be seen in Figure 3.12. Some of these carriers enter the substrate increasing the substrate current. A small fraction of these carriers may have sufficient energy to cross the oxide barrier and enter the oxide causing damage. Because these carriers have a high mean equivalent temperature, they are referred to as hot carriers. However, HCI becomes worse as ambient temperature decreases due to the corresponding increase in carrier mobility.

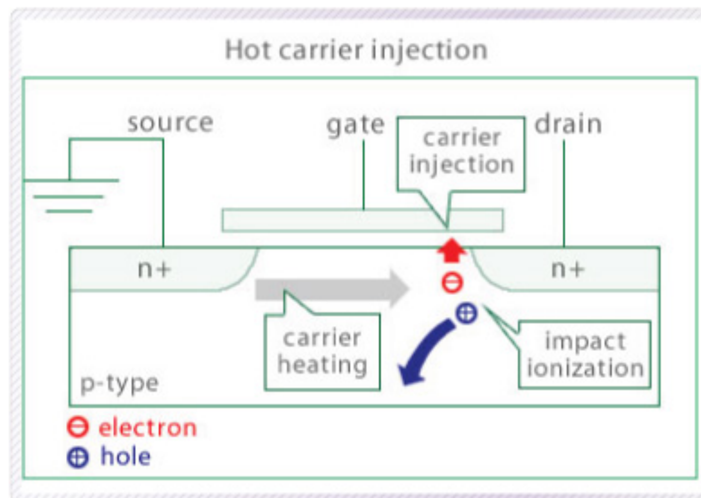


Figure 3.12: HCI Effect

The drain saturation current (I_{Dsat}) degradation is used to measure HCI degradation as is one of the key transistor parameters that most closely approximates the impact on circuit speed and because HCI damage occurs only when the transistor is in saturation.

Frequency guard banding is a typical measure to cope with HCI related degradation. The expected lifetime of a chip is often between 5 and 15 years, and the frequency degradation during the expected lifetime is between 1% and 10%. Hence, the chips are rated to run at a few percentage points below what they actually run at, calling this reduction as frequency guard band.

Transistor lifetime degradation (τ) due to HCI can be specified with the following equation:

$$\tau = Constant \frac{\frac{W}{I_D}}{(\frac{I_{sub}}{I_D})^{3'}}$$

Where W is the transistor width, I_D is the drain current, and I_{sub} is the substrate current. The I_D and I_{sub} parameters are estimated for the use condition of the chip.

3.8. NBTI/PBTI Aging

Negative Bias Temperature Instability (NBTI) [16], like HCI, causes degradation of the maximum frequency of the chip. However, while HCI can affect both nMOS and pMOS transistors, NBTI only affects short channel pMOS transistors. Under stress, like high temperatures, highly energetic holes bombard the channel-oxide interface, electrochemically react with the oxide interface, and release hydrogen atoms by breaking the silicon-hydrogen bonds. These free hydrogen atoms combine with oxygen or nitrogen atoms to create positively charged traps at the oxide-channel interface.

NBTI causes a reduction in mobility of holes and a shift in the pMOS threshold voltage towards the more negative direction. These effects cause the transistor drive current to degrade, slowing down the transistor device. The term “instability” refers to the variation of threshold voltage with time. There is active research to look for models that can predict how NBTI will manifest in future process generations.

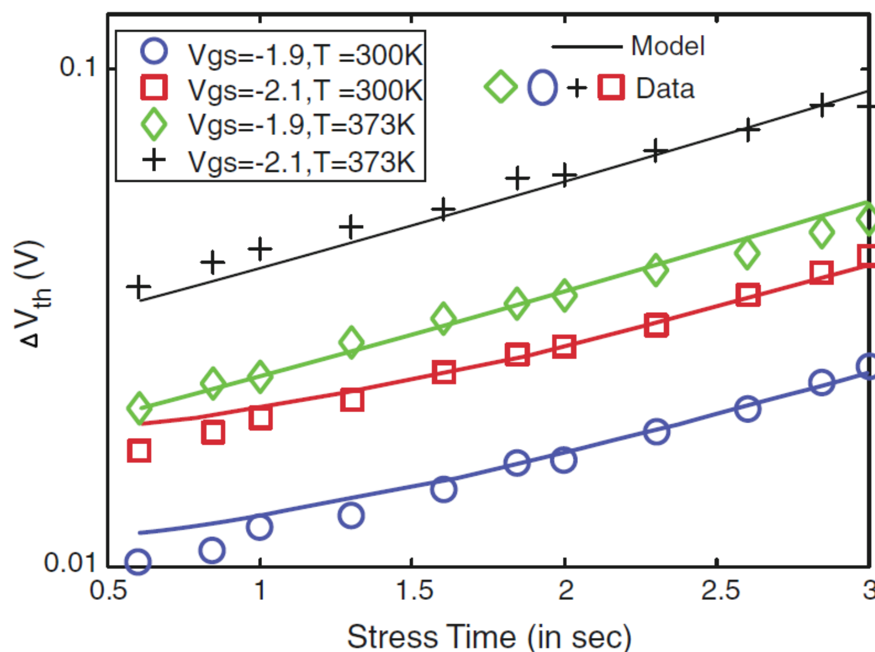


Figure 3.13: Vth degradation under static NBTI for different temperatures and Vgs for 90nm technology

Figure 3.13 shows V_{th} degradation under static NBTI for 90nm technology at different temperature and voltage conditions. NBTI shift recovers slightly after the stress condition is removed. There are some models for V_{th} shift that take account of recovery and dynamic stress.

For newer technologies using high-K dielectrics, nMOS devices suffer from a similar reliability problem due to Positive Bias Temperature Instability (PBTI).

3.9. Radiation Induced Faults (RIF)

Radiation induced transient faults [16][19] can be produced due to different types of sources: alpha particles from packaging and neutrons from the atmosphere. Most of the faults described in this chapter can be taken care before a chip is shipped. In contrast, radiation faults are addressed with fault detection and error correction circuitry.

An alpha particle consists of two protons and two neutrons bound together into a particle. Alpha particles are emitted by radioactive nuclei, such as uranium or radium, in a process known as alpha decay. Alpha particles have kinetic energies of a few MeV, which is lower than those of neutrons that affect CMOS chips. Nevertheless, alpha particles can affect semiconductor devices because they deposit dense track of charge and create electron-hole pairs as they pass through the substrate. Alpha particles can arise from radioactive impurities used in chip packaging such in the solder balls or contamination of semiconductor processing materials. Alpha particles are difficult to eliminate completely from the chip so chips need fault detection and error correction techniques.

The neutron is one of the subatomic particles that make up an atom. Atoms are considered the basic building blocks of matter and consists of three types of subatomic particles: protons, neutrons and electrons. A proton is positively charged, a neutron is neutral and an electron is negatively charged. An atom consists of an equal number of protons and electrons and hence it is neutral itself. The neutrons that cause soft errors arise when atoms break apart into protons, electrons and neutrons. Protons have a long half-life so can persist for long durations before decaying and constitute the majority of the primary cosmic rays that bombarded the earth's outer atmosphere. When these protons and associated particles hit atmospheric atoms, they create a shower of secondary particles named secondary cosmic rays. Untimely, the particles that hit the earth's surface are known as terrestrial cosmic rays.

Alpha particles and neutrons slightly differ in their interactions with silicon crystals. Charged alpha particles interact directly with electrons. In contrast, neutrons interact with silicon via inelastic or elastic collisions. Inelastic collisions cause the incoming neutrons to lose their identity and create secondary particles, whereas elastic collisions preserve the identity. Inelastic collisions cause the majority of the soft errors due to neutrons.

When an alpha particle penetrates a silicon crystal, it causes strong field perturbations, creating electron hole-pairs in the substrate of a transistor. The electric field near the p-n junction, the interface between the bulk and diffusion, can be high enough to prevent the electron-hole pairs from recombining. Then, the excess carriers could be swept into the diffusion regions and eventually to the device contacts, registering an incorrect signal.

One key concept to explain the interaction of alpha particles with silicon is the stopping power. Stopping power is defined as the energy lost per unit track length, which measures the energy exchanged between an incoming particle and electrons in a medium. Stopping power quantifies the energy released from an interaction between alpha particles and silicon crystals, which in turn can generate electron-hole pairs. About 3.6 eV of energy is required to create one such pair. Whether the generated charge can actually cause a malfunction or a bit flip depend on two factors named charge collection efficiency and critical charge of the circuit that will be explained later.

Neutrons do not directly cause a transient fault because they do not directly create electron hole-hole pairs in silicon crystals (their stopping power is zero). Instead, these particles collide with the nuclei in the semiconductor resulting in the emission of secondary nuclear fragments. These fragments could consist of particles such as pions, protons, neutrons, deuteron, tritons, alpha particles and others. These secondary fragments can cause ionization tracks that can produce a sufficient number of electron-hole pairs to cause transient faults in the device. The probability of a collision that produces these secondary fragments is very small so a greater

number of neutrons is necessary than alpha particles to produce the same number of transient faults.

Stopping power explains why and how many electron-hole pairs may be generated by an alpha or a neutron strike, but it does not explain if the circuit will malfunction. The charge accumulation needs to cross a certain threshold before an SRAM cell, for example, will flip the charge stored in the cell. This minimum charge necessary to cause a circuit malfunction is termed as the critical charge of the circuit represented as Qcrit. Typically, Qcrit is estimated in circuit models by injecting different current pulses till the circuit malfunctions.

Hazucha and Svensson [18] proposed the following model to predict neutron induced Soft Error Rate (SER):

$$Circuit\ SER = Constant\ x\ Flux\ x\ Area\ x\ e^{-\frac{Qcrit}{Qcoll}}$$

Constant is a constant parameter dependent on the process technology and circuit design style, Flux is the flux of neutrons at the specific location, Area is the area of the circuit sensitive to soft errors, and Qcoll is the charge collection efficiency, which is the ratio of collected and generated charge per unit volume. Qcoll depends strongly on doping and Vcc and is directly related to the stopping power, so the greater is the stopping power, the greater is Qcoll. Qcoll can be derived empirically using either accelerated neutron tests or device physics models, whereas Qcrit is derived using circuit simulators. This equation can also be used to predict the SER of alpha particles. Figure 3.14 shows a diagram illustrating the effects of soft errors.

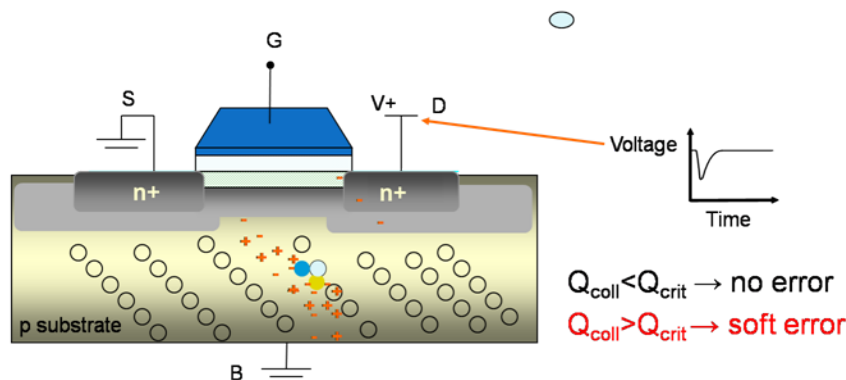


Figure 3.14: Diagram of soft errors effects

With every process generation, the area of the same circuit goes down, so this should reduce the effective SER from one process generation to the next. However, Qcrit also decreases because the voltage of the circuit goes down across process generations. Therefore, for some elements like latches and logic, this effect appears to cancel each other out, resulting in a constant SER across generations. However, if Qcrit is sufficiently low, such in SRAM devices, then the impact of the area begins to dominate. This is referred as saturation effect, where the SER decreases with process generations. However, the circuit is highly vulnerable to soft errors in the saturation region. In the extreme case, as Qcrit approaches to zero, almost any amount of charge produced by alpha or neutron strikes will result in a transient fault.

When a charge produced by an alpha particle or neutron strike is sufficient to overwhelm a circuit, then it may malfunction. At the gate or cell level, this malfunction appears as a bit flip. For storage devices, when a bit residing in a storage cell flips, a transient fault is said to have occurred. For logic devices, a change in the value of the input node feeding a gate or output node coming out of a gate does not necessarily mean a transient fault has occurred. Only when this fault propagates to a forward latch or storage cell does one say a transient fault has occurred.

3.10. SOI Self-Heating (SHE)

Silicon on insulator (SOI) [20] technology possesses some advantages over bulk silicon technology such as the reduction of parasitic capacitance, excellent, sub-threshold slope, elimination of latch up and resistance to radiation. Hence, it is preferred for high speed, high temperature and low power devices by some manufacturers.

SOI MOS devices employ a buried insulating thin layer usually made of silicon dioxide to electrically isolate the devices from the bulk of the semiconductor. Due to the poor conductance of SiO2, the buried dielectric layer also thermally insulates the MOSFETs from the bulk. Consequently, the heat generated in the SOI MOSFETs causes a larger temperature rise than in bulk devices under similar conditions, and the self-heating effect that results in reduced carrier mobility and corresponding decrease in the drain current transconductance and speed becomes an inherent issue for MOSFETs built in SOI. As the device geometries diminish and transconductance as well as current density increase with MOS scaling, the self-heating effect becomes more pronounced. There are some theoretical models to evaluate the effect of self-heating in SOI which are used by some simulators. Figure 3.15 and Figure 3.16 shows the effect of SOI self-heating with the ATLAS simulator.

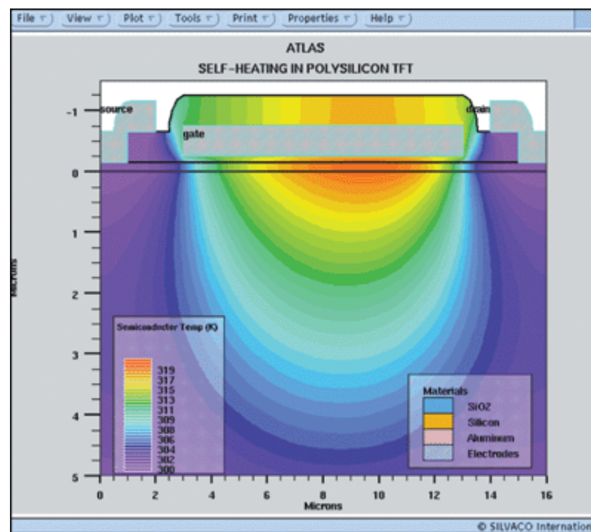


Figure 3.15: Self-heating in SOI transistors [21]

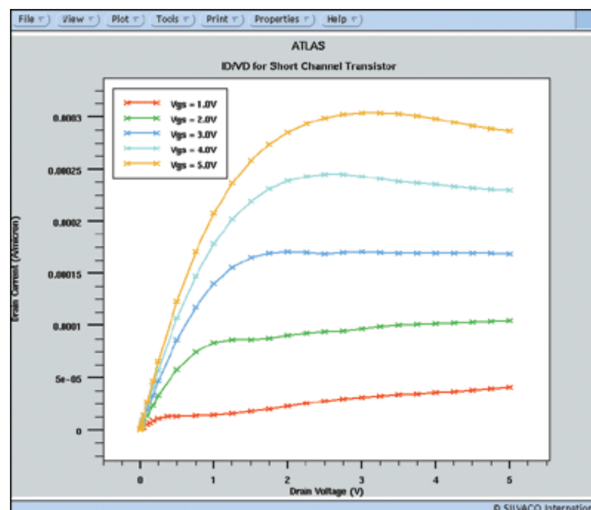


Figure 3.16: Effect of Self-heating on output characteristics [21]

3.11. Other Sources

RDF and LER are currently dominant sources of process variations but there are several other sources, which may become important for future technologies. Below there is a list of other sources of variations:

- **Oxide Charges Variation**: Interface charges can also cause V_{th} variations that may be significant with the recent adoption of high-K gates.
- **Mobility Fluctuation**: Variations in a transistor's drive current can be caused by mobility fluctuations. Mobility fluctuations can arise from several complex mechanisms such as fixed oxide charges, doping or inversion layer, among others.
- **Gate Oxide Thickness Variation**: Any variation in oxide thickness affects many electrical parameters, especially V_{th}.
- **Channel Width Variation**: Due to lithography limitations, transistor channel width also varies similarly to LER variations. Width variations can cause V_{th} variations, but as W is 2-4 times larger than L, its impact on V_{th} is smaller than the impact due to L variation.

3.12. Sources of Failure mapped with technologies

The described sources of failures can affect different technologies in different ways. For example, RDF and LER are critical for CMOS SRAM cells while FinFETs are more resistant to RDF but adds fin thickness variations [22], even some of them may only affect specific technologies. In Table 3.1, there is a summary of the sources of failure described, with their type, and the technologies that may be most affected by these failures.

Sources	Fault Type	Technology
Random Dopant Fluctuations (RDF)	Permanent	All
Line Edge Roughness (LER)	Permanent	All
Random Telegraph Noise (RTN)	Intermittent	All
Metal Stress Voiding (MSV)	Permanent	All
Electromigration (EM)	Permanent	All
Hot Carrier Injection (HCI)	Intermittent/Permanent	All
Gate Oxide Wearout (GOW)	Intermittent/Permanent	All
NBTI/PBTI Aging	Intermittent/Permanent	All
Radiation Induced Faults (RIF)	Transient	All
Self-Heating (SHE)	Intermittent/Permanent	SOI

Table 3.1: Sources of failure mapped with technologies

4. Characterization of Different Sources of Failure

In this chapter, we present the characterization of the previously described sources of failure. This characterization includes data that will be used within the project. Before that, some general considerations are made below.

The circuit components being tested in this project have been previously listed in Table 2.1 of chapter 2. These components have been modeled and analyzed with SPICE. For this purpose, we developed a description of the necessary circuits at transistor level and use a predictive technology model (PTM) of the technology node be analyzed.

Depending on the circuit, some transistors may need to be resized for correct operation or better performance. In the case of Planar CMOS, the resizing means to specify the length and the width of the transistor in lambdas or nanometers. Examples of most circuits for 32nm or higher technology nodes can be found in the literature. These examples have been used as starting point and then linearly scaled down at the technology nodes that we want to analyze. In a similar way, the transistors for FinFETs have been resized in terms of number of fins from a starting point taken from the literature.

Environmental factors can also impact the characteristics or behavior of a source of failure. Table 4.1 shows different environmental factors and describes how these factors impact on the different types of errors. In this project, we have used a variety of temperatures and voltages to take into account some of these factors.

Factors	Impact on transient errors	Impact on Intermittent errors	Impact on Permanent errors
Temperature	Increase in transient failures with higher temperatures due to higher energetic particles and increased leakage	Increase in intermittent failures due to device degradation (e.g. NBTI effects) and thermal stress	Increase in permanent failures due to device degradation effects (e.g. Electromigration effects) and thermal stress(e.g. wear out effects)
Humidity / Dust / Acid / Salt	N/A	N/A	Increase in permanent failures due to corrosion/shorting on contacts
Vibration / Shock / Pressure / Gravity / Explosion	N/A	May cause intermittent failures depending on the strength of the effect	Increase in permanent failures due to mechanical stress and contact/solder breaks
EMC / EMI / Radiation / Altitude	Increased soft errors due to increased interferences (e.g. IR effects, magnetic storage technologies)	May cause intermittent failures for unshielded components that last throughout the exposure period (e.g. solar EMP)	Oxide failure or metal melt due to ESD; power surges due to HEMP and HPM; device degradation effects (Total Ionizing Dose) and destructive effects (Single-Event Latch-Up)

Table 4.1: Environmental factors and their effects on different types of errors

The rest of this chapter is mainly focused on the characterization of soft errors, since their impact on the reliability of new systems is increasing and are becoming a major concern in the industry. We also show a worst case analysis that combines soft errors and variability due to aging. Below we report the studies performed for the circuits and technologies mentioned in chapter 2.

4.1. Methodology to Characterize Soft Errors

This section describes the methodology used to compute Soft Error Rates (SER) for different hardware blocks and technologies. First of all, we do an evaluation of the most important models and methods to compute the SER. Then, we justify our decision of developing our own methodology based on some of these models. Finally, our own methodology and the tools needed are described. Before entering in the methodology, some general considerations are made below.

As described in chapter 3.9, for an alpha particle or a neutron to cause a soft error, the strike must flip the state of a bit. Whether the bit flip eventually affects the final outcome of a program depends on whether the error propagates without being masked, and whether there is some error detection and correction scheme. Architecturally, the error detection and correction mechanisms create two categories of errors: Silent Data Corruption (SDC) and Detected Unrecoverable Error (DUE) [16].

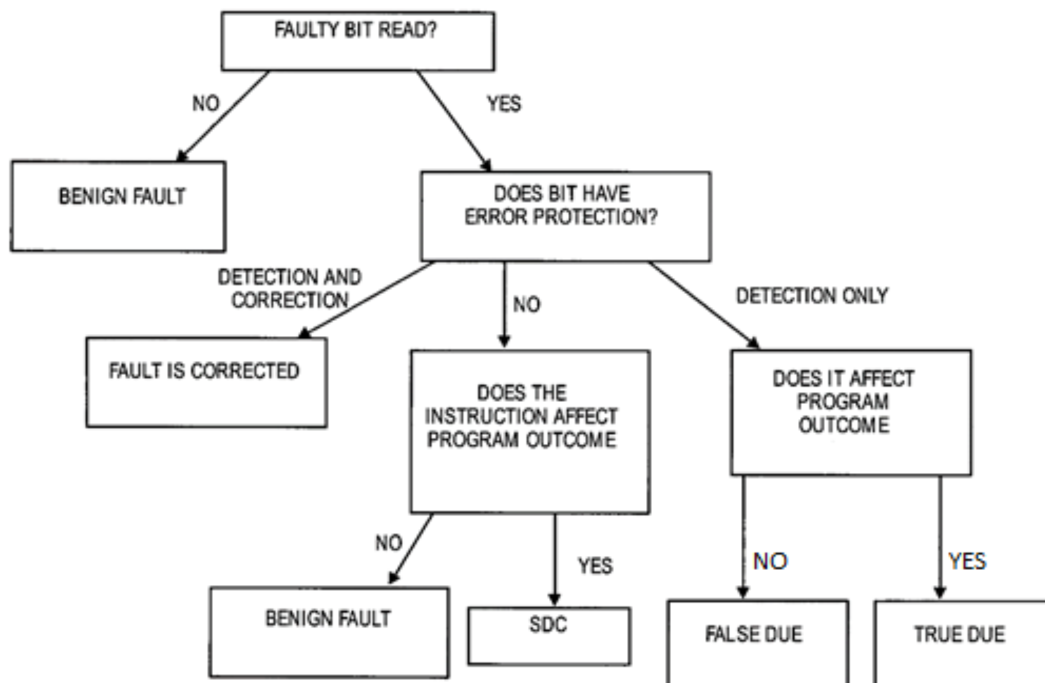


Figure 4.1: SDC and DUE Scheme

Figure 4.1 shows the different outcomes of a bit flip. The most insidious form of error is SDC since a fault induces the system to generate erroneous outputs. SDC rates can be expressed as either Failure in Time (FIT) or Mean Time to Failure (MTTF). FIT rates are the number of failures in one billion (10^9) device-hours of operation while MTTF describes the expected time to failure for a non-repairable system.

To avoid SDC, designers use basic error detection mechanisms, such as parity. The ability to detect a fault but not correct it avoids generating incorrect outputs, but prevents from finalizing the task. Therefore, simple error detection does not reduce the overall error rate but provides fail-stop behavior and avoids data corruption. Errors in this category are called DUE, and it can also be quantified using FIT and MTTF. DUE events are further divided according to whether the detected fault would have affected the final outcome of the execution or not, calling them true and false DUE respectively. In following sections, SERs are expressed in FIT rates.

4.1.1 Modeling Circuit Level Soft Error Rates (SER)

Computing the SER of a microprocessor requires the analysis of two areas: the raw SER of the circuits comprising the chip (technology vulnerability) and the corresponding derating factors [16]. Computing the raw SER of a circuit element is generally done in a two-step process: first one must compute the critical charge (Qcrit) that the charge released by a neutron strike must overcome to cause a malfunction. Thereafter, the Qcrit must be mapped to a corresponding SER for the circuit element. The general procedure to compute the SER applies to memory elements, latches and logic gates.

Once the raw SER is computed, it needs be derated by a variety of vulnerability factors. For example, if a latch is not vulnerable 50% of the time, then the raw SER needs to be multiplied by 0.5 to compute the derated SER. Later in this chapter, a description of such vulnerability factors and masking effects and how are they taken into account in our results is included.

4.1.2 Critical Charge (Qcrit)

An alpha particle or a neutron strike typically manifests itself as a transient disturbance that would usually last less than 100 picoseconds. If this charge disturbance is smaller than the noise margin, the circuit will continue to operate correctly. Otherwise, the disturbed voltage may invert the logic state.

Figure 4.2 shows an SRAM cell made of a pair of cross-coupled inverters. When the wordline is low, the cell holds data in the inverters and the bitlines are decoupled. If a particle strike causes one of the sensitive nodes to transition, then the disturbance may propagate through the inverter and cause a transient disturbance on the second sensitive node. This will cause the second node to propagate the incorrect value, thereby causing both nodes to flip. This results in flipping the state of the bit held in the SRAM cell. Other circuit elements, such as register files, latches and logic gates, are affected in similar ways by particle strikes.

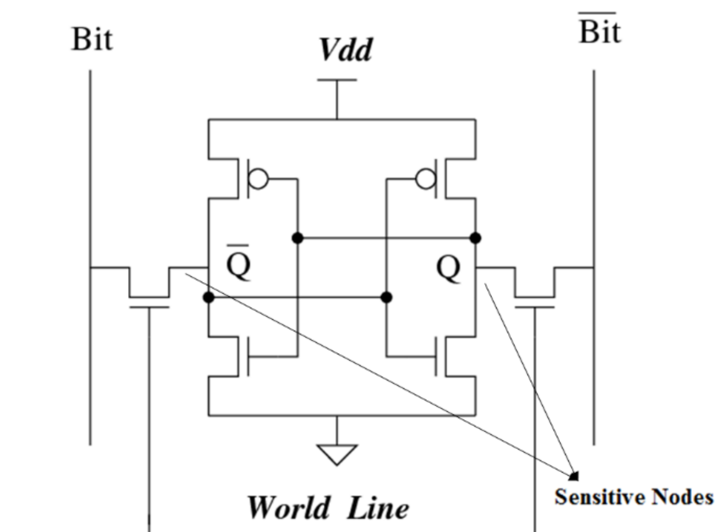


Figure 4.2: A transistor-level diagram of an SRAM cell

Critical charge (Qcrit) [16] is defined as the minimum charge that must be deposited by a particle strike to cause a circuit malfunction. Qcrit is usually computed using integrated circuit simulators, such as SPICE, by injecting current pulses into the sensitive nodes of a circuit as can be seen in Figure 4.3.

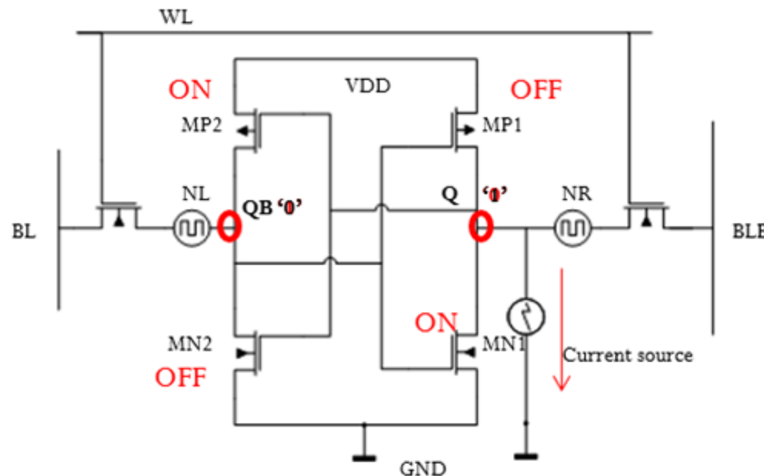


Figure 4.3: Current pulse injected in a 6T SRAM Cell sensitive node

The current pulses represent the current generated from electron-hole pairs created by a neutron strike. The smallest charge corresponding to an injected current pulse that inverts the state of a circuit element is the Qcrit of the circuit. However, there are many factors that impact the critical charge [23]. Because *charge = capacitance x voltage*, Qcrit depends on the supply voltage. Qcrit is also weakly dependent on temperature and strongly dependent on the shape of the current pulse injected.

The pulses in general have a rapid rise followed by a slow decay, and are characterized by their time constants. A circuit which recovers quickly from a disturbance may have a lower Qcrit for a spike of current than for a slower pulse. A high number of current models have been proposed in the literature [24] over the years and they are used to characterize Qcrit by performing SPICE simulations. The most common pulses are:

- **Roche Model**: Qcrit can be found by integrating an exponentially decaying current ($I_0 . exp(-\tau)$) with small time constants which are less than 20ps.
- **Diffusion Model**: Qcrit can be found with a diffusion collection model where t_{max} represents the instant when the maximum value of the current is reached, and it can be represented by the following equation:

$$I(t) \ = \ I_{max} \, [exp(t_{max}/t)]^{3/2} \, [exp(-3t_{max}/2t)]$$

- **Freeman Model**: Current is defined in terms of total charge deposited (Q) by the ion and a single timing parameter τ by the following equation:

$$I(t) \ = \ (2/\sqrt{\pi}).(Q/\tau).(\sqrt{(t/\tau)}).exp(-t/\tau)$$

- **Double Exponential Model**: The most commonly used model by the community is a double exponential pulse with two timing parameters representing the rising and falling time constants of the exponentials. The following equation is used:

$$I(t) = (Q/(\tau f - \tau r) \, [exp(-t/\tau f) \ - \ exp(-t/\tau r)]$$

Figure 4.4 shows a plot with an example of each of these current pulses. The current pulse rise and fall times strongly affect the characterization of Qcrit, to the point where each pulse model results in its own Qcrit value.

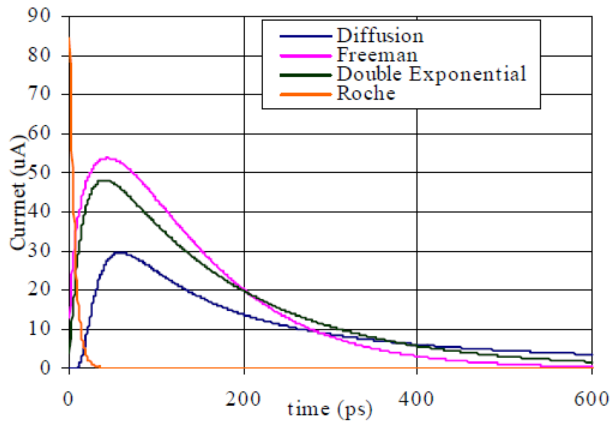


Figure 4.4: Current Pulse Profiles

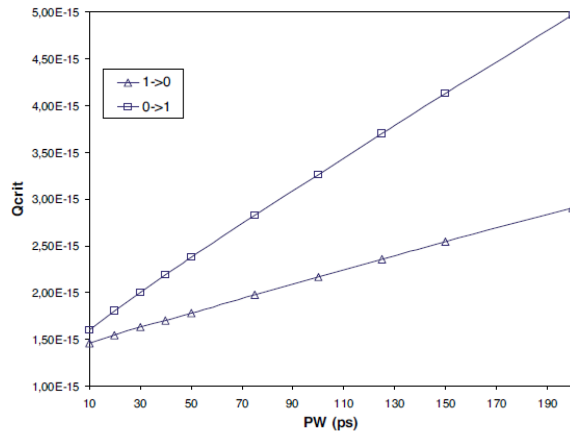


Figure 4.5: Pulse Width dependence of Qcrit

Another factor that strongly affects the value of Qcrit is the pulse width, as can be seen in Figure 4.5, which determines the range of the integral from where Qcrit is computed [25]. Some empirical approximations have been used in the literature to select values for these parameters. However, there is not a unique way to make the computation of Qcrit. Therefore, multiple voltages, temperatures, types of current pulses and parameters for these pulses can be tested. Section 4.1.6 describes how we compute the Qcrit and which parameters are used.

4.1.3 Mapping Qcrit to SER

Once Qcrit is computed for a specific circuit element, it needs to be mapped into a SER expressed in FIT. This mapping can be derived by combining physics-based models and experimental data. There are different models and methods to do this mapping. Three of these models are especially relevant as the rest are based on them by adding extensions or adjusting parameters, and there is also the option to use model simulations [16]. These models and methods are described below:

- **Hazucha and Svensson Model**: One can start from an equation such as the one proposed by Hazucha and Svensson [18]:

  $$Circuit\ SER = Constant\ x\ Flux\ x\ Area\ x\ e^{-\frac{Qcrit}{Qcoll}}$$

  Flux is the neutron flux experienced by the circuit, *Area* is the effective diffusion area, and Qcoll is the collection efficiency. The parameters of the equation (e.g., *Constant*, *Qcoll*) can be derived empirically using accelerated tests. Such empirical mapping is a popular method to compute the SER of CMOS circuits. However, the equation must be calibrated for each new technology generation.

- **Burst Generation Rate (BGR) Method**: The BGR method proposed by Ziegler and Lanford [26] is based on two key parameters: the sensitive volume (SV) and neutron-induced recoil energy (E-recoil). An upset is said to occur if the burst of charge generated by neutron-silicon interactions within the SV of a device is greater than *Qcrit*. E-recoil is expressed as:

  $$E - recoil = Qcrit\ x\ 22.5,$$

  Then, the upset rate is computed as:

  $$Upset\ rate = Qcoll\ x\ SV\ x \int_{E-neutron} \left(BGR(E - neutron, E - recoil) \frac{dN}{dE} \right) dE$$

Where *dN/dE* is the differential neutron flux, *E-neutron* is the neutron energy, the BGR function is the energy deposited in silicon by neutron interactions, and *Qcoll* is the collection efficiency. Empirical heavy ion testing is used to obtain and tabulate the BGR values and the integration is performed numerically using the experimental BGR data.

- **Neutron Cross-Section (NCS) Method**: To compute the device upset rate using the BGR method, one must compute the SV of the device, which is often difficult to compute. Instead, the NCS method proposed by Taber and Normand [27] avoids the use of the SV parameter (as well as Qcrit), by correlating the neutron environment parameters, such as flux and energy, with the device upset rate. NCS expresses the upset rate as:

$$Upset\ rate = \int_{E-neutron} (\sigma \frac{dN}{dE}) dE$$

This equation replaces Qcoll, SV and the BGR function, with a single variable denoting the neutron cross section. The neutron cross section is defined as the probability that a neutron with energy *E-neutron* will interact and produce an upset. These probabilities are generated for specific device types using accelerated neutron tests.

- **Simulation Models**: Murley and Srinivasan proposed to model the charge collection phenomenon simulating neutron strikes from first principles [28]. In cases where simulations result in a collected charge greater than Qcrit, the circuit is assumed to malfunction. This gives the probability of an upset given a certain neutron flux, and it can be easily converted into FIT rate. However, this methodology requires a detailed knowledge of the process technology and how that it interacts with neutrons.

Soft error models must be calibrated and validated with measurements. Because soft errors typically occur once in several years in a single chip, the occurrence of errors needs to be accelerated to measure them in a short period of time. This can be accomplished either by collecting data from numerous chips and computers or by increasing the flux of the generated alpha particles and neutrons. For neutrons, the accelerated neutron tests can be performed in particle accelerators. Thus, soft errors can be captured easily by exposing the test chips to a neutron beam. All these models have advantages and disadvantages, Table 4.2 summarizes them:

Model	Pros	Cons
Hazucha and Svensson	Qcrit and Area can be easily computed Popular and widely used in the literature	Constant and Qcoll derived empirically
Burst Generation Rate (BGR)	Qcrit can be easily computed	Qcoll and BGR require empirical tests SV difficult to compute without good knowledge of the chip's layout
Neutron Cross-Section (NCS)	Sensitive Volume not required	Requires experimental tests Probabilities for specific devices Qcrit not used
Simulation Models	Once the probability of an upset is obtained it can be easily converted into FIT	Requires detailed knowledge of the technology and its interaction with neutrons

Table 4.2: Pros and Cons of each model

In our methodology, the Hazucha and Svensson model is used because it is the most common in SER studies [3][29] and it has been validated with experimental data [18][15]. Moreover, most of the required parameters can be computed with the tools and resources that we have. Qcrit can be obtained with SPICE simulations and the Area can be easily computed since the dimensions of the transistors are specified. Moreover, we can deal with the parameters derived empirically scaling them as is described later in this chapter. The other methods require de-

tailed knowledge in fields that are out of our specialization and tools that are out of our possibilities.

4.1.4 Neutron Flux

The reference neutron flux commonly used in the SER computation is from New York City at sea level. However, neutron flux depends on the location and is mainly affected by two parameters: Altitude and Vertical Cutoff [31]. Neutron flux increases exponentially with the altitude while the vertical cutoff is a parameter of the magnetic field of the earth, which depends on the coordinates as can be seen in Figure 4.6. The earth magnetic field maximum is in the poles while the minimum is in the equator. Therefore, the neutron flux decreases when approaching the equator and increases in the poles. The neutron flux also depends on the solar activity.

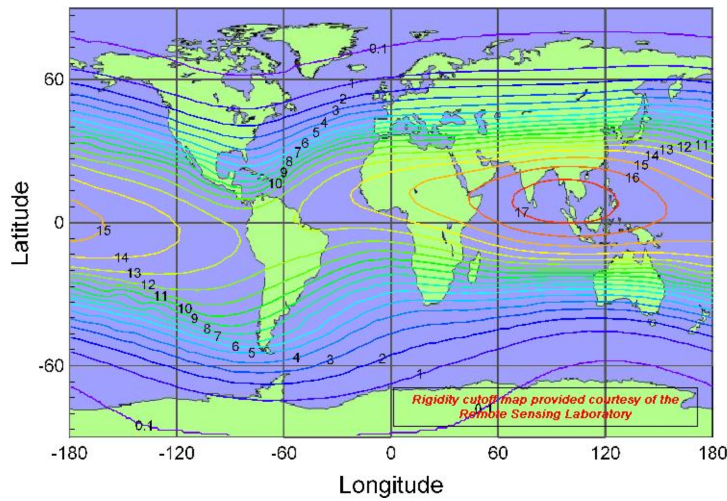


Figure 4.6: Vertical Cutoff Map

There are two main ways to compute the flux considering the location. First one involves using the methodology described in Annex A of the JEDEC standard [32]. Alternatively, one can use the online calculator from [33] which is compatible with the JEDEC standard and outputs the flux relative to the flux from NYC. The second method involves the use of a model tested and corrected with empirical data, which has been proposed by Gordon, et al. [34], and it has the following high-level form:

$$F = F_{ref} \; x \; F_{alt}(d) \; x \; F_{BSYD}(Rc, \; d, \; I)$$

Where F_{ref} is the flux at a reference location (i.e.: Flux of New York City at sea level), F_{alt} is the function describing the dependence on altitude, F_{BSYD} is the function describing the dependence on geomagnetic location and solar activity, d is the atmospheric depth, Rc is the vertical cutoff and I is the relative count rate of a neutron monitor measuring solar modulation. Both ways are equally good and they can be used to obtain a relative flux that can be directly multiplied by the SER computed with the reference flux. Table 4.3 shows the flux in some locations of the United States:

Locations	Altitude (m)	Cutoff (GV)	Relative Flux	Total Flux
Fremont Pass, CO	3450	2,94	12,58	0,07
Leadville, CO	3150	2,97	9,56	0,05
Mt. Wash., NH	1905	1,58	4,70	0,027
Yorktown Hts., NY	167	2	1,20	0,007
Houston, TX	14	4,68	0,91	0,005

Table 4.3: Neutron Flux in USA Locations [34]

Total flux has been obtained experimentally by Gordon, et al. and then fitted to their model [34]. Table 4.4 shows the neutron flux of different coordinates and altitudes:

Coordinates	Altitude (m)	Cutoff (GV)	Relative Flux	Total Flux
19N, 127W	20300	12	217,07	1,28
54N, 117W	20000	0,8	1495,42	10,2
56N, 121W	16200	0,7	1070,18	10
38N, 122W	11900	4,5	301,16	3,4
37N, 76W	0	2,7	0,99	0,0122

Table 4.4: Neutron flux at high altitudes [35]

In this case, total flux has been obtained by measurements aboard an ER-2 high-altitude airplane [35]. The flux observed increases between 200x-1500x, and the effect of the vertical cut-off can be observed as the first location has a big cutoff and the flux is reduced around 7x compared with the second location which is at a similar altitude but has a low cutoff. In chapter 6.5, some examples of relative fluxes for different locations are given using the online calculator, including a SER example.

4.1.5 Time Vulnerability Factor and Masking Effects

Once the raw SER of a circuit is computed, it must be derated by the appropriate vulnerability factors to compute the circuit-level SER [16]. Timing Vulnerability factor (TVF) is the fraction of time a circuit is vulnerable to upsets. An SRAM cell usually has a TVF of 100% because any strike during a clock cycle can change the value stored in the SRAM cell. However, flip-flops and latches are clocked elements and have a TVF less than 100%.

Figure 4.7 shows a latch and its corresponding timing diagram. When the clock transitions from high to low the data at input D is latched. During the low phase of the clock, the latch is in the hold mode, maintaining the value at the output Q. The storage nodes of the latch are vulnerable to soft errors when the latch is holding data at the low phase of the clock. When the clock phase is high, the latch is in transparent mode driving data to the next stage, and is able to recover from a particle strike. Consequently, latches TVF is roughly 50% (half of the clock).

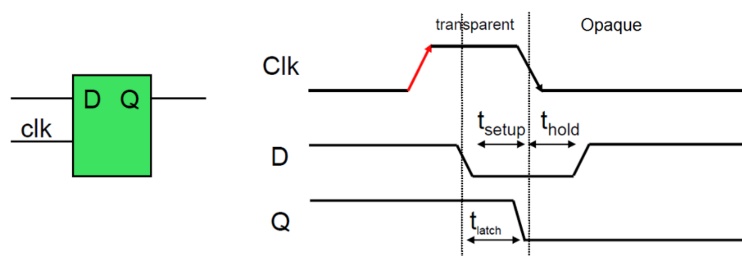


Figure 4.7: Latch Timing Diagram

In modern microprocessors, latches start to driven data during the hold mode so upsets must occur early in the low clock phase for the signal to propagate to the next element. Hence, TVF is usually smaller than 50% and also depends on different components, besides propagation delay, such as setup time, clock rise and fall time.

Logic gates are the building blocks of modern silicon chips. A malfunction due to a particle strike in one logic gate must reach and be captured in the forward memory element for the malfunction to cause an error. Otherwise, the effects are masked. Thus, evaluating the SER of a

logic gate consists of evaluating the Qcrit of each gate, mapping the Qcrit to the appropriate SER and evaluating if the fault introduced in the gate will be masked or reach the forward latch.

In today's microprocessors, more than 90% of the radiation induced faults in logic gates can be masked. Nevertheless, faults in logic gates cannot be ignored for three main reasons. First, modern microprocessors are composed of tens to hundreds of millions of logic gates. Second, the masking effects decrease with new technology generations. Third, it is more difficult to protect logic gates compared to SRAM cells because ECCs are difficult to implement for logic blocks. There are three kinds of masking commonly observed in logic blocks:

- **Logical Masking**: A strike can be logically masked if it affects a portion of the circuit that does not logically affect the final outcome of the circuit.
- **Electrical Masking**: A strike can be electrically masked if the pulse created by the strike attenuates before it reaches the forward latch.
- **Latch-Window Masking**: A strike can also be masked if the resulting pulse does not reach the forward latch at the clock transition where the latch captures its input value.

To accurately compute the SER of logic blocks, it is essential to model each of these masking effects. Electrical masking and latch-window masking can be taken into account at the technology layer, but for logical masking is required to know the function of the circuit and occurs one layer above. Consequently, all these effects are integrated in the upper layer of the project.

4.1.6 Evaluation Framework and Tools

After considering the main models to compute the raw SER and their parameters, we defined our methodology, which follows the workflow of Figure 4.9. As we target an exhaustive design space characterization, we wrote a python script for each component that is analyzed. Each script defines a collection of loops to simulate an element with a variety of configuration parameters, such as temperatures and voltages, and different technology models. In the inner loop, a function call is made. This function defines another loop to iterate the current injected with the pulse until a flip or glitch is detected measuring the stored value (SRAM) or the output (Logic Gates). An example of one script is shown in the following pseudocode:

```
for all Technologies do
  for all Voltages do
    for all Temperatures do
      for all Inputs or Stored values do
        for all Sensitive Nodes do
          Current = 0;
          Flipped = 0;
          while not flipped do
            Increase Current Injected;
            Generate SPICE Files;
            Simulate Element in HSPICE;
            Read Simulation Results;
            if Flip Detected then
              Flipped = 1;
              Write Results in Database;
            Clean Simulation Files;
End all loops
```

Figure 4.8: Script Pseudo-code

To make the SPICE simulations, HSPICE [36], which is a commercial circuit simulator from Synopsis, is invoked in a subprocess. The charge generated from a pulse that causes a malfunction is

stored and defined as the Qcrit of that element in a specific state. Finally, for each Qcrit, a raw SER is computed using the model in [18] and stored into an Excel file, including the parameters that define the state.

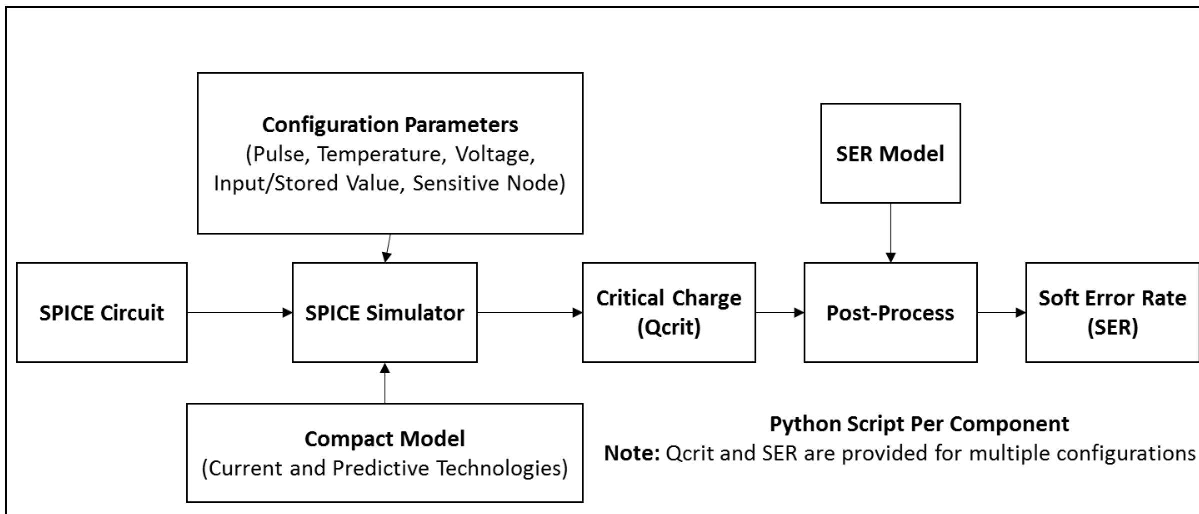


Figure 4.9: Workflow schema

As it has already been commented, there are many factors that affect the Qcrit. Because of that, we decided to test a variety of parameters and compute a Qcrit for each combination. Voltage ranges from 0.7V to 1.2V which can be used to distinguish between high performance and low power processors. Temperatures tested include 25, 50, 75 and 100 C° which can be used to map idle, typical and extreme conditions. Stored values 0 and 1 have been tested for SRAM cells, and for logic gates all the input combinations have been analyzed. Moreover, each element may have more than one sensitive node so all nodes are considered.

A double exponential pulse is used since HSPICE only has this type. The shape of the current pulse also strongly affects Qcrit. For that reason, multiple rise time constants used in the literature (2ps, 16ps, 33ps and 90ps) have been tested but maintaining a falling time constant of 200ps [24][25]. Pulse width also has a strong effect on Qcrit affecting the integral range. Looking at the literature, there is not a clear way to define the pulse width so we decided to define it from the start of the pulse until the pulse decreases an 80% of its maximum which represents the spike of the pulse. Then, Qcrit is computed by doing the integral of the current pulse in that range as can be graphically seen in Figure 4.10.

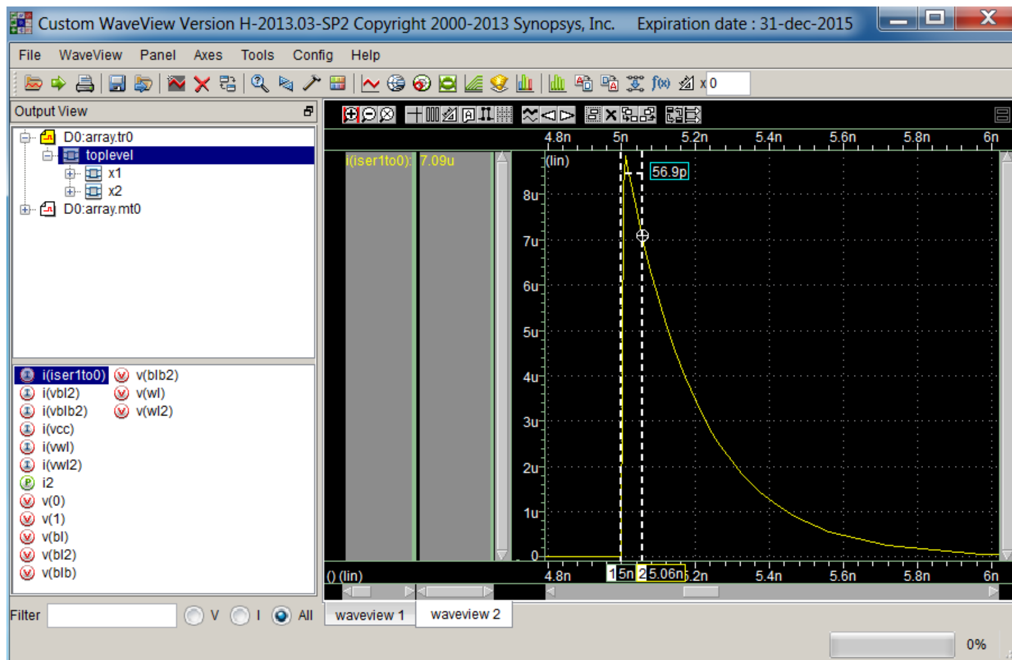


Figure 4.10: Qcrit Measurement

SER is computed using the Hazucha and Svensson model [18]:

$$SER_{raw} \; \alpha \; Constant \; x \; Flux \; x \; Area \; x \; e^{-\frac{Qcrit}{Qs}}$$

The area sensitive to neutron strikes is the drain area of the transistors which is defined in the SPICE circuits, so it can be easily obtained. The constant is a technology independent parameter which was computed by Hazucha and Svensson and it has a value of $2.2*10^{-5}$. The exponential part of the formula is the technology vulnerability factor (TVF). If the charge collected (Qcoll) by a particle is greater than Qcrit a soft error is produced. Charge Collection Efficiency (Qs) is the mean of Qcoll in a range of energy particles and a parameter dependent of the technology which is usually computed experimentally. However, Qs scales approximately linear with the Length Gate (Lg), so Qs has been scaled down with a linear regression from experimental data [18] for CMOS technology. In the case of newer technologies, a study of how Qcoll changes has been done, and an approximate technology factor has been extracted from previous works [37][38][39].

We can also combine the previous formula with the neutron flux model from Gordon to compute the neutron flux dependent of the location:

$$F = F_{ref} \; x \; F_{alt}(d) \; x \; F_{BSYD}(Rc, \; d, \; I)$$

Concluding this section, multiple SER values are obtained for each state, which is represented by the combination of parameters. However, the SER of the circuit or element is the sum of the SER from all sensitive nodes [29]. Therefore, SERs from different sensitive nodes but same conditions are summed. For example, a 6T SRAM cell has two sensitive nodes which are symmetric. Therefore, the SER of the cell can be computed as the sum of the SER of one node storing a 1 and the SER of the other node storing a 0. Then, depending on the element, SERs are derated by a timing factor, such as the latch where a factor of 50% is applied. Finally, a weighted average can be done with the SERs of different states to give a unique SER for the element. This is the case of logic gates where the SER can be averaged by the SERs of the different inputs, but still there will always be multiple SERs for the different voltages, temperatures and current pulses.

4.2. Multi Cell Upsets (MCU) Model

Multi Cell Upset (MCU) events consist in flipping the value of multiple SRAM cells, latches or logic gates from one single strike. MCU effect occurs when the charge cloud produced from one strike in an element is large enough to affect the elements that are near. MCU becomes worst when technology shrinks since elements are closer. Therefore, we are studying MCU to give the probabilities of this effect to occur in function of the distance between elements.

We have found in [48], experimental and simulated data of the probabilities that two latches flipped from one single strike for 65nm bulk planar technology in function of the distance between the two latches. They used the model from Hazucha and Svensson with the parameters in Table 4.5, which is similar to our methodology, to compute the MCU rate from the critical charge that makes the two latches flipped simultaneously.

F	0,00565
Qs (fC)	5,72
K	0,000022
A	1,2675E-10
FIT	3,6E+12

Table 4.5: SER Model Parameters

From their data and the model we obtained Table 4.6, which shows the MCU probability depending on the distance between two elements. The MCU (FIT) has been computed using their same SER model parameters and assuming a minimum area transistor. The SEU (FIT) has been computed from their MCU/SEU percentage and the MCU (FIT). Finally, the MCU probability is computed by dividing the MCU (FIT) with the total of the addition of SEU and MCU ratios.

D (um)	Qcrit (fC)	MCU/SEU (%)	MCU (FIT)	SEU (FIT)	MCU Probability	Estimated Probability
0,5	8,31	50,18	1,33E-05	2,7E-05	0,3295	0,3680
0,6	9,95	37,7	9,96E-06	2,7E-05	0,2695	0,3108
1	11,3	29,7	7,87E-06	2,7E-05	0,2256	0,1580
1,5	17,2	10,59	2,80E-06	2,7E-05	0,0941	0,0678
2	26,9	1,94	5,14E-07	2,7E-05	0,0187	0,0291
2,5	30,9	0,97	2,56E-07	2,7E-05	0,0094	0,0125
3	32,4	0,74	1,97E-07	2,7E-05	0,0072	0,0054
4	51,5	0,026	6,98E-09	2,7E-05	0,0003	0,0010
4,5	61,5	0,0046	1,21E-09	2,7E-05	0,0000	0,0004
5	75,4	0,0004	1,07E-10	2,7E-05	0,0000	0,0002

Table 4.6: MCU Probabilities depending on the distance

We have also done a regression model to obtain the following exponential equation:

$$MCU\ Probability = 0{,}8572 \times e^{-1{,}691 \times D}$$

Where D is the distance between latches in micrometers. The model has a coefficient of determination greater than 95% which means that this model is a good fit of the MCU probabilities of at least the two latches. Figure 4.11 shows the estimated probabilities using the model which are also shown in Table 4.6.

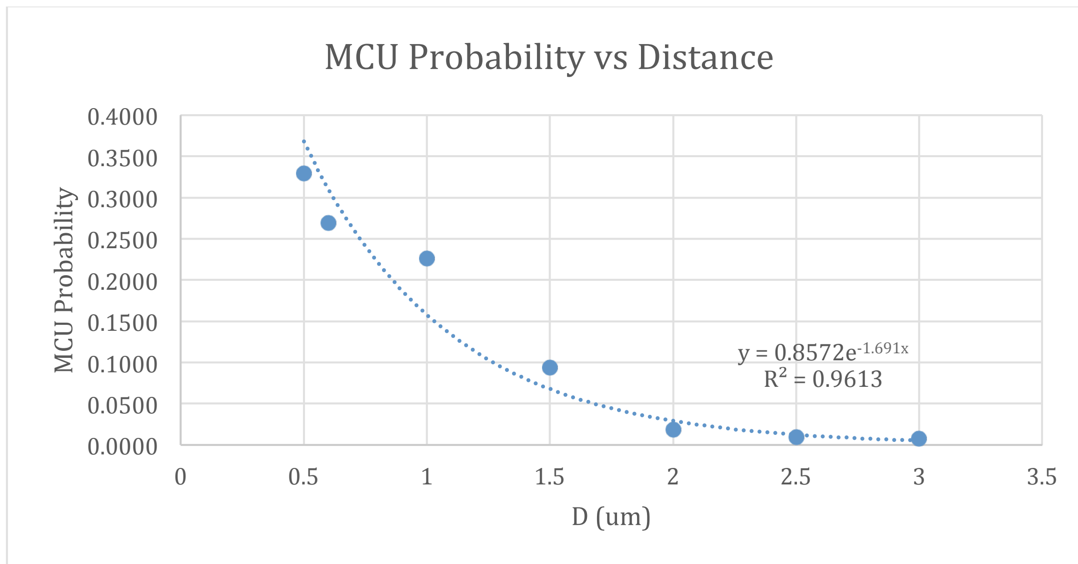


Figure 4.11: MCU Probability vs Distance

We have found data in the literature [49][50] about MCU probabilities of a 6T SRAM cell using different technologies, as can be seen in Figure 4.12.

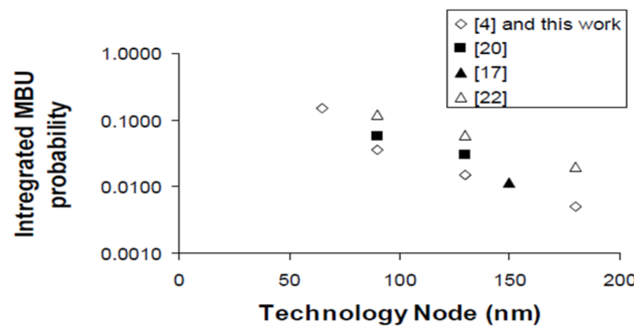


Figure 4.12: Published MCU Probabilities through different technologies of 6T SRAM cell

Then, we have compared these data with the probabilities obtained with our model.

Component	Technology Node (nm)	MCU Probability (Literature)	Estimated Probability
SRAM 6T	180	0,02	0,02
	130	0,05	0,06
	90	0,10-0,15	0,14
	65	0,2-0,25	0,23
	40	0,35	0,38

Table 4.7: SRAM MCU Probabilities

Table 4.7 shows the results of applying the model using 24 lambdas of distance between 6T SRAM cells and different technology nodes of bulk planar. The distance has been taken from real memory layouts considering that one lambda is half the technology node. We can see that the results applying the model with this distance have a good fit of the values from the literature.

Finally, we have searched for information about how this effect behaves in SOI and FinFET technologies [51][52], which can be seen in Figure 4.13. In the case of SOI, MCU probabilities are reduced around 0.25x, while in FinFET MCU probabilities are reduced 0.5x. However, this data doesn't seem very accurate since some papers contradict the others saying that FinFET rises the MCU probabilities [53] and there is not much more information. Moreover, taken into account how these technologies are built and their geometry, the MCU effect should not be much different from the bulk planar technology. Therefore, we could consider that our model can also be applied to these newer technologies without applying any other factor.

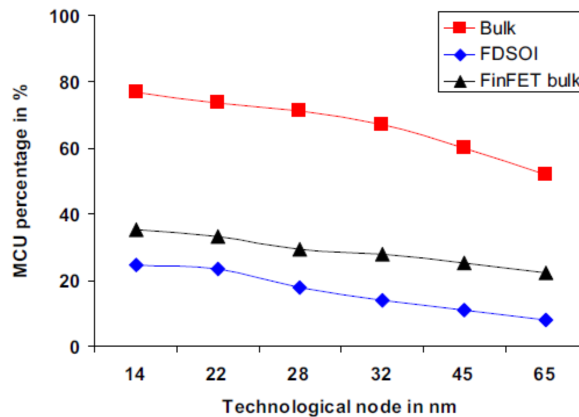


Figure 4.13: MCU % of SER for different technologies

4.3. SER and Aging Combined Effects

This section describes the methodology of how the Soft Error Rate (SER) is affected by aged transistors, as semiconductor aging effects shift the devices behavior.

NBTI and PBTI, which have been already described in section 3.8, are currently the most important aging effects in semiconductor devices. Both aging effects impact the transistors with an increase of the threshold voltage, NBTI affecting the PMOS transistors (more negative Vth) and PBTI affecting the NMOS transistors. Therefore, we could measure the impact of transistors aging on SERs by performing the same simulations and using the same methodology as for the nominal SERs but increasing the Vth of the transistors that we want to test.

We have looked at the literature previous works related to the combination of soft errors and aging to see their methodology and results. In [54] and [55], which use a similar methodology applying Vth shifts, they show that the impact of aging on SERs is very low, around 5-10% respect the nominal values. However, we have conducted our own experiments since we are using newer technology models.

The most important part of the characterization of these aging effects is the model to compute the voltage threshold shifts (ΔV_{th}). There are different models that can be used to compute the variation of Vth over time. NBTI degradation occurs in two phases, namely stress and recovery. Periods of stress are caused due to generation of interface traps while the recovery is made possible by the application of Vdd to the gate that temporarily inhibits further generation of interface traps. The well-established Reaction-Diffusion (R-D) [57] model is the most common way to compute the Vth variation. Figure 4.14 shows the different parameters influencing the amount of threshold voltage shift during stress and recovery phases.

	$\Delta\lvert V_{th}\rvert$ under NBTI
Stress	$\sqrt{K_v^2.(t-t_0)^{0.5}+\Delta V_{TH0}+\delta_v}$
Recovery	$(\Delta V_{TH0}-\delta_v).\lfloor 1-\sqrt{\eta(t-t_0)/t}\rfloor$
K_v	$A.t_{ox}.\sqrt{C_{OX}(V_{GS}-V_{TH}).}$ $[1-V_{DS}/\alpha(V_{GS}-V_{TH})].$ $exp(E_{OX}/E_0).exp(E_a/kT)$

Figure 4.14: Reactive-Diffusion Model Based Threshold Voltage Shift

A simplified version of the RD model can be found in [58] combining both phases. This model has the following form:

$$\Delta Vth = Sp \times K_{DC} \times t^n$$

Where Sp denotes the signal probability of the transistor also named activity factor ((i.e., effective ON time of PMOS/NMOS transistor), K_{DC} is a constant parameter that strongly depends on the technology, voltage and temperature, t is the time in seconds and n is the time exponent which is also dependent on the technology and is obtained experimentally, usually ranging between 0.14-0.5. Therefore, this model have many parameters that are difficult to obtain and require a good knowledge of the technology model to be used accurately.

Figure 4.15 and Figure 4.16 show different examples of Vth shifts at 85ºC and 125ºC using the 32nm PMOS transistor from PTM and different activity factors. We can see in the plots that there is a difference of almost 3x due to the temperature and around 1.2x due to the activity factor. At 85ºC the Vth shifts range between 8-20mV while at 125 ºC the Vth shifts range between 15-50mV. Therefore, the temperature and the activity factor have an important effect on the degradation of the transistors.

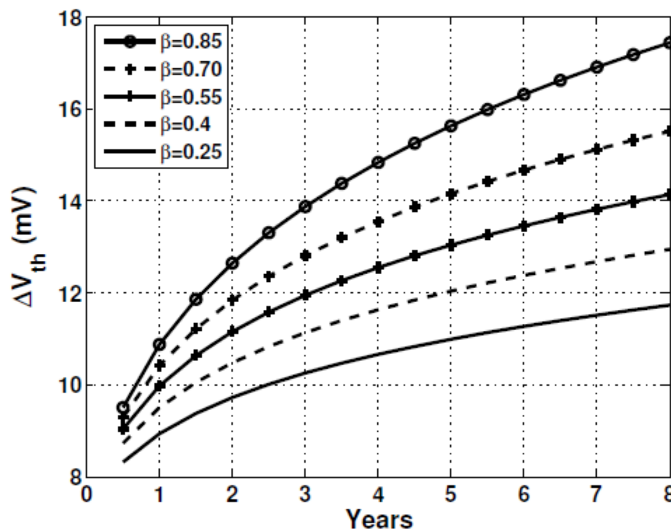


Figure 4.15: Vth shift in 32nm PMOS device operating at 85ºC [57]

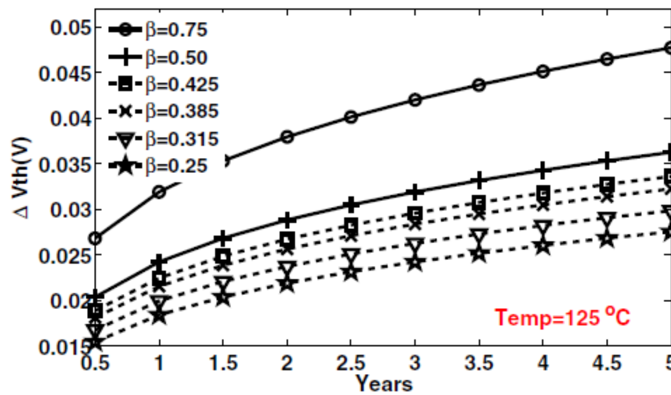


Figure 4.16: Vth shift in 32nm PMOS device operating at 125ºC [59]

We have found previous works in the literature reporting higher Vth shifts [22][55][58] as is shown in Figure 4.17 where the mean Vth shift after 3 years is around 50mV for the same technology model. The main explanation for these higher values is that these works are also taking into account process variations such as Random Dopant Fluctuations (RDF) in their reported values and not only aging. In addition, they are also adding a way to compute the statistical variations of these effects. Moreover, they use the RD model with different time exponent and with extreme conditions such as higher temperatures and voltages.

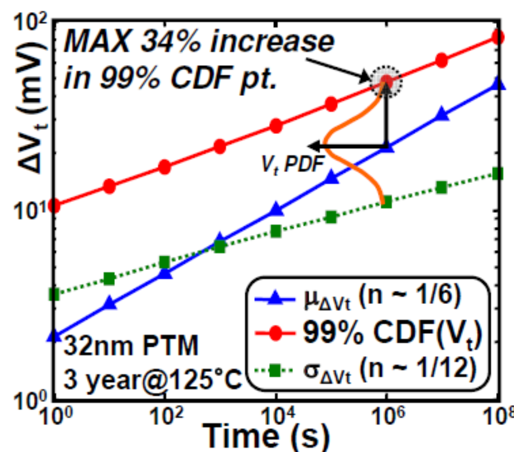


Figure 4.17: Vth shift Adding Variability

The tests performed and the results obtained are shown in section 5.4.

5. Analysis of Basic Components

As has been already described, to obtain the Soft Error Rate (SER) of a component, the critical charge (Qcrit) is required. Qcrit is obtained doing simulations with HSPICE by inserting a current pulse in the sensitive nodes of the component, where the current pulse represents the charge produced by the impact of a neutron strike. This chapter describes how the SER has been computed for each component and it summarizes some of the results obtained.

5.1. Analysis of SRAM Cells

To obtain the Qcrit of the SRAM cell, current pulses are inserted in the storage node Q since the other node (Qb) is symmetric. The values of Qcrit obtained for the 6T SRAM cell have been summarized in Table 5.1, showing the maximum, the minimum and the average Qcrit from all

the environmental parameters (i.e. voltage, temperature and stored value). The latest technology nodes have usually lower critical charge than their predecessors. However, recent technologies, such as FinFETs and SOI improve this aspect and have a higher Qcrit.

6T SRAM Cell			
Technology	**Minimum Qcrit (fC)**	**Maximum Qcrit (fC)**	**Average Qcrit (fC)**
22nm Bulk Planar	0,24	4,91	1,86
22nm SOI Planar	0,67	5,30	2,51
20nm Bulk FinFET	2,12	7,80	4,55
16nm Bulk Planar	0,14	3,08	1,21
14nm Bulk FinFET	2,78	12,52	6,79

Table 5.1: Qcrit values of a 6T SRAM Cell

A Soft Error Rate (SER) is computed from each Qcrit. Then, as the total SER of the element is the sum of the SER from all sensitive nodes, in the case of SRAM cells, SERs from the same environment (voltage, temperature and pulse) but different stored values (0 and 1) are added. That is because the cell has two sensitive nodes and each one always will store the inverse of the other node. Moreover, we could also weight SER values depending on the state of the cell (i.e. holding, reading and writing), but as most of the time cells are holding a value, only the holding mode is considered. Table 5.2 shows the total SER of the 6T and 8T cells build with different technologies, and simulated with typical environmental parameters (1V, 50ºC) and the pulse of 2ps, which is the worst case.

SRAM Cells with Typical Conditions		
Technology	**6T Total SER (FIT)**	**8T Total SER (FIT)**
22nm Bulk Planar	2,04E-05	1,95E-05
22nm SOI Planar	1,25E-06	1,11E-06
20nm Bulk FinFET	1,98E-07	1,78E-07
16nm Bulk Planar	1,09E-05	1,05E-05
14nm Bulk FinFET	8,55E-09	7,76E-09

Table 5.2: 6T and 8T SERs

Both cells have similar SER values, and similar values are obtained with the 10T cell, as the core of all the cells is the 6T and they all have the same sensitive nodes. The highest SERs are with bulk planar and the lowest with bulk FinFET, which corresponds with the highest and lowest Qcrit values. In the case of bulk planar, the 16nm node has lower SERs than the 22nm node. That is because the reduction in the area has more effect when the Qcrit values are already very low, overcoming the reduction of Qcrit.

5.2. Analysis of a Latch

The methodology used to compute the Qcrit of the latch is similar to the methodology used for SRAM cells. A current pulse is injected in the sensitive nodes of the latch, which in our design are the intermediate node and the output node. The latch can be in two modes, transparent which is when the latch transfers the input value to the output or holding the value, being 50% of the time each one. Only the hold mode is considered on the following results since in transparent mode the flipped value is usually rewritten and it can be only propagated if the flip happens in a very specific moment (setup time). Table 5.3 shows the Qcrit values obtained for the latch, showing the maximum, the minimum and the average Qcrit for all the combinations

of parameters. Results are similar to those of SRAM cells, being FinFET and SOI technologies more robust by having a higher Qcrit.

Latch			
Technology	**Minimum Qcrit (fC)**	**Maximum Qcrit (fC)**	**Average Qcrit (fC)**
22nm Bulk Planar	0,48	2,36	1,25
22nm SOI Planar	0,58	2,22	1,21
20nm Bulk FinFET	1,93	5,50	3,39
16nm Bulk Planar	0,39	1,69	0,92
14nm Bulk FinFET	2,70	9,35	5,47

Table 5.3: Latch Qcrit values

Similarly to SRAM cells, SERs from both sensitive nodes are added for each environmental setup (i.e. Temp, V…). Then, SERs of different inputs (0 and 1) are weighted considering equal probabilities. Finally, since we are only considering the holding mode, we are assuming that 50% of the time the latch is not sensitive to particle strikes, so we apply a 0.5x derating factor to the SERs. Table 5.4 shows the total SERs of a latch build with different technologies, and simulated with typical environmental parameters (1V, 50ºC) and the pulse of 2ps, which is the worst case.

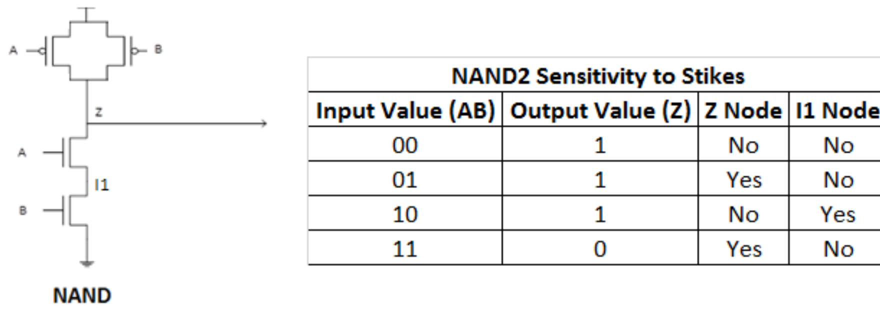
Latch	
Technology	**Total SER (FIT)**
22nm Bulk Planar	5,94E-06
22nm SOI Planar	4,45E-07
20nm Bulk FinFET	1,01E-07
16nm Bulk Planar	3,06E-06
14nm Bulk FinFET	2,53E-09

Table 5.4: Latch SERs

The critical charges of the latch are similar to ones of the SRAM cells. However, since the SERs of the latch are derated by a time vulnerability factor of 50%, the final results are lower. The technology comparison is still the same, being SOI and FinFET more robust to soft errors. Flip flop results are similar since it is composed of two latches being each one vulnerable 50% of the time.

5.3. Analysis of Logic Gates

Logic gate SER analysis is done by injecting the current pulses in the internal nodes that are sensitive to neutron strikes, which depends on the inputs and gate type. Understating that a particle strike activates an off transistor, we can analyze which nodes are sensitive to particle strikes and inject the current pulse in these nodes. Choosing a NAND of 2 inputs as example, we obtained Figure 5.1, which shows which nodes are sensitive for each combination of inputs. This analysis is done for each gate to simulate the strikes only in the sensitive nodes.



NAND2 Sensitivity to Stikes			
Input Value (AB)	Output Value (Z)	Z Node	I1 Node
00	1	No	No
01	1	Yes	No
10	1	No	Yes
11	0	Yes	No

Figure 5.1: NAND2 Sensitivity Analysis

Qcrit values for the NAND2 are shown in Table 5.5.

NAND2			
Technology	Minimum Qcrit (fC)	Maximum Qcrit (fC)	Average Qcrit (fC)
22nm Bulk Planar	0,67	21,28	6,13
22nm SOI Planar	0,72	8,25	4,26
20nm Bulk FinFET	3,33	27,28	13,54
16nm Bulk Planar	0,63	10,73	4,58
14nm Bulk FinFET	4,48	39,65	20,62

Table 5.5: NAND2 Qcrit values

A SER is computed for each input combination by adding the SERs of all the sensitive nodes. Then, the SERs of each input combination are weighted by the probability of the input to occur, but for now, equal probabilities are considered. SER results for the NAND2 are shown in Table 5.6.

NAND2	
Technology	Total SER (FIT)
22nm Bulk Planar	1,72E-06
22nm SOI Planar	5,39E-08
20nm Bulk FinFET	1,37E-09
16nm Bulk Planar	7,57E-07
14nm Bulk FinFET	2,19E-12

Table 5.6: NAND2 SERs

SER values of logic gates are even lower than SRAM cells and latches. That is because analyzing each combination of inputs, a gate usually has one or even none sensitive nodes to strikes. In contrast, the SRAM cells and the latch always have two sensitive nodes. Therefore, when the average is done the total SER of the gate is reduced and lower than in other components.

5.4. Analysis of SRAM Cells with Aging

We have performed a variety of tests with the methodology described in chapter 4.3. We have tested the 6T SRAM cell with three of the technology models that we have: 22nm bulk planar, 20nm bulk FinFET and 22nm SOI planar. The Vth shifts have to be introduce in different ways depending on the technology model. For the bulk planar technology, the BSIM model has an instance parameter named DELVTO which adds directly a shift to the threshold voltage of the transistor. Therefore, for this model we can add the shifts easily computing them with the equa-

tion described in chapter 4.3. For the other two technology models we didn't find such parameter so we have added the Vth shifts as a source of voltage at the gate of the transistors as is explained in [56].

The bulk planar technology is mainly affected by the NBTI effect so we have only applied the Vth shifts at the two PMOS transistors of the SRAM cell. On the other hand, bulk FinFET and SOI planar are affected by both NBTI and PBTI effects so the shifts have been applied at the four transistors of the inverters that compose the SRAM cell. The Vth shifts applied correspond to those published in [57]. The results obtained for similar conditions of temperature and voltage (75ºC and 1V) are shown in Table 5.7 and plotted in Figure 5.2.

Technology	Nominal SER	1 Year SER	3 Years SER	5 Years SER	8 Years SER	Maximum Difference
22BP	2,16E-05	2,16E-05	2,16E-05	2,16E-05	2,16E-05	0,1%
20BF	3,86E-07	3,97E-07	3,97E-07	3,97E-07	3,97E-07	3%
22PSOI	1,33E-06	1,33E-06	1,33E-06	1,33E-06	1,40E-06	6%

Table 5.7: SER compraison throught years

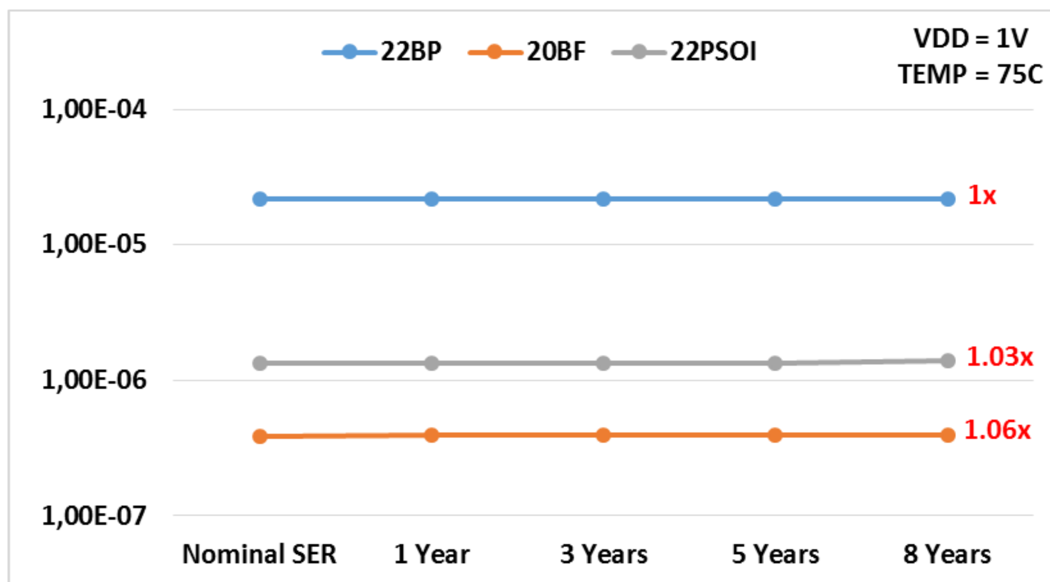


Figure 5.2: SER compraison throught years

Our results shows that for all the technologies, the impact on SERs due to aging is almost negligible, less than 1.1x, which agrees with the results shown in the literature [54][55]. We have also tested a worst case scenario by injecting a higher voltage threshold shift of 50mV and using a higher temperature, as this is the Vth shift reported in most of the works that we found in the literature [22][55][58]. The results of these tests are shown in Table 5.8 using a temperature of 100ºC and 1V.

Technology	Nominal SER	Worst Case Aging SER	Difference
22BP	2,24E-05	2,30E-05	3%
20BF	2,14E-07	2,46E-07	15%
22PSOI	1,41E-06	1,59E-06	12%

Table 5.8: SER in Worst Case Aging Scenario

As shown in Table 5.8, the differences are slightly higher, up to 1.15x for the FinFET technology, which is the most affected technology in our results. However, the effect is almost negligible still agreeing the results shown in the literature [54][55].

6. Trends

In previous chapter we have analyzed some of the components giving some clues on which technologies are more robust to radiation. This chapter provides more data and plots to show different trends. In section 6.1, there is a global comparison between the technologies and components described. Sections 6.2 and 6.3 show the impact of increasing the voltage and the temperature, respectively. Section 6.4 compares the SERs of a logic gate using different fanouts. Finally, section 6.5 shows SERs in different locations and the impact of the neutron flux.

6.1. Technology Trend

Gathering the results of the previous chapter we can compare the SERs of different components and technologies. These results are plotted in Figure 6.1, where each color represents a technology and each group of bars a component.

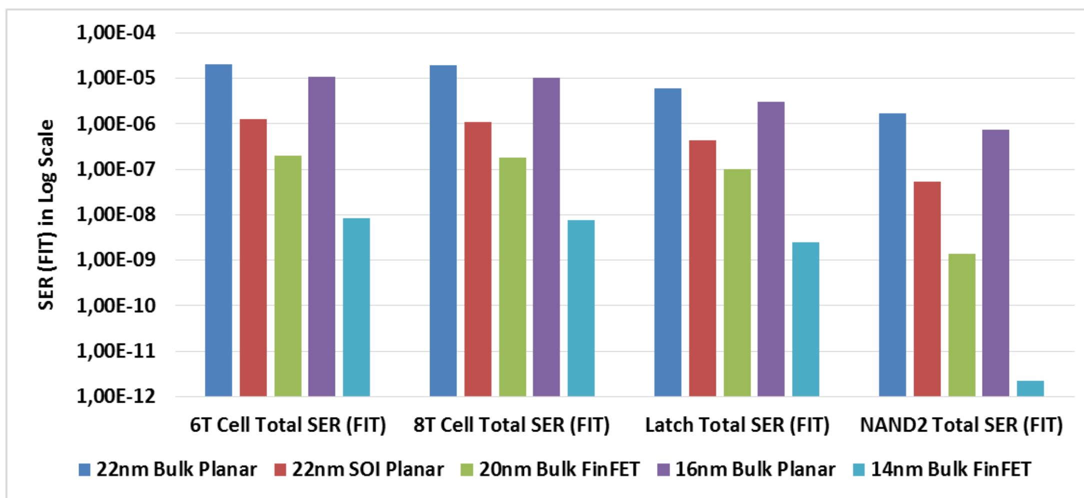


Figure 6.1: Technology Comparison

SERs are in logarithmic scale and when looking at the bars of a component, such as the 6T cell, the higher SERs are for bulk planar and the lower ones are for bulk FinFET with SOI planar in the middle. Therefore, the most vulnerable technology is the bulk planar while bulk FinFET and SOI planar can reduce SERs up to 100x or even more in their lower technology nodes, which makes sense since the sensitive area and the collected charge are bigger in bulk planar.

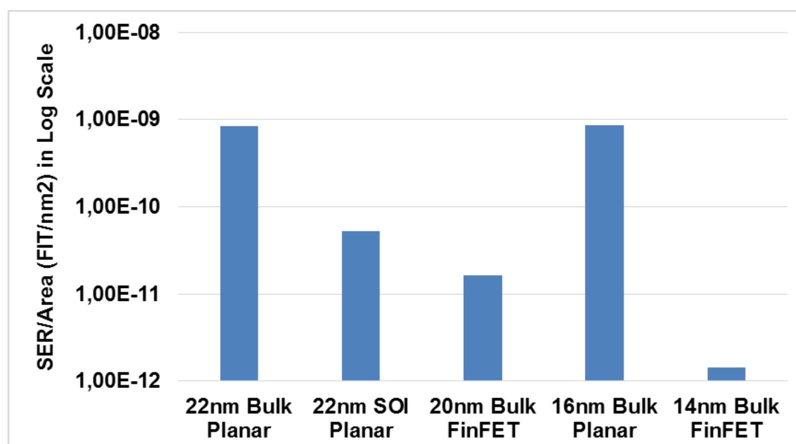


Figure 6.2: SER/Area of a 6T SRAM Cell

Between components, both memory cells have similar results, the latch is a bit more reliable as it is vulnerable only 50% of the time (transparent mode) and the NAND2 has the lower SERs. Typical logic gates (NAND, NOR and NOT) usually have less sensitive nodes to strikes for each input combination, resulting in a total SER lower than other components. In addition, in bulk technology, lower technology nodes have lower SERs which may seem contradictory as in lower nodes Qcrit is usually reduced. However, the reduction in area has a stronger effect when the critical charge is already very low. Therefore, if we look at the SER/Area in

Figure 6.2, both nodes of bulk planar are quite similar, being slightly higher the node of 16nm. As an example, if we consider an SRAM chip with constant die area of 1.5 cm^2, the approximately SER of the 16nm chip would be 128694 FIT and 127549 FIT for the 22nm chip. In the case of FinFETs, our results show that the critical charge of the 14nm node is lower than the 20nm one. Therefore, adding the lower critical charge, the reduction in the sensitive area and the reduction in the collection efficiency, results in much lower SER values.

6.2. Voltage Trend

Table 6.1 compares the SERs of a 6T cell though increasing voltages and different technologies.

Technology	6T SRAM Cell Total SER (FIT)					
	(0,7V|50C)	(0,8V|50C)	(0,9V|50C)	(1V|50C)	(1,1V|50C)	(1,2V|50C)
22nm Bulk Planar	2,46E-05	2,29E-05	2,16E-05	2,04E-05	1,88E-05	1,76E-05
22nm SOI Planar	3,70E-06	2,55E-06	1,78E-06	1,25E-06	8,27E-07	5,81E-07
20nm Bulk FinFET	6,60E-07	4,41E-07	2,96E-07	1,98E-07	1,70E-07	1,42E-07
16nm Bulk Planar	1,33E-05	1,22E-05	1,17E-05	1,09E-05	1,01E-05	9,44E-06
14nm Bulk FinFET	9,52E-08	4,88E-08	2,31E-08	8,55E-09	3,55E-09	1,23E-09

Table 6.1: Voltage comparison

Figure 6.3: Voltage Comparison Plot

Figure 6.3 shows the plot of these results in logarithmic scale where lower values are better. SERs increase with lower voltages since the critical charge becomes smaller. Therefore, it is easier to flip the value and the variation may be as high as 70x as can be seen with the red lines of the plot.

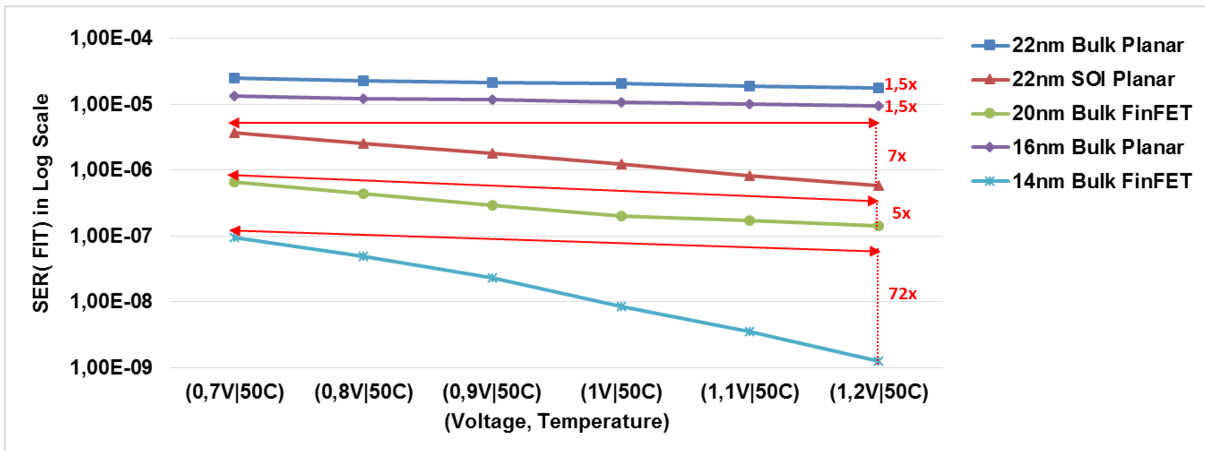


Figure 6.3: Voltage Comparison Plot

6.3. Temperature Trend

Table 6.2 compares the soft error rates of a 6T SRAM cell though increasing temperatures and different technologies.

Technology	6T SRAM Cell Total SER (FIT)			
	(1V|25C)	(1V|50C)	(1V|75C)	(1V|100C)
22nm Bulk Planar	1,84E-05	2,04E-05	2,16E-05	2,24E-05
22nm SOI Planar	1,14E-06	1,25E-06	1,33E-06	1,41E-06
20nm Bulk FinFET	3,75E-07	3,76E-07	3,86E-07	3,85E-07
16nm Bulk Planar	9,69E-06	1,09E-05	1,14E-05	1,20E-05
14nm Bulk FinFET	8,72E-09	8,55E-09	1,23E-08	1,31E-08

Table 6.2: Temperature Comparison

Figure 6.4 shows the plot of these results in logarithmic scale where lower values are better. SER increases with higher temperatures since the critical charge becomes smaller. Even if seems that the variation is low it can be greater than 20% as can be seen with the red lines of the plot, but still has a low effect compared with the voltage variation. In the case of FinFET technology, the models used do not model the temperature accurately [47] so the variations are very low and slightly oscillating. Therefore, only the results from the nominal temperature (25ºC) should be used for FinFET technology.

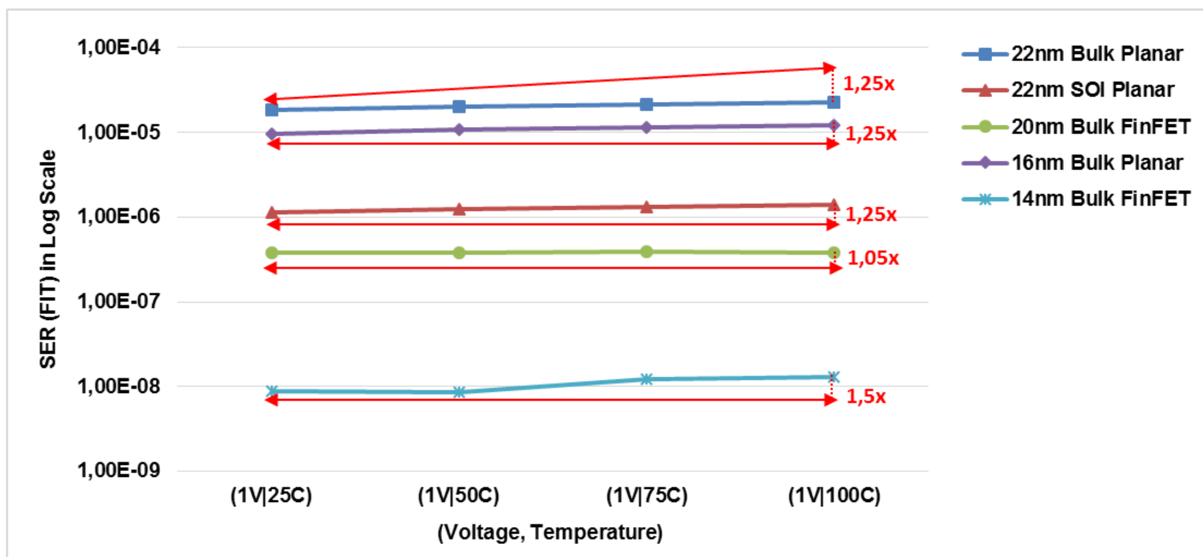


Figure 6.4: Temperature Comparison Plot

6.4. Fanout Trend

Table 6.3 compares the soft error rates of the logic gate NOT build in 22nm bulk planar technology with different fanouts.

NOT (22nm Bulk Planar)	
Fanout	**Total SER (FIT)**
1	2,94E-06
2	2,85E-06
3	2,79E-06
4	2,71E-06
5	2,63E-06
6	2,61E-06
7	2,55E-06
8	2,48E-06
9	2,48E-06
10	2,40E-06

Table 6.3: Fanout and Current Pulse Comparison

In Figure 6.5 SERs are slightly reduced with higher fanouts as there is more capacity in the output and the critical charge increases, with a variation that can be up to 1.5x. We can also see a linear model that is a good fit of these values, with a squared R of 98%.

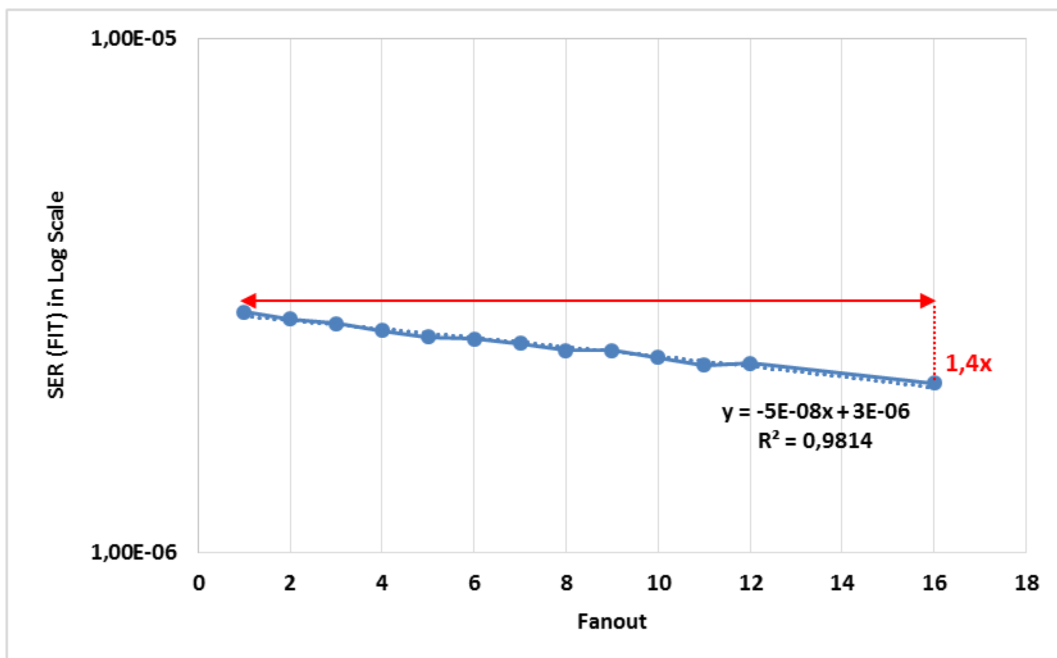


Figure 6.5: Fanouts Comparison Plot

6.5. Location Trend

Table 6.4 shows neutron fluxes of different European locations relative to the reference flux from New York City at sea level.

Reference Flux = 0,00565 neutrons/cm2*s (NYC, SL)				Note: Flux At Medium Solar Modulation Neutron Flux Relative to the Reference Flux					
Location	Coordinates (Grades)	Vertical Cutoff (GV)	Mean Altitude (m)	Sea Level	Base Altitude	2000m	4000m	8000m	12000m
Turin	45N, 7E	5	239	0,87	1,07	4,29	15,56	98,62	296,19
Barcelona	41N, 2E	6	12	0,8	0,81	3,78	13,25	79,41	228,2
Athens	37N, 23E	8	170	0,72	0,83	3,27	11,06	62,78	172,64
Västerås (Sweden)	59N, 16E	1	17	1,01	1,03	5,38	21,06	153,56	527,47
Berlin	52N, 13E	2	34	0,97	1	5,01	19,07	131,78	428,58
London	51N, 0W	3	24	0,97	0,99	4,99	18,96	130,7	424,01
Moscow	55N, 37E	2	150	0,99	1,13	5,18	19,96	141,23	469,94

Table 6.4: Relative Fluxes of different locations

Figure 6.6 shows the relative fluxes plotted. The higher neutron fluxes are located in Västerås as is closer to the pole while the lower is in Athens which is nearer the equator. These relative fluxes have been computed using the online calculator [33], which uses the JEDEC standard, with a medium solar activity (50%).

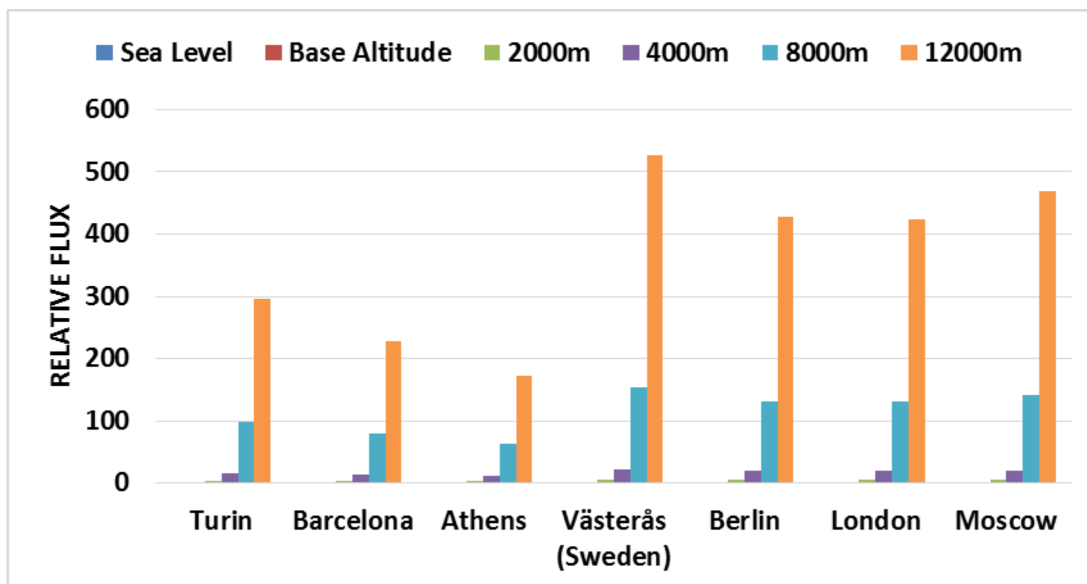


Figure 6.6: Relative Neutron Fluxes of different locations

The relative fluxes can be multiplied directly by the Soft Error Rates (SER) obtained with the reference flux to obtain the SER of the desired location. We have computed the SERs of a 6T SRAM cell at different locations and altitudes as is shown in Figure 6.7.

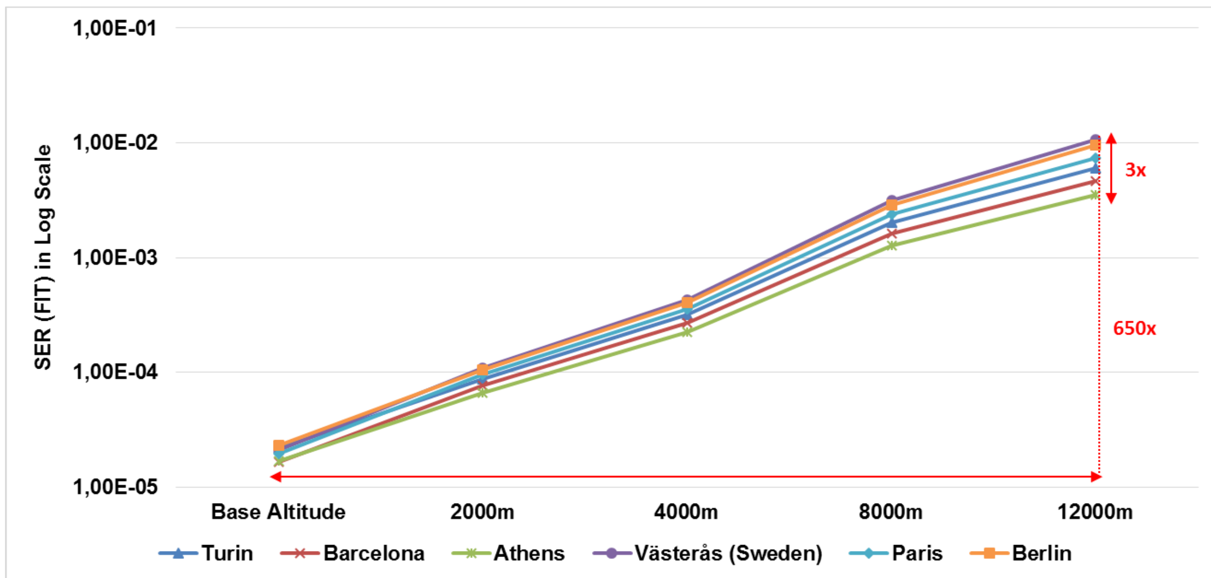


Figure 6.7: SERs depending on the Location and Altitude of a 6T SRAM Cell in 22nm Bulk Planar

The difference between cities is due the influence of the magnetic field of the earth, where cities near the equator have lower SERs. Moreover, there is an exponential increase of the SER when varying the altitude that can be as high as 650x.

7. Conclusions

The work has been focused on the characterization of soft errors due neutron strikes, which have been the major reliability concern of the industry in the last years and are also expected to be in the near future. In this document, technologies being used or expected to be used in the near future have been reviewed. In addition, different possible source of failures that may be critical for these technologies have been described.

Looking at the results, it is obvious that as bulk planar technology scales down Qcrit is lower so the elements may become more vulnerable to soft errors. However, the scaled area and collection efficiency overcomes the reduction of Qcrit making the SER almost constant or even a bit lower. Nevertheless, taking into account the increase in the number of elements integrated in a chip when the technology scales down, the SER increases and becomes an important issue for the reliability of the device.

On the other hand, newer technologies, such as multi-gate FinFETs, and newer materials, such as SOI, are more resistant to radiation effects. In addition, we have showed that environmental parameters, such as temperature and voltage, and the location, may have a huge impact on the soft error rates. Specially the altitude, which may increase the SERs up to 650x. In conclusion, this study suggests that newer technologies can reduce soft error rates up to 100x whereas planar CMOS is becoming more vulnerable due to the scaling down of its components and the increased number of elements.

8. Acronyms

The following table shows a list of the acronyms used in this document and their meaning:

Acronym	Definition
TVF	Technology Vulnerability Factor
CVF	Circuit Vulnerability Factor
CMOS	Complementary Metallic Oxide Semiconductor
FinFET	Fin-Shaped Field Effect Transistor
PTM	Predictive Technology Model
ITRS	International Technology Roadmap for Semiconductors
RDF	Random Dopant Fluctuations
LER	Line Edge Roughness
RTN	Random Telegraph Noise
EM	Electromigration
MeTTF	Median Time to Failure
MSV	Metal Stress Voiding
MTTF	Mean Time to Failure
GOW	Gate Oxide Wearout
HCI	Hot Carrier Injection
NBTI/PBTI	Negative/Positive Bias Temperature Instability
RIF	Radiation Induced Faults
SER	Soft Error Rate
Qcrit	Critical charge
SOI	Silicon On Insulator
SHE	Self-Heating
SDC	Silent Data Corruption
DUE	Detected Unrecoverable Error
FIT	Failure In Time

9. Bibliography

[1] R. Baumann, "Soft Errors in Advanced Computer Systems", IEEE Design & Test of Computers, vol. 22, no. 3, pp. 258-266, May/June, 2005

[2] S. Borkar et al., "Design and Reliability Challenges in Nanometer Technologies", IEEE DAC, pp. 75-75, 2004

[3] P. Shivakumar, M. Kistler, "Modeling the effect of technology trends on the soft error rate of combinational logic", IEEE DSN, 2002

[4] S. S. Mukherjee, C. Weaver, J. Emer, S.K. Reinhardt, T. Austin, "A systematic methodology to compute the architectural vulnerability factors for a high-performance microprocessor", MICRO, pp. 29-40, 2003

[5] D. Ernst et al., "Razor: circuit-level correction of timing errors for low-power operation", IEEE MICRO, Vol. 24, no. 3, pp. 10-20, 2004

[6] R. Vadlamani et al., "Multicore soft error rate stabilization using adaptive dual modular redundancy", IEEE DATE, pp. 27-32, 2010

[7] Kaliorakis, M.; Tselonis, S.; Foutris, N.; Gizopoulos, D., "D3.1 – Report on major classes of hard- ware component"

[8] M. H. Abu-Rahma, M. Anis. "Variability in Nanometer Technologies and Impact on SRAM", Springer New York, 2013

[9] P. Mishra, A. Muttreja, N. K. Jha, "FinFET Circuit Design", Springer Science, 2011

[10] Berkeley Predictive Technology Model and BSIM, http://www-device.eecs.berkeley.edu/bsim/

[11] Arizona State University (ASU) PTM, http://ptm.asu.edu/

[12] International Technology Roadmap for Semiconductors (ITRS) projections, http://www.itrs.net/

[13] Y. Cao, "Predictive Technology Model for Robust Nanoelectronic Design", Integrated Circuits and Systems, Springer, 2011

[14] D. Lu, "PhD Dissertation: Compact Models for Future Generation CMOS", Electrical Engineering and Computer Sciences University of California at Berkeley, 2011

[15] Shrikanth, "Reliability In The Face of Variability in Nanometer Embedded Memories", UPC Thesis, March, 2014

[16] S. Mukherjee. "Architecture Design for Soft Errors", Morgan Kaufmann, 2008

[17] Ralph Group nanoscale Physics, http://people.ccmr.cornell.edu/~ralph/projects/emig_movies/

[18] P. Hazucha and C. Svensson, "Impact of CMOS Technological Scaling on the Atmospheric Neutron Soft Error Rate", IEEE Transactions on Nuclear Science, Vol. 47, No. 6, pp.2586-2594, December 2000

[19] Q. Ding, R. Luo, H. Wang, H. Yang, Y. Xie, "Modeling the Impact of Process Variation on Critical Charge Distribution", IEEE, 2006

[20] Paul K. Chu, "Novel Silicon-on-Insulator Structures for Reduced Self-Heating Effects", IEEE Circuits and Systems magazine, 2005

[21] A. Armstrong, Self-Heating in SOI, http://www.ee.qub.ac.uk/nisrc/simulations/shsoi.htm

[22] X. Wang, A. R. Brown, B. Cheng, A. Asenov, "Statistical Variability and Reliability in Nanoscale Fin-FETs", IEEE, 2011

[23] T. Heijmen, D. Giot and P. Roche, "Factors that impact the critical charge of memory elements", IEEE International On-Line Testing Symposium, 2006

[24] R. Naseer, S. DasGupta. "Critical Charge Characterization for Soft Error Rate Modeling in 90nm SRAM", IEEE, 2007

[25] S. Bota, G. Torrens, et al., "Critical Charge Characterization in 6-T SRAMs During Read Mode", IEEE, 2009

[26] J. F. Ziegler and W. A. Lanford, "Effect of cosmic rays on computer memories", Science, Volume 206, Issue 4420, pp. 776-788, 1979

[27] A. Taber and E. Normand, "Single Event Upset in Avionics", IEEE Transactions on Nuclear Science, Vol. 40, No. 2, April 1993

[28] P. C. Murley and G. R. Srinivasan, "Soft-error Monte Carlo modeling program, SEMM", IBM, 1996

[29] Tino Heijmen, "Analytical semi-empirical model for SER sensitivity estimation of deep-submicron CMOS circuits", IEEE International On-Line Testing Symposium, 2005

[30] P. Hazucha, C. Svensson, and S. A. Wender, "Cosmic ray soft error rate characterization of a standard 0.6 μm CMOS process," IEEE J. Solid-State Circuits, Oct. 2000

[31] J. F. Ziegler, "Terrestrial cosmic rays", IBM journal of research and development, 1996

[32] JEDEC Standard, "Measurement and Reporting of Alpha Particle and Terrestrial Cosmic Ray-Induced Soft Errors in Semiconductor Devices", JESD89A, 2006

[33] Neutron Flux Calculator (SEUTEST), http://www.seutest.com/cgi-bin/FluxCalculator.cgi

[34] M. S. Gordon, P. Goldhagen, et al., "Measurement of the Flux and Energy Spectrum of Cosmic-Ray Induced Neutrons on the Ground", IEEE Transactions on Nuclear Science, Vol. 51, No. 6, December 2004

[35] P. Goldhagen, et al., "Measurement of the energy spectrum of cosmic-ray induced neutrons aboard an ER-2 high-altitude airplane", Nuclear Instruments and Methods in Physics Research, 2002

[36] Synopsys HSPICE, https://www.synopsys.com/tools/Verification/AMSVerification/CircuitSimulation/HSPICE/Pages/default.aspx

[37] E. H. Cannon, et al., "SRAM SER IN 90,130 AND 180 NM BULK AND SOI TECHNOLOGIES", IEEE, 2004

[38] D.R. Ball, et al., "Comparing Single Event Upset Sensitivity of Bulk vs. SOI Based FinFET SRAM Cells using TCAD Simulations", IEEE, 2010

[39] Yi-Pin Fang and Anthony S. Oates, "Neutron-Induced Charge Collection Simulation of Bulk FinFET SRAMs Compared With Conventional Planar SRAMs", IEEE Transactions On Device And Materials Reliability, 2011

[40] Leti UTSOI Model, http://www-leti.cea.fr/en/How-to-collaborate/Focus-on-Technologies/UTSOI

[41] TRAMS European Project, http://trams-project.upc.edu/en

[42] F. M. Yigletu, et al., "Compact Charge-Based Physical Models for Current and Capacitances in AlGaN/GaN HEMTs", IEEE TED, 2013

[43] SPICE Simulation Fundamentals, National Instruments, http://www.ni.com/white-paper/5413/en/#toc4

[44] Neil H. E. Weste, David Money Harris, "CMOS VLSI Design: A Circuits and Systems Perspective", Addison Wesley Publication, 2010

[45] Z. Jaksic and R. Canal, "Enhancing 6T SRAM Cell Stability by Back Gate Biasing Techniques for 10nm SOI FinFETs under Process and Environmental Variations", 19th International Conference "Mixed Design of Integrated Circuits and Systems", May 24-26, 2012

[46] S. A. Tawfik, V. Kursun, "Characterization of New Static Independent-Gate-Biased FinFET Latches and Flip-Flops under Process Variations", 9th International Symposium on Quality Electronic Design, IEEE, 2008

[47] S. Sinha, G. Yeric, et al., "Exploring Sub-20nm FinFET Design with Predictive Technology Models", ACM, 2012

[48] K. Zhang, J. Furuta, K. Kobayashi, and H. Onodera, “Dependence of Cell Distance and Well-Contact Density on MCU Rates by Device Simulations and Neutron Experiments in a 65-nm Bulk Process”, IEEE Transactions on Nuclear Science, 2014

[49] Anand Dixit and Alan Wood, “The Impact of New Technology on Soft Error Rates”, IEEE, 2011

[50] N. Seifert, et al., “RADIATION-INDUCED SOFT ERROR RATES OF ADVANCED CMOS BULK DEVICES”, IEEE, 2006

[51] G. Hubert, L. Artola, D. Regis, “Impact of scaling on the soft error sensitivity of bulk, FDSOI and FinFET technologies due to atmospheric radiation”, INTEGRATION the VLSI journal, 2015

[52] Ethan H. Cannon, et al., “MULTI-BIT UPSETS IN 65NM SOI SRAMS”, IEEE, 2008

[53] Yi-Pin Fang and A. S. Oates, “Cell Level Soft Error Rate Simulations of Planar and FinFET Processes”, TSMC, 2014

[54] Ethan H. Cannon, et al., “The Impact of Aging Effects and Manufacturing Variation on SRAM Soft-Error Rate”, IEEE TDMR, 2008

[55] M. Bagatin, et al., “Impact of NBTI Aging on the Single-Event Upset of SRAM Cells”, IEEE Transactions on Nuclear Science, 2010

[56] T. C. Carusone, et al., “Monte Carlo Simulation in HSPICE”, Analog Integrated Circuit Design, 2007

[57] S. Ganapathy, et al., “iRMW : A Low-Cost Technique to Reduce NBTI-Dependent Parametric Failures in L1 Data Caches”, IEEE, 2014

[58] K. Kang, et al., “Estimation of Statistical Variation in Temporal NBTI Degradation and its Impact on Lifetime Circuit Performance”, IEEE ICCAD, 2007

[59] A. Ricketts, et al., “Investigating the Impact of NBTI on Different Power Saving Cache Strategies”, DATE, 2010