1. Executive Summary

This document, deliverable D3.6, presents the results of the work performed in Work Package 3 (WP3) on the analysis of the effect of environmental (V, voltage fluctuations, T, temperature and SEU, single even upsets) as well as process parameter (P) and degradation mechanisms on the performance variability of the memories at cell and system levels. This deliverable at circuit and system level completes Milestone MS2 (M12) together with deliverable D1.1 at device level. The title of MS2 is “Variability and Reliability analysis for bulk CMOS”. The work is entirely inside Task 3.1; the analysis of the impact of PVT variations is a required precedent to the task of mitigation techniques (T3.1) and countermeasure techniques (T3.2) corresponding to the second year work (milestones 4, 7 and 8), see Figure 1.

Both, static and dynamic memory cells and circuits, have been contemplated in this work, considering as key cells the 6T for SRAM [1] and the 3T1D [2] (and partially 1T1C [3]) for DRAM. To evaluate the impact of the PVT variations, Hspice tool has been used. The MOS models taken into account have been the Si bulk CMOS for 45, 32, 22 and 16 nm nodes (Predictive Technology Models [4], PTM) with specific variability scenario described in section 3.2, and the 18 and 13 nm technologies result of WP1 (models specific of the TRAMS project). Due to the fact that Tasks 1.1 and 1.2 already work with the Carbon Nanotube Field Effect Transistor Technology Variability analysis as main objective (with the use of the CNTFET Hspice model from Stanford [5]) a section to compare the impact of PVT in Silicon technology and CNTFET has been included in this document (the comparison is concentrated only on the 6T cell) in Section 8.

![Figure 1. Organization of Work Package 3 (WP3)]
2. Introduction

The aim of WP3 is to deliver key mechanisms to mitigate and reduce variability and increase reliability at layout, circuits and system level, as well as to determine and propose innovative compensating and fault tolerant techniques considering the PVT variability and corresponding yield impact. The key effects of that environmental fluctuations and process variations are exposed in this deliverable. These objectives will be developed during the second and third year of the project, milestones 4 to 8. Figure 2 shows the global framework of technologies and cell type objective of Work Package 3 (WP3). At device level we differentiate between devices modelled by TRAMS whose characteristics as well as variability and reliability performances are a goal of the project (Bulk CMOS is scheduled for M12 (Milestone MS2), Finfets for M18 (Milestone MS3), CNT for M18, the rest for M30 (Milestones MS5,6)) and medium/long term technology devices with promising characteristics for memory systems that although they are not objective of device modelling in TRAMS they will be considered at a exploratory level at circuit and system level (Task 2.3, M36). In this last set of technologies we will include Metal-Insulator-Metal devices (MIMs) [6], RRAM [7], electromechanical CNT arrays (MCNT) [8], Nanobridge devices [9], Metal-Insulator-Semiconductor (MIS) [10], Phase change memories (PCM) [11], Ferromagnetic RAMs [12], MRAM [13] and other devices. The objective of T2.3 will be the analysis of new memory cells such as 1T-SRAM and CB. In this first year and in relation with deliverable D3.6, Task 3.1 has been dedicated to evaluate the impact of the PVT variations in a set of SRAM and DRAM cells (red boxes in Figure 2), using device models previously available (PMT for Silicon and Stanford for CNTFET) and including the new and original results from WP1 with the device modelling and variability evaluation from 18 and 13 bulk CMOS technologies.

![Figure 2. Framework of technologies and memory cells considered in WP3. Red boxes show the technologies and cell considered in this deliverable.](image-url)
2.1. Scope of this document

The document analyses the impact of Voltage and Temperature fluctuations as well as Process variations for different technology nodes. The document covers both SRAM and DRAM type of memories, comparing results among them. Simulations have been done both at basic cell circuit and system (32KB first level cache and 4MB last-level cache) levels and they are also compared with CNTFET technologies (the model of the devices is from Stanford and the variability models are preliminary results from WP1 TRAMS project). Additionally specific analysis of the impact of single event upsets (SEU) and device degradation (BTI) are also included in Sections 7 and 5.4 respectively.

![Figure 3. Scope of technologies and memory cells in this deliverable (D3.6)](image)

3. Objectives and variability scenario

3.1. Objectives and introduction

The aim of this document is the analysis of the environmental (power supply voltage and temperature) fluctuations, the process variability for different technology nodes including sub-22nm as well as BTI degradation and SEU impact on memory circuits. We will evaluate basic 6T, 1T1C and 3T1D bit cells, and 32KB and 4MB cache memory circuits for 6T and 3T1D and we consider the following Si-bulk CMOS technologies 45, 32, 22, 18, 16, 13 and CNT (equivalent to 16nm node). Device models for 45, 32, 22 and 16 nm are the ones known as Predictive Technology Models (University of Arizona [4]), the models for 18 and 13 nm are results of WP1, and the CNT analysis uses a modification of the models of Stanford (see section 8) with preliminary results about variability from TRAMS WP1.

Section 4 is dedicated to SRAM memories characterized by the 6T memory cell. Section 4.1 analyses the impact of VT and node variations on speed parameters and energy consumption, and in section 4.2 the robustness of the cell in front of process parameter variations is presented.

Section 5 analyses DRAM memories, characterized by 1T1C in section 5.1 and 3T1D in the rest. Section 5.2 analyses the impact of VT and node variations on speed parameters. In section 5.3 the robustness of the 3T1D cell in front of process variation is investigated and in section 5.4 the analysis of the impact of BTI degradation of 3T1D on memory performances and yield is presented. The impact of the process variation on the cache memory performances (both 6T and 3T1D) are analysed in Section 6. In section 7 the impact of SEU on the memories reliability is investigated, and in Section 8 the performances of CNT in comparison with the rest of Si-bulk technologies are presented (for the 6T cell).
3.2. Variability scenarios

The margin of temperature variation considered in this document is, in general, the range 25 °C to 110 °C. The margin of V_{DD} variations due to Rl and RdI/dt has been considered as a +/-10% of the nominal power supply used in each technology. For process variation we have considered the following different models:

Process variation model used for PTM technologies

For the four Si-bulk CMOS technologies, 45, 32, 22 and 16 nm, covered by PTM we have considered the process variations of the threshold voltage of the devices (V_{th}) and the device geometry (L and W).

For the V_{th} we have assumed a Gaussian distribution and independent components for random variation (due to random dopants distribution, RDD and line edge roughness, LER) and correlated Gaussian for systematic variations. Geometry variations have been modelled as systematic Gaussian distributions. In all the analysis at system level (cache) both systematic and random variations have been considered and in the case of analysis at cell level, only V_{th} random variations are contemplated. For each technology we have considered different variation scenarios, standard for 45nm, moderated and high for 32nm and moderated, high and very high for 22 and 16nm. Table 1 shows the standard deviations or second moment of the respective distributions. The levels of variability assumed in the high and very high variability scenarios are consequent with that observed and deduced for 18 and 13 nm technologies, result of Work Package 1.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Scenario</th>
<th>total systematic</th>
<th>random(*)(**)</th>
<th>Geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100 x 1 $\sigma$/nominal</td>
<td>100 x 1 $\sigma$/nominal</td>
<td>100 x 1 $\sigma$/nominal</td>
</tr>
<tr>
<td>45 nm</td>
<td>standard</td>
<td>V_{th}</td>
<td>V_{th}</td>
<td>L,W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2%</td>
<td>4%</td>
<td>2%</td>
</tr>
<tr>
<td>32 nm</td>
<td>moderated</td>
<td>3%</td>
<td>6%</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>4%</td>
<td>15%</td>
<td>2%</td>
</tr>
<tr>
<td>22 nm</td>
<td>moderated</td>
<td>4%</td>
<td>8%</td>
<td>2.5%</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>4%</td>
<td>15%</td>
<td>2.5%</td>
</tr>
<tr>
<td></td>
<td>very high</td>
<td>5%</td>
<td>30%</td>
<td>2.5%</td>
</tr>
<tr>
<td>16 nm</td>
<td>moderated</td>
<td>5%</td>
<td>10%</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>5%</td>
<td>20%</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>very high</td>
<td>6%</td>
<td>40%</td>
<td>3%</td>
</tr>
</tbody>
</table>

(*) (random dopants distribution, RDD, and line edge roughness, LER), non correlate (***) for minimum size, for general case correct with /sqrt(WL)

Table 1. Process variation model for the analysis with PTM technologies
Process variation model used for WP1 technologies

Devices models for 18 and 13 nm technologies provided by WP1 present a very high variability on $V_{th}$, caused by RDD and LER mechanisms. The standard deviations have been obtained from WP1 analysis and are given in Table 2.

Process variation model used for CNT technology

The process variation model for CNTFET technology is part of the work done in WP1 (Task 1.1), an introduction to the variation model used is presented in Section 8.

<table>
<thead>
<tr>
<th>device</th>
<th>$\sigma$ $V_{th}$</th>
<th>100$\times$ $\sigma$/nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>18nm NMOS</td>
<td>66.7 mV</td>
<td>33%</td>
</tr>
<tr>
<td>18nm PMOS</td>
<td>116 mV</td>
<td>58%</td>
</tr>
<tr>
<td>13nm NMOS</td>
<td>78.8 mV</td>
<td>39%</td>
</tr>
<tr>
<td>13nm PMOS</td>
<td>116 mV</td>
<td>58%</td>
</tr>
</tbody>
</table>

(*) (random dopants distribution, RDD, and line edge roughness, LER), non correlate (**) for minimum size, for general case correct with $/sqrt{WL}$

Table 2. $V_{th}$ process variation model for analysis with 18 and 13 nm CMOS devices (VDD=0.9 volts).
3.3. The 32KB cache memory under analysis

A 32KB L1 cache macro as shown below has been designed and implemented in HSPICE. The cache block takes advantage of array sub-blocking to reduce the impact of variations [14]. Each of the 32 sub-blocks is organized into 128 columns by 64 rows (Figure 4). Every sub-block is decoded using a pre decoder and an address decoder decodes the row within the decoded sub-block. The global and local controllers generate synchronization timing signals for the following: Address Generation, Pre Charging and Read/Write Enable.

![32 KB CACHE ORGANISATION](image1)

![1 KB ARRAY SUB-BLOCK](image2)

Figure 4. 32KB First-level cache and internal organization

The array and row decoder are designed with dynamic CMOS, the column multiplexer is a pass-transistor based tree design and a differential sense amplifier capable of amplifying very low voltage swings is designed.

---

page 6/98
4. Si-MOS Static RAM (SRAM) cells and systems.

This section analyses the cache structure shown in section 3 when implemented in Si-MOS (or bulk) technology. We conduct the study on the most-used static RAM cell: the 6T cell.

The 6T cell stores the value in a loop of 2 inverters. This value is read (or written) through the differential bitlines. The pass transistors connected to each bitline will connect the value stored in the cell to these bitlines when the signal word (or wordline) is activated otherwise they isolate the value stored in the cell from the bitline. This behaviour makes it possible to reuse the wordline and bitline signals across an array of 6T cells as shown in section 3.

The next subsection analyses a 6T SRAM based L1 cache. Section 4.1 describes the modelling of the cache structure plus the analysis of delay and energy consumption of such memory structure under process, variation and temperature variations. Then, Section 4.2 provides a robustness characterization through static and dynamic noise margins analysis for the 6T at cell level.

4.1. Modeling and analysis of delay and energy behavior of 6T SRAMs under spatio-temporal variations (PVT)

This subsection considers a 6T SRAM in the 32KB cache previously introduced. We first model its behaviour to speed-up the process of exploring different alternatives and configurations. Then, we analyse the behaviour of the cache under process, voltage and temperature variations. Finally, we analyse the effect of fine-grain $V_{th}$ tuning in the cache structure to alleviate the widespread in terms of delay and, specially, energy caused by process, voltage and temperature variations.

4.1.1. Delay and Energy Modeling

Simulation of memory structures implemented in future technologies is one of the main contributions of this project. Conventional tools, such as Hspice, provide high levels of accuracy at a big cost in terms of simulation time. This trade-off can be alleviated by developing fast yet accurate models of the behaviour of the defined memory structures. In this section, we will briefly describe the rationale behind the modelling and we will provide the error analysis. Overall, the use
of these models speed-up the exploration of the design-space by 1x10^5 while providing a median error of 5%.

**Delay Modelling**

Statistical Static Timing Analysis (SSTA) [15] based on Figure 6(a) assumes that all transistors within the gate have similar $L_{eff}$ and $V_{th}$ values. This assumption leads to a single distribution for delay as a function of varying parameters. Also, the impact of delay due to temperature variations on delay is higher when compared to other physical parameters. In reality, what happens is that gates have multiple distributions depending on their current state as shown in Figure 6(b). With increase in temperature, the probability density Path Based Delay Calculation function broadens and the number of samples tending towards the mean decreases with more number of samples towards right hand side of the mean.

Consider the case of 2 NOT gates connected in series. For any input, both pull-up and pull-down network of adjacent gates would actively take part in the output transition. Then again, this model would only assume a variation in the inputs and it will not account for individual variations across the transistors that constitute the path between the input and output. If we are able to determine a path for every possible input and model the parameters of every transistor along the path as a random Gaussian function, then this would result in an ideal delay model that is aware of the characteristics of every transistor that makes up the path.

Assuming a linear dependence between spatial parameters and delay, we can derive a generalized equation for delay $D_i$ for any path $i$ composed of $j$ transistors given by

$$D_i = \sum_{a=1}^{j} \left( D_{nominal}^{pmos} + m_{l_{eff}}^{pmos} \delta_{l_{eff}}^{p(a)} \right)$$

$$+ m_{v_{th}}^{pmos} \delta_{v_{th}}^{p(a)} + m_{v_{temp}}^{pmos} \delta_{v_{temp}}^{p(a)}$$

$$+ \sum_{a=1}^{j} \left( D_{nominal}^{nmos} + m_{l_{eff}}^{nmos} \delta_{l_{eff}}^{n(a)} \right)$$

$$+ m_{v_{th}}^{nmos} \delta_{v_{th}}^{n(a)} + m_{v_{temp}}^{nmos} \delta_{v_{temp}}^{n(a)}$$

For devices maintained at same temperature, the above equation can be approximated into the equation below:

![Figure 6. Example of SSTA (a) and Path-based Delay calculation (b)](image-url)
TRAMS project, FP7-INFSO-IST-248789

\[ D_i = \sum_{a=1}^{i} (D_{\text{nominal}}^{\text{pmos}} + m_{\text{eff}}^{\text{pmos}} \cdot \delta_{\text{eff}}^{(a)}) + m_{\text{th}}^{\text{pmos}} \cdot \delta_{\text{th}}^{(a)} + \sum_{a=1}^{i} (D_{\text{nominal}}^{\text{nominal}} + m_{\text{eff}}^{\text{nominal}} \cdot \delta_{\text{eff}}^{(a)} + m_{\text{th}}^{\text{nominal}} \cdot \delta_{\text{th}}^{(a)}) + f(m, n) \cdot (D_{\text{nominal}}^{\text{pmos}} + D_{\text{nominal}}^{\text{nominal}}) + m_{\text{temp}} \cdot \delta_{\text{temp}} \]

Solving the above equation for the solution of the form \( Y = Xb + c \) and estimating \( b = YX^{-1} \)

\[
D_i = \begin{bmatrix}
\delta_{\text{eff}}^{(1)} & \delta_{\text{th}}^{(1)} & \delta_{\text{eff}}^{(1)} & \delta_{\text{th}}^{(1)} \\
\delta_{\text{eff}}^{(2)} & \delta_{\text{th}}^{(2)} & \delta_{\text{eff}}^{(2)} & \delta_{\text{th}}^{(2)} \\
\vdots & \vdots & \vdots & \vdots \\
\delta_{\text{eff}}^{(n)} & \delta_{\text{th}}^{(n)} & \delta_{\text{eff}}^{(n)} & \delta_{\text{th}}^{(n)}
\end{bmatrix}
\begin{bmatrix}
m_{\text{eff}}^{\text{pmos}} \\
m_{\text{th}}^{\text{pmos}} \\
m_{\text{eff}}^{\text{nominal}} \\
m_{\text{th}}^{\text{nominal}}
\end{bmatrix}
\begin{bmatrix}
X
\end{bmatrix}
\]

Energy Modelling

The total energy for any operation (read, write & precharge) can be derived as:

\[
\text{Energy}_{\text{Total}} = \text{Energy}_{\text{dynamic}} + \text{Energy}_{\text{static}} = (\text{Energy}_{\text{switching}} + \text{Energy}_{\text{shortcircuit}} + \text{Energy}_{\text{glitch}} + \text{Energy}_{\text{leakage}} + \text{Energy}_{\text{pseudo-nmos}}) + \text{Energy}_{\text{control}}
\]

\[
\text{Energy}_{\text{Total}} = V_{dd} \int I_{\text{supply}} dt + \int I_{\text{device}} dt
\]

\[
+ \int \text{end} \cdot \text{glitch} \cdot \text{start} \cdot \text{transition} \cdot \text{idle} \cdot \text{time} \cdot \text{Val}=\text{V}_{dd} - \text{V}_{th}
\]

\[
= V_{dd} \left( \int I_{\text{supply}} dt + \int I_{\text{device}} dt \right)
\]

\[
+ \int \text{end} \cdot \text{glitch} \cdot \text{start} \cdot \text{transition} \cdot \text{idle} \cdot \text{time} \cdot \text{Val}=\text{V}_{dd} - \text{V}_{th}
\]

Circuit level simulators like HSPICE do not provide a direct method to estimate switching energy. We compute the integral of the current through the supply over the entire time period and over the time period there is zero-activity (idle period) and we subtract the latter from the former to compute the switching energy. This method also provides means of computing the short-circuit and static energy components accurately. The total energy of the cache is derived as:

\[
\text{Cache}_{\text{Energy}} = \left[ (E_{\text{precharge}} + E_{\text{column-active}} + E_{\text{Driver}} + E_{\text{Decoder}} + p \cdot E_{\text{(active/ce)}}) + (n-p) \cdot E_{\text{(write-active/ce)}} + (m-1) \cdot E_{\text{(idle-active/ce)}} + E_{\text{contr}} \right] + f(m,n) \cdot E_{\text{inactive-block}}
\]

where \( m \) is the number of rows, \( n \) the number of columns and \( p \) the number of active cells and \( f(m,n) \) is a function of \( m \) and \( n \).

The empirical results extracted from the simulated are then fitted to a polynomial best-fit. This is a cumbersome process as the total dimensions of the regression are very large. The dimensions are then reduced through a process called main-effect analysis as shown in the equation below.
Error analysis

The error is computed between energy and delay estimates obtained using the proposed models and Hspice. In the case of delay, the maximum percentage error oscillates between 6 and 10% while the median error is less than 2% in most cases. It has to be mentioned that the percentage error is independent of the temperature. At higher temperatures the number of access time failures increases thereby reducing the total number of successful samples for post-simulation processing.

\[
\delta_{\text{Energy}} = \left[ f(\tilde{x}_0, V_{dd-nom}, T) - f(\tilde{x}_0, V_{dd-nom}, T_{nom}) \right] \\
+ \left[ f(\tilde{x}_0, V_{dd}, T_{nom}) - f(\tilde{x}_0, V_{dd-nom}, T_{nom}) \right] \\
+ \left[ f(\tilde{x}, V_{dd-nom}, T_{nom}) - f(\tilde{x}_0, V_{dd-nom}, T_{nom}) \right]
\]

As energy has a very strong dependence on supply voltage, the simulation was performed for 500 samples across a range of 5 supply voltages and 9 temperatures. In most cases the error is well within 5% indicating the goodness of the fit. Only in the 0.7V range the percentage error is unusual which can be attributed to the non-linearity. This can be eliminated by using cubic splines of higher order polynomials.
4.1.2. Impact of process, voltage and temperature variations on delay and energy

Impact of Temperature on Access Time Delay

The first simulation (doing both Hspice and the previous macromodel) is performed only for temporal variations of temperature and spatial variations of process parameters are not taken into account. It can be seen that, with increase in temperature the difference in access time widens for every process generation. This behaviour is due to the decreasing difference in values of $V_{dd}$ and $V_{th}$ with every process generation that results in an increase in overall delay. The maximum difference in Access times is observed for 16nm, as validated by theory [16], and it is about 67% (Figure 9). This would mean that increase in temperature in the data cache would result in cycle misses thereby affecting the entire pipeline. While this would not occur always, the probability would only increase with further increase in temperature. This is when techniques like Thermal throttling and Clock gating would have to be employed to reduce the temperature which seems to take on a more ominous role when compared to power.

![Figure 9. Variation of Write Access Time versus Temperature](image)

As for the effect of temperature on Write and Read assessing them separately, in the case of a read operation (Figure 10), 16nm designs behave very close to 22nm and this would mean that if we are able to develop remedy methodologies for 22nm, we would be in a position to extend the same for 16nm designs also. However in the case of write operation there is a difference of nearly 15 per cent, considering only temperature variations, response mechanisms will have to be more aggressive with every process generation. Nevertheless, given that the write operation does not lie on the critical path, we could employ effective guard banding techniques to negate the effects of temperature variations.

![Figure 10. Variation of Read Access Time versus Temperature](image)

The next set of results in this section show the behaviour of the cache designed under the 16nm technology node.
Impact of Process Variations and Temperature on Access Time Delay

The black dots indicate the variation of access time for a gold-memory-system under temperature variations and the results for access time variation under spatio-temporal variations is indicated with coloured lines. From the above Figure 11, it can be clearly seen that less than 1% die perform better than no-variation dies under temperature constraints. While this could also point the deficiencies in the distribution adopted, delving deeper into the issue, we realize that the access time does not depend on SRAM cell variation alone but also the peripheral circuitry. It just means that, even if the SRAM array is designed to be variation-aware, it might not be the same with the peripheral circuitry. Every single block’s distribution needs to be understood at a more fine grain level to delve further deep into the distribution. Another observed phenomenon is the difference in access times for the same dies across different temperature. At lower temperatures the worst case difference is only 15% as opposed to the nearly 50% deviation at maximum temperatures. This would only mean that techniques such as speed binning which are currently used will not cater to the requirements posed by these technologies. No technique proposed at the moment would be able to compensate for a difference of the order of 50%.
At lower temperatures, each of the components is fast to respond and this can be validated by the fact that the control circuitry contributes to nearly 50% of the delay. This does not mean that the control circuitry is easily affected by the temperature. It means on a relative basis, the control circuitry which is made up of a very small number of devices decides the overall delay. As the temperature decreases, the contribution to access times due to control circuitry decreases while the other circuits are easily affected by temperature variations. When we normalized all the bars to 1, not shown here, the contribution of control circuit was only 36% at 110 °C when compared to the 50% contribution at 30 °C. It is clear that the Pre Decoder is the most affected by temperature in both cases - write and read. In the case of the read cycle, as the temperature increases the effect of temperature on the Sense Amplifier becomes important. From the above 2 inferences, we can draw the conclusion that structures that are made up of smaller number of transistors driving large loads are likely to be more susceptible to Temperature variations when compared to structures made up of more number of Transistors.

**Energy Analysis**

Figure 13 and Figure 14 present energy estimates for different technologies and different supply voltages, normalized to the value at 0.6V of 22nm tech node. It is evident from the plot that the potential benefits of scaling supply voltage is a quantum reduction in the energy consumption and with reducing feature sizes the reduction is greater.
It should be noted in this case that a significant portion of energy consumption is due to leakage and not accounting for temperature variations results in under estimation of leakage energy. While dynamic power to great extent is dependent on the supply voltage, not accounting for variation in sub-threshold leakage due to temperature variations leads to gross underestimation of total energy. While, procedures like burn-in and post-fabrication tuning have the same functionality, they require prototype chips which are seldom available at early design stages. This analysis will help alleviate this issue by providing accurate energy estimates for chips in future technologies.

4.1.3. Effect of fine-grain Vth tuning
While the previous section has analyzed the performance (in terms of delay and energy) of the cache design under process, voltage and temperature variations; in this section we analyze the use of Vth fine tuning as a first approach to reduce the spread in delay and energy consumed.

Using the energy and delay models proposed, we performed statistically 2 different circuit optimizations to evaluate their energy-delay tradeoffs. In the first case, different threshold voltage were assigned to delay and energy critical regions. While the peripheral circuitry was assigned a lower threshold, the threshold of memory array was not changed to make sure the leakage is minimal. While the delay improvement was minimal at lower temperatures, at higher temperature the speed was improved by 18%.
In the other case, the standby supply voltage of unaccessed blocks was reduced by nearly 40% while assigning multiple threshold to the periphery and memory access. Under worst case process variations while the chip is maintained at 80 degrees, the energy per access was reduced by nearly 50%.

This analysis suggests that finetunning the Vth may drastically reduce energy consumption for unused blocks (i.e. leakage) and it may help mitigate the widespread of access time delays caused by temperature variations. This opens a promising set of techniques that will be analyzed and reported in the next deliverables.

4.1.4. Conclusions

An accurate macro-model to evaluate delay and energy in a complete 32kB memory has been implemented and applied to the case of a 6T bit-cell SRAM. The impact of temperature is very high, in general for all technologies and specifically for deeper technologies. For a device sized design the read and write access time for the memory increases a 60% per a change of temperature from 25 to 110 °C. Supply voltage are also impacting, around a 15% of delay for a 10% of VDD fluctuation. Process variability, both systematic and random have been taken into account, shown a very high impact to delay. Energy consumption has also been evaluated for all the set of fluctuating variables and it has been shown that a fine-tuning of Vth may drastically reduce energy consumption.

4.2. Robustness analysis of 6T SRAM cell

4.2.1. The 6T SRAM bit-cell

A typical 6T SRAM cell uses two identical crossed coupled inverters and two access transistors as shown in Figure 16. The access transistors allow access to the cell during read and write operations and assure cell isolation during the data retention mode.

The parametric failures in SRAMs are due to systematic and random process parameter variations. The systematic variation in a parameter modifies the value of that parameter for all transistors in a die in the same direction. The random variations shift the process parameters of different transistors in a die in different directions. While circuit techniques have been developed to
compensate for the global systematic die-to-die variations [17], the local random variations cannot be as easily compensated and they have a significant effect on the SRAM yield because of the asymmetry introduced between the matched transistors of an SRAM cell [18], [19], [20]. Among the different sources of random intra-die variations, the most significant one are the threshold voltage (VTH) due to Random Dopant Fluctuation (RDF) and Line Edge Roughness (LER).

Physical failure mechanisms caused by process variations (parametric failures) in a SRAM bit-cell are [21]: hold failure, read failure, access failure and write failure. In Figure 16a, the transistors that affect each of these failures are marked and in Figure 16b, the failure mechanisms together with the correct operation are illustrated.

A **Hold Failure (HF)** can occur in data retention mode, when the supply voltage of the cell is decreased to reduce the leakage power consumption. If lowering the VDD causes the data stored in the cell to be destroyed, the cell is said to have hold failure. Asymmetry in the cross coupled inverters increases the failure probability, the hold failure probability depends on variation in any of the NL, PL, NR and PR transistors.

The **Read Failure (RF)** occurs if the data stored by the cell is destroyed during the read mode, i.e. the zero level degradation at the R node, becomes larger than the trip voltage of the inverter (NL&PL). Hence the read failure probability is affected by the variability in the pull down (NR) and access transistor (NaR) at the ‘0’ storing node as they directly affect the zero level degradation, and the pull down (NL) and pull up (PL) transistors at the ‘1’ storing node as they affect the trip voltage of the resulting inverter.

An **Access Failure (AF)** occurs during the read operation if the differential bit line voltage (∆BL = VBL – VBLB) does not reach the desired value (typically 10%VDD) during the access time (Taccess) of the cell. In read mode, the bit line BLB is partially discharged through the access transistor NaR and the pull down transistor NR (VBLB = VDD – ∆V), while the bit line BL maintains the pre-charged value (VBL = VDD). The differential bit line voltage and hence the access failure probability is affected by the variability in the two transistors.

An unsuccessful writing of the cell is referred to as a **Write Failure (WF)**. This occurs when the voltage at the ‘1’ storing node (L) does not decrease enough to flip the state of inverter (NR&PR). The write failure probability is determined by the variability in the pull up (PL) and access transistor (NaL) at the ‘1’ storing node as they affect the voltage level at node L and the pull down (NR) and pull up (PR) transistors at the ‘0’ storing node as they affect the trip voltage of the resulting inverter.

---

**Figure 16** a. The standard 6T SRAM bit – cell, b - 6T SRAM operation modes, illustrating the failures in each mode: read/access/write/hold
The failure probability of an SRAM cell \( P_{\text{cell}} \) is given by the probability that at least one of the above mentioned failures occur. Since only the failures due to random variations are considered in this analysis, it can be assumed that the failure of any cell in the memory array is independent of the failure of any other cell in the array. Hence, the SRAM array failure probability \( P_{\text{array}} \) is given by:

\[
P_{\text{array}} = 1 - \left(1 - P_{\text{cell}}\right)^N
\]

where \( P_{\text{cell}} \) is the failure probability of a single cell and \( N \) is the number of cells in the array. The parametric yield of the SRAM array is:

\[
Y_{\text{param}} = 1 - P_{\text{array}} = \left(1 - P_{\text{cell}}\right)^N
\]

### 4.2.2. Statistical analysis of the SRAM cell

Increased process variability in nano-scaled technologies is becoming a critical challenge for CMOS design. Process variability can be classified as inter-die and intra-die, and it is due to the fabrication process and to the non-uniform conditions during dopant deposition or diffusion resulting in high variability in transistor parameters.

Because of its speed and compatibility with standard logic process, SRAM is the embedded memory of choice for many VLSI systems [22]. According to ITRS 2009 [23], the area of the 6T SRAM cell will decrease from 0.11\( \mu \text{m}^2 \) to 0.012 \( \mu \text{m}^2 \) between 2010 and 2015 (SRAM density can reach over 5 billion transistors / \( \text{cm}^2 \), by 2015) [23]. In order to meet the small area requirement, SRAM bit-cells are designed with minimum (or near minimum) sized transistors, which increases their sensitivity to process variations. The major concerns regarding embedded SRAM memory as technology scales are increased static power, lower cell stability, and reduced operating margins, robustness and reliability [22]. The increase in variability aggravates all these problems, and hence statistical analysis methods considering process parameter variation become mandatory for memory robustness estimation.

Following, an overview of widely used techniques for statistical failure analysis of circuits under process variability will be presented, with special emphasis on those suitable for SRAMs.

The most common statistical failure analysis methodology for circuits under process variability is based on Monte Carlo (MC) simulation. Monte Carlo methods are a class of computational algorithms that rely on repeated random sampling to determine failure statistics. The accuracy of the statistical failure analysis depends on the sample size, so for extremely small failure probability events (as in the case of an idle SRAM cell at nominal supply voltage) the number of random experiments to accurately estimate this probability is extremely large (the sample size must be quadrupled to achieve twice the accuracy) [24].

Extensive research has been devoted to the reduction of sample size for speed improvement while maintaining the accuracy of standard Monte Carlo simulations. One of the resulting methods is the Stratified Sampling technique [25], which consists in stratifying the sample space by choosing a partition of the input parameter space. The integrals in each stratum are then estimated and combined to obtain the overall integral. Another common method of SRAM analysis is Importance Sampling [26], [27]. It is based on the fact that certain values of the input random variables have more impact on results than others. In Importance Sampling the statistical distribution function is transformed to increase the probability of occurrence of significant values. The use of biased distributions will result in a biased estimator. The simulation outputs are weighted to correct the biased distribution and ensure that the new importance sampling estimator is unbiased. The issue in implementing importance sampling simulation is the choice of the biased distribution which
encourages the important regions of the input variables. A good sample distribution can result in run-time savings and accuracy improvement [24], [26], [27], [28]. F. Gong et al. in [28] present a survey of parametric yield estimation methods applied on SRAM memories. They state that for a yield estimation close to 99.9%, assuming 3 parameters subjected to process variability, the speed up obtained by implementing Stratified Sampling vs. extensive Monte Carlo is 3.6X, while the speed up when Importance Sampling is used is 61X.

The work of R. Kanj, R. Joshi and S. Nassif [26] presents a methodology for statistical SRAM design and analysis based on Mixture Importance Sampling. Random variables using mixtures of distributions focusing on the failure region are generated. This is an adaptive technique which achieves about 100X gain in simulation speed when compared to standard Monte Carlo for yield estimation with a 95% confidence interval. The method is computationally intensive and requires complex sampling algorithms and post processing. The work of T.S. Doorn et al. [27] shows how the Importance Sampling Monte Carlo method can be used to estimate the yield loss for an SRAM array with less implementation effort by sampling more in the tails of parameter distribution. A Gaussian distribution with higher than nominal standard deviation is used to enhance the low probability tails. The density functions and distributions are corrected by a transformation based on the ratio of the original and importance sampling distribution.

The statistical failure analysis of circuits under process variability can be performed either in the performance domain or in the parameter domain.

*Performance domain* failure analysis methods obtain the metrics of interest by circuit simulations on different random parameter samples. The failure probability in the performance domain is obtained by comparing the resulting metrics with desired performance constraints and determining the percentage of failing samples from the total number of samples.

The ever increasing number of simulations required to perform an accurate failure analysis of circuits under process variability in the performance domain has motivated alternative methods based in parameter domain.

*Parameter domain* failure analysis methods estimate the acceptance region in the space of circuit parameters. In the parameter domain, the border between the acceptance region (where performance constraints are met) and the failure region is estimated, and the failure probability is determined as the ratio between the hyper-volume occupied by the failure region and the hyper-volume of the parameter domain [28].

A widely-used parameter domain technique for failure analysis is the *Most Probable Point Analysis*. It consists in dividing the parameter domain into acceptance and failure regions and finding a particular point in the parameter domain that can be related to the probability of system failure, defined by a limit state (i.e. the Most Probable Point – *MPP*) [29]. The failure probability estimation problem can be stated as finding the most likely point that causes the circuit to fail; that is, the point of maximum probability satisfying a certain failure criterion (*MPP*). The border between the acceptance and failure regions is approximated by interpolation. It can be done by a first order approximation, finding the tangent line to the border that passes through the MPP, second order approximation or spline interpolation. Another approach based on this method is the *Most Probable Failure Point* (MPFP) which was presented by D.E. Khalil et al. in [30]. The SRAM yield is analysed and the failure probability is approximated as the probability that each of the variation variables is equal or larger than its value at the most likely point. Once the MPFP is found, the acceptance region is approximated as the rectangle defined by the maximum variation, as only independent random threshold voltage variability is considered. This technique starts from
the hypothesis that SRAM cell failure is due only to independent random process variations and the failure metric is monotonic with each parameter variation. The method requires finding the combination of input variations maximizing the failure probability.

A recent proposal for statistical failure analysis in the parameter domain is the *Yield Estimation Nonlinear Surface Sampling* technique (YENSS). It was first presented by S. Srivastava and J. Roychowdhury in [31] and then improved by C. Gu and J. Roychowdhury in [32]. The method locates the failure region boundary in the parameter domain and determines the failure probability as the ratio between the area (or volume) outside the bounded region and that of the parameter domain. The boundary points are determined by a local search algorithm. This technique assumes the partition of the parameters space into $2^N$ regions ($N$ being the number of variable parameters, i.e., the dimension of the parameter domain), and uniform distribution of process parameters. It can be used for non-uniform distributions as well, by converting them into uniform distributions by the corresponding Cumulative Distribution Function (CDF). The method can also be extended to correlated parameters if principal component analysis is applied first.

In [33], F. Gong et al. present a technique to improve yield estimation efficiency, the *QuickYield* method, which proposes a yield surface boundary determination by surface-point finding and global search. The performance constraints are included in the differential algebra equation that describes the circuit, resulting in an augmented system equation.

We propose a different approach to SRAM cell statistical failure analysis in the parameter domain. The following section gives a detailed description of the method.

### 4.2.3. The Satisfiability Boundary – Statistical Integration (SB-SI) Method

With the proposed method of failure probability estimation, the boundary separating the acceptable performance and failure regions – *Satisfiability Boundary (SB)* – is found in the parameter domain, and then *Statistical Integration (SI)* is performed over the failure region to estimate the failure probability. It is a general methodology which can be applied to any circuit with a known distribution of process parameters and environmental variables.

The essence of the method consists in the decoupling between performance and statistics in the parameter domain. The failure probability is estimated in two steps: first, the Satisfiability Boundary separating the *Acceptance Region* from the *Failure Region* is found and second, a Statistical Integration over the two regions is performed for probability estimation. That is why the method is hereafter referred to as **SB-SI (Satisfiability Boundary – Statistical Integration)**. The next subsections summarize important issues and describe the above two steps.

#### A. Problem statement

Process variability has a strong influence on circuit performance and failure probability. All circuits on a die are subject to both systematic and random process variations. Systematic variations are mainly introduced during the photolithography process and affect the dimensions and threshold voltage of all transistors on a die in the same direction following a certain distribution. Random variability further affects the threshold voltage mainly because of the random nature of dopant deposition (RDF). When analysing a device under process variability, systematic variation is considered first and then random variation is superimposed. Hence, the distribution of the process parameter of the device under analysis results from the combination of systematic and random variation.
Assuming a certain device with a parameter affected by process variability, its failure probability is given as the probability of not properly performing its specified function, while the correct operation of the device is reflected by performance satisfiability.

In order to illustrate the concepts of Acceptance and Failure regions in the parameter space, let us consider parameter spaces of one single parameter (N=1) and two parameters (N=2) before presenting the general case with N parameters. Assuming only one device parameter is affected by process variability, there is a range of values for which the device meets the required performance (acceptance region in the parameter domain) whereas for the remaining values the system fails (failure region in the parameter domain). Once the two regions are identified, the failure probability is determined considering the statistical distribution of the process parameter (upper part of Figure 17).

The same applies to the two-dimensional case, in which two of the circuit parameters are subject to process variation affecting its performance (lower part of Figure 17). It also applies to the general N dimensional case.

Figure 17. One & two device parameters affected by process variability (Acceptance and Failure Regions)

Assuming a circuit with N parameters ($p = [param_1 \ param_2 \ ... \ param_N]$) subject to process variability and whose performance metric must have larger values than a given limit ($\text{Perf}(p) > P_{\text{min}}$), the acceptance and failure regions are defined as follows:

$$AR = \left\{ p \mid \text{Perf}(p) > P_{\text{min}} \right\} \quad \text{and} \quad FR = \left\{ p \mid \text{Perf}(p) < P_{\text{min}} \right\}$$

Equation 3

where $p$ is the $N$ dimensional vector of process parameters: $p = [param_1 \ param_2 \ ... \ param_N]$. The Satisfiability Boundary is the hyper-surface that separates the Acceptance Region (AR) from the Failure Region (FR). The Satisfiability Boundary (SB) is defined as:

$$SB = \left\{ p \mid \text{Perf}(p) = P_{\text{min}} \right\}$$

Equation 4

In the parameter domain, AR and FR are $N$ dimensional hyper-volumes and SB is composed of $N-1$ dimensional hyper-surfaces. The Satisfiability Boundary is searched for in the parameter domain while for the Statistical Integration the actual parameter distribution is taken into account. Parameter distribution depends on the technology node, fabrication process and design. Therefore it is important to separate the search for the Satisfiability Boundary from the Statistical Integration.
B. Satisfiability Boundary

The Satisfiability Boundary search can be very expensive and time consuming if performed by simulation, so we propose a method to estimate the SB by finding a set of boundary points (Significant Points – SP) and then interpolating to approximate the boundary surface with controllable error.

Two-Dimensional Analysis

To simplify the explanation of the method, a circuit with two \( (N = 2) \) parameters \( (p = [\text{param}_1, \text{param}_2]) \) subject to variability is considered. For illustration purposes, a 6T SRAM bit cell in data retention mode is analysed assuming that only two transistor parameters are affected by process variability.

In the parameter domain, the Satisfiability Boundary will be approximated by a polygon whose vertices (hereafter referred to as Significant Points (SP)) are obtained by simulation. An increase in the number of vertices obviously leads to a more precise boundary approximation. The first polygon that can be obtained to approximate the Satisfiability Boundary in the 2D parameter domain is a quadrilateral. In this case the Significant Points are determined by the intersection between the SB and the parameter domain axes. If the obtained quadrilateral is not a good estimate an extra set of SPs is obtained from the intersection between the SB and the lines bisecting the angles previously formed – First Angular Bisection (FAB) – (Figure 18). By further bisecting the resulting angles, more Significant Points can be determined for a more accurate approximation of the SB – Second Angular Bisection (SAB), Third Angular Bisection (TAB) and so on. Assuming the parameter domain space defined by the parameter variation, the nominal parameter point \( (p_{\text{nom}} = [\text{param}_{1\text{nom}}, \text{param}_{2\text{nom}}]) \) defines the origin of the system axes, i.e. the [0,0] point. The maximum and minimum values on the two axes are given by \((+\Delta_1, -\Delta_1)\) and \((+\Delta_2, -\Delta_2)\) respectively.

In First Angular Bisection, a set of points is generated in the parameter domain by all possible combinations of \( \Delta_{\text{param}_1} = [-\Delta_1, 0, \Delta_1] \) and \( \Delta_{\text{param}_2} = [-\Delta_2, 0, \Delta_2] \), resulting a total of \( 3^2 = 9 \) points (3 is the number of values in \( \Delta_{\text{param}}, 2 \) is the number of parameters), including the [0,0] point. The search directions for the SPs are given by the vectors connecting the origin of the system axes with the resulting 9 points. In this case, the number of directions is \( 3^2-1 = 8 \) (left side of Figure 4.2.3). The same algorithm applies for the Second Angular Bisection. The points are obtained by all possible combinations of \( \Delta_{\text{param}_1} = [-\Delta_1, -\Delta_1/2, 0, \Delta_1/2, \Delta_1] \) and \( \Delta_{\text{param}_2} = [-\Delta_2, -\Delta_2/2, 0, \Delta_2/2, \Delta_2] \), resulting a total number of \( 5^2 = 25 \) points. To avoid redundancy, the inner points are eliminated and the number of directions is given by \( 5^2-3^2 = 16 \) (right side of Figure 18). The algorithm can be extended to higher-levels of angular bisection.

Once the directions are established, the intersection points are found using a search algorithm. For this particular case, the bisection method was chosen for its robustness and simplicity. The equation \( \text{Perf}(p) - P_{\text{min}} = 0 \) is solved for the variable \( p \) (\( p \) is given by a point in the parameter domain \( p = [p_1, p_2] \)). Two initial points \( p_1 \) and \( p_2 \) are required such that \( \text{Perf}(p_1) - P_{\text{min}} \) and \( \text{Perf}(p_2) - P_{\text{min}} \) have opposite signs (in other words, one of the points must be in the Acceptance Region and the other in the Failure Region). The method divides the vector \( [p_1 \ p_2] \) into two equal parts by computing \( pm = (x_1 + x_2)/2 \) and verifies if \( pm \) satisfies \( \text{Perf}(pm) - P_{\text{min}} = 0 \). Otherwise, according to the sign of \( \text{Perf}(pm) - P_{\text{min}} \), \( p_1 \) or \( p_2 \) takes the value of \( pm \) and the algorithm continues until the desired precision is reached. The maximum absolute geometric error \( (\varepsilon_g) \) after \( n \) steps is

\[
\varepsilon_g = \left| \frac{p_1 - p_2}{2^n} \right|
\]

Equation 5
where $|p_1-p_2|$ indicates the module of the vector $p_1-p_2$. The maximum number of iterations required for certain accuracy ($e_s$) is given by

$$n = \left\lfloor \log_2 \frac{|p_1 - p_2|}{e_s} \right\rfloor - 1 \quad \text{Equation 6}$$

In Equations 5 and 6, $p_1$ and $p_2$ are the vectors with which the algorithm starts (initial points).

![Figure 18. Directions for Significant Point search in First and Second Angular Bisection](image)

The bisection method is applied separately for each direction to find all the Significant Points of the Satisfiability Boundary.

For a given error, the maximum number of simulations for each point is given by Equation 6. In all cases, the point $p_1$ is chosen to be the origin of the system axes ($p_1=[p_{11} \; p_{21}]=[0,0]$) and the vectors $p_2$ ($p_2=[p_{12} \; p_{22}]$) are given by the extremes. The first column of Tables 1 and 2 (Figure 19) contains the initial points $p_1$, $p_2$ for the bisection method; the second, the resulting Significant Points (which are also illustrated in the left figures) and the third, the modules $|p_1-p_2|$, which determine the number of simulations needed to obtain the corresponding SPs with a certain accuracy. The points marked by circles (SP1÷SP4) and triangles (SP5÷SP8) are obtained at the intersection between SB and the system axes and by first angular bisection, respectively.

The number of Significant Points, and consequently the number of simulations, increases with increasing the number of parameters, i.e. the dimension of the parameter space (N). In the 2D case under study, the Satisfiability Boundary can be approximated by the polygon obtained when applying an interpolation algorithm on the set of Significant Points (Figure 19b). The first step to obtain the polygon is to find the adjacent points, which is intuitive for the two-dimensional case but gets challenging for higher dimensions. The algorithm used in this work, which is extensible to the general N dimensional case, is described below for two dimensions.

In the quadrilateral approximation, the interpolation is straightforward, as the SPs are given by the intersection of the SB with the axes. For the first- and higher-order angular bisection approximations, the problem becomes more complex and an algorithm must be implemented. To begin with, the extreme points in the parameter domain are mapped (the +Δ values are mapped to 1 and the -Δ values are mapped to -1), as shown in Table 2 of Figure 19b. Starting the search from the [1, 1] point, half of the adjacent points are found by decrementing by 1 for each variable and the other half is found starting from the [-1,-1] point and incrementing by 1 in each direction (diagram of Figure 19b). Two points form a straight line. Thus, by connecting the adjacent points two by two the polygonal estimate of the Satisfiability Boundary is obtained.

The general equation of a straight line is given by
\[ a \cdot p_1 + b \cdot p_2 = 1 \]  

Equation 7

The \( a \) and \( b \) coefficients for each of the straight lines forming the polygon are determined by solving the system

\[
\begin{bmatrix}
P_{11} & P_{21} \\
P_{12} & P_{22}
\end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\]

Equation 8

where \((p_{11},p_{21})\) and \((p_{12},p_{22})\) are the coordinates of two of the adjacent points. After obtaining the entire set of straight lines, the boundary polygon defined by the set of all \( a \) and \( b \) coefficients is obtained.

An increase in the number of Significant Points leads to improved accuracy of the boundary approximation but also to a larger number of simulations, that is, a longer estimation time. As can be seen, there is a tradeoff between accuracy and speed.

\[
\begin{align*}
\text{Table 2 – Octagon Approximation} \\
[0,0] & \rightarrow SP1 \bullet p_0 + p_2 = \Delta_1 \\
[0,0] & \rightarrow SP2 \bullet p_0 + p_2 = \Delta_2 \\
[0,0] & \rightarrow SP3 \bullet p_0 + p_2 = \Delta_3 \\
[0,0] & \rightarrow SP4 \bullet p_0 + p_2 = \Delta_4 \\
[0,0] & \rightarrow SP5 \bullet p_0 + p_2 = \Delta_5 \\
[0,0] & \rightarrow SP6 \bullet p_0 + p_2 = \Delta_6 \\
[0,0] & \rightarrow SP7 \bullet p_0 + p_2 = \Delta_7 \\
[0,0] & \rightarrow SP8 \bullet p_0 + p_2 = \Delta_8 \\
\end{align*}
\]

Figure 19a. Choice of the Significant Points on the Satisfiability Boundary: Quadrilateral and Octagon Approximations

Once the polygonal approximation of the \( SB \) is obtained, the acceptance and failure regions are found. The condition that a random point in the parameter domain is in the acceptance region or the failure region is given by

\[
(x, y) \in AR \text{ if for all } (a, b) \quad a \cdot p_1 + b \cdot p_2 - 1 < 0 \\
(x, y) \in FR \text{ if for all } (a, b) \quad a \cdot p_1 + b \cdot p_2 - 1 > 0
\]

Equation 9

Based on the above parameter domain partition, the failure probability is determined by Statistical Integration (SI). A brief analysis of the Satisfiability Boundary approximation in the general \( N \) dimensional case is presented next.

**N Dimensional Analysis**

The steps described in the two dimensional case are also followed for the \( N \) dimensional (\( N \) parameters subject to variability, \( p = [param_1 \ param_2 \ ... \ param_N] \)) approximation of the Satisfiability Boundary. The Significant Points on the \( SB \) are determined by the same bisection method. The first approximation is performed by intersecting the axes with the \( SB \), which results in a \( 2N \) vertex hyper-polygon. Then, the angular bisection starts. Following the algorithm explained for the two-dimensional case in the previous subsection, and illustrated in Figure 18, the number of Significant Points depending on the level of angular bisection for the \( N \) dimensional approximation of the \( SB \) can be determined (Table 3). The number of elements in each \( A_{param} \) is \( (2^M+1) \), \( M \) being the order of angular bisection. The resulting number of points assuming \( N \)
parameters subjected to variability is \((2^M+1)^N\). To avoid redundancy the inner points are eliminated \((-2^{M-1})^N\) and the total number of search direction for the SPs is obtained. The algorithm of finding the adjacent Significant Points on the Satisfiability Boundary for the first angular bisection in a 3D analysis is the same as for the 2D analysis. The search starts from the \([1,1,1]\) point, half of the adjacent points are found by decrementing by 1 for each variable and the other half are found starting from the \([-1,-1,-1]\) point and incrementing by 1 for each variable (diagram in Figure 20). The hyper-polygon estimation of the Satisfiability Boundary for different levels of angular bisection in a 3D analysis is illustrated in Figure 21.

Once the Significant Points are found, the SB can be approximated by a hyper-surface. The adjacent points must be connected similarly to obtain the hyper-surface estimating the SB. Equations 7 and 8 become:

\[
a_1 \cdot p_1 + a_2 \cdot p_2 + \cdots + a_N \cdot p_N = 1
\]

\[
\begin{bmatrix}
p_{11} & \cdots & p_{N1} \\
\vdots & \ddots & \vdots \\
p_{1N} & \cdots & p_{NN}
\end{bmatrix} \cdot \begin{bmatrix}
a_1 \\
\vdots \\
a_N
\end{bmatrix} = \begin{bmatrix}
1 \\
\vdots \\
1
\end{bmatrix}
\]

The rules in Equation 7, for the N dimensional case become:

\[
(x_1 \cdots x_N) \in AR \text{ if for all } (a_1 \cdots a_N) \sum_{i=1}^{N} a_i \cdot p_i - 1 < 0 \tag{12}
\]

\[
(x_1 \cdots x_N) \in FR \text{ if for all } (a_1 \cdots a_N) \sum_{i=1}^{N} a_i \cdot p_i - 1 > 0
\]

As previously mentioned, the proposed method of failure probability estimation consists in finding the Satisfiability Boundary and then Statistically Integrating the probability density function over the failure region to estimate the failure probability. The statistical integration is described in the following subsection.

<table>
<thead>
<tr>
<th>Approximation</th>
<th>Number of Significant Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>2N vertices hyper-polygon</td>
<td>2N</td>
</tr>
<tr>
<td>M-th Angular Bisection</td>
<td>((2^{M+1})^N - (2^{M-1})^N)</td>
</tr>
<tr>
<td>2D</td>
<td>8</td>
</tr>
<tr>
<td>3D</td>
<td>26</td>
</tr>
<tr>
<td>4D</td>
<td>80</td>
</tr>
<tr>
<td>6D</td>
<td>728</td>
</tr>
<tr>
<td>8D</td>
<td>6560</td>
</tr>
<tr>
<td>10D</td>
<td>59048</td>
</tr>
<tr>
<td>First Angular Bisection</td>
<td>16</td>
</tr>
<tr>
<td>2D</td>
<td>80</td>
</tr>
<tr>
<td>3D</td>
<td>544</td>
</tr>
<tr>
<td>4D</td>
<td>14897</td>
</tr>
<tr>
<td>6D</td>
<td>384065</td>
</tr>
<tr>
<td>8D</td>
<td>9706577</td>
</tr>
<tr>
<td>Second Angular Bisection</td>
<td>32</td>
</tr>
<tr>
<td>2D</td>
<td>242</td>
</tr>
<tr>
<td>3D</td>
<td>4160</td>
</tr>
<tr>
<td>4D</td>
<td>102024</td>
</tr>
<tr>
<td>6D</td>
<td>5374176</td>
</tr>
<tr>
<td>8D</td>
<td>272709264</td>
</tr>
</tbody>
</table>

Table 3. Number of SP in N Dimensional Analysis for Different Levels of Angular Bisection
C. Statistical Integration

The parameter joint probability distribution is used to estimate the probability of satisfying the specifications. Assuming a multivariate distribution of $N$ random variables, the joint (cumulative) distribution function is a positive, real-valued function, given by:

$$F(x_1, x_2, \cdots, x_N) = P(p_1 \leq x_1, p_2 \leq x_2, \cdots, p_N \leq x_N)$$  \hspace{1cm} \text{Equation 13}
where \( p_1, \ldots, p_N \) are the random variables under analysis. The probability density function is given by \( f(p_1 \ldots p_N) \). The probability that the parameter variables lie in predefined ranges \( x_i < p_i < y_i, \ i=1 \ldots N \) between certain limits is given by

\[
P(x_1 \leq p_1 \leq y_1, \ldots, x_N \leq p_n \leq y_N) = \int_{x_1}^{y_1} \cdots \int_{x_n}^{y_n} f(p_1 \cdots p_n) \, dp_1 \cdots dp_n
\]

Equation 14

The parameter domain is an N-dimensional hyper-rectangle. Given the complex shapes of the two regions (AR and FR) a computational strategy in needed to determine the value of \( P \) in Equation 12. As integration over a regular, well defined space is straight forward and it is relatively easy to consider correlation [34], the parameter domain must be divided into hyper-rectangles. Three regions are thus obtained, i.e. Hyper-rectangle Acceptance Region (HAR), Hyper-rectangle Failure Region (HFR) and Hyper-rectangle Satisfiability Boundary (HSB) as shown in Figure 21 for the two-dimensional case. The partition is performed using a bisection-like method. The left side of Figure 21 illustrates the first step of space division: each edge of the initial rectangle is divided in half and, by connecting the resulting points, four rectangles are obtained.

In order to determine to which of the three regions these rectangles belong, the positions of the vertices are checked using Equation 7 (or Equation 12 for the N-dimensional case):

- if all vertices are inside the acceptance region, the rectangle (R) is in HAR;
- if all vertices are inside the failure region, the rectangle (R) is in HFR;
- if there are vertices both in the acceptance and failure regions, the rectangle (R) is in HSB.

Rectangles in HAR and HFR are left untouched whereas rectangles in HSB are further divided (Figure 21).

Once the space division achieves the required accuracy, the statistical integration is performed by integrating hyper-rectangles in the parameter domain using the statistical probability density function. By integrating the probability density function on each hyper-rectangle, the following probabilities are computed:

\[
P_{\text{fail}} = \sum_{i=1}^{n_f} P(R_i \in \text{HFR})
\]

\[
P_{\text{accept}} = \sum_{i=1}^{n_a} P(R_i \in \text{HAR})
\]

\[
P_{\text{SB}} = \sum_{i=1}^{n_b} P(R_i \in \text{HSB})
\]

Equation 15

where \( n_f \) is the number of hyper-rectangles in the Failure Region, \( n_a \) is the number of hyper-rectangles in the Acceptance Region and \( n_b \) is the number of hyper-rectangles on the Satisfiability Boundary.
The probability $P_{SB}$ determines the accuracy of the statistical integration. The space division continues until the desired accuracy is achieved (i.e. 0.1% accuracy is obtained when $P_{SB} = 0.1\%P_{accept}$). Different applications have different requirements in terms of accuracy and speed of the failure probability estimation. For example, when comparing designs the failure probability must be estimated as fast as possible without very strong restrictions on accuracy, as only relative values are important. On the other hand, for yield loss estimation, the failure probability must be determined both quickly and very accurately. There are two ways to improve the accuracy of the SB-SI method: by finding more points on the Satisfiability Boundary and by making the HSB region as small as the accuracy requires. The following section describes the application of the SB-SI method for SRAM failure probability estimation in data retention mode. Accuracy and speed results are compared with Monte Carlo simulation.

4.2.4. Statistical Robustness Analysis of the 6T SRAM Cell

The parametric yield estimation of a SRAM array is based on determining the failure probability in hold/read/write/access modes using dynamic failure criteria assuming random threshold voltage variation ($p = \Delta V_{TH}$). In read mode, the access transistors are turned on by increasing the word line voltage to VDD. The bit lines (BL and BLB) are pre-charged to VDD. During the read operation the bit line (BLB) corresponding to the ‘0’ sorting node discharges through the access transistor (NaR) and the pull down transistor (NR) – Figure 16. In the read mode, two failures can occur: read failure and access failure. The Read Failure (RF) occurs if the data stored by the cell is destroyed during the read mode (Figure 19). An Access Failure (AF) occurs if the differential bit line voltage ($\Delta BL = VBL – VBLB$) does not reach the desired value (typically 10%VDD) during $T_{access}$ (Figure 16b).

In write mode, the access transistors are turned on by increasing the word line voltage to VDD. The bit line voltage of BL is set to 0V, while the voltage of BLB is set to VDD. During the write operation, the node storing ‘1’ (L) discharges through the bit line (BL). Eventually the data stored by the cell flips – Figure 16a. An unsuccessful writing of the cell is referred to as a Write Failure (WF).

In data retention mode, the access transistors are turned off, and the supply voltage is decreased down to a certain value (VDDlow) in order to reduce static power consumption – Figure 16b. If lowering the VDD causes the data stored in the cell to be destroyed, the cell is said to have Hold Failure (HF).

The performance metric in read, write and hold failure analysis is given by the voltages at the nodes L and R (VL and VR): $\text{Perf} = VL – VR$. The performance metric in access failure analysis is
given by the bit line voltages (VBL and VBLB): Perf = VBL – VBLB. Assuming that the cell stores a ‘0’ at the R node the failure criteria is given by:

- Hold Failure: Perf(ΔVTH) < 0
- Read Failure: Perf(ΔVTH) < 0
- Write Failure: Perf(ΔVTH) > 0
- Access Failure: Perf(ΔVTH) < 10%VDD

Once the failure criteria are established, the Satisfiability Boundary is determined for each of the four failure mechanisms.

### 4.2.5. Results

The simulation environment is HSPICE and the SRAM cell is designed using Predictive Technology Model (PTM) transistors [4] and TRAMS WP1 transistors. As the SRAM cell design is symmetrical, the application of certain design rules reduces systematic process variability significantly. That is why only random process variability is considered in the present analysis. The combination of random dopant distribution and line edge roughness has an impact on threshold voltage variability. A 6σ upper bound in the variability range of the threshold voltage is a realistic assumption for CMOS technologies (max(|ΔV_{TH}|)=6σ). The transistors dimensions, their threshold voltage and the random threshold voltage variability are summarized in Table 4, for the technology nodes considered for this analysis.

<table>
<thead>
<tr>
<th></th>
<th>Pull - up MOS</th>
<th></th>
<th>Pull - down MOS</th>
<th></th>
<th>access MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>45nm</td>
<td>4</td>
<td>1.1</td>
<td>94</td>
<td>0.587</td>
<td>4</td>
</tr>
<tr>
<td>32nm</td>
<td>6</td>
<td>1</td>
<td>67</td>
<td>0.581</td>
<td>6</td>
</tr>
<tr>
<td>22nm</td>
<td>8</td>
<td>15</td>
<td>46</td>
<td>0.637</td>
<td>8</td>
</tr>
<tr>
<td>16nm</td>
<td>10</td>
<td>30</td>
<td>33</td>
<td>0.686</td>
<td>10</td>
</tr>
<tr>
<td>18nm</td>
<td>33</td>
<td>20</td>
<td>40</td>
<td>0.202</td>
<td>33</td>
</tr>
<tr>
<td>13nm</td>
<td>39</td>
<td>58</td>
<td>20</td>
<td>0.202</td>
<td>39</td>
</tr>
</tbody>
</table>

Table 4. Variability Table for different technology node and different variability degrees – the threshold voltage standard deviation in volts for each transistor in each variability scenario is marked in blue; the situations when 6σ variability margin exceeds the nominal value of the threshold voltage are marked in red. Values correspond for minimum size devices (W/L=1), so results are and overestimation of properly sized cell.

### A. Static analysis

A common way to reduce the static power consumption of an SRAM array is to decrease its supply voltage when in memory retention mode (idle). The value to which the supply voltage can be reduced is limited by the minimum voltage at which a cell retains stored data – Data Retention Voltage (DRV). However, process variability affects the DRV value, limiting the possibility to decrease static power consumption. Another problem that can arise is the temperature effect on cell stability in low leakage power mode.
Different proximities of the idle memory block to the active logic units lead to significant spatial temperature variations. This section analyses SRAM robustness at different supply voltages and temperatures.

The conventional way to analyse the robustness of an SRAM bit cell is to quantify its immunity to noise. The noise margin is commonly defined as the maximum noise signal tolerated by a device used in a system while still operating correctly. If the noise is represented by two opposite sign DC voltage sources at the internal nodes of the SRAM cell, the Static Noise Margin (SNM) is implied [35].

Graphically, the SNM is determined by drawing and mirroring the Voltage Transfer Characteristics (VTCs) of the two cross-coupled inverters in the 6T SRAM cell, thus obtaining the so-called Butterfly Curve. The maximum square which can be inscribed in the loops of the butterfly gives the value of the SNM (Figure 23a) [35]. Under process variability, an asymmetric transistor configuration causes the butterfly curve to become asymmetrical, and in this case the SNM is given by the smaller of the two maximum squares (Figure 23b). The supply voltage also has a strong influence on the SNM as the butterfly curve shrinks with decreasing the supply voltage, rendering the cell more sensitive to noise (Figure 23c). The effect of the operating temperature is also illustrated in Figure 23d, where a narrowing of the butterfly curve can be observed when the temperature increases. Figure 23 shows the importance of analysing the robustness of the SRAM bit cell in data retention mode considering the joint effects of process variability, supply voltage scaling and temperature variation.

The simulations are performed on a 6T SRAM bit cell like the one illustrated in Figure 16. In data retention mode the access transistors are off as the bit lines (BL and BLB) and the word line (WL) are connected to ground. In order to achieve low static leakage power, the supply voltage must be decreased. However, this leads to a reduction in robustness, and hence an increase in failure probability. The SB-SI method is applied to determine the robustness and failure probability of an SRAM cell for different values of the supply voltage. The performance metric under study is the Static Noise Margin $Perf = SNM$. The number of Significant Points in the first angular bisection approximation for the four-dimensional case is 80. These points are determined with a maximum absolute error of 1% for different supply voltages and temperatures. Starting from these points, the hyper-plane approximations of the SBs are determined by linear interpolation. Applying Equation 10 to the four-dimensional case the hyper-rectangular division is performed with a minimum absolute error of 1% (given by the maximum size of the hyper-rectangles on the SB). Figure 24 illustrates the dependence of SRAM cell failure probability for different values of the supply voltage in data retention mode only for 45nm, 32nm, 22nm and 16nm Predictive Technology Model transistors.
The probabilities $P_{\text{no\_acc}} (\text{Perf} = \text{SNM}_{\text{min}} = 10\% V_{\text{DD}})$ and $P_{\text{fail}} (\text{Perf} = \text{SNM}_{\text{min}} = 0)$ are determined and illustrated in Figure 4.2.9 as a function of supply voltage decrease and temperature variation for a 16nmPTM SRAM cell.

Following, the SB-SI method is used for failure probability estimation in read/access/write/hold modes, for the six technologies in Table 4.

**B. Dynamic Analysis**

Given a 6T SRAM cell, the performance metric can be estimated depending on the threshold voltage variation of the six transistors: $\text{Perf}(\Delta V_{th})$. For illustration purposes, the 6T SRAM bit cell is analysed assuming only two transistor parameters are affected by process variability (NR&PL in Figure 16). Simulations are performed in the parameter domain for Satisfiability Boundary
estimation. The cell access time is \( T_{\text{access}} = 1\text{ns} \) and the supply voltage in memory retention mode (data retention voltage) is \( V_{\text{DDlow}} = 500\text{mV} \). The cell access time determines the operation speed of the SRAM. The \( V_{\text{DDlow}} \) is the value of the cell’s supply voltage in data retention mode and determines the degree of static power reduction.

Transient simulations are performed for each of the operation modes and the Significant Points are obtained by the bisection method. The Satisfiability Boundary is approximated using the first angular bisection (FAB). The points are obtained with a maximum geometrical error of 0.1%. The obtained Satisfiability Boundaries in read, write, access and hold modes are illustrated in Figure 26. An SRAM cell fails if at least one of the above mentioned failures occur. Hence from the individual Satisfiability Boundaries the overall Satisfiability Boundary can be obtained. For each of the search directions the Significant Point closest to the origin of the parameter domain is part of the overall Satisfiability Boundary (Figure 26). The same applies for the six-dimensional case when all transistors in the SRAM cell are submitted to threshold voltage variation.

a. – Special case: only NR & PL transistors (Figure 4.2.1) subjected to process variability

b. – Special case: only NR & NL transistors (Figure 4.2.1) subjected to process variability

Figure 26. Satisfiability Boundaries in Write/Access/Hold/Read modes and the Overall Satisfiability Boundary – illustrated for 45nm PTM SRAM cell (2D case)
The failure probability of an SRAM cell is dependent on speed ($T_{\text{access}}$) and leakage power ($V_{\text{DDlow}}$) requirements. Figure 27 illustrates the Satisfiability Boundaries in read/access/write/hold operation modes when the access time ($T_{\text{access}}$) increases from 500ps to 1ns and the data retention voltage ($V_{\text{DDlow}}$) increases from 300mV to 500mV. The arrows show the direction towards failure probability increase.

Figure 27 Satisfiability Boundaries in Write/Access/Hold/Read modes for different values of $T_{\text{access}}$ and $V_{\text{DDlow}}$ – Special case: only NR & PL transistors (Figure 4.2.1) subjected to process variability

Figure 28 illustrates the overall satisfiability boundary in a 2D case (only the NR & PL transistors subjected to process variability) for four technology nodes (45nm, 32nm, 22nm, 16nm PTM). It can be observed that the Acceptance Region decreases with technology scaling. This, together with the higher variability associated to smaller technologies (Table 4) results in a higher cell failure probability.

Figure 28. Overall SRAM cell Satisfiability Boundary for 16nm/22nm/32nm/45nm (PTM) technology nodes – Special case: only NR & PL transistors (Figure 4.2.1) subjected to process variability
Using the SB-SI algorithm the failure probabilities of six SRAM cells (45nm PTM, 32nm PTM, 22nm PTM, 16nm PTM, 18nm TRAMS WP1, 13nm TRAMS WP1) are determined, assuming the same access time $T_{\text{access}} = 1\text{ns}$, and the same data retention voltage $V_{\text{DDlow}} = 0.5V$. The results are summarized in Table 5. For the PTM models, the low power transistors are chosen. The transistor sizes and the threshold voltage variability are the ones in Table 4.

<table>
<thead>
<tr>
<th>PTM</th>
<th>Hold ($P_{\text{hold}}$)</th>
<th>Write ($P_{\text{write}}$)</th>
<th>Read ($P_{\text{read}}$)</th>
<th>Access ($P_{\text{access}}$)</th>
<th>Cell ($P_{\text{cell}}$)</th>
<th>Array ($P_{\text{array}}$)</th>
<th>Yield (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>45nm</td>
<td>4%</td>
<td>0</td>
<td>1.03e-9</td>
<td>0</td>
<td>2.41e-9</td>
<td>2.87e-9</td>
<td>1.5e-3</td>
</tr>
<tr>
<td>32nm</td>
<td>6%</td>
<td>9.73e-8</td>
<td>2.91e-8</td>
<td>3.34e-8</td>
<td>7e-8</td>
<td>1.02e-7</td>
<td>5.21e-2</td>
</tr>
<tr>
<td>22nm</td>
<td>8%</td>
<td>1.06e-7</td>
<td>1.32e-7</td>
<td>9.47e-8</td>
<td>1.31e-7</td>
<td>2.16e-7</td>
<td>0.107</td>
</tr>
<tr>
<td>16nm</td>
<td>10%</td>
<td>5.14e-7</td>
<td>6e-7</td>
<td>2.91e-7</td>
<td>3.33e-7</td>
<td>1.11e-6</td>
<td>0.441</td>
</tr>
<tr>
<td>18nm</td>
<td>33%</td>
<td>9.81e-6</td>
<td>2.43e-5</td>
<td>1.06e-5</td>
<td>1.53e-5</td>
<td>4.81e-5</td>
<td>1</td>
</tr>
<tr>
<td>58%</td>
<td>6.3e-3</td>
<td>1e-2</td>
<td>4.34e-3</td>
<td>6.25e-3</td>
<td>1.42e-2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13nm</td>
<td>39%</td>
<td>2.62e-5</td>
<td>4.19e-5</td>
<td>2.83e-5</td>
<td>3.33e-5</td>
<td>8.32e-5</td>
<td>1</td>
</tr>
<tr>
<td>58%</td>
<td>4.66e-2</td>
<td>7.1e-2</td>
<td>2.04e-2</td>
<td>5e-2</td>
<td>1.21e-1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5. Failure Probability in hold/write/read/access modes of the 6T SRAM cell for different technology nodes, the overall SRAM cell failure probability, the array failure probability (4.2.1) and the corresponding parametric yield (4.2.2) – the zero parametric yield scenarios are marked in red.

The data in Table 5 shows an important decrease in parametric yield when the variability in threshold voltage increases. For instance, a 512k SRAM array designed in a 22nm PTM technology node has 89.29% parametric yield for moderate variability ($\sigma = 8\%$), while for very high variability ($\sigma = 30\%$) the parametric yield decreases to 0. Figure 29 illustrates the parametric yield for the SRAM array assuming the four selected PTM models and moderate threshold voltage variability.

![Figure 29. – Parametric Yield for different technologies and threshold voltage variability at different scanarios.](image)

The number of simulations needed to obtain the Satisfiability Boundary approximation and hence the failure probability for each operation mode for each of the 6 technologies is lower than 6000 while using standard Monte Carlo for failure probability estimation, more than 1e7 simulations are needed to obtain the same probability. Another important advantage of the SB-SI method is that
only one Satisfiability Boundary needs to be estimated regardless of the threshold voltage level of variability, while is standard MC is used, the simulations have to be repeated for each variability scenario.

4.2.6. Conclusions

We present the statistical SB-SI method to estimate circuit failure probability by finding the boundary (SB) separating the acceptable performance region (AR) from the failure region (FR) in the parameter domain, and then by statistically integrating (SI) the probability density function over the failure region. The method is applied for failure probability and robustness estimation of a 6T SRAM cell in data retention mode, under process, voltage and temperature variation and also for parametric yield estimation.

The accuracy of the proposed SB-SI method can be easily improved by increasing the number of Significant Points on the Satisfiability Boundary to be determined and the precision with which these points are found (though this method offers only slight a increase in accuracy), and by further dividing the HSB region. The proposed method is considerably faster than the standard Monte Carlo, especially when very small failure probabilities have to be determined, as is the case with the SRAM bit cell.

As an application of the SB-SI method, the failure probability and robustness of a 16nm PTM 6T SRAM cell in data retention mode are obtained for different supply voltages and temperatures (Figure 24). The minimum supply voltage ensuring data retention and robustness can be accurately determined. For the particular case of the 16nm PTM 6T SRAM cell used in this analysis, under process and temperature variability the minimum supply voltage for which data is retained is 0.68V (75.5% of nominal V_{DD}). However, to ensure a noise margin higher than 10%V_{DD}, the maximum allowed decrease in the supply voltage is 8.9% from nominal V_{DD} to 0.82V.

Also, the parametric yield of several SRAM arrays was determined. For a 45nm PTM SRAM array, the parametric yield is 99.85% when the access time is 1ns ad the data retention voltage is 0.5V. For a 32nm PTM SRAM array, under the same conditions, a 5.06% decrease in parametric yield is observed for the moderate threshold variability scenario (σ = 6%) and a 33.88% decrease for the high variability scenario (σ = 15%). The parametric yield decreases even more as the technology scales down to 22nm. A decrease of 10.6% is observed for the moderate variability scenario (σ = 6%) while for the high variability scenario (σ = 15%) the decrease is of 93.06%. With technology scaling the SRAM parametric yield decreases rapidly.
5. Si-MOS Dynamic RAMs (DRAM) cells and systems

This section is devoted to the analysis and evaluation of the impact of variations on DRAM circuits. Two DRAM cells have been selected for such analysis: the 1T1C and the 3T1D cells. Section 5.1 analyses the effect on access times and energy consumption of a complete set of 32 cells of 1T1C memory. Section 5.2 introduces the basics of the 3T1D memories, a promising candidate to replace 6T-SRAM cells in L1 data caches. Section 5.3 analyses the robustness of the 3T1D cell, following a parallel analysis that the one performed in section 4.2. Section 5.4 introduces reliability issues in the 3T1D cell, analysing the effect of aging in the access and retention time to the cell as well as the impact on the memory yield.

5.1. Modelling and analysis of energy and delay of 1T1C DRAM including technology nodes, process, voltage and temperature variations.

Dynamic RAM is increasing its use because of high component density, low error rate and low power consumption. Examples of current use of eDRAM are in the IBM’s BlueGene/L supercomputer, XBOX 360, PlayStation, Wii and iPhone.

The main advantage of the eDRAM over the rest of memory cells is the highest capacity in the same area. In the case of eDRAM, it provides from four times up to eight times higher capacity in the same area because of its smaller one transistor, one capacitor structure compared to the six-transistor SRAM cell.

The drawback of logic eDRAM is the need of additional masks, but recent designs obtain a right compromise between added process cost and memory density, and logic based DRAM allows the use of the standard cell libraries and cores, which is essential for SoC. Also, logic eDRAM permits system designers to use high bandwidth, due to the gap-reduction in the speed between SRAM and logic DRAM.

PVT variations, as well as wear-out continue to strongly influence delay, and their impact on power is increasing dramatically as leakage dominates. That is, variability is a key limiting factor to this type of memories too. In this section the effect of the PVT variation in logic eDRAM in the foreseeable future technologies is studied.

5.1.1. Circuit description

For the analysis of the eDRAM it has been selected a circuit published by J. Barth et al. in [3] which fulfils the specifications listed above. Although this proposal was made on SOI, the study has been designed on bulk technology.

One difficulty has been the adequate sizing of each one of the transistors for all the technologies used guaranteeing an acceptable performance in every node. The size was obtained for nominal conditions, and allowed large margins to ensure proper operation under hostile conditions.

As transistor model it has been used the available high performance Predictive Model Technology from Arizona State University [4], with models for 45 nm, 32 nm, 22 nm and 16 nm technology. The variability levels considered are the indicated in the introduction: moderate, high and very high variability.
To size the transistors, the delays for 65nm published by J. Barth [3] have been taken as reference. Using 65nm PTM models, the delays have been matched, and as additional rule, minimal size has been forced into the cell transistors, which allows for minimum cell size. For 45nm and below technologies, linear scaling rules have been used for sizing transistors and parasitic. To maximize retention time in each technology, capacitor cell size has been optimized individually.

To minimize the effect of the smaller sense signal available in DRAM, the circuit incorporates three sense amplifier levels, which also allows reducing bitline length to an equivalent of 32 cells. The low level sense amplifier (µSA) is composed by only 3 transistors, which allows minimal overhead.

The schematic uses three levels of sense amplifier, µSA, SSA and a third sense level (TSA), not shown in Figure 30. The TSA sense amplifier is a classical sense amplifier, and very similar to the used in the 6T SRAM circuits. For similar arrays and loads, all other delay components (address-in, decode, wordline drive, signal routing, and I/O bus) are fundamentally not very different. For this reason, we only incorporate in this study the circuit related to one TSA, which drives eight SSA, and each SSA drives eight µSA, and for each SSA are available eight cells, given a total of 512 cells simulated. Also, we have included the parasitic effects in the main signals, and an RC model for the capacitor cell extracted from [36]. All parasitic effects have been scaled for each of the technologies.

An additional improvement in this circuit is the non-destructive read cycle due to the action of the µSA transistors in the last period of the read cycle for both logical values.

5.1.2. Effects of temperature
This section shows the impact of the temperature obtaining the results from electrical simulations using HSPICE as tool. A margin from 25 °C to 110 °C has been considered. In every point temperature-technology (45nm, 32nm, 22nm and 16nm), Gaussian Monte Carlo analysis has been used, and 500 simulations were obtained to guarantee a good confidence.

5.1.3. Read-Write access time and retention time
The circuit studied uses several synchronous signals, which implies a continuous re-dimension of the time slots.
Figure 31 shows the results obtained for both basic access times (read and write), and the retention time for the four technologies versus temperature, without changing any other parameter. The expected results were that as technology shrinks, both cycles reduce their access time, but surprisingly, the 16 nm has the worst behaviour in all cases. The reason can be explained due to the trade-off between the access time and the retention time, both given by the cell capacitor, the size of the transistor cell and the transistors size of the µSA.

The size of the capacitor cell has been chosen taking into account minimum size for transistors of the cell and µSA, and maximizing the retention time up to the value that the increase of retention is no or almost negligible. As technology dimension reduces, retention time decreases, due to the increase of the leakage power, as we can see in the next section.

It can be observed a great dependence of the three measures with the temperature, especially with the retention time, which can change from 0.267ms at 25ºC, to 39.1 µs at 100ºC in 45nm.

**5.1.4. Power consumption**

Figure 32 shows the average and static power for 512 cells of eDRAM (16 sets of 32 bits each one), with all the sense amplifiers included. The rest of the circuit is similar to that used in SRAM and has not been included.

The measure of average power consumption was obtained by performing two consecutive full read and write cycles, for both logical values, and measuring the power delivered by all the sources.
According to the expected results, the global average power decreases as technology shrinks, but leakage power increases.

![Average Power](image1)
![Static Power](image2)

*Figure 32. Impact of Temperature on the power consumption for a 1T1C circuit*

Low dependence of the average power and high dependence of the static power with the temperature can be observed, especially in 22nm and 16nm.

### 5.1.5. Effects of Process Variability

In this section the effects of variability in access times and energy are analysed. As mentioned above, three scenarios have been taken into account, moderate variability, high and very high.

#### 5.1.5.1 Read-Write access time and retention time

In Figure 33, the effects of process variability for read and write delays, and retention time are shown. All the graphs show the relative variance. That is, the result of the variance, divided by the nominal value, depending on the temperature ($\sigma$/nominal).
As expected, the smaller variability in the delay occurs at the 45nm node for the three cases. Considering only the moderate scenario, the result is spectacular for 22nm in the retention time, which becomes 60% for one $\sigma$. Obviously, for the 16nm node variability in all access times is very large, reaching values of 50%, 90% and 200% for one $\sigma$.

In the high variability scenario, the effect of variability increases. We highlight the write cycle graph for 16nm node, which gives an irregular result because there are values that failed, and have not been included in the graph.

The worst case occurs for the high variability, with errors observed over 60% in read, 40% write, and 20% in retention time. As an example, the failure rate for read access in the following table (using a 500 sample Montecarlo experiment). It is emphasized that despite the extremely delayed control signals, in these cases the memory cell will not work properly.
### 5.1.6. Effects of Voltage variation

The measurement of voltage variation has been made without including the process variation. It has been considered a +/-10% effect due to RI and Ldl/dt drops. The following figures show the results for the four nodes, also depending on the temperature. The nominal voltage for the 45nm node, 32nm, 22nm and 16 nm are: 1 volts, 0.9 volts, 0.9 volts and 0.7 volts, respectively.

---

<table>
<thead>
<tr>
<th>Temp . (°C)</th>
<th>45nm</th>
<th>32nm</th>
<th>22nm</th>
<th>16nm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>standard</td>
<td>moderated</td>
<td>high</td>
<td>moderated</td>
</tr>
<tr>
<td>25</td>
<td>0 %</td>
<td>0 %</td>
<td>0.2 %</td>
<td>0 %</td>
</tr>
<tr>
<td>40</td>
<td>0 %</td>
<td>0 %</td>
<td>0.2 %</td>
<td>0 %</td>
</tr>
<tr>
<td>60</td>
<td>0 %</td>
<td>0 %</td>
<td>0.2 %</td>
<td>0 %</td>
</tr>
<tr>
<td>80</td>
<td>0 %</td>
<td>0 %</td>
<td>0.2 %</td>
<td>0 %</td>
</tr>
<tr>
<td>100</td>
<td>0 %</td>
<td>0 %</td>
<td>0.2 %</td>
<td>0 %</td>
</tr>
</tbody>
</table>

Table 6. Failure rate for read access
The most important change is due to temperature, especially for the write cycle. The peaks obtained in read charts of 16nm and 22nm, are caused by the precharge cycle, which if set correctly, would produce the same trend as in the rest of the graph. The most striking result is the u-shaped write graph for 45nm. It is because the capacitor is considered correctly loaded when the cell voltage reaches 80% of Vdd, and in this case, the correct supply voltage matches the nominal value.
5.1.7. Conclusions

The effects of PVT on foreseeable future technologies on 1T1C cell have been investigated. As MOS device models PTM have been used. Finally the global effects are presented.

The next graphs summarize the results. In all of them the additional delays (in percentage) for the write and read cycles induced by the PVT effects have been shown. It has also been included as reference, the nominal value. Both scenarios, moderated and high, are drawn in different charts, but the scenario “very high” has not been drawn due to the high failure rate. In Figure 37 the worst case has been drawn for temperature, process and voltage variability, but as process and voltage variability have low dependence on temperature in read/write cycles (see Figure 33), it can be used as upper limit.

In the worst case, for a 16nm technology and a moderate scenario, the delay in the write cycle increases by 155%. The case of read cycle is worse, for the same technology, surprisingly increased by 500%, and only in the moderated scenario.

![Figure 35. PVT effects on Write cycle for moderate scenario](image)

![Figure 36. PVT effects on Write cycle for high scenario](image)
The effects on the retention time have not been included, but the retention time variation could get to changes of 197%, 297%, 488% and 626% for 45nm, 32nm, 22nm and 16nm, respectively.

If we consider the "very high" scenario, the failure rate grows up to 26% for 22nm, and 59% for 16nm in the read cycle. Clearly, this failure rate is not acceptable.
5.2. Introduction to the 3T1D DRAM cell

The 3T1D cell was proposed by Luk and Dennard (IBM) in 2005 [2] [37], as a promising DRAM memory, compatible with conventional CMOS technologies. In fact an original evolution of the earlier version of the 3T cell was presented by Intel in 1970 [38].

![Schematic of the 3T1D DRAM cell.](image)

In the case of the 3T1D the specific storage capacitor of 3T is substituted with a MOS capacitance incorporated as gated-diode. Figure 39 shows the scheme of this basic cell. The basis of the storage system is the charge placed in node S, written from BLwrite line when T1 is activated. Consequently it has a dynamic nature, but allows a non-destructive read process (an advantage in comparison with 1T1C) and high performance read and write operations, comparable with 6T. Because the whole cell has only 4 NMOS transistors (T1 and T3 work as accessing devices), in comparison with 6T with 6 transistors, the area of a 3T1D is around 40% lower than the corresponding for the 6T with equivalent performances, so 3T1D is more compact. In order to write the cell at the BLwrite line level it is only required to activate T1 through the word line write WL_write. Hence, the S node stores a 0 voltage or a \( V_{DD} - V_{TH} \) voltage depending on the logic value. The voltage value is stored on the D1 and T2 gate capacitances.

In order to read the cell, the read bitline BL_read has to be previously precharged at \( V_{DD} \) level. Then T3 is activated from WL_read line. From a 1 (high) level stored in S, transistor T2 turns on and discharges the bitline. If a 0 (low) level is stored in S, transistor T2 does not reach enough conduction level. The objective of the gated diode D1 is to improve access time. When a 1 (high) level is stored in S, D1 connected to WL_read line causes a boosting effect on the voltage level in node S. This voltage is close to VDD voltage causing a fast discharge of the parasitic capacitance in BL_read.

5.3. Modelling and analysis of access and retention time of a 3T1D 32kB cache memory including technology nodes and temperature variations.

This section analyses the impact of temperature, voltage and technology node for 45, 32, 22 and 16 PTM CMOS technologies and 18 and 13 nm from WP1 on a 32KB L1 cache complete memory...
system, in parallel to what was done in Section 4.1 with the 6T cell. The cache structure is the one presented in 4.1 where all the signals controlling the basic cell are compatible, with the only difference that in read operation the sense amplification procedure is of single-ended type for the 3T1D, in comparison with the differential technique in the 6T. No process variations are considered in this section.

*Hspice* simulations at different temperatures (30, 50, 70, 90 and 110 °C), with nominal parameter MOS devices for 45, 32,22 and 16 (PTM, @0.9volt) and 18 and 13 (WP1, @0.7volts), have been performed for 3T1D 32KB L1 cache memory. Table 7 shows the numeric results for Read_Access, Write_Access and Retention time. Both access times increases with temperature with an average increase of a 60% from 30 to 100 °C, the slope of this increase augments with the technology shrinking and access times decreases with it.

Retention time decreases abruptly with temperature due to the increase of leakage currents, about 10X for the same temperature margin. The technology shrinking also affects severely the retention time, caused by the corresponding increase of leakage currents. For the 18 nm and 13 nm technologies, retention time decreases about 100 times respect to 16nm, the reason is that those two technologies have a higher sub threshold current level (the threshold voltage for PTM technologies is around 450 mV while in the case of 18 and 13 nm is close to 200 mV. Figure 40, Figure 41 and Figure 42 display these results.

<table>
<thead>
<tr>
<th>Tecno/Temperature</th>
<th>Read_Access_Time (ns)</th>
<th>Write_Access_Time (ns)</th>
<th>Retention time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30 50 70 90 110</td>
<td>30 50 70 90 110</td>
<td>30 50 70 90 110</td>
</tr>
<tr>
<td>45nm PTM</td>
<td>2.1 2.6 3.0 3.2 3.5</td>
<td>2.0 2.6 2.9 3.0 3.3</td>
<td>64 35 21 19 10</td>
</tr>
<tr>
<td>32nm PTM</td>
<td>1.8 2.3 2.7 3.1 3.5</td>
<td>1.7 2.2 2.4 3.0 3.3</td>
<td>28 20 14 10 6</td>
</tr>
<tr>
<td>22nm PTM</td>
<td>1.6 1.9 2.3 2.7 2.9</td>
<td>1.6 1.7 2.1 2.5 2.6</td>
<td>21 14 9 5.5 3</td>
</tr>
<tr>
<td>16nm PTM</td>
<td>1.4 1.7 2.0 2.3 2.6</td>
<td>1.4 1.5 1.9 2.1 2.2</td>
<td>11 8.4 5.1 2.3 1.9</td>
</tr>
<tr>
<td>18nm WP1</td>
<td>0.9 1.2 1.8 2.0 2.3</td>
<td>1.0 1.2 1.8 2.1 2.4</td>
<td>0.1 0.09 0.08 0.05 0.01</td>
</tr>
<tr>
<td>13nm WP1</td>
<td>0.8 1.1 1.6 1.9 2.1</td>
<td>0.9 1.1 1.4 2.0 2.1</td>
<td>0.09 0.07 0.06 0.03 0.007</td>
</tr>
</tbody>
</table>

Table 7. Results of 3T1D 32KB cache for different technology processes and temperatures. No process variation is considered. Observe that 18 and 13 nm (WP1) technologies are HP (high performance) devices while PTM are LP (low power).
Figure 40. Impact of temperature in 3T1D 32KB cache for PTM technologies on access and retention time

Figure 41. Impact of PTM technologies in Access and Retention time
Figure 42. Comparison of read access time and retention time for 16nm PTM and 18nm WP1 technologies

5.3.1. Conclusions

The read and write cycle times for the 3T1D is very close to the times for the 6T SRAM, with a 40% of silicon area reduction, and lower consumption. The impact of temperature on those times is also similar to the obtained for the 6T (around a 60% of increase from 15 to 110 °C). Charge retention time is a key factor in dynamic RAMs. For the small size of the 3T1D the retention time is much lower that the calculated in the 1T1C case (100 times lower), and as it happened in the case of the 1T1C the retention time is very sensitive to temperature with drastic reductions in the temperature margin considered.
5.4. Robustness analysis of the 3T1D cell

In this section we proceed with the analysis of the robustness of 3T1D following a method parallel to the one presented in section 4.2. The objective is to determine the Satisfiability Boundary for a variation map of threshold voltages of the four transistors that constitute the basic 3T1D cell (T1, D1, T2 and T3, see Figure 39). In this section only a 2D map has been considered, so for each of the three basic operations the two more significant transistors are considered. The rest of parameters are considered nominal.

Read operation

In this operation only transistor T2 and D1 has been considered with variable threshold voltage.

Write operation

Transistor T1 and D1 are considered in the write analysis.

Hold or Retention operation

In this operation only transistor T2 and D1 has been considered with variable threshold voltage.

The analysis has been done in each case with incremental values of the threshold voltage (0.05 volts per step) for both devices independently, following 4 lines (vertical and horizontal, only one voltage varies and the two 45° and 135° lines). Each point has been simulated determining if the point verifies or not the acceptable condition generating a first approach to the acceptance region. Table 8 shows the results obtained with the 3T1D.

Assuming Gaussian independent distribution for the two variables the two-dimensional independent distributions are integrated inside the acceptance region (Matlab).

<table>
<thead>
<tr>
<th>Model</th>
<th>Tech.</th>
<th>Vth Variability</th>
<th>Retention (P&lt;sub&gt;RE&lt;/sub&gt;) T&lt;sub&gt;nominal&lt;/sub&gt;</th>
<th>Write (P&lt;sub&gt;W&lt;/sub&gt;) T&lt;sub&gt;nominal&lt;/sub&gt;</th>
<th>Read (P&lt;sub&gt;R&lt;/sub&gt;) T&lt;sub&gt;nominal&lt;/sub&gt;</th>
<th>Cell (failure P&lt;sub&gt;CELL&lt;/sub&gt;)</th>
<th>Yield(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>P&lt;sub&gt;WP1&lt;/sub&gt;=0.2</td>
<td>P&lt;sub&gt;WP1&lt;/sub&gt;=2</td>
<td>P&lt;sub&gt;WP1&lt;/sub&gt;=2</td>
<td>P&lt;sub&gt;WP1&lt;/sub&gt;=2</td>
<td>32k</td>
</tr>
<tr>
<td>PTM</td>
<td>45nm</td>
<td>4%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>32nm</td>
<td>6%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>22nm</td>
<td>15%</td>
<td>8.0e-8</td>
<td>3.0e-8</td>
<td>4.0e-8</td>
<td>5.0e-8</td>
<td>99.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8%</td>
<td>4.0e-7</td>
<td>1.0e-8</td>
<td>8.0e-8</td>
<td>1.0e-8</td>
<td>99.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15%</td>
<td>1.0e-7</td>
<td>3.0e-7</td>
<td>1.0e-7</td>
<td>2.0e-7</td>
<td>99.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30%</td>
<td>1.25e-5</td>
<td>2.0e-6</td>
<td>3.2e-6</td>
<td>8.0e-6</td>
<td>77.41</td>
</tr>
<tr>
<td></td>
<td>16nm</td>
<td>10%</td>
<td>2.0e-7</td>
<td>6.0e-7</td>
<td>1.5e-7</td>
<td>2.21e-7</td>
<td>99.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20%</td>
<td>1.0e-6</td>
<td>2.0e-6</td>
<td>1.21e-7</td>
<td>9.0e-7</td>
<td>97.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40%</td>
<td>3.0e-4</td>
<td>2.9e-4</td>
<td>6.4e-5</td>
<td>1.0e-4</td>
<td>4.07</td>
</tr>
<tr>
<td>WP1</td>
<td>18 nm</td>
<td>58%</td>
<td>3.0e-4</td>
<td>1.0e-4</td>
<td>4.0e-5</td>
<td>1.42e-4</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>13 nm</td>
<td>58%</td>
<td>1.0e-3</td>
<td>1.0e-3</td>
<td>3.0e-4</td>
<td>9.0e-4</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 8. Failure probability for different technologies and variability scenario for the 3T1D cell. This analysis has been done for minimum size devices (W/L=1), so the results are an overestimation of sized cell.

Very high variability scenario represents yield near 0% for WP1 technologies, 4% for 16nm and 77.41% for 22nm. For moderate variability 3T1D exhibits an acceptable yield. In general the failure probability of 3T1D is lower than the one with 6T, and consequently presents a higher yield.
5.4.1. Conclusions
Following a parallelism with the SB technique explained in 4.2 for the 6T the robustness of the 3T1D has been analysed. Only two devices variation analysis has been performed for each of the three time cycles (write, read and retention times). The results are similar to the ones presented for the 6T. Although in the case of the 3T1D a higher robustness has been shown compared to 6T, showing low but acceptable yields for high variability scenario in 22 and 16 (where in the case of 6T we found negligible yield). The yield obtained for very high variability for 22, 16, 18 and 13 nm, is practically zero.

5.5. Reliability aspects for 3T1D DRAM cells

5.5.1. Introduction
System reliability is arising as a major concern in next-generation designs and implementations of memory systems. The situation is aggravated in the case of small transistors to minimize area, such as the ones used in 6T SRAM and 3T1D DRAM cells. Higher variation of characteristic parameters like threshold voltage leads to less predictable devices. Random behaviour requires larger guard bands to satisfy the requirements of reliability. This trend compromises the gain in performance expected for new technology generations.

Reviewing the literature we can find many papers on the reliability issues of 6T SRAM cells. In a work from S. Mukhopadhyay [39] the failure probabilities associated to access-time, read/write failure and hold failure in 6T SRAMs are analysed and modelled against process-parameter variations. In the work from M. Meterelliyouz [16] thermal studies of 6T and 8T SRAMs cache memories are performed and the stability, noise immunity and performance parameters are extracted and analysed. There are also studies on the effects of NBTI in 6T SRAMS like the paper from K. Kang [40]. However, despite the abundance of works for 6T, there are few articles on 3T1D. In this subsection we review an unexplored reliability problem of 3T1D cell. We analyse the effects of the aging of 3T1D cell due to Positive Bias Temperature Instability (PBTI). PBTI aging mechanism has not received much attention so far due to its relative small impact with respect to other ways of aging. However, since the introduction of high-k dielectrics, PBTI is growing in importance and it is expected to dominate NMOS aging in future technology nodes [41] [42] [43]. We end the subsection with a comparison between 6T and 3T1D cell.

Since the first proposal of 3T1D cell, from W. Luk et al. [37], this new memory cell is being carefully investigated in order to explore its capabilities and precisely compare them with 6T. In the paper from K. Lovin et al. [44] a performance model is developed to study 3T1D memories with accurate results and fast simulations. 3T1D cell does not have a symmetrical structure like 6T cell. For this reason, the problems related to instability and bit flipping, usual in 6T cells, do not affect 3T1D. All possible variations in device behaviour on a 3T1D cell can be lumped into variations of data retention times. In this sense, 3T1D is more robust against variability than 6T cell. In the literature we can find specific comparison between these two main candidates for data caches memories. In the paper from X. Liang et al. [45] different schemes for the L1 data cache are analysed under different levels of variation. They obtain that 6T memories suffer 40% reduction in frequency when variability is severe while 3T1D cell loses only 4% of performance with the same level of variability.
Another source of variability that gains relevance with the ultimate CMOS generations is PBTI aging. However, up to now there is still no work considering the effects of PBTI on 3T1D cell. For this reason, and because PBTI aging will be a serious limiting factor of lifespan for future CMOS technology, we analyse during the rest of this subsection this specific degradation mechanism.

PBTI manifests as an increase in the threshold voltage due to electron trapping in the high-k dielectric layer. There are many studies on NBTI and well-developed models because of its larger effect on current technology. In the case of PBTI, there are still some discrepancies in the literature. In this work, we consider a PBTI degradation model with similar trends shown for NBTI but with a lower impact on Vth. PBTI degradation under DC stress increases the threshold voltage of NMOS transistors asymptotically as $\Delta V_{th}(t) \propto t^{1/6}$ [43]. While Vth shift due to NBTI in 45nm PMOS transistors reaches 50mV after $10^8$ seconds (= three years), the shift obtained for NMOS devices due to PBTI for the same period of time is only 30mV. Figure 43 shows the relationship between Vth shift due to NBTI and PBTI and DC stress time.

![Figure 43. Effect of NBTI/PBTI on Vth against continuous stress time for 45nm high-k transistors](image)

Another important aspect of PBTI degradation is its dependence on the duty-factor (DF). It has been shown that the effective shift in threshold voltage due to BTI degradation in general is a strongly non-linear function of DF [46]. This is due to the existence of a component of degradation that quickly recovers during periods of no stress. This effect should be taken into account when analysing PBTI-tolerance of circuits as it greatly varies the results of lifespan.

Figure 44 shows a typical characteristic of the gate capacitance versus Vgs for D1 gated diode at different states of PBTI degradation ($\Delta V_{th} = 0$mV and 30mV ). The difference between charges transferred in reading operations ($Q_1 - Q_0$) diminishes as D1 degrades; this implies worse performance. In the case of Figure 44, the charge transferred when reading 1 ($Q_1$) diminishes more than 11.3%, and that of 0 ($Q_0$) has no significant change: less than 0.8% increase.
5.5.2. Objectives

Our goal is to analyse the main trends of 3T1D degradation due to PBTI and see the potential impact of this aging mechanism in the reliability of the memory cell. In order to perform this study, we first define a fault model based on critical parameters. Then, analysing the performance of the 3T1D cell by means of these critical parameters, we are able to evaluate the robustness of the 3T1D cell against PBTI degradation. Among all possible modes of operation of a memory cell we identify as the most critical both hold and read. Timing condition for write operation is not as critical as for read. Read operation requires charging or discharging an entire BLread line and it has an equivalent capacitor bigger than the equivalent parallel D1 and T1 capacitances of a single 3T1D cell. In hold mode, we have to consider the leakage that will discharge continuously the accumulated charge in the node S. Let us note here, before starting with the critical parameter definitions, that critical situations for both hold and read operations occur only when the data stored in the cell corresponds to high level (1). Otherwise, no leakage will discharge node S and no timing problem will affect our reading, as BLread does not discharge.

- Hold Operation:

When a bit is stored in a 3T1D cell, both T1 and T3 access transistors are switched off. In these conditions there is a subthreshold leakage flowing from node S through T1 and a smaller current of gate leakage flowing through T2 and D1. As a consequence, node S discharges over time as long as no other write or refresh operation is performed. After a certain time, when the voltage in node S decreases under a certain level V5min, data in node S is lost because it cannot be read within the slot time available according to the operating clock frequency.

A proper way to quantify how critical is the hold operation is to measure the time needed to discharge the node S from the nominal value Vdd−Vth, achieved after a writing or refresh operation, to V5min, the moment when data is lost.

- Read Operation:

When reading, T3 transistor is activated and BLread is connected through T2 to the ground. Diode D1 receives a pulse as WLread is activated and the boosting effect activates T2, which starts to
discharge BLread line. The time needed to get access to the data depends directly on how fast BLread discharges.

We define the read access time as the time needed to discharge 20% of BLread line given an initial nominal charge Vdd. We consider this level of voltage drop enough to activate the sense amplifier.

5.5.3. Retention and Read Access time Analysis
Here we present the analysis results of 3T1D-DRAM cell’s critical parameters VSmin, retention time and read access time against PBTI degradation effects. Results are derived from HSPICE electrical simulations using 45nm PTM device models (only 45nm has been included in this deliverable from brevity reason). We perform the measurements at three different temperatures (25 °C, 60 °C and 100 °C) and three different source voltages (the nominal value Vdd and the corners of ±10%) in order to get an idea of the effect of environmental variations. The modelling of PBTI in HSPICE, as a Vth shift, is performed by adding a DC voltage source (ΔVth) in series with the gate terminal of the transistor. Nominal values to model the cell at 45nm are extracted from the work by K. Lovin et al. [44].

5.5.3.1 VSmin Characterization
Before starting with the retention time analysis we need to characterize the VSmin value for different environmental conditions. VSmin highly depends on the working temperature as it affects directly the amount of leakage current. Figure 45 depicts the read access times of a 3T1D cell against different initial charges in node S. Read access times are computed according to the previous definition and for a cell working at different temperatures. The VSmin parameter can be found at each temperature as the voltage in node S with exactly the same read access time as the available slot time fixed by the clock frequency. In this paper we will consider an operating frequency of 3.5GHz, then, adding the corresponding guard bands, we have as a requirement the maximum admissible read access time t_access ≤ 250ps.

![Figure 45 Read access times against voltage in node S of 3T1D cell working at different temperatures (VSmin condition in slashed line)](image)
We observe in Figure 45, apart from the negative impact of reducing the voltage in node S, that variations in working temperature cause a small and quite stable effect on the read access time through all charges considered in node S. Actually, any variation of temperature within the range of 60 °C to 100 °C implies a variation in the read access time of about 0.4ps/°C, given a 3T1D cell with the nominal charge Vdd−Vth.

5.5.3.2 Retention time

Using the previous definition of retention time and the VSmin parameter characterization, we perform simulations of 3T1D cell at different temperatures and different states of PBTI degradation (∆Vth). We assume the same degradation for both D1 and T2 transistors because they share the same gate to source voltage Vgs with only small differences during short periods of time (reading operations). T1 degradation occurs only scarcely and during a short period of time (writing operations of low value (0)). Moreover, it has no effect on read access time and a positive effect in retention time as it reduces the leakage current. Figure 46 shows the resulting retention times when considering different states of PBTI degradation (∆Vth). It is clear that a temperature increase implies significant reduction in the retention time. For example: increasing temperature from 60 °C to 100 °C implies about halving the retention time in any case.

![Figure 46 Impact of PBTI on retention times at different working temperatures. Error bars correspond to ±σ voltage variations (3σ=10%Vdd)](image)

Another relationship detected in Figure 46 is that at temperatures above 60 °C the effect of PBTI degradation becomes irrelevant. Only for values of temperature under 60 °C PBTI has a significant positive effect on the retention time. For example: at 25°C, an increase in the threshold voltage ∆Vth of 30mV implies about 7.6% increase in the retention time. From these results we can conclude that refreshing techniques in 3T1D-DRAM memories do not need to take into account the effects of PBTI because they are negligible at high temperatures and positive at lower temperatures.

5.5.3.3 Read Access Time

The time required to access the data in a cell depends a lot on the amount of charge remaining in the node S (see Figure 47). Depending on the time between refreshing operations we will have different probabilistic distributions of voltage in node S. We represent in the following simulation,
in Figure 47, the read access times of a 3T1D cell at different temperatures against different states of PBTI degradation.

![Figure 47 Impact of PBTI on read access time at different working temperatures. Error bars correspond to ±σ voltage variations (3σ=10%Vdd)](image)

We can see in this case a significant negative impact of PBTI on 3T1D performance. Read access time significantly increases with small Vth shifts. From the above results we obtain that a Vth increase of 30mV implies a 26% increase in read access time of a 3T1D cell working at 60ºC. As a final observation in this subsection, we see that PBTI should be taken into account when studying the timing conditions of memory cells with small NMOS transistors, especially after the introduction of high-k dielectrics.

5.5.4. **Yield Consideration**

Here we present a study on the manufacturing yield of 2kB cache memory blocks compounded of 3T1D memory cells. We assume a reconfigurable structure with 32 cells per column, 512 columns and 24 redundant columns. Taking into account all possible sources of variability (process, voltage, temperature and aging) we perform 10,000 Monte Carlo simulations. Figure 48 shows the yield of this memory against the level of PBTI degradation at three different temperatures.

![Figure 48 Yield of 2kB 3T1D memory blocks](image)
The basis for regarding a cell as defective is to check if the read access time holds the requirement presented previously: $t_{\text{access}} \leq 250\text{ps}$. From the figure we see how important is the temperature combined with PBTI degradation. If we need a 90% yield in our manufacturing process, we cannot admit more than 22mV, 33mV and 43mV of Vth shift when working at 100 °C, 60 °C and 25 °C respectively.

Regarding the relationship between Vth shift and stress time presented in Figure 43, the above result allows us to make a rough prediction of the lifespan of our simulated 2kB memory blocks. When doing so, it has to be considered the particular application of our circuit because the relationship between Vth shift and stress time presented highly depends on the duty-factor. When there are periods with no stress, the effects of relaxation significantly vary the relation between Vth shift and time. In the particular case of 3T1D cells used in L1 cache memory, the storage of high (1) logic values, the ones that produce PBTI degradation in T2 and D1, statistically occur only half of the time.

5.5.5. Comparisons and partial conclusions

Contrasting the above results for yield of 2KB 3T1D memory blocks and a previous work on 6Ts from S. Drapatz et al. [47], we observe that the impact of PBTI in both memory structures is relevant and within the same order of magnitude.

![Figure 49 Yield including global and local variations after long hold of one data with NBTI and PBTI degradation for a single Bit cell (Extracted from [47]).](image)

Figure 49 depicts the yield for the hold operation of a 6T cell against different supply voltages and different levels of NBTI-PBTI degradation. The increase in the threshold voltage with BTI drives the 6T cell structure to a less stable point and it becomes less reliable. 6T SRAM yield decreases quickly when the effect of PBTI increases. In the referred study, the authors conclude that the combination of PBTI with NBTI is much worse for 6T SRAM reliability than considering only NBTI. During hold state the two effects are adding, so that the cell weakens drastically in terms of read stability.

We conclude from the previous yield considerations of both 3T1D and 6T cells that PBTI degradation mechanism should not be ignored in memory system designs. In order to be able to produce long-life memory structures it is very important to consider the effects of PBTI and NBTI. The combination of these aging mechanisms with other sources of variability gradually weakens the robustness and the reliability of memory systems at specific conditions of usage for both 6T SRAM and 3T1D DRAM memory cells.
6. Impact of process variation on memory systems, evaluation of cache memories with 6T and 3T1D cells.

In this section the impact of the process variations on the 32KB L1 cache system performances is analysed. The structure of the memory is the one described in Section 4, and used also in Section 5. The variability scenario is the model described in Table 1 (section 3) for technologies of 32, 22 and 16 nm, modelled with PTM, and the model in Table 2 for 18 and 13 nm which models are results of WP1.

6.1. The analysis

The analysis has been performed with Hspice, applying systematic geometric variations and threshold voltage systematic and random (as stated in the variability model). Analysis has been performed at three temperatures, 25, 60 and 100 °C. Power supply voltage is the nominal in all the simulations (0.9 volts for 32,22,16nm and 0.7 volts for 18,13nm).

<table>
<thead>
<tr>
<th>Technology</th>
<th>Scenario</th>
<th>25 C (%)</th>
<th>60 C (%)</th>
<th>100 C (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>45 nm</td>
<td>Standard</td>
<td>3T1D</td>
<td>6T</td>
<td>3T1D</td>
</tr>
<tr>
<td>32 nm</td>
<td>Moderate</td>
<td>3.5</td>
<td>3.4</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>10</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>22 nm</td>
<td>Moderate</td>
<td>4.5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>11</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Very high</td>
<td>18</td>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.5)</td>
<td>(0.0)</td>
<td>(1.0)</td>
<td>(2.0)</td>
</tr>
<tr>
<td>16 nm</td>
<td>Moderate</td>
<td>5.2</td>
<td>4.1</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>15</td>
<td>14</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.4)</td>
<td>(0.0)</td>
<td>(0.5)</td>
<td>(0.0)</td>
</tr>
<tr>
<td></td>
<td>Very high</td>
<td>21</td>
<td>20</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.0)</td>
<td>(0.4)</td>
<td>(2.0)</td>
<td>(0.5)</td>
</tr>
<tr>
<td>18 nm</td>
<td>WP1</td>
<td>25</td>
<td>X</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.0)</td>
<td></td>
<td>(2.5)</td>
<td></td>
</tr>
<tr>
<td>13 nm</td>
<td>WP1</td>
<td>25</td>
<td>X</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.5)</td>
<td></td>
<td>(2.8)</td>
<td></td>
</tr>
</tbody>
</table>

Table 9. Write-Read cycle variations for 32KB with 6T and 3T1D and different technologies and scenarios. Number into brackets means percentage of write-read failures.
Table 9 shows the percentage of the sigma write-read cycle time for 1000 Montecarlo samples (half of them writing a 1 and half a 0). For high and very high scenarios a bold number is indicated between parentheses, which indicates the percentage of the cases where the write-read failed (bit flip in the case of a 6T or a failure in 3T1D). For 6T the failures are approximately equi-probable while in the case of 3T1D only flips from 0 to 1 appear, due to the characteristic of the cell. Figure 50 and Figure 51 show graphically the write-read access time standard deviation for high and very high variability scenarios. Variability increases substantially with temperature and the variation is higher for 6T but with the same order.

Figure 50. Standard deviation for access time in cache memory for different technologies and high variability scenario.
Finally Figure 52 shows the percentage of failures observed in the experiment. All these failures correspond to an incorrect writing-read cycle (a 1 is written but a 0 is read and vice versa). For moderate variability there is no error, but there are several for high and very high variability levels. Observe that for the worst case (6T 13nm 100 °C) the percentage of failure cycles is 3.8%, which is an unacceptable value for reliable memories. 3T1D memories appear more tolerant to variations (in the case of failures): this is due to the characteristics of the cell. The 6T cell is more sensitive because of the balance requirement in the transistor of the cell; this fact does not apply for the dynamic cell. Additionally in the case of the 3T1D only single transition failures are observable due to the single end pre-charged bit read line.

The experiment shows that process variability not only affects the speed performances of memories, but also affects the reliability of the memory in a very substantial level.

6.2. Conclusions

At cache level we have analysed the variability on the write-read cycle for both 6T and 3T1D. As the robustness analysis at bit-cell level already shown the impact of variability on the performances of the cache circuit are drastic. The 3T1D shows a lower impact, in comparison with 6T, but for both of them we find also catastrophic faults in the write-read cycle. As we concluded in previous section, with this percentage of catastrophic faults the yield expected at cache level is practically zero for the very high variability scenario in 22 and 16 nm as well as in the case of 18 and 13nm technologies investigated in TRAMS WP1.
7. Robustness of memory cells in front of SEU: SRAM

Reducing leakage power and improving the reliability of data stored in the memory cells are both becoming challenging as technology scales down. While the smaller threshold voltages causes increased leakage, smaller supply voltages and node capacitances can be a problem for soft errors. In modern processors SRAMs occupy more than 50% of the chip area. Hence it is more likely that the single-event upset (SEU) occur in the SRAMs. It is crucial to guarantee the reliability of the SRAM cells as the technology scales down.

The objective of this work is to carry out a detailed analysis of the scaling trends of the soft error problem for the traditional 6T SRAM memory cell for different level of voltages and for different process nodes.

7.1. Introduction

Due to current technology scaling trends, digital designs are becoming more susceptible to radiation induced particle hits resulting from radioactive decay and cosmic rays, than all other factors [48]. A low energy particle that before had no effect on a circuit can now flip the output of a gate. Such a bit-flip of a gate is called a single-event transient (SET) or a glitch. A single-event upset (SEU) or a soft error occurs if the SET is propagated to an output and latched into a memory element. These failures are called soft errors because they do not permanently damage the circuit but do destroy data.

The rate at which soft errors occur is referred to as soft error rate (SER). The level of soft error vulnerability is important since it directly influences the error detection and the correction mechanisms employed in the circuitry.

Soft errors are caused by two types of radiation alpha particles emitted by radioactive impurities in IC and package materials and high-energy neutrons resulting from the interaction of cosmic radiation with the earth’s atmosphere [49]. Moreover, the soft error rate is inherently related to the device parameters (e.g., \( L \) and \( V_{th} \)).

The report is organized in the following sections. We first provide a brief introduction to the soft error problem and the study of the state of the art. It is followed by the discussion of the simulation platform and the particle strike modelling and the injection of the soft-errors. Finally the results and concluding remarks are given.
7.2. Soft error: Background and state of the art

As shown in Figure 53, the 6T SRAM memory cell has two sensitive nodes, the drain of the OFF-NMOS transistor and the drain of the OFF-PMOS transistor, \( Q \) and \( \bar{Q} \) respectively as shown in the figure. The drain and substrate of the OFF transistor create a reverse-biased junction. The reverse-biased junctions of the cell are most sensitive nodes to the particle strikes.

Let us assume that a particle hits our silicon. Immediately followed by the particle strike, charges generation and collection occur. The generated charges are collected at the opposite voltage terminals of the reverse-biased. The movements of charges cause a current pulse at the struck node. The memory cell flips when the collected charge, \( Q \), is more than the charge stored at the struck node.

A method for estimating SER in CMOS SRAM circuits was recently developed by [50]. This model estimates SER due to atmospheric neutrons (neutrons with energies >1MeV) for a range of submicron feature sizes. It is based on a verified empirical model for the 600nm technology, which is then scaled to other technology generations. The basic form of this model is:

\[
SER \propto N_{flux} \cdot A_{diff} \cdot \exp \left( \frac{Q_{crit}}{C_{coll}} \right) \quad \text{Equation 16}
\]

As it is evident from the equation, SER sensitivity of SRAMs depends on: (i) the critical charge \( Q_{crit} \), which is the minimum amount of charge that needs to be injected into the critical node in order to flip the state of the memory, (ii) the diffusion area \( A_{diff} \) of the sensitive drain, which implies that the probability of a soft-error is linearly related to the probability that an alpha-particle hits the sensitive drain area, (iii) the charge collection efficiency \( C_{coll} \), which is a measure for the amount of deposited charge that a circuit node collects after an alpha-particle impact and is a constant for the given technology and the device, and (iv) \( N_{flux} \), which is the neutron flux with the energy >1MeV.

From equation 16, it is obvious that changes in the value of \( Q_{crit} \) relative to \( C_{coll} \) will have a very large impact on the resulting SER. The SER also decreases proportional to the square of the device size. Hazucha & Svensson [50] used this model to evaluate the effect of device scaling on the SER of memory circuits. They concluded that SER-per-chip of SRAM circuits should increase at most linearly with decreasing feature sizes.

Because \( Q_{crit} \) is directly related to the soft-error rate, understanding the effects of the factors affecting \( Q_{crit} \) is important to estimate the SER of a circuit. The higher the \( Q_{crit} \) the more robust the circuit is.
Figure 54 Various contributions to the error rate for a small area SRAM cell, designed with 350nm CMOS cell vs. Critical Charge

Figure 54 shows how various contributions to the terrestrial error rate are affected by critical charge [51]. The work was done with SRAM cells fabricated with a 0.35 μm CMOS process. For critical charge < 35 fC it is possible to upset the cell with alpha particles. Another interpretation is that by increasing the \( Q_{\text{crit}} \), it is possible to make the memory cell robust to withstand the alpha-particle induced SEUs. However, cell still remains vulnerable to the neutron strikes.

A number of ways have been proposed to measure \( Q_{\text{crit}} \) accurately using 3D device simulator. \( Q_{\text{crit}} \) can be calculated efficiently using a circuit simulator. However, seeing the trade-offs between the computation time and the accuracy it was desirable to model the particle strike into an equivalent current pulse. \( Q_{\text{crit}} \) depends on both the collected charge and the shape of the current pulse. The current pulse is represented by an equivalent current source between the drain and the substrate of the transistor. A number of current models have been proposed in the literature over the years and they are used by the circuit community for the precise characterization of \( Q_{\text{crit}} \) by performing circuit level simulations.

A simplified model was proposed later [52] using 3D device simulations in 0.35μm technology.

\[
Q_{\text{crit}} = C_N \cdot VDD + I_{DP} \cdot T_F \quad \text{Equation 17}
\]

This shows that \( Q_{\text{crit}} \) depends on the nodal capacitance (\( C_N \)), supply voltage (VDD), the drain current (\( I_{DP} \)) and the flipping time (\( T_F \)) of the cell. This work mainly focuses on observing the variations in \( Q_{\text{crit}} \) by varying the supply voltage, changing the properties of the current pulse (e.g., pulse width) and by altering the nodal capacitance (i.e., by down-scaling the technology nodes).

The current model which is adapted to simulate the equivalent current source is as follow,

\[
I(t) = \frac{Q}{\tau_f - \tau_r} \left[ \exp\left(-\frac{t}{\tau_f}\right) - \exp\left(-\frac{t}{\tau_r}\right) \right] \quad \text{Equation 18}
\]

It is a double exponential in nature where: \( Q \) is the charge deposited as the result of a particle strike, \( \tau_f \) is collection time (fall time) constant of the junction, and \( \tau_r \) is the ion-track establishment time (rise time) constant. Both are constants and depend on several process-related factors. Their values affect the magnitude and the severity of the SEUs for a given charge. Such current model with very fast timing characteristics will provide a conservative estimation of the \( Q_{\text{crit}} \).
7.3. Simulation Platform

Hspice circuit level simulator is used to carry out all the circuit level experiments. The objective is to analyse the $Q_{\text{crit}}$ of a 6T SRAM cell by implementing a 16-way last-level cache (LLC) with the capacity of 4MB (similar to modern processors like Intel® core architecture®) including all the necessary peripherals, using 65nm, 45nm and 32nm predictive technology models [53] with the supply voltage varying from 1.1V to 0.6V.

![Figure 55 Architectural organization of the 4MB LLC cache including all the peripheral circuit.](image)

For the accurate and comparable simulations of the LLC with all the necessary peripherals, CACTI [54] was used to determine the minimum dimensions of the basic block to be implemented and simulated in HSPICE. As one can see in Figure 55, CACTI [54] divides the 16-way, 4MB LLC into the basic blocks called sub-arrays with the capacity of 32Kb each to optimize the energy-delay performance by dividing the bit-lines and word-lines into smaller segments. Such one sub-array with all the necessary peripherals is implemented and simulated to obtained necessary results.

We assume that the bit-line capacitance for 65nm technology node is 58fF [55] [56] including the sense amplifier. The bit-line capacitance scales down by 25% [55] [56] with each technology generation.

7.3.1. Particle strike modelling and soft-error injection

We use equation 3 as described in [52] for modelling the current source in HSPICE. It has been previously observed that the particle strike may result into current pulse of different widths. The $Q_{\text{crit}}$ can also be defined as the total charge deposited under the area of the curve $I(t)$ of equation 3. Analytical equation of $Q_{\text{crit}}$ for a given pulse can be given by,

$$Q_{\text{crit}} = I(t_0)(\tau_f - \tau_r)$$
Figure 56 (a) shows the upset mechanism of logic ‘1’ and (b) shows the upset mechanism of logic ‘0’ in 6T SRAM cell.

Figure 56 shows the mechanism implemented for inserting the particle strike in HSPICE. NL and NR are the noise sources representing the variation in the supply voltage. In Figure 56(a) we show the mechanism to upset logic ‘1’ that is the flipping of the logic ‘1’ state of the memory cell to logic ‘0’. In Figure 56(b) the mechanism to flip logic ‘0’ to logic ‘1’ is shown.

In the operation of a 6T SRAM cell, the stored information at the storage nodes Q and QB is read and written by activating the bit-lines (BL and BLB in Figure 56); while the state of operation of the cell is determined by the word-line (WL in Figure 56). When the word-line is active the cell is open for read/write operations. And when the word-line is in-active the cell is in “hold” mode. The memory cell is most vulnerable when it is in “hold” mode. In the experiments, all the particle strike events are carried out in the hold mode.

Figure 57 Transient simulation of upset logic ‘1’ and upset logic ‘0’ during the ‘hold’ mode of the SRAM cell.

Figure 57 shows two such particle strikes upsetting the logic ‘1’ and the logic ‘0’ respectively. First waveform shows the word-line and the second and the third waveform show the values stored at the node Q in the 6T SRAM cell. Sequence of the transient operation is write ‘1’, hold ‘1’ and then read ‘1’ followed by the write ‘0’, hold ‘0’ and then read ‘0’ and again write ‘1’, hold ‘1’ and then read ‘1’. It is important to observe that, for successful bit flips and the accurate estimation of the Q_{crit}, both the particle strike events occur during the hold mode and the sense amplifier will read the new, flipped values.
7.4. Results

Previous studies [52] [57] [58] [59] show that both the shape of the pulse (which is double exponential as shown in the equation 3) and the pulse width are very important for a precise estimation of the $Q_{\text{crit}}$.

Pulse width is defined as the time between the start of the pulse and the instant when it drops below 0.67% of the $I(t)$ value referring to equation 3. The approximated width of the resulting current pulse due to the particle strike can be given as $W_{\text{pulse}} = 0.7 \times \tau_f$. The width of the current pulse induced due to the particle strike ranges from few pico-seconds to hundreds of pico-seconds [52].

Current pulse rise and fall time and the full-width half-maximum (FWHM) significantly impact the estimation of $Q_{\text{crit}}$. It was observed that pulse characteristics from one technology model cannot be extrapolated to another due to the characteristic timing parameters of the resulting current pulse. Therefore, in this work we consider two different scaling options; we assume that the width of the current pulse resulting from a particle strike will scale down the fall time ($\tau_f$ in equation 3) of the current pulse by 10% and 25% for each technology generation, keeping the 65nm generation as the reference.

Table 10 and Table 11 show the obtained $Q_{\text{crit}}$ assuming that the pulse width scales down 10% and 25% for every technology node respectively. The values of $Q_{\text{crit}}$ are obtained for nominal supply voltage (1.1V) operations.

<table>
<thead>
<tr>
<th>Technology node (1.1V)</th>
<th>Timing parameters</th>
<th>$Q_{\text{crit}}$(fC) (upset ‘1’)</th>
<th>$Q_{\text{crit}}$(fC) (upset ‘0’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>65nm</td>
<td>$\tau_r = 0.05ps; \tau_f = 0.2ps$</td>
<td>69.5</td>
<td>297.6</td>
</tr>
<tr>
<td>45nm</td>
<td>$\tau_r = 0.05ps; \tau_f = 0.18ps$</td>
<td>30.9</td>
<td>287.5</td>
</tr>
<tr>
<td>32nm</td>
<td>$\tau_r = 0.05ps; \tau_f = 0.15ps$</td>
<td>21</td>
<td>180</td>
</tr>
</tbody>
</table>

Table 10 Obtained $Q_{\text{crit}}$ for upsetting logic ‘1’ and ‘0’ for various technology nodes by scaling down the fall time by 10% with each technology generation.

<table>
<thead>
<tr>
<th>Technology node (1.1V)</th>
<th>Timing parameters</th>
<th>$Q_{\text{crit}}$(fC) (upset ‘1’)</th>
<th>$Q_{\text{crit}}$(fC) (upset ‘0’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>65nm</td>
<td>$\tau_r = 0.05ps; \tau_f = 0.2ps$</td>
<td>69.5</td>
<td>297.6</td>
</tr>
<tr>
<td>45nm</td>
<td>$\tau_r = 0.05ps; \tau_f = 0.15ps$</td>
<td>23.5</td>
<td>222</td>
</tr>
<tr>
<td>32nm</td>
<td>$\tau_r = 0.05ps; \tau_f = 0.1125ps$</td>
<td>10.5</td>
<td>101</td>
</tr>
</tbody>
</table>

Table 11 Obtained $Q_{\text{crit}}$ for upsetting logic ‘1’ and ‘0’ for various technology nodes by scaling down the fall time by 25% with each technology generation.

We observe that the amount of charge required to flip the logic “1” is less than to flip the logic “0” in 6T SRAM cells for the given technology and the given current pulse characteristic. Such behavior
occurs because of the stronger influence of the N+ diffusion (stronger pull down transistors, Mn1 and Mn2) responsible for the discharging compared to P+ diffusion (pull up transistors MP1 and MP2). Hence, we conclude that a 1-0 flip is more likely than a 0-1 flip. Because of this reason it is also important to note that the sizing (W/L ratio) of the 6T SRAM cell hugely impacts the Qcrit.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Operation</th>
<th>Qcrit (fC) obtained in this work.</th>
<th>Critical charge in literature(fC) [58]</th>
</tr>
</thead>
<tbody>
<tr>
<td>65nm</td>
<td>Read ‘1’ upset</td>
<td>69.5</td>
<td>15-65</td>
</tr>
<tr>
<td></td>
<td>Read ‘0’ upset</td>
<td>297.6</td>
<td>100-350</td>
</tr>
</tbody>
</table>

Table 12 Obtained Qcrit for upsetting logic ‘1’ and ‘0’ for various technology nodes compared with the critical charge published earlier for 6T SRAM cell.

We made an effort to calibrate our results. Table 12 shows the summary of the obtained Qcrit for upsetting logic ‘1’ and logic ‘0’ compared to the data available in the literature (unfortunately, only 65nm results are available). The values obtained for 65nm nodes fit in the range, and as mentioned above the small discrepancies in the measured values of Qcrit is because of the differences in the models of the devices, current pulse, and also due to difference in the measurement techniques and the architecture (e.g., bitline capacitance, etc).

Next point in our study was to assess the impact of low voltages on Qcrit (and therefore, SER rate). To observe this effect the supply voltage is reduced ensuring the reliable operation of the entire cache module.

![Figure 58](image)

Figure 58 (a) Variation of Qcrit for upsetting logic ‘1’, with supply voltage scaling for different technology models. (b) Variation of Qcrit for upsetting logic ‘0’, with supply voltage scaling for different technology models. Assuming the pulse width scales down by 10% with each technology generation.

Figure 58 (a) and (b) show how the Qcrit of the given 6T SRAM changes with the supply voltage assuming a 10% scaling of the pulse-width. The results support the assumption that the Qcrit has a linear relation with the supply voltage.

For logic ‘1’ upset as it is clear from the result shown in Figure 58 (a) that for 65nm and 45nm technologies when the supply voltage drops from 1.1V to 0.6V the Qcrit to create upset ‘1’ drops by 63% and 72%. For 32nm the reduction in Qcrit at 0.7V is 80%. As shown in Figure 58 (b) the Qcrit for
creating upset ‘0’ drops by 53% and 62% at 0.6V for 65nm and 45nm respectively, and in 32nm the reduction in $Q_{\text{crit}}$ is 64%.

Figure 59 (a) Variation of $Q_{\text{crit}}$ for upsetting logic ‘1’, with supply voltage scaling for different technology models. (b) Variation of $Q_{\text{crit}}$ for upsetting logic ‘0’, with supply voltage scaling for different technology models. Assuming the pulse width scales down by 25% with each technology generation.

Figure 59(a) and (b) show the variation in $Q_{\text{crit}}$ for different supply voltages and different technology nodes. In these results the current pulse width is scaled down by 25% with each technology generation. For upsetting logic ‘1’ the $Q_{\text{crit}}$ drops by 63%, 73% and 65% when moving from 1.1V to 0.6V for 65nm, 45nm and 32nm respectively. As shown in Figure 59 (b) the $Q_{\text{crit}}$ for creating upset ‘0’ drops by 53%, 65% and 54%.

In both sets of results it is evident that reduced supply voltage to save the leakage power compromises the robustness of the cell.

### 7.5. Conclusions

This work presented a study of the circuit level estimation of $Q_{\text{crit}}$. From the results discussed above it is easy to infer that $Q_{\text{crit}}$ linearly depends on the supply voltage. Moreover, we conclude that it is easier to generate a logic ‘1’ upset due to lower $Q_{\text{crit}}$ than logic ‘0’ upset. We have also studied the impact of both the waveform and the pulse width induced from the particle strike; results show that they have a significant impact on the resulting value of the $Q_{\text{crit}}$. 
8. Carbon Nanotube technology: Modeling and analysis of basic SRAM cells with CNTFETs.

In silicon bulk CMOS technology the variability of the device parameters is a key drawback and it may be a limiting factor for further miniaturizing nodes. In this section, the variability in Carbon nanotube Field Effect Transistor (CNFET) is evaluated as well as its real capability to be a promising alternative to Si-CMOS technology.

In a first part the impact of carbon nanotube (CNT) diameter variations and the presence of metallic CNTs (main sources of variability) in the transistor are analysed (device level). Both, different diameters and metallic CNTs are due to the lack of chirality control during the CNT growth process.

In a second part, a comparison between Si-CMOS (32nm, 22nm and 16nm Predictive Technology Models, and 18nm and 13nm models developed by UoG in TRAMS WP1) and CNFET 6T SRAM cells is realized (circuit level). First, a performance analysis is presented. All 6T cells are compared in terms of write time, access time, area, power consumption and read stability. Then an analysis introducing random $V_{th}$ variations for different scenarios and temperatures is presented.

8.1. Introduction

Carbon Nanotubes (CNTs) have been attracting much attention in recent years due to its small dimension, unusual geometry and their extraordinary electronic properties [60].

A CNT is a graphene sheet rolled up to form a hollow cylinder (Figure 60). A CNT can be semiconducting or metallic depending on the angle of the atom arrangement along the tube. This is referred as the chirality vector and is represented by the integer pair $(n, m)$: $C_h = na_1 + ma_2 = (n, m)$, where $n$ and $m$ (wrapping indices) are integers and $a_1$ and $a_2$ are the unit vectors of the graphene lattice. A simple method to determine whether a CNT is metallic or semiconducting is based on considering the indexes $(n, m)$, i.e. the nanotube is metallic if $n=m$ or $n-m=3i$, where $i$ is an integer. The diameter of the CNT also depends on the chirality. It can be calculated as $d = \frac{\sqrt{3}}{\pi} a_{cc} \sqrt{n^2 + m^2 + nm}$, where $a_{cc}$ is the nearest-neighbour carbon atom distance of 0.142 nm.

Transistors made of CNTs or Carbon Nanotube Field Effect Transistors (CNFETs) are promising candidates to replace silicon CMOS due to its high performance; mainly a high current driving capability, tolerance to temperature drift, leakage currents avoidance and self assembly growth [61]. A CNFET is essentially a conventional MOSFET, except that its channel is made up of one or more CNTs. In a MOSFET-like CNFET the CNTs present two un-gated heavily doped segments between the gate and the source/drain and an intrinsic gated region as is shown in Figure 61(left).

Stanford University has developed a compact SPICE model for MOSFET-like CNFET simulation [5] [62]. This model is implemented in three levels (Figure 61,right). Level 1, CNFET-L1, models the intrinsic behaviour of CNFET. The second level, CNFET-L2, includes the device non-idealities. The first two levels deal with only one CN under the gate. The top level, CNFET-L3, models the interface between the CNFET device and CNFET circuits. This level deals with multiple CNTs per device and includes the parasitic gate capacitance and screening due to adjacent CNTs. We have used this model for our HSPICE simulations.
Exceptional I-V characteristics have been demonstrated in “ideal” CNFETs (meaning all CNTs are semiconducting, have the same diameter and are aligned and well-positioned) [63]. However, in reality, at least from the research results obtained up to now, there are some imperfections inherent to CNT synthesis and CNFET manufacturing process that may eclipse the expectations. So, as mentioned before, in this first section we analyse the main source of variability: the lack of chirality control during the CNT growth process. This limitation is responsible for the percentage of metallic CNTs and the random distribution of CNT diameters.

8.2. Sources of variation for CNFET

Although CNFETs posses extraordinary properties that make them suitable for digital circuit applications, there are several limitations in CNT production methods and CNTFET manufacturing process that affect its performance and must be taken into account.

8.2.1. CNTs growth process limitations

The most critical challenge in the CNTs growth process is to control the diameter and the proportion of metallic CNTs. In short, the control of chirality. With the main growth methods used
today, laser ablation, chemical vapour deposition (CVD), arc discharge and high-pressure CO decomposition (HPCO), a polydisperse mixture of CNTs with various chirality is obtained [64].

Metallic CNTs (m-CNTs) should not be used to make CNTFETs. They present a high conductivity that makes that its current cannot be controlled by the gate, causing source-drain shorts in the transistor. On the other hand, diameter variations, caused by the respective growth method affect the electrical properties of the transistors such as the drive current and the threshold voltage.

1. Percentage of m-CNTs: A typical synthesis process produces 1/3 of m-CNTs and 2/3 of semiconducting (s-CNTs). Although recent CNT growth techniques can achieve as high as 90% of s-CNTs [65] and even 96% [66] there is not technique available today to grow 100% s-CNTs. Therefore, different m-CNTs removal techniques after growth, such as electrical burning and selective chemical etching are necessary.

2. Diameter variations: As in the previous case, there is not still a process that synthesizes nanotubes of only a unique diameter. Depending on the production method of CNTs, different mean diameters and diameter distributions can be observed [64].

### 8.2.2. CNFETs fabrication process limitations

An "ideal" MOSFET-like CNFET (Doped-S/D CNFET) is composed by one or more perfectly aligned carbon nanotubes whose section under the gate is intrinsic and the source/drain extension regions are n/p doped. There are also some challenges in the CNFET manufacturing process.

1. Doping variations: Controlled doping is very difficult and doping S/D variations cause drive current variations in CNFETs [67] [68].

2. Misaligned CNTs: Mispositioned and misaligned CNTs may cause shorts in CNFETs. Impressive achievements have been obtained in this field through the use of quartz substrate. More than 99% of the CNTs lie along the direction of the quartz [69]. Even for CNTs on quartz substrates, a non-negligible fraction of CNTs are misaligned. It is really difficult to position and align all the CNTs for all CNFETs properly.

In the next section we will focus on the CNFET growth process limitations. We examine how the proportion of m-CNTs affects the CNFET performance (in terms of transistor parameter variability) for a given diameter distribution and a number (N) of nanotubes per transistor.

### 8.3. CNFET Model, CNT growth variations and simulation methodology

#### 8.3.1. CNFET model

The CNFET structure is shown in Figure 62. The main device parameters used in this work are summarized in Table 13, with the aim of defining a CNFET with a size between that of 16nm and 32nm silicon bulk transistors. As can be observed, we assume that each CNFET is composed by N CNTs (N-tubes CNFET). We consider a nominal value of N=8, following the selection of the number of CNTs exposed in [70] by reliability reasons. The inter-CNT spacing is 4 nm (parameter affecting the charge screening effect) and the length of the S/D extensions and the gate is 16 nm. The width of the metal gate is 36 nm for the nominal value of N.
8.3.2. CNT growth variations

The chirality is a key factor in a CNFET, because it is responsible of the CNT behaviour, diameter and the threshold voltage of the 1-tube transistor. It has to be noted that the threshold voltage of a CNFET can be calculated from the chirality as follows.

\[ V_{TH} = \frac{E_g}{2q} = \frac{aV_\pi}{eD_{CNT}} \]

where the parameter \( a \) \((\sim 2.49\text{Å})\) is the carbon to carbon atom distance, \( V_\pi \) \((\sim 3.033 \text{ eV})\) is the carbon \( \pi-\pi \) bond energy in the tight bonding model, \( e \) is electronic charge and \( D_{CNT} \) is the CNT diameter.

As mentioned before, the mean diameters and the diameter distribution depend on the CNT growth method [64]. For example, using laser ablation process (in general) a Gaussian diameter distribution is observed with a diameters mean of 1.4 nm and a standard deviation (STD) of 0.3 nm, whereas with methods such as CVD and HPCO the CNT diameters are found to be spread over a wide range with a minimum diameter of 0.7 nm, and maximum diameter of 6 nm and a decaying probability for higher diameters [64]. This occurs because the diameter distribution depends on the density and the diameter of the catalytic nanoparticles as well as other factors such as growth conditions and growth temperature. We assume an asymmetrical Chi distribution of diameters with a range between 1 nm and 6 nm (Table 13). The mean and the STD of this function are 2.9 nm and 0.7 nm respectively. This distribution covers 10 possible diameters of s-CNTs (see Figure 63). Then, we also consider a hypothetical manufacturing process with a narrower diameter distribution. In this case we use a Gaussian distribution of diameters with a range between 1 nm and 2 nm.

---

### Table 13. Device parameters and process assumptions.

<table>
<thead>
<tr>
<th>Fixed parameters</th>
<th>Variable parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power supply</td>
<td>Number of CNTs per device (N)</td>
</tr>
<tr>
<td>Oxide thickness</td>
<td>Nominal: 8</td>
</tr>
<tr>
<td>Gate/Source/Drain length (CNT)</td>
<td>Range: 4-12</td>
</tr>
<tr>
<td>Width of the metal gate</td>
<td>CNT diameter (D)</td>
</tr>
<tr>
<td>CNT pitch</td>
<td>1 nm-6 nm</td>
</tr>
<tr>
<td></td>
<td>Chi distribution</td>
</tr>
<tr>
<td></td>
<td>1 nm-2 nm</td>
</tr>
<tr>
<td></td>
<td>Gaussian distribution</td>
</tr>
<tr>
<td></td>
<td>Metallic CNT proportion (T_M)</td>
</tr>
<tr>
<td></td>
<td>0% - 33%</td>
</tr>
</tbody>
</table>

Figure 62. CNFET device model [5] [62]
We assume that all m-CNTs can be broken by a perfect removal process, not being m-CNTs actives in the N-tubes CNFET. In our analysis the probability of a CNT be metallic is between 0% (all CNTs are semiconducting) and 33% (typical synthesis process).

Due to both limitations, proportion of m-CNTs and diameter variations, it has to be noted that it is also important to determine the number of CNTs per transistor (N) in order to guarantee a good CNFET performance. In [70], a number of tubes N=8 is proposed. In our work, we consider this number but we also analyze a range of N between 4 and 12.

### 8.3.3. The simulation methodology

Figure 63 shows the simulating procedure used to evaluate the distributions of the threshold voltage ($V_{th}$) and the equivalent transconductance factor ($K$) in a scenario of growing variability as the discussed above. The factors in the analysis are:

- The number of CNTs used in the CNFET implementation, $N$.
- The proportion of m-CNTs in the growing process, $T_M$.
- The diameter distribution of semiconductor CNTs, $D$, a random variable.

Given a set of $N$, $T_M$ and $D$ distribution, the analysis process is displayed in a simplified way in Figure 63. A sample of CNFET is obtained getting N nanotubes in an ordered extraction. Some of them are m-CNTs (M), randomly selected following the proportion $T_M$, and the rest are s-CNTs (S). M-CNTs are assumed to be perfectly broken by a burn removal technique, whereas a diameter is assigned to each s- CNTs extracted from the respective diameter distribution. So, the result is a sample of N-tubes CNFET, with M no active m-CNTs and S active s-CNTs.

For each CNFET sample the $I-V$ characteristic is obtained. This is done through the summation of the $N$ $I_{DS}$ current components (each CNT that forms the CNFET). The m-CNTs do not contribute ($I_{DS} = 0$). S-CNTs are evaluated using the HSPICE CNFET model developed by Stanford University [5] [62] taking into account its diameter, the charge screening effects and the position of the CNT in the transistor (edge or middle).

Finally, we perform a 1000 samples Monte Carlo analysis obtaining an $I_{DS} - V_{DS}$ characteristic distribution (see Figure 64). We also obtain the equivalent $V_{TH}$ and $K$ from each sample and calculate the mean and the standard deviation, STD. It has to be noted that the $V_{TH}$ of each CNFET sample is the minimum $V_{TH}$ of the single-tube components of the transistor. The $K$ is evaluated from the $V_{DD}$ saturated current level using the expression of the Sah [71] model for a equivalent Si-MOS transistor neglecting channel modulation and carriers saturation effects (as correspond to a CNFET).
Figure 63. Monte Carlo experiment and simulation methodology used.

Figure 64. Example of $I_{DS}$-$V_{DS}$ distribution for 50 CNFET samples.
8.4. Results of CNFET variation analysis

The following figures show the mean and standard deviation for the $V_{TH}$ and the equivalent transconductance factor $K$ obtained using the procedure previously presented. Both parameters characterize the device variability in CMOS technology; the first as threshold voltage and the second as a factor that includes the impact of the geometric parameters of the layout (transistor channel width and length).

Figure 65 and Figure 66 show the mean of $V_{TH}$ and $K$ respectively. Mean of $V_{TH}$ of the CNFET increases with the probability of metallic tubes in the manufacturing process and decreases as the number of tubes per transistor grows. Figure 66 shows the effect on the mean of $K$, increasing with the number of tubes, as expected, but decreasing with the metallic probability. All these considerations are useful in the design phase but now we concentrate our attention on the parameters variability that is a key drawback for modern CMOS technologies.

Figure 65. Simulation results. Mean of $V_{TH}$.

Figure 66. Simulation results. Mean of $K$. 
Figure 67 and Figure 68 show the impact of \( N \) and \( T_M \) on the STD of \( V_{TH} \) and equivalent \( K \). In Table 14 the percentages of variation (for 3 sigma) of \( V_{TH} \) and \( K \) for \( N \) and \( T_M \) corners are shown, including the results for the nominal value of \( N=8 \). For the threshold voltage, the maximum and the minimum percentage of variation are 67 % and 35 % that correspond to the corners \((N = 4, T_M = 33\%)\) and \((N = 12, T_M = 0\%)\) respectively. These values imply that even for current CNFET manufacturing capabilities, the variation in the threshold voltage is much lower than the expected for 32 and 16 Si-CMOS technologies. Because the ITRS [23] predicts that they may reach 70% and 100% of percentage variation. In the case of \( K \), observe that the higher is the number of tubes \( (N) \) the lower the variability for both \( T_M \) probabilities. Consequently there is a trade-off between device size and variability (caused by \( N \)) that designers should take into account.

![Figure 67. STD of \( V_{TH} \).](image1)

![Figure 68. STD of \( K \).](image2)
For an ideal case of no m-CNTs in the manufacturing process ($T_M = 0\%$), we obtain a percentage of variation for $V_{TH}$ and $K$ that goes from 35\% to 48\% and from 35\% to 69\% respectively, that is much lower than the expected for 32nm and 16nm conventional technologies. So, the reduction of the metallic proportion is a key factor for the development of CNFET circuits.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$V_{TH}$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of tubes</td>
<td>$T_M=0%$</td>
<td>$T_M=33%$</td>
</tr>
<tr>
<td>$N=4$</td>
<td>48%</td>
<td>67%</td>
</tr>
<tr>
<td>$N=8$</td>
<td>39%</td>
<td>46%</td>
</tr>
<tr>
<td>$N=12$</td>
<td>35%</td>
<td>39%</td>
</tr>
</tbody>
</table>

Table 14. $V_{TH}$ and $K$ variation for 1000 samples and a Chi distribution of diameters.

For this last case ($T_M = 0\%$) the only reason of variability is the diameter distribution of the CNTs. If we consider a hypothetical manufacturing process with a narrower distribution (Gaussian diameter distribution with a mean and STD of 1.5 nm and 0.16 nm respectively) the percentage of variation would be between 16\% (N=12) and 23\% (N=4) for $V_{TH}$ and between 17\% and 25\% for $K$ (see Table 15). This would allow a design scenario where variability would not be a critical factor as it is nowadays in conventional CMOS.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$V_{TH}$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of tubes</td>
<td>$T_M=0%$</td>
<td>$T_M=33%$</td>
</tr>
<tr>
<td>$N=4$</td>
<td>23%</td>
<td>42%</td>
</tr>
<tr>
<td>$N=8$</td>
<td>21%</td>
<td>24%</td>
</tr>
<tr>
<td>$N=12$</td>
<td>16%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Table 15. $V_{TH}$ and $K$ variation for 1000 samples and a Gaussian distribution of diameters.

So, considering a range of metallic tubes from 33\% (current growth methods) to 0\% (perfection) and a realistic distribution of diameters, it has been shown that the variability of both $K$ factor and $V_{TH}$ is lower than CMOS for transistors with just 8 nanotubes, and much better for 12 tubes. Even for the case of only 4 tubes per transistor the observed variability is similar to the expected for CMOS. In a future scenario with a narrower distribution of diameters, variation for both parameters could reach levels from 16\% to 25\%, fact that would allow a design procedure without the stress caused by variability in current conventional technology.
8.5. Variability analysis and performance in CMOS and CNFET SRAM 6T cells

In order to evaluate the potential of CNFETs as possible alternative to Si-MOSFETs technology at circuit level, a comparison between Si-CMOS and CNFET 6T SRAM cells has been realized. First, we compare them in terms of Write time, Access time, area, power consumption and read stability (nominal comparison). Then, we investigate the effects of $V_{TH}$ and temperature on the characteristics of the different SRAM cells (variability comparison).

8.5.1. Model and parameters definition for 6T cell

A typical 6T SRAM cell is shown in Figure 69. This cell has been implemented using different technologies: 16nm, 22nm, 32nm, and 18nm, 13nm CMOS technologies and CNFET technology. The width ($W$, in nm) and length ($L$, in nm) ratio of each transistor for CMOS technologies are also shown in Figure 69 (in red), where $T$ is the Technology parameter and depends on the technology size. It takes the values shown in Table 16. For CNFET technology, all transistors have the same size, $W=32$nm ($8$ tubes) and $L=16$nm. HSPICE simulations are performed using Berkeley high-performance Predictive Technology Models (PTM) [4], 13nm and 18nm CMOS models developed by the UoG (TRAMS WP1), and carbon nanotube technology model [5] [62]. These models and the supply voltage for each one are shown in Table 16.

![Figure 69. Schematic of a typical 6T SRAM cell.](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>PTM</th>
<th>TRAMS WP1</th>
<th>CNFET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology size</td>
<td>32nm</td>
<td>22nm</td>
<td>16nm</td>
</tr>
<tr>
<td>$T$</td>
<td>16nm</td>
<td>11nm</td>
<td>8nm</td>
</tr>
<tr>
<td>$V_{DD}$</td>
<td>1V</td>
<td>1V</td>
<td>1V</td>
</tr>
</tbody>
</table>

Table 16.Models, technologies and supply voltage.
The value of bitline capacitances depends on the technology. The bitline capacitances ($C_{BL}$, $C_{BLB}$) have been calculated as the sum of the junction capacitance of the access transistors ($C_j$) of each cell plus the wire capacitance ($C_{wire}$). We consider 256 cells.

$$C_{BL} = C_j \cdot cells + C_{wire} = C_j + C_{wire}$$

For CMOS technology we assume that the wire capacitance is approximately 5 times $C_j$. So, the final expression for the $C_{BL}$ ($C_{BLB}$) is

$$C_{BL} = C_j + C_{wire} = C_j + 5C_j = 6C_j$$

For CNFETs the junction capacitance is negligible so we just take into account the wire capacitance that is similar to 16nm.

The value of the bitline capacitances for each technology is shown in Table 17.

<table>
<thead>
<tr>
<th>Capacitance</th>
<th>PTM (W=64nm)</th>
<th>TRAMS WP1 (W=32nm)</th>
<th>CNFET (W=36nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_j$ (aF)</td>
<td>38.48</td>
<td>17.35</td>
<td>negligible</td>
</tr>
<tr>
<td>$C_{wc}$ (fF)</td>
<td>9.85</td>
<td>4.44</td>
<td>2.64</td>
</tr>
<tr>
<td>$C_{wire}$ (fF)</td>
<td>49.26</td>
<td>22.21</td>
<td>22.21 (similar to 16nm)</td>
</tr>
<tr>
<td>$C_{BL}$ (fF)</td>
<td>59.11</td>
<td>26.65</td>
<td>15.85</td>
</tr>
</tbody>
</table>

Table 17. Bitline capacitance values for each technology.

### 8.5.2. CNFET SRAM cell versus Si-MOSFET SRAM cell (nominal comparison)

#### 8.5.2.1 Write time and Access time

Write time and Access time are two important metrics for assessing the performance of a SRAM cell. In this document, the Write time is defined as the time required for changing the cell contents (store node reaches 0.6$V_{DD}$ for write ‘1’ or 0.4 $V_{DD}$ for write ‘0’) after the wordline is turned on (Voltage(WL)=0.5$V_{DD}$). The Access time is defined as the time required for producing a prespecified voltage difference (0.1V) between BL and BLB after the wordline is turned on. This voltage difference is sensed by sense amplifier.

The results for both times are shown in Figure 70.

As can be observed both times decrease as the technology size decreases for PTM and TRAMS WP1 CMOS technologies. CNFET technology shows the least Write time (7.77ps) and Access time (57.42ps).
8.5.2.2 Area, Power Consumption and Read Stability

The area of the 6T cell for each node technology has been calculated basing on [16], obtaining the following size (Table 18). Obviously the area decreases as the technology size decreases. The area of CNFET is similar to 16nm area.

Static power consumption is very low for CNFET. This is because the CNFET has a significantly higher ON-OFF current ratio compared to the MOSFET in deep-submicron range. Dynamic power consumption is also less in CNFET than in PTM Si-MOSFET models but it is higher than in TRAMS WP1 models.

Read stability is measured by the Static Noise Margin (SNM), that is defined as the maximum value of DC noise voltage that can be tolerated by the SRAM cell without changing the stored bit. So, as higher is the SNM more robust is the cell. As is shown in Table 18 the most stable cell is the CNFET.
8.5.3. CNFET SRAM cell versus Si-MOSFET SRAM cell (variability comparison)

In this section we consider random VTH variations in 6T SRAM cells. For CMOS technologies we take into account the scenarios shown in Table 19 (moderated, high and very high). In column 3 the percentage of VTH variation (100 x σ/μ) for minimum transistor size is shown; in order to adapt it to our transistor size, we have corrected these values with percentage/√WL. Final values of VTH variability used in this document are shown in column 4. For CNFET we consider 15.58% of VTH variation (Table 20). This value has been obtained from the mean (0.12V) and the STD (18.7mV) of VTH extracted following the method presented in section 6.3.3 for 8 tubes and 33% of m-CNTs probability. We have also analysed the impact of these variations for three different temperatures: 25 °C, 60 °C and 100 °C.

<table>
<thead>
<tr>
<th>Si-Bulk Technology</th>
<th>Scenario</th>
<th>Random(min.size)</th>
<th>Random (corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100 x 1 σ/average</td>
<td>percentage / √WL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VTH</td>
<td>VTH</td>
</tr>
<tr>
<td>32nm</td>
<td>moderated</td>
<td>6%</td>
<td>4.24%</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>15%</td>
<td>10.61%</td>
</tr>
<tr>
<td></td>
<td>moderate</td>
<td>15%</td>
<td>10.61%</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>30%</td>
<td>21.21%</td>
</tr>
<tr>
<td>22nm</td>
<td>moderate</td>
<td>10%</td>
<td>7.07%</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>20%</td>
<td>14.14%</td>
</tr>
<tr>
<td></td>
<td>very high</td>
<td>40%</td>
<td>28.28%</td>
</tr>
<tr>
<td>16nm</td>
<td>moderate</td>
<td>10%</td>
<td>7.07%</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>20%</td>
<td>14.14%</td>
</tr>
<tr>
<td></td>
<td>very high</td>
<td>58%</td>
<td>41.01%</td>
</tr>
<tr>
<td>18nm</td>
<td>very high</td>
<td>58%</td>
<td>41.01%</td>
</tr>
<tr>
<td>13nm</td>
<td>very high</td>
<td>58%</td>
<td>41.01%</td>
</tr>
</tbody>
</table>

Table 18. 6T cell performance for each technology.

Table 19. Scenarios considered for each technology.

<table>
<thead>
<tr>
<th>CNFET Technology</th>
<th>NCNFET VTH</th>
<th>PCNFET VTH</th>
<th>Random 6T cell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>μ (V)</td>
<td>σ(V)</td>
<td>μ(V)</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.0187</td>
<td>-0.12</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.0187</td>
<td>15.58%</td>
</tr>
</tbody>
</table>

Table 20. VTH variability for CNFET technology.

The mean (μ) and the STD (σ) of the Write time and Access time have been measured for each scenario and each temperature. These values have been obtained for 500 samples. Also, the variability for 3σ and the percentage of errors are shown in the next tables. Noted that the percentage of errors include flipping errors and errors due to an excessive time. The percentage of flipping errors is shown in brackets.
### 8.5.3.1 Write time

#### 32nm PTM

<table>
<thead>
<tr>
<th>Random 6T cell 100 x 1σ/average</th>
<th>25°C</th>
<th>60°C</th>
<th>100°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_{TH} ) variation</td>
<td>( \mu ) (ps)</td>
<td>STD (ps)</td>
<td>( \mu ) (ps)</td>
</tr>
<tr>
<td>4.24% (6%)</td>
<td>22.36</td>
<td>0.82</td>
<td>26.66</td>
</tr>
<tr>
<td>10.61% (15%)</td>
<td>22.77</td>
<td>2.33</td>
<td>27.69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>25°C errors (%)</th>
<th>60°C errors (%)</th>
<th>100°C errors (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.24% (6%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10.61% (15%)</td>
<td>0</td>
<td>0.20 (0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( V_{TH} ) variation</th>
<th>25°C 100x3σ/( \mu ) (%)</th>
<th>60°C 100x3σ/( \mu ) (%)</th>
<th>100°C 100x3σ/( \mu ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.24% (6%)</td>
<td>11.03</td>
<td>13.96</td>
<td>15.53</td>
</tr>
<tr>
<td>10.61% (15%)</td>
<td>30.72</td>
<td>47.63</td>
<td>55.27</td>
</tr>
</tbody>
</table>

Table 21. Write time variation for 32nm PTM.

#### 22nm PTM

<table>
<thead>
<tr>
<th>Random 6T cell 100 x 1σ/average</th>
<th>25°C</th>
<th>60°C</th>
<th>100°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_{TH} ) variation</td>
<td>( \mu ) (ps)</td>
<td>STD (ps)</td>
<td>( \mu ) (ps)</td>
</tr>
<tr>
<td>5.66% (8%)</td>
<td>19.37</td>
<td>1.03</td>
<td>23.58</td>
</tr>
<tr>
<td>10.61% (15%)</td>
<td>19.66</td>
<td>1.99</td>
<td>24.23</td>
</tr>
<tr>
<td>21.21% (30%)</td>
<td>21.82</td>
<td>6.82</td>
<td>28.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>25°C errors (%)</th>
<th>60°C errors (%)</th>
<th>100°C errors (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.66% (8%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10.61% (15%)</td>
<td>0.4 (0)</td>
<td>1.2 (0)</td>
</tr>
<tr>
<td>21.21% (30%)</td>
<td>11 (0.2)</td>
<td>17 (0.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23.4 (0.4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( V_{TH} ) variation</th>
<th>25°C 100x3σ/( \mu ) (%)</th>
<th>60°C 100x3σ/( \mu ) (%)</th>
<th>100°C 100x3σ/( \mu ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.66% (8%)</td>
<td>15.89</td>
<td>18.81</td>
<td>21.57</td>
</tr>
<tr>
<td>10.61% (15%)</td>
<td>30.50</td>
<td>37.07</td>
<td>52.06</td>
</tr>
<tr>
<td>21.21% (30%)</td>
<td>93.72</td>
<td>123.87</td>
<td>107.38</td>
</tr>
</tbody>
</table>

Table 22. Write time variation for 22nm PTM.

#### 16nm PTM

<table>
<thead>
<tr>
<th>Random 6T cell 100 x 1σ/average</th>
<th>25°C</th>
<th>60°C</th>
<th>100°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_{TH} ) variation</td>
<td>( \mu ) (ps)</td>
<td>STD (ps)</td>
<td>( \mu ) (ps)</td>
</tr>
<tr>
<td>7.07% (10%)</td>
<td>16.99</td>
<td>1.11</td>
<td>20.94</td>
</tr>
<tr>
<td>14.14% (20%)</td>
<td>17.57</td>
<td>2.38</td>
<td>22.13</td>
</tr>
</tbody>
</table>
Table 23. Write time variation for 16nm PTM.

<table>
<thead>
<tr>
<th>V_{TH} variation</th>
<th>25ºC errors (%)</th>
<th>60ºC errors (%)</th>
<th>100ºC errors (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.07% (10%)</td>
<td>1.2 (0)</td>
<td>5.6 (0)</td>
<td>12.2 (0)</td>
</tr>
<tr>
<td>14.14% (20%)</td>
<td>11.6 (0)</td>
<td>20.2 (0)</td>
<td>26.6 (0.2)</td>
</tr>
<tr>
<td>28.28% (40%)</td>
<td>30.4 (2)</td>
<td>40 (3.4)</td>
<td>47 (4.6)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V_{TH} variation</th>
<th>25ºC 100x3σ/µ (%)</th>
<th>60ºC 100x3σ/µ (%)</th>
<th>100ºC 100x3σ/µ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.07% (10%)</td>
<td>19.65</td>
<td>22.72</td>
<td>26.69</td>
</tr>
<tr>
<td>14.14% (20%)</td>
<td>40.63</td>
<td>58.65</td>
<td>70.68</td>
</tr>
<tr>
<td>28.28% (40%)</td>
<td>165.39</td>
<td>132.69</td>
<td>137.68</td>
</tr>
</tbody>
</table>

Table 24. Write time variation for 18nm WP1 TRAMS.

<table>
<thead>
<tr>
<th>V_{TH} variation</th>
<th>25ºC errors (%)</th>
<th>60ºC errors (%)</th>
<th>100ºC errors (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.01% (58%)</td>
<td>10.8 (1.6)</td>
<td>10.2 (1.4)</td>
<td>11.2 (1.8)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V_{TH} variation</th>
<th>25ºC 100x3σ/µ (%)</th>
<th>60ºC 100x3σ/µ (%)</th>
<th>100ºC 100x3σ/µ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.01% (58%)</td>
<td>149.59</td>
<td>139.57</td>
<td>129.29</td>
</tr>
</tbody>
</table>

Table 25. Write time variation for 13nm WP1 TRAMS.
Table 26. Write time variation for CNFET.

### 8.5.3.2 Access time

#### 32nm PTM

<table>
<thead>
<tr>
<th>Random 6T cell 100 x 1σ/average</th>
<th>25°C</th>
<th>60°C</th>
<th>100°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{TH}$ variation</td>
<td>$\mu$ (ps)</td>
<td>STD (ps)</td>
<td>$\mu$ (ps)</td>
</tr>
<tr>
<td>4.24% (6%)</td>
<td>105.25</td>
<td>3.81</td>
<td>123.82</td>
</tr>
<tr>
<td>10.61% (15%)</td>
<td>100.75</td>
<td>21.72</td>
<td>119.28</td>
</tr>
</tbody>
</table>

| 25°C errors (%)                  | 0     | 0     | 0     |
| 60°C errors (%)                  | 0.8 (0.6) | 1.8 (1.6) | 3.4 (3.2) |

| 25°C 100x3σ/$\mu$ (%)            | 10.87 | 11.65 | 12.38 |
| 60°C 100x3σ/$\mu$ (%)            | 64.68 | 69.61 | 82.45 |

Table 27. Access time variation for 32nm PTM.

#### 22nm PTM

<table>
<thead>
<tr>
<th>Random 6T cell 100 x 1σ/average</th>
<th>25°C</th>
<th>60°C</th>
<th>100°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{TH}$ variation</td>
<td>$\mu$ (ps)</td>
<td>STD (ps)</td>
<td>$\mu$ (ps)</td>
</tr>
</tbody>
</table>

| 4.24% (6%)                       | 10.87 | 11.65 | 12.38 |
| 10.61% (15%)                     | 64.68 | 69.61 | 82.45 |
16nm PTM

<table>
<thead>
<tr>
<th>Random 6T cell</th>
<th>25°C</th>
<th>60°C</th>
<th>100°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 x 1σ average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V&lt;sub&gt;TH&lt;/sub&gt; variation</td>
<td>μ (ps)</td>
<td>STD (ps)</td>
<td>μ (ps)</td>
</tr>
<tr>
<td>7.07% (10%)</td>
<td>73.25</td>
<td>6.089</td>
<td>92.29</td>
</tr>
<tr>
<td>14.14% (20%)</td>
<td>72.28</td>
<td>26.17</td>
<td>87.95</td>
</tr>
<tr>
<td>28.28% (40%)</td>
<td>71.33</td>
<td>38.89</td>
<td>89.48</td>
</tr>
<tr>
<td>25°C errors (%)</td>
<td>1.8 (1.8)</td>
<td>5.6 (5.6)</td>
<td>14.2 (13.6)</td>
</tr>
<tr>
<td>60°C errors (%)</td>
<td>14.2 (13)</td>
<td>22.4 (20.6)</td>
<td>29 (25.6)</td>
</tr>
<tr>
<td>100°C errors (%)</td>
<td>34.6 (25.8)</td>
<td>42.4 (30.8)</td>
<td>46.2 (33)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>25°C 100x3σ/μ (%)</th>
<th>60°C 100x3σ/μ (%)</th>
<th>100°C 100x3σ/μ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.07% (10%)</td>
<td>24.94</td>
<td>67.32</td>
</tr>
<tr>
<td>14.14% (20%)</td>
<td>108.62</td>
<td>101.06</td>
</tr>
<tr>
<td>28.28% (40%)</td>
<td>163.56</td>
<td>202.64</td>
</tr>
</tbody>
</table>

Table 29. Access time variation for 16nm PTM.

18nm WP1 TRAMS

<table>
<thead>
<tr>
<th>Random 6T cell</th>
<th>25°C</th>
<th>60°C</th>
<th>100°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 x 1σ/average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V&lt;sub&gt;TH&lt;/sub&gt; variation</td>
<td>μ (ps)</td>
<td>STD (ps)</td>
<td>μ (ps)</td>
</tr>
<tr>
<td>41.01% (58%)</td>
<td>153.23</td>
<td>55.37</td>
<td>166.96</td>
</tr>
<tr>
<td>25°C errors (%)</td>
<td>14 (11.4)</td>
<td>18.4 (16.6)</td>
<td>23.8 (20.8)</td>
</tr>
</tbody>
</table>
8.5.4. Results and discussion

In general in Si-CMOS technologies, for both read and write times the variability (100x3σ/μ) increases as the scenario gets worse variability and as the temperature increases. In CNFETs they practically remain constant with the temperature because temperature has a low effect in CNFET technology as can be observed in Figure 71. The oscillation frequency (for a 11 stages oscillator) decreases as temperature increases for a 16nm PTM technology whereas it remains constant for CNFET technology.
Regarding the percentage of errors, it also increases with worse scenarios and higher temperatures for all technologies. In order to analyse in detail the values shown in the tables of this section the following graphs have been realized. The six technologies considered are represented in the X-axis. Observe that the points obtained for different models (PTM and TRAMS WP1 models) are linked by a line.

In Figure 72 the nominal values of the Write time (ps) an Access time (ps) for each technology at 25ºC are shown. Both times decreases as the technology size decreases for Predictive Technology Models and for WP1 TRAMS models. CNFET technology presents the best Write time and Access time.

In Figure 73 and Figure 74 percentage of variability at 3σ for Write time and Access time at 100ºC (worst temperature case) and for different scenarios is shown. In general, the variability in both times increases as the technology size decreases (PTM and WP1 TRAMS). In the case of CNFET technology, variability for Write time (28.49%) is similar to 16nm PTM in a moderated scenario, whereas variability in Access time is much lower in CNFET (10.33%) than in the rest of technologies.

In Figure 75 and Figure 76 the total percentage of errors for Write time and Access time at 100ºC and for different scenarios is shown. The percentage of flipping errors is shown in brackets. It can be observed that the percentage of errors in both times increases as the technology size decreases and as the scenario gets worse. The maximum percentage of total errors in Write time is for 16nm PTM (47%) whereas is 13nm WP1 TRAMS model that presents more flipping errors (7.2%). In Access time, the maximum percentage of total errors and flipping errors is also for 16nm PTM in a very high variability scenario. Finally it has to be noted that the percentage of errors for CNFET technology is 0 (no errors) for Write time and Access time.
Figure 72. Nominal Write time and Access time at 25°C for each technology.

Figure 73. Write time variability at 100°C for each technology.
Figure 74. Access time variability at 100°C for each technology.

Figure 75. Write time. Percentage of errors at 100°C. In brackets the percentage of flipping errors is shown.
8.6. Conclusions

Carbon nanotubes FET (CNFET) technology exhibits a set of attractive characteristics in order to be considered as a potential alternative to CMOS in memory systems. Its low leakage current reduces the static consumption and it would be a key factor in dynamic memories. It sensitivity to temperature is practically null, this is an interesting factor compared with the very high sensitivity bulk technology present. Additionally is faster, not substantially faster if we compare with the high performance 18 and 13 nm technologies. However the implementation maturity of such a technology sill requires several years of development. It as a technology compatible with bulk CMOS but it will introduce complementary and expensive new processing steps. The variability analysis shows, as a promising perspective, that even for todays performance of growing carbon nanotubes the variability is comparable with the scenario of bulk technology considered in this work as moderate. With the improvements on the chirality control the potential variability would reach levels of a very controlled technology with variability lower than moderate.
9. General conclusions

In this document the effects of environmental and process parameters fluctuations on the performance variability of memory cells and systems have been investigated. The work has concentrated mainly on bulk-CMOS technologies (key objective of the milestone MS2, M12) and, in order to compare them with a potential future alternative technology the work has been extended to CNFET technology.

The bulk-CMOS technologies can be divided into two groups, one is the set of open source technologies known as Predictive Technology Models (PTM), facilitated publically by the University of Arizona, covering 45, 32, 22 and 16 nm technology nodes. The second group of special interest are the bulk-CMOS 18 and 13 nm technologies developed by TRAMS in WP1 (Deliverable D1.1) on both variability and reliability aspects. For this last group we have considered in our analysis the nominal and variability models offered in WP1. In the case of the first group (PTM) we have used the nominal models provided by the University of Arizona and we have assumed three levels, or scenarios, of variability: moderate, high and very high. After the accurate results given from WP1 for 18 and 13 nm, we conclude that the most probable variability scenarios for PTM technologies are the corresponding to the high and manly very high levels. Regarding CNFET we have considered transistors composed by multiple nanotubes, between 4 and 12, with mean value 8 that is the number of nanotubes providing a good reliability level for such technology as showed the works of the University of Stanford.

The analysis at circuit level, performed from accurate electrical models using the previous device models and Hspice simulations, has been applied at two different levels: at specific bit-cell levels, that gives us the essential characteristics of each memorising technique, and at system level, understanding a complete 32KB cache memory system design. At bit-cell levels we have selected, in order to get complete and representative results, the 6T cells for SRAM and the 1T1C and 3T1D cells for DRAM. The analysis at system (cache) level has been performed for both 6T and 3T1D type of cells. The environmental fluctuations considered are the silicon surface temperature and the local $V_{DD}$ voltage supply. Specific analysis about the impact of Single Event Upsets impact analysis has been developed for the 6T SRAM memories and an analysis about the aging effects caused by BTI has been also performed for 3T1D DRAM.

Among the main conclusions, we find that the main environmental factor is temperature, for all the type of circuits and technologies not using CNFETs (in fact the deeper node technology the more aggressive the effect). For sub-22nm technologies a typical factor is a decrease of 60% of the speed (read, write, access) of the memories, both bit-cell or system level, for an increase of temperature between room temperature and 110 °C. In the case of the dynamic memories the effects of the temperature on the retention time is dramatic with impressive reductions (1/100) for the same temperature increase. Also $V_{DD}$ fluctuations affect in the same way all type of memories, around a 15% of speed (the speed decreases when the voltage decreases) for 10% of voltage fluctuation.

The robustness of the 6T bit-cell for process variations (only random independent variations) has been performed using the concept of Satisfiability Boundary. The results show a dramatic drop of
yield for the variability levels determined in WP1 for 18 and 13 nm, as well as for the PTM technologies considering high and very high variability scenarios. Even for only 32KB the calculated yield is unacceptable and even null. The same results have been obtained for the case of 3T1D, perhaps more robust in comparison with 6T but also unacceptable or null yield. In the case of 1T1C results are even more dramatic. The reason for this is that in fact in the analysis of the 1T1C at bit (32 bit) cell level we have contemplated the modern circuit proposed by IBM for the eDRAM, where a sophisticated complementary circuit (to the basic 1T1C cell) is included in order to allow an automatic data recovery in read cycles. In general similar results have been obtained in the analysis of the 32KB memories for both 6T and 3T1D. For that case a percentage between 2 to 3% of the independent write-read cycles causes a data retention failure. These faults are on top of unacceptable decreases of speed (with sigmas of 34% of the nominal value).

About the analysis of reliability drop caused by BTI aging it has been shown that for both 3T1D (analysis performed in this deliverable) and 6T (obtained from publications from other researchers) degradation is a key limiting factor, causing dramatic yield drops if no countermeasure is implemented.

In the analysis of the 32nm technology in front of SEU (SRAM) it has been shown that the critical charge required to cause flips drops from 60 to 15 (fC) when we consider 65 and 32 nm technologies, and the effect is even more drastic when we reduce V\text{DD} levels in stand-by operations. This reduction of the critical charge is expected to follow the same dramatic trend for sub-22nm technologies.

![Figure 77. Overall impact of variations on performances and reliability for sub-18nm technologies](image)

In Section 8 a complete comparison analysis at bit-cell level for bulk-CMOS and CNFET technologies has been developed, only for the 6T cell. Results are interesting and could imply a
hope for the future of memories implemented in emerging technologies (CNTFETS and presumably others). With similar speed and cell area than the corresponding to bulk-16nm, SRAM cells implemented with CNTFETS exhibit very interesting results: practically an independence of temperature fluctuations (similar $V_{DD}$ sensitivity), much higher SNM, much lower static consumption ($1/100$ less of leakage current) with similar dynamic consumption and finally a much higher robustness (and consequently theoretical yield) in front of process variations, conditioned to the future development of the technology. Even for the poor capability to grow nanotubes we have today the scenario of variability is between moderate and high, but if we accept assumable improvements on the control of chirality we obtained variations effects better than moderate.

Figure 77 shows as a matter of general picture, the speed performance variability levels for the different fluctuating causes. Temperature and random process variability are the two key impacting mechanisms. Although temperature has more impact on speed degradation, random process variations are the main cause of yield drop with the inclusion of catastrophic failures. Device degradation and sensitivity to SEU are also significant reliability limiting factors. Systematic process variations and power supply voltage fluctuations are sources with lower impact due to foreseeable improvements in conventional design and implementation techniques.

In the second year the TRAMS project will work on the analysis of alleviating mechanisms for systematic process variations and voltage fluctuations, at layout and circuit level. Regarding temperature we will investigate on innovative adapting and compensation mechanisms, with the use of sensors and compensating knobs. Similar solutions will be applied to attenuate the effect of aging caused by BTI. For process variations we will investigate the trade-offs between mitigation solutions at design level and the resulting performance penalizations. For moderate and high catastrophic and transient failures we will investigate redundancy and error detection and correction codes. For high and very high catastrophic and transient failures we will investigate hierarchic highly redundant fault tolerant architectures.
10. Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>Access Failure</td>
</tr>
<tr>
<td>BTI</td>
<td>Biasing Temperature Instability</td>
</tr>
<tr>
<td>CMOS</td>
<td>Complementary Metal Oxide Semiconductor</td>
</tr>
<tr>
<td>CNT</td>
<td>Carbon NanoTube</td>
</tr>
<tr>
<td>CNFET</td>
<td>Carbon Nanotube Field Effect Transistor</td>
</tr>
<tr>
<td>CVD</td>
<td>Chemical Vapour Deposition</td>
</tr>
<tr>
<td>DRAM</td>
<td>Dynamic Random Access Memory</td>
</tr>
<tr>
<td>DNM</td>
<td>Dynamic Noise Margin</td>
</tr>
<tr>
<td>eDRAM</td>
<td>embedded Dynamic Random Access Memory</td>
</tr>
<tr>
<td>HF</td>
<td>Hold Failure</td>
</tr>
<tr>
<td>HFB</td>
<td>Hyper-rectangle Failure Boundary</td>
</tr>
<tr>
<td>HSB</td>
<td>Hyper-rectangle Satisfiability Boundary</td>
</tr>
<tr>
<td>HPCO</td>
<td>ITRS International Technology Report for Semiconductors</td>
</tr>
<tr>
<td>LER</td>
<td>Line Edge Roughness</td>
</tr>
<tr>
<td>LLC</td>
<td>Last Level Cache</td>
</tr>
<tr>
<td>L1</td>
<td>cache Level 1</td>
</tr>
<tr>
<td>m-CNT</td>
<td>metal Carbon NanoTube</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo method</td>
</tr>
<tr>
<td>MCNT</td>
<td>Mechanical Carbon NanoTube</td>
</tr>
<tr>
<td>MIS</td>
<td>Metal Insulator Semiconductor</td>
</tr>
<tr>
<td>MPFA</td>
<td>Most Probable Failure Analysis</td>
</tr>
<tr>
<td>MPPA</td>
<td>Most Probable Point Analysis</td>
</tr>
<tr>
<td>NBTI</td>
<td>Negative Biasing Temperature Instability</td>
</tr>
<tr>
<td>PBTI</td>
<td>Positive Biasing Temperature Instability</td>
</tr>
<tr>
<td>PTM</td>
<td>Predictive Technology Model</td>
</tr>
<tr>
<td>PVT</td>
<td>Process, Voltage, Temperature variability</td>
</tr>
<tr>
<td>RDD</td>
<td>Random Doping Distribution</td>
</tr>
<tr>
<td>RF</td>
<td>Read Failure</td>
</tr>
</tbody>
</table>
SB  Satisfiability Boundary
SB-SI  Satisfiability Boundary Statistical Integration method
s-CNT  semiconductor Carbon NanoTube
SET  Single Event Transient
SEU  Single Event Upset
SNM  Static Noise Margin
SoC  System on Chip
SRAM  Static Random Access Memory
SSTA  Statistical Static Time Analysis
YENSS  Yield Estimation Non-linear Surface Sampling method
1T1C DRAM  1 Transistor 1 Capacitor Dynamic Random Access Memory
3T1D DRAM  3 Transistors 1 gated Diode Dynamic Random Access Memory
6T SRAM  6 Transistors Static Random Access Memory
11. References


