Moniteurs de Vieillissement In Situ : méthodologie d'intégration et application à la gestion dynamique de la fiabilité

Ahmed Benhassain

▶ **To cite this version:**

Ahmed Benhassain. Moniteurs de Vieillissement In Situ : méthodologie d'intégration et application à la gestion dynamique de la fiabilité. Micro et nanotechnologies/Microélectronique. Université Grenoble Alpes, 2017. Français. NNT : 2017GREAT126 . tel-03367612

HAL Id: tel-03367612

https://theses.hal.science/tel-03367612

Submitted on 6 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de
DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES
Spécialité : **Micro et nano électronique**
Arrêté ministériel : 25 mai 2016

Présentée par
Ahmed Benhassain

Thèse dirigée par **Lorena Anghel, professeur à l'institut polytechnique de Grenoble,** et
codirigée par **Florian Cacho, ingénieur à STMicroelectronics**
préparée au sein du **Laboratoire TIMA**
dans **l'École Doctorale « Electronique, Electrotechnique, Automatique et traitement de signal »**

Moniteurs de Vieillissement in-situ: Méthodologie d'intégration et application à la gestion dynamique de la fiabilité

Soutenance de thèse prévue le **29 Mai 2017**,
devant le jury composé de :
M. Michael Nicolaïdis,
Directeur de recherche CNRS, TIMA-Grenoble, Président du jury
Mme Cristelle Maneux
Professeur des universités, IMS-Bordeaux, Rapporteur
M. Giorgio Di Natalie
Directeur de recherche CNRS, LIRMM-Montpellier, Rapporteur
M. Olivier Heron
Ingénieur de recherche, CEA, Examinateur
Mme Lorena Anghel
Professeur des universités, TIMA-Grenoble, Co-Directeur de thèse
M. Florian Cacho
Ingénieur, STMicroelectronics-Crolles, Co-Directeur de thèse

Remerciements

LISTE DES ACRONYMES

ABB	Adaptive Body Bias
AC	Alternative Current
AOCV	Advanced On Chip Variation
API	Application Programmer Interface
AVS	Adaptive Voltage Scaling
BCH	Bose, Ray-Chaudhuri et Hocquenghem
BTI	Bias Temperature Instability
CAD	Computer Assisted Design
CCSM	Composite Current Source Model
CDF	Cumulative Distribution Function
CMOS	Complementary Metal-Oxide-Semiconductor
CPR	Critical Path Replica
DC	Direct Current
DIBL	Drain Induced Barrier Lowering
DVFS	Dynamic Voltage Frequency Scaling
ECSM	Effective Current Source Model
EOT	Effective Oxide Thickness
FBB	Forward Back Bias
FDSOI	Fully Depleted Silicon on Insulator
FF	Fast Fast
HCI	Hot Carrier Injection
HK	High-K
HVT	High Vth
IC	Integrated Circuit
IL	Interfacial Layer
IO	Input/Output
ISM	In-Situ monitor

ITRS	International Technology Roadmap for Semiconductors
LP	Low Power
LVT	Low Vth
MGC	Metal Gate Granularity
MC	Monte-Carlo
MOSFET	Metal-Oxide-Semiconductor Field Effect transistor
N/PBTI	Negative/Positive Bias Temperature Instability
OCV	On Chip Variation
RBB	Reverse Back Bias
RDD	Random Discrete Dopants
RO	Ring Oscillator
RVT	Regular Vth
SCE	Short Channel Effect
SPICE	Simulation Program with Integrated Circuit Emphasis
SRAM	Static Random Acces Memory
SS	Slew Slew
SP	Signal Probability
STA	Static Timing Analysis
SSTA	Statistical Static Timing Analysis
POCV	Parametric On Chip Variation
POFF	Point Of First Failure
STI	Shallow Trench Isolation
PDSOI	Partially Depleted Silicon on Insulator
PVT	Process, Voltage, Temperature
TDDB	Time Dependent Dielectric Breakdown
TT	Typical Typical
TWC	Time Windows Coverage
TW	Time Windows
UTB	Ultra-Thin Body

RESUME

Dans les technologies avancées, la course à la haute performance ainsi que la limitation de contraintes de consommation de puissance obligent les designers à mettre en place des nouvelles solutions pour optimiser ces paramètres des circuits intégrés. Une solution possible serait l'adaptation dynamique de la tension d'alimentation (AVS) et éventuellement la fréquence de fonctionnement selon les conditions d'utilisations du circuit. Elle utiliserait des capteurs de violation de délai, nommés aussi moniteurs in-situ. En pratique, il s'agit d'un circuit qui détecte des pré-erreurs, c'est-à-dire des transitions d'un chemin de propagation qui arrivent près du front d'horloge. En se basant sur ces moniteurs de pré-erreurs, la tension pourrait être ajustée et ceci bien avant qu'une erreur se produise. Ainsi, la consommation est optimisée en exploitant les marges de tension inutilisées, utilisées de façon traditionnelle dans le milieu industriel, par la conception d'un circuit dans le pire cas de procédé, tension, et température.

Ce travail effectué au sein de STMicroelectronics est développé sur quatre chapitres. Le premier chapitre présente les évolutions technologiques au sein de STMicroelectronics, la variabilité technologique, ainsi que les mécanismes de dégradation monotone de type Bias Temperature Instability (BTI) et sous injections de porteurs chauds (HCI), ou le vieillissement. Aussi, il aborde les impacts liés à ces mécanismes de dégradation, ainsi que les variations du procédé de fabrication, la tension et la température (PVT) sur le délai des portes et des chemins logiques. Le second chapitre est consacré d'une part à l'analyse de l'impact de diverses variations de type PVTs et du vieillissement sur un circuit numérique. D'autre part, le chapitre 2 présente également l'état d'art des moniteurs publiés dans la littérature, ainsi que la caractérisation du moniteur in-situ utilisés dans les circuits de test. Dans le troisième chapitre, nous présentons d'abord la méthodologie d'insertion des moniteurs dans le flot de conception des circuits digitaux. Dans un second temps, nous présentons la méthodologie de choix des chemins critiques qui seront observés par des moniteurs, en se basant sur les délais des chemins obtenus par l'analyse statique temporelle (STA) avant et après vieillissement. Le quatrième chapitre présente des résultats expérimentaux des circuits contenant des moniteurs, fabriqués dans les technologies 45nm, 28LP, et 28FDSOI. Ces circuits sont testés dans différents conditions d'utilisations : tensions, températures et usages du circuit mettant en œuvre la capacité des moniteurs de faire face à ces types de variations. Une proposition de méthodologie d'adaptation des paramètres de fonctionnement du circuit tels que la tension d'alimentation y est aussi présentée.

Table des matières

Liste des acronymes 3
Résumé 5
Introduction 9
1. Contexte global de la thèse 9
2. La fiabilité en microélectronique 10
3. La conception en vue de l'amélioration de la fiabilité 12
4. L'objectif de la thèse 15
5. Plan de la thèse 16
Chapitre 1 : de la physique du CMOS au circuit numérique 18
1.1 Introduction 18
1.2 Transistor MOS 18
1.2.1 Principe de fonctionnement 18
1.2.2 Fonctionnement de la capacité MOS 19
1.2.3 Paramètres caractéristiques du transistor MOS 21
1.3 Impact de la réduction des dimensions de la grille : effets secondaires 24
1.4 Evolution technologique du 40nm vers le 28nm sur substrat massif 26
1.5 Evolution technologique du 28nm sur substrat massif vers le 28nm sur substrat SOI 28
1.5.1 Avantages de la technologie sur substrat SOI 28
1.5.2 Tension de seuil du transistor FDSOI 30
1.6 Mécanismes de dégradation du transistor MOS 32
1.6.1 Mécanisme de dégradation BTI 32
1.6.2 Modélisation de la dégradation par NBTI 35
1.6.3 Mécanisme de dégradation par porteurs chauds HCI 36
1.7 Impact des mécanismes de dégradation sur le délai des portes logiques 38
1.7.1 Effet AC et DC sur un chemin de propagation à haute température 38
1.7.2 Effet AC sur un chemin de propagation à faible température 40
1.8 Impact des sources de variation sur le délai des chemins logiques 41
1.8.1 Les types de variabilités 41
1.8.2 Variations de la tension d'alimentation 44
1.8.3 Variation de la température 45
1.9 Conclusion 46

Chapitre 2 : Moniteurs in-situ et la gestion dynamique de la fiabilité 48
2.1 Introduction 48
2.2 Evaluation de l'impact des PVTs et du vieillissement sur les portes logiques d'une technologie numérique 49
2.2.1 Méthodes d'analyse de délais 49
2.2.2 Prise en compte des variations locales 52
2.2.3 Prise en compte du vieillissement 53
2.3 Impact global sur le délai de portes 60
2.4 Moniteurs in-situ : état de l'art 62
2.4.1 Oscillateur en anneau. 62
2.4.2 Moniteur de détection d'erreurs 63
2.4.3 Utilisation des détecteurs d'erreurs 65
2.4.4 Moniteurs de détection de pré-erreur 67
2.4.5 Comparaison des moniteurs de pré-erreur et des moniteurs d'erreur 70
2.5 Techniques de masquage d'erreur 71
2.6 Mesure des performances avec des moniteurs in-situ 72
2.7 Caractérisation des moniteurs de pré-erreur 73
2.7.1 La variabilité de la fenêtre de détection 75
2.8 Conclusion 76
CHAPITRE 3 : Stratégies de sélection et d'intégration des moniteurs in-situ 78
3.1 Introduction 78
3.2 Revue des méthodes de sélection des chemins critiques 79
3.3 Méthodologie d'intégration des moniteurs in-situ 82
3.3.1 Flot de conception des circuits numériques 82
3.3.2 Conception des moniteurs in-situ 84
3.3.3 Bilan d'intégration des moniteurs in-situ 86
3.3.4 Couverture des chemins critiques par les moniteurs in-situ 90
3.3.5 Evolution des chemins monitorés en prenant compte les PVTs 92
3.3.6 Méthodologie d'insertion des moniteurs basée sur les délais des chemins sous-critiques 94
3.3.7 Méthodologie de calcul de la criticité des bascules 98
3.3.8 Evaluation de l'approche statistique 100
3.3.9 Impact du vieillissement sur la sélection des chemins critiques à monitorer 102
3.4 Conclusion 106

CHAPITRE 4 : Application des moniteurs in-situ à l'adaptation de la tension d'alimentation et la fréquence 108
4.1 Introduction 108
4.2 Véhicule de test considéré 109
4.3 Implémentation du moniteur in-situ 111
4.4 Résultats expérimentaux 113
4.4.1 Détermination de la plage de fonctionnement du BCH à l'aide des moniteurs 113
4.4.2 Comportement du circuit face aux conditions d'usage 115
4.4.3 Comportement du circuit face à la température 119
4.4.4 Moniteurs in-situ et impact de la fenêtre de détection 120
4.4.5 Moniteur in-situ et impact des conditions d'usage 122
4.5 Etude des dispersions à l'aide des moniteurs in-situ 126
4.6 Moniteurs in-situ et adaptation dynamique de la tension 130
4.7 Moniteur in-situ et impact du vieillissement 133
4.8 Conclusion 134
Conclusion et perspectives 137
Bibliographie de l'auteur 139
Bibliographie 140

INTRODUCTION

1. Contexte global de la thèse

Le développement d'applications électroniques portables et communicantes pour le grand public est un des leviers actuels de la croissance de l'industrie de semi-conducteurs, avec une croissance du marché attendue autour de 25% en 2018 [1]: il s'agit non seulement d'ordinateurs portables, mais aussi de téléphones cellulaires et de smart phones, de tablettes internet et e-book, de disques SSD, et de jeux vidéo portables.

Pour ces applications mobiles, la tendance actuelle va vers l'introduction d'une nouvelle génération de processeurs mobiles à hautes performances qui forment le cœur des nouveaux produits, tels que les smart phones et les tablettes. Ces deux produits possèdent de nombreuses similarités en termes d'architecture autorisant les constructeurs à centraliser la conception autour d'un processeur exécutant un système d'exploitation de son choix (Android, MacOS, Windows). Toutefois, il est important de noter que le même processeur pourra se retrouver embarqué dans d'autres produits à l'utilisation très différente, dont la garantie en termes de performance et de fiabilité devra néanmoins être maintenue.

Cette course à la performance et à la faible consommation doit s'accompagner par la prise en compte d'une nouvelle contrainte relative à l'allongement de la durée d'utilisation des batteries et du temps de vie d'un circuit, ce qui implique entre autres l'utilisation d'une tension d'alimentation de plus en plus faible. Cette tendance à baisser la tension d'alimentation va dans le même sens que les techniques de base de l'électronique dite à "basse puissance" (*Low-Power*, [2]) qui impliquent, elles aussi, une baisse de la tension d'alimentation.

De plus, les applications portables grand public ont besoin de grandes quantités de mémoire vive, pour y stocker dynamiquement des logiciels de plus en plus sophistiqués et leurs plages de données, en particulier celles qui sont dédiées à gérer des interfaces de plus en plus complexes (écrans tactiles haute résolution, modem, audio, wi-fi). Ces énormes quantités de mémoire, toutefois, doivent occuper le moins de place possible, à l'intérieur des puces qui, tout en incluant aussi des processeurs, des circuits de traitement numérique et des circuits d'interface avec le reste du monde, doivent avoir la plus petite surface possible, pour être

économiquement rentables. D'où la nécessité pour les mémoires vives d'utiliser des transistors parmi les plus miniaturisés du marché.

Dans l'électronique à basse puissance, les mémoires vives sont basées souvent sur les mémoires statiques à accès direct (*Static Random Acces Memory, SRAM*), qui sont parmi les circuits les plus sensibles à la fois à la miniaturisation, à la réduction de tension [3] mais aussi au vieillissement.

Malheureusement, la miniaturisation extrêmement poussée de tous ces transistors s'accompagne de plus en plus par de phénomènes indissociables à leur fabrication comme l'augmentation de la variabilité du processus de fabrication, liée directement à des barrières économiquement et techniquement infranchissables dans l'amélioration des procédés industriels de fabrication [4]. Ceci est dû aux fluctuations stochastiques des dopants dans le canal des transistors [5] [6], ou encore aux mécanismes intrinsèques de vieillissement des transistors [7].

2. La fiabilité en microélectronique

La fiabilité peut être définie comme « la probabilité qu'un dispositif exécute une certaine fonction dans certaines conditions d'utilisation pendant une période déterminée ». L'étude de la fiabilité intègre donc trois notions fondamentales : le temps, les conditions d'utilisation et le critère de défaillance.
L'intérêt de l'analyse de fiabilité est donc de prédire, d'un point de vue statistique et ce dès l'étape de création du dispositif, si celui-ci sera en mesure de satisfaire les spécifications qui lui sont attribuées. Cette évaluation est basée sur le calcul de la durée de vie du dispositif ou le temps à la défaillance (TTF : Time To Failure). Cette durée représente le temps nécessaire au dispositif pour atteindre un critère de défaillance donné. Au niveau produit, la fiabilité peut être interprétée schématiquement en traçant le taux de défaillance en fonction du temps. Cette courbe en baignoire est représentée par la figure 1 et fait apparaitre trois zones, chacune étant associée à des mécanismes de défaillance différents :

- Zone A : Elle est souvent désignée comme la mortalité infantile. Elle correspond à une perte de fonctionnalité du produit dès les premiers instants de fonctionnement. Dans cette zone, la défaillance est due à des défauts extrinsèques qui sont directement liés aux étapes de fabrication (défauts d'oxyde, de masques ou contaminations).

- Zone B : Le taux de défaillance se stabilise à un niveau globalement constant. Le produit fonctionne normalement dans cette zone, mais la défaillance peut apparaître précocement à cause de la présence de défauts extrinsèques qui entrainent un vieillissement anormal du composant.
- Zone C : Elle correspond aux défaillances induites par un vieillissement normal des transistors. Le taux de défaillance augmente progressivement avec le temps, les performances du produit sont peu à peu dégradées par son vieillissement naturel et la défaillance est ainsi liée à des mécanismes d'usures telles que le Bias Temperature Instability (BTI), le porteur chaud Hot Carrier Injection (HCI) ou l'électromigration. Nous parlons alors de vieillissement (wear-out) du produit. Aussi comme représenté sur ce graphique, cette zone a tendance à apparaitre de plus en plus tôt avec l'évolution des nœuds technologiques.

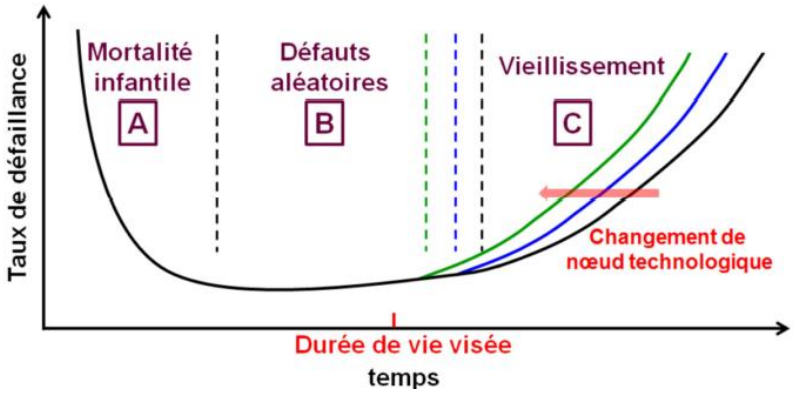


Figure 1 : Evolution temporelle du taux de défaillance d'un produit en fonction du temps. Cette courbe en baignoire fait apparaitre trois régions principales chacune associée à divers mécanismes de défaillance.

Afin de garantir et de prédire le vieillissement d'un transistor avec une certaine précision, il est nécessaire de connaître les mécanismes physiques de dégradation d'un transistor, ainsi que l'impact de cette dégradation sur la perte de fonctionnalité d'un circuit. L'approche traditionnelle de réduction des risques repose principalement sur l'optimisation des procédés de fabrication de manière à obtenir un dispositif qui vieillit le moins possible. Dans les technologies avancées, il apparait désormais nécessaire d'optimiser la fiabilité des produits tout au long de la chaine d'intégration dans une approche ascendante, du défaut au produit (cf. figure 2).

Cette prise en compte verticale de la fiabilité dans la chaine d'intégration du circuit nécessite non seulement une bonne connaissance de l'impact de la fiabilité aux différents niveaux hiérarchiques (défaut, transistor, circuit, système et produit) [9] [10] [11] mais surtout de

développer des passerelles conceptuelles, méthodologiques et pragmatiques entres ces niveaux. Néanmoins, seul le lien entre le transistor et le circuit est majoritairement étudié aujourd'hui par les nombreuses équipes de recherche.

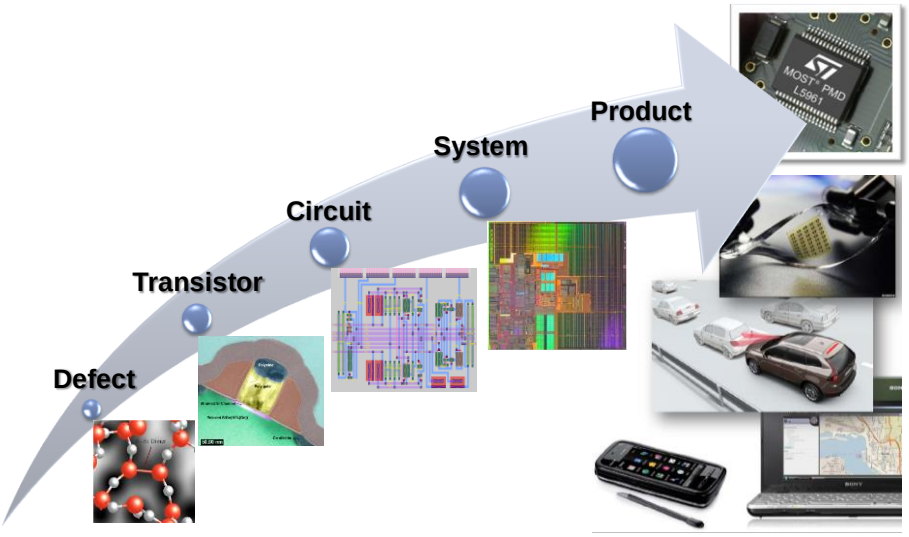


Figure 2 : Les différents niveaux de fiabilité Front End

Dans cette chaine d'intégration, il apparait indispensable de suivre expérimentalement le comportement en fiabilité du transistor dans son environnement réel qui est d'ailleurs très complexe (circuit, système et produit). Pour ce faire, il faut développer des solutions de design dédiées au suivi du comportement de transistors face au vieillissement, communément appelé moniteurs. Les moniteurs de vieillissement sont un sujet très étudié pour les technologies avancés par les organismes de recherche industriels et universitaires, il en existe un très grand nombre : des moniteurs pour suivre la dégradation induite par le HCI pour les applications analogiques [12], des moniteurs embarqués dans les produits capable de suivre la dégradation induite par le NBTI, sous un stress en mode AC ou DC [13], des moniteurs à base de SRAM pour suivre les phénomènes NBTI et PBTI [14], mais aussi récemment le NBTI/HCI en technologie C040 [15] [16] et même le claquage d'oxyde (TDDB) [17]. Afin d'optimiser la performance tout en garantissant un niveau de fiabilité élevé, l'alimentation ou la charge d'un système peut être parfois asservie sur les résultats fournis par les moniteurs.

3. La conception en vue de l'amélioration de la fiabilité

Dans le flot de développement des circuits intégrés, une attention particulière est portée sur l'évaluation de la fiabilité à partir du développement de la technologie, en passant par le transistor MOS jusqu'à la qualification d'un circuit de référence ou le prototype par des tests

spécifiques. Au niveau du transistor MOS, les différents mécanismes de dégradation sont caractérisés, par des lois qui dépendent du temps de fonctionnement, de la tension et/ou de la température. A l'autre bout de la chaîne pour le circuit prototype, le produit est testé et caractérisé dans des conditions accélérées de dégradation, par ses entrées et ses sorties, mais seuls les paramètres macroscopiques, tels que la fréquence, la consommation peuvent être analysés. Ces deux approches classiques chez les fabricants de circuits intégrés souffrent de plusieurs limitations. En effet, pour certains circuits, il est indispensable de pouvoir évaluer in-situ et durant le fonctionnement du circuit, l'impact du vieillissement, afin d'alerter l'utilisateur, ou de prendre une décision corrective afin de garantir sa fonctionnalité. Par ailleurs, certains circuits mixtes analogues-digitaux sont connus pour être particulièrement sensibles à des dégradations diverses dues au vieillissement (I/O, lien série, VCO…). Pour garantir le fonctionnement de ces circuits, la stratégie généralement adoptée est d'agir à l'étape de conception et de validation et de développer des outils de simulation spécifiques pour prédire le vieillissement afin de le rendre fonctionnel face à un comportement temporel défaillant.

Des moniteurs capables de mesurer les chutes de la tension d'alimentation ou la température locale d'un circuit, ont été proposés il y a quelques années [18] [20]. Leur rôle est, outre la détection du dépassement d'un seuil, d'asservir le circuit en fréquence, ou tension afin de revenir dans les paramètres nominaux. Leur intégration dans les circuits permet de réduire les marges temporelles prise lors de la conception du circuit, mais présentent le désavantage d'une latence importante face à un évènement rapide, comme une chute de tension [21]. De plus, à cause des dispersions locales, le temps de calibration est conséquent. D'autres solutions plus prometteuses sont apparues par la suite, telles que la détection en ligne d'erreurs transitoires et des erreurs de délais, tout en prenant en compte l'environnement de fonctionnement, voire le vieillissement [22] [23]. Leur utilisation a permis non seulement de détecter, mais aussi de corriger, les fautes induites par les particules alpha et la radiation cosmique énergétique. Il est tout à fait possible d'associer ces détecteurs d'erreurs avec un système de correction de faute afin de rendre le circuit robuste. Ainsi toutes les variations induites par l'environnement, pouvant générer des fautes de délais, sont détectées et corrigées afin de garder toute la fonctionnalité du circuit [24] [25]. Toutefois, il est indispensable d'introduire non seulement de la surface additionnelle dans le circuit, mais aussi de cycles d'horloge supplémentaires lorsqu'une erreur est détectée de manière à pouvoir ré-exécuter correctement l'instruction. Dans les premières versions de ce type d'approche [26], on observe un surcoût non négligeable de la consommation sur l'arbre de l'horloge, à cause de la

redondance introduite par un élément de stockage supplémentaire, mais il a été reporté également des soucis de métastabilité [26]. Des versions récentes de ces détecteurs d'erreurs ont été proposées et ont démontré, d'un point de vue théorique, l'efficacité de l'approche [27]. Le développement et l'utilisation sur un circuit industriel est une tâche bien plus complexe à mettre en œuvre puisqu'en principe, à chaque fois qu'une erreur est détectée, il faut ré-exécuter l'instruction qui a présenté une erreur. Ceci se fait en modifiant non seulement l'architecture du circuit, mais également en encaissant un surcout de surface et de consommation significatif.

Finalement une dernière approche consiste à détecter non pas des erreurs, mais des pré-erreurs. Ces pré-erreurs sont des alertes avertissant qu'une erreur pourrait se produire si les conditions de fonctionnement ne changent pas, et donc qu'une décision doit être prise en avant de phase. Il n'y a donc pas besoin de circuiterie de correction d'erreur ni de surcout excessif de conception. Généralement, le signal de pré-erreur est utilisé comme un indicateur de délai de chemins de propagation plus long que prévu, ce qui permet au circuit d'adapter ses paramètres d'environnement, d'utilisation ou de vieillissement. En absence de pré-erreur, il est toujours possible de baisser la tension d'alimentation et attendre une consommation minimaliste, tandis qu'en présence des signaux de pré-erreur la tension d'alimentation sera augmentée, ce qui induira un délai plus court, au détriment d'une consommation légèrement plus importante. La stratégie de circuit adaptatif utilisant les sorties de ces moniteurs consiste à prendre une décision (sur la tension ou la fréquence) en fonction du nombre de pré-erreurs tout en prenant en compte le nombre de transitions enregistrés par les moniteurs. Une démonstration sur silicium [31] [32] a montré que ces solutions permettent une économie de plus de 40% d'énergie par rapport à une solution de conception de circuit avec des contraintes de timing pire-cas appliquées à tout le circuit. Reste toutefois une décision importante qui est de bien choisir le placement de ces monitors, ainsi que leur nombre, ce qui est déterminant pour une bonne efficacité énergétique et un faible surcout en surface [33]. Un deuxième problème est la mise en place de la gestion dynamique au niveau architecture de la boucle de contrôle des paramètres du circuit. Ceci est un sujet très largement discuté dans la littérature. Récemment, il a été présenté des stratégies de réallocation adaptative de tâches afin de minimiser le vieillissement provoqué par le NBTI dans les processeurs graphiques [34]. Ces actions système adaptatifs doivent également prendre en compte la sollicitation du circuit (Workload) [35]. De la même manière pour le vieillissement provoqué par le mécanisme HCI, les solutions d'atténuation peuvent être pensées au niveau de la microarchitecture [36] avec une meilleure distribution de l'activité dans un système multicores basés sur des réseaux sur

puce. En effet, nombreuses stratégies innovantes d'adaptation dynamique de la tension du circuit sont proposées pour garantir la fiabilité du circuit, y compris dans un contexte multicore [37] [38] et même pour les mémoires [39].

4. L'objectif de la thèse

L'objectif de la thèse consiste à communiquer l'état du vieillissement des transistors MOS entre les différents niveaux d'abstraction, présentés dans la figure 2. En effet, il est très difficile d'accéder à des informations locales sur la fiabilité d'un transistor MOS si cela n'a pas été pensé préalablement lors de la conception d'un circuit intégré. Pour ce faire, il s'agit de développer des moniteurs dédiés, capables de mesurer précisément les paramètres de vieillissement que l'on cherche à suivre dans l'environnement réel d'utilisation (Workload) d'un circuit numérique. Le plus souvent, on cherche à mesurer un allongement de délai d'un chemin de propagation d'un circuit. Il est possible d'implémenter ces moniteurs à l'intérieur de la logique combinatoire et à chaque niveau de la profondeur de la logique séquentielle. Afin de ne pas impacter le flot de conception, ceci doit être fait de manière automatique lors des étapes de placement et de routage du circuit. Une partie du travail de recherche consiste à identifier les meilleurs moniteurs pour les mécanismes de vieillissement que l'on cherche à suivre (BTI, HCI et TDDB). Ces moniteurs devront être capables de mesurer le mécanisme voulu avec une grande précision tout en minimisant leurs impacts sur le circuit (surface, consommation et délai,…). Il s'agira d'une part de travailler sur l'implémentation de ces moniteurs dans des circuits spécifiques, décider de leurs nombres et de leurs emplacements, qui sont déterminants pour la précision de la décision macroscopique à prendre à l'échelle d'un grand circuit. D'autre part, il faut démontrer la robustesse de ces moniteurs en les testant dans différentes technologies de fabrication, tensions, températures et conditions d'usage. Ces moniteurs fonctionnant sur un principe de pré-erreur, une information est transmise lorsqu'un allongement de retard est détecté à un moment donné avec la résolution que l'on souhaite. Dès lors, plusieurs scénarii sont possibles : une boucle d'asservissement hardware est possible pour adapter la tension V_{DD} ou la fréquence, ou alors une boucle complexe, à la fois hardware et software peut être envisageable. Il est possible de jouer sur la tension du domaine en question ou alors sur les caissons du substrat VDDS et GNDS notamment pour la technologie 28FDSOI.

5. Plan de la thèse

Le premier chapitre introduit rapidement le transistor MOS ainsi que ses évolutions en termes de procédés de fabrication, de la technologie 40nm LP (Low Power) à la technologie 28nm 28FDSOI et les conséquences par rapport à la fiabilité. Dans une seconde partie, les mécanismes de dégradation monotones du transistor MOS sont explicités. Il s'agit des mécanismes de vieillissement BTI et HCI. Nous présentons d'un point de vue global, les méthodes de caractérisation, la compréhension des mécanismes physiques et la modélisation de ces deux modes de dégradation au moment où cette étude avait débuté. Dans une troisième partie, nous allons nous intéresser aux impacts des variations des procédés de fabrication, de la tension et de la température (PVTs) et des mécanismes de vieillissement sur le délai des portes et des chemins logiques.

Dans le deuxième chapitre, nous allons commencer par présenter les méthodologies de prise en compte des variations globales, locales et du vieillissement pendant la phase de conception du circuit. Nous présentons alors l'approche Multi-Mode Multi-Corners (MMMC) et la modélisation de la dégradation des délais des portes logiques. Ensuite, nous allons appliquer ces méthodologies sur les chemins d'un circuit numérique afin de quantifier l'impact de chaque type de variation mentionnée. Dans un second temps, une revue des moniteurs in-situ dans la littérature sera présentée. Nous présenterons également les moniteurs in-situ qui ont été étudiés durant cette thèse ainsi que leurs grandeurs caractéristiques.

Le troisième chapitre sera consacré dans un premier temps à la méthodologie d'insertion des moniteurs in-situ dans le flot de conception de circuits digitaux. Dans un temps second, nous détaillerons différentes approches qui permettent de sélectionner les chemins potentiellement critiques d'un circuit numérique qui seront équipés de moniteurs. Ces approches sont basées sur les dégradations induites par les variations, le vieillissement et les délais des chemins critiques et sous-critiques. Nous présenterons également les résultats de l'application de ces approches à quelques circuits de benchmarks.

Dans le quatrième chapitre, nous aborderons les résultats expérimentaux des circuits avec les moniteurs in-situ. Nous commencerons par présenter les circuits numériques implémentés. Ensuite, nous allons démontrer d'une part la robustesse des moniteurs face à la variation de la tension d'alimentation et de la température et d'autre part leurs capacités à réagir aux conditions d'usages et aux types de variabilités. Aussi, nous présenterons les résultats d'utilisations des moniteurs dans une application de régulation dynamique de la

tension d’alimentation et de recherche des points optimales de fonctionnement. Finalement, nous allons nous intéresser aux moniteurs dans les conditions de vieillissement.

CHAPITRE 1 : DE LA PHYSIQUE DU CMOS AU CIRCUIT NUMERIQUE

1.1 Introduction

Le but du chapitre est de rappeler les mécanismes importants qui influencent la fonction du circuit. Ces mécanismes sont dus aux imperfections de fabrication (les dispersions), au vieillissement et à l'impact des procédés, de la tension et de la température (PVT).

Dans un premier temps, les principaux régimes de fonctionnement d'un transistor MOS conventionnel sont rappelés ainsi que ses équations caractéristiques. Ensuite, les motivations pour le changement technologique majeur d'une architecture sur substrat massif vers une architecture sur substrat SOI sont présentées. Nous définissons certains aspects de la fiabilité du transistor, entre autre les mécanismes de dégradations tels que Bias Temperature Instability (BTI) et l'injection par les porteurs chauds Hot Carrier Injection (HCI) ainsi que leurs impacts sur un oscillateur en anneau. Ensuite, nous abordons les variations locales et globales dues à la fabrication et pour finir nous évaluons les variations des PVTs sur chemin de propagation d'un circuit numérique.

1.2 Transistor MOS

1.2.1 Principe de fonctionnement

Le transistor Métal Oxyde Semi-conducteur (MOS) est le composant de base des circuits intégrés dans le monde de l'électronique digitale. Il est constitué d'un empilement de Métal, de l'oxyde, et du semi-conducteur ainsi que les électrodes appelées grille, source, drain (cf. figure 1.1) et bulk qui permettent d'appliquer les potentiels V_g, V_d, V_s et V_b respectivement.

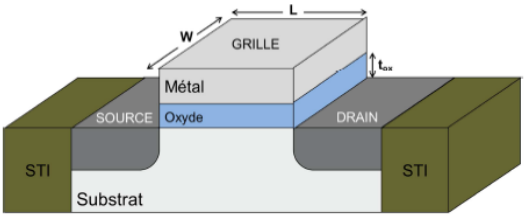


Figure 1.1 : Schéma en coupe du transistor MOS standard. W est largeur du transistor et L sa largeur. t_{ox} est l'épaisseur de l'oxyde de grille. Les tranchées de type STI permettent d'isoler la zone active des transistors adjacents.

Son fonctionnement est semblable à un interrupteur commandé en tension. On distingue deux états pour un NMOS:

- L'état bloqué (cf. figure 1.2(a)) : Lorsque la tension V_{gs} = 0V, le champ vertical est faible, donc les porteurs ne se déplacent pas entre la source et le drain, et par conséquent, aucun courant ne circule entre ces deux électrodes.
- L'état passant (cf. figure 1.2 (b)): Cet état est obtenu en appliquant les tensions V_{gs} > 0 et Vds > 0V. Le champ vertical module la densité des porteurs dans le canal et le champ longitudinal permet aux porteurs de se déplacer entre la source et le drain. Ainsi, un canal d'inversion se crée dans le substrat à proximité de l'oxyde. Les charges présentes dans le canal sont soumises à un l'effet de champ (FET). Lorsque la tension V_d devient plus forte, le canal est pincé : le transistor devient passant et le courant I_{DS} sature en fonction de la tension V_{gs} appliquée. L'entrée du régime saturé dépend de la tension dite de saturation V_{ds}

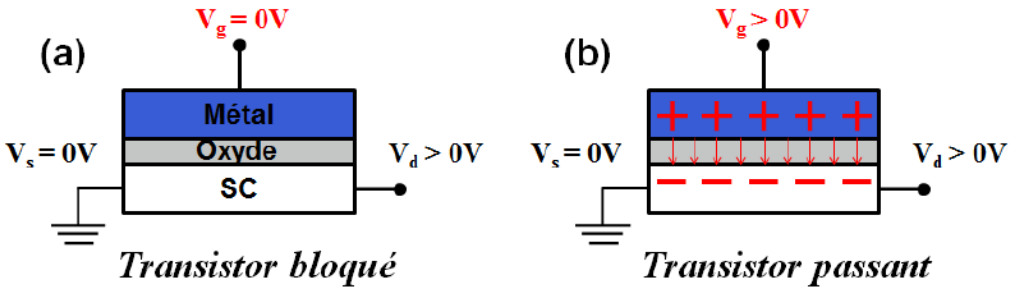


Figure 1.2 : Principe de fonctionnement d'un transistor NMOS : dans l'état bloqué, le transistor ne délivre aucun courant entre la source et le drain (a) et l'application du couple de polarisations (V_g > 0V, V_{ds} > 0V) permet au courant d'électrons de circuler du drain vers la source

1.2.2 Fonctionnement de la capacité MOS

Avant de présenter les régimes de fonctionnement du MOS, il est nécessaire de rappeler le régime de fonctionnement de la capacité MOS. En effet, L'empilement Métal – Oxyde – Semi-conducteur est équivalent a une capacité plane dont les électrodes de grille (G) et de bulk (B), respectivement le métal et le semi-conducteur, sont séparées par un oxyde d'épaisseur t_{ox} qui joue le rôle de matériau diélectrique. Le métal correspond ainsi à la grille du transistor et le semi-conducteur constitue la zone active dans laquelle le canal de conduction se forme sous une polarisation adaptée. La capacité MOS varie en fonction de la valeur du potentiel de surface ΦS de l'interface oxyde/silicium. Cette valeur de Φ_S varie en fonction de la tension V_g appliquée au niveau de la grille. Nous présenterons ce régime de fonctionnement pour un transistor NMOS (substrat de type P, les trous sont les porteurs

majoritaires et les électrons sont les minoritaires). On dénombre trois régimes de la capacité MOS :

- **Le régime d'accumulation $\Phi_S < 0$**

Sous l'application d'une tension $V_g < 0$ et la mise à la masse du substrat, les trous sont attirés vers l'interface oxyde/semi-conducteur et s'accumule. On parle du régime d'accumulation. La capacité résultante *C* est la somme des capacités séries de l'oxyde C_{ox} et d'accumulation C_{acc} :

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{acc}} \qquad \text{EQU 1-1}$$

La capacité d'accumulation est très faible, la capacité totale se réduit à : $C = C_{ox}$:

$$C_{ox} = \frac{\varepsilon_0 \varepsilon_{ox}}{t_{ox}} \quad (F.cm^{-2}) \qquad \text{EQU-1-2}$$

ε_0 représente la permittivité du vide et ε_{ox} et t_{ox} représente respectivement les permittivité relative et l'épaisseur de l'oxyde de grille.

- **Le régime de déplétion $0 < \Phi_S < \Phi_F$**

Lorsqu'on applique un $V_g > 0$, le potentiel de surface devient positif et les trous se déplaçant vers l'interface oxyde/semi-conducteur laisse une zone sans charge appelée zone de déplétion. Cette déplétion s'étend sur une profondeur T_{dep} et la charge associé est proportionnelle à $\sqrt{\Phi S}$.

$$T_{dep} = \sqrt{\frac{2\varepsilon_0 \varepsilon_{SI}.\Phi_s}{q.N_a}} \qquad \text{EQU 1-3}$$

$$Q_{dep} = -q.N_a.T_{dep} = \sqrt{2.q.N_a.\varepsilon_0 \varepsilon_{SI}.\Phi_s} \qquad \text{EQU 1-4}$$

Où q représente la charge élémentaire et ε_{SI} est la permittivité du silicium. Φ_F est le potentiel de fermi. La capacité équivalente dans ce mode de fonctionnement est l'association des capacités d'oxyde C_{ox} et de déplétion C_{dep} :

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{dep}} \qquad \text{EQU 1-5}$$

Cette capacité s'exprime en fonction de Φ_s :

$$C = \frac{C_{ox}}{1 + \sqrt{\frac{2\Phi_s C_{ox}^2}{q.N_a.\varepsilon_0 \varepsilon_{SI}}}} \qquad \text{EQU 1-6}$$

- **Régime d'inversion $\Phi_s > \Phi_F$**

Lorsque le potentiel de surface Φ_s augmente avec le champ vertical, la concentration des trous (P) dans l'interface démunie tandis que celle des électrons (n) augmente exponentiellement en fonction de Φ_s :

$$P = \mathrm{N_a}.\exp(-\frac{q\Phi_s}{kT}) \qquad \text{EQU 1-7}$$

$$n = \frac{n_i^2}{\mathrm{N_a}}.\exp(\frac{q\Phi_s}{kT}) \qquad \text{EQU1-8}$$

Les deux concentrations sont égales lorsque $\Phi_s = \Phi_F$:

$$\Phi_F = \frac{KT}{q}\ln(\frac{N_a}{n_i}) \qquad \text{EQU1-9}$$

Pour $\Phi_s > \Phi_F$ les porteurs minoritaires induisent une variation de charge dans le canal, on parle du régime d'inversion où la charge d'inversion Q_{inv} est proportionnelle à $\exp(\frac{q\Phi_s}{kT})$. On distingue deux sous régimes :

- Le régime à faible inversion pour $\Phi_s < \Phi_F < 2\,\Phi_F$
- Le régime à forte inversion $\Phi_s \geq 2\,\Phi_F$

La transition entre les deux régimes est pilotée par la tension de la grille, aussi appelée la tension de seuil V_{th} qu'on va introduire dans la prochaine section.

1.2.3 Paramètres caractéristiques du transistor MOS

- **La tension de seuil V_{th}**

Le passage de l'état bloqué à l'état passant se fait lorsque la tension de la grille V_g égale à la tension de seuil V_{th}, L'état passant correspond au régime d'inversion de la capacité MOS et est garanti pour. $\Phi_s = 2\Phi_F$. La tension de seuil dans le cas d'un NMOS s'exprime par [40] :

$$V_{th} = V_{th0} + K_b[\sqrt{2\Phi_s - V_b} - \sqrt{2\Phi_F}] \qquad \text{EQU 1.10}$$

V_{th0} représente la tension de seuil obtenu sous une polarisation V_b = 0V. et K_b est une constante qui dépend du dopage et de la capacité d'oxyde :

$$K_b = \frac{\sqrt{2q\varepsilon_0\varepsilon_{SI}.N_a}}{C_{ox}} \qquad \text{EQU 1.11}$$

Cette expression de la tension de seuil est valable pour les transistors ayant une longueur de grille L élevée : les transistors longs. Pour les transistors courts, il est indispensable de prendre en considération les effets secondaires liés à la réduction d'échelle. Ce point sera abordé dans les prochains paragraphes, mais on s'intéresse dans un premier temps à la description des autres paramètres caractéristiques du transistor MOS

- **Paramètres en régime sous seuil : le courant I_{off} et la pente sous seuil**

Le régime sous seuil du transistor MOS correspond au régime à faible inversion de la capacité MOS, i.e. $V_g < V_{th}$. Dans cette configuration la barrière énergétique Φ est très haute et seulement quelques porteurs minoritaires parviennent à la franchir par effet thermique. Ce courant de diffusion évolue exponentiellement par rapport à tension de la grille (cf. figure 1.3) et est donné par l'équation [41] :

$$I_{ds} = \mu_0 \frac{W}{L_{eff}} \left(\frac{KT}{q}\right)^2 d \left(1 - \exp\left(-\frac{qV_d}{KT}\right)\right) \exp(q\frac{V_g - V_{th}}{nKT}) \qquad \text{EQU 1.12}$$

Avec $n = 1 + d = 1 + \frac{K_B}{2\sqrt{2\Phi_F - V_b}}$, d étant le facteur d'effet de substrat et L_{eff} est la longueur effective du canal.

La Figure 1.3 illustre la dépendance en V_g de ce courant I_{ds} dans une échelle logarithmique. Le coefficient directeur de cette droite notée S^{-1} correspond à l'inverse du paramètre appelé pente sous-seuil (S) et qui s'exprime en mV/dec selon l'équation 1.13 [42].

$$S = \frac{\partial V_g}{\partial \ln(I_{ds})} = \frac{KT}{q} \ln(10)(1 + \frac{C_{dep} + C_{it}}{C_{ox}}) \qquad \text{EQU 1.13}$$

Ce paramètre est particulièrement important dans l'évaluation des performances d'un transistor puisqu'il traduit l'augmentation de la tension de grille V_g nécessaire pour augmenter le courant I_{ds} d'une décade. Sa valeur théorique idéale est de 60mV/dec à une température de 300K.

Au regard de cette équation, la valeur de la pente sous le seuil est améliorée lorsque l'épaisseur de l'oxyde t_{ox} est réduite et lorsque les capacités de déplétion et d'états d'interface sont réduites.

Pour un transistor de dimensions W et L, on définit un paramètre électrique I_{th} [43] selon l'équation 1.14 qui correspond à la valeur du courant au seuil, c'est à dire pour laquelle $V_g = V_{th}$

$$I_{th} = 10^{-7} \frac{W}{L_{eff}} \, [A] \qquad \text{EQU 1.14}$$

Ce paramètre est indiqué sur la Figure 1.3 pour W = L et permet d'extraire la valeur de la tension de seuil V_{th}.

Dans le régime sous seuil, il existe un courant de fuite noté I_{off} (qui correspond au courant de drain lorsque $V_g = 0$ et $V_{ds}=V_{dd}$. Ce courant est également indiqué dans la figure 1.3. La

consommation statique dissipée des circuits intégrés dépend de ce courant. Il est donc indispensable de le réduire au minimum pour des applications à base consommation. Ce dernier peut être extrait analytiquement de la caractéristique Log(I_d)=f (V_g), selon l'équation 1.15 :

$$\log(I_{off}) = \log(I_{th}) - \frac{V_{th}}{S} \qquad \text{EQU1.15}$$

Cette relation montre clairement l'intérêt de disposer de dispositifs MOS avec une valeur de pente sous le seuil S la plus faible possible pour réduire le courant de fuite I_{off}.

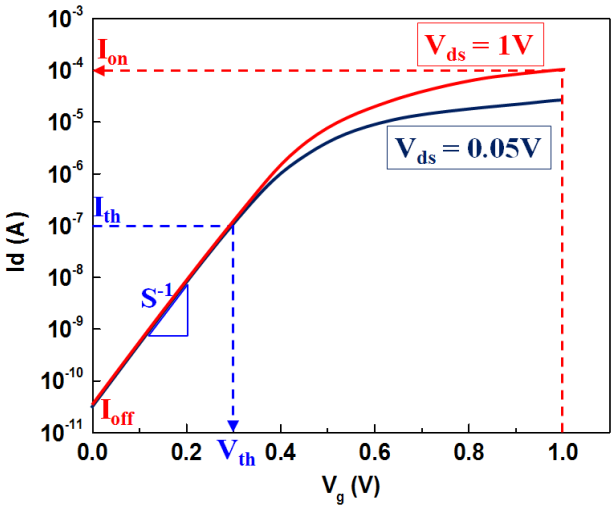


Figure 1.3 : Représentation schématique de la caractéristique Id(Vg), en échelle logarithmique, d'un transistor NMOS long avec W = L en régime linéaire (Vds = 0,05V) et en régime sature (Vds = 1V). Les principaux paramètres caractéristiques Ioff, I_{th}, S^{-1}, I_{on}, V_{th} [53].

- **Régime de conduction**

Le régime de conduction correspond au régime de fort inversion de la capacité MOS et est obtenu lorsque $V_g > V_{th}$. Dans cette configuration, le courant I_{ds} est dû aux porteurs minoritaires qui sont mis en mouvement par le champ électrique longitudinal crée au travers du potentiel V_{ds}. On parle alors de courant de dérive qui se décline en deux régimes de conduction délimités par le paramètre $V_{dsat} = \frac{V_g - V_{th}}{n}$. n ayant été définit par l'équation 1.12.

- Régime ohmique ou linéaire défini pour les $V_{ds} < V_{dsat}$ ou $\Phi_s = 2\Phi_F$.
- Régime de saturation défini pour $V_{ds} > V_{dsat}$

Dans le régime ohmique, on peut exprimer la charge d'inversion en une position *x* du canal suivant la relation 1.16 avec pour condition aux limites au niveau de la source et du drain les équations 1.17 et 1.18 :

$$Q_{inv}(x) = C_{ox}(V_g - V_{th} - \frac{x}{L}V_d) \quad \text{EQU 1.16}$$

$$Q_{inv}^{S}(x = 0) = (V_g - V_{th}) \quad \text{EQU 1.17}$$

$$Q_{inv}^{D}(x = L) = (V_g - V_{th} - V_d) \quad \text{EQU 1.18}$$

Par intégration le long du canal, le courant linéaire du drain s'exprime par l'équation 1.10

$$I_{dlin} = \mu_{eff} C_{ox} \frac{W}{L_{eff}} (V_g - V_{th} - \frac{1}{2} n V_d) V_d \quad \text{EQU1.19}$$

Où μ_{eff} est la mobilité effective des porteurs dans le canal. Ce courant atteint sa valeur maximale lorsque $V_{ds} = V_{dsat}$. A cette tension Q_{inv} dans le drain devient nulle, ce qui se traduit par le pincement du drain. Quand on augmente la tension V_{ds} au-delà de V_{dsat}, le point de pincement se déplace vers la source. Ainsi, la charge ne peut plus augmenter, les porteurs dans le canal atteignent leur vitesse maximale, le courant du drain ne dépend plus de V_{ds} et sature. On parle du régime de saturation. Le courant I_d s'exprime dans ce régime selon l'équation 1.20.

$$I_{dsat} = \frac{1}{2} \mu_{eff} C_{ox} \frac{W}{L_{eff}} \frac{(V_g - V_{th})^2}{n} \quad \text{EQU1.20}$$

On en déduit le courant de saturation I_{on} du transistor MOS pour $V_g = V_d = V_{dd}$:

$$I_{on} = \frac{1}{2} \mu_{eff} C_{ox} \frac{W}{L_{eff}} \frac{(V_{dd} - V_{th})^2}{n} \quad \text{EQU1.21}$$

Au vu de cette équation, on peut remarquer qu'à tension d'alimentation V_{dd} fixée, ce courant I_{on} est d'autant plus élevé que la tension de seuil V_{th} est faible. Un transistor MOS est considéré comme performant lorsqu'il possède un courant de fuite I_{off} faible, et un courant de saturation I_{on} élevé. Ainsi cette tension de seuil apparait comme le paramètre clef pour satisfaire au mieux le compromis I_{on}/I_{off}, en fonction de l'application visée. La réduction du courant est possible par la miniaturisation de la technologie. Cependant, cette miniaturisation s'accompagne des effets indésirables que nous allons décrire dans le prochain paragraphe avant d'aborder le passage de la technologie 40nm sur substrat massif au 28FDSOI.

1.3 Impact de la réduction des dimensions de la grille : effets secondaires

Comme mentionné dans le paragraphe 1.2, les équations décrites précédemment ne sont valables que pour les transistors longs et pour une tension V_{ds} faible (V_{ds} << V_{dsat}). Cette configuration est illustrée sur la Figure 1.4 (a). Lorsque la longueur de grille est réduite, il convient d'intégrer des effets supplémentaires. En effet, à cause des jonctions n/p formées par le substrat avec la source et le drain respectivement, une zone de déplétion également appelée zone de charge d'espace (ZCE) se forme à chaque interface des jonctions et s'étend du côté où le dopage est le plus faible, c'est à dire dans le substrat. Ces ZCE sont délimitées par des lignes pointillées sur la Figure 1.4. Dans ces zones de charge d'espace proches de l'oxyde de grille, le potentiel n'est plus complétement contrôlé par la grille. Pour les transistors courts, ces ZCE peuvent s'étendre jusqu'au centre du canal et abaissent ainsi la barrière du potentiel entre la source et le drain (Figure 1.4 (b)) et donc la tension de seuil mesurée à faible tension V_{ds}. Ce phénomène est appelé effet de canal court (SCE pour Short Channel Effect). Par ailleurs, sous l'application d'une polarisation V_{ds} élevée, la zone de charge d'espace côté drain s'étend (Figure 1.4 (c)) et contribue à abaisser d'avantage cette barrière de potentiel. Ce phénomène est intitulé DIBL (Drain Induced Barrier Lowering).

Les valeurs des potentiels de SCE et DIBL sont données par les équations 1.22 et 1.23 [44] :

$$SCE \approx 0.64 \frac{\varepsilon_{Si}}{\varepsilon_{ox}} EI.\phi_d \qquad \text{EQU 1.22}$$

$$DIBL \approx 0.8 \frac{\varepsilon_{Si}}{\varepsilon_{ox}} EI.V_{ds} \qquad \text{EQU 1.23}$$

Le paramètre EI est appelé facteur d'intégrité électrostatique. Il dépend de la géométrie du MOS et traduit l'influence que le drain a sur le canal de conduction. ϕ_d représente la tension de la diode, formé par la jonction source-canal et canal-drain.

Dans ces conditions, la tension de seuil V_{th} s'écrit

$$V_{th} = V_{th,long} - SCE - DIBL \qquad \text{EQU 1.24}$$

$V_{th,long}$ est la tension du transistor long.

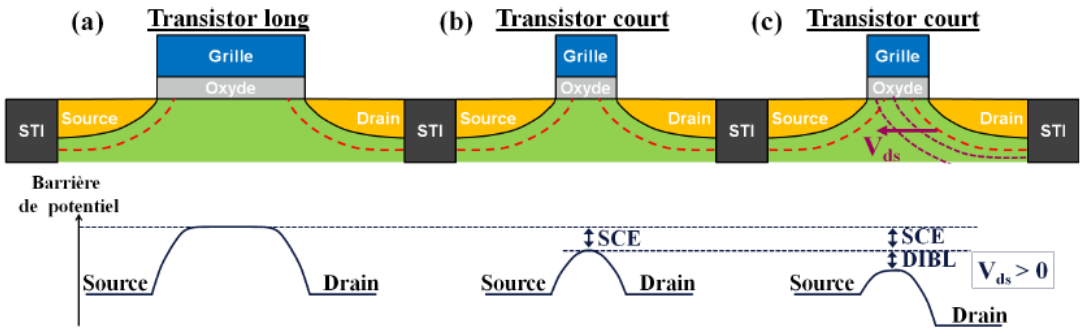


Figure 1.4 : Illustration de l'extension de la zone de charge d'espace cotés source et drain pour un transistor à canal long (a), un transistor court avec une polarisation V_{ds} négligeable

(b) et un transistor court avec une polarisation V_{ds} plus élevée (c). Les lignes pointillées indiquent l'extension de la zone de charge d'espace. En dessous est illustrée l'abaissement de barrière de potentiel correspondante entre la source et le drain et par conséquent de la tension de seuil due à la réduction de la longueur de grille (b) : effet SCE (c) : effet DIBL

1.4 Evolution technologique du 40nm vers le 28nm sur substrat massif

La réduction de la taille des transistors a pour but d'optimiser les performances électriques de celui-ci, c'est-à-dire l'augmentation du courant I_{on} et la réduction du courant de fuite ainsi que les effets parasites SCE et DIBL. Ces performances sont liées à la capacité de l'oxyde de grille C_{ox} qui dépend de la propriété de l'isolant, et aux dimensions de la grille : longueur L_G et la largeur W_G. (cf. équation 1.25). En effet, Plus l'épaisseur du diélectrique est fine plus la capacité C_{ox} est grande et plus faible sera le courant de fuite. Meilleur sera l'isolant du point de vue diélectrique plus fort sera le courant de drain qui pilote le fonctionnement du transistor

$$C_{ox} = \frac{\varepsilon_0 \varepsilon_{ox}}{t_{ox}} W_G . L_G \qquad \text{EQU 1.25}$$

Où ε_0 et ε_{ox} sont respectivement les permittivités relatives du vide et de l'oxyde, t_{ox} son épaisseur.

Cependant, pour des épaisseurs d'oxyde suffisamment faibles (typiquement tox <3nm), un courant dit de fuite noté I_g apparait entre la grille et le substrat. Ce dernier peut même devenir du même ordre de grandeur que le courant I_{off} ce qui induit, au niveau circuit, une augmentation de la consommation statique. Une illustration de cette augmentation drastique du courant de fuite avec la réduction de l'épaisseur t_{ox} de l'oxyde de grille en SiO_2 est donnée sur la Figure 1.5 [45].

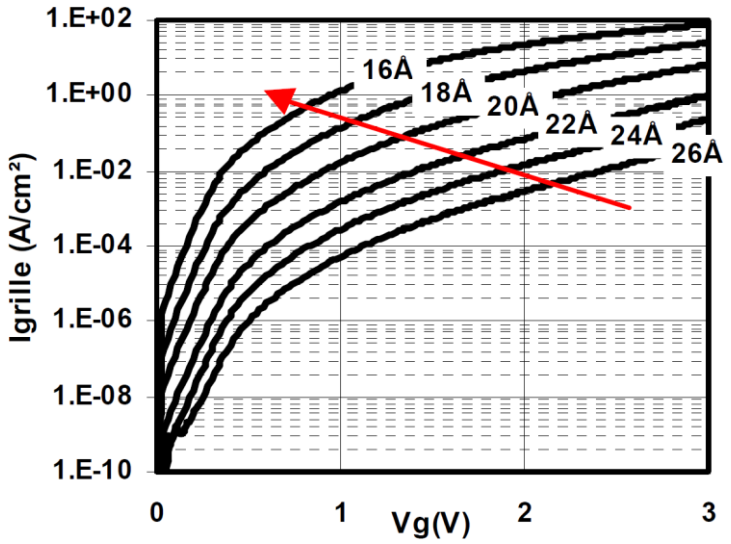


Figure 1.5 : Caractéristiques Ig(Vg) pour des épaisseurs physiques d'oxydes SiO_2 allant de 16Å à 26Å. La réduction de cette épaisseur tox résulte en une augmentation très significative du courant de grille qui peut se répercuter directement sur le courant de fuite Ioff [45].

Pour les technologies 40nm sur substrat massif et antérieurs, la grille est constituée de polysilicium. Ce choix a été principalement fondé sur sa valeur de travail de sortie ϕ_M qui permet d'obtenir une faible tension de seuil V_{th}. Cependant, lorsque le transistor fonctionne en régime d'inversion, la charge ainsi créée dans le canal est responsable de la formation d'une charge de déplétion dans la grille au niveau de l'interface avec l'oxyde de grille. Ce phénomène s'intitule la polydéplétion de la grille. Cette charge de déplétion s'étend sur quelques Angströms et induit alors un effet capacitif parasite qui devient non négligeable lorsque l'épaisseur de l'oxyde diminue. Cela se traduit par une augmentation effective de l'épaisseur de l'oxyde de grille t_{ox} et donc une réduction des performances électriques (réduction I_{on} voir équation 1.21). Une solution efficace pour réduire ce phénomène de déplétion de la grille consiste en l'utilisation de grilles métalliques et c'est pour cette raison que la grille en TiN a été introduite à STMicroelectronics dès le nœud 28 sur substrat massif en remplacement de la grille en polysilicum. En effet, le matériau TiN est un matériau de type 'mid-gap', c'est à dire choisi pour son travail de sortie qui place le niveau de Fermi a la moitié du gap du silicium. C'est donc un matériau spécifiquement choisi pour faciliter l'obtention des tensions de seuils sensiblement symétriques pour les transistors NMOS et PMOS [46].

Le passage de technologie 40nm au 28nm sur substrat massif (cf. figure 1.6) s'est caractérisé par l'utilisation des matériaux à forte permittivité High-K conjointement à la grille métallique TiN afin de palier au problème du courant de gille I_g et donc du courant de fuite I_{off}. La particularité des matériaux High-K est la constante diélectrique plus élevé (de 25 à 30) que

celle du polysilicium (3.9). le matériau High-K devait satisfaire un certain nombre de conditions, notamment :

- La stabilité vis-à-vis des différentes étapes de fabrication des transistors (en particulier durant les étapes de recuit)
- La compatibilité avec les aspects industriels (temps de dépôts, cout, budget thermique)
- Une faible quantité de défauts intrinsèques et un bon contrôle de son épaisseur déposé.
- La compatibilité avec les exigences en termes de fiabilité.

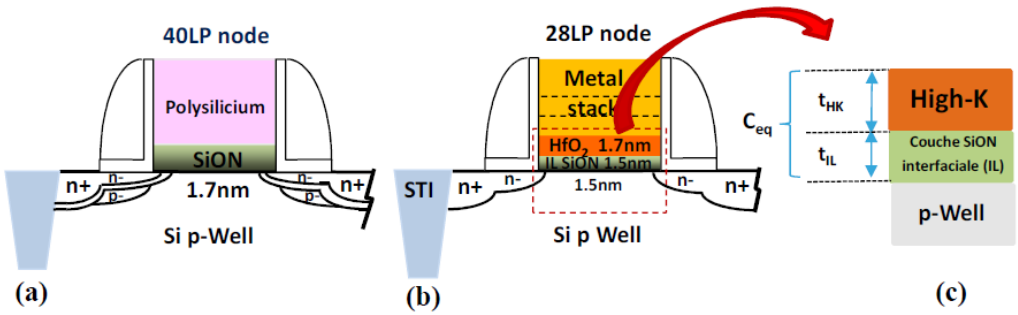


Figure 1.6: Evolution de la structure MOS entre (a) la technologie low power (40LP) à grille en poly Si sur silicium p-well (b) la structure 28LP avec une grille métallique, le diélectrique High-K+IL (c) Zoom sur l'empilement des différentes couches utilisé pour le calcul de l'EOT (Equivalent Oxide Thickness).

Pour quantifier le bénéfice apporté par le replacement du polysiliclium SiO_2 par le matériau High-K, on définit le paramètre EOT (Equivalent Oxide Thickness) qui correspond à l'épaisseur de SiO_2 qu'il serait nécessaire pour obtenir un couplage équivalent au système bicouche SiO_2/High-K (cf. figure 1.6(c)). Ce paramètre EOT s'écrit selon la relation 1.26 en considérant que l'IL est purement en SiO_2.

$$EOT = t_{IL}\frac{\varepsilon_{SiO_2}}{\varepsilon_{HK}}t_{HK} \qquad \text{EQU 1.26}$$

Où t_{IL} et t_{HK} sont les épaisseurs respectives de la couche interfaciale et du High-K.

1.5 Evolution technologique du 28nm sur substrat massif vers le 28nm sur substrat SOI

1.5.1 Avantages de la technologie sur substrat SOI

La technologie de type Silicon On Insulator (SOI) a été développée dans les années 60 et 70, pour répondre à des contraintes liées aux applications spatiales, en particulier l'exposition aux

effets de radiations naturelles et ionisantes [47]. Son principe repose sur l'ajout d'une couche d'oxyde (BOX) entre la zone active, dite Ultra-Thin-Box (UTB) et le substrat, comme illustré sur la figure1.7.

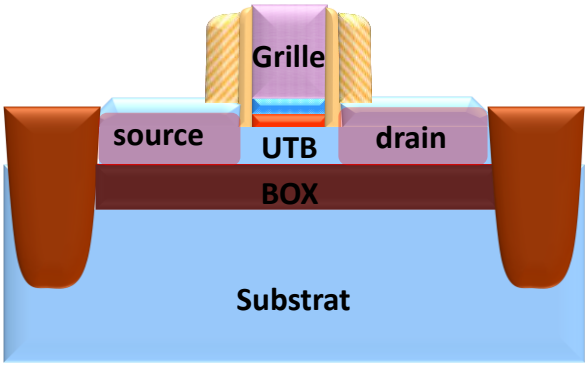


Figure 1.7: Architecture du transistor Fully-Depleted-SOI (FDSOI)

Cette technologie présente plusieurs avantages par rapport à la technologie sur substrat massif, le premier est la présence du l'oxyde épais enterré BOX (cf. figure 1.7) qui entraine une réduction de la surface de jonction et l'élimination des capacités parasites drain/substrat et source/substrat. Par conséquent, les performances électriques en régime statique sont améliorées ce qui permet au transistor le fonctionnement à très haute fréquence et dans une gamme de température très large.

Le deuxième avantage est lié à l'utilisation de cette structure en double grille commandée. En effet, la tension du substrat bulk V_{BS} est commandée par une autre électrode de commande Back-Plane (BP), qui est une couche fortement dopé (N ou P selon le type du transistor) en dessous du BOX, ce qui implique une polarisation opposée dans le régime d'inversion vis-à-vis de la couche active de silicium intrinsèque UTB. Ceci offre des meilleures performances en courant et une excellente réduction des dimensions sans les effets du canal court. Ce meilleur contrôle électrostatique du FDSOI vient du fait que la charge dans la zone active ne varie plus avec la tension de grille mais à travers le couplage capacitif entre les deux grilles. Il en résulte que le potentiel de la face avant FG devient couplé au potentiel de la face arrière BG, rendant les caractéristiques électriques du canal dépendantes de la polarisation de la face opposée.

L'épaisseur de la couche UTB joue un rôle important dans les performances du transistor. Lorsque celle-ci est fine (6nm à 10nm), la zone de déplétion couvre totalement le film de silicium, on parle d'une architecture SOI Fully Depleted (FDSOI). Lorsque la couche UTB est plus importante, la zone de déplétion ne s'étend pas sur toute la profondeur du silicium, on parle alors d'une architecture SOI Partially Depleted (PDSOI). Le PDSOI présente des inconvénients tels que le substrat flottant, ce qui rend le transistor plus vulnérable aux effets

de latch-up (haute impédance entre le V_{DD} et le Gnd) et kink (effet mémoire lié à la tension de seuil) [48].

En conclusion, cette architecture améliore l'isolation entre les dispositifs et permet de réduire de façon significative les capacités de source et de drain mais également d'éliminer la majeure partie des effets parasites apparus dans la technologie sur substrat massif [49]. Nous pouvons résumer les avantages de la technologie SOI par les points suivants :

- Meilleure pente sous-seuil et donc une commutation ON/OFF plus rapide
- Effets DIBL et SCE réduits par rapport aux technologies sur substrat massif à cause de la couche UTB qui assure la conduction des porteurs entre le drain et la source très proche de la grille.
- Faible résistance d'accès source/drain, faible réduction des capacités de couplage et une bonne immunité face aux effets de latch-up et kink .
- Consommation statique plus faible, bénéfique pour les applications des circuits à basse tension V_{DD}
- Meilleur comportement du canal court malgré la réduction poussée des dimensions du canal L_G et W_G
- Performance en courant plus élevée de 20 à 30% par rapport au 40nm sur substrat massif pour une même tension V_{DD}
- Meilleur fonctionnement à haute température (125°-300°C)
- Plus résistant aux irradiations naturelles et ionisantes pour les applications spatiales

1.5.2 Tension de seuil du transistor FDSOI

Un des avantages de la technologie FDSOI est son application en multi-V_{th} possible suivant les matériaux employés entre la face avant et la face arrière pilotée par le back plane (BP). En effet, les potentiels développés en face avant FG et en face arrière BG sont couplés de telle sorte que les caractéristiques du canal en surface, commandé en premier lieu par la tension $V_{G,FG}$, varie avec la polarisation appliquée sur la grille face arrière $V_{G,BG}$. Ceci traduit la propriété d'utilisation du FDSOI en double grille (2G) où la tension substrat (V_{BS}) considérée pour les structures standards sur substrat massif, devient une tension de polarisation de commande supplémentaire face arrière $V_{G,BG} = V_{BS}$ qui peut polariser la face avant en mode d'inversion, de déplétion et d'accumulation suivant le signe et l'amplitude de $V_{G,BG}$ [50]. Il en résulte que suivant le signe de la tension $V_{G,BG}$ qui polarisera en régime d'accumulation ou d'inversion, le canal de conduction à travers l'UTB aura une profondeur verticale qui dépendra du couplage entre les deux grilles de commande [52][53]. Ceci va modifier les

caractéristiques habituelles du transistor MOS classique en introduisant le terme de couplage capacitif *C* qui est fortement dépendant de la topologie de la structure et des polarisations utilisées.

L'application de V_{BS} (cf. figure 1.8 (a)) lors du fonctionnent des transistors FDSOI module la tension de seuil à travers le couplage entre le FG et BG comme l'effet substrat dans le MOS classique. Sur la figure 1.8 (b-c), ceci peut être observé par la variation des tensions de seuil V_{th} pour le NMOS et le PMOS (VTN, VTP) avec la tension V_{BS} entre la polarisation inverse (RBB) et directe (FBB) pour chaque transistor. En effet, les transistors NMOS ont la couche BP non polarisé avec V_{BS}=0 tandis que pour l'utilisation haute performance (LVT), les PMOS sont à V_{BS}= -V_{DD} (FBB). L'utilisation des modes FBB et RBB permet de garantir et rétablir une symétrie dans les valeurs de V_{th} comme observé par les caractéristiques $I_D(V_G)$ en mode de saturation qui traduit la commutation entre l'état Off et l'état On.

Ainsi, nous pouvons obtenir trois types de tension de seuil V_{th}:

- les transistors LVT (Low Vth) à faible V_{th} pour les applications hautes performances.
- les transistors RVT (Regular Vth) ou SVT (Standard V_{th}) à V_{th} standard.
- les transistors HVT (High Vth) à fort V_{th} pour les applications basses consommations.

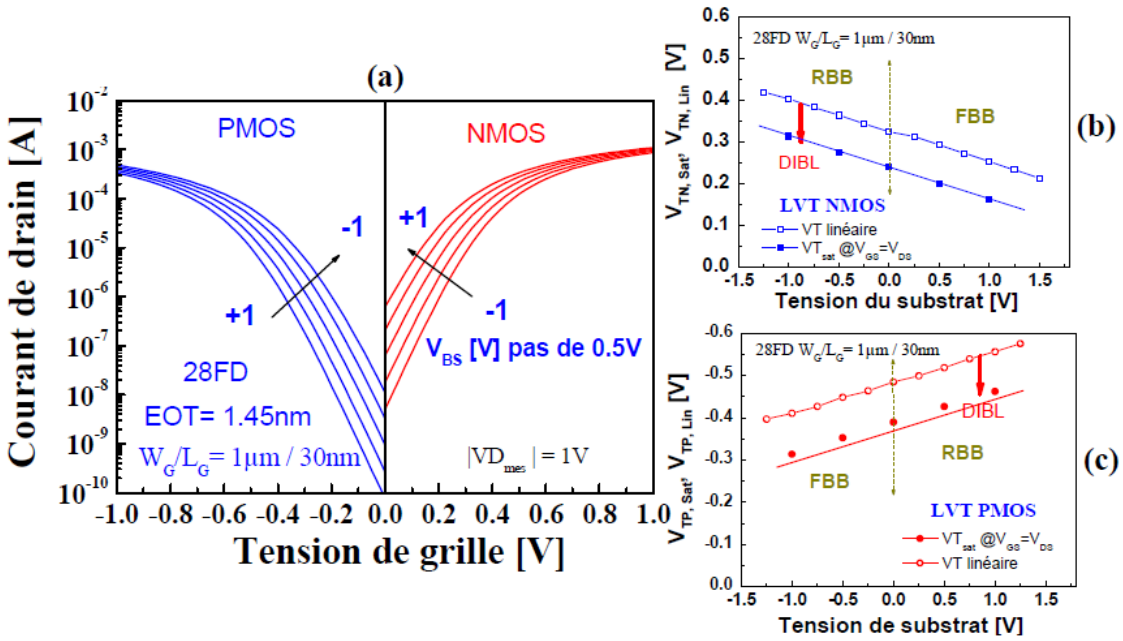


Figure 1.8 : (a) Courant sous-seuil en mode de saturation (V_{DS}= ±1V) des transistors NMOS et PMOS FDSOI à BP de type P polarisés en fonction de V_{BS} entre le mode RBB (reverse) et FBB (forward). Evolutions des tensions de seuils en mode linéaire et saturé laissant apparaitre l'effet DIBL (b) pour les transistors NMOS 28FD LVT (c) pour les transistors PMOS 28FD LVT [54].

Du point de vue procédé de fabrication, la variabilité importante dans la filière 40nm jusqu'au 28nm sur substrat massif est fortement réduite dans le 28FDSOI en raison de l'absence de dopage de la couche UTB nécessaire pour la réalisation des poches (ou halos). En effet, le dopage de la zone active (UTB) a pour but l'ajustement de la tension de seuil dans les filières CMOS sur substrat massif, ce qui induit plusieurs valeurs de V_{th}. Donc, l'absence de dopage de la couche UTB dans la technologie FDSOI élimine tous les problèmes majeurs de fluctuation statistique et aléatoire de la tension de seuil V_{th} à dimension fixée.

En conclusion, la technologie 28FDSOI montre des performances similaires que la technologie 20nm sur substrat massif, pour un coût de fabrication bien plus faible, fournissant une meilleure efficacité, que la technologie 28nm bulk, en performance et en consommation. Alors que les transistors PDSOI sont couramment employés pour l'utilisation dans certains microprocesseurs à haute performance, les transistors FDSOI ont un net avantage pour une meilleure intégration plus poussée en dimensions. Le fait que le transistor fonctionne en déplétion complète (la zone de charge d'espace couvre tout l'UTB), la charge d'espace du transistor FDSOI montre une faible variation avec la tension de grille, ce qui renforce le contrôle de la grille sur la charge d'inversion [42].

1.6 Mécanismes de dégradation du transistor MOS

1.6.1 Mécanisme de dégradation BTI

La dégradation de type Negative Bias Temperature Instability (NBTI) est provoquée par l'application d'une tension négative V_{GS} à haute température [55] (cf. figure 1.9). Elle concerne les transistors PMOS de façon beaucoup plus marquée que les transistors NMOS [56] [57]. Cette dégradation a pour conséquence la contamination dans les oxydes de grille sous l'effet de l'application d'un champ électrique [58] [59]. En effet, lorsque la tension $V_{GS} < 0$ est appliquée, les trous de la couche d'inversion se trouvent piégés à proximité de l'interface oxyde-canal et interagissent alors avec les atomes à l'interface. Une fraction d'entre eux est injectée vers la grille sous l'action du champ vertical ce qui crée des défauts dans l'oxyde. L'apparition des défauts induit une dérive de tous les paramètres électriques du transistor. Il faut noter aussi que la dégradation par NBTI est un mécanisme de dégradation statique, puisqu'il intervient en absence d'un courant de porteurs à travers le canal (V_{DS} = 0V).

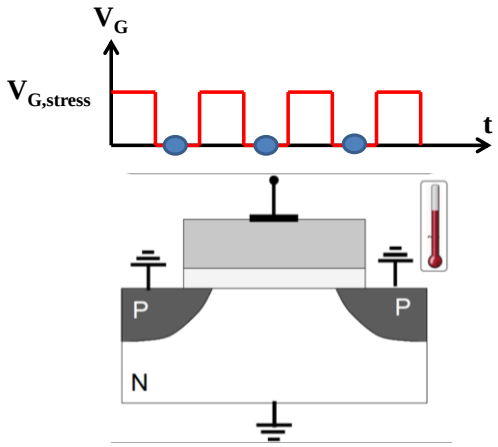


Figure 1.9: Configuration de contrainte NBTI à une tension $V_{G,stress}$ et à haute température, a contrainte étant interrompue pendant les phases de caractérisation électrique des paramètres du MOS (ronds bleu dans le chronogramme).

La conséquence de la dégradation NBTI dans le transistor PMOS est un décalage important de la tension de seuil vers les tensions V_{GS} négatives, ce qui impacte la caractéristique $I_D(V_G)$ par un décalage quasi parallèle de la pente sous-seuil et au-dessus du seuil de cette caractéristique. La particularité de cette dégradation est qu'elle augmente très fortement à haute température comme nous le constatons dans la figure 1.10. Cette accélération de la dégradation avec la température peut se quantifier facilement grâce à la loi d'Arrhenius, donné par l'équation 1.27

$$\Delta V_{th} = A.\exp\left(\frac{E_a}{K_B.T}\right).V_{Gstress}^{b}.{t_{stress}}^{n} \qquad \text{EQU 1.27}$$

T est la température (°K) et A une constante dépendante du procédé (nature de l'oxyde et t_{ox}). E_a est l'énergie d'activation. b et n sont des paramètres de fit. V_G et t_{stress} représente la tension et le temps de stress respectivement, n est le facteur d'accélération.

L'énergie d'activation dépend de la nature du diélectrique et de l'épaisseur d'oxyde. Pour les oxydes épais, nous trouvons des intervalles de l'énergie d'activation E_a entre 0.6eV a 0.35eV [59, 60], puis dans les oxydes minces a ultra minces entre 0.35 et 0.1eV [61]. K_B = 8.617 10^{-5} eV/K

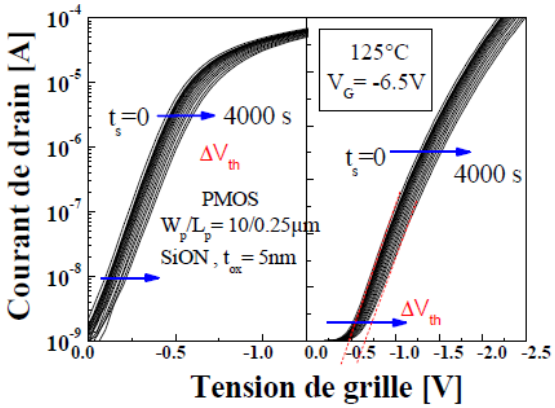


Figure 1.10 : Caractéristiques électriques $I_D(V_G)$ lors de l'application d'une contrainte NBTI à une tension $V_{Gstress}$ à haute température. La mesure du courant linéaire montre un décalage parallèle due à la dégradation ΔVth [54]

La dégradation du transistor PMOS soumis à une contrainte NBTI, et observée par la réduction du courant et l'augmentation de sa tension de seuil en valeur absolue (cf. figure 1.10), résulte de la génération de défauts uniformément sur toute la surface active du transistor, à l'interface oxyde /substrat et dans le diélectrique. Les principaux défauts sont :

- Des défauts qui créent des charges à l'interface Si/IL, ils ont une densité d'état d'interface N_{it} de type donneur. Ces défauts sont chargés positivement dans le gap du semi-conducteur en fonction de $(\phi_c - \phi_F)$ [60]
- Les charges positives piégées dans le volume de l'oxyde N_{ox} (cm^{-2}) [61]
- Les charges piégées dans l'oxyde mais à proximité de l'interface oxyde / substrat qui peuvent se dépieger quand le champ diminue et s'annule : ils sont nommés états lents N_{ss} ou *border traps*

Les deux premiers types de défauts sont des défauts permanents, c'est-à-dire que la réduction de la tension montre leur effet persistant, tandis que N_{SS} est qualifié en défauts recouvrables car ces derniers conduisent aux phénomènes de relaxation [60]. La relaxation est plus marquée dans les oxydes de grille minces à ultra minces, pour $t_{ox} < 3.5$nm, c'est-à-dire quand la composante d'effet tunnel direct devient importante. Cet aspect a donné lieu à la mise au point de mesures rapides à la volée (OTF) pour pouvoir capturer la dégradation permanente avant relaxation [60] [55].

Un phénomène similaire existe pour les transistors NMOS: dans ce cas, ce sont les électrons du canal injectes dans l'oxyde de grille sous l'application d'une tension de grille positive qui remplissent des défauts permanents. Il s'agit du mécanisme PBTI (Positive Bias Temperature Instability). La particularité de la dégradation PBTI est qu'elle se manifeste avec des niveaux de dégradation qui, extrapolés aux conditions nominales de fonctionnement, sont bien plus

faibles que dans le cas du NBTI et ce, même pour les nœuds technologiques les plus avancées [62] [63] [64].

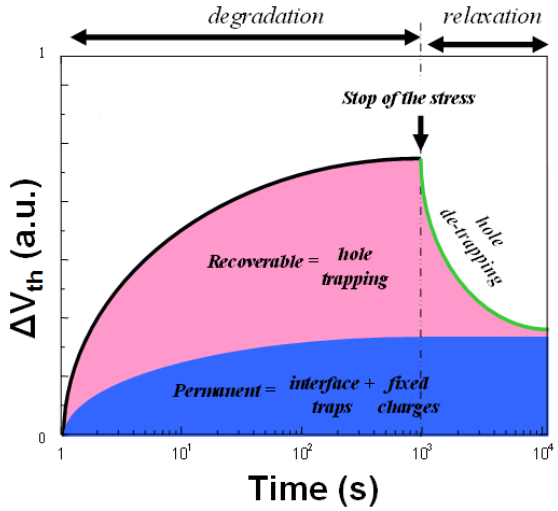


Figure 1.11: Les deux composantes de dégradation NBTI sont illustrées : la partie permanente de la dégradation est due à la création d'états d'interfaces N_{IT} et de charges fixes dans l'oxyde N_{ox+}, la partie réversible est associée à un mécanisme de piégeage/ dépiégeage de trous [61].

1.6.2 Modélisation de la dégradation par NBTI

Le mécanisme de dégradation BTI est décrit analytiquement par divers modèles physiques, notamment ceux de Reaction Diffusion [65] [66] et de Reaction Limitée [61]. Ces derniers reposent sur le même fondement selon lequel la dégradation NBTI est associée à un mécanisme de rupture de la liaison Si – H, qui induit le transfert de l'atome d'hydrogène H de son état fondamental vers son état de transport [67]. Ce changement d'état est assisté à la fois par le champ vertical dans l'oxyde de grille et par les trous injectés du canal vers la grille. Les ruptures des liaisons Si – H donnent alors naissance à des états d'interface conjointement à une charge positive dans l'oxyde [55] ainsi qu'à un mécanisme de piégeage/dépiégeage impliqué comme une autre facette du même mécanisme suivant le niveau d'énergie atteint, entre la configuration permanente et recouvrable par capture et émission de trous pouvant commuter entre ces deux états, qualifiant ainsi ces pièges comme des *switching traps* [69].

La description complète de ce mécanisme de dégradation donne encore lieu à de nombreux débats [70] [71] [72] mais la communauté scientifique commence à s'accorder sur l'existence de ces deux composantes indépendantes qui sont illustrées schématiquement sur la figure 1.11. En pratique, si nous appliquons une contrainte DC sur la grille d'un transistor pendant une durée donnée (1000a.u. sur la figure 1.11) et que nous interrompons brusquement cette contrainte, tout en suivant l'évolution de la dégradation ΔV_{th}, nous constatons qu'une partie

de la dégradation disparait progressivement au cours de la phase de relaxation, c'est la partie recouvrable**.**

1.6.3 Mécanisme de dégradation par porteurs chauds HCI

La dégradation par porteurs chauds (HC) est historiquement associée à l'action de porteurs énergétiques du canal qui, sous la polarisation V_{DS} élevée, induisent un mécanisme d'ionisation par impact à l'intérieur de la zone de charges en mode de saturation, générant des paires électrons-trous localisées à proximité du drain comme illustré dans la figure 1.12. Les électrons énergétiques du transistor NMOS peuvent être alors injectés dans l'oxyde de grille en fonction de la distribution du champ électrique latéral et vertical dans cette zone (IG_{inj}), alors que les trous sont collectés par le courant du substrat (I_{Sub}). La principale différence entre les dégradations HCI et NBTI consiste à appliquer une tension de drain en plus de la tension de grille, ce qui rend le mécanisme fortement dépendant du courant dans le canal. La génération de porteurs chauds est une conséquence de la réduction de la longueur effective ($I_{DS} \approx 1/L_{eff}$) qui est due à l'augmentation du champ effectif dans le canal pour les porteurs incidents.

Les premières conséquences, dépendant de V_{GS} (V_{DS}), se traduisent par la génération d'états d'interface et de charges piégées dans l'oxyde de grille jusque dans la zone de recouvrement grille-drain, qui dégradent alors les paramètres caractéristiques du transistor MOS:

- l'augmentation de la tension de seuil V_{th}
- Réduction des courants I_{Dlin}, I_{DSat}, due à la réduction de la mobilité $\Delta\mu_{eff}$ et de ΔV_{th} [74]
- Réduction correspondante de la transconductance Gm, de la conductance Gd et de la pente sous seuil SS
- Augmentation de la résistance série au drain R_{GD}

Cette création d'états d'interface N_{it} est associée à la rupture des liaisons Si – H à l'interface Si/SiO_2. Les paramètres couramment utilisés pour l'étude sont les courants en régime linéaire I_{Dlin} et en régime de saturation I_{Dsat}. Le premier courant est très sensible aux défauts générés pendant la contrainte, et le second en raison de son lien direct avec la commutation des transistors lors du fonctionnement réel dans les cellules [75], alors que le décalage de V_{th} est généralement moins utilisé, à l'inverse de la dégradation NBTI, car le V_{th} reste moins affecté par la dégradation due à la localisation de la région des défauts au drain.

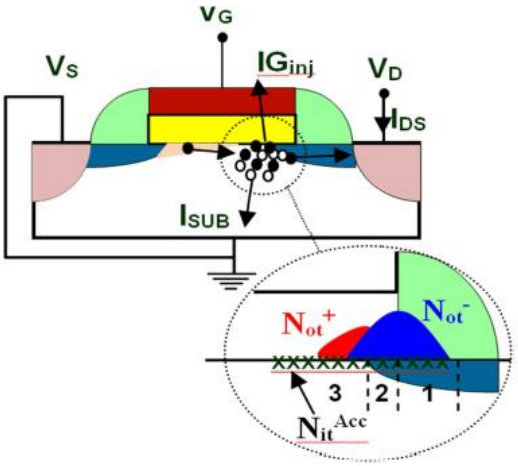


Figure 1.12 : Configuration électrique pour appliquer une contrainte par injection de porteurs chauds (HC) avec les tensions V_{GS} et V_{DS}, à 25°C, ce qui Illustre la création d'états d'interface Nit et de charges piégées Nox+, Nox- dans l'oxyde de grille en fonction de VGS. Le mécanisme HC conduit à l'existence du courant injecté (IG_{inj}) et du courant substrat (ISub) issu de l'ionisation par impact [73].

Traditionnellement, les effets de porteurs chauds étaient expliqués par le modèle des "électrons chanceux" [76], où la dégradation était accélérée par le champ électrique latéral, ce qui rendait la modélisation valide pour les technologies dont les tensions d'alimentations étaient moyennes à élevées ($V_{DD} \geq 3.3V$). Pourtant ce modèle s'est révélé inadapté pour expliquer les effets porteurs chauds dans les nœuds technologiques avancés avec des tensions d'alimentations plus basses $V_{DD} < 3V$ [77], c'est-à-dire dans des domaines d'énergies inférieurs à l'énergie de rupture ou de dissociation des liaisons atomiques. Plus récemment, de nouvelles théories ont été introduites [78] [79] [80], d'une part pour dépasser les limitations du modèle de l'électron chanceux et d'autre part pour expliquer le mécanisme de dégradation HCI dans les transistors MOS à canaux courts, alimentés à basse tension (V_{DD}= 1V à 1.2V). En effet, le formalisme a été transféré à une modélisation en énergie à l'aide de fonctions de distributions, puis à tenir compte des mécanismes d'acquisitions de l'énergie par les interactions des porteurs entre eux (électron-électron) dans le domaine des moyennes énergies [78] [79] [80], puis récemment, grâce aux interactions avec les phonons, où le nombre de porteurs du canal à l'aide des modes de vibrations des atomes, qui mettent la liaison Si-H en résonance (MVE) jusqu'à la désorption de l'atome d'hydrogène.

L'impact des mécanismes de dégradation monotone BTI et HCI sur le délai des portes logiques sera abordé dans le prochain paragraphe.

1.7 Impact des mécanismes de dégradation sur le délai des portes logiques

1.7.1 Effet AC et DC sur un chemin de propagation à haute température

Les résultats que nous allons présenter dans ce paragraphe résultent d'un travail commun réalisé avec [81]. L'objectif recherché est de mettre en évidence l'impact des mécanismes de dégradation BTI et HCI sur le délai des portes logiques. Plus particulièrement, nous allons appliquer des signaux AC et DC sur une structure de chaine de porte en 28FDSOI, et nous allons évaluer la dégradation en mesurant la fréquence d'oscillation. L'étude est menée à deux températures : 25°C et 125°C. Il faut noter aussi que nous ne nous intéressons qu'à la partie permanente du BTI. Entre chaque contrainte et chaque mesure, nous imposons un temps de relaxation pour nous affranchir de la partie recouvrable de la dégradation BTI [61]. De plus, nous préférons tracer l'évolution de la dérive du délai de propagation de chaque porte $\Delta\tau$, plutôt que la fréquence d'oscillation f : ces deux paramètres sont reliés par l'expression : $f = \frac{1}{N*\Delta\tau}$ N étant le nombre d'étage.

Le schéma de la structure utilisée est donné dans la figure 1.13. Deux chemins de buffers symétriques, peuvent être bouclés en oscillateur en anneau notés RO1 et RO2 pour la mesure de fréquence d'oscillation, et la caractérisation de leurs temps de propagation. Un des deux chemins dispose d'une alimentation dédiée V_{DD2} alors que le reste de la structure est alimenté à V_{DD}. Il est également possible de sélectionner une fréquence parmi 2GHz, 1.5GHz, 1GHz et 500MHz, à l'entrée des chemins de buffers (RO en boucle ouverte).

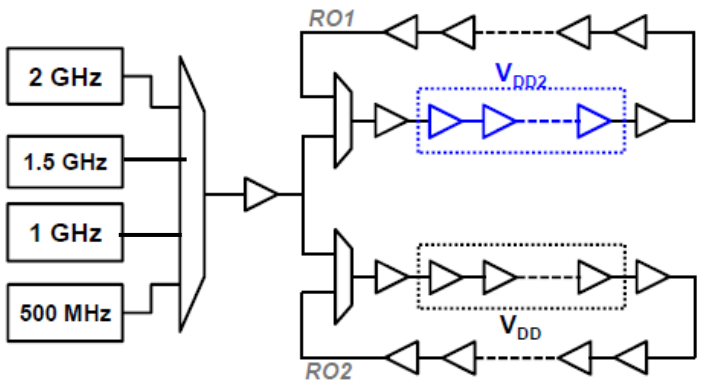


Figure 1.13 : Schéma de la structure de test. Les deux chemins de buffers sont symétriques à deux alimentations différentes. Ils peuvent être bouclés en oscillateur en anneau pour mesurer leur fréquence.

Dans cette application, nous pouvons mesurer soit directement les fréquences des RO1 et RO2 soit la différence de leurs fréquences d'oscillations. C'est la raison pour laquelle nous disposons de deux oscillateurs en anneau et une alimentation V_{DD2} différente de V_{DD}. Le but de posséder deux alimentations, est pour imposer deux signaux de stress différents, afin d'obtenir deux dégradations de fréquences différentes. Cette technique différentielle de mesure de fréquence est très précise pour les faibles différences de fréquence.

Nous rappelons que la structure possède des signaux de contrôles qui permettent de la tester soit en boucle ouverte, soit en boucle fermée.

Nous appliquons une contrainte V_{DD2} = 1.8V pour différentes conditions AC et DC en entrée des ROs. Concrètement cela signifie que la chaine de buffers est alimentée à V_{DD2} = 1.8V mais également qu'elle est en boucle ouverte et qu'à son entrée nous imposons un signal AC de fréquence égale à 2GHz, 1GHz ou 500MHz ou un signal DC. Pour la mesure, la chaine de buffers est bouclée en RO et un signal logique nous permet de démarrer les oscillations afin d'en mesurer la fréquence et la dérive de fréquence après le stress.

L'évolution de la dérive du temps de propagation de chaque porte (notée $\Delta\tau$) au cours de la contrainte est mesurée pour différentes conditions AC et DC en entrée du chemin sur la figure 1.14 à 125°C. Il faut tout d'abord remarquer qu'il n'y a pas de fort effet AC sur la dégradation du délai de chaque porte [82]. Le mécanisme de dégradation BTI est dominant et la contribution HCI est très faible à 125°C.

Nous considérons ici le cas particulier d'un RO de buffer (ou d'inverseur). Au cours d'une contrainte de type AC à une tension V pendant un temps t, toutes les portes sont contraintes à la tension V mais durant le temps t/2. La différence entre différentes conditions AC est le nombre de transitions du signal de contrainte en entrée du chemin. Au cours d'une contrainte DC à une tension V pendant un temps t, seule une porte sur deux est contrainte à la tension V mais cette fois ci durant le temps t. Dans ce cas précis, cette différence est quantifiable par un facteur 2^{1-N}, directement lié au facteur d'accélération en temps (N). Sinon, la différence entre les dégradations d'un chemin induites au cours de contraintes AC et DC est directement liée aux portes utilisées et à la façon dont elles sont connectées.

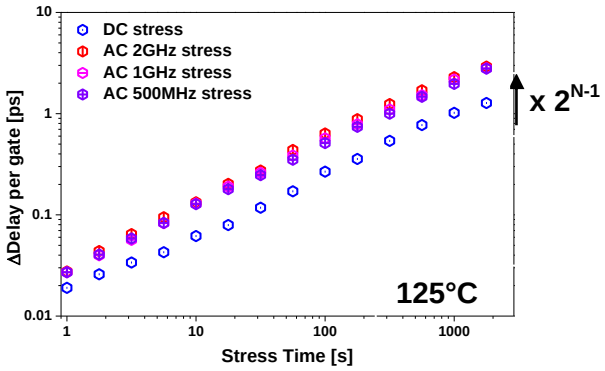


Figure 1.14 : Evolution de la dérive du temps de propagation de chaque porte d'un chemin de buffers au cours de la contrainte V_{DD2} = 1.8V sous différentes conditions AC et DC à 125°C en technologie 28nm FDSOI.

En conclusion, la dégradation induite par une contrainte AC constitue le pire cas par rapport à une contrainte DC. Elle est également plus représentative, car toutes les portes du chemin contribuent à cette dégradation contrairement au cas DC. De plus, la dégradation BTI est dominante à 125°C et la fréquence du signal de type AC n'a pas d'effet sur sa partie permanente [84]. En effet, la dégradation dépend du temps du stress. Par ailleurs, nous remarquons que le délai du buffer à t_0 est 18ps en 28FDSOI. Après l'application d'une contrainte V_{DD2}= 1.8V à 2000s et à 25°C, la dégradation par effet BTI représente 2.2ps (cf. figure 1.14), soit 12.2% de délai supplémentaire.

1.7.2 Effet AC sur un chemin de propagation à faible température

Dans ce paragraphe, nous nous intéressons à la dégradation du délai d'un chemin de porte à la température ambiante, 25°C. Pour cela, nous allons utiliser la structure et les mêmes conditions de test présentées dans le paragraphe 1.7.1 à l'exception de la température qui est fixé à 25°C.

La figure 1.15 donne l'évolution de la dérive du délai par porte (notée $\Delta\tau$) au cours de la contrainte pour différentes conditions AC, à 25°C. Un effet AC sur la dégradation du délai de chaque porte [83] est observable. Cet effet est lié à la fréquence du signal d'entrée. Plus la fréquence du signal d'entrée AC est grande, plus la dégradation induite est importante. Ceci est dû au fait que le mécanisme de dégradation par porteurs chauds est à priori dominant à 25°C. En effet, la dégradation par HCI survient durant les transitions du signal AC en entrée du chemin. Cela signifie à priori que plus la fréquence du signal d'entrée AC ou d'oscillation d'un circuit sera élevée, plus les transitions seront nombreuses et plus la dégradation induite sera importante. C'est ce qui est montré dans la figure 1.15, qui met en évidence un effet sur la dégradation induite liée à la fréquence du signal d'entrée du chemin.

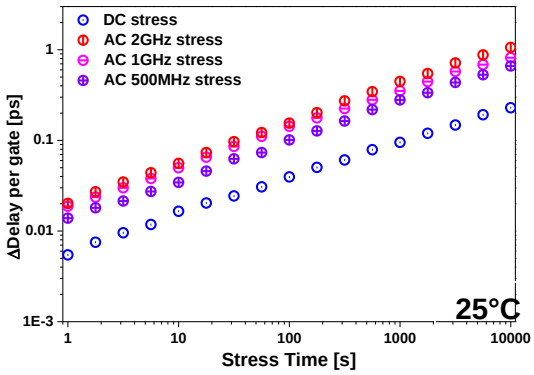


Figure 1.15: Evolution de la dérive du délai de chaque porte d'un chemin de buffers au cours de la contrainte V_{DD2} = 1.8V sous différentes conditions AC à 25°C en technologie 28nm FDSOI.

Dans le prochain paragraphe nous nous intéressons à d'autres sources de variation qui sont le précédé de fabrication, la tension et la température. Nous mettons en œuvre leurs impacts sur le délai des portes logiques.

1.8 Impact des sources de variation sur le délai des chemins logiques

1.8.1 Les types de variabilités

Une conséquence non désirée de la réduction des dimensions des transistors est l'augmentation de la variabilité. Elle se traduit par une augmentation de la dispersion des valeurs des paramètres électriques des transistors tels que la tension de seuil V_{th}, les courants I_{off}, I_{on} et la pente sous-seuil S. Une variabilité non contrôlée peut alors, à terme, être préjudiciable au bon fonctionnement d'un circuit.

L'étude de la variabilité occupe donc une place essentielle dans le développement des technologies avancées. Elle se manifeste à différents niveaux définis en terme de distance spatiale *d* entre deux transistors "identiques" et se décline en deux sous-familles : la variabilité globale et la variabilité locale. Ces deux types de variabilités sont présentés schématiquement sur la figure 1.16 [85] [86].

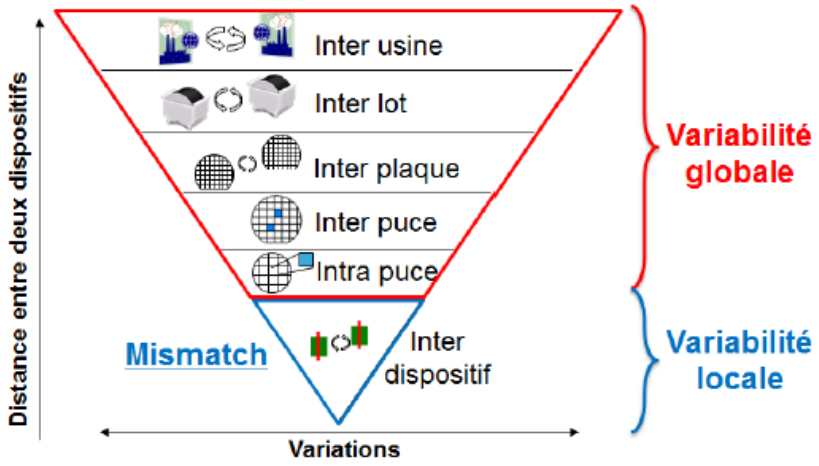


Figure 1.16 : Différents niveaux de variabilité des paramètres électriques des transistors : on parle de variabilité globale lorsqu'on compare ces paramètres entre usines, lots, plaques, puces ou à l'intérieur d'une même puce et de variabilité locale lorsque la comparaison se fait entre deux transistors MOS indépendants et intègres au sein d'une même paire de transistors

- **Variabilité globale**

La variabilité globale fait référence à la variabilité qui existe lorsqu'on compare les paramètres des transistors MOS entre plusieurs usines utilisant les mêmes procédés de fabrication, entre lots, entre plaques ou entre puces. Elle est donc principalement liée aux équipements qui peuvent différer d'une usine à l'autre et à l'enchainement des étapes de fabrication qui peut introduire certaines inhomogénéités. Cette variabilité se traduit en une variation de la tension de seuil V_{th} dont les principales contributions :

- les variations de l'oxyde de grille : les variations de l'épaisseur et de la permittivité de l'oxyde [87], charges fixes [88], charges pièges [89]
- la rugosité de bord de lignes : (LER/LWR : Line Edge/Width Roughness) [90]
- les variations liées aux étapes de gravure de la grille [91] et des STI (Shallow Trench Isolation) [92].
- la granularité de la grille (MGG : Metal Gate Granularity) [93] [94]
- les fluctuations aléatoires des dopants (RDD : Random Discrete Dopants) [95] [96] [97]
- les variations associées aux étapes d'implantations et de recuit [98] [99]
- les variations des épaisseurs des films minces (t_{Si}, t_{BOX} notamment en technologie UTBB-FDSOI) [100]

La variation des procédés a pour conséquence la génération des transistors plus ou moins rapide. Elle est traduite dans les modèles sous la forme de corner. Nous distinguons trois types

de corners : corner Fast Fast (FF) dans lequel le V_{th} et la mobilité des transistors NMOS et PMOS sont élevés et donc ils sont considérés comme des transistors rapides. Le Corner Slow Slow (SS) où les transistors sont lents et le corner Typical Typical (TT) qui est considéré comme le cas nominal. L'évaluation des délais des portes dans ces corners fera l'objet d'étude dans le paragraphe 2.2.1 du chapitre 2.

- **Variabilité locale**

La variabilité locale porte sur les mêmes paramètres ou valeurs mais intervient lorsque la distance entre les transistors MOS est la plus faible possible du point par rapport aux règles de dessin. Pour étudier cette variabilité locale, des structures spécifiques sont utilisées dans lesquelles deux transistors identiques et indépendants sont appariés et situés dans un environnement identique avec des connexions électriques symétriques. Cette étude permet de capturer les variations discrètes de la matière et son origine, et s'explique par des variations stochastiques liées à la nature discrète des défauts, impuretés et dopants dus aux étapes de fabrication mêmes. Cette variabilité est donc couramment intitulée sous le nom de 'mismatch' stochastique.

La conséquence de la variabilité locale est la fluctuation aléatoire la tension de seuil $\sigma_{\Delta Vth}$ qui a tendance à augmenter avec la réduction des nœuds technologiques [101]. Du point de vue du circuit, l'impact de cette variation sur le délai chemins peut être décrit par un modèle stochastique de n portes logiques qui ont un temps de commutation t_{gate} et une déviation standard $\sigma_{t,gate}$. La déviation standard du délai du chemin induite par la variation locale est proportionnelle au nombre de portes (n) des chemins et s'écrit [30]:

$$\sigma_{t,d} = \sqrt{n}.\sigma_{t,gate} \qquad \text{EQU1.28}$$

Notons que l'équation 1.28 n'est valable que si les portes sont toutes identiques, sinon la déviation standard est égale à la somme quadratiques des sigmas des portes du chemin de propagation.

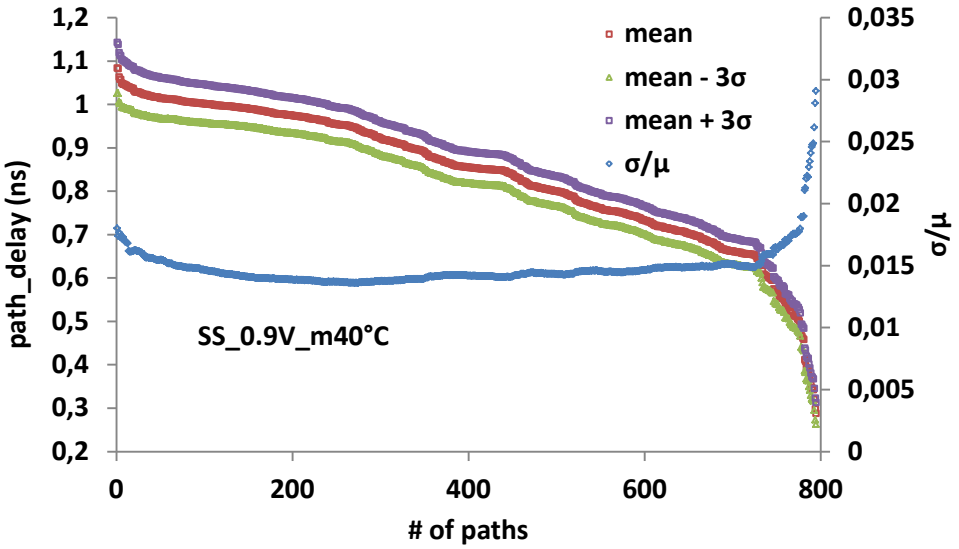


Figure 1.17 : Evaluation des délais et des déviations standard avec la simulation Monte Carlo. Le délai présenté en rouge représente la valeur moyenne de 500 tirages Monte Carlo dans le corner SS à 0.9V et une température de -40°C

Nous avons vérifié l'impact de la variabilité locale sur des chemins de propagation d'un circuit numérique en 28FDSOI. Le corner considéré est le SS, à une tension de 0.9V et une température de -40°C. Nous avons effectué des simulations Monte-Carlo en considérant la variation de la tension de seuil V_{th} et la mobilité μ. Le résultat est illustré dans la figure 1.17, il présente la valeur moyenne de 500 tirages Monte Carlo ainsi que les délais à ±3 σ. Nous remarquons que la déviation standard est plus importante pour les chemins lents et diminue de plus en plus pour les chemins moins lents. Ceci est en adéquation avec l'équation 1.28 car dans le circuit numérique utilisé, plus les chemins sont lents plus ils contiennent de portes logiques. Il est important de souligner que la déviation standard normalisé σ/μ diminue pour les chemins lents

1.8.2 Variations de la tension d'alimentation

Nous nous intéressons maintenant à l'impact de la variation de la tension sur le délai de portes logiques. Cette variation peut être provoquée d'une part par la chute de la tension ohmique à travers les résistances parasites de la grille d'alimentation du circuit, connu sous le nom de IR Drop, et d'autre part par la variation du courant di/dt survenue du réseau R-L qui se forme entre le boitier circuit et le régulateur de la tension [30]. La variation de ces tensions affecte le délai des portes logique. En effet, la relation qui lie le délai d'une porte et la tension V_{DD} est donnée par la loi en puissance [102] :

$$t_{gate} \sim \frac{V_{DD}}{b(V_{DD}-V_{th})^{a}} \qquad \text{EQU 1.29}$$

a et b sont des paramètres de fit.

Ainsi, comme le délai d'un chemin composé de plusieurs portes est obtenu en sommant les délais de chaque porte, ce délai sera plus faible à fortes tensions V_{DD}, selon l'équation 1.19. Nous avons vérifié cette dépendance du délai à la tension V_{DD} pour un chemin composé de 30 portes logiques en 28FDSOI. Nous avons relevé le délai pour chaque tension que nous avons fait varier de 0.7V à 1.25V. Le résultat est donné dans la figure 1.18 qui montre que le délai devient plus faible avec l'augmentation de la tension.

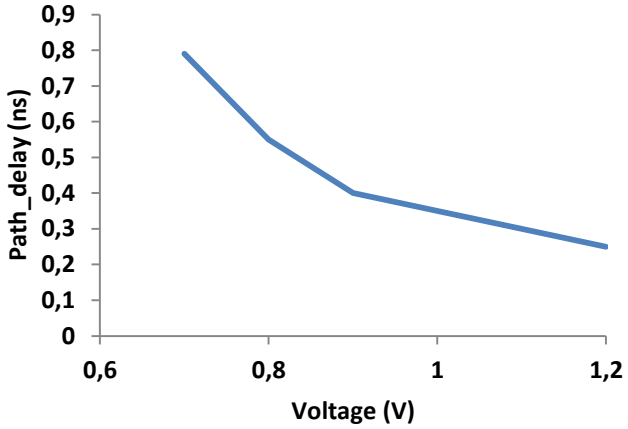


Figure 1-18 : L'impact de la tension sur le délai d'un chemin de propagation

1.8.3 Variation de la température

Comme pour la tension, la variation de la température impacte le délai des portes logiques. Cette variation est causée par la puissance dissipée par effet Joule ainsi que par les fluctuations locale dans les régions où l'activité est très forte, appelé point chaud ou Hot-Spots en anglais. Par ailleurs, la fluctuation de la température n'est pas rapide et possède une constante de temps de l'ordre de la milliseconde à la seconde [103]. Et donc une augmentation de la température réduit la vitesse du circuit à cause de la réduction de la mobilité des porteurs.

Nous avons vérifié l'impact de la température sur le délai du chemin du circuit. En effet, nous avons appliqué sur un chemin composé de 30 portes trois tensions V_{DD} : 0,6V, 0,8V et 1.2V. Pour chaque tension nous avons fait varier la température de -40°C à 100°C et nous avons mesuré le délai. Le résultat est illustré dans la figure 1.19, il montre que pour les tensions élevées (V_{DD} = 1.2V et V_{DD} = 0.9V), le délai du chemin de propagation augmente légèrement avec la température. Cependant, pour les faibles tensions, i.e. V_{DD} = 0.6V, le circuit fonctionne en dessous du point d'inversion [30] et donc la tension de seuil V_{th} décroit plus

rapidement que la mobilité. En conséquence, le délai des chemins montre une caractéristique inverse en fonction de la température. Dans le 28FDSOI, le point d'inversion est établi à 0.8V.

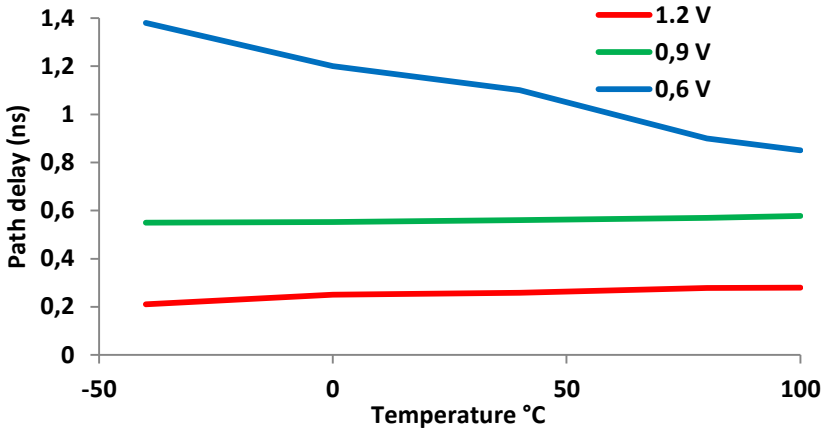


Figure 1.19 : Impact de la température sur le délai du chemin

1.9 Conclusion

Nous avons montré dans ce premier chapitre les enjeux rencontrés suite à la poursuite de la miniaturisation du CMOS à basse consommation et hautes performances, à partir de la technologie 40nm sur substrat massif vers la technologie 28nm sur substrat SOI. Cela a été réalisé tout d'abord par l'intégration des diélectriques à haute constante diélectrique high-K puis dans un second temps par l'introduction de la structure FDSOI. Nous avons mis l'accent sur l'utilisation de la tension substrat V_{BS} qui dans la technologie FDSOI permet de transformer le transistor commandé en double grille face avant et face arrière par le couplage capacitif à travers le BOX et ground plane (GP). Cet effet est mis à profit afin d'obtenir plusieurs tensions de seuil et par conséquent plusieurs niveaux de performances pour les applications dédiées à basse consommation ou à hautes performances.

Dans la deuxième partie, nous nous sommes attachées à présenter les principaux modes de dégradations progressives que sont les mécanismes BTI et HC, qui limitent les performances des transistors NMOS et PMOS à moyen et à long termes. Le phénomène NBTI survient en absence de courant direct, mais il est fortement activé en température pour le PMOS, tandis que le phénomène HCI intervient au niveau du courant I_{DS}. Il est donc fortement lié aux performances et à l'effet de la longueur du canal malgré une activation en température plus faible. Nous avons évalué l'impact de chacun de ces mécanismes, particulièrement les effets AC et DC, sur la dégradation des délais d'un chemin de portes en 28FDSOI. Nous avons

montré que le BTI est dominant à 125°C et que la dégradation due à l'effet AC est très légère à cette température. Ceci est dû à la faible contribution des porteurs chauds. Nous avons montré aussi qu'il y a une différence entre les dégradations induites par l'effet AC et DC, qui s'explique par le fait que le fonctionnement AC implique une contrainte sur toutes les portes de durée t/2 alors que DC implique une contrainte alternativement sur une porte sur deux et une contrainte 'distincte' sur les autres portes pendant la période t. Dans le cas particulier des ROs d'inverseurs, cette différence est quantifiable par un facteur 2^{1-N} qui est directement lié au facteur d'accélération en temps (N).

Le mécanisme de dégradation par porteurs chauds est dominant à 25°C. Au cours des différentes contraintes de type AC en entrée du chemin à 25°C, les transistors MOS du chemin étudié ont été soumis au mécanisme de dégradation par porteurs chauds. Cela se traduit par un effet AC au niveau de la dérive du délai de chaque porte du chemin. Cette dérive est directement liée à la fréquence du signal sous contrainte AC en entrée du chemin et donc au nombre effectif de transitions.

Finalement, nous avons abordé l'impact des différents types de variations sur le délai d'un chemin de portes logiques. La variabilité globale liée à la qualité du procédé de fabrication affecte des transistors, qui peuvent être alors lents ou rapides. La variabilité locale quant à elle, se traduit par la variation de la tension de seuil V_{th} du transistor. Concernant l'impact de la tension, nous avons montré que le délai d'un chemin de propagation est plus court pour les tensions élevées. L'effet de la température sur le délai illustre deux comportements : en dessous du point d'inversion, l'impact de la température est plus conséquent car l'effet de la réduction de la tension de seuil V_{th} est plus importante que la dégradation de la mobilité. Au-dessus, du point d'inversion, le délai du chemin est quasi constant à haute température à cause de l'effet de la mobilité qui l'emporte sur le V_{th}.

Dans le prochain chapitre, nous allons nous intéresser à la modélisation de ces variations pendant la phase de conception du circuit. Nous allons quantifier la contribution de chaque type de variation sur le délai. Aussi, nous allons proposer un circuit qui permet d'adapter dynamiquement la tension d'alimentation afin de réduire sa consommation dynamique du circuit.

CHAPITRE 2 : MONITEURS IN-SITU ET LA GESTION DYNAMIQUE DE LA FIABILITE

2.1 Introduction

Les variations des procédés de fabrication, de la tension et de la température (PVT) ont toujours posé des problèmes lors de la conception des circuits intégrés en technologie avancées. Ces variations impactent la vitesse de la commutation des transistors et par conséquent le délai des chemins de propagation à travers les portes logiques. Afin de garantir un fonctionnement sans erreurs pour une période d'horloge spécifiée, les concepteurs doivent quantifier ces variations avec précision. Traditionnellement, ceci est effectué en rajoutant des marges à la tension d'alimentation [30] afin de garantir le fonctionnement correct dans le pire cas. De plus, dans les technologies avancées, les transistors sont de plus en plus impactés par les effets de vieillissement de type BTI et HCI qui dégradent la tension de seuil et le courant de drain. Ces dégradations sont-elles aussi prises en compte par des marges de tension supplémentaire comme le montre la figure 2.1.

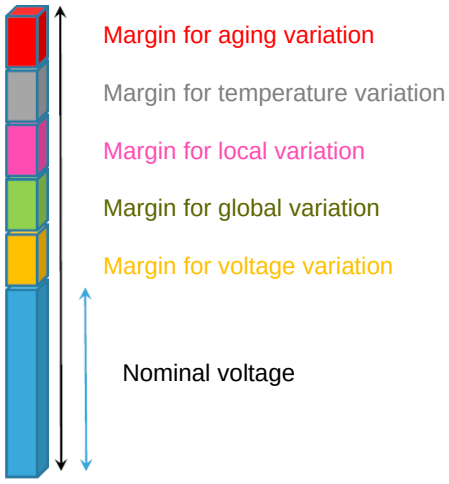


Figure 2.1 : illustration de la prise en compte des PVTs et du vieillissement.

A cause de ces marges supplémentaires de tension qui se traduisent en pratique par des marges de délais plus courtes, et par conséquent l'utilisation de portes logiques très rapides, il en résulte un circuit surdimensionné en surface avec une consommation plus importante, voire excessive. Dans ce contexte, nous allons proposer des circuits de monitoring qui sont insérés dans le circuit digital initial permettant de réduire les marges de tension en fonction des

conditions d'utilisation du circuit, et ce à travers l'utilisation des techniques adaptatives telles que AVS (Adaptive Voltage Scaling) [106], ABB (Adaptive Body Bias) [107] [108] et DFVS (Dynamic Frequency Voltage Scaling) [109]. Il s'agit du moniteur in-situ de mesure de dépassement de délai nominal.

Dans ce chapitre, nous allons commencer par détailler les méthodologies d'évaluation des PVTs et du vieillissement utilisées par les concepteurs dans une technologie donnée, typiquement le 28FDSOI fabriqué par STMicroelectronics. Nous allons appliquer ces méthodologies pour quantifier les marges de délai (ou de tension) imposées par chaque type de variation : P, V, T et vieillissement.

Dans un second temps, nous allons expliciter les approches qui permettent la réduction des marges. En effet, nous allons comparer des méthodes de monitoring basées sur la duplication des éléments logiques du circuit original avec les moniteurs in-situ. Ces derniers se subdivisent en deux catégories : les moniteurs de détection d'erreur et de pré-erreur. Nous détaillons le fonctionnement de chaque type et mettons l'accent sur leurs avantages et inconvénients.

Nous allons ensuite présenter les résultats d'implémentation des moniteurs de pré-erreurs issues de la littérature dans des barrettes de silicium en 28FDSOI, nous allons rappeler les résultats de caractérisations qui concernent la fenêtre de détection.

2.2 Evaluation de l'impact des PVTs et du vieillissement sur les portes logiques d'une technologie numérique

Dans ce paragraphe, nous allons présenter la méthode MMMC (Multi Mode Multi Corner) qui permet de prendre en compte les variations dans leurs globalités. Ensuite nous présenterons les approches d'estimation des délais induits par la variation locale, utilisées dans l'analyse statique du timing (STA). Finalement, nous détaillerons le flot de simulation du vieillissement prenant en compte l'impact du vieillissement.

2.2.1 Méthodes d'analyse de délais

L'impact des PVTs sur les délais a été explicité dans le chapitre1. Nous avons illustré qu'un chemin de propagation est plus rapide dans le corner FF (Fast Fast) qu'en TT (Typical

Typical) et en SS (Slow Slow). De la même manière, il est d'autant plus rapide que la tension d'alimentation est importante. Enfin, les portes commutent plus rapidement aux températures en dessus du point d'inversion qui est de 0.8V pour la technologie 28FDSOI.

Compte tenu de toutes ces variations, le fonctionnement du circuit doit être validé dans cette large gamme de conditions de PVT avant la fabrication. La méthode préconisée est dite MMMC (Multi Mode Multi Corner), elle consiste à vérifier la propagation des délais dans des cas extrêmes, notamment dans le meilleur et le pire des cas. Si le circuit fonctionne dans ces deux conditions, il est supposé alors fonctionner dans toutes les autres configurations intermédiaires.

Cette méthode repose sur la modélisation des délais des portes pour chaque combinaison réaliste de procédés, tension et température. En effet, une rampe de signal (transitions de '0' vers '1' ou de '1' vers '0') est appliquée sur les entrées de la porte (cf. figure 2.2) et la pente du signal de sortie est mesurée entre 5% et 95% du V_{DD} pour différentes capacités de charge C_L.

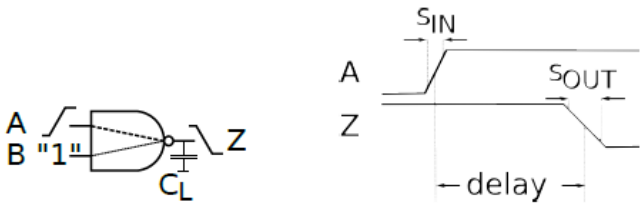


Figure 2.2 : Illustration du calcul de délai d'une porte logique : Une transition est appliquée en entrée et le temps que met cette transition pour atteindre la sortie de la porte constitue le délai de la porte.

Ces délais sont stockés dans une table de recherche LUT (Look-up Table) qui est une table à deux dimensions contenant les valeurs des délais pour chaque pente de transition en entrée et capacité de charge en sortie. Les délais pour d'autres valeurs intermédiaires de transition-capacité sont obtenus soit par interpolation soit par une méthode analytique qui se base sur des équations du temps de propagation. Donc, plus le nombre de points de caractérisation est grand plus le délai calculé est précis.

Cependant, cette méthode de caractérisation n'est plus utilisée dans les technologies avancées, notamment en dessous du 90nm. En effet, dans ces technologies, les temps de propagation des interconnexions peuvent dépasser celles des cellules, ce qui induit une difficulté d'estimation des délais. C'est la raison pour laquelle la méthode Current Source Model (CSM) a été introduite.

Le principe de CSM est de modéliser la porte comme une tension contrôlée par une source de courant qui charge un réseau d'interconnexion complexe et des portes à grand fan-out. A cet effet, Synopsys [110] propose l'approche dite CCSM (Composite Current Source Model) qui stocke les valeurs du signal temps-courant dans les LUTs. Cadence [111] propose une approche dite ECSM (Effective Current Source Model) qui stocke le signal temps-tension qui est converti en source de courant. Les deux méthodes, CCSM et ECSM, ont l'avantage d'être compatible avec les outils de la CAD et largement adoptés par l'industrie. Une telle technique de caractérisation apporte un gain considérable en termes de temps CPU et permet la réalisation d'analyses temporelles de circuits complexes en des temps relativement courts. Toutefois, elle a un coût dans la mesure où la caractérisation des cellules de bibliothèque peut être longue.

Cette technique de caractérisation permet le calcul du temps de propagation dans les portes logiques du circuit dans les différents PVTs ou des corners. Cette validation est effectuée par la méthode dite STA (Static Timing Analysis). La STA a pour objectif principal de vérifier que les données sont correctement acquises par les éléments séquentiels, c'est à dire que les contraintes de temps de propagation et d'établissement de maintien des données soient respectées en entrée des bascules (Flip-Flop) d'entrée vers la sortie. En effet, l'architecture des circuits numérique synchrone est constitué des éléments séquentiels (bascules) et combinatoire (portes logiques) (cf. figure 2.3 (a)). A chaque coup d'horloge, la donnée est transmise depuis une bascule de lancement (Startpoint) vers une bascule de capture (Endpoint) via la logique combinatoire. Le délai de propagation du signal à travers la logique combinatoire doit être inférieur à la période du signal d'horloge et la donnée doit arriver suffisamment tôt et être stable avant le temps de setup de la bascule de capture. Cette première contrainte temporelle est appelée contrainte de 'setup'. La deuxième contrainte est relative au temps pendant lequel la donnée doit être stable après le front d'horloge. Cette contrainte est appelée contrainte de 'hold' (cf. figure 2.3 (b)).

Dans la STA, nous définissions le slack par l'équation :

$$Slack = RT - AT \qquad \text{EQU 2.1}$$

RT (Required Time) est le temps nécessaire pour la propagation de la donnée depuis la bascule de lancement vers la bascule de capture. AT est le temps d'arrivée de la donnée. Un slack négatif veut dire que le délai de propagation de la donnée est soit trop lent soit trop rapide et donc les contraintes de setup et du hold sont violées.

Les calculs des slacks se font donc dans chaque corner PVT et le circuit est dimensionné de façon que le slack du plus long chemin soit supérieurs à zéro.

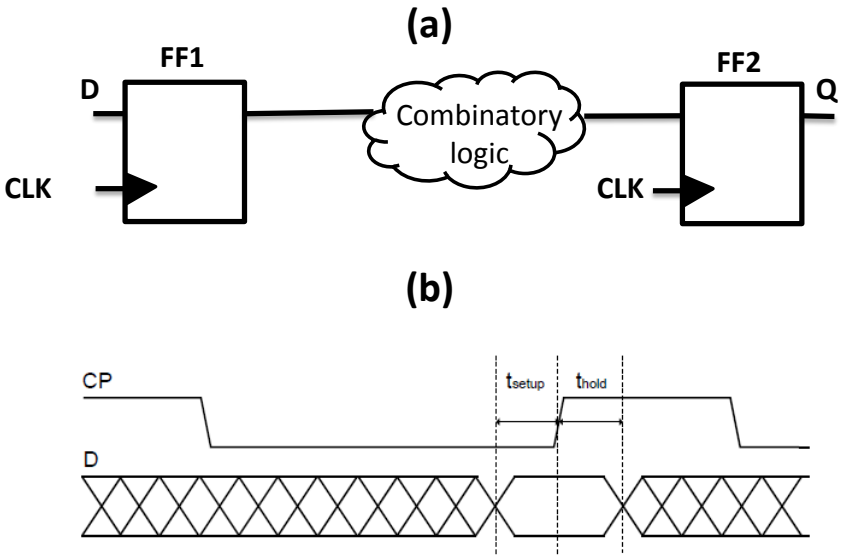


Figure 2.3 : Illustration d'un chemin de donnée (a) : la donnée se propage de la bascule FF1vers la bascule FF2 après front montant d'horloge. (b) Diagrammes présentant les cas où le temps de setup (à droite) et le temps de hold (à gauche) sont respectés

2.2.2 Prise en compte des variations locales

Initialement, la prise en compte de variations locales dans la phase de conception se réalisait en rajoutant une marge de délai supplémentaire à chaque porte logique du circuit ou au chemin de propagation. Cette approche est appelé OCV (On Chip Variation). Les données des fondeurs ont montré que la méthode surestime les marges de délai et mène à un surdimensionnement du circuit. En effet, la dégradation des paramètres des transistors tels que le V_{th} dépend de leurs emplacements physique [112] et des effets layout [113] ; les transistors regroupés dans une surface bornée du circuit ne subissent pas les mêmes dégradations du V_{th} par rapport aux autres transistors. OCV ne tient pas compte de cette corrélation et applique une marge fixe de délai à la totalité des portes du circuit.

Afin de réduire ces marges excessives, la méthode dite SSTA (Statistical Static Timing Analysis) a été proposée [114] [115]. Avec cette approche, la méthode de calcul des délais des portes utilise une variation stochastique de la transition du signal à l'entrée de la porte vers sa sortie, représentée par μ (moyenne) et σ (dispersion). Pour chaque transition d'un signal d'entrée, le délai de la porte est calculé avec les paramètres μ et σ en plus. L'inconvénient de la SSTA est le temps d'exécution pour la caractérisation des portes, qui est relativement long à cause du nombre de transistors dans les circuits actuels. Ce temps est jugé inacceptable par les concepteurs, c'est pourquoi son application se limite qu'aux portes de l'arbre d'horloge.

Une alternative à la SSTA a été proposée par Synopsys [116], il s'agit de l'AOCV (Advanced On Chip Variation). Selon la méthode OCV, les portes se trouvant bien au milieu d'un chemin de propagation, sont moins affectées par les variations et donc une marge moins importante peut être appliquée sur les portes se trouvant sur ces positions. Des simulations Monte-Carlo (MC) peuvent être effectuées pour calculer la valeur de la variation à appliquer sur chaque porte selon sa position dans la profondeur du chemin et sa corrélation spatiale.

Dans la même logique, l'approche statistique dite POCV (Parametric On Chip Variation), considérée plus précise que l'AOCV, a été proposé en [116]. Elle calcule le délai des portes en prenant en compte la distribution μ et σ des interconnexions ; i.e. résistances, capacité, et charge des fils de connexions.

Le tableau 2.1 donne une comparaison des différentes approches permettant l'estimation des délais en présence des variations locales.

Dans la suite de cette thèse, et pour plus de précision, la variation locale est estimée en utilisant la méthode Monte-Carlo (MC). Dans les simulations MC, nous considérons la variation de la tension de seuil V_{th} des transistors ainsi que la mobilité μ.

	OCV	AOCV	POCV
Approche	Application globale des marges	Adaptation des marges en fonction de la position dans le chemin et dans le circuit	Propagation statistique des délais des interconnexions
Avantage	Facile à appliquer	Etape intermédiaire vers la SSTA	Plus précise
Défis	pessimiste	Exige des données technologiques	Exige beaucoup de données du fondeur pour une analyse précise

Tableau 2.1 : Comparaison des différentes méthodes de l'analyse

Nous allons maintenant nous intéresser à la méthodologie d'estimation du vieillissement : Nous allons présenter dans un premier temps le flot du vieillissement mis en place par STMicroelectronics. Dans un second temps, nous établirons un modèle de dégradation d'une porte de type inverseur. Ensuite, nous montrons le résultat de l'application du flot de vieillissement sur le délai de plusieurs chemins logiques.

2.2.3 Prise en compte du vieillissement

Nous allons maintenant expliquer le flot d'évaluation du vieillissement. Comme illustré dans la figure 2.4, le flot est composé de deux étapes. Dans une première étape (cf. figure 2.4(a)),

une simulation transitoire est réalisée et selon l'activité de chaque MOS composant le circuit, une dégradation est calculée à l'aide des équations du modèle 'Reliability API'. Ceci a été réalisé dans des conditions d'activité, de température et de tension les plus défavorables qu'un produit peut rencontrer lors de son utilisation. Dans une seconde simulation (cf. figure 2.4(b)), le circuit est analysé dans des conditions de P, V et T du pire cas de fonctionnement. Lors de cette simulation, la dégradation des transistors MOS calculée précédemment est mise à jour dans la carte des modèles des MOS. Grâce à ce flot, il est possible de décrire le comportement de toutes les portes quel que soit la situation du vieillissement et pour toutes les conditions d'utilisation.

Cette méthodologie est appliquée sur toutes les portes et dans tous les PVTs de librairies. Les délais vieillis sont stockés dans des LUTs afin qu'ils puissent être utilisés dans la STA.

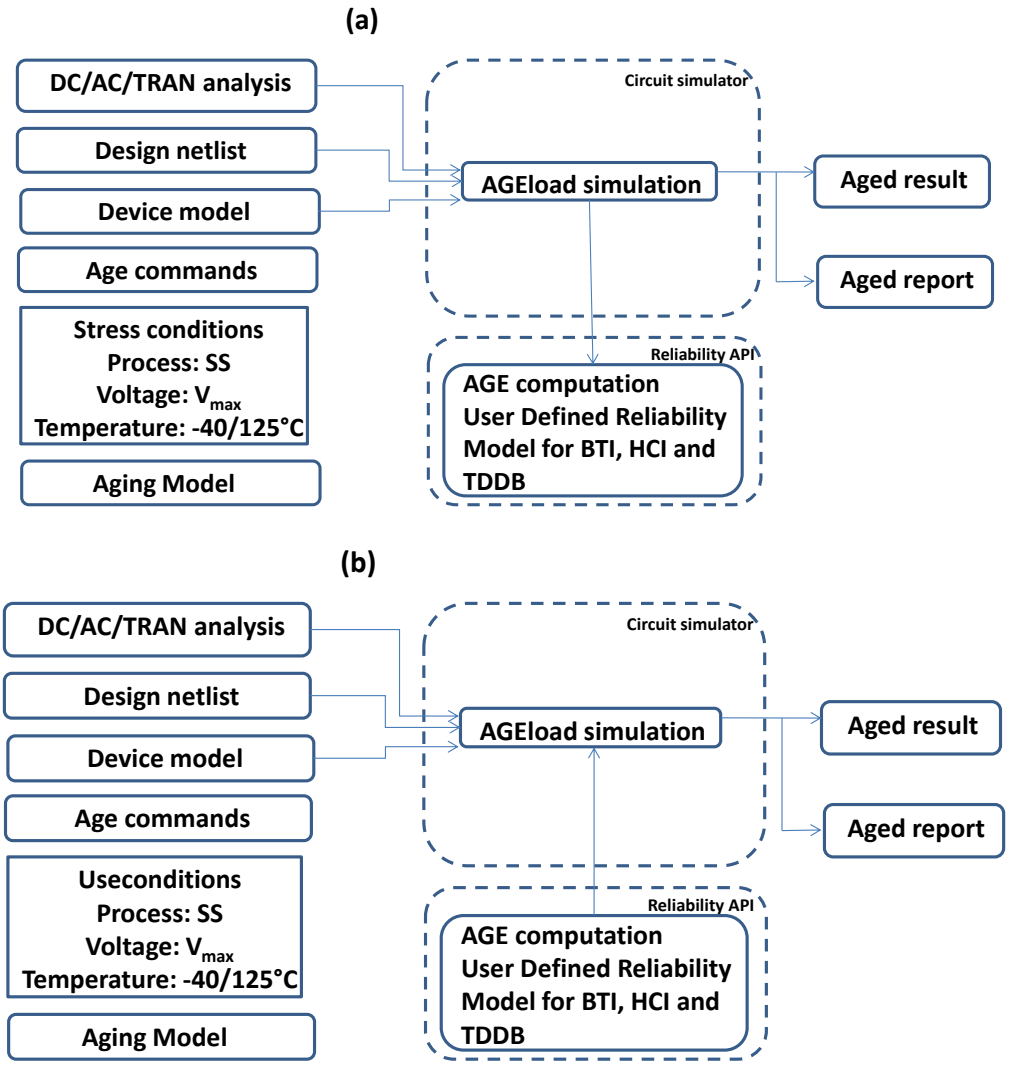


Figure 2.4 : Illustration du flot de simulation du vieillissement : (a) illustration du stress (b) illustration de la caractérisation

- **Modèle de vieillissement au niveau porte logique**

Afin de bien comprendre le vieillissement d'une porte logique, nous nous intéressons au comportement des inverseurs lors du vieillissement par BTI et HCI. Nous cherchons à établir un modèle analytique afin de s'affranchir des simulations SPICE. Pour ce faire, nous considérons trois inverseurs dans la technologie 32nm sur substrat massif (bulk) avec des différentes capacités à conduire le courant (Drive), en l'occurrence INVX8, INVX33 et INVX67. Sur ces inverseurs, nous appliquons le flot de la figure 2.4 en considérant différentes pentes d'entrée S_{in} et capacités de charge C_{load}. Le stress est réalisé avec un signal DC de 1.15V à 125°C à différents temps de stress et de Signal Probability (SP). Le choix de ces conditions de stress et de caractérisation est justifié par l'hypothèse d'un fonctionnement à tension maximale V_{max} dans toute la vie de la porte logique et pour laquelle nous devons vérifier le pire PVT fonctionnelle. Dans ces conditions, nous mesurons la dégradation de la pente de sortie et le délai des transitions du signal de l'entrée de la porte vers sa sortie. Les données de stress et de caractérisation appliquées sont résumées dans les tableaux 2.2 et 2.3.

Transition S_{in} (ns)	0.002, 0.003, 0.0045, 0.006, 0.01, 0.014, 0.018, 0.02, 0.025, 0.03, 0.034, 0.037, 0.04, 0.045
Capacité de charge (pF) C_{load}	0.0005, 0.001, 0.0014, 0.0018, 0.0025, 0.0035, 0.0069,0.0085, 0.0104, 0.0138, 0.0162, 0.0207, 0.0350, 0.0413, 0.11, 0.21, 0.6407
Tension de stress (V)	1.15
Temps de stress (y)	0.2, 0.02, 2
Température (°C)	125
Fréquence (Ghz)	1
Forme du signal : probabilité d'observer des transitons	SP100, SP90, SP75, SP50, SP25, SP10, SP0

Tableau 2.2 : Les conditions de stress

Transition (ns)	0.002, 0.003, 0.0045, 0.006, 0.01, 0.014, 0.018, 0.02, 0.025, 0.03, 0.034, 0.037, 0.04, 0.045
Capacité de charge (pF)	0.0005, 0.001, 0.0014, 0.0018, 0.0025, 0.0035, 0.0069,0.0085, 0.0104, 0.0138, 0.0162, 0.0207, 0.0350, 0.0413, 0.11, 0.21, 0.6407
Tension de caractérisation (V)	0.9
Température de caractérisation (°C)	-40 °C

Tableau 2.3 : Les conditions de caractérisations

Les résultats sont donnés dans la figure 2.5. Ils montrent la dégradation des délais des inverseurs à 10 ans de fonctionnement en fonction de la dégradation de la pente du signal à la

sortie des inverseurs. Nous remarquons qu'il y a une faible corrélation entre le délai de la porte et la dégradation de la pente de sortie. En effet, pour une faible dégradation de la pente de sortie, i.e. 1%, le délai de l'inverseur IVX67 présente une dégradation de 4% alors que pour 5% de dégradation de la pente de sortie, le délai est dégradé d'un très faible pourcentage, i.e. 2%. Par contre, pour les dégradations de la pente de sortie supérieures à 10%, la dégradation du délai présente une forte dispersion. Donc, il n'est pas possible de relier la dégradation de la pente de sortie au délai de la porte. Il faudra développer deux modèles.

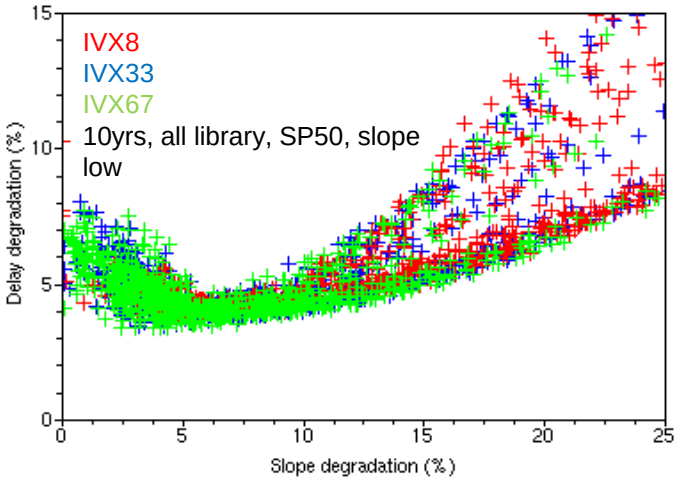


Figure 2.5 : Dégradation du délai en fonction de la dégradation de la pente de sortie pour trois inverseurs après un vieillissement de 10 ans.

Afin d'avoir un modèle le plus générique possible, nous nous proposons de faire une séparation de variable : la dégradation du délai ou la pente de sortie peuvent s'exprimer par le produit de différents fonctions indépendantes dont les paramètres sont SP, S_{in} et C_{load}.

Nous nous intéressons maintenant à la première composante de la fonction du modèle, en l'occurrence le signal SP. Nous considérons un inverseur INVX33 et un temps de stress de 0.02 année (72 jours). La dégradation de la pente de sortie de l'inverseur est calculée pour chaque probabilité du signal SP avec différentes transitions S_{in} et capacités de charges C_{load} du tableau 2.2. Le résultat est illustré dans la figure 2.6, il montre que plus les transitions S_{in} et les capacités de charge C_{load} sont grandes plus la dégradation est importante. Ceci est valable pour un signal SP allant jusqu'à 20% d'activité. A partir de SP=20%, la dégradation de la pente de sortie devient quasi constante. Elle est exprimée par l'expression suivante:

$$g(SP) = \left(\frac{100-SP}{SP}\right)^{a} \qquad \text{EQU2.1}$$

a est un paramètre de fit, il vaut 0.11 pour l'inverseur IVX33. Néanmoins, la dépendance exprimée par l'équation 2.1 est vérifiée pour différents inverseurs de la librairie.

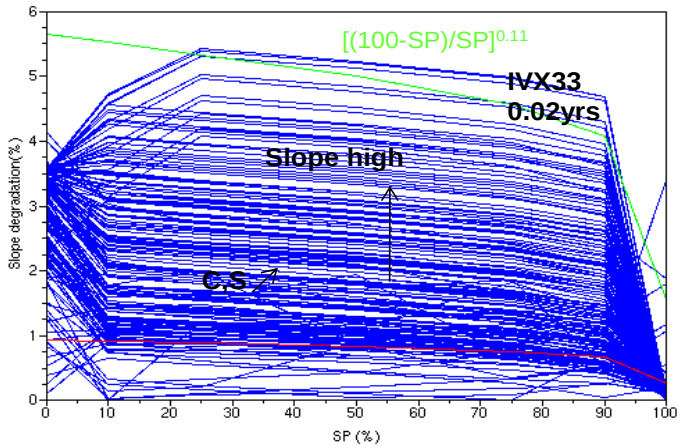


Figure 2.6 : Dégradation de la pente de sortie en fonction de la probabilité du signal SP pour un inverseur IVX33 à 0.02 année de fonctionnement

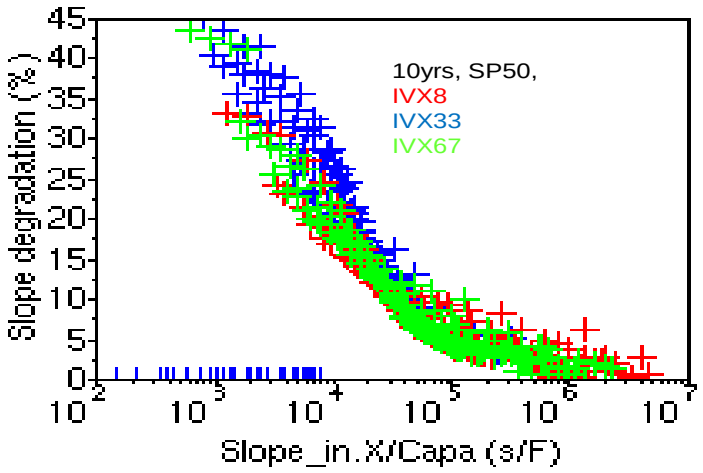


Figure 2.7 : Dégradation de la pente de sortie en fonction rapport S_{in}/C_{load} pour 3 inverseurs à 10 ans.

Nous nous intéressons au second terme de la fonction du modèle qui est la dépendance du délai à la transition S_{in} et la capacité de charge C_{load}. En effet, nous avons calculé la dégradation de la pente de sortie en fonction du rapport S_{in}/C_{load} pour trois inverseurs, à 10 ans, en appliquant un signal SP=50%. Le résultat est donné dans la figure 2.7, elle montre une normalisation de la dégradation de la pente de sortie avec l'expression $(S_{in}/C_{load})*X$. X étant la capacité à conduire le courant (drive) de l'inverseur. La fonction qui relie la dégradation de la pente de sortie au rapport S_{in}/C_{load} est donnée par l'équation 2.2 :

$$h\left(X.\frac{S_{in}}{C_{load}}\right) \qquad \text{EQU 2.2}$$

Ainsi, le modèle de dégradation totale de la pente de sortie qui tient compte de ces paramètres notamment S_{in}/C_{load}, et SP s'écrit :

$$\frac{\Delta\delta}{\delta_0} = f(t, V_{stress}, T) * g(SP) * h \qquad \text{EQU 2.3}$$

$f(t_{stress}, V_{stress}, T)$ est exprimé par la loi d'Arrhenius et donc nous avons :

$$\frac{\Delta\delta}{\delta_0} = A.\exp\left(\frac{E_a}{K_B.T}\right).V_{Gstress}^{b}.{t_{stress}}^{n}.\left(\frac{100-SP}{SP}\right)^{a} * h\left(X.\frac{S_{in}}{C_{load}}\right) \qquad \text{EQU 2.4}$$

A, b, n, a sont des paramètres de fit. E_α est l'énergie d'activation, K_B est la constante de Boltzmann. T est la température. $V_{Gstress}$ et t_{stress} sont la tension et le temps de stress respectivement.

A l'aide de cette équation, nous avons calculé la dégradation de la pente de sortie de l'inverseur IVX33 pour plusieurs valeurs SP et l'avons comparé à la dégradation obtenue par la simulation. Le résultat est donné dans la figure 2.8, il montre une corrélation relativement bonne des deux méthodes d'évaluation. En effet, la dégradation de la pente de sortie en fonction de S_{in}/C_{load} suit la même évolution pour les deux méthodes de calcul : modèle et simulation.

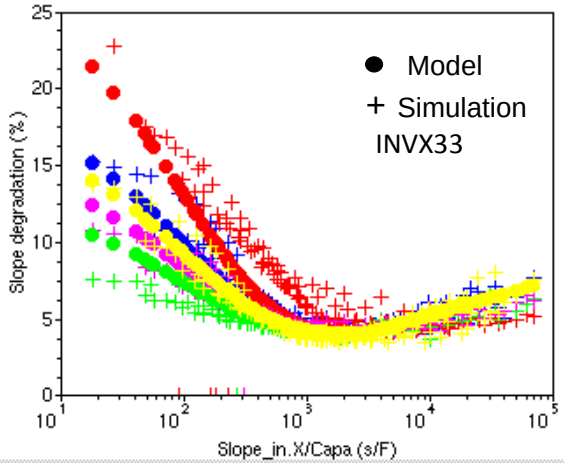


Figure 2.8 : Comparaison de la dégradation de la pente de sotie d'un inverseur IVX33pour plusieurs SP en fonction de rapport S_{in}/C_{load} avec les deux méthodes : simulation et modèle.

Avec le modèle de l'équation 2.4, nous pouvons ainsi estimer l'impact du vieillissement sur la dégradation des délais. Pour vérifier la précision du modèle, nous l'avons comparé aux dégradations obtenues par le modèle Design in Reliability (DiR) mis en place par STMicroelectronics. Pour cet exercice, nous avons considéré tous les inverseurs de la librairie du 32nm et différents S_{in}, C_{load} et SP. Le résultat de la dégradation du délai de ces inverseurs pour les deux méthodes (modèle donné par l'équation 2.4 et modèle DiR) est donné dans la figure 2.9, il montre une imprécision dans l'estimation des délais, surtout pour la dégradation de la pente de sortie. En effet, pour une dégradation de pente sortie de 30% estimé par le modèle de l'équation 2.4, le modèle DiR montre une dégradation qui varie de 15% à 40% (courbe verte de la figure 2.9). La dégradation des délais quant à elle suit la même tendance (courbe rouge) : par exemple, pour une dégradation de 10% donnée par le modèle, le modèle DiR montre une variation de délai entre 5% et 15%.

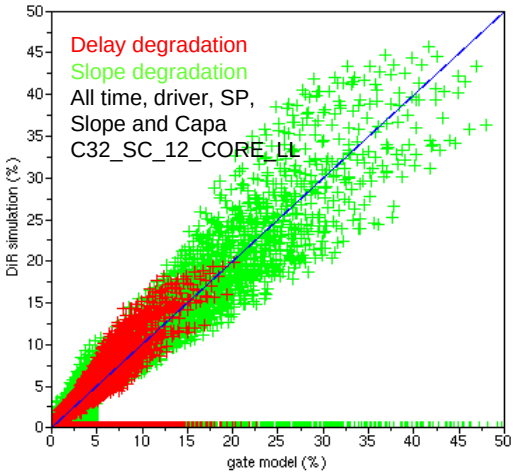


Figure 2.9 : Comparaison des délais obtenus par le modèle DiR et le modèle de l'inverseur

Bien qu'un inverseur soit la porte la plus simple de la bibliothèque, nous ne pouvons pas estimer avec précision la dégradation du délai avec un modèle qui tient compte des variations des S_{in}, C_{load} et SP. Pour faire face à cette situation, la dégradation des délais des portes logiques est calculée en prenant en compte une seule transition du signal d'entrée S_{in} et une capacité de charge C_{load}. La valeur de S_{in} et C_{load} est égale à la valeur moyenne des transitions des capacités contenues dans LUT à t_0.

Nous nous proposons maintenant de calculer les délais induits par le vieillissement de 650 chemins de propagation d'un circuit numérique en procédant de deux manières :

- Nous avons effectué une STA avec des librairies vieillies pour les portes se trouvant sur ces chemins.
- Nous avons extraits la netlist SPICE des chemins de propagation, nous les avons simulés en vieillissement, en appliquant un signal de stress de 1.15V à 125°C et une activité de 30%. Ensuite, nous avons caractérisé ces chemins dans le corner SS_0.9V_m40°C pour un vieillissement à 10 ans.

Le résultat des comparaisons est donné dans la figure 2.10 (a). En comparant avec les délais à t_0, nous remarquons d'une part que les délais des chemins se sont dégradés à cause du vieillissement et d'autre part que les deux méthodes de calcul montrent une bonne corrélation (STA et SPICE). Nous remarquons aussi que l'impact du vieillissement à 10 ans est quasiment constant pour tous les chemins, i.e. 10ps de dégradation moyenne pour des chemins à 30 portes en moyenne. Ceci s'explique par le fait que nous avons appliqué le même signal (Stimuli) de caractérisation sur tous les chemins. Cette approche est pessimiste et ne reflète pas forcement la réalité car pendant l'usage du circuit, les chemins logiques reçoivent des signaux avec différentes activités. La STA ne peut pas calculer les délais en tenant compte

du ‘vrai’ vecteur d’entrée, elle reste donc une méthode très approximative pour estimer le vieillissement.

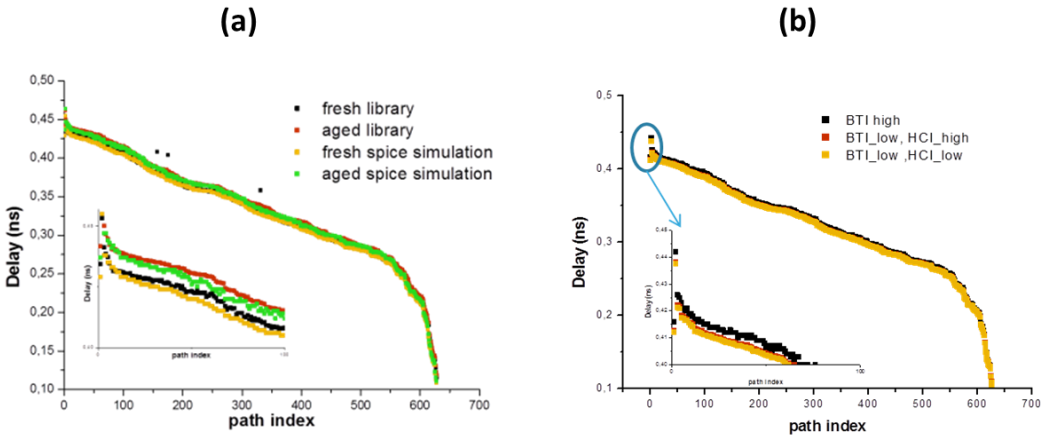


Figure 2.10 : (a) Comparaison des délais à t_0 et après vieillissement de 2 ans aves la STA et le SPICE : l’impact du vieillissement est bien mis en œuvre. (b) contribution de chaque mécanisme de vieillissement.

Notons aussi que ce flot permet de sélectionner la nature du mécanisme de vieillissement (BTI, HCI, TDDB) à considérer dans le calcul des délais à l’aide d’une interface graphique. Ainsi, nous avons évalué la contribution de chaque mécanisme de dégradation en effectuant les mêmes simulations mais en tenant compte seulement du BTI, ou du HCI seul et les deux à la fois. La figure 2.10(b) montre que le BTI est le premier contributeur du vieillissement du circuit. Si on tient compte seulement du HCI, les délais à t_0 et après vieillissement sont quasiment identiques, soit 2ps de dégradation due au porteurs chauds pour un vieillissement de 10 ans.

2.3 Impact global sur le délai de portes

Les méthodologies présentées dans le paragraphe 2.3 permettent d’évaluer le délai supplémentaire induit par la variation des PVTs et le vieillissement. Nous nous proposons d’appliquer ces méthodologies sur 100 chemins d’un circuit numérique en 28FDSOI qui contiennent en moyenne 30 portes par chemin et dont le plus lent a une valeur de délai de 0.4ns. Le but est de quantifier l’impact de chaque type de variations sur le délai à savoir le procédé de fabrication, la tension, la température, la variation locale et le vieillissement.

Pour ce faire, la première étape de la démarche consiste à calculer les délais dans un corner de référence. Dans notre cas, nous avons choisi un corner typique à 0.9V et 125°C noté aussi TT_0.9V_125°C. Ensuite, nous avons fait varier un seul paramètre du corner PVT en fixant les deux autres. Par exemple, pour quantifier l’impact du procédé de fabrication, nous avons

calculé le délai dans les corners FF_0.9V_125°C et SS_0.9V_125°C, la température restant constante et nous avons comparé les délais avec le corner de référence. Il faut noter que l'évaluation du vieillissement et de la variation locale sera faite avec les librairies vieillies et la simulation Monte-Carlo respectivement.

Le résultat de ces variations est donné dans la figure 2.11. En se référant au corner TT_0.9V_125C, le type de procédé induit une variation supplémentaire de délai allant jusqu'à 50ps (différence moyenne des délais dans les corners SS_0.9V_125°C et FF_0.9V_125°C par rapport au TT_0.9V_125C). Cette variation est plus importante lorsqu'on analyse la variation de la tension, elle représente 100ps de dégradation lorsque la tension est augmentée ou réduite de 0.1V (cas des corners SS_0.8V_1250C ou FF_1.1V_125°C). La température quant à elle induit le même ordre de grandeur de variation que le procédé, approximativement 50ps en l'occurrence. Nous constatons le même ordre de grandeur de variation pour le vieillissement. La dispersion locale représente la plus petite variation de délai : environ 10ps à 3σ.

Les délais de la collection des chemins de référence ont été initialement classés par ordre décroissant de slack. En faisant varier les paramètres des PVTs, le classement des chemins subit une modification. Il s'avère que l'ordre n'est plus respecté déjà d'un corner à un autre et la tendance générale est une perte de classement des chemins si on rajoute les variations de la température et de la tension.

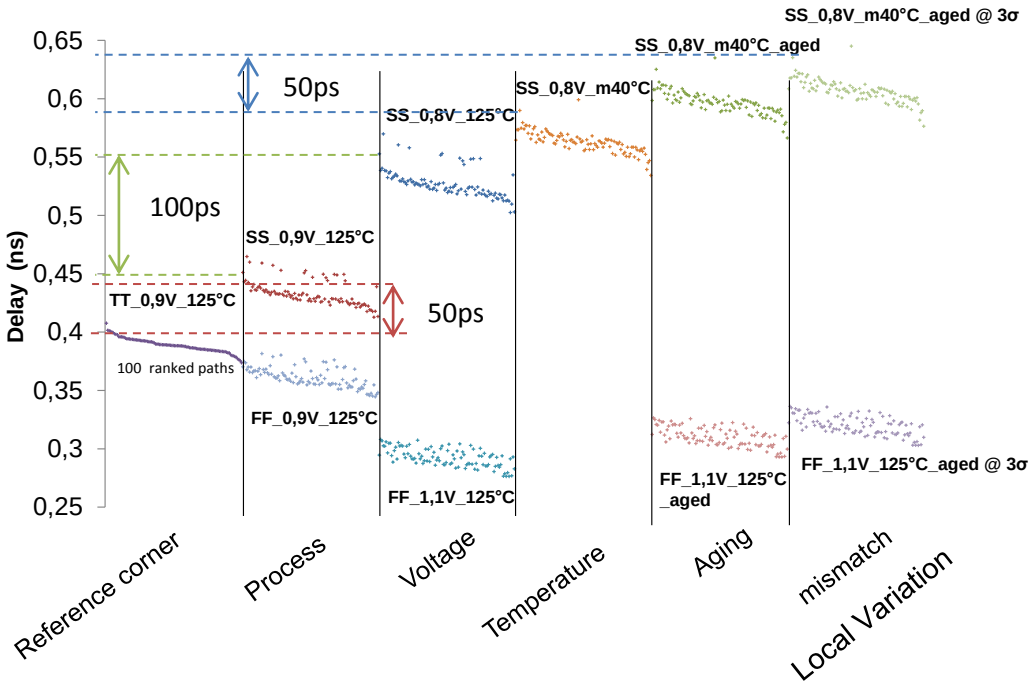


Figure 2.11 : L'impact de la variation des PVTs et du vieillissement sur les délais des chemins dans la technologie 28FDSOI.

En conclusions, pour prendre en compte tous les types de variation, i.e. locale, globale et le vieillissement, des marges importantes de tension ou de délai sont à considérer. Par exemple il faut compter 100ps pour une variation de 100mV, ce qui présente 1ps par 1mV (d'après les corners ss_0.9V_125°C et ss_0.8V_125°C), ou 50ps de plus pour accommoder les variations de la température et encore 50ps pour compenser le vieillissement.

La conception des circuits dans ces conditions peut s'avérer complexe et les contraintes de fréquence peuvent être inatteignables. Nous proposons de réduire les marges relatives au vieillissement et à la variation locale en utilisant les moniteurs in-situ. Ces marges étant d'environ 50ps, représenté par la différence moyenne des délais des corners SS_0.8V_m40C et SS_0.8V_m40C_aged@3σ. En termes de tension d'alimentation, cela peut représenter 50mV. Comme la puissance dynamique est proportionnelle au carré de la tension: $P \approx V_{DD}^2$, la réduction de la tension V_{DD} peut s'avérer intéressante avec un réel impact sur la puissance consommée. L'objet donc du prochain paragraphe est la présentation des moniteurs in-situ qui permettent de réduire ces marges.

2.4 Moniteurs in-situ : état de l'art

La réduction des marges de tension est possible à l'aide des techniques adaptatives telles que Adaptive Body Bias (ABB), Adapative Voltage Scaling (AVS) et Dynamic Voltage and Frequency Scaling (DVFS).

AVS et DVFS consiste à ajuster dynamiquement la tension d'alimentation et la fréquence en fonction de l'usage du circuit afin de moduler la vitesse de commutation des portes et tenir les mêmes spécifications. ABB quant à elle consiste à ajuster la tension des caissons du transistor MOS afin de moduler la tension de seuil V_{th} et rendre le circuit plus rapide en commutation. L'information pour effectuer ces régulations provient des circuits dédiés qui mesurent le temps de dépassement de propagation dans les chemins du circuit. Parmi ces structures dites moniteurs, on trouve les chemins redondants et les moniteurs in-situ que nous allons détailler dans les sections 2.4.1 et 2.4.2

2.4.1 Oscillateur en anneau.

L'oscillateur en anneau, appelé aussi Ring Oscillator (RO), est un parmi les types de moniteurs in-situ. Il s'agit d'un chemin redondant composé d'un certain nombre pair d'inverseurs rebouclés [117]. Son délai total doit être égal au délai du chemin critique du circuit. La fréquence du fonctionnement du circuit est obtenue en mesurant directement la

fréquence du RO. Cette structure présente quelques inconvénients : elle ne peut couvrir seulement les variations globales [117], mais pas la variante locale. De plus, le RO et le chemin critique ne sont pas impactés de la même manière par les variations de type PVT[82], ceci se traduit par une dérive de mesure de la fréquence du circuit face à des PVTs différentes Quoique utilisé dans certains travaux et en vue des mesures expérimentales de délais obtenus avec une structure redondante (cf. paragraphe 2.6.1), nous n'allons pas retenir cette structure dans les circuits utilisés dans le cadre de cette thèse.

2.4.2 Moniteur de détection d'erreurs

Compte tenu des inconvénients des circuits redondants présentés dans la paragraphe 2.4, un circuit de détection d'erreur appelé Double-Sampling a été proposé par *Nicolaïdis* [118] [119]. Il constitue la base de tous les circuits résilients de détection d'erreur et de pré-erreur existant dans la littérature. Son principe consiste à détecter des valeurs erronées en observant les sorties des signaux de chaque étage du circuit synchrone (pipeline) à deux instants différents, d'où vient l'appellation Double-Sampling. Ceci est réalisable dans la figure 2.12 :

- On rajoute un élément redondant de type latch ou Flip-Flop (FF) à la sortie de chaque chemin combinatoire.
- On retarde la capture de la donnée (Out) de l'élément redondant à l'aide d'un élément de délai δ ajouté à son arbre d'horloge. Donc, la bascule de référence est pilotée par le signal CK et l'élément redondant par le signal CK+δ.
- On utilise un élément comparateur pour comparer les sorties de la bascule et l'élément redondant.
- On obtient un signal issu de la comparaison en rajoutant un élément de rétention (Latch) qui indique une erreur.

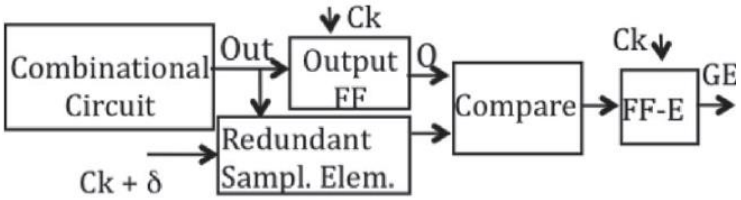


Figure 2.12 : Circuit de détection d'erreur 'Double-Sampling' [118]

Le 'Double-Sampling' peut être affecté par la métastabilité qui surgit lorsque les contraintes de setup ou de hold sont violées. Ainsi, la tension de sortie de la bascule de référence peut correspondre soit à '0' soit '1'. Il y a donc une probabilité non-nulle que ce niveau de tension soit mal interprété de façon hiérarchique par les pipelines ultérieures, ce qui peut induire des

erreurs non détectées. La solution est l'utilisation d'un détecteur de métastabilité, composé généralement d'une bascule FF ou un latch à la sortie du comparateur. Il y a des inconvénients de cette technique tels que les pénalités de surfaces et de consommation induites par le layout des éléments qui peuvent compromettre l'efficacité de la détection d'erreur de la technique Double-Sampling. Pour y faire face, en [118] [119], on se propose de rajouter un inverseur à la sortie de FF dont le délai est supérieur à celui de la zone de la métastabilité de la FF. Cette solution est efficace dans une application où on monitore les erreurs qui peuvent apparaitre au cours du temps. Cependant, dans une application où on doit réguler dynamiquement la tension/fréquence afin d'opérer autour du Point Of First Failure (POFF), c'est-à-dire le couple F(V) proche de la zone de dysfonctionnement du circuit, le taux d'erreurs peut devenir très grand (1 erreur tous les 10000 cycle [26]). Par conséquent, la probabilité d'observer la métastabilité est élevée. Donc, l'utilisation d'un détecteur de métastabilité est nécessaire dans ce cas.

Un autre inconvénient du Double-Sampling est l'utilisation d'une horloge retardée de δ pour l'élément retardé, cette horloge est obtenue en insérant un élément de délai dans l'arbre d'horloge. Ceci nécessite une implémentation de deux arbres d'horloges ce qui induit une augmentation du *Skew* entres les signaux délivrés par l'arbre du circuit et l'arbre des moniteurs.

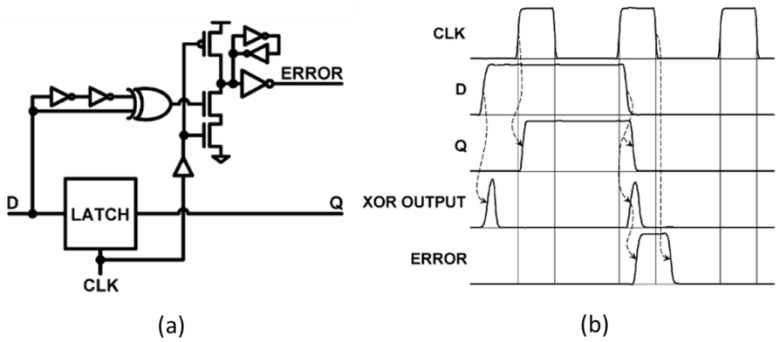


Figure 2.13 : (a) schéma de principe de TDTB et son diagramme de fonctionnement (b) [27].

Compte tenu des inconvénients de l'approche Double-Sampling, un détecteur de transition a été proposée en [27] , appelé TDTB (Transition Detector With Time Borrowing) (cf. figure 2.13), il est considéré comme immune à des problèmes de la métastabilité à condition d'imposer des contraintes de délai telles que le 'min_delay' et 'max_delay' pour maitriser le temps d'arrivé du signal à la bascule de capture. Son fonctionnement se résume comme suit :

- Lorsqu'une transition se produit à l'entrée D du latch et en avance du front d'horloge, un pulse est généré à la sortie du XOR. Ce pulse se propage à la sortie de la porte dynamique et maintient le signal ERROR à '0' (cf. figure 2.13 (b)).

- Lorsqu'une transition se produit près de l'entrée D du latch et près du front de d'horloge, un pulse est généré à la sortie du XOR. Ce pulse se propage à la sortie de la porte dynamique entrainant une mise '1' du signal ERROR (cf. figure 2.13 (b)). Lorsque CLK passe de '1' à '0', la sortie des portes dynamique se charge, ce qui met le signal ERROR '0'.

Une autre approche proposée par [26] [118] [119], appelé DSTB (Double-Sampling with Time Borrowing), exploite le fait que les délais des chemins du circuit ne dépassent pas la durée de la demi-période (niveau haut) du signal d'horloge. Ainsi, la donnée est stable dans les chemins de propagation durant cette phase d'horloge. Le schéma du circuit consiste à combiner les avantages du Double-Sampling et du TDTB. De ce fait, la FF de sortie est remplacé par un latch (cf. figure 2.14). Le latch, supposé stable dans le niveau haut de CLK, transmet sa valeur à chaque front montant d'horloge permettant au pipeline ultérieur de démarrer l'exécution. Ce circuit est considéré comme immunisé à la métastabilité car les erreurs, dont la durée ne dépassant pas le niveau haut de l'horloge, vont disparaitre avant le front descendant de l'horloge [118].

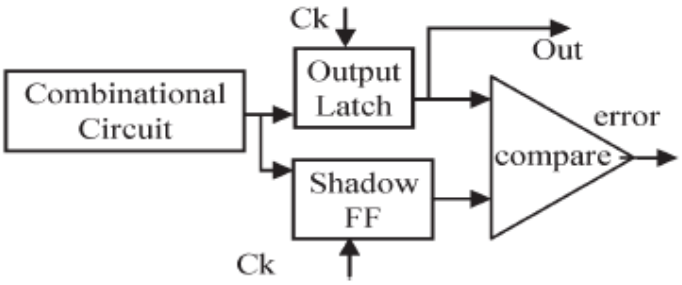


Figure 2.14 : Circuit Double Sampling with time Borrowing : la FF du Double Sampling est remplacée par un Latch [118].

Dans le prochain paragraphe, nous présentons quelques applications des moniteurs d'erreur. Ces applications constituent des circuits les plus connus et cités dans la littérature.

2.4.3 Utilisation des détecteurs d'erreurs

- **Razor I**

Le schéma du Double-Sampling a été utilisé [25] [26] [27] pour concevoir un circuit de détection et correction d'erreur. Il s'agit du circuit du Razor I. le schéma de principe est présenté dans la figure 2.15, Il est constitué d'un FF de référence, un latch et un détecteur de métastabilité. Afin d'éviter la modification du signal d'horloge et donc l'implémentation de deux arbres d'horloges, les auteurs ont proposé de capturer la donnée sur un front montant

d'horloge par la FF de référence et de la libérer sur un front descendant par le latch. Ainsi, le latch aura un temps additionnel d'une demi-période pour détecter et restaurer la donnée.
Lorsqu'une erreur se produit dans la demi-période (niveau bas de l'horloge), le signal de la sortie de la porte OR passe à '1' et la correction de l'erreur est assurée par le signal Restore qui instaure la donnée du latch dans le FF de référence.
Bien que le Razor I soit une approche fiable pour la détection et la correction d'erreurs, il présente de nombreux inconvénients tels que la génération et la propagation du signal Restore. En effet, le signal Restore est évalué à la sortie d'un arbre de portes OR à très grand fan-in, il doit être convenablement routé et bufférisé pour que la valeur correcte du signal soit instaurée dans le registre du pipeline, avant le prochain front montant de l'horloge. Ceci impose des contraintes de délai très importantes. D'autre part, la conception du détecteur de métastabilité s'avère complexe car il doit faire face à des FFs métastables dans les différents PVTs. L'utilisation d'un FF avec une grande capacité de conduire le courant (Drive) est donc fortement recommandée dans ce cas, ce qui peut impacter la surface et la consommation. Par ailleurs, le problème du signal Restore peut avoir des conséquences dramatiques car elle peut entrainer la propagation d'un mauvais signal dans les pipelines et donc induire des erreurs de fonctionnement.

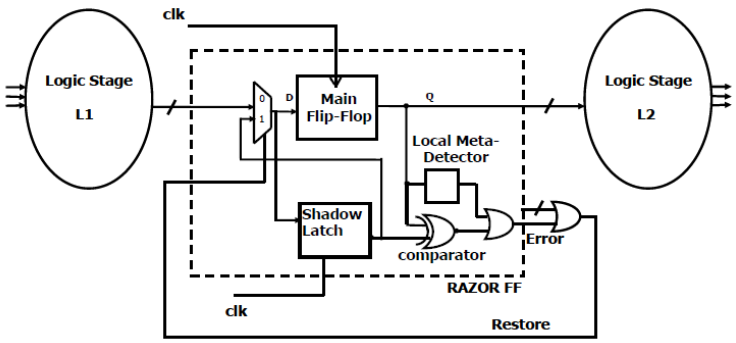


Figure 2.15 : Schéma de principe de l'architecture Razor I [25] [26]

Le Razor I présente aussi le problème des chemins courts. En effet, la nouvelle donnée, provenant du précédent Pipeline, capturée par la FF de référence peut être propagée à travers un chemin court et atteint le latch avant la fin de temps de son hold ce qui entraine des erreurs de violation de la contrainte de hold. Comme la donnée capturée par la registre de référence est libérée après un temps δ (le durée de demi-période dans ce cas), il faut contraindre les chemins courts de telle manière à ce que $\delta < D_{min} - t_{hold}$ (t_{hold} est le temps de hold du Latch et

D_{min} est le délai du chemin court). Cette contrainte implique l'insertion d'un grand nombre de buffers dans les chemins courts, ce qui impacte la surface et la consommation.

- **Razor II**

Compte tenu de tous ces inconvénients que nous avons présentés, une alternative au Razor I a été proposée en [120], il s'agit du Razor II (cf. figure 2.16). Son principe de fonctionnement est similaire à TDTB. La différence entre le Razor I et II est que le Razor II n'effectue que la détection des transitions afin d'éviter l'utilisation du détecteur de métastabilité ainsi que la complexité liée au mécanisme de correction (signal *Restore*) et donc une relaxation significative des contraintes de délai. La correction d'erreur, quant à elle, est assurée par une architecture de recouvrement d'erreur. Il s'agit d'une mémoire qui contient l'état des registres des pipelines en temps réels. Si une erreur se produit, les pipelines sont mis à '0' et la réexécution est recommencé depuis l'instruction qui a causé l'erreur. Les valeurs des registres avant l'apparition d'erreur sont instaurées depuis la mémoire avant le re-lancement de l'exécution. Cette opération est répétée autant de fois que les erreurs apparaissent avec la réduction de la tension.

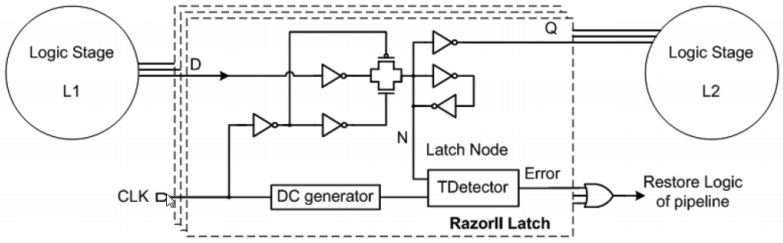


Figure 2.16 : Schéma de principe du Razor II [120]

2.4.4 Moniteurs de détection de pré-erreur

Afin d'éviter les limitations que présentent les techniques de détection d'erreurs, telles que la métastabilité, les contraintes de délai des chemins courts ou bien l'utilisation de l'architecture de correction d'erreur, le schéma de Double-Sampling a été légèrement modifié en rajoutant un élément de délai entre le registre de référence et l'élément redondant [121] [30][121]. Ainsi, l'élément redondant détecte toutes les transitions qui se produisent avant le front montant d'horloge et dans un intervalle de temps déterminé par le délai de l'élément du délai T_{pre}. Ces transitions ne vont pas conduire à des erreurs puisque la valeur mémorisée est toujours correcte, d'où l'appellation moniteur de pré-erreurs.

Le schéma de principe de fonctionnement du moniteur de pré-erreur et le diagramme des signaux sont illustrés dans la figure 2.17. Il est composé d'un FF, un élément de délai et un comparateur (une porte XOR). Le signal qui arrive à l'entrée D du registre original est retardé avec un temps noté $T_{pre.}$ Pour les transitions qui arrivent en dehors de T_{pre}, le FF du moniteur capture et libère le même signale que le FF de référence (cf. cycles 1,2 et 4 de la figure 2.17). Cependant, pour les transitions qui se produisent dans le laps de temps T_{pre}, la bascule du moniteur ne détecte pas la valeur correcte et donc la sortie du registre de référence Q et celle du moniteur Q_{pre} ne capturent pas la même valeur du signal. Cette différence est détectée par le comparateur et entraine une mise à '1' du signal à la sortie du XOR (Pre-Error) (cf. cycles 3 de la figure 2.17). Les sorties Pre-Error des moniteurs peuvent constitués les entrées d'un arbre de XOR afin d'avoir un signal de pré-erreur qui est considéré comme un indicateur pour effectuer AVS, DFVS ou ABB.

Il faut noter que l'intervalle dans lequel les transitions sont détectées est appelé fenêtre de détection, il est déterminé par le délai de l'élément du délai (T_{pre} dans notre cas).

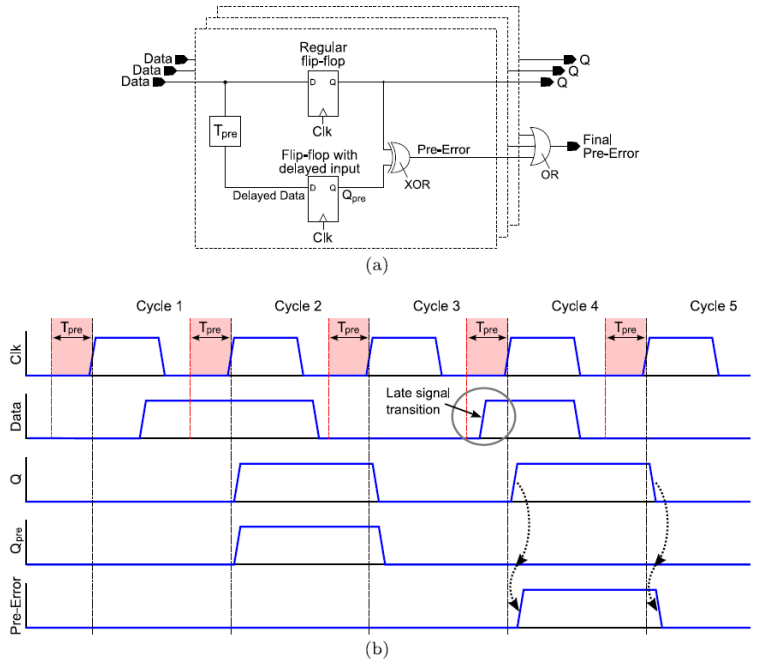


Figure 2.17 : Schéma de base du moniteur de pré-erreur (a) et son diagramme correspondant (b). Les transitions qui se produisent pendant l'intervalle T_{pre} sont détectées [30].

Le schéma de détection de pré-erreur que nous avons présenté détecte seulement les transitions dans la fenêtre de détection précise, ce schéma a été modifié en [121] afin de pouvoir détecter toutes les transitions dans le chemin de propagation à monitorer et donc

monitorer l'activité du chemin. L'approche est connue sous le nom de Crystall-ball. Son schéma de fonctionnement et son diagramme sont illustrés dans la figure 2.18.

Pour la détection de transitions dans la fenêtre de détection (partie basse du circuit), le principe est identique à celui de la figure 2.17. En dehors de cette zone, la détection est assurée par le bloc de la partie haute du circuit (Transition Detector). En effet, le comparateur XOR 2 compare le signal capturé à la sortie Q avec le signal à l'entrée du registre de référence : dans le cycle 1, la sortie Q est à l'état bas depuis la phase précédente de l'horloge. Si une transition se produit dans la fenêtre de détection, les deux signaux n'ont pas la même valeur et la sortie X2 du XOR passe à l'état '1'. La porte XOR reste dans cet état jusqu'à ce que le FF de référence recalcule une valeur au prochain coup d'horloge. Ceci entraine la génération d'une impulsion à la sortie X2.

Nous remarquons dans la figure 2.18 (b), que la largeur de l'impulsion dépend des instants des transitions. Ce signal doit être stable avant le temps de setup du FF du détecteur de transition. Sinon, il ne va pas être correctement capturé par le FF et par conséquent l'observation de l'activité dans le chemin va être erronée, ce qui conduira à une mauvaise décision de choix des FFs à équiper avec un moniteur.

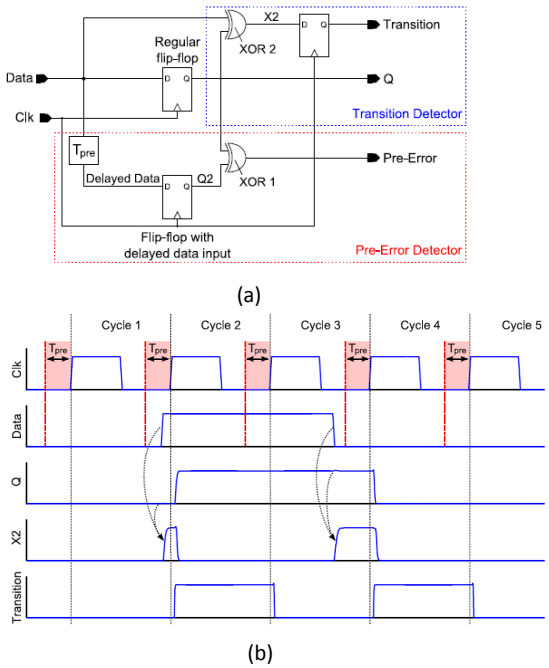


Figure 2.18 : (a) Schéma de principe du Crystall-ball et son diagramme correspondant (b) [121]

Les moniteurs que nous avons présentés se basent sur l'élément du délai qui retarde le signal d'une part et constituent la fenêtre de détection d'autre part. Il existe une autre version de

moniteur basée sur le rapport du cycle de d'horloge [125] [124][25]. En effet, la fenêtre de détection est constitué par la demi-période du signal d'horloge et non pas par l'élément de délai. Le fonctionnement du moniteur, dit statique, (cf. figure 2.19 (a)) est identique au Crytall-ball. Pour le moniteur de type dynamique (cf. figure 2.19(b)) [124], une modulation de la fenêtre de détection est faite en changeant le rapport cyclique de l'horloge [123], soit avec une horloge multi-phase soit avec un PWLC (Pulse-With -Control-Loop) [125]. Ainsi, la détection des transitions de nature rapides et lentes peut s'effectuer séparément.

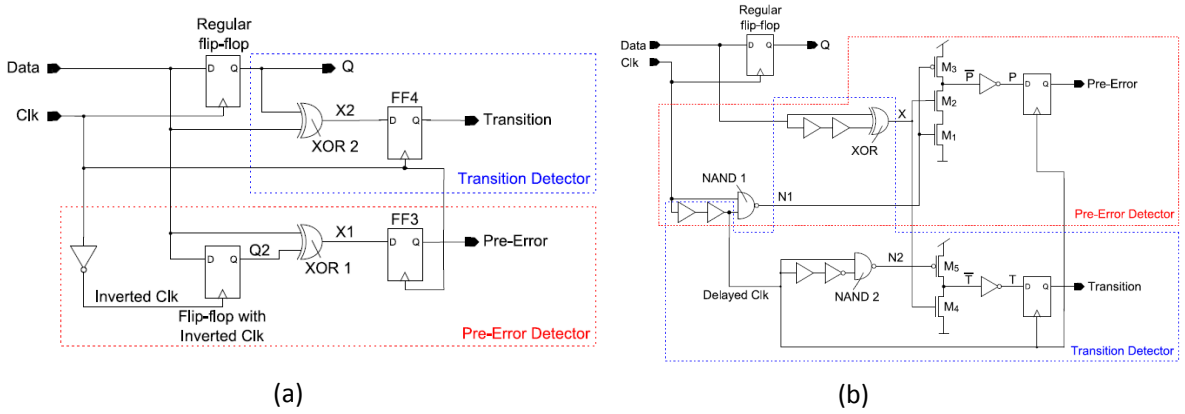


Figure 2.19 : moniteur basé sur le rapport cyclique : (a) statique [125], (b) dynamique [124]

2.4.5 Comparaison des moniteurs de pré-erreur et des moniteurs d'erreur

Les moniteurs de pré-erreur détaillés dans ce paragraphe présentent un avantage significatif qui est la détection des pré-erreurs avant l'apparition des 'vrais' erreurs dans les bascules. Ainsi, nous évitons l'utilisation d'une architecture de correction d'erreur. De plus, cette approche peut être utilisée pour monitorer les dégradations qui agissent sur le long terme, telles que celles produites par le vieillissement. En effet, les chemins dont le délai se dégrade à cause de vieillissement, et dont le slack est inférieur à la fenêtre de détection, seront détectés à l'aide des alertes des moniteurs.

Par ailleurs, les moniteurs d'erreur [25] [26] [27] ont l'avantage de pouvoir fonctionner autour du Point Of First Failure (POFF). Ainsi, le circuit opère dans des zones de tension/fréquence très réduites, ce qui permet de réduire significativement la consommation. Les auteurs des articles [25] [26] [27] ont démontré la capacité des moniteurs d'erreur de fonctionner très proche du POFF et ont reporté une réduction de consommation de 52% dans une architecture ARM fonctionnant à 1Ghz.

2.5 Techniques de masquage d'erreur

Afin de pouvoir fonctionner autour du POFF sans l'utilisation d'une architecture de correction d'erreur, une technique dite de masquage d'erreur est proposée dans la littérature. Son principe consiste à masquer les erreurs en faisant suivre un cycle d'horloge plus tard la valeur correcte du signal dans le pipeline (Time Borrowing). La motivation s'explique par le fait que très peu de bascules constituent à la fois des registres de lancement (Startpoint) et de capture des chemins critiques. Ainsi, les erreurs qui affectent seulement un étage du Pipeline, peuvent être masquées en injectant la valeur correcte dans l'étage de Pipeline suivant.

Les circuits permettant le masquage d'erreur sont classés en deux catégories : logiques et temporelles. Le masquage logique [142] utilise une logique redondante et un délai pour évaluer la valeur à la sortie des registres avec un retard exercé sur le chemin (similaire au Double-Sampling). Le masquage temporel quant à lui masque les erreurs en retardant l'arrivée de la donnée correcte au pipeline suivant. Dans le contexte du masquage temporel, une approche a été proposée qui consiste à bloquer l'horloge d'un cycle après avoir détecté une erreur et de la corriger pendant le même cycle [143][127]. Les auteurs assument que la latence pour corriger l'erreur dure moins d'un cycle d'horloge dans le pire cas avant que cette dernière ne soit propagée dans les pipelines. Cependant, cette pratique peut être difficile à achever dans les circuits complexes en technologie avancées car d'une part, le temps de cycle d'horloge est très court et d'autre part la latence pour instaurer la valeur correcte du signal dans le pipeline peut être longue. Donc il y a un risque de propager des valeurs erronées dans les registres.

Le tableau 2.2 résument les techniques discutées dans ce chapitre.

Caractéristiques	Détection d'erreur	Prédiction d'erreur	masquage d'erreur	
			logique	temporelle
Mécanisme de détection	FF/ Latch dupliqué [25] détection de transition	FF/ Latch dupliqué [121]	Logique redondante [142]	FF/ Latch dupliqué [143] détection de transition [144]
Quand (par rapport au front d'horloge)	après	avant	-	après
Architecture de recouvrement	circuit de correction d'erreur	-	-	-
Impact sur l'arbre d'horloge	oui	oui	non	oui
Impact des chemins courts	oui	non	non	oui
Logique séquentielle supplémentaire	large	large	-	large
Logique combinatoire supplémentaire	peu	peu	modéré	peu
Recouvrement des délais	total	partiel	total	total
Technique	Razor [25] TDTB [27] et DSTB [25]	Canary [121] TRC [146], CPR [127]	circuit dédié [142]	Timber [145] PEDFF [143] DCFF [144]

Tableau 2.2 : comparaison des techniques de détection d'erreurs, de pré-erreur et du masque d'erreur

Compte tenu des différents moniteurs de délai que nous avons présentés, notre choix s'est porté sur le moniteur de pré-erreur. Le prochain paragraphe sera dédié à la présentation des résultats expérimentaux, concernant la caractérisation des différents moniteurs de pré-erreur.

2.6 Mesure des performances avec des moniteurs in-situ

Nous avons évoqué dans le paragraphe 2.4.1 le fait que les structures redondantes ne permettent pas de détecter toutes les variabilités. Nous avons vérifié cette propriété. En effet, nous avons implémenté sur silicium un circuit redondant, appelé réplica de chemin critique CPR (Critical Path Replica), d'un circuit numérique Bose-ray Choudhary and Hoquenheim (BCH) qui sera détaillé dans les prochains chapitres. Nous rappelons que le principe du CPR consiste à dupliquer le layout du chemin critique, portes logiques et interconnexions afin d'avoir un comportement fonctionnel identique au chemin critique. Le but de cette manipulation est de réaliser une mesure comparative de la fréquence maximale d'un circuit numérique et celle donnée par le CPR.

Le schéma de principe du CPR utilisé est illustré dans la figure 2.20. Il est composé des éléments du chemin critique du circuit numérique. En plus de ces éléments, nous avons ajouté un comparateur pour comparer les sorties du chemin critiques et du CPR.

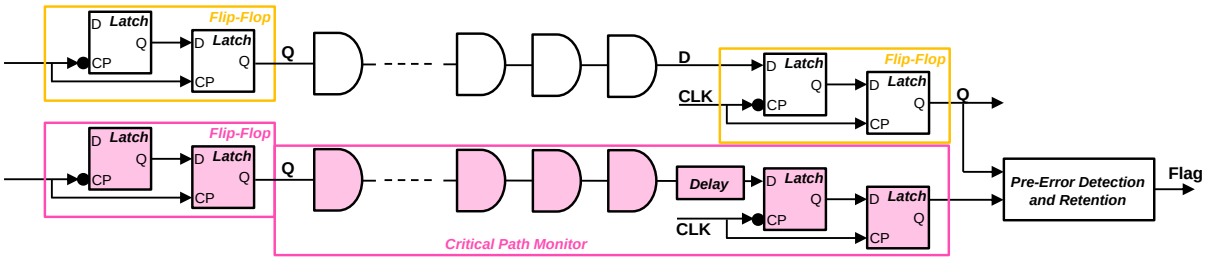


Figure 2.20 : Illustration du CPR implémenté dans le circuit numérique. La logique du chemin critique est dupliquée.

Pour pouvoir comparer la vitesse du circuit et celle du CPR, nous avons mesuré la tension minimale V_{DDmin} du CPR à la fréquence du sign-off, soit 500MHz et l'avons comparé à celle du circuit numérique BCHs dans 3 plaques de silicium, soit une comparaison sur près de 1500 circuits. Le résultat est donné dans la figure 2.21 et il montre une dispersion de mesures des V_{DDmin} du circuit numérique, principalement à cause du phénomène de la variabilité. Nous remarquons aussi un écart entre les mesures des V_{DDmin} du CPRs en le comparant à celui du

circuit BCH. Ils différent de 20% en moyenne. Ceci s'explique par le fait que le vecteur de test injecté n'a pas activé en priorité le chemin dupliqué. De plus, la figure 2.21 montre aussi que les tensions mesurées sont très supérieurs à celle du sign-off. Ceci laisse une très grande marge de compensation en tension ou en fréquence pour chaque puce en utilisant les techniques adaptatives.

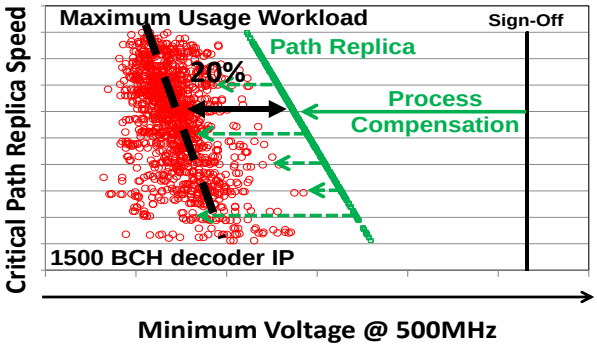


Figure 2.21 : Mesure des fréquences maximales du BCH et du CPR

Compte tenu de ces résultats, nous démontrons que la structure CPR n'est pas adéquate car elle ne permet pas de mesurer avec précision la vitesse du circuit. D'où la nécessité de monitorer le délai de chaque chemin du circuit à l'aide des moniteurs in-situ. Ces moniteurs feront l'objet de la caractérisation dans la suite du document.

2.7 Caractérisation des moniteurs de pré-erreur

Nous nous proposons maintenant de caractériser trois types de moniteurs in-situ :

- Moniteur in-situ, dont la fenêtre de détection est basée sur un élément de délai (figure 2.22 (a)),
- Moniteur in-situ, dont la fenêtre de détection est basée sur un élément passif (figure 2.22 (b))
- Moniteur in-situ composé de Latch (figure 2.22 (c)), appelé Master Delay. La fenêtre de détection est basée sur le délai du maitre de la bascule de référence

Ces moniteurs ont été implémentés dans une barrette de silicium en 28FDSOI.

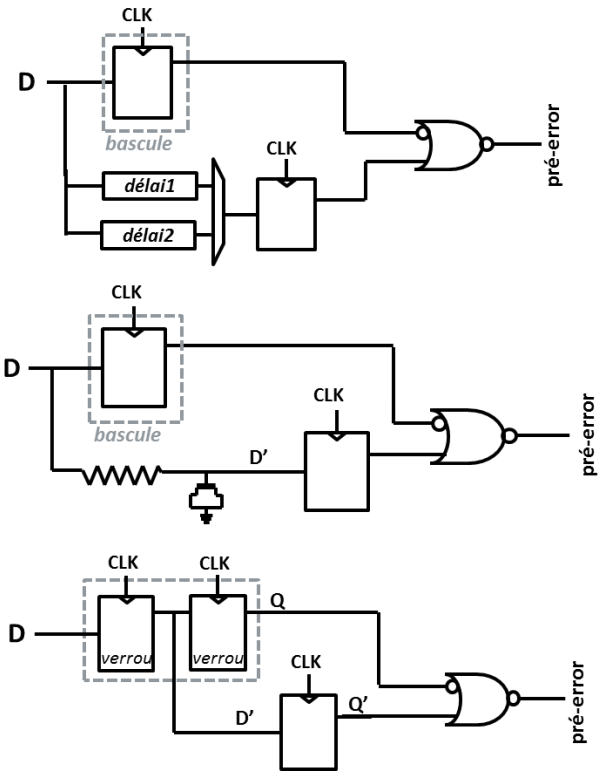


Figure 2.22 : (a) Schéma du moniteur in-situ basé sur un élément de délai à partir de buffers. Deux valeurs de délais, obtenues par deux buffers, sont ici proposées : délai 1= 70ps et délai 2 = 100ps à 1V. (b) schéma du moniteur in-situ basé sur un élément de délai à partir d'un élément passif. (c) Schéma du moniteur Master Delay basé sur le délai du maitre.

Pour ces moniteurs, nous avons mesuré la fenêtre de détection expérimentalement en faisant varier la tension de V_{DD} de 0.8V à 1V jusqu'à ce que le signal de pré-erreur se lève.

Un échantillon de 192 puces d'une même plaque a été testé. L'évolution de la fenêtre de détection avec la tension d'alimentation V_{DD} est mesurée à 25°C [81] (cf. figure 2.23). Des simulations ont également été réalisées (tracées en lignes noires sur la figure 2.21) dans le même corner que la plaque. Ces simulations sont en adéquation avec les mesures. Les longueurs des fenêtres de détection des moniteurs considérés varient entre 28ps et 48ps à la tension nominale (V_{DD}=1V). Le moniteur avec l'élément du délai variable présente la plus forte dépendance en tension et celui avec l'élément de délai passif la plus faible : pour respectivement une déviation de la fenêtre de détection égale à 46ps et 11ps lorsque V_{DD}=0.7V. En effet, la longueur de la fenêtre de détection dépend directement de l'élément de délai lui-même. Or, dans le cas d'un élément passif, la grandeur de celui-ci ne varie pas avec la tension d'alimentation. De ce fait la fenêtre de détection dépend très peu de la tension d'alimentation, contrairement au cas des délais 1&2 dont les temps de propagation sont directement liés à la tension d'alimentation.

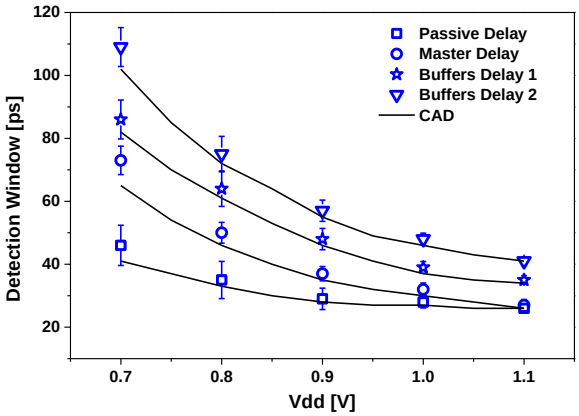


Figure 2.23 : Mesures (symboles bleus) et simulations (lignes noires) des fenêtres de détection des différents moniteurs in-situ considérés pour différentes tensions d'alimentation VDD à 25°C.

2.7.1 La variabilité de la fenêtre de détection

Nous nous intéressons maintenant à l'impact de la variabilité sur les distributions des fenêtres de détection pour illustrer les effets de variations des PVTs. La figure 2.24 et 2.25 [81] présentent les distributions des fenêtres de détection du moniteur Master Delay et celui basé sur l'élément du délai, pour différentes tensions d'alimentation V_{DD}. Les distributions en bleu représentent seulement la variabilité des moniteurs alors que les distributions en magenta montrent la variabilité des moniteurs cumulée avec celle des éléments logiques du chemin.

Sur la figure 2.24 qui concerne les moniteurs in-situ basés sur le cycle d'horloge, la variabilité induite par le chemin de délai peut être observée pour des tensions d'alimentation V_{DD} inférieures à la tension d'alimentation nominale (de 1V). La fenêtre de détection et l'effet de la variabilité locale croit de manière significative avec la diminution de la tension d'alimentation V_{DD}.

Les moniteurs à base du délai des buffers sont considérés ici comme le pire cas, car les buffers utilisés pour constituer l'élément de délai contribuent à augmenter l'impact de la variabilité et ainsi la dispersion des distributions pour les deux types de moniteurs in-situ. Des simulations Monte Carlo (symboles noirs) ont été réalisées avec le simulateur Eldo pour reproduire la dispersion liée à la variabilité pour les tensions d'alimentation V_{DD} de 0.7V et 1V à 25°C. Effectivement, les simulations Monte Carlo permettent de reproduire la dispersion des distributions liée à la variabilité, en particulier dans le cas du moniteur Master Delay. Dans le cas du moniteur basé sur le délai des buffers (cf. figure 2.25), la dispersion est sous-estimée avec les simulations Monte Carlo. En effet, les mesures donnent une déviation standard de la fenêtre de détection de 6.18ps pour V_{DD} = 0.7V alors que les simulations Monte Carlo donnent 4.9ps et respectivement 1.87ps et 1.52ps pour V_{DD} = 1V et V_{DD} = 1.1V.

Bien que le moniteur in-situ avec l'élément de délai passif présente la plus petite fenêtre de détection et la plus faible dépendance en tension, il ne pourra pas être utilisé ni intégré dans des circuits numériques. Donc, la solution la plus approprié restera celle du moniteur basé sur l'élément du délai.

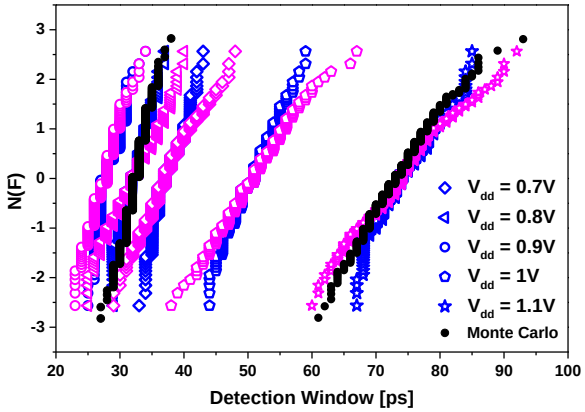


Figure 2.24 : Distributions des fenêtres de détection du moniteur in-situ basé sur le délai du cycle d'horloge (Master Delay) pour différentes tensions d'alimentation à 25°C en considérant soit la variabilité du moniteur (en bleu) soit la variabilité du moniteur in-situ et du chemin (en magenta). Les simulations Monte Carlo réalisées avec Eldo sont représentées par des symboles noirs.

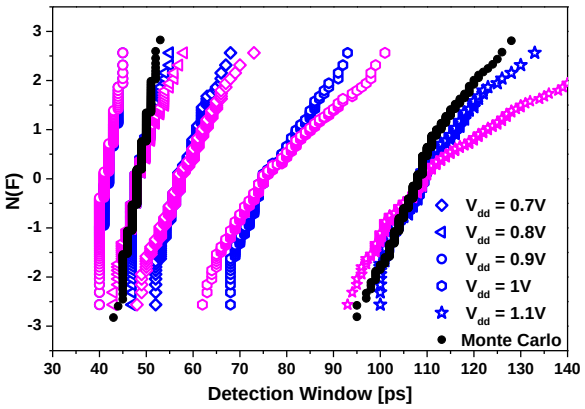


Figure 2.25 : Distributions des fenêtres de détection du moniteur in-situ basé sur l'élément du délai de buffers pour différentes tensions d'alimentation à 25°C en considérant soit la variabilité du moniteur (en bleu) soit la variabilité du moniteur in-situ et du chemin de délai (en magenta). Les simulations Monte Carlo réalisées avec Eldo sont représentées par des symboles noirs.

2.8 Conclusion

Dans ce chapitre, nous avons évalué l'impact des PVTs sur les paramètres de fonctionnement des circuits numériques avec la méthode MMMC et du vieillissement à l'aide d'un flot de simulation du vieillissement.

L'impact du vieillissement est obtenu en appliquant un signal de stress sur une entrée de porte logique puis en caractérisant cette porte dans des conditions de type PVTs données. Nous avons appliqué ce flot sur des inverseurs en considérant différentes transitions S_{in} et capacité de charges C_{load}. Un modèle relativement simple établi dans ces conditions s'avère très imprécis car il ne permet pas d'évaluer la dégradation dans toutes les situations. Par conséquent, le délai induit par le vieillissement sera calculé avec une seule valeur de S_{in} et C_{load}.

Nous avons appliqué la méthode MMMC et le flot du vieillissement sur des chemins d'un circuit numérique afin d'évaluer l'impact des PVTs. Les résultats ont montré que la variation de la tension induit plus d'impact sur délai, ensuite le procédé, la température, le vieillissement puis la variation locale. Les marges induites par ces variations peuvent être compensées par des méthodes adaptatives, à l'aide des circuits tels que les CPRs et les moniteurs in-situ car ils permettent de mesurer les délais du circuit en temps réel.

Nous avons d'abord implémenté les CPRs dans le BCH afin de mesurer la tension minimale de fonctionnement. Les résultats ont montré une imprécision de mesure du F_{max} par le CPR qui est de 20% en moyenne par rapport au F_{max} du circuit. Nous avons donc opté pour les moniteurs in-situ, ces circuits se présentent sur deux formes : les moniteurs de détection d'erreurs et de pré-erreurs.

Les moniteurs de détection d'erreurs permettent de détecter des erreurs ou des transitions qui surviennent après le front montant d'horloge. L'inconvénient de ce schéma est la propagation éventuelle des signaux métastables ainsi que la nécessité d'utiliser des circuits de recouvrement, ce qui induit des pénalités de surface.

Compte tenu de ces inconvénients, nous retenons dans la suite des études le moniteur de pré-erreur qui consiste à détecter des transitions dans une fenêtre de temps juste avant le front d'horloge. Cette fenêtre de temps, aussi appelée fenêtre de détection, est obtenue soit avec des éléments des délais, soit avec le délai du maitre, soit avec un élément passif. Nous avons caractérisé ces trois types de moniteurs notamment la fenêtre de détection, les résultats ont montré que le moniteur basé sur l'élément passif a une faible dépendance à la tension et présente moins de variabilité mais il n'est pas adéquat pour les applications numériques. Pour les moniteurs basés sur l'élément du délai et le cycle d'horloge, la fenêtre de détection dépend de la tension V_{DD} et présente une forte variabilité à faible tension.

CHAPITRE 3 : STRATEGIES DE SELECTION ET D'INTEGRATION DES MONITEURS IN-SITU

3.1 Introduction

Les variations des procédés de fabrication, de la tension, et de température (PVT) ainsi que les mécanismes de vieillissements tels que le BTI, HCI et TDDB doivent être évaluées avec précision aux étapes de la conception afin de garantir à la fois la fonctionnalité du circuit en post-fabrication et pendant toute la vie du produit. Ces variations impactent les délais différemment car elles dépendent de plusieurs conditions, notamment des modes d'opération et des corners de fabrication.

L'utilisation des moniteurs in-situ (ISM) est une solution prometteuse pour monitorer le dépassement des délais induits par ces variations lors de l'utilisation du circuit car ils émettent des alertes dans le cas où des transitions critiques se produisent dans la fenêtre de détection. Ainsi, nous pouvons prévoir le temps restant avant l'apparition d'une erreur, qui est une violation du temps setup. Toutefois, leurs intégrations nécessitent de mener une étude pour sélectionner les chemins qui doivent être monitorer afin de limiter les pénalités de surface et de consommation engendrées par le rajout des éléments des moniteurs in-situ tels que la bascule et le buffer.

Dans ce chapitre, nous allons présenter quelques méthodologies de sélection des chemins critiques parues dans la littérature. Ensuite, nous allons présenter le flot de conception des circuits digitaux de STMicroelectronics et nous évaluerons l'impact de l'intégration des moniteurs in-situ (puissance, surface, délai…) dans les différentes étapes du flot. Nous mettons en œuvre une version optimisée d'un moniteur qui engendre moins de pénalité ainsi qu'une méthodologie qui maximise le taux de couverture des chemins critiques par des moniteurs.

Aussi, nous allons proposer plusieurs approches de sélection des chemins critiques se basant sur :

- La sélection des pires chemins issus de l'analyse statique temporelle des délais STA (Static Timing Analysis) dans le pire cas de fonctionnement représenté par P, V, T.

- La sélection des registres se trouvant en bout des plus lents chemins sous-critiques.
- La sélection des chemins qui présentent les plus longs délais après la prise en compte de la dégradation induite par le vieillissement.

3.2 Revue des méthodes de sélection des chemins critiques

L'identification des chemins de propagation potentiellement critiques dans la phase de la conception est un sujet d'étude de la communauté scientifique depuis des nombreuses années. Une première approche proposée par [128] [129] consiste à identifier les chemins testables qui traversent le maximum de portes logiques à l'aide de la méthode de la réduction des graphes. Cette approche ne considère que les délais des portes dans les conditions nominales de fonctionnement. Le même principe a été utilisé par [130] [131] mais en considérant l'impact des procédés sur le délai des portes logiques.

Lai et al [132] propose de monitorer les chemins qui se situent dans une fenêtre de détection dite Opportunism Window (cf. figure 3.1 (a)). La taille de cette fenêtre de détection est définie par la différence des délais du pire chemin dans le pire corner (SS) et le pire chemin dans le corner typique (TT). Nous rappelons que SS et TT désignent le centrage du procédé du transistor : SS signifie procédé lent pour le NMOS et le PMOS. TT signifie procédé rapide pour le NMOS et le PMOS.

Ce nombre de chemins de propagation peut être important et donc une optimisation a été proposée en appliquant le critère suivant : pour chaque chemin de propagation *i* se situant dans la fenêtre de détection, les délais des chemins *j*, qui passent à travers les portes du chemin *i*, doivent être inférieurs ou égales au délai du chemin *i*.

La figure 3.1 (b) montre une illustration. Les chemins A et B présentés dans la figure se trouvent dans la fenêtre de détection, le chemin B passe à travers la porte G4 du chemin A et son délai est inférieur à celui du chemin A. D'après le critère, le chemin B doit être exclu de la sélection.

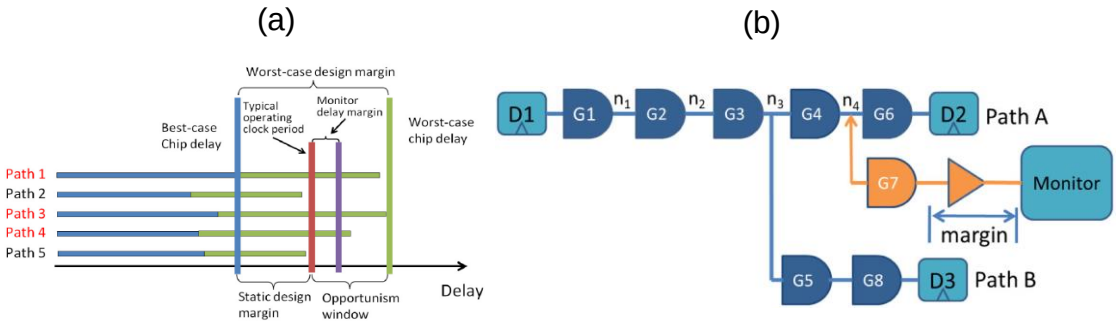


Figure 3.1 : Méthodologie de sélection des chemins critiques proposée par [132], les chemins se situant dans la fenêtre de détection 'Opportunism Window' doivent être monitorés(a). (b) Illustration du critère d'optimisation de la sélection des chemins.

Une approche similaire a été proposée [133] [134], i.e. choix des chemins critiques qui arrivent dans la fenêtre de détection délimitée par les délais des pires chemins dans les corners FF et SS. La différence est que cette dernière tient compte aussi des délais induits par le vieillissement en plus de la variation des procédés. En effet, les chemins, dont les délais après vieillissement se situent dans la fenêtre de détection, doivent aussi être monitorés. Ce vieillissement est pris en considération en rajoutant 20% de délai supplémentaire à chaque chemin. Les auteurs proposent aussi une optimisation de la sélection en se basant sur la dispersion locale des chemins qui peut être évaluée par l'analyse SSTA (Statistical Static Timing analysis). Les chemins qui présentent un recouvrement des délais à 3 σ sous l'effet de la variation locale peuvent être filtrés. La figure 3.2(a) illustre un exemple de deux chemins qui présentent un recouvrement de délai. Dans ce cas, un des deux chemins doit être sélectionné pour qu'il soit monitoré. La figure 3.2 (b) montre que les distributions des délais des deux chemins ne présentent pas de recouvrement de délai et donc ils doivent être considérés tous les deux.

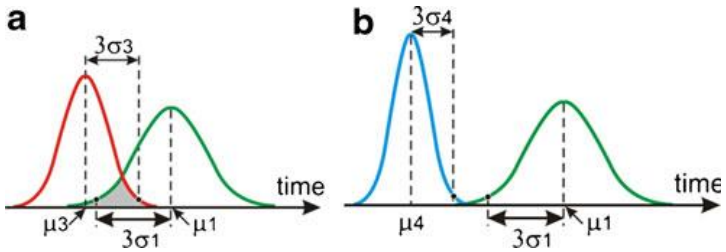


Figure 3.2 : Distribution statistique des délais de deux chemins en utilisant la SSTA (a) cas de recouvrement des délais à 3σ. (b) pas de recouvrement de délai

Une autre méthode qui tient compte de la criticité des délais et l'activité des chemins a été proposée en [135]. Les chemins à monitorer sont initialement choisis par leurs délais dans les

conditions nominales et dans une fenêtre de détection ΔT (cf. figure 3.3 (a)). Ces chemins sont dénotés PP (Potential Path). Ensuite, la probabilité d'avoir une transition critique, au niveau des bascules de capture des chemins se situant dans ΔT, est calculée à l'aide d'un vecteur de test et une simulation fonctionnelle RTL (cf. figure 3.3 (b)). Cette probabilité désigne le taux d'activité des chemins.

Afin de tenir compte des PVTs, une fonction dite 'poids' définit par l'équation 3.1 est appliqué sur chaque chemin de la fenêtre de détection ΔT (cf. figure 3.2) et les chemins retenus sont ceux qui présentent une valeur importante de $W_{m,PP}$ c'est-à-dire un taux d'activité élevé et des délais critiques dans les PVTs.

$$W_{m,PP} = \sum i * t_{d,norm,i} \qquad \text{EQU 3.1}$$

$i \in \{P, V, T, A, SA, D\}$ est un facteur pour considérer les critères du design (process, tension, température, vieillissement, activité, délai). $t_{d,norm,i}$ est le délai normalisé dans un PVT donné.

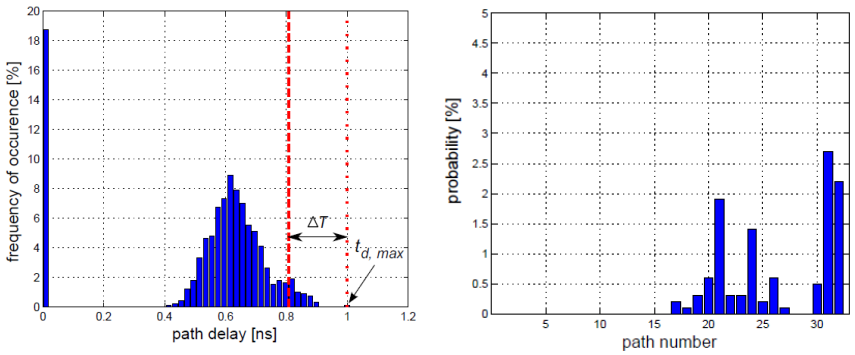


Figure 3.3 : Illustration de la méthodologie de sélection proposée en [135]. (a) Les chemins sélectionnés dans la fenêtre de détection ΔT. (b) La probabilité d'observer des transitions critiques des chemins sélectionnés dans la fenêtre ΔT.

Les méthodes de sélection des chemins critiques que nous avons citées, présentent quelques inconvénients. En effet, les méthodologies proposées dans [132] [133] [134] ne tiennent pas compte de l'activité des chemins. Ainsi, l'application des algorithmes de réduction du nombre de moniteurs peuvent entrainer l'élimination des chemins à forte activités. Concernant la méthodologie proposée dans [135], le poids qui détermine la criticité du chemin est calculé en tenant compte de tous les types de variations simultanément. Ceci est imprécis car les variations de type PVTs, vieillissement et activité n'ont pas le même impact sur le délai, et doivent être évaluées séparément.

Compte tenu de ces constatations, nous nous proposons une méthode de sélection des chemins critiques, qui combine les avantages des méthodologies de l'état de l'art et qui limitent leurs inconvénients. En effet, nous allons attribuer à chaque bascule du circuit BCH un poids qui détermine sa criticité. Ce poids dépend des délais de tous les chemins sous-critiques qui arrivent sur la bascule et de l'activité. Ensuite, nous allons comparer cette méthodologie à une méthode de sélection des chemins dans une fenêtre de temps.

3.3 Méthodologie d'intégration des moniteurs in-situ

Avant de présenter la méthodologie d'intégration des ISMs, nous proposons d'abord de détailler le flot de conception des circuits digitaux utilisé au sein de STMicroelectronics. Les moniteurs in-situ sont intégrés à l'aide de ce flot.

3.3.1 Flot de conception des circuits numériques

Le flot d'implémentation des circuits digitaux de STMicroelectronics est inspiré du flot classique de la conception et donné par la figure 3.4**.** Il commence par l'ensemble des étapes appelées Front-End, il s'agit en effet de l'implémentation du code (RTL) décrivant le comportement HDL (Hardware Description Language) du circuit. Le circuit est ensuite synthétisé en utilisant un outil de synthèse (Design Compiler de Synopsys dans notre cas) qui prendra en compte les cellules standards d'une technologie propriété de STMicroelectronics ainsi que les contraintes de délai, de surface, de puissance.

Il peut y avoir d'autres contraintes selon le produit ou les attentes du client. Dans cette thèse, nous allons utiliser principalement ces contraintes :

- La période et la latence des horloges du circuit.
- Budget des délais des modules hiérarchiques entre les entrées/sorties.
- Délai des transitions sur les entrées des portes et la capacité maximale sur leurs sorties.

Le flot s'enchaine par les étapes de l'implémentation physique dites Back-End. La première étape de ce flot est dite Floorplaning, elle consiste à fixer la taille de circuit et à mettre en place la grille d'alimentation générale (V_{DD}, GND, vdds, gnds) et celle de différentes macros personnalisées (Mémoires, blocs analogiques…). Ensuite, les portes logiques contenues dans

la netlist issue de la synthèse sont physiquement placées et l'arbre d'horloge CTS (Clock Tree Synthesis) est généré. A ce stade, une STA peut être effectuée en prenant compte des corners car les délais de propagation du signal dans les fils et des portes logiques sont connus. L'étape dite CTS est ensuite optimisée, il s'agit de l'optimisation de tous les délais de setups, holds et les transitions sur les entrées des portes logiques et les bascules. Il s'agit du bloc « Optim CTS » de la figure 3.4.

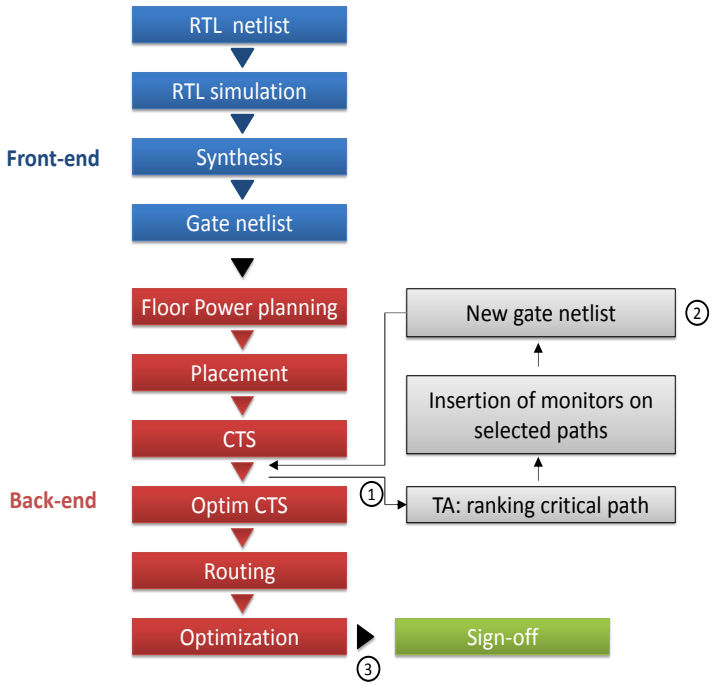


Figure 3.4 Flot d'insertion des moniteurs in-situ. Les moniteurs sont placés dans les chemins critiques repérés dans le Back-end après CTS. Le flot s'achève par le routage et l'optimisation du routage.

Le flot continue avec l'étape du routage, i.e. la connexion physique des portes logiques. La particularité du routage dans cette étape est qu'il respecte tous les contraintes physiques de la technologie, i.e. espacements entre les niveaux de métaux. Par conséquent, les délais peuvent subir une dégradation, car la forme et la longueur des fils est modifiée, ce qui induit des nouvelles valeurs des parasites RC et du délai. Dans cette nouvelle configuration, une ré-optimisation des délais, telle qu'en post-CTS, est effectuée. Il s'agit de la dernière étape du flot d'implémentation physique qui est l'optimisation du routage.

Après le parcours du flot, une étape dite ' sign-off ' consiste à vérifier les règles suivantes :

- Les règles de la technologie DRC (Design Rule Check) c'est-à-dire les distances des fils, les vias, les niveaux de métaux et le respect des règles de fabrication formalisées dans l'ensemble des règles DRC.
- Les courts-circuits et les circuits ouverts par la vérification LVS (Layout VS Schematic)
- Les délais des chemins dans tous les modes de fonctionnement et les corners afin de vérifier les violations du setup et du hold.
- Le couplage entre les fils (Cross Talk) qui peut induire des modifications des valeurs logiques des signaux
- La consommation statique et dynamique du circuit ainsi que la chute de tension brutale dans les rails d'alimentation connue par IR-Drop.
- Vérification des règles de l'électro-migration
- La preuve formelle, c'est-à-dire la comparaison entre les netlist après synthèse et placement-routage afin de s'assurer qu'il n'y a aucune modification de la logique de fonctionnement du circuit.

3.3.2 Conception des moniteurs in-situ

Dans ce paragraphe, nous présentons les stratégies de conception du layout des moniteurs que nous avons implémentés. En effet, nous avons confronté deux méthodes pour la réalisation des layouts des moniteurs in-situ présentés en chapitre 2:

- **ISM layout A**

La première solution dite layout A (LA) (cf. figure 3.5) consiste à embarquer la bascule de référence et les éléments du moniteur (bascule, buffer et le bloc de rétention) dans une cellule standard. Nombreuses vues sont développées pour rendre la cellule standard intégrable dans le flot digital de STMicroelectronics. Ce layout présente l'avantage d'avoir une distance optimisée entre la bascule à monitorer et celle du moniteur, ce qui nous évite les longs fils de connexions. Néanmoins, ce layout devrait être développé et caractérisée en prenant en compte chaque type de bascule utilisé dans le design : bascule, verrou (Latch) avec et sans pin de test, et pour différentes capacités à conduire le courant (Drive) disponible dans la plateforme de conception.

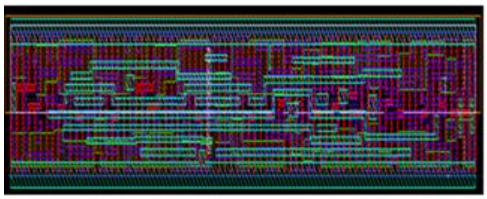


Figure 3.5: moniteur de type layout A LA. La bascule de référence ainsi que les éléments du moniteur sont intégrés dans une cellule standard.

- **ISM Layout B**

La deuxième solution du layout dite layout B (LB) (cf. figure 3.6) consiste à construire le layout du moniteur à partir des cellules standard de la librairie sans avoir à tout caractériser. Le placement des éléments du moniteur et les connexions se font à l'aide des commandes de l'outil de placement-routage. La proximité entre le moniteur et la bascule de référence se gère soit avec un placement personnalisé en spécifiant les coordonnées précises des cellules du moniteur soit en imposant des contraintes de délai de type max/min délai vers la bascule du moniteur. Cette solution est très facilement intégrable dans un layout complexe et offre un degré de liberté supplémentaire qui est de moduler la largeur de la fenêtre de détection en modifiant le nombre de buffers et/ou leurs dimensions. Toutefois, le plus grand désavantage est qu'elle demande une attention particulière pour connecter le moniteur à l'arbre d'horloge : une ré-optimisation est nécessaire pour équilibrer l'arbre d'horloge sur toutes les bascules y compris celles des moniteurs. Si la bascule du moniteur se trouve loin de la cellule de référence, la résolution de la détection sera dégradée. Donc c'est un point à contrôler dans le placement.

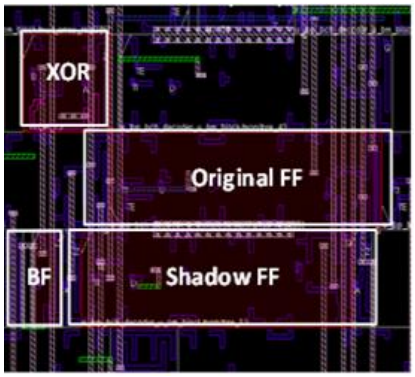


Figure 3.6 : moniteur in-situ basé sur le layout de type LB, il est construit à partir des portes logiques de la bibliothèque. Le placement des éléments du moniteur est effectué avec les commandes de l'outil de placement-routage.

3.3.3 Bilan d'intégration des moniteurs in-situ

Dans cette section, nous allons comparer les performances du circuit complet implémentés avec des moniteurs basés sur les deux layouts. Le circuit étudié est un circuit digital de correction d'erreur « Bose, Ray-Choudhary, Hocquenheim » (BCH). Son fonctionnement complet est détaillé dans le chapitre 4. Le circuit a été implémenté dans les technologies 45nm sur substrat massif et 28FDSOI de STMicroelectronics, à des fréquences de fonctionnement à 500MHZ et 1Ghz respectivement. Le fonctionnement à ces fréquences est garanti dans les corners SS_0.9V_m40°C (procédé lent pour le NMOS, et le PMOS, à une tension de 0.9V et une température de -40°C) et FF_1.3V_125°C (procédé rapide pour le NMOS et le PMOS à une tension de 1.3V et une température de 125°C).

- **Bilan d'intégration du moniteur basé le layout LA**

L'intégration des moniteurs devrait se faire à la sortie des chemins les plus lents du circuit. Toutefois, le délai de ces chemins est obtenu avec précision après l'optimisation de routage. Idéalement, les moniteurs doivent être insérés à cette étape. Or, après le routage, l'emplacement des bascules et la position des interconnexions sont déjà fixés. Le fait d'intégrer de nouveaux éléments nécessitent un re-placement et un re-routage. Ces opérations ont un coût significatif tel que la dégradation de tous les paramètres parasites (RC, couplage entre les fils…) et les performances (délai, consommation, intégrité du signal). Nous devons donc choisir une autre approche dans laquelle nous pouvons insérer les moniteurs sans compromettre la performance. Ainsi, nous pouvons sélectionner les moniteurs dans l'étape de la synthèse car une STA peut être effectuée à ce niveau. Cependant, cette STA est imprécise car l'arbre d'horloge et les fils de connexions sont considérés idéals, i.e. le délai des fils est nul. De plus, le circuit subit plusieurs modifications depuis la synthèse ce qui entraine des modifications des délais. Compte tenu de ces conditions, nous avons décidé d'insérer les moniteurs après l'étape CTS car l'arbre est déjà propagé et les délais sont réels.

Nous avons appliqué le flot décrit dans la figure 3.4 sur le circuit BCH. Après l'étape de la synthèse et du placement, le circuit compte environ 10000 portes logiques et 1000 chemins. Nous décidons d'insérer les moniteurs dans 10% des chemins les plus lents, marqués par les pointillés dans la figure 3.7(a), soit dans 100 chemins. Ces chemins sont obtenus après l'étape de la CTS et une STA dans le pire corner (SS_0.9V_m40°C).

Le choix de la sélection des 10% des chemins critiques se base sur le résultat du vieillissement d'un chemin d'inverseur présenté dans le chapitre 1 [81]. En effet, nous avons

démontré d'une part que le chemin d'inverseur subit une dégradation d'environ 10%, après l'application de la contrainte de type AC pendant 2000s pour une tension de 1.8V. Ces conditions de stress correspondent à un vieillissement de 10 ans. D'après les analyses STA, une plage de délai de 10% de chemin critiques correspond à une sélection de 100 chemins de propagation (cf. figure 3.7 (a)).

Nous nous sommes intéressés à une autre dimension qui est la précision de la STA dans l'étape de la CTS. En effet, nous avons recalculé les délais des chemins critiques sélectionnés après CTS à l'aide des simulations SPICE. Le résultat est donné dans la figure 3.7 (b), il montre une bonne corrélation entre les deux méthodes, nous observons un écart de 5% en moyenne sur une période de 1GHz. Nous concluons que nous pouvons utiliser les résultats provenant de la STA en post-CTS.

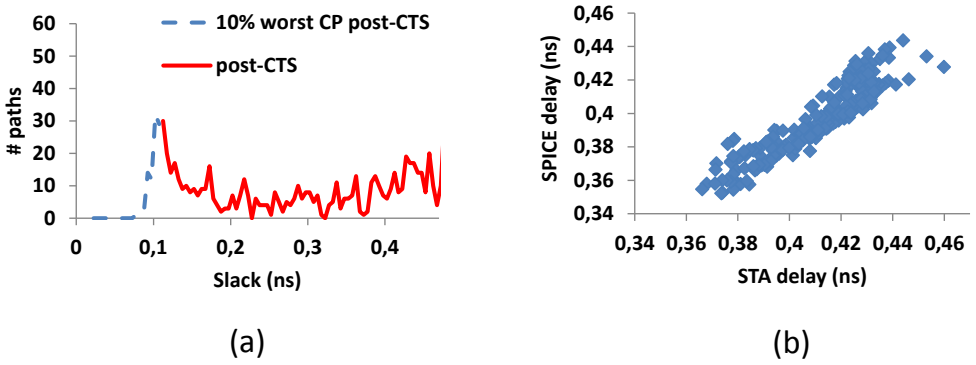


Figure 3.7 : (a) Histogramme de distribution des délais des chemins du BCH après CTS, les moniteurs sont intégrés dans les 10% des chemins les plus lents (pointillé). (b) corrélation SPICE/STA des délais après CTS.

Nous avons intégré les moniteurs avec la méthode LA dans la collection des chemins discutés ci-dessus. L'intégration est faite avec la méthode ECO (Engineering Change Order) après la CTS, c'est-à-dire par le remplacement de la bascule de capture d'un chemin par une cellule standard contenant le moniteur. Ensuite, le flot s'achève par l'optimisation de la CTS après l'insertion des moniteurs, routage et optimisation du routage.

La figure 3.8 présente la comparaison de distributions des chemins critiques pour le circuit BCH de référence et celui avec des moniteurs avec la méthode LA. Pour le circuit avec les moniteurs, la dégradation temporelle de tous les chemins est bien mise en évidence. Particulièrement, elle est de 90ps pour le chemin le plus critique (marqué en étoile). Cette dégradation est due à la congestion lors du routage, i.e. insuffisance des ressources de routage. L'origine de cette congestion est la version du moniteur non optimisée que nous avons évalué.

En effet, les entrées/sorties de la bascule du moniteur n'étant pas alignées avec les points de passage du routeur (Routing Track), l'outil alloue plus de ressources pour accéder à ces entrées/sorties. Ceci conditionne la qualité du placement routage du circuit.

La cellule a été réalisée dans un contexte de recherche et engineering et de nombreuses optimisations sont bien possible, notamment sur la taille et la performance. Dans un contexte industriel, le développement de cette cellule aurait surement de meilleures performances.

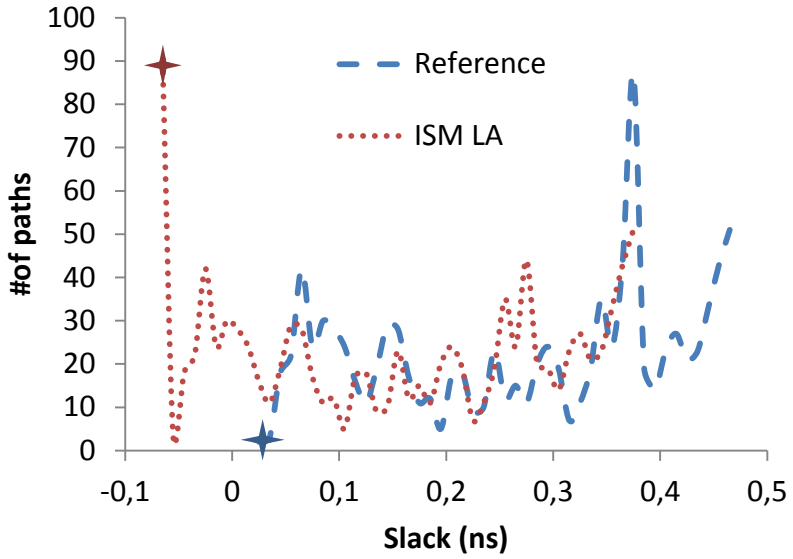


Figure 3.8 : Distribution des délais après optimisation du routage pour un BCH de référence (sans moniteurs) et avec moniteur

- **Bilan d'intégration du moniteur avec la méthode layout LB**

Nous nous proposons maintenant de présenter les résultats d'implémentation des moniteurs issus du layout LB. Nous allons étudier l'impact de l'insertion d'un grand nombre de moniteurs. Pour ce faire, nous avons implémenté deux BCHs avec des moniteurs de type LB dans 10% et 30% (dénotée 10%_ISM et 30%_ISM) des chemins les plus lents, soit 100 et 300 moniteurs.

L'insertion et la connexion des moniteurs sont effectuées avec un programme développé en langage Tcl dans l'environnement de simulation de Cadence. La proximité entre la bascule de référence et celle du moniteur est établie en imposant plus précisément des contraintes de délai entre la bascule de référence et le moniteur afin de minimiser le skew. Nous rappelons que le skew est défini par la différence des délais entre les entrées CP des bascules.

La distribution des chemins critiques après l'optimisation du routage est présentée dans la figure 3.9 pour trois circuits BCHs. Macroscopiquement, les circuits n'ont pas subi de dégradation à l'issue de l'intégration des moniteurs car les slacks des plus lents chemins sont

environ les mêmes pour les trois configurations du BCH. Si on s'intéresse au chemin critique du BCH de référence, marqué en étoile, il n'est plus critique dans les circuits avec les moniteurs. Initialement, son slack est de 20ps pour un BCH de référence, puis il passe à 50ps pour 10%_ISM, ensuite il devient 40ps pour 30%_ISM. Cette observation est contre intuitive car ce chemin doit subir plus de dégradation en ayant intégré un grand nombre de moniteurs. Ceci s'explique d'une part par une nouvelle répartition de la position des bascules dans les trois circuits et d'autre part par la modification de leurs skews pendant les étapes des optimisations (post-CTS et post-routage), ce qui se traduit par un retard supplémentaire pour que le signal d'horloge arrive à l'entrée CP des bascules de capture.

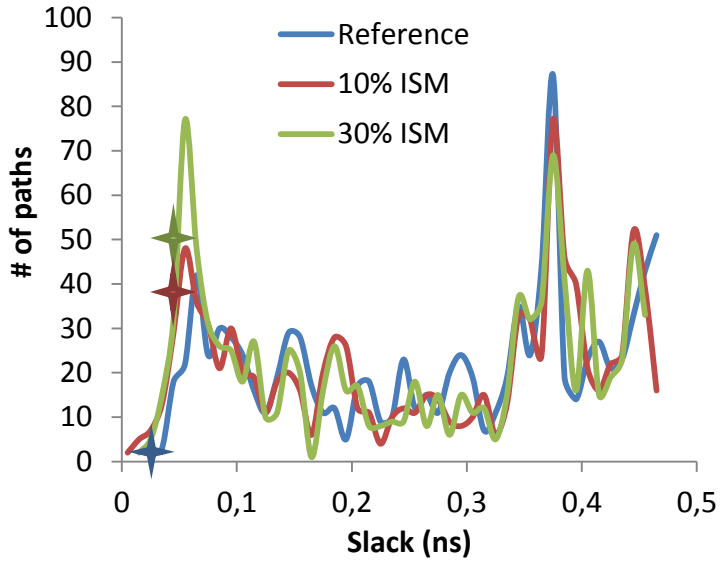


Figure 3.9 : Distribution des chemins critiques après placement-routage pour trois BCHs : référence (sans moniteurs), avec 10% et 30% des chemins critiques monitorés.

La dégradation des performances temporelles induite par l'intégration des moniteurs selon la méthode LB est très faible, voire négligeable à l'issue d'insertion des ISMs, même avec un grand nombre de chemins critiques monitorés, en l'occurrence 30%. L'impact des moniteurs sur les chemins se limite au délai induit par la capacité du fil du moniteur appliqué sur l'entrée D de la bascule de référence. Pour montrer cet impact, nous avons comparé les délais de 100 chemins monitorés avant et après insertion des moniteurs. Le résultat est présenté dans la figure 3.10, il montre une bonne corrélation des deux délais, soit 10ps de dégradation moyenne supplémentaire sur des chemins d'environ 500ps de délai, ce qui reste négligeable. Notons que ce résultat est établi en effectuant une STA en post-CTS.

Compte tenu des résultats présentés dans ce paragraphe, nous pouvons conclure que la conception du moniteur avec la méthode de type LB offre de meilleur performance en matière de délai et son intégration est très simple en comparant avec le moniteur conçu avec la méthode LA. Cependant, des contraintes de délais doivent être imposées pour gérer la proximité avec la bascule de référence, ce qui nécessite un re-placement des circuits logiques. Ceci n'est pas acceptable dans les circuits complexes où la contrainte sur la période est difficilement atteignable. Afin d'éviter l'utilisation des contraintes supplémentaires, il faut guider l'outil dans le placement des moniteurs en imposant des cordonnées bien précises dans le circuit. L'utilisation du moniteur de type LA quant à lui peut être recommandée lorsque nous connaissons dès l'étape de la synthèse les chemins à monitorer et donc ils peuvent être insérés même durant l'étape de synthèse. Ainsi, nous n'avons pas à modifier la topologie du circuit au cours du flot à l'issue de leur insertion.

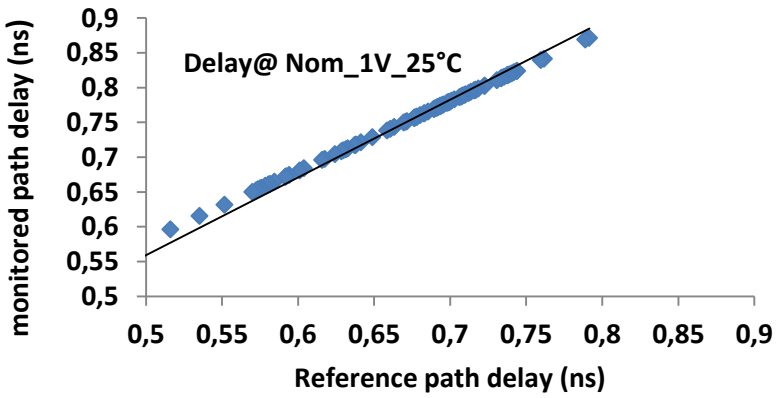


Figure 3.10 : Comparaison des délais des chemins avant et après insertion des moniteurs.

3.3.4 Couverture des chemins critiques par les moniteurs in-situ

Nous avons expliqué dans le paragraphe 3.3.3 la nécessité d'insérer les moniteurs in-situ après la CTS. Or, lorsque nous parcourons le flot depuis la CTS, des modifications de délais sont observées après chaque étape à cause des optimisations du temps de setup et du hold. Il faut donc s'assurer que les 10% des chemins les plus lents après l'optimisation du routage sont bien équipés de moniteur. Nous nous proposons donc de calculer le taux de couverture des moniteurs défini par le rapport des chemins monitorés après optimisation du routage et le nombre de moniteurs initialement insérés en post-CTS. En effet, après l'optimisation du routage, la STA a montré que 60% (sur les 10% initialement choisis) des chemins ne sont pas

équipés avec des moniteurs (ce qui veut dire que seulement 40% des chemins identifiés par la STA pendant la CTS ont été couvert). Un taux de 60% n'est pas acceptable, nous proposons les deux approches suivantes pour améliorer ce taux de couverture.

- **Méthode dite Cell-Padding :**

Cette méthode consiste à dérouler le flot d'implémentation jusqu'à l'étape de la CTS afin de de repérer les chemins critiques. Ensuite, nous recommençons une fois de plus le flot depuis l'étape du placement mais en réservant un espace (Cell-Padding) à côté des bascules des chemins critiques pour l'insertion des moniteurs, sans imposer de contraintes de délai afin de laisser à l'outil plus de degrés de libertés pendant les optimisations sans se préoccuper des moniteurs. Cette approche nous a permis d'améliorer le taux de couverture à 50% après optimisation du routage. Nous signalons que l'inconvénient de cette méthode est la répétition du flot.

- **Insertion après l'optimisation de la CTS.**

La particularité de l'étape post-CTS est que la position des registres est fixée contrairement à la position des cellules combinatoires qu'on peut encore toujours manipuler. Nous avons profité de cette propriété pour insérer les moniteurs sans contraindre de délai puisque nous allons imposer les cordonnées d'insertion. De plus, nous faisons en sorte que la distance entre le moniteur et la bascule de référence soit toujours la même et la plus petite possible afin d'éviter les longs fils de connexions.

Ce placement personnalisé a été établi à l'aide d'un algorithme que nous avons mis en place. Il consiste à chercher des emplacements 'légitimes', dans un périmètre donné, autour des bascules de référence, afin d'insérer les moniteurs. Cela consiste en d'autre terme de trouver des cellules de la logique combinatoire qui peuvent être déplacées sans impacter la performance. Ensuite, nous insérons les moniteurs à la place de ces cellules. Cette opération peut entrainer un chevauchement des moniteurs et les cellules combinatoires du circuit qui peut être facilement résolu et à moindre coût en 'raffinant' le placement. La dégradation des délais induite par le déplacement de la logique combinatoire est corrigée dans l'étape en post-routage.

Le figure 3.11 montre la distribution des délais après l'insertion des moniteurs, elle met en évidence un taux de couverture de 80% sans induire des pénalités. Il est important donc de souligner que les délais des 20% des chemins non couvert par des moniteurs sont inférieurs aux autres chemins monitorés et que tous les chemins dits sous-critiques sont couverts de moniteurs dans ce cas d'étude.

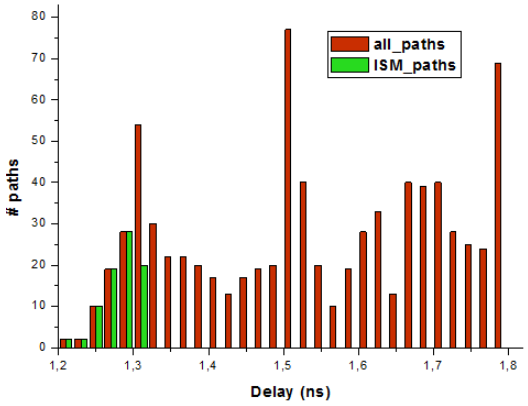


Figure 3.11 : Distribution des délais des chemins du BCH après insertion des moniteurs: la courbe en rouge représente les chemins de référence et la courbe verte représente les chemins monitorés

3.3.5 Evolution des chemins monitorés en prenant compte les PVTs

Dans l'approche précédente, nous avons choisi de monitorer 10% des plus lents chemins sélectionnés dans le pire corner d'implémentation. Comme le circuit est censé fonctionner dans plusieurs PVTs, nous devons nous assurer que nous maintenons une bonne couverture des chemins critiques à travers ces PVTs. Pour cela, nous avons mené une étude qui consiste à choisir une collection de chemins dans un corner de référence (cf. figure 3.12) et nous nous sommes focalisés sur le taux de présence de cette collection de chemin dans les pires chemins des autres PVTs caractérisés dans la plateforme de la technologie 28FDSOI.

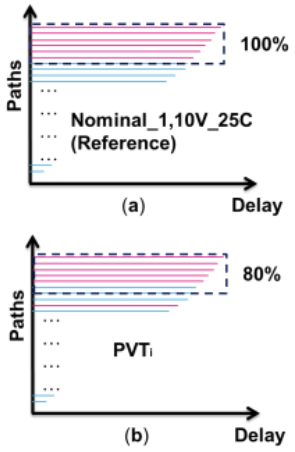


Figure 3.12: (a) Illustration des chemins monitorés dans un corner de référence. (b) taux de présence des chemins monitorés dans un corner arbitraire PVT_i.

Cette étude a été menée sur le circuit BCH, il a été implémenté et caractérisé dans tous les corners de la plateforme du 28FDSOI, à savoir trois types de procédés (nominal, SS, FF), une gamme de tension variant de 0.6V à 1.3V avec un pas de 0.05V et des températures de 0°C, 125°C, -40°C. Le corner de référence choisi est le Nominal_1.10V_25°C (typique à une tension de 1.1V et une température à 25°C) et la fréquence d'implémentation est de 1 GHz.
Après implémentation du circuit dans ces corners, nous avons effectué une analyse temporelle STA et nous nous sommes intéressés au pourcentage de présence des 10% des chemins les plus lents de la collection de référence dans tous les autres corners qui sont indépendants les uns des autres.
Le résultat du taux de présence de la collection de chemins de référence dans les autres corners est reporté dans la figure 3.13 pour chaque type de procédé, en l'occurrence le procédé nominal (figure 3.13 (a)), FF (figure 3.13(b) et SS (figure 3.13 (c)) ainsi que pour les tensions et les températures qui leur sont associées. Nous remarquons que le taux de présence de la collection de référence varie en moyenne entre 80% et 90% pour les corners nominaux, FF et SS pour les tensions au-delà de 0.8V. Ce pourcentage est plus réduit pour les faibles tensions du corner SS en dessous de la tension du point d'inversion (0.8V). Particulièrement, nous avons obtenu respectivement 30% et 60% de chemins de la collection de référence dans les corners SS_0.65_0C et SS_0.65_m40C. Cette différence s'explique d'une part par le fait qu'à faible tension, i.e. en dessous de 0.75V, la fréquence du circuit n'est pas atteignable et d'autre part par l'effet de la température sur les délais et la redistribution des chemins. Cet effet a été explicité dans le chapitre 2.

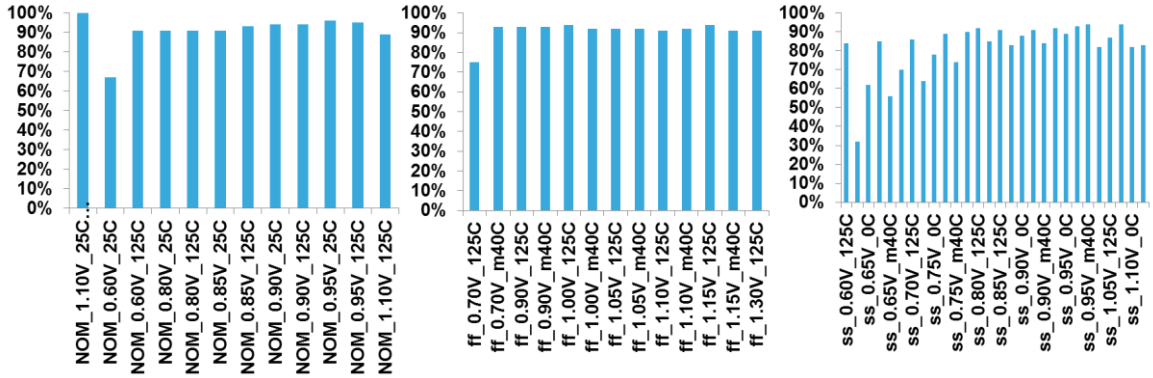


Figure 3.13 : Pourcentage de présence d'une collection de chemins d'un corner de référence (Nom_1.1V_25°C) dans les corners de la plateforme du design en technologie 28FDSOI.

Pour la fréquence de 1GHz et les tensions en dessous de 0.8V, le circuit BCH est non fonctionnel car les slacks des chemins sont négatifs. Nous considérons donc qu'un taux de

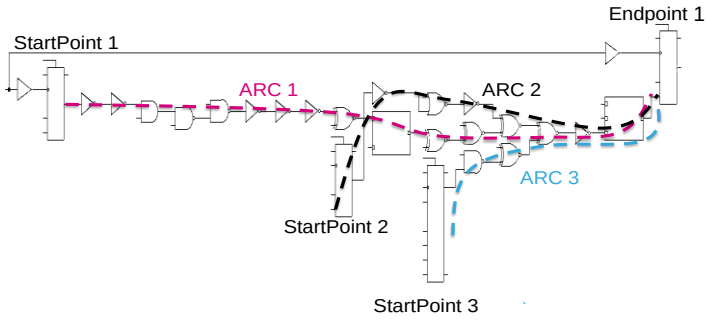
couverture de 80% de la sélection de référence est acceptable pour monitorer la variation des délais induite par les PVTs et que nous pouvons choisir les chemins à monitorer dans le pire corner.

3.3.6 Méthodologie d'insertion des moniteurs basée sur les délais des chemins sous-critiques

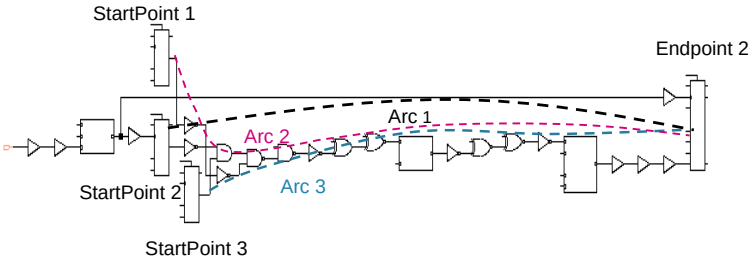
Dans ce paragraphe, nous proposons d'améliorer la méthodologie présentée dans le paragraphe 3.3.1, concernant l'insertion des moniteurs. L'objectif recherché est d'une part réduire le nombre des moniteurs à insérer dans un circuit qui contient un nombre important des bascules et d'autre part la considération des délais des chemins sous-critiques. En effet, les chemins sous-critiques peuvent avoir une probabilité de violation de setup lors du vieillissement. Cette situation dépend des conditions d'usage et fera l'objet de discussion dans le chapitre 4.

Pour ce faire, nous allons établir une approche qui consiste à identifier des bascules dans une fenêtre de temps. Le but est d'anticiper la dégradation des chemins et d'utiliser les moniteurs in-situ pour remplacer les corners 'vieillis'. Comme on ne sait pas comment les chemins vont vieillir les uns vis-à-vis des autres, on s'intéresse à identifier les chemins qui vont rentrer dans une fenêtre de couverture temporelle TWC (Time Windows Coverage). Dans cette étude, nous n'aborderons pas la notion de la fenêtre du temps TW du moniteur.

Tout d'abord, nous rappelons que les transitions de signaux se transmettent via la logique combinatoire à partir d'un registre de lancement (Startpoint) vers un registre de capture (Endpoint). Le registre de capture peut être la destination de plusieurs transitions provenant de plusieurs registres de lancement. Nous désignons par le terme 'arc', le délai cumulé sur un chemin de propagation pour toute transition depuis une bascule de lancement vers une bascule de capture. Le délai des arcs varie en fonction du nombre de portes logiques qui le composent, le délai de chaque porte, la capacité de charge des portes ou bien la manière dont l'arbre d'horloge est propagé. La figure 3.14 met en œuvre, pour deux bascules de capture, la propagation du signal à travers trois arcs différents. Pour le registre 1 (Endpoint 1), l'arc1 est plus critique que les arcs 2 et 3 à cause du nombre de portes logiques qu'il traverse. Les 3 arcs du registre 2 (Endpoint 2) sont tous de même ordre de grandeur de délai vue l'architecture du cône logique vers la bascule 2.



(a)



(b)

Figure 3.14 : Chemin de propagation vers deux bascules de capture (Endpoint 1&2), à partir de trois bascules de lancement (Startpoint 1,2 et 3)

Au vue de cette observation, il est important de s'intéresser à la distribution des délais des arcs pour toutes les bascules du circuit. En effet, dans la figure 3.15, nous présentons la distribution de 50 arcs pour les deux registres de capture de la figure 3.14, faisant partie du circuit BCH utilisé pour toutes les analyses de ce chapitre. Cette figure montre que pour l'Endpoint 1, l'écart entre le délai du chemin critique et le 50ème est de 40 ps alors qu'elle est de 20ps pour le registre 2. Ce qui montre que la bascule 2 est à considérer car les délais de ses arcs sont très proches du chemin le plus lent. Cette observation démontre la criticité de la bascule et peut nous aider à parvenir à une décision plus rapide pour sélectionner les bascules à monitorer.

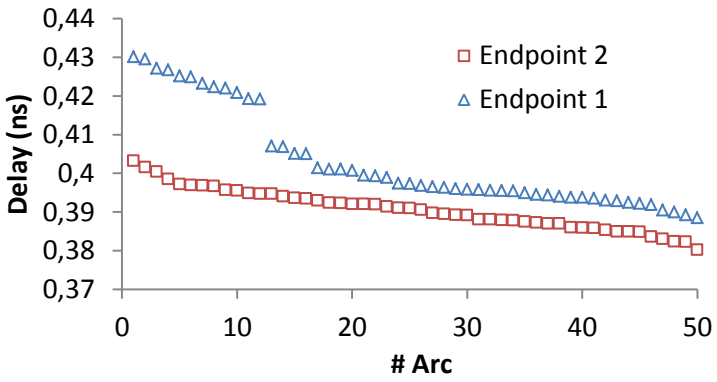


Figure 3.15 : Distribution des délais de cinquante arcs pour deux bascules.

Nous allons à présent nous intéresser à l'étude de la distribution des arcs pour un grand nombre de bascules de capture du circuit BCH afin de savoir s'il est possible de modéliser la distribution des arcs statistiquement. Nous avons étudié les 50 premiers arcs pour 300 bascules. Nous nous proposons de présenter les déviations standards normalisées des délais en fonction de leurs moyennes des délais par bascule (cf. figure 3.16).

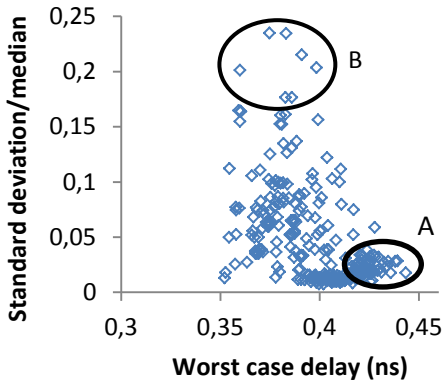


Figure 3.16 : La déviation standard normalisée des délais des arcs de chaque bascule du BCH

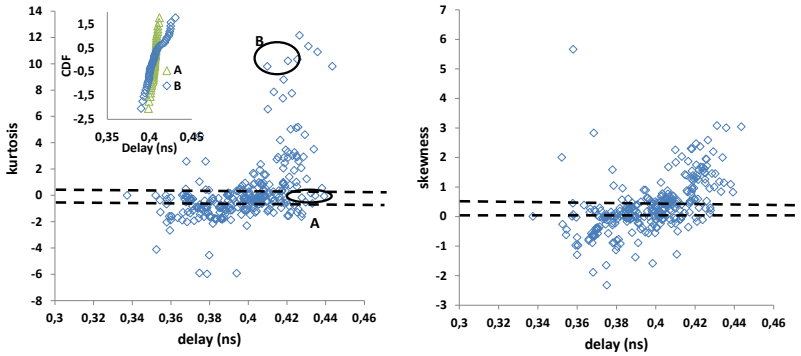
Nous remarquons que les chemins de type A, dont le délai est proche du chemin critique du BCH (0.45 ns), ont une faible déviation comme le cas de l'Endpoint1 de la figure 3.15. Les chemins de type B présentent une forte déviation (cas de Endpoint 2 de la figure 3.15).

En poussant plus loin l'analyse par un test de normalité sur les distributions des délais (cf. figure 3.16), les coefficients de l'aplatissement (cf. figure 3.17 (a)) et d'asymétrie (cf. figure 3.17(b)) montrent que la distribution de délai des chemins de type A, suit une loi normale car l'aplatissement et l'asymétrie sont nuls (représenté par les pointillés dans la figure 3.17). Cette distribution n'est pas normale pour les chemins en dehors de cette zone, i.e. cas du chemin B.

En conclusion, les bascule de capture des chemins de type A seront choisis en priorité pour l'insertion des moniteurs car les délais des arcs sont plus proches du délai du chemin critiques.

L'analyse précédente portait sur des délais des chemins dans les conditions nominales. Il est indispensable de prendre en compte la variabilité locale et son impact sur les délais des chemins. Pour cela, nous avons repris les mêmes bascules de la figure 3.15 avec 50 arcs, et nous avons simulé l'impact de la variabilité avec la méthode Monte-Carlo (MC). Chaque délai est calculé après 500 tirages MC dans le corner typique. Vu le temps de simulation, nous nous sommes intéressés aux deux bascules de la figure 3.15. Les valeurs moyennes des arcs ainsi que les dispersions sont présentés dans les figures 3.18 et 3.19 respectivement, ils

montrent que les déviations standards σ des arcs sont similaires quelle que soit la bascule et varient entre 2 et 2.5ps. Ces déviations standards vont servir pour la sélection des chemins à monitorer en appliquant l'approche statistique présentée ci-dessous. Par la suite, une valeur moyenne de 2.3ps sera considérée pour prendre en compte la variabilité locale stochastique



(a) (b)

Figure 3.17 : Test de normalité des délais des arcs de chaque bascule du BCH : (a) le coefficient d'asymétrie, (b) le coefficient d'aplatissement. Les chemins délimités par les pointillés suivent une loi normale.

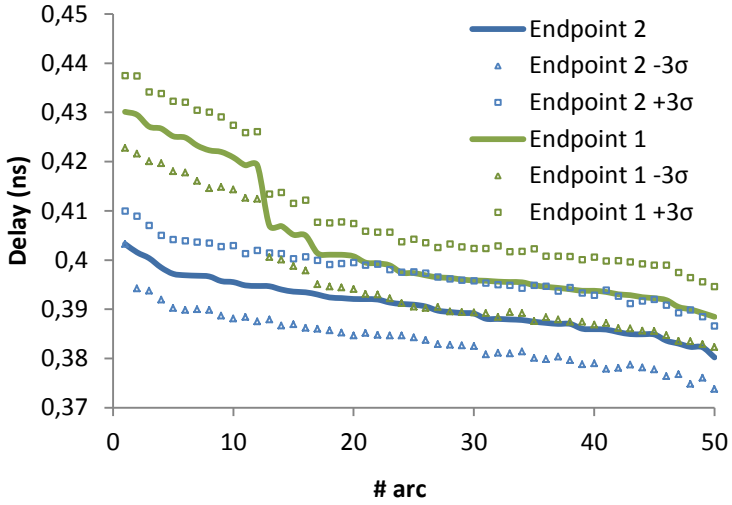


Figure 3.18 : Représentation des délais moyens de cinquante arcs pour deux bascules ainsi que leurs délais à +3σ et -3σ dûs à la variabilité locale.

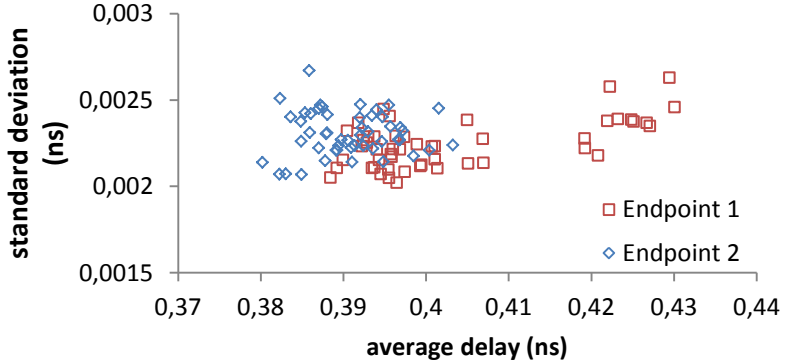


Figure 3.19 : Déviation standard des délais de chaque arc pour les deux bascules.

3.3.7 Méthodologie de calcul de la criticité des bascules

En utilisant l'analyse des délais des chemins ainsi que leurs déviations standards, nous allons attribuer à chaque bascule une probabilité P_{Di} qui détermine sa criticité. Les bascules ayant un poids fort sont susceptibles d'être critiques et donc il faut les monitorer.

Pour calculer ce poids, nous définissons une fenêtre de couverture temporelle TWC dont le délai est égale à la différence entre la période T et un coefficient ε (cf. figure 3.20 (a)). Cette plage de délai pourrait être choisie de telle manière à couvrir les délais induits par les variations de types locales ou le vieillissement. Quant à la fenêtre de détection du moniteur (TW), elle pourrait être exprimée par :

$$TW = T - D_{CP} + m \qquad \text{EQU 3.1}$$

D_{CP} est le délai du chemin sous critique est *m* est une marge de délai.

Dans cette étude, nous allons considérer aussi les délais des arcs des bascules contenues dans la TWC. En effet, nous associons à chaque arc j qui arrive sur un registre D_i la probabilité P_{Dij} pour que le délai de chaque arc j soit compris dans la fenêtre de couverture temporelle en tenant compte de la dispersion locale σ. Nous faisons l'hypothèse ici que la distribution stochastique suit une loi normale. (cf. figure 3.20(b)), soit :

$$P_{Dij} = P_{(T-D_{ij}>\varepsilon)} \qquad \text{EQU 3.2}$$

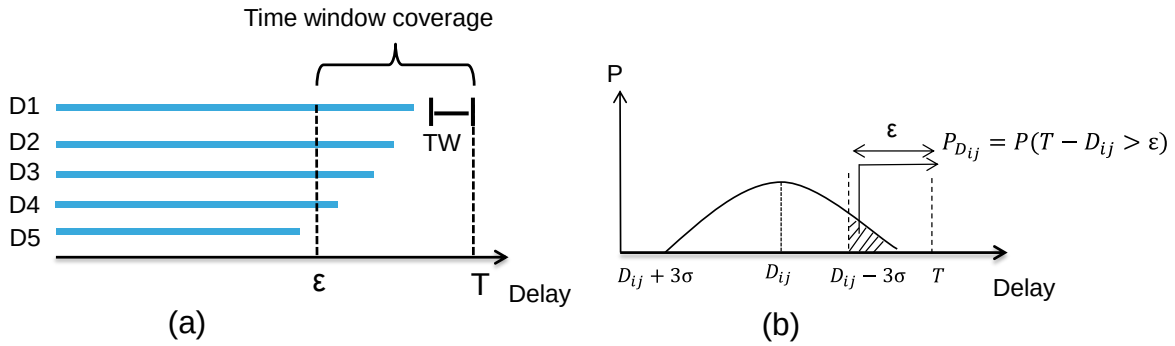


Figure 3.20 : (a) illustration de la fenêtre de détection. (b) la région d'intégration des chemins compris dans la fenêtre de détection en tenant compte de la variabilité locale

La criticité d'une bascule est donnée par la probabilité qu'au moins un arc se trouve dans la fenêtre de couverture temporelle. Pour les bascules dont plusieurs arcs pourraient se retrouver dans la TW, il faudrait pondérer ces probabilités par le nombre d'arcs. L'activité des arcs n'est pas prise en compte dans cet exercice, nous ferons donc l'hypothèse de l'équiprobabilité de l'activation des arcs.

Deux critères sont évalués ici:

- Critère 1 : Pour chaque bascule, nous considérons que tous les arcs sont indépendants, et le calcul de criticité prendra en compte la contribution de tous les arcs
- Critère 2 : seulement le pire arc est considéré pour chaque bascule.

Pour le critère 1, nous définissons P_{Di} comme étant la moyenne des probabilités P_{Dij} de tous les arcs (cf. EQU3.2). P_{Di} s'écrit :

$$P_{D_i} = \sum_j \frac{P_{D_{ij}}}{N_j} \qquad \text{EQU 3.3}$$

N_j étant le nombre d'arc par registre.

Pour le critère 2, cette moyenne se limite à un chemin, i.e. le pire arc et donc P_{Di} s'écrit :

$$P_{D_i} = P_{D_{i,j=1}} \qquad \text{EQU 3.4}$$

Nous avons calculé la probabilité P_{Di} de chaque bascule du BCH en appliquant les deux critères et en choisissant différentes fenêtres de couverture temporelle (TWC), en l'occurrence 20ps, 30ps et 40ps. Ces valeurs ont été choisies selon les résultats présentés dans la figure 2.11 du chapitre 2 qui permettent d'évaluer l'impact des PVTs et du vieillissement sur le délai.

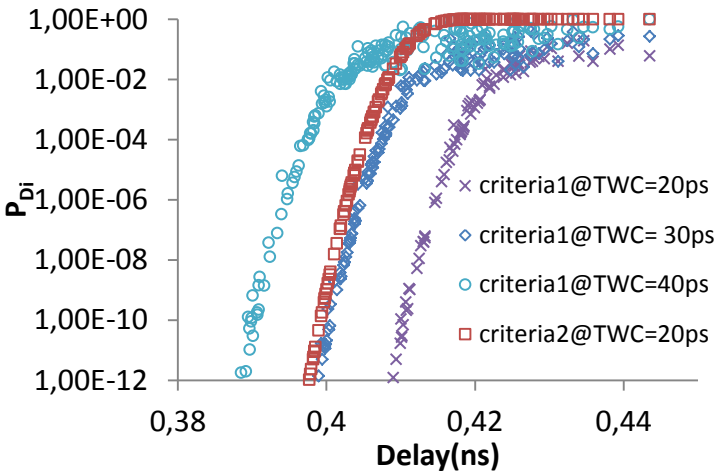


Figure 3.21: Evolution du coefficient P_{Di} de chaque registre en fonction de délai à t_0 pour trois largeurs de TWC en considérant le critère 1 et le critère 2 (courbe rouge)

Le résultat du calcul du poids P_{Di} est donné dans la figure 3.21, il montre que les registres des chemins les plus lents présentent les poids les plus forts, ce qui prouve que tous leurs arcs sont critiques. Cette constatation est valable pour tous les TWCs. La figure 3.21 montre aussi que plus la fenêtre de couverture est large plus le nombre de registres ayant un fort P_{Di} est grand.

Ce résultat est attendu car pour une fenêtre de couverture temporelle large, c'est-à-dire un ε grand, le nombre de chemins vérifiant l'équation EQU3.1 est plus important. Le résultat met aussi en évidence l'avantage de l'identification des chemins critiques par le critère 1. En effet, pour la même fenêtre de couverture temporelle, i.e. TWC = 20ps, le nombre de bascules ayant un poids très fort obtenu par le critère 2, i.e. P_{Di} = 1 est très supérieure à celui obtenu par le critère 1. Par conséquent, en utilisant le critère 2 tous ces chemins doivent être retenus pour les monitorer, ce qui surestime et peut induire des pénalités telles que la difficulté d'intégration. En utilisant le critère 1, le nombre de moniteurs sera réduit car moins de bascules sont à prendre en compte. En justifiant à priori, nous allons démontrer dans le chapitre 4 que la probabilité d'activation du pire chemin par registre est faible, ce qui renforce notre conclusion de choix de critère 1 et mène à une sélection pertinente de bascules à monitorer sans gaspillage de ressources supplémentaires. Avec ce résultat, nous démontrons que la méthode classique ou le critère 2 est pessimiste.

3.3.8 Evaluation de l'approche statistique

Nous nous proposons maintenant d'appliquer les étapes l'identification des bascules susceptible d'être critique détaillées dans le paragraphe 3.3.6 sur les circuits B19, B15, B14 du benchmark ITC'99 et le circuit BCH. Aussi, nous allons comparer les résultats obtenus avec l'approche de sélection des 10% des plus lents chemins.

Tout d'abord, nous commençons par illustrer l'approche d'identification du paragraphe 3.3.6 dans l'algorithme ci-dessous. Nous signalons que la largeur de la fenêtre de couverture temporelle TWC, le nombre d'arcs par bascule à considérer et le critère d'acceptation pour qu'une bascule se trouve dans TWC, définit par $P_{Di,crit}$, sont les entrées de l'algorithme. Pour chaque bascule, nous identifions seulement les arcs dont le délai se situe dans la fenêtre de couverture temporelle en tenant compte de la variabilité locale. Les registres identifiés sont ceux dont P_{Di}, défini par la moyenne des probabilités P_{Dij}, qui sont supérieure à un $P_{Di,crit}$.

Nous avons comparé l'approche statistique avec celle de la sélection des 10% des plus lents chemins. Dans cette expérimentation, nous avons considéré les circuits benchmarks B19, B15, B14 de ITC'99 et le circuit BCH. Nous avons implémenté ces circuits dans la technologie 45nm en considérant les corners SS_0.9V_m40C et FF_1.3V_125°C et à une fréquence de 500MHz. Les données d'implémentations sont listées dans le tableau 3.1. Nous remarquons que le circuit B19 contient le plus grand nombre de cellules et des bascules. Il s'agit donc

d'un bon exemple pour l'application de l'algorithme. D'autre part, nous n'avons pas pu implémenter le circuit B14 à 2ns (500Mhz), le chemin critique présente un délai de 1.97ns. Concernant l'application de l'algorithme, le nombre d'arcs considéré est de 50, la fenêtre de couverture temporelle est fixée à 30ps et $P_{Di,crit}$ égale à 0.01. Notons que d'après la STA du circuit BCH, une sélection de 10% des plus lents chemins présente une fenêtre de temps de 30ps.

***Algorithm Criteria 1**: Potential critical path identification*

Input: *Ranking paths with STA*
N_{ARC} *: Maximum arc number*
ε *: Time windows coverage size*
$P_{Di,crit}$ *: acceptance criteria for an Endpoint to be in detection window*

Output: *Potential critical path to be monitored*

```
For each i
    For each j
        P_Dij = P(T-Dij>ε)
    End for
  P_Di= mean (P_Dij)
if  P_Di > P_Di,crit
    select Endpoint
else
    Discard Endpoint
End if
End for
```

Le tableau 3.2 résume le résultat de l'application des deux méthodologies sur les circuits de benchmarks, il confirme le résultat obtenu dans la figure 3.21. Une réduction du nombre de registres à sélectionner est observée pour tous les benchmarks. Nous avons obtenu un rapport de réduction de nombre de chemins de 2 par rapport à l'approche 10% pour les circuits B19 et B15. Pour les deux autres circuits, la réduction du nombre de bascules est moins spectaculaire.

Benchmark	# of gates	# of Flip-Flop	Clock period Fast corner	Clock period at worst-case corner
B19	36211	2180	1,75 ns	2 ns
B14	7405	222	1,38 ns	1,97 ns
B15	14194	446	1,42ns	2 ns
BCH	8469	869	0,49 ns	2 ns

Tableau 3.1 : Les informations post-implémentation des circuits implémentés

Benchmark	Approach 10% CP	Statistical approach algorithm
B19	218	105
B15	44	17
B14	22	19
BCH	86	73

Tableau 3.2 : Comparaison de nombre de chemins critiques monitorer avec l'application de la méthodologie de sélection des 10% des chemins les plus lents et de l'approche statistique

Nous nous intéressons à une autre dimension qui est l'évaluation du nombre de bascules calculés par l'algorithme en fonction des différents paramètres c'est-à-dire la fenêtre de couverture temporelle et $P_{Di,crit}$. Pour cela, nous avons appliqué l'algorithme sur le circuit BCH en faisant varier TWC et en considérant deux valeurs de $P_{Di,crit}$. Le résultat est donné dans la figure 3.22, il présente l'évolution du nombre de moniteurs pour les différents critères d'étude. Nous remarquons une évolution significative du nombre de moniteurs pour les deux valeurs $P_{Di,crit}$. Ceci s'explique par le fait que plus la fenêtre de couverture temporelle est large plus il y a des chemins à observer. Un choix pertinent des paramètres de l'algorithme est alors nécessaire pour un compromis optimal entre la couverture des chemins et les pénalités induites.

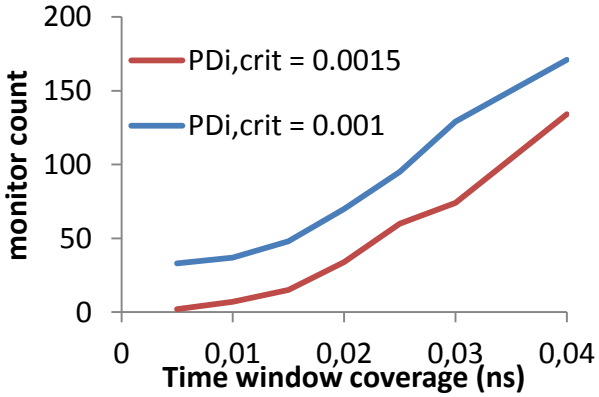


Figure 3.22 : Evolution du nombre de moniteur en fonction de la fenêtre de couverture temporelle pour deux critères de $P_{Di,crit}$.

3.3.9 Impact du vieillissement sur la sélection des chemins critiques à monitorer

Dans cette étude, nous allons appliquer l'algorithme de l'approche statistique mais en tenant compte de la dégradation des chemins du circuit BCH pour adresser les problématiques liés au vieillissement. Pour cela, nous avons utilisé le flot de simulation de vieillissement présenté dans le chapitre 2 en considérant différentes activités, afin de modéliser cette dégradation selon les conditions d'usages.

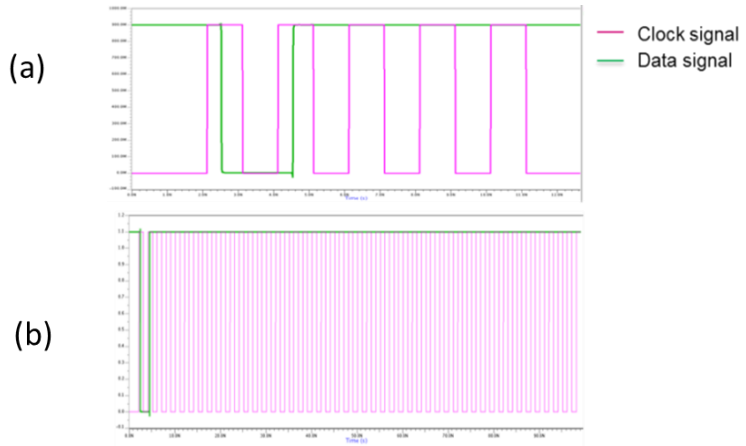


Figure 3.23 : Simulation temporelle du signal AC à 30% d'activité (a) et à 5% d'activité (b)

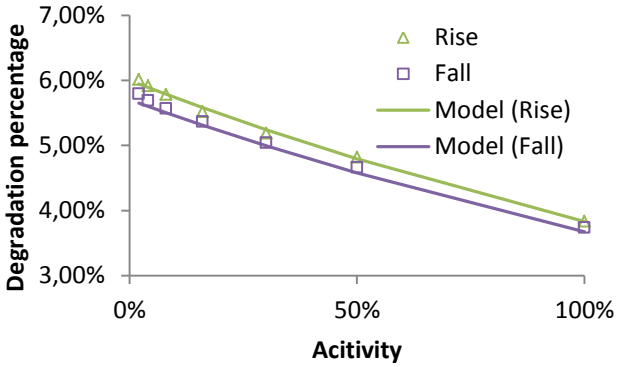


Figure 3.24 : Pourcentage de dégradation du délai en fonction de l'activité en considérant la capture sur front montant (rouge) et front descendant (bleu).

Les stimuli choisis sont des signaux AC avec une activité de 2%, 4%, 8%,16%, 30%, 50%, 100%. Nous présentons dans la figure 3.23 deux cas d'activité : 5% et 30%. Ces stimuli ont été appliqués sur un seul chemin du BCH en considérant deux modes de capture, en l'occurrence sur front montant et descendant. Une étude montre que le comportement du vieillissement est au premier ordre dépendant des portes logiques. Donc l'étude en vieillissement sur un chemin peut être appliquée à l'ensemble des chemins.

La tension de stress choisi est de 1V et la caractérisation a été effectuée à 0.9V, 125°C. Ces conditions représentent les conditions de caractérisations du vieillissement des portes logiques. La figure 3.24 présente le pourcentage de dégradation en fonction de l'activité pour les deux modes de capture. Cette dégradation suit l'équation 3.6 de type :

$$D_g = ae^{-b\alpha}.t^n.V_{DD}^n \qquad \text{EQU3.6}$$

α est l'activité, a et b sont des paramètres de fit.

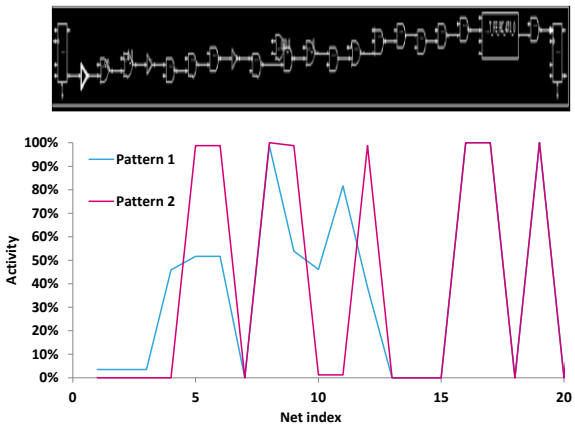


Figure 3.25 : Variation d'activité des portes d'un chemin dans le cas d'application de deux vecteurs

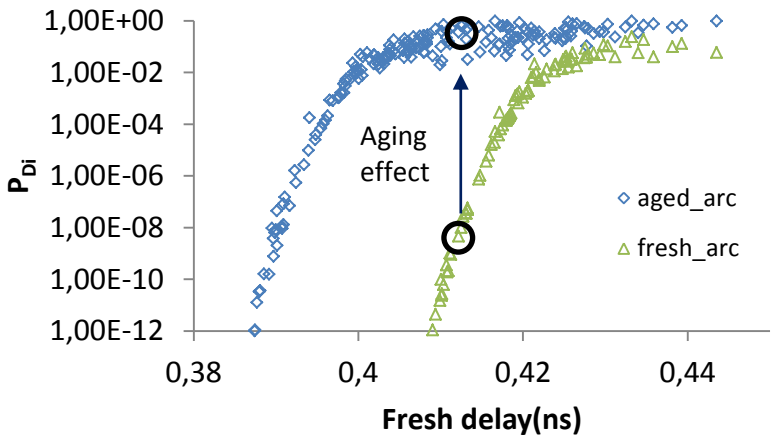


Figure 3.26 : Evolution du coefficient P_{Di} à t_0 et en tenant compte du vieillissement

Par la suite, nous avons utilisé cette loi pour calculer la dégradation de tous les chemins du BCH. L'activité de ces chemins est extraite à l'aide de l'outil Spyglass Physical pour un vecteur de test particulier. Nous avons obtenu une activité variable tout au long des portes du chemin comme le montre la figure 3.25. Pour simplifier, nous avons considéré seulement l'activité de l'interconnexion à l'entrée D de la bascule de capture.

La figure 3.26 présente l'évolution de P_{Di} en considérant les délais des chemins au temps initial et après vieillissement. Le poids des bascules P_{Di} est obtenu par l'application de l'équation 3.6. Elle met en œuvre l'effet du vieillissement sur le poids des registres. En effet, le nombre de registre ayant un poids fort est plus conséquent dans le cas des chemins dégradés. Ceci est attendu car les délais des chemins augmentent sous l'effet du vieillissement et par conséquent un grand nombre de chemin se retrouve dans la fenêtre de couverture temporelle, i.e. (T - ε).

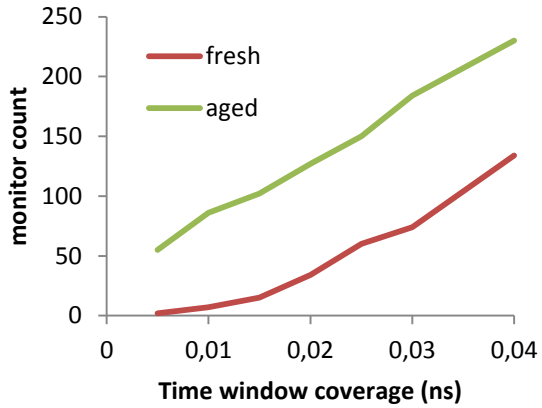


Figure 3.27 : Evolution du nombre de moniteurs en fonction de la fenêtre de détection, en prenant compte les délais à t_0 (rouge) et vieillis (vert) pour $P_{Di,crit} = 0.01$

La figure 3.27 présente le nombre de moniteurs que nous devrions insérer dans le cas d'application de l'algorithme de l'approche statistique pour garantir une couverture du vieillissement à 10 ans. Ce calcul est établit en faisant varier la fenêtre de couverture temporelle et en garantissant une couverture de 2σ, ce qui correspond à $P_{Di,crit} = 0.01$. Dans ces conditions, une augmentation de nombre de moniteur par 3 est observée par rapport au délai à t_0 pour le circuit BCH. Ce grand nombre de moniteurs supplémentaires s'explique par l'hypothèse pessimiste sur le choix de la valeur de l'activité sur le chemin que nous avons choisi. En effet, nous avons appliqué la valeur de l'activité sur l'entrée D de la bascule sur tout le chemin alors que certains fils présentent peut être une activité nulle (cf. figure 3.25). L'approche la plus adéquate serait de considérer l'activité de chaque nœud du chemin. Or, lorsque nous appliquons un vecteur, nous ne connaissons pas l'arc activé et donc nous ne pouvons pas estimer la vraie valeur de l'activité des fils. Ceci est donc une observation préliminaire, car en effet, une analyse plus poussée de l'impact de l'activité sur le vieillissement doit être menée, afin de pouvoir intégrer une dégradation de délais en fonction d'une activité moyenne, qui proviendrait éventuellement du profil d'utilisation, inconnu pour nous à cette étape d'étude.

3.4 Conclusion

Dans ce chapitre, nous nous sommes intéressés à la stratégie de sélection des chemins critiques à monitorer et la méthodologie d'intégration des moniteurs.

Nous avons commencé par expliciter les différentes méthodologies de sélection des chemins critiques issues de la littérature, en l'occurrence des méthodes basées sur le choix des chemins critiques dans une fenêtre de temps et les chemins susceptible d'être critiques sous l'effet de la variation locale ainsi que le vieillissement. Ensuite nous avons présenté notre approche initiale basée sur l'intégration des moniteurs dans 10% des plus lents chemins dans le pire corner et avons évalué les pénalités induites. Elles présentent 90ps de dégradation du chemin critique pour le moniteur basé sur la méthode dite layout A. Cette pénalité nous a poussé à mettre en place une autre manière d'intégrer les moniteurs à partir des cellules standard de la librairie. Il s'agit de moniteur basée sur la méthode dite layout B. Ce nouveau schéma a donné des résultats satisfaisants en matière de pénalité de délai induite car ils se limitent à la capacité du buffer. Dans ce contexte, nous avons démontré une bonne corrélation des délais entre le chemin de référence et le chemin monitoré.

Nous avons étudié aussi le taux de couverture des moniteurs définit par le rapport du nombre de chemins monitorés présents après l'optimisation du routage et le nombre de chemins initialement sélectionnés. Avec le flot d'insertion en post-CTS, nous avons obtenu un taux de couverture de 40%. Ce taux a été amélioré en proposant d'autres méthodes d'intégrations : méthodes Cell-padding et l'intégration après l'optimisation de la CTS, nous avons atteint successivement 50% et 80%. De plus, nous avons étudié la présence de ces chemins dans les autres PVTs, elle représente une moyenne de 80% pour les corners au-dessous du point d'inversion.

Nous avons proposé une autre méthode d'identification basée sur le délai des chemins sous-critiques, i.e., les arcs. En effet, nous avons attribué à chaque registre un poids en fonction de la criticité de ces arcs qui se terminent sur une bascule de capture, plus le poids est fort plus les arcs de la bascule sont critiques. Cette approche a été appliquée aux benchmarks de ITC'99 en fixant des paramètres tels que la fenêtre de détection et $P_{Di,crit}$: nous avons obtenu une réduction par 2 du nombre de moniteurs insérés pour les circuits B19 et B15 mais moins pour d'autres circuits. Aussi, nous avons considéré un autre aspect qui est le vieillissement des chemins. En effet, la dégradation des chemins du BCH a été calculée à l'aide d'un modèle

établi en fonction de différentes activités qui simulent l'usage du circuit. Le nombre de bascules dont la probabilité P_{Di} élevé devient important avec cette approche à cause de la valeur pessimiste de l'activité que nous avons appliquée.

Le principe de notre approche statistique ressemble à celui proposé dans [135]. La différence est que nous considérons les délais de tous les arcs ainsi que le vieillissement calculé à l'aide d'un modèle obtenu par un vecteur de test. Dans [135], ces paramètres sont considérés par des paramètres de fit qui peuvent être imprécis.

Dans le prochain chapitre, nous allons vérifier la validité de ces approches en présentant des résultats expérimentaux des circuits implémentés dans les technologies 45nm, 28LP et 28FDSOI, contenant des moniteurs in-situ. Ces circuits sont testés dans différents PVTs et conditions d'usages. Nous allons nous intéresser aussi à l'application de AVS à l'aide des moniteurs.

CHAPITRE 4 : APPLICATION DES MONITEURS IN-SITU A L'ADAPTATION DE LA TENSION D'ALIMENTATION ET LA FREQUENCE

4.1 Introduction

Dans les nœuds technologiques avancés, pour réaliser des produits ayants des hautes performances, il faut satisfaire les contraintes fonctionnelles dans un grand nombre de PVT avec un budget de consommation limité. L'approche traditionnelle pour garantir la fonctionnalité dans ces PVTs consiste à ajouter des tensions supplémentaires à la tension nominale dans les pire cas (cf. chapitre 2). Cette tension est définie comme la tension nécessaire pour le fonctionnement du circuit dans les conditions nominales. Si cette approche permet de garantir un bon fonctionnement dans certains scénarios, elle entraine néanmoins une surconsommation dans les conditions de fonctionnement standard car la tension d'alimentation (V_{DD}) est plus importante. De plus, comme l'impact de la variabilité est plus important dans les nœuds technologiques avancés, ces marges de tension doivent augmenter en conséquence, ce qui impacte l'efficacité énergétique du circuit.

Il est donc intéressant de chercher à supprimer les marges inutilisées dans certaines conditions de fonctionnement afin de gagner en consommation énergétique, sans perdre de vue la performance et la fiabilité fonctionnelle. Nous rappelons que la puissance dynamique décroît quadratiquement avec la tension d'alimentation et la puissance statique en dépend linéairement. Une solution passera par des moniteurs in-situ au vue d'une adaptation de tension ou de fréquence (AVS, DFVS, ABB comme expliqué dans le chapitre 2).

Dans ce chapitre, nous nous sommes intéressés à l'application de la méthode AVS à l'aide de la réponse des moniteurs in-situ. Nous commençons par présenter le circuit BCH étudié et rappeler les moniteurs de pré-erreur choisis. Par la suite, nous allons expliquer les résultats obtenus suite à l'application de l'AVS notamment sur la plage de fonctionnement du BCH et les réponses des moniteurs en fonction de la variation de la tension. Aussi, nous mettons en œuvre les problématiques des dispersions locales et globales en mesurant les tensions minimales V_{DDmin} et les fréquences qui garantissent un fonctionnement correct du circuit dans la même puce (intra-puce) ainsi que dans différentes puces (inter-puce).

Nous allons présenter aussi le résultat de l'étude du comportement du circuit dans différentes conditions de fonctionnement. En effet, d'une part, nous allons injecter plusieurs vecteurs de test et nous allons analyser la réaction du circuit en fonction de leurs complexités et ce à travers le comptage du nombre des alertes issues des moniteurs à chaque opération. Cette étude va être menée en faisant varier la température et la fenêtre de détection. Aussi, nous avons développé une méthode de test qui permet d'extraire l'index des moniteurs activés et de connaitre donc précisément le moniteur qui se déclenche. Par ailleurs, nous allons nous intéresser aux moniteurs mêmes en analysant leurs comportements en fonction des conditions d'usage, la température et la fenêtre de détection. Cette technique a permis de mesurer expérimentalement les délais des chemins critiques et puis déceler leurs classements dans la STA. Nous allons expliciter la contribution de chaque source de variabilité à l'aide des comparaisons des mesures expérimentales et de la simulation CAD. Finalement, nous aborderons le comportement des moniteurs dans des applications de vieillissement.

4.2 Véhicule de test considéré

Le premier circuit implémenté puis testé est le BCH (reprenant les initiales de ses inventeurs : Bose, Ray-Chaudhuri et Hocquenghem). Il s'agit d'un code correcteur utilisé pour corriger des erreurs aléatoires, permanentes ou transitoires. Ce circuit possède 7 signaux d'entrées et 5 signaux de sorties, il utilise un générateur des trames pseudo aléatoire PRBS (Pseudo Random Binary Sequence) pour générer des trames qui peuvent éventuellement contenir des erreurs. Ces trames sont ensuite encodées puis décodées par l'algorithme du décodage du Syndrome [135]. Aussi, le circuit permet de détecter le coefficient d'erreur et sa position à l'aide des algorithmes Berlekamp-Massey (BM) [137] et Chien [137].

Une vue d'ensemble du circuit BCH est présentée dans la figure 4.1 (a). Les signaux *nb_error_autotest* et *error_found_sel* permettent de déterminer le nombre d'erreurs à injecter dans la trame et leurs positions. *restart_autotest, input_en_hold_test* et *start_block_ext* sont des signaux de contrôle qui arrêtent ou redémarrent le processus de génération de trames et du décodage. Le signale *autotest_mode* contrôle le bloc hiérarchique *autotest_block* (cf. figure 4.1 (b)). Lorsque le signal *autotest_mode* est à '1', le générateur PRBS1 génère une trame qui est ensuite encodée. Au même temps, le générateur PRBS2 génère des erreurs dans des positions bien déterminées, indiquées par les signaux d'entrées. Ces deux trames sont fusionnées et injectées dans le décodeur qui effectue le décodage de la trame résultante (cf.

figure 4.1(b)). La trame initialement générée par le PRBS et celle issue après décodage sont comparées. Si elles sont identiques, cela indique que les processus de décodage et de détection d'erreur se sont effectués correctement. Ainsi, le signal *autotest_ok_o* se met à '1'. Le décodeur renseigne le nombre d'erreurs détectées et leurs positions par les intermédiaires des signaux de sortie *error_pos_o* et *error_number_o* (cf. figure 4.1(a)). Le signal *error_ready_o* quant à lui indique la fin du processus de décodage.

Nous signalons que le circuit BCH prend en charge un autre mode de fonctionnement qui est la prise en compte des trames injectées par l'utilisateur. Ceci est fait par le signal d'entrée *input_ext.* Dans notre cas, nous n'avons pas considéré ce mode, nous avons utilisé les trames générés du PRBS.

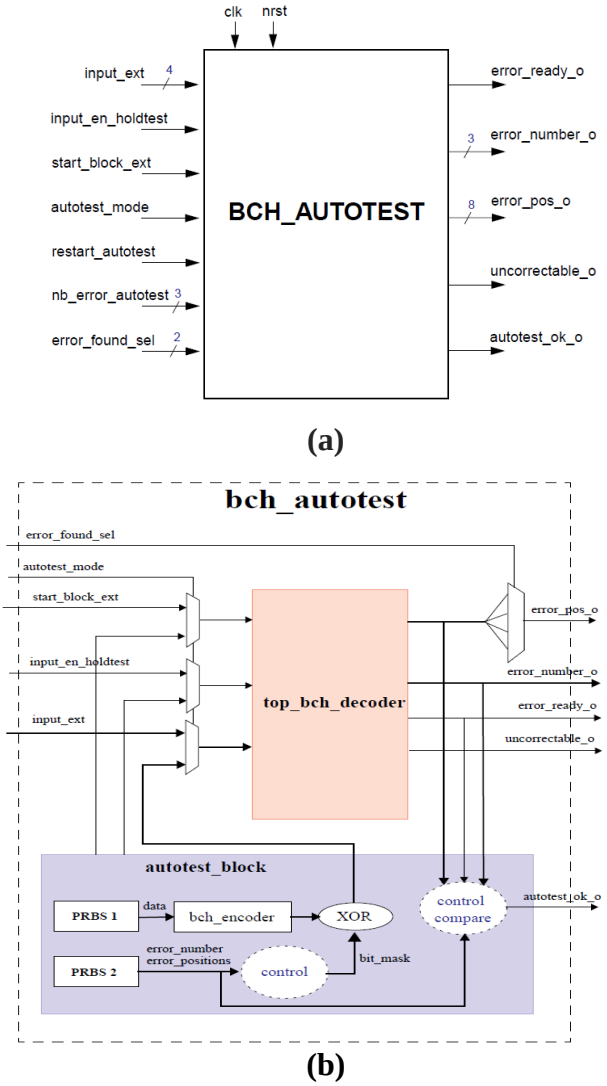


Figure 4.1 : (a) Vue d'ensemble du circuit BCH (b) Vue détaillée du Top level : les erreurs sont générées dans le bloc autotest_block et sont décodées dans le bloc top_bch_decodeur.

L'avantage du circuit BCH est la possibilité de contrôler les conditions d'usages du circuit en fixant le nombre d'erreurs à injecter dans les trames. Ainsi, nous pouvons mener une étude sur l'impact des conditions d'usage sur le fonctionnement du circuit.

Par ailleurs, nous avons développé un deuxième circuit, appelé PSR. Il s'agit d'un circuit qui réalise des opérations arithmétiques sur 4 étages (pipeline). Son principe est illustré dans la figure 4.2 (a). Il est composé de deux PRBS qui génèrent des trames A et B sur 32 bits. Des opérations arithmétiques sont effectuées dans chaque pipeline (cf. figure 4.2) et le résultat du dernier pipeline (B) est comparé à la trame généré par le PRBS 2. Un signal Autotest_OK indique que la comparaison a été effectuée correctement.

La motivation pour laquelle nous avons développé ce circuit, est la possibilité de contrôler les délais et le nombre de portes des chemins logiques du circuit. En effet, le dernier pipeline va contenir les retenues des opérations arithmétiques des précédents pipelines et donc les délais de ces chemins vont être les plus lents. Ainsi, il serait intéressant d'équiper les chemins du dernier pipeline avec des moniteurs.

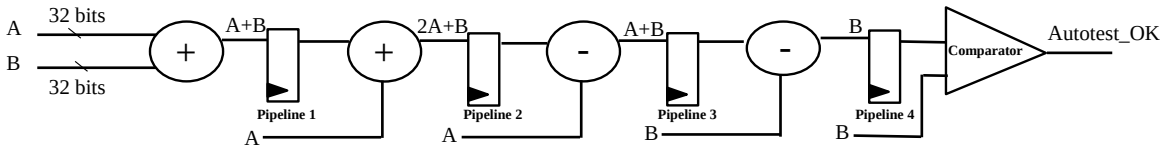


Figure 4.2 : Schématique du circuit PSR : des opérations arithmétiques sont effectuées sur les trames générées par les PRBSs.

4.3 Implémentation du moniteur in-situ

Le moniteur in-situ utilisé dans cette application est un moniteur de pré-erreur (cf. chapitre 2) basé sur l'élément de délai (cf. Figure 4.3). Nous rappelons qu'il est composé d'un élément de délai, d'une bascule dupliquée et identique à la bascule de référence pour éviter les problèmes de métastabilité ainsi qu'un comparateur pour comparer un signal avec sa valeur retardée par le passage dans un buffer de délai (TW). Si une transition arrive dans la fenêtre de détection TW, les signaux issus des bascules de référence et du moniteur vont être différents, ce qui entrainera une mise à '1' du signal Flag (cf. figure 4.4 (a)). Cependant, pour les transitions qui s'effectuent en dehors de la fenêtre de détection TW, elles ne seront pas détectées par le moniteur, ce qui maintient le signal Flag à '0' (cf. figure 4.4 (b)).
La largeur de la fenêtre de détection TW est déterminée par le délai du buffer. Plus elle est grande plus on détecte les transitions. Dans notre cas, nous explorons trois fenêtres de

détections : TW1= 60ps, TW2=100ps, TW3= 130ps sélectionnés à l'aide d'un multiplexeur. Ces délais sont donnés par la CAD à une tension de 0.9V dans le procédé SS.

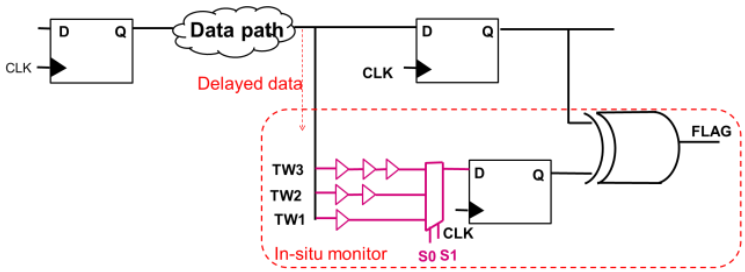


Figure 4.3 : Schéma du moniteur in-situ : il est composé d'un élément de délai (buffer), une bascule et un comparateur. Ici trois délais de la TW sont explorés.

Nous avons choisi la fenêtre de détection TW1 de 60ps car il s'agit du délai du plus petit buffer dans la librairie. Ainsi, nous limitions l'impact du moniteur sur le chemin monitoré car la capacité du fil est très faible. D'autre part, nous avons montré dans le chapitre 2 que la dégradation induite par le vieillissement à 10 ans présente 10% de délai supplémentaire. Pour le circuit BCH, 10% de délai supplémentaire représente 30ps. Nous avons décidé de couvrir cette dégradation par la deuxième fenêtre de détection TW2. Finalement, afin de prendre en compte la dispersion de mesure due à la variabilité locale et globale, nous avons aussi choisi un buffer plus grand correspondant à la taille de la fenêtre de détection de 130ps.

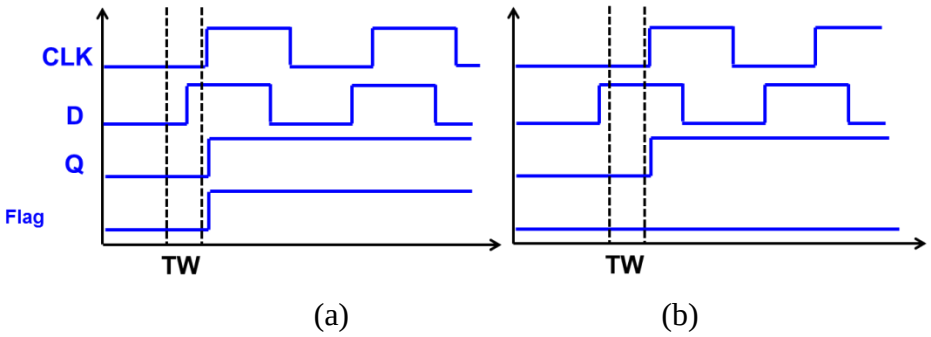


Figure 4.4 : Chronogramme du moniteur : (a) le signal Flag se lève car une transition du signal D se passe dans la fenêtre de détection. (b), la transition se passe en dehors de la fenêtre de détection, donc le signal Flag est maintenu à '0'.

Le moniteur de pré-erreur de la figure 4.3 a été intégré dans le circuit BCH. Nous avons implémenté sur silicium ce circuit dans les technologies 28 nm Low Power (LP) et 45nm. Pour ces deux technologies, nous avons choisi le corner FF_1.15V_125°C et SS_0.9V_m40°C pour considérer le meilleur cas et le pire cas respectivement. Les chemins

critiques monitorés sont sélectionnés à partir de la STA dans le pire corner SS_0.9V_m40°C. Concernant les moniteurs, nous avons utilisé la méthode d'implémentation dite layout LA et nous les avons intégré dans 10% des chemins critiques avec la méthode de ECO après la CTS. Dans la réalité, il faut mener une étude qui consiste à vieillir tous les chemins du circuit et considérer la valeur moyenne des délais induits par le vieillissement. Dans notre cas, nous avons choisi 10% des plus lents chemins car nous avons démontré dans le chapitre 2 qu'un chemin composé de 30 portes subit une dégradation de délai 10%. Ce pourcentage peut être appliqué sur le circuit BCH car les chemins critiques du circuit comptent en moyenne 30 portes logiques.

Les données d'implémentation sont présentées dans le tableau 3.1 du chapitre 3.

Nous signalons que le sign-off du circuit BCH de référence (sans moniteur) dans les technologies 28LP et 45nm est réalisé à des fréquences de 1Ghz et 500 MHz respectivement dans le pire corner. Par ailleurs, nous avons montré dans le chapitre 3 que la méthode dite layout LA induit une dégradation des délais des chemins du circuit. C'est pourquoi nous avons effectué le sign-off du circuit BCH avec les moniteurs à 400Mhz dans les technologies 45nm.

Concernant le circuit PSR, il a été implémenté dans la technologie 28FDSOI dans les corners FF_1.15V_125°C et le SS_0.9V_m40°C. La fréquence du sign-off dans le pire corner est de 1Ghz. Après implémentation, nous comptons 44327 cellules standard et 8900 bascules. Nous avons inséré 250 moniteurs dans les chemins les plus lents obtenus après optimisation de la CTS, ces moniteurs sont basés sur la méthode dite layout LB.

Les résultats expérimentaux de ces deux circuits avec les moniteurs sont présentés dans les prochains paragraphes.

4.4 Résultats expérimentaux

4.4.1 Détermination de la plage de fonctionnement du BCH à l'aide des moniteurs

Dans ce paragraphe, nous étudions la plage de fonctionnement du BCH à l'aide des moniteurs dans la technologie 45nm dans différentes conditions d'usage et de températures.

Dans un premier temps, nous nous intéressons à la plage de fonctionnement du circuit BCH dans une seule puce, à température ambiante. En effet, nous avons fait varier à la fois la tension et la fréquence, et pour chaque couple tension-fréquence, nous avons compté le

nombre de mise à '1' du signal Flag. Aussi, nous avons mis en œuvre des conditions de fonctionnement extrême du circuit BCH en le poussant à des tensions/fréquences au-delà des conditions du sign-off. Les fréquences considérées varient de 250 Mhz à 650 Mhz et de 0.8V à 1.2V pour la tension.

Le résultat de cette expérience est présenté dans le Shmoo de la figure 4.5, il a fait l'objet d'une publication à IRPS 2016 [139]. La plage de la tension/fréquence qui garantit un fonctionnement correct est schématisé par la couleur verte, i.e. Autotest_OK à '1'. La couleur rouge indique la zone de dysfonctionnement du circuit, i.e. Autotest_OK à '0'. Dans les conditions du sign-off (400Mhz, 0.9V) obtenues par la CAD, le fonctionnement du circuit est correcte. Si on se positionne dans cette tension et on augmente la fréquence, nous remarquons que le nombre de Flags à '1' devient important. La première fréquence de dysfonctionnement est reportée à 600Mhz.

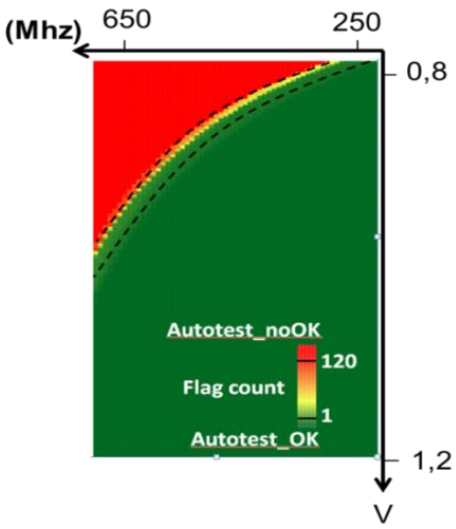


Figure 4.5 : Shmoo des signaux Flags en fonction de la tension et la fréquence : la plage de détection des moniteurs est plus abrupte pour les forts V_{DD}. Le nombre de Flags croit avec la réduction de la tension (partie délimité par les pointillés) [139].

La partie délimitée par les traits en pointillé représente donc la plage tension-fréquence dans laquelle le BCH fonctionne mais avec une présence des Flags. Cette plage garantit toujours un fonctionnement correct avec des performances optimales car il s'agit d'une zone proche du dysfonctionnement.

Par conséquent, nous pouvons extraire des conditions d'opération juste de type tension-fréquence sans qu'il y ait besoin de rajouter de marges de tension/fréquence supplémentaires.

Le comportement des moniteurs observé par le biais de la figure 4.5, c'est-à-dire l'apparition des signaux Flags avant le dysfonctionnement du circuit, doit être similaire pour toutes les puces. Nous avons vérifié cette propriété sur 160 puces. Dans cette expérience, nous avons

fixé la fréquence à 400MHz et avons réduit la tension avec un pas de 5 mV en allant de 1V jusqu'au dysfonctionnement du circuit. Pour chaque tension, nous avons relevé les éventuelles tensions minimales V_{DDmin} du premier, 10ème et 50ème Flags comptés ainsi que celle du signal Autotest_fail (signal qui montre le dysfonctionnement). Intuitivement, le nombre de Flags devrait augmenter jusqu'au dysfonctionnement du circuit. Ceci est illustré dans la figure 4.6 [139] qui présente les distributions cumulées CDF en fonction des V_{DDmin} pour 160 puces. Nous remarquons que pour toutes les puces, le V_{DDmin} du premier Flag est toujours inférieur au 10ème, au 50ème, et au signal Autotest_Fail. Ceci est une preuve que le circuit ne présente jamais de dysfonctionnement avant l'apparition des alertes des moniteurs. Nous remarquons aussi une dispersion de mesure des V_{DDmin} qui est due aux variabilités locales et globales. Ceci justifie et renforce la conclusion que les moniteurs sont une bonne solution pour compenser les effets de la variabilité sur le circuit.

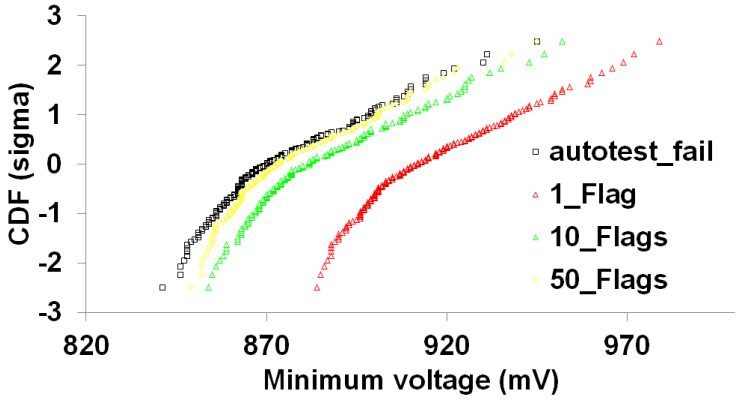


Figure 4.6 : Distribution expérimentale des V_{DDmin} du 1^{èr} ,50ème Flag et du signal Autotest_fail du circuit BCH [139].

4.4.2 Comportement du circuit face aux conditions d'usage

Dans ce paragraphe, nous proposons d'étudier le comportement du circuit en fonction des conditions d'usage du circuit, i.e. le vecteur de test (Workload). Pour ce faire, nous avons utilisé 3 circuits BCH dans le Testchip en 45nm (cf. figure 4.7(a)) et nous avons testé les BCHs sous différentes conditions d'usage. Ces conditions d'usage correspondent aux différents nombres d'erreurs injectés dans les trames du circuit. Nous désignons par Workload A, B, C, D, E les trames contenant 1, 2, 3, 4, 5 erreurs respectivement. Ces erreurs doivent être encodées, décodées et puis corrigées.

Dans un premier temps, nous nous proposons de faire varier la tension d'alimentation afin de mesurer les V_{DDmin} qui garantissent un fonctionnement correct des BCHs. Ces mesures ont été effectuées à une fréquence de 400Mhz et en tenant compte de deux vecteurs : A et B. La tension d'alimentation, quant à elle, a été réduite avec un pas de 5mV à partir de la tension 1V jusqu'à ce que le signal Autotest_Fail se lève. Ces mesures ont été appliquées sur 160 puces qui contiennent 3 BCHs chacune. Le résultat des mesures est illustré dans la figure 4.6 (a) et a fait l'objet d'une publication à la conférence IVSW 2016 [140]. Il montre que pour les points qui se situent dans le cadrant A, les V_{DDmin} du BCH2 sont supérieurs à ceux du BCH1. Alors que pour ceux qui se situent dans le cadrant B, il s'agit des V_{DDmin} du BCH1 qui sont supérieures.

Par ailleurs, nous remarquons que pour le Workload_A les V_{DDmin} sont corrélés alors qu'ils le sont moins pour le Workload_B (cf. figure 4.7(a)). En effet, la dispersion des V_{DDmin} est plus importante pour le Workload_B, elle représente 60mV alors qu'elle est de 10 mV pour le Workload_A.

Nous signalons que les mêmes tests ont été effectués sur le BCH3 et ont montré le même comportement illustré dans la figure 4.7(b).

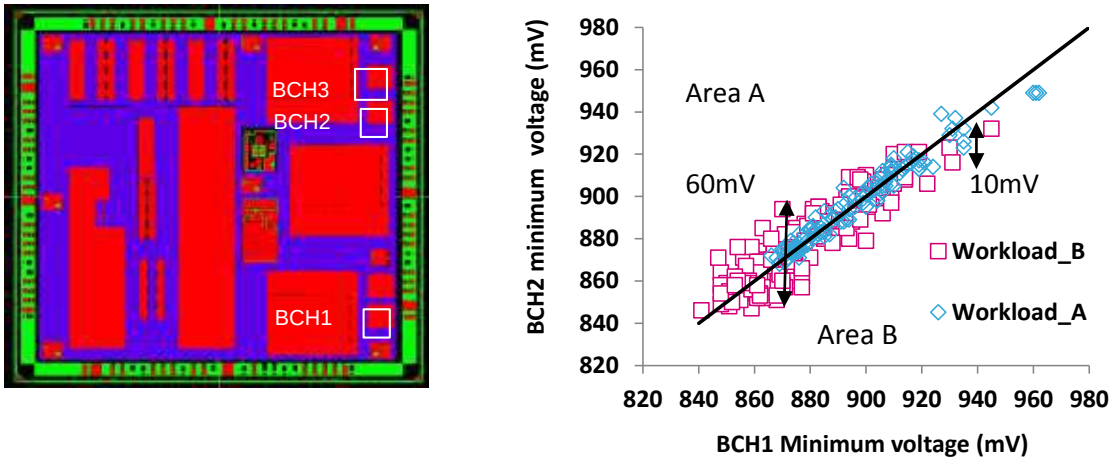


Figure 4.7 : (a) Le Floorplan du Testchip en 45nm. Le BCH est placé à trois endroits dans le TOP afin d'étudier l'impact de la variation locale. (b)mesures des V_{DD} de deux BCHs pour deux workloads

Nous pouvons donc conclure que la complexité du vecteur contribue à l'accentuation de la dispersion et que la nécessité d'utiliser des moniteurs in-situ évoqué dans le chapitre 2 est encore une fois justifiée par ces mesures. L'usage des moniteurs in-situ ne peut être que

bénéfique pour compenser ces marges car pour prendre en compte les dispersions, il faut considérer des marges de V_{DD}.

Nous nous proposons maintenant de vérifier le comportement des moniteurs obtenu dans la figure 4.6 c'est-à-dire l'apparition du signal Flag avant le dysfonctionnement du circuit, mais cette fois en tenant compte des conditions d'usage. Pour ce faire, nous avons adopté la même expérience présentée au début du paragraphe, i.e. réduction du V_{DD} jusqu'au dysfonctionnement du circuit à la fréquence de 400Mhz. Au cours de la réduction de tension, nous avons relevé le V_{DDmin} du premier Flag pour les deux vecteurs. Le résultat est présenté dans la figure 4.8 et a été publié à IRPS 2016 [139], il montre que tous les points de mesure se situent en dessous de la droite de corrélation. Ceci est une preuve que le V_{DDmin} du signal Flag est toujours inférieurs à celui du signal Autotest_Fail pour les deux vecteurs et donc nous pouvons conclure que le signal Flag se lève toujours avant le dysfonctionnement du circuit et quel que soit le vecteur.

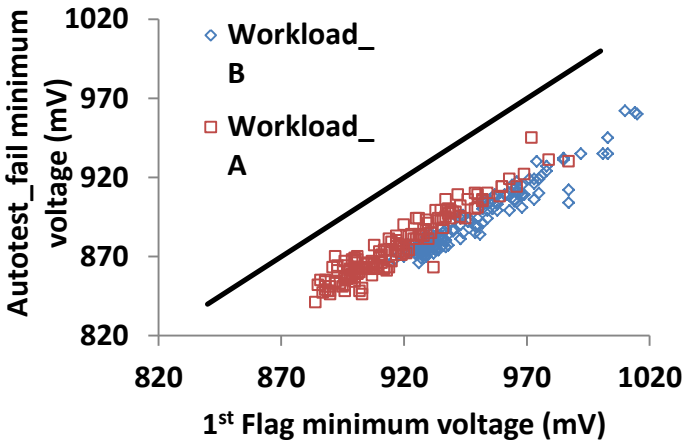


Figure 4.8 : Mesures des V_{DDmin} du 1^{èr} Flag et du signal Autotest_fail, pour 160 puces, suite à l'application de deux vecteurs.

Nous nous intéressons maintenant à l'étude du comportement des moniteurs en fonction des conditions d'usage lorsqu'on applique la réduction de la tension. Pour cela, nous avons utilisé 5 vecteurs dans le circuit BCH de type Workload. Pour chacun, nous avons réduit la tension jusqu'au dysfonctionnement du circuit, i.e. Autotest_fail = '1' et avons compté le nombre de Flags. Le résultat est présenté dans la figure 4.9 (a) [139], il montre que la tension V_{DDmin} des Flags dépend du vecteur. En effet, les V_{DDmin} sont plus importants pour le Workload_E. Particulièrement pour le premier Flag, les V_{DDmin} sont de 0.84V et 0.86V pour les Workload A et E respectivement. Ce résultat est attendu car le Workload_E contient un nombre important d'erreurs et donc un grand nombre de chemins dans le BCH sont activés pour la correction de ces erreurs. Ceci se répercute sur la mesures des tensions minimales.

Nous remarquons aussi que le nombre de Flag à '1' dépend du vecteur, il varie entre 20 et 40. Aussi, il connait une fluctuation (cf. figure 4.9 (b)): il est de 40 à 0.81V pour le Workload_E, il diminue puis il augmente avec la réduction de la tension.

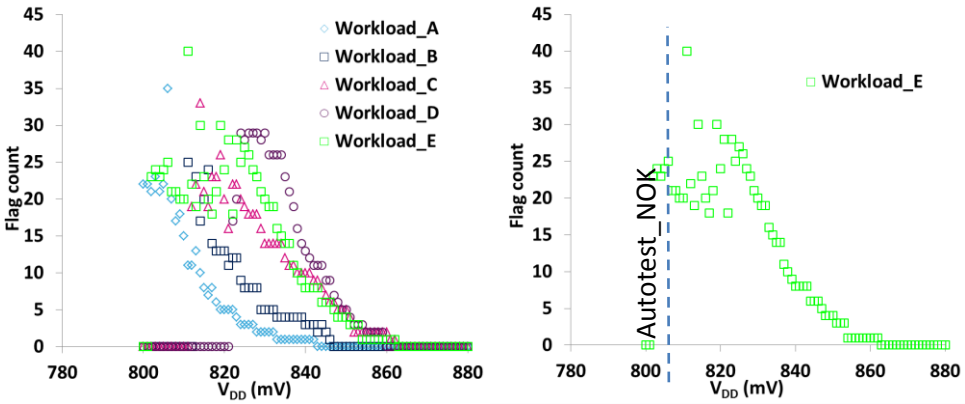
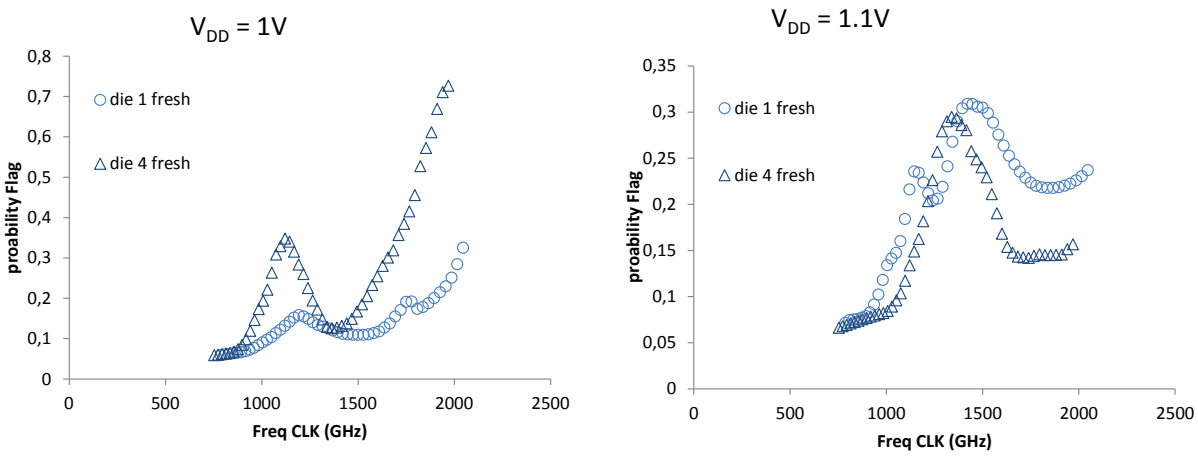


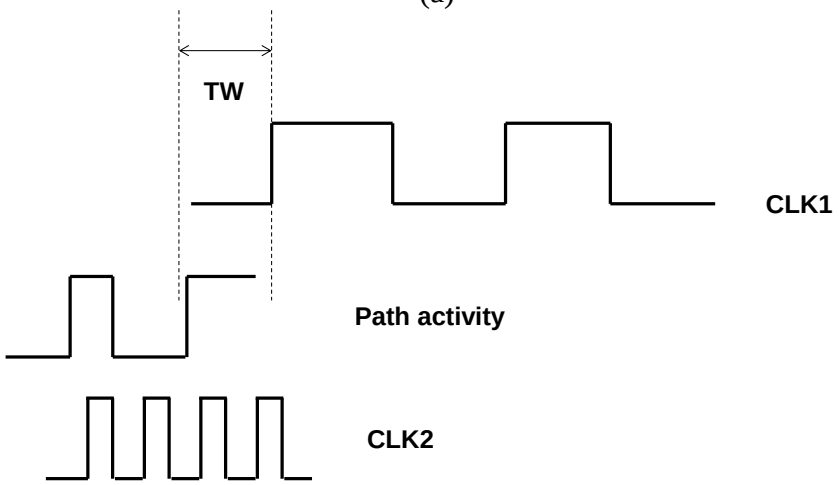
Figure 4.9 : Calcul du nombre de Flags en fonction de la tension V_{DD} pour différentes conditions d'usage (workload) [139].

Ce comportement a été observé aussi dans le circuit PSR en 28FDSOI (cf. figure 4.10 (a)) à différentes tensions de fonctionnement comme le montre la figure 4.10. Les mesures reportées dans cette figure ont été obtenues en réduisant la fréquence et en calculant la probabilité d'observer des Flags. L'explication de cette observation est illustrée dans la figure 4.10(b). En effet, pour les fréquences qui garantissent un fonctionnement correct, i.e. sans erreurs de setup, nous comptons des Flags lorsque des transitions arrivent dans la fenêtre de détection. Ce qui justifie l'accroissement du nombre des Flags. Quand la fréquence est réduite à tel point que les erreurs de setup se produisent, les transitions ne sont plus détectées, ce qui explique la diminution du nombre de Flags comptés. Par ailleurs, l'augmentation du nombre de flags comptés à haute fréquence s'explique par le fait que la détection et le comptage des Flags sont fait sur plusieurs cycles d'horloge.

D'une manière générale, lorsque la tension est trop réduite, la fenêtre de détection du moniteur se situe dans la zone d'activité du chemin et donc nous détectons des transitions qui constituent une fausse alerte par les moniteurs. Pour une application de type AVS ou DVFS, la fréquence minimale à considérer est celle qui génère le premier pic du nombre de Flags comptés.



(a)



(b)

Figure 4.10 : (a) probabilité d'observation des alertes des moniteurs lors de la réduction la fréquence pour deux tensions. (b) Explication de la fluctuation du nombre de Flags lors de la réduction de la fréquence

La figure 4.9 montre aussi que le nombre maximum de Flags que nous avons compté est d'environ 45 alors que nous avons inséré les moniteurs dans 100 chemins critiques. A l'issue de ce résultat, nous nous posons les questions suivantes :

a. Quels sont les index des moniteurs activés et ceux qui ne sont pas activés par un vecteur spécifique ?
b. Quels sont les chemins qui ont plus d'influence sur le circuit ?
c. Est-ce que les chemins activés durant le mode 'test et mesure' ont été identifiés par l'algorithme détaillé dans le chapitre 3 ?

Les réponses à ces questions seront discutées lors de présentation des résultats dans le paragraphe 4.4.5.

4.4.3 Comportement du circuit face à la température

Nous nous proposons maintenant d'étudier l'impact de la température sur le comportement du circuit à l'aide des moniteurs in-situ. Pour cela, nous avons mené la même étude présentée dans le paragraphe 4.4.2, à savoir la réduction de la tension V_{DD} et le calcul du nombre de Flags. Aussi, nous avons considéré deux vecteurs (Workload A et B) à deux températures : 30°C et 125°C. Le résultat est illustré dans la figure 4.11 [140], il montre deux dépendances : au vecteur de test et à la température. La dépendance au vecteur a été discutée dans la figure dans le paragraphe 4.4.2 (cf. figure 4.9). Concernant la dépendance à la température, la figure 4.11 montre que plus la température est élevé plus les V_{DDmin} des Flags comptés sont importants et ce pour les deux vecteurs. Particulièrement, pour le premier Flag, les V_{DDmin} sont à 0.89V et 0.85V respectivement pour les températures 125°C et 30°C. Ce constat est cohérent car lorsqu'on augmente la température, le délai des portes logique devient plus grand (cf. chapitre 1) ainsi que les délais des chemins de propagation du circuit et donc les V_{DDmin} deviennent plus importants. Par ailleurs, la constatation sur la fluctuation du nombre de Flags est toujours vérifiée pour les deux températures.

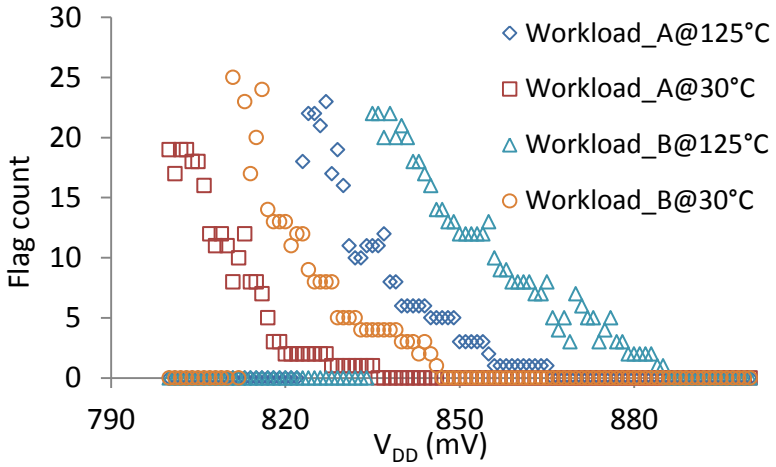


Figure 4.11 : Calcul du nombre de Flags à différentes températures et conditions d'usage

4.4.4 Moniteurs in-situ et impact de la fenêtre de détection

Dans le paragraphe 4.3 de ce chapitre, nous avons justifié le choix de l'utilisation des trois fenêtres de détection du moniteur, il est important d'étudier l'impact de la taille des fenêtres de détection sur la détection des signaux Flags. Pour ce faire, nous avons considéré l'expérimentation précédente qui consiste à relever les V_{DDmin} des Flags qui se déclenche en réduisant la tension d'alimentation. Les mesures ont été effectuées en considérant un seul vecteur, Workload_B en l'occurrence et les trois fenêtres de détection. La détection des pré-erreurs est donnée par la mesure de la tension minimale des Flags. En effet, les chemins

monitorés qui se déclenchent en premier, voit des V_{DDmin} plus élevé. Cette détection est impactée par la taille de la fenêtre de détection (TW).

Le résultat des mesures est donné dans la figure 4.12 [139], il montre que les V_{DDmin} des Flags les plus élevés sont observés pour la fenêtre de détection la plus large, i.e. TW3. Ce constat est attendu car plus la fenêtre est large plus, rapidement on détecte les transitions. Cependant, pour les fenêtres de détection TW1 et TW2, nous observons des tensions V_{DDmin} quasi identique. Ceci est dû probablement au fait que pour l'activité du vecteur de test considéré, les transitions arrivent dans un temps couvert par les deux fenêtres de détection.

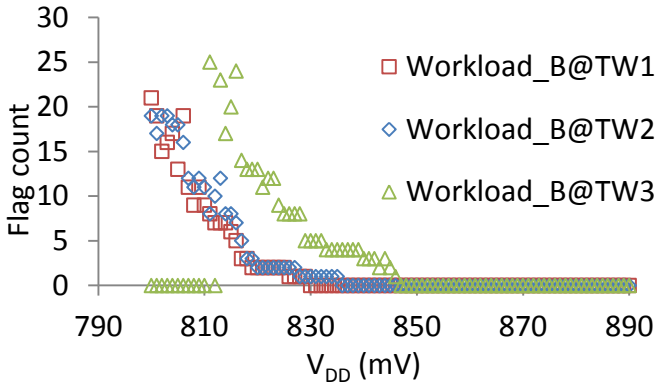


Figure 4.12 : Calcul du nombre de Flags en fonction de la taille de la fenêtre de détection. Ici, nous considérons trois TWs : TW1= 60ps, TW2=100ps, TW3=130ps

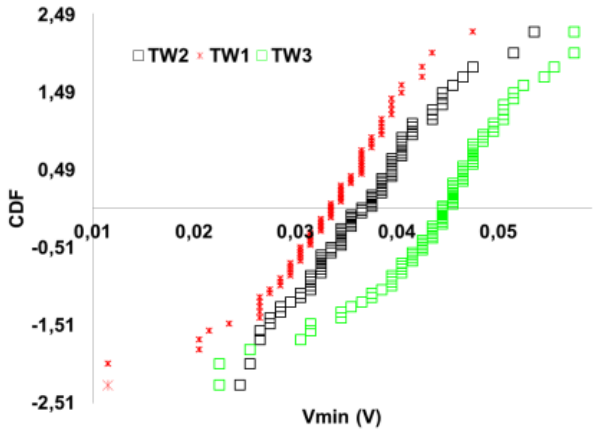


Figure 4.13 : Mesures expérimentales des fenêtres de détection du BCH.

Nous nous proposons maintenant de mesurer expérimentalement la fenêtre de détection afin d'étudier l'impact de la dispersion sur sa largeur. Ces mesures sont effectuées en réduisant la tension et en relevant le V_{DDmin} du premier Flag et du signal Autotest_OK. La fenêtre de détection correspond à la différence de tension de ces deux signaux. Les résultats des mesures sont illustrés dans la figure 4.13 [139], ils concernent 160 puces et sont présentés sous forme d'une distribution cumulée. Ils montrent que la dispersion est plus grande pour la fenêtre de

détection TW3. Ceci est un comportement attendu car plus la TW est large plus la probabilité d'observer des Flags est élevée.

Au vu des résultats présentés dans la figure 4.12 et 4.13, il en résulte une meilleure utilisation des moniteurs avec une grande fenêtre de détection car elle permet de prendre en considération la dispersion des mesures dues aux variabilités locales et globales.

Par ailleurs, les petites fenêtres de détections permettent de détecter des transitions qui arrivent près du front d'horloge. Ainsi, le réglage de la tension sera très fin. Il s'agit certes de la meilleure solution pour réduire la consommation mais son inconvénient est le fonctionnement près du POFF.

En conclusion, pour les circuits à faible slack, il faut considérer une marge de sécurité afin de ne pas opérer dans une zone très proche de POFF lorsqu'on applique AVS. Ceci est possible avec une grande fenêtre de détection.

4.4.5 Moniteur in-situ et impact des conditions d'usage

Dans le paragraphe précèdent, nous avons présenté les mesures des paramètres du circuit BCH dans différents PVTs à l'aide des moniteurs in-situ. A présent, nous allons nous intéresser à l'impact de l'activité sur le circuit à l'aide des moniteurs in-situ dans les mêmes corners PVTs. En effet, nous allons mesurer expérimentalement la tension V_{DDmin} de chaque moniteur avec les mêmes conditions décrites dans le paragraphe précédent. Ce calcul est réalisé à l'aide d'une méthode de test qui permet de donner l'index de chaque moniteur lorsque le signal Flag correspondant passe à '1'. Aussi, nous avons calculé ces tensions V_{DDmin} avec le simulateur SPICE afin de confronter les résultats des mesures à la simulation. Nous signalons que cette étude est réalisée en réduisant la tension avec un pas de 5mV à partir de la tension 1V jusqu'au dysfonctionnement du circuit et en considérant deux usages du circuit BCH : standard et maximal. Ces derniers correspondent aux trames contenant 0 et 3 erreurs respectivement.

Le résultat de la variation de la tension est présenté dans la figure 4.14 (a) [140], il montre les signatures des moniteurs et met en évidence leurs capacités à réagir en fonction de l'usage du circuit : pour un usage maximal du BCH, le nombre de moniteurs activés ayant un grand V_{DDmin}, est nettement supérieur à ceux obtenus pour un usage standard. Les faibles V_{DDmin} signifient que les chemins ne sont pas très activés. Ce résultat est attendu car plus l'usage du circuit est complexe, plus le nombre de chemins sollicités est important. Cela a été démontré

aussi dans le paragraphe 4.5 et a été reconfirmé à travers les mesures de V_{DDmin} pour un vecteur donné.

Ces résultats de test ont permis de déceler entre autre les positions des moniteurs dans les différents blocs du BCH (cf. figure 4.14 (b)). Nous constatons que la plupart des chemins se trouvent dans le décodeur, plus précisément dans le bloc Chien pour un usage critique. Ce qui prouve que l'opération de recherche de la position d'erreur dans la trame est la plus délicate et nécessite beaucoup de ressources.

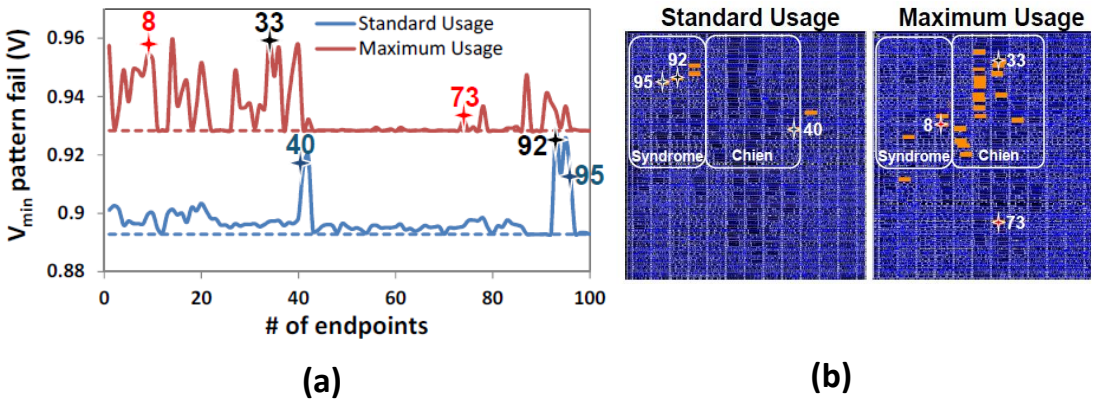


Figure 4.14 : (a) mesures expérimentales du V_{DDmin} en fonction des signatures des moniteurs pour un usage standard et maximal du BCH : l*a dépendance activité-Vecteur est mise en avant* (b) *Les positions des moniteurs dans le BCH pour les deux usages [140].*

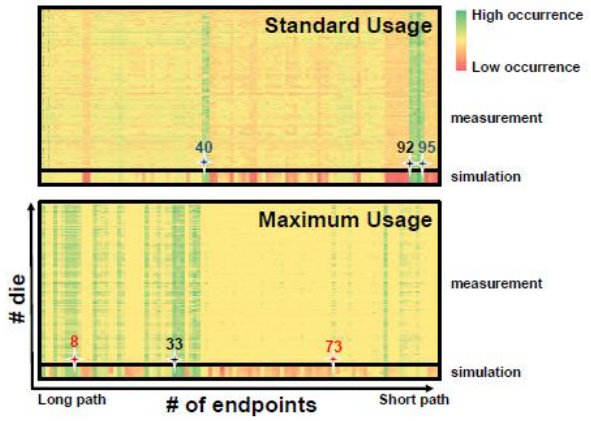


Figure 4.15 : Mesure des V_{DDmin} des moniteurs dans 160 puces pour différents usages et comparaison avec la simulation SPICE.

Nous avons vérifié les réponses des moniteurs obtenues dans la figure 4.14 pour 160 puces et deux usages (cf. figure 4.15) [140]. Les moniteurs des chemins 40, 92 et 95 sont toujours les plus actifs pour l'usage standard alors que les moniteurs 8 et 33 le sont pour l'usage maximal. Les mêmes mesures ont été validées avec la simulation SPICE et montrent une parfaite

corrélation avec les mesures expérimentales. Il faut aussi noter que plus le chemin est critique plus le V_{DDmin} est élevé.

Nous nous intéressons maintenant à l'impact de la fenêtre de détection et de la température sur les moniteurs. Nous avons effectué les mêmes mesures reportées sur la figure 4.14 mais cette fois-ci en prenant en compte les trois fenêtres de détection ainsi que l'usage maximale du circuit. Concernant l'étude en température, elle a été appliquée sur le circuit BCH fonctionnant dans les conditions standards. Les températures considérées sont 25°C et 125°C.

Le résultat de la variation de la tension dans ces conditions est donné dans la figure 4.16. Il met en évidence l'impact de la fenêtre de détection (cf. figure 4.16(a)) discuté dans le paragraphe 4.4.4. En effet, nous observons que plus la fenêtre de détection est large, plus le V_{DDmin} est élevé. Ce résultat est en adéquation avec les mesures présentées dans la figure 4.12. Ceci s'explique par le fait que les transitions sont détectées plus rapidement lorsque la fenêtre de détection est large. Ce qui justifie la valeur élevé du V_{DDmin} pour les larges fenêtres de détection.

Par ailleurs, l'effet de la température sur le V_{DDmin} est illustré dans la figure 4.16(b). En effet, nous remarquons que les V_{DDmin} les plus élevés sont observés pour la température 125°C. Ce résultat est attendu car la température augmente les délais des chemins (cf. chapitre 1) et ceci se répercute sur la mesure des V_{DDmin}. Il faut aussi noter que la signature des moniteurs, étant dépendante de l'usage, reste identique quelle que soit la température.

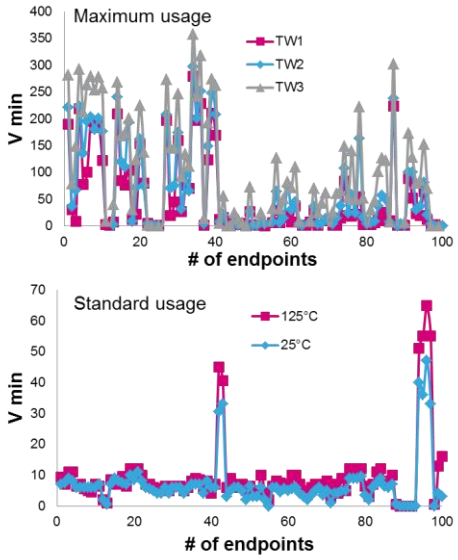


Figure 4.16 : Mesure du V_{DD} des moniteurs en fonction de la fenêtre de détection (a) et de la température (b)

Nous proposons maintenant de comparer les délais des chemins du BCH obtenus par simulation et ceux mesurés à l'aide des moniteurs. Pour cela, nous avons commencé par

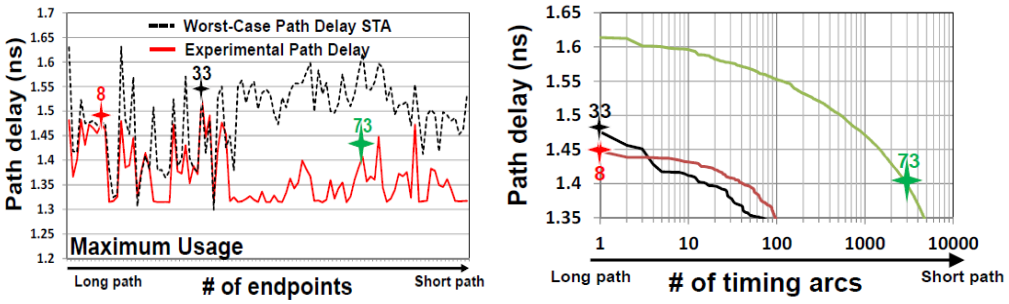
extraire la dépendance F(V) du circuit BCH en mesurant le V_{DDmin} de chaque moniteur à trois fréquences de PLL. Ensuite, nous avons déduit le délai du chemin en soustrayant le délai de la fenêtre de détection (T_{TW}) au délai du moniteur ($T_{moniteur}$). Nous rappelons que la relation qui lie la fréquence maximale du chemin (F_{max}) et le délai du moniteur s'exprime par l'équation 4.1 :

$$F_{max} = \frac{1}{T_{moniteur}} = \frac{1}{T_{chemin} + T_{TW}} \qquad \text{EQU4.1}$$

Ainsi nous pouvons confronter les délais expérimentaux à ceux de la STA.

Le résultat qui illustre les délais expérimentaux et ceux calculés par la STA est donné dans la figure 4.17 (a). Notons que pour les délais calculés par la STA, nous avons tracé seulement le délai du pire arc par bascule.

La figure montre que pour certains chemins, i.e. chemins 8 et 33, les délais mesurés et calculés par la STA corrèlent parfaitement. Ceci prouve que pour ces bascules, il s'agit des pires chemins qui sont activés. Pour d'autres chemins, par exemple le chemin 73, nous remarquons une différence entre la mesure et la STA. Ceci est dû au fait qu'il ne s'agit pas du pire chemin sur la bascule activée. Afin de trouver exactement le rang du chemin monitoré 73 activé, nous avons extrait pour ce chemin les délais de tous les arcs par la STA. Le résultat est donné dans la figure 4.17(b), il montre que pour le chemin 73, il s'agit du 3000ème arc qui est activé.



(a) **(b)**

Figure 4.17 : (a) délais des chemins monitorés mesurés expérimentalement et obtenus par la STA.(b), les délais des arcs des chemins 33, 8 et 73 extraits par STA.

Ce résultat est important car il démontre qu'il ne s'agit pas toujours du chemin le plus lent par bascule qui est activé, mais des chemins sous-critiques dont le délai est bien inférieur au critique. Si l'usage du circuit était connu à la phase du design, des marges de délai pourraient être appliquées sur les chemins activés, plutôt que sur le circuit en entier. Aussi, les efforts d'optimisation (setup, hold) seraient concentrés uniquement sur ces chemins, évitant les pénalités de surface et de consommation engendrées par l'application des marges excessives

sur tout circuit. Ceci prouve aussi que l'utilisation du critère 2 de l'algorithme présenté dans le chapitre 3 est pessimiste car elle prend en compte toujours le pire chemin alors qu'il n'est pas forcément le plus active.

Ce résultat nous permet aussi de raffiner le choix de la taille de la fenêtre de détection TW du moniteur. En effet, elle doit dépendre de la fréquence ciblée f_{target} à laquelle le circuit doit toujours fonctionner. Afin de garantir le fonctionnement à cette fréquence, la taille de la fenêtre de détection TW doit être égale à $1/(f_{target} - f_{path})$. f_{path} étant la fréquence du chemin activé qui peut être obtenue par la STA. Ce délai de la fenêtre de détection calculé est valable pour un usage particulier. Comme il doit prendre en compte le multi usage du circuit ainsi que les dispersions, nous devons considérer une fenêtre de détection suffisamment grande.

A l'issue de la méthode de test développée et des résultats obtenus concernant les chemins critiques activés et détectés, nous pouvons répondre aux questions posées au début du paragraphe, à savoir l'identification des chemins activés en fonction de l'usage du circuit et la sélection des chemins susceptible d'être critique dans la phase de la conception. Pour cette dernière, nous constatons que tous les chemins activés dans la figure 4.16, ont été identifiés et considérés comme critiques selon l'algorithme du chapitre 3. Toutefois, on remarque que le nombre de chemins identifiés par l'algorithme de l'approche statistique est supérieur à ceux extraits expérimentalement. Nous rappelons que nous avons proposé une sélection de 73 registres avec l'algorithme de l'approche statistique alors que nous comptons 2 et 50 moniteurs activés avec l'usage standard et maximal respectivement. Cet écart s'explique par l'hypothèse de l'équiprobabilité des arcs que nous avons considérée pour attribuer le poids à chaque registre. Ce problème de sur-sélection aurait été résolu si nous avions considéré l'activité de chaque arc.

A travers les résultats présentés, nous avons démontré la capacité des moniteurs à fonctionner dans différentes conditions environnementales. Dans le prochain paragraphe, nous allons nous servir des moniteurs pour mettre en œuvre les problématiques de dispersion.

4.5 Etude des dispersions à l'aide des moniteurs in-situ

Afin de mettre en œuvre les problématiques des dispersions locales et globales, nous proposons une série des mesures des tensions minimales du circuit BCH lorsque nous

réduisons la tension d'alimentation. Les mesures ont été effectuées sur des BCHs implémentés dans les technologies 28LP et 45nm.

Dans un premier temps, nous étudions la dispersion globale en mesurant le V_{DDmin} de 24 tuiles de BCH fabriquées en 28LP. Dans cette application, nous avons démarré d'une tension de 1V et une fréquence de 1GHz et avons ensuite réduit la tension jusqu'au dysfonctionnement du circuit, i.e. la mise à '0' du signal Autotest_OK. Le résultat est donné dans la figure 4.18 [141], et il montre que pour la première puce, le signal Flag se lève à 0.9V. A cette tension, 99% des puces n'enregistrent pas de signal Flag. Lorsque nous continuons à réduire la tension, de plus en plus de Flag apparaissent dans un grand nombre de puces et le premier dysfonctionnement (violation de setup) est reporté à 0.85V. Notons qu'à cette tension, 95% des puces fonctionnent correctement.

La figure 4.18 montre aussi que la marge de tension entre le premier Flag qui se lève et le dysfonctionnement du circuit est d'environ 100mV en moyenne. Cette marge peut être exploitée par l'utilisation de la technique AVS à l'aide des moniteurs in-situ.

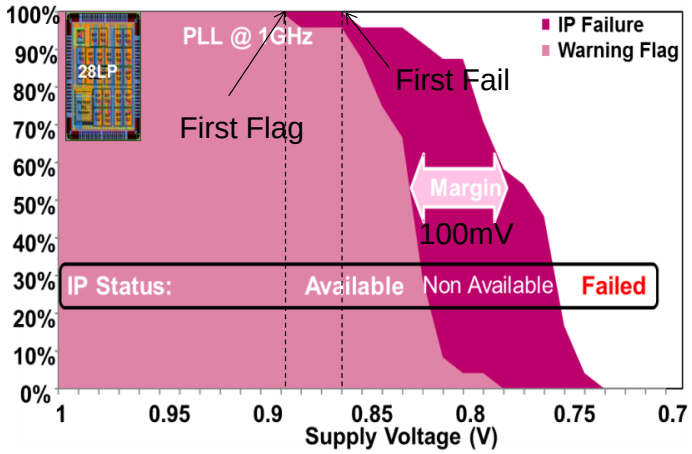


Figure 4.18 : Distribution des V_{DDmin} des BCHs : Lorsqu'on réduit la tension les Flags apparaissent avant le dysfonctionnement du circuit.

La dispersion des mesures des tensions minimales du BCH dans les différentes tuiles est probablement due à la dispersion des délais présentés par les chemins du circuit. Cette dispersion peut être d'origine locale. Nous avons vérifié cette hypothèse en mesurant les délais de 20 chemins du circuit BCH dans 24 différentes tuiles. Les mesures de ces délais sont présentées dans la figure 4.19 sous forme d'un pourcentage de déviation par rapport au chemin critique de chaque tuile. Nous observons que la déviation du délai du premier chemin dans les 24 tuiles varie entre 2% et 8%. Par ailleurs, concernant la dispersion inter-tuile, des chemins peuvent être critique dans certaines tuiles et deviennent moins critique dans d'autres. C'est le cas des chemins 10 et 15 dans les tuiles 8 et 11. Afin d'analyser les fréquences des

chemins mesurées dans les différentes tuiles, nous présentons dans la figure 4.20 la distribution cumulée des fréquences de deux chemins dans les 24 tuiles. Il s'agit du chemin le plus rapide et le plus lent dans la tuile 1. Ces chemins présentent une dispersion d'environ 50MHz.

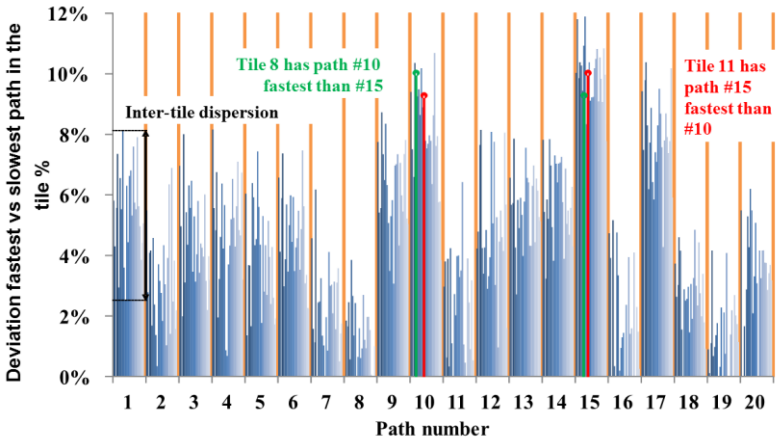


Figure 4.19 : Mesure des délais normalisés de 20 chemins dans les 24 tuiles.

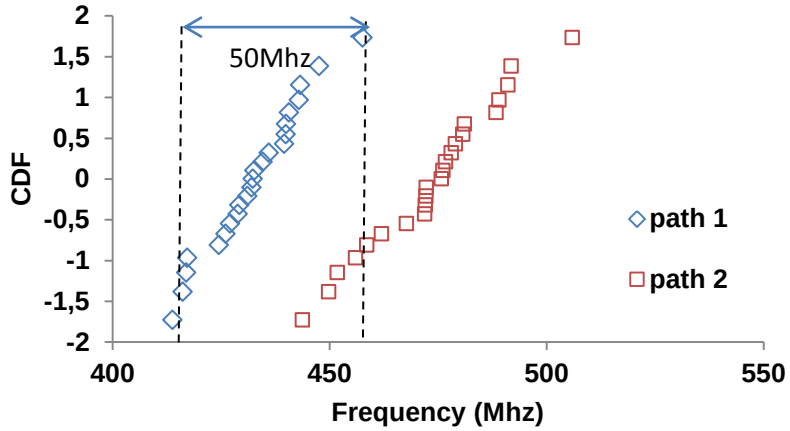


Figure 4.20 : distribution cumulée des fréquences dans 24 tuiles pour deux chemins

Jusqu'ici nous avons présenté chaque type de dispersion, i.e. locale et global, à part. La question qu'on peut se poser est de savoir si les deux types de dispersions ont la même contribution dans le circuit mesuré.

Afin de répondre à cette question, nous avons mesuré les délais de deux chemins dans le BCH en 45nm dans trois plaques de silicium. Il s'agit des chemins monitorés 95 et 40 qui sont les plus actifs dans le cas d'usage standards (cf. figure 4.15). Notons que les mesures effectuées sur ces plaques incluent les deux types de variations, locales et globales. Par ailleurs, nous avons aussi simulé l'impact de la variation locale avec la méthode Monte-Carlo dans les corners TT, FF et SS. Le résultat est donné dans la figure 4.21, il montre une grande

dispersion de mesures centrée autour du corner typique. Cette dispersion est d'environ 1.7ns. D'autre part, la dispersion locale ne dépasse pas 0.2ns pour les trois types de corner. Nous remarquons aussi que les dispersions des mesures sont plus importantes que celle simulées avec Monte-Carlo, nous pouvons conclure que la part de la variation globale est dominante.
Malgré la dispersion de mesures observée dans la figure 4.21, le flot de conception et la modélisation des PVTs utilisés par STMicroelectronics permettent de couvrir ces dispersions car les mesures sont contenus entre les délais simulés des corners SS et FF. Ceci justifie d'une certaine façon la pertinence et la fiabilité de la méthode MMMC.

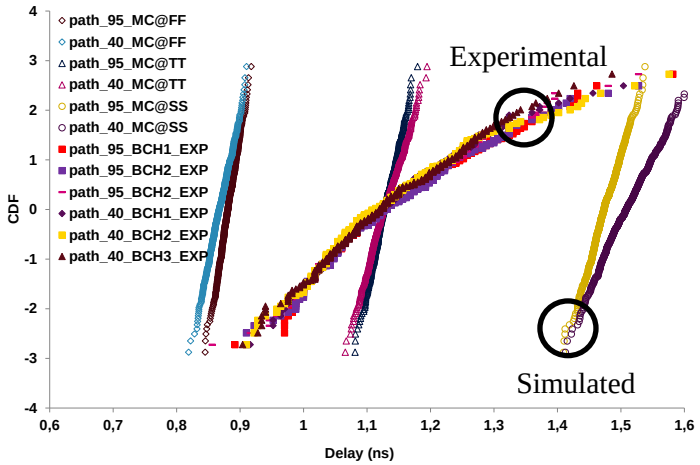


Figure 4.21 : Mesures expérimentales des V_{DDmin} des chemins 95 et 40 dans 160 puces et comparaison avec les délais issus de la simulations MC dans les différents procédés

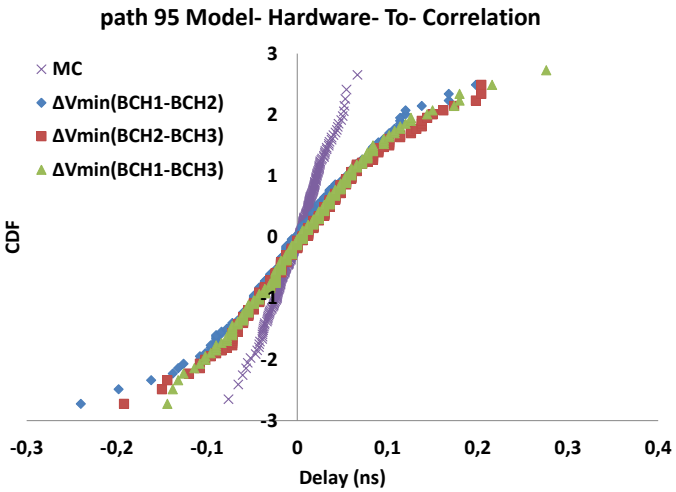


Figure 4.22 : Comparaison des dispersions locales du chemin 95 obtenues par la mesure et par la simulation MC.

Nous nous intéressons dans cette étude à extraire la contribution de la variation locale des résultats mesurés afin de la comparer à la simulation MC. Pour ce faire, nous avons supprimé la composante de la variation globale des mesures en calculant les ΔV_{min} du chemin 95 entre

les différents BCHs. Les opérations effectuées sont de type : BCH1-BCH2, BCH2-BCH3, et BCH1-BCH3.

Les composantes locales expérimentales et simulés sont présentées dans la figure 4.22. Nous observons que la dispersion expérimentale est deux fois supérieure à la dispersion simulée avec des valeurs respective de 0.4ns et 0.2ns environ. Cette différence est peut être due à une autre contribution qui n'est pas prise en compte dans le modèle du transistor utilisé dans les simulations MC.

Afin de comprendre ce désaccord, nous avons mesuré la variation relative du courant I_d d'une centaine de transistors fabriqués dans la technologie 45nm et dans un procédé typique. Nous avons simulé cette variation à l'aide de la méthode Monte-Carlo et du modèle que nous avons utilisé pour les simulations de la figure 4.22. Le résultat de cette expérience est présenté dans la figure 4.23, il montre une bonne corrélation entre la mesure et la simulation, ceci signifie que le modèle n'est pas à l'origine de la différence entre la mesure et la simulation de la variabilité locale observée dans la figure 4.22. L'origine de cette différence est peut être liée aux paramètres RC des fils. En effet, la dispersion de type RC n'est pas prise en compte dans les simulations MC.

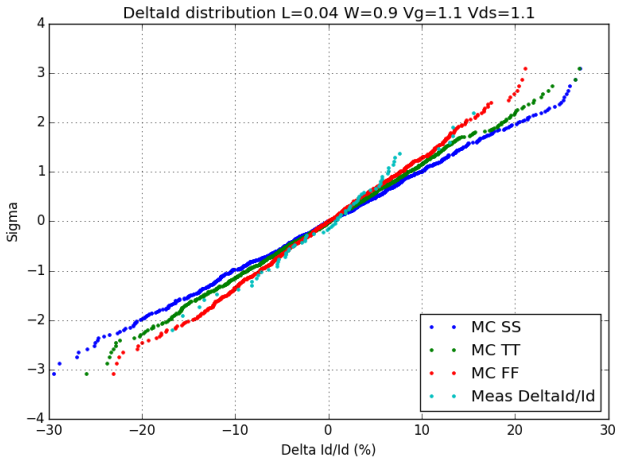


Figure 4.23 : Mesure de la dégradation du courant I_d du transistor et comparaison avec le modèle SPICE.

4.6 Moniteurs in-situ et adaptation dynamique de la tension

Dans les paragraphes précédents, nous avons testé les comportements des moniteurs dans différents conditions de PVTs où la tension est réduite jusqu'au dysfonctionnement du circuit.

Dans ce paragraphe, nous allons nous intéresser à deux aspects :

a. L'adaptation dynamique de la tension V_{DD} (DVS) à l'aide des moniteurs in-situ.
b. La recherche du point de performance optimale OPP (Operation Performance Point).

Concernant la DVS, Il s'agit d'une technique qui permet d'alimenter le circuit avec la tension V_{DD} optimale de telle manière de s'éloigner du mode d'opération dans lequel les moniteurs signalent des alarmes. Quant à la recherche des OPPs, elle consiste à trouver le couple-tension F(V) que le microprocesseur doit utiliser dans l'application de la DVFS. Ce couple est attribué au circuit par le microprocesseur en fonction des conditions d'usages.

Le principe de la DVS utilisé est illustré dans la figure 4.24. Les sorties des moniteurs sont connectés à un compteur qui compte toute mise à '1' du signal Flag. La sortie du compteur est branchée à un régulateur de tension qui régule cette dernière selon l'apparition des signaux Flags.

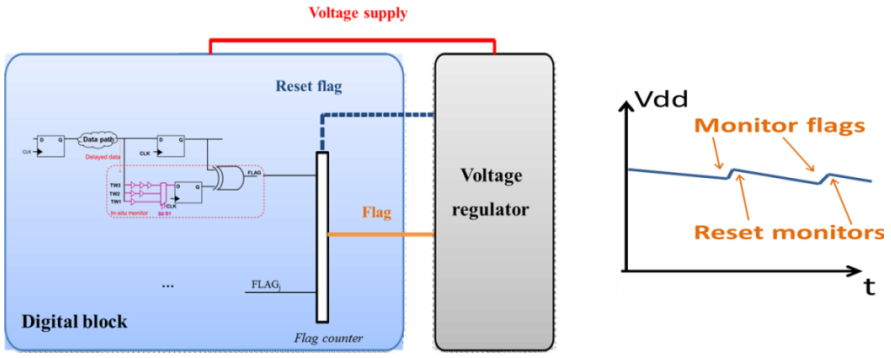


Figure 4.24 : illustration du principe de l'adaptation de la tension d'alimentation : La tension est régulée en fonction de l'apparition des Flags.

Nous avons appliqué la DVS et la recherche de OPPs au circuit BCH dans la technologie 28FDSOI. En effet, dans un premier temps, nous avons considéré 8 fréquences entre 0.27 et 2.3 GHz que nous avons fait varier. Pour chaque fréquence, nous avons réduit la tension V_{DD} jusqu'à l'apparition du premier signal Flag, il s'agit de OPP dans notre cas. Nous continuons ensuite de réduire la tension jusqu'au dysfonctionnement. Les tensions V_{DDmin} du premier Flag et de dysfonctionnement ainsi que la fréquence constituent des informations qui pourraient être utilisées dans la DVFS.

Dans un second temps, pour chaque fréquence, nous avons réduit la tension avec un pas de 1mV jusqu'à l'apparition du premier Flag. Ensuite, la tension est augmentée jusqu'à ce qu'on n'observe plus le signal Flag et nous procédons à la régulation dynamique afin d'opérer autour de la tension V_{DDmin} du premier Flag pendant un temps *t*.

Le résultat de cette manipulation est illustré dans la figure 4.25 [141]. Il montre la fiabilité des moniteurs à adapter dynamiquement la tension V_{DD} ainsi que dans la recherche de OPP. De plus, nous avons répété ces opérations en faisant varier la fréquence dans un ordre croissant et décroissant pendant plusieurs cycles d'horloge qui constituent un temps de test de 75ms. En guise de résultat, nous avons toujours obtenu les même OPPs.

Ainsi, nous démontrons la capacité des moniteurs de trouver la tension/fréquence optimale dans n'importe quelle condition d'opération, chose difficile à réaliser durant les étapes de conception car le sign-off devrait être réalisé dans chacune de ces opérations. Ce qui rend le nombre de cas possibles très important.

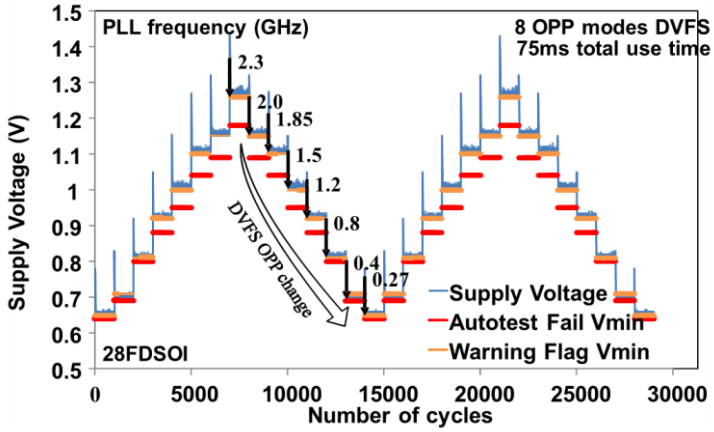


Figure 4.25 : Recherche de OPP pour différentes conditions d'opérations : la tension est ajustée afin de maintenir le V_{DDmin} proche du V_{DDmin} du premier Flag.

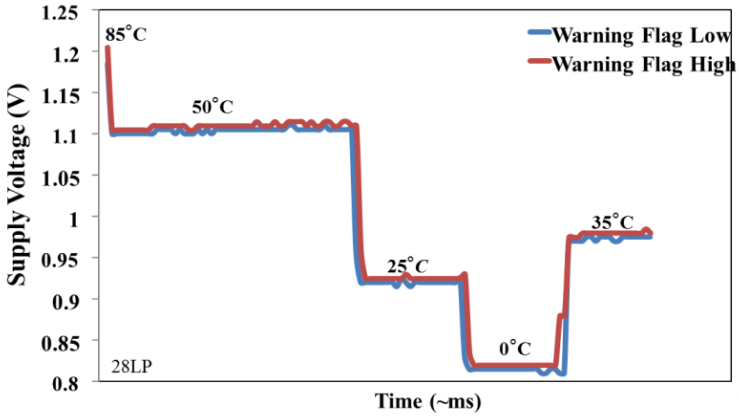


Figure 4.27 : Comportement des moniteurs in-situ face aux variations brutales de tensions et de température

Dans les paragraphes 4.3 et 4.4, nous avons démontré la capacité des ISMs à opérer dans différentes conditions environnementales. En termes de tension, ces changements environnementaux représentent des marges ne dépassant pas 50mV dans le pire des cas. Du point de vue fiabilité la question reste à savoir si les moniteurs in-situ suivent le

comportement du circuit dans le cas des changements brutaux induisant des fortes variations de tensions et de températures. Pour y répondre, nous avons fait varier la tension de 1.2V à 0.8V et la température de 85°C à 0°C avec un pas de 35°C pour le circuit BCH fabriqué dans la technologie 28LP. Ces changement sont fait par palier et d'une manière brutale. En effet, dans un premier temps, nous avons imposé une réduction de tension de 100mV et la température de 35°C dans un temps très court (quelques millisecondes). Nous avons maintenu cette nouvelle tension (1.1V) et température (50°C) et ensuite nous avons imposé un autre changement de température et de tension encore plus brutal qui représente une variation de tension de 200mV et 25°C de température. Le résultat est présenté dans la figure 4.26 [141]. Nous observons que pour toutes ces conditions, les moniteurs suivent correctement le comportement du circuit lors des changements imposés. Ce résultat prouve l'efficacité des moniteurs in-situ pour faire face également au changement de la tension et de la température dans un régime brutale.

4.7 Moniteur in-situ et impact du vieillissement

Dans ce paragraphe, nous nous intéressons à une autre dimension qui est le vieillissement. La question qu'on doit se poser concerne la manière dont la fréquence maximale des moniteurs se dégrade et si ces derniers suivent correctement le comportement du circuit après le vieillissement. Pour répondre à ces questions, nous avons mesuré les F_{max} du circuit BCH et du moniteur dans deux conditions de vieillissements. En effet, nous avons appliqué un premier stress de type AC à une tension V_{DD} = 1.1V et une température de 25°C et un second stress AC à V_{DD} = 1.2V et une température de 125°C afin d'accélérer le vieillissement. Les mesures des F_{max} du circuit et du moniteur sont effectuées dans les conditions nominales de fonctionnement à savoir le corner TT à une fréquence de 1Ghz et une tension V_{DD}=1V.

La figure 4.28 montre une comparaison de la dégradation du F_{MAX} du moniteur et du BCH dans deux conditions vieillissement pour 25 puces. Cette dégradation est calculée par rapport au F_{max} à t_0. Nous remarquons que la dégradation du F_{max} du moniteur et du circuit est corrélée pour les deux conditions. Ceci prouve que les moniteurs se dégradent de la même manière que le circuit et que nous ne risquons pas d'observer un défaut de détection à l'issue du vieillissement.

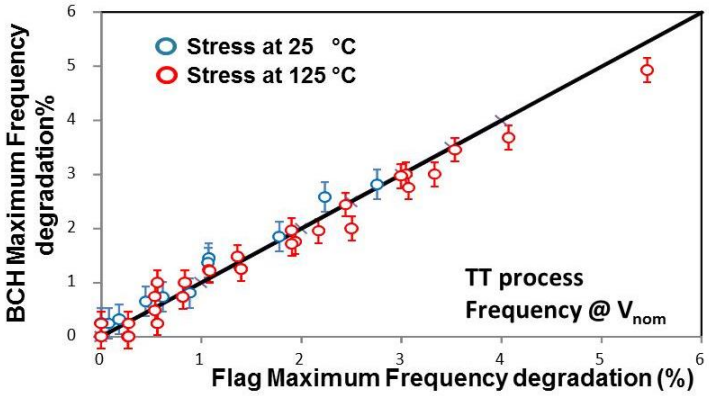


Figure 4.28 : Comparaison de la dégradation des F_{max} du circuit BCH et des moniteurs [141] après vieillissement

Nous nous proposons maintenant de compter le nombre des moniteurs lors de l'application d'une contrainte de vieillissement accélérée. En effet, nous avons appliqué une tension de stress V_{DD} = 1.5V et une température de 125°C au circuit PSR en 28FDSOI. Ensuite nous avons fait varier la fréquence et avons compté le nombre de moniteurs pour deux puces. Le résultat est illustré dans la figure 4.29 et est donné sous forme de probabilité d'observer des Flags. Nous remarquons que la probabilité d'observer des Flags à l'issue du vieillissement est importante. Ceci est attendu car après vieillissement, un grand nombre de signaux se trouvent dans la fenêtre de détection. Nous observons aussi que le profil de l'évolution de la courbe du nombre de Flags à t_0 et après vieillissement est identique. L'explication de cette évolution a été discutée dans le paragraphe 4.4.2.

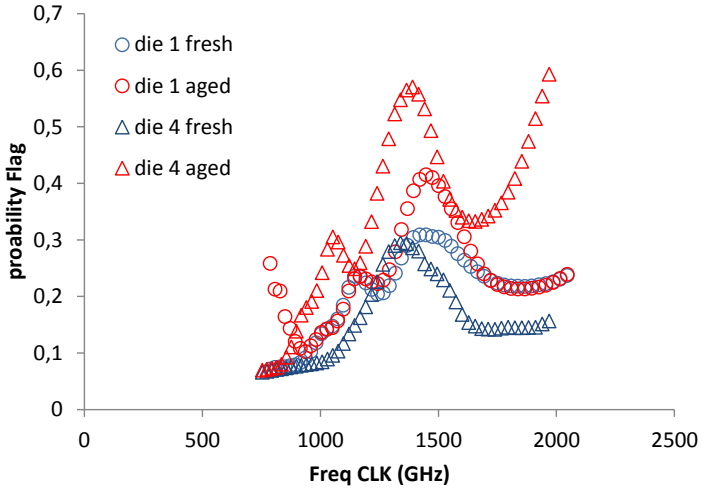


Figure 4.29 : Mise en œuvre des comportements des moniteurs après vieillissement et à t_0 en fonction de la fréquence.

4.8 Conclusion

Dans ce chapitre, nous avons mis à l'épreuve la robustesse et la capacité des moniteurs in-situ à faire face à différents changements environnementaux tels que les variations des PVTs et les conditions d'usage. Nous avons commencé par étudier la réponse des moniteurs en faisant varier la tension et la fréquence. Il en résulte une carte de tension-fréquence qui montre les zones dans lesquelles le circuit peut fonctionner correctement sans alertes provenant des moniteurs, correctement mais avec alertes et une zone de non fonctionnalité.

Par ailleurs, nous avons compté le nombre de Flags lorsqu'on réduit la tension d'alimentation, et ce sous l'impact des PVTs et des conditions d'usage. Dans toutes ces conditions, les signaux Flag se lèvent toujours avant le signal *Autotest_OK*. Aussi, nous avons constaté une dépendance de la tension minimale V_{DDmin} aux conditions environnementales. En effet, plus le vecteur est complexe plus le V_{DDmin} du circuit est élevé. Ceci est lié au grand nombre de chemins activés lors de l'exécution des vecteurs complexes. Ce constat est valable aussi pour la température : le V_{DDmin} est d'autant plus élevé que la température augmente. Ceci est dû à l'impact de la température sur le délai des portes logiques que nous avons détaillé dans le chapitre 1.

Dans la deuxième partie du chapitre, nous nous sommes intéressés au comportement des moniteurs dans les mêmes conditions de test citées précédemment. Nous avons constaté que les V_{DDmin} des moniteurs suivent la même tendance des V_{DDmin} du circuit, c'est-à-dire la dépendance aux conditions environnementales et à la largeur de la fenêtre de détection.

Par ailleurs, nous avons pu extraire les indexes des moniteurs activés lors de l'exécution des différents vecteurs, grâce à une technique de test. Ainsi, nous avons pu faire une comparaison entre les bascules sélectionnées par l'algorithme proposé dans le chapitre 3 et les moniteurs activés en fonction des vecteurs exécutés. Certes, le nombre de moniteurs sélectionnés par l'algorithme est supérieur à ceux activés expérimentalement, mais nous obtenons une bonne couverture des chemins avec les moniteurs, même dans le pire cas, i.e. le vecteur contenant trois erreurs. D'autre part, la technique de test nous a permis de calculer les délais des chemins activés. En comparant ces délais avec ceux obtenue avec la STA, nous constatons qu'il ne s'agit pas toujours des pires chemins par bascule qui sont critiques.

La troisième partie du chapitre a été consacré à l'étude de la fiabilité des moniteurs in-situ lorsque le circuit est soumis à des variations brutales. En effet, nous avons imposé des changements brutaux de tensions et de températures, et dans tous les cas, les V_{DDmin} des moniteurs suivent la même tendance du V_{DDmin} du circuit. Ainsi, nous prouvons la capacité des moniteurs à faire face à des variations environnementaux extrêmes.

Par ailleurs, nous avons utilisé les moniteurs dans une application d'adaptation dynamique de la tension. Nous avons réduit la tension jusqu'à l'apparition du premier Flag et nous avons opéré autour de cette tension pendant une certaine durée. Ainsi, nous démontrons leurs capacités de fonctionner autour de la tension minimale du premier Flag qu'on considère comme un OPP.

Finalement, nous nous sommes intéressés au comportement des moniteurs après vieillissement. Nous avons montré une bonne corrélation de la dégradation du F_{max} du circuit et du moniteur, ce qui prouve que nous ne risquons pas d'obtenir de fausses alertes à l'issue du vieillissement. Avec ce résultat, nous pouvons donc utiliser les moniteurs in-situ pour monitorer les mécanismes de dégradation BTI et par porteurs chauds (HCI). Etant donné qu'au cours de son fonctionnement normal, un circuit vieillit, cela se traduit principalement par une dégradation des délais intrinsèques des portes logiques qui constituent ce circuit. Les moniteurs in-situ peuvent être un bon indicateur du vieillissement de certains chemins critiques car lorsque le signal Flag se lève, nous obtenons une marge temporelle correspondant à la fenêtre de détection.

CONCLUSION ET PERSPECTIVES

Durant cette thèse, nous nous sommes intéressés à la fiabilité de circuits intégrés numériques à l'aide de l'utilisation des moniteurs in-situ. Ces moniteurs sont capables de remonter des alertes relatives à l'état du vieillissement du circuit dans l'environnement réel d'utilisation.

Les impacts des mécanismes de dégradation BTI et HCI ont été présentés dans le chapitre 1. Ces dégradations se traduisent par un décalage de la tension de seuil du transistor et par conséquent de la dégradation des délais des portes logiques qui à leur tour peuvent provoquer des violations de timing sur certains chemins du circuit. Par ailleurs, la variation du procédé joue un rôle important sur les délais des chemins: nous avons démontré que le délai d'un chemin est plus court dans un procédé de type FF que dans les procédés TT et SS. De la même manière, les délais sont plus courts dans les fortes tensions d'alimentation et les basses températures.

Traditionnellement, les variations PVTs et le vieillissement sont prises en compte en rajoutant des marges de tension supplémentaires durant la phase design. Dans le chapitre 2, nous avons quantifié ces marges pour un circuit numérique en technologie 28FDSOI. Nous estimons une marge de tension supplémentaire d'environ 200mV pour couvrir l'occurrence de toutes les variations. Ces marges excessives (plus de 20% de V_{DD}) peuvent être réduites en utilisant des moniteurs in-situ et une technique adaptative afin de réguler la tension d'alimentation, la fréquence ou la tension des caissons dans les technologies FDSOI en fonction de besoins applicatifs.

Le moniteur in-situ que nous avons choisi dans cette thèse est basé sur une détection de pré-erreur. Ce schéma est avantageux car il détecte des pré-erreurs et non pas d'erreurs. Donc nous n'avons pas à utiliser une circuiterie additionnelle ou du temps supplémentaire pour la correction d'erreurs. Toutefois, même l'insertion des moniteurs de pré-erreur dans les circuits peut présenter des pénalités de surface et de consommation. Ainsi, il faut trouver la méthode optimale pour limiter ces pénalités. Ceci a fait l'objet du chapitre 3. En effet, nous avons proposé un layout du moniteur et un flot d'insertion qui minimise l'impact sur les délais des chemins du circuit. Concernant les pénalités de surface, nous avons établi un algorithme qui sélectionne les chemins potentiellement critiques. Nous avons pu obtenir une réduction de nombre de chemins d'un facteur 2 dans le meilleur cas, par rapport à la méthode basée sur la sélection des 10% des chemins critiques.

Dans le chapitre 4 nous avons présenté les résultats de test de deux types de circuits numériques. Ces tests ont été effectués dans le but de réduire la tension d'alimentation au minimum, pour un circuit encore fonctionnel. Ces tests ont mis en évidence la dépendance de la tension minime, de températures et des conditions d'usage. Une étude de la variation de la fenêtre de détection du moniteur a aussi été menée. Pour chacune de ces variations, les moniteurs se comportent différemment. En effet, les mesures des V_{DDmin} des moniteurs dépendent de la tension d'alimentation même, de la température, de l'usage et de la position de la puce au sein de la plaque. Par ces mesures, nous démontrons non seulement l'intérêt de moniteurs in-situ mais également leur robustesse face à variations environnementales. Par ailleurs, nous avons utilisé les moniteurs in-situ dans une application d'adaptation dynamique de la tension, et nous avons montré leurs capacités de fonctionner autour du point optimal de fonctionnement (OPP). En prenant en compte aussi le vieillissement d'un circuit complexe, nous avons obtenu une bonne corrélation du F_{max} du circuit et des moniteurs, ce qui prouve qu'ils vieillissent de la même manière que le circuit même et donc que nous ne risquons pas d'obtenir des fausses alertes dans le temps.

L'algorithme de sélection des chemins potentiellement critiques développé dans le chapitre 3 a permis une sélection de 76 chemins sur lesquelles il faudrait insérer des moniteurs. Or, les résultats expérimentaux de test montrent que le nombre de chemins activés varient entre 2 et 50, selon les conditions d'usage. Cette différence vient du fait que dans nos analyses, nous avons fait l'hypothèse de l'équiprobabilité des arcs alors que dans la réalité, les arcs ne sont pas équiprobables. Donc, une étude sur l'activité des chemins devrait être menée durant les phases Front-End du design afin de sélectionner d'une manière plus exacte les chemins qui seront activé pendant l'utilisation du circuit. Les outils de simulation peuvent estimer l'activité de chaque fil du circuit mais ne permettent pas d'extraire avec précision l'arc activé. Dans cette situation, il faut peut-être se tourner vers des méthodes de type Machine Learning pour déceler les chemins activés du circuit, face à une application donnée. Ainsi, seuls ces chemins seront équipés des moniteurs et donc les pénalités de surface et de consommation liées aux moniteurs seront réduites.

BIBLIOGRAPHIE DE L'AUTEUR

A. Benhassain, F. Cacho, V. Huard, S. Mhira, L.Anghel, C. Parthasarathy, A.Jain, A.Sivadasan, "Robustness of timing in-situ monitors for AVS management", Internationnal Reliability Physics Symposium (2016)

A. Benhassain, S. Mhira, F. Cacho, V. huard, L.Anghel, "In-situ slack monitors: Taking up the challenge of on-die monitoring of variability and reliability", International Verification and Security Workshop (2016)

A. Benhassain, F. Cacho, V. Huard, M.Saliva, L. Anghel, C. Parthasarathy1, A. Jain, F. Giner, "Timing in-situ monitors: implementation strategy and applications results", Custom Integrated Circuits Conference (2015)

A. Benhassain, F. Cacho, V, Huard, L.Anghel, "Early failure prediction by using in-situ Slack Monitors: Implementation and application results", ERMAVSS (DATE WORKSHOP), 2016

M. Saliva, F. Cacho, V. Huard, X. Federspiel, D. Angot, **A. Benhassain**, A. Bravaix, L. Anghel, ''Digital Circuit Reliability with In-Situ Monitors in 28nm Fully Depleted SOI'', IEEE Design Automation & Test in Europe 2015

Sivadasan, S. Mhira, A. Notin, **A. Benhassain**, V. Huard, E. Maurin, F. Cacho, L. Anghel, A. Bravaix, "ARCHITECTURE- AND WORKLOAD-DEPENDENT DIGITAL FAILURE RATE", IRPS 2017

A. Sivadasan, , F. Cacho, **A. Benhassain**, V. Huard, L. Anghel, "Study of workload impact on BTI HCI induced aging of digital circuits", DATE 2016

V. Huard, F. Cacho, X. Federspiel, M. Saliva, **A. Benhassain**, C. Parthasarathy1, A. Jain, F. Giner, "Adaptative wear out management with in-situ management", Internationnal Reliability Physics Symposium (IRPS), pp. 6B.4.1 - 6B.4.11, 2014

F. Cacho, **A. Benhassain**, R. Shah, S. Mhira, V. Huard, L. Anghel, "Investigation of Critical Path Selection for In-Situ Monitors Insertion", IOLTS 2017

F.Cacho, **A. Benhassain**, S. Mhira, A. Sivadasan, V. Huard, P. Cathelin, V. Knopik, A. Jain, C. Parthasarathy, L. Anghel, "Activity Profiling: review of different solutions to develop reliable and performant design", IOLTS 2016

V. Huard, F. Cacho, **A. Benhassain**, C. Parthasarathy, "Aging-aware Adaptive Voltage Scaling of Product Blocks in 28nm Nodes", IRPS 2016

BIBLIOGRAPHIE

[1] http://www.eeherald.com/section/news/nws201106221.html

[2] A.P. Chandrakasan et al. “Low-power CMOS digital design”; Solid-State Circuits, IEEE Journal of, Volume: 27 , Issue: 4 1992 , pp. 473 - 484

[3] R. Kapre et al. “SRAM Variability and Supply Voltage Scaling Challenges”, IEEE International Reliability Physics Symposium, 2007, pp. 23-28

[4] R. Doering et al. “Limits of integrated-circuit manufacturing”, Proceedings of the IEEE, Mar 2001 Volume : 89 , Issue:3, pp. 375 - 393

[5] D. Burnett et al. “Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits”, VLSI Technology, 1994. Digest of Technical Papers. 1994, pp. 15 - 16

[6] A.J. Bhavnagarwala et al. “The impact of intrinsic device fluctuations on CMOS SRAM cell stability”, Solid-State Circuits, IEEE Journal of, Volume: 36 , Issue: 4, pp. 658 - 665

[7] V. Huard et al. “A bottom-up approach for System-On-Chip reliability”, Microelectronics Reliability (2011), doi:10.1016/j.microrel.2011.07.086

[8] V. Huard et al. “A bottom-up approach for System-On-Chip reliability”, Microelectronics Reliability (2011), doi:10.1016/j.microrel.2011.07.086

[9] V. Huard et al. “A Predictive Bottom-up Hierarchical Approach to Digital System Reliability”, In IEEE International Reliability Physics Symposium, 2012, pp. 1–10

[10] D. Lorenz et al. “Aging analysis of circuit timing considering NBTI and HCI” In IEEE International On-Line Testing Symposium. Ieee, 2009, pp. 3–8.

[11] B.C. Paul et al. “Impact of NBTI on the Temporal Performance Degradation of Digital Circuits”, IEEE Electron Device Letters, 26(8), pp.560–562.

[12] D. Dubois_et al. “Ageing sensor for analog application”, Circuits and Systems and TAISA Conference, 2009

[13] J. Keane, “An On-Chip NBTI Sensor for Measuring pMOS Threshold Voltage degradation”, TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, VOL. 18, NO. 6, JUNE 2010.

[14] Q. Zhenyu, “SRAM-Based NBTI/PBTI Sensor System Design”, Design Automation Conference (DAC), 2010.

[15] S. Prashant, “Dynamic NBTI Management Using a 45 nm Multi-Degradation Sensor”, TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS 2011.

[16] K. Kyung, “On-Chip Aging Sensor Circuits for Reliable Nanometer MOSFET Digital Circuits”, TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, VOL. 57, NO. 10, OCTOBER 2010.

[17] S. Prashant, “Compact Degradation Sensors for Monitoring NBTI and Oxide Degradation” TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, 2011

[18] T. Fischer, “A 90-nm variable frequency clock system for a power-managed itanium architecture processor,” IEEE J. Solid-State Circuits, pp. 218–228, Jan. 2006.

[19] R. McGowen et al., “Power and temperature control on a 90-nm itanium family processor,” IEEE J. Solid-State Circuits, pp. 229–237, Jan. 2006.

[20] J. Tschanz et al., “Adaptive frequency and biasing techniques for tolerance to dynamic temperature-voltage variations and aging,” in IEEE ISSCC Dig. Tech. Papers, Feb. 2007, pp. 292–293.

[21] A. Muhtaroglu, “On-die droop detector for analog sensing of power supply noise”, IEEE J. Solid-State Circuits, pp. 651–660, Apr. 2004.

[22] P. Franco et al. “Delay testing of digital circuits by output waveform analysis”, in Proc. IEEE Int. Test Conf., Oct. 1991, pp. 798–807.

[23] P. Franco et al. “On-line testing of digital circuits,” in Proc. IEEE VLSI Test Symp., Apr. 1994, pp. 167–173.

[24] M. Nicolaidis, “Time redundancy based soft-error tolerance to rescue nanometer technologies,” in Proc. IEEE VLSI Test Symp., Apr. 1999, pp. 86–94.

[25] D. Ernst et al., “Razor: A low-power pipeline based on circuit-level timing speculation,” in Proc. IEEE/ACM Int. Symp. Microarchitecture (MICRO-36), Dec. 2003, pp. 7–18.

[26] S. Das et al., “A self-tuning DVS processor using delay-error detection and correction”, IEEE J. Solid-State Circuits, pp. 792–804, Apr. 2006.

[27] K. Bowman et al. “Energy-Efficient and Metastability-Immune Resilient Circuits for Dynamic Variation Tolerance”, JOURNAL OF SOLID-STATE CIRCUITS, VOL. 44, NO. 1, JANUARY 2009

[28] N.Pour, et al. “Comparison of in-situ delay monitors for use in adaptative voltage scaling” Adv. Radio Sci., 10, 215-220, 2012

[29] D. Fick, “In Situ Delay-Slack Monitor for High-Performance Processors Using An All-Digital Self-Calibrating 5ps Resolution Time-to-Digital Converter”, IEEE International Solid-State Circuits Conference, 2010

[30] M. Wirnshofer, et al. “adapatative voltage scaling by in-situ delay monitoring for an Image Processing circuit”, Trans. VLSI, vol. 20, no. 2, 2012

[31] H. Fuketa, “Adaptative performance compensation with in-situ timing error predictive sensors for subthreshold circuits”, VLSI, vol. 20, no.2, 2012

[32] S. Abdollahi, “Power Reduction by Automatic Monitoring and Control System in Active Mode”, World Academy of Science, Engineering and Technology 61 2012

[33] L. Lai et al. “SlackProbe: A Low Overhead In Situ On-line Timing Slack Monitoring Methodology”, DATE ,2013

[34] A. Rahimi, “Aging-aware compiler-directed VLIW assignment for GPGPU architectures”, DAC 2013

[35] P. Mercati, “Workload and User Experience-Aware Dynamic Reliability Management in Multicore Processors”, DAC 2013

[36] D.M. Ancajas, “HCI-Tolerant NoC Router Microarchitecture”, DAC 2013

[37] H. Tajik, “VAWOM: Temperature and Process Variation Aware WearOut Management in 3D Multicore Architecture”, DAC 2013

[38] X. Chen, “Dynamic Voltage and Frequency Scaling for Shared Resources in Multicore Processor Designs”, DAC 2013

[39] A. Calimera, “Energy-Optimal SRAM Supply Voltage Scheduling under Lifetime and Error Constraints”, DAC 2013

[40] T. Skotnicki, T. & F. Boeuf, F, Introduction a la Physique du transistor MOS, 2002

[41] T. Skotnicki, Transistor MOS et sa technologie de fabrication. In Encyclopedie Techniques de l’ingenieur. p. cahier E 2430, 2002

[42] C. Colinge. Physics of semiconductor devices. Springer, 2002.28

[43] D. Fleury “Contribution à l’étude experimentale du transport dans les transistors de dimensions deca nanometrique des technologies CMOS sub 45nm’’. PhD thesis, Institut national polytechnique de grenoble, 2008. 28

[44] T. Skotnicki et al. “The voltage-doping transformation : a new approach to the modeling of mosfet short-channel eects”, Electron Device Letters, IEEE, 9(3) :109-112, March 1988. 33

[45] M. Bidaud et al. “Complete re-estimation of the gate leakage current limit for sub-0.12um technologies (eot= 1.8-2.8nm)”, In Solid-State Device Research Conference, 2000. Proceeding of the 30th European,pages 304-307, September 2000. 36.

[46] M. Rafik, “Caracterisation et modelisation de la _abilite des transistors avances à dielectriques de haute permittivite et a grille metallique”, PhD thesis, Institut national polytechnique de grenoble, 2008. 14, 23, 27, 38

[47] S. Cristoloveanu et al. “Electrical Characterization of Silicon on Insulator Materials and Devices”, 1995

[48] G.K Celler et al. “Frontiers of silicon-on-insulator”, Journal of Applied Physics, 93(9).

[49] A. Vandooren, et al. “ Fully-Depleted SOI Devices With TaSiN Gate , HfO 2 Gate Dielectric,and Elevated Source / Drain Extensions”, IEEE Electron Device Letters, 24(5), pp.342–344, 2003

[50] H.K. Lim et al. “Threshold Voltage of Thin-Film Silicon-on-Insulator (SOI) MOSFET’s”, IEEE Transactions on Electron Devices, 30(10), pp.1244–1251, 1983

[51] S. Cristoloveanu. “Monitoring Transistor Degradation” In Electrical Characterization of Silicon-On-Insulator Materials and Devices. p. 337. 1999

[52] J. Noel. Optimisation de dispositifs FDSOI pour la gestion de la consommation et de la vitesse application aux mémoires et fonctions logiques 2011.

[53] D. Angot Fiabilité et Variabilité Temporelle des Technologies CMOS FDSOI 28-20nm, du Transistor au Circuit Intégré. 2014

[54] M. Saliva. Circuits dédiés à l’étude des mécanismes de vieillissement dans les technologies CMOS avancées : Conception et mesures, 2015.

[55] M. Denais, Etude des phénomènes de dégradation de type negative bias temerature instability dans les transistors MOS submicroniques des filières CMOS avancées, 2005.

[56] D.K Schroder et al. “Negative bias temperature instability: What do we understand?”, Microelectronics Reliability, 47(6), pp.841–852, 2007

[57] T. Grasser et al. “NBTI in Nanoscale MOSFETs — The Ultimate Modeling Benchmark”, IEEE Transactions on Electron Devices, 61(11), pp.3586–3593. 2014

[58] B.E. Deal et al. “Characteristics of the Surface-State charge (QSS) of thermally oxidized silicon”, Solid State Science 114(3), pp.266–274. 1967

[59] K.O Jeppson et al. “Negative bias stress of MOS devices a high electric fields and degradation of MNOS devices”, Journal of Applied Physics, 48(5). 1977a

[60] M. Denais et al. “On-The-Fly characterization of NBTI in ultra-thin gate-oxide PMOSFET’s”, In IEEE International Electron Devices Meeting. pp. 109–112. 2004.

[61] V. Huard et al. “Design-in-Reliability Approach for NBTI and Hot-Carrier Degradations in Advanced Nodes”, IEEE Transactions on Device and Materials Reliability, 7(4), pp.558–570. 2007.

[62] X. Garros et al. “ IMPACT OF CRYSTALLINITY OF HIGH -K OXIDES ON VT INSTABILITIES OF NMOS DEVICES ASSESSED BY PHYSICAL AND ELECTRICAL MEASUREMENTS”, In IEEE International Reliability Physics Symposium. pp. 330–334. 2008

[63] M. Wang et al “Superior PBTI Reliability for SOI FinFET Technologies and Its Physical Understanding”, IEEE Electron Device Letters, 34(7), pp.837–839.

[64] K.T. Lee et al. “Technology Scaling on High-K & Metal-Gate FinFET BTI Reliability”, In IEEE International Reliability Physics Symposium.

[65] A. Alam et al. “A comprehensive model for PMOS NBTI degradation: Recent Progress”. Microelectronics Reliability, 47(6), pp.853–862.

[66] S. Ogawa et al. “Generalized diffusion-reaction model for the low-field chargebuildup instability at the Si-SiO2 interface”, Physics Review, 51, pp.4218–4230.

[67] B. Biswas, et al. “Enhanced stability of deuterium in silicon”. Applied Physics Letter, 72, pp.61–62.

[68] T. Aichinger et al. “Evidence for P b center-hydrogen complexes after subjecting PMOS devices to NBTI stress – a combined DCIV / SDR study”, In IEEE International Reliability Physics Symposium. pp. 1–6.

[69] T. Grasser et al. “A Two-stage Model for Negative Bias Temperature Instability”, In IEEE International Reliability Physics Symposium. pp. 33–44.

[70] T. Grasser et al. “The “ Permanent ” Component of NBTI : Composition and Annealing”, In IEEE International Reliability Physics Symposium. pp. 605–613, 2011

[71] T. Grasser et al. “The Paradigm Shift in Understanding the Bias Temperature Instability : From Reaction – Diffusion to Switching Oxide Traps”, IEEE Transactions on Electron Devices, 58(11), pp.3652–3666, 2011

[72] T. Grasser et al. “Hydrogen-Related Volatile Defects as the Possible Cause for the Recoverable Component of NBTI”, In IEEE International Electron Devices Meeting. pp. 409–412, 2011

[73] A. Bravaix, “Hot-carrier to cold-carrier issues in nanoscale cmos nodes: From energy driven to multiple particle regime”, In IEEE International Reliability Physics Symposium Tutorial.

[74] W. Arfaoui et al. “Energy-driven hot-carrier model in advanced nodes. In Reliability Physics Symposium”, 2014 IEEE International, pages XT.12.1-XT.12.5, June 2014. 85, 86, 89.

[75] A. Bravaix et al. “Hot-carrier damage in AC-stressed deep submicrometer CMOS techtnologies”, 1999, In IEEE International Integrated Reliability Workshop. pp. 1–5.

[76] S. Tam et al. “Lucky-Electron Model of Channel Hot-Electron Injection in MOSFET’s”, In IEEE Transactions on Electron Devices, ED-31(9), pp.1116–1125, 1984

[77] H. Kufluoglu, “MOSFET Degradation due to Negative Bias Temperature Instability (NBTI) and Hot Carrier Injection (HCI) and its implications for Reliability Aware VLSI Design”, Purdue University, 2007

[78] S.E. Rauch et al. “The Energy-Driven Paradigm of NMOSFET Hot-Carrier Effects”, In IEEE Transactions on Device and Materials Reliability, 5(4), pp.701–705.

[79] A. Bravaix et al. “Hot-Carrier Acceleration Factors for Low Power Management in DC-AC stressed 40nm NMOS node at High Temperature”, In IEEE International Reliability Physics Symposium. pp. 531–548, 2009

[80] C. Guérin et al. “The Energy-Driven Hot-Carrier Degradation Modes of nMOSFETs”, in IEEE Transactions on Device and Materials Reliability, 7(2), pp.225–235, 2007

[81] M.Saliva, A.Benhassain et al “Digital Circuits Reliability with In-Situ Monitors in 28nm Fully Depleted SOI” Date 2015.

[82] R. Fernández et al. “AC NBTI studied in the 1 Hz – 2 GHz range on dedicated on-chip CMOS circuits”, In IEEE International Electron Devices Meeting.

[83] C.H. Kim “In-Situ Monitoring of Reliability in Circuits”, Tutorial, 2013

[84] T. Grasser et al. “A Two-stage Model for Negative Bias Temperature Instability”, In IEEE International Reliability Physics Symposium. pp. 33–44, 2009

[85] C. Mezzomo, “Modeling and characterization of the random fluctuations in electrical parameters of advanced CMOS”, PhD thesis, Universite de Grenoble, 2010. 105.

[86] J. Croon. “Matching properties of deep sub-micron MOS transistor”, PhD thesis, Katholieke Universiteit Leuven, 2004. 105

[87] V.S. Kaushik, et al. “Estimation of fixed charge densities in hafnium-silicate gate dielectrics”, Electron Devices, IEEE Transactions on 53(10) :2627-2633, Oct 2006. 107.

[88] M. Koh, “Limit of gate oxide thickness scaling in mosfets due to apparent threshold voltage fluctuation induced by tunnel leakage current”, Electron Devices, IEEE Transactions on, 48(2) :259-264, Feb 2001. 107.

[89] H.C Wen, “On oxygen deficiency and fast transient charge-trapping eacts in high-dielectrics”, Electron Device Letters, IEEE, 27(12) :984-987, Dec 2006. 107.

[90] A. Asenov et al. “Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness”, Electron Devices, IEEE Transactions on, 50(5) :1254-1260, May 2003. 107.

[91] J.M. Steigerwald, “Chemical mechanical polish : The enabling technology”, In Electron Devices Meeting, 2008. IEDM 2008. IEEE International, pages 1-4,Dec 2008. 107.

[92] S. Nag et al. “Comparative evaluation of gap-_ll dielectrics in shallow trench isolation for sub-0.25 /spl mu/m technologies”, In Electron Devices Meeting, 1996. IEDM '96., International, pages 841-845, Dec 1996. 107

[93] AR. Brown et al. “Poly-si-gate-related variability in decananometer mosfets with conventional architecture”, Electron Devices, IEEE Transactions on, 54(11) :3056-3063, Nov 2007. 107.

[94] Z. Xiao et aL. “Physical model of the impact of metal grain work function variability on emerging dual metal gate mosfets and its implication for sram reliability”, In Electron Devices Meeting IEDM, 2009 IEEE International, pages1-4, Dec 2009. 107.

[95] P.A Stolk et al. “Modeling statistical dopant fluctuations in mos transistors. Electron Devices”, IEEE Transactions on, 45(9) :1960-1971, Sep 1998. 107, 108.

[96] K. Takeuchi et al. “Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies”, In Electron Devices Meeting, 2007. IEDM 2007. IEEE International, pages 467-470, Dec 2007. 107.

[97] A Asenov et al. “Random dopant induced threshold voltage lowering and fluctuations in sub-0.1µm m-mosfet's : A 3-d atomistic simulation study”, Electron Devices, IEEE Transactions on, 45(12) :2505-2513, Dec 1998. 107.

[98] T. Tanaka et al. “Vth fluctuation induced by statistical variation of pocket dopant profile”, In Electron Devices Meeting, 2000. IEDM 00. Technical Digest. International, pages 271-274, Dec 2000. 107.

[99] I Ahsan, et al. “Rta-driven intra-die variations in stage delay, and parametric sensitivities for 65nm technology”, In VLSI Technology, 2006. Digest of Technica Papers. 2006 Symposium on, pages 170-171, 2006. 107.

[100] O. Weber et al. “High immunity to threshold voltage variability in undoped ultra-thin fdsoi mosfets and its physical understanding”, In Electron Devices Meeting, 2008. IEDM 2008. IEEE International, pages 1-4, Dec 2008. 107, 109.

[101] T. Fischer et al. “A 90nm variable-frequency clock system for a power-managed Itanium-family processor”, in Proceedings of the IEEE International Solid- State Circuits Conference (ISSCC) (2005), pp. 294–295.

[102] T. Sakurai et al. “Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas”, IEEE J. Solid-State Circuits 25(2), 584–594 (1990).

[103] J. Choi et al. “Thermal-aware task schedulingat the system software level”, in Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED) (2007), pp. 213–218

[104] J. Keane, et al. “An All-In-One Silicon Odometer for Separately Monitoring HCI, BTI and TDDB”, IEEE Journal of Solid-State Circuits, 45(4), pp.817–829, 2010.

[105] S. Markov et al. “Drain Current Collapse in Nanosaled Bulk MOSFETs due to Random Dopant Compensation in the Source/Drain Extensions”, IEEE Transactions on Electron Devices, 58(8), pp.2385–2393, 2011.

[106] J. Tschanz et al., “Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low-power and high-performance microprocessors”, IEEE J. Solid State Circuits, Vol. 38, No. 5, May 2003.

[107] J. Tschanz et al., “Dynamic sleep transistor and body bias for active leakage power control of microprocessors”, IEEE J. Solid State Circuits, Vol. 38, No. 11, Nov 2003.

[108] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, “Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage”, IEEE J. Solid-State Circuits, Vol. 37, Issue 11, pp. 1396–1402, Nov. 2002.

[109] J. Tschanz et al., “Adaptive frequency and biasing techniques for tolerance to dynamic temperature-voltage variations and aging”, IEEE ISSCC Dig. Tech. Papers, Feb. 2007.

[110] http://www.synopsys.com/products/solutions/galaxy/ccs/cc_source.html, 2006

[111] http://www.cadence.com/Alliances/languages/Pages/ecsm.aspx, 2007

[112] N. Drego et al., "Lack of Spatial Correlation in MOSFET Threshold Voltage Variation and Implications for Voltage Scaling", IEEE Transactions on Semiconductor Manufacturing, May 2009.

[113] N. cheikh et al., "Reliability Compact Modeling approach for Layout Dependent Effects in advanced CMOS nodes ", In IEEE International Reliability Physics Symposium (IRPS), April 2017

[114] D. Blaauw et al., "Statistical timing anlaysis: From basic principles to state of the art", IEEE Transactions on Computeraided Design of Integrated Circuits and Systems, 2008.

[115] A. Agarwal et al. "Statistical timing analysis using bounds", In DATE 03.

[116] http://www.synopsys.com/tools/implementation/signoff/capsulemodule/primetime_advancedocv_wp.pdf

[117] V. Huard et al. "Adaptative wear out management with in-situ management", Internationnal Reliability Physics Symposium (IRPS), pp. 6B.4.1 - 6B.4.11, 2014

[118] M. Nicolaidis, "Time redundancy based soft-error tolerant circuits to rescue very deep submicron," in Proc. 17th IEEE VLSI Test Symp., Dana Point, CA, USA, Apr. 1999, pp. 86–94

[119] M. Nicolaidis, "Circuit logique protégé contre des perturbations transitoires,"Patent WO2000054410 A1, Mar. 9, 2000

[120] S.DAS et al. "Razor II: In Situ Error Detection and Correction for PVT and SER Tolerance" IEEE J. Solid-State Circuits, vol. 44, no. 1, pp. 32–48, Jan. 2009.

[121] M. Agarwal et al. "Circuit failure prediction and its application to transistor aging," in Proc. 5th IEEE VLSI Test Symp., Berkeley, CA, USA, May 6–10, 2007, pp. 277–286

[122] N. Pour "Comparison of in-situ delay monitors for use in Adaptive Voltage Scaling". Advances in Radio Science, 10, pp.215–220

[123] D. Bull et al. "A power-efficient 32 bit ARM processor using timing-error detection and correction for transient-error tolerance and adaptation to PVT variation", IEEE J. Solid-State Circuits 46(1), 18–31 (2011)

[124] M. Wirnshofer et al. "A variation-aware adaptive voltage scaling technique based on in-situ delay monitoring," in Proceedings of the 14th IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS) (2011), pp. 261–266

[125] M. Wirnshofer et al. "An energy-efficient supply voltage scheme using in-situ pre-error detection for on-the-fly voltage adaptation to PVT variations", in Proceedings of the 13th International Symposium on Integrated Circuits (ISIC) (2011), pp. 94–97

[126] M. R. Choudhury et al. "Masking timing errors on speed-paths in logic circuits," in Proc. Design Automation and Test in Europe, pp. 87–92 2009.

[127] M. Kurimoto et al. "Phase-adjustable error detection flip-flops with 2-stage hold driven optimization and slack based grouping scheme for dynamic voltage scaling," in Proc. Design Automation Conference, pp. 884–889, 2008

[128] W. Li et al. “On path selection in combinational logic circuits”.IEEE Trans Comput-Aided Des Integr Circ Syst 1989;26(1):56–63

[129] M. Sharma et al. “Finding a small set of longest testable paths that cover every gate”, In IEEE International test conference (ITC), December 2002. P.974–82.

[130] X. Lu et al. “Longest-path selection for delay test under process variation”. IEEE Trans Comput-Aided Des Integr Circ Syst 2005:1924–9.

[131] V. Zolotov et al. “Statistical path selection for at speed test”. IEEE Trans Comput-Aided Des Integr Circ Syst 2010:749–59

[132] L. Lai et al. “SlackProbe: A Low Overhead In Situ On-line Timing Slack Monitoring Methodology”, IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems, vol. 33 , pp.1168-1179,2014

[133] J.C Vazquez et al. “Process Variations-Aware Statistical Analysis Framework for Aging Sensors Insertion”, IEEE J Electron Test,2013

[134] A.F Gomez et al. “ Early Selection of Crititical Paths for Reliable NBTI Aging-Delay Monitoring”, IEEE VLSI 2016

[135] N. Pour, “Monitoring Concepts for Degradation Effects in Digital CMOS Circuits”, 2015

[136] K. Zeng et al. “An Improved Linear Syndrome Algorithm in Cryptanalysis With Applications”, 2001

[137] JL. Massey, “Shift-register synthesis and BCH decoding”. IEEE Trans Inform Theory 15:122–127, 1969

[138] H. Lee et al., “Implementation of the Chien search algorithm on application specific instruction set processor”, ICCE, 2012

[139] A. Benhassain et al. “Robustness of timing in-situ monitors for AVS management”, Internationnal Reliability Physics Symposium (2016)

[140] A. Benhassain et al. “In-situ slack monitors: Taking up the challenge of on-die monitoring of variability and reliability”, International Verification and Security Workshop (2016)

[141] A. Benhassain et al. “Timing in-situ monitors: implementation strategy and applications results”, Custom Integrated Circuits Conference (2015)

[142] M. R. Choudhury et al., “Masking timing errors on speed-paths in logic circuits,” in Proc. Design Automation and Test in Europe, pp. 87–92, 2009

[143] M. Kurimoto et al., “Phase-adjustable error detection flip-flops with 2-stage hold driven optimization and slack based grouping scheme for dynamic voltage scaling,” in Proc. Design Automation Conference, pp. 884–889, 2008

[144] K. Hirose et al., “Delay-compensation flip-flop with in-situ error monitoring for low-power and timing-error-tolerant circuit design,” Japanese Journal of Applied Physics, vol. 47, pp. 2779–2787, 2008

[145] M. R. Choudhury et al “TIMBER: Time borrowing and error relaying for online timing error resilience”

[146] J.Tshanz et al “Tunable Replica Circuit and Adaptive Voltage-Frequency Techniques for Dynamic Voltage, Temperature, and Aging Variation Tolerance”

Moniteurs de Vieillissement in-situ: Méthodologie d'intégration et application à la gestion dynamique de la fiabilité

Résumé

Dans les technologies avancées, la course à la haute performance ainsi que la limitation de contraintes de consommation de puissance obligent les designers à mettre en place des nouvelles solutions pour optimiser ces paramètres des circuits intégrés. Une solution possible serait l'adaptation dynamique de la tension d'alimentation (AVS) et éventuellement la fréquence de fonctionnement selon les conditions d'utilisations du circuit. Elle utiliserait des capteurs de violation de délai, nommés aussi moniteurs in-situ. En pratique, il s'agit d'un circuit qui détecte des pré-erreurs, c'est-à-dire des transitions d'un chemin de propagation qui arrivent près du front d'horloge. En se basant sur ces moniteurs de pré-erreurs, la tension pourrait être ajustée et ceci bien avant qu'une erreur se produise. Ainsi, la consommation est optimisée en exploitant les marges de tension inutilisées, utilisées de façon traditionnelle dans le milieu industriel, par la conception d'un circuit dans le pire cas de procédé, tension, et température.

Ce travail effectué au sein de STMicroelectronics est développé sur quatre chapitres. Le premier chapitre présente les évolutions technologiques au sein de STMicroelectronics, la variabilité technologique, ainsi que les mécanismes de dégradation monotone de type Bias Temperature Instability (BTI) et sous injections de porteurs chauds (HCI), ou le vieillissement. Aussi, il aborde les impacts liés à ces mécanismes de dégradation, ainsi que les variations du procédé de fabrication, la tension et la température (PVT) sur le délai des portes et des chemins logiques. Le second chapitre est consacré d'une part à l'analyse de l'impact de diverses variations de type PVTs et du vieillissement sur un circuit numérique. D'autre part, le chapitre 2 présente également l'état d'art des moniteurs publiés dans la littérature, ainsi que la caractérisation du moniteur in-situ utilisés dans les circuits de test. Dans le troisième chapitre nous présentons d'abord la méthodologie d'insertion des moniteurs dans le flot de conception des circuits digitaux. Dans un second temps, nous présentons la méthodologie de choix des chemins critiques qui seront observés par des moniteurs, en se basant sur les délais des chemins obtenus par l'analyse statique temporelle (STA) avant et après vieillissement. Le quatrième chapitre présente des résultats expérimentaux des circuits contenant des moniteurs, fabriqués dans les technologies 45nm, 28LP, et 28FDSOI. Ces circuits sont testés dans différents conditions d'utilisations : tensions, températures et usages du circuit mettant en œuvre la capacité des moniteurs de faire face à ces types de variations. Une proposition de méthodologie d'adaptation des paramètres de fonctionnement du circuit tel que la tension d'alimentation y est aussi présentée.

Mots clés : Moniteurs in-situ, AVS, DVFS, BTI, HCI, fiabilité, variabilité

Abstract

Increasing performance demands in advanced technology, together with limited energy budgets, force integrate circuit designers to think about new ways of saving power.

One innovative way of doing so is presented in this work. The newly developed Adaptive Voltage Scaling (AVS) scheme tunes the supply voltage of digital circuits according to the present Process, Voltage and Temperature variations as well as Aging (PVTA). The key components of the proposed approach are in-situ delay monitors, detecting late but still non-erroneous signal transitions (pre-errors). Based on the measured pre-error rate, the voltage is adjusted with a low-overhead control unit connected to the on-chip voltage regulator. This way power consumption is optimized, by exploiting unused timing margin, produced by state-of-the-art worst-case designs.

This work has been done in STMicroelectronics and has been developed over four chapters in order to integrate reliability progressively from devices to digital circuits: in the first chapter we are interested in technological developments that have been necessary from standard CMOS technologies (40LP, 28LP) to FDSOI technology for 28nm node as well as monotonic degradation mechanisms such as BTI and HCI. Then, the impact of these degradation mechanisms along with PVT on data-path timing is discussed. The second chapter summarizes the state-of-arts of in-situ monitors and shows the characterization results of three kinds of pre-error in-situ monitors. The third chapter looks first at the integration of in-situ monitor in the digital implementation flow. As a second step, a potential critical path selection algorithm is developed. This algorithm is based on sub-critical paths identified by Static Timing Analysis (STA) before and after aging. The fourth chapter exhibits the experimental results of digital circuits that contain in-situ monitors. These circuits are designed in 45nm, 28LP and 28FDSOI technology and tested under different PVTs along with workload conditions. The results demonstrate the robustness of in-situ monitor to cope with different kinds of variations. Also, the behavior of in-monitor using AVS scheme is shown.

Keywords : In-situ monitors, AVS, DVFS, BTI, HCI, Reliability, variability