Étude de la fiabilité et des mécanismes de dégradation dans les composants numériques de dernière génération

Julien Coutet

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DES SCIENCES PHYSIQUES
ET DE L'INGÉNIEUR

SPÉCIALITÉ : ÉLECTRONIQUE

Préparée au sein de la société **Thales SIX GTS FRANCE** à Toulouse
en collaboration avec le **laboratoire de l'Intégration du Matériau au Système** à Talence

Par **Julien COUTET**

Étude de la fiabilité et des mécanismes de dégradation dans les composants numériques de dernière génération

Soutenue le 7 octobre 2020

Membres du jury :

Lorena ANGHEL	Professeur, Grenoble-INP (Grenoble)	Rapporteur
Patrick GIRARD	Directeur de Recherche, LIRMM (Montpellier)	Rapporteur
Cristell MANEUX	Professeur, IMS (Bordeaux)	Président
François MARC	Maitre de Conférences HDR, IMS (Bordeaux)	Directeur de thèse
Christian MOREAU	Ingénieur, DGA-MI (Rennes)	Examinateur
Suzel LAVAGNE	Ingénieur, Thales (Toulouse)	Examinateur
Jean-Claude CLEMENT	Ingénieur, Thales (Palaiseau)	Invité

Remerciements

Une thèse est un travail qui s'inscrit dans la durée. Pendant plus de 3 ans on reste focalisé sur un même sujet. Il y a des moments de détentes, des moments de découvertes et des moments plus tendus. Je tiens à remercier toutes les personnes qui m'ont aidé dans chacune de ces trois parties.

Je tiens premièrement à remercier tout particulièrement François MARC qui a bien voulu être mon directeur de thèse et m'accorder sa confiance. Même si l'éloignement géographique entre Toulouse et Bordeaux n'était pas simple, il a toujours été disponible par téléphone et courriel en cas de besoin. Les derniers mois de rédaction ont été très particuliers. Il y a 3 ans jamais je n'aurais imaginé qu'un confinement national – voir même mondial – allait se produire. François a toujours fourni une relecture minutieuse de mon mémoire au fil des versions, et ce malgré l'enseignement à distance en parallèle, les enfants à la maison, les difficultés matérielles, etc.

Je tiens à remercier Mme Lorena ANGHEL et M. Patrick GIRARD d'avoir accepté d'être les rapporteurs de ma thèse, j'espère que la lecture de ce mémoire leur sera plaisante.

Cette thèse a pris place dans un projet de grande envergure nommé PISTIS. C'était un réel plaisir de travailler avec des personnes compétentes et passionnées. Pour cela je souhaite vivement remercier Jean-Claude CLEMENT ainsi que Patrick CARTON. Je remercie également Christian MOREAU pour ses relectures minutieuses des livrables ainsi que pour ses qualités humaines.

J'ai également eu la chance de travailler et d'être intégré dans une bonne équipe : celle du laboratoire « E^2 LAB » de THALES en partenariat avec le CNES. Pour cela je veux remercier Pierre LEBOSSE qui m'a encadré pendant deux ans - avant de nous quitter pour de nouveau horizons professionnels -, ainsi que Suzel LAVAGNE qui a parfaitement pris le relais. Je tiens également à remercier l'ensemble de mes collègues : Alexandre, Antonin, Aurélien, Flavien, Klára, Léo, Romain, ainsi que le reste de l'équipe.

Il me reste à remercier tous ceux qui m'ont soutenu de manière indirecte : Julien, Florian, Maeva, Benjamin, Juliette, Charlène, Nicolas, Manon, Charlotte et tous les autres. Je tiens

sincèrement à remercier Margaux d'avoir été à mes côtés malgré les moments difficiles. Pour finir, merci à ma mère pour l'éducation qu'elle m'a donnée, et qui me permet d'en être ici aujourd'hui. Merci également à ma petite sœur Tania pour son soutien.

« The weight is a sign of reliability. »
Borris the blade

Table des matières

Liste des abréviations

AC	Alternating Current
ADC	Analog to Digital Converter
ASIC	Application-Specific Integrated Circuit
ATE	Automated Test Equipment
BEOL	Back End Of Line
BL	Bit Line
BTI	Bias Temperature Instability
CHE	Channel Hot Electron
CLB	Configurable Logic Block
CMOS	Complementary Metal Oxide Semiconductor
COTS	Commercial Off-The-Shelf
CRAM	Configuration Random Access Memory
DAHC	Drain Avalanche Hot Carrier
DC	Direct Current
DRAM	Dynamic Random Access Memory
DSM	Deep Sub Micron
DSP	Digital Signal Processor
DUT	Device Under Test
E_a	Energie d'activation apparente
ECC	Error-Correcting Code
EEPROM	Electrically-Erasable Programmable Read-Only Memory
EPROM	Erasable Programmable Read-Only Memory
EM	ElectroMigration
FEOL	Front End Of Line
FG	Floating Gate
FIT	Failures In Time

F-N	Fowler-Nordheim
FPGA	Field Programmable Gate Array
HCI	Hot Carrier Injection
HDL	Hardware Description Language
HKMG	High- κ Metal Gate
IDE	Integrated Development Environment
IEEE	Institute of Electrical and Electronics Engineers
IHM	Interface Homme Machine
IO	Input Output
ISE	Integrated Synthesis Environment
k_B	Constante de Boltzmann ($k_B = 8{,}617330 \times 10^{-5} \mathrm{eV.K^{-1}}$)
LSB	Less Significant Bit
LUT	Look Up Table
MCTF	Mean Cycles To Failure
MLC	Multi-Level Cell
MLE	Maximum Likelihood Estimation
MOS	Metal Oxide Semiconductor
MRAM	Magnetic Random Access Memory
MSB	Most Significant Bit
MTTF	Mean Time To Failure
NBTI	Negative Bias Temperature Instability
OP	Over-Provisioning
P/E	Program/Erase
PBTI	Positive Bias Temperature Instability
PCB	Printed Circuit Board
PEA	Programme d'Etudes Amont
PLL	Phase-locked loop
RAM	Random Access Memory
RO	Ring Oscillator
RTD	Resistance Temperature Detector
SDRAM	Synchronous Dynamic Random Access Memory
SGHE	Secondary Generated Hot Electron

SILC	Stress Induced Leakage Current
SLC	Single-Level Cell
SRAM	Static Random Access Memory
SSD	Solid-State Drive
TAT	Trap Assisted Tunneling
TDDB	Time Dependent Dielectric Breakdown
UART	Universal Asynchronous Receiver-Transmitter
UTE	Union Technique de l'Electricité
VHDL	VHSIC Hardware Description Language
VHSIC	Very High Speed Integrated Circuit
VLSI	Very-Large-Scale Integration
WL	Word Line
WLR	Wafer Level Reliability
XIP	eXecute In Place
ZM	Zone Motif

Liste des tableaux

Liste des figures

Introduction

Dans le domaine de l'électronique, la recherche des performances est indissociable de l'adoption des nouvelles technologies. Ces avancées se font notamment par la diminution de la taille des transistors gravés sur un circuit intégré. Cette course à la finesse de gravure est aujourd'hui essentiellement motivée – et financée – par les besoins des composants grand public qui équipent par exemple les smartphones. Les équipements électroniques complexes dans les domaines avioniques, militaires et spatiaux utilisent également ces composants commerciaux à condition qu'ils soient éprouvés (technologies anciennes mais matures). Afin d'améliorer les possibilités, les performances et les coûts de ces équipements, l'utilisation des dernières technologies de composants numériques est incontournable.

Les dernières générations de composants à la pointe de la performance sont pour la plupart destinées au marché grand public (COTS, ou composants sur l'étagère). Ces composants standards sont dimensionnés pour des besoins classiques où la durée de vie est courte (5 ans tout au plus) comparés aux besoins de l'industrie aéronautique ou militaire (au-delà de 15 ans). De plus, les conditions environnementales subies par ces composants sont extrêmes comparées à leurs homologues commerciaux. Auparavant, ces composants grand public présentaient une fiabilité effective compatible avec ces marchés particuliers. Pour des missions spécifiques dans les domaines militaires, avioniques et spatiaux, des composants dédiés durcis (et par conséquent plus onéreux) étaient conçus ; c'est le cas par exemple des ASIC. Dorénavant, ces domaines utilisent des composants du marché grand public afin d'améliorer leurs performances. Il devient alors nécessaire d'en évaluer la fiabilité.

Sur les technologies standards, avec la poursuite de la réduction d'échelle, les dernières technologies de circuits intégrés utilisent des transistors de plus en plus petits (65nm, 45nm, 28nm, 20nm, etc.). Ces miniaturisations s'accompagnent de nouveaux matériaux. C'est par exemple le cas du nœud 45nm où l'oxyde de grille a changé au profit d'un isolant à haute permittivité diélectrique. Ce changement a été effectué afin de pouvoir supporter le fort champ électrique entre le canal et la grille dorénavant très proche.

Les évolutions structurelles impactent la fiabilité du composant. De nouveaux mécanismes de défaillance apparaissent ; et d'autres autrefois négligeables peuvent ne plus l'être. Cette baisse de fiabilité peut avoir de lourdes conséquences sur le maintien en condition

opérationnelle ainsi que sur la disponibilité des équipements industriels. C'est pourquoi il est indispensable d'évaluer la fiabilité des technologies au moins DSM (Deep Sub-Micron). Afin de modéliser au mieux cette fiabilité, il est important de comprendre chacun des mécanismes dégradant les performances. Cette compréhension physique permettra de remonter de la couche élémentaire (le transistor MOS principalement) au niveau système (un circuit numérique complexe par exemple).

Ces problématiques de fiabilité sont communes à tous les industriels intégrateurs d'électronique. De manière à mener des études conjointes et partageables entre industriels, universitaires et DGA (Direction Générale de l'Armement), une étude avancée est lancée sous l'égide de la DGA : le Programme d'Etude Amont « PISTIS ». Elle a démarré en 2015 pour une durée de 5 ans. Ce PEA permettra notamment de s'assurer que la fiabilité des composants DSM est compatible avec les contraintes de la défense, de l'avionique ou du spatial. Ce projet - découpé en plusieurs postes - aborde trois grandes familles de composants : les composants DSM, les composants de puissance standard et les composants de puissance RF. On y distingue des parties applicatives comme le développement de banc de test, et des parties plutôt théoriques comme l'élaboration de modèles de fiabilité. Les essais du poste DSM - dans lequel s'intègre ce mémoire - sont physiquement situés dans le centre d'expertise de composants de Thales (situé sur la plateforme commune du CNES à Toulouse) en collaboration avec l'IMS (Université de Bordeaux).

Le terme de composant DSM est vaste, c'est pourquoi un choix sur les composants à tester a été fait. Les deux catégories choisies pour représenter les technologies DSM sont des mémoires (Flash NAND en 20 nm, Flash NOR en 65 nm et DDR3 en 20 nm) et des FPGA en 28 nm : l'une représentant l'augmentation brute de la densité d'intégration et l'autre les performances.

Le stockage de masse est une problématique actuelle. Il y a un réel besoin d'augmentation de la quantité d'information contenue sur une même puce. L'émergence des mémoires Flash NAND dans les clefs USB et principalement les SSD de nos jours intéresse également les industriels. La possibilité d'intégrer ce type de composant dans des équipements en environnements sévères décuplerait de nombreuses capacités. Cependant, ce type de mémoire destiné majoritairement au grand public est particulière à utiliser. Ses nombreuses spécificités ont un impact direct sur sa fiabilité.

Contrairement au fabricant qui peut élaborer des architectures de test WLR (fiabilité au niveau wafer) précises, l'utilisateur ne peut pas faire ses propres structures de test pour la mesure du vieillissement des circuits numériques. Cependant les FPGA constituent une

exception car il est possible de les configurer pour constituer une structure de test à base d'oscillateurs en anneaux afin de générer et de mesurer simplement des dégradations au niveau transistors MOS. Cette structure permet de mesurer des dérives paramétriques dues au vieillissement. L'étude de ce type de circuit est ainsi le second point majeur de cette thèse.

Ce mémoire va donc tout naturellement se scinder en trois parties. Le premier chapitre concernera les définitions de base de la statistique. Ensuite, les technologies évoquées dans cette étude seront décrites une à une. Enfin, ce chapitre se terminera par la présentation des mécanismes de défaillance associés.

Le second chapitre abordera la fiabilité des mémoires Flash. Il présentera les essais mis en places afin d'identifier les mécanismes de défaillance. En plus de l'analyse de l'influence de la température de stockage et du nombre de cycle d'écriture-effacement, l'originalité de cette étude est la considération de plusieurs températures d'écriture. Les aspects « rétention de données » et « endurance » seront différenciés. Nous accorderons une attention particulière à la modélisation de la fiabilité dans le cadre d'une utilisation réelle type SSD. Par conséquent, notre méthode d'estimation de la fiabilité sera généralisée à plusieurs codes correcteurs d'erreurs ainsi qu'à différents niveaux de surdimensionnements.

Enfin, le dernier chapitre détaillera toute la démarche mise en œuvre pour mesurer et modéliser le vieillissement dans les FPGA. Cette partie sera divisée en 4 sous-parties. La première présentera la mise en place des essais. La seconde formalisera le traitement des données qui a été mis en place. Les méthodes d'extraction des mécanismes de vieillissement seront ensuite décrites. Nous aborderons uniquement les mécanismes de dégradation BTI et HCI car nous n'en avons pas observé d'autres. Finalement, une analyse fine de la cinétique des dégradations sera faite. De plus, l'aspect statistique des résultats sera également traité avec une méthodologie réaliste.

Chapitre 1

Etat de l'art

Dans ce mémoire, nous allons aborder fiabilité et mécanismes de dégradation au sein du monde du silicium. Ainsi, avant de détailler les études réalisées, ce chapitre récapitulera dans un premier temps les outils et définitions de base de la fiabilité au sens statistique. Puis dans un second temps, la technologie et le fonctionnement des différents composants étudiés durant cette thèse (mémoire et FPGA) seront décrits. Enfin un état de l'art des mécanismes pertinents de dégradation sera effectué. Ces mécanismes sont pour la plupart activement étudiés par la communauté scientifique et industrielle, avec régulièrement de nouvelles publications.

1.1 Approche probabiliste et statistique de la fiabilité

Cette section a pour objectif d'introduire toutes les notions liées à la fiabilité et le vocabulaire associé qui seront utilisés dans la suite du document.

1.1.1 Définitions probabilistes

Fiabilité

L'Union Technique de l'Electricité (UTE) a proposé la définition suivante :

> *« La fiabilité d'une entité est son aptitude à assurer une **fonction** requise pendant un **temps** donné dans des **conditions** données. »*

De même :

> « *Une défaillance est la cessation de l'aptitude à assurer une fonction requise.* »

Il existe deux types de défaillance:

- Soit la panne est soudaine, il s'agit alors de *défaillance catalectique*.
- Soit la dégradation d'un paramètre est progressive, on parle alors de *défaillance paramétrique*. On fixe alors un critère de défaillance à partir duquel on considère le composant comme inutilisable.

Un système peut être soit réparable, soit non réparable. Dans le premier cas, lorsqu'une défaillance intervient, une phase de maintenance est mise en œuvre afin de le rendre de nouveau opérationnel jusqu'à la prochaine défaillance. Dans le cas des systèmes non réparables, lorsque le composant est défaillant, aucune maintenance n'est prévue.
Ce mémoire traitera uniquement des composants sans maintenance, ce qui est par exemple le cas des systèmes spatiaux ou des composants électroniques au sein d'une carte.

Fonction de survie

Afin de décrire la fiabilité sous un aspect plus mathématique, on définit la fonction de survie ou fonction de fiabilité. Cette fonction notée *R(t)* (R pour « *reliability* ») correspond à la probabilité que le système soit encore fonctionnel à l'instant t.

$$R(t) = P[T > t] \tag{1}$$

Où *T* est la durée de vie de l'entité considérée. *T* est considéré comme une variable aléatoire.
Pour les systèmes non réparables, c'est une fonction décroissante prenant des valeurs comprises entre 0 et 1. Par convention elle a pour valeur 1 à t = 0 car le système fonctionne toujours à l'état initial.

Fonction de défaillance

A l'inverse, la fonction de défaillance (ou de défiabilité) notée *F(t)* (F pour « *failure* ») correspond à la probabilité que le système soit défaillant à l'instant t.

$$F(t) = P[T \leq t] \tag{2}$$

Du fait de cette définition, $F(t)$ est aussi la fonction de répartition de la variable aléatoire *T*.

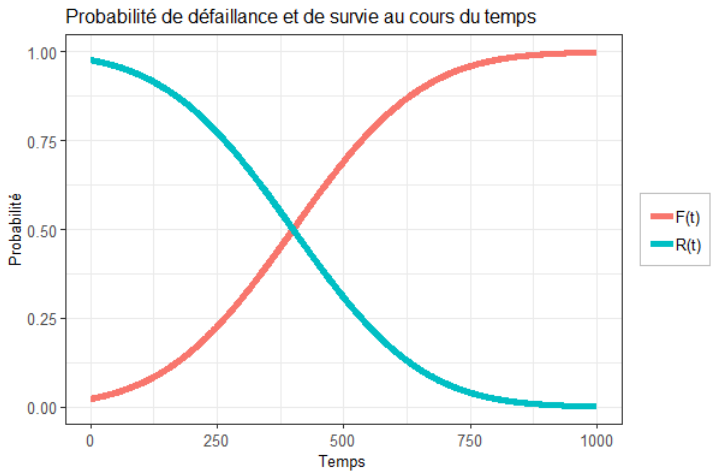


Figure 1 : Exemple fictif de la dualité entre la fonction de survie et la fonction de répartition

On remarquera que :

$$F(t) = 1 - R(t) \tag{3}$$

La Figure 1 illustre la dualité entre la fonction de survie (courbe bleue) et la fonction de répartition (courbe rouge). La courbe en rouge augmente avec le nombre de composants en panne au cours du temps alors que celle en bleu diminue.

Densité de probabilité de défaillance

La densité de probabilité de défaillance est définie par :

$$f(t) = \lim_{\Delta t \to 0} \frac{P[t<T\leq t+\Delta t\,]}{\Delta t} \tag{4}$$

Notons également que :

$$f(t) = \frac{dF(t)}{dt} = -\frac{dR(t)}{dt} \tag{5}$$

La densité de probabilité f correspond à la limite de l'histogramme des instants de panne (courbe bleue sur la Figure 2).

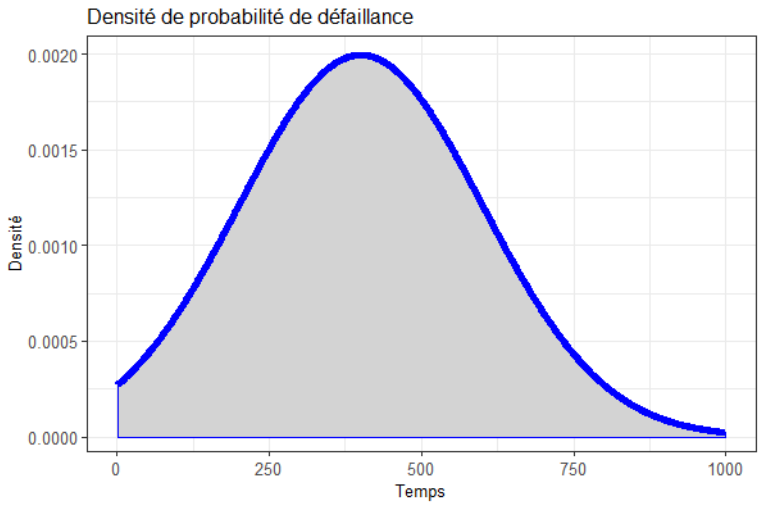


Figure 2 : Exemple de densité de probabilité

Taux de défaillance

Le taux de défaillance est la densité de probabilité que le système soit en panne à un instant $t + dt$ sachant qu'il était encore fonctionnel à l'instant t.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P[T\leq t+\Delta t \mid T>t]}{\Delta t} \tag{6}$$

Soit :

$$\lambda(t) = \frac{1}{P[T>t]} \times \lim_{\Delta t \to 0} \frac{P[t<T\leq t+\Delta t]}{\Delta t} \quad (7)$$

On peut ainsi exprimer plus simplement $\lambda(t)$ sous la forme :

$$\lambda(t) = \frac{f(t)}{R(t)} \quad (8)$$

Le taux de défaillance est exprimé en panne par heure. On peut également l'exprimer en FIT (Failures In Time).

$$1\ FIT = 10^{9}.heures^{-1} \quad (9)$$

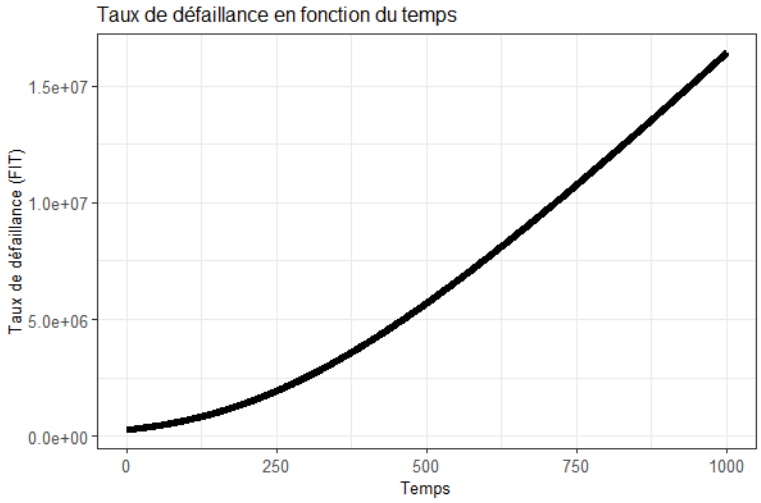


Figure 3 : Exemple du taux de défaillance

MTTF

Le *Mean Time To Fail* (MTTF) est le temps moyen de bon fonctionnement du composant. Il correspond à l'espérance mathématique *E[T]* des instants de pannes.

$$MTTF = E[T] = \int_0^{+\infty} t.f(t).dt \quad (10)$$

Afin d'éviter toute confusion dans ce mémoire, le *Median Time To Fail* sera plutôt noté $t_{50\%}$.

MTBF

Le *Mean Time Between Failures* (MTBF) correspond au temps moyen entre deux pannes consécutives d'un système réparable. Dans le cas d'un système non réparable, comme c'est ici le cas, le MTBF correspond au MTTF.

$t_{50\%}$ et $t_{0.1\%}$

Le $t_{50\%}$ correspond au temps où 50% des échantillons sont défaillants. Cette valeur s'appelle également la valeur médiane.

$$F(t_{50\%}) = 0{,}5 \tag{11}$$

De même, le $t_{0,1\%}$ correspond au temps où 0,1% des échantillons sont défaillants.

$$F(t_{0,1\%}) = 0{,}001 \tag{12}$$

1.1.2 Définitions statistiques

Cette section décrit toutes les fonctions statistiques analogues à celles de la partie probabiliste. Les fonctions statistiques sont construites au moyen d'observations parmi une population d'échantillons donnés.

Fonction de survie

La détermination statistique de la fonction de survie *R* à partir des résultats observés peut se faire simplement de la manière suivante :

$$R_{réel}(t_i) = \frac{N_{total} - N_{t_i}}{N_{total}} \tag{13}$$

Où :

- $R_{réel}$ est la fonction de survie observée durant le test
- t_i est un instant de test où une observation a été réalisée
- N_{t_i} est le nombre de composants défaillants entre l'instant t_0 et l'instant t_i
- N_{total} est le nombre total de composants sous test

La visualisation de cette fonction discrète est une « courbe » en escalier comme illustré Figure 4.

Cette fonction empirique a l'inconvénient de ne pas tenir compte des données censurées à droite (échantillons non tombés en panne à la fin de l'essai). Or à l'issue de campagnes de test, le taux de censure à droite est souvent relativement élevé, et donc non négligeable. En cas de données censurées, il est préférable d'utiliser l'estimateur de Kaplan-Meier présenté dans la section 1.1.3.

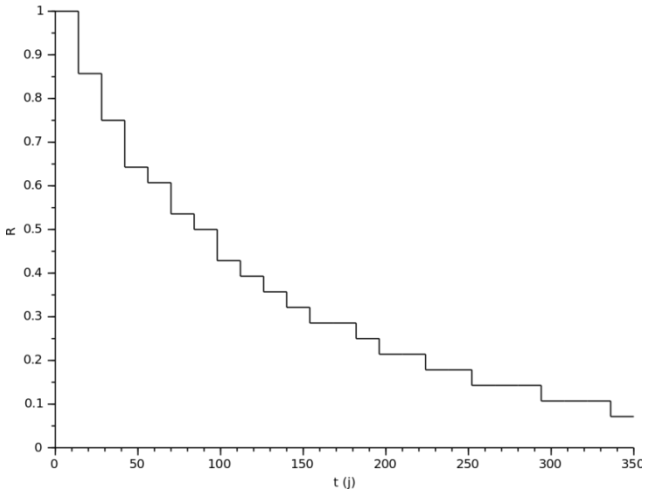


Figure 4 : Exemple de fonction de survie empirique en escalier

Fonction de défaillance

La fonction de défaillance empirique observée peut être écrite :

$$F_{réel}(t_i) = \frac{N_{t_i}}{N_{total}} \tag{14}$$

Où $F_{réel}$ est la fonction de défaillance observée durant le test.

Densité de probabilité de défaillance

La densité de probabilité de défaillance observée est l'histogramme des instants de panne.

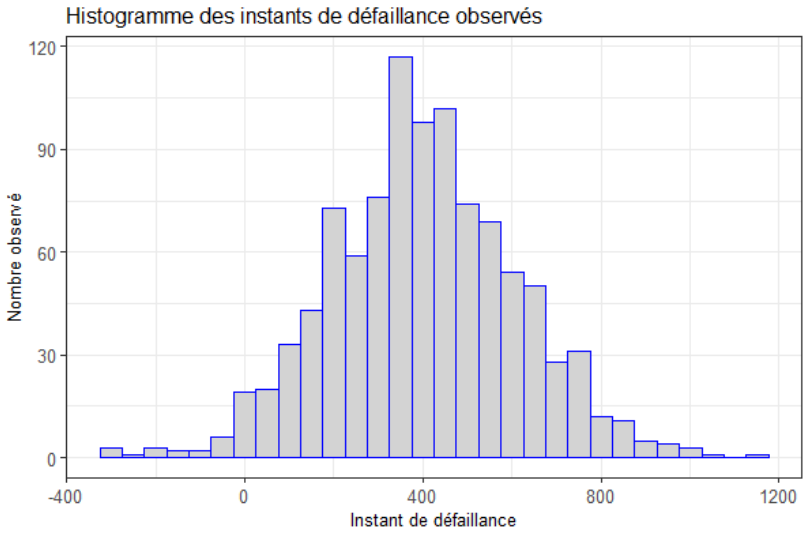


Figure 5 : Exemple d'histogramme expérimental des instants de panne

MTTF

Le temps moyen de bon fonctionnement observé est :

$$MTTF_{réel} = \frac{\sum_{i=1}^{N_{total}} t_{f_i}}{N_{total}} \quad (15)$$

Où :

- $MTTF_{réel}$ est le temps moyen de survie observé
- t_{f_i} est l'instant de défaillance de l'échantillon *i*

1.1.3 Censure

Une donnée est dite "censurée" si on ne connaît pas l'instant de défaillance exact, mais seulement une estimation inférieure ou supérieure. Il y a plusieurs types de censures. Nous allons uniquement aborder dans cette sous-partie les deux plus courantes qui seront utilisées dans ce mémoire.

Censure à droite

On parle de censure à droite lorsque des composants sont encore en vie à la fin du test. En pratique, on n'attend pas que toutes les pièces soient en panne pour faire des prévisions statistiques.

Taux de censure

On appelle taux de censure le ratio entre le nombre d'échantillon n'ayant pas présenté de défaillance à la fin des heures de test et la population totale.

$$cens_{rate} = R_{réel}(T_{final}) = 1 - F_{réel}(T_{final}) = 1 - \frac{N_{T_{final}}}{N_{total}} \quad (16)$$

Où :

- T_{final} est la durée totale du test
- $N_{T_{final}}$ est le nombre de composant défaillant à la fin du test

Un taux de censure trop élevé amènera une baisse de la précision et de la pertinence des calculs menés. Il faut ainsi dans l'idéal un grand nombre d'échantillons et un faible taux de censure.

Censure par intervalle

Il y a censure par intervalle lorsque l'on ne peut pas connaitre exactement chaque instant de défaillance mais que l'on peut uniquement l'encadrer dans un intervalle de temps. Ici, cela correspond à l'intervalle entre deux étapes de mesures consécutives.

Estimateur de Kaplan-Meier

Lorsque l'on est en présence d'un taux de censure élevé, la fonction de survie empirique $R_{réel}$ ne permet plus d'être représentative. En effet, elle ne tient pas compte des censures. Edward L. Kaplan et Paul Meier ont proposé en 1958 [1], [2] l'estimateur suivant :

$$\hat{R}(t_i) = \prod_{j=0}^{i} \frac{n_i - d_i}{n_i} \quad (17)$$

Où :

- $\hat{R}(\,)$ est la fonction de survie estimée
- t_i est un instant de test où une observation a été réalisée
- n_i est le nombre de survivant juste avant t_i (ou à l'instant t_{i-1}) moins le nombre de censure ; c.-à-d. le nombre de composant à risque observable à l'instant t_i
- d_i est le nombre de défaillance entre t_{i-1} et t_i

Cet estimateur largement utilisé porte le nom d'estimateur de Kaplan-Meier. Sa prise en compte de la censure en fait un bon estimateur, assez représentatif de la réalité. Ce dernier sera utilisé pour les représentations graphiques de toutes les fonctions de répartition empirique de cette thèse.

1.1.4 Lois de vieillissement dans les technologies siliciums

1.1.4.1 Courbe en baignoire

Le taux de défaillance des systèmes électroniques suit généralement la courbe représentée par la Figure 6. On la nomme communément « courbe en baignoire ».

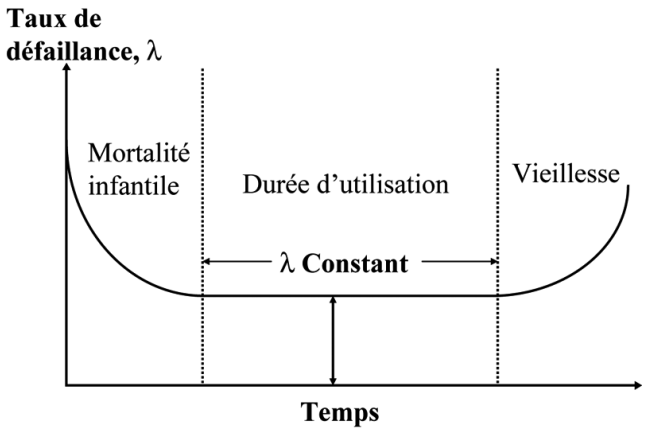


Figure 6 : Courbe en baignoire des systèmes électroniques

Elle se compose de trois parties :

- La période de jeunesse : pendant cette période, le taux de défaillance est décroissant avec le temps. Elle correspond à un manque de maturité du procédé de fabrication.
- La vie utile du composant : pendant cette période, le taux de défaillance est constant. Les pannes du fond de la baignoire correspondent principalement à des causes externes.
- La vieillesse du composant : c'est la période d'usure, le taux de défaillance croit avec le temps. Les différents mécanismes de vieillissement se cumulent jusqu'à déclencher une défaillance.

1.1.4.2 Distributions

Cette section a pour but de récapituler rapidement les fondements des distributions statistiques utilisées dans ce mémoire. Un soin particulier sera apporté aux distributions modélisant le vieillissement des technologies silicium.

1.1.4.2.1 Distributions classiques

Nous présentons les principales lois qui apparaitront dans ce mémoire dans le Tableau 1.

Nom	Grandeurs et expressions
Loi Normale Notée $\mathcal{N}(\mu, \sigma^2)$	$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$ $F(t) = \frac{1}{2} + \frac{1}{2}.erf\left(\frac{t-\mu}{\sigma.\sqrt{2}}\right)$ $F(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$
Loi Exponentielle	$f(t) = \lambda.e^{-\lambda.t}$ $F(t) = 1 - e^{-\lambda.t}$
Loi de Weibull Notée $\mathcal{W}(\beta, \eta)$	$f(t) = \frac{\beta}{\eta}.\left(\frac{t}{\eta}\right)^{\beta-1}.e^{-\left(\frac{t}{\eta}\right)^{\beta}}$ $F(t) = 1 - \exp\left(-\left(\frac{t}{\eta}\right)^{\beta}\right)$
Loi Log-normale	$f(t) = \frac{1}{\sigma.t.\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{ln(t)-\mu}{\sigma}\right)^2}$ $F(t) = \frac{1}{2} + \frac{1}{2}.erf\left(\frac{ln(t)-\mu}{\sigma.\sqrt{2}}\right)$ $F(t) = \Phi\left(\frac{ln(t)-\mu}{\sigma}\right)$

Loi Binomiale Notée $\mathcal{B}(n,p)$	$P[X=k]=\binom{n}{k}.p^k.(1-p)^{n-k}$

Tableau 1 : Distributions statistiques classiques

Parmi ces lois, seules les lois continues (Normale, Weibull, Exponentielle et Log-normale) peuvent avoir un temps comme variable aléatoire. Nous ne détaillons que les trois lois couramment utilisées pour modéliser les durées de vie. Les grandeurs caractéristiques y seront mentionnées ainsi que leurs expressions.

1.1.4.2.2 Distribution exponentielle

La loi exponentielle est la plus utilisée pour modéliser les pannes dans le monde du silicium. C'est une loi simple, très utilisée en fiabilité, dont le taux de défaillance est constant (noté λ) au cours du temps. Ce taux de défaillance constant traduit la modélisation de phénomènes sans mémoire. En effet, la probabilité de panne à $t+\varepsilon$, sachant qu'il n'y a pas eu de panne entre 0 et t, sera la même qu'entre 0 et ε. Cette distribution modélise donc les pannes catalectiques (parfois appelées pannes aléatoires par abus de langage) qui composent le fond plat de la courbe en baignoire (voir Figure 6).

Dans le cas d'une loi exponentielle, et seulement dans ce cas, on a :

$$\lambda(t)=\lambda=\frac{1}{MTTF} \tag{18}$$

1.1.4.2.3 Distribution de Weibull

La distribution statistique de Weibull est souvent utilisée pour modéliser le vieillissement des systèmes électroniques. Elle peut modéliser une fonction de risque décroissante, croissante ou constante, ce qui permet de décrire toutes les phases de la durée de vie d'un objet.
Sa fonction de répartition est la suivante :

$$F(t)=1-\exp\left(-\left(\frac{t}{\eta}\right)^{\beta}\right) \tag{19}$$

Où :

- F est la fonction de répartition
- t est le temps
- β est le paramètre de forme
- η est le paramètre d'échelle

La valeur du paramètre d'échelle correspond au temps où ~63% des échantillons sont défaillants, c.-à-d. $F(\eta)=1-e^{-1}\approx 0{,}63$.

L'espérance, ou le temps moyen avant défaillance (MTTF), est donnée par l'expression suivante :

$$MTTF = \eta.\Gamma\left(1+\frac{1}{\beta}\right) \quad (20)$$

Où Γ( ) est la fonction Gamma.

Le t_x est le temps au bout duquel une proportion x des échantillons est défaillante.

$$t_x = \eta.\left(-ln(1-x)\right)^{\frac{1}{\beta}} \quad (21)$$

La valeur médiane ($t_{50\%}$) est donnée par:

$$t_{50\%} = \eta.\left(ln(2)\right)^{\frac{1}{\beta}} \quad (22)$$

Le taux de défaillance d'une distribution de Weibull est le suivant :

$$\lambda(t) = \frac{\beta}{\eta}.\left(\frac{t}{\eta}\right)^{\beta-1} \quad (23)$$

Suivant la valeur de β, la distribution de Weibull peut décrire chacune des trois parties de la courbe en baignoire :

- Une valeur de β inférieure à 1 traduit une décroissance du taux de défaillance, ce qui correspond à la période de jeunesse d'un système.
- Une valeur égale à 1 correspond à la loi exponentielle avec un taux de défaillance constant.
- Un paramètre β supérieur à 1 correspond à la remontée du taux de défaillance, donc au vieillissement du système.

1.1.4.2.4 Distribution log-normale

Une variable aléatoire T suit une loi log-normale lorsque son logarithme est distribué selon une loi normale. Le paramètre μ est l'espérance de la variable aléatoire ln(T), et σ est son écart type.

$$ln(T) \sim \mathcal{N}(\mu, \sigma^2) \quad (24)$$

Par analogie avec la distribution de Weibull : μ est le paramètre d'échelle et σ est le paramètre de forme.

Sa fonction de répartition est de la forme suivante :

$$F(t) = \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right) \tag{25}$$

Où Φ est la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$.

On en déduit que le taux de défaillance d'une distribution Log-Normale est le suivant :

$$\lambda(t) = \frac{\frac{1}{t.\sigma.\sqrt{2\pi}}.e^{-\frac{1}{2}\left(\frac{ln(t)-\mu}{\sigma}\right)^2}}{1-\Phi\left(\frac{ln(t)-\mu}{\sigma}\right)} \tag{26}$$

Contrairement à la distribution normale, la distribution log-normale est asymétrique comme illustré sur la Figure 7.

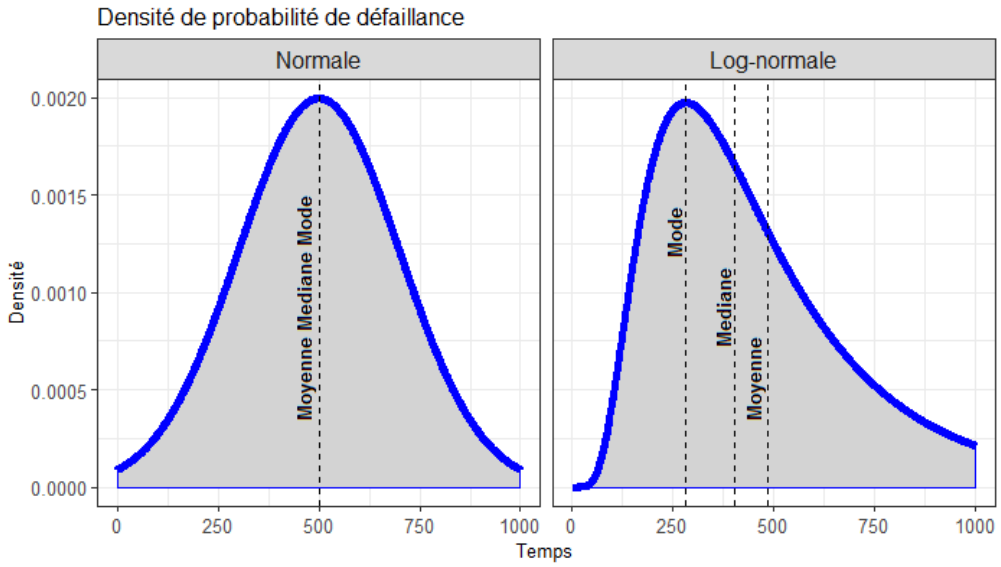


Figure 7 : Comparaison d'une distribution normale et d'une distribution log-normale

Là où pour la loi normale, la médiane et la moyenne sont confondues ; dans le cas de la loi log-normale la moyenne est d'autant plus éloignée de la médiane que la queue de la distribution est importante comme on peut l'observer sur la Figure 7.

Dans une distribution log-normale, le logarithme de la variable aléatoire suit une loi normale. Ainsi, le paramètre μ représente l'espérance et la médiane du logarithme de la variable. Le paramètre σ^2 représente la variance du logarithme de la variable.

Le lien avec la valeur médiane $t_{50\%}$ et t_x se fait à l'aide des expressions suivantes :

$$t_{50\%} = e^{\mu} \tag{27}$$

$$t_x = e^{\mu+\sigma*\phi^{-1}(x)} \tag{28}$$

Où Φ^{-o} est la réciproque de la fonction de répartition (fonction quantile) de la loi normale centrée réduite $\mathcal{N}(0,1)$.

L'espérance de la variable aléatoire T n'est pas « exp(μ) » mais :

$$E(T) = MTTF = exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{29}$$

Souvent Weibull et Log-Normale donnent des résultats assez similaires lorsque le taux de défaillance a une allure monotone. Lorsque celui-ci a une allure présentant un maximum, la loi de Weibull n'est pas utilisable contrairement à la loi log-normale.

La loi log-normale est le modèle de distribution le plus fréquemment utilisé pour de nombreuses applications. Elle est basée sur le modèle de croissance multiplicatif. Cela signifie qu'à n'importe quel moment le procédé subit une augmentation aléatoire de la dégradation proportionnelle à son état actuel. L'effet multiplicatif de toutes ces croissances indépendantes aléatoires s'accumule pour déclencher une défaillance. Par conséquent, cette loi est souvent utilisée pour modéliser des pièces ou des composants présentant une défaillance principalement due à la contrainte ou à la fatigue. Elle est historiquement utilisée pour les mécanismes de défaillance dans les semi-conducteurs (hors brasure et PCB).

1.1.4.3 Facteurs d'accélération

Les défaillances observées des composants sont souvent la résultante d'une accumulation de plusieurs mécanismes. Chacun de ces mécanismes suit une cinétique propre vis-à-vis des différents facteurs physiques vus par le composant. Ces différents cofacteurs sont par exemple la température (température boitier, température de jonction, …), le champ électrique généré par l'alimentation du composant, la fréquence de fonctionnement, etc. Ce sous-chapitre va donc lister les différentes lois mathématiques régissant classiquement ces cinétiques. Elles seront réutilisées dans la suite de ce mémoire dans l'élaboration des modèles.

Un facteur d'accélération est noté *AF*. Dans le cas d'une durée de vie, il intervient comme suit :

$$MTTF_{op} = \frac{MTTF_{ref}}{AF} \tag{30}$$

Où :

- $MTTF_{op}$ est la durée de vie dans des conditions opérationnelles.
- $MTTF_{ref}$ est la durée de vie dans des conditions de références.

Dans le cadre d'un taux de défaillance constant, on a :

$$\lambda_{op} = AF \times \lambda_{ref} \tag{31}$$

Où :

- λ_{op} est le taux de défaillance dans des conditions opérationnelles.
- λ_{ref} est le taux de défaillance dans des conditions de références.

1.1.4.3.1 Loi d'Arrhenius

La loi d'Arrhenius permet de décrire la vitesse d'une réaction chimique en fonction de la température. Elle fut énoncée par Svante August Arrhenius en 1889 [3]. Cette loi est largement utilisée pour calculer le facteur d'accélération thermique lors des tests de vieillissement accélérés [4].

$$AF_{Th} = e^{\frac{E_a}{k_B}\cdot\left(\frac{1}{T_{ref}}-\frac{1}{T_{op}}\right)} \tag{32}$$

Où :

- AF_{Th} est le facteur d'accélération thermique
- E_a est l'énergie d'activation apparente en eV
- K_B est la constante de Boltzmann
- T_{ref} est la température de référence en Kelvin
- T_{op} est la température du composant dans son utilisation normal en Kelvin

Le seul paramètre à déterminer de cette loi est l'énergie d'activation apparente. Le paramètre E_a est supposé constant (indépendant de la température) sur une plage de température limitée. Ses valeurs typiques varient entre -0,2 et 1 eV en fonction du/des mécanisme(s) de dégradations observés et de la technologie du composant silicium.

1.1.4.3.2 Loi d'accélération électrique en exponentielle

Ceci est la loi classique de description de l'accélération électrique selon la JEDEC [4]. Le facteur d'accélération en exponentiel des grandeurs électriques peut être exprimé de la manière suivante [5, p. 138] :

$$AF_{\exp\_V} = e^{\gamma\left(V_{op}-V_{ref}\right)} \tag{33}$$

Où :

- AF_{exp_V} est le facteur d'accélération électrique
- γ est le paramètre d'accélération électrique en 1/V
- V_{op} est la tension nominale du composant en Volt
- V_{ref} est la tension de référence du composant en Volt

L'accélération du vieillissement par augmentation de sa tension d'alimentation permet de réduire considérablement le temps des tests. Cependant, il faut veiller à ne pas trop augmenter cette tension afin de ne pas faire intervenir de nouveaux mécanismes de défaillance absents dans une condition usuelle du composant.

1.1.4.3.3 Loi d'accélération électrique en puissance

Le facteur d'accélération des grandeurs électriques en puissance peut être exprimé de la manière suivante [5, p. 138]:

$$AF_{pwr\_V} = \left(\frac{V_{op}}{V_{ref}}\right)^{\alpha} \tag{34}$$

Où :

- AF_{pwr_V} est le facteur d'accélération électrique
- α est le paramètre d'accélération électrique

Cette loi en puissance permet notamment de décrire l'influence de la tension sur la dégradation de la tension de seuil des transistors due au mécanisme de défaillance NBTI.

1.1.4.3.4 Loi d'accélération électrique de Takeda

Le modèle suivant a été proposé par Takeda et Suzuki pour la première fois en 1983 [6].

$$AF_{takeda} = e^{\alpha.\left(\frac{1}{V_{ref}} - \frac{1}{V_{op}}\right)} \tag{35}$$

Où :

- AF_{takeda} est le facteur d'accélération
- α est le paramètre d'accélération électrique

Ce modèle est particulièrement adapté pour modéliser le facteur d'accélération électrique dans le cas de mécanismes de dégradation par électron chaud (ou HCI).

1.1.5 Principe des méthodes d'estimation des paramètres

Lorsque l'on veut modéliser un phénomène, il faut identifier le modèle le plus adapté puis déterminer les valeurs des paramètres de celui-ci. Pour cette seconde étape, il existe des

méthodes qui permettent d'estimer les paramètres qui génèreront au mieux les résultats observés. Si la forme du modèle est erronée, ces méthodes estimeront tout de même des valeurs pour « coller » au mieux aux mesures mais ce modèle divergera en dehors de la plage des valeurs observées. Or la modélisation est justement utile pour estimer des valeurs en dehors de cette plage. Il est donc primordial d'avoir une forme du modèle pertinente (ayant un sens physique si possible).

Cette section résumera succinctement les principales méthodes d'estimation utilisées dans ce mémoire. Ces dernières seront confrontées en incluant leurs avantages et leurs inconvénients.

1.1.5.1 Estimateur des moindres carrés

La méthode d'estimation des moindres carrés permet de comparer des données expérimentales avec un modèle. Le but étant de chercher le jeu de paramètres du modèle ayant la plus faible différence au carrée. Pour un ensemble de N observations $\mathbf{y} = (y_i)_{i\in[\![1;N]\!]}$ de p cofacteurs respectifs $\mathbf{x_i} = \left(x_{k,i}\right)_{k\in[\![1;p]\!]}$, on construit le modèle f à m paramètres $\boldsymbol{\theta} = \left(\theta_j\right)_{j\in[\![1;m]\!]}$ de la manière suivante :

$$\widehat{\boldsymbol{\theta}} = argmin_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(y_i - f_{\boldsymbol{\theta}}(\mathbf{x_i})\right)^2 \tag{36}$$

Cette minimisation se fait de manière numérique pour des fonctions non linéaires.

$$y_i = \underbrace{f_{\widehat{\boldsymbol{\theta}}}(\mathbf{x_i})}_{\hat{y}_i} + \varepsilon_i \; ; \; \forall i \in [\![1, N]\!] \tag{37}$$

Où les ε_i sont les erreurs résiduelles du modèle.

Ces résidus sont supposés suivre une loi normale identique pour tout i ; et dans ce cas : les estimateurs $\widehat{\boldsymbol{\theta}}$ construits sont sans biais et de variance minimale.

Cette méthode d'estimation d'un modèle a de nombreux avantages : elle n'est pas biaisée – sous conditions d'une même loi normale sur les erreurs -, elle est applicable sur des fonctions non linéaires et elle est simple à mettre en application. Cependant, dans le cas de l'estimation des paramètres d'une distribution, cette méthode ne tient pas compte de la censure (contrairement au maximum de vraisemblance) ce qui limite son champ d'utilisation.

<u>Cas de la régression linéaire sur des variables transformées</u>

L'une des méthodes les plus courantes dans l'estimation des paramètres d'un modèle est la régression linéaire (qui peut être graphique dans le cas d'une seule variable). Afin de pouvoir appliquer cette méthode, il faut mettre le modèle sous la forme d'une somme de fonctions

(d'un cofacteur à la fois) comme sur l'équation 38. La régression linéaire est une méthode des moindres carrés sur un modèle f linéaire [7].

$$f_{\boldsymbol{\theta}}(\mathbf{x_i}) = \theta_0 + \sum_{j=1}^{j=p} \theta_j \times x_{j,i} \tag{38}$$

Un modèle est de bonne qualité si [8]:

- les vraies valeurs y_i sont globalement proches des valeurs ajustées $\widehat{y_i}$ (l'erreur résiduelle moyenne est faible)
- il n'y a aucun point aberrant ou trop influent (point levier), ceci pouvant fausser le modèle
- la liaison entre les variables est bien linéaire (ce qui se traduit par un coefficient de corrélation élevé)
- l'analyse des résidus ε_i ne laisse apparaître aucune structure identifiable (ils suivent une même loi normale)

Dans le cas de variables transformées afin de rendre le modèle linéaire, l'hypothèse où les erreurs résiduelles suivent une même loi normale n'est plus valide. Par exemple lors d'un passage en logarithme sur les valeurs, les erreurs ε_i des petites valeurs vont être grandement amplifiées par rapport aux grandes valeurs.

Pour que la représentation soit graphique, on utilise usuellement un seul cofacteur par régression linéaire. Pour cela, il faut placer tous les autres cofacteurs dans les mêmes conditions afin de les rendre comparables. Prenons par exemple l'estimation graphique d'une énergie d'activation à partir de plusieurs instants de défaillances avec leur température de vieillissement respective. Après passage en logarithme, nous avons un modèle de forme linéaire (voir équation 39) avec des variables transformées. Dans ce cadre les erreurs résiduelles ne suivent plus une loi normale identique pour tout T.

$$ln\big(MTTF(T)\big) = E_a.\frac{1}{k_B.T} + ln(MTTF_0) \tag{39}$$

Ainsi, une régression linéaire sur le logarithme des temps moyens de défaillance par température en fonction de $1/(K_B \times T)$ nous permettra d'extraire la valeur de l'énergie d'activation. Pour une même tension, par exemple 1.4V, l'étude fournit des pièces vieillies à cinq températures de 25°C à 125°C comme illustré sur l'exemple fictif de la Figure 8.

Cependant, cette méthode n'est valable que si l'énergie d'activation est indépendante de la température et si l'on est bien en présence d'une loi d'Arrhenius. Ainsi les points tracés seront plus ou moins alignés. Cette méthode est peu précise mais permet de donner une

première valeur approximative d'E_a ainsi que la concordance des mesures avec le modèle en fonction de l'alignement des points.

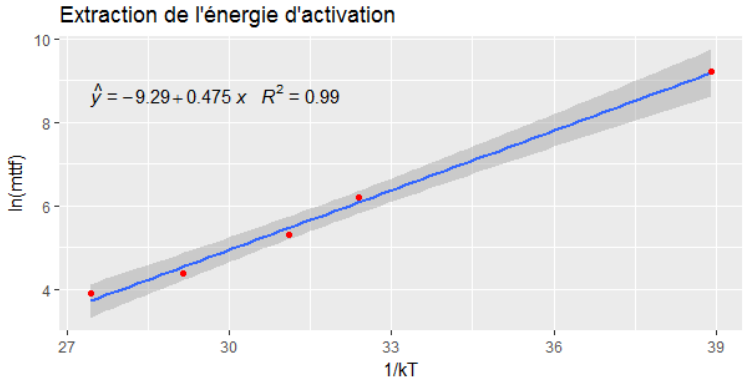


Figure 8 : Exemple fictif de régression pour estimer l'énergie d'activation

L'avantage de la régression linéaire est bien sûr sa simplicité et sa rapidité de mise en œuvre, mais elle apporte aussi beaucoup d'inconvénients. La présence de points dits levier (particulièrement aux extrémités) fausse l'estimation en leur donnant un poids trop important. De plus, la transformation des variables modifie les distributions des erreurs résiduelles ce qui invalide le caractère « sans biais » de la méthode des moindres carrés.

1.1.5.2 Méthode du Maximum de vraisemblance

La méthode du maximum de vraisemblance est utilisée pour estimer les paramètres d'une loi de probabilité d'une population donnée en recherchant le jeu de valeurs des paramètres maximisant la fonction de vraisemblance, c.-à-d. pour lequel les données observées présentent la plus grande probabilité [9], [10]. Elle fut élaborée en 1922 par le statisticien Ronald Aylmer Fisher. Les résultats obtenus par cette méthode sont beaucoup plus précis et fiables à condition que le modèle soit bien défini. De plus, elle prend en compte l'ensemble des informations disponibles.

Pour commencer, il faut définir la vraisemblance noté $\mathcal{L}_\theta$ (« Likelihood ») de l'estimation des paramètres θ. Dans la mesure où les données sont censurées (par intervalles), la vraisemblance des paramètres de la distribution statistique choisie est la probabilité conjointe d'apparition des évènements observés:

$$\mathcal{L}_\theta(\mathrm{t},\mathrm{X},\mathrm{Y}) = [\prod_{\mathrm{i}}(P_\theta[t_i < T \leq t_{i+1}])^{\mathrm{X_i}}].\left[\prod_{\mathrm{j}}\left(P_\theta\left[T > t_j\right]\right)^{\mathrm{Y_j}}\right] \qquad (40)$$

Ou de manière plus explicite :

$$\mathcal{L}_\theta(t, X, Y) = \left[\prod_i \left(F_\theta(t_{i+1}) - F_\theta(t_i)\right)^{X_i}\right] . \left[\prod_j R_\theta(t_j)^{Y_j}\right] \quad (41)$$

Où :

- t_i sont les différents temps où une mesure a été effectuée
- X_i est le nombre de pièces défaillantes entre ti et ti+1. (Censuré par intervalle)
- Y_j est le nombre de pièces encore vivante à l'instant t_j, t_j étant la dernière étape de mesure de la pièce. (Censuré à droite)

Après passage en logarithme pour la simplicité des calculs par ordinateur:

$$ln\left(\mathcal{L}_\theta(t, X, Y)\right) = \left[\sum_i X_i . \ln\left(F_\theta(t_{i+1}) - F_\theta(t_i)\right)\right] + \left[\sum_j Y_j . \ln\left(R_\theta(t_j)\right)\right] \quad (42)$$

Ensuite, il suffit d'utiliser des méthodes d'optimisation afin de trouver le jeu de paramètre ayant la valeur de vraisemblance la plus forte.

Exemple de l'estimation des paramètres d'une distribution de Weibull

Nous remplaçons par l'expression de la distribution de Weibull :

$$ln\left(\mathcal{L}_{\beta,\eta}(t, X, Y)\right) = \left[\sum_i X_i . \ln\left(e^{-\left(\frac{t_i}{\eta}\right)^\beta} - e^{-\left(\frac{t_{i+1}}{\eta}\right)^\beta}\right)\right] - \left[\sum_j Y_j . \left(\frac{t_j}{\eta}\right)^\beta\right] \quad (43)$$

1.2 Description des technologies des composants électroniques étudiés

Afin de comprendre au mieux les mécanismes de dégradations des différents composants étudiés dans cette thèse, il faut tout d'abord définir les technologies dont il est question. La structure élémentaire subissant le vieillissement étudié est toujours en finalité le transistor dans cette étude. Cependant, les architectures des systèmes varient. Il convient ainsi d'expliciter les structures silicium VLSI de type FPGA et mémoire Flash.

Cette étude ne prend en compte que les effets du vieillissement du silicium et des transistors et élimine les problématiques de défaillance des boitiers, cartes électroniques ou autres éléments du banc de test.

1.2.1 Premier composant à l'étude : le FPGA

Les FPGA (Field Programmable Gate Array) sont des circuits logiques configurables rapidement et simplement. Ils ont été introduits en 1985 par Xilinx [12]. La configuration se fait via un langage HDL (Hardware Description Language) tel que le VHDL (Europe) ou le Verilog (USA). Là où les premiers FPGA embarquaient moins de 10 000 portes logiques, avec les progrès technologiques, les dernières générations peuvent en comporter jusqu'à plusieurs millions afin d'accomplir des fonctions de plus en plus complexes.

Un FPGA est composé de blocs logiques configurables organisés en matrice (voir Figure 9). Ces blocs logiques appelés CLB (Configurable Logic Block) permettent de réaliser des fonctions logiques simples. Ils sont interconnectés entre eux via des commutateurs programmables afin de créer des systèmes plus complexes. Des blocs IO (Entrées Sorties) permettent d'interagir avec d'autres systèmes externes au FPGA en s'adaptant à différents niveaux de tension (1.5V, 1.8V, 3.3V, etc.). Les FPGA embarquent également d'autres blocs non configurables comme des blocs mémoires, PLL (Phase-locked loop) ou DSP (Digital Signal Processor). Là où l'architecture et les fonctionnalités des ASIC (Application-Specific Integrated Circuit) sont fixés après fabrication car spécialisés pour une fonction fixe, les FPGA sont reconfigurables. Cette configuration se traduit par la modification des valeurs inscrites dans les mémoires configurant chaque CLB.

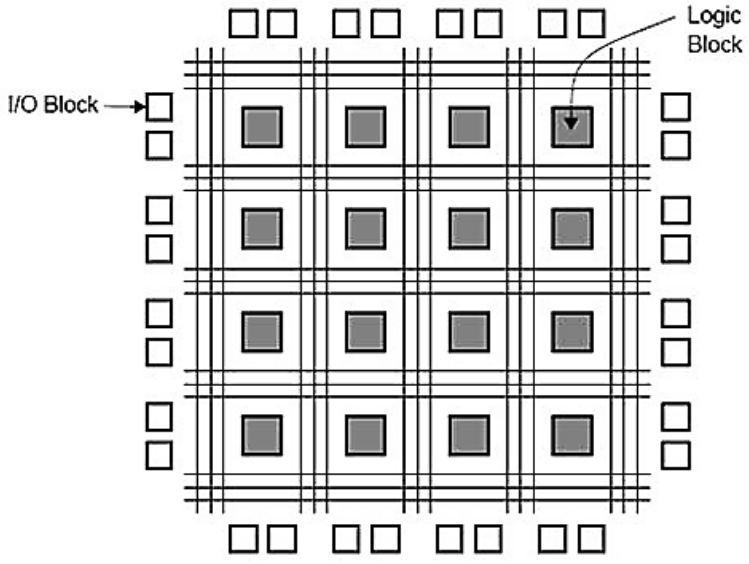


Figure 9 : Structure d'un FPGA

Le processus de conception d'un FPGA (voir Figure 10) est :

1. Spécification du besoin
2. Ecriture par le concepteur du code HDL pour remplir les fonctionnalités requises
3. Synthèse de la logique en fonction du HDL effectuée automatiquement par ordinateur
4. Implémentation automatisée de la logique dans le canevas du FPGA cible par ordinateur
5. Configuration des éléments programmables via SRAM

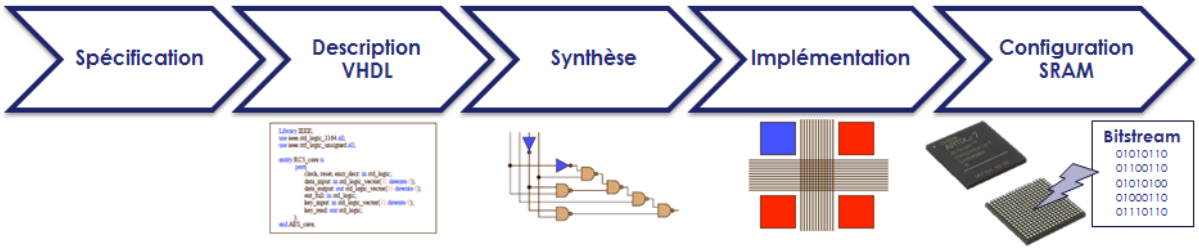


Figure 10 : Processus de conception d'un FPGA

1.2.1.1 Bloc logique configurable (CLB)

Les CLB sont usuellement constitués d'une LUT (Look-Up Table), d'un multiplexeur et d'une bascule D (appelée Flip-Flop en anglais). Cette structure est illustrée par la Figure 11.

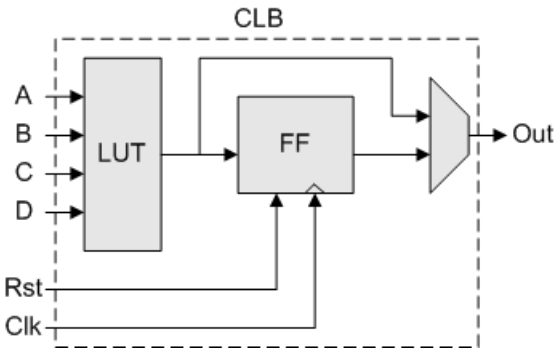


Figure 11 : Structure d'un bloc logique

La LUT est un circuit combinatoire totalement configurable. La table de vérité la régissant est stockée dans un bloc mémoire. La bascule D permet de maintenir la donnée entre deux fronts d'horloge. Le multiplexeur permet de choisir entre un CLB séquentiel ou combinatoire en contournant ou non la bascule. La configuration d'une LUT se fait par l'intermédiaire d'une mémoire CRAM (Configuration Random Access Memory). Une modification du contenu de cette CRAM implique une modification de la fonction logique équivalente comme illustré dans la Figure 12.

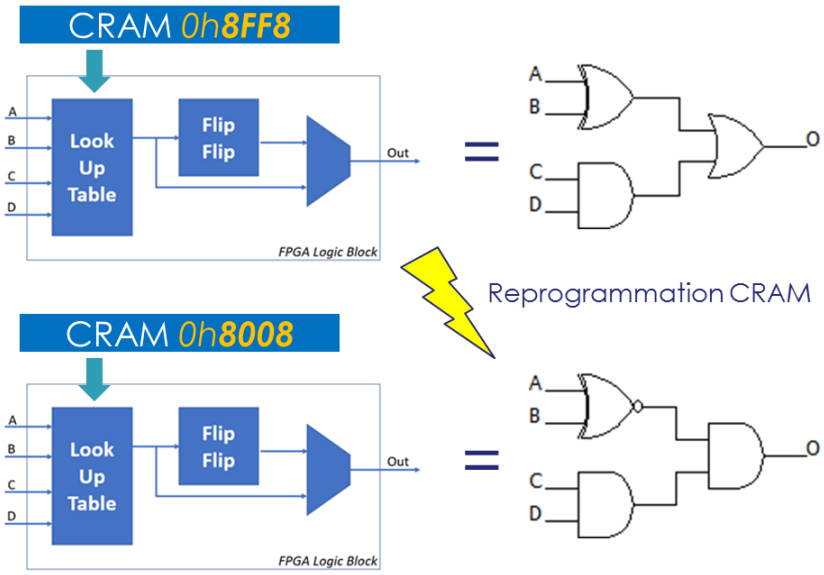


Figure 12 : Configuration d'une LUT

1.2.1.2 Mémoire de configuration

Les blocs CRAM permettent de configurer chaque CLB afin d'avoir une fonction logique donnée. Ils sont pour la plupart constitués de cellules SRAM (Static Random Access Memory). La structure d'une cellule SRAM est donnée Figure 13.

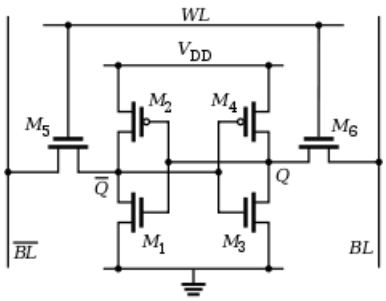


Figure 13 : Structure d'une cellule SRAM

Les SRAM sont des mémoires volatiles, aussi l'information stockée est perdue lorsque le FPGA est mis hors tension. De ce fait, les FPGA ont besoin d'une mémoire externe non volatile afin de stocker ces informations en vue de les charger à chaque démarrage. Xillinx a fait le choix de ne pas embarquer cette mémoire non volatile là où par exemple Altera (Intel aujourd'hui) en embarque le plus souvent une en interne.

1.2.1.3 Blocs I/O

Les blocs I/O permettent d'adapter en tension et courant les signaux d'entrées/sorties du FPGA. Ils permettent aussi une configuration en haute impédance (voir Figure 14). Ils

comprennent plusieurs banques d'alimentation dites d'I/O afin de pouvoir gérer plusieurs niveaux de tension.

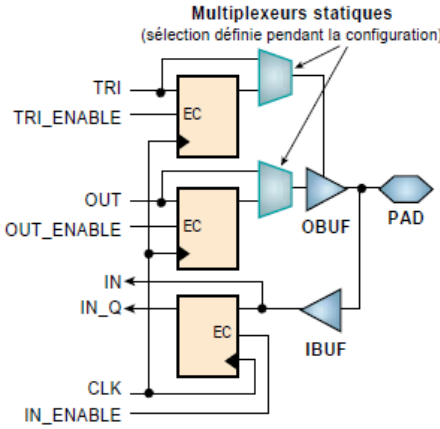


Figure 14 : Schéma d'un bloc I/O

1.2.2 Deuxième composant à l'étude : les mémoires FLASH

Les mémoires non-volatiles n'ont pas besoin de rester alimentées pour conserver l'information stockée. Ces composants sont surtout utilisés pour du stockage à moyen ou long terme de données contrairement aux mémoires volatiles de type DRAM (Dynamic Random Access Memory) ou SRAM par exemple. Les mémoires FLASH-EEPROM (Electrically-Erasable Programmable Read-Only Memory) sont aujourd'hui les mémoires non-volatiles les plus utilisées en succession aux UV-EPROM, lesquelles devaient être sorties du circuit puis placées sous UV afin d'être totalement effacées.

Le principe des mémoires FLASH-EEPROM (appelées Flash par souci de simplification) repose sur les transistors MOS à grille flottante. Les transistors MOS à grille flottante ont une tension de seuil ajustable grâce à l'ajout d'une grille flottante comme illustré sur la Figure 15. La valeur de cette tension de seuil code la donnée stockée. Ainsi, pour connaitre l'état d'une cellule ('0' ou '1'), il suffit de mesurer le courant I_{ds} pour une tension V_{gs} fixe. Cet ajustement se fait en injectant des charges négatives (des électrons) dans la grille flottante. Cette nouvelle grille a pour rôle de piéger des électrons afin d'augmenter la tension de seuil du transistor [13]. L'isolant couramment utilisé pour séparer la grille de contrôle de la grille flottante est l'oxyde nitruré (ONO); celui pour séparer la grille flottante du substrat est souvent de l'oxyde de silicium SiO_2.

Sur la Figure 16, on retrouve bien que pour une cellule écrite, correspondant à un '0' logique, la tension de seuil est plus élevée que pour une cellule effacée, correspondante à un '1' logique. Ainsi, pour une tension de lecture notée ici V_{read}, appliquée entre la grille et la

source, le courant I_{ds} sera plus important pour une cellule effacée ayant peu de charges dans sa grille flottante.

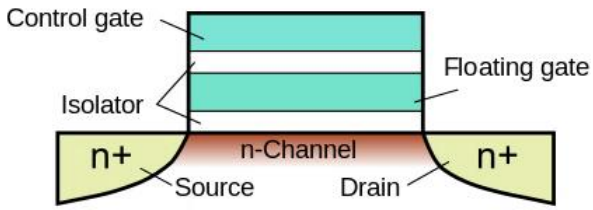


Figure 15 : Schéma d'un MOS à grille flottante

Pour les FLASH-EEPROM, afin de piéger des charges négatives dans la grille flottante et ainsi de modifier la tension de seuil, on utilise l'effet tunnel Fowler-Nordheim [14] ou le mécanisme d'injection d'électrons chauds pour que les électrons traversent la couche d'isolant entre le canal et la grille flottante.

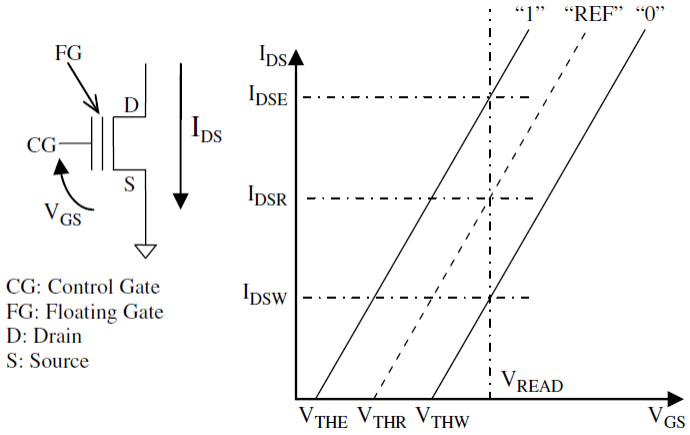


Figure 16 : Lecture d'une cellule mémoire avec grille flottante [13, Fig. 2.1]

Le choix entre les deux approches dépend du fabriquant : l'architecture NAND décrite dans la section 1.2.2.2 utilise plutôt l'effet tunnel Fowler-Nordheim (F-N) alors que l'architecture NOR utilise plutôt l'injection d'électrons chauds (CHC) [13], [15]. L'injection de charge par effet tunnel F-N permet de programmer plusieurs cellules en parallèle, alors que l'injection de charge par électron chaud nécessite plus de courant et ne permet de programmer qu'une cellule à la fois. Notons que les tensions appliquées pour le CHC sont de l'ordre de ~15V là où l'effet tunnel nécessite plutôt ~12 V. Comme on peut le voir sur la Figure 17, le choix entre les deux méthodes de programmation se fait en jouant sur les niveaux de tension aux bornes du transistor. Les tensions appliquées dépendent du nœud technologique afin d'avoir un bon compromis entre rapidité et fiabilité. Pour les deux types d'architecture, l'émission de

charge (lors de l'effacement) se fait via l'effet tunnel. A noter que les tensions de programmation sont générées en interne de la puce sur les Flash, contrairement aux anciennes générations EPROM ou EEPROM.

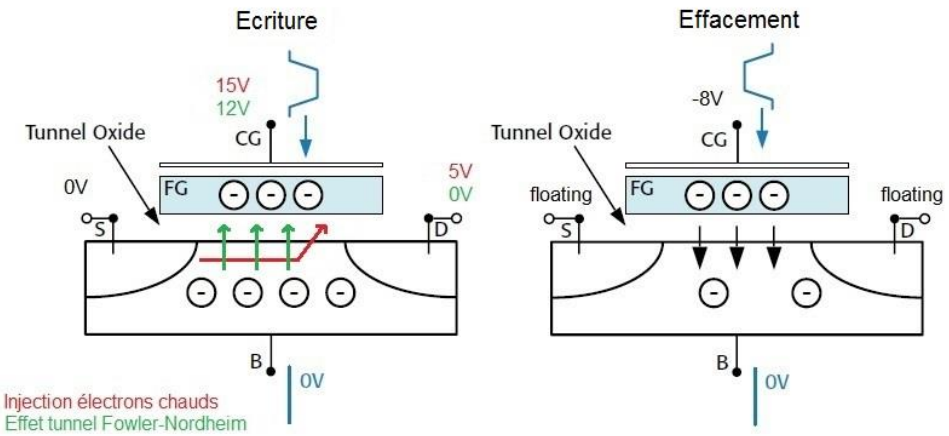


Figure 17 : Programmation ou effacement d'une cellule mémoire à grille flottante

Le mécanisme d'écriture permet exclusivement de passer un point mémoire de '1' à '0', seul l'effacement permet de repasser un point mémoire à '1'. Pour effacer une cellule on applique une tension de l'ordre de -8V sur la grille et 0V sur le substrat. Ainsi, la différence de potentiel créée entre la grille flottante et le canal permet d'éjecter des charges. Cette éjection permet d'abaisser la tension de seuil, et ainsi d'effacer la cellule.

Les matrices de cellules mémoires peuvent être organisées de plusieurs façons. Sur le marché actuel, seules deux architectures sont principalement utilisées : l'architecture NOR développée par Intel en 1988, et l'architecture NAND développée par Toshiba en 1989.

1.2.2.1 Flash NOR

Dans les mémoires Flash à architecture NOR, toutes les cellules mémoires (composées d'un transistor à grille flottante) sont accessibles individuellement. Comme schématisé par la Figure 18, toutes les cellules mémoires d'une même « word line » ont leurs grilles de contrôle interconnectées. Toutes celles sur une même « bit line » ont leurs drains interconnectés [16]. Pour accéder à un point mémoire spécifique - pour le lire ou le programmer - toutes les autres cellules de la même « bit line » sont mises dans un état bloqué. Cela permet ainsi d'accéder à une seule cellule à la fois par « bit line », on peut donc lire toutes les cellules d'une même « word line » en même temps du fait de leurs parallélisassions. La lecture d'un point mémoire se fait comme décrit dans la section précédente. Pour écrire une cellule mémoire, on applique une tension V_{write} (typiquement 15V) sur la grille ainsi qu'une tension V_{ds} d'environ 5V du point mémoire voulu. Notons que

toute écriture sur une zone non vierge doit être précédée d'un effacement du bloc. Cette architecture permet aux Flash-NOR d'avoir une vitesse de lecture élevée et un accès aléatoire sur chaque point mémoire, mais au détriment des temps d'écriture et d'effacement. A cela s'ajoute une grande fiabilité de ces mémoires. Cependant, le coût par point mémoire est élevé du fait de son manque d'intégration.

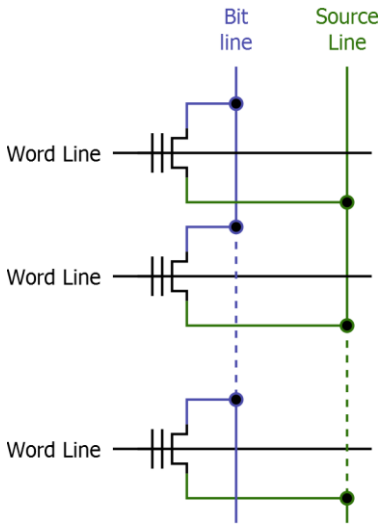


Figure 18 : Architecture mémoire Flash NOR

L'interface de ce type de mémoire permet d'exécuter du code directement depuis celle-ci, elles sont dites XIP (eXecute In Place). C'est pourquoi elles sont principalement utilisées pour le stockage d'un système exécutable, typiquement l'OS d'un téléphone portable ou d'un système embarqué.

1.2.2.2 Flash NAND

Les mémoires Flash NAND sont quant à elles beaucoup plus adaptées pour le stockage de masse avec un coût par point mémoire faible. On les retrouve aujourd'hui notamment dans les clefs USB ou les disques SSD. Leur structure Figure 19 permet d'être plus compacte que les NOR Flash. En effet les transistors à grille flottante sont cette fois-ci en série sur la « bit line ». Pour lire une cellule mémoire, on applique une tension V_{read} (typiquement 0V) sur la grille du point mémoire voulu et on applique une tension V_{gs} d'environ 5V sur les autres points mémoire de la ligne afin de les rendre passants quel que soit leur état. Ainsi on peut mesurer le courant de la cellule désirée et connaitre sa valeur [13].

Les NAND Flash possèdent des temps de lecture rapides [16]. Les temps d'effacement et écriture sont cependant importants (plusieurs centaines de millisecondes). Afin d'assurer des vitesses de transfert en phase avec le marché, les fabricants ont donc tendance à augmenter

les tailles des blocs et des pages. Le fonctionnement des Flash NAND est particulier en raison de leur forte densité et de l'optimisation interne du composant. La plus petite entité lisible pour une NAND est communément appelée une « page ». Une page correspond à 64 octets pour les anciennes générations et à 8192 octets les nouvelles mémoires NAND de grande capacité. La plus petite entité pouvant être effacée correspond à un « bloc ». Un bloc est un regroupement de 64 à 256 pages. La Figure 19 illustre cette organisation interne des mémoires NAND.

Notons que toute écriture sur une zone non vierge doit être précédée d'un effacement du bloc. De plus, les pages doivent obligatoirement être écrites par ordre croissant au sein d'un même bloc. Une page ne peut également pas être effacée de manière unitaire, tout le bloc doit être également effacé.

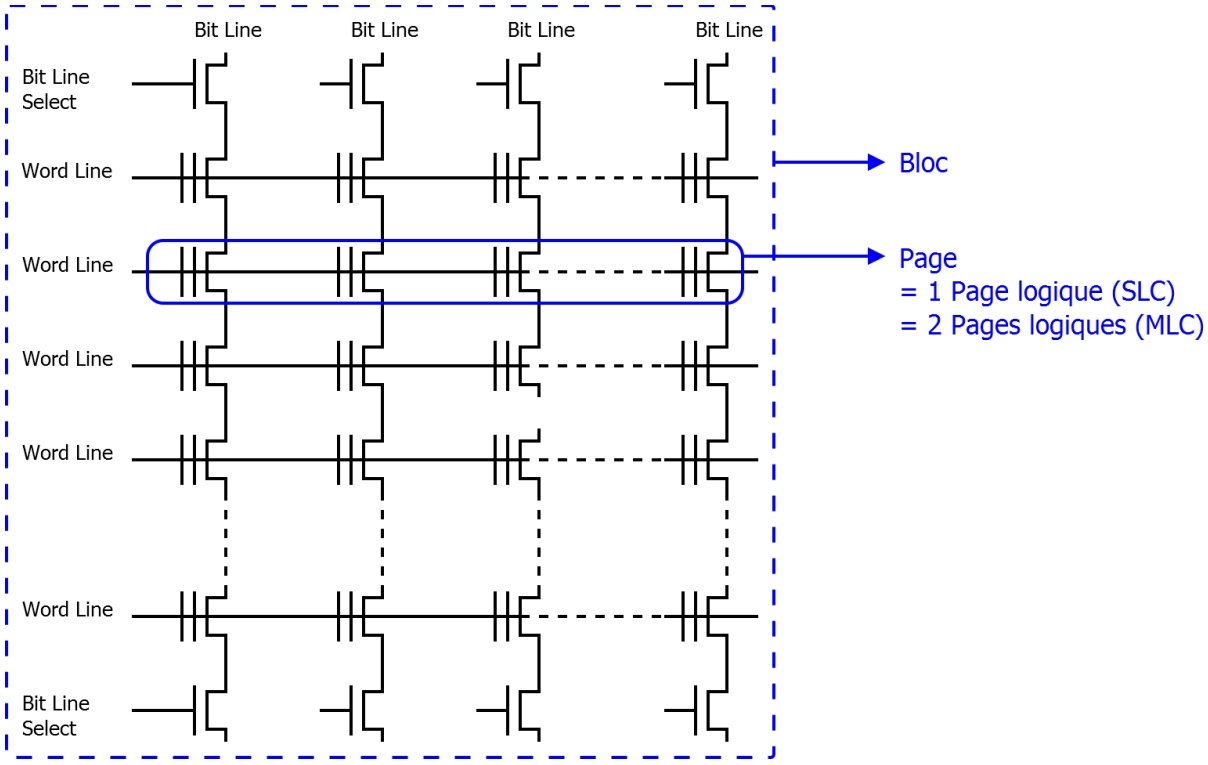


Figure 19 : Organisation logique des pages dans les mémoires NAND

Les NAND possèdent une fiabilité moyenne. De ce fait, chaque page doit comporter des bits d'ECC (Code Correcteur d'Erreurs). Avec la réduction de la taille des transistors et l'augmentation du nombre de bit par cellule, la fiabilité en rétention a diminué. Cette diminution contraint les constructeurs à utiliser des ECC corrigeant un plus grand nombre de faute comme présenté sur la Figure 20 [17], [18]. Les courbes en pointillées représentent le nombre de bits récupérables par ECC requis pour les SLC en jaune et pour les MLC en bleue.

Les courbes pleines représentent quant à elles l'endurance en PE des SLC en violet et des MLC en rouge.

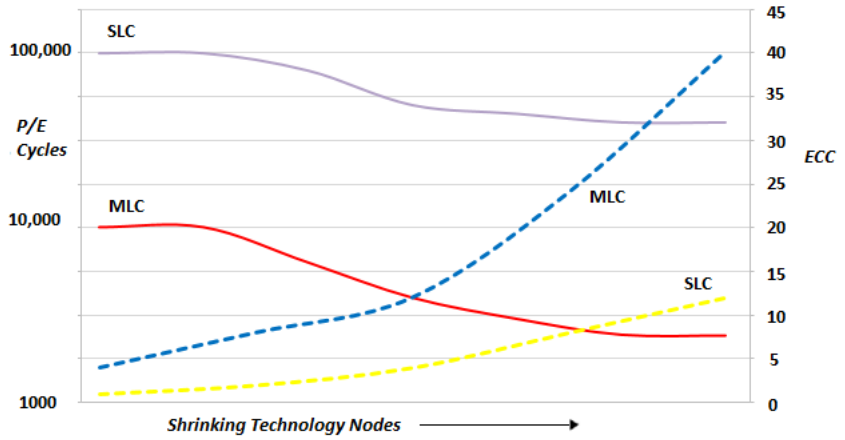


Figure 20 : Impact du nœud technologique sur les ECC requis dans les NAND [18, p. 41]

Dans l'optique de baisser les coûts des NAND pour des capacités toujours plus grandes, les mémoires considérées comme bonnes en sortie de fonderie possèdent un rendement inférieur à 100%. Ces zones initialement défaillantes sont appelées Bad Blocks. Les constructeurs garantissent aux clients un nombre minimal de blocs utilisables. Selon certains tests, le nombre de Bad Blocks initial est plus important aux adresses limites de la mémoire [19].

1.2.2.2.1 SLC et MLC

Jusque dans les années 2000, toutes les mémoires EEPROM étaient de type SLC (Single-Level Cell). Les SLC possèdent un seul bit par transistor à grille flottante. La marge de bruit est donc assez importante (voir Figure 21). En effet, lorsque la grille flottante perd des charges négatives, la tension de seul du transistor se rapproche de celle du transistor effacé. Une cellule effacée est notée « ERS » et une programmée est notée « P ».

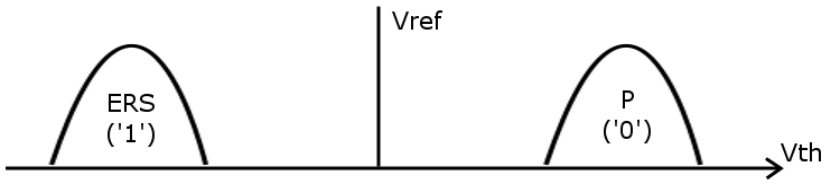


Figure 21 : Distribution théorique des tensions de seuil des cellules mémoires SLC

Les MLC (Multiple Levels Cell) contiennent deux bits par cellule mémoire. Pour ce faire, le nombre de niveaux de tension de seuil est passé de deux (SLC) à quatre (MLC). La marge entre deux niveaux logiques est ainsi réduite (voir Figure 22).

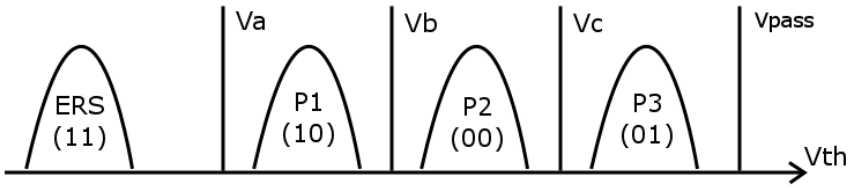


Figure 22 : Distribution théorique des tensions de seuil des cellules mémoires MLC

Cette réduction de marge a fait chuter le nombre de cycles d'écriture/effacement garanti par le fabriquant de 100 000 pour les SLC à 3000 pour les MLC [18], [20] (voir Tableau 2). Le temps de rétention des données est également impacté négativement (voir Figure 23).

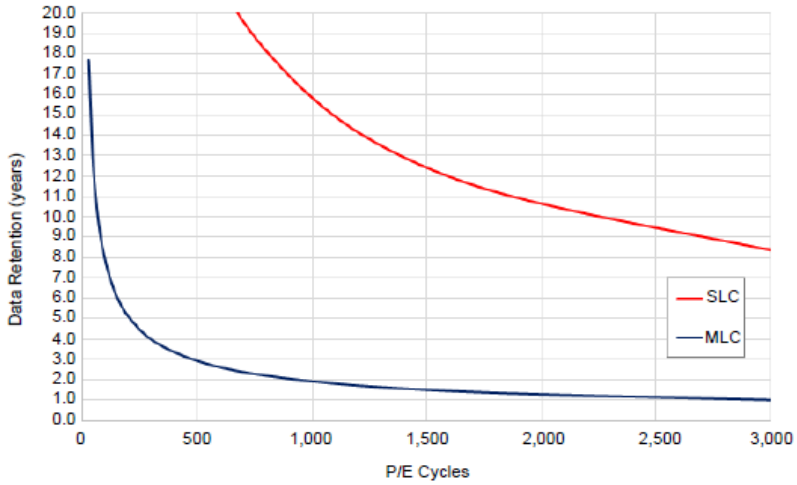


Figure 23 : Comparaison du temps de rétention entre les SLC et les MLC [20, p. 5]

Type de mémoire NAND	Nombre de niveau de tension	Nombre moyen de cycles garanti par le constructeur
SLC	2	100 000
MLC	4	1000 à 3000 (haute capacité, SSD modernes) 5000 à 10 000 (capacité moyenne)
TLC	8	1000 en 2D, 1000 à 3000 en 3D
QLC	16	100 à 1000 (mémoire uniquement en 3D)

Tableau 2 : Evolution de nombre de PE maximal des Flash NAND avec les MLC

Toshiba a annoncé en 2009 la conception de nouvelles mémoires Flash NAND avec trois bits par cellule (TLC) – avec huit niveaux de tension de seuil - et même d'autres avec quatre bits par cellule (QLC) [21]. Sandisk a produit des cartes mémoires Flash en QLC en 2009, et

Samsung a lancé la commercialisation de SSD TLC en 2010. Les problèmes des MLC sont amplifiés pour ces nouvelles mémoires mais leur coût par bit en est d'autant réduit. Il est important de mentionner que l'introduction des TLC s'est notamment accompagnée d'un changement d'architecture passant de puce 2D à des puces 3D comme mentionné dans le Tableau 2. Cette thèse traitera uniquement des architectures 2D.

1.2.2.2.2 Read Retry

Au vue des dérives rapides des tensions de seuil des nouvelles NAND au cours de leur utilisation, les constructeurs ont dû trouver une solution pour augmenter la fiabilité de leurs produits. Compte tenu de la marge de bruit de plus en plus faible entre deux niveaux, les dérives dues aux P/E ou à la perte de charges durant la rétention sont visibles plus rapidement, se traduisant ainsi par une perte de donnée (voir Figure 24).

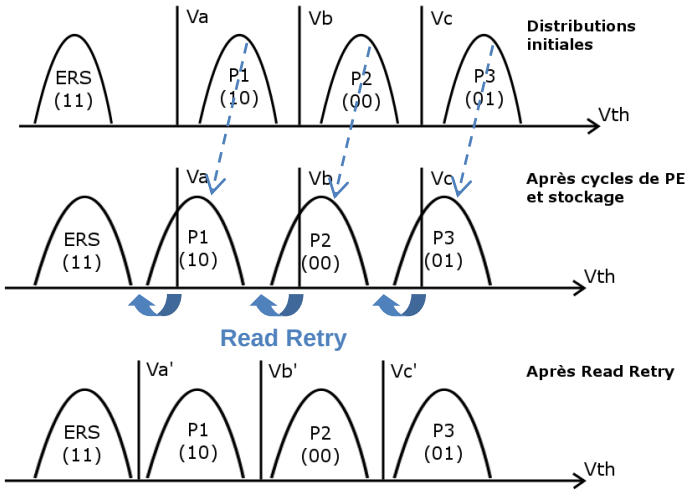


Figure 24 : Dérive de la distribution des tensions de seuil en fonction du cyclage et du temps, et utilisation du Read-Retry

Une des solutions mise en œuvre, nommée « Read Retry », permet de faire varier les niveaux des tensions de référence utilisés lors de la lecture des cellules. Ainsi, lorsqu'une donnée lue est détectée incohérente par l'ECC, le système la relit en modifiant légèrement la tension de référence pour compenser les dérives [22], [23]. L'algorithme à utiliser pour lire une page avec le Read Retry est donné sur la Figure 25.

Cette nouvelle option permet certes de pouvoir augmenter la fiabilité des mémoires, mais ce au détriment du temps de lecture (nombreuses relectures nécessaires jusqu'à atteindre le bon niveau de référence) et en nécessitant un circuit de pilotage de la mémoire bien plus conséquent.

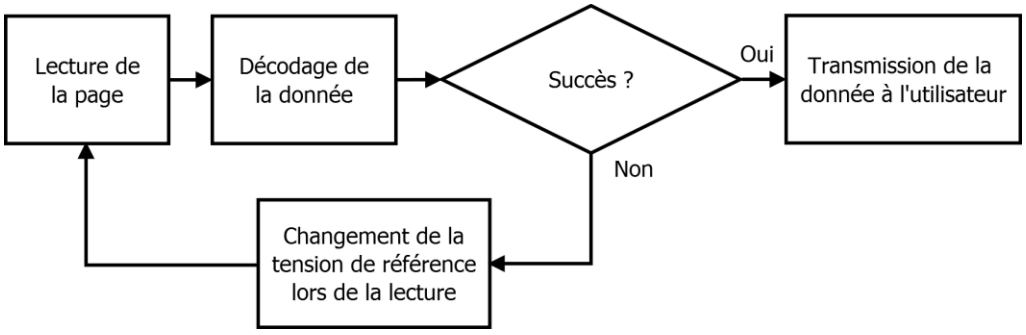


Figure 25 : Principe du Read Retry

1.3 Mécanismes de dégradation

Les technologies qui seront étudiées dans ce mémoire ont été décrites dans la section 1.2, cette présente section d'état de l'art s'intéressera aux mécanismes de dégradation. Dans le monde des semi-conducteurs à applications numériques, on peut considérer 4 mécanismes de dégradation principaux : le Hot Carrier Injection (HCI), le Time Dependent Dielectric Breakdown (TDDB), l'électromigration (EM) et le Bias Temperature Instability (BTI). Chacun de ces mécanismes sera décrit dans cette partie. Le cas particulier des mémoires sera également abordé, tant sur le plan rétention de donnée que sur le plan de l'endurance.

1.3.1 Hot Carrier Injection

L'injection de porteurs chauds (HCI) est l'un des mécanismes de dégradation des transistors les plus étudiés. Il se produit lorsqu'une forte tension V_{gs} est imposée entre la grille et la source et qu'une différence de potentiel existe entre le drain et la source. Les électrons dans le canal acquièrent une énergie cinétique suffisante via le champ électrique drain source. Ils entrent alors en collision avec des ions près de l'interface au niveau du drain, créant ainsi des défauts dans l'oxyde [24], [25]. Les électrons ont une énergie cinétique suffisante qui, combinée avec le champ électrique crée par la grille, passent la barrière de potentiel du diélectrique de grille. Des charges sont donc injectées dans l'oxyde comme illustré sur la Figure 26.

Ce phénomène se manifeste sous forme d'une lente dégradation de la tension de seuil du transistor (augmentation de la tension de seuil avec le nombre de charges injectées dans l'oxyde), d'une baisse de la mobilité μ des électrons dans le canal et d'une augmentation des courants de fuite dans le substrat [6], [26]. Cette dégradation se mesure par une diminution du courant I_{ds}. Pour les circuits numériques, cela se traduit par une diminution de la fréquence maximale de fonctionnement du système.

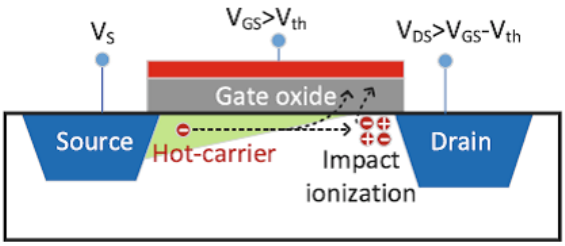


Figure 26 : Principe du HCI [27, p. 292]

La dégradation due au HCI est amplifiée par une augmentation de la tension d'alimentation (V_{GS} et V_{DS}). Elle est plus importante à basse température, contrairement à la plupart des autres mécanismes de dégradation [26], [28], [29]. On peut expliquer ceci par le fait que, plus la température est basse, moins la matière est agitée, et donc le risque de collision des électrons accélérés dans le canal avec un ion est diminué. Il est donc plus probable qu'ils atteignent une énergie cinétique suffisante pour créer une avalanche. Pour les circuits numériques, les injections de charge se produisent à chaque circulation de courant dans le canal, soit à chaque commutation du MOS. La fréquence de fonctionnement est donc également un facteur clef de ce mécanisme. Plus la fréquence est élevée, plus les dégradations le sont également [24], [30], [31].

Du point de vue des circuits numériques dans la mesure où le HCI se produit à chaque passage de courant, on peut exprimer les dégradations en fonction du nombre de commutations [32], [33] :

$$A \propto (f.t)^m \tag{44}$$

Où :

- f est la fréquence en Hz
- t est le temps de stress

Selon Takeda, qui est l'un des premiers contributeurs à l'étude du HCI avec Suzuki, il y a trois types de mécanismes sous-jacent principaux [24], [28], [29] :

- Channel hot electron (CHE) injection
- Drain avalanche hot carrier (DAHC) injection
- Secondary generated hot electron (SGHE) injection

Le CHE est composé « d'électrons chanceux » qui obtiennent suffisamment d'énergie pour passer la barrière de potentiel du Si-SiO_2. Cette acquisition d'énergie se fait en l'absence de collision lors de la traversée du canal. Le CHE est très présent à basse température. Le DAHC est causé par ionisation via impact près de l'interface lors de la commutation. La pire

condition pour cette injection est la température ambiante car à la fois les électrons et les trous sont injectés dans l'oxyde de grille. Le SGHE provient des porteurs minoritaires issus de secondes ionisations par impact.

De plus, Takeda a proposé un modèle empirique des dégradations dues au HCI de la forme [6], [34] :

$$\Delta V = A.t^n.exp\left(\frac{-\alpha}{V_D}\right) \quad (45)$$

Où A, α et n sont des paramètres empiriques estimés pour chaque technologie.

Hu décrit ce mécanisme par le modèle de l'électron chanceux ("Lucky Electron Model") [35]–[37]. Un électron chaud du canal peut être injecté dans l'oxyde de grille s'il acquiert suffisamment d'énergie dans le canal (via le champ électrique Drain-Source), ne la perd pas lors d'une collision élastique et qu'il reparte bien en direction de l'oxyde. Ce modèle a été amélioré en ajoutant - à l'injection par porteur chaud - la création de défaut à l'interface suite à la rupture de liaison Si-H. Ce deuxième mécanisme analogue au NBTI n'est valable que lorsque la tension V_G est importante [37], [36], [38]. Selon Hu, l'étude du courant de fuite dans le substrat I_{sub} est un bon indicateur de l'évolution du HCI [39].

D'après Hu, le temps de vie moyen des MOS sous HCI peut être donné par [37], [40], [41]:

$$MTTF = b.{I_{sub}}^{-N}.e^{\frac{E_a}{k_B.T}} \quad (46)$$

Où :

- I_{sub} est le courant de substrat en Ampères
- *b* est un paramètre d'ajustement dépendant des dimensions, du dopage du transistor
- *N* est un paramètre dépendant de la technologie du transistor.
- *T* est la température absolue en Kelvin
- *Ea* est l'énergie d'activation apparente de la loi d'Arrhenius

La littérature semble s'accorder sur le fait que la dégradation due au HCI est régie par une loi puissance [25], [40], [42], [43] de la forme :

$$\Delta P = a.{V_{GS}}^m.e^{\frac{-E_a}{k_B.T}}.t^n \quad (47)$$

Où :

- *P* est un paramètre tel que la tension de seuil *Vt*, le courant *Ids* ou la transconductance *gm*
- *a* est un paramètre d'ajustement dépendant des dimensions du transistor
- *n* est un paramètre dépendant de la technologie du transistor
- *m* défini la loi puissance régissant l'accélération en tension

Les valeurs constatées de *n* sont autour de 0,5 pour des MOS PolySi/SiO_2 [6], [32], [44]. Les publications plus récentes sur les MOS avec diélectriques High- κ donnent des valeurs dans le même ordre de grandeur [43], [45]. Une étude du HCI décorrélé du BTI sur du HKMG (High- κ Metal Gate) donne un exposant *n* d'environs 0,35 [46]. L'énergie d'activation apparente pour le HCI est de -0,2 à 0,4 eV. Cette faible valeur pouvant même être négative confirme le fait que la température n'est pas un accélérateur du HCI, bien au contraire [5, p. 210], [40], [47], [48, p. 126]. Le paramètre de la loi d'accélération électrique *m* varie de 8 à 13 [44], [47].

Avant l'introduction du HKMG, les NMOS était moins sensibles au HCI que les PMOS en raison d'une barrière de potentiel plus haute du SiO_2 [49].

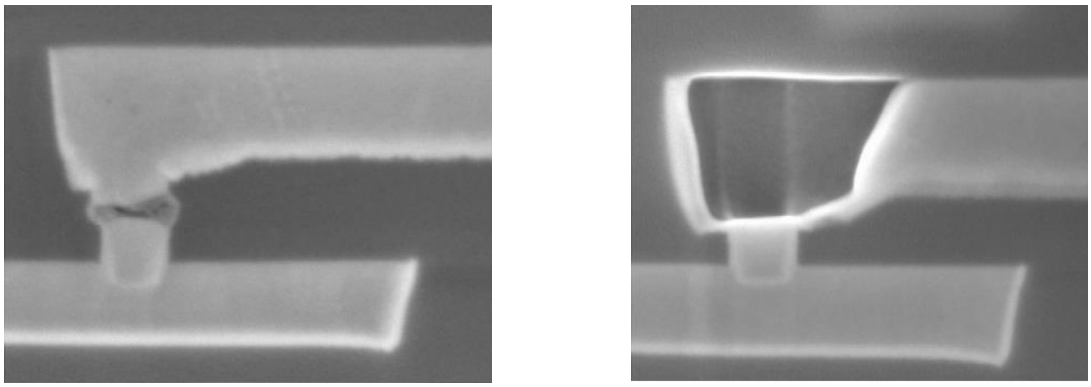
Sur les dernières technologies où les gammes de tension sont autour du Volt, des phénomènes de guérison peuvent être générés comme présenté par Bravaix. Ces phases de guérison nécessitent cependant de haute température ainsi qu'une polarisation particulière [50], [51].

Toutes ces études ont été faites pour la plupart au niveau wafer. Or dans notre cas, nous n'aurons pas accès aux caractéristiques unitaires des transistors. Les paramètres des modèles tels que V_D ou V_{GS} seront simplifié à $V = V_{dd}$. Cette équivalence se justifie dans le cadre des portes logiques CMOS où V_G égal V_{dd} ou V_{ss}.

1.3.2 Electromigration

L'électromigration (EM) est une usure au niveau des pistes d'interconnexion (BEOL, Back End Of Line) au sein d'un circuit intégré. Sous l'effet du courant et du champ électrique, une diffusion de la métallisation des pistes s'amorce. Cela conduit à un déplacement de matière dans la piste créant des zones de vide [5, p. 165‑171] comme observé sur la Figure 27. Ce mécanisme de dégradation est non négligeable si les conditions d'utilisation du composant sont au-delà de 125°C [52]. Il a entrainé une limitation de la densité de courant à 2×10^5 A/cm² dans les règles de conception jusqu'à l'introduction du cuivre.

Les conséquences de l'EM sont une augmentation de la résistivité de la piste allant jusqu'au circuit ouvert, ou un court-circuit entre pistes proches (accumulation de matières du côté du potentiel positif). Cette détérioration, bien que progressive, peut s'accélérer rapidement du fait de la concentration des lignes de courant à proximité d'un vide. Ceci provoque un saut important de résistance.



(a) Défaut rapide au niveau d'un via (b) Défaut tardif sur une piste

Figure 27 : Coupes transversales de phénomènes d'électromigration [53, Fig. 5]

Ce mécanisme d'usure est modélisé par l'équation de Black [54] ci-dessous :

$$MTTF = \frac{A}{J^N} . e^{\frac{E_a}{k.T}} \quad (48)$$

Où :

- *A* est un paramètre empirique
- *J* est la densité de courant en A/cm²
- *N* est un paramètre extrait de la modélisation

Le paramètre d'échelle *N* est usuellement autour de 2. L'énergie d'activation *Ea* est selon la littérature comprise entre 0.85 eV et 0.9 eV [55], [56]. Elle dépend principalement de l'épaisseur des pistes et de leur matériau (aluminium, cuivre, etc.). Cependant, les intermétalliques en cuivre semblent bien plus fiables que ceux en aluminium-cuivre vis-à-vis de ce phénomène [52].

1.3.3 Bias Temperature Instability

Le Bias Temperature Instability (BTI) est le principal mécanisme de vieillissement des transistors MOS en VLSI pour les nouvelles générations. Il est dû à la tension appliquée sur la grille et à la température du transistor. Ce mécanisme est un phénomène de diffusion. Ainsi il est fortement accéléré par une augmentation de la température. Une élévation de la tension de grille en valeur absolue ou une diminution de l'épaisseur d'oxyde de grille sont également des facteurs aggravant de ce mécanisme. Les dégradations interviennent lorsque le transistor est passant, aussi le rapport cyclique est déterminant dans le vieillissement [46]. Plus le transistor est passant, plus ses dégradations sont importantes. La fréquence de commutation n'a quant à elle que peu d'importance sur le BTI jusqu'à 100 Hz [57].

Il y a deux types de BTI : le NBTI pour les PMOS et le PBTI pour les NMOS. Ces deux différents mécanismes sont détaillés dans les deux sous-sections suivantes.

1.3.3.1 Negative Bias Temperature Instability

Le Negative Bias Temperature Instability (NBTI) concerne les transistors PMOS. Lorsque la tension V_{gs} est négative (état passant), les liaisons Si-H à l'interface se rompent suite au stress électrique vertical créant ainsi une diffusion des H^+ , H_2 dans l'isolant [58], [59] comme illustré sur la Figure 28. Cette migration crée des pièges à électron dans la zone d'inversion, diminuant ainsi la tension de seuil du PMOS. Cela conduit par conséquent à une diminution du courant I_{ds} [41].

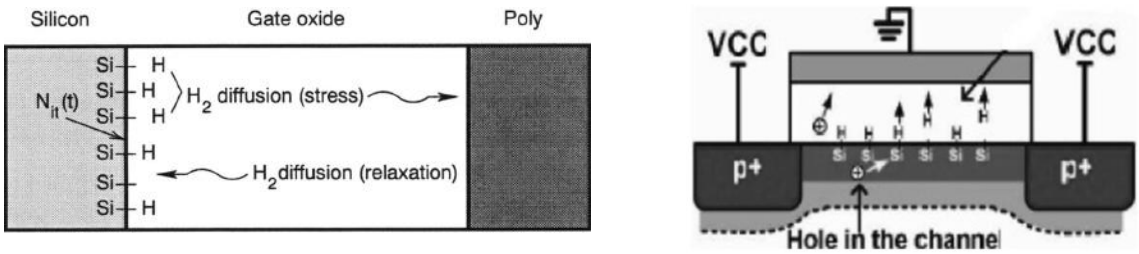


Figure 28 : Diffusion lors du NBTI [60, Fig. 2], [61, p. 609]

La dérive de la tension de seuil semble suivre une loi puissance telle que [30], [40], [62], [63]:

$$\Delta Vt = a.{V_{GS}}^{\alpha}.e^{\frac{-E_a}{k.T}}.t^n \tag{49}$$

Où :

- *Vt* est la tension de seuil du MOS, on peut aussi utiliser le courant *Ids* ou la transconductance du transistor *gm*
- *a* est un paramètre d'ajustement dépendant des dimensions du transistor
- *n* est un paramètre dépendant de la technologie du transistor
- *α* définit la loi puissance régissant l'accélération en tension

La valeur de *n* varie entre 0,22 et 0,29 en accord avec les différentes valeurs empiriques trouvées dans la littérature pour des oxydes de grille en SiO_2 [30], [47], [64]–[66]. Pour des oxydes en High- κ, la valeur est plutôt aux alentours de 0,12 à 0,17 [67]–[69]. Cet exposant ne semble pas varier en stress DC ou en stress AC [68]. L'énergie d'activation apparente définissant l'accélération en température de la diffusion varie de 0,12 à 0,19 eV [47], [68]. L'exposant *α* définissant la loi puissance de l'accélération électrique est de l'ordre de 4 à 5 [40], [47].

Cependant, en fonction de l'épaisseur de l'oxyde de grille, la loi d'accélération électrique peut plutôt suivre une loi exponentielle telle que [40], [63], [68], [70] :

$$\Delta Vt = a.e^{\gamma.V_{GS}}.e^{\frac{-E_a}{k.T}}.t^n \quad (50)$$

avec γ définissant la loi exponentielle régissant l'accélération en tension, avec des valeurs pour le HKMG autour de 0,14 V^{-1} [68].

Une augmentation de la fréquence de stress pour un même rapport cyclique diminue les dégradations selon une publication de 2005 [60].

La distribution statistique correspondant le mieux à ce mécanisme de dégradation est une loi log-normale vis-à-vis de la dispersion des instants de défaillances en fonction d'un critère de dérive donné [64], [71].

1.3.3.2 Positive Bias Temperature Instability

Le Positive Bias Temperature Instability (PBTI) concerne quant à lui les transistors NMOS. Il se produit lorsque la tension Vgs est positive (état passant). Le mécanisme de dégradation PBTI est analogue au NBTI. Cependant, le PBTI a toujours été considéré comme négligeable par rapport au NBTI pour les MOS avec un oxyde de grille en SiO_2. Cependant avec les nouveaux oxydes High- κ utilisés en dessous du nœud technologique des 45 nm, le PBTI devient de moins en moins négligeable [67], [72], [73].
Les équations du PBTI sont à l'état de connaissance actuel les mêmes que le NBTI. Les paramètres estimés sont toutefois légèrement différents. L'énergie d'activation de la loi d'Arrhenius est plutôt de l'ordre de 0,11 à 0,15 eV [47]. L'exposant temporel *n* reste dans le même ordre de grandeur (~ 0,16) que pour le NBTI [68], [69], [74]. L'exposant α définissant la loi d'accélération électrique semble être de l'ordre de 5,5 pour des oxydes High- κ [67], et entre 8 et 9 pour des oxydes SiO_2 [47]. Sous la forme de loi exponentielle, pour du High- κ, la valeur du paramètre γ serait de 0,17 V^{-1} [68].

Selon une publication de 2018 [74] sur une technologie en High- κ, les dégradations dues au PBTI deviendraient dominantes sur celle générées par du NBTI après 10 ans. Ainsi le NBTI reste dominant sur des durées de vie usuelles (~ 8 à 10 ans) mais il est à prendre en considération au-delà.

Comme pour le NBTI, le niveau de dégradation dépend du rapport cyclique. Concernant la fréquence de stress, une séparation du PBTI en trois sous mécanismes (deux dus à des

défauts préexistants dans l'oxyde et un dû à des défauts générés) permet de modéliser les dégradations via des stress DC ou même AC [74].

1.3.3.3 Cas des portes logiques CMOS

Dans un circuit numérique, les transistors sont utilisés en configuration CMOS afin de former des portes logiques. Ces portes logiques comportent à la fois des NMOS et des PMOS.
Dans le cas de l'inverseur (voir Figure 29), le cas le plus dégradant est le cas où le rapport cyclique du signal d'entrée tend vers 0, avec le cas extrême où l'entrée reste constante à 0 (il n'y a pas de récupération dans ce cas et la dégradation peut être 50% plus importante comparé à un rapport cyclique de 10%). Ce pire cas ne prend en compte que les dégradations du PMOS, le PBTI étant considéré dans ce cas comme négligeable. Une dégradation asymétriques des PMOS et des NMOS amènera par conséquence une dissymétrie entre le temps de montée et le temps de descente de la porte logique.

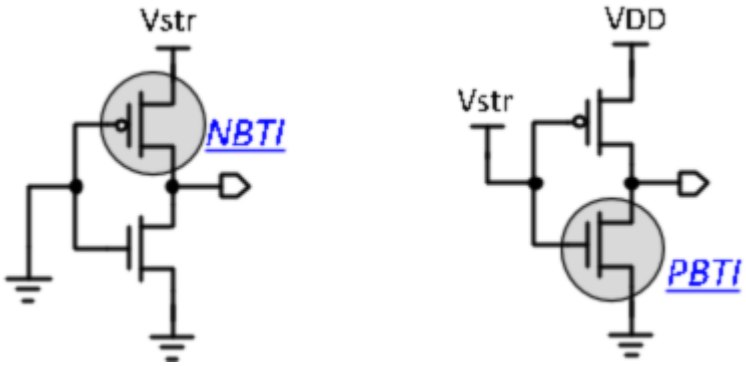


Figure 29 : Cas de l'inverseur pour le BTI

On pourra donc modifier les équations précédentes afin de prendre en compte le rapport cyclique dans le cas où l'on ne considère que le NBTI de la manière suivante :

$$\Delta Vt = a.{V_{GS}}^{\alpha}.e^{\frac{-E_a}{k.T}}.((1-r).t)^n \quad (51)$$

Où *r* est le rapport cyclique du signal d'entrée.

1.3.3.4 Guérison partielle

La grande particularité du BTI est sa phase de guérison partielle [75]. Lorsque le transistor est bloqué ou que le circuit n'est pas alimenté il se produit une reformation partielle des liaisons Si-H à l'interface [76], [77]. Notons qu'après une phase de stress, il est impératif de mesurer les dégradations le plus rapidement possible afin de limiter cette phase de récupération. Avec l'apparition de nouveaux instruments de mesure plus performants, il est aujourd'hui possible d'effectuer les mesures en moins de 10 µs, ce qui permet de limiter la

récupération [68], [69], [74], [78], [79, p. 78]. Notons également qu'en stress AC, une phase de récupération intervient à chaque cycle [66, p. 796].

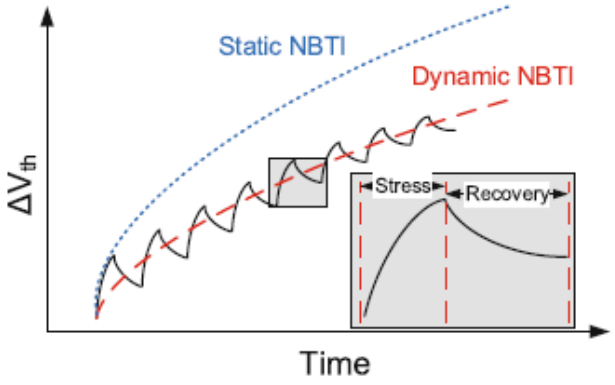


Figure 30 : Dégradation due au NBTI en AC ou en DC [27, p. 285]

Comme le montre la Figure 30, dès que le transistor est au repos, la tension de seuil se rétablit partiellement mais la dégradation globale est irréversible. D'après une étude complète menée sur ce phénomène de récupération, la guérison serait meilleure à haute température pendant 10 secondes, puis devient indépendante de la température [80].

Un modèle d'estimation de l'importance de la phase de « recovery » a été proposé [57], [77] :

$$RecoveryFraction = \frac{\Delta Vt_{recovery}}{\Delta Vt_{stress}} = \left[1 + A.\left(\frac{t_{recovery}.e^{\frac{E_a}{k}.\left(\frac{1}{T_{stress}}-\frac{1}{T_{recovery}}\right)}}{t_{stress}}\right)^{\beta}\right]^{-1} \quad (52)$$

Où :

- ΔVt_{stress} est la dégradation de la tension de seuil à la fin de la période de stress
- $\Delta Vt_{recovery}$ est la dégradation de la tension de seuil à la fin de la période de « recovery »
- A est un paramètre dépendant de la tension durant la phase de « recovery »
- T_{stress} est la température durant la phase de stress en Kelvin
- $T_{recovery}$ est la température durant la phase de « recovery » en Kelvin
- t_{stress} est la durée du stress
- $t_{recovery}$ est la durée du « recovery »
- β est un paramètre dépendant de la technologie du transistor, autour de 1.2 à 1.6

1.3.4 Time Dependent Dielectric Breakdown

Le Time Dependent Dielectric Breakdown (TDDB) est une accumulation de charges piégées dans l'oxyde qui finissent par créer un chemin conducteur entre la grille et le substrat. On appelle communément cela un claquage d'oxyde. L'énergie nécessaire pour passer la barrière

de potentiel de l'oxyde devient de plus en plus petite jusqu'à devenir insuffisante pour garantir l'isolation comme on peut le voir sur la Figure 31. Le TDDB se caractérise par un saut de courant de fuite au niveau de la grille et une perte de contrôle du courant I_{ds}.

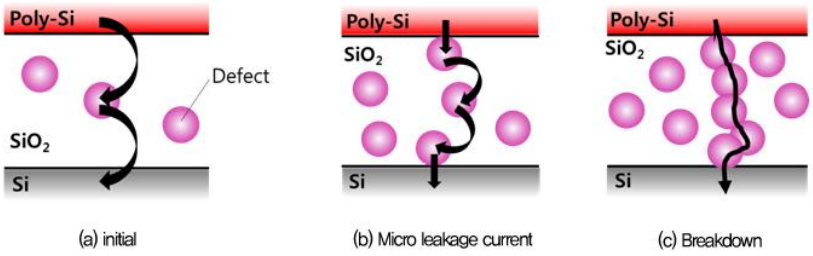


Figure 31 : Evolution de la percolation du TDDB

L'origine de ces défauts est multiple. Certains sont issus d'imperfections lors du processus de fabrication du wafer. Les autres proviennent d'une percolation dans l'oxyde (par exemple injection de porteurs chaud [29] ou bien de mécanismes type BTI). Les facteurs d'accélération majeurs sont une diminution de l'épaisseur de l'oxyde de grille, une augmentation de la température et une augmentation de la tension de grille [81], [82].

Cette dégradation est stochastique, comme l'EM, il faut par conséquent un grand nombre d'échantillon lors des tests afin d'en extraire une loi précise [52]. Le modèle du TDDB ne fait pas encore consensus. Même si le fait que la fonction de défaillance empirique de ce phénomène suit une loi de Weibull est unanime [52], [76], [82]–[84], plusieurs modèles se sont succédés dans la littérature sur le paramètre d'échelle $\eta = t_{63\%}$ [41].

$$F(t_{BD}) = 1 - e^{-\left(\frac{t_{BD}}{t_{63\%}}\right)^{\beta}} \tag{53}$$

Modèle 1/E :

$$t_{63\%} = A.e^{G.\frac{t_{ox}}{V_{GS}}}.e^{\frac{E_a}{k.T}} \tag{54}$$

Modèle E :

$$t_{63\%} = A.e^{-\gamma.\frac{V_{GS}}{t_{ox}}}.e^{\frac{E_a}{k.T}} \tag{55}$$

Modèle Root E :

$$t_{63\%} = A.e^{-\gamma.\sqrt{\frac{V_{GS}}{t_{ox}}}}.e^{\frac{E_a}{k.T}} \tag{56}$$

Modèle Power Law E :

$$t_{63\%} = A.\left(\frac{V_{GS}}{t_{ox}}\right)^{-n}.e^{\frac{E_a}{k.T}} \quad (57)$$

Où :

- β est le paramètre de forme de la loi de Weibull
- A est un paramètre d'ajustement dépendant des dimensions du transistor
- G est le facteur d'accélération électrique du modèle 1/E
- t_{ox} est l'épaisseur de l'oxyde de grille
- γ est le facteur d'accélération électrique de la loi exponentielle
- V_{gs} est la tension de grille
- n défini la loi puissance régissant l'accélération en tension
- t_{BD} est l'instant de claquage d'oxyde

Le modèle 1/E est un des premiers modèles utilisés. Il a vite été remplacé par le modèle E qui présentait des résultats plus proches de la réalité pour des épaisseurs d'oxyde inférieures à 10 nm [85]–[87]. Aujourd'hui, le modèle le plus largement utilisé est le modèle « Root E » [88] même si le modèle « Power Law E » pourrait également décrire le TDDB pour les technologies sous les 15 nm (avec tox < 2 nm) [89], [90].
Les valeurs lues de γ dans le cas du modèle « Root E » sont de l'ordre de 45 à $75(nm/V)^{0.5}$ [88], [89]. L'énergie d'activation de la loi d'Arrhenius est aux alentours de 0,85 eV [26, p. 38], [70], [86]. Concernant le modèle émergeant « Power Law E », le paramètre n semble avoir comme ordre de grandeur 20 [89].

Notons que le TDDB peut aussi intervenir dans les diélectriques intermétalliques.

1.3.5 Mécanismes de défaillance des mémoires Flash

Les défaillances des mémoires non volatiles sont de deux types : des problèmes de rétention, ou bien des problèmes d'endurance, c.-à-d. une usure liée à un trop grand nombre d'écriture-effacement. En rétention, une défaillance est une perte de la donnée écrite initialement. Un bit est dit défaillant si la valeur lue est différente de sa dernière valeur écrite (programmation d'un '0' ou effacement donnant un '1'). En endurance, c'est une impossibilité d'écrire correctement la donnée désirée.

Lorsqu'une valeur est écrite dans un point mémoire, sa tension de seuil est modifiée de manière à être comprise entre deux bornes de référence comme illustré sur la Figure 22. Plus le nombre de bit contenu par chaque transistor à FG (« Floating Gate », ou grille flottante) augmente (SLC → MLC → TLC → QLC), plus les intervalles de tension de seuil pour chaque état sont de plus en plus étroits.

L’accumulation des cycles d’écriture-effacement (endurance) provoque une augmentation du nombre de défauts dans l’oxyde. Cette usure entraine une baisse de la fiabilité en rétention, voir même une impossibilité d’utiliser le point mémoire. Au fil du temps et de la température, des pertes de charges se produisent au sein de la grille flottante, ce qui donne lieu à une diminution de la tension de seuil du point mémoire. Au contraire, lors de diverses opérations au sein de la mémoire, de légères injections de charges accidentelles peuvent l’augmenter. Lorsque cette dérive est assez importante pour outrepasser la marge de bruit entre deux états voisins, le point mémoire change de valeur et la donnée est perdue. Ces différents mécanismes de perte de donnée sont appelés Read Disturb, Pass Disturb et Program Disturb. Ils seront détaillés dans la suite de la section.

1.3.5.1 Usure des mémoires Flash

On parle de défaillance en endurance lorsque des données écrites sur des mémoires ayant subi un grand nombre de cycles de Programmation/Effacement (P/E) sont illisibles après une relecture immédiate. Le point mémoire est ainsi inutilisable du fait de son usure trop importante. Sinon, lorsque les dégradations ne sont pas trop importantes (peu de P/E) ce mécanisme impacte sur la fiabilité en rétention. Cette sous-section liste les mécanismes entrainant une dégradation de l’oxyde des cellules mémoires et ceux jusqu’à la casse du point mémoire.

1.3.5.1.1 Mécanismes de dégradation pour la rétention

La principale cause d’usure des transistors à grille flottante est l’accumulation de défauts dans l’oxyde. Plusieurs mécanismes distincts sont responsables de cette dégradation. Ils apparaissent à différents moments lors de la rétention, de l’écriture ou de l’effacement. Ils ont également une participation plus ou moins importante selon les conditions d’utilisation. La Figure 32 résume les mécanismes impliqués dans chacune des phases de dégradation des mémoires.

Detrapping

Au cours de cycles de P/E, des charges sont piégées dans l’oxyde. Lors de la rétention, l’énergie thermique permet à ces charges d’être dé-piégées et de partir vers le substrat comme illustré Figure 33. Cela induit une variation de la tension de seuil. On appelle ce mécanisme le « Detrapping » [91]–[93].

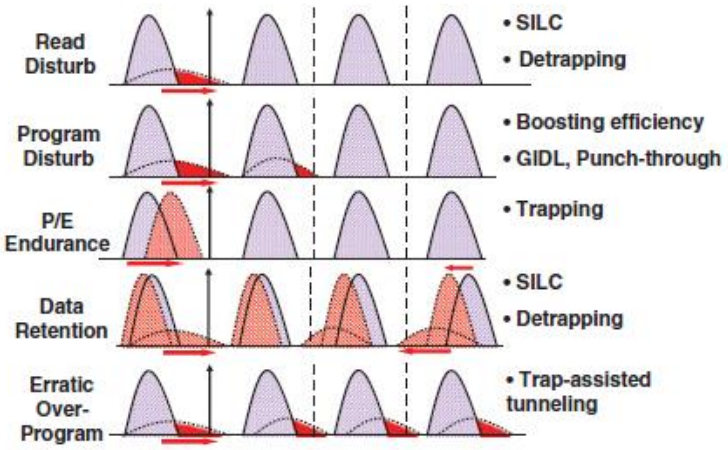


Figure 32 : Illustration des mécanismes de dégradation en fonction de l'utilisation de la mémoire [15, Fig. 6.84]

Son énergie d'activation est comprise entre 0,7 et 1,2 eV [91], [92], [94]. Ce mécanisme est grandement accéléré par une forte température. Dans la mesure où les charges piégées sont la conséquence des P/E, il y a un nombre de charge à dé-piéger croissant avec le nombre de cycle de P/E. Une augmentation du nombre de P/E augmente donc légèrement l'impact de ce mécanisme.

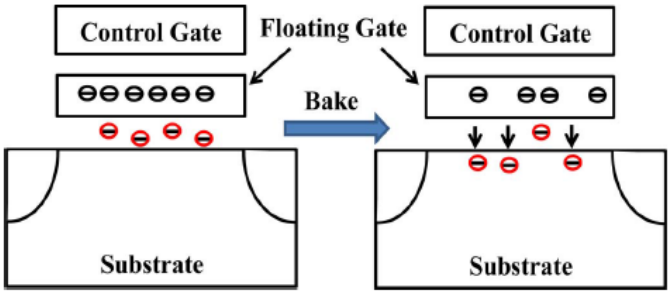


Figure 33 : Mécanisme de Detrapping des mémoires FLASH

Trap Assisted Tunneling

Le Trap Assisted Tunneling (TAT ou SILC) est le mécanisme de défaillance prédominant dans la rétention des mémoires Flash [91]–[93]. Les charges piégées dans l'oxyde baissent la barrière de potentiel entre le canal et la grille flottante. Ainsi, plus le nombre de P/E est important, plus l'énergie nécessaire aux électrons pour passer à travers l'oxyde est faible (voir Figure 34). La température permet d'augmenter dans une moindre mesure la perte de charge dans l'oxyde de grille via ce phénomène.

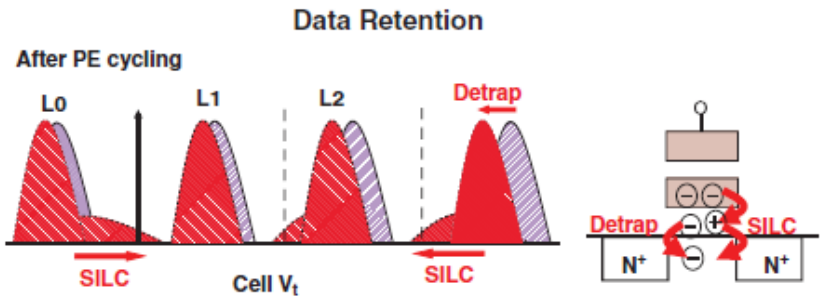


Figure 34 : Intervention des pièges dans l'oxyde sur la fuite de charge des mémoires [15, Fig. 6.2]

L'énergie d'activation de ce mécanisme n'est que de l'ordre de 0,3 à 0,6 eV en fonction du nombre de P/E [91], [92], [94]. Dans le cas d'une cellule mémoire effacée, le TAT peut permettre de faire gagner des électrons à la grille flottante [95].

Interface Trap Recovery

Lors des cycles de P/E, les liaisons Si-H se brisent durant l'effet F-N créant des pièges à l'interface et dans l'oxyde. Lors de la rétention sans stress, des liaisons Si-H de reforment. Cela entraine une modification de la tension de seuil des cellules mémoires. Ce mécanisme appelé ici Interface Trap Recovery (Nit) correspond à la phase de guérison du BTI vu précédemment [91], [92]. Cette dégradation de la tension de seuil est minoritaire par rapport aux deux précédentes, à tel point qu'elle est difficilement observable lors de tests au-delà de 117°C. Par ailleurs, elle est amplifiée par le nombre de P/E effectué [92].
Son énergie d'activation est de l'ordre de 0,1 à 0,5 eV [91], [92], [94].

1.3.5.1.2 Endurance avec les cycles de « Programmation effacement »

Lorsque la dégradation de l'oxyde est trop importante, le point mémoire n'est plus utilisable. Cette problématique est nommée « endurance ». Cette défaillance fonctionnelle est la résultante de l'accumulation des mécanismes cités en 1.3.5.1.1. La présente sous-section concerne ce point particulier.

La diminution des épaisseurs d'oxyde ainsi que la diminution de la marge de bruit entre deux états voisins ont dégradé la robustesse des mémoires vis-à-vis de la dégradation des oxydes. Là où les Flash SLC étaient garanties par le constructeur jusqu'à 100 000 cycles de programmation / Effacement (P/E) par bloc, les nouvelles MLC ne sont plus garanties qu'à moins de 10 000 P/E. Les cycles P/E constituent donc un facteur majeur dans le vieillissement par endurance d'une mémoire Flash [96].

Plusieurs mécanismes sont responsables de la détérioration de l'oxyde entrainant l'augmentation du SILC de grille (Stress Induced Leakage Current) au cours des P/E. Les cycles P/E répétés augmentent le nombre de charges piégées dans le diélectrique [97]. La succession des programmations (injection d'électron dans la grille flottante) accumule une dérive permanente (augmentation) de la tension de seuil qui ne peut pas être récupérée par l'effacement de la cellule (extraction partiel des charges vers le canal) [98]–[100]. Ainsi la tension de seuil des cellules ERS (effacées) augmente avec le nombre de P/E et la tension de seuil des cellules P3 diminue [101] comme on peut l'observer sur la Figure 35. L'usure de l'oxyde amplifie également les problématiques de Program Disturb et de Read/Pass Disturb [102].

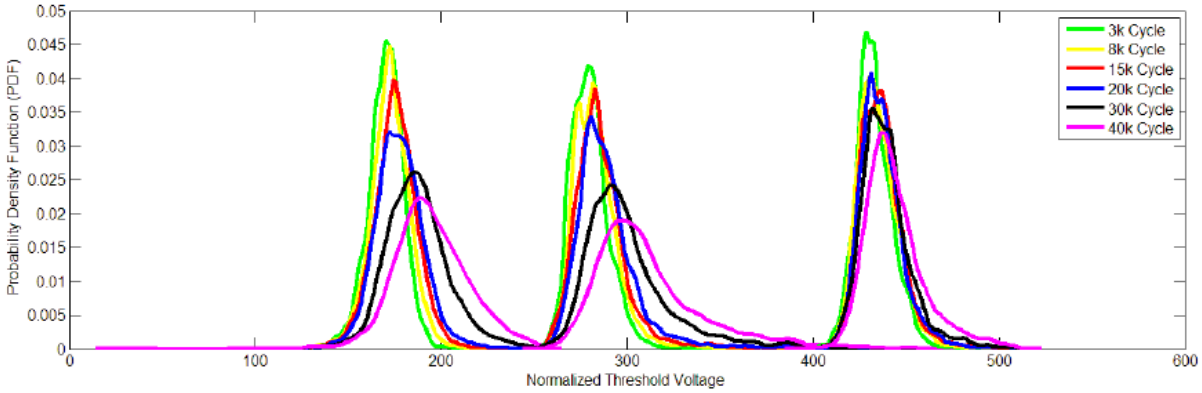


Figure 35 : Influence du nombre de P/E sur les tensions de seuil d'une cellule mémoire FLASH [103, Fig. 11]

Selon une étude récente présentée en 2019, la fiabilité en rétention des données écrites sur des points mémoires ayant subis un grand nombre de PE est grandement améliorée si l'on introduit un temps de repos entre le dernier PE et le suivant [104]. La publication annonce une réduction du taux d'erreur de l'ordre de 60% avec un temps d'attente de 6 heures (entre le dernier PE et l'écriture de la donnée observée) comparé à une écriture immédiatement après le dernier PE.

1.3.5.2 Rétention des données en stockage

Le temps de rétention est une caractéristique importante des mémoires. En règle générale, les fabricants garantissent la rétention des données jusqu'à 10 ou 20 ans. En stockage, les grilles flottantes des cellules mémoires perdent peu à peu leurs charges au cours du temps. Ces pertes de charges se traduisent par une dérive (diminution) des tensions de seuil des cellules mémoires. L'état ayant la tension de seuil la plus élevée est noté « P3 » et est le plus touché par ces fuites [93].

Le temps de rétention d'une donnée est d'autant réduit que le nombre de PE enduré par le point mémoire est grand. Lors de tests de vieillissement accéléré, un pré-cyclage jusqu'à la limite de PE (Programmation Effacement) donnée par le fabriquant permet d'accélérer l'apparition des défaillances en rétention. Ainsi en utilisant plusieurs températures de rétention, on peut en estimer une énergie d'activation [105], [106]. Les fuites de charges étant favorisée par la température, la durée de rétention des données est dégradée par une élévation de la température de stockage de la mémoire après écriture. L'influence de la température lors de la phase d'écriture sur la rétention est quant à elle peu étudiée. On trouve cependant quelques publications à ce sujet : la fiabilité diminue si la différence entre la température d'écriture et la température d'activation est grande [106], [107]. Ainsi une programmation à froid puis une lecture à chaud (avec donc un grand écart de température) induirait une fiabilité en rétention moindre [20]. Cet aspect sera discuté dans la section 2.5.1.1.2.
De plus, la répartition des erreurs au sein d'une même puce n'est pas uniforme. Certaines cellules sont plus enclines à perdre ou gagner des charges au cours de la rétention. Cette variation d'une cellule à l'autre serait due à des problèmes de désalignement des contacts lors de la fabrication du wafer [108], [109].

Afin de prédire le temps de rétention des données dans une mémoire Flash, plusieurs modèles ont été proposés. Ces modèles estiment la perte de charge de la grille flottante jusqu'à ce que la tension de seuil de celle-ci soit confondue avec un autre état. Le modèle en puissance et Arrhenius [40], [96] est :

$$t_R = t_0.(cycles)^{-n}.e^{\frac{E_a}{k.T}} \tag{58}$$

Où :

- t_R est le temps de rétention moyen des données
- t_0 est une constante dépendant de la technologie
- $cycles$ est le nombre de P/E initial effectué
- n est un l'exposant de la loi puissance prenant en compte le cyclage P/E

Les valeurs de *n* selon la *JEDEC* sont comprises entre 0,4 et 0,7 [40]. Les énergies d'activation apparentes estimées par les différentes publications sont comprises entre 0,3 à 1,9 eV, et de 1,1eV pour des NOR [105], [110], [111]. Cette valeur estimée semble être fonction de la plage de température considérée dans les mesures comme on peut le voir sur la Figure 36. Cette figure présente les énergies d'activations extraites par publication en faisant bien apparaitre les conditions thermiques utilisées.

Ce modèle est également particulièrement peu adapté aux mémoires avec des diélectriques High-κ selon cette publication [112].

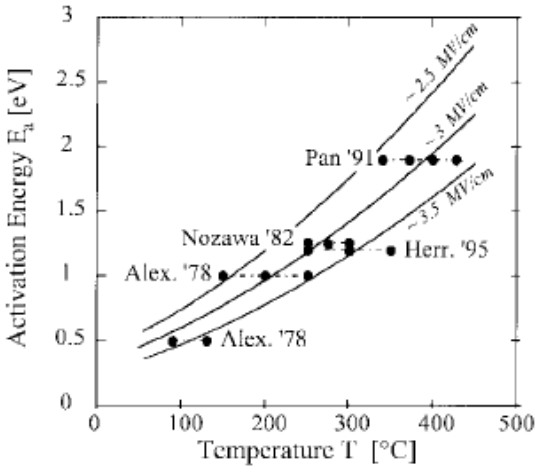


Figure 36 : Energies d'activation des Flash estimées par la littérature [113, Fig. 8]

Si l'énergie d'activation semble varier avec la température, alors on ne peut pas considérer qu'il n'y a qu'un seul mécanisme et que la loi d'Arrhenius s'applique ici. Il semblerait coexister plusieurs mécanismes distincts de conduction dans l'oxyde ONO (effet tunnel direct avec l'oxyde supérieur, Poole–Frenkel, effet tunnel assisté par multi-phonons, etc.), chacun d'eux présentant une énergie d'activation qui lui est propre. Devant cette incertitude, le modèle alternatif T a été présenté [113].

$$t_R = t_0 . e^{\frac{-T}{T0_{DR}}} \tag{59}$$

Où :

- t_R est le temps de rétention moyen des données
- t_0 est une constante dépendant de la technologie
- $T0_{DR}$ est une température caractéristique de la rétention de donnée en Kelvin
- T est la température absolue

La valeur de $T0_{DR}$ est aux alentours de 21 K. Le lien avec le modèle « Modèle puissance et Arrhenius » peut être fait via la formule suivante :

$$E_a(T) = \frac{k.T^2}{T0_{DR}} \tag{60}$$

Ce type de composant est très sensible aux tensions appliquées, on n'utilise généralement pas la tension comme facteur d'activation car elle modifie le comportement du circuit de manière incompatible avec sa fonctionnalité normale [96, p. 32].

En plus du temps, de la température, et du nombre de PE effectué, d'autres mécanismes peuvent participer à la perte d'information des mémoires Flash. Ces mécanismes sont liés à l'architecture condensée des mémoires Flash NAND. Lors d'une lecture ou d'une écriture sur une page, les pages voisines subissent également un stress électrique qui peut amener à une injection de charges dans la grille flottante des points mémoires non sélectionnés.

1.3.5.2.1 Pass Disturb

Dans une architecture NAND, tous les transistors à grille flottante d'une même ligne de bit sont en série. Ainsi pour lire ou écrire une page, les autres transistors de la ligne de bit doivent être passants. Pour ce faire, l'application d'une tension V_{pass} (voir Figure 22) permet de rendre la cellule passante quel que soit l'état de la cellule. V_{pass} est plus élevée que la tension de seuil d'une cellule au niveau P3. Les cellules non sélectionnées de la même Bit Line subissent ainsi une légère programmation du fait de la tension V_{pass} appliquée sur leur grille afin de les rendre passantes. On appelle ce phénomène le « Pass Disturb ».

Alors qu'il fallait lire une page un million de fois pour voir apparaitre un Pass disturb sur les anciennes technologies SLC, il suffit de la lire moins de cent mille fois sur les nouvelles MLC pour obtenir du Pass disturb [19]. Ce phénomène prend donc de plus en plus d'importance avec les réductions de marges de bruit en augmentant le nombre de bits par cellule mémoire et en diminuant l'épaisseur d'oxyde de grille avec les nouvelles technologies. Les cellules les plus touchées par le Pass Disturb sont celles étant effacées [102] comme illustré sur la Figure 37.

(a) Sans « Pass Disturb » (b) Après de nombreux « Pass Disturb »

Figure 37 : Dérive des tensions de seuil liée au Read/Pass Disturb

1.3.5.2.2 Program Disturb

Lorsque l'on programme une cellule d'une Flash, la tension V_{prog} appliquée sur toutes les cellules de la même Word Line que celle à programmer peut générer par effet Fowler-Nordheim une légère programmation de celles qui étaient auparavant effacées [114]. En effet une cellule effacée contient peu de charges négatives dans sa grille flottante, et possède donc une tension de seuil basse. Ainsi le champ électrique (V_g-V_t) appliqué sur le canal est plus important pour une cellule effacée que pour une cellule programmée au niveau P3 (cas extrême possédant une tension de seuil bien plus élevée). La Figure 38 illustre le cas de l'écriture d'une cellule. Les cellules mémoires de la même page subissent du « Program disturb » et toutes les cellules de la même ligne de bit subissent du « Pass disturb ».

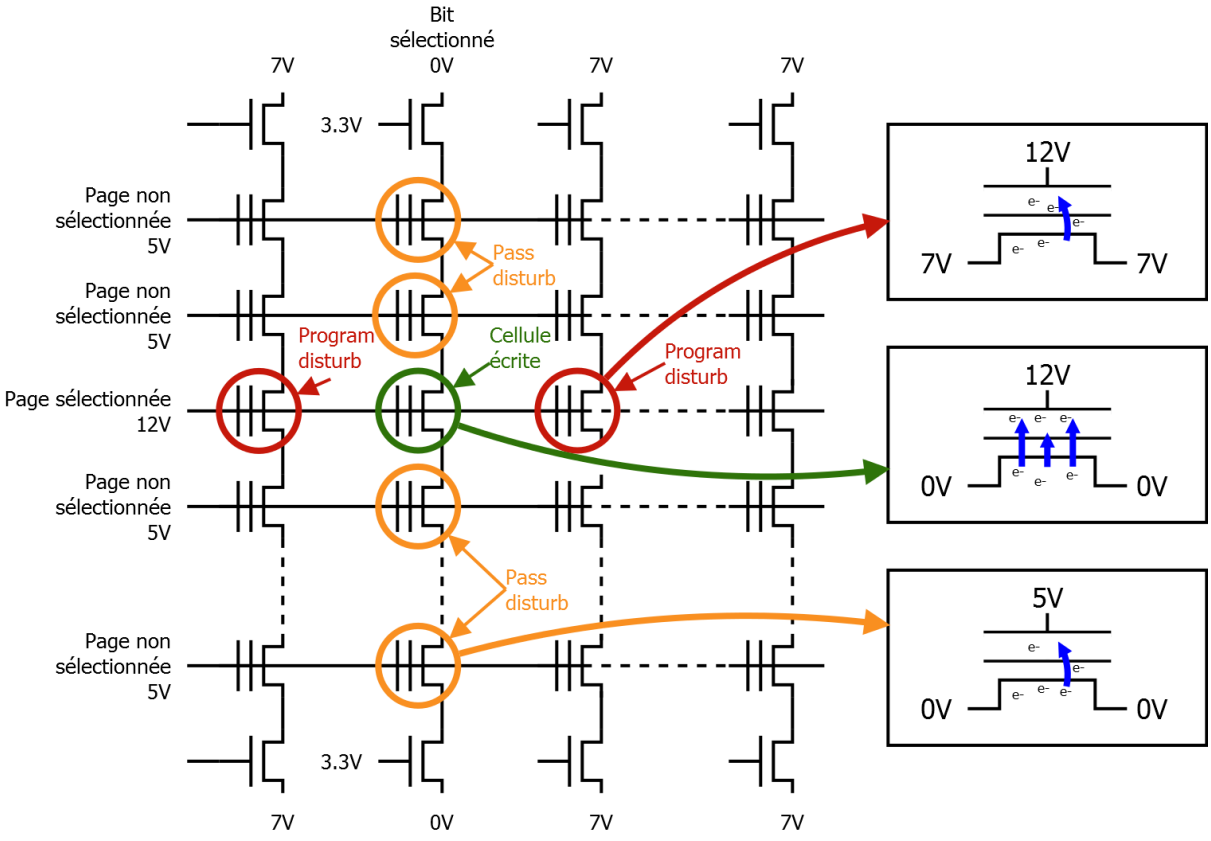


Figure 38 : Program Disturb dans les Flash NAND

1.3.5.2.3 Read Disturb

Lors de la lecture d'une cellule sur une mémoire Flash NOR ou NAND, une tension V_{ref} (V_a, V_b ou V_c) est appliquée sur toutes les grilles de la même WL (page). Les grilles de la BL à lire sont soumises à une différence de potentiel d'environ un volt afin d'estimer la tension de seuil de la cellule à lire et donc son contenu. Mais les autres cellules de la même WL ont leur BL à la masse ou flottante. Ainsi encore, une légère écriture est possible. Quelques charges

négatives peuvent passer dans la grille flottante par effet F-N et donc augmenter la tension de seuil. Ce phénomène touche également principalement les cellules à l'état effacés [114]. L'injection « accidentelle » de charge dans la grille flottante est favorisée par la présence de pièges dans l'oxyde (SILC). Ces pièges sont la résultante de défauts initiaux et une accumulation de cycles de PE [15]. La Figure 39 montre l'exemple où une page est lue. Toutes les cellules sur les mêmes BL subissent du « Pass disturb ».

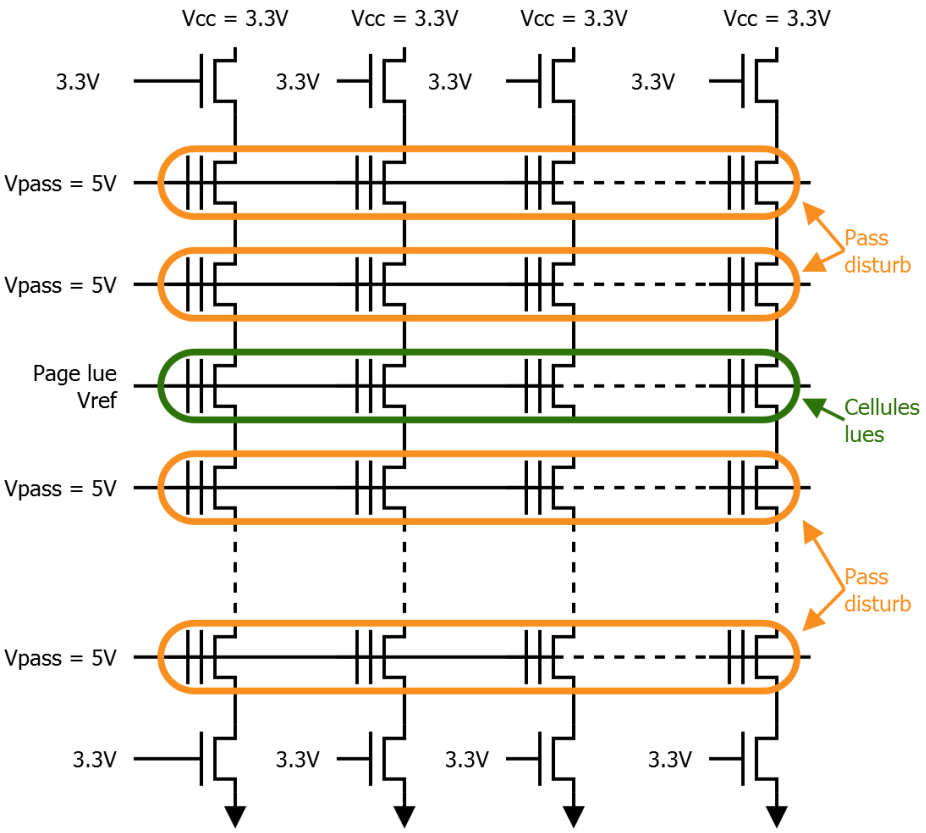


Figure 39 : Read Disturb dans les Flash NAND

1.4 Synthèse

Ce premier chapitre vient de parcourir toutes les notions qui sont nécessaires pour appréhender les deux chapitres suivants. Il a en effet balayé les aspects technologiques, mécanismes de dégradation et les méthodes d'analyse des résultats d'un point de vue probabiliste.

Les FPGA sont des circuits logiques configurables comprenant des technologies de pointes. Cette flexibilité nous permettra de créer des structures de test dédiées à la mise en évidence de mécanismes de dégradation spécifiques au niveau transistor MOS. Les quatre principaux

sont : l'injection de porteurs chauds, le BTI, le TDDB et l'électromigration. Le HCI est principalement actif à basse température et haute fréquence de fonctionnement du circuit. Le BTI est l'un des principaux mécanismes de dégradation des performances des circuits. Il est amplifié par une élévation de la température et par un rapport cyclique élevé du transistor à l'état passant. La présence de TDDB sera très dépendante de la bonne maîtrise du procédé de fabrication du nœud technologie étudié. Ce mécanisme peut être une accumulation de HCI et de BTI. Ces trois mécanismes sont également amplifiés par une augmentation des champs électriques, et conduisent à une diminution de la fréquence de fonctionnement du circuit jusqu'au claquage d'oxyde (TDDB). L'EM est significative à haute température (plusieurs centaines de degrés Celsius) et sous une forte densité de courant dans les pistes du circuit intégré. Contrairement aux précédents mécanismes, cette dégradation ne concerne pas l'oxyde entre le canal et la grille, mais les interconnexions. Cette dégradation conduit à une augmentation de la résistivité des pistes impactées.

Nous avons également développé la description de l'autre famille de composant de notre étude - les mémoires Flash – dans ce chapitre. Il existe deux principales architectures internes : la NOR et la NAND. Chacune possède ses avantages et ses inconvénients. Les Flash NAND permettent une forte densité de stockage à moindre coût, mais au détriment de la vitesse de lecture et de la fiabilité. Les Flash NOR quant à elle ont un coût élevé mais une fiabilité bien meilleure et des temps de lecture courts. Toutes ces architectures reposent sur le transistor à grille flottante pour stocker l'information sous forme de tension de seuil. Un même point mémoire peut contenir plusieurs bits (on les appelle MLC dans ce cas). Cette augmentation du nombre de niveau logique est au détriment de la marge de sécurité entre deux états voisins, et par conséquent de la fiabilité. La fiabilité des mémoires est à considérer sous deux aspects : la rétention et l'endurance. Les cycles d'écriture-effacement usent le point mémoire au niveau de l'oxyde notamment. Cette usure mène à une baisse de la fiabilité en rétention, voir à l'impossibilité d'utiliser la cellule. La rétention de données est affectée par beaucoup de facteurs : l'usure, la température, les lectures répétées mais aussi les écritures répétées.

Toutes études de fiabilité impliquent une part d'aléatoire. La modélisation de cette statistique qui va être appliquée à la dispersion des instants de défaillances sera élaborée au moyen d'une modélisation physique de la cinétique des dégradations, puis - par la méthode du maximum de vraisemblance - les paramètres de la distribution statistique seront estimés.

Le chapitre suivant concernera l'étude des mémoires Flash. Il décrira les moyens mis en place afin de mettre en évidence les différents types de dégradation. Ensuite les résultats

observés seront analysés. Puis pour finir, la fiabilité de ce type de composant sera évaluée tant sur l'aspect endurance que rétention de données.

Chapitre 2

Etude des mémoires Flash

2.1 Introduction

2.1.1 Plan général

Les composants de type mémoires sont aujourd'hui utilisés dans de nombreuses applications. La quantité d'information stockable sur une même puce ne fait qu'augmenter avec les avancées technologiques. Cependant, les Flash NAND qui composent la plupart des supports de stockage de masse actuels sont quasi exclusivement utilisées dans le domaine grand public. En plus des améliorations technologiques, une surcouche de micro-gestion de ce type de mémoires est dorénavant à prendre en compte. Cette étude permet d'approfondir la connaissance de la fiabilité de cette filière en technologie DSM.

Les campagnes de test décrites dans ce chapitre ont été réalisées sur des mémoires NAND (20 nm MLC), sur des mémoires NOR (65nm « Mirror bit » SLC). L'étude de la fiabilité des DDR3 (en 20nm) fait également partie du cadre initial de ce projet.

Ce chapitre présente les essais mis en places afin de mettre en évidence différents mécanismes de défaillance. En fonction de ces tests, un certain nombre de résultats ont été mis en lumière. L'originalité la plus marquante de cette étude est la considération de plusieurs températures d'écriture pour les mémoires Flash. Une hypothèse fondée sur nos résultats permet ici de montrer que les dérives de fiabilité à basse température proviennent du circuit de gestion périphérique et non de la grille flottante en elle-même. Les aspects « rétention de données » et « endurance » seront différenciés. Une analyse statistique des résultats sera ensuite proposée. Une méthode de généralisation de ces résultats à plusieurs ECC ainsi qu'à différents niveaux de sur-approvisionnement (OP) sera également étudiée. En effet, les algorithmes correcteurs d'erreurs (ECC) implémentés dans les SSD ne font que s'améliorer. Ces améliorations ont une importance cruciale sur leur fiabilité en rétention et doivent donc être pris en compte.

Cette étude se limitera à l'étude de la fiabilité du silicium. Pour ce faire, des bancs ont été développés par les équipes Thales ou des sous-traitants (pour des problématiques de coûts). Les phénomènes de défaillance issus des boitiers ou banc de test ne sont pas pris en compte dans l'analyse des résultats de cette thèse. De même, la description des bancs ne sera pas exhaustive car non développée par mes soins. L'objet de cette thèse est bien le recueil des résultats de test et l'exploitation des données afin de construire des modèles prédictifs sur les nouvelles technologies.

2.1.2 DDR3

L'étude de la fiabilité de mémoires DDR3 était initialement prévue. Les plans de test comprenaient : une étude de l'influence de la température de fonctionnement (de 25°C à 140°C) sur la rétention de données, différentes période de rafraichissement des points mémoires afin d'accélérer l'apparition des pertes de données, différentes fréquences de fonctionnement (400MHz et 800MHz) et de la réécriture en boucle d'une données jusqu'à défaillance du point mémoire.

A ces fréquences de fonctionnement, la conception d'une carte de test n'est pas triviale. C'est pourquoi cette tâche a été sous-traitée à une entreprise externe au projet. De nombreux problèmes de conception ont été constatés tout au long du projet avec ces cartes. Il y avait un sérieux problème de contact entre les DDR3 et le socket, des manques de bonnes pratiques dans la conception du FPGA chargé de piloter les mémoires, une stratégie d'asservissement thermique imparfaite. De nombreuses actions ont été entreprises afin de corriger les défauts de ce banc. Sa fiabilité est cependant restée malgré tout bien inférieure comparé à celle des composants testés. Ces retards n'ont pas permis de réaliser de durées de vieillissement suffisamment longues. C'est pourquoi aucune présentation ou analyse de ces résultats ne sera faite dans ce mémoire.

2.1.3 Flash NOR

A la fin du vieillissement des mémoires NOR, aucune erreur n'a été observée – toutes files d'essai confondues. Ces différentes files d'essai - analogues à celle décrite par la suite pour les mémoires Flash NAND - comportaient :

- Une étude de l'influence de la température de stockage ou d'utilisation sur la rétention. Les températures appliquées ont été de 25°C, 85°C et 125°C.
- Une étude de l'influence de la température d'écriture. Les températures d'écritures étaient de -40°C, -10°C, 25°C et 85°C.
- Différents motifs de données écrits dans les mémoires.
- Un vieillissement des mémoires avec des données initialement écrites soumis à un grand nombre de lecture à haute température.

Cette étude a concerné 86 mémoires distribuées sur les différentes files de test. Elles ont été vieillies pendant plus de 20 000 heures. A l'issue de ce stress très longue durée, aucune perte de donnée n'a été observée. Cette architecture NOR, même en technologie DSM, a ici montré une grande fiabilité en rétention. Cependant, en l'absence de défaillances, aucun calcul n'est possible, ni même aucune analyse des mécanismes de défaillance. C'est pourquoi ce mémoire détaillera uniquement les résultats sur les Flash NAND où de nombreuses erreurs ont été constatées.

2.1.4 Flash NAND

L'étude des mémoires Flash NAND constituent un enjeu majeur. Ce type de composant est pour l'instant très peu utilisé dans les domaines spatial, militaire ou avionique. Toute la suite de ce chapitre sera consacrée au Flash NAND. D'une part, c'est la file de composant qui a présenté le plus de résultats en terme de défaillance, et d'autre part son utilisation n'est pas commune – comparé aux générations précédentes de mémoire EEPROM – cet aspect doit être correctement pris en compte dans l'étude de sa fiabilité.

2.2 Détail de l'architecture des données dans la Flash NAND

D'une référence de Flash NAND à une autre, les dimensionnements des zones mémoires sont très fluctuantes comme décrit dans la section 1.2.2.2. La mémoire choisie pour cette étude est architecturée comme illustré dans la Figure 40. Une page est ici composée de 8 kilo octet de données. Les données s'écrivent par page entière. Elles sont regroupées par groupe de 256 pour former un bloc. La mémoire s'efface par bloc entier. Dans les plans de test décrits en 2.4, les ZM (zones motifs) sont composées de plusieurs blocs consécutifs. Dans l'exemple de la Figure 40, une ZM est un groupement de 43 blocs.

Dans la suite du document, par abus de langage, l'écriture d'une page NAND supposera implicitement : un effacement préalable du bloc, puis une programmation de toutes les pages du bloc.

2.3 Stratégie des essais

Deux types de vieillissement ont été appliqués sur les Flash NAND : du vieillissement dynamique avec le composant alimenté et du vieillissement de stockage. Le vieillissement composant non alimenté permet de simuler des phases de stockages, comme par exemple une

clef USB dans un tiroir, là où l'activation alimentée simule une application de type serveur ou calculateur embarqué.

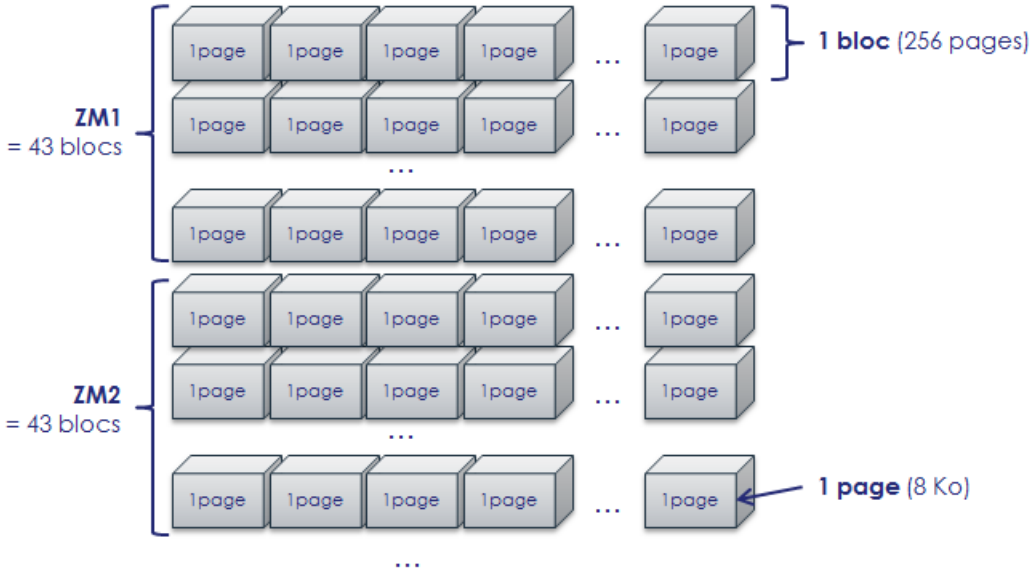


Figure 40 : Organisation de la mémoire Flash NAND de l'étude

2.3.1 Choix du composant

Les mémoires NAND sont caractérisées par une grande densité de données et un protocole de lecture complexe. La référence étudiée ici est une mémoire FLASH NAND MLC du fabricant Micron ayant une finesse de gravure de 20 nm. Sa densité est de 64 Gbit pour une surface de moins de 2 cm². Nous avons choisi d'approvisionner des composants de gamme industrielle, ayant des conditions d'utilisation de -40°C à +85°C.

2.3.2 Choix des scenarii de vieillissement

Le but de ces essais est de provoquer des défaillances suivant plusieurs conditions afin d'isoler les mécanismes de dégradation. Pour commencer, il existe différents types de défaillances possibles pour une mémoire (hors boîtier) :

- La mémoire ne peut plus être écrite correctement, c'est donc une défaillance due à l'endurance du point mémoire
- La donnée écrite initialement (qui a donc été relue correctement au moins une fois) n'est plus lisible sans erreur à un certain moment au cours du temps, on a donc dans ce cas affaire à une défaillance en rétention
- Une défaillance peut également survenir dans les circuits périphériques chargés de gérer les points mémoires à proprement parler. Ces circuits peuvent être dégradés par d'autres mécanismes de dégradation comme le HCI par exemple.

- Les intermétalliques au niveau des entrées/sorties peuvent aussi présenter des circuits ouverts (ou courts circuits avec d'autres pistes) ce qui empêchera les communications de la mémoire, et conduira donc à une défaillance.

Les différents facteurs d'accélération du vieillissement sont habituellement la température, de nombreux cycles d'écriture-effacement (PE : Program Erase) et un nombre important de lectures. Ce chapitre va donc présenter les plans de test mis en œuvre, ainsi que les moyens matériels déployés, afin d'observer ces mécanismes. Le Tableau 3 présente le lien entre les différentes files d'essai et les facteurs de stress appliqués.

	NAND1	NAND2	NAND3	NAND4	NAND 5
Température de jonction	X	X	X	X	X
Température de programmation	X			X	
Durée de stockage initial		X			
Nombre de cycles de P/E initial			X		X
Fréquence de lecture				X	
Fréquence de cycles de P/E					X
Composant alimenté pendant le vieillissement				X	X

Tableau 3 : Synthèse de files de test NAND

Les cinq types de conditions de vieillissement concernant les mémoires ont été nommés respectivement NAND 1 à NAND 5. Les quatre premières permettent de mesurer la fiabilité en rétention, alors que la NAND 5 permet de tester l'endurance pure. Les tests de rétention vont donc viser les mécanismes de vieillissement de SILC et de dé-piégeage des charges dans la grille flottante présenté dans la section 1.3.5. Tandis que les tests d'endurance recherchent les mécanismes de piégeage persistant de charge dans l'oxyde. Ces points sont synthétisés dans le Tableau 4.

File d'essai	Mécanismes recherchés	
NAND 1	SILC	
NAND 2	SILC	Usure initiale des matériaux due au stockage
NAND 3	SILC	Dé-piégeage de charges
NAND 4	SILC	Read-Disturb, Pass-Disturb
NAND 5	Piégeage de charges	

Tableau 4: Synthèse des mécanismes de vieillissement recherchés par file

2.4 Objectifs des essais

Cette partie présente pour chaque essai de NAND 1 et NAND 5, les conditions mises en place pour révéler un mode de défaillance.

2.4.1 NAND 1 : Rétention en fonction de la température d'écriture et de la température de stockage

L'objectif de cet essai est d'analyser l'influence de la température de programmation et de la température de stockage sur la rétention de données écrites au début du test dans la mémoire. Comme vu dans l'état de l'art en 1.3.5.2, certaines publications mentionnent que la programmation à basse température engendre une baisse de la fiabilité. Cette étude permet donc d'approfondir la connaissance dans ce domaine.

Phase 1 : Ecriture des données initiales

Plusieurs mémoires ont été écrites à des températures de -40°C ; -10°C ; 25°C ou 85°C. Cette gamme de température balaye l'ensemble de la gamme industrielle approvisionnée. La manipulation des mémoires ainsi que l'écriture ont été effectuées au moyen d'un testeur type ATE Mu-test (voir Figure 41). La régulation thermique est assurée en laboratoire à l'aide d'une tête thermique avec flux d'azote.

Les mémoires étant MLC (2 bits par grille flottante), plusieurs motifs de données ont été écrits de façon à programmer différentes tensions de seuil dans les cellules. Pour ce faire, plusieurs motifs de données ont été mis en œuvres dans différentes zones mémoires (ZM). Au nombre de douze, ils sont listés dans le Tableau 5. Chaque motif est répété sur 43 blocs consécutifs.

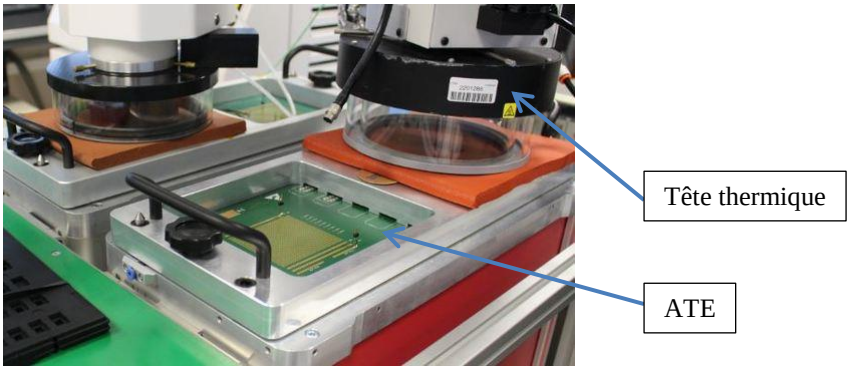


Figure 41 : Testeur ATE avec tête thermique

Zone Motif mémoire	Motif stocké
ZM1	Toutes les cellules MLC à ‘11’
ZM2	Toutes les cellules MLC à ‘01’
ZM3	Toutes les cellules MLC à ‘10’
ZM4	Toutes les cellules MLC à ‘00’
ZM5	Blocks à 25% à ‘00’
ZM6	Blocks à 50% à ‘00’
ZM7	Blocks à 75% à ‘00’
ZM8	Damier 1 : AA55
ZM9	Damier 2 : 55AA
ZM10	Damier 3 : 00FF
ZM11	Quelques 00,01,10 entourés de 11
ZM12	Séquence aléatoire d’une longueur de 8k-octets (une page)

Tableau 5 : Zones Mémoires NAND sans polarisation

Les ZM1 à ZM4 correspondent aux différents niveaux de tension de seuil. La ZM1, remplie uniquement de ‘11…11’, est donc laissée vierge. Pour les ZM2 à 4, la correspondance entre les niveaux de tension de seuil et l’état logique dépend des choix faits par le fabricant. On peut cependant faire l’hypothèse d’une disposition selon un code de Gray comme sur la Figure 42.

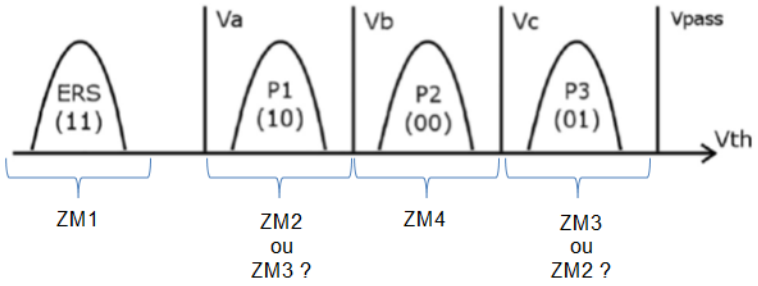


Figure 42 : Hypothèse initiale de correspondance entre les états MLC et les ZM

Les ZM5 à ZM7 correspondent respectivement à 25%, 50% et 75% de pages consécutives à '00' par bloc. Le cas de 100% est géré par la ZM4. La Figure 43 illustre le cas de la ZM5. Pour chaque bloc de la ZM5, les 64 premières pages (25% de 256) sont composées uniquement de '0', et les suivantes restent vierges (donc composées de '1').

	Bloc ZM5
Page 0	00…00
Page 1	00…00
Page 2	00…00
…	…
Page 63	00…00
Page 64	11…11
…	…
Page 253	11…11
Page 254	11…11
Page 255	11…11

25%

75%

Figure 43 : Illustration de la structure d'un bloc de la ZM5

La ZM8 est composée de motifs en damier 'AA55', la ZM9 de damier '55AA' et la ZM10 de damier '00FF'. La ZM11 est composée de points mémoires chargés '00' entourés de point mémoires à '11' non chargés. Ce motif est potentiellement plus sensible au read-disturb. Pour finir, la ZM12 est un motif aléatoire de longueur 8 k-octets (une page) répété sur toutes les pages. Cette dernière ZM est la plus proche d'une application réelle.

Phase 2 : Vieillissement accéléré

Une fois les mémoires écrites aux différentes températures, elles sont placées en étuve pour toute la durée du vieillissement. Les températures consignes des étuves sont 25°C ; 85°C ou 125°C. Cette phase de vieillissement simule un stockage des mémoires non polarisé.

Trois mémoires sont vieillies dans chaque condition de test, soit un total de 36 DUT (composant sous test) pour cette file.

NAND1	T° vieillissement		
T° programmation	**25°C**	**85°C**	**125°C**
-40°C	3	3	3
-10°C	3	3	3
25°C	3	3	3
85°C	3	3	3

Tableau 6 : Nombre de mémoires par condition de vieillissement, NAND1

Phase 3 : Mesures de reprise

Durant le vieillissement, des relectures sont réalisées de manière périodique. Ces caractérisations permettent d'avoir un suivi régulier des dégradations des mémoires au cours du temps. En moyenne, il y a une relecture toutes les 1000 heures d'étuve. Lors des relectures, les DUT sont lues à 20°C (zone laboratoire régulée en température) sur testeur. Lors de ces tests, le nombre de page en erreur est enregistré, ainsi que le nombre de Read-retry nécessaire pour lire correctement les données. Une page est déclarée défaillante si aucune méthode de Read-retry et aucun ECC ne permet de retrouver la donnée écrite. Le code correcteur d'erreur (ECC) considéré (conseillé par le fabricant) permet de récupérer 40 bits erronés dans une zone de 1 117 octets consécutifs, sachant qu'une page dans cette mémoire Flash NAND est de 8 Ko. Ensuite les pièces retournent en étuves jusqu'à la prochaine mesure de reprise.

2.4.2 NAND 2 : Rétention en fonction de la température de stockage et du temps de stockage initial.

Dans le secteur industriel, les composants mémoires peuvent être stockés des dizaines d'années avant d'être utilisés pour la première fois. Ceci concerne par exemple les stocks stratégiques de composants (pour pallier au manque d'approvisionnement d'une référence obsolète une fois sa production arrêtée), ou bien le cas des systèmes de type missiles qui peuvent être utilisés pour la première - et unique – fois au bout de plusieurs années. Il est donc important de savoir si l'utilisation d'anciennes mémoires neuves peut avoir un impact sur la rétention à cause du vieillissement des matériaux.

L'objectif de cette file est ainsi l'étude de l'influence du temps de stockage avant la première utilisation du composant et de la température de stockage sur la rétention des données. Afin de simuler cette phase de pré-stockage (avant écriture d'une donnée) en un temps raisonnable, les mémoires ont été chauffées à plus de 100°C pendant quelques centaines d'heures. L'hypothèse d'une fourchette d'énergie d'activation entre 0,5eV et 0,725eV nous

permet d'estimer une simulation de pré-stockage équivalente à 22°C de l'ordre de 10, 20 ou 30 ans suivant les durées appliquées. Le détail de ces différentes sous-files de test est donné Tableau 7. Une fois ce pré-vieillissement terminé, les mémoires ont été écrites à température ambiante avec les contenus du Tableau 5. Les mémoires ont ensuite été vieillies pendant environs 20 000 heures à 85°C et à 125°C. Des caractérisations (avec relecture du contenu écrit initialement) ont été réalisées périodiquement.

Avant écriture des données				**Après écriture des données**	
Durée de pré-vieillissement	Température de pré-vieillissement	Estimation du temps équivalent de préstockage à 22°C	Valeur de l'Ea utilisée	Température de vieillissement	
				85°C	125°C
229 h	100°C	10 ans	0,725 eV	2 DUT	2 DUT
229 h	150°C	10 ans	0,5 eV	2 DUT	2 DUT
458 h	100°C	20 ans	0,725 eV	2 DUT	2 DUT
458 h	150°C	20 ans	0,5 eV	2 DUT	2 DUT
687 h	100°C	30 ans	0,725 eV	2 DUT	2 DUT
687 h	150°C	30 ans	0,5 eV	2 DUT	2 DUT
Durée de vieillissement en rétention de donnée				**20 512 h**	**18 788 h**

Tableau 7 : Nombre de mémoires par condition de vieillissement, NAND2

A l'issue de ce vieillissement, aucune défaillance ou perte de donnée n'a été observée sur ces 24 NAND. On peut donc considérer que cet aspect n'est pas problématique sur cette technologie.

2.4.3 NAND 3 : Rétention en fonction de la température de stockage et du nombre de PE initial.

L'objectif de cette file est d'étudier l'influence du nombre de cycles d'écriture-effacement (PE) avant programmation des mémoires pour les tests de rétention. Comme on peut le lire dans la littérature (voir 1.3.5.1), plus une mémoire est usée, plus la rétention est mauvaise. Cette usure provient d'un grand nombre de PE. La mémoire choisie accepte 3000 PE au maximum d'après le fabriquant.

Phase 1 : Accumulation de PE avant vieillissement

On réalise pour chaque composant 500, 1500 ou 3000 PE. Certains concepteurs de systèmes, utilisateurs de ces mémoires, se limitent à 1500 voire 500 PE pour les plus conservateurs. Les cycles de PE ont été réalisés sur 2 blocs par zone mémoire, soit un total de 24 blocs.

Phase 2 : Ecriture des données initiales

Ensuite, les mémoires ont été écrites à température ambiante avec les contenus du Tableau 5.

Phase 3 : Vieillissement accéléré

Les conditions de vieillissement sont identiques à la file NAND1. La répartition des composants est donnée dans le Tableau 8.

NAND3	T°C vieillissement		
Nb cycles PE	**+25°C**	**+85°C**	**+125°C**
500	1	3	3
1500	1	3	3
3000	1	3	3

Tableau 8 : Nombre de mémoires par condition de vieillissement, NAND3

Phase 4 : Mesures de reprise

Les mesures de caractérisation intermédiaires sont identiques à celles décrites dans la file NAND1.

2.4.4 NAND 4 : Rétention en fonction de la température d'écriture, de la température d'activation et du nombre de lecture

L'objectif de cet essai est d'évaluer la rétention de données des mémoires préalablement écrites à diverses températures, lorsqu'elles sont alimentées et régulièrement relues. La fréquence de lecture soutenue permettra d'activer au maximum les mécanismes de « read disturb » ou de « pass disturb » en plus de ceux mis en œuvre dans la file NAND1.

Phase 1 : Ecriture des données initiales

Comme pour la NAND1, plusieurs mémoires ont été écrites sous des températures de -40°C ; 25°C ou 85°C. La manipulation des mémoires et l'écriture a été effectuée au moyen d'un testeur type ATE Mu-test. La régulation thermique s'est faite en laboratoire à l'aide d'une tête thermique avec flux d'azote. Contrairement aux files d'essai sans polarisation pendant le vieillissement, il n'y a dans les files dynamiques (composants polarisés et activés pendant le vieillissement) que 5 Zones Motifs (ZM). La séquence aléatoire de la ZM5 n'a que 8 octet de longueur. Cette limitation est liée à la complexité de programmation et de vérification des contenus en continu. Chaque motif est répété sur 600 blocs consécutifs.

Zone Motif mémoire	Motif stocké
ZM1	Toutes les cellules MLC à ‘11’
ZM2	Toutes les cellules MLC à ‘01’
ZM3	Toutes les cellules MLC à ‘10’
ZM4	Toutes les cellules MLC à ‘00’
ZM5	Séquence aléatoire d’une longueur de 8 octets

Tableau 9 : Zones Mémoires NAND Dynamiques

Phase 2 : Vieillissement accéléré et activation

Une fois les mémoires écrites aux différentes températures, elles sont placées sur le banc d’activation dynamique - illustré Figure 44 - pour la durée du vieillissement. Les températures consignes des réchauffeurs sont 25°C ; 85°C ou 110°C.

Figure 44 : Photo du banc d’activation dynamique des mémoires NAND

L’architecture système du banc d’activation dynamique est illustrée par la Figure 45. Ce banc est basé sur un FPGA « maître » chargé de gérer les différentes cartes auxiliaires et l’activation des DUT (Device Under Test). La partie alimentation permet de régler la tension d’alimentation des DUT voir même de les mettre hors tension. Une résistance de 0,15 ohm placée en série permet également de mesurer la consommation des mémoires. Le réchauffage des DUT se fait au moyen de réchauffeurs locaux (un par DUT). L’asservissement thermique est réalisé à l’aide de PT100 (RTD) entre le réchauffeur et le boitier du DUT. Notons qu’une correction affine est appliquée sur l’asservissement pour obtenir une température de jonction et non de boitier. Afin de calibrer les coefficients de la correction du contrôle en puissance des réchauffeurs, des mesures de température à l’aide des diodes de protection des broches ont été initialement réalisées. Cela permet donc de caler la température de jonction avec la puissance injectée sur les réchauffeurs en tenant compte des résistances thermiques.

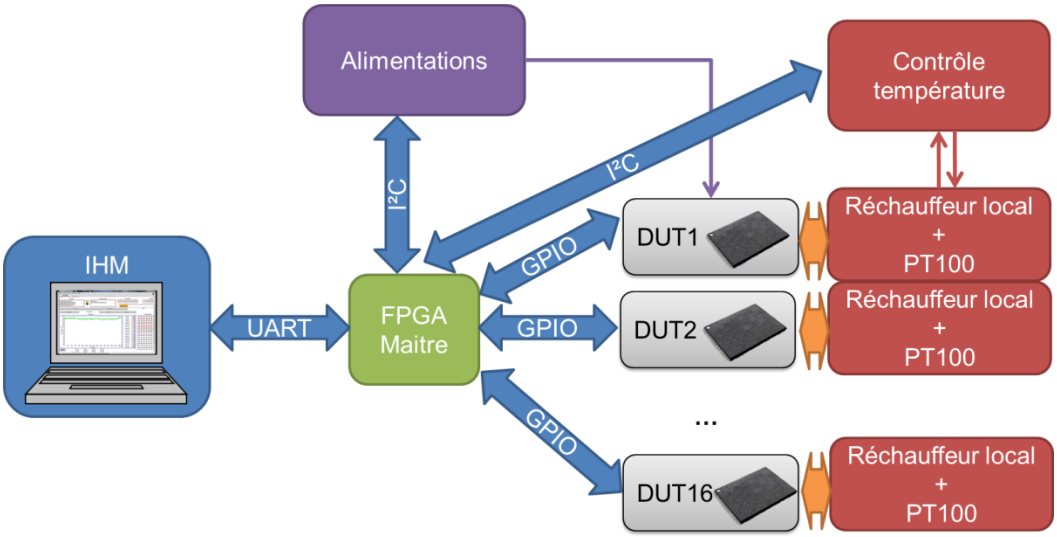


Figure 45 : Schéma du banc d'activation

Concernant l'activation des mémoires à proprement parler, 50% des blocs écrits initialement sont relus de manière continue. Notons que par soucis d'uniformité du nombre de lectures enduré par chaque page, toutes les méthodes de read retry sont à chaque fois réalisées. En pratique, chaque point mémoire est relu toutes les 45 minutes. Le nombre de pages défaillantes, malgré l'ECC et la meilleure méthode de read retry, est enregistré pour chaque ZM et pour chaque DUT. Une fois par jour, l'ensemble de la mémoire est relu et comparé avec son contenu initial. Cela permet ainsi d'avoir 50% de la mémoire lue toutes les 45 minutes et 50% relue toutes les 24 heures. Toutes ces données sont ensuite récupérées à intervalle régulier par l'IHM (interface home machine écrite en programme LabVIEW).

La répartition des DUT par condition pour cette file est dans le Tableau 10.

NAND4 (dyn)	T°C vieillissement		
T° programmation	**+25°C**	**+85°C**	**+110°C**
-40°C	1	2	2
+25°C	1	2	2
+85°C	1	2	2

Tableau 10 : Nombre de mémoires par condition de vieillissement, NAND4

2.4.5 NAND 5 : Endurance en fonction de la température d'activation et du nombre d'écriture/effacement

L'objectif de cet essai est d'étudier l'influence du nombre de cycles de programmation / effacement et de la température d'activation sur l'endurance de la mémoire. Le but est ici non

pas d'estimer la rétention de donnée, mais uniquement sa capacité à l'écrire correctement malgré l'usure du point mémoire.

Lecture → Effacement → Ecriture

L'ensemble du processus de vieillissement de cette file est réalisé sur le banc de test générique décrit dans la section 2.4.4 précédente. Seul le « bitstream » (configuration du FPGA) de test a été modifié. Ainsi, les mémoires Flash NAND sont ici testés vis-à-vis de leur endurance aux cycles de PE. Pour cela, les DUT subissent continuellement la séquence suivante : Lecture → Effacement → Ecriture → Attente. Trois durées d'attente ont été implémentées afin d'évaluer l'influence de la fréquence de cyclage PE sur la fiabilité. La gamme de température d'activation utilisée pour cette file est également de 25°C à 110°C. De plus cinq différents motifs de donnés sont utilisés (voir Tableau 9). Chaque motif est répété sur 600 blocs consécutifs.

La répartition des DUT par condition pour cette file est dans le Tableau 11.

NAND5 (dyn)	T°C vieillissement		
Durée cycle P/E	+25°C	+85°C	+110°C
3 heures	1	2	2
6 heures	1	2	2
12 heures	1	2	2

Tableau 11 : Nombre de mémoires par condition de vieillissement, NAND5

2.5 Résultats du vieillissement des mémoires Flash NAND

Les plans de test décrits précédemment ont été mis en œuvre sur une période de plus de 19 000 heures (plus de 2 ans) de vieillissement. Dans cette section nous étudierons - à partir des résultats obtenus - l'action de chaque facteur de stress sur la fiabilité en rétention ou en endurance.

2.5.1 Rétention de données

La perte de données lors des tests de rétention peut être considérée comme une défaillance dans la mesure où la mémoire ne peut plus remplir sa fonction de restitution de l'information enregistrée. Pour étudier cette rétention, on s'intéressera aux conditions vues par le point mémoire au moment de son écriture (usure initiale du point) et pendant son vieillissement

(condition de stockage ou d'activation).
Deux méthodes vont nous permettre d'observer le vieillissement :

- L'évolution du nombre de pages illisibles (malgré l'ECC et le read retry)
- L'évolution du nombre de read retry nécessaire pour lire la donnée

Une page est déclarée défaillante si aucune méthode de Read-retry et aucun ECC ne permet de relire la donnée écrite.

2.5.1.1 Influence des facteurs de stress

2.5.1.1.1 Température en fonctionnement ou en stockage

La température de vieillissement est le facteur de stress le plus fréquent. La plupart des qualifications HTOL (« High Temperature Operating Life ») se basent uniquement sur du vieillissement accéléré par la température pendant rarement plus de 2000 heures.
Dans notre cas, la température est bien un facteur aggravant de la perte de donnée. Plus la température est élevée, plus la rétention est mauvaise. Ce constat est vrai quelle que soit l'usure initiale ou la température de programmation. En vieillissement sans polarisation plusieurs températures d'écriture ont été testées. Seules les mémoires à la plus haute température de stockage (125°C) avec une programmation à -40°C ont présenté des pages illisibles au cours du vieillissement.

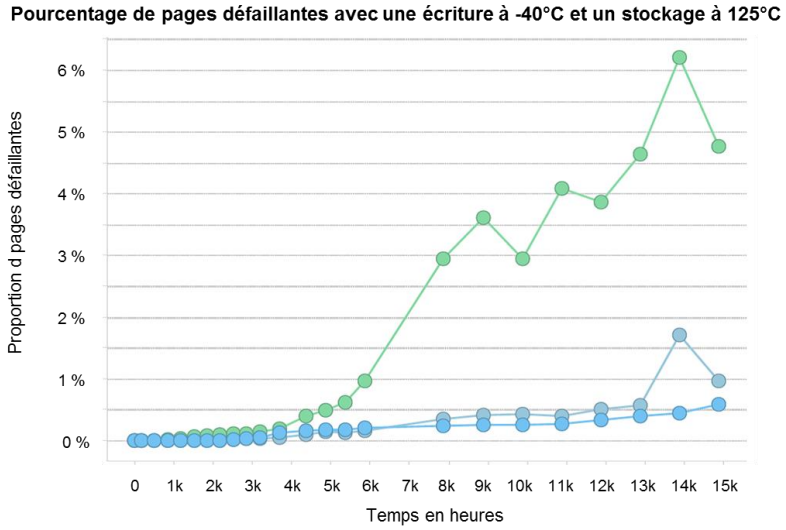


Figure 46 : Evolution du nombre de pages défaillantes en fonction du temps et de la température pour les 3 mémoires écrites à -40°C et stockée à 125°C

La Figure 46 présente la proportion de pages défaillantes (toutes zones motif mémoire confondues) pour les DUT en erreur de la file NAND1. On constate premièrement que les

trois mémoires ont un comportement globalement similaire, même si une (en vert) sort nettement du lot. Cette variation peut s'expliquer par une dispersion technologique entre les pièces. Une autre hypothèse serait qu'une température d'écriture soit légèrement différente des deux autres, comme par exemple une position du composant lors de l'écriture initiale en coin de la carte de test (la tête thermique projette un flux azote à -40°C au centre de la carte).

Cependant, des dégradations sont tout de même mesurables. En effet, afin de lire correctement les données écrites initialement sans altération, un certain nombre de Read retry est nécessaire. Ces différentes méthodes de Read retry modifient légèrement la référence de tension utilisée en interne lors de la lecture afin de suivre les dérives des tensions de seuil au cours du temps. Ainsi, une dérive du nombre de Read retry utilisé (de 1 à 8) permet de suivre une dérive de la tension de seuil établie initialement. Cependant, nous ne savons pas précisément à quoi correspondent ces méthodes puisqu'elles sont propres à l'architecture du fabricant. Nous savons tout de même que pour les composants considérés, les 4 premières méthodes correspondent à une modification de la tension de référence utilisée en lecture, et que les 4 autres utilisent de plus une augmentation des temps d'intégration lors de la lecture.

La Figure 47 présente cette évolution du nombre de Read retry nécessaire pour lire correctement les données lisibles au cours du vieillissement. Chaque couleur correspond à une température de stockage. Chaque ligne correspond à un DUT. Chaque sous-graphique correspond à une température d'écriture des données en initial.
On observe très nettement que plus la température de vieillissement est élevée, plus les dérives sont grandes. Du point de vue de la rétention, il est tout à fait cohérent que la fuite de charge des grilles flottantes des points mémoires soit aggravée par la température. Ce résultat est en accord avec la littérature citée en 1.3.5.2.

Plus le nombre de Read Retry nécessaire pour lire une donnée correcte est élevé, plus le temps de lecture total sera long. En effet, chaque lecture avec les méthodes de Read retry de 1 à 4 prend 100 µs et les méthodes 5 à 8 prennent quant à elles 285 µs chacune. Ainsi le temps de lecture peut varier d'un facteur 10 car toutes les méthodes doivent être tentées à la suite. Notons que le temps de lecture à l'état initial est de 101,64 µs.

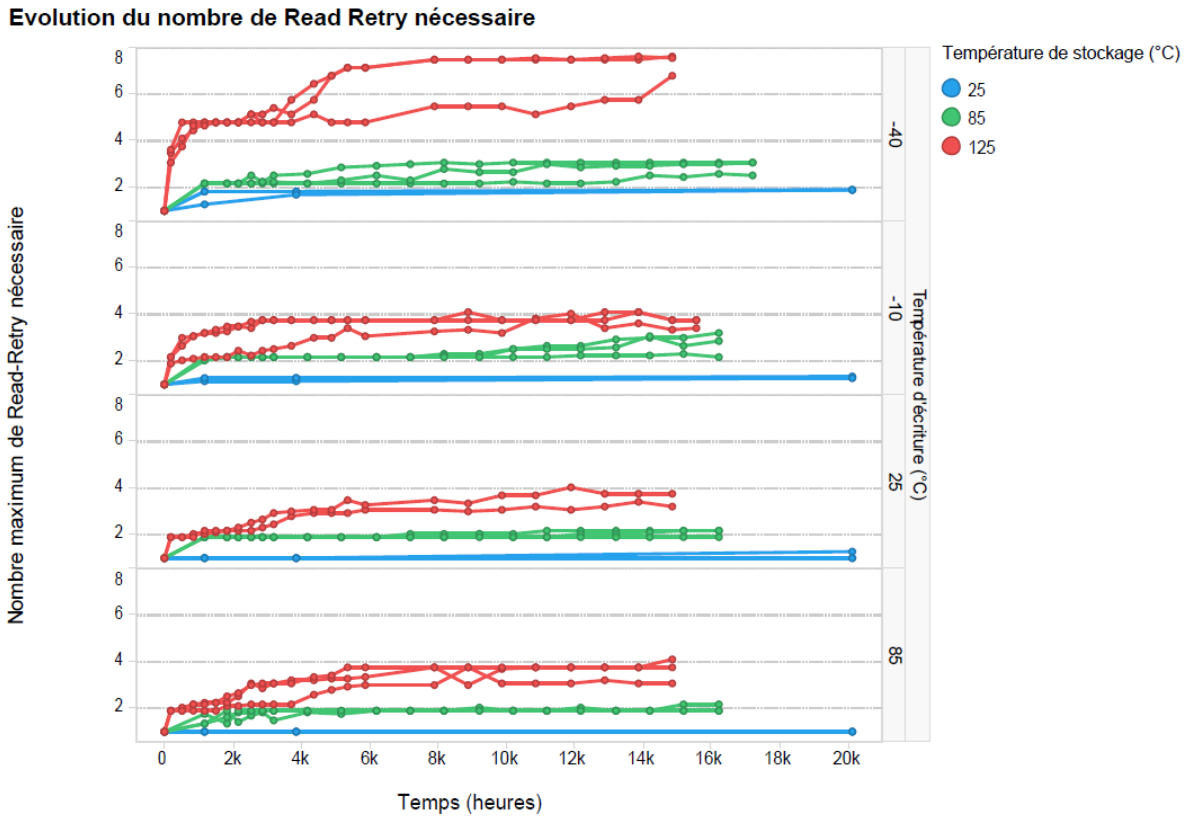


Figure 47 : Evolution du nombre de Read retry pour les mémoires écrites à différentes températures et vieillies sans polarisation

La Figure 48 présente l'augmentation des temps de lecture normalisés par rapport au temps de référence (celui à t_0 de 101,64 µs). Ainsi par exemple, après 12 000 heures à 125°C avec une programmation à -10°C, le temps de lecture sera multiplié en moyenne par 4.

Comme expliqué dans certaines publications sur le sujet (voir 1.3.5.1), un grand nombre de PE avant la phase de vieillissement permet de considérablement augmenter le nombre de défaillances. La Figure 49 trace le nombre d'erreur pour les mémoires de la file d'essai NAND3 pour plusieurs températures. Quel que soit le nombre de PE, la température reste un facteur aggravant sur le nombre de pages défaillantes. Les mémoires stockées à 25°C ne présentent pas de défaillance (en rétention) malgré les 15 000 heures de vieillissement, celles à 85°C présentent des erreurs à partir de 1500 PE en initial, et enfin celles à 125°C présentent le plus d'erreur.

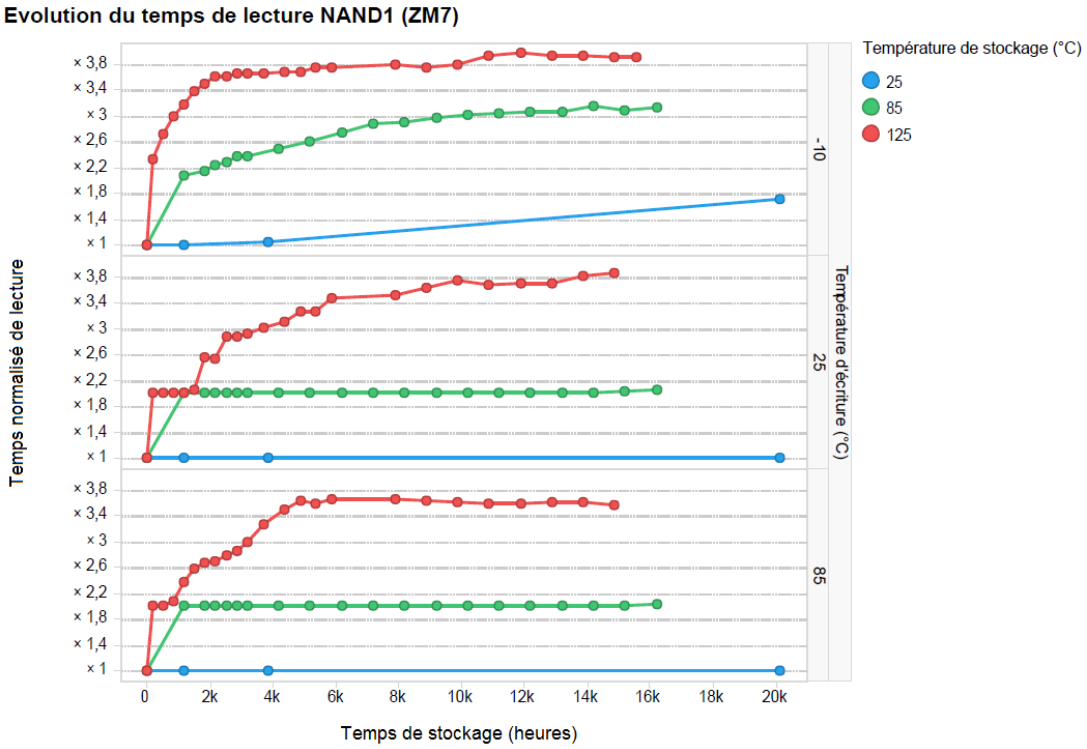


Figure 48 : Evolution du temps normalisé de lecture en fonction du temps de stockage

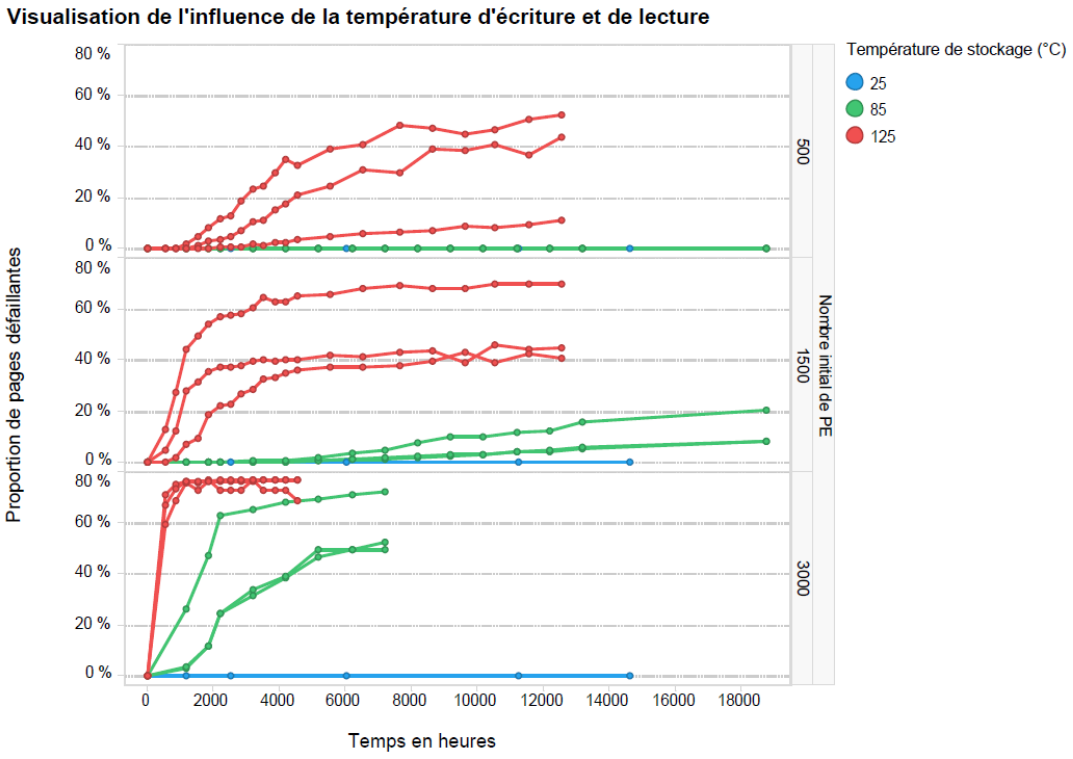


Figure 49 : Evolution du nombre d'erreurs pour les mémoires ayant subies un grand nombre de PE initiale avant écriture

Il en est de même pour les mémoires de la file d'essai NAND4 qui sont quant à elle vieillies sur un banc d'activation dynamique (avec une lecture toutes les 45 minutes).

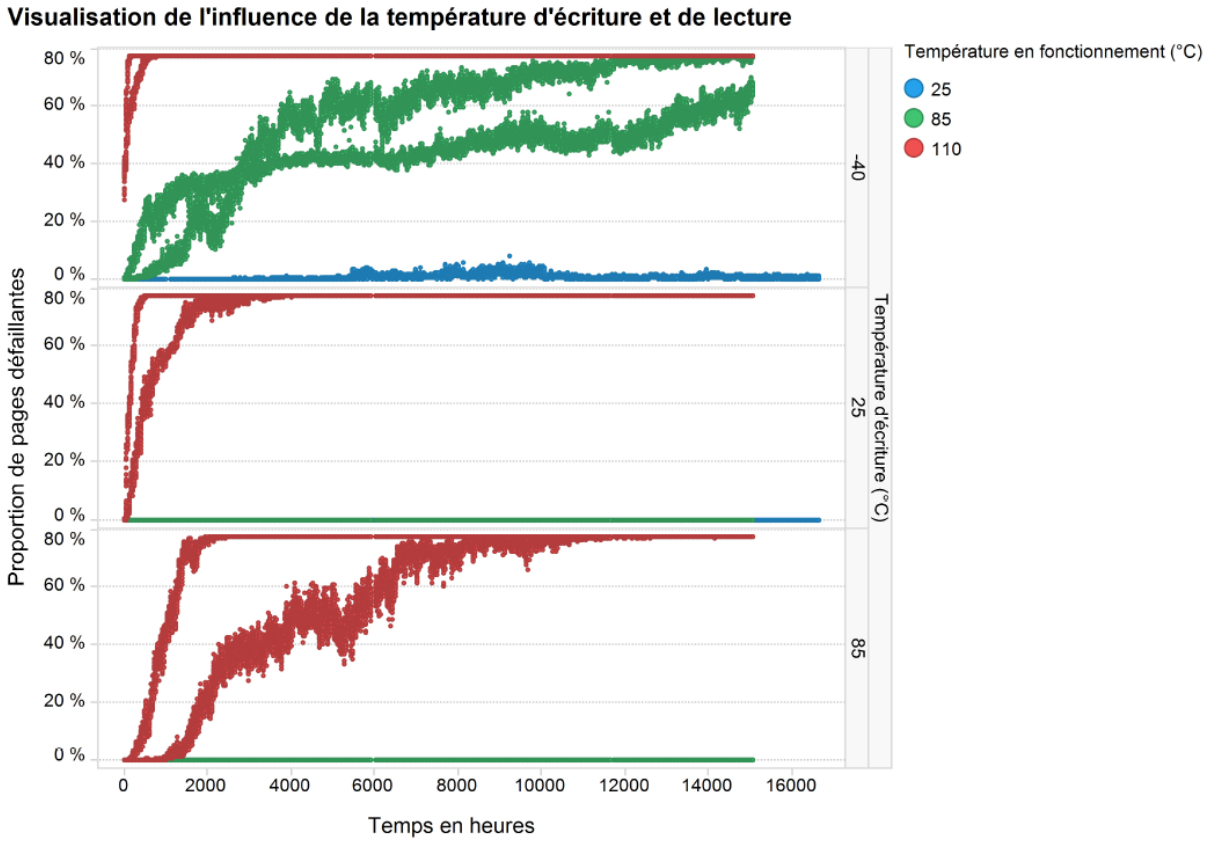


Figure 50 : Evolution du nombre d'erreur des mémoires du banc de vieillissement dynamique en fonction de la température

Notons que dans la pire condition de la Figure 50, c.-à-d. écriture à -40°C et fonctionnement à 110°C, la proportion de pages défaillantes est limitée à 80%. Ce résultat est cohérent car 20% de la mémoire est composé par la ZM1 qui correspond à un contenu effacé composé uniquement de '11'.

2.5.1.1.2 Température programmation

L'étude de l'influence de la température de la mémoire au moment de son écriture sur la fiabilité en rétention est l'une des particularités de ce mémoire. En effet, comme précisé dans l'état de l'art, au moment de la mise en place des essais, aucun papier ne traitait finement ce point.

Pour rappel, dans le cadre de cette étude, différentes températures d'écriture ont été considérées, allant de -40°C à 85°C. La Figure 51 trace l'évolution du nombre de Read retry

nécessaires en fonction de la température de stockage (ce qui correspond à chaque sous-graphique) et de la température d'écriture (qui correspond à une couleur) pour la file NAND 1.

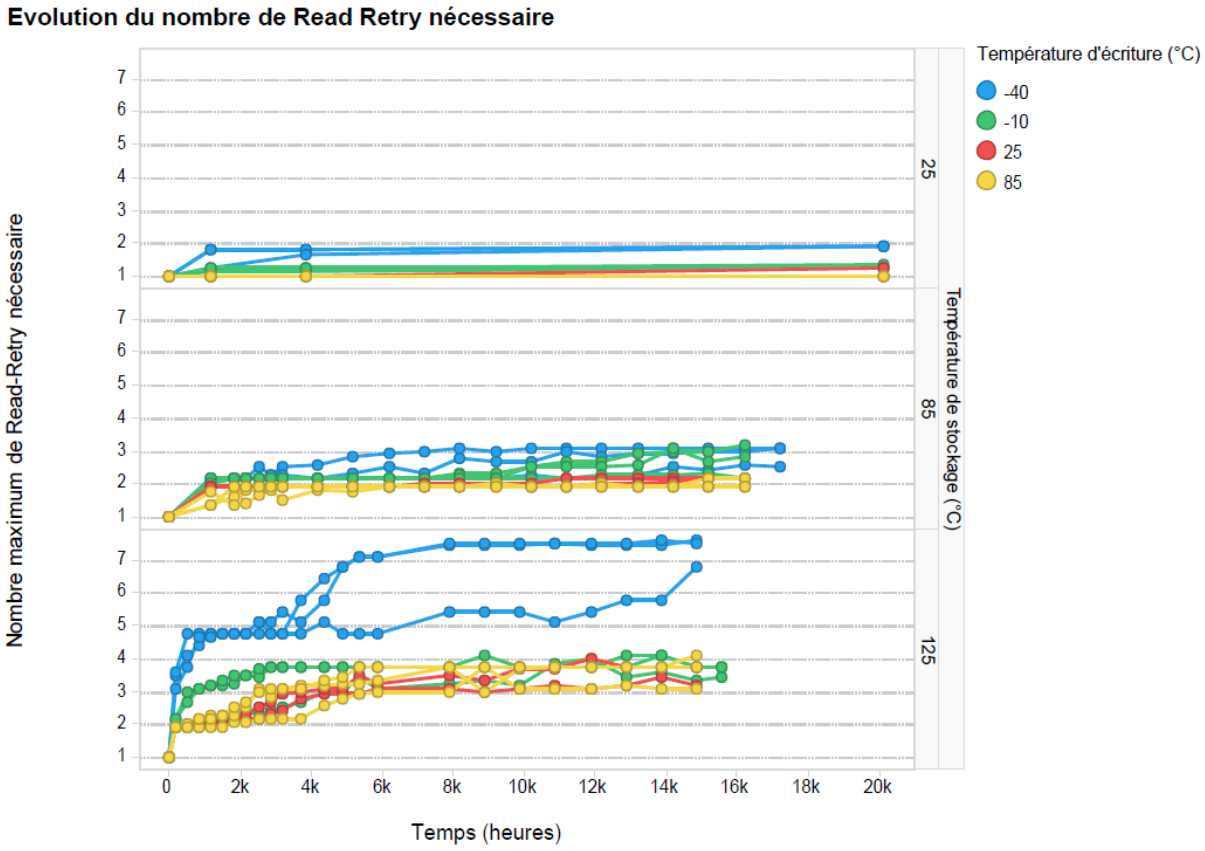


Figure 51 : Impact de la température d'écriture sur le nombre de Read retry nécessaire des mémoires en vieillissement sans polarisation NAND 1

Pour chaque température de vieillissement, les mémoires présentant la plus mauvaise rétention sont toujours celles ayant subies une écriture à -40°C. On constate le même phénomène sur les DUT du banc d'activation dynamique NAND 4 (vieillissement polarisé avec des lectures régulières). La Figure 52 présente le nombre de défaillance en rétention des mémoires NAND vieillies en dynamique.

Sur cette file d'essai, nous avons observé des pages défaillantes dans plusieurs conditions de stress. Ce banc de vieillissement n'enregistre malheureusement pas le nombre de Read retry nécessaire. Le nombre de défaillance est ici bien plus conséquent que celui de la file de vieillissement sans polarisation en étuve. Toutes les mémoires activées à 110°C présentent presque 100% de défaillance après 1 an (~ 9000 h) pour toutes les températures d'écriture. Les 20% de pages encore lisibles concernent la ZM1 qui est initialement effacée puis laissée

tel quel. Toutes les mémoires écrites à -40°C présentent des défaillances dans les trois températures d'activation.

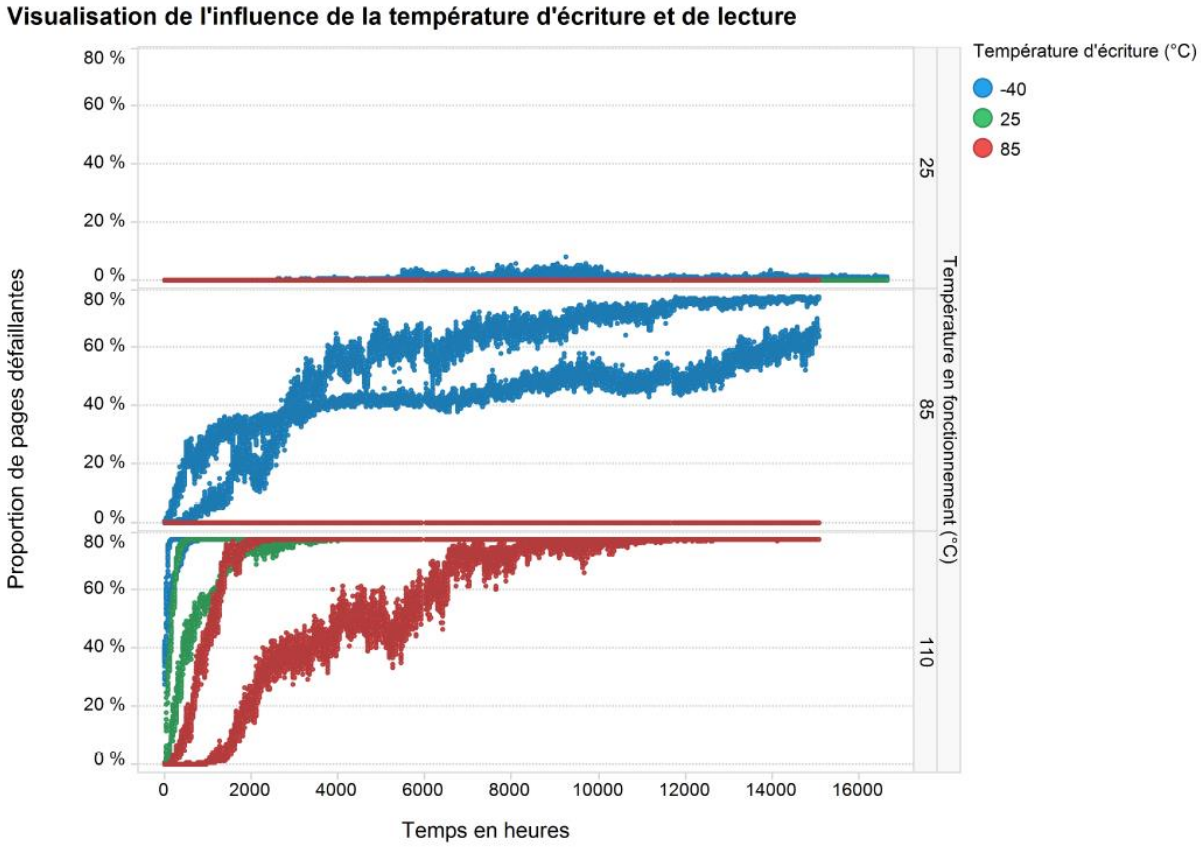


Figure 52 : Impact de la température d'écriture sur le nombre d'erreurs des mémoires en vieillissement dynamique (NAND 4)

Ces résultats sont en accord avec ceux évoqués dans la littérature. La condition la plus critique du point de vue de la fiabilité en rétention est une écriture à basse température et une rétention à haute température. Sous l'effet de la polarisation et des lectures continues à haute température, les défaillances surviennent encore plus tôt. La faible fiabilité en rétention à haute température des mémoires Flash est bien connue et abordée par nombre de publications scientifiques (voir 1.3.5.2). Cela est dû à des fuites de charges hors de la grille flottante sous l'effet de la température. Ces fuites induisent une diminution de la tension de seuil, et donc un changement d'état du point mémoire. Il est donc tout à fait logique que la ZM1 restée vierge ne présente aucune défaillance.

L'influence de la température pendant la phase d'écriture est bien moins connue. A partir de nos résultats, nous proposons une explication dans ce chapitre. La plupart des mémoires NAND Flash utilisent l'effet tunnel de Fowler-Nordheim pour injecter des charges au sein de

la grille flottante pendant l'opération d'écriture. Le diagramme du processus d'écriture est donné en Figure 53. Afin de programmer la bonne tension de seuil après chaque injection de charge, un circuit contrôle si suffisamment de charges ont été injectées. Si non, une autre phase d'injection est faite jusqu'à avoir écrit la bonne donnée.

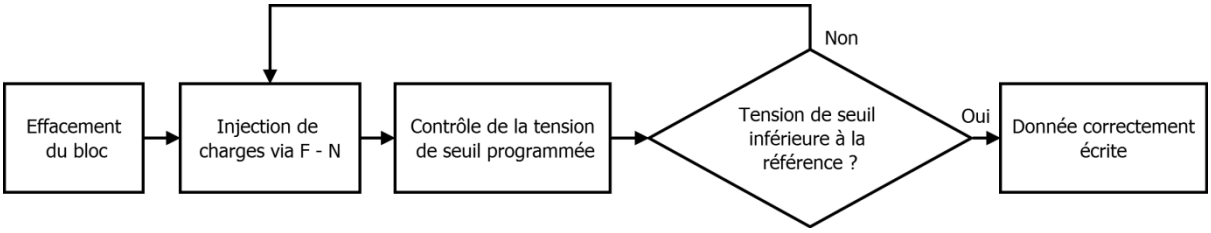


Figure 53 : Diagramme des phases d'écriture dans les mémoires NAND Flash

Si ce circuit de contrôle dérive avec une basse température, alors la donnée écrite sera mauvaise. Cette dérive concerne plus particulièrement les tensions de références utilisées pour délimiter les bornes de chaque état du point mémoire, comme on peut le voir sur la Figure 54. Ainsi, à l'issue des itérations lors de la phase d'écriture, la distribution résultante des tensions de seuil programmées est décalée. Par conséquent à plus haute température les tensions de seuils écrites à plus basse température sont très proches des limites entre deux niveaux adjacents. Ces limites ont été dimensionnées pour une utilisation de la mémoire à température ambiante. De là, la quantité de charge à perdre pour corrompre la donnée est d'autant réduite que la tension de seuil écrite est proche des limites de l'état attendu.

Pour corroborer cette hypothèse, certains DUT qui ont été programmés à -40°C n'avait pas d'erreur juste après leur écriture (relecture immédiate à -40°C pour confirmer que la donnée initiale a été correctement écrite), mais présentait des pages défaillantes dès leur première lecture à 110°C. Plus encore, après 16 000 heures de vieillissement sur le banc d'activation dynamique, 8 mémoires ont été re-testées sur testeur ATE à température ambiante (20°C). Seulement une d'entre elle présentait des pages défaillantes lors de la lecture à 20°C (mais toutes présentaient des dérives de Read retry en cohérence avec la file de vieillissement sans polarisation) alors qu'elles étaient défaillantes à leur température de vieillissement (110°C).

La Figure 55 confirme ainsi que le circuit périphérique ne présente pas un comportement similaire à basse ou à haute température. Lors de l'écriture, il ne détecte aucun problème à cause de sa propre dérive.

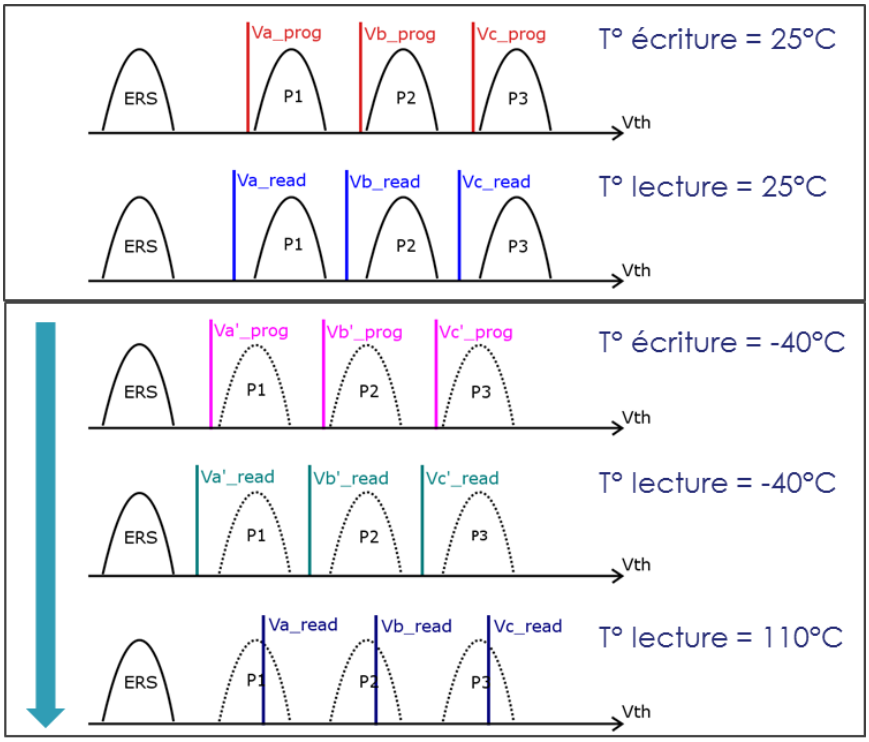


Figure 54 : Effets de la dérive des tensions de référence pendant les opérations d'écriture et de lecture à différentes températures

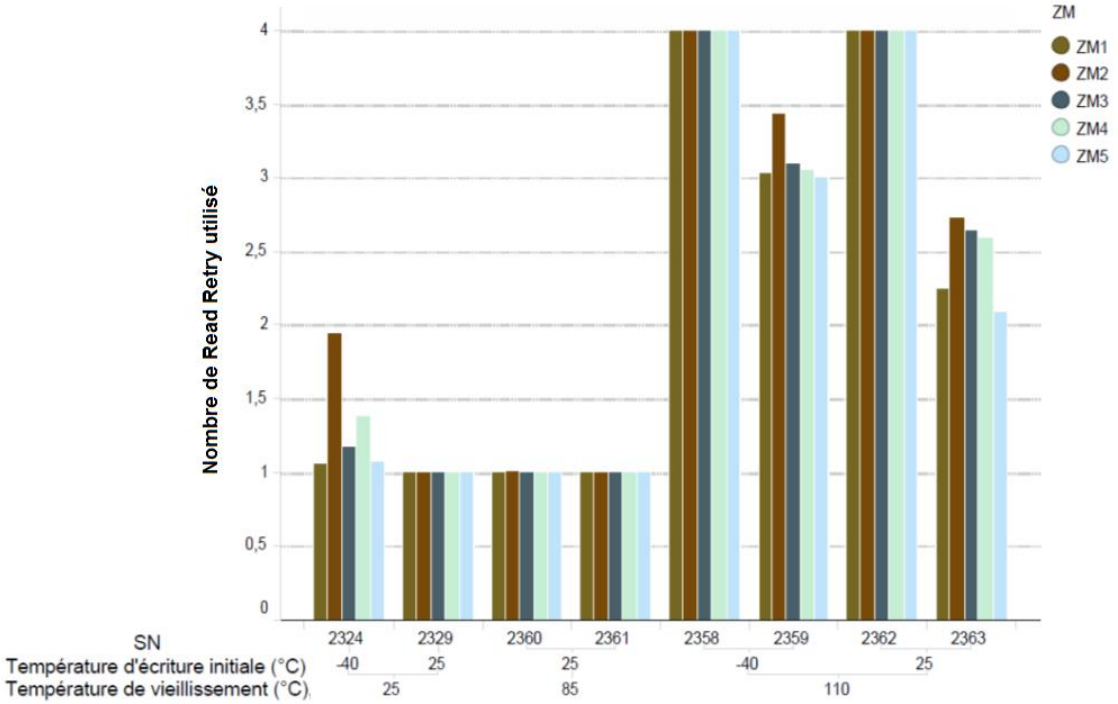


Figure 55 : Nombre moyen de Read Retry nécessaires par DUT lors de la relecture finale sur testeur

En conclusion, la prise en compte de l'influence de la température de programmation dans le plan de test nous a permis de mettre en évidence une limite du composant souvent négligée dans les autres études : la rétention est très mauvaise en présence d'un écart de température important entre la phase d'écriture et la phase de lecture. L'analyse des résultats a montré que ce phénomène n'est pas physique au niveau matériau, mais provient d'une dérive du circuit périphérique dans le composant. La référence de tension utilisée lors de la lecture et de l'écriture est modifiée par la température, ce qui conduit à une programmation « décalée » des tensions de seuil qui sont ainsi très proches des limites. La corruption de données est par conséquent d'autant plus rapide que la marge de bruit est réduite par ce phénomène.

2.5.1.1.3 Fréquence de lecture

Afin d'évaluer les problématiques de Read/Pass Disturb, deux fréquences lectures avaient été introduites dans le banc dynamique NAND 4 : une lecture toutes les 45 minutes et une lecture une fois par jour. Dans les pires conditions de stress (écriture à -40°C et activation à 110°C), la différence est flagrante au bout de 1 000 heures de vieillissement : plus de 95% des pages relues toutes les 45 min sont défaillantes là où aucune ne l'est avec uniquement une lecture par jour. La Figure 56 présente la proportion de pages défaillantes pour ces deux fréquences de lecture en fonction du temps. Elle contient uniquement les mesures jusqu'à 1500 heures - sur les 14 000 h effectuées - car il n'y a plus aucune évolution au-delà de 1000 h pour cette condition de test.

Notons que les instabilités (entre 0 et 1000 h) du nombre d'erreur des pages lues une fois par jour proviennent de sa méthode d'extraction. Notons Q_{45m} la première moitié de la mémoire constituée des pages lues toutes les 45 minutes, et notons Q_{24h} la dernière moitié qui n'est lue qu'une fois par jour. Le banc lit en continu Q_{45m} et enregistre à chaque fois le nombre d'erreurs par DUT et par ZM par comparaison avec la valeur programmée. Une fois par jour, 100% de la mémoire est relue et seul le nombre d'erreurs global - sans distinction parmi Q_{45m} et Q_{24h} - est retenu lors de cette mesure. Ainsi pour obtenir uniquement le nombre d'erreurs de Q_{24h}, on soustrait à la mesure sur l'ensemble des pages la précédente mesure de Q_{45m} (intervenue 45 minutes plus tôt). Le problème de cette méthode est que la fluctuation de la fiabilité des pages Q_{45m} parasite l'extraction, c.-à-d. que le bruit de mesure de Q_{45m} est ajouté aux mesures extraites de Q_{24h}. Ainsi lorsqu'il n'y a pas de défaillance sur Q_{24h}, les seules erreurs résiduelles proviennent de ce bruit.

Aucune défaillance en rétention n'a été observée sur les ZM1 (vierge) et ce malgré les lectures en boucle. Ainsi ces tests n'ont pas réussi à injecter suffisamment de charge par Read Disturb sur les cellules effacées pour les faire changer d'état vers « P1 ».

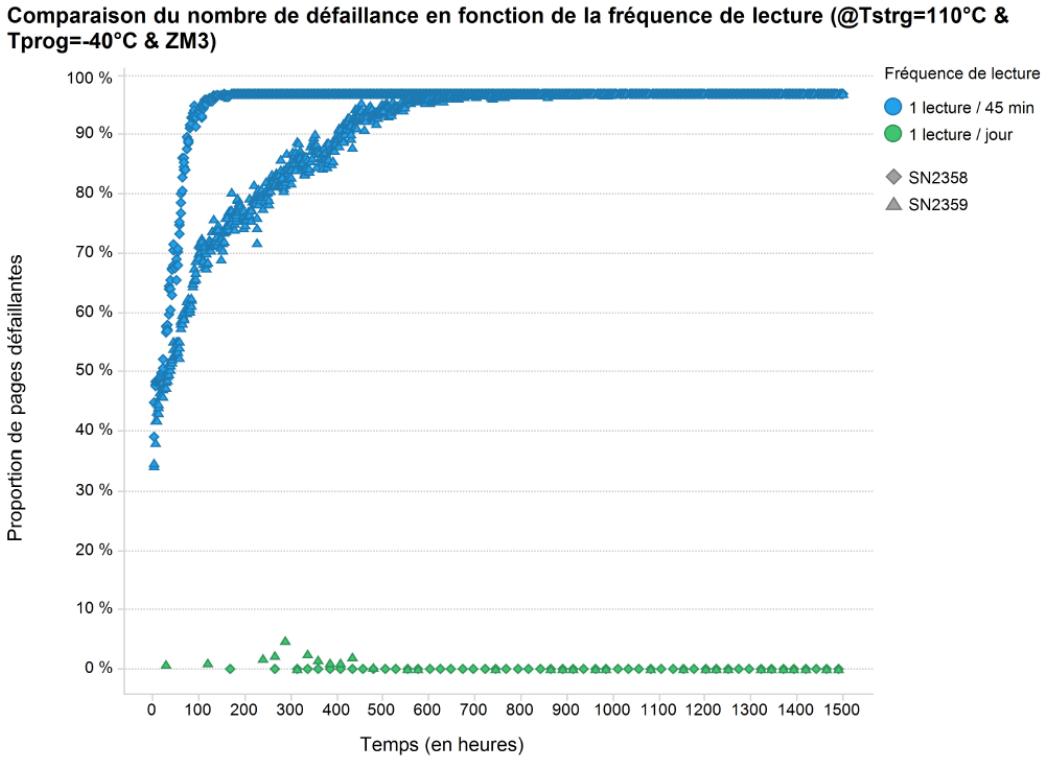


Figure 56 : Influence de la fréquence de lecture sur la rétention, mémoires lues à 110°C et écrites à -40°C entre 0 et 1500 heures

2.5.1.1.4 Motif écrit initialement

Initialement, plusieurs motifs différents de données avaient été écrits afin de voir d'éventuelles différences comportementales, le but étant de solliciter plusieurs niveaux de tension de seuil dans les cellules MLC. A l'issue des 15 000 heures de vieillissement non polarisé, on distingue en effet plusieurs comportements. La Figure 57 présente la proportion moyenne de blocs défaillants (c.-à-d. présentant au moins une page défaillante) des mémoires Flash NAND écrites à -40°C et vieillies à 125°C sans polarisation pendant la phase de vieillissement. Chaque couleur correspond à une zone motif mémoire.

Le motif de données écrit a bien un impact significatif sur la fiabilité en rétention de la mémoire. Les motifs les plus critiques (ZM5, ZM6 et ZM7) sont ceux ayant 25% à 75% de pages à '00' en début de bloc puis les dernières restées vierges. Il vient ensuite la ZM4 (contenant uniquement des '00'). Les ZM 2 et ZM 3 (avec respectivement que des '10' et de '01') ont un comportement également similaires. La ZM 1 restée vierge ne présente aucune erreur comme attendu car ne comportant aucune charge dans la grille flottante.

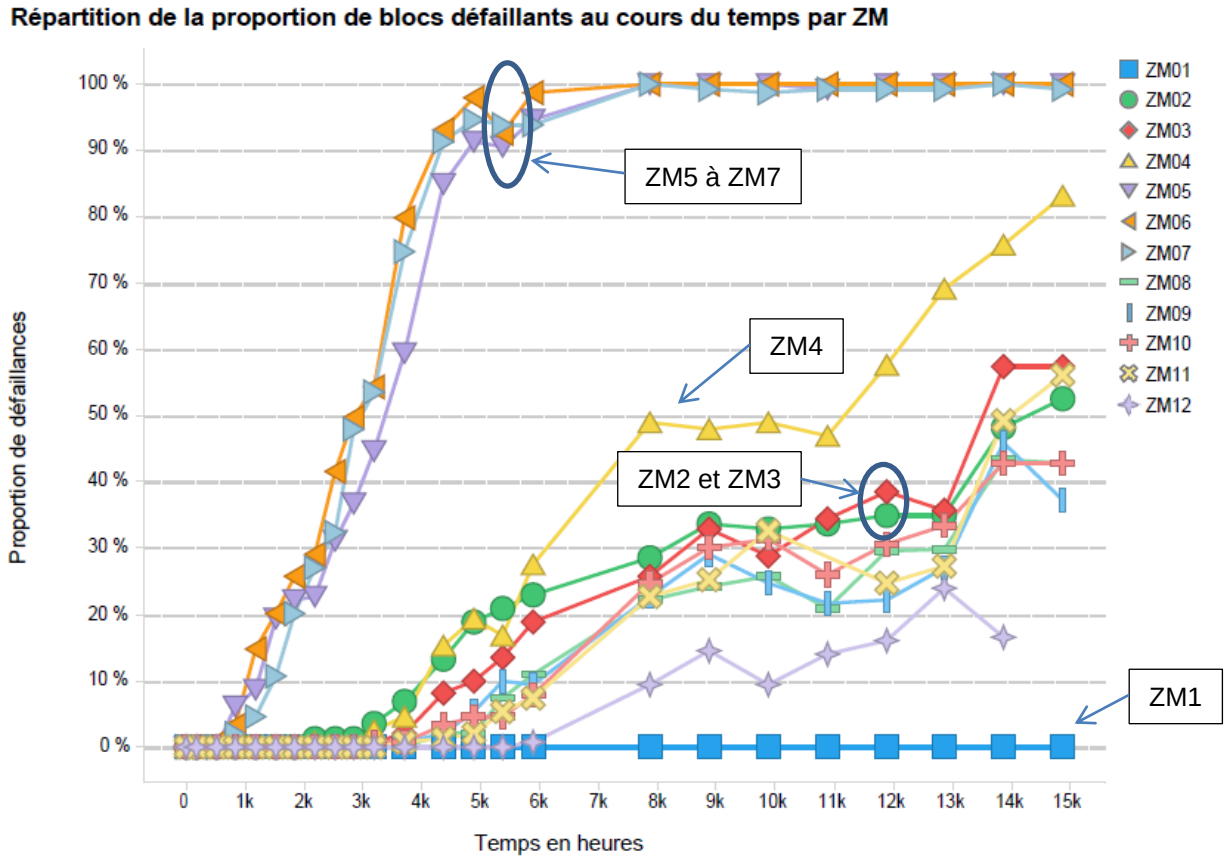


Figure 57 : Proportion de blocs défaillants par ZM au cours du temps pour les mémoires écrites à -40°C et stockées à 125°C, essais sans polarisation

Au lancement des essais, la ZM4 (avec uniquement des '00') était attendue comme étant la plus critique, or ce n'est pas le cas. Les ZM5 à 7 présentent plus d'erreurs que la ZM4, ainsi, le nombre de MLC à '00' n'est pas directement corrélé à la fiabilité du point mémoire. Notons d'ailleurs que l'on considère ici le nombre de blocs défaillant et non le nombre de page car les ZM5 à 7 ne possèdent pas le même motif sur toutes les pages, mais juste le même motif par bloc.

L'analyse de la distribution des numéros de pages défaillantes des ZM5 à 7 souligne un résultat intéressant. Ces zones ne contiennent que des pages à '11' ou '00'. Après le vieillissement, seulement les pages écrites à '00' présentent des erreurs (voir Figure 58). Ce résultat était attendu, cependant pour chaque ZM (5, 6 et 7) la huitième et la seconde page avant l'interface entre les pages '00' et '11' portent à elles seules une grande partie des erreurs.

Dans la ZM4 (toutes les pages à '0'), il n'y a aucune page qui sorte du lot. Le nombre global

d’erreur est bien inférieur aux trois autres ZM : c’est l’interface entre les pages à ‘0’ et celles à ‘1’ qui génère cette baisse de fiabilité.

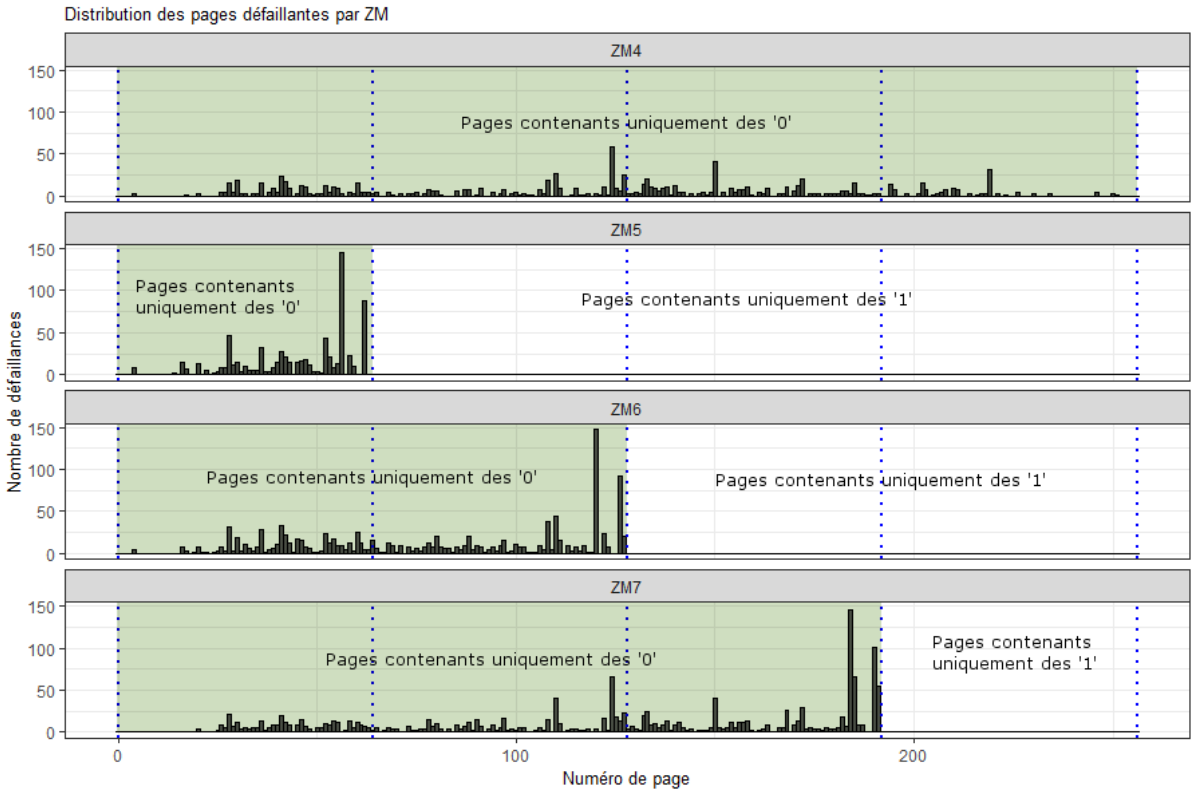


Figure 58 : Distribution des pages défaillantes pour les ZM 4 à 7 à la fin du vieillissement

Cette caractéristique des défaillances nous a amené à étudier plus en détail la notice du fabricant de la mémoire Flash NAND testée. Il est apparu que l’organisation des bits de données n’est pas celle qui avait présidé à l’élaboration des motifs ZM2, ZM3 et ZM5. En effet pour ces motifs nous supposions qu’une MLC stockait deux bits consécutifs d’un même octet (voir Figure 42). Cependant, en parcourant la notice, on peut lire que les deux bits de donnée de chaque MLC sont distribués sur 2 pages NAND et que les pages au sein d’un même bloc doivent être programmées par ordre croissant. Une même word line (ligne de transistor à grille flottante) contient donc deux pages. Un tableau donne la correspondance entre les pages partagées. Ainsi, par exemple, le premier transistor de la première word line contiendrait le premier bit de la page 0 et le premier bit de la page 6, le second transistor contiendrait le second bit de la page 0 et le second bit de la page 6, etc. De plus, il est bien précisé que la page de poids faible doit être écrite nécessairement avant celle de poids fort. C’est pourquoi les pages doivent être écrites par ordre croissant.

Ce partage des pages dans les MLC nous apporte une particularité contenue dans les ZM5, ZM6 et ZM7. Sur le Tableau 12 représentant la ZM7, les pages de 0 à 191 sont programmées (à '000…000'), puis les pages de 192 à 255 sont laissées vierges (à '111…111'). Les deux premières colonnes correspondent aux pages partagées fournies dans la notice.

Pages partagées		**MSB**	**LSB**
6	**0**	0	0
7	**1**	0	0
..	**...**		
190	**184**	0	0
191	**185**	0	0
194	**188**	**1**	**0**
195	**189**	**1**	**0**
198	**192**	1	1
199	**193**	1	1
202	**196**	1	1
...	**...**		

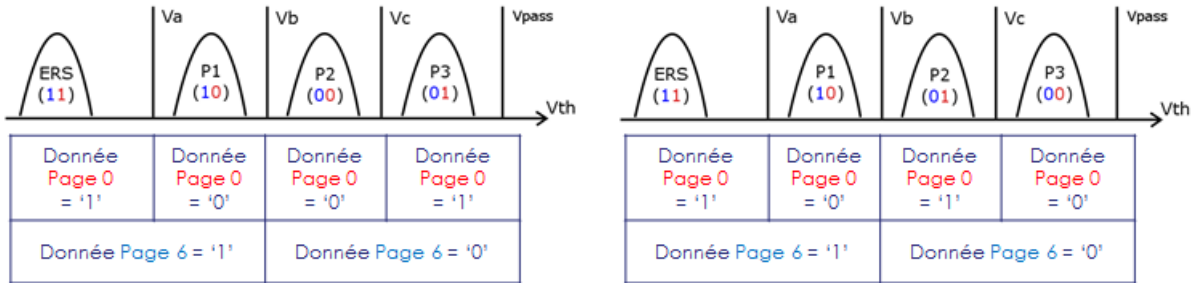
Tableau 12 : Extrait des pages partagées de la ZM7

Deux word lines (celle contenant les pages 188 et 194 ; ainsi que celle contenant les pages 189 et 195) sont différentes des autres. Ces cellules mémoires contiennent l'information '10' au lieu de '11' ou '00' comme toutes les autres. Cependant, les deux pages (visibles sur la Figure 58) présentant le plus d'erreur sont les pages 184 et 190. Ces deux pages partagent un même point mémoire. La raison de ce comportement nous échappe, cependant il doit être lié à l'architecture interne de la mémoire. Les pages stockées sur les deux MLC à '10' (les pages 188 et 189 pour le cas de la ZM7) ne présentent pas d'erreur.

La ZM4 n'est composée que des pages à '0', donc toutes les MLC des blocs de cette ZM sont à l'état « P2 ». Les ZM2 et ZM3 ne contiennent que des pages à '10' ou à '01', donc 50% des MLC de ces blocs sont à l'état « ERS » et 50% sont à l'état « P2 ». Ceci explique la raison pour laquelle la ZM4 est légèrement plus critique que les ZM2 et 3.

A partir de ce mode de fonctionnement en pages partagées et des résultats observés dans nos essais, on peut faire quelques hypothèses quant à l'échelonnement des niveaux logiques implémentés dans ces MLC. Cet échelonnement doit permettre une programmation de la page de poids faible puis la page de poids fort sans effacement, c.-à-d. uniquement en

injectant des charges dans la grille flottante. Il y a plusieurs solutions de distribution possibles : on peut utiliser un codage de Gray (Figure 59.a), ou bien utiliser un codage binaire naturel (Figure 59.b).



(a) Code de Gray (b) Codage binaire

Figure 59 : Distributions possibles des pages partagées sur les MLC. La sous figure (a) présente le codage de Gray, la (b) le codage binaire naturel.

La Figure 59 explique ce fonctionnement. Dans le cas du code de Gray les états « ERS » et « P1 » correspondent au cas où la page LSB (page 0 dans l'exemple) est à '1', et les états « P2 » et « P3 » au cas où la page est à '0'. Le bit MSB (de la page 6 de l'exemple) est ensuite défini parmi les deux sous-états. Comme décrit dans la publication de Liu et Zou [115], cette distribution en codage Gray permet de programmer les pages en deux séquences :

- On écrit le LSB (page 0). S'il est égal à '0', on programme la MLC à « P1 », sinon on ne fait rien.
- On écrit ensuite le MSB (page 6). S'il est égal à '0', on programme la MLC à « P2 » ou « P3 » (suivant la valeur du bit [0]), sinon on ne fait rien. Dans tous les cas à ce stade, on n'a pas besoin d'effacer la MLC pour écrire ce bit.

Notons que l'écriture d'un '01' nécessitera un temps de programmation relativement plus grand lors de la phase 2 (écriture de la page de poids fort) car il faudra passer de l'état « ERS » à l'état « P3 ».

Dans le cas d'un codage binaire (Figure 59.b), les significations des états « P2 » et « P3 » sont inversées par rapport au codage de Gray présenté. Ce codage permet également de programmer les pages en deux séquences. La Figure 60 illustre le cas où l'on écrit un '00' dans une MLC en codage binaire. Cet autre codage plus naturel a ici l'avantage de proposer des temps d'écriture plus réguliers. En effet, quelle que soit la donnée à écrire, on n'augmente jamais de plus de deux états par passe (écriture d'une page partagée) contrairement au codage de Gray. Cependant, lors du passage de l'état « P2 » à « P1 » au

cours de la rétention (fuite de charges), on perd deux bits d'un seul coup là où en code de Gray on n'en perd qu'un seul.

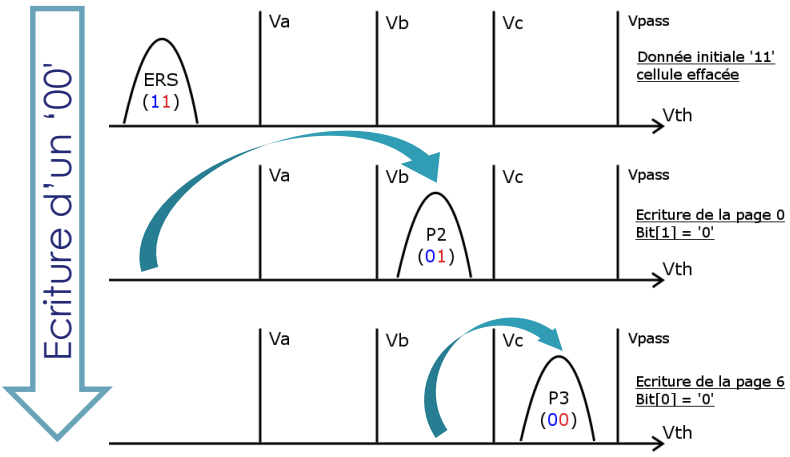


Figure 60 : Processus d'écriture de '00' dans une MLC avec un codage binaire

Lors de nos essais, nous avons observé les cas suivants:

- Une défaillance sur la page LSB alors que la page MSB était encore correcte, ainsi la donnée initialement écrite ('00') est devenue '01' sur la MLC. Il s'agit du cas apparaissant le plus souvent et le plus rapidement.
- Une défaillance sur la page MSB alors que la page LSB était encore correcte, ainsi la donnée initialement écrite ('00') est devenue '10' sur la MLC. Ce cas minoritaire apparait après bon nombre d'heure de test.
- Les deux pages (MSB et LSB) partagées sont illisibles. La MLC est donc passée à '11'.

Avec un codage de Gray, en fuite de charge, on n'aurait jamais pu avoir le cas où une MLC passe de l'état « P2 » ('00') à l'état « P3 » ('01'). Un codage binaire semble être dans notre cas le plus cohérent. En effet, le passage de « P3 » ('01') à « P2 » ('00') est bien le plus probable.

Les résultats du banc de vieillissement dynamique sont également expliqués par ces pages partagées. Pour rappel seules 5 ZM ont été implantées sur ces bancs :

- ZM 1 : contenu vierge
- ZM 2 : ~~tout à '10'~~ → la moitié des MLC à 11, l'autre moitié des MLC à 00
- ZM 3 : ~~tout à '01'~~ → la moitié des MLC à 00, l'autre moitié des MLC à 11
- ZM 4 : tout à '00',
- ZM 5 : Séquence aléatoire d'une longueur de 8 octets. Donc des MLC à 00 et des MLC à 11.

La Figure 61 présente l'évolution du pourcentage de pages défaillantes pour toutes les mémoires écrites à -40°C et activées à 85°C ou 110°C en fonction du temps. Le pourcentage en ordonné correspond à la totalité de la mémoire testée (chaque Zone Mémoire intervient dans 20% du calcul).

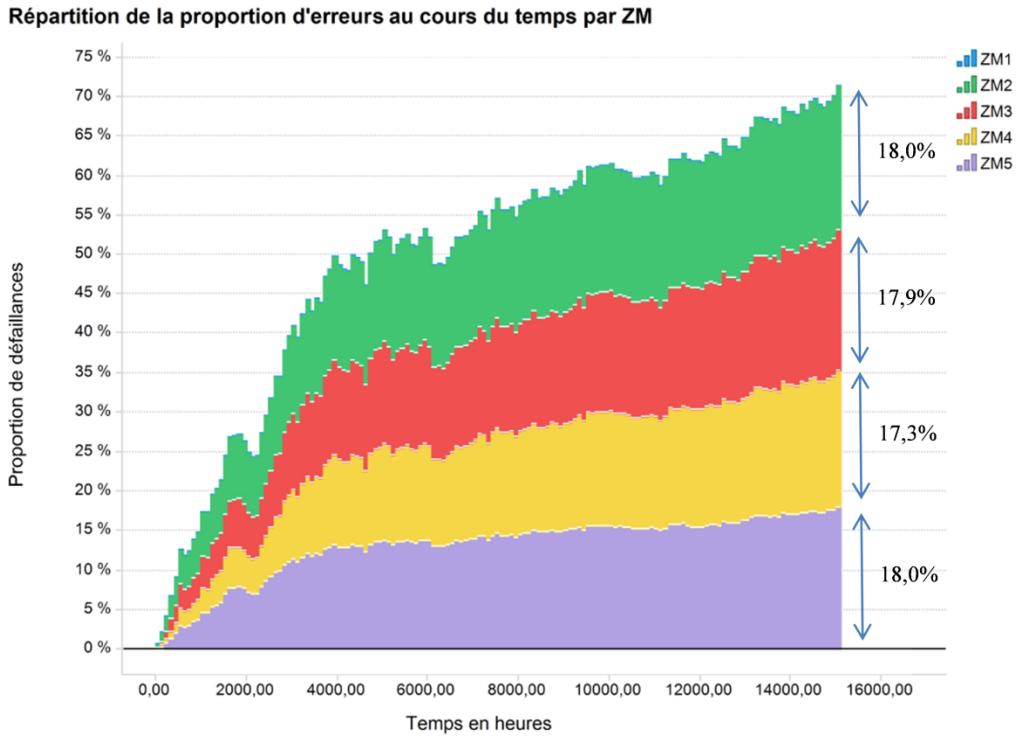


Figure 61 : Répartition du nombre d'erreurs par ZM de la file d'essai dynamique des DUT écrits à -40°C et activés au-dessus de 85°C (NAND 4)

A l'exception de la ZM1 qui ne présente pas d'erreurs, les quatre autres ont un comportement similaire. Ces comportements similaires s'expliquent par le fait que toutes les pages d'un même bloc sont identiques. Ainsi seuls deux états sont implémentés dans les MLC : '00' ou '11'.

2.5.2 Endurance au fil des écritures

L'autre aspect important vis-à-vis de la fiabilité des mémoires est l'endurance - qui qualifie la capacité à écrire correctement une donnée sur un point mémoire après un grand nombre d'écritures. Si le point mémoire est trop usé, il sera alors très difficile (voire même impossible) d'écrire la donnée sans erreurs immédiates. Afin d'évaluer cette endurance, pour rappel, une file d'essai (NAND 5) sur banc dynamique a appliqué un grand nombre de PE, avec relecture « immédiate » pour suivre l'évolution des dégradations.

Trois périodes de cycle « Lecture → Effacement → Ecriture » ont été implémentées : 3 ; 6 et 12 heures. A l'issue des 18 808 heures de vieillissement, la file la plus rapide a atteint les 6 341 PE. Sachant que le fabricant garantit 3000 PE pour cette référence, cette limite est largement dépassée.

Le vieillissement de cette file se faisant à chaque Programme/Effacement, il est cohérent de considérer une échelle horizontale en PE (c.-à-d. « temps » × « fréquence de cyclage ») plutôt qu'en temps. Ainsi les figures de cette section auront comme abscisse le nombre de PE accumulé et non le temps.
Les résultats sont particulièrement cohérents pour cette file, les premières défaillances à haute température interviennent autour de la limite du fabricant. La visualisation Figure 62 trace la proportion des pages en erreur en fonction du nombre de cycles de programmation/effacement. Chaque colonne correspond à une température d'activation. Chaque ligne correspond à un des différents motifs de données programmés (ou Zone Mémoire). Les défaillances commencent à survenir autour des 3000 cycles de programmation/effacement à +110°C, voir même à +85°C.

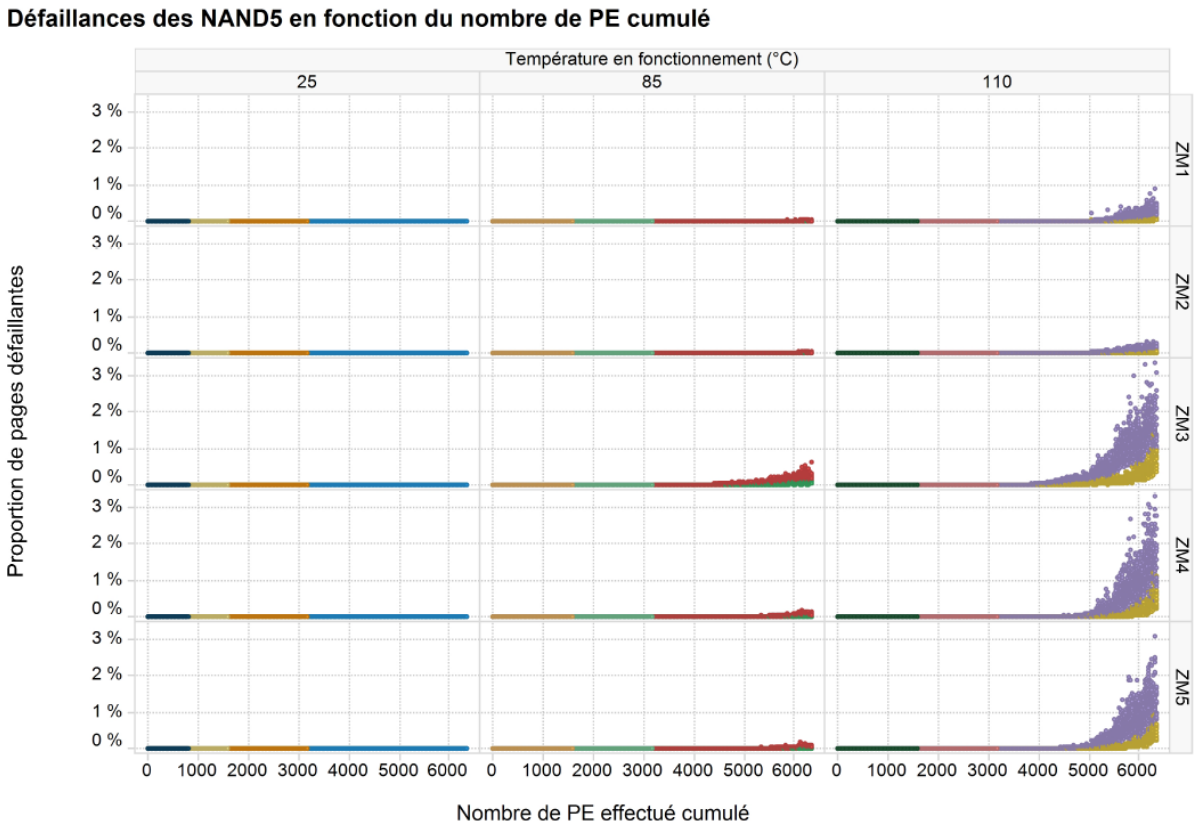


Figure 62 : Evolution des erreurs en fonction du nombre de PE, de la température et des Zones Mémoire

La même visualisation est présentée en échelle logarithmique Figure 63 pour plus de lisibilité sur les petites valeurs.

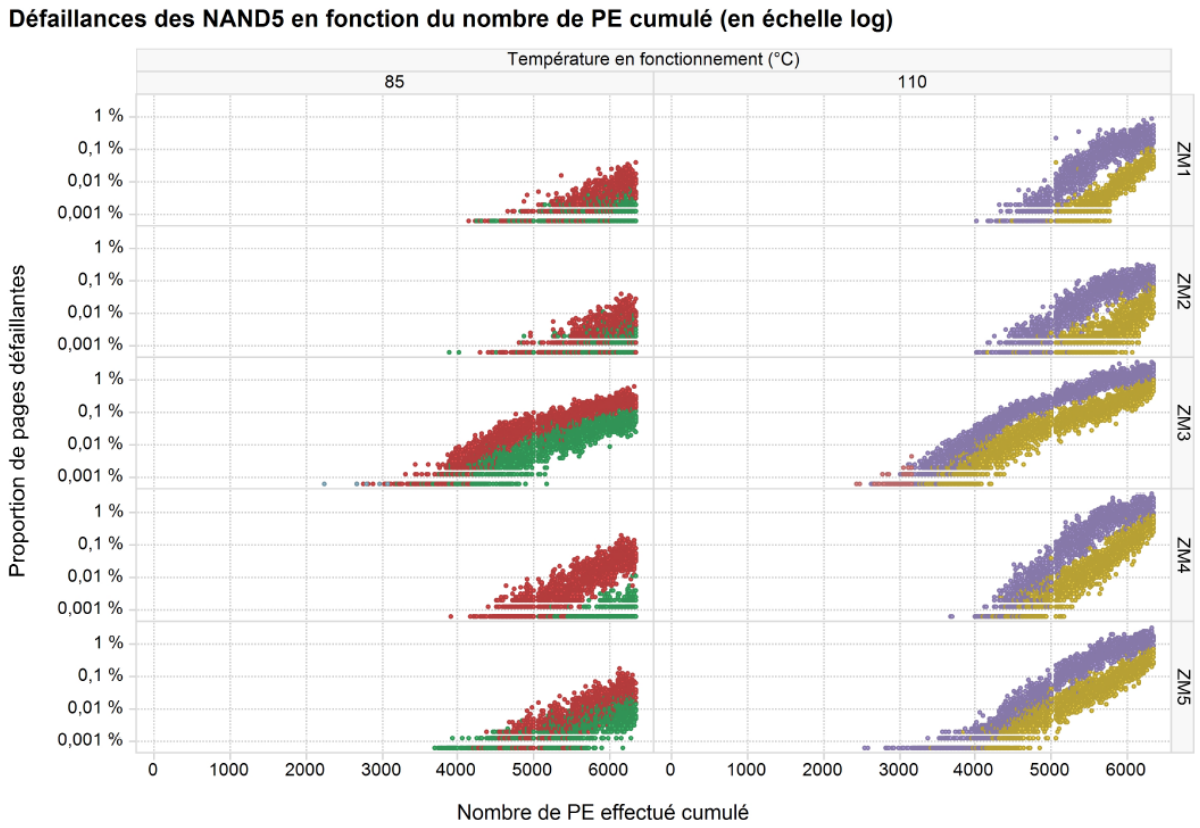


Figure 63 : Evolution des erreurs en fonction du nombre de PE, de la température et des ZM (échelle LOG)

Comme attendu, la température est bien un facteur aggravant. Les mémoires ayant subies des cycles de PE à 25°C ne présentent aucune défaillance, alors que toutes celles activées au-dessus de 85°C commencent à présenter des problèmes entre 3000 et 4000 PE. Toutes les ZM sont impactées de manière similaire. Seul la ZM3 sort légèrement du lot.

2.6 Estimation de la fiabilité des NAND MLC

La section précédente a décrit les essais réalisés ainsi que les résultats associés. Ces observations nous donnent une meilleure compréhension des mécanismes de vieillissement impliqués dans cette technologie. Le second aspect qui est maintenant important d'analyser est la dispersion statistique des instants de défaillance. En effet, la finalité de cette étude est la connaissance de la fiabilité de ce type de composant. La fiabilité des mémoires Flash NAND ne repose plus seulement sur la physique des cellules mémoires, mais également sur

les stratégies des circuits de gestion (codes correcteur d'erreurs, Read retry, surdimensionnement ou « Over-provisionning », uniformisation d'usure ou « Wear leveling »). Ces méthodes permettent de compenser la perte de fiabilité technologique. En effet, le coût unitaire par point mémoire ayant fortement diminué, les fabricants peuvent se permettre de fournir plus de capacité qu'annoncé afin d'uniformiser l'usure sur davantage de cellules. De plus, aujourd'hui sans code correcteur d'erreur, les mémoires MLC ne fonctionneraient pas.

Cette partie du chapitre va présenter la démarche utilisée pour estimer la fiabilité en rétention ou en endurance des mémoires NAND. Une fois la fiabilité d'une page estimée, des méthodes de prise en compte de l'ECC et de la redondance dans le calcul de la fiabilité d'après nos tests de vieillissement seront proposés. Ces méthodes ont pour but de prendre en compte le maximum d'information afin de proposer un modèle pertinent pour ce type de composant. La non prise en compte de ces méthodes d'amélioration conduit à des prévisions faussées car fortement sous-estimée.

2.6.1 Définition du critère de défaillance

Dans l'évaluation de la fiabilité des mémoires Flash NAND MLC, nous avons retenu le critère de défaillance suivant :

- *Une page est déclarée défaillante si aucune méthode de Read-retry et aucun ECC ne permet de relire la donnée écrite*.

Le code correcteur d'erreur (ECC) considéré (conseillé par le fabricant) permet de récupérer 40 bits en erreur tous les 1 117 octets consécutifs. La taille mémoire d'une page d'une Flash NAND est 8 Ko.

2.6.2 Rétention de données

Les essais de rétention de données effectués nous permettent d'estimer la fiabilité en rétention de ces mémoires selon plusieurs facteurs de stress. Le Tableau 13 résume l'ensemble des conditions qui ont été réalisées.

File d'essais	Température vieillissement	Température écriture	Cyclage P/E initial	Fréquence de lecture	Vieillissement polarisé ?
NAND 1	T_{act}	T_{prog}	1	1/{Temps moyen entre deux caractérisations}	Non
NAND 3	T_{act}	25°C	N_{PE}	1/{Temps moyen entre deux caractérisations}	Non
NAND 4	T_{act}	25°C	1	f_{read}	Oui

Tableau 13 : Synthèse des files d'essai NAND avec leurs cofacteurs

2.6.2.1 Modélisation de l'effet des cofacteurs

Cette étude prend en compte quatre cofacteurs : la température d'écriture, la température de lecture, la fréquence de lecture des données et le nombre de cycle de PE enduré par le point mémoire avant écriture de la donnée. Chacun de ces paramètres physiques doit être modélisé avant de considérer l'aspect statistique. En accord avant l'état de l'art section 1.3.5, ces différents facteurs d'accélération seront modélisés par l'équation 61. L'influence de la température d'écriture et de lecture sera modélisée par deux lois d'Arrhenius distinctes. Le nombre de PE avant écriture de la donnée suivra un modèle en loi puissance. La fréquence de lecture sera aussi modélisée en loi puissance.

$$t_{50\%} = e^{\frac{E_a\_a}{k_b \cdot T_{act}}} . (f_{read})^{-\alpha_{read}} . e^{\frac{E_a\_p}{k_b \cdot T_{prog}}} . (N_{PE})^{-\alpha_{PE}} . K \quad (61)$$

Où :

- K est un pré-facteur dépendant de la distribution statistique des instants de défaillance
- T_{act} est la température d'activation en Kelvin
- T_{prog} est la température de programmation en Kelvin
- f_{read} est la fréquence de lecture en lecture / heure
- α_{read} est le paramètre du facteur d'accélération en puissance de la fréquence de lecture
- N_{PE} est le nombre de Program/Erase avant écriture de la donnée
- α_{PE} est le paramètre du facteur d'accélération en puissance du nombre de PE initial
- E_a_a est l'énergie d'activation en activation en eV
- E_a_p est l'énergie d'activation en programmation en eV

Cette modélisation des facteurs d'accélération sera validée par la suite dans la section 2.6.2.3. Les valeurs estimées de ces paramètres seront données dans le Tableau 14.

2.6.2.2 Comparaison des modèles statistiques

La population considérée est composée de 9 568 256 échantillons (pages) répartis sur 109 Flash NAND. De nombreuses conditions de stress sont totalement censurées à droite : malgré les nombreuses heures de stress accéléré, les pages concernées n'ont pas perdu la donnée stockée à la fin du test. Le taux de censure est très variable d'une condition à l'autre. Le taux de censure total est de 84%. Cependant, comme nous avons tout de même 1 565 837 défaillances observées dans plusieurs conditions, une modélisation est envisageable. Le détail de la population est fourni en Annexe A.

Les deux distributions statistiques les plus utilisées pour modéliser ce type de défaillance sont les lois de Weibull et log-normale. La distribution donnant la plus haute vraisemblance – via estimation par la méthode du maximum de vraisemblance - sera retenue. Ce test a été

appliqué séparément sur les différentes files de vieillissement dans un premier temps. Toutes les files en rétention ont montré la même tendance.
La distribution de Weibull donnait -6487713 (exemple de la NAND4) de maximum de log-vraisemblance contre -6293302 pour la distribution log-normale. Si l'on regarde la différence de vraisemblance par échantillon entre les deux, on obtient un écart de probabilité de 4% entre les deux distributions. De plus, visuellement sur la Figure 64 (où le taux de censure est nul), on peut voir que l'alignement des points est sensiblement meilleur dans un papier log-normale que dans un papier de Weibull. Le modèle choisi pour la modélisation des instants de défaillance des pages des mémoires NAND, vis-à-vis de la rétention de donnée, est donc une distribution log-normale.

2.6.2.3 Application du modèle sélectionné

Le modèle sélectionné pour la modélisation des instants de défaillance des pages des mémoires NAND vis-à-vis de la rétention de donnée est une distribution log-normale.
La liste de tous les cofacteurs, ainsi que leurs valeurs par file d'essai sont recensés dans le Tableau 13. La forme du modèle est la présentée équation 62. Le lien avec les facteurs d'accélération définis dans l'équation 61 est donnée dans l'équation 63.

$$F(t) = \Phi\left(\frac{log(t)-\mu}{\sigma}\right) \tag{62}$$

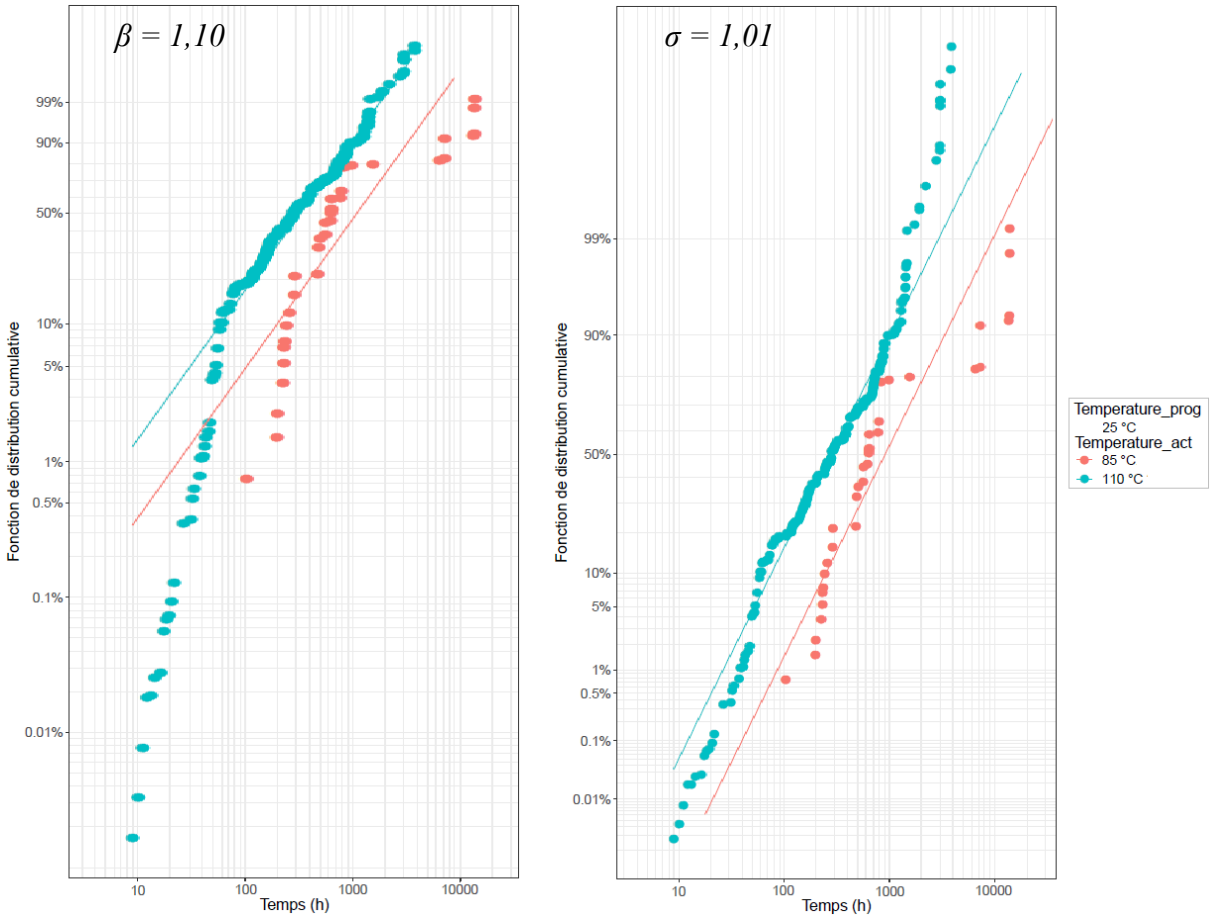
$$\begin{cases} \mu = ln(t_{50\%}) \\ K = \mu_0 \end{cases} \tag{63}$$

Où :

- $F(\,)$ est la probabilité de défaillance
- Φ est la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$
- t est le temps en heure depuis la dernière écriture
- σ est l'écart type du logarithme de t
- μ est le paramètre d'échelle
- μ_0 est le pré-facteur du paramètre d'échelle

Les valeurs de ce modèle estimées par MLE pour nos résultats sont dans le Tableau 14.

La Figure 65 est la fonction de défaillance observée où toutes les conditions de stress ont été ramenées à des conditions opérationnelles (T_j=85°C, T_{prog}=55°C, 1 lecture/mois et 500 PE initiaux) en utilisant les facteurs d'accélération du modèle choisi.



(a) Weibull (b) Log-normale

Figure 64 : Comparaison de la distribution de Weibull (a) et de la distribution log-normale (b) pour les mémoires en rétention pour deux températures d'activation, NAND 4

Paramètre	Inf_5%	Standard	Sup_95%
$exp(\mu_0)$	1,337 × 10^{-8} h	1,364 × 10^{-8} h	1,378 × 10^{-8} h
σ	1,646	1,647	1,648
E_{a_a}	1,022 eV	1,022 eV	1,022 eV
E_{a_p}	-0,1698 eV	-0.1695 eV	-0,1692 eV
α_{read}	1,384	1,385	1,386
α_{PE}	0,9540	0,9552	0,9565

Tableau 14 : Estimation par MLE des paramètres du modèle NAND Rétention (avec intervalles de confiance à 90%)

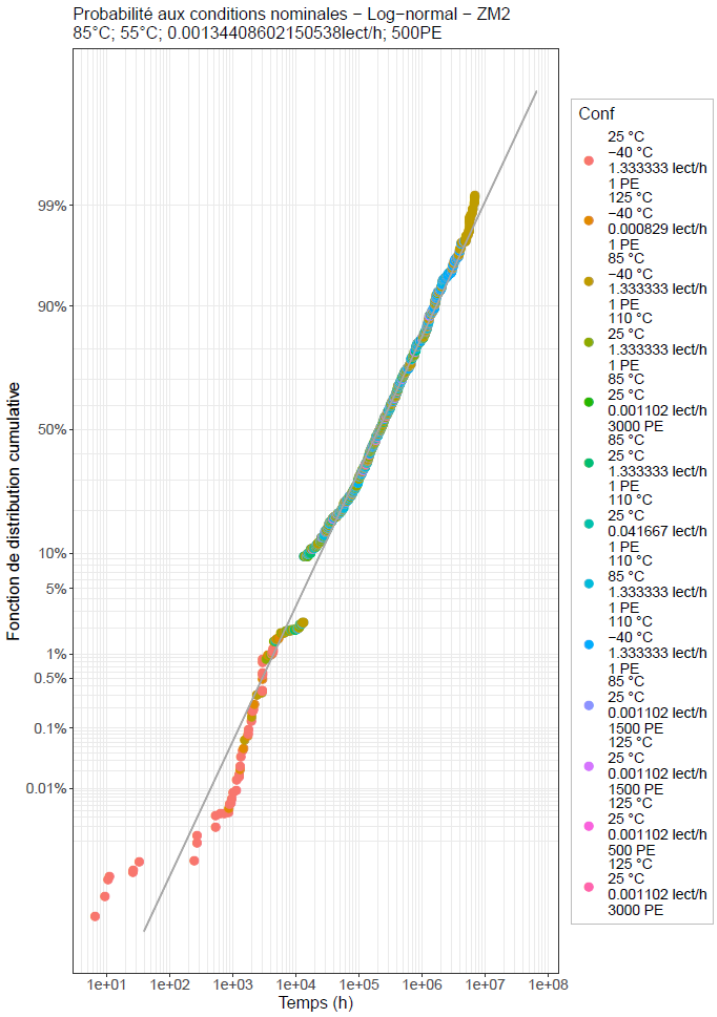


Figure 65 : Fonction de répartition des observations rapportées à la condition T_j=85°C, T_{prog}=55°C, 1 lecture/mois et 500PE initiaux, NAND Rétention

Cette visualisation permet de valider en première approche de manière visuelle et globale l'ensemble des facteurs d'accélération estimés. Ici, à part certaines files d'essais (en rouge) programmées à T_j = -40°C, l'alignement est plutôt satisfaisant.

Concernant les estimations de MTTF, voici les formes analytiques du modèle du temps moyen jusqu'à défaillance :

$$MTTF(T_j) = t_{50\%}.e^{\frac{\sigma^2}{2}} \quad (64)$$

On peut maintenant calculer le MTTF aux conditions {T_j=85°C, T_{prog}=55°C, 1 lecture tous les mois et 500 PE initiaux}. Ces résultats avec leur intervalle de confiance à 90% sont consignés dans le Tableau 15.

	Inf_5%	Standard	Sup_95%
MTTF (années)	720	729	739
$t_{50\%}$ (années)	477	483	489
$t_{10\%}$ (années)	58	59	59
$t_{1\%}$ (années)	10.3	10.5	10.6
$t_{0.1\%}$ (années)	2.9	3.0	3.0

Tableau 15 : Durées de vie projetées à {T_j=55°C, T_{prog}=55°C, 1 lecture par mois et 500 PE initiaux}, pour les NAND vis à vis de la rétention

Notons que la notice du composant garantit jusqu'à 3000 PE, jusqu'à 85°C et pour une rétention de 10 ans. Or il n'est pas fait mention de l'influence de la température de programmation ou du nombre de lectures. On prendra donc une seule lecture pour valider la valeur fabricant. Ainsi en appliquant notre modèle avec les conditions {T_j=85°C, T_{prog}=-40°C, 1 lecture tous les 10 ans et 3000 PE initiaux}, on obtient un $t_{10\%}$ de 66 ans et un $t_{0.1\%}$ de 4.4 ans.

2.6.2.4 Extension à l'utilisation de différents ECC

Comme vu dans la section 1.2.2.2, les codes correcteurs d'erreur font partie intégrante de la technologie Flash NAND. L'augmentation du nombre de bit par transistor à grille flottante a entrainé une perte de fiabilité en rétention telle que l'utilisation d'ECC est maintenant devenue indispensable. Sans eux, les mémoires ne pourraient pas garantir une durée de rétention acceptable. Ainsi le nombre de bits corrigeable par ECC est une donnée très importante à prendre en compte pour estimer de manière plus réaliste la fiabilité en condition opérationnelle. Un meilleur ECC augmentera drastiquement la fiabilité de la mémoire d'un point de vue rétention. C'est pourquoi cette section va présenter une méthodologie pour généraliser nos essais à différents ECC.

Le code correcteur d'erreur considéré (conseillé par le fabricant) lors de nos tests permet de récupérer k_{ECC} = 40 bits en erreur tous les n = 1117 octets = 8936 bits consécutifs. La Figure 66 schématise la structure d'une page. Une page contient $N_B \times n = 8 \times 1117$ octets soit 71 488 bits. Dans la pratique, sur 8936 bits, 8192 bits sont utilisables pour les données « utilisateurs », et 744 bits sont réservés pour le stockage de l'ECC associé.

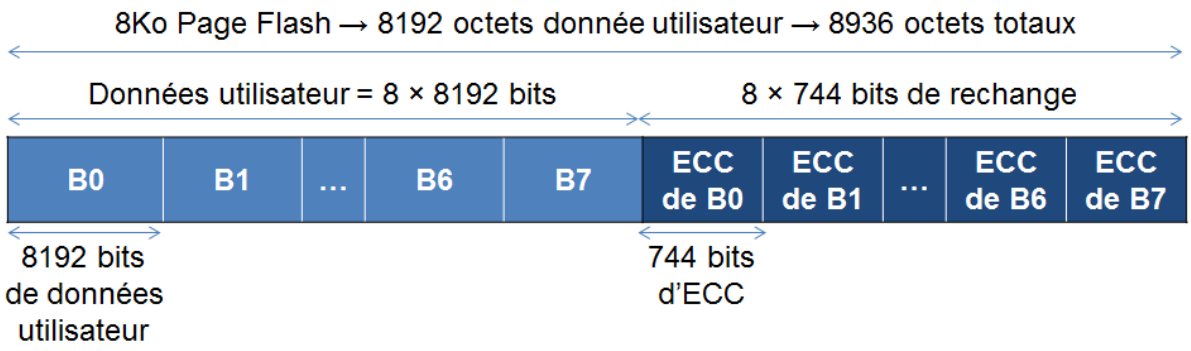


Figure 66 : Structure d'une page FLASH NAND vis-à-vis de l'ECC

Par analogie, on peut considérer la fiabilité d'une page comme le système de la Figure 67.

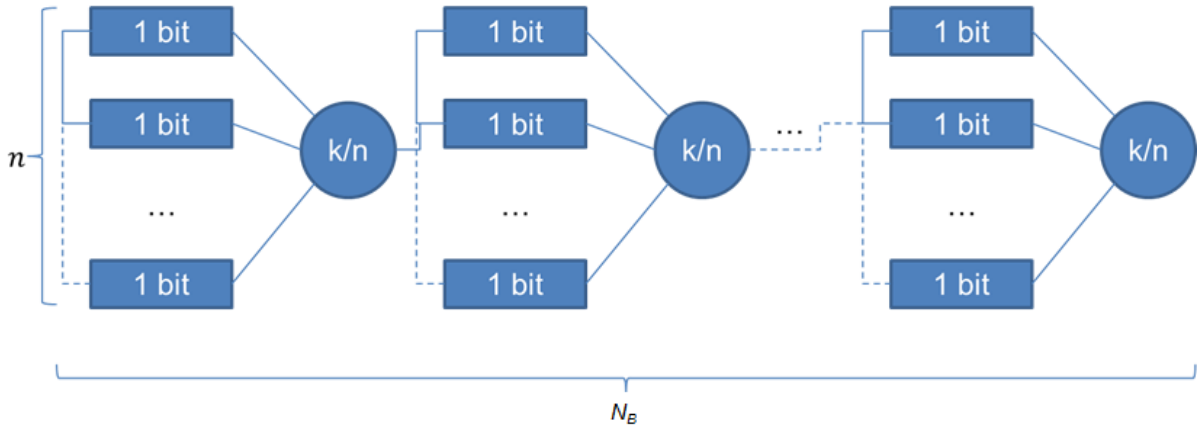


Figure 67 : Schéma de redondance du principe de l'ECC dans une page NAND FLASH

Où $\begin{cases} n = 8936 \\ \mathrm{k} = n - k_{ECC} = 8936 - 40 = 8896 \end{cases}$

En effet, si l'on sectionne une page en $N_B = 8$ sous-parties, on a 8936 bits sur lesquelles seulement 40 bits peuvent être défaillants sans rendre la page entière défaillante (après ECC). Le sous-système (ou sous-partie de page) est opérationnel si :

- Exactement k bits sont bons et (n-k) sont défaillants,
- Ou exactement k+1 bits sont bons et (n-k-1) sont défaillants,
- Ou exactement k+2 bits sont bons et (n-k-2) sont défaillants,
- …
- Ou exactement n bits sont bons et 0 sont défaillants.

Ainsi, la fiabilité du morceau de page suit une loi binomiale $B\big((n-i), R_{bit}(t)\big)$.

On peut définir la probabilité de survie d'une page comme :

$$R_{Page}(t) = \left(\sum_{i=0}^{k_{ECC}} \binom{n}{n-i} . R_{bit}(t)^{n-i} . \left(1 - R_{bit}(t)\right)^{i}\right)^{N_B} \quad (65)$$

Où :

- R_{Page} est la fiabilité d'une seule page
- t est le temps en heure
- R_{bit} est la fiabilité d'un seul bit
- k_{ECC} est le nombre de bits récupérables par ECC par tranche de n bits.
- $\binom{n}{k}$ est le nombre de combinaison de k pages parmi toutes les pages (pages de spare comprise)

Or on ne connaît pas $\mathrm{R_{bit}(t)}$ car le testeur n'a pas été configuré pour enregistrer le nombre de bits défaillants. Partant de l'équation 65 et connaissant $R_{Page}(\mathrm{t_i})$, on peut approcher numériquement $\mathrm{R_{bit}(t_i)}$ pour un ensemble de valeurs discrètes de t. Cependant, le calcul numérique de cette valeur peut vite devenir laborieux car R est inférieur à 1 et k très grand. Heureusement, il existe plusieurs approximations de la loi Binomiale pour des valeurs de n assez grandes.

On peut utiliser l'approximation de la loi binomiale par la loi normale. Cette approximation a l'avantage de nous fournir une solution analytique de $\mathrm{R_{bit}}$. La loi Binomiale $\mathcal{B}\left(n, R_{page}(t)\right)$ peut-être approximée par la loi normale $\mathcal{N}\left(n.R_{page}(t) - \frac{1}{2}, n.R_{page}(t).\left(1 - R_{page}(t)\right)\right)$ lorsque : $n \geq 30$, $n.R_{page}(t) \geq 5$ et $n.\left(1 - R_{page}(t)\right) \geq 5$. L'erreur de cette approximation est d'autant plus faible que *n* est grand et que $\mathrm{R_{page}(t)}$ est proche de 0,5.

Ainsi la fiabilité d'une page après ECC est approximée par :

$$R_{Page}(t) \simeq \left(1 - \phi\left(\frac{n - k_{ECC} - n.R_{bit}(t) + \frac{1}{2}}{\sqrt{n.R_{bit}(t).\left(1 - R_{bit}(t)\right)}}\right)\right)^{N_B} \quad (66)$$

Nous pouvons donc retrouver R_{bit} grâce au polynôme d'ordre 2 suivant:

$$\begin{cases} a.R_{bit}(t)^2 + b.R_{bit}(t) + c = 0 \\ a = n^2 + n.\left(\phi^{-1}\left(1 - \left(R_{Page}(t)\right)^{\frac{1}{N_B}}\right)\right)^2 \\ b = -n.\left(\phi^{-1}\left(1 - \left(R_{Page}(t)\right)^{\frac{1}{N_B}}\right)\right)^2 - 2.n.\left(n - k_{ECC} + \frac{1}{2}\right) \\ c = \left(n - k_{ECC} + \frac{1}{2}\right)^2 \end{cases} \quad (67)$$

Où ϕ^{-1} est la fonction quantile de la loi normale centrée réduite $\mathcal{N}(0,1)$.

En résolvant cette équation on obtient finalement :

$$R_{bit}(t) = \frac{-b+\sqrt{b^2-4.a.c}}{2.a} \quad (68)$$

On a choisi cette solution parmi les deux possibles car $R_{bit}(t)$ est nécessairement croissante avec $R_{Page}(t)$.

La fonction de fiabilité (ou survie) d'une page en fonction de celle d'un bit est présentée sur la Figure 68. La courbe rouge correspond à la solution utilisant la loi binomiale. La courbe bleue correspond à l'approximation par la loi normale.

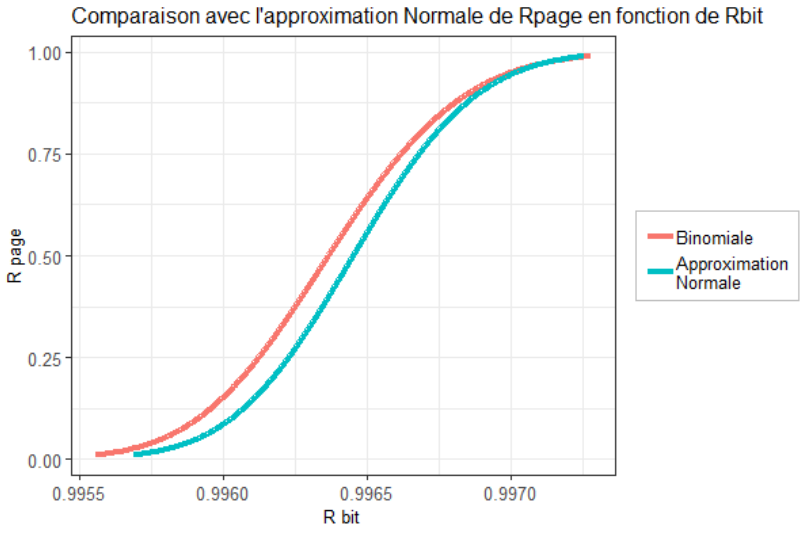


Figure 68 : Fonction de fiabilité d'une page en fonction de la fiabilité d'un bit, avec et sans approximation par la loi normale

Pour avoir une fiabilité correcte au niveau page, il faut impérativement avoir une fiabilité au niveau bit de l'ordre de 99%. L'approximation par la loi normale n'est pas parfaite - car l'on est éloigné de 0,5 pour une fiabilité de R_{page} non nulle – mais convenable pour notre besoin. La différence entre les deux courbes provient essentiellement de l'échelle sur l'axe des abscisses entre 0,9955 et 0,9970.

A partir du modèle défini section 2.6.2.3, on peut tracer (à l'aide de l'approximation par la loi normale) la fonction de fiabilité (ou survie) d'un seul bit écrit à 25°C, activé une fois par jour à 110°C et écrit sur une mémoire vierge (1 PE) comme illustrée à la Figure 69.

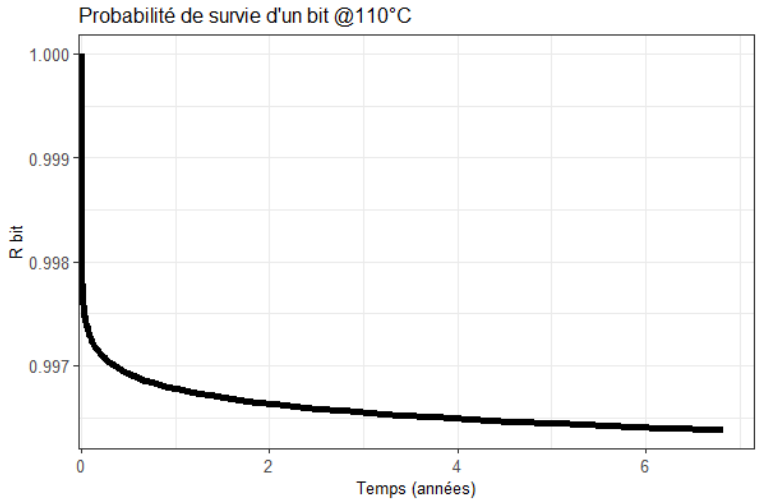


Figure 69 : Fiabilité calculée numériquement d'un bit

A partir de la modélisation de la fiabilité d'un bit il est possible d'estimer de manière plus réaliste la fiabilité d'une mémoire en fonction du code correcteur d'erreur. Dans le cadre de notre étude, l'ECC permet de récupérer 40 bits pour 1117 octets. La Figure 70 présente la fiabilité d'une page NAND en fonction du temps pour 3 capacités de récupération par ECC. On voit très nettement qu'un bon ECC améliore de façon plus que notable la fiabilité des mémoires. C'est pour cette raison qu'ils sont de plus en plus utilisés et performants. Les algorithmes utilisés – qui ne sont pas l'objet de ce mémoire – sont améliorés par une augmentation de la puissance de calcul externe ainsi que par l'augmentation de la taille des pages et des blocs (cela permet de corréler de plus en plus de données dans les codes ECC écrits en fin de page).

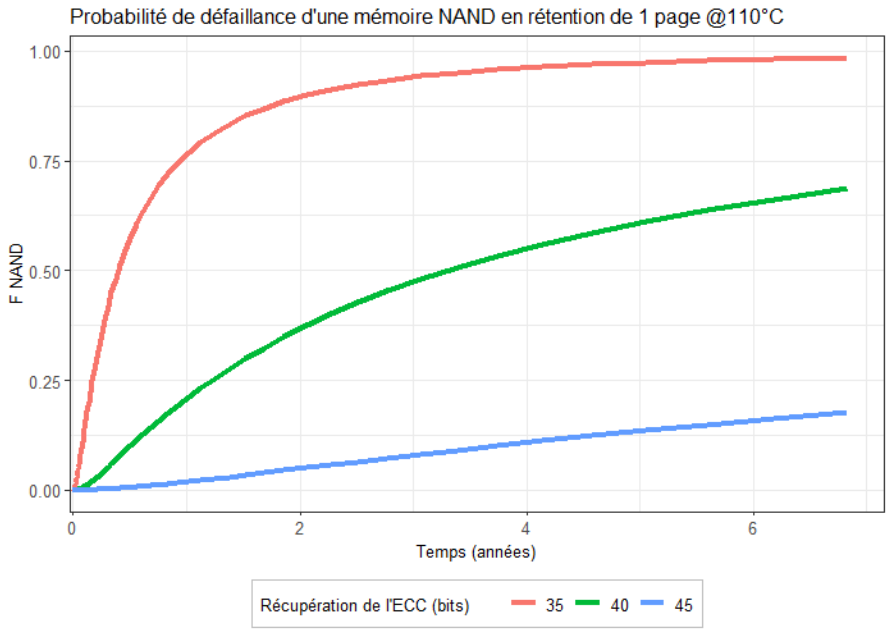


Figure 70 : Probabilité de défaillance en rétention d'une seule page selon plusieurs ECC

2.6.3 Endurance

L'autre aspect important dans l'analyse de la fiabilité des mémoires est l'endurance au fil des cycles d'effacement/écriture. En effet, si un point mémoire ne peut plus être écrit correctement, la notion de rétention devient caduque.

La population considérée est composée de 12 288 000 échantillons (pages) répartis sur 16 Flash NAND. Beaucoup de conditions de stress sont totalement censurées à droite. Le taux de censure total est de 99%. Cependant, comme nous avons tout de même 15 784 défaillances observées dans plusieurs conditions, une modélisation est envisageable. Le détail de la population est donné dans le Tableau 16.

2.6.3.1 Modélisation du vieillissement

A la fin des 18 808 heures de vieillissement, seules les mémoires vieillies à 85°C et à 110°C présentent des erreurs. Même si la proportion d'erreur est faible (au plus 3% pour la ZM3), une modélisation peut être construite. La montée des erreurs est intervenue peu après les 3000 PE (Progam/Erase). Cette valeur est cohérente avec ce qu'annonce la notice du composant. Toutes les ZM ont un comportement similaire, on ne fait donc pas de distinction.

Le seul facteur d'accélération de ce modèle est la température de fonctionnement de la mémoire. Ce cofacteur sera régi par une loi d'Arrhenius (voir équation 70).

Température d'activation	Intervalle de cyclage (h/PE)	Nombre de PE effectués à 18 808 h	Nombre d'échantillons	Nombre de défaillances observées	Ratio de défaillance
25°C	3	6269	768000	0	0,0%
25°C	6	3134	768000	0	0,0%
25°C	12	1567	768000	0	0,0%
25°C	24	783	768000	0	0,0%
85°C	3	6269	1536000	1128	0,1%
85°C	6	3134	1536000	0	0,0%
85°C	12	1567	1536000	0	0,0%
110°C	3	6269	1536000	14654	1,0%
110°C	6	3134	1536000	2	0,0%
110°C	12	1567	1536000	0	0,0%
		TOTAL	12288000	15784	**0,1%**

Tableau 16 : Répartition de la population des pages en endurance NAND

Dans la mesure où l'on observe un mécanisme d'usure, les distributions de Weibull et log-normale ont été confrontées. Les deux donnent des résultats identiques dans la gamme d'échelle observée : autant sur le plan de la valeur de log-vraisemblance que des facteurs d'accélérations estimés, des durées de vie estimées, ou même visuellement. Ce document traitera de la distribution de Weibull. Ce choix provient des facilités de calcul offertes par cette loi ainsi que son utilisation usuelle pour ce type de phénomène.

Le mécanisme d'usure considéré dans cette section est directement corrélé au nombre de PE enduré par les points mémoires. Dans la modélisation de ce mécanisme, on ne considère pas directement le temps mais un nombre de PE cumulé. Le passage de l'un à l'autre s'effectue via une multiplication/division par la fréquence de cyclage f_{PE}.

La forme du modèle en distribution de Weibull est la suivante :

$$F(N_{PE}) = 1 - exp\left(-\left(\frac{N_{PE}}{\eta}\right)^{\beta}\right) \quad (69)$$

$$\eta = \eta_0 . e^{\frac{E_a}{k_b \cdot T}} \quad (70)$$

Où :

- $F()$ est la probabilité de défaillance
- N_{PE} est le nombre de cycles d'écriture effacement effectué. Il est égal à t × f_{PE}

- T est la température absolue de jonction en Kelvin
- η est le paramètre d'échelle de la distribution de Weibull
- η_0 est le pré-facteur du paramètre d'échelle de la distribution de Weibull
- E_a est l'énergie d'activation en eV
- β est le paramètre de forme de la distribution de Weibull

Les valeurs de ce modèle estimées par MLE pour nos résultats sont dans le Tableau 17. Le paramètre de forme β est de 10. Cette valeur bien supérieure à 1 atteste d'un fort mécanisme de vieillissement situé à droite de la courbe en baignoire.

Paramètre	Inf_5%	Standard	Sup_95%
η_0	302,8 PE	302,8 PE	302,8 PE
β	10,46	10,58	10,69
E_a	0,1148 eV	0,1149 eV	0,1151 eV

Tableau 17 : Estimation par MLE des paramètres du modèle NAND endurance (avec intervalles de confiance à 90%)

Visuellement, sur la Figure 71 ayant des échelles adaptées (transformation de Weibull ou « Weibit »), l'alignement des points est très bon. De plus, le paramètre de forme (pente des courbes) est constant au travers des conditions de vieillissement. Cette homogénéité des pentes confirme la présence d'un seul mécanisme apparent de dégradation.

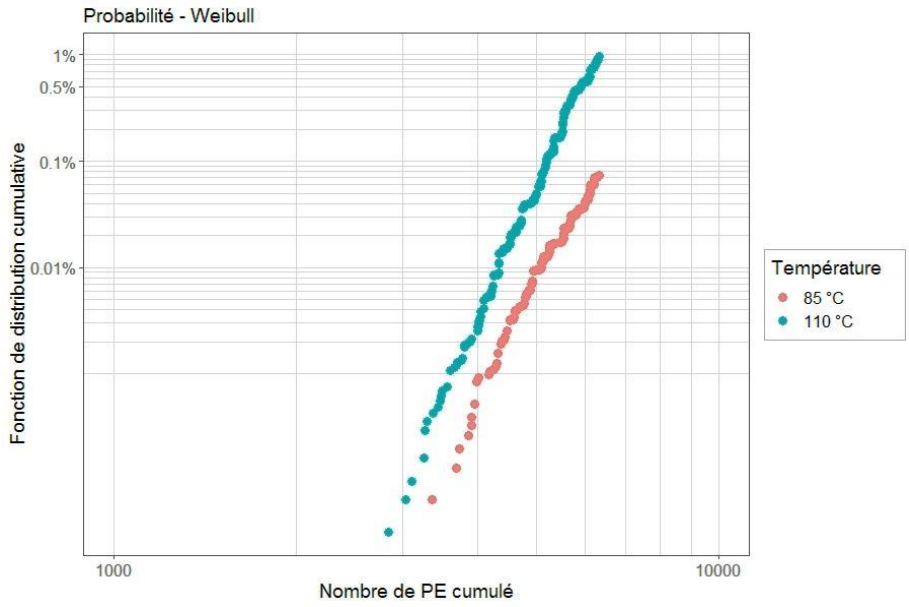


Figure 71 : Fonctions de répartition empiriques par condition de test des NAND 5 endurance

Le *test du rapport de vraisemblance* permet de tester statistiquement un modèle paramétrique contraint contre un autre non contraint. L'intérêt de ce test est ici de savoir si le fait de laisser le paramètre de forme β libre pour chaque condition de test apporte un gain significatif sur la vraisemblance du modèle. Si le test est négatif, on peut utiliser un paramètre β commun à toutes les conditions de test. Le test est ici satisfaisant comme on peut le lire sur le Tableau 18. Le niveau de confiance choisi est de 95%. Le modèle non contraint a 2 degrés de liberté (noté DL) car il y a deux conditions présentant des défaillances (85°C et 110°C). Le paramètre *T* correspond à la statistique du test [116, p. 164].

$$T = -2.log\left(\frac{\mathcal{L}(\beta_0)}{\mathcal{L}(\beta)}\right) \tag{71}$$

Où $\mathcal{L}(\beta_0)$ est la vraisemblance du modèle avec un paramètre β_0 commun aux conditions de test et $\mathcal{L}(\beta)$ est la vraisemblance du modèle non contraint avec un paramètre β propre à chaque condition de test.

La grandeur T suit une loi du χ^2 avec un degré de liberté égal au nombre de contraintes imposées moins 1. On rejette l'hypothèse du test – qui est que le modèle contraint et le modèle non contraint ne diffèrent pas significativement – si T est supérieur au quantile d'ordre $1-\alpha$ de la loi du χ^2.
Dans notre cas $T = 0{,}777$ est inférieur à $(\chi_1^2)^{-1}(0{,}95) = 3{,}84$. Cela signifie qu'il n'y a pas de variations significatives du paramètre de forme entre les différentes conditions de vieillissement. Ainsi on observe bien un même mécanisme de dégradation apparent.

Test du rapport de vraisemblances

Label	Valeur
Niveau de confiance (NC)	0.9500000
DL	2.0000000
T	0.7771082
Khi2 (NC, DL−1)	3.8414588

Variation du paramètre de forme entre les groupes de test significative au niveau de risque 5% ? : FALSE

Tableau 18 : Test du rapport de vraisemblances, NAND endurance

2.6.3.2 Application du modèle choisi

Le MCTF (Mean Cycles To Failure) est le nombre moyen de cycles avant défaillance. Il est ici de la forme :

$$MCTF(T_j) = \eta_0 . exp\left(\frac{E_a}{k_b.T}\right).\Gamma\left(1+\frac{1}{\beta}\right) \quad (72)$$

Le nombre de PE médian jusqu'à défaillance :

$$C_{50\%}(T_j) = \eta_0 . exp\left(\frac{E_a}{k_b.T}\right).\left(ln(2)\right)^{1/\beta} \quad (73)$$

Où Γ( ) est la fonction gamma.

On peut maintenant calculer le nombre moyen de cycle d'écriture/lecture effectuable avant défaillance du point mémoire aux conditions T_j=85°C (température maximale donnée par le constructeur). Ces résultats avec leur intervalle de confiance à 90% sont consignés dans le Tableau 19.

	Inf_5%	**Standard**	**Sup_95%**
Nombre moyen de PE avant panne	11 904	**11 965 PE**	12 029
Nombre médian de PE avant panne	12 058	**12 120 PE**	12 185
Nombre de PE jusqu'à 0.1% de probabilité de panne	6 515	**6 530 PE**	6 544

Tableau 19 : Nombre de d'écriture/lecture avant défaillance projetées à 85°C, NAND 5

Notons que la notice du composant garantit le composant jusqu'à 3000 PE et à 85°C. En appliquant notre modèle, au bout de 3000 PE à 85°C, 0,00003% des pages seront défaillantes.

On peut calculer le taux de défaillance en connaissant la loi du modèle ainsi que ses paramètres. En effet, contrairement à la distribution exponentielle (utilisé pour modéliser le fond plat de la courbe en baignoire), celui-ci n'est pas constant. Pour notre modèle, le taux de défaillance λ par rapport au temps est de la forme :

$$\lambda(t) = -\frac{1}{R(t.f_{PE})}.\frac{dR(t.f_{PE})}{dt} \quad (74)$$

D'où :

$$\lambda(t) = \frac{\beta.{f_{PE}}^{\beta}.t^{\beta-1}}{(\eta)^{\beta}} \quad (75)$$

Ce taux de défaillance est illustré avec un exemple à 85°C et avec 1 cycle de PE toutes les 4 heures sur la Figure 72. L'axe des abscisses est divisé en deux : la première ligne correspond au nombre de PE et la seconde au temps en heures. On observe bien un taux de défaillance croissant du fait du paramètre de forme supérieure à 1.

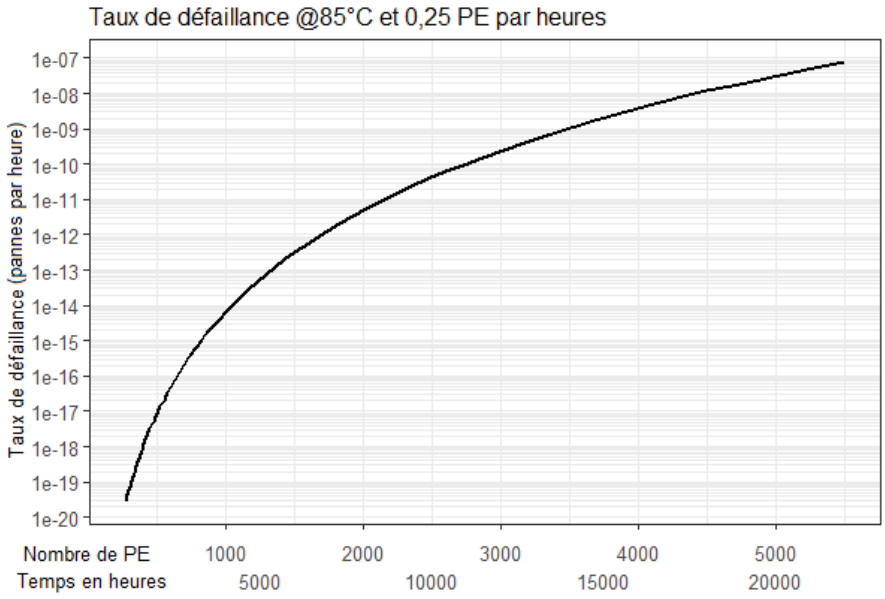


Figure 72 : Taux de défaillance à 85°C, NAND endurance

2.6.3.3 Prise en compte du surdimensionnement sur la fiabilité

Le nombre de cycle d'écriture est la principale limite d'usure sur les mémoires MLC. Les avancées technologiques – hors mémoires 3D NAND – permettent augmenter la densification du nombre de points mémoires. Le nombre maximal de PE admissible par cellule diminue quant à lui. Afin de réduire la sollicitions de chaque point mémoire, diverses techniques sont utilisées : l'usure est uniformisée sur l'ensemble de la mémoire (« wear leveling »), la capacité réelle de stockage est supérieure à celle utilisable simultanément afin de limiter artificiellement l'usure par point mémoire. Dans le calcul de la fiabilité d'un tel support, on ne peut plus se contenter de ne considérer que l'usure physique des cellules. Ces méthodes doivent être prises en compte pour obtenir une estimation réaliste de durée de vie.

Afin d'augmenter la durée de vie des SSD notamment, les fabricants embarquent de 7 à 28% de capacité de rechange (ou plutôt diminuent la capacité utilisateur dans les faits). Cette pratique est appelée surdimensionnement (« over-provisioning » en anglais) (noté OP). Le but de cette section est de généraliser la fiabilité estimée d'une seule page à un SSD en tenant compte de l'OP.

Notons :

- *k* le nombre de page utilisable donné par le fabricant
- *s* le nombre de page de rechange (de « spare »)

Dans le cas où l'on a 7% de rechange, $s = \frac{0,07 \times k}{1-0,07}$. (Valeur usuelle des fabricants)

Les pages de redondance ne sont utilisées que lorsque les pages redondées sont défaillantes.

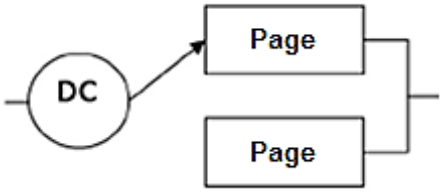


Figure 73 : Schéma d'une redondance passive d'un système avec 2 pages

La fiabilité d'une redondance passive est [117, p. 51]:

$$R_{Sys}(t) = R_{Page}(t) + \int_0^t f_{Page}(\tau).R_{Page}(t-\tau).d\tau \quad (76)$$

Ainsi la fiabilité d'une mémoire peut être vue comme une redondance passive sur les k pages utilisateurs (c.-à-d. que quand une des k pages tombe en panne, elle est remplacée par une des *s* pages de rechange; lorsqu'une seconde page tombe en panne en incluant celle de remplacement, elle est remplacée par une deuxième page, etc.). Ainsi pour garantir l'intégrité de la mémoire, il suffit que k pages soit opérationnelles.

Exemple :
Dans le cas où l'on a 7 pages dont 2 en redondance passive, le système est opérationnel (5 pages opérationnelles à t) si :

- Les 5 pages du début sont encore opérationnelles à t,
- 4 pages du début sont opérationnelles à t (soit 4 parmi 5 choix), et l'une a été remplacée (à l'instant x) par une page de rechange qui n'est pas tombée en panne (soit 1 parmi la seule page restante),
- 4 pages du début sont opérationnelles à t (soit 4 parmi 5 choix), et l'une a été remplacée à l'instant *x* par une page de rechange qui a elle-même été remplacée à l'instant *u* par la dernière page de rechange (soit 1 parmi la seule page restante),
- 3 pages du début sont encore opérationnelles à t (soit 3 parmi 5 choix), et deux ont été remplacées par une page de rechange (soit 2 parmi les deux pages restantes).

On obtient ainsi la formule :

$$\begin{aligned} R_{SYS}(t) = \; & R(t)^5 + \\ & \binom{5}{4}.R(t)^4.\binom{5-4}{1}.\int_0^t f(x).R(t-x).dx + \\ & \binom{5}{4}.R(t)^4.\binom{5-4}{1}.\int_0^t f(x)\left[\int_x^t f(u-x).R(t-u).du\right].dx + \\ & \binom{5}{3}.R(t)^3.\binom{5-3}{2}.\left[\int_0^t f(x).R(t-x).dx\right]^2 \end{aligned} \tag{77}$$

Où R_{sys} est la fiabilité du système, et $R = R_{Page}$ la fiabilité d'une seule page.

<u>Généralisation</u> :

Notons k le nombre de pages utilisateur et s le nombre de pages de rechange. La généralisation à k pages utilisateur et s pages de rechange est une expression de la forme :

$$R_{SYS}(t) = \sum_{i=0}^{s} R_{i\,fail}(t) \tag{78}$$

Où:

- $R_{SYS}(t)$ est la fiabilité du système à l'instant t
- $R_{i\,fail}(t)$ est la fiabilité dans le cas où i pages (parmi les k pages mises en service à t=0) sont défaillantes à t

Dans le cas où toutes les pages n'ont pas eu de panne jusqu'à t, on a :

$$R_{0\,fail}(t) = R(t)^k \tag{79}$$

Si une page est tombée en panne, la redondance passive s'active. On inclut ici la redondance des pages de rechange jusqu'à épuisement des pages de rechange.

$$\begin{aligned} R_{1\,fail}(t) = \; & R(t)^{k-1}.\binom{k-1}{1}.\int_0^t f(x_1).R(t-x_1).dx_1 + \\ & R(t)^{k-1}.\binom{k-1}{1}.\int_0^t f(x_1)\left[\int_0^{t-x_1} f(x_2).R(t-x_1-x_2).dx_2\right].dx_1 + \\ & \ldots + \\ & R(t)^{k-1}.\binom{k-1}{1}.\int_0^t f(x_1).\left[\int_0^{t-x_1} f(x_2).\cdots.\left[\int_0^{t-x_1-\cdots-x_s} f(x_s).R(t-x_1-\cdots-x_s).dx_s\right].\cdots.dx_2\right].dx_1 \end{aligned} \tag{80}$$

On peut simplifier cette expression en posant la suite récursive $R_j(t)$ définie comme suit :

$$\forall(t) \in \mathbb{R}^+;\; \begin{cases} R_0(t) = R_{page}(t) \\ R_j(t) = \int_0^t \left[f(x).R_{j-1}(t-x).dx\right] \end{cases} \tag{81}$$

Et on obtient :

$$R_{1\,fail}(t) = \mathrm{R_0(t)}^{k-1}.\sum_{j=1}^{s}\binom{\mathrm{k-1}}{1}.\mathrm{R_j(t)} \tag{82}$$

La généralisation à *k* pages utilisateur et *s* pages de rechange est une expression de la forme :

$$\mathrm{R_{sys}(t)} = \sum_{\begin{cases}[\sum_{i=0}^{s}(n_i)]=k\\ [\sum_{i=1}^{s}(i.n_i)]\leq s\end{cases}}\left[\prod_{j=0}^{s}\binom{\sum_{i=j}^{s}(n_i)}{n_j}.\left(\mathrm{R_j(t)}\right)^{n_j}\right] \tag{83}$$

n est une suite de nombres *(n_0, n_1, … , n_k)* où chaque n_i est le nombre de pages ayant subi *i* remplacements. La condition « $[\sum_{i=0}^{s}(n_i)] = k$ » permet d'assurer qu'exactement *k* pages sont fonctionnelles pour rendre le système opérationnel. La condition « $[\sum_{i=1}^{s}(i.n_i)] \leq s$ » permet quant à elle d'assurer qu'il n'y ait pas plus de pages remplacées que de pages de remplacement pour rendre le système opérationnel. La somme « $\sum_{i=j}^{s}(n_i)$ » traduit le nombre de page encore à remplacer.

Notons que les termes de combinaison « k parmi n » se simplifient comme suit :

$$\binom{k}{n_0}.\binom{k-n_0}{n_1}.\binom{k-n_0-n_1}{n_2}.\cdots.\binom{n_s}{n_s} = \frac{k!}{n_0!\times(k-n_0)!}\times\frac{(k-n_0)!}{n_1!\times(k-n_1)!}\times\cdots\times\frac{(n_s)!}{n_s!\times(0)!} = \frac{k!}{n_0!n_1!\cdots n_s!} \tag{84}$$

L'expression devient donc en l'appliquant à notre cas :

$$R_{NAND}(n_{PE}) = k!\times\sum_{\begin{cases}[\sum_{i=0}^{s}(n_i)]=k\\ [\sum_{i=1}^{s}(i.n_i)]\leq s\end{cases}}\left[\prod_{j=0}^{s}\frac{\left(\mathrm{R_j}(n_{PE})\right)^{n_j}}{n_j!}\right] \tag{85}$$

Où :

- R_{NAND} est la fiabilité de toute la mémoire
- n_{PE} est le nombre de PE cumulé
- R_{page} est la fiabilité d'une seule page
- *k* est le nombre de pages garanti
- *n=k+s* le nombre total de page dans le composant

On peut approximer numériquement (méthode de Simpson, de Gauss-Legendre,…) cette expression. Cependant, la génération des couples $\begin{cases}[\sum_{i=0}^{s}(n_i)] = k\\ [\sum_{i=1}^{s}(i.n_i)] \leq s\end{cases}$ est très fastidieuse lorsque *k* devient grand. Pour information, le nombre de composition d'un entier *n* est 2^{n-1}. Malgré le fait, de par la condition sur la somme des $i\times n_i$, que l'on ne considère pas toutes les combinaisons, cette expression est inutilisable en l'état avec les moyens de calcul dont nous

disposons pour *n* supérieur à 15. L'algorithme de calcul utilisé pourrait être amélioré dans une poursuite de cette étude.
Pour gagner également un peu en temps de calcul, on peut utiliser la formule de Stirling modifiée (approximation de Gosper) :

$$n! \sim \sqrt{2\pi . n}. \left(\frac{n}{e}\right)^n . \left(1 + \frac{1}{12.n}\right) \quad (86)$$

On peut approximer la valeur R_{NAND} numériquement pour 15 pages en un temps de calcul raisonnable. Appliquons maintenant cette redondance à notre étude. Pour rappel, la fiabilité en endurance d'une seule page NAND est donnée dans la section 2.6.3.1 où la modélisation de Weibull a été établie. Considérons une température de 85°C, la probabilité de défaillance d'une seule page est représentée Figure 74.

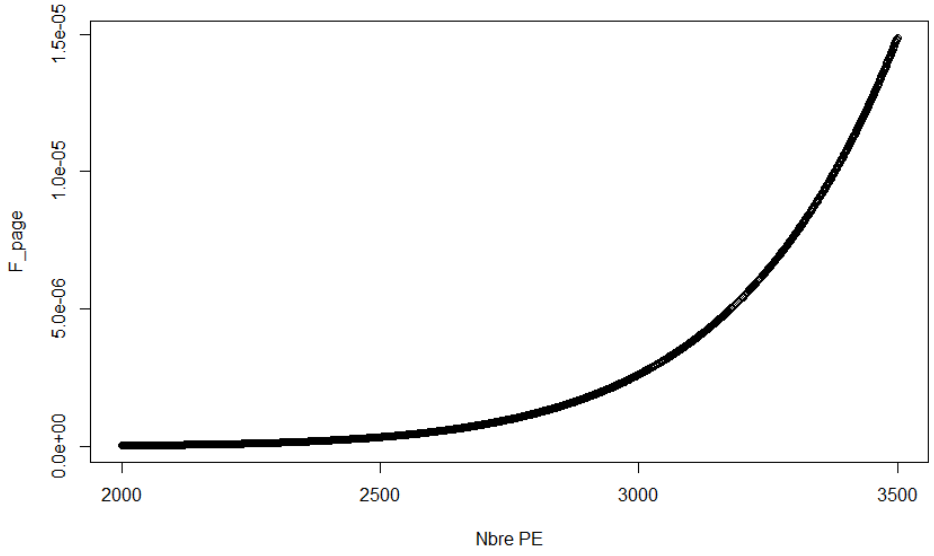


Figure 74 : Probabilité de défaillance d'une page en endurance à 85°C

Notons que la fiabilité sans surdimensionnement est donnée par :

$$R_{NAND}(t) = R_{page}(t)^n \quad (87)$$

Si l'on considère maintenant 15 pages utilisables de notre mémoire pour différents niveaux de pages de rechange, on obtient la Figure 75.

Ainsi, l'augmentation du pourcentage de page de rechange augmente bien la fiabilité de la mémoire au détriment de la capacité de stockage. Sur cet exemple, rien que 7% de redondance permet de repousser le C_{50} de plus de 500 PE – ce qui n'est pas négligeable. Notons, que ce raisonnement est uniquement valable si un système d'uniformisation de

l'usure est en place (à l'aide du contrôleur de la mémoire) afin de répartir l'usure sur l'ensemble des points mémoires.

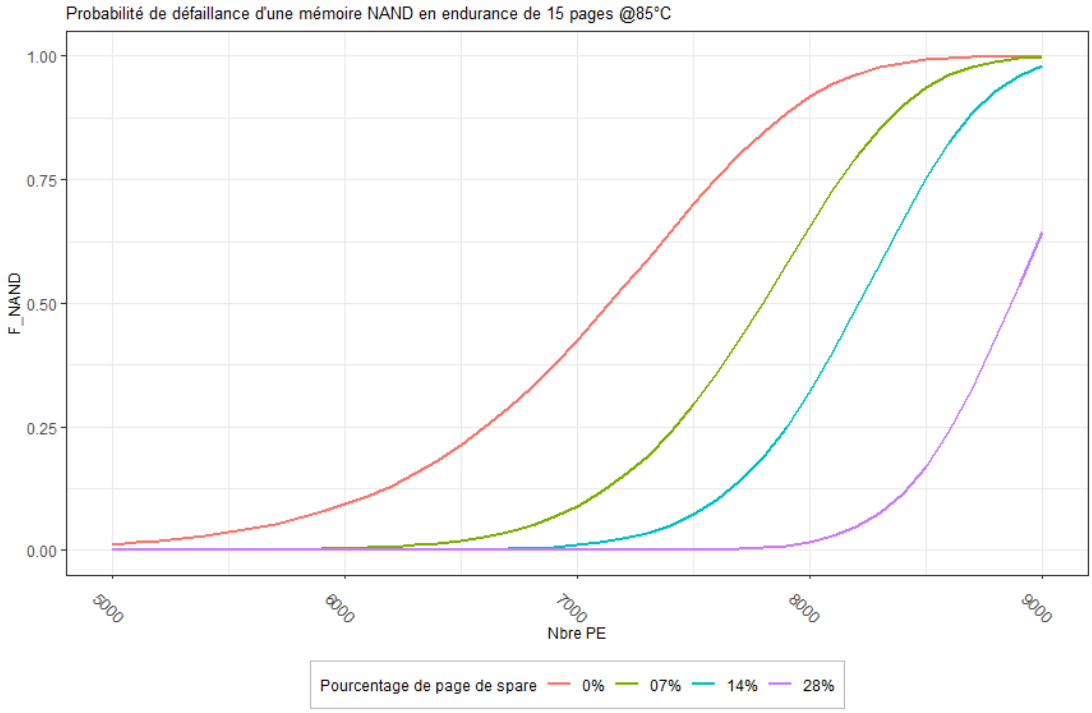


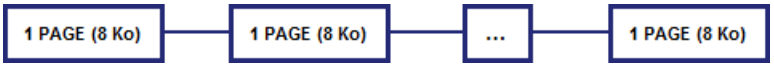
Figure 75: Probabilité de défaillance d'une mémoire de 15 pages complète incluant différent pourcentage de zone de rechange

2.6.4 Combinaison des deux mécanismes

La fiabilité des mémoires Flash NAND en rétention et en endurance ont été estimées précédemment séparément. Il est maintenant pertinent de proposer un modèle général. L'endurance impacte la fiabilité au moment de l'écriture de la donnée alors que la fiabilité en rétention concerne la période de stockage ou d'activation. La fiabilité globale (hors boîtier) est donc la somme des deux.

$$\lambda_{NAND,1page}(t, T_j, T_{prog}, f_{read}, Nbre_{PE}) = \lambda_{\text{rétention}}(t, T_j, T_{prog}, f_{read}, Nbre_{PE}) + \lambda_{\text{endurance}}(Nbre_{PE}, T_j) \quad (88)$$

Le taux de défaillance calculé ci-dessus est valable pour la fiabilité d'une seule page. Dans le cas de la fiabilité d'une mémoire complète (sans redondance pour simplifier ici les calculs), cela revient au système série suivant :



La fiabilité du système total peut donc s'écrire de la forme :

$$\mathrm{F_{mémoire}} = 1 - \left(1 - \mathrm{F_{1\ page}}\right)^{N} \tag{89}$$

Où :

- N est le nombre de pages
- $F_{mémoire}$ est la probabilité de défaillance de la mémoire complète
- $F_{1\ page}$ est la probabilité de défaillance d'une seule page

Le taux de défaillance pour une mémoire de N pages est :

$$\lambda_{NAND}(t, T_j, T_{prog}, f_{read}, Nbre_{PE}, Taille) = \frac{(Taille\ mémoire\ en\ Mo)}{8\ Ko} \times \lambda_{NAND,1\ page}(t, T_j, T_{prog}, f_{read}, Nbre_{PE}) \tag{90}$$

La suite de cette sous-section présente deux exemples d'utilisation d'une mémoire : du stockage longue durée et une utilisation dynamique de la mémoire.

2.6.4.1 Cas d'une rétention longue durée

Prenons le cas d'une clef USB pour un exemple d'application en stockage longue durée. Une donnée peut être écrite sur une mémoire « usée » ayant subi un grand nombre de PE. La donnée écrite sera ensuite très peu lue, mais on veut pouvoir garder l'information écrite sur une longue durée - de l'ordre de la dizaine d'années. Ainsi, comme vu dans ce chapitre, la fiabilité se décompose en deux parties : les chances de réussite d'écriture sans échec de la donnée en initial puis la non perte de l'information au cours du temps. Le taux de défaillance est représenté Figure 76. On considère que la donnée est bien contrôlée (c.-à-d. relue au moins une fois correctement) après écriture, ainsi le taux de défaillance lié à l'écriture n'est pas représenté ici. Dans les conditions de la Figure 76, à savoir à 20°C, une mémoire de 16Mo, une lecture de la donnée tous les ans et une mémoire à mi-vie (1500 PE sur les 3000), on observe bien la montée strictement croissante du taux de défaillance en rétention.

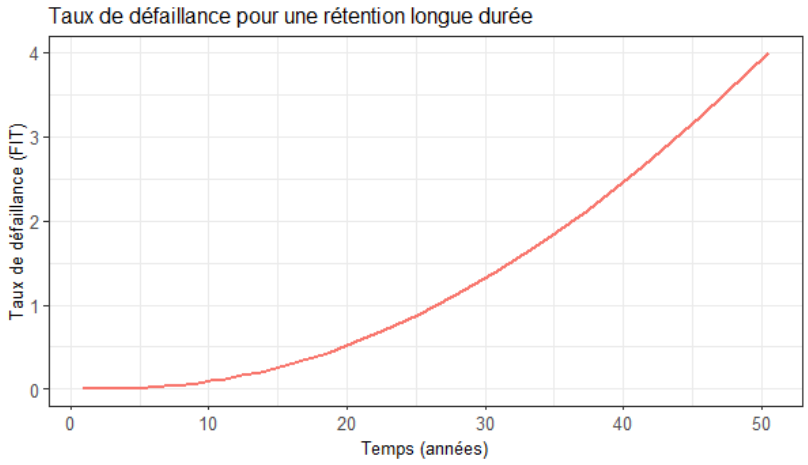


Figure 76 : Taux de défaillance en fonction du temps à $T_j = T_{prog} = 20°C$, taille de 16Mo, une lecture par an et 1500 PE

2.6.4.2 Cas d'une utilisation dynamique en PE

Le second cas d'utilisation d'une mémoire est beaucoup plus dynamique. Les données sont conservées sur des courtes périodes avec de nombreuses réécritures. Prenons par exemple un calculateur avionique. A chaque arrêt, il doit sauvegarder tous ses paramètres courants dans un espace mémoire afin de la recharger au prochain démarrage. Notons qu'il est préconisé de ne pas toujours réécrire aux mêmes adresses pour limiter le nombre de PE accumulés, mais pour cet exemple, cette bonne pratique n'est pas appliquée.

On part d'une mémoire vierge dans un environnement à 20°C, on la lit toutes les 6 heures et on réécrit une donnée également toutes les 6 heures. Ainsi le taux de défaillance en rétention redescend toutes les 6 heures. Or avec l'accumulation croissante du nombre de PE, le taux de défaillance de tous les mécanismes de dégradation augmente. Le calcul du taux de défaillance au cours du temps sera donc :

$$\begin{aligned}\lambda_{NAND}\left(t, T_j = 20^\circ C, T_{prog} = 20^\circ C, f_{read} = \frac{1}{6}, Nbre_{PE} = \left\lceil \frac{t}{6} \right\rceil\right) = \\ \lambda_{\text{rétention}}\left(t \bmod 6, T_j = 20^\circ C, T_{prog} = 20^\circ C, f_{read} = \frac{1}{6}, Nbre_{PE} = \left\lceil \frac{t}{6} \right\rceil\right) + \\ \lambda_{\text{endurance}}\left(Nbre_{PE} = \left\lceil \frac{t}{6} \right\rceil, T_j = 20^\circ C\right)\end{aligned} \quad (91)$$

La Figure 77 illustre cette formule sur quelques cycles de réécriture de notre cas exemple. La courbe verte correspond au taux de défaillance en endurance. Ici le nombre de PE effectué est très faible, donc quasi nul. La courbe rouge correspond au taux de défaillance lié à la rétention. La courbe bleue correspond à la somme des deux mécanismes. Dans ce cas, les

courbes rouge et bleue sont confondues. Le taux de défaillance en rétention (courbe rouge et bleue) redescend après chaque nouvelle écriture.

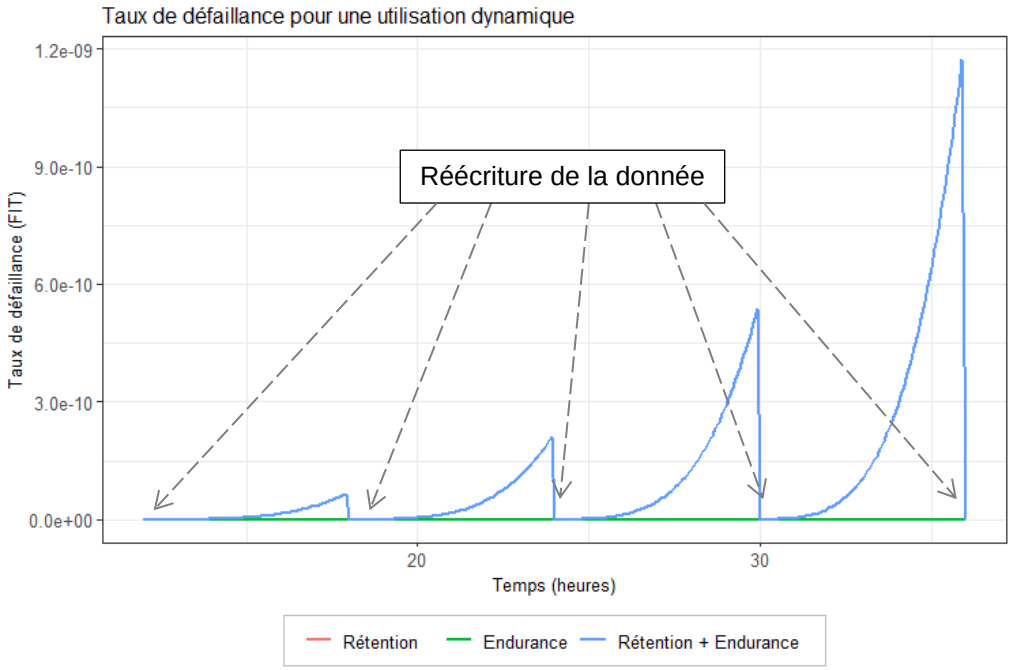


Figure 77 : Taux de défaillance en fonction du temps à $T_j = T_{prog} = 20°C$, une lecture toutes les 6h et une réécriture de la donnée (PE) toutes les 6 heures

Si l'on veut présenter ce type de graphique pour une durée assez longue, ces dents de scie rendent illisible la visualisation. Une autre approche peut être de remplacer chaque « dent de scie » par son intégrale. Ce choix se justifie car la durée de chaque pic qui ne dépasse pas ici les 6 heures. La valeur à la fin de chaque réécriture est l'intégrale entre deux écritures successives. La Figure 78 compare le tracé du taux de défaillance en « dent de scie » et l'approche « intégrale ». La valeur intégrale suit bien la même tendance sans donner trop d'importance au taux de défaillance nul après chaque réécriture, ni à la valeur extrême juste avant. Cette approche va donc être gardée pour représenter la fiabilité globale tenant compte à la fois de la rétention et de l'endurance.

Dans la vie d'une mémoire utilisée de la sorte, la fiabilité en rétention entre deux écritures devient de plus en plus problématique. Cette perte de fiabilité provient principalement de l'augmentation de l'usure du point mémoire. En parallèle, la fiabilité en endurance (intervenant lors des phases d'écriture) ne cesse de décroitre également. La Figure 79 présente les taux de défaillance associés à chacun de ces deux mécanismes ainsi que la somme des deux. La fiabilité en endurance (courbe verte) diminue après chaque réécriture jusqu'à devenir significative après 10 ans dans cet exemple. De plus, cette remontée est bien

plus rapide que pour la rétention. Ainsi il est bien important de considérer ces deux mécanismes pour évaluer correctement la fiabilité d'une mémoire utilisé de manière dynamique.

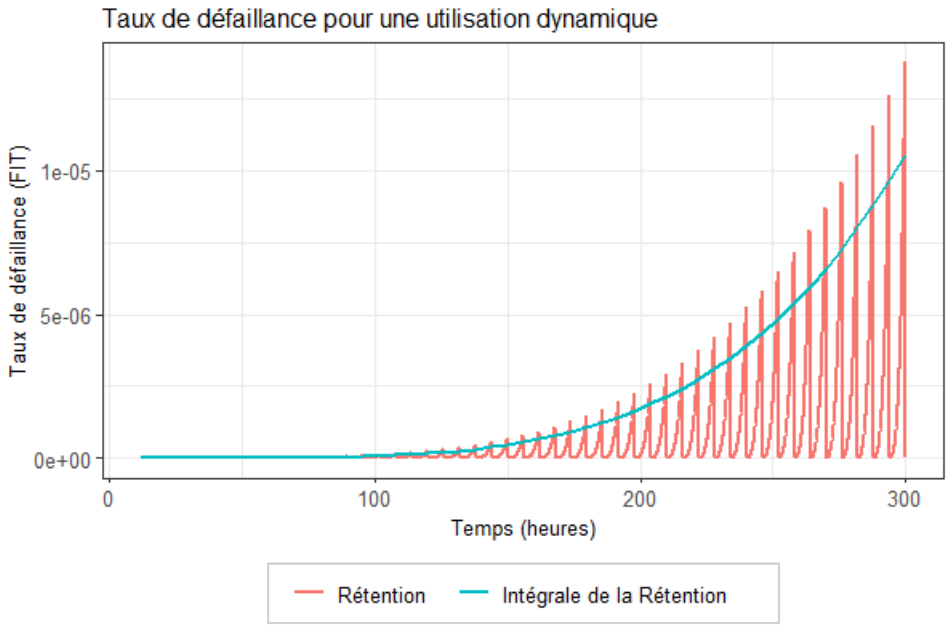


Figure 78 : Taux de défaillance en rétention avec approximation par la valeur intégrale pour une utilisation dynamique

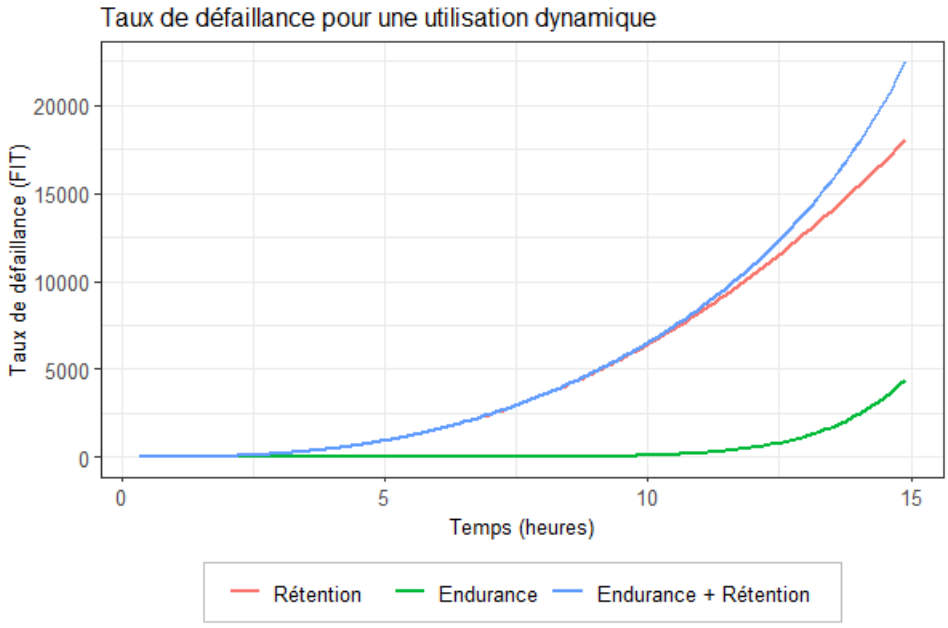


Figure 79 : Taux de défaillance pour une utilisation dynamique

2.7 Conclusion sur la fiabilité des mémoires flash NAND

La fiabilité des mémoires Flash NAND est bien un sujet à part entière. Cette technologie MLC soulève beaucoup de questions au niveau de sa fiabilité. Nous en avons balayé un certain nombre. Comme observé par la littérature, la température de stockage ainsi qu'une forte usure – beaucoup de cycle de Programmation/Effacement – réduisent nettement la fiabilité en rétention de ces mémoires en accord avec nos essais.

La particularité de notre étude a été l'investigation poussée de l'influence de la température pendant la phase d'écriture sur la durée de rétention de la donnée. Un grand écart entre celle d'écriture et celle de lecture mène à une forte perte de fiabilité. Ce problème n'est pas lié à la physique du point mémoire à basse température mais à une dérive externe. Cette anomalie provient du circuit périphérique gérant l'écriture et la lecture des cellules. La référence de tension utilisée dérive sous l'effet de la température. Ainsi une lecture à plus haute température entraine un désalignement entre les limites de tension de seuil de deux états voisins et les valeurs écrites à basse température. La marge de bruit prévue initialement à température ambiante par le fabricant est d'autant plus réduite que l'écart thermique est important.

L'aspect endurance face au cycle de Programmation/Effacement est quant à lui en total accord avec les valeurs annoncées par le fabricant. Nous avons également mené une analyse statistique dans ce chapitre. Concernant la rétention de données, les mémoires Flash NAND ont une fiabilité en rétention très mauvaise – voir inexploitable dans des conditions extrêmes – sans systèmes externes. Dans ce chapitre, nous avons montré comment les algorithmes codes correcteurs d'erreurs, les mécanismes d'uniformisation d'usure ou bien le surdimensionnement permettent de considérablement augmenter la fiabilité de la mémoire en partant d'une fiabilité physique au niveau page très mauvaise. Elle est en effet très vite non admissible pour une application critique lorsque les conditions thermiques deviennent sévères ou bien que la mémoire est utilisée en écriture ou lecture de manière continue. Nous avons proposé des méthodes d'estimation de la fiabilité en tenant compte de tous ces aspects afin de calculer des prévisions plus réalistes.

Chapitre 3

Etude des FPGA

L'objectif global de ce mémoire est - pour rappel - de connaitre la fiabilité des technologies DSM par rapport aux générations antérieures. Les deux principales familles de composant sont les mémoires Flash (avec transistors à grilles flottantes) et les circuits logiques (en cellules CMOS). La fiabilité des mémoires Flash a été abordée au Chapitre 2. La seconde famille – les circuits logiques complexes – va être traitée dans le présent chapitre. Les composants choisis sont des FPGA. Ils permettent de créer des fonctions complexes en bénéficiant d'une intégration et de performances de plus en plus poussées au fur et à mesure des évolutions technologiques.

Les FPGA sont les composants idéaux pour effectuer des tests de vieillissement au niveau silicium. De par leur flexibilité, on peut rapidement et à moindre coût mesurer des dégradations des transistors sans faire de test WLR (test sous pointes effectué au niveau wafer en sortie de fabrication). Le sous-chapitre suivant détaillera la structure de test (autour d'oscillateurs en anneau) utilisée dans les FPGA. Le composant choisi est un « Artix7 » de Xilinx. Cette référence est réalisée dans une technologie DSM gravée en technologie 28 nm avec diélectrique high-κ (HKMG) en accord avec le besoin de l'étude. De plus c'est un bon compromis entre coût unitaire et performance.

Les principaux mécanismes de dégradation dans les circuits intégrés au niveau CMOS sont le BTI, le HCI, le TDDB et l'EM. Tous ces mécanismes ont été détaillés dans le chapitre « Etat de l'art». Les tests effectués visaient à priori à faire apparaitre tous ces mécanismes. Dans les faits, nous n'avons observé aucune variation brutale des grandeurs mesurées permettant d'identifier de l'électromigration ou du TDDB. Il est probable que les températures mises en jeu dans nos essais étaient trop basses pour faire apparaitre de l'électromigration pendant la durée de nos essais, et les tensions appliquées aux grilles des transistors pas assez élevées pour générer rapidement du TDDB. Ainsi, ce chapitre abordera uniquement le BTI et le HCI. Dans un profil de vie classique des FPGA, le HCI et le BTI sont les mécanismes de dégradation - au niveau logique transistor CMOS - les plus visibles.

Une analyse fine de la cinétique des dégradations sera faite. De plus, la statistique de la dispersion des pièces sera également traitée comme pour le chapitre sur le mémoire.

Cette partie sera divisée en 4 sous-parties. La première présentera la mise en place des essais. La seconde formalisera le traitement des données qui a été mis en place. Les méthodes d'extraction des mécanismes de vieillissement seront ensuite décrites. Finalement, l'analyse de chacun d'entre eux sera présentée en détail.

3.1 Organisation des essais

3.1.1 Stratégie des essais

Le HCI et le BTI sont des mécanismes distincts qui sont favorisés dans des conditions qui leur sont propres. Le HCI est actif à basse température et à haute fréquence de fonctionnement. Le BTI quant à lui est aggravé par les hautes températures et un rapport cyclique élevé, le pire cas étant un état passant constant. Pour chacun, une hausse de la tension d'alimentation est traduite par une augmentation du vieillissement. La Figure 80 présente les zones où le HCI et le BTI sont les plus attendues en fonction de la température de vieillissement (ordonnées) et de la fréquence de stress (abscisses).

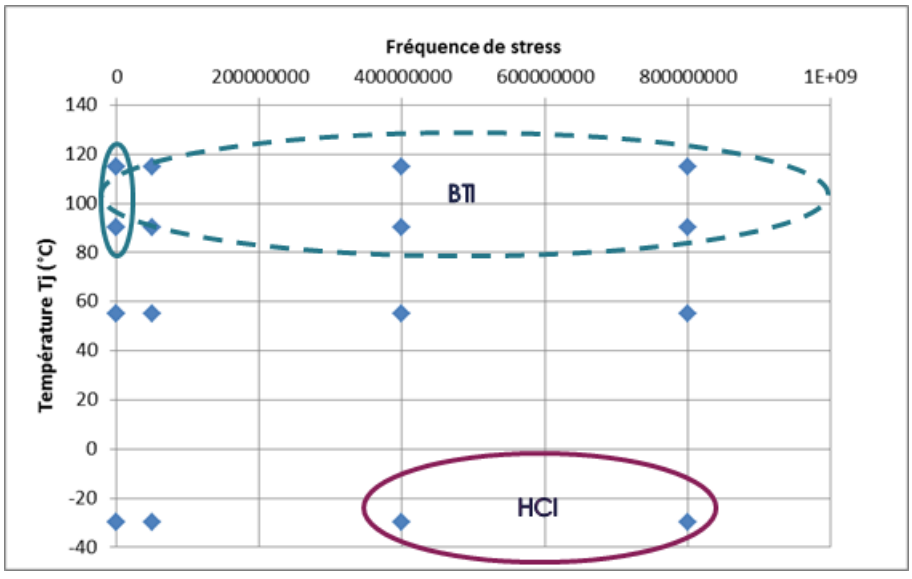


Figure 80 : Localisation du HCI et du BTI sur la carte des conditions de stress

La stratégie adoptée pour le vieillissement des FPGA est d'utiliser une structure à base d'oscillateurs en anneau (« Ring Oscillators », ou RO). Les RO ainsi que leur circuit de mesure associé sont embarqués dans chaque DUT. Cette partie sera décrite dans la

section 3.1.2. Concernant les stress thermiques et électriques (plus classiques dans les HTOL ou LTOL), un banc de test permettra de mettre en œuvre les conditions voulues. Celui-ci est conçu sur la même base que celui utilisé pour les essais dynamiques de mémoires présenté dans le chapitre précédent.

Les différentes conditions de vieillissement choisies sont détaillées dans les tableaux suivants. Les températures de jonction atteintes vont de -30°C à 115°C. Les tensions d'alimentation - cœur logique - appliquées aux DUT vont de la tension nominale V_{nom} (1V) à V_{nom}+50% (1,5V).

FPGA RO1 (HOT)	T° vieillissement		
Tension	25°C	90°C	120°C
1 V		2	3
1,1 V			3
1,2 V			2
1,4 V		3	
1,5 V	2		
Nb heures par composant	5 038	5 038	5 038

Tableau 20 : Nombre de DUT par condition, FPGA HOT V1

FPGA RO3 (HOT+FCT)	T° vieillissement		
Tension	25°C	90°C	115°C
1 V	1		
1,1 V		2	2
1,3 V		2	2
1,5 V		2	4
Nb heures par composant	8 123	8 123	8 123

Tableau 21 : Nombre de DUT par condition, FPGA HOT V2

FPGA RO2 (COLD)	T° vieillissement			
Tension	-30°C	25°C	90°C	115°C
1 V			1	
1,1 V	2		1	1
1,2 V		1	2	
1,3 V	2	2		
1,4 V	1			
1,5 V	1	1		
Nb heures par composant	11 650	11 650	11 650	11 650

Tableau 22 : Nombre de DUT par condition, FPGA COLD

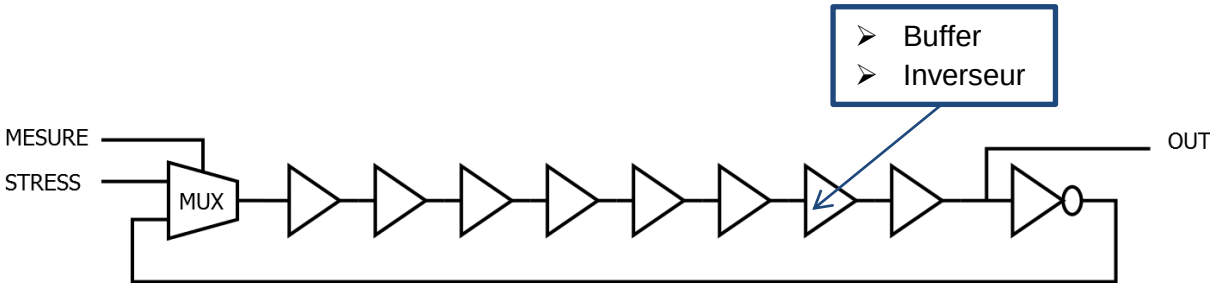
3.1.2 Description des oscillateurs en anneau implémentés

La structure de test élémentaire est ici à base d'oscillateurs en anneau (RO). Chaque FPGA en contient 96 ou 192 suivant la matrice de vieillissement considérée. Le principe de fonctionnement du RO est le suivant : on met en cascade un nombre impair d'étages inverseurs, puis on reboucle la chaîne sur elle-même. La chaîne est ainsi instable : un front montant puis descendant se propage, créant ainsi une oscillation libre.
La fréquence d'oscillation dépend du temps de propagation des étages, donc de la tension de seuil des transistors impliqués. Une dérive d'au moins un transistor de la chaîne générera une dérive de la fréquence du RO. Cette structure simple à mettre en œuvre permet d'observer des dérives paramétriques bas niveau.

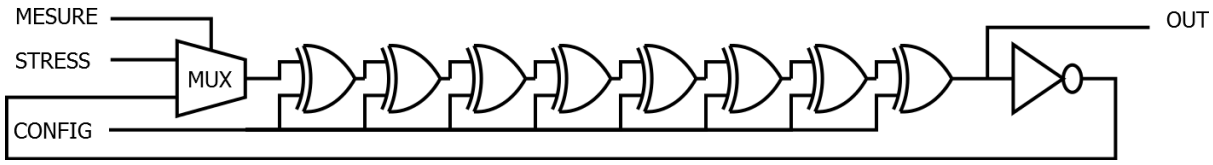
La Figure 81 schématise les différents types de RO implémentés dans nos véhicules de test. Chaque porte représentée est implémentée par une LUT. Notons qu'il peut y avoir des étages types « buffer » dans les RO en plus du nombre impair d'inverseur comme dans la chaîne de buffer en Figure 81.a.
Le multiplexeur permet ici de basculer entre deux états :

- Le mode stress : la chaîne est ouverte et un signal de stress externe (AC ou DC) est appliqué ;
- Le mode mesure : la chaîne est fermée et l'oscillation libre se produit.

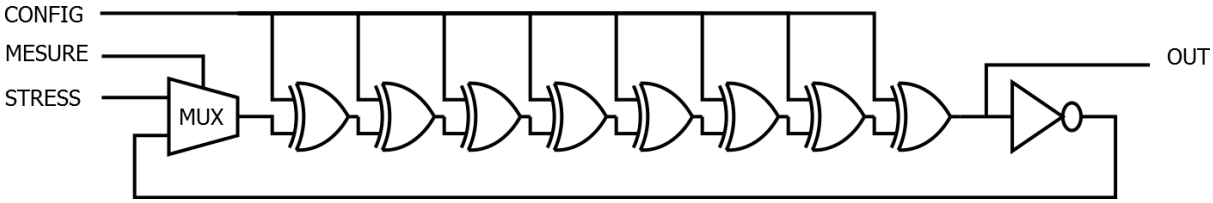
Les différents signaux de stress sont générés à partir du signal 25 MHz issu du banc de test via les PLL internes des DUT. Afin d'étudier le HCI, plusieurs fréquences de stress sont utilisées. Ces fréquences vont du signal continu à 800 MHz. Dans l'étude du BTI, plusieurs rapports cycliques de stress sont implémentés pour différentes fréquences de stress: 0% (DC0), 25%, 50%, 75% et 100% (DC1). Quatre types de chaîne de RO ont également été considérés : Buffer (voir Figure 81.a), Inverseur (voir Figure 81.a), XOR (voir Figure 81.b) et LUT2 (voir Figure 81.c). Même si les XOR sont également une porte LUT2, cette appellation permet de distinguer les deux dans la suite de cette étude. La différence entre XOR et LUT2 est l'ordre des entrées de chaque étage (permutation de la sortie de l'étage précédant et du signal de commande). La motivation de cette variation sur l'ordre des entrées est de forcer l'outil de placement et routage à utiliser différentes entrées sur les LUT6 fixes du FPGA. Les types XOR et LUT2 sont stressés avec un signal de commande à '0', mais mesurés une fois sur deux avec ce signal à '1'. Il y a donc une alternance entre le type buffer et le type inverseur. Deux longueurs de chaînes ont été implémentées. La plupart des RO sont de taille 9 LUTs, mais certains sont composés de 17 LUTs. La liste complète des structures des 96 RO implémentés ainsi que leurs différents signaux de stress est donnée en Annexe B. Cette architecture d'autotest est basée sur les travaux effectués par Mohammad NAOUSS [30].



(a) Chaîne de type Buffer ou inverseur



(b) Chaîne de type XOR



(c) Chaîne de type LUT2

Figure 81 : Structure des différents RO implémentés.

En résumé, les cofacteurs de stress sont les suivants :

- Température de vieillissement : -35°C, 55°C, 90°C et 115°C
- Composants alimentés pendant le vieillissement
- Tension de vieillissement cœur (V_{nom}=1V) : 1V ; 1,1V ; 1,2V ; 1,3V ; 1,4V et 1,5V
- Type de RO : Buffers, Inverseurs, XOR et LUT2
- Fréquence de stress du RO : 0Hz, 25Hz, 50MHz, 400MHz et 800MHz
- Rapport cyclique de stress : 0%, 25%, 50%, 75% et 100%

3.1.3 Circuit de mesure

Une fois par heure, les RO sont rebouclées sur eux-mêmes. A ce moment-là, une mesure de la fréquence d'oscillation libre ainsi qu'une mesure du rapport cyclique d'oscillations du RO est effectuée. Les signaux de sorties des RO sont multiplexés vers le circuit de mesure afin d'optimiser les ressources du FPGA. Il n'y a qu'un circuit de mesure pour 96 RO. Ils sont ainsi mesurés de manière séquentielle. Les mesures des RO sont faites dans les conditions thermiques et électriques de stress du composant. Le schéma du circuit de mesure (embarqué dans les DUT) est donné Figure 82. Une version plus détaillée est en Figure 83.

Ce circuit comporte deux compteurs. Le premier permet pour compter le nombre de front montant du RO pendant une seconde ; et ainsi obtenir directement la fréquence en Hertz codé sur 30 bits (jusqu'à 1 GHz). Le second compteur permet de mesurer le rapport cyclique. Le nombre de coups d'horloge d'un signal à 300 MHz est comptabilisé lorsque le signal de sortie du RO est à l'état haut. Il suffit ensuite de diviser la sortie du compteur (codée sur 29 bits) par 3×10^8 pour obtenir le rapport cyclique. Notons que ce circuit a été dimensionné pour supporter une fréquence de fonctionnement de maximale 300 MHz. Cette contrainte est également valide pour le circuit de mesure de la fréquence des RO.

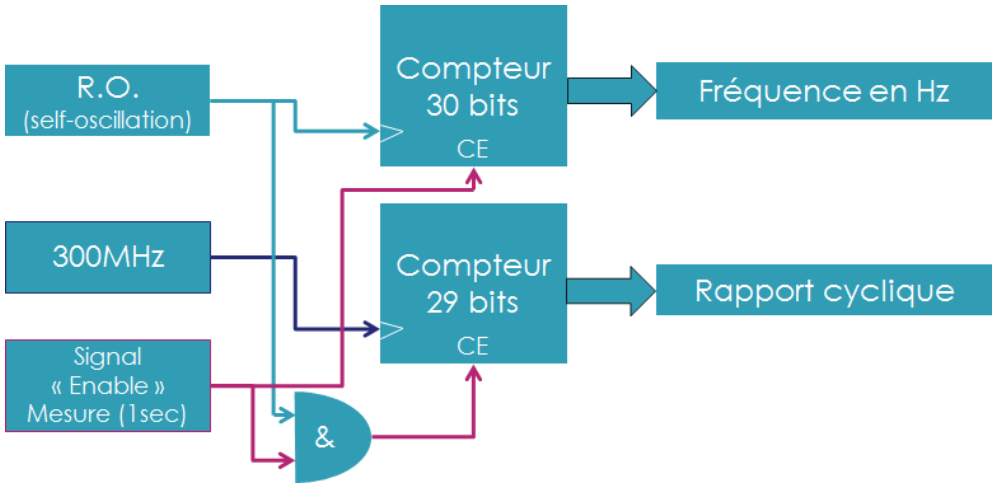


Figure 82 : Diagramme du circuit de mesure des RO

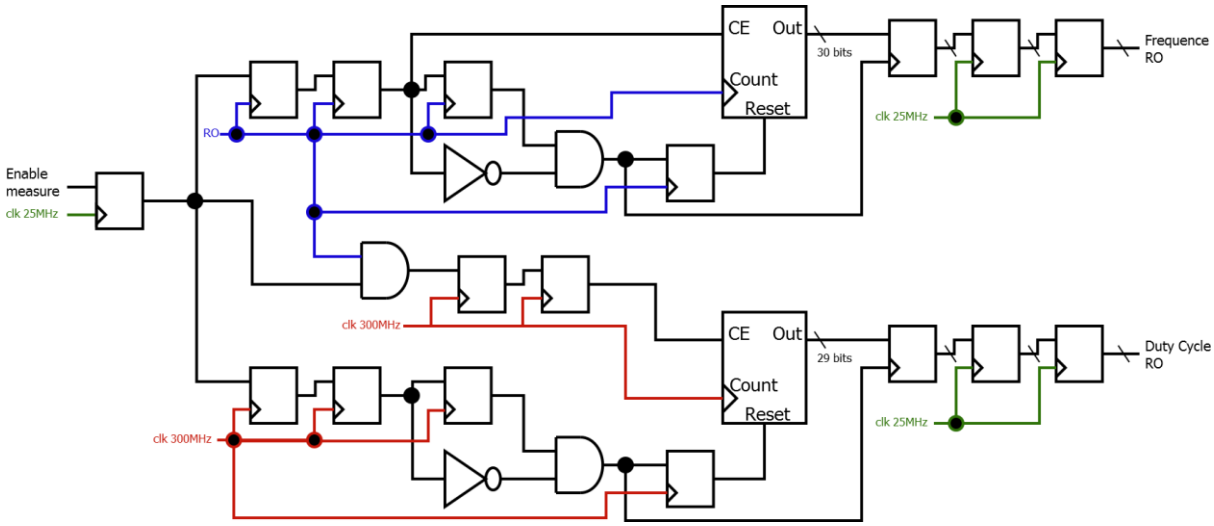


Figure 83 : Schéma détaillé du circuit de mesure des RO

La Figure 84 présente l'évolution de la fréquence d'oscillations initiale des différents types de RO en fonction de la tension et de la température. Les RO de 17 étages sont bien les plus

lents. Les chaînes de XOR et LUT2 sont également plus lentes que celle en LUT1 – pour un même nombre d'étages. Même avec une tension d'alimentation V_{nom}+50%, les fréquences maximales d'oscillations libres observées ne dépassent pas cette limite (300 MHz) comme on peut le voir en Figure 84.La marge entre la fréquence de fonctionnement maximale théorique de l'architecture et les fréquences d'oscillation des RO à froid est limitée. Mais ce n'est pas un problème dans la mesure où la contrainte donnée au logiciel de placement et routage - fourni par le fabricant - suppose une température ambiante de fonctionnement. En effet, à basse température, les temps de propagation sont raccourcis donc la fréquence de fonctionnement réelle est améliorée par rapport à l'estimation.

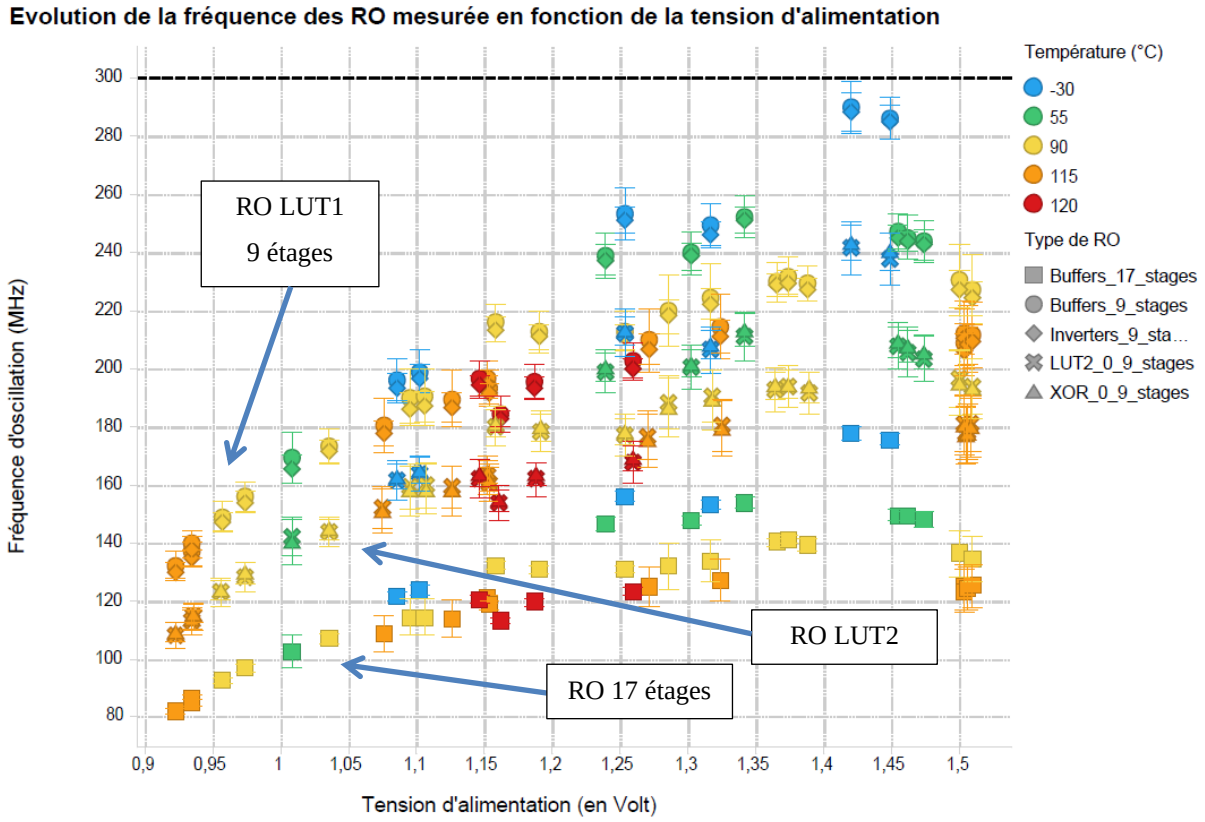


Figure 84 : Répartition des fréquences d'oscillation initiales des RO

Le bruit de mesure du circuit intégré dans les FPGA est majoré par 850 kHz. Pour des oscillateurs en anneaux oscillant à environ 200 MHz, on a donc une précision relative de 0,42% sur les mesures de dérives.

3.1.4 Description du banc FPGA

La première version du banc de vieillissement est basée sur un FPGA central (un MAX10 d'Altera) chargé de gérer les différentes cartes auxiliaires et la communication avec les DUT. Elle comporte une carte de gestion de l'alimentation permettant de régler les diverses tensions d'alimentation des DUT. Cette carte peut également gérer le basculement ON / OFF des composants sous test. Une seconde carte embarquant des ADC permet de mesurer les tensions et le courant à la sortie des cartes d'alimentation. Le réchauffage des DUT se fait au moyen de réchauffeurs locaux (un par DUT). La mesure de la boucle de contre réaction est réalisée au moyen d'une PT100 (RTD) entre le réchauffeur et le boitier du DUT. Notons qu'une correction affine est appliquée sur l'asservissement afin d'obtenir une température de jonction et non une température boitier. Les coefficients de cette correction ont été déterminés en utilisant les valeurs renvoyées par le capteur de température interne au FPGA. Les mesures des fréquences des oscillateurs réalisées dans les DUT sont transmises au banc via une communication UART, le FPGA maître se charge uniquement de centraliser les résultats pour les envoyer à l'ordinateur de contrôle. Le signal d'horloge envoyé au DUT est de 25 MHz. Le schéma bloc du banc est donné sur la Figure 85.

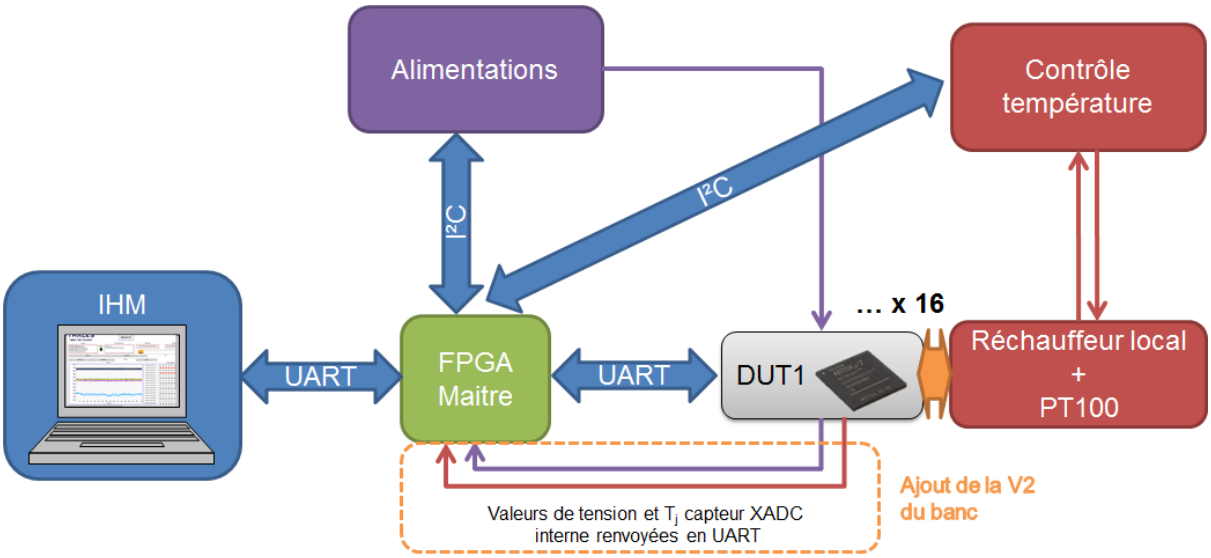
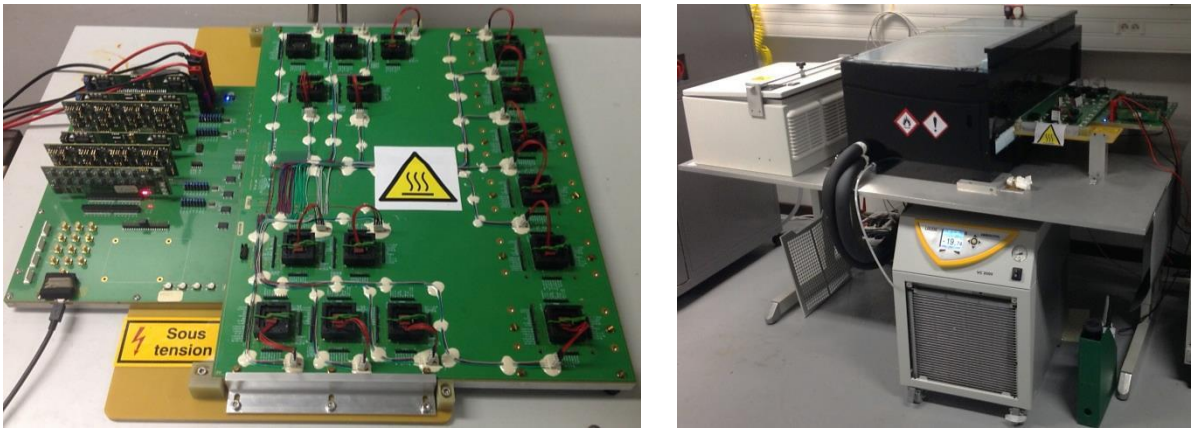


Figure 85 : Schéma du banc d'activation FPGA intégrant les ajouts de la V2 en pointillés

Pour les positions « froides », ayant une température négative, une machine à froid a été mise en place (Figure 86, b). Elle est composée d'un caisson isotherme, de refroidisseurs thermoélectriques à effet Peltier pour refroidir les FPGA, d'un groupe à eau glacée - appelé également « chiller » - pour dissiper la puissance émise par les modules et d'un sécheur d'air

pour évacuer l'humidité dans le caisson. Une température de consigne des modules Peltier à -55°C, ainsi qu'un circuit de liquide de refroidissement des radiateurs des Peltiers à -20°C, permettent d'atteindre des températures de jonction DUT entre -28°C et -35°C.



(a) Banc des positions chaudes (b) Banc des positions froides

Figure 86 : Photo des bancs de vieillissement FPGA

3.1.5 Stratégie d'asservissement en tension et température

Cette matrice de vieillissement a été soumise à différents tests de robustesse et de fiabilité. Un pré-lancement d'une durée de 1000 h a montré quelques faiblesses au niveau de la répétabilité du circuit de mesure. Des modifications ont été effectuées et validées pour corriger ce problème. Ainsi avec cette première version du banc, les tests de vieillissement accéléré ont démarré pour une durée initialement prévue de 1 an.

Les dérives observées semblaient extrêmement volatiles et dépendantes des conditions extérieures aux composants (température du local notamment). Après investigation, il est apparu que la tension d'alimentation - au plus proche du composant - subissait également ces variations. Une résistance de 0,15 Ω est placée en série afin de mesurer la consommation du composant. Cependant, avec une forte suralimentation du FPGA combinée à une haute température d'activation, le courant traversant le FPGA peut aisément atteindre 2 ampères. Cette consommation induit une tension non négligeable aux bornes de la résistance série citée précédemment. Ainsi, la tension d'alimentation du FPGA chute (phénomène appelé « IR drop » d'augmentation de la tension aux bornes d'une résistance série parasite). On obtient par exemple entre 1,2V et 1,45V au lieu de 1,75V suivant la température d'activation. Sur la Figure 87, on observe bien l'impact combiné des variations thermiques et électriques sur la fréquence d'oscillation du RO pris en exemple.

Par ailleurs, la température de jonction obtenue est très dépendante de l'auto-échauffement du FPGA ainsi que de la position spatiale du socket sur la carte de test (due à la forte dissipation thermique des FPGA dans le PCB support). La méthode d'asservissement simple à mettre en application ne tient pas compte des variations de la puissance dissipées par les DUT.

Au bout de 5000 heures de vieillissement, les tests associés à la première matrice à chaud ont été arrêtés. Ce premier essai nous a montré l'importance de la maîtrise fine des conditions de stress lors de chaque mesure. Les données collectées seront traitées afin de compenser ces variations dans la section 3.2.1.

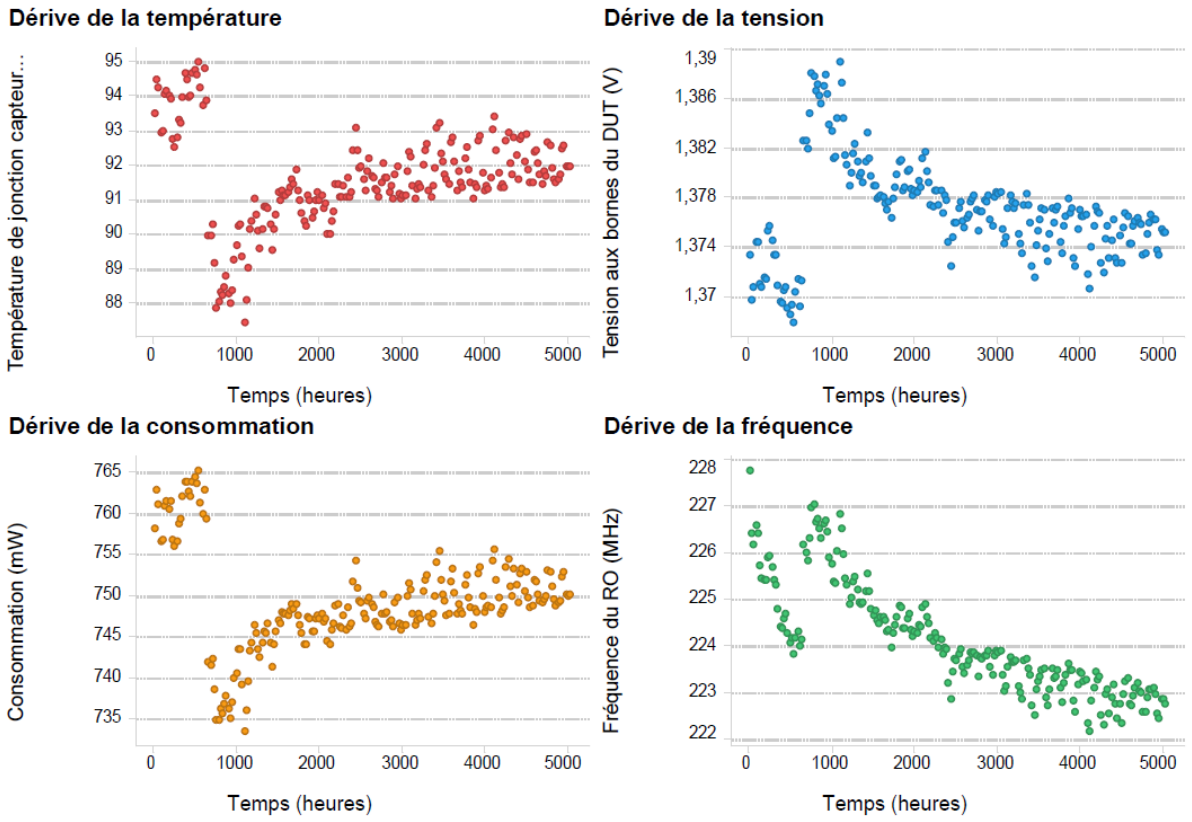


Figure 87 : Impact des fluctuations thermiques sur la fréquence d'un RO

Afin d'éviter d'avoir de nouveau ces écueils, la stratégie d'asservissement a été repensée pour concevoir une seconde version du banc de vieillissement. L'asservissement de la température de jonction est directement basé sur la valeur renvoyée par le XADC (capteur interne de température et de tension du FPGA sous test) au lieu de la PT100 du réchauffeur local. L'asservissement de la tension d'alimentation du FPGA est également en fonction de la valeur renvoyée par le XADC.

La seconde version du banc de test est ainsi architecturée comme illustré en pointillé sur la

Figure 85. De plus, cette matrice bénéficie de l'ajout d'une partie test fonctionnel au design (PLL, DSP, BRAM, IO), et d'un nombre de RO doublé via une meilleure conception de l'architecture.

Caractérisation du capteur embarqué dans le DUT

Selon la notice du FPGA, la précision du capteur de température embarqué est de ±4°C et celle de la mesure de la tension cœur est garantie à ±1% (soit ±15mV dans notre cas extrême de 1,5V). Afin de vérifier que la valeur de température XADC renvoyée par le FPGA sous test ne dérivait pas avec les conditions extrêmes de test - même à haute température et suralimentation électrique - on a comparé cette valeur avec la valeur théorique issue d'un calcul thermique utilisant la résistance thermique $R_{JA\_effective}$ issue de la notice du FPGA (voir équation 92).

$$T_j = T_{ambiant} + R_{JA\_effective} * P_{FPGA} \quad (92)$$

Où :

- T_j est la température de jonction du FPGA en °C
- $T_{ambiant}$ est la température ambiante en °C
- R_{JA} est la résistance thermique équivalente entre la puce et l'air ambiant avec un PCB standard 4 couches en °C/W
- P_{FPGA} est la puissance dissipée par le FPGA en Watt

La notice du fabricant pour le package utilisé indique que $R_{JA\_effective}$ est de 15,4°C/W pour un flux d'air de 6L/s soufflé sur 50 cm². Cette condition de débit d'azote correspond à la tête thermique utilisée dans le laboratoire pour cette caractérisation. A l'aide d'un testeur ATE pour générer les conditions électriques on obtient le Tableau 23. Les valeurs de T_j calculées et celles renvoyées par le capteur du DUT sont très proches. Ainsi, même en suralimentation électrique et à haute température, la sonde de température intégrée au FPGA reste valide.

Température du flux d'azote	Tension appliquée	Consommation	Température renvoyée par le XADC	T_j calculée
25°C	1 V	100 mA	33°C	26,5°C
125°C	1,4 V	1130 mA	150°C	149,4°C

Tableau 23 : Résultat de la caractérisation du capteur de température

Cette non déviation du capteur valide notre nouvelle stratégie d'asservissement qui a par conséquent été mise en œuvre. Elle permet enfin d'avoir des mesures plus stables. Pour ce faire, une meilleure maîtrise de la température de jonction et de la tension aux bornes du

FPGA est appliquée au cours de son vieillissement et surtout lors de chaque phase de mesure de la fréquence des différents RO.

3.2 Traitement des mesures FPGA

La fréquence d'oscillation libre des RO est très sensible aux variations de température ou de tension. Ce point a été abordé dans la section 3.1.5. Si l'on veut observer de faibles dérives, ou bien tout simplement avoir des mesures moins parasitées, il est important de supprimer ce bruit de mesure extrinsèque. L'approche développée dans cette section consiste dans un premier temps à corriger les mesures effectuées en compensant les fluctuations dues aux variations thermiques ou électriques afin de ne garder que la composante « vieillissement ». Par ailleurs, dans la manière d'exprimer ensuite les dérives relatives, la première mesure - qui contient une erreur de mesure - peut fortement biaiser l'allure des dérives associées. Ainsi une méthode sera également proposée pour atténuer ce biais. Une fois ces résultats traités, le passage des mesures des temps de propagation dans tout le RO à un seul étage sera abordé - d'un point de vue critique - via à vis de notre architecture.

3.2.1 Compensation des mesures fréquentielles en fonction de la température et de la tension

Lors de chaque mesure, la tension aux bornes du DUT est enregistrée, ainsi que la température de jonction. Avec la connaissance de ces éléments, on peut ainsi extraire – et donc supprimer – la perturbation environnementale extrinsèque au vieillissement. Cette réduction du bruit permettra non seulement de pouvoir utiliser les données du banc de vieillissement V1, mais également d'affiner celle du banc V2. Ainsi, afin de supprimer ces variations lors des mesures, une caractérisation des variations fréquentielles a été menée sur testeur ATE et tête thermique. Plusieurs couples {Température; Tension} allant de -40°C à 125°C et de 1V à 1,6V ont été caractérisés. Ces plages de valeurs permettent de couvrir l'ensemble de nos conditions de vieillissement. Par rapport à des conditions données (de références), une variation de tension ou de température induit un décalage (addition d'un delta fréquentiel) [118]. Ainsi pour des conditions de références de 25°C et V_{nom}, on peut écrire :

$$\Delta f_{25°C;1V}(T,V)_{\{r;\,s\}} = f(T,V)_{\{r;\,s\}} - f(25°C,1V)_{\{r;\,s\}} \quad (93)$$

Où :

- $\Delta f_{25°C;1V}(T,V)_{\{r;\,s\}}$ est le delta fréquentiel du RO numéro *r* du FPGA numéro *s* entre une oscillation libre à {T ;V} et {25°C ;1V}

- $f(T,V)_{\{r;\,s\}}$ est la fréquence d'oscillation libre du RO numéro r du FPGA numéro s à la température T et à la tension V
- T est la température jonction du FPGA
- V est la tension d'alimentation cœur du FPGA

Ce terme peut être modélisé de manière empirique par un polynôme de degré 2 pour chaque type de RO :

$$\Delta f_{25°C;1V}(T,V) = a_0 + a_1.T^2 + a_2.T + a_3.V^2 + a_4.V + a_5.T.V \qquad (94)$$

Où a_i sont les coefficients du modèle empirique estimés par régression.

Nous avons mené une étude afin de vérifier que l'on pouvait négliger l'influence intrinsèque du composant et la position physique du RO sur la puce dans la modélisation de la grandeur $\Delta f_{25°C;1V}(T,V)$. Il n'y a pas de différence significative dans le comportement des RO d'un même type dans leurs variations en tension/température. Le modèle trouvé est satisfaisant pour notre besoin comme on peut le voir sur la Figure 88 puisque la corrélation entre les valeurs observées et les valeurs prédites par le modèle est bonne.

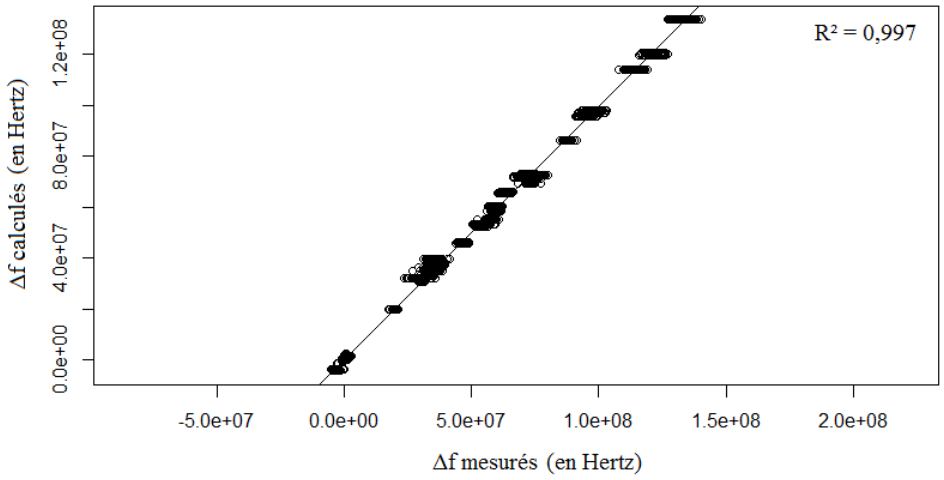


Figure 88 : Corrélation du modèle de compensation des mesures fréquentielles

Le processus de validation – décrit en 1.1.5.1– a été le suivant :

1. Analyse de la forme des résidus : L'analyse des résidus ne révèle aucune forme particulière. Seule l'homoscédasticité (distribution uniforme du bruit) n'est pas parfaite, avec une légère forme en « trompette ».
2. Analyse de la normalité des résidus : Comme souvent en pratique, les extrémités du tracé des résidus du modèle en fonction des quantiles de la loi normale s'éloignent un peu de la distribution théorique. Cependant, ce résultat reste dans notre cas satisfaisant.
3. Analyse des points leviers : On considère qu'un point est levier si sa distance de Cook est supérieure à 1, voir même 0,5 [119, p. 366‑367], [120], [121]. On n'a ici aucun point levier.

Ainsi pour chaque mesure de fréquence d'un RO, en connaissant la tension du FPGA et la température T_j (via capteurs internes), nous pouvons compenser les dérives extrinsèques et se ramener aux conditions moyennes de vieillissement. La Figure 89 présente le résultat de la compensation définie. Le RO considéré ici est issu du DUT de la Figure 87.

La courbe en bleu représente les mesures de fréquences brutes, la courbe verte représente les mesures après compensation des dérives externes. Cette compensation permet ici de réduire le bruit de mesure. De plus, elle permet de considérer des dégradations à tension et température constante. Sans cela toutes modélisations en fonction de ces paramètres seraient moins pertinentes. Au cours du vieillissement, on observe des dérives sur la consommation des DUT ce qui induit une variation de la tension via l'IR drop.

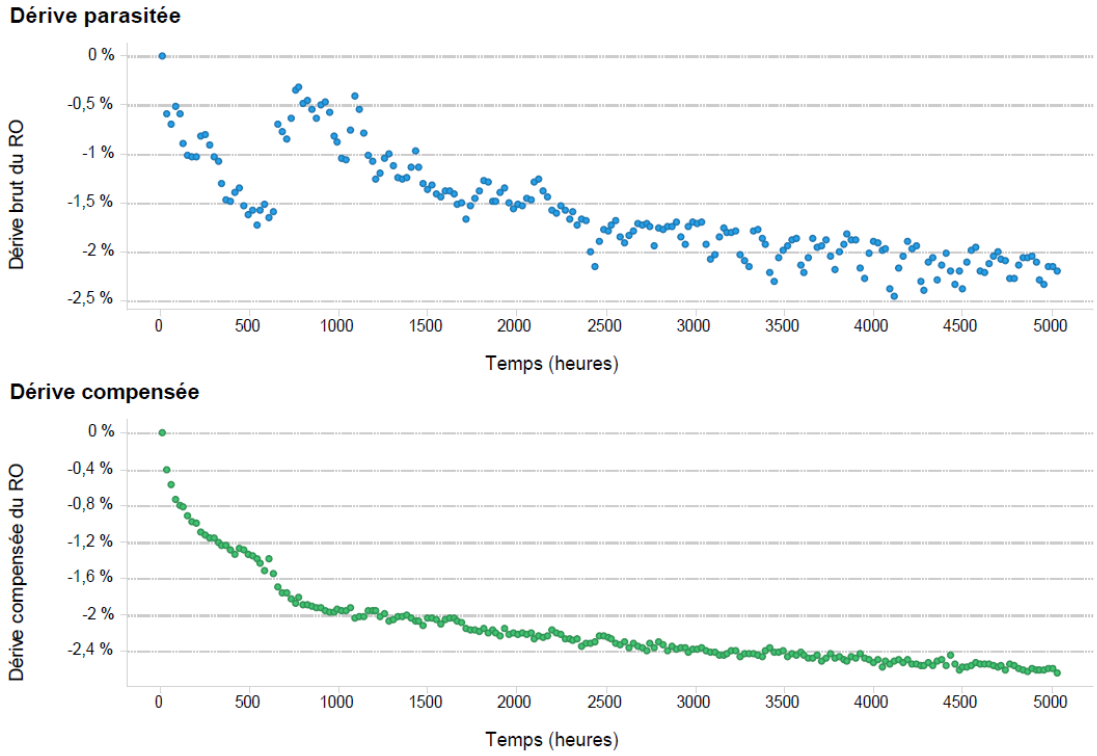


Figure 89 : Illustration de l'effet de la compensation des mesures de fréquence

3.2.2 Ajustement de la première mesure dans la définition de la dérive

Toutes les mesures de fréquences d'oscillation vont être transformées sous forme de dérive normalisée par rapport à leur valeur initiale. L'expression de ce passage en dérive fréquentielle relative est :

$$vf(t) = \frac{f(t)-f(0)}{f(0)} \quad (95)$$

La valeur de la mesure initiale $f(0)$ conditionne ainsi toutes les autres dérives. Si cette mesure est erronée, toute la suite le sera également. Il est donc primordial de soigner cette première valeur. Lors de la mesure de cette valeur, le bruit de mesure est pourtant bien présent.

L'idée développée dans cette sous-section est de modifier cette première valeur tout en restant dans la marge du bruit de mesure afin d'optimiser une allure en loi puissance des dérives. La première valeur est ajustée dans la marge de bruit estimée afin d'obtenir un écart au carré minimum (méthode des moindres carrés) entre la courbe des dérives mesurées et une loi en puissance.

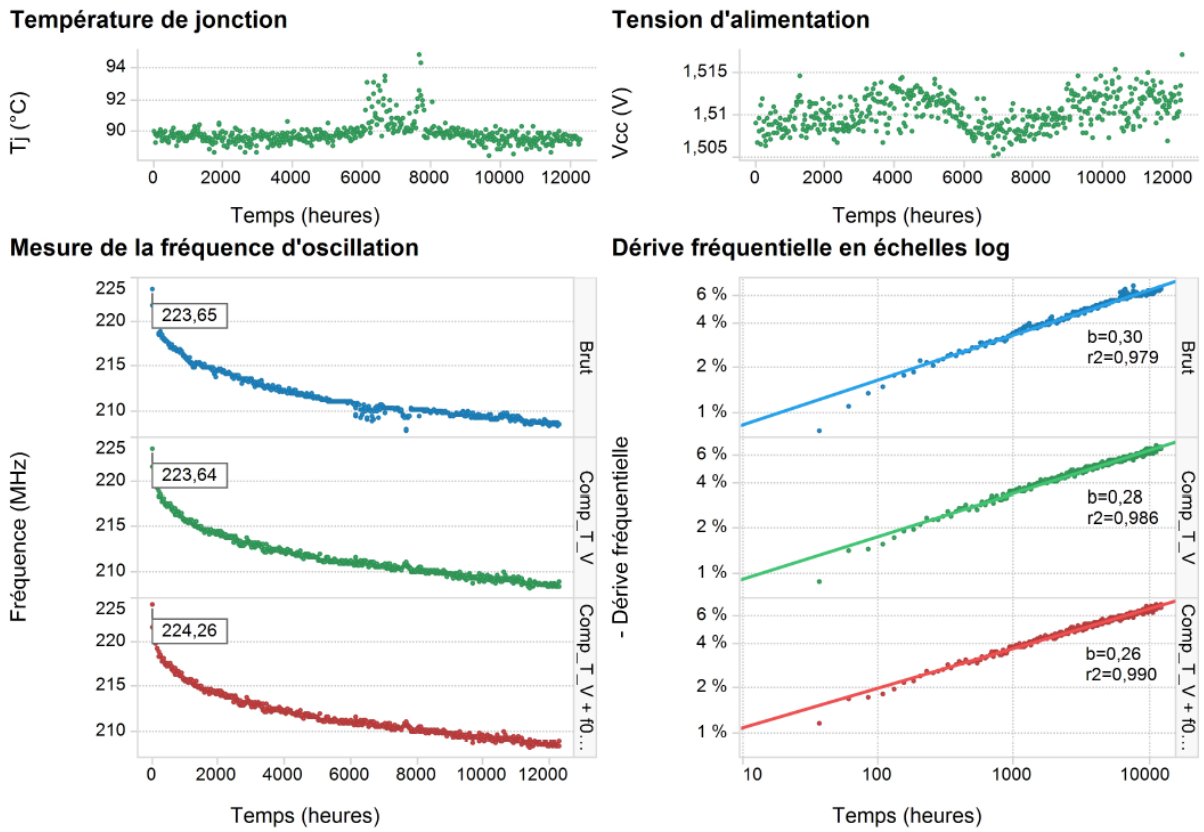


Figure 90 : Traitement des résultats d'un RO

La Figure 90 illustre un cas exemple. Les deux graphiques supérieurs présentent les variations en température et en tension vues par le DUT. Le graphique en bas à gauche présente la fréquence du RO en fonction du temps en échelle linéaire. Le graphique en bas à droite présente quant à lui les dérives en échelles logarithmiques. Dans ces deux graphiques, chaque ligne correspond à une étape de traitement des données. La première ligne « Brut » correspond aux résultats mesurés sans traitement. La seconde ligne notée « Comp_T_V » correspond à la compensation des mesures en fonction des conditions extérieures à chaque

mesure décrite dans la partie 3.2.1. La dernière ligne « Comp_T_V + f0_tunning » correspond à une dernière étape de traitement qui est l'ajustement de la première valeur $f(0)$. Le graphique de l'évolution des dérives représente l'amélioration des résultats au fur et à mesure des différentes étapes du traitement. Cette dernière étape de traitement qui peut paraître insignifiante permet relativement simplement d'éviter un biais de mesure qui a tendance à augmenter artificiellement la pente des dérives, c.-à-d. l'indice de la loi en puissance.

3.2.3 Extraction des temps de propagation par étage

Afin d'améliorer notre compréhension de la dégradation des portes logiques (ici des LUT), il peut être intéressant de différencier le temps de montée du temps de descente. L'avantage de mesurer la fréquence ainsi que le rapport cyclique des RO est que l'on peut donc distinguer les temps de propagation moyens en montée et en descente d'un étage du RO.

3.2.3.1 Chaîne de Buffers

Pour rappel, les chaînes de buffers implémentées sont de la forme présentée en Figure 91. Cette figure comporte également les noms des signaux représentés dans le chronogramme Figure 92.

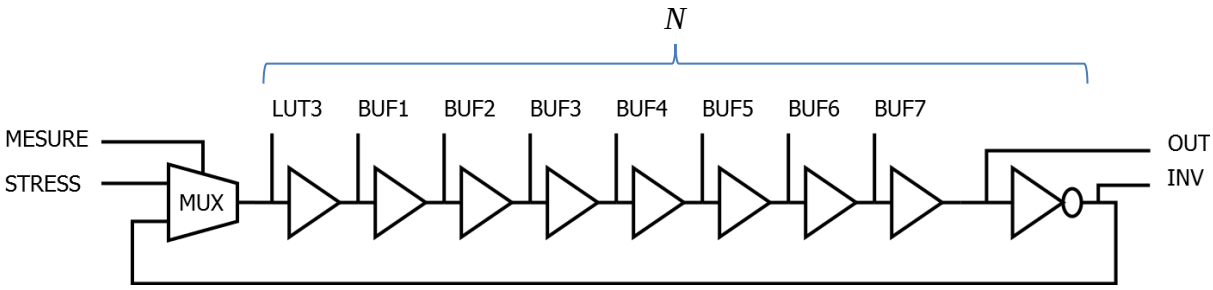


Figure 91 : Schéma d'un RO de buffers

On considère ici que tous les étages Buffer ont le même temps de propagation en montée - au niveau de la sortie de l'étage - tr_{buf} et en descente tf_{buf} . De même, les étages Inverseurs et MUX ont respectivement leur temps de propagation noté tr_{inv} et tf_{inv} ainsi que tr_{mux} et tf_{mux}. Ces différents temps de propagation sont illustrés sur le chronogramme Figure 92.

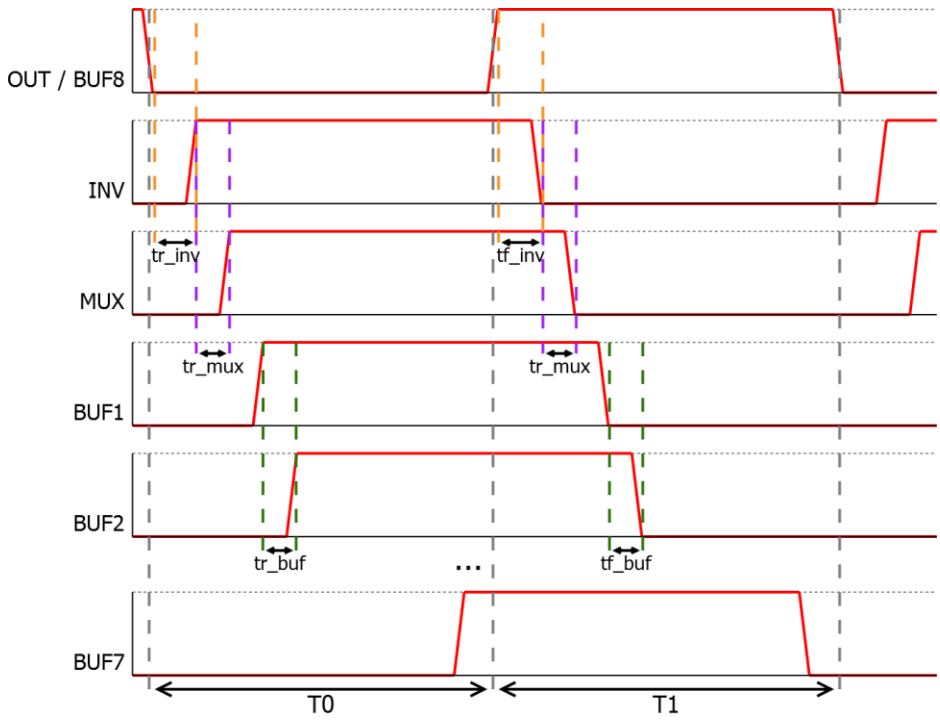


Figure 92 : Chronogramme d'un cycle d'une chaîne de buffers

Le système d'équations 96 permet de calculer le temps de propagation de chaque étage (ici des buffers).

$$\begin{cases} T_0 = (N-1).t_{r_{buf}} + t_{r_{inv}} + t_{r_{mux}} \\ T_1 = (N-1).t_{f_{buf}} + t_{f_{inv}} + t_{f_{mux}} \end{cases} \quad (96)$$

Où :

- T_0 est le temps à l'état bas de la sortie du RO
- T_1 est le temps à l'état haut de la sortie du RO
- N est le nombre d'étage du RO sans compter le MUX

Notons que :

$$\begin{cases} T_0 = \frac{1-r}{f} \\ T_1 = \frac{r}{f} \end{cases} \quad (97)$$

Où :

- f est la fréquence du RO
- r est le rapport cyclique de la sortie OUT du RO

Si l'on considère N assez grand, les différences de temps de propagation entre les Buffers et les étages Inverseurs ou MUX sont négligeables devant le nombre d'étage de Buffer. En faisant l'approximation que tous les étages sont identiques en temps de propagation, on peut exprimer le système d'équations 98.

$$\begin{cases} t_{r_{buf}} \approx \frac{1-r}{(N+1).f} \\ t_{f_{buf}} \approx \frac{r}{(N+1).f} \end{cases} \quad (98)$$

3.2.3.2 Chaîne d'Inverseurs

Dans le cas des chaînes d'inverseurs, la démarche reste la même. La Figure 93 présente la structure du RO inverseur avec la correspondance des noms représentés dans le chronogramme associé (Figure 94).

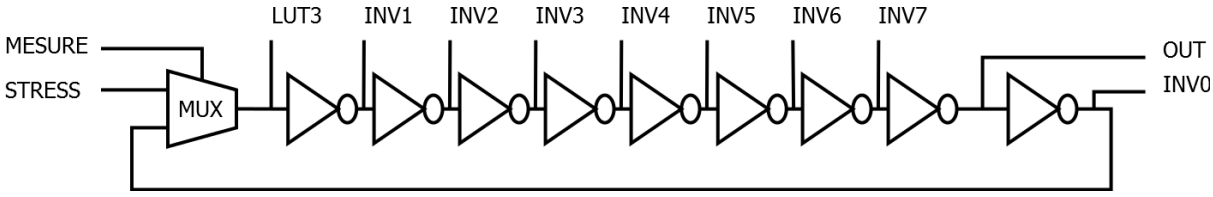


Figure 93 : Schéma d'un RO d'inverseurs

Contrairement à une chaîne de buffer, lors de la propagation de l'instabilité, les étages inverseurs passent alternativement d'une monté (sortie 0 → 1) à une descente (1 → 0). Lorsque la sortie (notée « OUT ») est à '0', il y a 5 (soit $\frac{N+1}{2}$) étages inverseurs qui ont une propagation montante (sur leur sortie) et 4 (soit $\frac{N-1}{2}$) qui ont une propagation descendante. L'étage « MUX » effectue une propagation montante. Puis la sortie du RO passe à '1'. Cette séquence ainsi que la suite du cycle complet d'oscillation sont illustrées sur la Figure 94.

Pour calculer le temps de propagation de chaque étage (ici des inverseurs), on part du système suivant :

$$\begin{cases} T_0 = \frac{N+1}{2}.t_{r_{inv}} + \frac{N-1}{2}.t_{f_{inv}} + t_{r_{mux}} \\ T_1 = \frac{N-1}{2}.t_{r_{inv}} + \frac{N+1}{2}.t_{f_{inv}} + t_{f_{mux}} \end{cases} \quad (99)$$

La différence $T_1 - T_0$ n'est que $t_{f_{inv}} - t_{r_{inv}} + t_{f_{mux}} - t_{r_{mux}}$. Dans ce cas contrairement à la chaîne de buffers, on ne peut pas considérer tous les étages égaux au niveau des temps de propagation. En effet, l'étage MUX a ici beaucoup trop de « poids » quel que soit N.

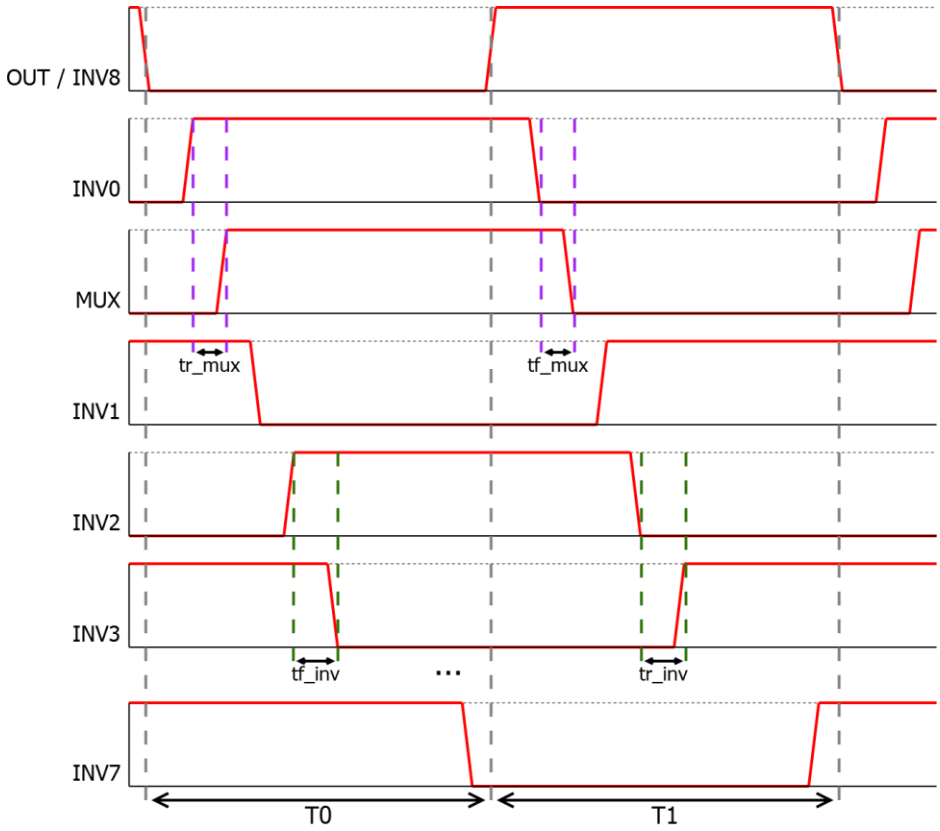


Figure 94 : Chronogramme d'un cycle d'une chaîne d'inverseurs

3.2.4 Effet de la tension sur le rapport cyclique mesuré

Nous avons vu dans la section précédente l'effet de la tension d'alimentation et de la température sur les fréquences d'oscillation initiales des RO (voir Figure 84). De même, la Figure 95 présente les rapports cycliques des RO mesurés en fonction de la tension d'alimentation du FPGA. Chaque couleur correspond à une température de mesure ; et chaque forme correspond à un type de RO.

Pour des tensions d'alimentation « cœur » entre V_{nom} et V_{nom}+45%, les valeurs sont encore relativement proches. Cependant pour V_{nom}+50%, les rapports cycliques mesurés présentent une rupture de pente. Globalement, le rapport cyclique des RO diminue avec l'augmentation de la tension. Cette dissymétrie s'explique par le fait que les concepteurs du FPGA ont prévu les portes logiques comme ayant des temps de commutation en montée et en descente quasi-identiques à V_{nom}. Avec l'augmentation de la tension, les différences de tensions de seuil entre le PMOS et les NMOS s'affirment.

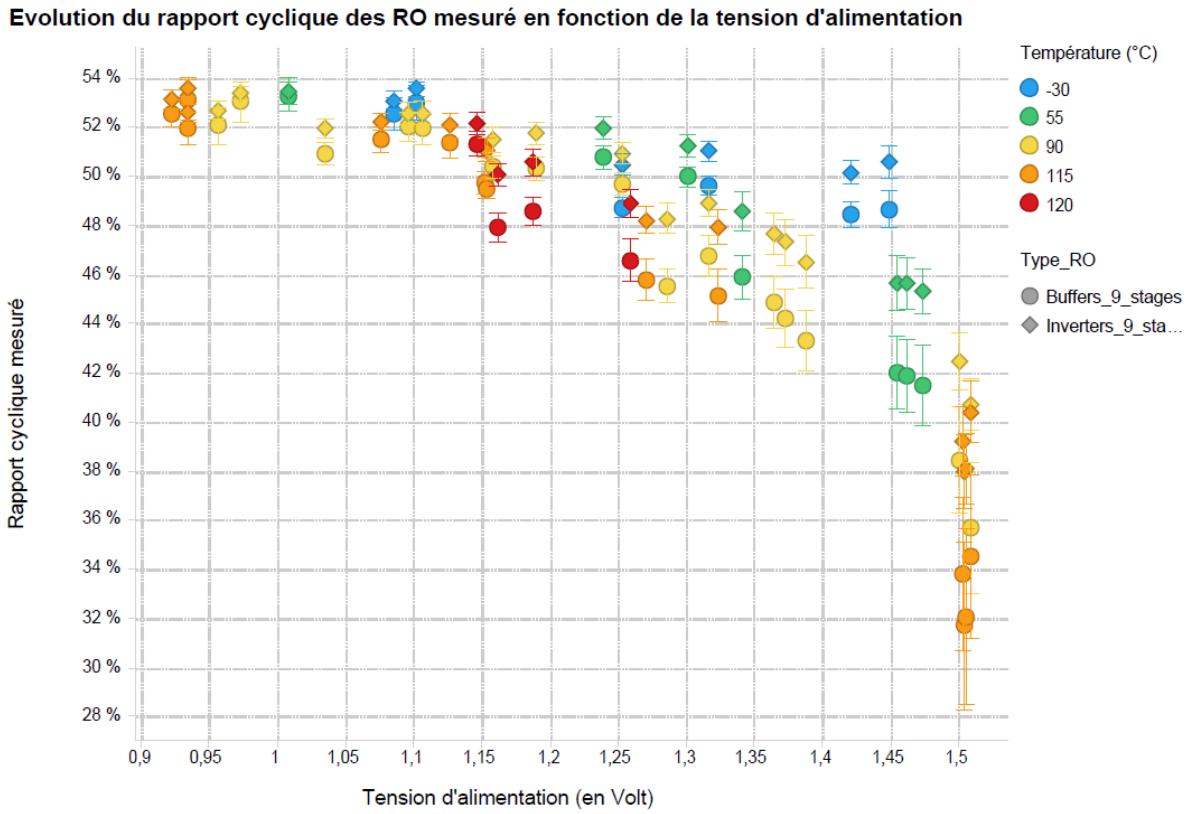


Figure 95 : Evolution des rapports cycliques mesurés initialement en fonction de la tension

3.3 Analyse des dégradations

Le BTI et le HCI apparaissent majoritairement à des conditions différentes : haute température et rapport cyclique élevé pour le BTI ; basse température et haute fréquence pour le HCI. L'objectif développé dans cette section est la présentation des dérives observées et ensuite la séparation des deux mécanismes. Une fois chacun d'eux isolés, la modélisation se fera plus simplement.

3.3.1 Evolution de la consommation des DUT

Dans la plupart des systèmes embarqués, la consommation est un sujet capital. Lors de la conception, le bloc alimentation est dimensionné pour délivrer une certaine quantité d'énergie à l'ensemble du système. Si lors du vieillissement la consommation augmente, l'incapacité de fournir assez d'énergie à l'équipement mènera à une défaillance certaine. Il est donc important d'analyser cet aspect.

La consommation des FPGA augmente légèrement avec le vieillissement. Cette dérive est d'autant plus importante que la tension de suralimentation est grande. Dans la pire condition

(V_{nom}+50% et T_j = 115°C) la consommation relative a dérivé d'environ 5% en 10 000 heures de vieillissement accéléré. Cependant, nos données sont peu précises à ce sujet. Ces mesures ont été réalisées à l'aide d'un ADC et d'une résistance série montée sur la carte. Elles permettent néanmoins d'observer cette tendance.

3.3.2 Etude du BTI

La condition où le BTI est le plus actif est à haute température, haute tension et rapport cyclique élevé - jusqu'au signal constant. C'est pourquoi dans cette sous-section nous allons particulièrement nous intéresser aux DUT du banc de test à chaud V2 (hautes températures et avec des mesures plus précises).

3.3.2.1 Mesures et analyse

Les résultats de mesures issus de cette étude dépendent de nombreux paramètres différents. Le premier défi est de parcourir toutes ces données simplement. Pour ce faire, on considérera les cofacteurs de stress un par un : la température, la tension, le rapport cyclique de stress et la fréquence de stress.

3.3.2.1.1 Influence de la température

Le BTI est un mécanisme analogue à la diffusion qui est aggravé par une augmentation de la température. Pour vérifier qualitativement ce point, nous allons considérer les résultats des RO ayant eu un signal de stress DC. Le facteur d'accélération électrique permet d'aggraver les dégradations pour tous les mécanismes considérés ici. Ainsi les figures présentées seront souvent en régime de suralimentation électrique. Nous allons tout de même prendre le soin à chaque fois de vérifier si les résultats annoncés sont également valables à plus basse tension. De plus, avec un signal de stress DC, il n'y a pas de commutation des transistors pendant les phases de vieillissement, ainsi les dégradations dues au HCI deviennent par conséquent négligeables par rapport au BTI.

La Figure 96 présente les dérives relatives de la fréquence d'oscillation (ordonnées) en fonction du temps (abscisse). Chaque couleur correspond à une température de jonction. Dans cette figure, la tension d'alimentation considérée est V_{nom}+30%. La figure (a) présente les dérives en échelle linéaire de tous les RO. La figure (b) présente en échelle logarithmique la moyenne des dérives en valeur absolues. Les moyennes sont prises sur les différents RO des différents DUT dans les mêmes conditions. Les barres d'erreur correspondent à l'écart type empirique sans biais des valeurs mesurées.

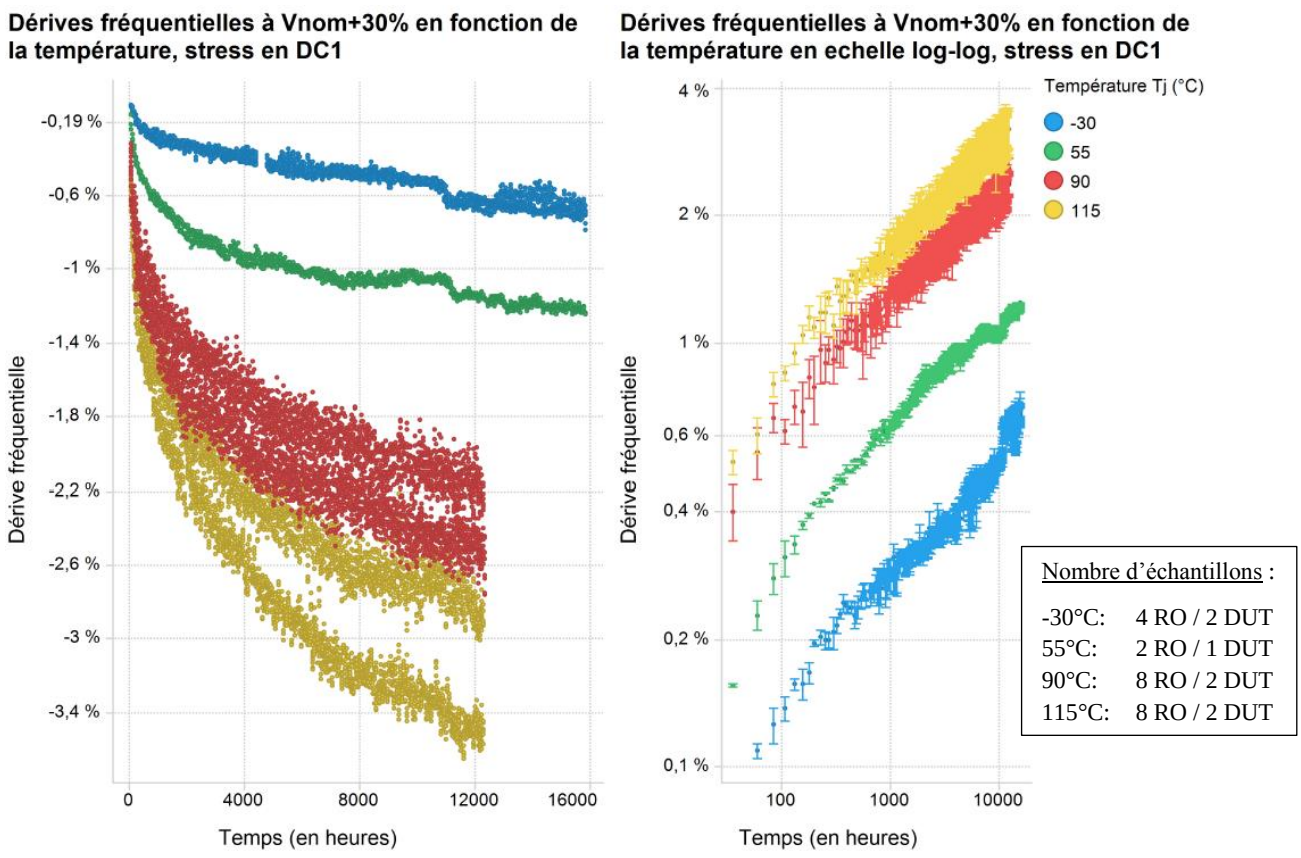


Figure 96 : Dérives fréquentielles des RO stressés en DC1 des DUT à V_{nom}+30% en échelle linéaire (a) et en échelle logarithmique (b)

Sur la sous-figure (a), on distingue deux groupes de points à 115°C (en jaune). Ces deux groupes correspondent à deux DUT où la tension d'alimentation est légèrement différente (1,27 V pour le moins dégradé et 1,32 V pour l'autre). La tension d'alimentation étant un facteur d'accélération important, une faible différence comme celle observée ici peut générer une différence significative des dérives. Ces différences seront bien prises en compte dans l'élaboration du modèle dans les sections suivantes.

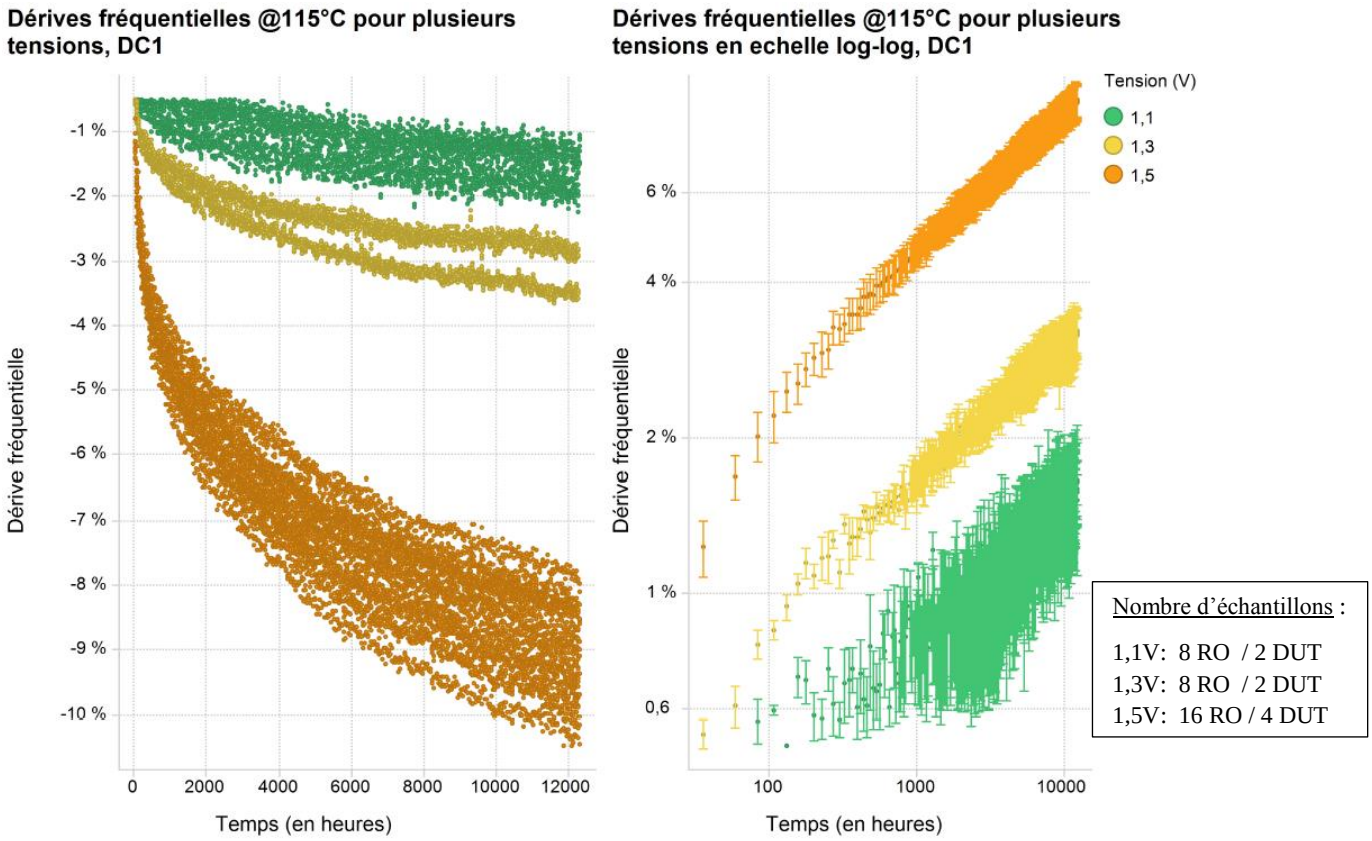
Plus la température de vieillissement est élevée, plus les dégradations sont importantes. Les RO stressés ici à T_j=115°C et V_{nom}+30% atteignent 3,5% de dérive au bout de 12 000 heures ; là où ceux vieillis à température ambiante présentent moins de 1,5% de dérive à 16 000 h. Ces fortes dégradations pour des RO stressés par un signal constant sont vraisemblablement dues à du BTI. De plus, une visualisation des dérives en échelle log-log permet de visualiser le bon alignement des points. Cet alignement confirme la loi régissant

cette dégradation : une loi puissance (avec un exposant compris entre 0,25 et 0,30) comme décrit dans la littérature. Cette représentation nous permet d'estimer une première énergie d'activation comprise entre 0,05 et 0,10 eV.

3.3.2.1.2 Influence de la tension

Différentes tension d'alimentation cœur ont été appliquées dans les files de vieillissement. Ces tensions d'alimentation vont de V_{nom} à V_{nom}+50% afin d'avoir des dérives au-delà du bruit de mesure et de pouvoir également modéliser le facteur d'accélération électrique.

La Figure 97 présente les dérives de la fréquence d'oscillation en fonction du temps. Chaque couleur correspond à une tension d'alimentation cœur. La température de jonction considérée est 115°C. Une représentation en échelle linéaire et une seconde en échelle logarithmiques sont présentées.



(a) (b)

Figure 97 : Dérives fréquentielles des RO stressés en DC1 des DUT à 115°C en échelle linéaire (a) et en échelle logarithmique (b)

La dispersion des mesures à V_{nom}+50% provient des variations de tension d'alimentation d'un FPGA à l'autre. Sur le banc de test, chaque alimentation est distribuée sur deux DUT. Ces deux DUT sont mis dans des conditions thermiques identiques. Le but étant de limiter les différences de consommation qui amènent des écarts d'IR-Drop. Cependant ce système n'est pas parfait : des faibles différences sont constatées mais elles sont néanmoins bien prises en compte.

On observe le phénomène recherché par le plan de test considéré. La tension d'alimentation est un levier majeur de l'augmentation de la dégradation des RO. Cette augmentation de la tension d'alimentation correspond à une augmentation du champ électrique entre la grille et le canal des transistors, ce qui aggrave le BTI. On arrive ainsi dans la pire condition de stress (V_{nom}+50% et T_j=115°C) à 10% de dérive en fréquence au bout de 12 000 heures de vieillissement.

3.3.2.1.3 Influence du rapport cyclique du signal de stress

L'aspect le plus caractéristique du BTI est l'influence du rapport cyclique sur les dégradations. En effet, comme détaillé dans l'état de l'art, le taux de dégradation est directement relié au temps passé en mode « Passant » par les transistors, c.-à-d. au rapport cyclique. Nous allons donc nous assurer qu'ici il a bien une influence significative.

La Figure 98 présente les dérives de la fréquence d'oscillation en fonction du temps. Chaque ligne correspond à une température de vieillissement au niveau jonction. Chaque colonne correspond à une tension de vieillissement au niveau cœur. Chaque couleur correspond un rapport cyclique du signal de stress. Les fréquences du signal de stress considérées ici sont 0 Hz (DC) et 25 Hz.
La dispersion des mesures d'une même pièce est faible. Cela est dû aux conditions de tension et de température maîtrisées et stabilisées d'une mesure à l'autre. La dispersion globale de cette figure provient comme expliqué précédemment des petites variations de tension d'une pièce à l'autre.

La Figure 99 reprend les mêmes caractéristiques que précédemment, mais en affichant uniquement les dérives moyennes parmi les FPGA dans des conditions similaires au cours du temps. L'ensemble des conditions de cette figure suit une dégradation en loi puissance. Les exposants de ces lois sont tous compris entre 0,2 et 0,3 (ce point sera détaillé dans la suite du mémoire). Ces valeurs nous confortent encore une fois dans la bonne observation d'un mécanisme de dégradation type BTI.

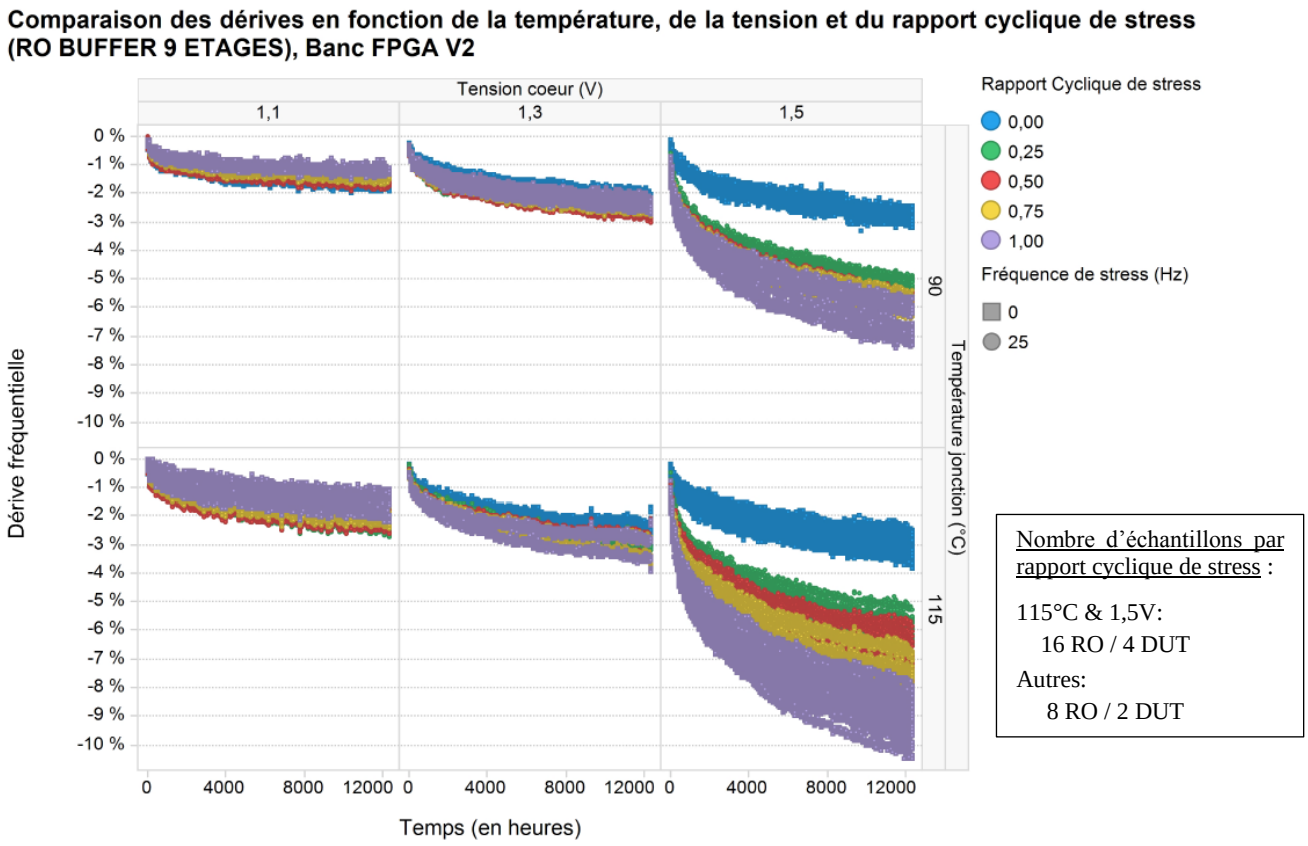


Figure 98 : Dérives fréquentielles en fonction de la température, de la tension et du rapport cyclique

D'autre part, à haute tension plus le rapport cyclique du signal de stress est élevé, plus les dégradations sont importantes. Inversement, à basse tension plus le rapport cyclique est faible, plus les dégradations sont importantes - avec des dérives beaucoup plus faibles qu'à haute tension. La meilleure condition de stress est ici un signal de stress constant comme attendu dans l'observation du BTI.

Ce phénomène peut s'expliquer par l'intervention combinée du NBTI et du PBTI. Un signal d'entrée à '1' n'implique pas que tous les PMOS de la chaîne RO sont activés. Le seul constat que l'on puisse faire est qu'il y a une dissymétrie entre le nombre de PMOS et de NMOS activés lorsque le signal de stress est à '1' ou à '0'. Ces deux mécanismes ayant chacun des comportements physiques différents (différente pente, énergie d'activation thermique et facteur d'accélération électrique), ils ne réagissent pas de la même façon aux conditions de stress. Cependant, ne connaissant pas l'architecture du FPGA, il est difficile de distinguer les deux. Ainsi, cette différence s'explique certainement par des amplitudes de dégradation différentes entre le NBTI et le PBTI, ainsi que des facteurs d'accélération en

tension différents. De plus, en fonction du nombre de PMOS et de NMOS dégradés impliqués dans chaque porte logique, les temps en montées et en descentes deviennent inégaux avec le vieillissement.

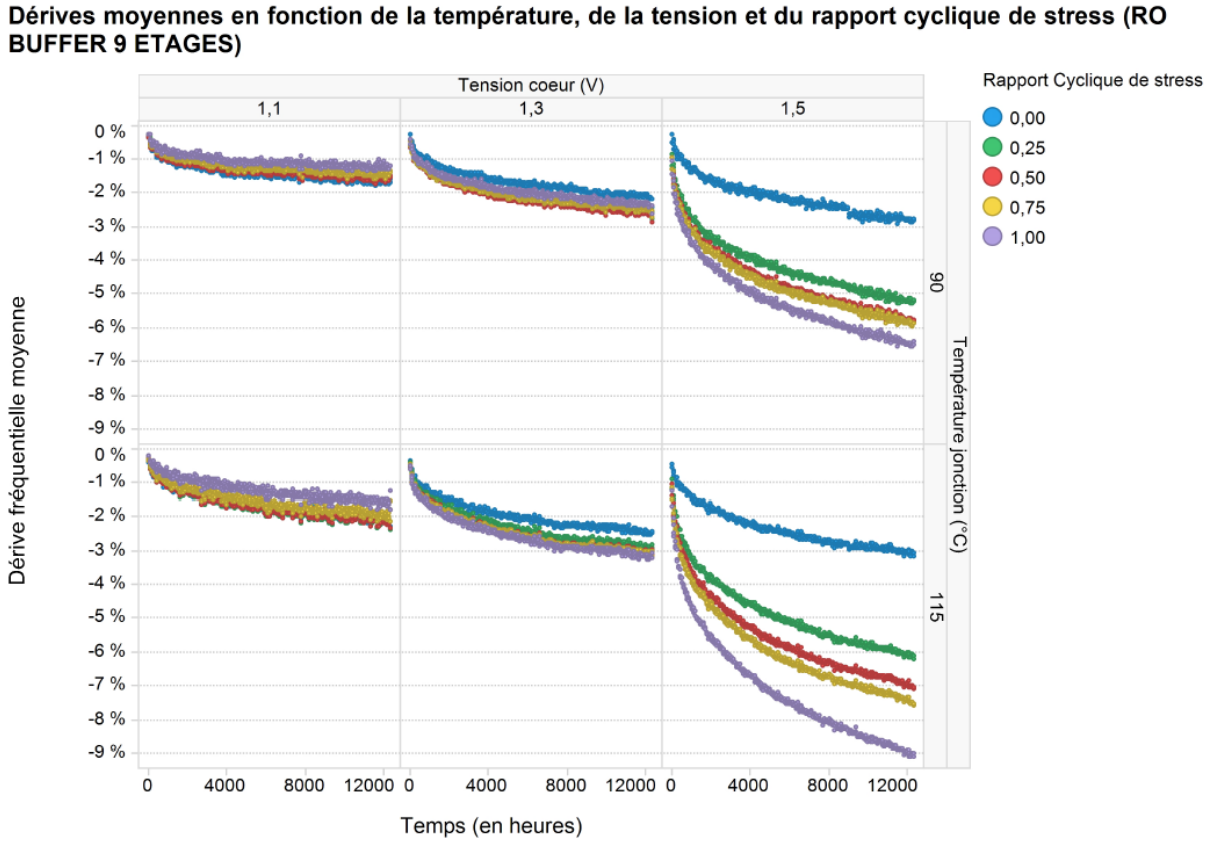


Figure 99 : Dérives fréquentielles moyennes en fonction de la température, de la tension et du rapport cyclique

L'analyse des différentes structures de RO implémentées dans le plan de test permet également d'appréhender les différences de dérive entre la vision globale du RO et chaque étage qui le compose (de différents types dans cette figure). La Figure 100 présente les dérives de la fréquence d'oscillation en fonction du temps. Chaque couleur correspond à un type de RO. Chaque forme (carré ou triangle) correspond à un rapport cyclique de 0 ou de 1.

Tous les types de RO hormis la chaîne d'inverseurs purs (nommée « Inverters_9_stages ») se comportent d'une manière similaire. Ce constat est valable d'un signal de stress DC0 à DC1. Dans les autres chaînes RO, il y a 9 étages de type « Buffer » et un seul inverseur en fin de chaîne. Ainsi, la plupart des étages sont dans la même configuration par RO durant le stress comme illustré sur le schéma « Buffers_9_stages » de la Figure 101. La chaîne d'inverseurs est quant à elle peu sensible au rapport cyclique. Dans une chaîne d'inverseurs, le nombre

d'inverseurs ayant une sortie à '0' et d'inverseurs ayant une sortie à '1' sont quasi identiques à un près. Cette différence est schématisée sur le RO « Buffers_9_stages » de la Figure 101. Ceci peut expliquer la légère différence de dégradation entre les RO stressés à DC0 et ceux à DC1 représentés par les points rouges sur la Figure 100.

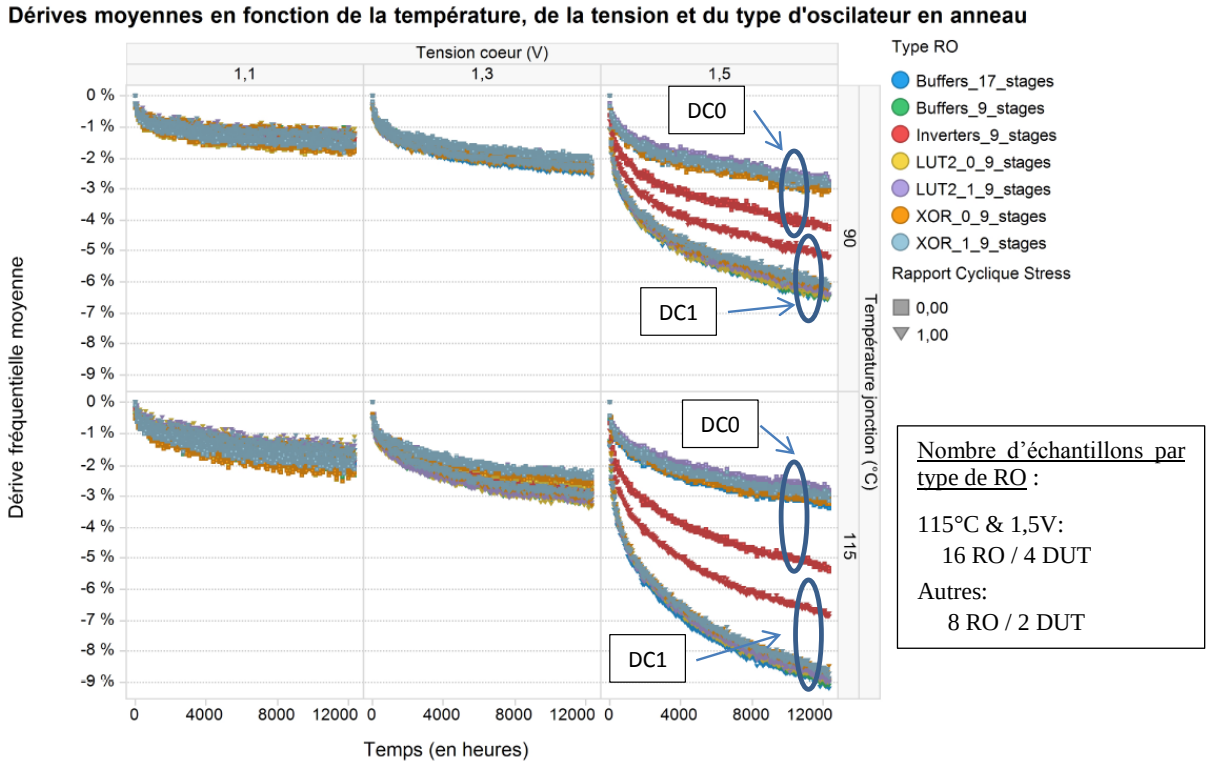


Figure 100 : Dérives fréquentielles moyennes en fonction de la température, de la tension et du type de RO

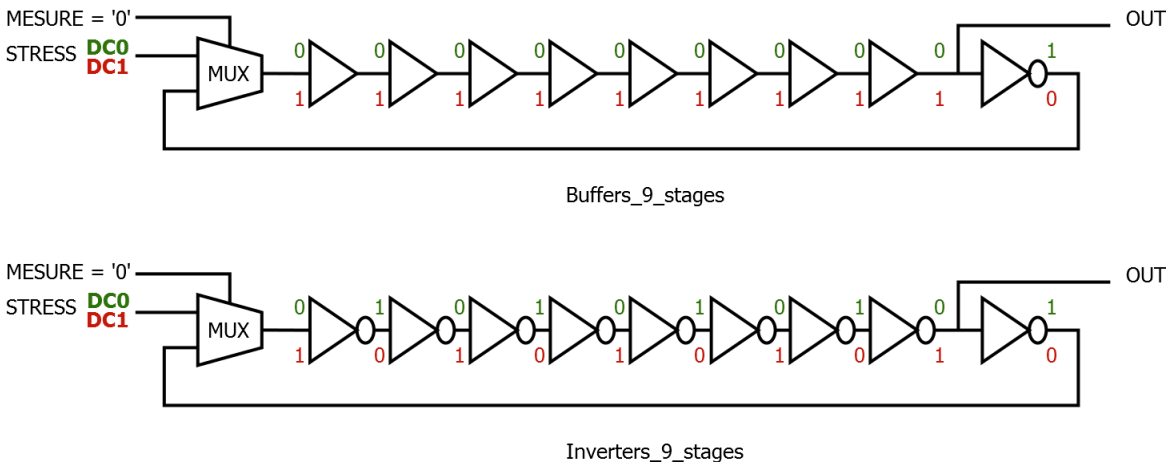


Figure 101 : Différence entre le stress DC0 et DC1 sur les RO de type buffer et de type inverseur

3.3.2.1.4 Influence de la fréquence de stress

Plusieurs fréquences de stress ont été implémentées dans ces vieillissements. Malgré le fait que les signaux de stress constants DC sont les plus représentatifs du BTI, l'analyse des fréquences de stress dynamiques est importante. En effet dans la réalité il y a très peu d'applications numériques – hormis les mémoires volatiles SRAM – restant toujours dans le même état. Une fréquence de stress de l'ordre de 50 MHz est bien plus proche de l'utilisation usuelle d'un circuit numérique.

La Figure 102 présente les dérives de la fréquence d'oscillation en fonction du temps en heures. Chaque ligne correspond à une gamme de température de vieillissement au niveau jonction. Chaque couleur correspond une fréquence de stress. Le rapport cyclique du signal de stress considéré ici est de 50% pour les signaux AC. Toutes les mesures de RO sont ici représentées par des points.

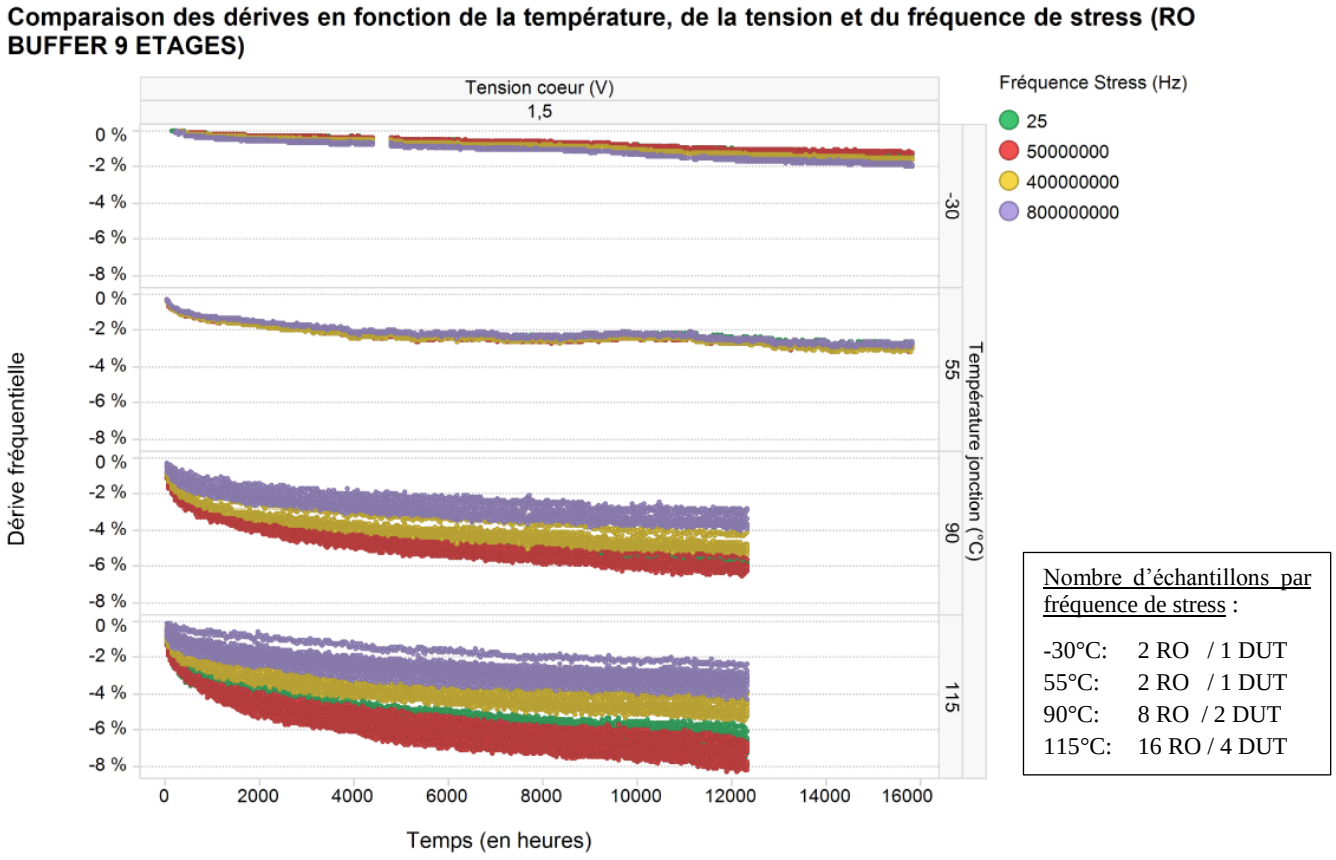


Figure 102 : Dérives fréquentielles en fonction de la température et de la fréquence de stress pour tous les RO à V_{nom}+50%

La Figure 103 reprend les caractéristiques de la figure précédente, mais en affichant uniquement les dérives moyennes parmi les FPGA dans des conditions similaires au cours du

temps. Les échelles verticales varient pour chaque ligne, cela permet de visualiser plus aisément les dérives relatives à basse température.

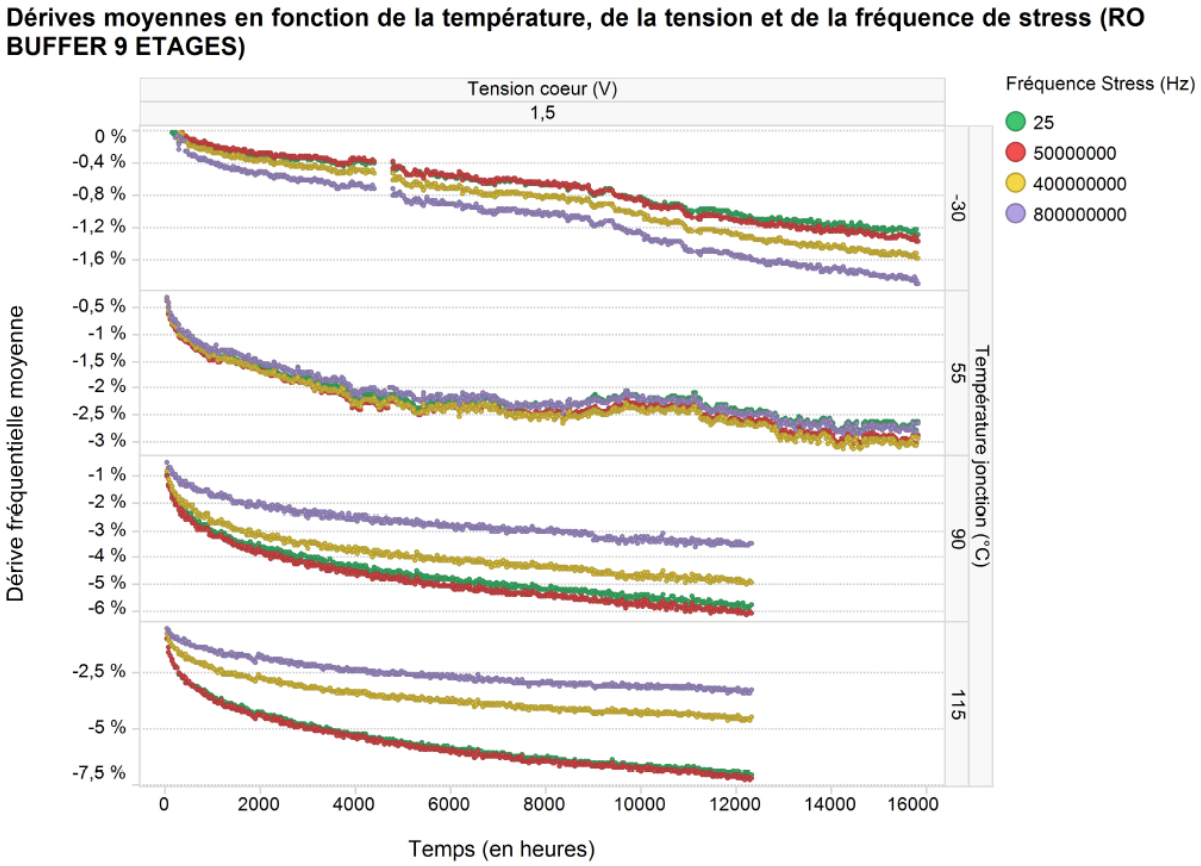


Figure 103 : Dérives fréquentielles moyennes en fonction de la température, de la tension et de la fréquence de stress pour tous les FPGA à V_{nom}+50%

On observe ici encore qu'une élévation de la fréquence de stress a une influence positive sur la dégradation à haute température, mais une influence négative à froid. Ce constat est attendu pour les températures négatives via l'intervention du HCI. Ce point sera traité par la suite dans la section 3.3.3. La diminution des dégradations en fonction de la fréquence de stress semble surtout intervenir lorsque cette dernière dépasse les 100 MHz. Ce phénomène peut s'expliquer par le mécanisme de guérison partiel du BTI lorsqu'il n'est plus activé. La dégradation et la récupération ayant chacun leurs dynamiques propres, une augmentation de la fréquence peut engendrer plus ou moins de guérison. Pendant les phases de stress, la température locale de la puce est plus élevée au niveau des RO activés à 800MHz qu'au niveau de ceux activés à une fréquence moindre. Cette différence thermique peut également en partie expliquer les différences de dégradation constatées. Dans ce sens, une publication

de 2014 [80] mentionne que la guérison du BTI est grandement améliorée – au début tout du moins – en présence de haute température – comme c'est le cas ici.

L'hypothèse avancée ici est illustrée par la Figure 104. Cette analyse va dans le sens de la publication de Wittmann et al. [60] obtenant des dégradations BTI en puissance inverse de la fréquence de stress. La cinétique de guérison est rapide au début puis stagne, selon un modèle en puissance du temps de repos [74]. Les dégradations en loi puissance interviennent à chaque état passant des transistors. Ainsi pour des fréquences plus élevées (courbe en bleue sur la figure), la part de guérison est plus importante et par conséquent les dégradations globales plus faibles.

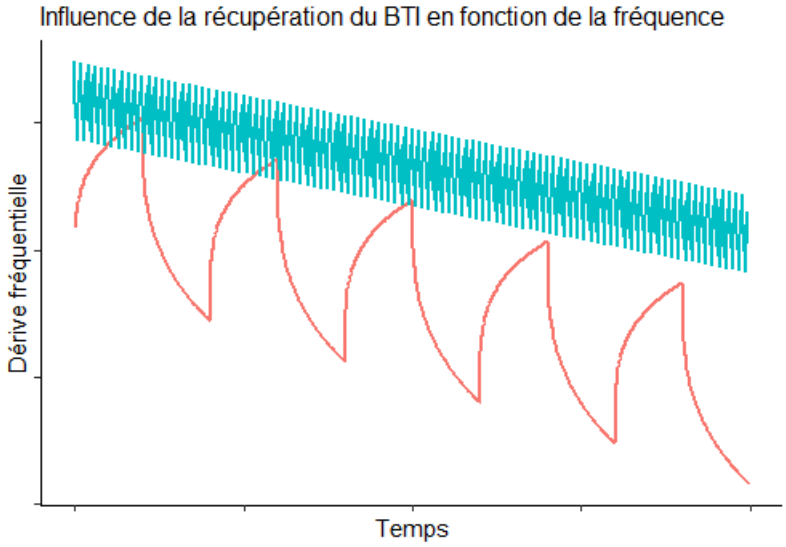


Figure 104 : Illustration de l'intervention de la guérison en fonction de la fréquence de stress

3.3.2.2 Observation des dérives des temps de propagation par étage

La section 3.2.3 a présenté une méthode d'extraction des temps de propagation en montée et en descente des RO par étage. Cette méthode est cependant limitée puisqu'elle moyenne le temps de propagation sur tous les étages. La Figure 105 présente les temps moyens de propagation par étage d'une chaîne de buffer à 9 étages en fonction du temps. La colonne de gauche contient les temps de propagation en montée ; celle de droite les temps en descente. Tous les temps sont ici exprimés en picoseconde. Chaque ligne correspond à une tension d'alimentation. Tous les RO présentés sont à 115°C. Chaque couleur correspond à un rapport cyclique de stress (signal DC ou à 25Hz).

A basse tension, les RO ayant subis un stress DC0 ont leur temps de propagation principalement dégradé en descente ; ceux ayant subi un stress DC1 ont plutôt leur temps en

montée dégradé. Ce constat est cohérent : si les tensions de seuil des transistors augmentent (en valeur absolue) avec le BTI, alors la transition des étages du RO sera plus longue pour retourner dans leur état stressé.

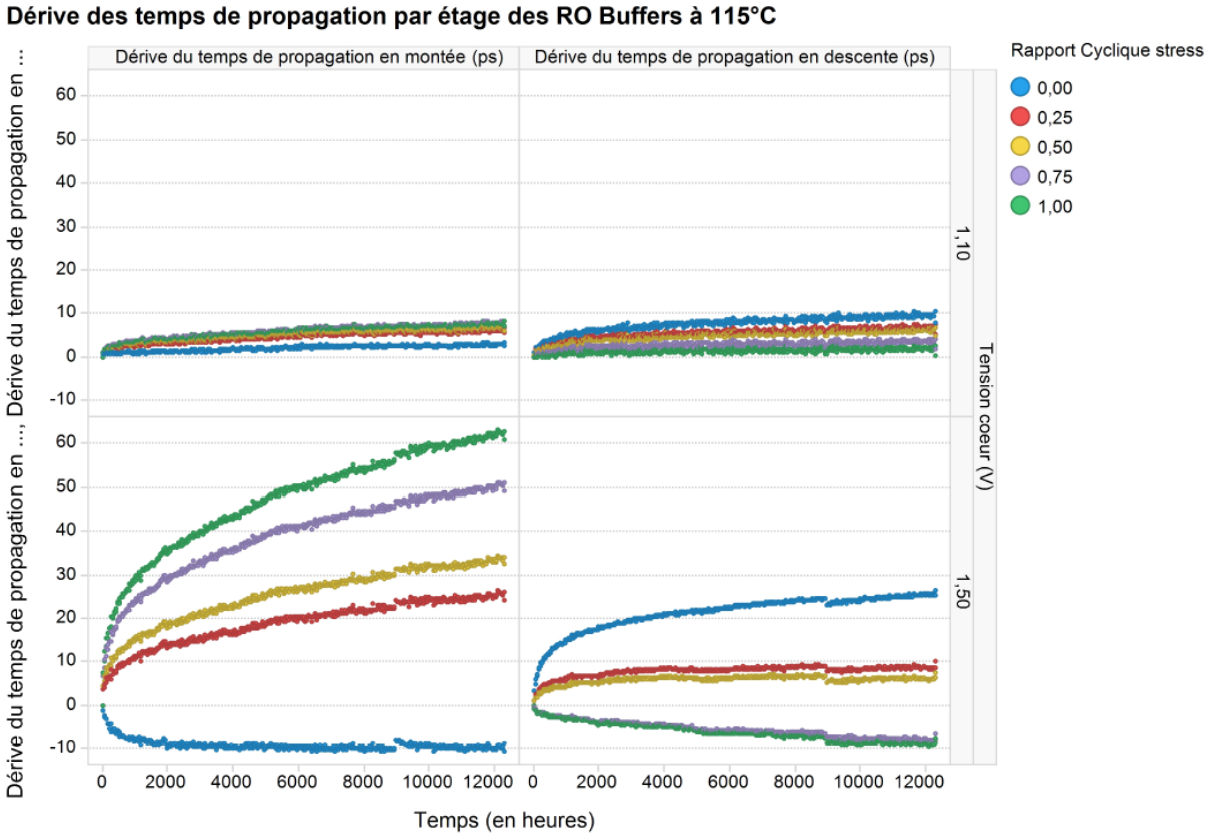


Figure 105 : Dérive des temps de propagation par étage des RO buffers à 115°C

Par exemple, si une porte buffer a vieilli avec sa sortie à '1', la tension de seuil du NMOS a subit une augmentation de sa tension de seuil à cause du PBTI. Ainsi le temps de transition 0→1 sera plus dégradé que la transition 1→0 comme représenté sur la Figure 106 car la porte logique n'est plus symétrique en montée (NMOS) et en descente (PMOS).

Par contre, à haute tension ces résultats sont bien plus contrastés. La dégradation du temps en montée avec un stress DC1 est bien plus importante que celle en descente stressé en DC0. De plus, certains temps de transition s'améliorent même avec le vieillissement. Ce résultat ne signifie pourtant pas que les performances du circuit s'améliorent, bien au contraire. La dissymétrie des temps de propagation en montée et en descente des portes logiques à haute tension (V_{nom}+50%) présenté en section 3.2.4 s'aggrave avec le vieillissement.

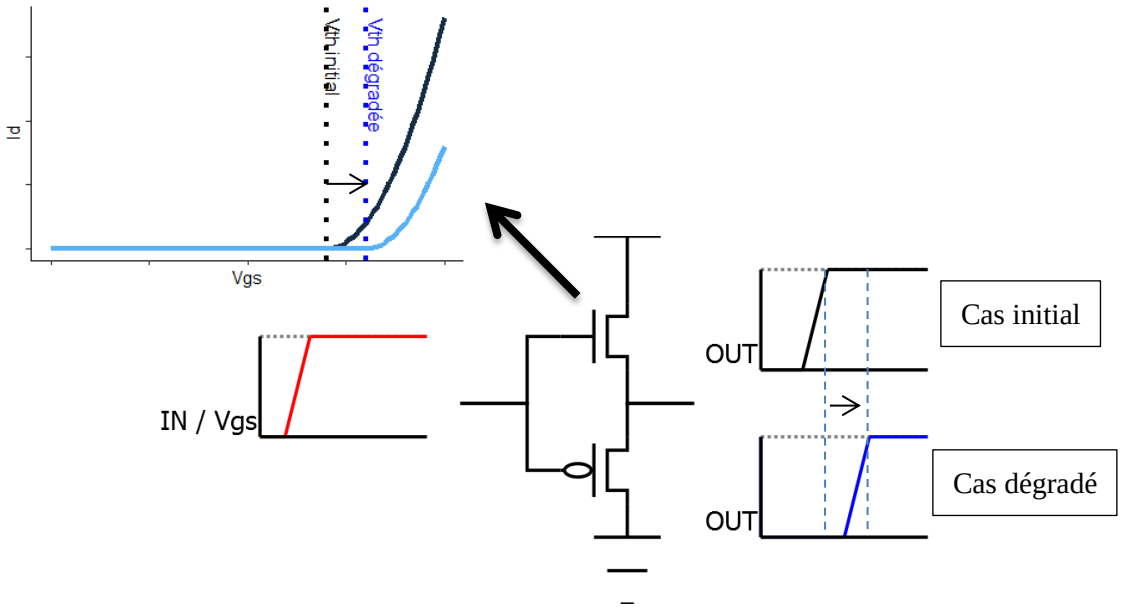


Figure 106 : Illustration de l'influence de la dégradation de la tension de seuil du NMOS sur le temps de propagation d'un buffer

3.3.2.3 Recherche des dérives des temps de commutation NMOS et PMOS au cours du vieillissement

Jusqu'à présent dans ce mémoire, nous n'avons raisonné qu'au niveau RO, voir étage. Or dans un FPGA, un étage – même LUT1 – est en réalité un ensemble de plusieurs dizaines de transistors. La brique élémentaire de ce composant est une architecture LUT6 (2 LUT5 multiplexées pour être précis) utilisée au besoin en LUT1, LUT2 ou LUT3. Nous n'avons pas la connaissance précise de l'architecture de cette brique. De plus, la structure finale d'un étage de RO est également dépendante des entrées de la LUT6 utilisée. En fonction de l'entrée sélectionnée, le temps de propagation sera plus ou moins long. Ainsi chaque étage d'un même oscillateur en anneau est en réalité différent (avec un nombre de PMOS et de NMOS vieilli par conséquent différent).

Afin de distinguer l'impact du NBTI par rapport à celui du PBTI dans les dégradations observées, une méthode va être présentée dans cette section. Le principe général sera de compter le nombre de PMOS et de NMOS directement impliqués dans le vieillissement du RO. Pour cela, nous avons dans la mesure du possible reconstitué la schématique de chaque RO. Elle prend en compte la manière dont on a été implémenté les différentes LUT constituant chaque RO. Nous utilisons pour cela des information extraites du synthétiseur et

un brevet de schématique de LUT déposé par Xilinx (US6809552) [122]. Il reste cependant important de garder ici à l'esprit que cette architecture n'est qu'une hypothèse.

Pour commencer, considérons uniquement la chaîne de buffer décrite en Figure 91 de la section 3.2.3. Elle cascade une LUT3 et neuf LUT1. Sur cet exemple nous allons compter le nombre de PMOS et de NMOS dégradés par BTI. Ensuite la méthode de comptabilisation exacte va être présentée. Pour finir, nous analyserons les résultats issus de cette extraction.

3.3.2.3.1 Présentation du principe de comptabilisation sur un cas simple

Cette méthode consiste à compter le nombre de MOS impliqués dans le vieillissement au travers du brevet Xilinx. Elle sera dans cet exemple uniquement appliquée aux cas des RO constitués de buffers et stressés par un signal DC0.

Dans les slices du FPGA Artix7 de Xilinx, la LUT élémentaire semble être la LUT5. Cependant, dans notre chaîne de buffer, chaque étage est constitué d'une LUT1. On peut donc considérer que la LUT5 est ici configurée pour ne former qu'une LUT1 comme sur la schématique Figure 107. On suppose ici par souci de simplicité que l'entrée utilisée est la première (notée A1 par la suite) même si cela correspond au temps de propagation le plus long. Le chemin de propagation d'un front montant est dessiné en violet, celui d'un front descendant est en vert.

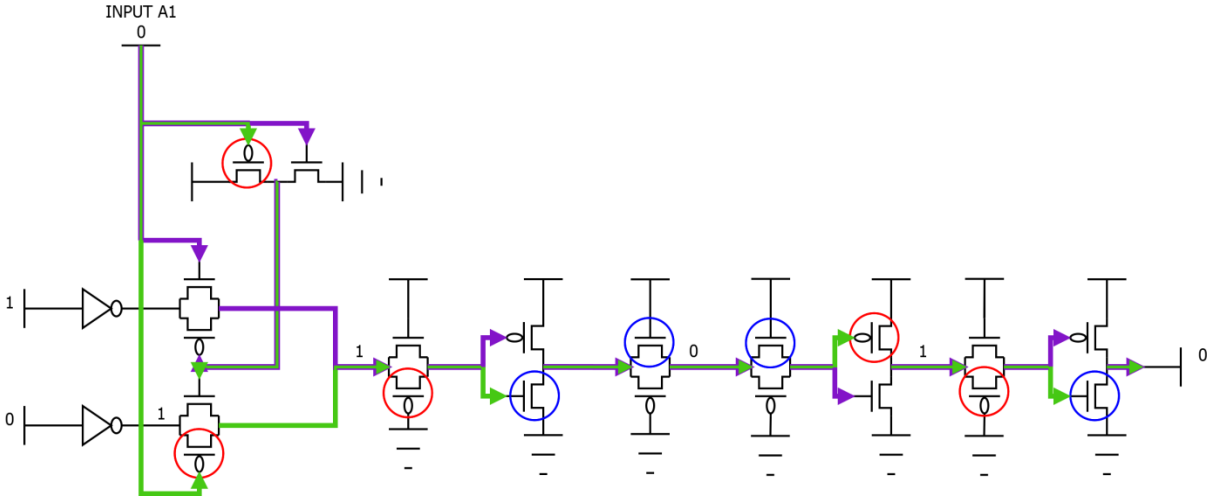


Figure 107 : LUT configurée en buffer subissant un stress DC0

On ne considère que les transistors subissant du BTI et changeant d'état lors des oscillations libres. Ainsi un MOS est comptabilisé comme vieilli s'il est ON pendant la phase de stress, si ses tensions Drain et Sources sont différentes de sa tension de grille et s'il commute de OFF à ON au moins une fois pendant les oscillations libres.

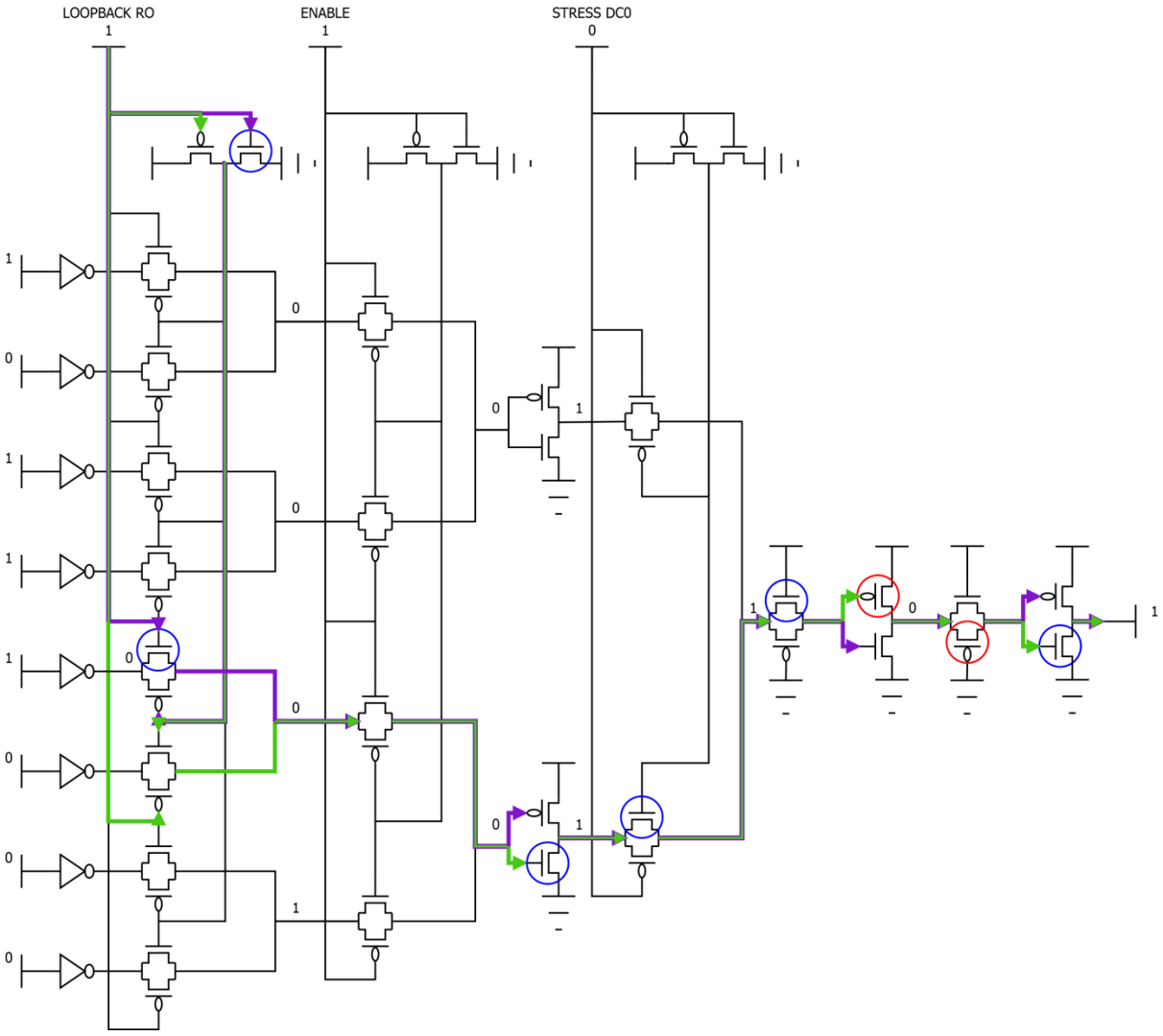


Figure 108 : LUT configurée en MUX subissant un stress DC0 et en mode mesure sur une propagation montante

Dans chaque étage de buffer, 5 PMOS subissent du NBTI (cercle rouge) et 4 NMOS subissent du PBTI (cercle bleu). Ces transistors vont donc rallonger le temps de propagation en montée et en descente de l'étage. Afin d'osciller, la chaine de buffer doit obligatoirement finir par un étage d'inverseur. Le même raisonnement est donc à appliquer à l'étage inverseur.

Pour permuter entre le mode « Stress », ici DC0, et le mode « RO » un MUX a été intégré. Ce MUX composé d'une LUT3 fait partie intégrante du RO et doit aussi être considéré. Les transistors du MUX dégradés par le BTI pendant le stress sont ceux avec les cercles de couleur. Cependant, lors des oscillations, les transistors impliqués dans une propagation montante ou dans une propagation descendante ne sont pas les mêmes. Sur la Figure 108, on

constate en suivant le chemin de propagation violet que seulement 2 NMOS dégradent la propagation en montée à cause du BTI. De même, lors de la propagation en descente (chemin vert) 4 NMOS et 2 PMOS sont responsables d'une dégradation BTI.

Il suffit d'additionner le nombre de MOS dégradés comme présenté dans le Tableau 24. Ce tableau – ici non représentatif de la réalité car il s'agit d'un cas exemple simplifié – nous permettra d'obtenir les dégradations du temps de propagation en montée et en descente pour les NMOS et les PMOS de manière distincte.

Propagation dans le RO sortie OUT	Porte	Transition de la sortie de la porte	NMOS impactés	PMOS impactés
Montante (0 → 1)	Inverseur	0 → 1	0	0
	MUX	0 → 1	2	0
	Buffers (x8)	0 → 1	0	0
	TOTAL	**0 → 1**	**2**	**0**
Descendante (1 → 0)	Inverseur	1 → 0	4	5
	MUX	1 → 0	4	2
	Buffers (x8)	1 → 0	4	5
	TOTAL	**1 → 0**	**40**	**47**

Tableau 24 : Nombre de MOS subissant du BTI dans un RO Buffer stressé en DC0

3.3.2.3.2 Mise en place de l'extraction des circuits exacts pour chaque RO

Maintenant que l'approche utilisée ici a été présentée de manière simplifiée, nous allons pouvoir expliciter la méthode complète. La nuance réside dans la manière de comptabiliser les MOS impliqués dans le vieillissement. Au lieu de faire des hypothèses, nous allons maintenant récupérer dans les résultats de « Place & Route » du logiciel de synthèse les véritables IO utilisées par chaque LUT.

Le logiciel ISE - de Xilinx « Vivado » - est un outil de conception de circuit pour FPGA. Il permet essentiellement d'effectuer les différentes étapes propres à la synthèse de circuit numériques à partir du VHDL sur le canevas matériel. Il est alors possible d'en faire l'implémentation sur les différentes familles de puces fournies par Xilinx. Lorsque l'outil ISE effectue la synthèse et surtout le placement-routage de nos éléments logiques sur le FPGA, les LUT1, LUT2, etc. sont dispatchées dans les LUT5/6 immuables en optimisant les temps de propagation sur les interconnexions. D'après le brevet US6998872B1 de Xilinx [123], les LUT6 natives sont câblées comme sur la Figure 109.

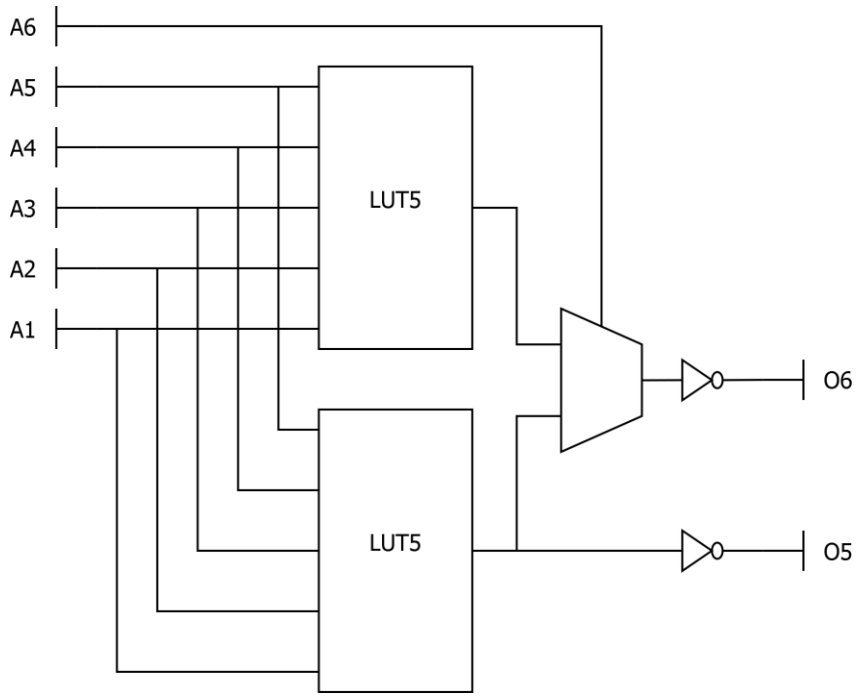


Figure 109 : Schématique d'une LUT6 à partir de 2 LUT5

Ces variations de branchement – notamment sur les entrées – entre deux RO censés être identiques peuvent induire des différences sur le nombre de NMOS/PMOS impliqués dans le vieillissement. Ces différences expliquant par conséquent la variation de la dégradation globale entre deux RO des mêmes types stressés dans les mêmes conditions. C'est pourquoi, une extraction du nombre précis de NMOS et de PMOS vieillis pour chaque RO doit être considérée.

Afin d'estimer au mieux ce nombre de MOS vieillis, le processus décrit sur la Figure 110 a été mis en œuvre. Nous allons parcourir ces différentes étapes. La première est de trouver l'architecture de la LUT6, ce point a été fait – sous réserves – en utilisant des brevets du fabricant.
La seconde étape consiste en l'utilisation de l'ISE afin de récupérer le « plan de câblage » des LUT6 de tous les étages des RO. On obtient ici concrètement la liste des LUT6 utilisées dans nos RO avec leurs entrées associées comme illustré sur la Figure 111. Lorsque deux étages sont hébergés dans une même LUT6 (ils occupent une LUT5 chacun car il y a deux sorties pré-câblées), ils possèdent le même identifiant de LUT6 mais avec des IO différentes.

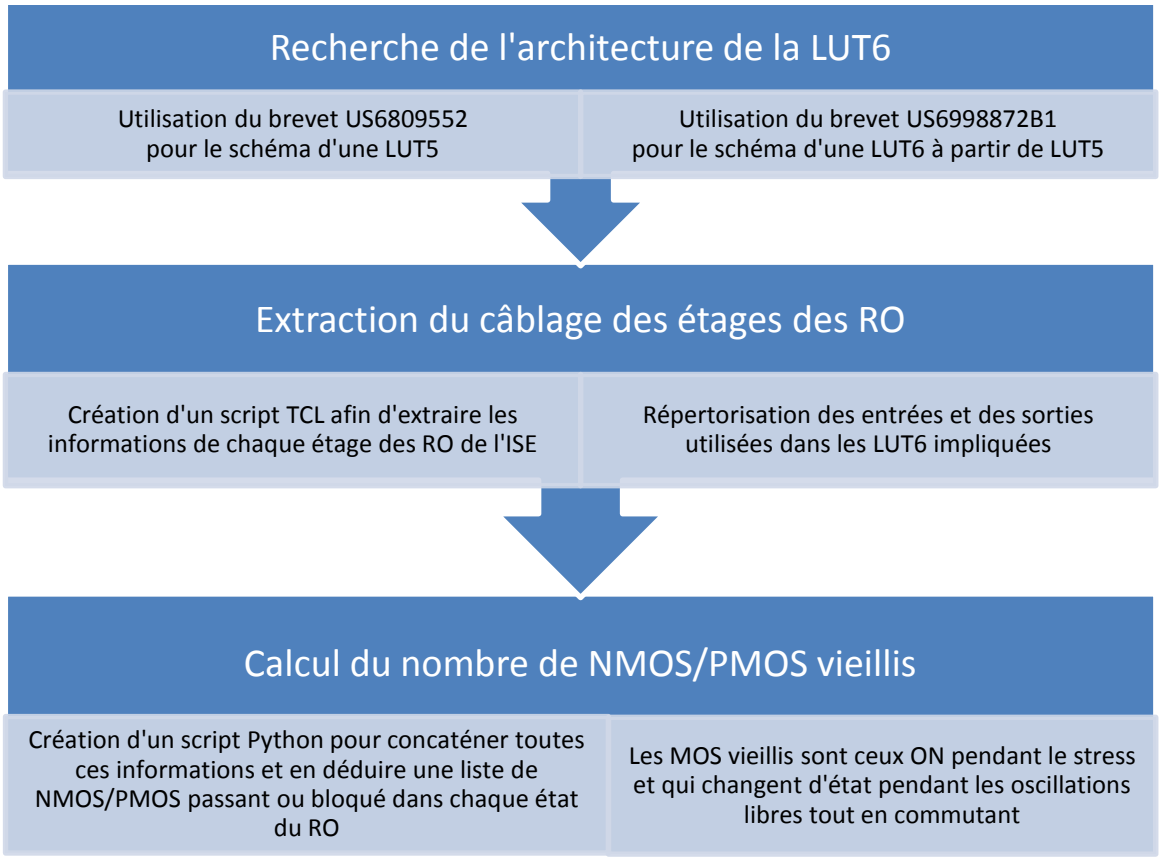


Figure 110 : Processus de la méthode de comptabilisation précise du nombre de MOS vieillis

Type RO	Signal Stress	id RO	NMOS off→on re	NMOS on→off re	NMOS off→on fe	NMOS on→off fe	PMOS off→on re	PMOS on→off re	PMOS off→on fe	PMOS on→off fe
buf9	DC0	1	1	24	24	1	1	37	37	1
buf9	DC0	2	1	24	24	1	1	37	37	1
buf9	DC0	3	1	24	24	1	1	37	37	1
buf9	DC0	4	1	22	22	1	1	34	34	1
buf9	DC1	1	37	1	1	37	24	1	1	24
buf9	DC1	2	38	1	1	38	25	1	1	25
buf9	DC1	3	37	1	1	37	24	1	1	24
buf9	DC1	4	33	1	1	33	20	1	1	20

Tableau 25 : Extrait de la table de sortie script Python

La dernière étape est la concaténation des informations des deux précédentes étapes. A l'aide d'un script nous cartographions les NMOS et PMOS subissant du vieillissement dans chaque RO. Le script écrit et interprété sous Python3.6 permet de générer une table hiérarchisée de la forme *[[« NMOS », « ON », « BTI »], [[« PMOS », « ON », « NO_BTI »], ...]*. Ainsi, une fois cette table générée pour la phase de stress, la phase d'oscillation libre montante et la

phase d'oscillation libre descendante, on peut en déduire le nombre de NMOS et de PMOS vieillis et ayant une conséquence sur les temps de propagation du RO. Un MOS est considéré comme subissant du BTI s'il est passant en mode stress, si ses tensions Drain et Sources sont différentes de sa tension de grille et s'il commute de OFF à ON au moins une fois lors des oscillations libres du RO. Le tableau généré par le script en sortie est ordonné comme sur le Tableau 25.

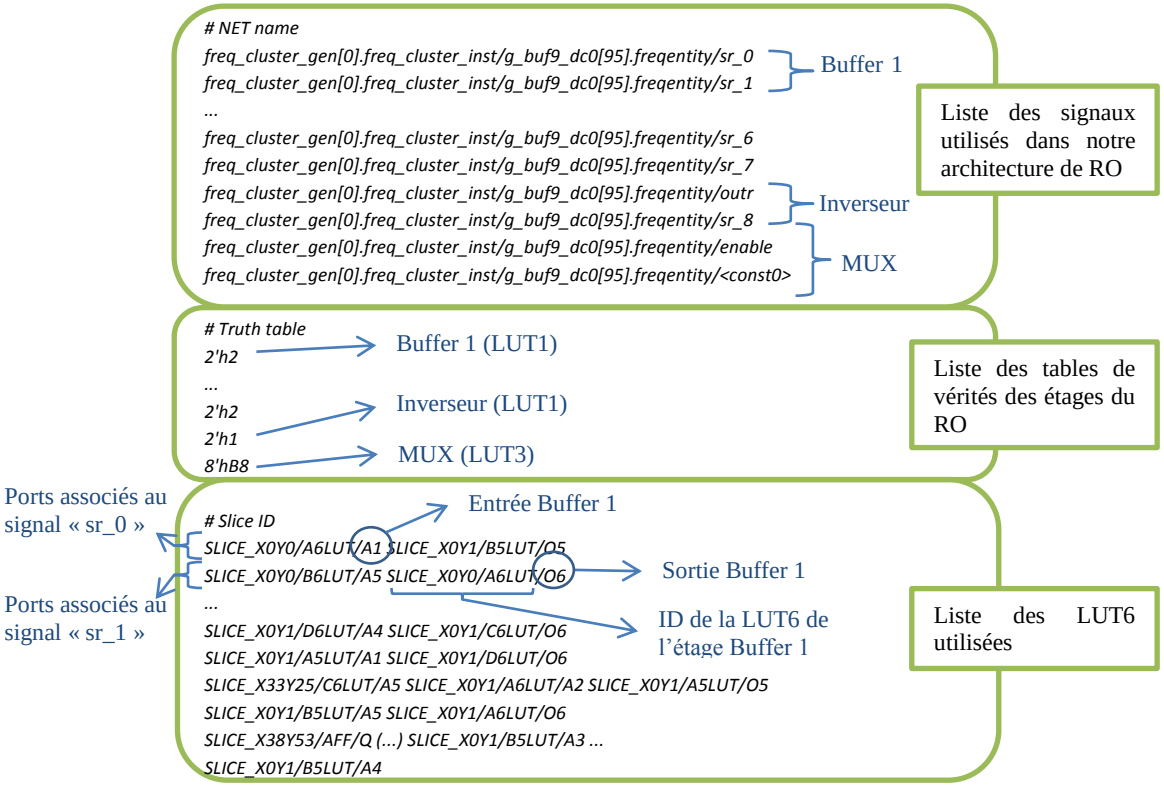


Figure 111 : Extrait des informations de sortie de l'ISE

3.3.2.3.3 Extraction de la dérive sur chaque transistor

Maintenant que nous connaissons le nombre de NMOS et de PMOS dégradés par BTI pour chaque RO, nous allons utiliser nos résultats afin d'en déduire une dégradation au niveau temps de propagation par transistor. Les résultats issus de cette étape sont soumis au fait que l'architecture de LUT6 considérée est bien valide. Cependant, cette démarche apporte néanmoins une méthodologie qui pourrait être réutilisée en parallèle d'une analyse technologique de la structure interne d'un FPGA.

En reprenant les notations T_0 et T_1 de la section 3.2.3, nous pouvons maintenant écrire en connaissant le nombre de N/PMOS vieillis le système suivant :

$$\begin{cases} T_0(t) = T_0(0) + N_{P,ON\to OFF,r}.\tau_{P,ON\to OFF}(t) + N_{N,ON\to OFF,r}.\tau_{N,ON\to OFF}(t) + \\ \qquad N_{P,OFF\to ON,r}.\tau_{P,OFF\to ON}(t) + N_{N,OFF\to ON,r}.\tau_{N,OFF\to ON}(t) \\ T_1(t) = T_1(0) + N_{P,ON\to OFF,f}.\tau_{P,ON\to OFF}(t) + N_{N,ON\to OFF,f}.\tau_{N,ON\to OFF}(t) + \\ \qquad N_{P,OFF\to ON,f}.\tau_{P,OFF\to ON}(t) + N_{N,OFF\to ON,f}.\tau_{N,OFF\to ON}(t) \end{cases} \quad (100)$$

Où :

- $N_{P,ON\to OFF,f}$ est le nombre de PMOS commutant de ON à OFF lors de la propagation du front descendant dans le RO
- $\tau_{N,OFF\to ON}$ est la dégradation temporelle à l'instant t des NMOS lors de la commutation OFF à ON

Nous avons donc 4 inconnues pour 2 équations par RO. Nous pouvons regrouper les différents RO d'un même type en sommant les équations, puis utiliser les RO subissant un stress DC0 ou DC1 uniquement. On obtient donc le système suivant :

$$A \times X = B \quad (101)$$

$$A = \begin{bmatrix} N_{P,ON\to OFF,r,DC0} & N_{N,ON\to OFF,r,DC0} & N_{P,OFF\to ON,r,DC0} & N_{N,OFF\to ON,r,DC0} \\ N_{P,ON\to OFF,f,DC0} & N_{N,ON\to OFF,f,DC0} & N_{P,OFF\to ON,f,DC0} & N_{N,OFF\to ON,f,DC0} \\ N_{P,ON\to OFF,r,DC1} & N_{N,ON\to OFF,r,DC1} & N_{P,OFF\to ON,r,DC1} & N_{N,OFF\to ON,r,DC1} \\ N_{P,ON\to OFF,f,DC1} & N_{N,ON\to OFF,f,DC1} & N_{P,OFF\to ON,f,DC1} & N_{N,OFF\to ON,f,DC1} \end{bmatrix}$$

$$X = \begin{bmatrix} \tau_{P,ON\to OFF}(t) \\ \tau_{N,ON\to OFF}(t) \\ \tau_{P,OFF\to ON}(t) \\ \tau_{N,OFF\to ON}(t) \end{bmatrix} \qquad B = \begin{bmatrix} T_{0,DC0}(t) - T_{0,DC0}(0) \\ T_{1,DC0}(t) - T_{1,DC0}(0) \\ T_{0,DC1}(t) - T_{0,DC1}(0) \\ T_{1,DC1}(t) - T_{1,DC1}(0) \end{bmatrix}$$

Où $N_{P,ON\to OFF,r,DC0}$ est ici la somme des $N_{P,ON\to OFF,r}$ pour chaque RO en DC0 (il y a quatre RO dans la même configuration à chaque fois par FPGA).

En résolvant ce système à chaque instant t, on obtient la Figure 112. Celle-ci représente les dégradations des temps de commutation OFF→ON des transistors impliqués dans le vieillissement du RO. Chaque ligne correspond à une tension et chaque colonne à une température de stress. Les points en bleu correspondent au NMOS, donc aux dégradations dues au PBTI ; et ceux en rouge correspondent au PMOS, donc aux dégradations dues au NBTI. Les transistors considérés ici proviennent des chaînes d'inverseurs et des chaînes de buffers. On ne constate pas de corrélation entre les dégradations et le type d'étage d'origine, cela nous conforte dans l'hypothèse de l'architecture de la LUT6.

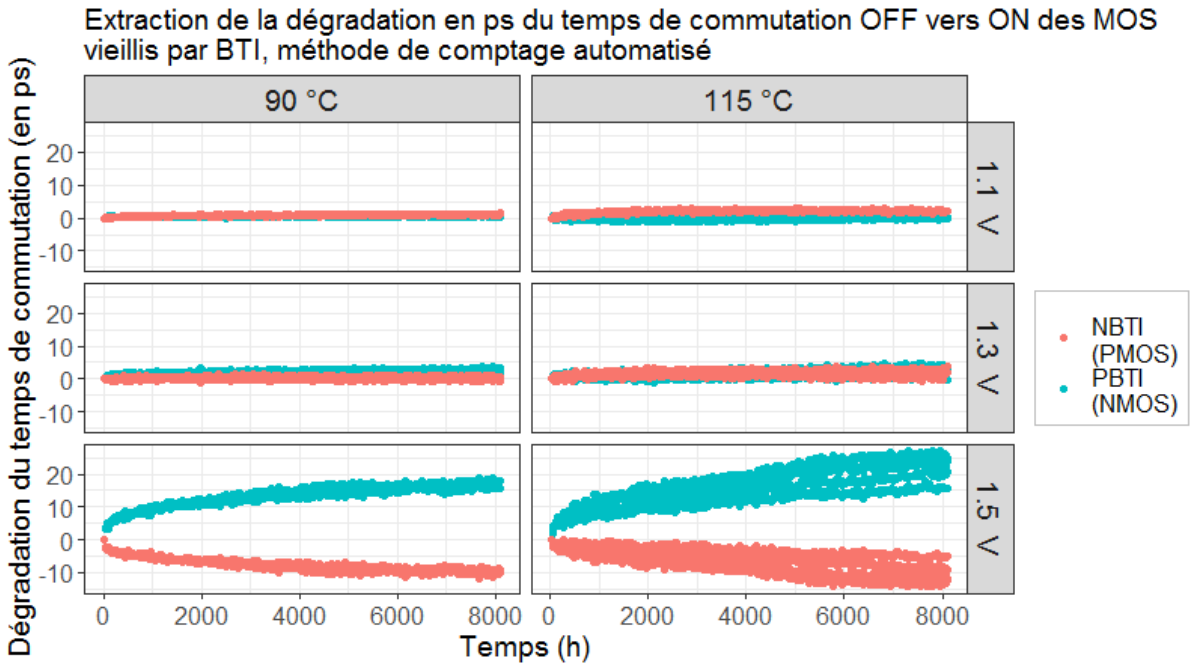


Figure 112 : Dégradation des temps de commutation OFF vers ON des MOS extraite des chaînes de Buffers et d'Inverseurs

Nous observons ici encore une rupture de comportement entre les RO stressés à faible suralimentation et les autres. A basse tension, les dégradations semblent plus importantes sur les PMOS, alors qu'à haute tension ce sont les NMOS les plus impactés. Cela laisse sous-entendre que le NBTI serait dominant à basse tension puis le PBTI deviendrait dominant à haute tension sur un circuit numérique DSM. Cette hypothèse est tout de même émise sous réserves de l'architecture des LUT6 qui est ici supposée. Cette architecture est propriété de Xilinx et nous est partiellement inconnue.

La Figure 113 présente les dégradations des temps de commutation des MOS de ON vers OFF. Dans la plupart des cas, on observe bien une opposition par rapport à la Figure 112. Quand le temps de mise en conduction d'un MOS est dégradé, son temps de coupure est amélioré (sa tension de seuil a augmenté avec le vieillissement). Cependant à forte suralimentation – 1,5V et 115°C – nous observons de fortes instabilités (descente des dégradations puis remontée). Ces variations proviennent très probablement des limites du circuit de mesure du rapport cyclique évoquées dans la section 3.2.4. Pour rappel, ce circuit semble dériver à forte suralimentation à cause de la dysmétrie des portes logiques qui ont été prévues pour fonctionner à 1V. Dans la mesure où nous ne mesurons que l'état haut de la sortie du RO, nous ne pouvons pas comparer avec une mesure de l'état bas pour confirmer ou infirmer ce point.

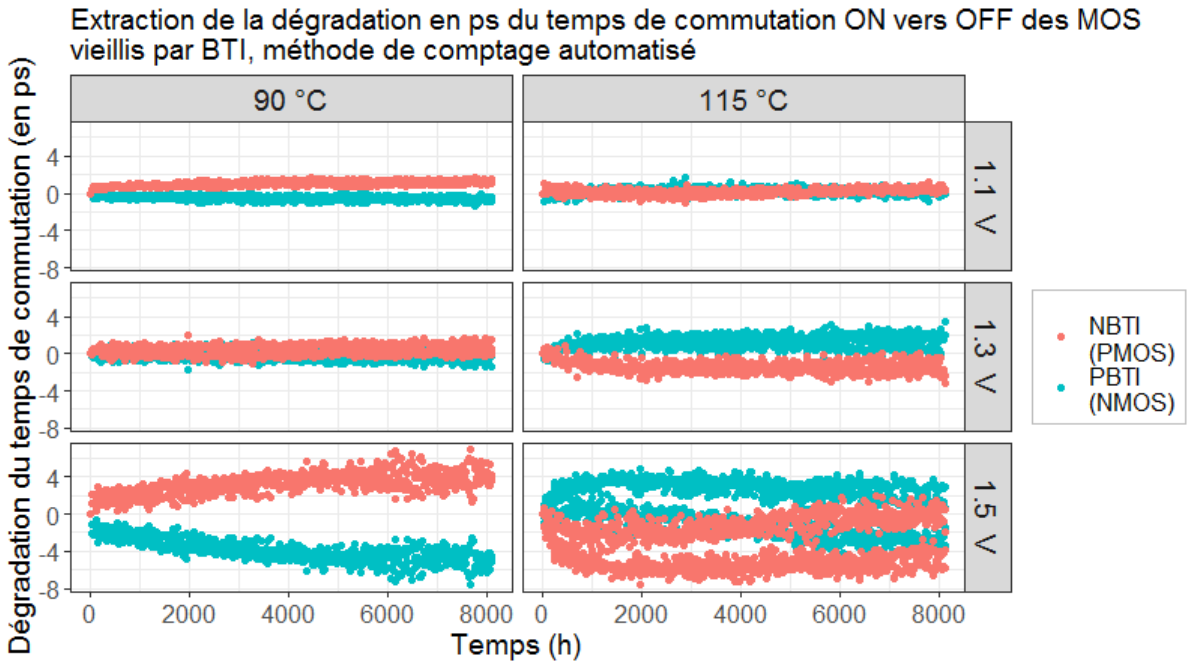


Figure 113 : Dégradation des temps de commutation ON vers OFF des MOS extraite des chaînes de Buffers et d'Inverseurs

Le point le plus important ici est la faisabilité de la méthode proposée. Cette méthode utilisant les données de « Place & Route » permet d'être plus précise en prenant en compte à la fois la structure interne du FPGA ainsi que les optimisations faites par le logiciel. Les résultats obtenus restent en accord avec ceux présentés dans la section 3.3.2.1.3. L'hypothèse où le PBTI deviendrait dominant sur le NBTI à haute tension en technologie DSM est uniquement avancée et justifiée sous l'hypothèse de l'architecture de la LUT6 présentée. A cela s'ajoute l'imprécision du circuit de mesure du rapport cyclique. C'est pourquoi dans la suite de ce mémoire, nous n'allons pas poursuivre avec cette méthode mais uniquement nous focaliser sur la dérive de la fréquence d'oscillation du RO qui est beaucoup plus précise.

3.3.3 Extraction du HCI

Cette section va considérer uniquement le mécanisme de dégradation HCI. Pour ce faire, les résultats obtenus à basse température vont être analysés, puis une méthode d'extraction du HCI sera proposée et mise en pratique.

Le facteur de stress qui nous permet le plus rapidement d'appréhender la présence de dégradation dues aux porteurs chauds est la fréquence de stress des transistors. La Figure 114 présente les dérives moyennes de la fréquence d'oscillation en fonction du temps. Chaque ligne correspond à une gamme de température de vieillissement au niveau jonction. Chaque

colonne correspond à une gamme de tension de vieillissement au niveau cœur. Chaque couleur correspond une fréquence de stress. Le rapport cyclique du signal de stress considérés ici est 50%. Nous avons observé la même tendance pour les rapports cycliques 25% et 75%.

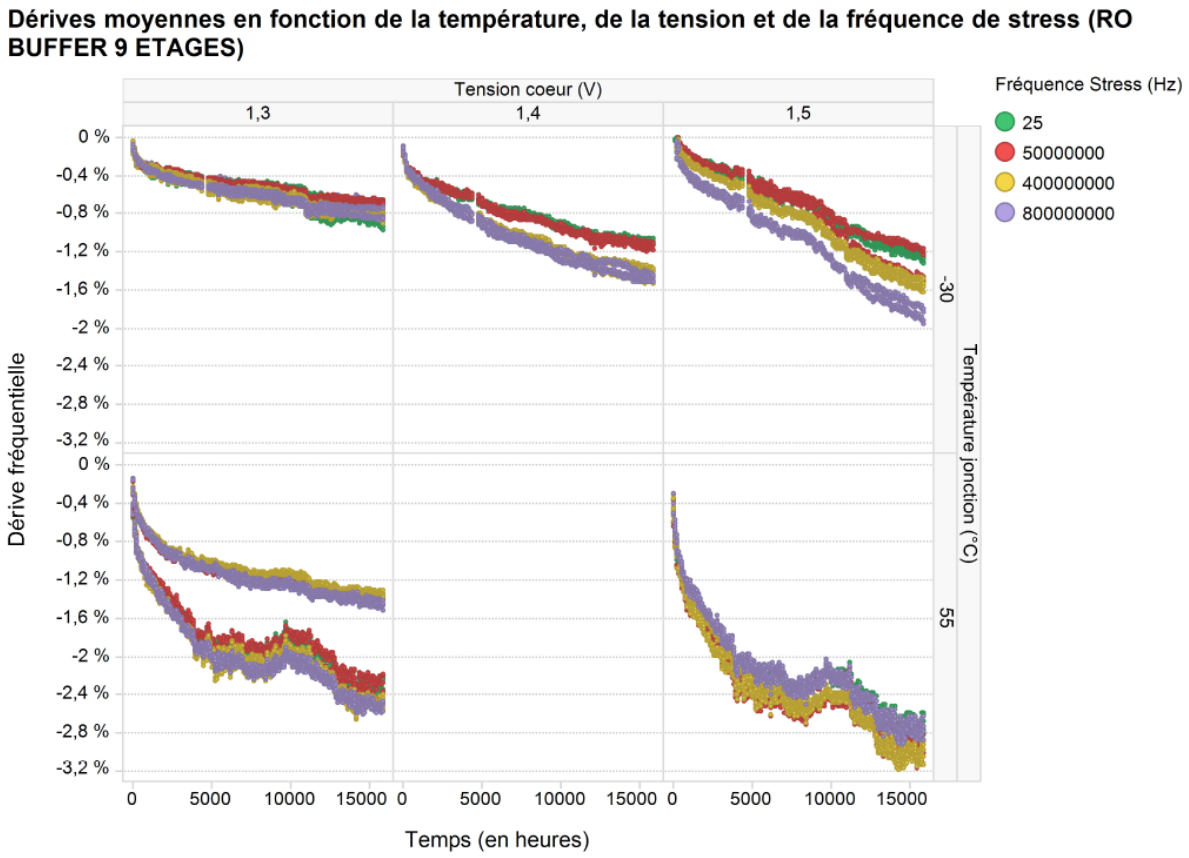


Figure 114 : Dérives fréquentielles moyennes en fonction de la température, de la tension et de la fréquence de stress pour tous les FPGA à basse température

Le BTI est souvent prépondérant sur le HCI dans nos mesures. De ce fait les dégradations du HCI sont diluées dans les dégradations globales. L'extraction proposée dans cette section permet de supprimer les dégradations dues au BTI pour une meilleure analyse du HCI seul. Pour soustraire les dégradations dues au BTI et ne garder que le HCI, on a soustrait les dégradations des RO stressés à 25Hz à ceux ayant une fréquence de stress plus élevée dans les mêmes conditions. En effet, on considère que le HCI est négligeable à 25Hz, là où le BTI est présent. Cette soustraction est résumée par la formule 102. Les dégradations ainsi filtrées sont présentées sur la Figure 115.

$$\upsilon f_{HCI}(\alpha_{stress}, f_{stress}) = \upsilon f(\alpha_{stress}, f_{stress}) - \upsilon f(\alpha_{stress}, 25Hz) \qquad (102)$$

Où :

- vf_{HCI} est la dérive fréquentielle du RO due au HCI
- α_{stress} est le rapport cyclique de stress du RO
- f_{stress} est la fréquence de stress du RO
- vf est la dérive fréquentielle du RO

On retrouve bien le résultat précédant : plus la fréquence de stress est élevée, plus les dégradations sont importantes. Ce constat nous conforte dans le fait d'avoir effectivement extrait des dégradations dues au HCI. L'influence de la tension d'alimentation sur les dégradations est aussi ici un bon signe d'observation de HCI.

Les dérives dues uniquement au HCI sont très faibles ici. Même après 16 000 heures de vieillissement à -30°C, les dérives maximales sont inférieures à 0,7%. Or le bruit de mesure à froid est d'environ 0,3% (avec une fréquence moyenne d'oscillation libre à 275 MHz). On remarque cependant que les mesures sont un peu plus bruitées par rapport à la Figure 114. Ce fait provient de la méthode d'extraction : en soustrayant les dérives du BTI à 25Hz on a également rajouté son bruit associé.

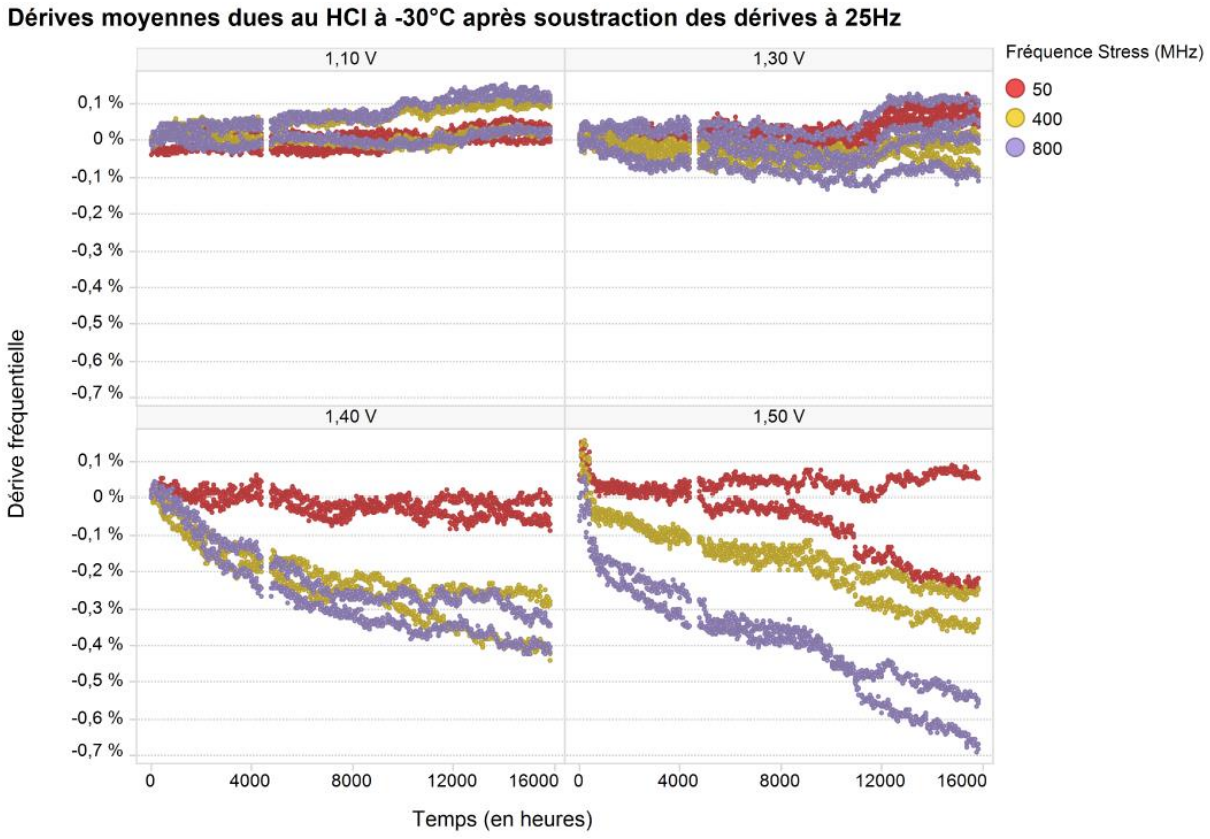


Figure 115 : Dégradation due au HCI des RO de type Buffer_9 avec un rapport cyclique de stress de 50%

Comme précisé dans l'état de l'art, les dégradations dues au HCI doivent être considérées en fonction du nombre de commutations accumulées au cours du temps. Ainsi, le temps sera par la suite multiplié par la fréquence de stress de chaque RO. La Figure 116 présente ainsi le logarithme des dérives en valeur absolue extraites en fonction du logarithme du nombre de commutation subi par le RO. Chaque ligne correspond à une gamme de tension de vieillissement au niveau cœur. Chaque couleur correspond à une fréquence de stress.

Après transformation des abscisses (temps × fréquence de stress), les mesures sont alignées – dans la limite du bruit de mesure – pour les différences fréquences de stress. Les dégradations sont donc ici dues essentiellement au nombre de commutations endurées par les transistors. Les pentes en échelle logarithmique des dégradations observées sont autours de 0,4. Cet alignement témoigne donc bien du mécanisme de dégradation HCI.

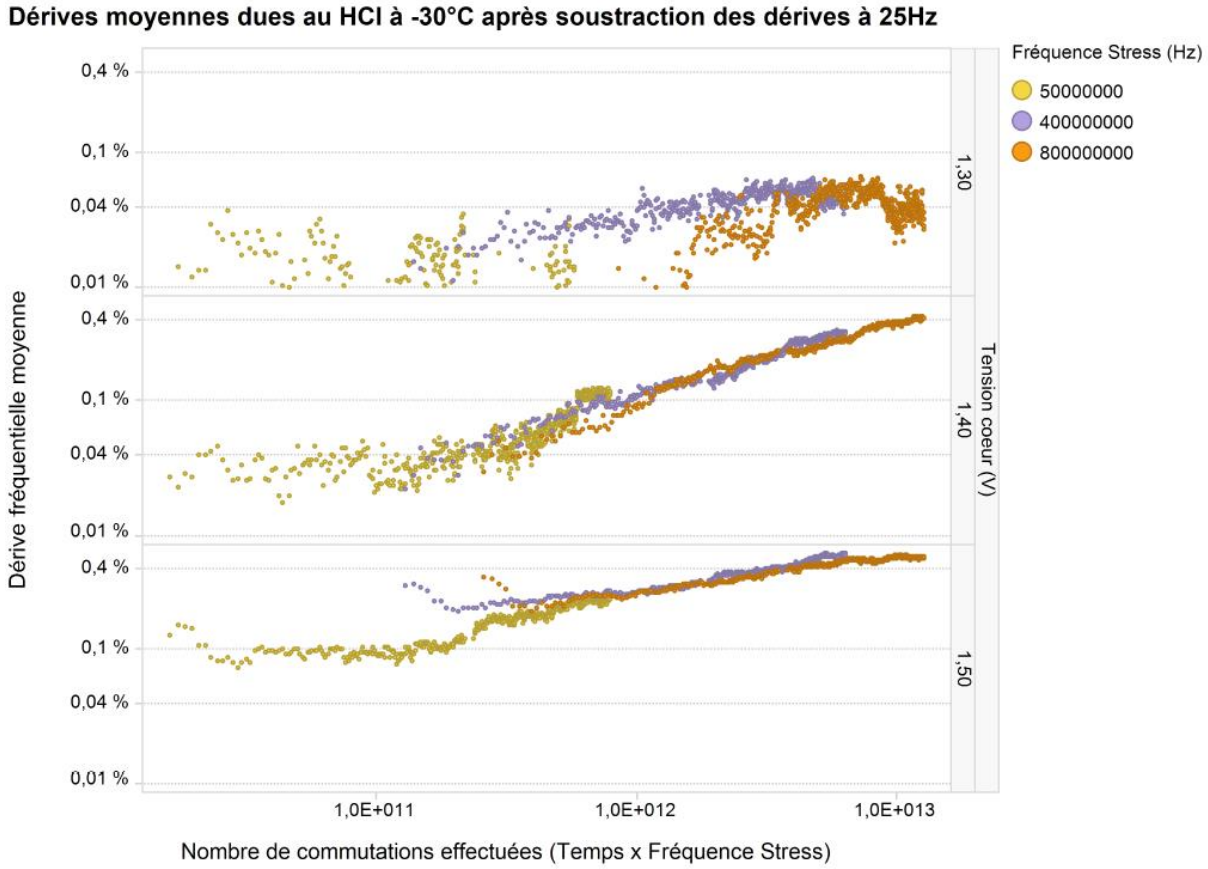


Figure 116 : Dérives fréquentielles moyennes dues au HCI des FPGA à -30°C en fonction du nombre de commutation, de la tension et de la fréquence de stress en échelle LOG-LOG

Les dégradations mesurées à basse température étaient déjà très faibles (au plus 1,5%). Après extraction, la pire dérive - dorénavant due principalement au HCI - ne dépasse pas 0,7%.

Cependant, en analysant nos résultats nous retrouvons tous les signes de la présence de ce mécanisme (lien avec le nombre de commutations, l'effet de la basse température et la pente des dérives plus élevée que pour le BTI). La précision des valeurs sont limitées par le bruit de mesure et par la méthode d'extraction où le BTI n'est pas parfaitement indépendant de la fréquence de stress. Néanmoins nous avons confirmé la présence de ce mécanisme dans le cadre de nos essais et nous l'avons mesuré.

3.4 Modélisation des dérives

Nous avons mis en évidence des deux mécanismes de dégradation – le BTI et le HCI –au moyen de nos plans de test. Les résultats ont été exploités en 3.3. De là, nous avons appliqué une méthode d'extraction de chacun des mécanismes.
Dans cette section nous allons proposer un modèle des dégradations pour chaque mécanisme ainsi qu'une analyse statistique permettant d'estimer le MTTF ou taux de défaillance utilisé dans le secteur industriel. Nous allons dans un premier temps analyser le BTI, puis dans un second temps le HCI. Pour conclure ce chapitre, nous allons finalement proposer une approche de la prise en compte des deux mécanismes combinés.

3.4.1 Analyse de la dérive du BTI

Dans l'analyse du BTI, deux aspects sont importants : l'évolution des dérives et leur dispersion statistique associée.

3.4.1.1 Modélisation de la variation de fréquence

Dans cette sous-section, deux cas vont être abordés : le stress par un signal DC1 ou DC0. La partie 3.3.2.1.3 a en effet montré que la cinétique des dégradations était différente pour ces deux conditions de stress. L'une est plutôt teintée de NBTI et l'autre de PBTI, cependant nous ne pouvons pas les distinguer avec certitude dans notre étude. Une piste a été explorée en 3.3.2.3.2 avançant que le PBTI pourrait être majoritaire sur les RO stressés en DC1 à haute tension.

Nous allons présenter pour ces deux cas une modélisation de la dégradation relative de la fréquence d'oscillation en fonction du temps, de la température et de la tension dans ce mémoire. Les paramètres de ces modèles seront estimés par la méthode des moindres carrés présentée en 1.1.5.1. Le modèle de dégradation sera toujours en loi puissance par rapport au temps avec un facteur d'accélération thermique en Arrhenius. Le facteur d'accélération électrique sera de deux formes : en exponentiel (équation 103) ou en puissance

(équation 104). Les deux aboutissent en effet à des résultats très proches sur la gamme de tension considérée.

$$vf = \mathrm{A}.\exp(\gamma.V)\,.\,exp\left(-\frac{E_a}{k_B.T}\right).t^n \quad (103)$$

$$vf = \mathrm{A}.\mathrm{V}^{\alpha}.\,exp\left(-\frac{E_a}{k_B.T}\right).t^n \quad (104)$$

Où :

- vf est la dérive fréquentielle relative du RO
- A est une constante dépendante de la technologie
- t est le temps en heures
- n est l'exposant de la loi puissance
- γ est le paramètre de la loi exponentielle en tension (en V^{-1})
- α est l'exposant de la loi puissance en tension
- E_a est l'énergie d'activation apparente en eV
- V est la tension en V
- T est la température absolue de jonction en Kelvin

Nous allons maintenant appliquer cette modélisation sur nos dégradations en stress DC1, puis sur celles en stress DC0.

3.4.1.1.1 Cas des RO stressés en DC1

Dans cette partie nous allons considérer les RO dans les conditions suivantes :

- Signal de stress DC1
- Température de jonction au cours du vieillissement supérieur à 0°C
- RO formé d'une chaîne de buffer

Nous considérons ainsi qu'ils vont essentiellement présenter des dérives de type BTI. Le signal de stress DC1 est de loin le pire cas mesuré au niveau des dérives à haute température.

Comme présenté dans les résultats en 3.3.2.1.1, une dégradation en loi puissance du temps semble tout à fait convenir ici. Cette loi est en accord avec ce type de dégradation. Le facteur d'accélération en température sera modélisé par la loi d'Arrhenius. Le facteur d'accélération en tension (noté AFV) est variable d'une technologie à l'autre. C'est pourquoi nous avons testé sur ce point une loi en puissance (voir équation 104) ainsi qu'une loi exponentielle (voir équation 103). Un facteur d'accélération (AF) en tension en loi puissance a obtenu une erreur au carré minimale de *6520,* soit une erreur quadratique moyenne par point de 0,369%. Une forme en exponentielle est descendue à *5765,* soit une erreur quadratique moyenne par point de 0.347%. Les deux formes donnent ici des résultats similaires (les autres paramètres estimés ne varient pas comme donné dans le Tableau 26). Notons que sur la gamme de

variations considérées (les tensions allant de 1 V à seulement 1,5 V) ces deux lois se comportent quasi identiquement. Même si la loi en exponentiel est légèrement meilleure ici, nous ne pouvons pas raisonnablement en choisir une. Dans ce sens, cette section va présenter les deux modèles. Les paramètres estimés pour ces modèles sont listés dans le Tableau 26.

Les valeurs estimées sont cohérentes avec la littérature pour la modélisation du BTI. Pour ce mécanisme, l'exposant de la loi puissance sur le temps est bien autour de 0,25. Les paramètres des facteurs d'accélération correspondent aux gammes de valeurs estimées par la littérature pour du PBTI (E_a de 0,11 à 0,15 eV et un exposant α de 5,5 à 9). Ce dernier point nous conforte dans la conclusion de la section 3.3.2.3.2 : il y a plus de NMOS impliqués dans le vieillissement que de PMOS, et donc qu'on observe une majorité de PBTI dans le cas DC1.

Facteur	Paramètre	Valeur estimée modèle AFV exponentiel	Valeur estimée modèle AFV puissance
Constante	$\boldsymbol{A}$	-0,06610 h^{-n}	-6,733 $V^{-\alpha}.h^{-n}$
Temps	$\boldsymbol{n}$	0,2645	0,2652
Température	$\boldsymbol{E_a}$	0,1604 eV	0,1621 eV
Tension	/	$\boldsymbol{\gamma}$ = 4,804 V^{-1}	$\boldsymbol{\alpha}$ = 6,474

Tableau 26 : Paramètres des modèles de dégradation BTI en stress DC1

La Figure 117 présente les données utilisées ainsi que les prévisions du modèle proposé. Elle présente le logarithme des valeurs absolues des dérives relatives de la fréquence d'oscillation en fonction du logarithme du temps. Chaque couleur correspond à une tension d'alimentation. Chaque sous-figure correspond à une gamme de température de jonction. Les points correspondent aux valeurs mesurées. Les lignes colorées sont les prévisions du modèle avec facteur d'accélération en tension en loi exponentielle. Les lignes noires sont les prévisions du modèle avec facteur d'accélération en tension en loi puissance. La sous-figure en bas à droite correspond à un agrandissement des points à 115°C et autour de 1,5V.

Pour une même couleur, nous pouvons distinguer différents groupes de mesure. Ces différences proviennent des légères variations de tension d'un DUT à l'autre dans une même gamme de tension. Ces différences ont été prises en compte dans l'estimation du modèle comme nous pouvons l'observer sur l'agrandissement des mesures à 115°C et 1,5V. Cependant, les prévisions (ligne droite) ont été faites pour la valeur moyenne constatée dans chaque gamme de tension et de température. Visuellement, les prévisions faites par les

modélisations sont bonnes au regard des valeurs mesurées. Ainsi la modélisation est correcte et pertinente sur la gamme considérée (de T_j = 55°C à 115°C).

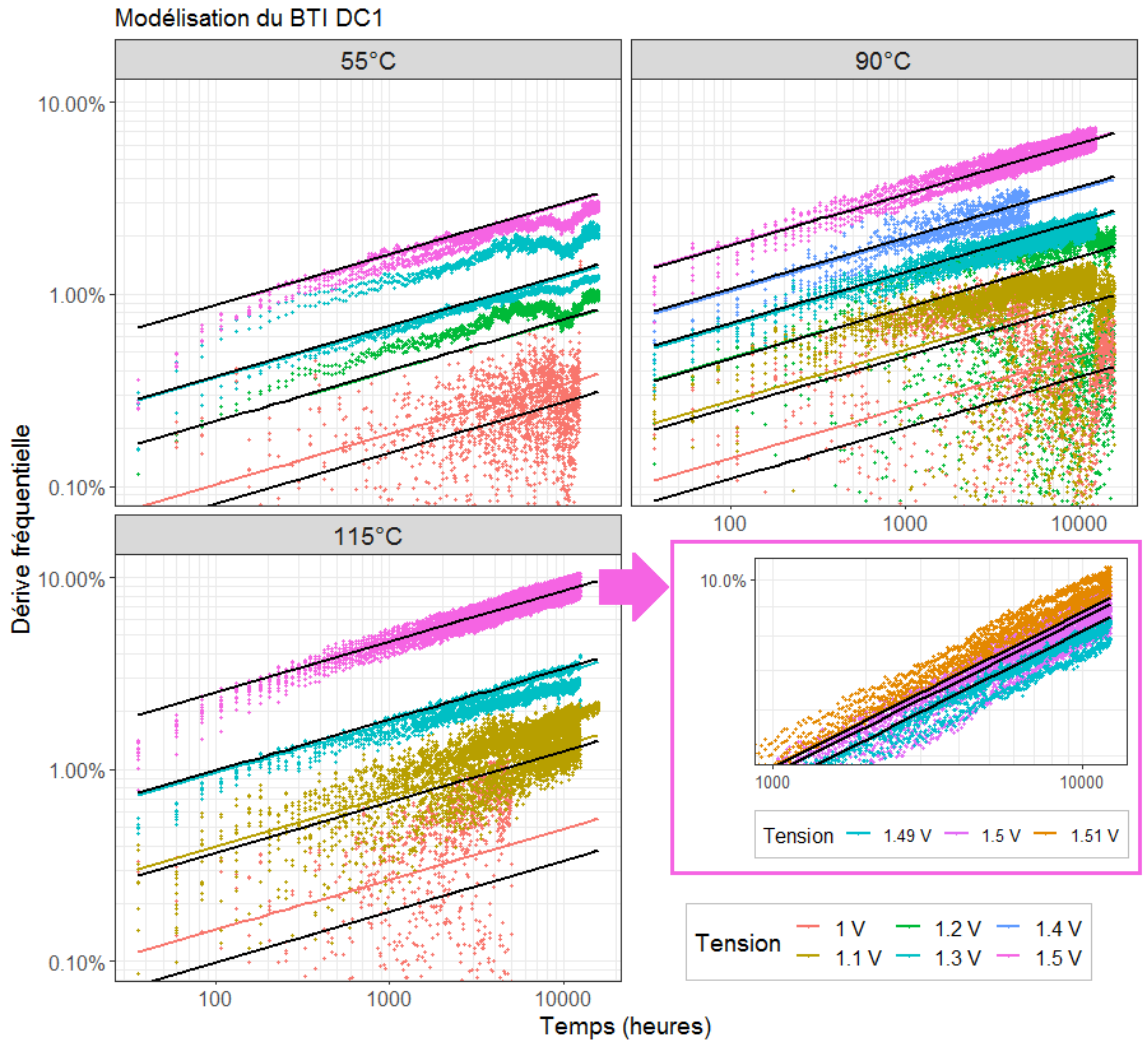


Figure 117 : Modélisation des dérives BTI en stress DC1, paramètres du modèle comparés avec les mesures

3.4.1.1.2 Cas des RO stressés en DC0

Dans cette seconde partie nous allons considérer les RO dans les conditions suivantes :

- Signal de stress DC0
- Température de jonction au cours du vieillissement supérieur à 0°C
- RO formé d'une chaîne de buffer

Nous considérons ainsi qu'ils vont essentiellement présenter des dérives de type BTI. Pour les basses tensions d'alimentation – même si les dérives sont faibles – le signal de stress DC0

génère plus de dégradation que le signal DC1. Ainsi il est également important d'étudier ce cas.

Le choix de la forme du modèle est analogue au cas DC1 présenté dans la partie précédente. Un facteur d'accélération (AF) en tension en loi exponentielle a obtenu une erreur au carré minimale de *5381*, soit une erreur quadratique moyenne par point de 0,332%. La forme en puissance est descendue à *5266*, soit une erreur quadratique moyenne par point de 0,329%. Les deux formes sont ici encore relativement proches. Comme pour la section précédente sur le stress DC1, nous allons présenter les deux modèles faute de pouvoir en discréditer un. Les paramètres estimés pour ces modèles sont listés dans le Tableau 27.
Les valeurs estimées sont - comme pour le cas DC1 - cohérentes avec la littérature pour la modélisation du BTI. Les paramètres des facteurs d'accélération correspondent ici plutôt aux gammes de valeurs estimées par la littérature pour du NBTI (E_a de 0,01 à 0,4 eV et l'exposant α de 3 à 4).

Facteur	Paramètre	Valeur estimée modèle AFV exponentiel	Valeur estimée modèle AFV puissance
Constante	**A**	-0,5915 h^{-n}	-2,005 $V^{-\alpha}.h^{-n}$
Temps	$\boldsymbol{n}$	0,2620	0,2609
Température	$\boldsymbol{E_a}$	0,08870 eV	0,08916 eV
Tension	/	$\boldsymbol{\gamma}$ = 1,231 V^{-1}	$\boldsymbol{\alpha}$ = 1,594

Tableau 27 : Paramètres des modèles de dégradation BTI en stress DC0

La Figure 118 illustre visuellement la pertinence des modèles (AFV en puissance en lignes noires et modèle AFV en exponentiel en lignes colorées) en confrontant les prévisions aux mesures. Les dérives mesurées sont ici beaucoup plus faibles que pour le cas d'un stress DC1, ainsi la plupart des mesures sont encore proches de la marge de bruit de mesure. Malgré ce bruit, nos modèles restent cohérents avec les mesures. La différence entre les deux facteurs d'accélération est perceptible à basse tension (là où l'on est encore dans le bruit de mesure), sinon les deux formes du modèle sont quasi confondues.

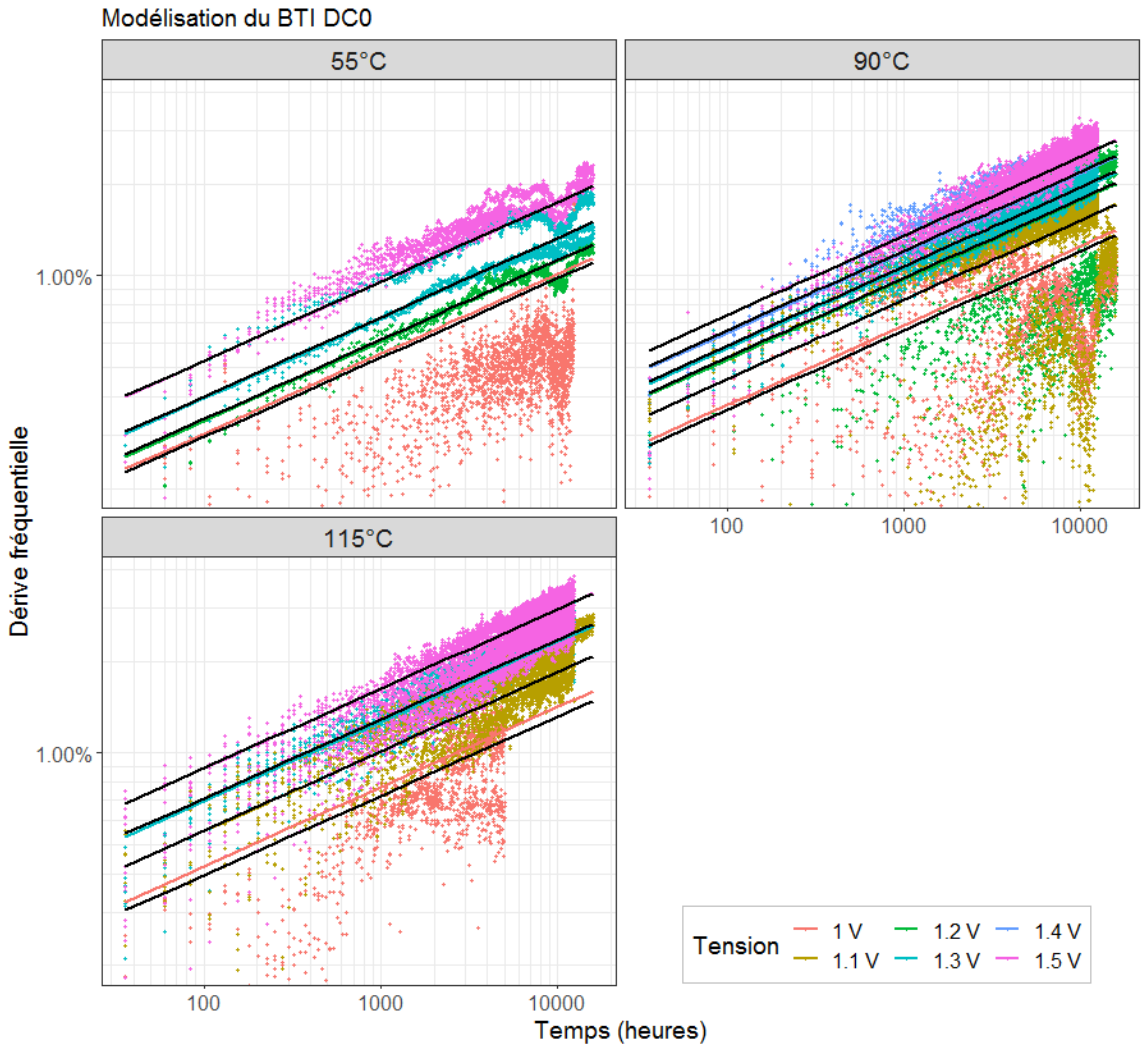


Figure 118 : Modélisation des dérives BTI en stress DC0, paramètres du modèle comparés avec les mesures

3.4.1.2 Analyse statistique des durées de vie

Maintenant que les dégradations ont été modélisées, il est important d'évaluer la dispersion statistique de ces résultats. Dans cette section, nous allons détailler les étapes qui mèneront à l'obtention d'un MTTF avec intervalle de confiance.

Pour ce faire nous procèderons comme suit :

1. Fixer un critère de défaillance,
2. Extrapoler les dérives jusqu'au critère de défaillance pour obtenir les durées de vie,
3. Choisir la distribution statistique la plus adaptée,
4. Estimer les paramètres de la distribution des durées de vie en fonction des mesures.

Cette méthodologie d'analyse des données et d'estimation du temps moyen de bon fonctionnement a été validée sur les données d'un projet antérieur. Il s'agissait d'un PEA destiné à évaluer la fiabilité d'un nœud technologique d'un fabricant spécifique. Cette étude a cumulé 2 ans de vieillissement continu de 76 ASIC (en technologie 65nm) à application cryptographique en conditions extrêmes (jusque V_{nom}+40% et 125°C pour la pire condition de stress). A partir de ces résultats nous avons réalisé une modélisation de sa fiabilité vis-à-vis du BTI. Cette étude a fait l'objet d'une publication [124] et d'une présentation en symposium scientifique.

3.4.1.2.1 Choix du critère de défaillance

Lors de la plupart des campagnes de vieillissement accéléré, la majorité des échantillons ne présente pas de panne franche. Cet aspect est problématique pour estimer correctement la fiabilité et la cinétique des défaillances. Les mécanismes recherchés ici génèrent de lentes dérives sur des paramètres précis (la tension de seuil des transistors, les temps de propagation, la fréquence de fonctionnement maximales du système, etc.). Ainsi, il est courant dans le domaine de prendre comme critère de défaillance une dérive paramétrique relative de 10%. Ce seuil correspond notamment à la marge de conception usuellement pris.

Le mécanisme de dégradation BTI induit une augmentation de la valeur absolue de la tension de seuil des transistors MOS. Ainsi, le temps de commutation augmente avec le vieillissement. Lorsque le temps de propagation du chemin critique (chemin logique ayant le temps de propagation le plus long, définissant par conséquent la fréquence maximale de fonctionnement de l'ensemble de l'architecture) dépasse la marge de conception pour une fréquence donnée, on peut considérer le système comme défaillant. Les concepteurs prennent habituellement une marge de 10% par rapport à la fréquence de fonctionnement du FPGA. C'est pourquoi on choisira ici comme critère de défaillance une dérive fréquentielle de 10%.

La file d'essai FPGA nous permet de suivre une dérive paramétrique : la fréquence des RO. Ce paramètre traduit directement une dérive du temps de propagation du signal au travers des différents étages de transistors. Considérons que chaque étage du RO (possédant un nombre forcement impair d'inverseurs pour pouvoir osciller) dérive de manière identique, et que les temps de propagation en montée et en descente sont aussi identiques. La dérive d'un étage traduit la dérive du RO de la manière suivante :

$$\upsilon f_{RO} = \frac{f_{RO} - f_{RO\,0}}{f_{RO\,0}} = \frac{\frac{1}{2.N.\tau} - \frac{1}{2.N.\tau_0}}{\frac{1}{2.N.\tau_0}} = \frac{\frac{1}{\tau} - \frac{1}{\tau_0}}{\frac{1}{\tau_0}} = \upsilon f_{stage} \qquad (105)$$

Où :

- vf_{RO} est la dérive fréquentielle relative d'un RO
- f_{RO} est la fréquence d'oscillation libre du RO
- $f_{RO\,0}$ est la fréquence d'oscillation libre du RO au début du vieillissement
- N est la longueur du RO
- τ est le temps de propagation d'un étage
- τ_0 est le temps de propagation d'un étage au début du vieillissement
- vf_{stage} est la dérive fréquentielle d'un étage du RO

Ainsi on peut étendre ce résultat : si la fréquence d'un RO de test dérive de 10%, alors la fréquence maximale de fonctionnement de n'importe quel chemin (dont le chemin critique) d'un système dérivera aussi de 10%. On peut donc choisir ici comme critère de défaillance une dérive fréquentielle d'un RO de 10%.

3.4.1.2.2 Extrapolation des dérives

Une fois le critère de défaillance choisi, il faut que nos dérives atteignent ce dit critère. Dans notre cas, seulement une condition d'essai (V_{nom}+50% et 115°C) atteint 10%. Cependant, nos mesures de dérive étant fines et le comportement des dérives monotone et régulier, on peut extrapoler les dérives en utilisant une modélisation. En accord avec la littérature et nos résultats, on utilise une modélisation en puissance temporelle comme définie dans les équations 103 et 104. Le modèle utilisé est de la forme :

$$vf = A_0 . t^n \tag{106}$$

Où A_0 est une constante.

Les paramètres de ce modèle ont été estimés pour chaque RO et le temps de défaillance TTF calculé de la sorte :

$$TTF_{RO} = \left(\frac{10\%}{A_{0_{RO}}}\right)^{1/n_{RO}} \tag{107}$$

Si ce temps est supérieur à 300 000 heures, la donnée restera censurée pour éviter d'avoir des temps de défaillance totalement incohérents avec quelqu'application que ce soit. Cette censure concerne surtout les pièces avec les conditions de vieillissement aggravées les plus faibles. Lorsque la modélisation est mauvaise (c.-à-d. ayant un R^2 inférieur à 0,95) on censure tout simplement la donnée au temps de test réel. De plus, l'instant de défaillance extrapolé est calculé sous forme d'intervalle de confiance.

3.4.1.2.3 Modélisation statistique des durées de vie

Afin d'obtenir la fiabilité du composant vis à vis du mécanisme de dégradation BTI, on a comme à la section 3.3.2 considéré uniquement les FPGA ayant une température de vieillissement supérieure à 0°C et un rapport cyclique élevé. On utilise uniquement les cofacteurs suivants : température et tension. Les autres cofacteurs seront choisis afin d'être dans la démarche « pire cas », c.-à-d. ici un signal de stress DC1. Nous avons également porté cette étude uniquement sur les chaînes de RO contenant une majorité d'étages de buffers. Ce choix provient du fait qu'une chaîne d'inverseur pure sera moins sensible au rapport cyclique de stress qu'une chaîne de buffer (où tous les étages sont dans le même état).

La population considérée est composée de 432 échantillons (RO) répartis sur 39 FPGA. Les conditions de stress totalement censurées concernent uniquement les conditions les moins agressives. Le taux de censure total est de 54%, ce qui est plus que suffisant pour mener une étude pertinente. De plus, les défaillances sont bien réparties sur plusieurs conditions. Ceci permet d'obtenir de manière précise les paramètres des facteurs d'accélération. Le détail de la population est donné dans le Tableau 28.

Le choix du meilleur modèle se fera parmi quatre possibilités. La distribution statistique sera soit de Weibull, soit Log-normale. Dans le monde du silicium, ce sont les distributions les plus classiques et les plus pertinentes. Le facteur d'accélération thermique suivra la loi d'Arrhenius. Le facteur d'accélération électrique sera soit en exponentiel, soit en puissance. L'estimation des paramètres du modèle se fera par la méthode du maximum de vraisemblance présenté en 1.1.5.2.

Le choix de la forme du facteur d'accélération en tension est ici évident. Pour chaque distribution, la forme exponentielle donne de bien meilleurs résultats de vraisemblance. De plus, on est en accord avec la forme de modélisation de dérive choisie précédemment. Le choix de la distribution est plus ambigu. Les scores de vraisemblances des deux distributions sont très proches.

Température de jonction	Tension	Nombre d'échantillons	Nombre de défaillances observées	Taux de censure
55°C	**1 V**	16	0	100%
55°C	**1,2 V**	8	0	100%
55°C	**1,3 V**	16	7	56%
55°C	**1,5 V**	24	1	96%
90°C	**1 V**	24	0	100%
90°C	**1,1 V**	40	0	100%
90°C	**1,2 V**	16	0	100%
90°C	**1,3 V**	32	23	28%
90°C	**1,4 V**	24	7	71%
90°C	**1,5 V**	32	32	0%
115°C	**1 V**	24	0	100%
115°C	**1,1 V**	48	14	71%
115°C	**1,3 V**	32	29	9%
115°C	**1,5 V**	64	64	0%
120°C	**1,1 V**	16	8	50%
120°C	**1,2 V**	16	12	25%
	TOTAL	**432**	**197**	**54%**

Tableau 28 : Répartition de la population des RO en BTI

La Figure 119 présente les fonctions de défaillance empiriques (via estimateur de Kaplan-Meier décrit en 1.1.2) dans une échelle de Weibull en (a) et dans une échelle log-normale en (b). Chaque couleur correspond à une condition de test (couple {Température ; Tension}). Visuellement, l'alignement des points - sur un graphique ayant des échelles adaptées - est bon dans les deux cas. Le paramètre de forme (pente des courbes) est constant au travers des conditions de vieillissement. Cette homogénéité des pentes est signe de la présence d'un seul mécanisme apparent de dégradation.

Nous avons choisi dans ce mémoire une distribution log-normale pour deux raisons :

- l'alignement visuel est légèrement meilleur,
- la distribution log-normale est plus pessimiste, ce qui nous place dans un pire cas.

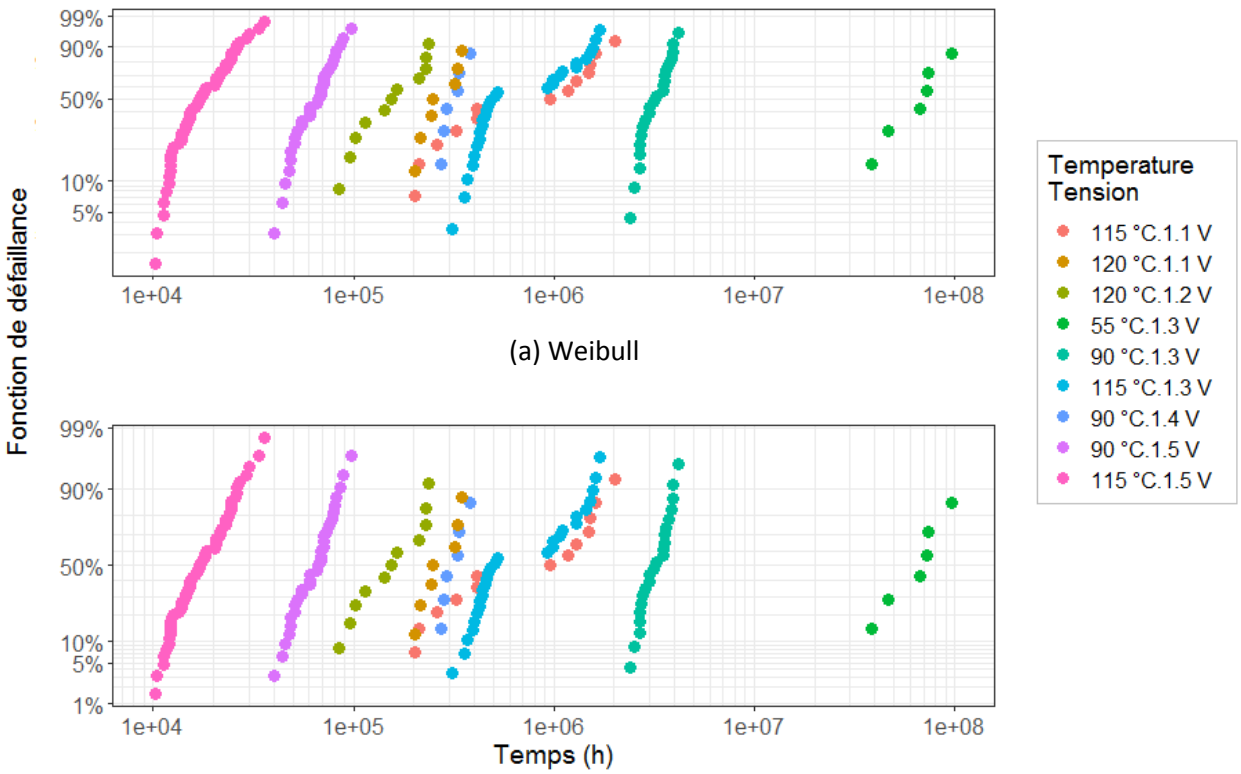


Figure 119 : Fonction de défaillance empirique des FPGA en BTI en échelle Weibit et log-normal

La forme du modèle log-normale choisi est la suivante :

$$F(t,T,V) = \Phi\left(\frac{\ln(t) - \mu_0 + \gamma_{STAT}.V - \frac{Ea_{STAT}}{k_B.T}}{\sigma}\right) \quad (108)$$

Où :

- Φ est la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$
- μ_0 est le pré-facteur de l'espérance du logarithme de t
- σ est l'écart type du logarithme de t
- γ_{STAT} est le paramètre du facteur d'accélération en tension en exponentiel

Les valeurs de ce modèle estimées par MLE pour nos résultats sont dans le Tableau 29. Pour information, l'estimation d'une distribution de Weibull donnait un paramètre de forme β entre 1,44 et 1,77. Cette valeur supérieure à 1 atteste d'un mécanisme de vieillissement.

Paramètre	Inf_5%	Standard	Sup_95%
μ_0	1,505	3,449	5,393
σ	0,7329	0,8021	0,8713
$E_{a_{STAT}}$	0,7552 eV	0,8126 eV	0,8701 eV
γ_{STAT}	11,17 V^{-1}	11,88 V^{-1}	12,59 V^{-1}

Tableau 29 : Estimation par MLE des paramètres du modèle FPGA BTI (avec intervalles de confiance à 90%)

Notons que les valeurs des paramètres des facteurs d'accélération sont ici différentes de celles estimées dans la section 3.4.1.1. Cette différence vient du fait de considérer dans un cas des dérives, et dans l'autre cas des instants de défaillance. Nous allons démontrer cette relation.

Notons vf_{lim} le critère de défaillance retenu. A cette valeur est associé un temps t_{lim} au bout duquel ce critère sera atteint. En reprenant l'équation 103 on peut exprimer ce temps limite. On note ici avec l'indice « drift » les paramètres issus des modèles de dérive.

$$t_{lim} = \left(vf_{lim} . \frac{1}{A_0} . exp\left(-\gamma_{drift} . V\right) . exp\left(\frac{E_{a_{drift}}}{k_b . T}\right)\right)^{\frac{1}{n}} \tag{109}$$

On regroupe les termes par facteur d'accélération.

$$t_{lim} = exp\left(-\frac{\gamma_{drift}}{n} . V\right) . exp\left(\frac{E_{a_{drift}}}{n . k_b . T}\right) . \left(\frac{vf_{lim}}{A_0}\right)^{\frac{1}{n}} \tag{110}$$

Puis par identification on peut établir les relations suivantes.

$$\begin{cases} E_{a_{stat}} = \frac{E_{a_{drift}}}{n} \\ \gamma_{stat} = \frac{\gamma_{drift}}{n} \end{cases} \tag{111}$$

En prenant les valeurs de la section 3.4.1.1.1 on obtient $E_{a_{stat}} = 0{,}6\ eV$ et $\gamma_{stat} = 18\ V^{-1}$. L'ordre de grandeur est cohérent entre ces nombres et ceux du Tableau 29. La différence provient des approches qui sont différentes : dans un cas nous modélisons des dérives en tenant compte de l'ensemble de la pente des dégradations, dans l'autre cas nous travaillons avec uniquement des instants de défaillances dispersés et partiellement censurés à droite.

La Figure 120 représente la fonction de répartition empirique de nos observations où toutes les conditions de stress ont été ramenées à des conditions opérationnelles (T_j=100°C et V_{nom}+5%) en utilisant les facteurs d'accélération du modèle choisi. On observe un bon alignement des points suivant les prévisions (avec enveloppe de confiance 90%) de notre modèle. Ceci nous permet de renforcer notre confiance dans les paramètres des facteurs d'accélération estimée.

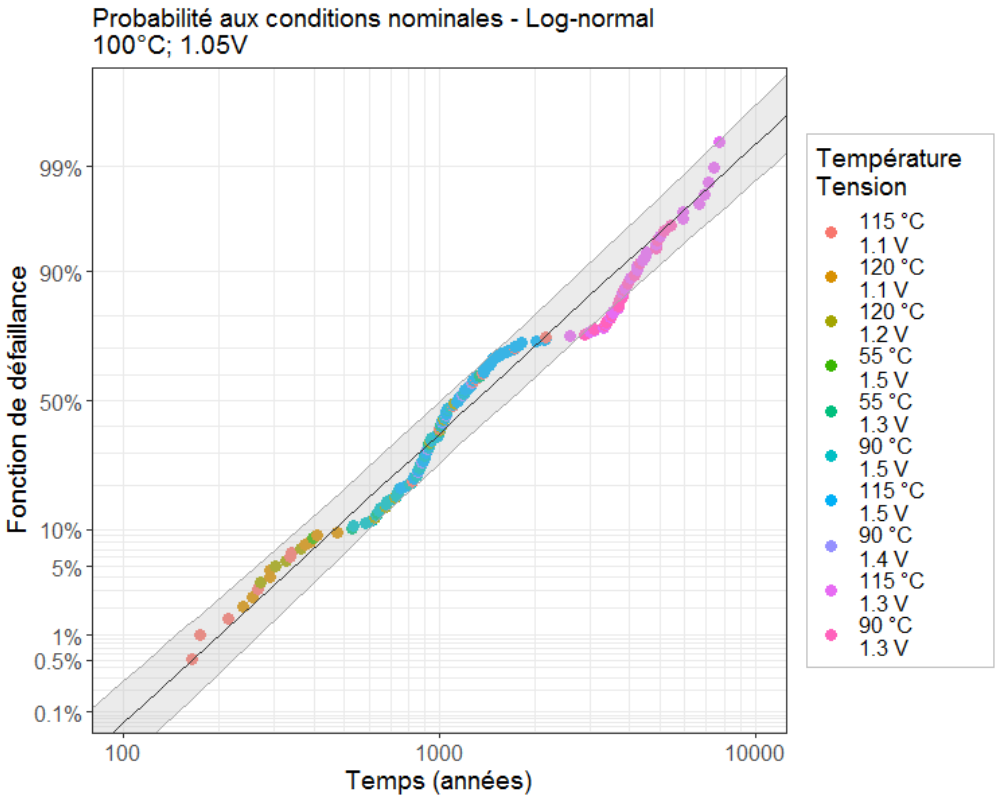


Figure 120 : Probabilité de défaillance des observations rapportées aux conditions 100°C et V_{nom}+5%, FPGA BTI

3.4.1.2.4 Application du modèle choisi

Maintenant que le modèle est défini, on peut estimer des grandeurs plus explicites utilisées dans le milieu industriel comme le MTTF (équation 112) ou le temps médian (équation 113).

$$MTTF(T,V) = e^{\mu_0 - \gamma.V + \frac{E_a}{k_B.T} + \frac{\sigma^2}{2}} \quad (112)$$

$$t_{50\%}(T,V) = e^{\mu_0 - \gamma.V + \frac{E_a}{k_B.T}} \quad (113)$$

Les composants considérés ayant été approvisionnés en gamme industrielle, leurs durées de vie projetées aux conditions maximales usuelles sont dans le Tableau 30. Ces valeurs sont au-delà du besoin opérationnel d'une application classique.

	Inf_5%	**Standard**	**Sup_95%**
MTTF (années)	1233	1586	2010
$t_{50\%}$ (années)	1007	1298	1646
$t_{10\%}$ (années)	355	464	593
$t_{1\%}$ (années)	147	201	267
$t_{0,1\%}$ (années)	76	109	148

Tableau 30 : Durées de vie projetées à 100°C et V_{nom}+5% (max gamme industrielle), FPGA BTI

Concernant le taux de défaillance, on peut le calculer en connaissant la loi du modèle ainsi que ses paramètres. Pour notre modèle, le taux de défaillance λ est de la forme :

$$\lambda(t) = \frac{\frac{1}{t.\sigma.\sqrt{2\pi}}.e^{-\frac{1}{2}.\left(\frac{ln(t)-\mu_0+\gamma.V-\frac{E_a}{k_B.T}}{\sigma}\right)^2}}{1-\Phi\left(\frac{ln(t)-\mu_0+\gamma.V-\frac{E_a}{k_B.T}}{\sigma}\right)} \quad (114)$$

Le taux de défaillance est tracé Figure 121 pour différentes températures de jonction. La tension d'alimentation est V_{nom}+5% et le critère de défaillance est 10%. L'enveloppe correspond à l'intervalle de confiance des valeurs estimées.

Ces intervalles de confiance sont obtenus par la méthode de Monte-Carlo. Dans la mesure où l'expression du taux de défaillance est une combinaison non linéaire de variables aléatoires - supposées normales -, le calcul analytique est laborieux. On génère alors 1000 tirages de chacun des paramètres pour calculer 1000 tirages de λ. On trie ces valeurs par ordre croissant. Puis pour approximer la borne inférieure et la borne supérieure de l'intervalle de confiance à 90%, on prend respectivement la 50^{ième} et la 950^{ième} valeur. Dans la pratique on regardera principalement la borne supérieure de l'intervalle de confiance.
Une erreur commune peut être de prendre les valeurs extrêmes des paramètres à 90% de confiance pour en déduire la borne supérieures à 90% de confiance du taux de défaillance. Cette pratique est erronée et mène à une surestimation excessive de cette borne.

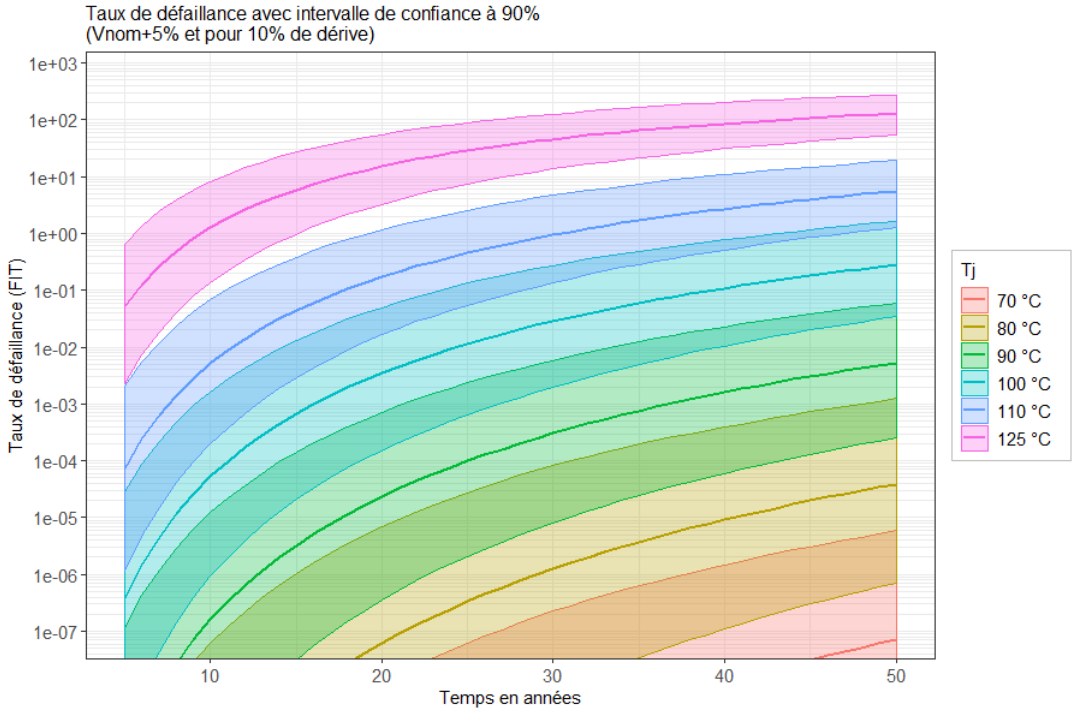


Figure 121 : Taux de défaillance à V_{nom}+5% et pour 10% de dérive, FPGA BTI

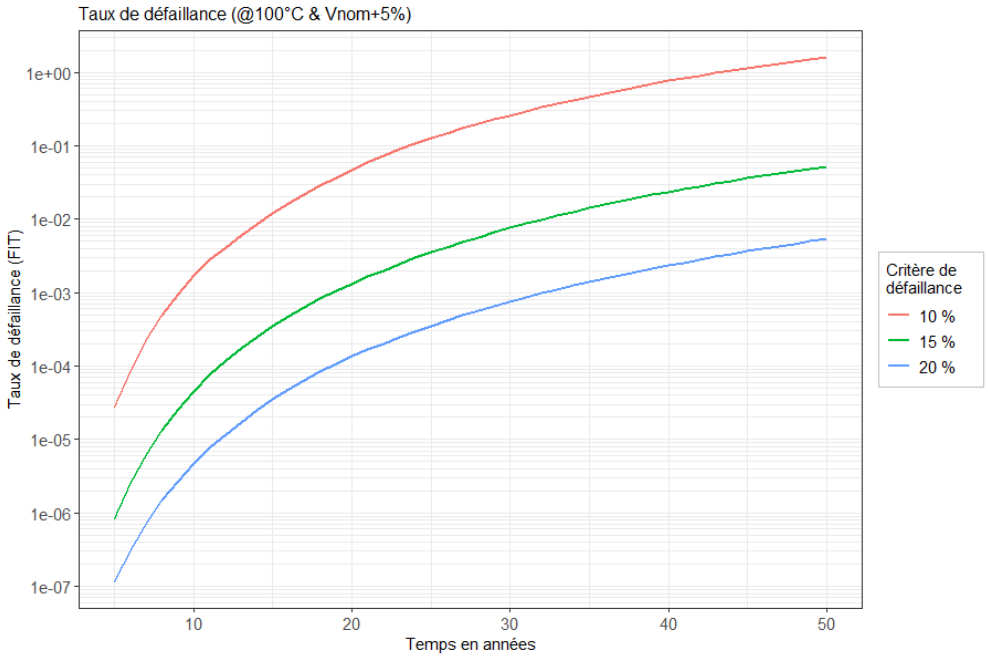


Figure 122 : Borne supérieure de l'estimation du taux de défaillance pour trois critères de défaillance à V_{nom}+5% et 100°C, FPGA BTI

Pour conclure sur ce sujet, il est important de noter que le choix du critère de défaillance est très important. Dans cette étude nous avons choisi 10% de dérive mais dans la réalité les

marges de fonctionnement sont bien plus importantes (à la marge de 10% prise par le concepteur s'ajoute la marge prise par le fabricant pour homogénéiser les performances garanties sur un même wafer). La Figure 122 présente ce taux de défaillance estimé pour plusieurs critères de défaillance de 10% à 20%. Un choix de critère pertinent en accord avec les besoins du circuit numérique est essentiel. La variation de ce paramètre peut amener des sous-estimations, ou des surestimations, importantes de la fiabilité du composant en conditions opérationnelles.

3.4.2 Analyse de la dérive du HCI

Les dérives dues aux porteurs chauds sont quant à elles beaucoup plus faibles que celles observées pour le BTI. Ce résultat a été détaillé en 3.3.3. Nous entreprenons une modélisation du phénomène dans cette partie.

3.4.2.1 Analyse du taux de dégradation

Dans cette partie nous allons considérer les RO dans les conditions suivantes :

- Signal de stress AC
- Température de jonction au cours du vieillissement inférieure à 60°C
- Mesure après extraction du HCI décrite en 3.3.3

Nous considérons ainsi qu'ils vont essentiellement présenter des dérives de type HCI. Les dérives ont été considérées en fonction du nombre de commutation, et non au cours du temps. La forme du modèle la plus adaptée dans notre cas comprend un facteur d'accélération électrique en suivant une loi de Takeda.
La modélisation des dégradations HCI d'un RO est de la forme :

$$vf = \mathrm{A}.\exp\left(-\frac{\alpha}{V}\right).exp\left(-\frac{E_a}{k_B.T}\right).(f.t)^n \qquad (115)$$

Où :

- α est le paramètre de la loi de Takeda en tension
- f est la fréquence de stress du RO en Hertz

Les paramètres estimés pour ce modèle sont listés dans le Tableau 31. Les valeurs estimées sont cohérentes avec la littérature pour la modélisation du HCI. L'exposant *n* est autour de 0,5. On retrouve ici une énergie d'activation très faible. Cette valeur d'E_a est négative dans ce cas, mais le résultat principal est que ce mécanisme est très peu sensible à la température. Il y a cependant un effet seuil de la température dans la mesure où l'on n'observe ce phénomène qu'à basse température.

Facteur	Paramètre	Valeur estimée
Constante	**A**	**$-2{,}7\times10^{-4}$**
Nombre de commutation	n	**0,46**
Température	E_a	**-0,019 eV**
Tension	α	**16 V**

Tableau 31 : Paramètres du modèle de dégradation HCI

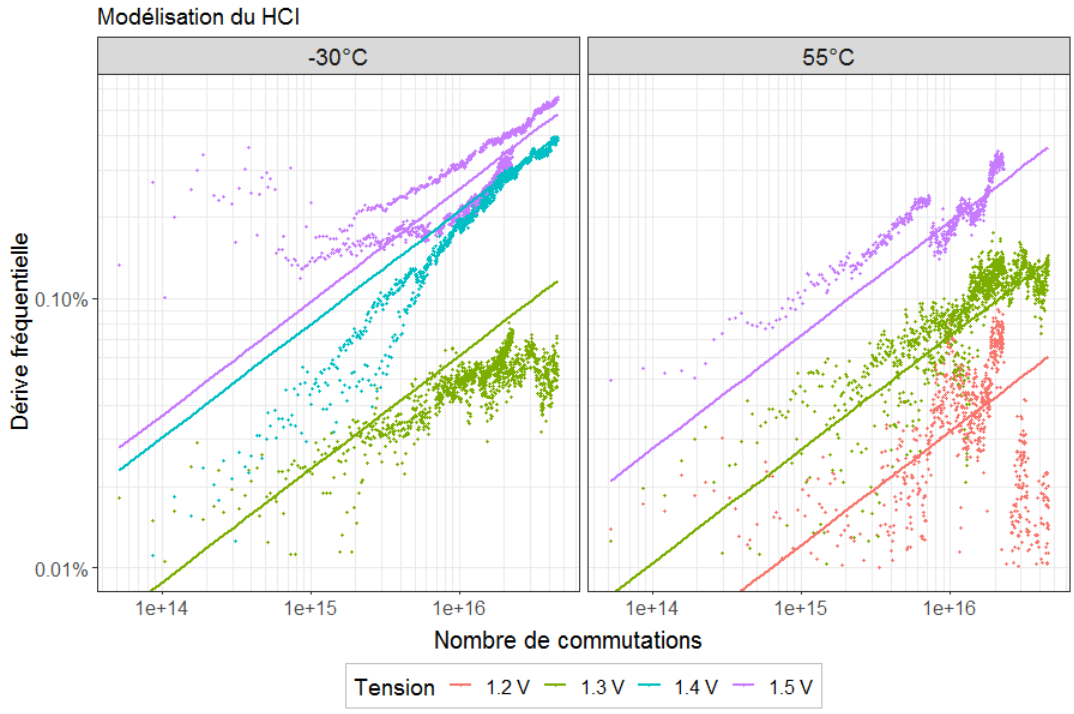


Figure 123 : Modélisation des dérives HCI

La Figure 123 illustre la modélisation en confrontant les prévisions (les droites de couleurs) aux mesures moyennes (les points). Les mesures sont très faibles (< 0,7%), et par conséquent très bruitées. La modélisation effectuée nous donne plus la tendance des dérives qu'un modèle utilisable tel quel.

3.4.2.2 Analyse statistique des durées de vie

Une analyse statistique a été menée sur le HCI. Cette section va décrire les résultats obtenus, ainsi que les limites de la méthode employée.

3.4.2.2.1 Extrapolation des durées de vie

Le choix du critère de défaillance est ici plus délicat. Les dégradations ne dépassent pas les 0,7% de dérive au bout de plus de 16 000 heures de vieillissement. Une extrapolation à 10% sur la même base que la section sur le BTI nous donnerait des instants de défaillance beaucoup trop lointains pour une vie réelle de composant électronique, ainsi qu'une précision incertaine. Afin d'évaluer l'influence de ce choix sur la pertinence du modèle statistique trouvé, plusieurs critères ont été évalués.

La même extrapolation que celle décrite dans le chapitre 3.4.1.2.2 sur les dérives HCI donne des instants de défaillances beaucoup trop dispersés. Pour un critère de défaillance de 10%, le paramètre σ estimé d'une distribution log-normale est autour de 2,9 (ou un paramètre de forme β en Weibull de 0,4).

Cette grande dispersion à deux causes :

- Les dérives sont extrapolées jusqu'à 10% sachant que les dérives maximales réellement observées ne dépassent pas les 0,7%
- Les dérives observées sont encore très proches du bruit de mesure. Ainsi, le bruit de mesure est par conséquent amplifié.

Une petite variation de la pente (exposant « *n* » de la loi en puissance des dérives) induit une grande variation de l'instant de défaillance extrapolé. De plus, là où une loi puissance fait l'unanimité pour le BTI, cette loi est moins valable pour le HCI. C'est pourquoi il a été mis en œuvre une pente commune par gamme de tension afin de limiter la dispersion par condition d'essai. C'est-à-dire que la pente « n » lors de l'extrapolation est commune à tous les RO proches soit de 1,2V, de 1,3V, de 1,4V ou soit de 1,5V. Ce paramètre « n » varie alors entre 0,3 et 0,6 suivant la condition en tension. En revanche, le pré-facteur est bien indépendant pour tous les RO. En comparaison, pour le BTI dans la section 3.4.1.2 nous avons extrapolé chaque RO indépendamment en utilisant une nouvelle pente à chaque fois. La nouvelle valeur du paramètre σ est maintenant 1,53 (ou un β en Weibull de 0,68). Bien que toujours très dispersé, cette amélioration est notable.

3.4.2.2.2 Tentative de modélisation des durée de vie

La population considérée est composée de 297 échantillons (oscillateurs en anneau) répartis sur 7 FPGA (voir Tableau 32). Le nombre d'échantillon non censuré est de 59 RO. Le taux de censure est donc de 80%, ce qui est élevé.

Température	Tension	Nombre d'échantillons	Nombre de défaillances observées	Taux de censure
-30	**1,1**	44	1	98%
-30	**1,3**	51	0	100%
-30	**1,4**	33	22	33%
-30	**1,5**	32	10	69%
55	**1,2**	23	0	100%
55	**1,3**	46	5	89%
55	**1,5**	68	21	69%
	TOTAL	297	59	**80%**

Tableau 32 : Répartition de la population des RO, FPGA HCI

Au regard de la vraisemblance, la distribution log-normale est nettement meilleure par rapport à celle de Weibull dans le cas de la modélisation de la fiabilité du HCI. On choisira donc cette dernière. Le choix de la forme du facteur d'accélération en tension est légèrement meilleur avec une forme en Takeda cette fois ci. La forme du modèle est donnée équation 116. Les valeurs estimées de ce modèle avec nos résultats sont dans le Tableau 33.

$$F(t.f,T,V) = \Phi\left(\frac{\ln(t.f) - \mu_0 - \frac{\alpha}{V} - \frac{E_a}{k_B.T}}{\sigma}\right) \quad (116)$$

Paramètre	Standard
μ_0	6,4
σ	1,5
E_a	-0,076 eV
α	19 V

Tableau 33 : Estimation par MLE des paramètres du modèle FPGA HCI pour un critère de défaillance de 10%

La Figure 124 est la fonction de répartition empirique de nos observations où toutes les conditions de stress ont été ramenées à des conditions opérationnelles (-40°C, V_{nom}+5% et

800MHz) en utilisant les facteurs d'accélération du modèle choisi. On visualise bien sur cette figure que l'incertitude – représentée par la surface grise – est très grande. La distribution log-normale est cependant adaptée pour ce mécanisme comme en atteste le relativement bon alignement des points.

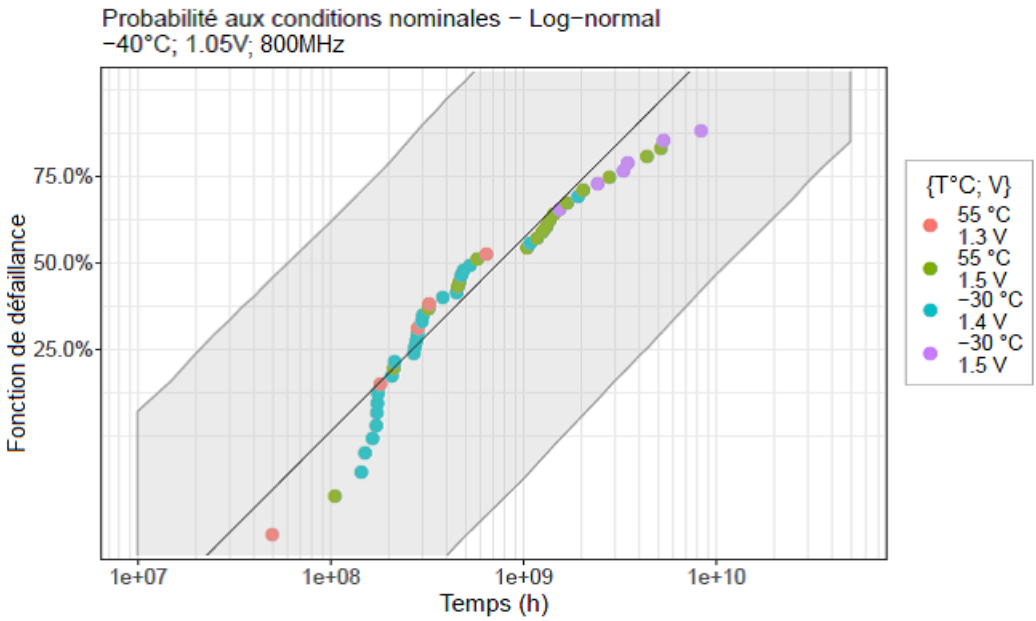


Figure 124 : Fonction de répartition des observations rapportées aux conditions -40°C ; V_{nom}+5% et 800 MHz, FPGA HCI

Comme cité plus haut, plusieurs critères de défaillances ont été considérés. Comme les dégradations observées sont très faibles, le critère le plus bas est de 1,5% de dérive, et le plus haut de 10%. Le Tableau 34 liste ces résultats.

Critère de défaillance	σ
1,5%	0,91
3%	1,11
5%	1,28
8%	1,44
10%	1,52

Tableau 34 : Influence du critère de défaillance sur la dispersion extraite des instants de défaillance, FPGA HCI

Une augmentation du critère de défaillance induit une augmentation directe du paramètre de forme σ. Ce paramètre traduit la dispersion des instants de défaillance extrapolés en fonction

des dérives mesurées. Cette dépendance de σ au critère montre que l'extrapolation amplifie le bruit de mesure ainsi que le bruit rajouté lors de la soustraction des dégradations du BTI. Elle est plus une conséquence du bruit de mesure qu'une propriété intrinsèque de la dispersion des composants. Cette variation est problématique et ne nous permet pas d'estimer avec précision la dispersion statistique des instants de défaillance due au mécanisme de dégradation HCI.

Les dégradations étant très faibles même après 16 000 heures dans les pires conditions de notre plan de test, nous allons tout de même nous assurer que le taux de défaillance induit par le HCI dans les pires conditions opérationnelles reste également faible. Le Tableau 35 récapitule les grandeurs de durée de vie intéressantes dans les pires conditions opérationnelles du HCI (T_j=-40°C, V_{nom}+5% et 800 MHz) pour un critère de dérive de 10%.

Critère de défaillance	10%	5%	3%
$\mathbf{t_{0,1\%}}$	758 ans	399 ans	241 ans
$\mathbf{t_{50\%}}$	84 935 ans	20 884 ans	7 449 ans
MTTF	124 425 ans	28 764 ans	9 833 ans
$\boldsymbol{\lambda(t_{max\ FIT})}$	1,216 FIT	/	/
$\mathbf{t_{max\ FIT}}$	10 784 ans	/	/
$\boldsymbol{\lambda(50\ ans)}$	1,83E-03 FIT	/	/

Tableau 35 : Taux de défaillance et durée de vie jusqu'à 10% de dérive à Tj=-40°C, Vnom+5% et 800 MHz, FPGA HCI

Les durées de vie moyennes estimées (même le $t_{0.1\%}$) sont bien au-delà des besoins. Le taux de défaillance dans les pires conditions est inférieur à 0,002 FIT pendant 50 ans. Le taux de défaillance maximal est atteint au bout de plus de 10 000 ans avec une valeur de 1,2 FIT.

3.4.3 Réalisation d'un modèle statistique couplé HCI+BTI

Dans une utilisation opérationnelle du composant, les deux mécanismes de dégradation étudiés dans cette thèse vont intervenir de manière concourante. Un certain nombre d'autres mécanismes (telle l'électromigration des pistes ou le TDDB) qui ont été abordés dans l'état de l'art peuvent également entrer en action pour d'autres technologies.

Cette section du mémoire va mettre en évidence les erreurs de la méthode usuellement mise en pratique dans l'estimation de la fiabilité globale, puis va décrire une autre approche statistiquement plus exacte.

3.4.3.1 Somme des taux de défaillance

Dans de nombreuses publications, on peut trouver la méthodologie suivante :

1. Réalisation de stress accélérés
2. Détermination d'un MTTF pour mécanisme
3. Calcul du taux de défaillance en 1/MTTF
4. Somme des taux de défaillance comme illustré ci-dessous

$$\lambda_{HCI+BTI} = \lambda_{HCI} + \lambda_{BTI} \tag{117}$$

Cette méthode simple à mettre en œuvre est - dans le cas des mécanismes d'usures BTI et HCI - erronée d'un point de vue probabiliste. Le taux de défaillance d'un mécanisme de vieillissement n'est en aucun cas 1/MTTF. Cette formule n'est valable que pour une loi exponentielle de mécanismes catalectiques indépendants. De plus le HCI et le BTI dégradent un même paramètre (la fréquence de fonctionnement maximale du système), leur taux de défaillance ne sont donc pas indépendants. Ainsi si chaque λ est estimé pour un critère de dérive de 10%, la somme des deux λ n'a plus aucun lien (ni même de sens) avec cette limite de 10%.

3.4.3.2 Détermination statistique du TTF avec HCI + BTI

Cette seconde approche du calcul de durée de vie pour la combinaison des deux mécanismes est différente. Il s'agit ici de calculer le taux de défaillance pour 10% de dégradation en sommant les dégradations au cours du temps des deux mécanismes. Cela nécessite donc une modélisation fine des dégradations pour chacun.

$$\upsilon f_{HCI+BTI}(t) = \upsilon f_{HCI}(t) + \upsilon f_{BTI}(t) \tag{118}$$

Ainsi, pour un critère de défaillance de 10%, il faut résoudre l'équation suivante :

$$\upsilon f_{HCI+BTI}(\mathrm{TTF}) = 10\% \tag{119}$$

L'algorithme de cette seconde méthode est :

1. Réalisation de stress accélérés.
2. Modélisation BTI $\upsilon f_{BTI}(t)$ et modélisation HCI $\upsilon f_{HCI}(t)$.
3. Détermination numérique du TTF par résolution de l'équation précédente.
4. Monté Carlo sur 1000 tirages de TTF. On prend en compte la moyenne et la variance de l'approximation normale de chaque paramètre des modèles.
5. Détermination de la distribution globale issue des tirages pour une condition donnée. On peut faire un test de Shapiro sur le logarithme des TTF pour valider une distribution log-normale par exemple.
6. Calcul du taux de défaillance en fonction de la distribution globale.

Si le but est d'estimer une borne inférieure avec 5% de risque du MTTF, on peut trier par ordre croissant les tirages et prendre la 50^{ième} valeur. C'est ce qui est réalisé dans la Figure 125. Cette méthode est plus exacte que la précédente, cependant elle est difficile à mettre en œuvre.

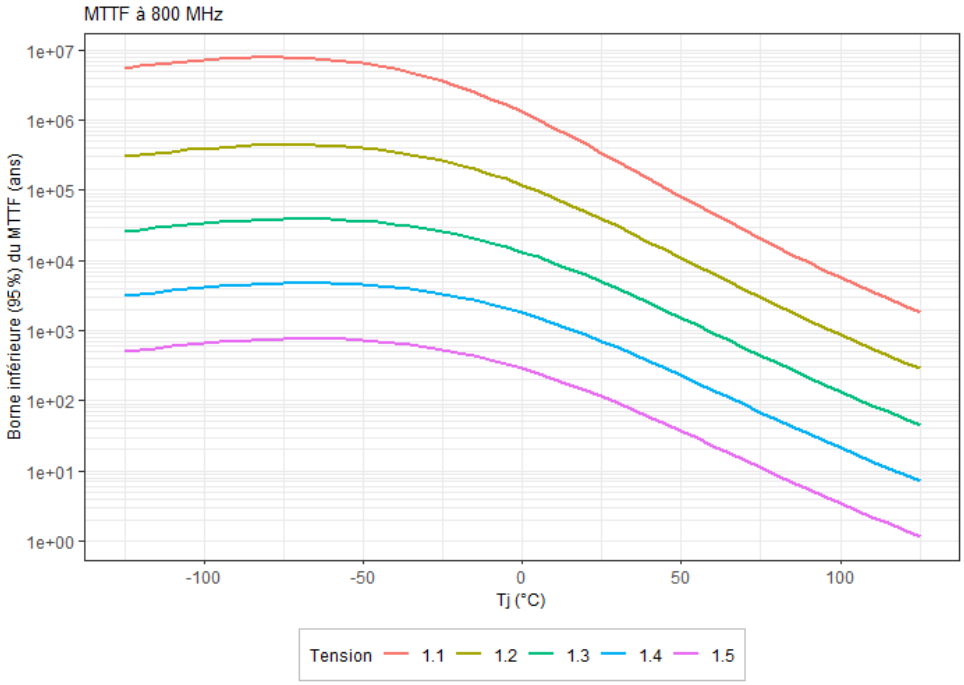


Figure 125 : Borne inférieure avec 95% de confiance du MTTF en utilisant le somme des deux modèles de dérive

3.5 Conclusion sur la fiabilité des FPGA

Le but de ce chapitre a bien été atteint. Nous avons évalué la fiabilité d'une technologie DSM sur FPGA. Les bancs de test mis en place ont permis de générer plusieurs conditions de stress à la fois thermique, électrique et fréquentiel. Une attention toute particulière a été apporté à l'asservissement en température et en tension afin d'avoir un minimum de bruit de mesure. La structure de test implémentée à base d'oscillateurs en anneau a facilité la mise en évidence ainsi que la mesure des dégradations des principaux mécanismes attendus. Nous n'avons pas observé d'électromigration ni de TDDB (claquage d'oxyde). En effet, nous n'avons mesuré aucune détérioration soudaine et brutale des performances des DUT. Les raisons de l'absence de ces deux phénomènes sont multiples : des températures trop faibles pour générer rapidement suffisamment de déplacement de matières dans les intermétalliques, un nœud technologique suffisamment maîtrisé par le fabricant pour avoir peu de défauts initiaux dans

l'oxyde de grille, etc. En revanche nous avons clairement mesuré du BTI et du HCI au cours de notre vieillissement.

En positionnant les DUT dans les conditions les plus favorables, nous avons réussi à générer du BTI. Ces conditions ont été plus de 12 000 heures de vieillissement à 115°C, une suralimentation au niveau cœur logique de V_{nom}+50% et un signal de stress constant. Dans cet état, les dégradations ont atteint -10%. En plus de cela, nous avons proposé plusieurs modèles de dégradation fréquentielle en accord avec la littérature. La particularité de cette étude est la démarche mise en œuvre pour extraire le maximum d'information au niveau du BTI tout en effectuant des mesures au niveau RO. Le défi est de passer d'un ensemble de LUT6 au niveau transistor unitaire. La finalité de cette opération est de pouvoir distinguer le NBTI du PBTI sans avoir recours au test sous pointes sur wafer. Pour cela une architecture du fabricant a été supposée sur la base des derniers brevets déposés. Ensuite les informations de placement et routage des RO dans les LUT6 (notamment les entrées utilisées) ont été extraites de l'ISE. Enfin, un script dédié a permis de comptabiliser le nombre de NMOS et le nombre de NMOS impactés par le vieillissement et ayant une implication dans la dégradation des performances. Dans les limites de notre analyse (soumise à l'hypothèse de l'architecture), il semblerait que le NBTI soit dominant à basse tension et que le PBTI le deviendrait quant à lui à plus haute tension (au-dessus de V_{nom}+30% dans notre cas). Ce constat est conforté par les valeurs des paramètres des modèles établies en stress DC1 et DC0 au niveau RO. En DC1, ces valeurs indiquent du PBTI, alors qu'en stress DC0 nous modélisons plutôt du NBTI.

La mise en place d'essais à -30°C au niveau température de jonction n'a pas été vaine : nous avons observé des porteurs chauds. Les dégradations sont très faibles, par conséquent les résultats sont très bruités. Nous avons tout de même entrepris une méthode d'extraction du HCI en déduisant la part de BTI dans les dégradations observées. L'analyse de ces dérives nous prouve bien l'intervention du mécanisme HCI en faisant clairement apparaitre le lien avec le nombre de commutation (c.-à-d. la fréquence de stress multipliée par le temps).

Contrairement à la plupart des contributions sur le sujet, l'aspect statistique de ces mécanismes n'a pas été négligé. Les données ont été extrapolées selon les lois connues du mécanisme afin d'avoir un nombre suffisant de défaillance paramétriques. Ainsi une modélisation statistique des instants de défaillance a été faite. Le BTI ne représente pas plus une menace vis-à-vis de la fiabilité en conditions opérationnelles pour les technologies DSM que pour les technologies précédentes. Le HCI quant à lui est négligeable pour une utilisation usuelle.

Enfin, nous avons présenté différentes manière de traiter la fiabilité d'un FPGA présentant les deux types de dégradation (HCI et BTI). La méthode simple (somme des deux taux de défaillance) est trompeuse et non représentative de la fiabilité globale dans ce cas. Une nouvelle méthode plus réaliste a été proposée dans ce mémoire. Cette méthode repose sur la somme des dérives allant jusqu'au critère de défaillance.

Conclusion

Les motivations de cette thèse sont d'établir une meilleure évaluation et une meilleure compréhension de la fiabilité des circuits intégrés numériques. Pour ce faire, nous avons étudié séparément les différents mécanismes de dégradation. Ensuite, une démarche statistique rigoureuse a été mise en place afin d'estimer une fiabilité réaliste. Nous avons concentré cette étude sur les composants DSM car ils intègrent un saut technologique de par l'introduction d'un nouvel isolant de grille à haute permittivité.

Une grande partie de ce mémoire est consacrée aux mémoires Flash NAND. Nos essais ont montré que la température de stockage ainsi que de nombreux cycles d'écriture-effacement réduisent significativement la fiabilité en rétention de ces mémoires. Une analyse poussée de l'influence de la température pendant la phase d'écriture sur la durée de rétention de la donnée a été effectuée. Elle a révélé qu'un grand écart entre celle d'écriture et celle de lecture mène à une forte perte de fiabilité. Ce problème n'est pas lié à la physique du point mémoire à basse température mais à une dérive du circuit périphérique gérant l'écriture et la lecture des cellules. Ce constat signifie que la technologie des transistors à grille flottante n'est pas la source de ce problème.

Globalement, avec ce type de composant, le manque de fiabilité en rétention avéré - sans systèmes externes - est dû à l'architecture MLC mémoire et non à la finesse du nœud technologique. Cette thèse a montré comment les algorithmes de codes correcteurs d'erreurs (ECC), les mécanismes d'uniformisation d'usure ou bien le surdimensionnement permettent de rendre la fiabilité système acceptable. Ces corrections externes font partie intégrante du fonctionnement de ces circuits et sont à prendre en compte dans le calcul de la fiabilité.

Le second cas d'étude de ce mémoire concerne la fiabilité des FPGA. La structure de test utilisée pour mesurer des dérives est à base d'oscillateurs en anneaux. Nous n'avons observé ni électromigration ni TDDB. Cependant nous avons mesuré du BTI et du HCI. Après plus de 12 000 heures (16 mois) de vieillissement à 115°C, à V_{nom}+50% et sous un signal de stress constant, les dégradations fréquentielles relatives dues au BTI ont atteint -10%. L'extraction des dégradations par transistor dans une LUT6 – en s'appuyant sur les descriptions trouvées dans les brevets d'architecture Xilinx et les informations de l'outil de conception de circuit

pour FPGA – ont permis de distinguer le NBTI du PBTI. Sous l'hypothèse de l'architecture, il semblerait que le NBTI soit dominant à basse tension et que le PBTI le deviendrait quant à lui à plus haute tension.

Cependant le BTI ne représente pas pour les technologies DSM étudiées une rupture de fiabilité par rapport aux technologies antérieures en condition d'utilisation usuelle. Par ailleurs, les dégradations observées induites par les porteurs chauds sont très faibles, et par conséquent négligeables pour une utilisation usuelle. L'analyse de ces dérives à basse température nous prouve bien l'apparition du mécanisme HCI en mettant clairement en évidence le lien avec le nombre de commutations.

Lorsque plusieurs mécanismes catalectiques apparaissent simultanément, on calcule le taux de défaillance global en effectuant la somme des deux. Dans le cas présent de mécanismes entrainant une dérive paramétrique, cette méthode est erronée et non représentative de la fiabilité. En effet, ces mécanismes induisent des dégradations sur un même paramètre ce qui invalide la somme des taux de défaillance. Nous avons proposé une autre méthode plus réaliste reposant sur la somme des dérives allant jusqu'au critère de défaillance dans ce mémoire.

Perspectives

Dans le cadre de la poursuite de cette étude, il serait judicieux d'apporter certaines améliorations. Certains choix qui ont été faits sur la base des connaissances et des moyens du début de la thèse, et qui pourrait être fait différemment aujourd'hui avec notre retour d'expérience.

Mémoire Flash

Le nombre de bits par cellule mémoire ne fait que croitre, par conséquent les ECC sont capitaux pour maintenir un niveau de fiabilité constant. Dans le cadre d'une nouvelle étude sur le sujet des mémoires, il serait pertinent de mesurer la fiabilité au niveau « bit » et non « page », c.-à-d. comptabiliser le nombre de bit défaillant en rétention et non le nombre de page. Ainsi les prévisions de temps de rétention estimées pourront être adaptées à plusieurs niveaux d'ECC.

De plus, l'enregistrement de chaque page lue permettrait également de suivre l'évolution des données lues au cours du temps. Ceci faciliterait ainsi la déduction de l'organisation des données MLC en fonction de la tension seuil. Pour ce faire, il suffirait de comparer les

données lues sur une page spécifique au cours du temps pour en déduire les différents états logiques implémentés dans la cellule mémoire. Cette modification engendrera cependant une quantité colossale de données à enregistrer lors de chaque mesure.

FPGA

La plus grande imprécision de notre étude sur les FPGA est la méconnaissance de l'architecture interne. Ce point est en effet difficile à combler sans avoir accès à des données relevant du secret industriel.

Dans notre première version du banc de vieillissement des FPGA, nous avons dû arrêter les essais à cause de la forte variabilité des mesures. Cette variabilité provenait d'une stabilité imparfaite de la tension d'alimentation et de la température du DUT. Il ne faut en aucun cas négliger la qualité de l'asservissement en température et en tension lors du vieillissement et de la mesure d'oscillateur en anneaux. Il faut bien prendre en compte le phénomène d'IR Drop en positionnant le capteur servant à régler la tension de vieillissement au plus près du FPGA. De même pour la température, il faut également ne pas négliger l'auto-échauffement du FPGA - sous une forte suralimentation électrique notamment.

Le circuit de mesure du rapport cyclique d'oscillation du RO peut gagner en précision. Une piste proposée est de mesurer le rapport cyclique d'un RO avant et après un étage inverseur. La mesure supplémentaire du signal complémentaire permettrait au moins de détecter un biais de mesure à haute tension du fait de la dissymétrie des portes logiques.

Le dernier point à améliorer sur le banc FPGA serait la précision du temps pendant lequel on comptabilise le nombre d'oscillation du RO afin d'en déduire sa fréquence. Il serait intéressant d'utiliser une horloge d'échantillonnage externe au DUT. Cela permettrait d'être certain d'avoir une horloge stable au cours du temps. Une autre approche actuellement en développement au laboratoire IMS est d'utiliser les signaux GPS pour obtenir une haute précision sur cette mesure temporelle.

Les études d'évaluation de la fiabilité sont longues, et par conséquent elles ont du mal à suivre la technologie. L'évolution des technologies fait de plus apparaître de nouvelles architectures qui obligent les modèles de fiabilité à évoluer. Les mémoires Flash sont aujourd'hui en 3D, ce qui amène des problématiques bien loin des technologies planaires DSM. Concernant les circuits numériques complexes, le nouvel horizon technologique semble être les transistors Fin FET. La structure même de ce nouveau gabarit amène des problématiques à comprendre et à évaluer de nouveau. Cependant, toute la méthodologie

développée dans ce mémoire peut être réutilisée sur ces nouvelles technologies afin d'en évaluer fidèlement leur fiabilité.

Publications

Colloques et congrès internationaux avec actes à diffusion publique :

- **J. Coutet**, F. Marc, F. Dozolme, R. Guétard, A. Janvresse, P. Lebossé, A. Pastre, et J.-C. Clement, « Influence of temperature of Storage, Write and Read operations on Multiple Level Cells NAND Flash memories », présenté à European Symposium on Reliability of Electron Devices, Failure Physics and Analysis, Aalborg, Denmark, oct-2018.
- **J. Coutet**, E. Doche, R. Guetard, A. Janvresse, S. Lavagne, P. Lebosse, A. Pastre, M. Sarlotte, C. Moreau, F. Marc, et F. Bayle, « Long term accelerated ageing of an ASIC dedicated to cryptographic application », présenté à European Symposium on Reliability of Electron Devices, Failure Physics and Analysis, Toulouse, France, sept-2019.

Articles dans une revue internationale avec comité de lecture :

- **J. Coutet**, F. Marc, F. Dozolme, R. Guétard, A. Janvresse, P. Lebossé, A. Pastre, et J.-C. Clement, « Influence of temperature of storage, write and read operations on multiple level cells NAND flash memories », *Microelectronics Reliability*, vol. 88‑90, p. 61 ‑ 66, 2018.
- **J. Coutet**, E. Doche, R. Guetard, A. Janvresse, S. Lavagne, P. Lebossé, A. Pastre, M. Sarlotte, C. Moreau, F. Marc, et F. Bayle, « Long term accelerated ageing of an ASIC dedicated to cryptographic application », *Microelectronics Reliability*, vol. 100‑101C, 2019.

Workshop national :

- **J. Coutet**, J.-C. Clement et P. Carton « Fiabilité composants numériques DSM », présenté à Normandy Reliability Technology Workshop, Saint Etienne du Rouvray, France, 10-oct-2019.

Rapports industriels de contrat :

- M. Sarlotte, **J. Coutet**, E. Doche, et J. Soufflet, « Rapport d'analyse des dérives et des défaillances », Thales, Etude de la durée de vie des technologies très submicroniques pour les applications cryptographiques (PEA FAST), 159 pages, 2018.
- J.-C. Clement, **J. Coutet**, A. Gigliati, F. Marc, R. Guétard, F. Bayle, et P. Carton, « Modèles de fiabilité FIDES - Poste 1 », Thales, ProgressIon deS Techniques de fiabIlité préviSionnelle (PEA PISTIS), 195 pages, 2020.

Bibliographie

[1] E. L. Kaplan et P. Meier, « Nonparametric Estimation from Incomplete Observations », *J. Am. Stat. Assoc.*, vol. 53, n° 282, p. 457‑481, 1958.

[2] D. Rullière et D. Serant, « Généralisation de l'estimateur de Kaplan-Meier d'une loi de durée de maintien en présence d'observations tronquées à gauche. Extension à l'étude conjointe de deux durées de maintien. », *Bull. Fr. Actuar.*, vol. 1, n° 2, p. 97‑114, déc. 1997.

[3] Svante Arrhenius, « Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren », *Zeitschrift für Physikalische Chemie*, vol. 4, n° 1, p. 226‑248, 01-janv-1889.

[4] « JESD91A Method for Developing Acceleration Models for Electronic Component Failure Mechanisms », JEDEC SOLID STATE TECHNOLOGY ASSOCIATION, Méthodologie JESD91A, août 2003.

[5] J. W. McPherson, *Reliability Physics and Engineering: Time-To-Failure Modeling*, Third Edition. Springer, 2019.

[6] E. Takeda et N. Suzuki, « An empirical model for device degradation due to hot-carrier injection », *IEEE Electron Device Lett.*, vol. 4, n° 4, p. 111‑113, avr. 1983.

[7] P.-A. Cornillon et E. Matzner-Løber, « La régression linéaire simple », in *Régression avec R*, P.-A. Cornillon et E. Matzner-Løber, Éd. Paris: Springer Paris, 2011, p. 1‑28.

[8] P.-A. Cornillon et E. Matzner-Løber, « Validation du modèle », in *Régression avec R*, P.-A. Cornillon et E. Matzner-Løber, Éd. Paris: Springer Paris, 2011, p. 67‑88.

[9] J. Aldrich, « R.A. Fisher and the making of maximum likelihood 1912-1922 », *Stat. Sci*, vol. 12, n° 3, p. 162‑176, sept. 1997.

[10] S. M. Stigler, « The Epic Story of Maximum Likelihood », *Stat. Sci*, vol. 22, n° 4, p. 598‑620, nov. 2007.

[11] Gordon E. Moore, « Cramming more components onto integrated circuits », *Electronics*, vol. 38, n° 8, 19-avr-1965.

[12] Xilinx, Inc., « Xcell 32 », *Xcell*, 1999.

[13] L. Crippa, R. Micheloni, I. Motta, et M. Sangalli, « Nonvolatile Memories: NOR vs. NAND Architectures », in *Memories in Wireless Systems*, Springer, 2008, p. 29‑53.

[14] R. H. Fowler et L. Nordheim, « Electron Emission in Intense Electric Fields », *Contain. Pap. Math. Phys. Character*, vol. 119, n° 781, p. 173‑181, mai 1928.

[15] Seiichi Aritome, « Reliability Of NAND Flash Memory », in *Nand Flash Memory Technologies*, First Edition., John Wiley & Sons, Inc., 2016, p. 195‑272.

[16] Arie Tal, « Two Flash Technologies Compared: NOR vs NAND », M-Systems, White Paper 91-SR-012-04-8L, oct. 2002.

[17] « Error Correction Code (ECC) in Micron Single-Level Cell (SLC) NAND », Micron, Technical Note TN-29-63, 2011.

[18] Varsha Regulapati, « Error Correction Codes in NAND Flash Memory », Report, University of Texas, Austin, 2015.
[19] Douglas Sheldon et Michael Freie, « Disturb Testing in Flash Memories », National Aeronautics and Space Administration, JPL Publication 08-7 3/08, 2008.
[20] Ian Olson, « NAND Flash Memory Reliability in Embedded Computer Systems », Schweitzer Engineering Laboratories, Inc., White Paper LWP0015-01, déc. 2014.
[21] Toshiba, « Toshiba Makes Major Advances in NAND Flash Memory with 3-bit-per-cell 32nm generation and with 4-bit-per-cell 43nm technology », 11-févr-2009. [En ligne]. Disponible sur: http://www.toshiba.co.jp/about/press/2009_02/pr1102.htm.
[22] Mark Hawes, « How Micron FortisFlash Technology Improves Performance and Endurance », Micron, TECHNICAL MARKETING BRIEF, 2015.
[23] Erich F. Haratsch, « NAND Flash Media Management Algorithms », présenté à Flash Memory Summit 2016, Santa Clara, CA, 2016.
[24] Eiji Takeda, Ryuichi Izawa, Kazunori Umeda, et Ryo Nagai, « AC hot-carrier effects in scaled MOS devices », *29th Annual Proceedings Reliability Physics 1991*, Las Vegas, NV, USA, USA, p. 118 ‑ 122, 1991.
[25] Stewart Rauch, « Considerations for the Reliability Estimation of Silicon CMOS », in *Extreme Environment Electronics*, 1st Edition., John D. Cressler et H. Alan Mantooth, Éd. CRC Press, 2013, p. 455 ‑ 458.
[26] Mark White, « Scaled CMOS Technology Reliability Users Guide », National Aeronautics and Space Administration, JPL Publication 09-33 01/10, 2010.
[27] S. Tan, M. Tahoori, T. Kim, S. Wang, Z. Sun, et S. Kiamehr, « Introduction », in *Long-Term Reliability of Nanometer VLSI Systems: Modeling, Analysis and Optimization*, S. Tan, M. Tahoori, T. Kim, S. Wang, Z. Sun, et S. Kiamehr, Éd. Cham: Springer International Publishing, 2019, p. 279‑304.
[28] E. Takeda, « Hot-Carrier and Wear-Out Phenomena in Submicron VLSI's », in *1985 Symposium on VLSI Technology. Digest of Technical Papers*, 1985, p. 2‑5.
[29] E. Takeda, Y. Ohji, et H. Kume, « High field effects in MOSFETS », in *1985 International Electron Devices Meeting*, 1985, p. 60‑63.
[30] Mohammad NAOUSS, « Conception et exploitation d'un banc d'auto-caractérisation pour la prévision de la fiabilité des circuits numériques programmables », Thèse, Université de Bordeaux, 2016.
[31] A. Bensoussan, « M-STORM reliability model applied to DSM technologies », *2016 IEEE Nanotechnology Materials and Devices Conference (NMDC)*, 2016.
[32] M. Naouss et F. Marc, « FPGA LUT delay degradation due to HCI: Experiment and simulation results », *Microelectronics Reliability*, vol. 64, p. 31‑35, sept-2016.
[33] A. Bravaix, M. Saliva, F. Cacho, X. Federspiel, C. Ndiaye, S. Mhira, E. Kussener, E. Pauly, et V. Huard, « Hot-carrier and BTI damage distinction for high performance digital application in 28nm FDSOI and 28nm LP CMOS nodes », in *2016 IEEE 22nd International Symposium on On-Line Testing and Robust System Design (IOLTS)*, 2016, p. 43‑46.
[34] E. Takeda, N. Suzuki, et T. Hagiwara, « Device performance degradation to hot-carrier injection at energies below the Si-SiO2energy barrier », in *1983 International Electron Devices Meeting*, 1983, p. 396‑399.

[35] Chenming Hu, « Lucky-electron model of channel hot electron emission », in *1979 International Electron Devices Meeting*, 1979, p. 22‑25.
[36] Chenming Hu, « Hot-electron effects in MOSFETs », in *1983 International Electron Devices Meeting*, 1983, p. 176‑181.
[37] Chenming Hu, Simon C. Tam, Fu-Chieh Hsu, Ping-Keung Ko, Tung-Yi Chan, et K. W. Terrill, « Hot-electron-induced MOSFET degradation—Model, monitor, and improvement », *IEEE Trans. Electron Devices*, vol. 32, n° 2, p. 375‑385, févr. 1985.
[38] C. Guerin, V. Huard, A. Bravaix, M. Denais, J.M. Roux, F. Perrier, et W. Baks, « Combined effect of NBTI and channel hot carrier effects in pMOSFETs », présenté à 2005 IEEE International Integrated Reliability Workshop, S. Lake Tahoe, CA, USA, 2005, p. 10 ‑ 16.
[39] T. Di Gilio, « Etude de la fiabilité porteurs chauds et des performances des technologies CMOS 0. 13 µm-2nm », 2006.
[40] « JEP122H Failure Mechanisms and Models for Semiconductor Devices », JEDEC SOLID STATE TECHNOLOGY ASSOCIATION, Méthodologie JEP122H, sept. 2016.
[41] Xiaojun Li, Jin Qin, et Joseph B. Bernstein, « Compact Modeling of MOSFET Wearout Mechanisms for Circuit-Reliability Simulation », *IEEE Transactions on Device and Materials Reliability*, vol. 8, n° 1, p. 98 ‑ 121, mars-2008.
[42] Eric S. Snyder, Ashok Kapoor, et Clint Anderson, « The impact of statistics on hot-carrier lifetime estimates of n-channel MOSFETs », *Microelectronics Manufacturing and Reliability*, vol. 1802, 1992.
[43] A. Bravaix, C. Guerin, V. Huard, D. Roy, J.M. Roux, et E. Vincent, « Hot-Carrier acceleration factors for low power management in DC-AC stressed 40nm NMOS node at high temperature », *2009 IEEE International Reliability Physics Symposium*, Montreal, QC, Canada, p. 531 ‑ 548, 2009.
[44] A. Kerber, W. McMahon, et E. Cartier, « Voltage Ramp Stress for Hot-Carrier Screening of Scaled CMOS Devices », *IEEE Electron Device Letters*, vol. 33, n° 6, p. 749 ‑ 751, juin-2012.
[45] A. Kerber, S. Cimino, F. Guarin, et T. Nigam, « Assessing device reliability margin in scaled CMOS technologies using ring oscillator circuits », *2017 IEEE Electron Devices Technology and Manufacturing Conference (EDTM)*, Toyama, Japan, p. 28 ‑ 30, 2017.
[46] Ketul B. Sutaria, Pengpeng Ren, Abinash Mohanty, Xixiang Feng, Runsheng Wang, Ru Huang, et Yu Cao, « Duty cycle shift under static/dynamic aging in 28nm HK-MG technology », présenté à 2015 IEEE International Reliability Physics Symposium, Monterey, CA, USA, 2015.
[47] A. Kerber, P. Srinivasan, S. Cimino, P. Paliwoda, S. Chandrashekhar, Z. Chbili, S. Uppal, R. Ranjan, M.-I. Mahmud, D. Singh, P.P. Manik, J. Johnson, F. Guarin, T. Nigam, et B. Parameshwaran, « Device reliability metric for end-of-life performance optimization based on circuit level assessment », *2017 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, CA, USA, 2017.
[48] C. Ndiaye, « Etude de la fiabilité de type negative bias temperature instability (NBTI) et par porteurs chauds (HC) dans les filières CMOS 28nm et 14nm FDSOI », 2017.
[49] E. Takeda, Y. Nakagome, H. Kume, N. Suzuki, et S. Asai, « Comparison of characteristics of n-channel and p-channel MOSFET's for VLSI's », *IEEE Trans. Electron Devices*, vol. 30, n° 6, p. 675‑680, juin 1983.

[50] A. Bravaix, F. Cacho, X. Federspiel, C. Ndiaye, S. Mhira, et V. Huard, « Potentiality of healing techniques in hot-carrier damaged 28nm FDSOI CMOS nodes », *Proc. 27th Eur. Symp. Reliab. Electron Devices Fail. Phys. Anal.*, vol. 64, p. 163‑167, sept. 2016.
[51] M. J. de Jong, C. Salm, et J. Schmitz, « Towards understanding recovery of hot-carrier induced degradation », *29th Eur. Symp. Reliab. Electron Devices Fail. Phys. Anal. ESREF 2018*, vol. 88‑90, p. 147‑151, sept. 2018.
[52] Fernando Guarin, « Failure Mechanisms in Modern Integrated Circuits and Industry Best Practices for Reliability Degradation Predictions », in *Extreme Environment Electronics*, 1st Edition., John D. Cressler et H. Alan Mantooth, Éd. CRC Press, 2013, p. 443 ‑ 449.
[53] C.J. Christiansen, Baozhen Li, J. Gill, R. Filippi, et M. Angyal, « Via-depletion electromigration in copper interconnects », *IEEE Transactions on Device and Materials Reliability*, vol. 6, n° 2, p. 163 ‑ 168, juin-2006.
[54] James R. Black, « Mass Transport Of Aluminum By Momentum Exchange With Conducting Electrons », *6th Annual Reliability Physics Proceedings*, p. 148‑159, 1968.
[55] L. Doyen, X. Federspiel, L. Arnaud, F. Terrier, Y. Wouters, et V. Girault, « Electromigration multistress pattern technique for copper drift velocity and Black's parameters extraction », *2007 IEEE International Integrated Reliability Workshop Final Report*, S. Lake Tahoe, CA, USA, p. 74‑78, 2007.
[56] L. Doyen, X. Federspiel, L. Arnaud, Y. Wouters, et S. Courtas, « Multistress Pattern for Electromigration Test on Advanced Copper Interconnect ». .
[57] Wenping Wang, Shengqi Yang, Sarvesh Bhardwaj, Sarma Vrudhula, Frank Liu, et Yu Cao, « The Impact of NBTI Effect on Combinational Circuit: Modeling, Simulation, and Analysis », *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, n° 2, p. 173 ‑ 183, févr-2010.
[58] S. Chakravarthi, A. Krishnan, V. Reddy, C.F. Machala, et S. Krishnan, « A comprehensive framework for predictive modeling of negative bias temperature instability », *2004 IEEE International Reliability Physics Symposium. Proceedings*, Phoenix, AZ, USA, USA, p. 273‑282, 2004.
[59] A. E. Islam, N. Goel, S. Mahapatra, et M. A. Alam, « Reaction-Diffusion Model », in *Fundamentals of Bias Temperature Instability in MOS Transistors: Characterization Methods, Process and Materials Impact, DC and AC Modeling*, S. Mahapatra, Éd. New Delhi: Springer India, 2016, p. 181‑207.
[60] R. Wittmann, H. Puchner, L. Hinh, H. Ceric, A. Gehring, et S. Selberherr, « Impact of NBTI-driven parameter degradation on lifetime of a 90nm p-MOSFET », présenté à 2005 IEEE International Integrated Reliability Workshop, S. Lake Tahoe, CA, USA, 2005, p. 99 ‑ 102.
[61] S. V. S. V. Prabhu Deva Kumar et S. Akashe, « Reliability Analysis of Comparator: NBTI, PBTI, HCI, AGEING », in *Communication, Networks and Computing*, 2019, p. 606‑619.
[62] R. Alves Fonseca, L. Dilillo, A. Bosio, P. Girard, S. Pravossoudovitch, A. Virazel, et N. Badereddine, « Detecting NBTI Induced Failures in SRAM Core-Cells », *2010 28th VLSI Test Symposium (VTS)*, Santa Cruz, CA, USA, p. 75‑80, 2010.

[63] A. Kerber et E. Cartier, « Bias Temperature Instability Characterization Methods », in *Bias Temperature Instability for Devices and Circuits*, T. Grasser, Éd. New York, NY: Springer New York, 2014, p. 3‑31.
[64] Hiroo Masuda, Donald G. Pierce, Kazunori Nishitsuru, et Ken Machida, « Assessment of a 90nm PMOS NBTI in the form of products failure rate », *Proceedings of the 2005 International Conference on Microelectronic Test Structures, 2005. ICMTS 2005*, Leuven, Belgium, p. 89‑94, avr-2005.
[65] Ryo Kishida, Takuya Asuke, Jun Furuta, et Kazutoshi Kobayashil, « Extracting BTI-induced Degradation without Temporal Factors by Using BTI-Sensitive and BTI-Insensitive ring Oscillators », présenté à 2019 IEEE 32nd International Conference on Microelectronic Test Structures (ICMTS), Kita-Kyushu City, Fukuoka, Japan, Japan, 2019.
[66] J. Martin-Martinez, R. Rodriguez, et M. Nafria, « Simulation of BTI-Related Time-Dependent Variability in CMOS Circuits », in *Bias Temperature Instability for Devices and Circuits*, T. Grasser, Éd. New York, NY: Springer New York, 2014, p. 783‑810.
[67] Andreas Kerber, Siddarth A. Krishnan, et Eduard Albert Cartier, « Voltage Ramp Stress for Bias Temperature Instability Testing of Metal-Gate/High- k Stacks », *IEEE Electron Device Letters*, vol. 30, n° 12, p. 1347 ‑ 1349, déc-2009.
[68] Subhadeep Mukhopadhyay, Nilesh Goel, et Souvik Mahapatra, « A Comparative Study of NBTI and PBTI Using Different Experimental Techniques », *IEEE Transactions on Electron Devices*, vol. 63, n° 10, p. 4038 ‑ 4045, oct-2016.
[69] X. Wang, Seung-hwan Song, A. Paul, et C. H. Kim, « Fast characterization of PBTI and NBTI induced frequency shifts under a realistic recovery bias using a ring oscillator based circuit », in *2014 IEEE International Reliability Physics Symposium*, 2014, p. 6B.2.1‑6B.2.6.
[70] Mark White, « A Study of Nanometer Semiconductor Scaling Effects on Microelectronics Reliability », Thèse, University of Maryland, College Park, 2009.
[71] Jin Qin, Baoguang Yan, Yossi Shoshany, Druker Roy, Hezi Rahamim, Haim Marom, et Joseph B. Bernstein, « Study of Transistor and Product NBTI Lifetime Distributions », *2008 IEEE International Integrated Reliability Workshop Final Report*, S. Lake Tahoe, CA, USA, p. 64 ‑ 67, 2008.
[72] W.-T. K. Chien, Y. A. Zhao, Y. Zhu, et Y. Song, « Early detection and prediction of HKMG SRAM HTOL performance by WLR PBTI tests », *Proc. 27th Eur. Symp. Reliab. Electron Devices Fail. Phys. Anal.*, vol. 64, p. 185‑188, sept. 2016.
[73] Saman Kiamehr, Abdulazim Amouri, et Mehdi B. Tahoori, « Investigation of NBTI and PBTI induced aging in different LUT implementations », *2011 International Conference on Field-Programmable Technology*, New Delhi, India, 2011.
[74] Rui Gao, Zhigang Ji, Jian Fu Zhang, John Marsland, et Wei Dong Zhang, « As-grown-Generation Model for Positive Bias Temperature Instability », *IEEE Transactions on Electron Devices*, vol. 65, n° 9, p. 3662 ‑ 3668, sept-2018.
[75] Tibor Grasser, Wolfgang Gos, Victor Sverdlov, et Ben Kaczer, « The Universality of NBTI Relaxation and its Implications for Modeling and Characterization », présenté à 2007 IEEE International Reliability Physics Symposium Proceedings. 45th Annual, Phoenix, AZ, USA, 2007.

[76] J. Martin-Martinez, B. Kaczer, J. Boix, N. Ayala, R. Rodriguez, M. Nafria, X. Aymerich, P. Zuber, B. Dierickx, et G. Groeseneken, « Circuit-design oriented modelling of the recovery BTI component and post-BD gate currents », *2009 Spanish Conference on Electron Devices*, Santiago de Compostela, Spain, p. 156‑159, 2009.
[77] S. Ramey, J. Hicks, L. S. Liyanage, et S. Novak, « BTI recovery in 22nm tri-gate technology », *2014 IEEE International Reliability Physics Symposium*, Waikoloa, HI, USA, 2014.
[78] S. Mahapatra, « A Comprehensive Modeling Framework for DC and AC NBTI », in *Bias Temperature Instability for Devices and Circuits*, T. Grasser, Éd. New York, NY: Springer New York, 2014, p. 349‑378.
[79] Souvik Mahapatra, *Fundamentals of Bias Temperature Instability in MOS Transistors: Characterization Methods, Process and Materials Impact, DC and AC Modeling*, vol. 52. Springer India, 2016.
[80] T. Aichinger, G. Pobegen, et M. Nelhiebel, « Application of On-Chip Device Heating for BTI Investigations », in *Bias Temperature Instability for Devices and Circuits*, T. Grasser, Éd. New York, NY: Springer New York, 2014, p. 33‑51.
[81] Z. Chbili et A. Kerber, « Self-heating impact on TDDB in bulk FinFET devices: Uniform vs Non-uniform Stress », *2016 IEEE International Integrated Reliability Workshop (IIRW)*, South Lake Tahoe, CA, USA, p. 45 ‑ 48, 2016.
[82] M. Arabi, X. Federspiel, F. Cacho, M. Rafik, A.-P. Nguyen, X. Garros, et G. Ghibaudo, « Temperature dependence of TDDB at high frequency in 28FDSOI », *Microelectron. Reliab.*, vol. 100‑101, p. 113422, sept. 2019.
[83] C. LaRow, Y. Liu, Z. Chbili, et A. Gondal, « Fast TDDB for early reliability monitoring », *2016 IEEE International Integrated Reliability Workshop (IIRW)*, South Lake Tahoe, CA, USA, p. 53 ‑ 56, 2016.
[84] R. Ranjan, Y. Liu, T. Nigam, A. Kerber, et B. Parameshwaran, « Impact of AC voltage stress on core NMOSFETs TDDB in FinFET and planar technologies », *2017 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, CA, USA, 2017.
[85] Fengming Lu, Jiang Shao, Xiaoyu Liu, et Xinghao Wang, « Validation test method of TDDB Physics-of-Failure models », *Proceedings of the IEEE 2012 Prognostics and System Health Management Conference (PHM-2012 Beijing)*, Beijing, China, 2012.
[86] Joe McPherson, Vijay Reddy, Kaustav Banerjee, et Huy Le, « Comparison of E and 1/E TDDB models for SiO2 under long-term/low-field test conditions », *International Electron Devices Meeting 1998*, San Francisco, CA, USA, USA, p. 171 ‑ 174, 1998.
[87] Dae-Hyun Kim et Linda Milor, « Memory reliability estimation degraded by TDDB using circuit-level accelerated life test », *2017 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, CA, USA, 2017.
[88] Tian Shen, Kong Boon Yeap, Cathryn Christiansen, et Patrick Justison, « Field acceleration factor extraction in MOL and BEOL TDDB », *2017 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, CA, USA, 2017.
[89] R. Muralidhar, E. Wu, T. Shaw, A. Kim, B. Li, P. Mclaughlin, J. Stathis, et G. Bonilla, « A stochastic model for impact of LER on BEOL TDDB », *2017 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, CA, USA, 2017.
[90] Tae-Young Jeong, Jinseok Kim, Myungsoo Yeo, Jonghyuk Park, Miji Lee, Sari Windu, Hyunjun Choi, Yuri Choi, Yunkyung Jo, Mi-ji Lee, et Sangwoo Pae, « Opportunities for further BEOL technology scaling using power-law IMD TDDB model on 10/14nm

BEOL process technologies and beyond », *2017 IEEE International Interconnect Technology Conference (IITC)*, Hsinchu, Taiwan, 2017.
[91] Kyunghwan Lee, Myounggon Kang, Seongjun Seo, Duckseoung Kang, Shinhyung Kim, Dong Hua Li, et Hyungcheol Shin, « Activation Energies (Ea) of Failure Mechanisms in Advanced NAND Flash Cells for Different Generations and Cycling », *IEEE Transactions on Electron Devices*, vol. 60, n° 3, p. 1099 ‑ 1107, mars-2013.
[92] Kyunghwan Lee, Myounggon Kang, Seongjun Seo, Dong Hua Li, Jungki Kim, et Hyungcheol Shin, « Analysis of Failure Mechanisms and Extraction of Activation Energies (Ea) in 21-nm nand Flash Cells », *IEEE Electron Device Letters*, vol. 34, n° 1, p. 48 ‑ 50, janv-2013.
[93] Yu Cai, Yixin Luo, Erich F. Haratsch, Ken Mai, et Onur Mutlu, « Data retention in MLC NAND flash memory: Characterization, optimization, and recovery », *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, Burlingame, CA, USA, 11-févr-2015.
[94] Kyunghwan Lee, Myounggon Kang, Yuchul Hwang, et Hyungcheol Shin, « Accurate Lifetime Estimation of Sub-20-nm NAND Flash Memory », *IEEE Transactions on Electron Devices*, vol. 63, n° 2, p. 659 ‑ 667, févr-2016.
[95] Kyunghwan Lee, Duckseoung Kang, Hyungcheol Shin, Sangjin Kwon, Shinhyung Kim, et Yuchul Hwang, « Analysis of failure mechanisms in erased state of sub 20-nm NAND Flash memory », *2014 44th European Solid State Device Research Conference (ESSDERC)*, Venice, Italy, p. 58 ‑ 61, 2014.
[96] Laurence Montagner Morancho, « Nouvelle méthode de test en rétention de données de mémoires non volatiles », Thèse, Institut National Polytechnique de Toulouse, 2004.
[97] K. M. Farhan Shahil, Md. Nayeem Arafat, Q. D. M. Khosru, et M. Rezwan Khan, « Study of charge trapping/detrapping mechanism in SiO2/HfO2 stack gate dielectrics considering two-way detrapping », *2007 International Workshop on Electron Devices and Semiconductor Technology (EDST)*, Tsinghua University, China, p. 117 ‑ 120, 2007.
[98] Jim Cooke, « The Inconvenient Truths of NAND Flash Memory », présenté à Flash Memory Summit 2007, Santa Clara, CA USA, août-2007.
[99] Robert Frickey, « Data Integrity on 20nm SSDs », présenté à Flash Memory Summit 2012, Santa Clara, CA, 2012.
[100] Chris, « Flash Endurance Testing », *Hypnocube*, 25-nov-2014. [En ligne]. Disponible sur: http://hypnocube.com/2014/11/flash-endurance-testing/.
[101] Jérémy Postel-Pellerin, « Fiabilité des Mémoires Non-Volatiles de type Flash en architectures NOR et NAND », Thèse, Universite D'Aix-Marseille 1, 2008.
[102] Yu Cai, Yixin Luo, Saugata Ghose, et Onur Mutlu, « Read Disturb Errors in MLC NAND Flash Memory: Characterization, Mitigation, and Recovery », *2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, Rio de Janeiro, Brazil, p. 438 ‑ 449, 2015.
[103] Yu Cai, Erich F. Haratsch, Onur Mutlu, et Ken Mai, « Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling », présenté à 2013 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, France, 2013.

[104] Debao Wei, Liyan Qiao, Xiaoyu Chen, Mengqi Hao, et Xiyuan Peng, « SREA: A self-recovery effect aware wear-leveling strategy for the reliability extension of NAND flash memory », *Microelectron. Reliab.*, vol. 100‑101, p. 113433, sept. 2019.

[105] Marcello Calabrese, Carmine Miccoli, Christian Monzio Compagnoni, Luca Chiavarone, Silvia Beltrami, Andrea Parisi, Sebastiano Bartolone, Andrea L. Lacaita, Alessandro S. Spinelli, et Angelo Visconti, « Accelerated reliability testing of flash memory: Accuracy and issues on a 45nm NOR technology », présenté à Proceedings of 2013 International Conference on IC Design & Technology (ICICDT), Pavia, Italy, 2013, p. 37‑40.

[106] Cristian Zambelli, Luca Crippa, Rino Micheloni, et Piero Olivo, « Cross-Temperature Effects of Program and Read Operations in 2D and 3D NAND Flash Memories », présenté à 2018 International Integrated Reliability Workshop (IIRW), South Lake Tahoe, CA, USA, USA, 2018.

[107] C. Zambelli, E. Ferro, L. Crippa, R. Micheloni, et P. Olivo, *Dynamic V TH Tracking for Cross-Temperature Suppression in 3D-TLC NAND Flash*. 2019.

[108] Jongwoo Park, Miji Lee, Hanbyul Kang, Wooram Ko, Eunkyeong Choi, Junsik Im, Minwoo Lee, Dohwan Chung, Jinchul Park, Sangchul Shin, et Sangwoo Pae, « Effects of front-end-of line process variations and defects on retention failure of flash memory: Charge loss/gain mechanism », *2015 IEEE International Reliability Physics Symposium*, Monterey, CA, USA, 2015.

[109] Pekka Muroke, « Flash Memory Field Failure Mechanisms », *2006 IEEE International Reliability Physics Symposium Proceedings*, San Jose, CA, USA, p. 313‑316, 2006.

[110] W.H. Lee, Dong-Kyu Lee, Young-Min Park, Keon-Soo Kim, Kun-Ok Ahn, et Kang-Deog Suh, « Data retention failure in NOR flash memory cells », in *2001 IEEE International Reliability Physics Symposium Proceedings. 39th Annual (Cat. No.00CH37167)*, Orlando, FL, USA, USA, 2001.

[111] Carmine Miccoli, Christian Monzio Compagnoni, Luca Chiavarone, Silvia Beltrami, Andrea L. Lacaita, Alessandro S. Spinelli, et Angelo Visconti, « Assessment of Distributed-cycling Schemes on 45nm NOR Flash Memory Arrays », présenté à 2012 IEEE International Reliability Physics Symposium (IRPS), Anaheim, CA, USA, 2012.

[112] Bogdan Govoreanu et Jan Van Houdt, « On the Roll-Off of the Activation Energy Plot in High-Temperature Flash Memory Retention Tests and its Impact on the Reliability Assessment », *IEEE Electron Device Letters*, vol. 29, n° 2, p. 177‑179, févr-2008.

[113] Barbara De Salvo, Gérard Ghibaudo, Georges Pananakakis, Gilles Reimbold, François Mondond, Bernard Guillaumot, et Philippe Candelier, « Experimental and theoretical investigation of nonvolatile memory data-retention », *IEEE Transactions on Electron Devices*, vol. 46, n° 7, p. 1518‑1524, juill-1999.

[114] Cherng-Ming Yih, Zhi-Hao Ho, Mong-Song Liang, et Steve S. Chung, « Characterization of hot-hole injection induced SILC and related disturbs in flash memories », *IEEE Transactions on Electron Devices*, vol. 48, n° 2, p. 300‑306, févr-2001.

[115] S. Liu et X. Zou, « QLC NAND study and enhanced Gray coding methods for sixteen-level-based program algorithms », *Microelectron. J.*, vol. 66, p. 58‑66, août 2017.

[116] L. Wasserman, « Hypothesis Testing and p-values », in *All of Statistics: A Concise Course in Statistical Inference*, L. Wasserman, Éd. New York, NY: Springer New York, 2004, p. 149‑173.
[117] D. J. Sherwin et A. Bossche, « Standby redundancy », in *The Reliability, Availability and Productiveness of Systems*, D. J. Sherwin et A. Bossche, Éd. Dordrecht: Springer Netherlands, 1993, p. 48‑56.
[118] Yousuke Miyake, Yasuo Sato, Seiji Kajihara, et Yukiya Miura, « Temperature and Voltage Estimation Using Ring-Oscillator-Based Monitor for Field Test », présenté à 2014 IEEE 23rd Asian Test Symposium, Hangzhou, China, 2014.
[119] R. M. Heiberger et B. Holland, « Multiple Regression—Regression Diagnostics », in *Statistical Analysis and Data Display: An Intermediate Course with Examples in R*, R. M. Heiberger et B. Holland, Éd. New York, NY: Springer New York, 2015, p. 345‑375.
[120] R. D. Cook, « Detection of Influential Observation in Linear Regression », *Technometrics*, vol. 19, n° 1, p. 15‑18, févr. 1977.
[121] H. Aguinis, R. K. Gottfredson, et H. Joo, « Best-Practice Recommendations for Defining, Identifying, and Handling Outliers », *Organ. Res. Methods*, vol. 16, n° 2, p. 270‑301, janv. 2013.
[122] Tao Pi et Patrick J. Crotty, « FPGA Lookuptable With Transmission Gate Structure For Reliable Low-Voltage Operation », US 6,809,552 B1, 26-oct-2004.
[123] Manoj Chirania et Venu M. Kondapalli, « Lookuptable Circuit Optionally Configurable As Two Or More Smaller Lookuptables With Independent Inputs », US 6,998,872 B1, 14-févr-2006.
[124] Julien Coutet, Emmanuel Doche, Romain Guetard, Aurelien Janvresse, Suzel Lavagne, Pierre Lebosse, Antonin Pastre, Michel Sarlotte, Christian Moreau, François Marc, et Franck Bayle, « Long term accelerated ageing of an ASIC dedicated to cryptographic application », *Microelectronics Reliability*, vol. 100‑101C, 2019.

Annexe A

Répartition de la population des pages en rétention NAND

File d'essai	Température		Période de lectures (en heures)	Nombre de PE initial	Nombre d'échantillons	Nombre de défaillances observées	Taux de censure
	vieillissement	écriture					
NAND 1	125°C	85°C	1206,27	1	589824	0	100,00%
	125°C	25°C	1206,27	1	368640	0	100,00%
	125°C	-40°C	1206,27	1	560128	27970	95,00%
	85°C	-10°C	1206,27	1	135168	0	100,00%
	85°C	85°C	1206,27	1	282624	0	100,00%
	85°C	25°C	1206,27	1	73728	0	100,00%
	85°C	-40°C	1206,27	1	380928	0	100,00%
	25°C	-10°C	1206,27	1	49152	0	100,00%
	25°C	85°C	1206,27	1	36864	0	100,00%
	25°C	25°C	1206,27	1	36864	0	100,00%
	25°C	-40°C	1206,27	1	135168	0	100,00%
	125°C	-10°C	1206,27	1	491520	0	100,00%
NAND 2	125°C	25°C	1416,43	1	1560576	0	100,00%
	85°C	25°C	1416,43	1	749568	0	100,00%
NAND 3	125°C	25°C	907,44	500	29696	13510	54,50%
	85°C	25°C	907,44	500	7168	0	100,00%
	25°C	25°C	907,44	500	512	0	100,00%

	125°C	25°C	907,44	1500	29696	16740	43,60%
	85°C	25°C	907,44	1500	19456	2401	87,70%
	25°C	25°C	907,44	1500	1024	0	100,00%
	125°C	25°C	907,44	3000	21504	20832	3,10%
	85°C	25°C	907,44	3000	13312	10862	18,40%
	25°C	25°C	907,44	3000	1536	0	100,00%
NAND 4	85°C	25°C	0,75	1	3530	3530	0,00%
	85°C	-40°C	0,75	1	305556	263895	13,60%
	25°C	-40°C	0,75	1	21934	21934	0,00%
	110°C	85°C	0,75	1	354030	354030	0,00%
	110°C	25°C	0,75	1	395656	395656	0,00%
	110°C	-40°C	0,75	1	451200	434299	3,70%
	85°C	85°C	24	1	614400	0	100,00%
	85°C	25°C	24	1	610870	0	100,00%
	85°C	-40°C	24	1	155244	0	100,00%
	25°C	85°C	24	1	460800	0	100,00%
	25°C	25°C	24	1	153600	0	100,00%
	25°C	-40°C	24	1	131666	0	100,00%
	110°C	85°C	24	1	106770	0	100,00%
	110°C	25°C	24	1	218744	178	99,90%
	110°C	-40°C	24	1	9600	0	100,00%
				TOTAL	9568256	1565837	83,6%

Annexe B

Tableau détaillé des différentes configurations de RO implémentées dans les FPGA

Le tableau suivant présente les différentes configurations de stress de RO implémentées. Il y a 96 RO dans ce tableau. Il correspond à la version 1 du banc de test (matrice à chaud 1 et matrice à froid). La version 2 du banc (matrice à chaud uniquement) comporte 96×2 RO, c.-à-d. les valeurs du tableau doublées.

N° RO	Type de RO	Longueur	Fréquence de stress	Rapport cyclique de stress
1	BUFFERS	9 LUTS	DC0	
2				
3			DC1	
4				
5			25 Hz	0,25
6				
7				0,5
8				
9				0,75
10				
11			50 MHz	0,25
12				
13				0,5
14				
15				0,75
16				
17			400 MHz	0,25
18				
19				0,5
20				
21				0,75
22				
23			800 MHz	0,5
24				
25	BUFFERS	17 LUTS	DC0	
26				
27			DC1	
28				
29			25 Hz	0,25
30				
31				0,5
32				
33				0,75
34				
35			50 MHz	0,25
36				
37				0,5
38				
39				0,75
40				
41			400 MHz	0,25
42				
43				0,5
44				
45				0,75
46				
47			800 MHz	0,5
48				

N° RO	Type de RO	Longueur	Fréquence de stress	Rapport cyclique de stress
49	INVERTERS	9 LUTS	DC0	
50				
51			DC1	
52				
53			25 Hz	0,25
54				
55				0,5
56				
57				0,75
58				
59			400 MHz	0,25
60				
61				0,5
62				
63				0,75
64				
65	XOR	9 LUTS	DC0	
66				
67			DC1	
68				
69			25 Hz	0,25
70				
71				0,5
72				
73				0,75
74				
75			400 MHz	0,25
76				
77				0,5
78				
79				0,75
80				
81	LUT2	9 LUTS	DC0	
82				
83			DC1	
84				
85			25 Hz	0,25
86				
87				0,5
88				
89				0,75
90				
91			400 MHz	0,25
92				
93				0,5
94				
95				0,75
96				

Titre : Étude de la fiabilité et des mécanismes de dégradation dans les composants numériques de dernière génération

Résumé : La réduction des tailles aux niveaux transistors des composants électroniques commerciaux est rendue possible par l'utilisation de nouveaux matériaux au niveau de l'oxyde de grille notamment pour les DSM. Afin de pouvoir utiliser en toute confiance ces composants pour des applications en environnements sévères, il est nécessaire d'en évaluer la fiabilité. Les deux catégories de composants choisies pour représenter les technologies DSM sont les mémoires Flash et les FPGA. Nous avons étudié les différents mécanismes de dégradation séparément au moyen de vieillissements accélérées. Ensuite, les résultats sont analysés et une démarche statistique rigoureuse a été mise en place afin d'estimer une fiabilité réaliste.

Nos essais ont montré que la température de stockage ainsi que les cycles d'écriture-effacement réduisent nettement la fiabilité en rétention des mémoires Flash NAND. Une particularité de cette étude est la mise en évidence qu'un grand écart entre la température d'écriture et celle de lecture mène à une forte perte de fiabilité en rétention qui n'est pas liée à la physique du point mémoire à basse température mais à une dérive du circuit périphérique gérant l'écriture et la lecture des cellules. D'autre part, nous avons montré que le manque de fiabilité en rétention avéré avec ce type de composant est dû à l'architecture MLC et non à la finesse du nœud technologique. Les codes correcteurs d'erreurs, les mécanismes d'uniformisation d'usure ou bien le surdimensionnement permettent de rendre la fiabilité acceptable. Ils font par conséquent partie intégrante du système et sont à prendre en compte dans le calcul de la fiabilité.

Pour mesurer des dérives dans circuits numériques, nous avons utilisé des oscillateurs en anneaux implantés dans des FPGA. Nous avons mesuré du BTI et du HCI mais nous n'avons pas observé d'électromigration ou de TDDB. Ces dérives ont été modélisées pour chaque mécanisme. Une extraction des dégradations par transistor sous l'hypothèse de l'architecture issue de brevets ont permis de distinguer le NBTI du PBTI. Par ailleurs l'analyse des dérives - très faibles - à basse température nous prouve l'intervention du mécanisme HCI en mettant clairement en évidence le lien avec le nombre de commutations. Enfin, pour évaluer la fiabilité de manière rigoureuse - la somme des deux taux de défaillance HCI et BTI n'étant pas représentative de la fiabilité - une autre méthode plus réaliste reposant sur la somme des dérives allant jusqu'au critère de défaillance a été proposée dans cette étude.

Mots clés : Fiabilité, DSM, FPGA, HCI, NBTI, PBTI, CMOS, VLSI, Vieillissement, Modélisation, Statistique, Flash NAND

Title: Study of the reliability and mechanisms of degradation in DSM digital integrated circuits

Abstract: The downscaling of transistors in commercial electronic circuits has been permitted by the use of new materials in gate oxide (especially for DSM bound). In order to fully trust this kind of chips in extreme environments, it is necessary to make a proper reliability assessment. We choose two types of integrated circuits to carry this study: NAND flash memories and FPGA. Accelerated ageing has been applied to distinctly activate the various degradation mechanisms. Then results are analyzed and a precise statistical approach leads us to a realistic estimation of its reliability.

Our ageing tests of NAND flash memories showed that storage temperature as well as many program-erase cycles significantly decrease the retention time of data. Our method demonstrates that a wide gap between the temperature of writing and the temperature of reading leads to a meaningful lack of reliability in retention of data. This issue is not due to the physic of memory cells at low temperature but it is due to drifts of the peripheral management circuit (in charge of writing and reading operations). However we have confirmed this weak reliability is caused by MLC design and not by the technologic node. Error code correction, wear leveling or even over-provisioning allow sufficient reliability. Therefore they are an important part of global memory and must be taken in account in the reliability calculation.

Digital circuit ageing has been monitored by means of ring oscillators embedded in FPGA. We well measured some BTI and few HCI, but we did not succeed in involving any electromigration or TDDB. Drifts of both mechanisms are modeled. A new methodology based on the assumption of LUT design from patents is described and applied in order to extract degradations at transistor level: NBTI and PBTI can be set apart.. Moreover the analysis of small drifts occurred at low temperature which is evidence that HCI was involved. These low temperature degradations are correlated with the number of commutations. Finally, the estimation of reliability based on the sum of HCI and BTI failure rates gives false result. We present a proper approach to estimate reliability. It considers the sum of both drifts until a user's determined criterion of failure is reached.

Keywords: Reliability, DSM, FPGA, HCI, NBTI, PBTI, CMOS, VLSI, Ageing, Modeling, Statistic, NAND flash

---

Unité de recherche

Laboratoire de l'Intégration du Matériau au Système, Talence, France