Etude d'architecture et circuiterie digitale dans le régime sous-le-seuil en technologie submicronique

F. Abouzeid

UNIVERSITÉ DE GRENOBLE
INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
978-2-84813-160-3

THESE

pour obtenir le grade de

DOCTEUR DE L'Université de Grenoble
délivré par L'Institut Polytechnique de Grenoble

Spécialité : *Micro et Nano Électronique*

préparée au laboratoire **TIMA**

dans le cadre de **l'École Doctorale Électronique, Électrotechnique, Automatique** & **Traitement du Signal**

présentée et soutenue publiquement

par

Fady Abouzeid

le 18 Novembre 2010

Étude d'architecture et circuiterie digitale dans le régime sous le seuil en technologie submicronique

Gilles Sicard
Marc Renaudin

JURY

M.	Sorin Cristoloveanu,	Directeur de Recherche, CNRS,	Examinateur
M.	Denis Flandre,	Professeur des Universités, UCL,	Rapporteur
M.	Christian Piguet,	Professeur des Universités, EPFL,	Rapporteur
M.	Philippe Maurine,	Maitre de Conférences, UM2,	Examinateur
M.	Gilles Sicard,	Maître de Conférences, UJF,	Directeur de thèse
M.	Sylvain Clerc,	Ingénieur, STMicroelectronics,	Co-encadrant
M.	Marc Renaudin,	Directeur Technique, TIEMPO,	Invité, Co-directeur
MM.	Edith Beigné,	Ingénieur, CEA-LETI,	Invité

Remerciements

Ces trois années ont été le théâtre de plusieurs rencontres, toutes marquantes car elles ont enrichi mon savoir, et mon enthousiasme. Je vais donc profiter de ce petit espace de liberté pour remercier ceux qui ont pris part à cette fructueuse aventure.

Je commencerai par Sylvain Clerc, directeur technique et chef bien aimé. Au cours de ces trois années il a partagé sans retenue ses connaissances, son expérience, et ses contacts, qui m'ont permis de donner une autre dimension au travail que j'ai fourni. Sur le plan personnel, c'est une véritable rencontre, et je profite de ces lignes pour lui transmettre ma sincère amitié.
Je remercie chaleureusement mes directeurs de thèse, Marc Renaudin et Gilles Sicard, pour avoir contribué au succès de ces travaux par leur expertise et leur savoir. Avoir su faire preuve de recul, tâche au combien difficile, est une de leurs nombreuses qualités.

Je remercie également Philippe Roche, le "chef du chef", qui a parfaitement orienté les travaux de cette thèse pour les maintenir dans le cadre industriel. Ses conseils avisés m'ont apporté crédibilité et visibilité, ce qui s'est traduit par mon embauche au sein de son équipe. Je souhaite également adresser quelques mots à Pascal Urard, qui fut l'un de mes "N+2" au cours de cette thèse, et avec qui il était possible d'échanger autant sur la technique que sur les loisirs. Je leur transmet toute mon amitié.

Il m'est impossible d'oublier mes compagnons de route dans ces remerciements. Tout d'abord à Fabian Firmin, collègue et voisin de bureau, devenu un ami précieux dont les jeux de mots resteront pour la prospérité. Il y a également Lahcen Hamouche, capable de vous surprendre en glissant quelques rondelles de citron cuit dans un tajine. Et que dire de Nicolas Hébert, le roi du système D et des ficelles économiques ? Il y a aussi Daniele D'Agostino, d'une sympathie sans limite, et qui, je le rappelle, me doit une pizza. Je souhaite également du courage à Julien Le-Coz et Dimitri Soussan dans la conclusion de leur thèse. Je remercie également mes camarades Slawosz Uznanski et Josep Torras. Slawosz m'a fait découvrir la sagesse et la gentillesse polonaise : je pourrai lui confier un lingot d'or et être sur de le retrouver entier à mon retour. Josep, quand à lui, en plus d'avoir un humour de haute volée, m'a permis de progresser au football, et il faut dire que j'en avais besoin.

Les autres membres de l'équipe sont également à l'origine du bon déroulement de cette thèse. Je pense notamment à Gilles Gasiot et Jean-Marc Daveau, que je remercie et qui sont désormais mes collègues

pour de nombreuses années. J’adresse également quelques mots à Dominique Zamora, qui a grandement contribué à une partie des travaux au cours de son stage. Enfin, je remercie Charly Chatelain dont la capacité à percevoir le tonnerre est sans précédent.

De nombreux collaborateurs au sein de STMicroelectronics ont également contribué aux travaux évoqués dans ce manuscrit. Je pense notamment à Etienne Maurin et Frédéric Avellaneda pour les aspects librairies. Je remercie Sylvain Landelle et Pierre-François Ollagnon pour la simulation assistée. Je remercie Manuel Sellier pour ses conseils sur la vérification temporelle. Je remercie Louis Rolindez pour la formation accélérée sur la synthèse et le placement/routage. Je remercie Vincent Heinrich pour la conception du démonstrateur numérique. Je remercie Andrea Veggetti, Abhishek Jain, Vivek Bathia, Rajesh Kumar, Balwant Singh et Philippe Flatresse pour la réalisation des nombreux véhicules de tests. Je remercie Dennis Crippa, Sebastien Haendler et Fabrice Argoud pour les mesures sur silicium qui ont grandement contribué à la validation des travaux. Enfin, je remercie David Turgis, Brice Lhomme et Olivier Callhen pour les conseils avisés sur tous les aspects mémoire.

Je terminerai en remerciant chaleureusement ma famille, car ils ont su me donner goût et intérêt pour les études d’autant plus que je n’ai, au cours de mon enfance, jamais manqué de rien. Je souhaite également remercier ma belle famille pour son soutient au cours des dernières années.

Je conclurai en remerciant ma femme Laetitia. Je la remercie pour son soutient sans faille, et pour ces années de bonheur passées et à venir. Je lui dédie ce manuscrit.

Résumé

L’alimentation des circuits à très faible tension, permettant une efficacité énergétique multipliée par 10, répond aux contraintes des applications mobiles, au prix d’une variabilité accrue limitant la prédiction des résultats et nécessitant des efforts et méthodologies de conception spécifiques. Cette thèse associe la conception à très faible tension aux exigences industrielles, et présente le développement de cellules digitales optimisées pour la très faible tension, par une méthodologie indépendante de la technologie. Ces cellules, validées par des mesures sur silicium en technologie CMOS 40nm, ont conduit à la fabrication d’un circuit numérique, dont le test met en évidence les adaptations permettant d’améliorer le rendement. Enfin, une cellule mémoire a été conçue et optimisée à très faible tension, ainsi que des solutions d’assistance en lecture et en écriture pour renforcer la tolérance à la variabilité. Un démonstrateur 128kb est fabriqué en 65nm pour valider ces développements.

Abstract

Ultra-low voltage enables to answer the limitations of the wearable mobile applications with an energy efficiency improved by a factor x10, at the price of an increased transistor variability limiting the predictability of the results.
In respect with the industrial requirements, this thesis presents the development of logical cells optimized at ultra-low voltage, using a technology independent methodology. These cells, certified then validated by silicon measurements in 40nm, led to the design of a digital circuit, fabricated on silicon, which analysis highlighted the adaptations needed to enhance the yield and the predictability of the results. At last, a memory cell was developed and optimized at ultra-low voltage. Read and write assist solutions were conceived in order to reinforce the tolerance to variability. A 128kb memory demonstrator was then fabricated in 65nm to validate these developments.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'industrie de la micro-électronique, motivée par les besoins du grand public, est engagée dans une course à la performance impliquant un rendement élevé et des coûts de production réduits. D'autres besoins se sont ensuite profilés, notamment l'électronique bas coût qui suite à l'émergence des étiquettes RFID fût à l'origine d'un développement accéléré de l'électronique organique [1; 2], qui offre des procédés de fabrication simplifiés. Les besoins spatiaux, médicaux, ou ayant un lien avec la sécurité des utilisateurs, ont nécessité le développement de technologies robustes [3], capables de limiter l'effet des perturbations induites par les particules chargées en provenance de l'environnement.

Enfin, l'économie d'énergie, née de la volonté d'améliorer l'autonomie des applications mobiles limitée par la durée de vie des batteries actuelles, est à l'origine du développement de solutions à faible consommation, nommé « Low Power » en langue anglaise.

1 De la faible consommation à la très faible tension

L'objectif du Low Power est de fournir un ensemble de mécanismes [4] permettant la réduction de la consommation énergétique d'une application, tout en préservant ses performances. Les systèmes mobiles haute-performance, ainsi que les applications à usage domestique, sont directement concernés par ces solutions. Les mécanismes utilisés sont dans la plupart des cas liés à l'architecture des systèmes [5]. On peut citer le parallélisme, qui consiste à exécuter une tache sur plusieurs unités de calcul à la fois, ou le pipelining, correspondant à l'introduction d'étages supplémentaires dans un circuit, pour exécuter différentes tâches en même temps. Le gain en temps de fonctionnement obtenu par l'usage de ces solutions permet de réduire la tension d'alimentation pour le convertir en un gain en consommation. Il est également possible d'optimiser la consommation en utilisant un code binaire différent, ou en modifiant l'organisation des horloges [6]. Cependant, le Low Power n'est pas conçu pour répondre aux exigences des systèmes dont la priorité est l'autonomie, au détriment des performances.

Ces exigences peuvent être adressées par une importante réduction de la tension d'alimentation, conformément à la dépendance quadratique en tension de la consommation énergétique. Les premiers écrits concernant l'utilisation de la très faible tension datent de 1972 [7], alors que l'idée d'une utilisation de la faible inversion est née des travaux du Dr. Vittoz en 1960 [8].

La très faible tension, ou Ultra-Low Voltage (ULV) en langue anglaise, est la solution choisie dans le cadre de cette thèse pour répondre aux besoins des applications dont la consommation énergétique est prioritaire sur les performances. Il apparait que les principales limitations de cette solution soient l'importante dégradation des performances, mais surtout l'augmentation de la sensibilité des circuits aux variations liées aux différentes étapes de fabrication [9].

2 Les domaines d'applications

Certaines applications mobiles, dont les besoins en termes de fréquence de fonctionnement sont faibles tels que les dispositifs intégrés à l'habillement [10], composent le marché tributaire de cette solution. De même que les réseaux de capteurs, fonctionnant à fréquence faible, mais de manière continue. Il apparait que le principal marché concerne les applications médicales portées, telles que les aides auditives [11] dont la faible autonomie est la conséquence d'une évolution lente des batteries. La très faible tension peut également être utilisée au sein d'une application à tension nominale, dans des modules faibles performances qui entourent un processeur standard, ou encore comme solution de maintient lorsque le taux d'activité est faible [12]. La particularité de ces applications est la nécessité d'un continuum de fonctionnalité entre la tension nominale et la très faible tension.

La source d'énergie est une donnée importante dans la mesure où il existe peu de sources d'alimentation fournissant une faible tension. En prenant pour exemple l'aide à l'audition, ces appareils fonctionnent avec des piles alimentées autour de 1V, et possèdent une durée de vie moyenne de deux jours : la tension aux bornes de la pile décroit continuellement, jusqu'à atteindre la tension minimale de fonctionnement du système, soit environ 0,8V. En offrant une fonctionnalité à très faible tension au système, son temps d'utilisation sera étendue pour une même batterie. Il est toutefois nécessaire de garantir un continuum de fonctionnalité sur l'ensemble de la plage de tension.

Pour les applications destinées à un fonctionnement à tension unique, il existe des convertisseurs de tension, dont le rendement en termes de

courant est entre 80% et 90% dans les développements les plus récents [13]. On note également la possibilité d'utiliser des sources d'alimentations dites « naturelles », tel que le solaire, où celles actuellement développées autour des technologies MEMS [14], permettant une autonomie complète du système. Ces solutions ne sont pas matures à ce jour.

3 Projet de recherche

Ce manuscrit présente les travaux de recherche et les développements réalisés sur la très faible tension dans le cadre de cette thèse CIFRE, pour le compte du laboratoire TIMA et de la société STMicroelectronics. L'objectif de cette thèse est d'amener la conception à très faible tension dans le milieu industriel avec pour finalité la commercialisation de cette technologie. Bien que la très faible tension concerne différents domaines de l'électronique, cette thèse adresse la conception numérique.

Dans un premier temps, une étude bibliographique présente les réalisations de la communauté scientifique sur les différents aspects de la conception à très faible tension : technologie, cellules logiques, circuits et mémoires.

Ces observations conduisent à la définition de directives permettant une entière compatibilité avec les standards de l'industrie. A la suite d'une calibration de la simulation avec l'état de l'art et le comportement sur silicium, le second chapitre décrit le développement d'une méthodologie de conception de cellules logiques optimisées pour le fonctionnement à très faible tension.

Ces cellules sont ensuite utilisées pour le développement d'un circuit numérique. Le chapitre 3 présente la réalisation de ce circuit par l'utilisation du flot de conception industriel usuel. Les mesures sur silicium de ce circuit sont alors analysées pour adapter le flot tout en maintenant la compatibilité industrielle.

Enfin, le dernier chapitre présente la conception d'une cellule mémoire à la suite d'un comparatif entre l'offre industrielle et les mémoires les plus avancées de l'étude bibliographique. Cette mémoire est ensuite intégrée dans un démonstrateur puis fabriquée sur silicum. Les résultats de ce démonstrateur sont ensuite analysés, avant de conclure sur l'ensemble de la contribution et les perspectives de développement.

Chapitre 1

État de l'art

1 Le transistor MOS

1.1 Principe de fonctionnement

Le terme MOS signifie Metal-Oxide-Semiconducteur . Le transistor MOS est ainsi composé d'une grille en polysilicium et/ou métal, d'une source et d'un drain dopés, et d'un substrat. Il se décline en deux versions selon le type de dopage : PMOS et NMOS (Figure 1.1).

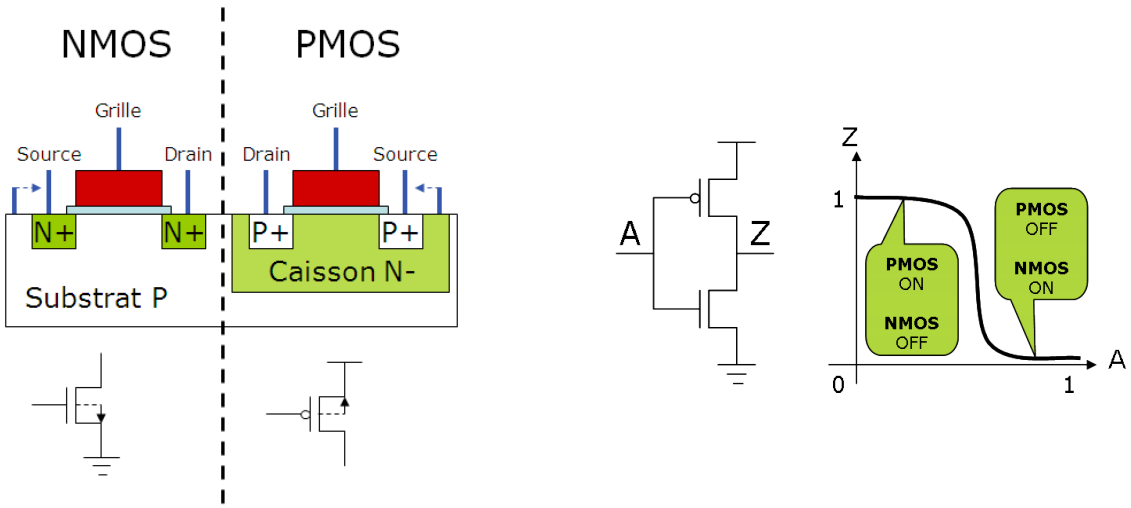

Figure 1.1 – Transistors MOS et exemple d'un inverseur CMOS.

Le NMOS présente une source et un drain dopés N+. Pour qu'il soit passant, la tension V_{GS}, entre la Grille et la Source, doit être positive pour créer une inversion des porteurs du substrat dans le canal, dont le dopage intrinsèque est de type P. Le NMOS permet de conduire la valeur numérique 0, correspondant à la masse.
Le PMOS présente une source et un drain dopés P+, et est placé dans un caisson dopé N-. Pour qu'il soit passant, la tension V_{GS} doit être négative pour créer une inversion des porteurs du caisson dans le canal.

Le PMOS permet de conduire la valeur numérique 1, correspondant à la tension d'alimentation.
Ces deux transistors sont complémentaires et l'utilisation de cette complémentarité est nommée logique CMOS. Ainsi, une porte logique composée d'un NMOS et d'un PMOS avec grilles communes et drains communs permet de réaliser la fonction d'inverseur.

Le transistor présente une quatrième prise pour le substrat, connectée par défaut à la source. Une alimentation dissociée permet de moduler le passage du courant dans le canal.

L'expression du courant dynamique d'un transistor en fonctionnement nominal est :

$$I_{saturation} = \frac{W}{L} \cdot \frac{\mu_{eff} \cdot C_{ox}}{2} \cdot (V_{GS} - V_{TH})^2 (1 + \lambda V_{DS}) \tag{1.1}$$

avec W la taille des drains/sources, L la longueur de grille, μ_{eff} la mobilité effective, C_{ox} la capacité de l'oxyde de grille, V_{TH} la tension de seuil, V_{DS} la tension entre le drain et la source, et λ le paramètre de correction de la longueur effective de grille, conséquence de l'effet Early.

Les effets de la réduction de la tension d'alimentation sont régis par les différentes composantes de l'équation énergétique (Équation 1.3). On distingue ainsi la composante dynamique et la composante statique, dite énergie de fuite. La première décrit l'énergie consommée par le dispositif en activité, soit l'énergie utile. L'énergie de fuite quant à elle correspond à l'énergie perdue, qui a été dissipée à travers des mécanismes de fuites.

$$E_{TOTALE} = E_{Dynamique} + E_{Statique} \tag{1.2}$$

$$= \frac{1}{2} \cdot C_S \cdot V_{DD} + I_{fuite} \cdot V_{DD} \cdot t_{délai} \tag{1.3}$$

1.2 Les courants de fuites

En opposition au courant dynamique, un courant de fuite est défini comme tout courant sans service rendu pouvant intervenir dans un transistor lorsque celui-ci se trouve dans un état bloqué ou passant. La Figure 1.2 rassemble les six courants de fuites majeurs présents dans le transistor. Chacun de ces courants de fuite correspond à un mécanisme différent [15].

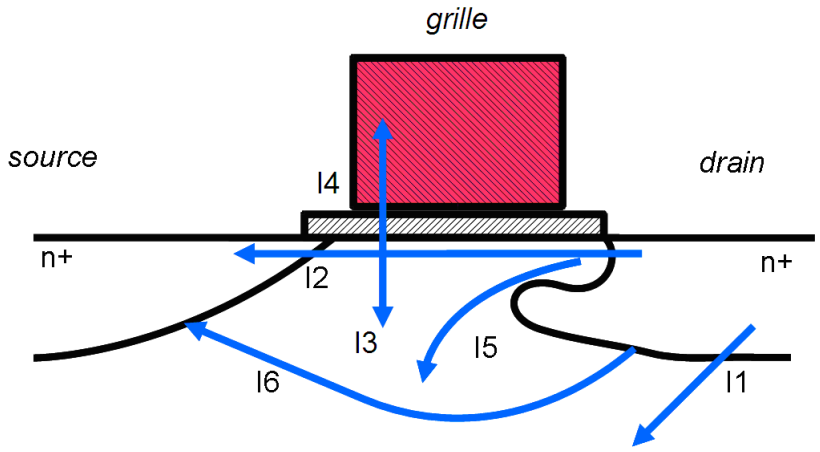

Figure 1.2 – Les courants de fuite du transistor MOS.

Le courant de fuite des jonctions P-N i_1

Lorsque deux jonctions de polarités différentes sont en contact, les porteurs proches de cette zone se déplacent par diffusion et se recombinent avec ceux de la polarité opposée. Il se crée alors, à la place de la jonction, une zone de charge d'espace, nommée zone de déplétion. Les ions présents, dépourvus de leurs porteurs, créent alors un champ qui s'oppose à ce phénomène, ce qui place la jonction dans un état stable. Cependant, certains porteurs minoritaires générés thermiquement, par exemple un trou dans la région dopée N+, passent alors d'une zone à l'autre, créant un courant inverse. Il s'agit du courant de fuite i_1. Lorsque la source et le drain sont fortement dopés, cela favorise le passage de ces porteurs par effet tunnel, phénomène appelé Band-to-Band tunneling [16] (BBT) en langue anglaise, dont le courant devient dominant.

Ces fuites sont sensibles au champ électrique à travers la jonction, et augmentent si la tension d'alimentation est plus élevée. La conception des sources et drains est également un facteur important dans l'apparition de ces courants. Plus le dopage sera élevé et abrupt, marquant ainsi fortement la délimitation entre la zone N et P, plus les fuites de jonctions seront importantes. Les solutions permettant de limiter ce courant de fuite sont les solutions technologiques autour du substrat, tels que le Halo doping [17] en langue anglaise, qui consiste à créer des zones dopées dans le canal, ou encore l'utilisation de plaques SOI, pour Silicon-on-Insulator en langue anglaise, qui élimine la présence du substrat sous le drain et la source, et donc une partie des jonctions. D'un point de vue conception, seules l'utilisation d'une tension d'alimentation modérée, ou de l'application d'une différence de potentiel entre la source et le substrat [18], permettent de le réduire. Ces fuites n'auront que peu d'incidence si l'on fonctionne à très faible tension.

Le courant de fuite sous-le-seuil i_2

Ce courant est au coeur du sujet de cette thèse. Il correspond au passage de porteurs alors que le transistor se trouve dans un état dit de faible inversion. Lorsque l'inversion des porteurs à la surface commence mais n'est pas suffisamment importante pour permettre au courant dynamique de passer, certains porteurs parviennent à se diffuser le long du canal, sous l'effet du champ électrique. Il s'agit, pour les technologies submicroniques, d'un des courants de fuite le plus important. Il s'accroit avec l'utilisation de tensions plus importantes, ainsi que l'augmentation de la température. L'énergie donnée aux porteurs facilite leur diffusion.

Les solutions permettant la réduction de ce courant de fuite sont multiples, notamment les solutions technologiques appliquées au canal et son dopage, ou l'augmentation de la longueur de grille. Cependant, ce courant correspond au courant dynamique mis en oeuvre à très faible tension : les optimisations technologiques destinées à le limiter peuvent avoir un impact négatif sur le fonctionnement à très faible tension du transistor. On notera le comportement en température de ce courant de fuite, opposé à celui du courant dynamique lors du fonctionnement à tension nominale.

Le courant de fuite à travers l'oxyde i_3

La réduction des dimensions du transistor, et donc de l'isolant, a provoqué l'apparition d'un courant de fuite à travers celui-ci. La tension appliquée sur la grille est la source d'un champ électrique, celui-ci pouvant provoquer le claquage de l'isolant si la tension est trop élevé, l'isolant perdant alors son efficacité. Toutefois, à tension nominale, le champ électrique aux bornes de l'isolant peut donner suffisamment d'énergie pour que les électrons finissent par traverser par effet tunnel la fine barrière imposée par l'isolant. Ce courant de fuite augmente avec l'utilisation d'une tension d'alimentation plus élevée, et d'un isolant plus fin.

Les solutions actuelles pour répondre à ces fuites sont l'utilisation de nouveaux matériaux. On trouve désormais des isolants avec des constantes diélectriques plus élevées [19], ou High-K en langue anglaise, en remplacement du traditionnel SiO2 dans les technologies 45nm et 32nm. Lorsque des tensions plus importantes sont nécessaires, une épaisseur d'isolant plus grande est utilisée. L'impact de ces fuites à très faible tension est limité.

La capture d'électrons chauds i_4

Ce phénomène correspond à un mécanisme de dégradation des performances du transistor. Lorsque le champ électrique dans le canal est suffisamment important, les électrons deviennent très énergétiques : on parle alors d'électrons chauds. Ces électrons parviennent alors à passer la barrière imposée par l'isolant pour se retrouver piégés dans celui-ci. Cette capture est peu probable à très faible tension.

Le courant de fuite du drain induit par la grille i_5

Le GIDL [20], pour Gate-Induced Drain Leakage en langue anglaise, est un phénomène observé dans la zone où la grille du transistor recouvre le drain. Lorsque la grille du transistor est alimenté de telle sorte que le canal soit en accumulation, phénomène opposé de l'inversion, la zone de recouvrement devient déplétée. Cette modification entraine l'intensification des champs entre cette jonction et la grille, jusqu'à provoquer des phénomènes dit d'avalanche. Les porteurs qui traversent l'isolant par effet tunnel depuis la grille sont alors balayés dans le substrat, moins dopé.

Cet effet est amplifié par l'utilisation de tensions plus importantes, et l'utilisation d'un isolant de faible épaisseur. Enfin, un faible dopage du substrat accentue l'effet d'accumulation de la polarisation, ce qui favorise le GIDL. Les solutions pour réduire cet effet sont identiques à celles évoquées précédemment. Des traitements et matériaux dédiés à l'interface substrat - isolant, et un dopage abrupt et important du drain, permettent d'en diminuer l'effet. Cette dernière solution induit toutefois d'autres courants de fuite, étudiés précédemment.

Le courant de fuite dit de perçage i_6

Le courant de perçage, « punchthrough [21] » en langue anglaise, intervient lorsque les zones de déplétion de la source et du drain se rejoignent. Cette zone entièrement déplétée incite le passage du courant dans le volume du substrat. Ce phénomène se produit lorsque la tension appliquée sur le drain augmente, jusqu'à atteindre une tension de claquage, qui va permettre le passage des porteurs. Pour empêcher le passage de ce courant de fuite, des implantations de dopants sont effectuées. Ces solutions permettent de supprimer complètement ce phénomène.

La réduction de l'ensemble de ces courants de fuites fait parti des objectifs du Low Power. La coupure de certains éléments d'un circuit lorsqu'ils ne sont pas utilisés en est un exemple [22].

2 Le courant de fuite sous-le-seuil

2.1 Expression

Lorsque la tension d'alimentation du transistor est réduite, l'expression quasi-linéaire du courant en fonction de la tension devient exponentielle (Figure 1.3).

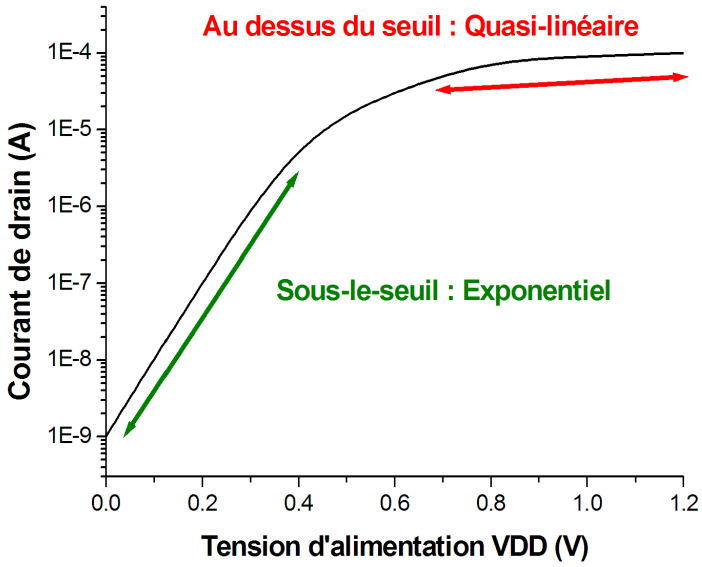

Figure 1.3 – Courant de drain en fonction de la tension V_{DD}

Il en résulte alors l'expression standard suivante [23] :

$$I_{subTH} = \frac{W}{L} \cdot \mu_{eff} \cdot C_{ox} \cdot (m-1) \cdot \nu_T^2 \cdot e^{\left(\frac{V_{GS}-V_{TH}}{m \cdot \nu_T}\right)} \cdot \left(1 - e^{-\frac{V_{DS}}{\nu_T}}\right) \qquad (1.4)$$

$$m = \left(1 + \frac{3 \cdot T_{ox}}{W_{dep}}\right)\left(1 + \frac{11 \cdot T_{ox}}{W_{dep}} \cdot e^{\frac{\pi \cdot L_{eff}}{2\left(W_{dep}+3 \cdot T_{ox}\right)}}\right)$$

où L_{eff} est la longueur efficace de la grille, m le facteur de la pente sous-le-seuil, et ν_T le coefficient thermique kT/q. Cette expression simplifiée permet d'observer les principales dépendances du courant sous-le-seuil.

Si à une tension nominale de 1,1V pour les technologies 40nm, ce courant est considéré comme pénalisant car perdu, il constitue le courant dynamique des transistors à très faible tension. Le courant de fuite pour un fonctionnement à très faible tension est le courant traversant un transistor lorsque celui-ci est bloqué.
Pour qu'une porte logique soit fonctionnelle, il est nécessaire que le

courant dynamique soit plus important que le courant de fuite, la logique CMOS faisant intervenir l'opposition entre un réseau de transistors passant et un réseau de transistors bloqué. Si le réseau de transistor bloqué a un courant de fuite égal ou supérieur au courant dynamique du réseau passant, on est en situation de court circuit, et l'information en sortie est erronée (Figure 1.4).

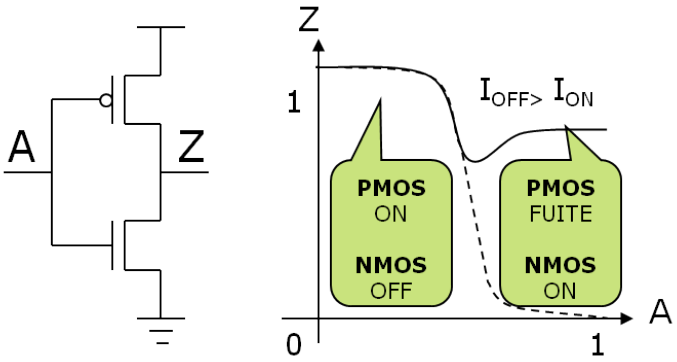

Figure 1.4 – Représentation de l'effet du rapport I_{ON}/I_{OFF} sur le fonctionnement d'un inverseur CMOS.

2.2 Dépendances

L'Équation 1.4 nous éclaire sur les dépendances linéaires du courant en fonction des dimensions du transistor. On constate que la variation du courant est exponentielle en fonction de la tension d'alimentation et de la tension de seuil V_{TH}. Les variations de V_{TH} induites par les variations de température et les défauts de fabrication se répercutent exponentiellement sur le courant. La conception à très faible tension est rendue complexe par cette sensibilité.

3 La puissance et l'énergie à travers le transistor MOS

L'objectif de la très faible tension est de réduire les dépenses énergétiques. On distingue deux notions permettant d'évaluer ces dépenses : La puissance et l'énergie consommée. On peut exprimer la puissance moyenne P ainsi :

$$P_{TOTALE} = P_{Dynamique} + P_{Statique} \tag{1.5}$$

$$P = \frac{1}{t}\int_0^t P_{TOT}(t)\cdot dt$$
$$P = \frac{1}{t}\int_0^t \left(P_{Dyn}(t)+P_{Stat}(t)\right)\cdot dt$$

P_{Stat} étant constant :

$$P = \frac{1}{t}\int_0^t \left(P_{Dyn}(t)\cdot dt\right)+I_{fuite}\cdot V_{DD}$$
$$P = \frac{1}{t}\int_0^t (I_{ON}(t)\cdot V_{DD}(t)\cdot dt)+I_{fuite}\cdot V_{DD}$$
$$P = \frac{1}{t}\int_0^t \left(C_s\cdot\frac{dV}{dt}\cdot V_{DD}(t)\cdot dt\right)+I_{fuite}\cdot V_{DD}$$
$$P=\frac{1}{2}C_s\cdot V_{DD}^2\cdot f+I_{fuite}\cdot V_{DD} \quad (1.6)$$

où C_s est la capacité commutée, et f la fréquence. On note que la puissance dynamique dépend quadratiquement de la tension appliquée au transistor.

Pour évaluer la consommation d'un dispositif, notamment pour les applications fonctionnant sur batteries, on utilise la notion d'énergie, ou de consommation pour un service rendu. On l'obtient en multipliant la puissance moyenne par le temps du service. Cette énergie dispose de deux composantes : dynamique et fuite. On peut alors l'écrire sous la forme [24] :

$$E_{TOT}=P\cdot t_{délai}$$
$$E_{TOT}=\frac{1}{2}\cdot C_S\cdot V_{DD}^2+I_{fuite}\cdot V_{DD}\cdot t_{délai} \quad (1.7)$$

On retrouve l'Équation 1.7. Si l'énergie dynamique est dépendante de manière quadratique de la tension appliquée, on observe que l'énergie de fuite est à la fois dépendante de la puissance dissipée et du délai, qui correspond au temps de commutation.

4 La tension d'énergie minimale V_{Emin}

On note que le délai est également dépendant de la tension :

$$t_{délai}=C_S\cdot\frac{V_{DD}}{I_{sous-le-seuil}} \quad (1.8)$$

La diminution de la tension augmente le délai, dominé par l'exponentielle induite par le courant. L'énergie perdue pendant ce délai augmentera de concert. Ainsi, la composante E_{fuite} de l'énergie va s'accroitre pour devenir plus importante que la composante dynamique (Figure 1.5). L'énergie totale connait donc une phase décroissante correspondant à la diminution de l'énergie dynamique, suivit d'une phase croissante associée à la domination de l'énergie de fuite. Le minimum d'énergie entre les deux phases correspond à une tension du minimum d'énergie V_{Emin}.

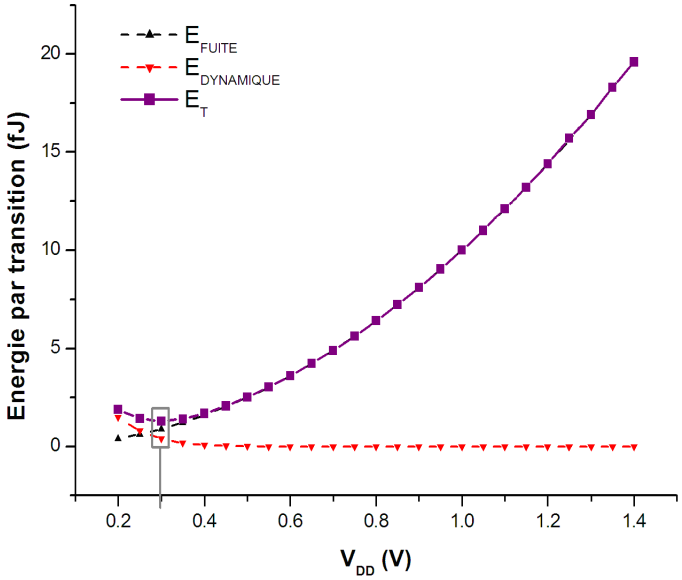

Figure 1.5 – Énergie totale par transition en fonction de la tension

L'objectif des dispositifs à très faible tension est de fonctionner autour de cette tension théorique V_{Emin} afin de profiter d'une consommation énergétique minimale. Le fonctionnement à une tension inférieure à cette valeur induit une augmentation de la consommation énergétique.

Plusieurs écrits sont consacrés à la modélisation de ce point d'énergie minimum. Son équation est exprimée ainsi [25] :

$$V_{Emin} = \left[1.587 ln\left(\frac{\eta}{U_{eff}}\right) - 2.355\right] \cdot m \cdot \frac{kT}{q} \tag{1.9}$$

où U_{eff} désigne l'activité du circuit, η sa profondeur logique. La pente sous-le-seuil, l'activité et la profondeur logique permettent donc de modifier V_{Emin}.
Cette tension V_{Emin} ne décrit pas la tension minimale de fonctionnement du circuit. Certaines publications ont également mis en équation cette tension d'alimentation minimale de fonctionnalité [26] :

$$V_{dd,lim} = 2 \cdot \frac{kT}{q} \left[1 + \frac{C_{fs}}{C_{ox} + C_d} \right] \cdot ln \left(2 + \frac{C_d}{C_{ox}} \right) \quad (1.10)$$

où C_{fs} est la capacité liée aux états de surface rapides à l'interface, et C_d la capacité de la zone déplétée.
Selon Meindl, la limite théorique de fonctionnement d'un inverseur de la technologie CMOS se situe autour de 36mV [27].

5 La variabilité

Comme observé précédemment, le courant sous-le-seuil est exprimé par une exponentielle de la tension de seuil, qui est sujette à différentes sources de variations, et notamment les variations induites par les procédés de fabrication [28; 29], qui regroupent les étapes technologiques dédiées à la réalisation d'un composant. La finesse de gravure réduite, atteignant la limite de résolution dont sont capables les appareils de lithographie actuels, introduit des imprécisions sur le résultat final, se traduisant par des variations sur les dimensions ou le dopage des transistors, faisant varier sensiblement la tension de seuil. Cette variation est amplifiée à très faible tension par la dépendance exponentielle.

Les variations sont de deux types :

- **Systématiques :** Les variations peuvent être modélisées.
- **Aléatoires :** Les variations ne peuvent être observées que par la simulation statistique.

Il est montré [30] que la principale source de variation aléatoire provient du taux de dopage, dont les fluctuations sont importantes. Dans des conditions de fonctionnement nominales, au dessus du seuil, ces variations sont d'importances égales à celles liées aux dimensions du transistor.

Lors de l'étude du courant sous le seuil, nous avons évoqué sa dépendance en température. Un contrôle de l'évolution de la fonctionnalité et des performances en fonction de la variation de la température est donc nécessaire. Les PMOS et NMOS n'ayant pas le même dimensionnement ni le même dopage, l'effet des variations peut-être différent.

Des recherches ont été effectuées pour limiter la sensibilité aux variations. Du coté de la conception, il est constaté que l'utilisation d'une profondeur logique plus importante conduit à une réduction de ces variations [31] car la composante aléatoire est moyennée le long du che-

min. De même, il est recommandé d'augmenter la longueur du canal [32]. Cela est confirmé par la modélisation de l'impact du dimensionnement sur la variabilité, réalisée par Pelgrom [33] :

$$\sigma_{Vth} = \frac{A_{Vth}}{\sqrt{WL}} \tag{1.11}$$

D'autres solutions, consistant en une alimentation dissociée du substrat et de la source, sont proposées [34; 35]. Enfin, il existe des solutions pour limiter les variations liées aux dimensions en technologie 40nm. On utilise une périodicité des grilles de transistors afin de maitriser la dispersion liée à la lithographie. Elle permet d'obtenir des motifs plus nets[36]. Cette solution n'est pas spécifique à la très faible tension.

On remarquera qu'à part le sur-dimensionnement et la profondeur logique, les solutions existantes permettent de corriger la variabilité d'un circuit dans son ensemble, ce que l'on appelle variations globales. Ces solutions ne permettent pas la correction directe des variations locales des transistors.

6 Les métriques de la très faible tension

L'utilisation d'une métrique standard permet d'évaluer les différents résultats de nos recherches, et de les comparer aux travaux effectués dans d'autres laboratoires et industries. Dans les publications concernant la très faible tension, on retrouve régulièrement la mesure de l'énergie, de la puissance, de la fréquence ainsi que la variabilité.

Plus précisément, l'énergie s'exprime en énergie par transition, ou Power Delay Product [37] dans la langue anglaise, telle que l'exprime l'Équation 1.7. Cette énergie, extraite en fonction de la tension d'alimentation, permet de retrouver graphiquement la tension V_{Emin}. On mesure ainsi le gain énergétique pour un même service rendu en fonction de la tension d'alimentation.

L'énergie par transition sera utilisée comme métrique principale pour nos études.

Certaines publications expriment l'énergie en fonction du temps de traversée plutôt que de la tension d'alimentation [38]. Cette métrique permet d'observer le passage du fonctionnement nominal au fonctionnement sous-le-seuil, par le comportement exponentiel du délai, mais également de visualiser les coûts et gains mutuels sur le couple énergie–vitesse qu'implique une réduction ou une augmentation de la tension [39]. Les courants de fuite sont également extraits pour mesurer l'impact des améliorations de conceptions sur la consommation lors des

états inactifs d'un circuit. L'énergie par transition totale varie alors en fonction du rapport entre les phases actives et statiques, appelé taux d'activité.

Le délai, correspondant à l'inverse de la fréquence maximale, exprime la performance du circuit en termes de vitesse.

On distingue également dans la plupart des publications une métrique usuelle de la variabilité en température, exprimée par l'énergie ou le délai en fonction de celle-ci. Concernant les variations liées aux procédés de fabrication, elles sont observées par des simulations Monte-Carlo. Ces simulations reproduisent les variations locales, en appliquant une modulation quasi-probabilistique des paramètres clés tels que les dimensions, le dopage et la mobilité. Il en résulte une distribution dont on extrait l'écart-type relatif. Ainsi, pour l'énergie ou le délai, on obtient une mesure à $3\sigma/\mu$ pour un intervalle de confiance à 99.7% ou σ / μ pour un intervalle de confiance à 95%, σ étant l'écart-type et μ la moyenne.

Enfin, la sensibilité aux variations des mémoires est évaluée par l'utilisation de métriques spécifiques [40] : la SNM, pour Static Noise Margin en langue anglaise, et la WM, pour Write Margin en langue anglaise. Ces métriques définissent les marges au-delà desquelles la mémoire sera incapable de lire (SNM) ou d'écrire (WM). L'ensemble des informations concernant la mémoire SRAM et ses paramètres sont décrits dans le Chapitre 4.

7 Les solutions actuelles

Le gain énergétique déjà établi par la diminution de la tension d'alimentation situe le challenge principalement dans la fonctionnalité des dispositifs à cette tension face à la sensibilité aux variations. Différentes solutions ont ainsi été développées dans la littérature.

7.1 Les solutions technologiques

Il existe désormais de nouveaux types de transistors qui permettent de répondre aux besoins de la très faible tension, bien qu'elle n'en soit pas l'objectif premier. La première solution consiste en l'utilisation d'un substrat SOI (Figure 1.6). Le SOI, Silicon-On-Insulator en langue anglaise, consiste à poser le substrat sur un isolant. Cela permet de réduire les capacités parasites introduites par les jonctions [41], et d'obtenir un meilleur contrôle des fuites, favorisant ainsi le fonctionnement

sous-le-seuil [42] en termes de consommation statique, de délai et de variabilité [43].

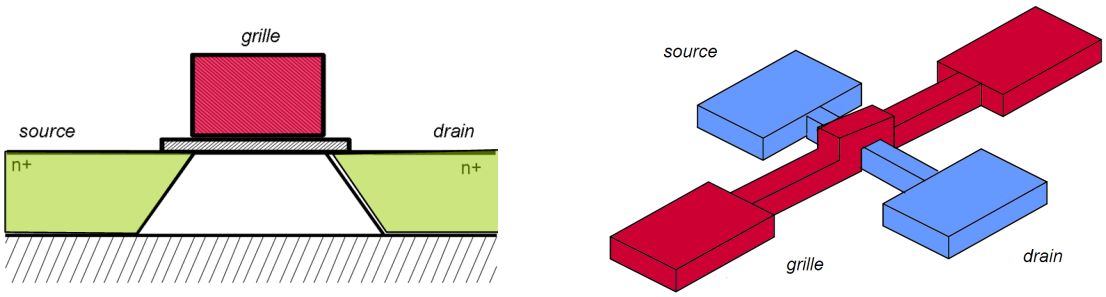

Figure 1.6 – Exemple d'un transistor sur substrat SOI (gauche) et FinFET (droite)

On retrouve deux solutions technologiques permettant un contrôle plus précis du canal. La première fait appel à des transistors multi-grilles, obtenus en entourant le canal de polysilicium. Ainsi, il existe des transistors double, voir triple grille, en fonction du nombre de cotés couverts du canal. Selon Kim et Roy [44], l'utilisation d'une double grille permet d'améliorer la pente sous-le-seuil du transistor et, par conséquent, sa capacité de fonctionnement à très faible tension. On observe qu'un courant dynamique plus important est obtenu.

L'autre solution est en principe identique mais dispose d'une architecture différente : le FinFET [45], qui consiste à entourer un silicium très fin auquel une source et un drain ont été aboutés. On obtient une amélioration de la sensibilité aux variations, et une réduction de l'énergie et de la tension de fonctionnement minimale.

Enfin, le HFETT, ou transistor à hétérojonction, propose un composant dont les sources et drains ont un dopage opposé. Le principe consiste à faire passer le courant par effet tunnel. L'avantage de ce transistor est de permettre un meilleur contrôle de la pente sous-le-seuil en modifiant la position du niveau de Fermi par le dopage utilisé [46].

Ces technologies ne sont pas supportées par STMicroelectronics.

Il est possible d'appliquer des options spécifiques sur les technologies usuelles. Elles consistent en la modification du dopage du canal des transistors, modulant ainsi la tension de seuil. On distingue alors :

- **le SVT :** Standard VT, correspondant à la technologie standard.
- **le HVT :** High-VT, correspondant à une tension de seuil plus élevée.

- **le LVT :** Low-VT, correspondant à une tension de seuil plus faible.

Le HVT permet de favoriser la diminution de la consommation statique au dépend d'une vitesse réduite. Le LVT permet inversement d'augmenter la vitesse au détriment de la consommation statique. Ces solutions permettent d'améliorer le comportement des transistors à très faible tension [47], mais limitent leur gamme de fonctionnalité en tension : le LVT est coûteux en énergie à tension nominale, tandis que le HVT est coûteux en vitesse à faible tension.

7.2 Les solutions de conception

Il existe de multiples solutions dites de conception permettant d'adapter les composants actuels à la très faible tension. Ces solutions ont pour objectifs de minimiser l'énergie, de maximiser les performances, et enfin de réduire la sensibilité des transistors aux variations.

La logique

L'avènement du CMOS fut le moteur de l'évolution des technologies actuelles. Il remplace une logique NMOS qui faisait appel à des résistances très couteuses en surface et en consommation. Toutefois, de nouvelles logiques affichent des résultats prometteurs pour la très faible tension.

On retrouve ainsi la logique Pseudo-NMOS, reprenant le concept de la logique NMOS [48]. La principale différence se situe dans l'absence de résistance, remplacée par un PMOS dont la grille est connectée à la masse. L'absence du réseau PMOS que contient la logique CMOS permet une importante réduction de la capacité de la cellule à commander. Bien qu'apportant de meilleures performances en vitesse pour une tension donnée, cette logique accuse également d'une consommation énergétique plus élevée.

La logique dynamique dite Domino [49] fait appel à des transistors de précharge et d'évaluation, commandés par une horloge. Ces transistors préparent le chemin au changement d'état. Cette logique permet d'augmenter la vitesse et d'obtenir une amélioration de l'efficacité énergétique, qui reste dépendante du taux d'activité en raison d'une consommation statique dégradée.

D'autres logiques permettent de réduire la variabilité des transistors [50]. Ces logiques sont nommées logiques variables et dynamiques. La logique variable VT-CMOS (Figure 1.7) fait appel à un circuit supplémentaire de stabilisation. Ce circuit relève les courants de fuites du

transistor, et communique les variations au reste du circuit de stabilisation, qui se charge d'appliquer une tension au substrat afin de les compenser. Cette logique nécessite ainsi d'une pompe de charge afin de fournir, ou récupérer, la tension nécessaire. Bien qu'efficace dans la résolution de la sensibilité aux variations, ce circuit fait appel à des systèmes complexes au coût en surface élevé.

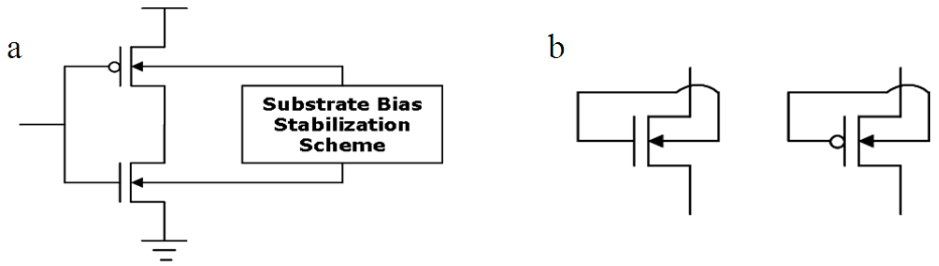

Figure 1.7 – Logique (a) VT-CMOS et (b) DT-CMOS

Le concept de la logique dynamique DT-CMOS est quant à lui basé sur une connexion entre le substrat et la grille du transistor. Cela améliore la transition d'un état à un autre et modifie la tension de seuil de façon dynamique. La sensibilité aux variations est alors réduite. Cette logique est coûteuse en surface car il est nécessaire d'isoler les transistors par l'insertion de caissons, ce qui la limite aux technologies SOI. Son fonctionnement est également limité à la très faible tension.

On distingue également la logique MCML[51] pour Mos Current Mode Logic en langue anglaise, où le fonctionnent en courant la distingue du fonctionnement en tension usuel. Cette logique proche de la logique NMOS permet une réduction de l'énergie par transition.

Enfin, on retrouve également l'utilisation, dans un même circuit, de transistors avec différentes tensions de seuil [52], ou encore différentes épaisseurs d'isolant. Sur les chemins critiques, où la vitesse est essentielle, on place des transistors plus rapides, avec une épaisseur d'isolant réduite.

Les cellules standards

La réalisation de circuits se fait traditionnellement à partir de blocs logiques élémentaires, tels qu'un inverseur, une porte logique ET ou une porte logique OU. Ces blocs, appelés communément cellules standards et regroupés dans une librairie, sont réalisés avec une hauteur de cellule fixe et une largeur qui est un multiple d'une valeur unitaire. Les cellules d'une même librairie peuvent ainsi être aboutées entre elles dans une configuration de voisinage arbitraire, et être placées automatiquement en minimisant les pertes de surface par un outil de concep-

tion de circuit.

Certaines recherches ont conduit à la modification de ces blocs élémentaires, donnant pour résultat de nouveaux types d'inverseurs ou de portes logiques. Les principales modifications effectuées concernent l'agencement des transistors, afin de privilégier la réduction de la consommation, ou l'augmentation de la vitesse de transition. En effet, la mise en parallèle permet d'augmenter le courant dynamique, tandis que la mise en série réduit le courant statique [53]. Certaines architectures favorisent ainsi le parallélisme sur les chemins critiques, et la mise en série sur les autres.

Le dimensionnement de ces cellules permet d'optimiser leur performance à très faible tension. Selon l'expression du courant (Équation 1.4), la réduction des dimensions de la source et du drain des transistors permet la réduction de la consommation énergétique [54]. Certaines stratégies de conception utilisent des transistors de petites tailles sur les chemins non critiques ayant de faibles besoins en délai. C'est cependant au détriment de la variabilité qui, comme étudié précédemment, augmente lorsque les dimensions sont réduites. En revanche, l'augmentation de L_{eff} conduit à une diminution de la consommation et une augmentation du délai [24; 55]. L'utilisation d'une longueur de grille plus importante permet de réduire la sensibilité aux variations du transistor.

Il existe une méthodologie permettant d'extraire le dimensionnement optimal d'une porte logique, appelé effort logique [56]. Elle tient compte du rapport de courant entre un NMOS et un PMOS et de l'architecture des réseaux. L'utilisation de cette méthodologie permet de définir un dimensionnement optimal permettant à une porte logique d'en commander quatre identiques avec un temps de traversée unitaire. Cette méthode, calibrée pour le fonctionnement à tension nominale, nécessite d'être ajustée lors de la réduction de la tension d'alimentation. La modification du comportement suite à la mise en parallèle ou la mise en série de transistors lors du fonctionnement à très faible tension se traduit ainsi par de nouvelles règles de dimensionnement [57].

La profondeur logique a également un impact sur la consommation énergétique d'un circuit [58]. Lorsqu'une porte logique commence à commuter, les portes qu'elle commande démarrent leur commutation. Ainsi, il existe une profondeur logique optimale correspondant à un couple [temps de commutation – consommation énergétique] optimal. Dépendante du type de cellules utilisé, la profondeur logique optimale est différente pour chaque chemin logique.

Enfin, les cellules standards ayant plusieurs entrées, telles que les

portes ET ou les portes OU, ont une consommation et un délai différent en fonction de la configuration des entrées [59]. Ce constat permet de cibler les transistors responsables des fuites ou de la dégradation des performances, ouvrant la voie à de nouvelles optimisations.

Les éléments séquentiels

Les éléments séquentiels ou à rétention permettent la mémorisation ou le séquencement des données. Les bascules, ou Flip-Flops en langue anglaise, ainsi que les mémoires, composent ces éléments. Il est observé [60] que leur fonctionnement à très faible tension est plus complexe, en raison d'une sensibilité aux variations plus importante.

Les flip-Flops sont usuellement composées d'un couple de bascule maître et esclave. Chaque bascule contient une boucle de mémorisation. La bascule maître commande, en fonction de l'horloge, la recopie de la donnée en entrée vers la bascule esclave. La bascule esclave communique la donnée précédente tant que la recopie n'est pas réalisée. Les boucles doivent ainsi permettre le maintient d'une donnée et son remplacement. Ce conflit entre stabilité et instabilité est amplifié à très faible tension.

Une solution [61] fait appel à un dimensionnement irrégulier des bascules pour limiter ce conflit. Une seconde architecture [62] introduit un étage supplémentaire permettant un contrôle dynamique du bistable. Différentes études comparatives [63; 64] confirment que la fonctionnalité logique des bascules à très faible tension, sans modifications et simulées dans des conditions d'utilisation usuelles, n'est pas remise en cause. Les optimisations sont réalisées dans le but de limiter l'impact de la variabilité et d'améliorer les performances.

Les mémoires statiques SRAM, dont la description détaillée est faite au Chapitre 4, sont couramment composées de deux inverseurs rebouclés, formant la boucle de mémorisation, et deux transistors d'accès (Figure 1.8). Ces transistors d'accès sont passants lorsque l'on souhaite effectuer une lecture ou une écriture. Les mémoires étant disposées en colonnes et en lignes, chaque cellule est exposée aux fuites des cellules voisines. Ainsi, lors d'une opération de lecture, la cellule accédée oppose son courant dynamique à ces fuites. A très faible tension, en raison de la réduction du rapport I_{ON}/I_{OFF}, la valeur lue peut être erronée.

La réduction de la tension d'alimentation affaiblie également la stabilité de la boucle de mémorisation. La chute de tension provoquée par l'ouverture des transistors d'accès peut faire basculer son contenu. Le dimensionnement est insuffisant pour compenser ces lacunes : le rap-

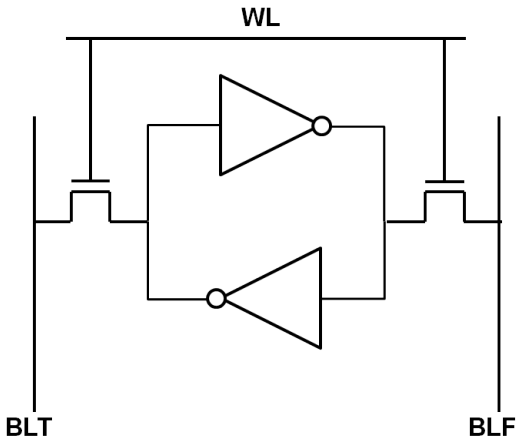

Figure 1.8 – Cellule SRAM à 6 transistors.

port entre la dimension des transistors d'accès et celle des inverseurs a un effet opposé sur les opérations de lecture et d'écriture : améliorer la lecture conduit donc à détériorer l'écriture.

Une des solutions observées pour rétablir la fonctionnalité à très faible tension est la séparation des accès de lecture et d'écriture [65]. En procédant ainsi, le dimensionnement peut être dissocié. Cette solution nécessite deux commandes additionnelles et augmente la complexité de la périphérie chargée de commander la SRAM. De nombreux travaux [66; 67] sont basés sur cette cellule en y ajoutant d'autres optimisations tels que l'insertion d'une pompe de charge pour augmenter la tension de certains noeuds, ou l'ajout de transistors additionnels permettant la réduction des fuites [68].
Il existe également d'autres architectures composées d'un nombre de transistors plus important[69; 70; 71]. L'objectif de ces réalisations est de protéger l'information contenue dans la cellule lors des phases de lecture et écriture.

La très faible tension, par les difficultés citées précédemment et la sensibilité aux variations, augmente la probabilité d'apparition d'une erreur dans une matrice mémoire pouvant contenir plusieurs millions de SRAM. Il existe des circuits en périphérie de la matrice mémoire permettant de détecter les erreurs et de les corriger [72]. Le principe de la redondance, qui consiste en la duplication des cellules jugées faillibles afin de les substituer en cas de dysfonctionnement [73], est également utilisé.

Les cellules d'interface

Les circuits à très faible tension peuvent être amenés à communiquer avec d'autres circuits alimentés à une tension différente. Les entrées et sorties, généralement composées de périphériques de contrôle ou de commande fonctionnant à une tension plus élevée, sont alors composées d'un convertisseur de tension, appelé "Level Shifter" en langue anglaise. On observe que de simples buffers, composés de deux inverseurs successifs, sont généralement utilisés en entrée des blocs à faible tension. Ces buffers, alimentés à la tension basse, sont commandés par un signal à tension nominale :
le NMOS du premier étage d'inverseur est bloqué lorsque l'entrée est à 0, et fortement passant lorsque l'entrée est à la tension nominale.
Le PMOS du premier étage d'inverseur est faiblement passant lorsque l'entrée est à 0, et fortement bloqué lorsque l'entrée est à la tension nominale.
Le courant dynamique du NMOS étant plus important que le courant dynamique du PMOS, il apparait un déséquilibre entre le délai du front montant et du front descendant de la sortie, ce qui se traduit par une dégradation du rapport cyclique pour les horloges ou les signaux périodiques. On ne trouve cependant pas d'information ni de travaux réalisés sur ces cellules dans la littérature.

Les principaux travaux de recherche ont pour axe le développement de level shifters en sortie des circuits à très faible tension, nécessaires pour commander la périphérie alimentée à tension nominale. Le principe est simple : il faut traverser deux inverseurs, un premier alimenté avec la tension d'entrée, et un second alimenté avec la tension de sortie. Cependant, dans le cas d'une entrée à 0, le NMOS du second inverseur recevant une entrée VDD_{ULV} conduit la valeur 0 avec un courant dynamique faible. Le PMOS alimenté à VDD_{HAUT} recevant une entrée VDD_{ULV} conduit la valeur 1 avec un courant dynamique faible, en raison de la différence de potentiel $V_{GS} < 0$. La sortie Z est en court-circuit (Figure 1.9).

La solution commune utilisée pour réduire le V_{GS} du PMOS de tête est l'insertion d'un second PMOS en série avec commande croisée : l'entrée à faible tension ne commande donc pas directement le PMOS relié à la source. La commande croisée du nouveau PMOS de tête limite l'apparition d'un court-circuit avec le NMOS. Cependant, lorsque la tension d'entrée est très réduite en regard de la tension de sortie, les fuites à travers les PMOS s'imposent de nouveau devant le courant dynamique des NMOS. Certaines solutions proposent un affaiblissement des PMOS en augmentant la tension à leur grille par l'utilisation d'une diode [74], limitant la plage de tension acceptée en entrée aux faibles tensions. Des level shifters dynamiques [75] ont été conçus, faisant appel à un procédé de précharge commandé par une horloge. Le syn-

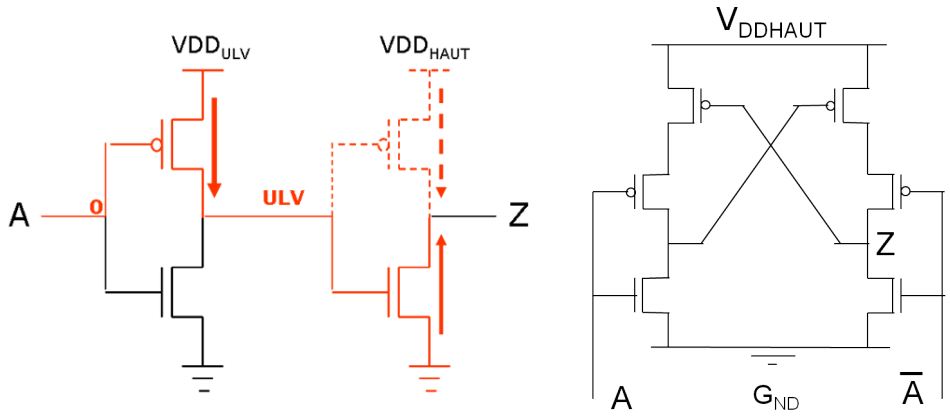

Figure 1.9 – Principe du Level Shifter et mise en évidence du court-circuit (gauche), et Level Shifter standard avec A l'entrée alimentée à la tension basse (droite).

chronisme de l'horloge nécessite toutefois un circuit additionnel.

7.3 Les solutions en tension

Il a été fait mention dans les solutions logiques de la possibilité de faire varier la tension dans le substrat pour contrôler la tension de seuil. Cette solution nommée "Body-Biasing", en langue anglaise, permet de compenser l'effet de la variabilité [76]. Si l'on prend l'exemple d'un NMOS, on relie généralement le substrat à la masse. Si l'on alimente négativement la prise substrat, la différence de potentielle V_{SB} devient positive, et la jonction source/substrat est en polarisation inverse : la zone bloquée s'étend dans le canal. La tension de seuil a virtuellement augmenté : il faut une tension plus élevée sur la grille pour un même courant de drain. Cette polarisation inverse, ou "Reverse Body Biasing" en langue anglaise, permet pour une même tension de grille de réduire les fuites à travers le canal au prix d'un courant dynamique plus faible. La polarisation directe, ou "Forward Body Biasing" en langue anglaise, permet d'augmenter le courant dynamique pour une même tension de grille, au prix d'une consommation statique plus élevée.

On observe plusieurs implémentations de cette solution pour la réduction de la consommation statique [77], ou l'adaptation dynamique des performances d'un circuit en fonction des besoins [18; 78]. Elle est également utilisée pour la compensation des variations [79] et notamment celles induites par la température [80], en addition d'un système de contrôle permettant la lecture de ces variations. L'application du body-biasing pour réduire la sensibilité aux variations à très faible tension est la solution la plus efficace à ce jour.

Pour un coût en surface réduit, le body-biasing doit être général : il s'applique sur un caisson N contenant un grand nombre de PMOS et sur le substrat contenant un grand nombre de NMOS. Pour une complexité réduite, il doit être statique : on définit une table de tensions de substrat, et on l'applique en sortie de fonderie de manière définitive. En dehors de ces deux cas, il est nécessaire d'implémenter des circuits de mesure permettant un ajustement dynamique de la tension pour chaque domaine d'alimentation.

D'autres approches dynamiques favorisent la modification de la tension d'alimentation plutôt que la tension du substrat. Cela nécessite un système de contrôle des variations accompagné d'un régulateur de tension [81]. On parle alors d'Adaptative Voltage Scaling (AVS) ou Dynamic Voltage Scaling (DVS) en langue anglaise. Généralement, la régulation de la tension d'alimentation est utilisée pour adapter la fréquence (DVFS) ou la consommation du circuit [82], permettant le passage d'un mode performance haute tension à un mode d'économie d'énergie basse tension [83], selon les besoins de l'instruction en cours d'exécution.

8 Les réalisations Ultra-Low Voltage

Plusieurs prototypes de circuits logiques complexes et de mémoires ont été développés et mesurés à très faible tension. Les paragraphes suivants présentent les réalisations les plus abouties.

8.1 Éléments logiques complexes

Un laboratoire américain est à l'origine d'un processeur capable de traiter l'information par la réalisation de transformées de Fourier [84] à une tension d'alimentation de 180mV. On trouve également des filtres numériques, tel qu'un filtre FIR [85], ainsi qu'un processeur RISC 32bit capable de fonctionner à très basse tension [86]. Il existe également des processeurs dédiés à la mise en réseau de capteurs [87]. Enfin, des unités de calcul plus ciblées ont été réalisées [88; 89; 90], et s'ajoutent à une gamme de réalisations en constante évolution.

Dernièrement, un micro-contrôleur avec mémoire intégrée, fonctionnant à 300mV a été présenté [91]. Il implémente une architecture 16bit RISC basée sur le jeu d'instructions du MSP430, un micro-contrôleur fabriqué par Texas Instrument. Il regroupe l'utilisation d'un ensemble de techniques, allant du dimensionnement à la conception d'une librairie de cellules dédiées, et se positionne comme la réalisation la plus abou-

tie à ce jour.

Ces circuits combinent également les solutions propres au low power, notamment la parallélisation. Enfin, l'utilisation du body-biasing est quasi-systématique dans l'ensemble de ces réalisations.

Il faut cependant noter que les tensions affichées par les diverses publications correspondent dans une grande partie des cas à la tension minimum de fonctionnalité et non à la tension V_{EMIN} d'énergie minimum. Le processeur FFT, bien que fonctionnant à 180mV, connait un minimum d'énergie à 350mV. Cette orientation dans la communication de la tension la plus faible pouvant être atteinte se confirme avec le MSP430 qui annonce une fonctionnalité à 0.3V, tandis que le convertisseur de tension embarqué génère une tension minimale de 0.5V.

On observe que la gamme de fréquence maximale de ces microprocesseurs est comprise entre 10kHz et 1MHz à très faible tension, avec une efficacité énergétique améliorée d'un facteur x10 comparé au fonctionnement nominal. A titre d'exemple, le processeur dédié aux réseaux de capteurs consomme 0.85 pJ par instruction, quand les processeurs fonctionnant à tension nominale en consomment plusieurs dizaines.

La majorité des processeurs cités sont réalisés dans des technologies de gravure de l'ordre de la centaine de nanomètres. Peu de publications présentent des résultats en 65nm ou encore 40nm. L'augmentation de la variabilité dans ces technologies rend la réalisation de dispositifs à très faible tension plus complexe.

Notons que dans la plupart des cas, ces processeurs sont accompagnés de mémoires fonctionnant à basse tension. On constate que ces mémoires limitent la réduction de la tension d'alimentation, leur tension minimale de fonctionnement étant plus élevée. Ainsi, les microprocesseurs ne fonctionnent pas à leur minimum d'énergie mais à celui imposé par la mémoire. Des sources de tensions différentes, en plaçant la logique et la mémoire dans des domaines de tension différents, permettent de s'affranchir de cette limitation au prix d'une plus grande complexité.

8.2 Les mémoires

Les mémoires sont indispensables au fonctionnement d'un microprocesseur puisqu'elles permettent la conservation des états et de l'information. L'exploitation de cellules mémoires se fait par une périphérie conçue pour réaliser les opérations de lecture et d'écriture. La majorité

des publications à très faible tension sur les mémoires SRAM proposent une amélioration de l'architecture de la cellule mémoire plutôt que de la périphérie. La communication est faite principalement sur la tension minimale de fonctionnement pouvant être atteinte, le nombre de transistors de la cellule, et la stabilité.

Les SRAM industrielles fonctionnant à tension nominale utilisent des cellules à six transistors (6T). Bien qu'il existe des propositions de SRAM 6T à très faible tension [92], la stabilité réduite de cette architecture se traduit généralement par la réalisation de cellules plus complexes [93], comptant jusqu'à douze transistors (12T). On notera que la capacité de ces mémoires reste limitée, la taille maximale observée étant de 480kb [46]. En moyenne seuls quelques kilobits composent les SRAM publiées. La technologie de gravure utilisée est en moyenne plus fine que celle des processeurs, entre 65nm et 130nm.

L'axe principale des recherches est l'amélioration de la stabilité face aux variations. Comme décrit précédemment, le rapport entre le courant dynamique et le courant de fuite régule la capacité d'exécution des opérations de lecture et d'écriture. Ce rapport étant dépendant du dimensionnement et de la tension de seuil, il est par conséquent sensible à la variabilité. Les principales solutions utilisées sont à la fois de nouvelles architectures plus stables par isolement de la boucle de mémorisation, l'utilisation du body biasing [94] ou encore l'utilisation du DVS [95]. Le body biasing permet d'augmenter efficacement la stabilité de la cellule mémoire [96].

Dans la structure traditionnelle 6T, réduire la tension favorise l'écriture plutôt que la lecture [69], en conséquence d'une plus forte dégradation du courant dynamique du PMOS. L'équilibre entre la lecture et l'écriture, déterminé par le dimensionnement des NMOS, PMOS et transistors d'accès, ne peut être obtenu qu'en adaptant la tension d'alimentation selon l'opération en cours [97]. Si le critère de surface est important en termes industriels, les structures usuelles ne sont pas adaptées à la très faible tension. L'utilisation de registres [98], c'est-à-dire de SRAMs avec lecture et écriture séparées permettant le dimensionnement optimal de chaque chemin dédié, apparait comme préférable.

L'architecture du point mémoire est donc le point clé permettant le fonctionnement à très faible tension. On note que la réalisation d'une SRAM basée sur une structure Schmitt Trigger [70], un type d'inverseur dont la stabilité du noeud de sortie est plus importante, permet un fonctionnement à 160mV. Concernant les performances en vitesse, certaines réalisations atteignent plusieurs centaines de kilohertz voir le mégahertz. La consommation de ces mémoires est réduite à quelques μA/MHz, contre quelques dizaines à tension nominale.

Les mémoires sont le sujet le plus délicat concernant la très faible tension, car contrairement à la logique où la variabilité est principalement la source de dégradation des performances, le point mémoire peut être affecté dans sa fonctionnalité, jusqu'à transmettre des données erronées.

9 L'axe de recherche de cette thèse

L'étude bibliographique nous éclaire sur la faible intégration industrielle de la très faible tension, malgré les nombreuses solutions proposées. Pour permettre le déploiement industriel, il est nécessaire de regrouper les critères suivants :

- La familiarisation de la très faible tension au sein de l'industrie.
- La mise en lumière du marché de la très faible tension.
- Le choix d'une plateforme technologique adaptée.
- Le développement de solutions dans le respect des possibilités industrielles en vigueur.
- Un coût de conception proche des coûts usuels.
- La validation des solutions sur silicium.
- Un rendement élevé qui traduit la fiabilité des solutions.

L'objectif de cette thèse est d'apporter une réponse à ces exigences. Pour cela, les règles suivantes ont été fixées :

- La logique utilisée est la logique CMOS.
- Les plateformes technologiques exotiques sont écartées.
- Seuls les outils et modèles en vigueur dans l'industrie sont utilisés.
- L'étude et le développement des solutions sont réalisés en suivant une méthodologie basée sur les résultats de simulation.
- La comptabilité industrielle des solutions sur l'ensemble du flot de conception, cellules/librairies/circuit/mémoire, est garantie.

Chapitre 2

Les Cellules Logiques

Ce chapitre, à la suite de la définition du contexte de simulation, met en évidence le comportement à très faible tension reporté lors de l'étude bibliographique. Ces observations conduisent à la construction d'une méthodologie de conception de cellules logiques adaptées et optimisées pour la très faible tension tout en limitant l'impact des performances à tension nominale. Des librairies de cellules ainsi optimisées sont conçues en respectant la compatibilité avec les outils industriels usuels. Un travail spécifique a été fourni sur les cellules sensibles pour répondre à l'augmentation de la variabilité. Enfin, ces cellules ont été fabriquées puis mesurées sur silicium permettant la validation de ces travaux.

1 Contexte

Le progrès de l'industrie du semiconducteur, dont la tendance est anticipée par l'ITRS, ou International Technology Roadmap for Semiconductors, s'articulait autour de deux critères clés : le coût et la vitesse. Le budget alloué à la consommation énergétique des circuits s'est ajouté suite au succès des applications mobiles, complété par la prise en compte de la variabilité en raison de son impact sur la prédictibilité des résultats. Chaque plateforme technologique, définie par la taille minimale des transistors et les procédés de fabrication, correspond à un quatuor [Coût - Vitesse - Consommation - Variabilité] spécifique.

1.1 Le cadre technologique

L'objectif fixé est la fonctionnalité à très faible tension au bénéfice d'une efficacité énergétique plus importante, et au prix d'une vitesse de fonctionnement moindre. Or, il a été montré dans l'étude bibliographique que la réduction des dimensions induit une augmentation de la

vitesse, de la consommation statique, et de la variabilité. Pour ces raisons, les travaux cités en bibliographie sont réalisés en grande majorité sur les plateformes 90nm et 65nm.

Cette thèse adresse en priorité les besoins du marché mobile médical. Ce marché nécessite une technologie mature afin de réduire les risques de dysfonctionnements, tout en permettant la réalisation de dispositifs à dimensions réduites pour garantir la mobilité et diminuer les coûts de fabrication. L'utilisation de la très faible tension remplie l'exigence de l'autonomie indépendamment de la plateforme technologique. Ainsi, Les technologies 90nm et 65nm répondent au critère de maturité, le 40nm et le 32nm à celui de la surface. Si les technologies plus fines permettent d'augmenter la vitesse de fonctionnement des dispositifs à très faible tension, elles conduisent à une augmentation de la variabilité.

L'intérêt de chacune des plateformes étant justifié, le choix se fera également en tenant compte des demandes client et des opportunités de fabrication sur silicium. La technologie 40nm étant en cours de maturation, elle requiert de nombreuses phases de tests permettant de confronter les résultats simulés aux résultats réels. Cette technologie a été sélectionnée dans un premier temps. Dans un second temps et à la suite d'une demande client, une partie des travaux a été menée sur la plateforme 65nm.

Une observation de l'impact de la faible tension et la portabilité des résultats sur les autres plateformes ont permis de compléter l'expertise.

1.2 Les cellules standards

Les librairies utilisées pour la conception de circuits sont composées de cellules dites standards. Elles possèdent les attributs suivants :

- Une hauteur fixe
- Un pas fixe en largeur, correspondant au pas de placement des cellules.
- Un pitch de routage, correspondant au routage des métaux.
- Une position fixe des rails de métaux dédiés à l'alimentation.

Le pas de placement est généralement défini par la distance minimale entre deux grilles de transistors. Ces attributs permettent l'aboutement arbitraire de l'ensemble des cellules d'une même librairie ou de librairies partageant les mêmes caractéristiques, principe nécessaire à la conception automatisée.

Des effets de dispersion sont observés lors de l'étape de lithographie utilisée pour l'impression des motifs de grille dans les technologies 40nm et 32nm. Pour garantir une impression fidèle des grilles, les motifs sont répétés à chaque pas de placement Figure 2.1.

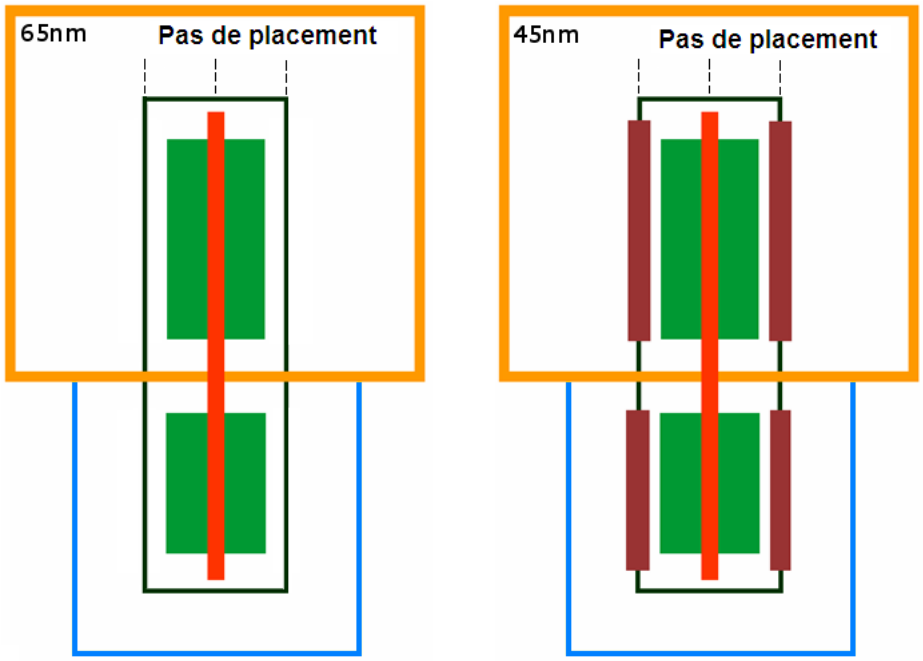

Figure 2.1 – Représentation de la répétition des motifs de grille en technologie 40nm. Ces motifs sont flottants.

Comme évoqué lors de l'étude bibliographique, les cellules standards remplissent les fonctions de base tels que l'inverseur, le ET, ou le OU. Les librairies industrielles comptent plusieurs centaines de cellules standards, qui sont dupliquées avec un dimensionnement différent pour répondre à plusieurs situations de charge de sortie. A titre d'exemple, on retrouve plusieurs portes logiques ET à deux entrées, dimensionnées différemment, chacune étant utilisée en fonction de la taille de la charge à contrôler.

1.3 Les modèles SPICE

Configuration de la simulation

Les simulations sont réalisées avec l'outil Eldo en utilisant des modèles de transistors SPICE, dans les configurations de tensions, de températures, et de procédés de notre choix. Les paramètres liés aux procédés émulent la qualité de l'impression du circuit sur silicium, et permettent d'en reproduire les variations. Usuellement, l'industrie encadre ces variations en définissant trois configurations différentes (Figure 2.2) :

- **TT :** Typical N, Typical P, pour le courant dynamique nominal.
- **FFA :** Fast N Fast P, pour un courant dynamique plus élevé.
- **SSA :** Slow N Slow P, pour un courant dynamique plus faible.

Ces configurations reproduisent la distribution du courant dynamique du transistor. La condition TT correspond au centre de la distribution, SSA et FFA correspondant à TT +/- 3σ. Il existe également des configurations intermédiaires permettant d'augmenter la couverture des cas possibles en dissociant les courants dynamiques des PMOS et NMOS :

- **FNSP :** Fast N Slow P.
- **SNFP :** Slow N Fast P.

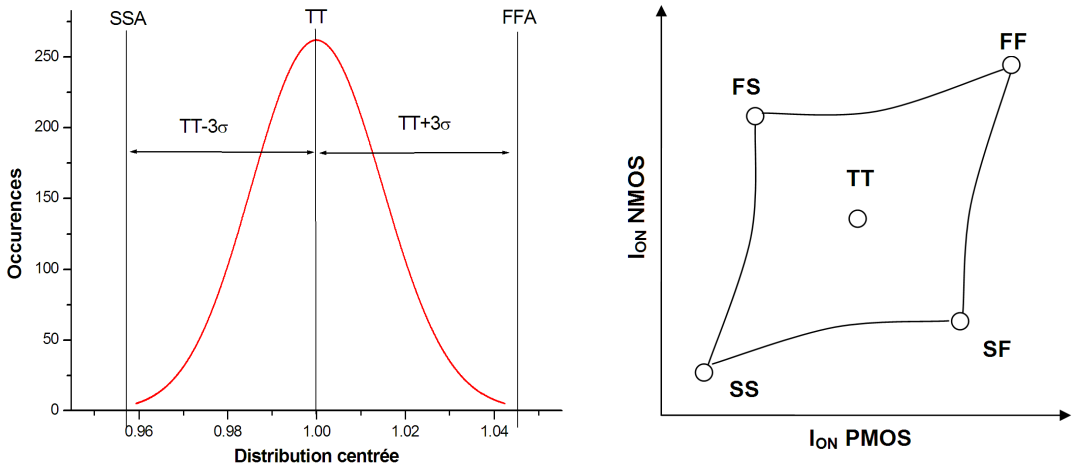

Figure 2.2 – Distribution des variations autour de la moyenne et espace des conditions du transistor.

A tension nominale, les conditions usuelles de simulation en température sont comprises entre -40°C et 125°C. A très faible tension, le courant dynamique diminue à mesure que la température diminue, comportement révélé lors de l'étude bibliographique et vérifié par simulation sur la Figure 2.3. Les applications visées ont cependant des contraintes en température moins étendues, la plage de fonctionnement étant de 0°C à 50°C.

Les configurations en température, tension d'alimentation, et procédés de fabrication sont regroupées dans des points de caractérisation. Quatre points de caractérisation ont été créés pour la conception à très faible tension :

- **Worst Case (WC) =** Pire cas en courant : SSA, 0°C.

- **Typical Case (TC) =** Point nominal : TT, 25°C et 37°C (température corporelle).
- **Best Case (BC) =** Meilleur cas en courant : FFA, 50°C.

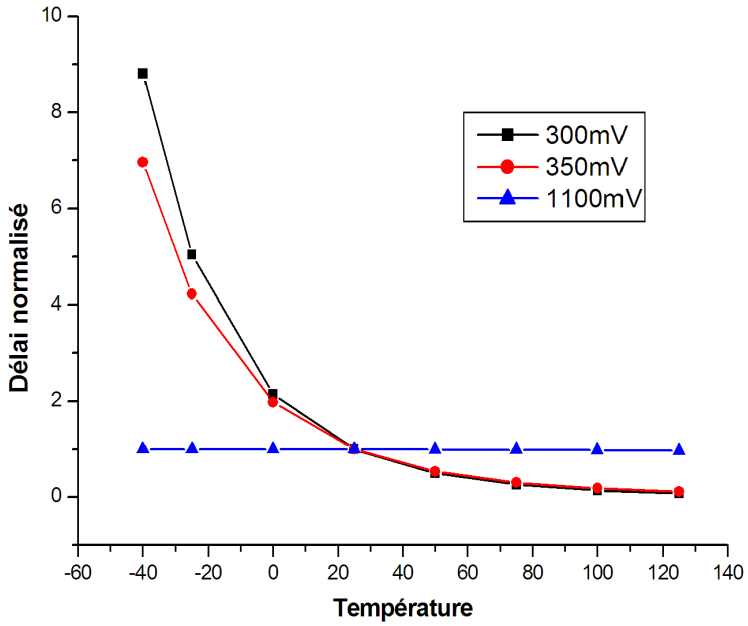

Figure 2.3 – Simulation du délai en fonction de la température.

Précision des modèles à très faible tension

L'étude par la simulation est dépendante de la précision des modèles utilisés. Cette précision peut être définie par deux points principaux :

- **Le centrage :** Écart de la valeur moyenne du courant entre sa mesure sur silicium et le résultat de la simulation.
- **La variation :** Évolution du courant autour de sa valeur moyenne.

Le centrage défini la prédiction des résultats, tandis que la variation définie l'évolution du courant d'un transistor face à la modification de ses caractéristiques. Pour étudier ces deux points, des mesures sur silicium ont été réalisées sur des transistors élémentaires aux dimensions variées.
En technologie 40nm, on observe un centrage pessimiste avec une réduction de 58% du courant dynamique à 0°C, amplifiée lorsque l'on extrait les capacités et résistances pour atteindre 85% (Figure 2.4).

L'évolution du courant dynamique est observée en fonction de trois paramètres différents : la température, la taille des sources/drains, et la longueur de grille. On note sur la Figure 2.4 que l'effet des résistances et capacités dégrade la précision de la variation du courant dynamique en fonction de la température. A température fixe, elles permettent toutefois d'améliorer la modélisation de la variation du courant en fonction des dimensions du transistor (Figure 2.5). On observe par ailleurs que l'évolution du courant dynamique en fonction de la dimension des sources/drains est correctement modélisée. Concernant la longueur de la grille du transistor, le modèle est optimiste.

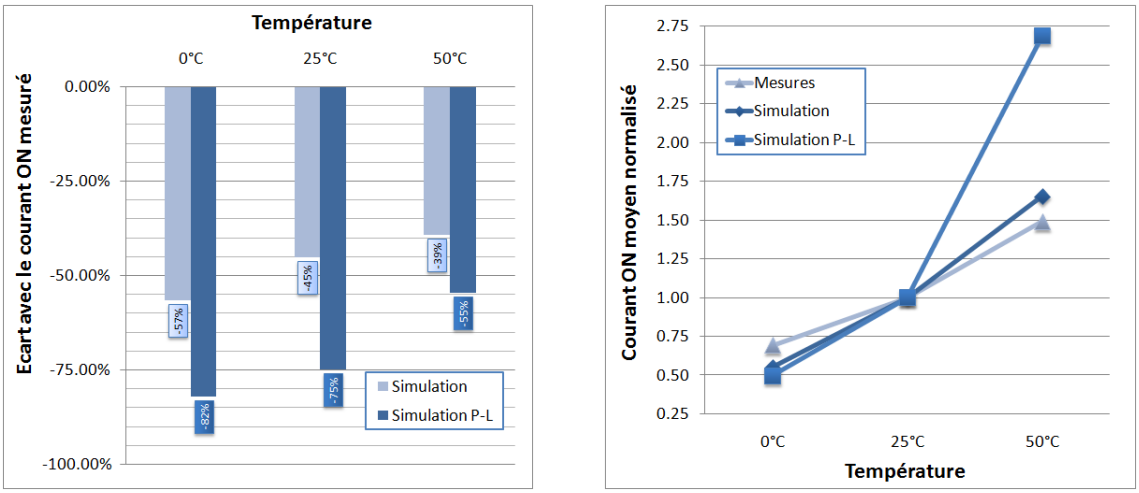

Figure 2.4 – Centrage et variation du courant dynamique en fonction de la température, mesurés à 350mV.

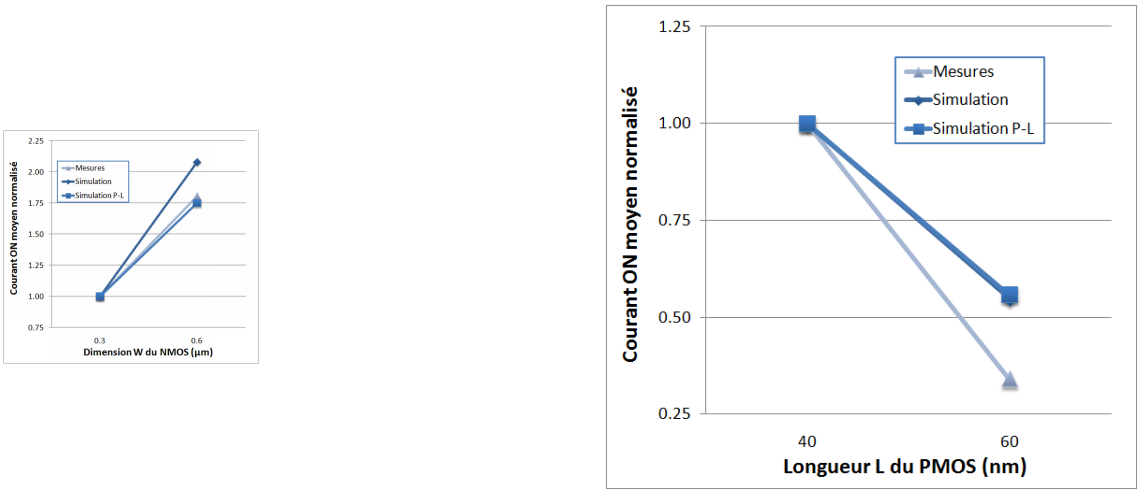

Figure 2.5 – Variation du courant dynamique en fonction de la dimension W des sources/drains et de la longueur de grille L. Mesurée à 350mV.

Conclusion sur les modèles

Malgré l'existence de différences, la modification des modèles n'est pas envisagée dans un premier temps. Cette procédure doit être réalisée et validée dans le cadre industriel, ce qui ne peut être déclenché qu'en réponse à une demande forte du marché.

Les choix de conception seront par conséquent adaptés en tenant compte des écarts précédemment observés :

- Le courant dynamique simulé est pessimiste.
- L'effet du dimensionnement des sources/drains est correctement modélisé.
- L'effet du dimensionnement de la grille est optimiste.

1.4 La rapport $I_{Dynamique}/I_{Statique}$

La conception de cellules logiques CMOS est basée sur l'existence d'un courant dynamique supérieur au courant statique, traduit par la fermeture et l'ouverture d'un réseau de transistor. A faible tension, ce rapport est réduit comme présenté sur la Figure 2.6. Cette réduction implique une limitation des possibilités d'opposition entre les réseaux NMOS et PMOS. A titre d'exemple, la porte logique ET présente un réseau de transistors PMOS en parallèle et un réseau de transistors NMOS en série. Le courant dynamique réduit du réseau NMOS doit rester supérieur au courant statique augmenté du réseau PMOS.

Le ratio $I_{Dynamique}/I_{Statique}$ entre le NMOS et le PMOS pour la technologie 40nm est divisé par un facteur x200 lorsque la tension d'alimentation est réduite de 1.1V à 0.35V. Ce nouveau rapport à très faible tension permet toutefois de maintenir le fonctionnement des cellules standards.

1.5 Définition de la méthodologie qFO4

L'étude par la simulation sur ces structures élémentaires présente une limitation en termes de reproduction des résultats. Les circuits numériques se composent de chemins réalisés à partir de portes logiques. Ces portes logiques présentent des réseaux de transistors PMOS et NMOS aux architectures variées. L'effet du chemin et des architectures de transistors ne sont pas visibles sur les structures de transistors élémentaires.

Des inverseurs mis en chaine sont généralement utilisés pour reproduire ces effets et émuler le chemin critique d'un circuit numérique

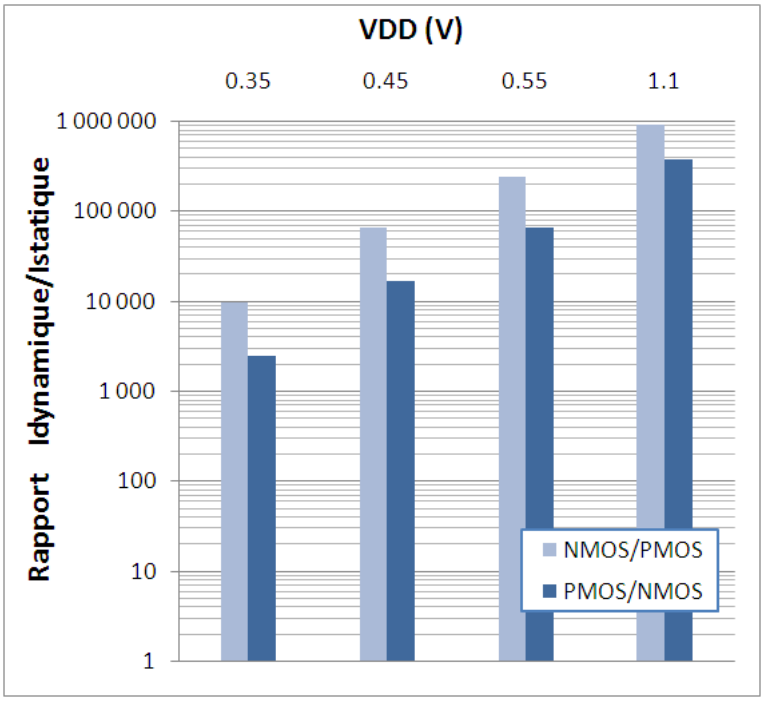

Figure 2.6 – Évolution du rapport $I_{Dynamique}/I_{Statique}$ entre le NMOS et le PMOS en fonction de la tension d'alimentation.

[26]. Pour obtenir des résultats prenant en compte les effets d'architecture des réseaux, un chemin logique spécifique a été conçu.

Le chemin est composé d'inverseurs, de portes logiques ET à deux et trois entrées, et de portes logiques OU à deux et trois entrées. Chaque porte logique charge l'équivalent de quatre fois sa propre capacité d'entrée, ce qui lui permet d'avoir un temps de traversé unitaire FO4, pour Fan Out of Four, en langue anglaise [56]. Les portes à plusieurs entrées ont toutes leurs entrées connectées entre elles, ce qui permet d'obtenir les extremums en délai et fuites comme présenté sur la Figure 2.7. Ainsi configuré, ce chemin logique permet d'obtenir une valeur moyenne de la caractéristique simulée en un cycle d'horloge. Par construction, le chemin logique élimine l'aspect arbitraire lié au dimensionnement par une normalisation naturelle de la charge de sortie de chaque étage, les modifications étant répercutées sur toutes les cellules.

Ce chemin composé de onze étages a été nommé qFO4, pour quasi-FO4. Idéalement, il doit être placé entre deux bascules pour révéler l'effet des contraintes temporelles, ce qui ne peut être réalisé car les bascules nécessitent un travail spécifique et couteux à très faible tension.

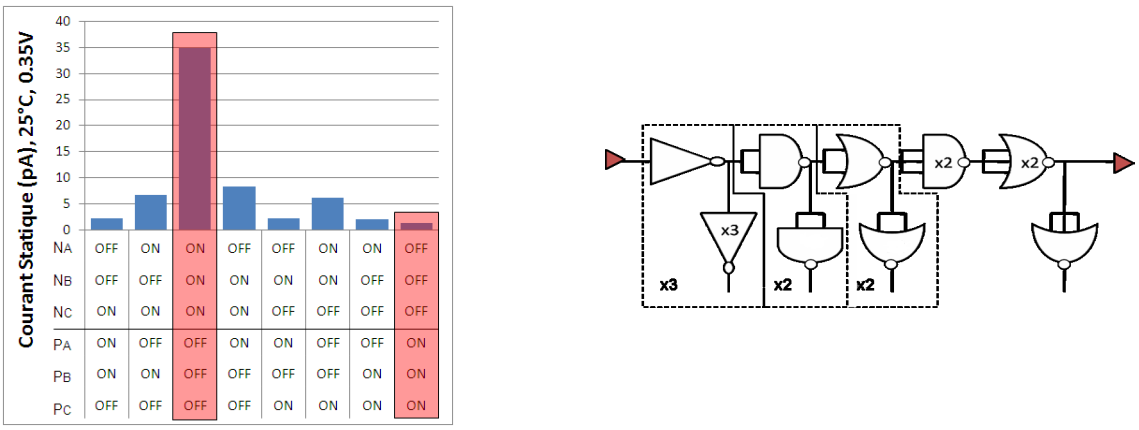

Figure 2.7 – Extraction des extremums en courant statique en fonction de la configuration des entrées (gauche) et chemin logique qFO4 (droite).

1.6 Définition de la méthodologie rFO4

Le qFO4 ne permettant pas d'observer l'effet du dimensionnement en raison de l'adaptation de la charge de sortie, une méthodologie a été définie pour ce type de mesure. Elle consiste à placer la cellule dont le paramètre varie dans un environnement figé. Pour cela, les charges d'entrée et de sortie sont réalisées en utilisant la cellule de référence. A titre d'exemple, un inverseur de dimension W, est placé dans un environnement composé de trois inverseurs de dimension W en entrée et quatre inverseurs de dimension W en sortie. Pour connaître l'effet de la variation de W, seule la dimension de l'inverseur simulé est modifiée, ceux qui composent les charges d'entrées et de sortie conservant le dimensionnement de référence.

Ce chemin est nommé rFO4, pour référence-FO4 (Figure 2.8).

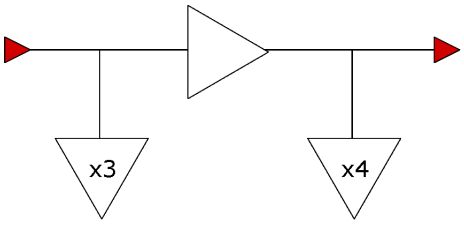

Figure 2.8 – Chemin logique rFO4 utilisant des charges fixées par la cellule de référence.

2 Simulation de la réduction de la tension d'alimentation

L'objectif premier est de confirmer les observations faites lors de l'étude bibliographique. En faisant appel aux différentes méthodologies définies précédemment, les métriques de la très faible tension seront extraites par l'intermédiaire de simulations SPICE. Ces observations sont réalisées au point de caractérisation TC, défini en page 32.

2.1 Le délai

Le délai correspond au temps de transition d'une porte logique. C'est un indicateur de la vitesse de commutation de la porte. L'étude bibliographique et l'expression théorique du délai révèle un comportement exponentiel à très faible tension :

$$t_{délai} = C_S \cdot \frac{V_{DS}}{I_{sous-le-seuil}} \tag{2.1}$$

avec l'Équation 1.4 on obtient :

$$t_{délai} \approx A \cdot \frac{V_{DD}}{e^{\left(\frac{Vgs-Vth}{m \cdot \nu_T}\right)} \cdot \left(1 - e^{-\frac{Vds}{\nu_T}}\right)} \tag{2.2}$$

Ce comportement est également observé par la simulation du qFO4 : le passage à la phase exponentielle est caractérisé par l'inflexion observée lorsque la tension est en deçà de 400mV (Figure 2.9).

2.2 La puissance de fuite

La puissance dissipée par le chemin logique qFO4 lorsqu'il n'y a pas d'activité est une puissance perdue. On observe que cette puissance peut être divisée par un facteur x10 lorsque l'on réduit la tension d'alimentation de 1,2V à 0,6V (Figure 2.10). A 200mV, correspondant au minimum fonctionnel simulé, les fuites sont divisées par un facteur x100.

2.3 L'énergie par transition

L'énergie par transition permet de mesurer l'efficacité énergétique d'une opération. On note que la réduction de la tension d'alimentation réduit la consommation par opération d'un facteur x6 sous la tension de seuil (Figure 2.11). On observe qu'elle atteint son minimum autour de

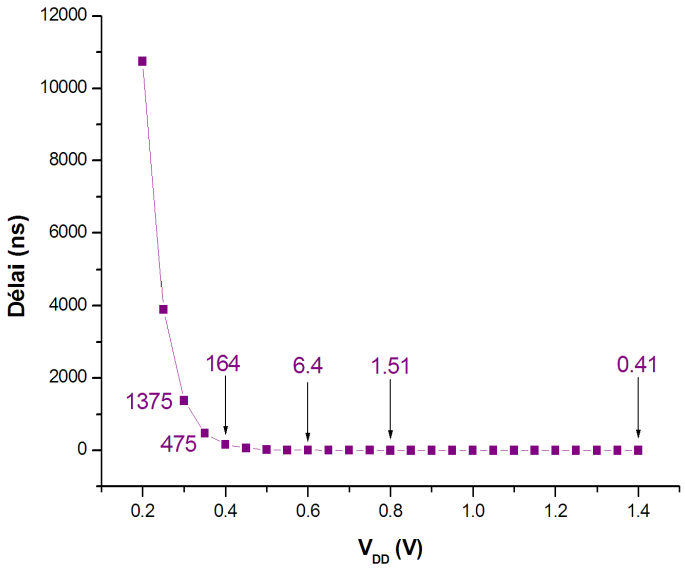

Figure 2.9 – Délai de la chaine qF04 en fonction de la tension d'alimentation.

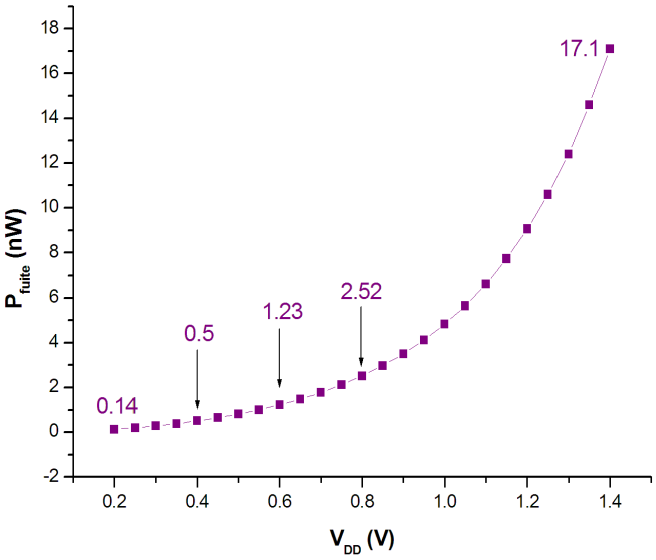

Figure 2.10 – Fuites de la chaine qF04 en fonction de la tension d'alimentation.

300mV, avec une division de la consommation énergétique d'un facteur x8.

2.4 Évolution en fonction du taux d'activité

L'énergie par transition traduit la consommation énergétique pour un taux d'activité de 100%. Afin d'observer l'évolution du minimum énergétique lorsque le taux d'activité se réduit, le chemin logique qFO4

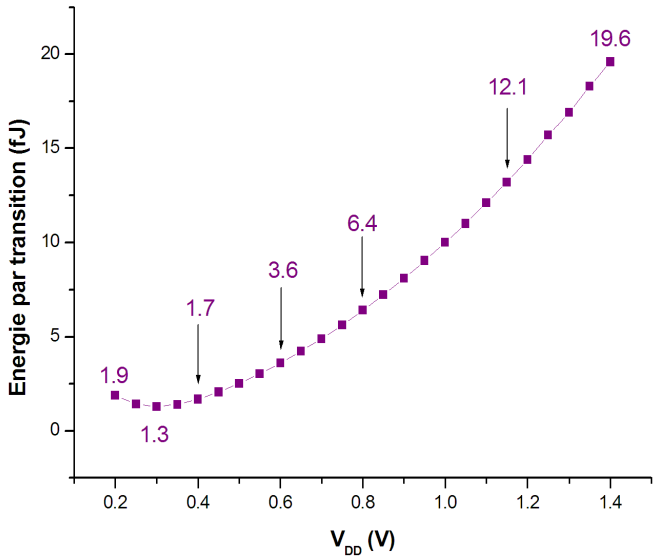

Figure 2.11 – Énergie par transition de la chaine qF04 en fonction de la tension d'alimentation.

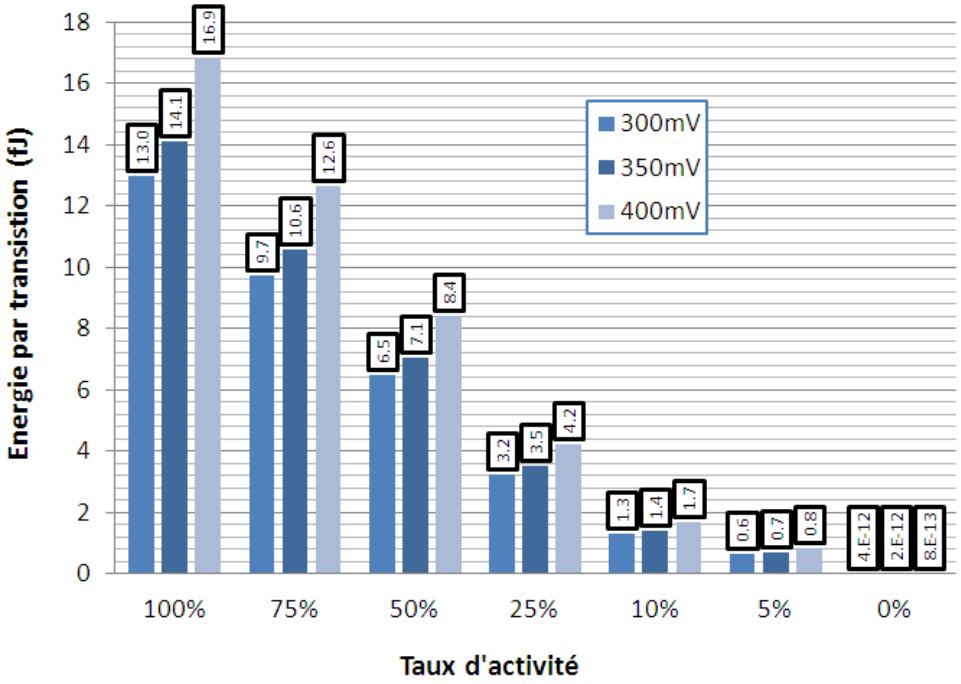

Figure 2.12 – Évolution de l'énergie par transition en fonction du taux d'activité, pour trois tensions différentes.

est simulé sur 10 cycles dont la proportion de phases actives et inactives varie. On constate sur la Figure 2.12 que la tension permettant d'obtenir le minimum d'énergie reste constante excepté à 0% où la tension d'énergie minimale est de 400mV, puisqu'elle correspond à l'énergie de fuite. Cependant les circuits présentent un taux d'activité toujours supérieur à 0%.

2.5 La variabilité

Des caractéristiques du transistor

Des simulations Monte-Carlo ont été réalisées pour exprimer l'effet d'une variation aléatoire des paramètres des transistors sur le délai de la qFO4. Le délai est observé car il est le garant de la commutation des cellules du chemin logique. Cette observation permet également d'évaluer la tolérance sur la prédiction des résultats post-silicium.

On observe une variation du délai de 11% entre le meilleur et le moins bon résultat à 1,0V (Figure 2.13). A 300mV, cette variation est de 100%.

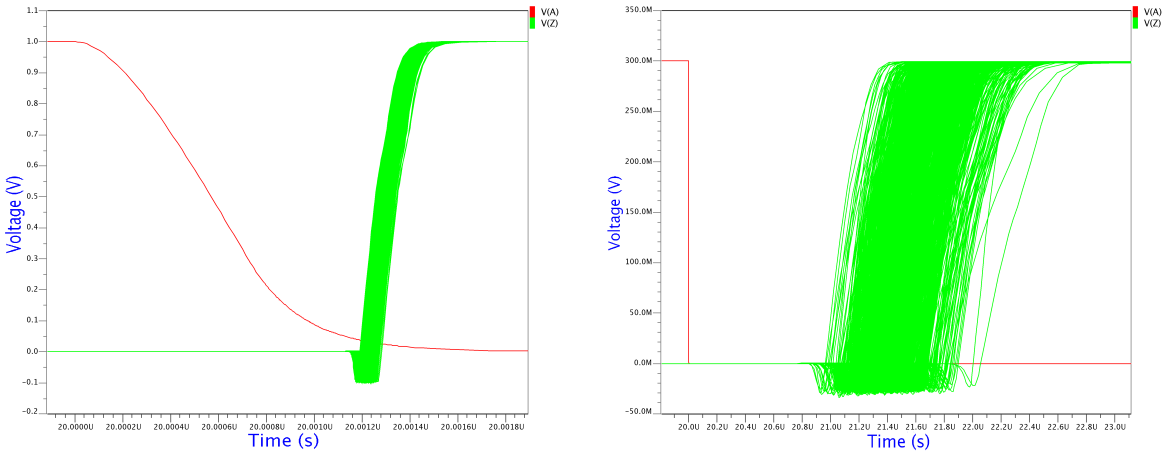

Figure 2.13 – Simulation Monte-carlo du délai sur le chemin logique qFO4 à 1.1V et 0.3V, à 25°

L'écart-type relatif σ/μ de la distribution du délai est multiplié par un facteur x5,5 entre ces deux tensions (Figure 2.14).

De la température

Lors de l'étude de la précision des modèles, on a observé l'amplification de l'impact de la température sur le délai à très faible tension. Le courant dynamique dans les conditions typiques connait une variation de -35% lorsque la température chute de 25°C, et +50% lorsque la température augmente de 25°C. On observe que le délai du chemin logique qFO4 (Figure 2.15) augmente de +115% à 0°C et diminue de 50% à 50°C, en comparaison de la référence à 25°C. Le délai à 0°C traduit la dégradation de la commutation le long du chemin logique.

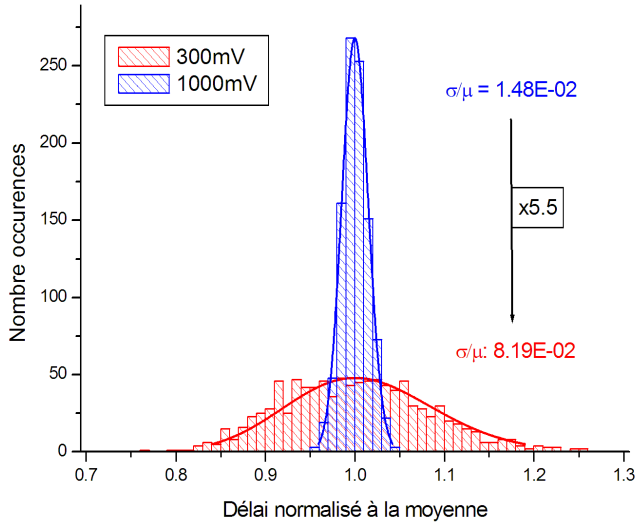

Figure 2.14 – Distribution du délai à tension nominale et à très faible tension, pour 1000 tirages Monte-Carlo.

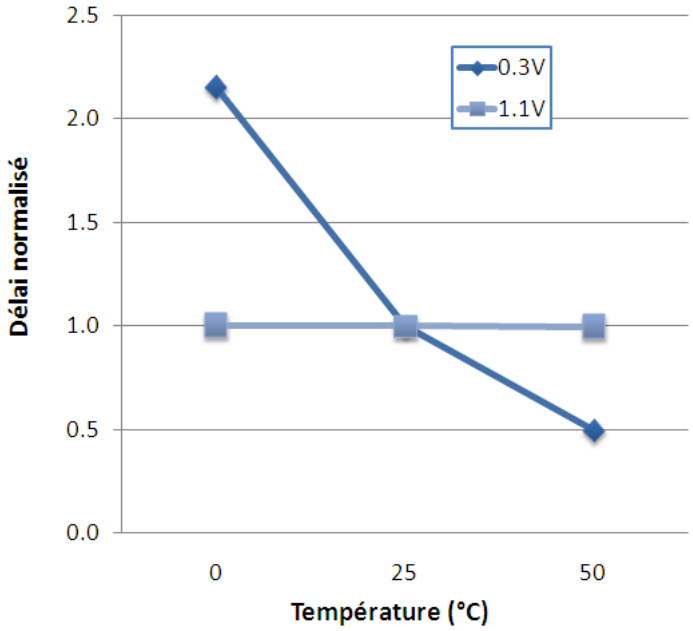

Figure 2.15 – Simulation du délai en fonction de la température du chemin logique qFO4.

2.6 Définition de la tension optimale

L'étude bibliographique a révélé trois notions de tension minimale :

- Le minimum de fonctionnalité, observé à 200mV.
- Le minimum d'énergie, observé à 300mV.
- Le minimum optimal.

La tension minimale optimale correspond à la mise en commun d'un ensemble d'objectifs. Dans le cadre de cette thèse, la réduction de la tension d'alimentation répond au besoin énergétique des applications nécessitant une fréquence de fonctionnement de l'ordre du mégahertz.

La tension d'alimentation optimale est un compromis entre l'énergie et le délai, visible sur la Figure 2.16. A V_{DD}=350mV, on observe un délai de l'ordre de 500ns, et une réduction de la consommation énergétique d'un facteur x8. A 400mV, on note un coût énergétique additionnel de 30%, tandis qu'à 300mV le délai est multiplié par un facteur x3. Par conséquent, la tension V_{OPT}=350mV correspond à notre tension optimale.

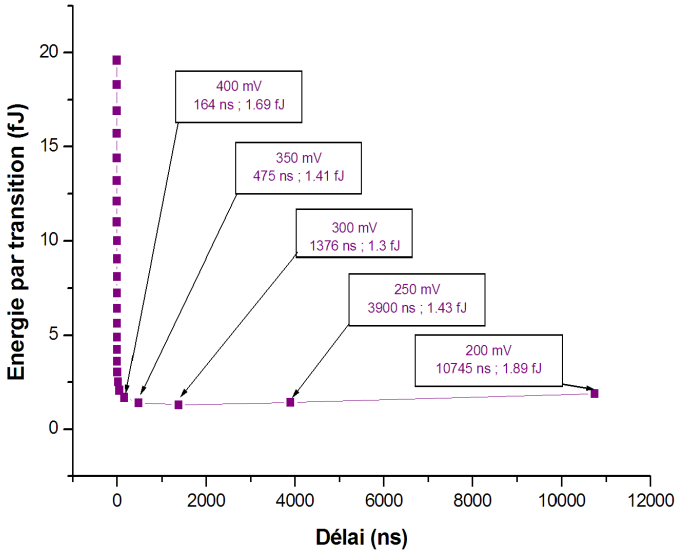

Figure 2.16 – Expression de l'énergie en fonction du délai pour la recherche du V_{OPT}.

2.7 Choix de l'option technologique

La plateforme 40nm propose les solutions technologiques LVT, SVT, et HVT. Le LVT permet d'atteindre des objectifs en vitesse en contrepartie d'une augmentation des fuites, tandis que le HVT permet de réduire ces fuites en contrepartie d'une vitesse moins élevée. Les librairies de cellules usuelles utilisent le SVT.

L'étude de l'effet de la tension de seuil sur le chemin logique qFO4 à faible tension révèle une augmentation d'un facteur x5,8 des fuites lors l'usage du LVT, permettant toutefois une augmentation nette de la vitesse (Figure 2.17). Le HVT augmente le délai d'un facteur x8,6, ce qui le rend inutilisable pour remplir l'objectif en vitesse d'un mégahertz. Finalement, le SVT propose le meilleur compromis énergie-vitesse. Cette

plateforme sera retenue pour l'étude et le développement des cellules, en notant toutefois la possibilité d'utiliser le LVT de manière ponctuelle pour combler d'éventuelles lacunes en vitesse, d'autant plus qu'elle permet une réduction de la sensibilité aux variations.

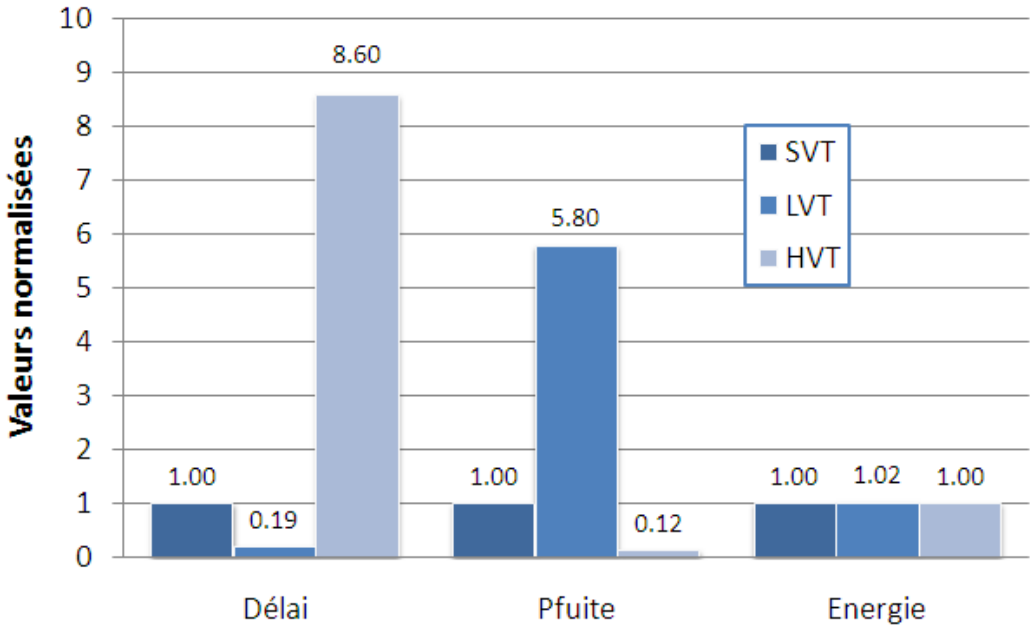

Figure 2.17 – Résultats de la simulation du qFO4 à 350mV pour plusieurs VT.

2.8 Directives architecturales

Les précédentes observations permettent de définir les directives architecturales suivantes pour le développement et l'optimisation de cellules logiques à très faible tension :

- La tension optimale est V_{OPT}=350mV.
- L'efficacité énergétique est améliorée d'un facteur x9.
- Le délai d'une porte logique en TC à 350mV est de 50ns.
- La variabilité du délai est 5,5 fois supérieure à V_{DD}=350mV.
- Le dopage offrant le meilleur compromis est le SVT.

Les solutions et optimisations appliquées aux cellules standards nominales doivent assurer le meilleur compromis entre l'énergie et le délai et maintenir ou améliorer la sensibilité aux variations.

3 Le dimensionnement des transistors

3.1 Les sources / drains

Le dimensionnement des sources et drains des transistors a été évoqué dans l'étude du courant dynamique effectuée en début de chapitre, dans le cadre de structures élémentaires. Le paramètre qui défini cette dimension est W. En utilisant la méthodologie rFO4, l'effet linéaire attendu de l'augmentation de W sur le délai est vérifié par la simulation (Figure 2.18).

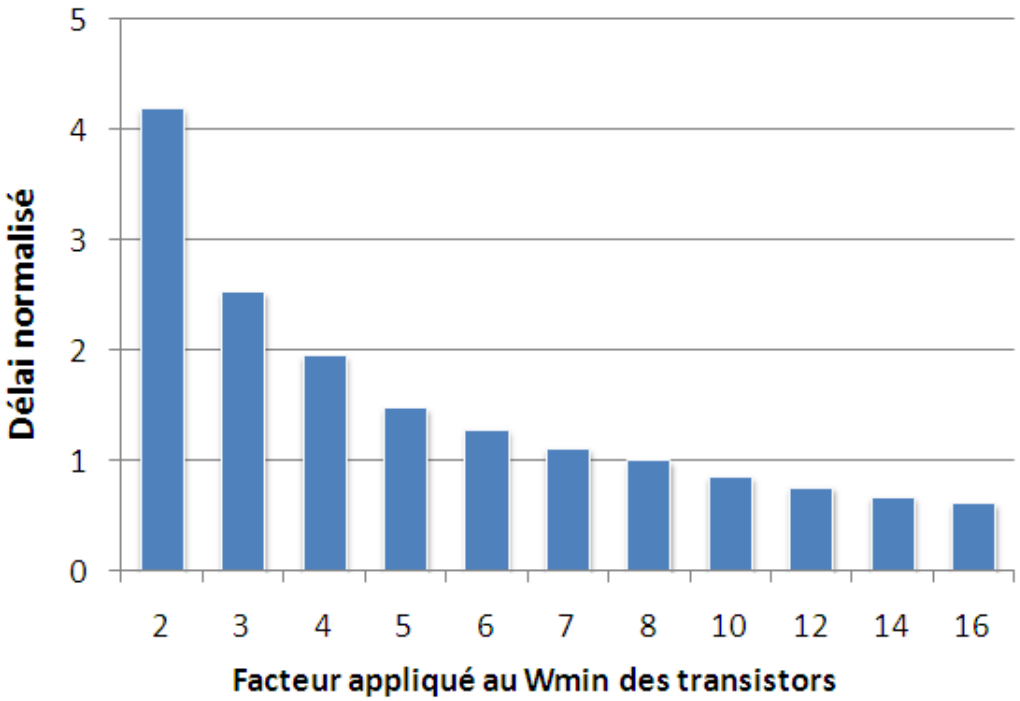

Figure 2.18 – Délai en fonction de W simulé sur rFO4.

3.2 Les grilles

L'effet du dimensionnement des grilles de transistors a déjà été évoqué précédemment. Cependant, il s'applique à l'ensemble des cellules qui composent la librairie. L'effet sur une cellule dépend donc de l'augmentation de la capacité d'entrée de l'étage suivant. Par conséquent, l'étude du dimensionnement de la longueur L des grilles est réalisée en utilisant la méthodologie qFO4.

L'augmentation de L en technologie 40nm implique une augmentation de la surface en raison des contraintes de conception évoquées précédemment. La surface du chemin logique qFO4 augmente de 36%

pour un L de 50nm, suite à l'ajout d'un pas de grille sur chaque cellule. En termes de performance, on observe sur la Figure 2.19 une augmentation du délai de 39% et une diminution des fuites de 60% pour L=50nm. L'énergie par transition est ainsi réduite de 6%. Le gain s'étend à 8% pour L=60nm. Étant donné l'optimisme des modèles, une limitation à L=50nm est définie pour modérer la dégradation du courant dynamique.

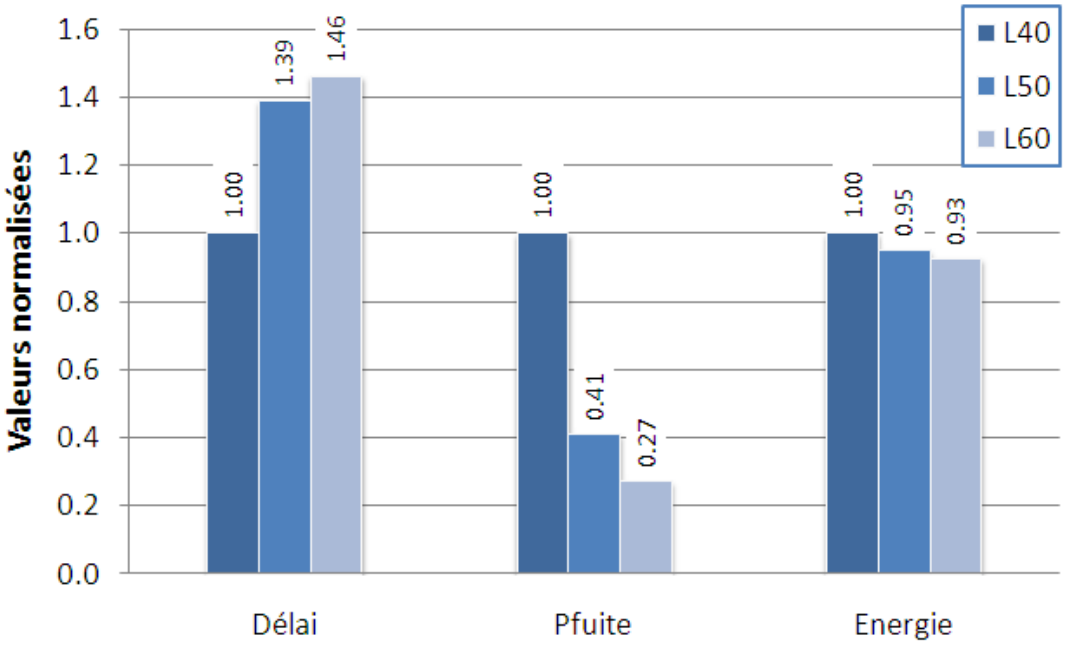

Figure 2.19 – Résultats de la simulation du qFO4 à 350mV pour plusieurs longueur de grille L.

3.3 Le β-ratio

La différence de mobilité entre les porteurs P et les porteurs N conduit à un courant différent des PMOS et NMOS pour un même W. L'équilibre du courant est atteint pour un dimensionnement du PMOS plus important que celui du NMOS. Le rapport entre W_{PMOS} et W_{NMOS} est le β-ratio, dont la valeur en technologie 40nm est de 1,4.

La modulation du β-ratio de l'ensemble des cellules du chemin logique qFO4 conduit à différents résultats en termes de délai et d'énergie par transition (Figure 2.20). Cette modulation est réalisée en conservant la taille totale $W_{T,MAX}=W_{N,MAX}+W_{P,MAX}$ dans le respect de la hauteur fixe des cellules standards. Le minimum d'énergie est observé pour un ratio de 1. Toutefois, le coût énergétique permettant de se situer à l'optimum de vitesse, obtenu pour un ratio de 1.25, est de 2%. Le β-ratio optimal représentant le meilleur compromis entre l'énergie et le

délai est β=1,25.

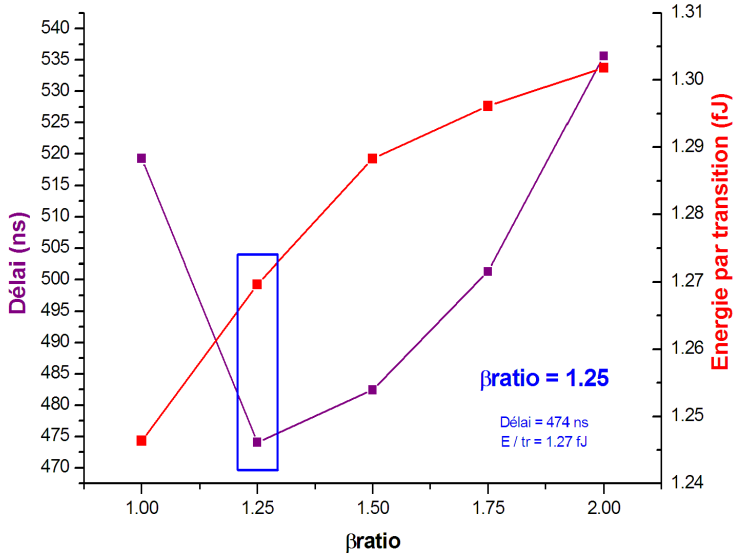

Figure 2.20 – Simulation de l'impact du β-ratio sur la fréquence et l'énergie, sur qFO4 à 350mV.

Comparativement au β-ratio standard utilisé à 1.1V en technologie 40nm, ce nouveau ratio affiche un gain en performance et en énergie par transition respectivement de 1.2% et 0.9%. Cette nouvelle valeur ne garantie pas un équilibre des temps de montée et de descente, mais le meilleur couple vitesse/énergie.

3.4 Les réseaux de transistors

La conception de cellules logiques telles que les portes ET et OU nécessite la mise en série ou parallèle d'un des réseaux de transistors pour la réalisation de l'opération logique (Figure 2.21).

Selon la théorie de l'effort logique, on observe que :

- le courant de deux transistors de taille W en parallèle est égal au courant d'un transistor de taille 2W
- le courant de deux transistors de taille W en série est égal au courant d'un transistor de taille W/2

Ces ratios sont vérifiés à 1,1V pour la mise en parallèle, et on observe une perte en délai de 27% pour la mise en série. Cette dégradation peut provenir d'une augmentation du V_{SB}, pour tension source-substrat, du transistor non relié à la source, entrainant une réduction

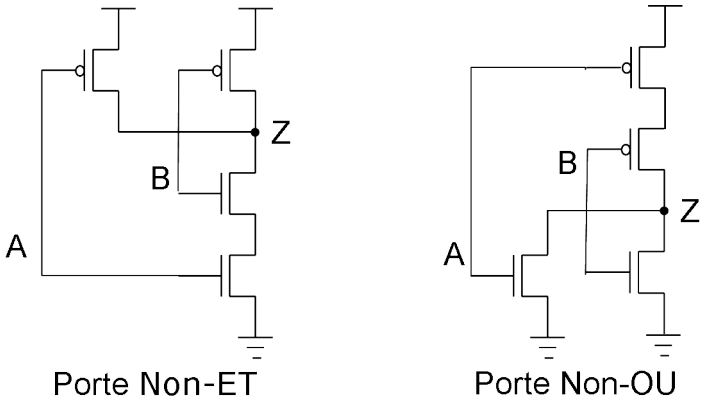

Figure 2.21 – Réseaux de transistors des portes logiques Non-ET et Non-OU.

du courant dynamique comme évoqué lors de l'étude bibliographique sur le body-biasing.

Ce facteur x2, selon l'étude bibliographique, évolue lorsque la tension d'alimentation diminue. La méthodologie rFO4 a été utilisée pour comparer l'effet du ratio sur ces trois types de cellules. L'objectif est de comparer le délai de chacune des valeurs de ratio pour charger la même capacité. On observe qu'en moyenne (Figure 2.22), un ratio de parallélisme R_P = 1,8 permet de maintenir un délai constant. Concernant la mise en série, le délai constant est estimé pour un ratio R_S=6. Cependant, ce ratio est ramené à R_S = 3 pour maintenir un coût en surface raisonnable. La perte en délai induite par ce ratio est de 47%.

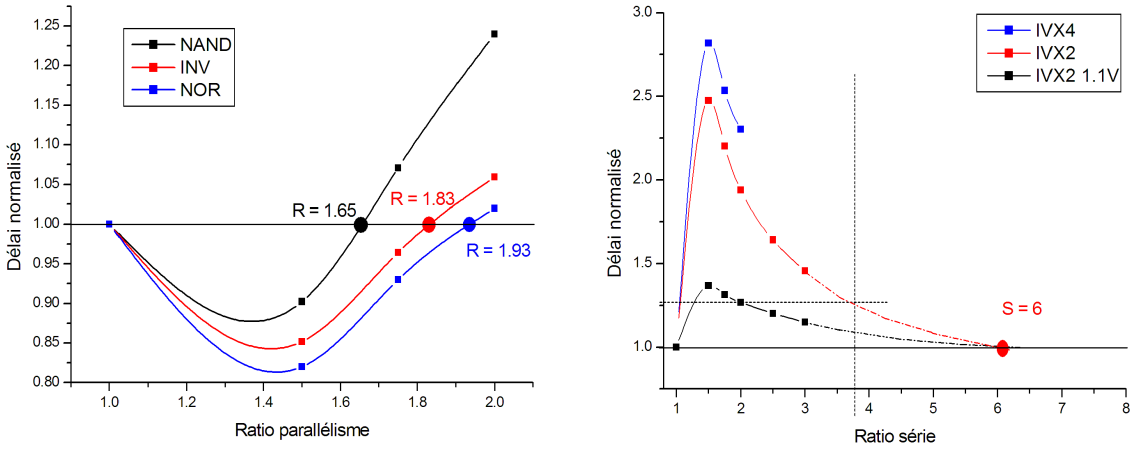

Figure 2.22 – Simulation du délai sur le chemin logique rFO4 à 350mV pour la mise en parallèle et la mise en série.

3.5 Récapitulatif et sélection

Le récapitulatif des précédentes observations est le suivant :
- L'évolution du courant en fonction de W est linéaire
- La longueur de grille allongée est L = 50nm
- Le β-ratio optimal est de 1,25.
- Le ratio de mise en parallèle est R_P = 1,8.
- Le ratio de mise en série est R_S = 3.

4 Les stratégies de conception

Le dimensionnement optimal obtenu précédemment est appliqué en complément d'une stratégie de conception généralisée. A titre d'exemple, les librairies haute performance et basse consommation disponibles pour le fonctionnement nominal se distinguent par l'utilisation de grilles plus longues dans le second cas. Elles partagent toutefois les mêmes règles de dimensionnement.

4.1 Les grilles longues systématiques

On observe l'effet de l'utilisation systématique d'une longueur L = 50nm sur l'ensemble des transistors du chemin logique qF04 en comparaison du chemin standard avec L = 40nm. A la différence de l'étude précédente sur l'effet de L sur le chemin logique qFO4, le β-ratio est égal à 1,25.

A la tension optimale V_{OPT}=350mV, un augmentation du délai de 49% est observée (Figure 2.23), avec en contrepartie une réduction des fuites de 58%. L'énergie par transition de cette stratégie est réduite de 11%. Concernant la variabilité, on note une diminution de l'écart-type relatif du délai de 8%.
On rappellera que cette solution impose des pertes en surface de 36% liées aux contraintes de conception. Ce coût reste toutefois 20% inférieur à celui du passage à la plateforme technologique 65nm.

4.2 Les grilles longues asymétriques

Le développement de la méthodologie qFO4 conduit à l'observation suivante : les réseaux de transistors en parallèle sont responsables du maximum de fuites, tandis que les réseaux de transistors en série sont responsables du maximum en délai. Une simulation des portes Non-ET et Non-OU permet de confirmer cette analyse (Figure 2.24).

En conséquence, une stratégie de grilles longues asymétriques a été développée : agrandir le L des transistors en parallèle sans modifier ce-

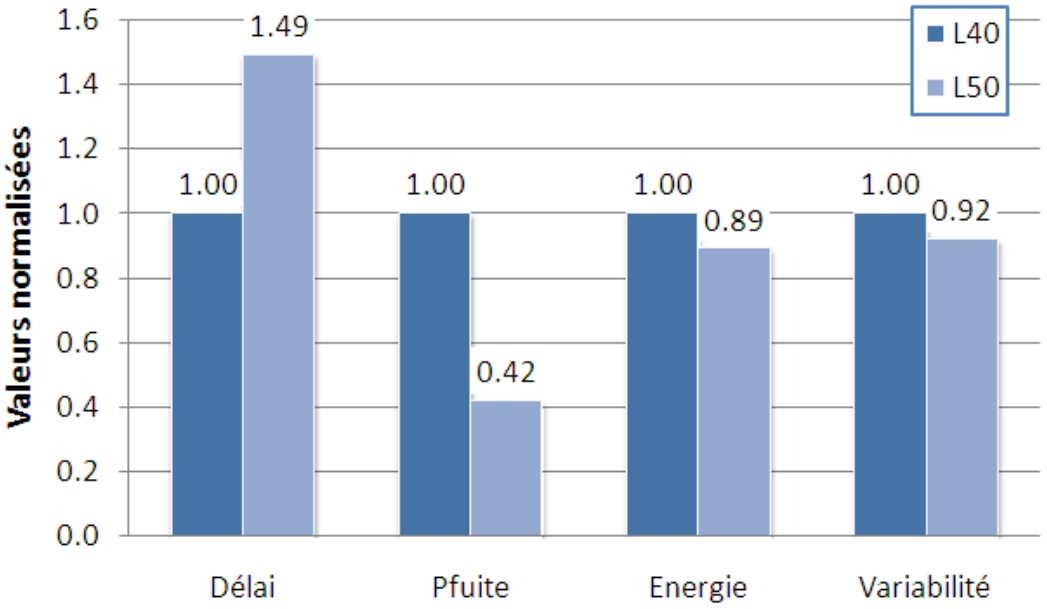

Figure 2.23 – Résultats de la simulation du qFO4 à 350mV pour les stratégies nominales et grilles longues systématiques.

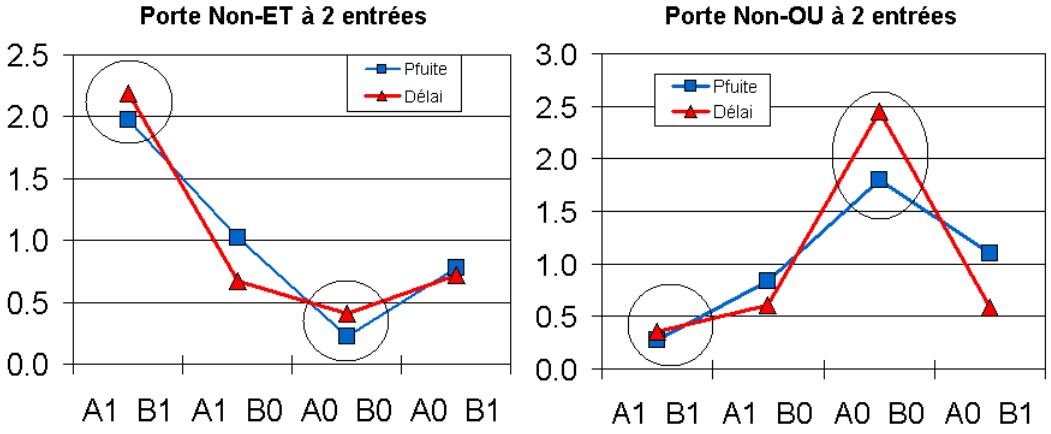

Figure 2.24 – Résultats de la simulation des portes Non-ET et Non-OU à 350mV en fonction des entrées.

lui des transistors en série. On préserve ainsi le courant dynamique du réseau série tout en réduisant le courant statique du réseau parallèle. Le postulat est le suivant :

"*Tout transistor d'un même réseau, ne constituant pas un inverseur, connecté directement à la source d'une part, et à la sortie ou à la grille d'un autre transistor d'autre part, utilise une grille allongée*".

Ce postulat permet d'écarter les transistors placés à la fois en parallèle et en série avec d'autres transistors du réseau. En comparaison

de la stratégie standard, pour un $L_{ASYMETRIQUE}$ de 50nm, on note une réduction de l'énergie par transition de 13% pour une augmentation du délai de 6% (Figure 2.25). Le gain en termes de variabilité est de 2%. Enfin, le coût en surface est ramené à 19%, confirmant le meilleur compromis offert en comparaison des grilles longues systématiques.

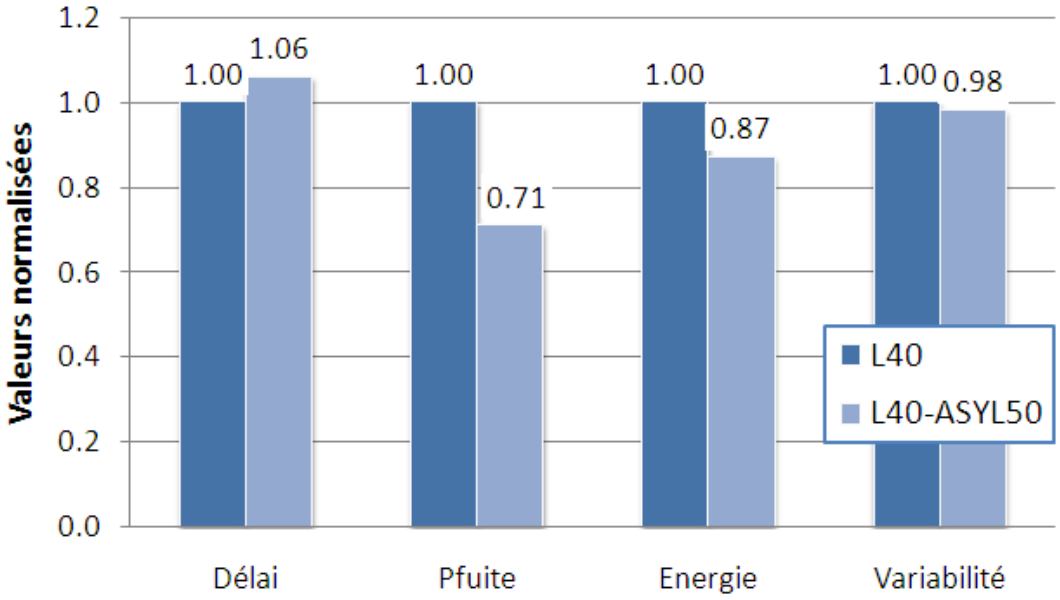

Figure 2.25 – Résultats de la simulation du qFO4 à 350mV pour les stratégies standards et asymétriques.

4.3 La mise en parallèle systématique

L'objectif de cette stratégie, qui consiste à remplacer systématiquement chaque transistor de taille W par deux transistors en parallèle de taille W/2, est de réduire la sensibilité aux variations selon l'hypothèse suivante :

- La variation de l'un des deux transistors est compensée par l'autre. Il existe alors neuf cas de fonctionnement différents (Figure 2.26). On devrait observer une réduction de la probabilité de se retrouver dans les cas extrêmes, qui se traduirait par un resserrement de la distribution.
- Cette réduction est en opposition à la réduction de la dimension de chaque transistor, qui conduit à une augmentation de l'écart-type d'un facteur $\sqrt{2}$.

On note une réduction de la dispersion de l'ordre de 3%. Concernant les performances, on constate un délai 8% supérieur à la structure standard, au prix d'une consommation énergétique plus importante (+22%)

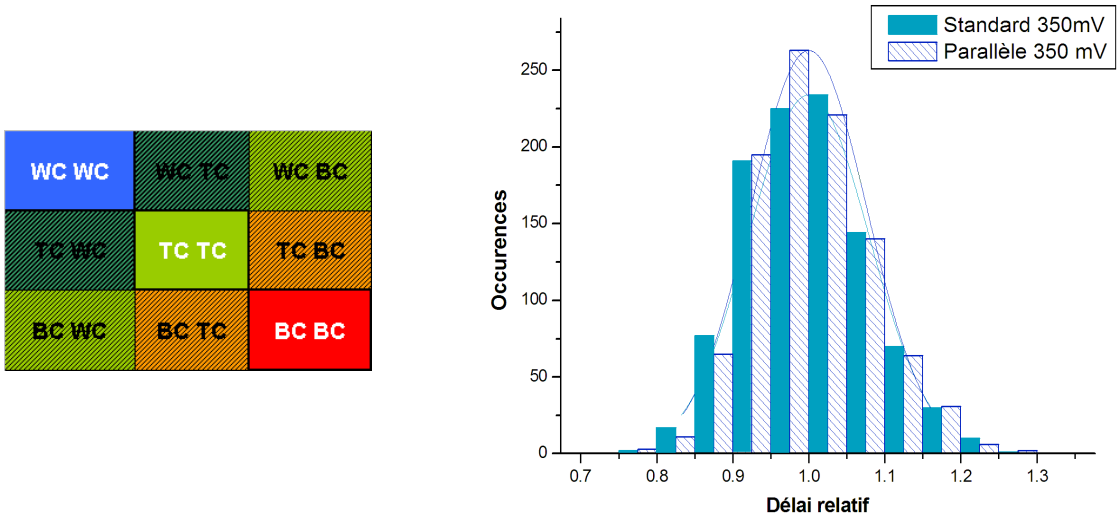

Figure 2.26 – Conditions du couple de transistor et simulation Monte-Carlo de la stratégie de mise en parallèle systématique à 350mV.

en raison des fuites induites par la mise en parallèle (Figure 2.27). L'augmentation de 63% de la surface limite l'intérêt de cette solution.

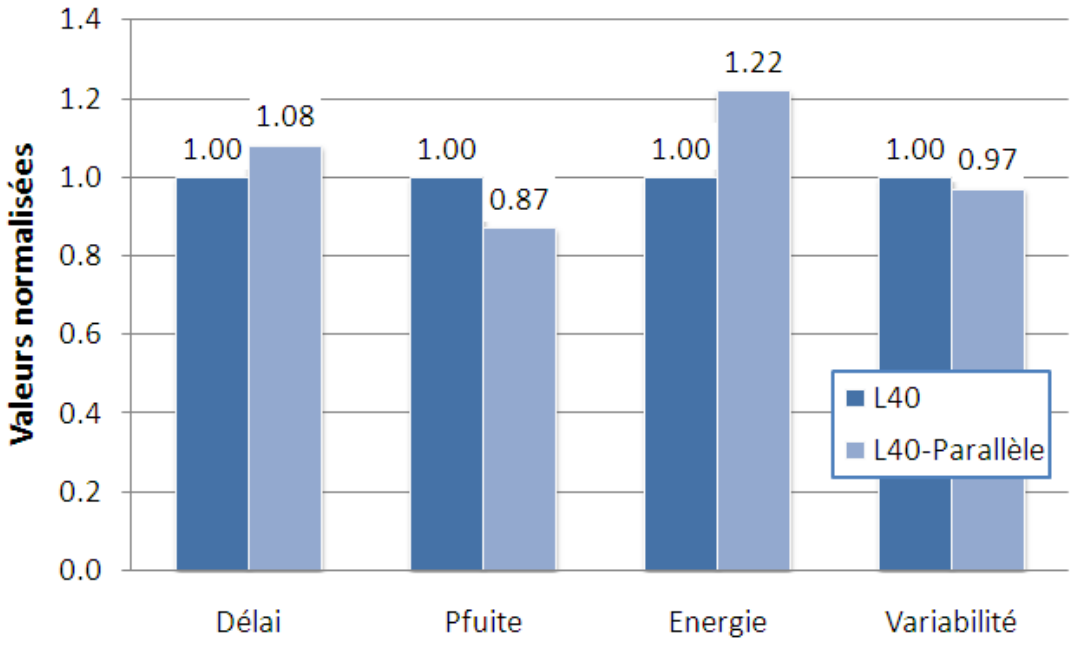

Figure 2.27 – Résultats de la simulation du qFO4 à 350mV pour les stratégies standards et mise en parallèle systématique.

4.4 La mise en série systématique

En connaissance de l'impact de la mise en série d'un transistor sur le courant dynamique, une stratégie de mise en série systématique a été testée sur le chemin logique qFO4. L'objectif est d'évaluer le gain en termes de consommation énergétique. La mise en application consiste à remplacer systématiquement chaque transistor de taille W par deux transistors de taille 3W.

Contrairement aux attentes, on constate une augmentation de l'énergie par transition, associée à l'augmentation de W et à la forte dégradation de la vitesse traduite par des fuites sur un temps plus long (Figure 2.28). En opposition à la mise en parallèle systématique, cette stratégie permet une réduction de la variabilité par l'augmentation de W, tout en induisant une contamination systématique du chemin par chaque transistor. L'écart-type relatif σ/μ est toutefois réduit de 11%. Enfin, les besoins en dimensionnement permettant de maintenir un $I_{Dynamique}$ constant conduisent à une augmentation de la surface de 300%. On conclura sur la nécessité de réduire la présence de structures en série dans la conception à très basse tension.

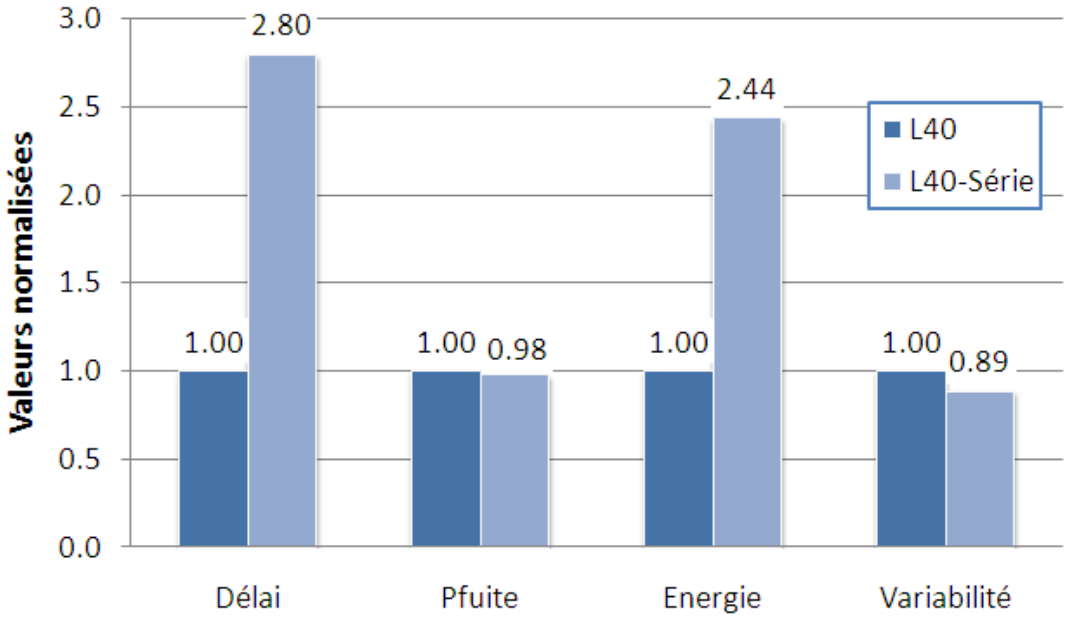

Figure 2.28 – Résultats de la simulation du qFO4 à 350mV pour les stratégies standards et mise en série systématique.

4.5 Récapitulatif et sélection

L'ensemble des résultats sont résumés dans la Figure 2.29. La stratégie de grilles longues asymétriques offre le meilleur compromis en termes de vitesse, d'énergie par transition, de variabilité et de surface.

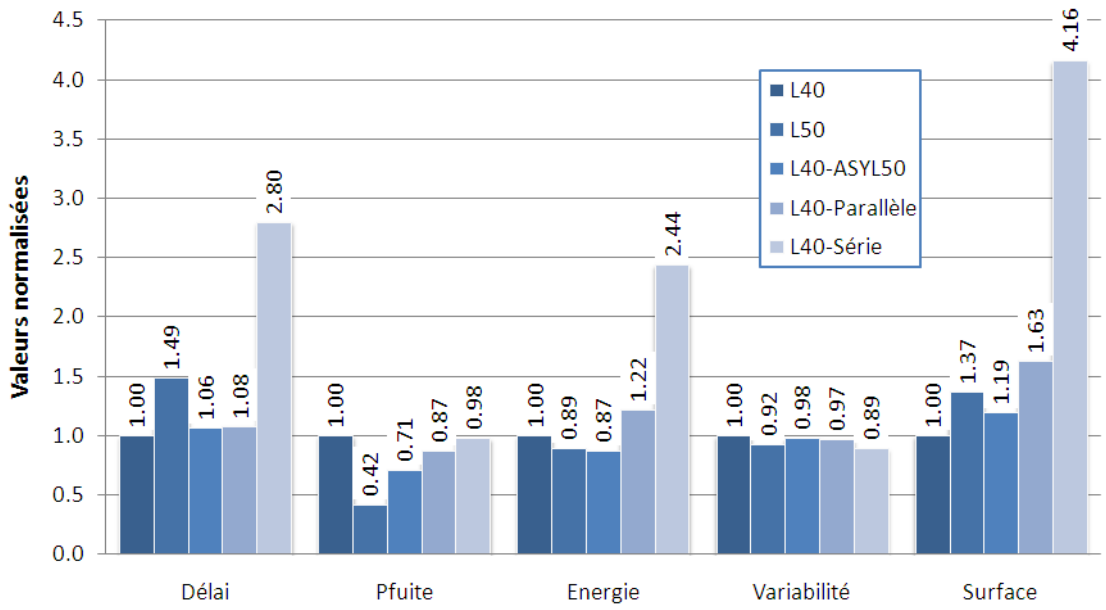

Figure 2.29 – Résultats de la simulation du qFO4 à 350mV pour les différentes stratégies.

5 Construction des librairies de cellules

Les études précédentes permettent d'engager la conception de librairies de cellules optimisées pour le fonctionnement à très faible tension tout en maintenant des performances raisonnables à tension nominale.

5.1 Méthodologie de conception

Si les librairies de cellules matures peuvent contenir plus de 800 cellules, comprenant une centaine de fonctions dupliquées avec plusieurs dimensionnements différents, la définition de la liste de cellules à concevoir a été faite à partir des demandes usuelles de clients lors de l'apparition d'une nouvelle plateforme technologique. Ces demandes consistent en un jeu de cellules réduit mais permettant la réalisation de circuit digitaux suffisamment denses, avec un taux d'occupation de la surface de 80%. Ce choix s'inscrit dans une logique de réduction des coûts de conception.

Ces librairies respectent les choix suivants :

- Un β-ratio de 1,25
- Un couple Rp = 1,8 et Rs = 3
- Un maximum de transistors en série fixé à 3
- L'utilisation de la stratégie de grilles asymétriques systématiques avec L_{ASY} = 50nm

- L'utilisation du SVT sur l'ensemble des librairies
- L'utilisation ponctuelle du LVT.

La définition de la sortance d'une cellule, i.e. de la capacité qu'elle est capable de charger en un délai unitaire FO4, dépend de son dimensionnement. Le nom de chaque cellule comprend une valeur qui caractérise sa sortance. On attribue à l'inverseur de taille maximal, c'est à dire dont le $W_T = W_{P,MAX}+W_{N,MAX}$, la valeur de sortance D8, D correspondant au terme anglais "Drive" qui caractérise la sortance. Ce choix réside dans l'observation suivante :

$$\begin{aligned} W_{N,MAX} &= 600nm \quad \text{pour D8} \\ W_N &= 300nm \quad \text{pour D4} \\ W_N &= 150nm \quad \text{pour D2} \end{aligned}$$

Le minimum de la technologie pour D1 étant à $W_N = 120$nm, dimensionnement qui ne sera pas utilisé dans la librairie en conséquence de l'effet sur la variabilité et le délai.

On défini une relation entre le "Drive" et la dimension des transistors, en fonction des différentes caractéristiques définies précédemment et de la linéarité du courant avec W :

$$W_P = \beta \cdot W_N \tag{2.3}$$

$$Drive = \frac{8 \cdot L_{min} \cdot (W_P \cdot m_P + W_N \cdot m_N)}{L_{eff} \cdot W_{T,MAX}} \tag{2.4}$$

avec $m_{P,N}$ un paramètre utilisé pour convertir un réseau de transistor en un transistor équivalent :

$m_{P,N} = 1$ Pour une structure simple grille,

$m_{P,N} = \frac{N \cdot 1,8}{2}$ Pour une structure avec N transistors en parallèle,

$m_{P,N} = \frac{2}{N \cdot 3}$ Pour une structure avec N transistors en série,

N étant le nombre de transistors.

5.2 Les cellules standards

Cette librairie regroupe l'ensemble des fonctions logiques suivantes :

- **Inverseur :** Réalise l'opération binaire $\overline{A}$.
- **Buffer :** Permet de régénérer une donnée en cas de dégradation de celle-ci sur de longues distances de fils.
- **Non-ET :** Réalise l'opération binaire $\overline{A \cdot B}$. Plusieurs nombres d'entrées sont proposés.
- **Non-OU :** Réalise l'opération binaire $\overline{A + B}$. Plusieurs nombres d'entrées sont proposés.

- **ET-OU :** Réalise l'opération logique (A · B) + C. Plusieurs nombres d'entrées sont proposés.
- **OU-ET :** Réalise l'opération logique (A + B) · C. Plusieurs nombres d'entrées sont proposés.
- **OU-ET-OU :** Réalise l'opération binaire ((A + B) · C) + D.
- **ET-OU Programmable :** En fonction de l'entrée P, réalise l'opération A + B (P = 1) ou A · B (P = 0).
- **OU Exclusif :** Réalise l'opération binaire A ⊕ B.
- **Non-OU Exclusif :** Réalise l'opération binaire $\overline{A \oplus B}$.
- **Multiplexeur :** Permet de sélectionner une entrée parmi plusieurs par sélection. Utilise des portes de transmission.
- **Multiplexeur Trois États :** Permet de sélectionner une entrée parmi plusieurs par sélection. Utilise des inverseurs trois états.

En comptant les variations du nombre d'entrée et du "Drive", cette librairie compte soixante cellules.

5.3 Les Flip-Flops

Les bascules D sont des éléments logiques permettant de mémoriser puis transmettre une donnée en fonction de l'état de l'horloge (Figure 2.30). Les Flip-Flops, composées de deux bascules D, assurent le séquencement dans un circuit logique. En réponse au front d'horloge montant ou descendant, elles transmettent de manière synchrone, avec l'ensemble des Flip-Flops utilisant la même horloge, la donnée présente en entrée. En addition du délai, correspondant au temps de propagation de la sortie lors du front d'horloge, elles présentent des contraintes temporelles définies par deux caractéristiques :

- **Le setup :** La donnée doit être présentée un temps T_{SETUP} avant le front d'horloge.
- **Le hold :** La donnée doit être maintenue un temps T_{HOLD} après le front d'horloge.

L'élément mémorisant de chaque bascule, constitué de deux portes logiques rebouclées, est dit bi-stable, car il permet de maintenir stable deux états logiques différents. Le temps T_{SETUP} correspond au temps nécessaire pour stabiliser le premier bi-stable, le temps T_{HOLD} correspond au temps nécessaire pour stabiliser le second. Lorsque la tension d'alimentation est réduite, la sensibilité du bi-stable augmente, ce qui réduit sa stabilité. L'effort consiste à renforcer le dimensionnement pour améliorer la stabilité, tout en limitant son impact sur le temps nécessaire à la modification de la donnée dans le bi-stable.

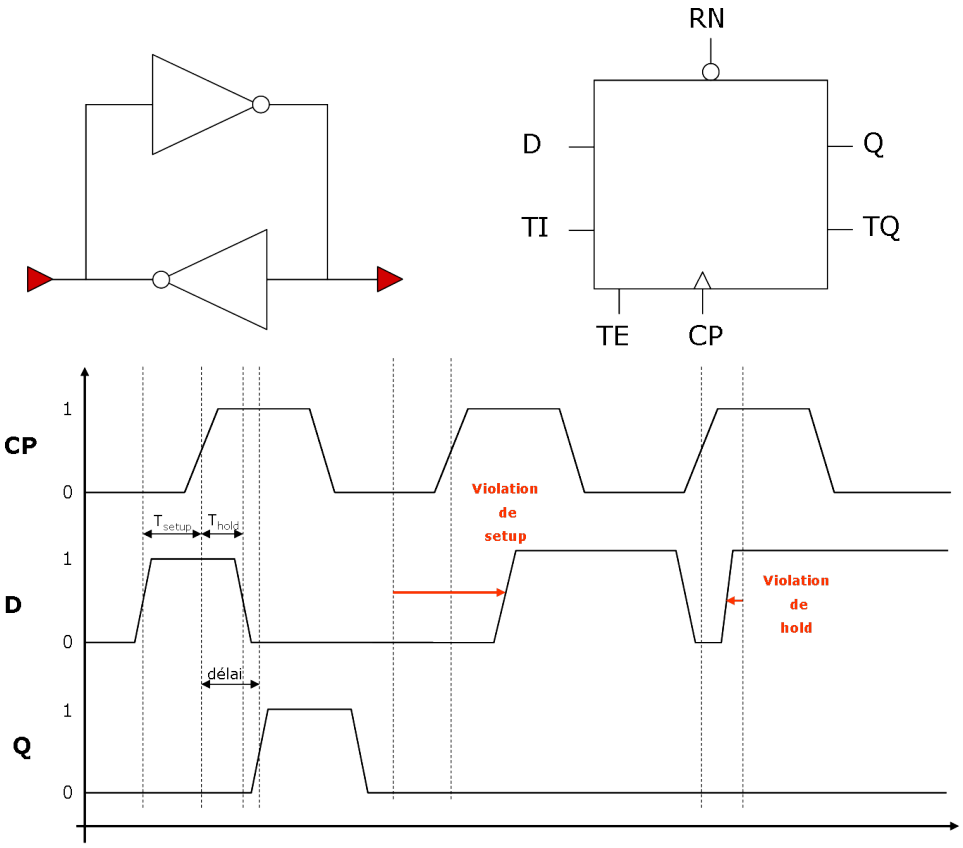

Figure 2.30 – Principe de la mémorisation (gauche), symbole d'une flip-flop (droite), et chronogramme de son fonctionnement.

Deux Flip-Flops ont bénéficié d'un redimensionnement pour le fonctionnement à faible tension. Elles ont été sélectionnées pour les raisons suivantes :

- **La DICE :** Une Flip-Flop CMOS, bénéficiant d'une architecture DICE [99] permettant une plus grande résistance aux radiations. Sa fiabilité est adaptée au marché médical.
- **La FAST :** Une Flip-Flop composée de deux bascules Svensson [100] offrant une plus grande vitesse et des contraintes de temps réduites.

La DICE est une Flip-Flop dite C2MOS, c'est à dire qu'elle utilise une logique CMOS fonctionnelle à faible tension et deux phases d'horloge φ et $\overline{\varphi}$. Elle a bénéficié d'un dimensionnement classique à l'image des cellules standards précédentes.
La FAST est une Flip-Flop utilisant une seule phase d'horloge, comportant par conséquent certains réseaux de transistors non-CMOS. Les chemins commandés par l'horloge rencontrent des difficultés à très faible fréquence liées à l'absence du réseau opposé. Des transistors additionnels permettant de contrer les phénomènes de court-circuit sont

utilisés. Un dimensionnement additionnel est fait pour garantir la fonctionnalité malgré la sensibilité aux variations. L'avantage de cette cellule est sa vitesse et son temps T_{HOLD} franchement négatif imposé par l'utilisation d'une phase d'horloge unique sur les deux bascules.

Étant donné les contraintes temporelles de ces cellules, la sensibilité de leur structure, et l'importance de leur fonctionnement au sein d'un circuit logique, leur dimensionnement a été affiné à la suite de simulations Monte-Carlo. Ces simulations ont été faites sur les quatre points de caractérisation, auxquels ont été ajoutés les deux points de simulations suivant pour émuler les cas les plus défavorables :

- FNSP 0°C
- SNFP 0°C

Ces deux Flip-Flops sont comparées à la Flip-Flop C2MOS présente dans la librairie de cellules nominale (Figure 2.31). Celle-ci conserve son dimensionnement d'origine, et n'est pas fonctionnelle sur l'ensemble des simulations Monte-Carlo.

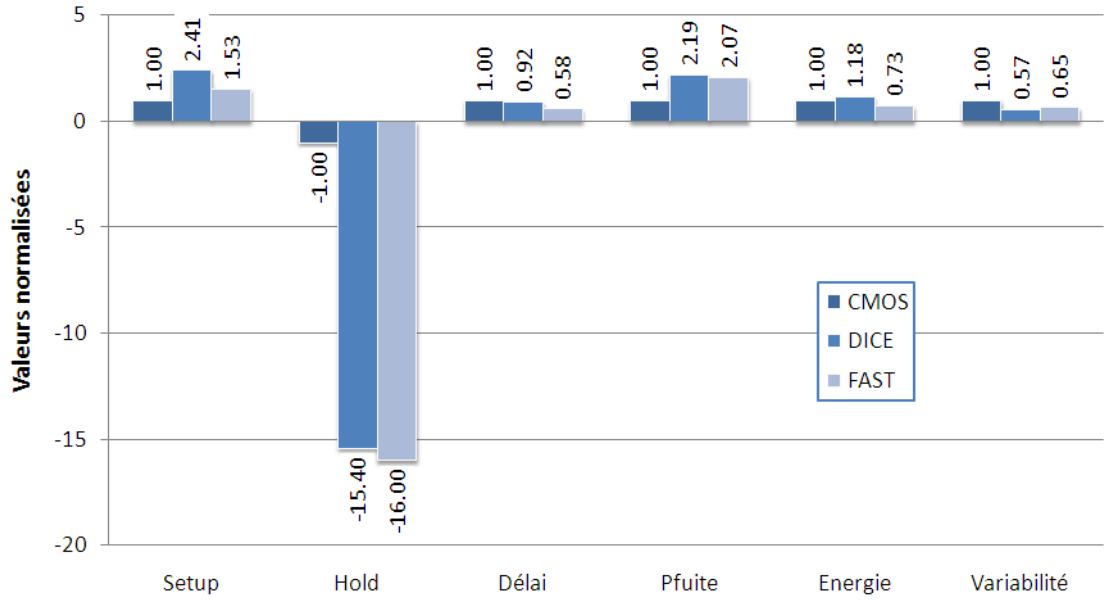

Figure 2.31 – Résultats de la simulation à 350mV pour les différentes Flip-Flops.

La FAST affiche une réduction du temps [délai + T_{SETUP}] de 2%. Ce faible gain est la traduction du dimensionnement réalisé pour garantir la fonctionnalité. Elle présente toutefois un gain énergétique de 27% et une contrainte T_{HOLD} nettement inférieure. Enfin, elle permet de réduire la sensibilité aux variations de 43%. La DICE présente un T_{SETUP} plus élevé tout en assurant un T_{HOLD} plus court que celui de la Flip-Flop

nominale. Bien que présentant des atouts face à cette dernière, ses performances en termes de vitesse, d'énergie et de variabilité, sont inférieures à celles de la FAST : elle ne sera utilisée qu'au bénéfice de sa robustesse aux radiations.

5.4 Les cellules d'interface

Les circuits faible tension peuvent être utilisés en tant que sous-parties d'un circuit fonctionnant à tension nominale. Par conséquent, il y a communication de données, en entrée et en sortie, entre la tension nominale et la faible tension.

De la tension nominale à la très faible tension

L'étude bibliographique a révélé peu de solutions dans ce domaine. Pourtant, il existe une dégradation importante de la sortie d'une porte logique alimentée à 0.35V et commandée à 1.1V. Si l'on prend pour exemple un inverseur qui serait une cellule d'entrée d'un bloc faible tension placé dans un environnement à tension nominale, on observe un comportement différent des deux transistors qui le composent :

- PMOS faible tension
  - IOFF très faible : Source à 0.35V, grille à 1.1V
  - ION faible : Source à 0.35V, grille à 0V

- NMOS tension nominale
  - IOFF standard : Source à 0V, grille à 0V
  - ION important : Source à 0V, grille à 1.1V

Ce qui crée un déséquilibre d'un rapport x50 entre le temps de montée et le temps de descente de l'inverseur (Figure 2.32). Le résultat sur un signal d'horloge est la dégradation du rapport cyclique.

VDD = 350mV, VG = 1.1V		
Porte Logique	Temps de monté	Temps de descente
Inverseur	1	0.02
Non-ET	1	0.04

Figure 2.32 – Résultats de la simulation pour V_{DD} = 350mV et V_G = 1.1V.

La solution proposée consiste à créer une librairie dédiée, composée de cinq cellules d'entrée classiques : ET, OU, NON-ET, NON-OU,

et Buffer. Ces cellules sont redimensionnées pour compenser le déséquilibre : on dégrade volontairement le courant dynamique du réseau NMOS par la mise en série, et on augmente le courant dynamique du réseau PMOS par la mise en parallèle. Le dimensionnement optimal correspond au compromis entre la compensation du déséquilibre à basse tension et le maintien d'un équilibre correct en cas de fonctionnement à tension nominale. Le dimensionnement choisi permet d'atteindre un rapport x10 pour V_{DD} = 0.35V, et x2 pour V_{DD}=1.1V.

De la faible tension à la tension nominale

L'étude bibliographique a permis de mettre en avant les difficultés de la réalisation d'un convertisseur basse tension vers tension nominale. La limite de la tension d'entrée du convertisseur standard est 500mV. Pour permettre la fonctionnalité jusqu'à 350mV, quatre optimisations principales ont été implémentées (Figure 2.33) :

- **Diode HVT :** Pour faciliter le blocage des PMOS de tête, par l'introduction d'une chute de tension entre VDD et leur source.
- **W_{PMOS} réduit :** Pour réduire le courant de fuite des PMOS de tête.
- **NMOS LVT :** Pour opposer un courant dynamique nettement supérieur au courant de fuite des PMOS de tête.
- **Second étage :** Pour récupérer la chute de tension induite par le premier étage et garantir le niveau de tension en sortie
- **Grilles longues distribuées :** Pour réduire le courant de fuite des PMOS de tête ainsi que des NMOS LVT.

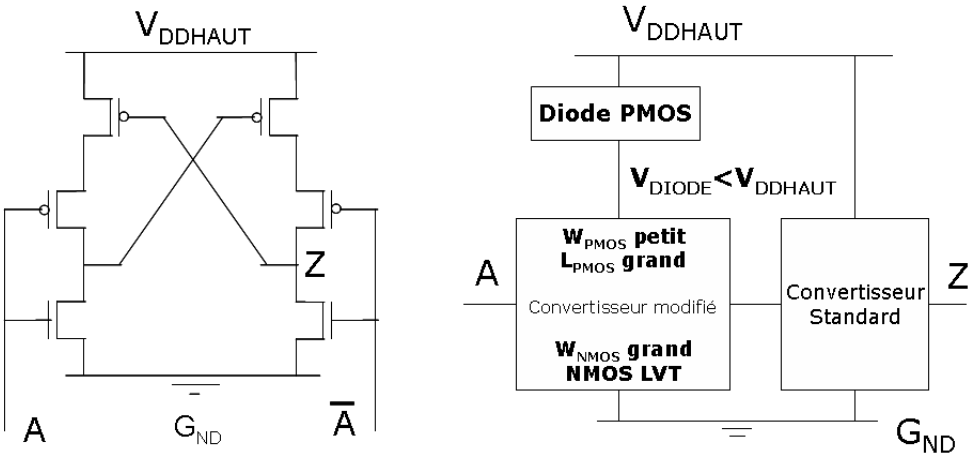

Figure 2.33 – Convertisseur de tension standard (gauche) et convertisseur de tension très faible tension (droite)

A l'image des Flip-Flops, le dimensionnement a été affiné au travers de simulations Monte-Carlo sur l'ensemble des points de caractérisation et les points de simulation spécifiques utilisés précédemment pour

les Flip-Flops. Le convertisseur de tension supporte une tension d'entrée minimale de 0.35V pour une tension de sortie maximale de 1.2V, et affiche un délai équivalent à celui d'une cellule standard.

5.5 Les cellules d'horloge

La construction de l'arbre d'horloge, qui distribue l'horloge à tous les éléments séquentiels du circuit, utilise des buffers pour régénérer le signal en plusieurs points de l'arbre, mais également pour répondre aux contraintes temporelles, cas étudié dans le chapitre suivant. La sensibilité aux variations à très faible tension rend difficile la prédiction du comportement temporel de cette arbre, ce qui peut conduire à la violation des contraintes des blocs séquentiels. Une librairie spécifique de cellules d'horloge a été développée en conséquence. L'objectif des cellules de cette librairie est de réduire la sensibilité aux variations tout en garantissant un courant dynamique élevé. Par conséquent, elles implémentent les solutions suivantes :

- **LVT :** Tous les transistors sont en LVT pour réduire la dispersion et augmenter le courant dynamique.
- **Mise en parallèle systématique :** Tous les transistors sont parallélisés pour réduire la dispersion.
- **Grilles très longues :** Tous les transistors ont une grille allongée de 100% pour compenser l'effet du LVT et de la mise en parallèle, réduire la dispersion, tout en ayant un impact limité sur le courant dynamique.

Le résultat de chaque optimisation sur le délai, les fuites et l'énergie par transition sont présentés sur la Figure 2.34. On remarque que le cumul de ces solutions permet de retrouver une distribution à très faible tension équivalente à celle obtenue à tension nominale (Figure 2.35).

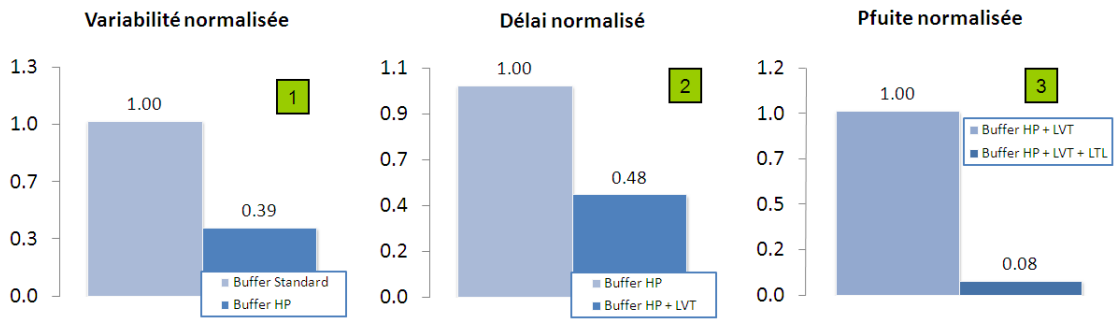

Figure 2.34 – Effet de chaque solution simulé à 350mV : HP = Hautement parallélisé, LTL = L Très Longue.

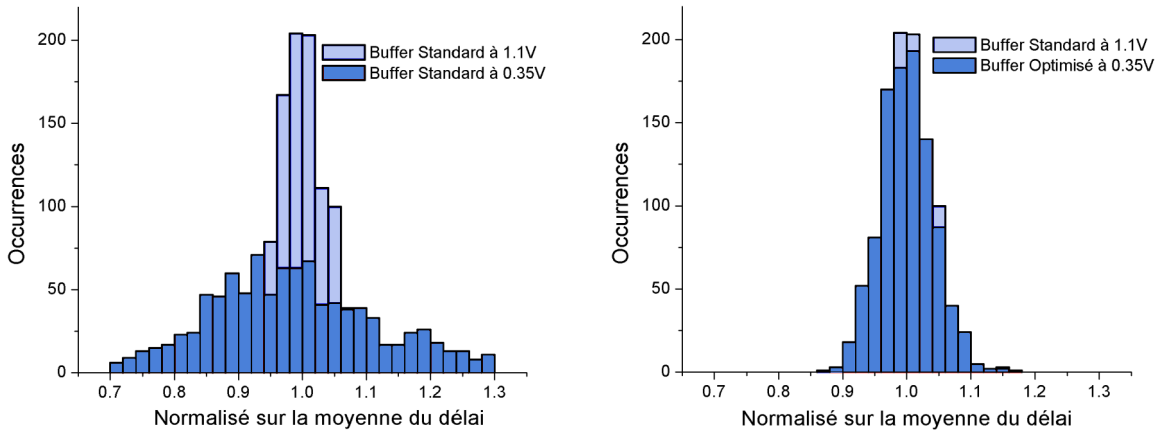

Figure 2.35 – Superposition des différentes distributions : Buffer standard avec V_{DD}=1.1V, Buffer standard avec V_{DD}=0.35V, et Buffer optimisé avec V_{DD} = 0.35V.

6 Caractérisation des librairies

La caractérisation d'une librairie consiste à reporter les performances en termes de vitesse et de consommation énergétique des cellules qui la composent, en fonction de la pente du signal d'entrée et de la charge en sortie de la cellule. Ces informations sont inscrites dans un fichier lisible par les outils de conception de circuits automatisés. L'objectif est de permettre à ces outils de déduire les performances d'un chemin logique sans passer par l'étape de simulation électrique. Cette caractérisation doit mettre en évidence l'ensemble des conditions de fonctionnement finales des cellules : gamme de températures, gamme de tensions, gamme de conditions liée aux procédés de fabrication, gamme de pentes d'entrée et gamme de charges en sortie. Il s'agit de la methodologie PVTSC, pour "Process", "Voltage", "Temperature", "Slope" et "Capacitance", en langue anglaise.

6.1 Définitions des points de caractérisation PVT

Les points de caractérisation ont été définis en début de chapitre. Ils reprennent les conditions de température, tension, et procédés de fabrication destinés à émuler le fonctionnement final des cellules à faible tension. L'objectif étant d'assurer un fonctionnement sur la gamme de tensions 0.35V à 1.2V, les points de caractérisation usuels à tension nominale ont été ajoutés (Tableau 2.1).

L'hypothèse faite étant que le fonctionnement aux tensions intermédiaires est garanti par la monotonicité observée sur les différentes simulations précédentes. VDDI correspond à l'alimentation nominale des cellules d'interfaces. On notera le choix d'une tension minimale fixe,

Table 2.1 – Points de caractérisation définis pour l'ensemble des cellules.

Nom	$VDD = VDDS$	$VDDI = VDDIS$	T	*Conditions*
$PssV0350T000$	$0.35V$	$1.20V$	$0C$	*SSA*
$PttV0350T025$	$0.35V$	$1.10V$	$25C$	*TT*
$PttV0350T037$	$0.35V$	$1.10V$	$37C$	*TT*
$PffV0350T050$	$0.35V$	$1.05V$	$50C$	*FFA*
$PssV1050Tm40$	$1.05V$	$1.05V$	$-40C$	*SSA*
$PssV1050T125$	$1.05V$	$1.05V$	$125C$	*SSA*
$PttV1100T025$	$1.10V$	$1.10V$	$25C$	*TT*
$PffV1200Tm40$	$1.20V$	$1.2V$	$-40C$	*FFA*

V_{DD}=0.35V étant la tension pire cas de fonctionnement souhaitée. Définir une variation de tension pour les points de caractérisation implique un coût en vitesse important lors de l'étape de conception de circuit, traitée dans le chapitre suivant.

6.2 Définition des tables de pentes d'entrée et capacités de sortie SC

La donnée en entrée d'une cellule sur un chemin logique transite d'un état à l'autre avec une pente définie par la cellule précédente. La valeur de la pente d'entrée influe sur le temps de commutation. Les cellules étant simulées seules, il est nécessaire de reproduire les valeurs de pentes d'entrée possibles en fonction des conditions de simulation.

La pente FO4 est extraite au milieu du chemin logique qFO4, simulé sur les différents points de caractérisation. On autorise, en prenant en compte la variation linéaire du délai en fonction de la charge de sortie, une dégradation et une amélioration de la pente d'un facteur x2 et x4. La définition de la gamme de valeurs de pentes est critique car il a été observé une non régénération de celle-ci par certaines cellules. A titre d'exemple, une porte logique ET ayant une pente d'entrée de 1300ns génère une sortie ayant une pente de 1400ns. Si la gamme de pentes définie est bornée à 1300ns, l'outil de conception automatisé sera incapable d'extraire le délai de la cellule commandée par la porte ET, la pente d'entrée de 1400ns n'étant pas définie. Le Tableau 2.2 résume les valeurs de pentes définies pour les différents points de caractérisations.

Chaque cellule possède une valeur de "Drive" équivalent à sa sortance. La définition de la table de capacités de sortie garantie la capacité de chaque porte logique à commander une charge sans détériorer la pente de sortie au-delà des bornes définies précédemment.

Table 2.2 – Gamme de pentes d'entrée définie pour chaque point de caractérisation

$PssV0350T000$	Pentes = 10, 75, 150, 300, 600, 1250	$Moyenne = 300$
$PttV0350T025$	Pentes = 1, 7.5, 15, 30, 60, 120	$Moyenne = 30$
$PttV0350T037$	Pentes = 3, 20.5, 41, 82, 164, 342.5	$Moyenne = 82$
$PffV0350T050$	Pentes = 0.1, 0.75, 1.5, 3, 6, 12	$Moyenne = 6$

Le tableau est ainsi construit : C_{MIN} C_{FO1} C_{FO2} C_{FO4} C_{2FO4} C_{4FO4}. Littéralement, cela signifie que chaque cellule est capable de régénérer sa pente d'entrée pour une capacité de sortie jusqu'à quatre fois supérieure à la capacité FO4. La charge minimum correspond à la capacité d'entrée de la plus petite cellule de la librairie, correspondant à l'inverseur de Drive "D2". La charge FO1 du Drive "D8" a été extraite en tant que référence : les autres capacités C_{FO2}, C_{FO4}, C_{2FO4} et C_{4FO4} ainsi que l'ensemble des autres Drives de la librairie ont été calculés à partir de cette référence en accord avec la dépendance linéaire.

A titre d'exemple, quelques définitions de capacités :

- **D2LOAD =** 0.0001 0.0001 0.0002 0.0005 0.0010 0.0019 ;
- **D8LOAD =** 0.0001 0.0005 0.0010 0.0019 0.0038 0.0076 ;
- **D21LOAD =** 0.0001 0.0012 0.0025 0.0050 0.0100 0.0200 ;
- **D86LOAD =** 0.0001 0.0051 0.0102 0.0204 0.0409 0.0817 ;

7 Validation de la méthodologie

7.1 Maintien des performances à tension nominale

On s'assure dans un premier temps que les optimisations réalisées permettent de maintenir la fonctionnalité et les performances à tension nominale.

On constate que la librairie de cellules standards optimisées à très faible tension affiche à tension nominale une diminution des performances en vitesse de 2% pour une réduction de la consommation statique et de l'énergie respectivement de 34% et 9%, comparativement à la librairie standard. Il existe donc un gain énergétique pour un coût en vitesse restreint, atteint principalement par l'utilisation de grilles asymétriques.

7.2 Vérification de la réplicabilité

La méthodologie qFO4 a été répliquée sur les plateformes technologiques 65nm et 32nm. On observe la validité de la stratégie de grilles

asymétriques sur l'ensemble de ces plateformes. Le compromis offert par cette stratégie est vérifié, comme reporté sur la Figure 2.36.

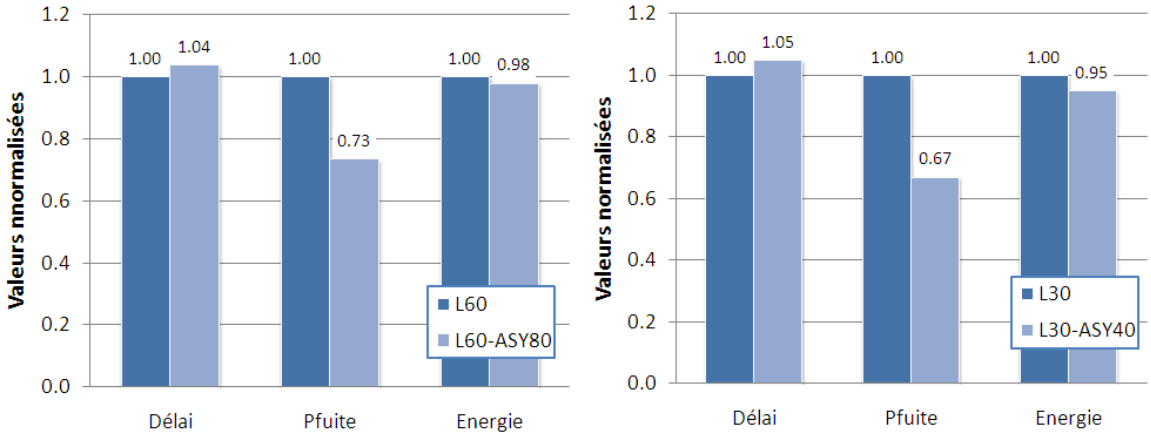

Figure 2.36 – Simulation du chemin logique qFO4 à 350mV, pour les technologies 65nm et 32nm.

7.3 Mesures sur silicium

Pour valider la méthodologie, les cellules ont été reproduites sur silicium sous forme d'oscillateurs, en technologie 90nm, 65nm, 40nm, et 32nm.

Conception d'oscillateurs

Plusieurs oscillateurs ont été conçus :

- un oscillateur de 50 qFO4 utilisant la stratégie standard
- un oscillateur de 50 qFO4 utilisant les grilles longues systématiques
- un oscillateur de 50 qFO4 utilisant les grilles longues asymétriques
- un oscillateur de 50 qFO4 utilisant la mise en parallèle systématique
- un oscillateur de 20 convertisseurs de tension
- un oscillateur de 60 Flip-Flops CMOS
- un oscillateur de 60 Flip-Flops FAST

Le principe de l'oscillateur consiste à créer une chaine rebouclée sur elle même, avec une sortie inversée à chaque traversée. Le schéma de ces oscillateurs est présenté en Figure 2.37.

Analyse des résultats

Les résultats des différents oscillateurs sont comparés aux résultats de simulations. On constate sur la Figure 2.38 que les oscillateurs

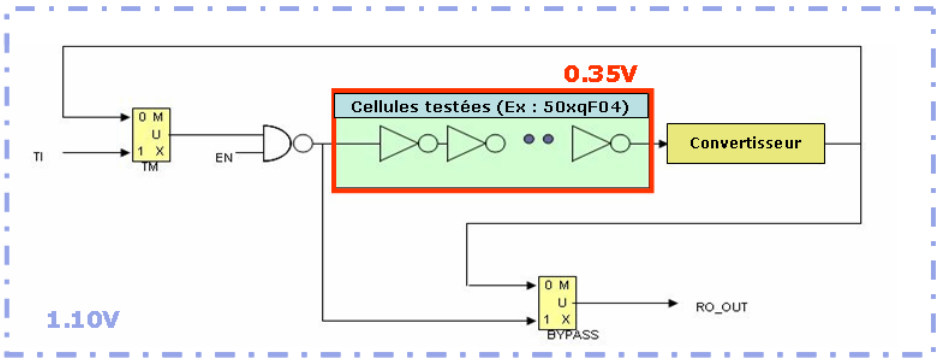

Figure 2.37 – Schéma de l'oscillateur conçu pour la validation sur silicium.

composés de Flip-Flops sont fonctionnels à très faible tension. Sur l'ensemble des oscillateurs, on observe une fonctionnalité sur la gamme de tension 0,35V à 1,2V, et notamment sur le convertisseur de tension. La monotonicité fréquentielle en fonction de la tension d'alimentation permet de valider la définition de points de caractérisation aux seuls extremums de tension.

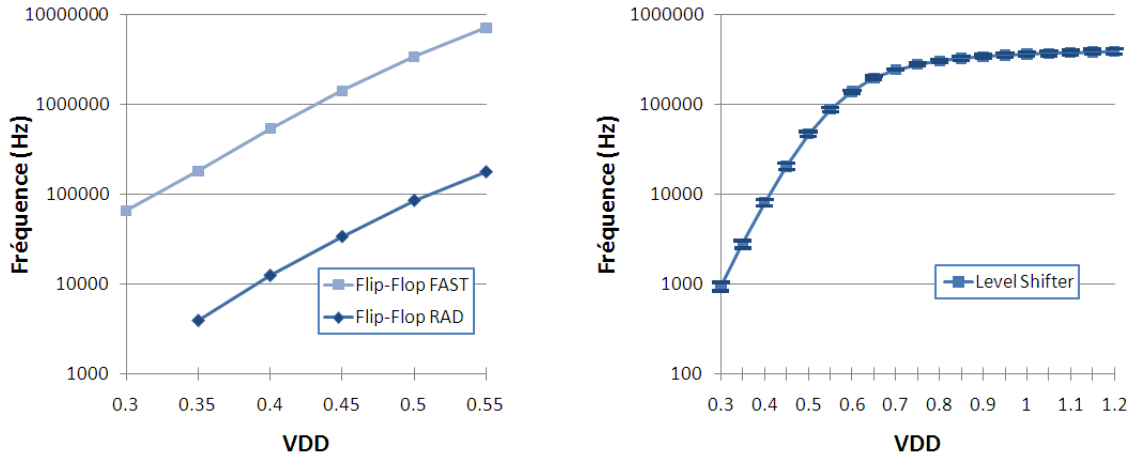

Figure 2.38 – Mesure sur silicium des oscillateurs de Flip-Flops (gauche) et de l'oscillateur composé de convertisseurs de tension (droite), à 25°C.

Les différentes stratégies affichent un comportement fréquentiel équivalent à celui observé lors de l'étude par la simulation sur le chemin logique qFO4 (Figure 2.39). En raison d'un partage des rails d'alimentation entre les différents oscillateurs, il est difficile de dissocier les contributions énergétiques de chacun, mais ces résultats confortent l'intérêt de la stratégie de conception asymétrique.

On s'intéresse cependant au gain énergétique permis par la réduction de la tension d'alimentation. On note sur la Figure 2.40 que l'énergie dynamique est réduite d'un facteur x6 entre 0,35V et 1,1V. Concernant la variabilité à très faible tension, elle augmente avec la température, partant d'un facteur x1 à 0°C pour atteindre un facteur x5,3 à

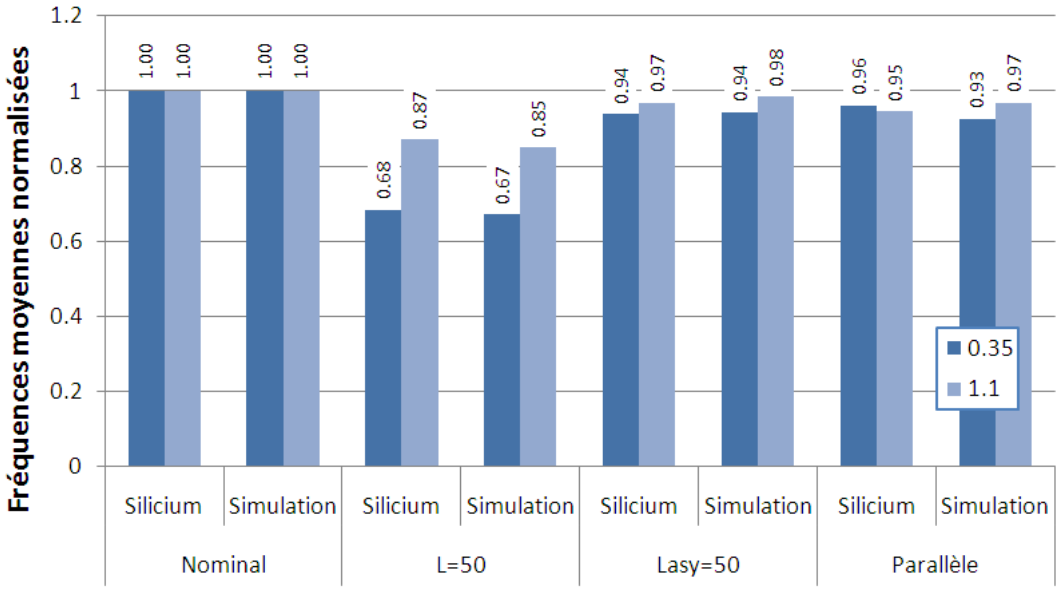

Figure 2.39 – Mesures sur silicium de la fréquence des différents oscillateurs, à 25°C.

50°C comparé à la tension nominale. Cette variabilité illustre les variations globales, les variations locales étant moyennées par les multiples étages de l'oscillateur.

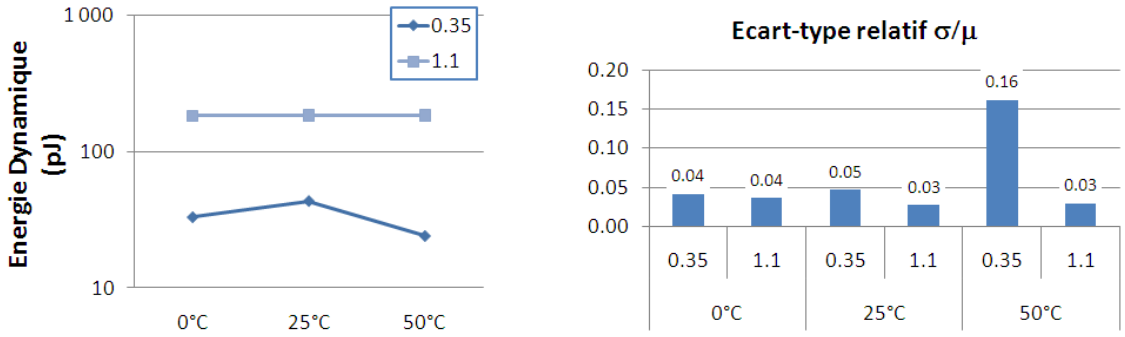

Figure 2.40 – Mesure sur silicium de l'énergie dynamique (gauche) et de la variabilité (droite) d'un oscillateur, en fonction de la température.

Enfin, le fonctionnement des cellules est vérifié lors d'une étape OLT, pour Operating Life Test. Cette procédure permet de mesurer les effets du vieillissement par des temps de mesure atteignant les 168 heures successives. Aucun effet n'a été observé lors de cette opération.

Ces différents résultats permettent de valider la méthodologie en confirmant le comportement fréquentiel, la tendance énergétique, l'évolution de la variabilité et la fonctionnalité à très faible tension, dans le cadre des observations réalisées par la simulation.

8 Récapitulatif de la contribution

Les développements très faible tension suivants ont été mis en avant dans ce chapitre :

- Une étude sur la précision des modèles
- Une méthodologie d'étude qFO4 réplicable
- La définition des paramètres de dimensionnement de cellule pour un couple énergie-délai optimal
- La conception d'une librairie de cellules standards optimisées
- La conception de deux flip-flops optimisées
- La conception de cellules d'interface
- La conception de cellules d'horloge à courant dynamique élevé et à faible variabilité
- La définition des points de caractérisation, pentes et capacités
- La validation de la méthodologie sur les plateformes technologiques 65nm et 32nm.
- La validation de la méthodologie par des mesures sur silicium d'oscillateurs.

Ces travaux permettent à un concepteur le développement de cellules optimisées pour la faible tension, avec des performances raisonnables à tension nominale, en faisant appel à une méthodologie indépendante de la plateforme technologique utilisée. Cette méthodologie, qui s'applique en faisant appel aux outils de conception usuels de l'industrie permettant le maintien des coûts de conception, éclaire le concepteur sur les axes d'optimisations disponibles et les gains permis par chacun d'eux. Les limitations liées à l'interface ou aux phénomènes de variabilité ont également été révélés, et une solution pour les limiter a été proposée. Enfin, des points de caractérisation ont été définis pour permettre la conception de circuit sur une gamme de tension étendue. Ces travaux ont été répliqués sur les plateformes technologiques 65nm, 40nm et 32nm, puis validé en 40nm par des mesures sur silicium.

9 Publications scientifiques

L'étude sur les cellules a donnée lieu à la publication de deux brevets [100; 101] et des articles suivants :

– **"Design Solutions for Ultra-Low Voltage** [102]
Beyond the reduction of supply voltage induced by the aggressive scaling strategy, wearable mobile and medical applications set energy needs in priority over perform-ance. Follows a concept of energy efficiency, where the supply voltage is reduced until the system operates with minimum energy consumption. The minimum is achieved in the subthreshold region, where the current is exponentially dependant of

the threshold voltage, resulting in a high sensitivity to variations. In order to go further in the energy reduction, and explore ways for variability mitigation, three design solutions are explored using SPICE simulations : Large L, Redundancy, and Stacking, under 45nm Low Power process, and 50°C temperature.

– **"45nm Ultra-Low Voltage Design"** [103]
To address the energy efficiency needs of wearable mobile and medical applications, the use of ultra-low voltage power supply is a promising strategy, while inducing variability and drastic performance drop issues. To manage those issues while going further in the energy consumption reduction, design solutions are explored using SPICE simulations under 45nm Low Power process, and 0°C to 50°C temperature range.

– **"Ultra-Low Voltage from 65nm to 32nm"** [104]
From 65nm to 32nm, the benefit of using an ultra-low voltage through device scaling is reported. The energy consumption and timing through voltage and device scaling are compared by simulating flip-flops and a combinatorial path including usual cells. This work highlights the better performance of the preliminary 32nm technology over previous ones, while showing a rise of variability.

– **"A 45nm CMOS 0.35v-optimized standard cell library for ultra-low power applications"** [105]
Ultra-low voltage is now a well known solution for energy constrained applications designed using nanometric process technologies. This work is focused on setting-up an automated methodology to enable the design of ultra-low voltage digital circuits exclusively using standard EDA tools. To achieve this goal, a 0.35V energy-delay optimized library was developed. This library, fully compliant with standard library design flow and characterization, was verified through the design and fabrication of a BCH decoder circuit, following a standard front-end to back-end flow. It performs at 457 kHz, with a total energy consumption of 2.9fJ per cycle.

Chapitre 3

Les Circuits Numériques

Les circuits numériques sont réalisés par la succession de différentes tâches et l'utilisation d'un ensemble d'outils industriels. Ce séquencement, appelé flot de conception, fait appel aux librairies développées précédemment.

A la suite d'une introduction sur les différentes tâches du flot de conception, la réalisation d'un circuit numérique, fabriqué en technologie 40nm, est décrite. Ce circuit intègre les librairies très faible tension tout en utilisant un flot de conception nominal. Suite à la mesure du circuit et l'analyse des résultats, des marges et adaptations ont été mises en lumière pour permettre la réalisation d'un circuit numérique fonctionnant à très faible tension et possédant un rendement et un coût de conception équivalents au nominal.

1 Le flot de conception industriel

Le processus de conception usuel d'un circuit numérique industriel est le suivant (Figure 3.1) :

- Conception logique
- Synthèse
- Placement et routage
- Vérification

La conception logique consiste en l'écriture de la fonctionnalité d'un circuit dans un langage de programmation de type VHDL ou VERILOG. Aucun travail spécifique n'a été réalisé sur cet aspect dans le cadre de cette thèse.

L'exécution des autres étapes nécessite que le contenu suivant soit intégré à la librairie :

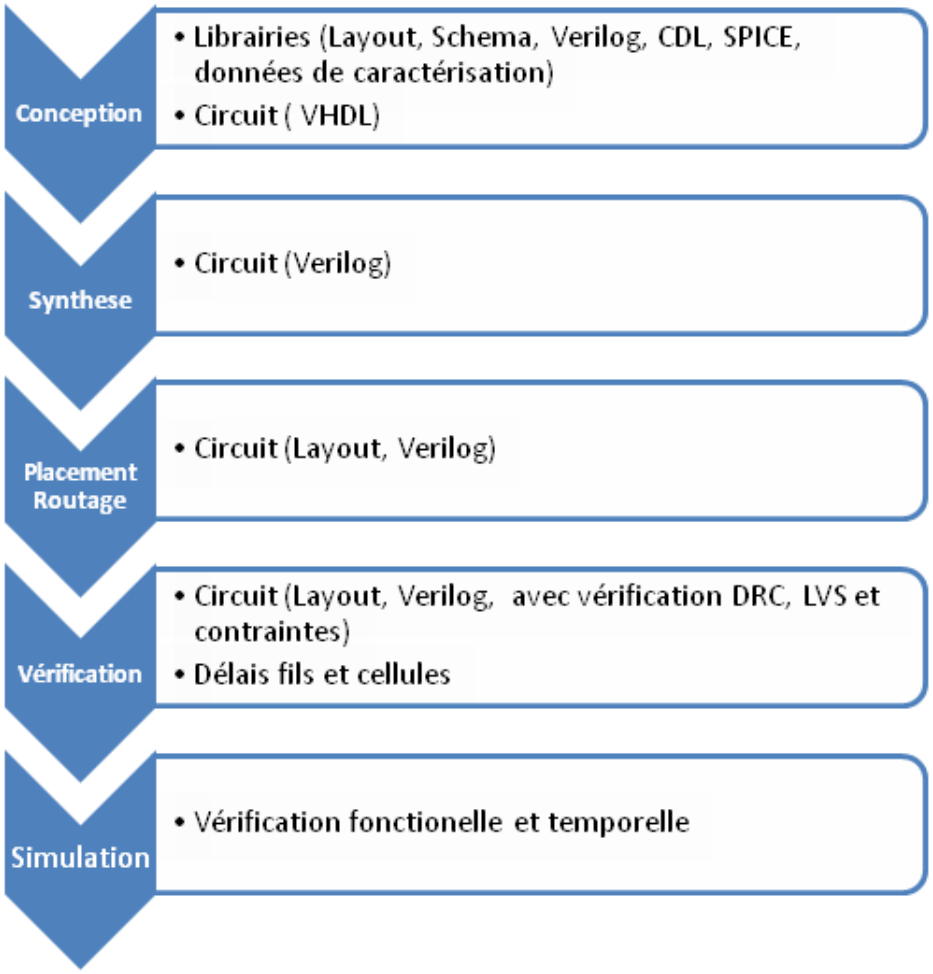

Figure 3.1 – Représentation du flot de conception d'un circuit numérique.

- **La description logique des cellules en VERILOG**

Ces données permettent la synthèse d'un circuit.

- **Les schémas et layouts des cellules**

Ces données permettent le placement et le routage d'un circuit.

- **Les données de caractérisation définissant les performances des cellules en vitesse et consommation en fonction des conditions PVTSC**

Ces données permettent la vérification temporelle et énergétique d'un circuit.

1.1 La synthèse

La synthèse consiste à traduire la fonctionnalité d'un circuit en utilisant les cellules contenues dans les librairies. Les fonctions logiques du circuit sont ainsi réalisées en utilisant les descriptions VERILOG des cellules. Le résultat de cette étape est la création d'une description VE-

RILOG du circuit intégrant les cellules (Figure 3.2).

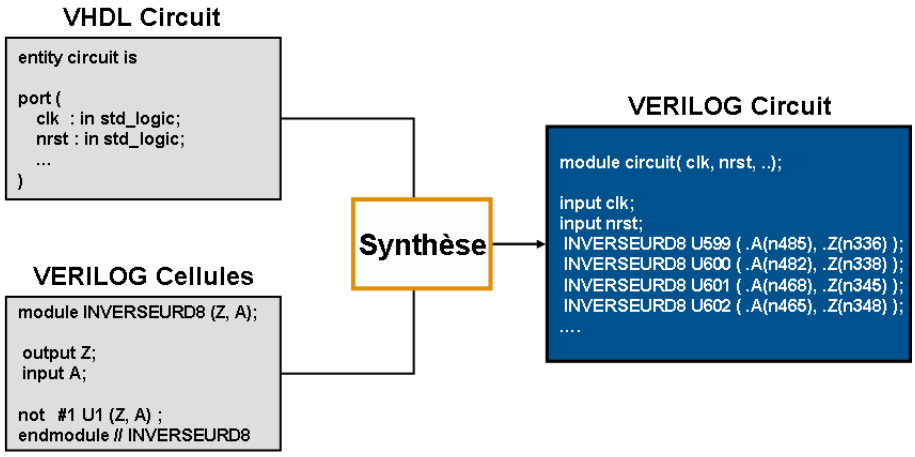

Figure 3.2 – Représentation de l'étape de synthèse d'un circuit.

1.2 Le placement-routage

Le placement

Cette étape réalise le placement des cellules contenues dans le VERILOG élaboré lors de la synthèse, grâce aux informations de surface, de position des entrées/sorties, et des coordonnées des métaux. Ces cellules sont placées selon l'espace alloué au circuit et le taux de remplissage défini par le concepteur (Figure 3.3).

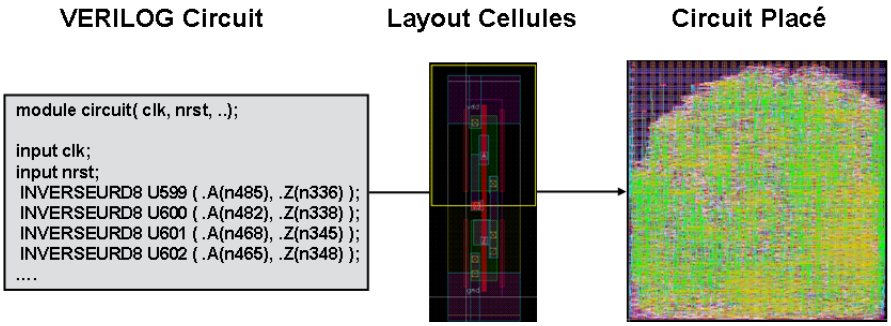

Figure 3.3 – Représentation de l'étape de placement d'un circuit.

C'est lors du placement que les domaines de tension sont définis. La Figure 3.4 présente deux placements différents : un premier placement mono-domaine, où l'ensemble de la surface est alloué au circuit, et un second placement multi-domaines. Dans ce cas où deux domaines de tension sont définis, la surface est divisée en deux parties, chaque domaine étant isolé de l'autre par séparation physique des rails d'alimentation.

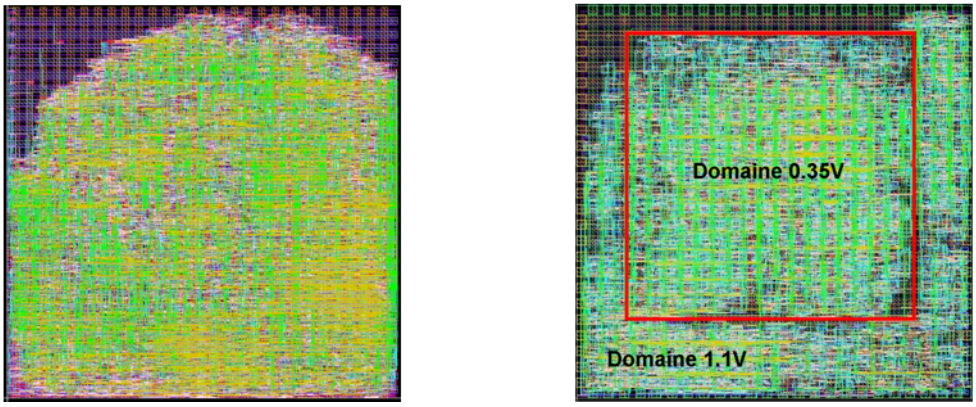

Figure 3.4 – Placement mono-domaine (gauche) et multi-domaines (droite).

C'est également au cours de cette étape que l'arbre d'horloge est élaboré. Il s'agit de la propagation de l'horloge principale à l'ensemble des cellules séquentielles qu'elle commande. Cette étape est fondamentale pour garantir la fonctionnalité dans les circuits synchrones.

Le routage

Le routage réalise la connexion métallique entre les cellules placées en accord avec le VERILOG. Ces fils introduisent des capacités et des résistances qui sont extraites et intégrées dans l'évaluation temporelle. Suite à l'évolution du comportement temporel du circuit, l'outil de routage effectue une vérification du respect des contraintes suivantes :

- **maxCap :** La valeur de capacité maximale en sortie d'une porte logique ne dépasse pas celle définie dans les données de caractérisation.
- **maxFanout :** Le nombre de cellules chargées par une porte logique ne dépasse pas le nombre fixé par le concepteur, ici 16 (4FO4).
- **maxTrans :** Les pentes en entrée de cellule sont inférieures à la pente maximale définie dans les données de caractérisation.
- **Setup :** Les temps T_{SETUP} ne sont pas violés.
- **Hold :** Les temps T_{HOLD} ne sont pas violés.
- **PWL = PWH :** Le rapport cyclique de l'horloge est équilibré.

Lorsqu'une contrainte n'est pas respectée, l'outil modifie le circuit en conséquence. Dans le cas d'une violation temporelle, un buffer sera introduit sur le chemin incriminé pour retarder le signal. Cette étape délivre en sortie le layout et le VERILOG quasi-final du circuit.

1.3 La vérification

L'étape de vérification permet de valider le layout et le VERILOG obtenus à la suite du routage. On vérifie alors :

- L'équivalence fonctionnelle entre le VERILOG d'origine et le VERILOG post-routage
- L'équivalence entre le VERILOG et le layout post-routage.
- Le respect des règles de conception de masques.
- Le respect des contraintes.
- L'absence d'effets de couplage entre domaines de tension différents.

1.4 La simulation back-annotée

Cette étape consiste à vérifier le fonctionnement du circuit en sortie de l'étape de vérification, qui génère un fichier contenant le temps de propagation de chaque porte et chaque fil composant le circuit. Le simulateur numérique réalise une simulation fonctionnelle du circuit, et ajoute les temps contenus dans le fichier cité précédemment. On peut ainsi vérifier le maintient de la fonctionnalité logique du circuit lorsque les délais sont propagés.

2 Réalisation d'un circuit numérique sur silicium

Un circuit numérique est réalisé en suivant le flot de conception nominal. L'objectif de cette étape est d'analyser les différents aspects du flot qu'il est nécessaire d'adapter pour atteindre une fonctionnalité de la très faible tension à la tension nominale associée à un rendement élevé. Ce circuit a été synthétisé avec les librairies développées au Chapitre 2.

2.1 Description du BCH

Le circuit numérique sélectionné est un décodeur d'erreur utilisé usuellement dans les produits de type récepteur vidéo (Figure 3.5). Ce décodeur accompagne les mémoires NAND flash à l'image des ECC utilisés pour les cellules mémoires usuelles. Basé sur l'algorithme de décodage Bose-Chaudhuri-Hoequenghem (BCH) [106], ce circuit permet de repérer jusqu'à quatre erreurs dans une trame de 252bits, et de reporter leur position. Il intègre un bloc permettant d'auto-tester la fonctionnalité du circuit dans le but de réduire le temps et la complexité de la phase de test post-silicium. Ce choix de circuit a été fait en raison

de sa complexité suffisante et de son domaine d'application, adapté au développement de mémoires traité au Chapitre 4.

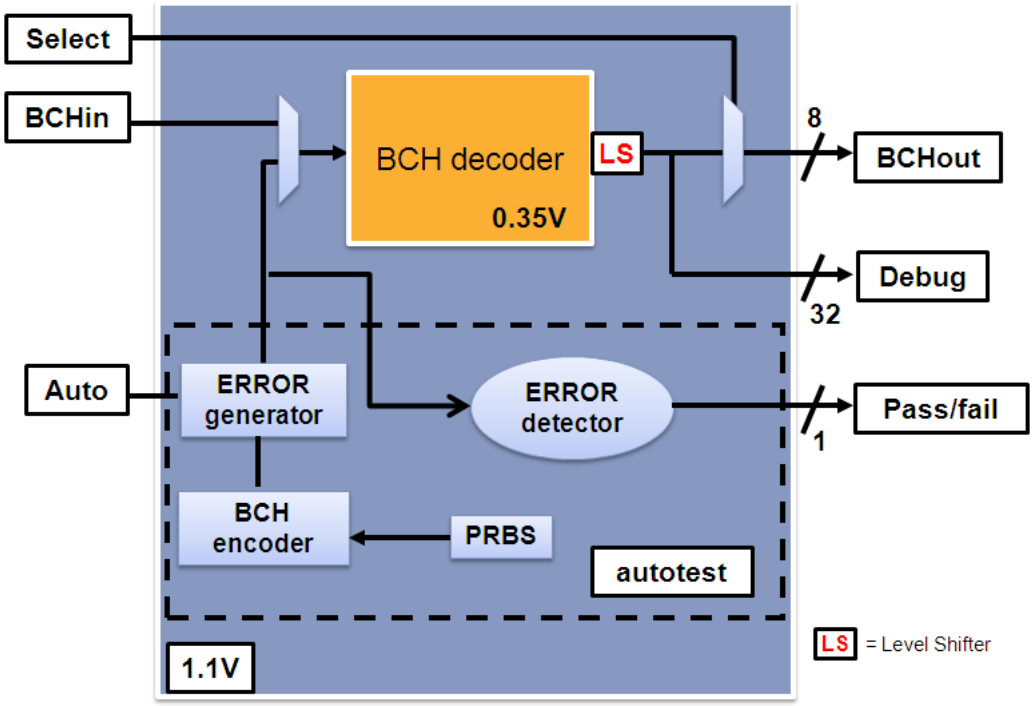

Figure 3.5 – Description schématique du décodeur d'erreur BCH. La sortie "pass" permet par un bit unique de vérifier la fonctionnalité du circuit.

Ce circuit est présenté en deux versions :

- **Une version multi-domaines :** Auto-test à tension nominale et décodeur à très faible tension.
- **Une version mono-domaine :** Auto-test et décodeur à très faible tension.

La version multi-domaines permet la réalisation d'un circuit utilisant simultanément les données de caractérisation nominale et très faible tension. Il est pour cela nécessaire d'avoir un alignement en température. Ce choix s'inscrit également dans la volonté de préserver la fonctionnalité de la partie auto-test, permettant de remonter les erreurs de fonctionnement à très faible tension au seul décodeur. La version mono-domaine sert de référence pour distinguer les effets liés à la présence de plusieurs tensions.

2.2 Conception des circuits

Les circuits présentent les caractéristiques suivantes :

– **Circuit Mutli-Domaines : QLIBD**

- Conception dans la technologie 40nm.
- Utilise la librairie de cellules standards nominale pour le bloc auto-test.
- Utilise la librairie de cellules standards très faible tension pour le bloc décodeur.
- Utilise une cellule d'interface en sortie du domaine à très faible tension.
- N'utilise pas de cellule d'interface en entrée du domaine à très faible tension.
- N'utilise pas de cellules d'horloge spécifiques.
- Utilise le réglage nominal pour la vérification des contraintes.
- La fréquence pire cas est réglée à 33kHz.
- Vérification temporelle avec les points de caractérisation SSA et FFA à très faible tension et à tension nominale pour le bloc décodeur.
- Vérification temporelle avec les points de caractérisation SSA et FFA à tension nominale uniquement pour le bloc auto-test.
- Fabriqué sur silicium 40nm TSMC.
- Permet le Body-Biasing.

– **Circuit Mono-Domaine : MERCURY**

- Conception dans la technologie 40nm.
- Utilise la librairie de cellules standards très faible tension.
- Utilise une cellule d'interface sur chaque sortie du bloc.
- N'utilise pas de cellule d'interface en entrée du bloc.
- N'utilise pas de cellules d'horloge spécifiques.
- Utilise le réglage nominal pour la vérification des contraintes.
- Utilise une fréquence pire cas de 33kHz.
- Vérification temporelle avec les points de caractérisation SSA et FFA à très faible tension et à tension nominale.
- Fabriqué sur silicium 40nm TSMC.

Ces circuits ont été fabriqués sur silicium après avoir passé l'étape de vérification et la simulation back-annotée sans erreur.

2.3 Mesures sur silicium

Tension minimale de fonctionnement

On observe sur le MERCURY une monotonicité de la fréquence sur la gamme de tension 1,1V à 0,55V (Figure 3.6). Le circuit n'est plus fonctionnel en dessous de cette tension, malgré la réduction de la fréquence d'horloge. Le QLIBD présente le même comportement, atteignant une tension V_{MIN} moyenne de 0,55V. Cette conformité des résultats obtenus sur les deux circuits marque l'absence d'effets liés à l'utilisation de deux domaines de tension à l'intérieur d'un même bloc.

Sur les 80 puces mesurées, on note une dispersion de V_{MIN} de 12%.

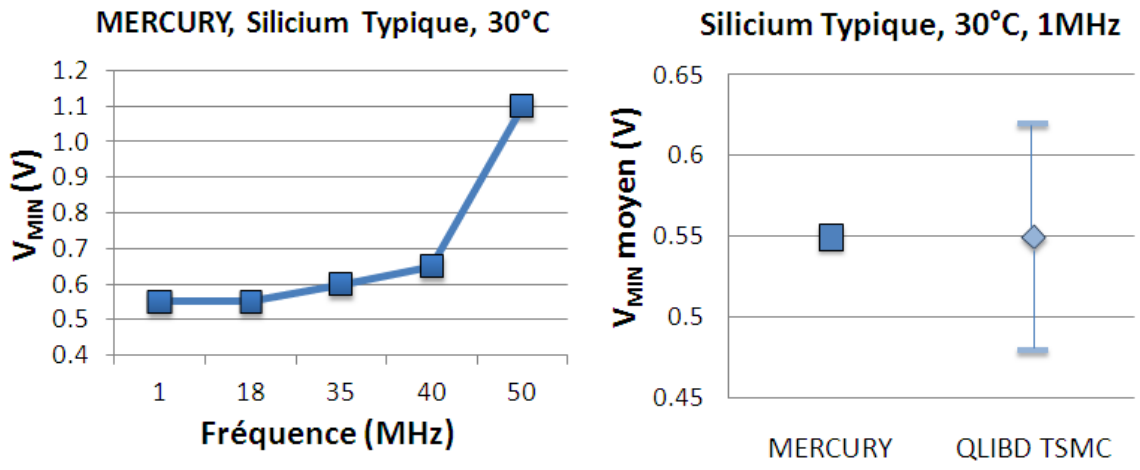

Figure 3.6 – Mesures des circuits MERCURY et QLIBD : Tension minimale en fonction de la fréquence (gauche) et Tension minimale à 1MHz (droite).

On observe que la tension d'alimentation minimale V_{MIN} diminue lorsque la température augmente (Figure 3.7).

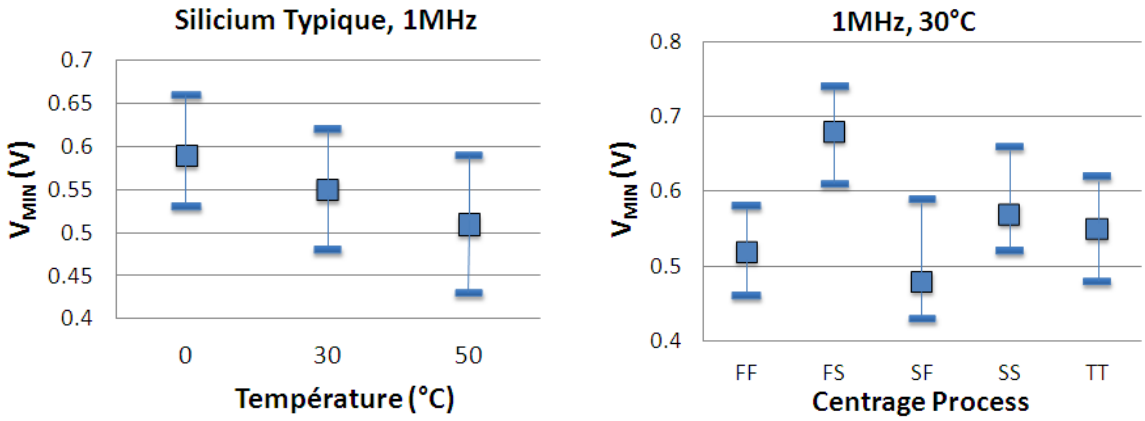

Figure 3.7 – Mesures sur silicium de V_{MIN} sur le QLIBD en fonction de la température (gauche) et en fonction du centrage process (droite).

Les conditions FF, SS, FS et SF ont été reproduites délibérément sur silicium, procédure usuelle de qualification des circuits de test dans l'industrie. Les résultats de ces différents centrages sont reportés sur la Figure 3.7. On observe une variation modérée de la tension minimale pour les siliciums FF, SS, et TT, à 30°C. En contrepartie, une importante dissymétrie est relevée pour les conditions FS et SF. La tension V_{MIN} moyenne la plus faible est obtenue au centrage SF, soit V_{MIN} = 0,47V, en opposition à la tension V_{MIN} moyenne la plus élevée obtenue au centrage FS, soit V_{MIN} = 0,67V.

Compensation par Body-Biasing

Évoqué dans la partie bibliographique, le Body-Biasing permet de reproduire l'effet d'une modification de la tension de seuil d'un réseau de transistor en appliquant un différentiel de tension entre la source du réseau et son substrat. L'application du Body-Biasing sur le circuit produit les résultats reportés sur la Figure 3.8. Dans notre cas, une amélioration est observée principalement en diminuant la tension de seuil du réseau PMOS. En moyenne, le Body-Biasing permet de réduire la tension d'alimentation V_{MIN} de 100mV. On constate que l'effet est inversé sur le circuit SF.

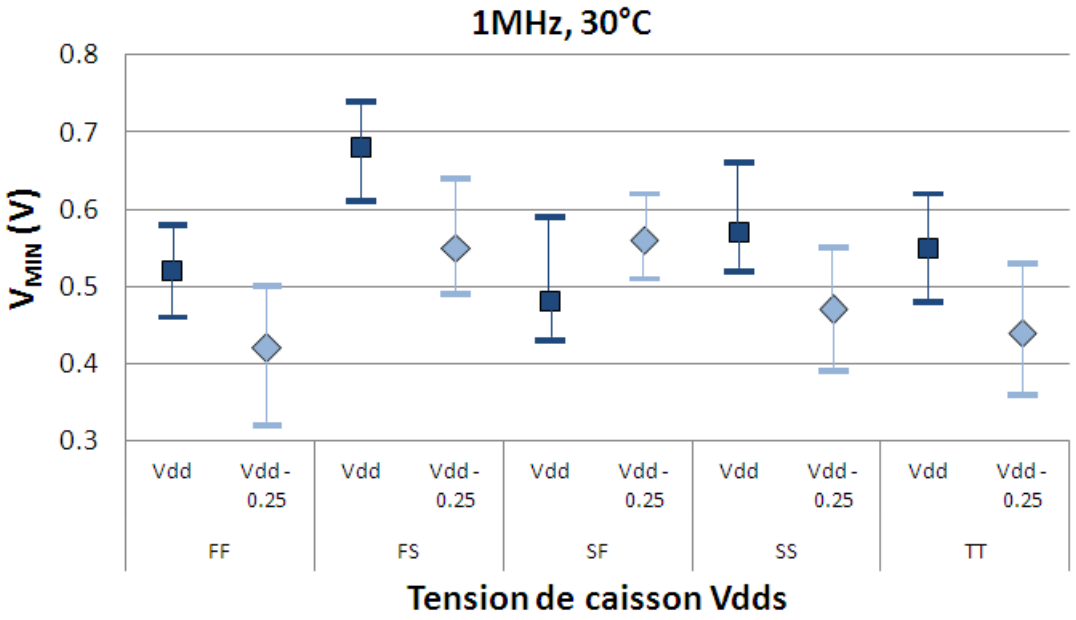

Figure 3.8 – Mesures sur silicium avec et sans Body-Biasing pour compenser les variations liées à la fabrication.

Cette solution peut également être utilisée pour compenser les effets de variation liés à la température (Figure 3.9). Une nouvelle fois, la tension minimale peut être réduite de 100mV. On observe également une importante amélioration du rendement du circuit à faible tension. Le Body-Biasing garanti un rendement de 93% à 0,5V pour les circuits ayant un centrage typique, au lieu de 30%.

Consommation dynamique

Les circuits ne bénéficient pas d'alimentations dédiées : les rails d'alimentation sont partagés avec les différentes contributions présentes sur le circuit de test. Par conséquent, la consommation statique mesurée sur les rails est globale et ne peut être dissociée. Il est cependant possible de mesurer la consommation dynamique en effectuant deux

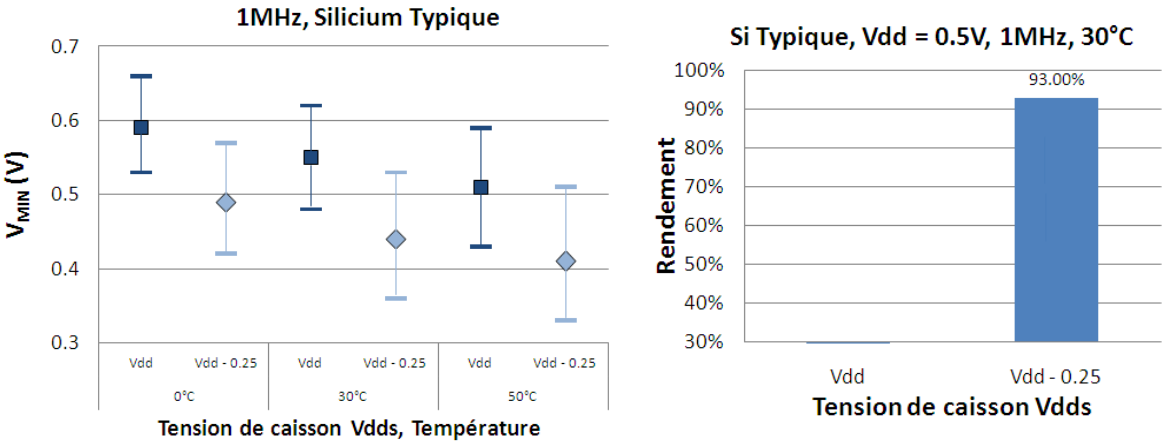

Figure 3.9 – Mesures sur silicium avec et sans Body-Biasing pour compenser les variations liées à la température (gauche), et observation de l'impact sur le rendement (droite).

mesures : une mesure sans activité, contenant l'ensemble des courants de fuite du circuit, suivie d'une mesure ou le BCH est en activité. Il suffit alors de soustraire les deux courants. Cette méthode n'est applicable qu'en l'existence d'un courant dynamique supérieur à la précision de l'appareil de mesure.

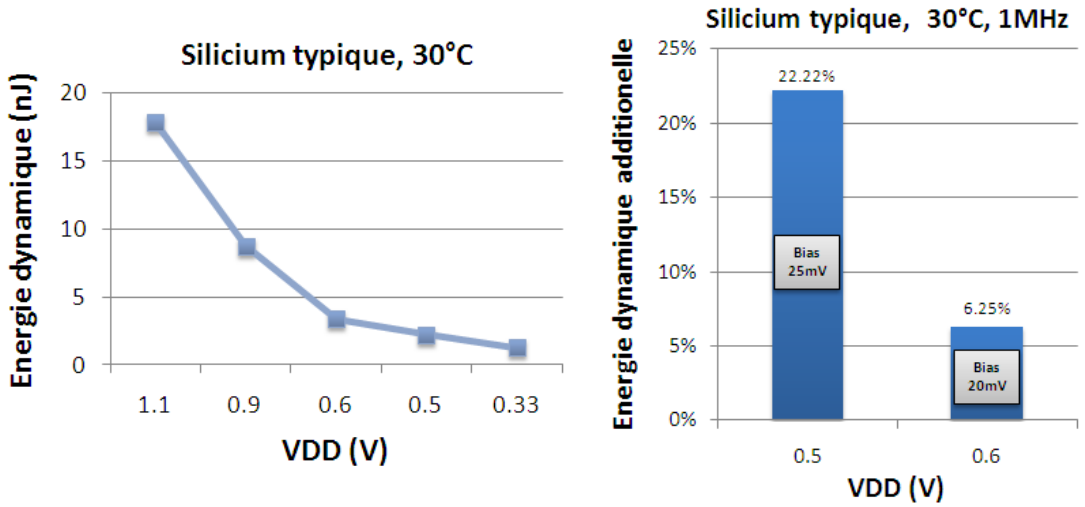

Figure 3.10 – Mesures sur silicium de l'énergie dynamique (gauche) et de l'impact du Body-Biasing (droite).

On mesure l'énergie dynamique sur une puce fonctionnelle jusqu'à 330mV (Figure 3.10). On observe une diminution de cette énergie d'un facteur x5 à 0,6V et jusqu'à x14 à 0,33V. Ces rapports sont cohérents avec ceux observés lors de l'étude sur les cellules réalisées au Chapitre 2. Concernant l'utilisation du Body-Biasing pour compenser les effets liés à la variabilité des procédés de fabrication et de la température,

on note une augmentation de 22% de la consommation à 0,5V pour un caisson alimenté à Vdd-25mV, et 6% à 0,6V pour un caisson alimenté à Vdd-20mV.

3 Analyses et adaptations

Ces résultats permettent de constater deux limitations principales : le plancher de V_{MIN} situé autour de 0,55V, soit 200mV plus élevé que les 0,35V souhaité, et la variabilité de V_{MIN} en fonction de la température et des procédés de fabrication. On a observé que le Body-Biasing apporte une réponse pour le traitement des effets liés à la variabilité, ce qui est en accord avec les conclusions du chapitre bibliographique. Il permet également de baisser la tension V_{MIN}, dans le cas d'un circuit alimenté à une tension fixe. Pour les circuits fonctionnant sur une gamme de tension, l'intérêt n'existe que si $E_{VMIN,Body-Biasing}$ est inférieure à la consommation énergétique E_{VMIN} de départ.

La limitation observée sur le plancher de V_{MIN} est critique puisqu'elle s'oppose à une prédictibilité des résultats requise pour un passage en production. Les circuits développés pour fonctionner à 0,35V doivent afficher un rendement supérieur à 80% à cette tension. Pour atteindre cet objectif, les différents aspects du flot de conception sont analysés en regard des mesures sur silicium évoquées précédemment, en mettant en lumière les adaptations permettant d'y parvenir.

Notons cependant que l'approche consistant en la définition de points de caractérisation aux extremums de la plage de tension parait valide en raison de la monotonicité observée sur les différents résultats ; les tensions inférieures au V_{MIN} mesuré sur chaque puce ne rétablissent pas la fonctionnalité. Le fonctionnement des cellules n'est également pas remis en cause, leur fonctionnalité sur silicium ayant été démontrée au chapitre précédent.

3.1 Alignement des modèles

Les oscillateurs du Chapitre 2 ont été reproduit au côté du BCH. L'alignement des modèles avec les mesures sur silicium est reporté sur la Figure 3.11. On observe une dégradation croissante lorsque la température diminue. On note également que si l'évolution de la fréquence en fonction de la tension d'alimentation est monotone sur les simulations comme sur les mesures, le rapport des deux ne l'est pas en dessous de 0,5V. Cependant, le comportement du silicium reste borné par les

points de caractérisation BC et WC.

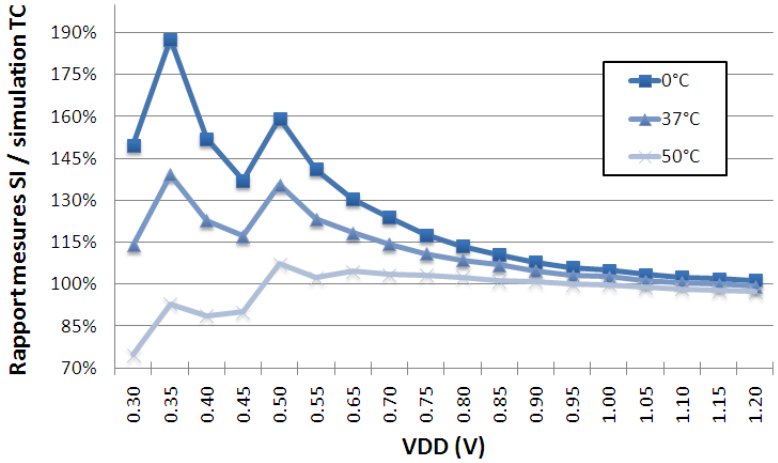

Figure 3.11 – Alignement des oscillateurs avec le modèle en fonction de la température.

Ainsi, le comportement du silicium est borné par les points de caractérisation, et il affiche une monotonicité fréquentielle. Ces observations mettent en lumière l'hypothèse d'une homothétie dissociée de la vitesse des chemins de données et du chemin d'horloge à très faible tension, liée à la composition de ces chemins, et amplifiée à mesure que la température décroit. Ce constat est conforté par l'augmentation de la variabilité lorsque la température décroit à faible tension, et est cohérent avec la variation du plancher de tension V_{MIN} en fonction de la température affichée précédemment sur la Figure 3.7.

3.2 Les marges temporelles

L'indépendance fréquentielle de V_{MIN} observée lors des mesures est caractéristique d'une violation de T_{HOLD} : à cette tension, le chemin de donnée commute plus rapidement que le front d'horloge, et cet écart de vitesse reste valide pour toute fréquence de l'horloge. Cette violation peut être associée à l'homothétie dissociée évoquée précédemment. Ces observations conduisent à la modification des marges utilisées lors de l'analyse temporelle du circuit.

Description de l'analyse temporelle nominale

Au cours du flot de conception, les opérations de vérification des contraintes temporelles sont effectuées sur les différents points de caractérisation. Usuellement, les points de caractérisations suivants sont utilisés :

- **WC :** Pour vérifier la non-violation des temps T_{SETUP}.
- **BC :** Pour vérifier la non-violation des temps T_{HOLD}.

Le point de caractérisation Worst Case réduit la vitesse du circuit. Les temps de traversée étant plus longs, la fréquence de l'horloge est réduite en conséquence. La fréquence maximale du circuit est obtenue durant cette étape.
Le point de caractérisation Best Case augmente la vitesse du circuit. La fréquence de l'horloge étant fixe, seuls les chemins de données sont accélérés, ce qui permet de remonter aux violations de T_{HOLD}. L'objectif est d'émuler les phénomènes de variation post-fabrication pour garantir le rendement.

Ces opérations de vérification utilisent des marges en addition des délais extraits et des contraintes simulées des Flip-Flops. Ces marges permettent d'émuler les variations aléatoires autour des points de caractérisation :

- **Temps d'incertitude :** Marge émulant une variation globale sur le chemin logique.
- **Facteur homothétique :** Marge émulant une variation locale sur le chemin logique.

L'incertitude, identifiée par le temps T_{INC}, permet d'augmenter virtuellement les temps T_{HOLD} et T_{SETUP} pour anticiper une variation du délai du chemin logique. Par conséquent les Flip-Flops nécessiteront que l'information soit présentée un temps T_{INC} plus tôt, et qu'elle soit maintenue un temps T_{INC} plus longtemps. Usuellement, T_{INC} est égal au temps de traversée d'une cellule standard tel qu'un inverseur.

Le facteur homothétique, identifié par le facteur X_H exprimé en %, permet de modifier virtuellement le temps d'arrivée de la donnée et le temps d'arrivée de l'horloge. Pour la vérification des temps T_{SETUP}, l'outil applique un facteur $+X_H$ sur la donnée, ce qui signifie qu'elle arrivera en retard, tout en appliquant un facteur $-X_H$ sur l'horloge, ce qui signifie que le front arrivera plus tôt (Figure 3.12). On se place ainsi dans le pire cas pour les violations de T_{SETUP}.
La vérification des temps T_{HOLD} est réalisée en appliquant les facteurs opposés : $-X_H$ sur la donnée, et $+X_H$ sur l'horloge. Dans ce cas on se place dans le pire cas pour les violations de T_{HOLD}. Usuellement, le facteur X_H correspond à l'enveloppe des variations autour du point de caractérisation.
Le réglage nominal dans les technologies 40nm est le suivant :

- T_{INC} = 20ps.
- X_H = 18%.

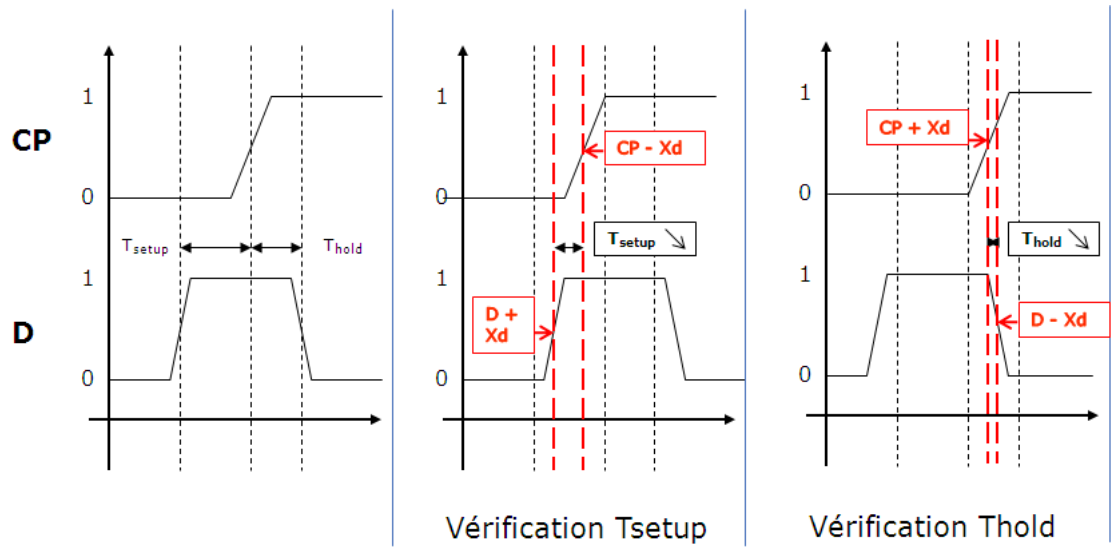

Figure 3.12 – Description du principe du facteur homothétique pour la vérification des contraintes.

Adaptation des marges à très faible tension

A très faible tension et au centrage typique, le temps de traversée d'un inverseur est de 150ns. Il est ainsi recommandé d'utiliser une valeur T_{INC} = 150ns. L'incertitude pouvant être définie de manière spécifique pour chacun des points de caractérisation, cette adaptation n'a aucune incidence sur la vérification temporelle à tension nominale. Lorsque cette marge est redéfinie, l'outil de vérification des contraintes met en lumière de nouvelles violations temporelles, ce qui confirme l'importance de la définition du temps d'incertitude.

Pour étudier le facteur X_H nécessaire à très faible tension, les mesures sur silicium des oscillateurs ont été reprises. La Figure 3.13 présente l'écart entre les simulations BC, WC, et les mesures sur silicium avec la simulation TC. Les mesures sur silicium typique dévient très peu de la simulation TC. On note cependant une importante dissymétrie entre l'écart TC-WC et l'écart TC-BC en dessous de 0,6V, qui traduit une déviation du réglage des points de caractérisation. Ces réglages étant figés dans le modèle, le champ d'action sur cette limitation est restreint. Toutefois, on vérifie dans le Tableau 3.1 que les points de caractérisation encadrent la distribution statistique à 3σ.

On note sur la Figure 3.13 que l'écart des points de caractérisation, soit le facteur X_H, est compris entre 70% et 350% à très faible tension, et autour de 35% à tension nominale. La distribution, telle que reportée dans le Tableau 3.1, conclut sur un facteur X_H de près de 24% dans le pire des cas.

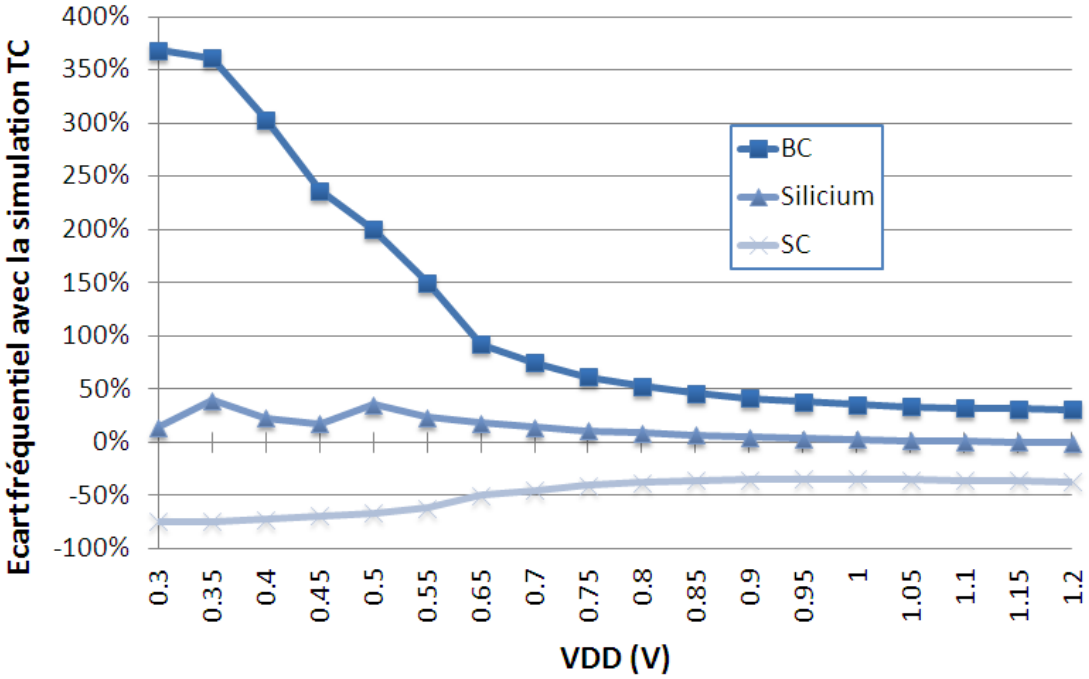

Figure 3.13 – Écart fréquentiel de la simulation BC et WC et des mesures silicium avec la simulation TC.

1000 Tirages	0°C	25°C	50°C
Distribution (ns)	1830 (Max)	626 (Median)	283 (Min)
Caractérisation (ns)	6995	600	70
μ (ns)	1390	629	330
3σ (ns)	339	133	60
XH	**24%**	**21%**	**18%**

Table 3.1 – Simulation Monte-Carlo d'une structure qFO4 et comparaison avec les points de caractérisation.

On définit alors quatre scénarios :

- Utiliser un facteur basé sur les points de caractérisation, soit $\mathbf{X_H = 35\%}$ à tension nominale, et $\mathbf{X_H = 350\%}$ à très faible tension.
- Utiliser un facteur basé sur la simulation statistique, soit $\mathbf{X_H = 7\%}$ à tension nominale, et $\mathbf{X_H = 24\%}$ à très faible tension.
- Utiliser le facteur usuel nominal recalculé, basé sur les points de caractérisation, soit $\mathbf{X_H = 18\%}$ à tension nominale, et $\mathbf{X_H = 118\%}$ à très faible tension.
- Utiliser le facteur usuel nominal recalculé, basé sur la simulation statistique, soit $\mathbf{X_H = 18\%}$ à tension nominale, et $\mathbf{X_H = 62\%}$ à très faible tension.

Les deux derniers sont obtenus en multipliant le facteur homothé-

tique par le rapport entre la variabilité à tension nominale et la variabilité à très faible tension.

Pour déterminer le choix à appliquer, la vérification temporelle du circuit BCH est relancée en modifiant le facteur homothétique. L'objectif est de reproduire les résultats observés sur silicium, c'est-à-dire observer des violations de T_{HOLD} à très faible tension pour une température de 0°C. En se positionnant dans le point de caractérisation WC, un facteur $X_H > 50\%$ est nécessaire pour voir apparaitre des violations. L'utilisation d'un facteur fixé à X_H=25% en accord avec les simulations Monte-Carlo est insuffisant : l'utilisation du ratio de 62% est alors recommandé.

On en déduit deux approches différentes :

- La définition d'un facteur différent pour les points de caractérisation à tension nominale et les points de caractérisation à très faible tension.
- La définition d'un facteur unique.

La première solution implique la réalisation de la vérification temporelle en deux passes, pour chaque tension. La correction de la violation des temps T_{HOLD} à très faible tension peut toutefois conduire à une altération indirecte des performances à tension nominale.
La seconde solution permet de faire l'économie de la seconde passe. Cependant elle introduit une altération directe des résultats obtenus à tension nominale, tout en introduisant un risque de dysfonctionnement à cette tension par la modification d'un réglage nominal validé par l'industrie.

Une autre approche consiste en l'utilisation d'un facteur homothétique nul, remplacé par une vérification aux points de caractérisation dissymétriques SF et FS. L'objectif est d'offrir une plus large couverture des variations à l'image de la Figure 2.2 présentée au Chapitre 2. Cela nécessite cependant la caractérisation de l'ensemble des cellules sur ces deux points.

L'utilisation d'une vérification en deux passes est recommandée en accord avec le principe de compatibilité et de qualité industrielle souhaité dans le cadre de cette thèse.

3.3 Construction de l'arbre d'horloge

Les précédentes observations confirment la nécessité d'une distribution correcte de l'horloge aux cellules séquentielles pour garantir la

fonctionnalité des circuits synchrones à très faible tension. Cela passe par la construction d'un arbre d'horloge tolérant aux variations. Il existe des recommandations pour les circuits multi-domaines [107] :

- La désynchronisation des domaines.
- Le non-croisement des arbres d'horloge ayant des tensions différentes.
- L'utilisation de cellules dédiées.

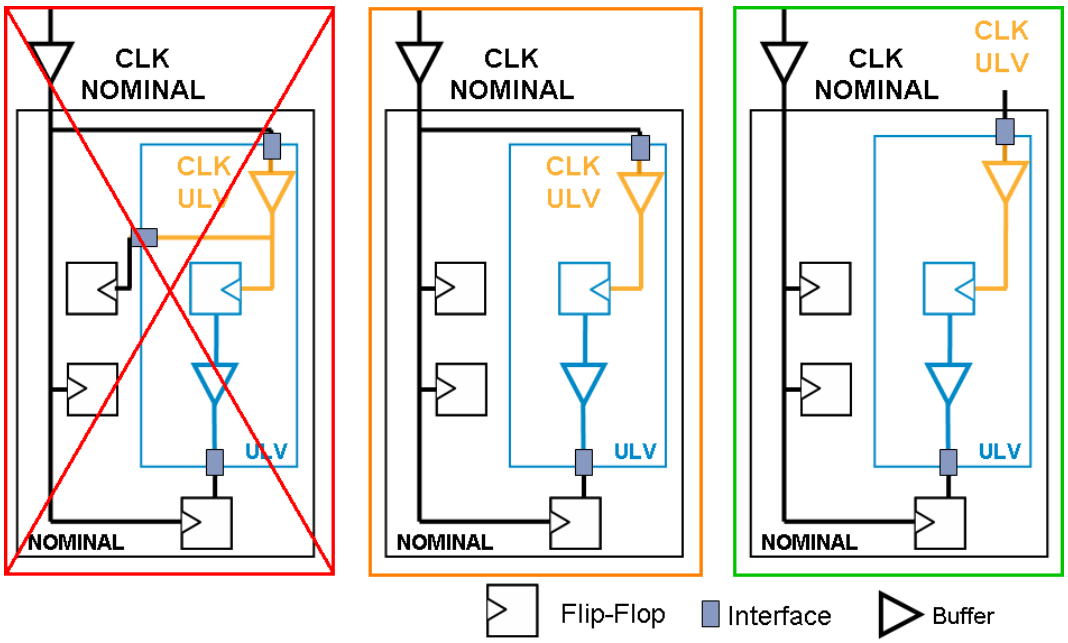

Figure 3.14 – Arbre d'horloge avec branches croisées (gauche), branches dédiées (centre), et arbres d'horloge dédiés (droite).

Le comportement de l'horloge et sa variabilité sont différents dans chaque domaine de tension. Si une même branche de l'arbre d'horloge traverse le domaine faible tension pour ressortir et commander une porte logique dans le domaine à tension nominale, la prédiction du comportement temporel de cette horloge est complexe : les effets de la variabilité des fils, des cellules, et cela pour chaque domaine doivent être pris en compte et croisés. Pour réduire la complexité, il existe deux approches synchrones (Figure 3.14) :

- **Une horloge unique dont les branches sont dédiées :** La même horloge commande l'ensemble des domaines de tension, cependant les branches sont dissociées en amont des domaines. L'incidence de la variabilité des branches d'horloge n'est observable qu'en sortie.
- **Une horloge par domaine :** Un arbre d'horloge est définit par domaine de tension. Cette approche permet de supprimer les corré-

lations en reportant la complexité sur la conception de l'arbre. Il est également possible de dissocier les fréquences de chaque horloge pour corriger les erreurs de synchronisation de données entre les domaines.

Des branches dédiées ont été utilisées dans les circuits mesurés précédemment. La solution consistant en l'utilisation d'une horloge par domaine parait comme permettant de combler les aléas induits par les phénomènes de variabilité. Cette dernière approche est par conséquent recommandée.

L'utilisation complémentaire de FIFO pour First-In-First-Out en langue anglaise, peut faciliter la synchronisation entre domaines. Ces structures permettent notamment de transmettre des données reçues à une fréquence F1 en données à fréquence F2. Il est également possible de traiter l'interaction des données entre domaines de manière asynchrone [108; 109]. Les blocs qui composent les circuits asynchrones communiquent par l'intermédiaire de zones tampons qui permettent de s'affranchir des effets de la variabilité, le principe étant d'attendre l'arrivée de la donnée pour poursuivre l'exécution des tâches. L'asynchronisme ne peut cependant pas être envisagé dans le cadre du flot de conception utilisé durant cette thèse.

L'arbre d'horloge est usuellement constitué d'inverseurs et de buffers. Les inverseurs permettent de propager $\overline{CLK}$ tandis que les buffers permettent de maintenir l'amplitude, le rapport cyclique, et les pentes du signal. Les cellules d'horloges développées et présentées dans le Chapitre 3 offrent une sortance plus grande ainsi qu'une sensibilité aux variations réduite par l'usage du dopage LVT, des grilles très larges et de la mise en parallèle systématique. Par conséquent, l'usage exclusif de ces cellules pour la construction de l'arbre d'horloge à très faible tension est recommandé.

3.4 L'interface nominal-très faible tension

Le circuit n'utilise pas de cellules d'interface spécifiques, à l'image des cellules dont la description a été faite au Chapitre 2. Ces cellules ont été développées à la suite de cette analyse.

Les mesures ont mis en évidence que les circuits SF offrent une tension V_{MIN} inférieure à celle obtenue sur les circuits FS. Cette limitation est cohérente avec l'analyse sur l'interface faite au Chapitre 2 concernant l'importante dissymétrie entre le courant passant du NMOS et celui du PMOS, entrainant une dégradation du rapport cyclique. Le centrage SF permet de rééquilibrer cette dissymétrie, tandis que le cen-

trage FS l'amplifie. De même, l'efficacité du Body-Biasing lorsque seul le courant des PMOS est augmenté confirme cette analyse.

Une autre hypothèse est une plus importante dégradation des performances du PMOS lorsque la tension est réduite. Cependant, cette hypothèse n'a pas été validée lors de l'étude sur les modèles réalisée au Chapitre 2.

L'utilisation de cellules d'interface et la caractérisation du comportement de ces cellules est fondamental.

3.5 Les effets de couplage

Le couplage correspond à la tension parasite générée par un fil parcouru par une transition rapide sur un autre fil à proximité, par effet capacitif (Figure 3.15). Ce phénomène peut être amplifié à très faible tension, puisque l'effet parasite d'un fil alimenté à tension nominale sur un fil parcouru par une donnée à très faible tension peut être suffisamment important pour faire basculer le niveau logique de celle-ci. Nommés "Crosstalk" en langue anglaise, ces effets de couplage sont détectés par l'outil de vérification.

Dans le cas du BCH, et par précaution, les fils parcourus par une transition rapide en provenance du domaine nominal, telle que l'horloge, ne survolent pas le domaine très faible tension. Ce résultat est obtenu en masquant le domaine très faible tension lors du routage des fils du domaine nominal.

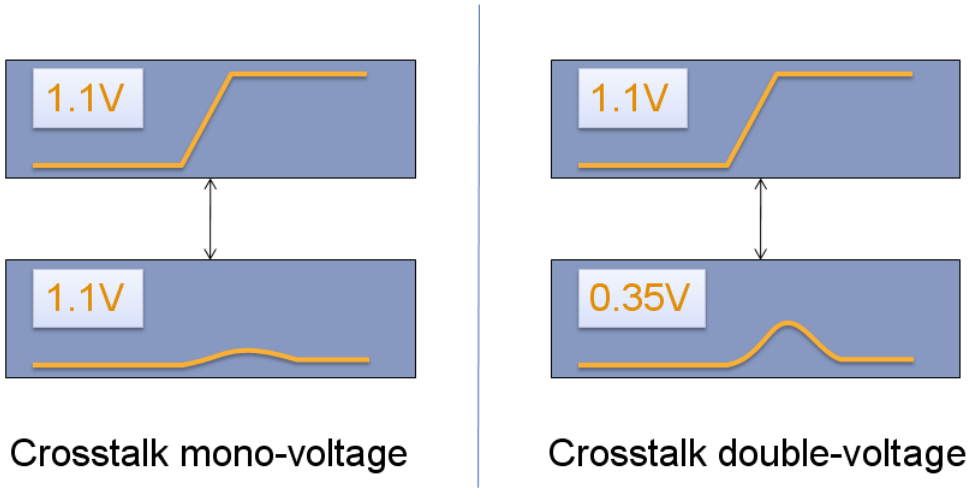

Figure 3.15 – Représentation du phénomène de Crosstalk.

4 Récapitulatif de la contribution

Les développements très faible tension suivants ont été mis en avant dans ce chapitre :

- La conception de circuits très faible tension utilisant le réglage nominal.
- La mesure sur silicium des circuits et la mise en lumière des limitations.
- L'étude de la validité des points de caractérisation.
- L'importance de la prédiction et de la tolérance face à la variabilité.
- L'adaptation des marges temporelles pour la très faible tension.
- L'importance de la construction de l'arbre d'horloge et de l'utilisation de cellules spécifiques.
- L'utilisation fondamentale de cellules d'interface.
- Le gain en énergie dynamique permis par la réduction de la tension d'alimentation.
- L'efficacité du Body-Biasing dans la compensation de la variabilité et dans l'amélioration du rendement.

Tout en maintenant une compatibilité complète avec le flot de conception usuel de l'industrie, ce chapitre met en évidence les points clés permettant de converger vers une conception de circuit à très faible tension offrant un rendement élevé. L'étape de vérification des contraintes temporelles s'étant révélée cruciale pour atteindre cet objectif, des adaptations sont proposées à partir d'une analyse des mesures sur silicium d'un circuit numérique. Ces résultats ont ensuite été insérés dans l'outil pour vérifier leur validité. Des recommandations sur les autres aspects du flot de conception sont proposées, et l'efficacité du Body-Biasing dans la compensation de la variabilité est vérifiée. Enfin, ce circuit confirme par des mesures de courant dynamique le gain énergétique permis par la réduction de la tension d'alimentation.

5 Publications scientifiques

L'étude sur le développement de circuit à très faible tension a donnée lieu à la publication de l'article suivant :

– **"A 40nm CMOS, 1.27nJ, 330mV, 600kHz, Bose Chaudhuri Hocquenghem 252 bits frame decoder"** [110]
Following the will to answer to the energy constrained applications requirements, an Ultra-Low Voltage (ULV) 40nm Bose - Chaudhuri - Hocquenghem (BCH) error-correcting circuit is presented. Mapped on

a ULV specific standard cells library, the circuit was designed following standard industrial implementation and verification flows. The BCH circuit runs at 0.330V, 600kHz frequency and needs 1.27nJ to decode a 252bits frame. With 14% of extra power compared to typical process, applying forward bias enables to compensate temperature and skewed process effects, regaining 150mV minimum operating voltage.

Chapitre 4

Les mémoires

1 Contexte

Les circuits numériques embarquent des éléments mémorisant permettant la réalisation d'opérations complexes. A titre d'exemple, les cellules mémoires servent à maintenir le résultat d'une opération pour son usage futur par un micro-contrôleur. Comme évoqué lors de l'étude bibliographique, le fonctionnement des cellules mémoires à très faible tension est délicat, à l'image des Flips-Flops : le dimensionnement de la boucle de mémorisation doit à la fois garantir la stabilité de l'information stockée dans la boucle, et la possibilité de modifier cette information.

Dans ce chapitre, une description générale des mémoires SRAM est faite, suivie d'une présentation de la métrique statique utilisée pour les caractériser. Une métrique dynamique est alors définie pour compléter l'étude. Plusieurs cellules mémoires sont ensuite comparées en faisant appel à ces différentes métriques. Le choix de cellule, composé des mémoires usuelles de l'industrie et des propositions les plus abouties de la littérature, permet de distinguer l'architecture offrant le meilleur comportement à très faible tension. Une cellule mémoire est alors conçue à la suite de ces observations, et est optimisée pour le fonctionnement à très faible tension, tout en maintenant la fonctionnalité à tension nominale. Ces étapes ont été automatisées par le développement d'un script permettant une plus grande productivité. Enfin, des solutions d'assistance sont développées pour renforcer le fonctionnement de la cellule mémoire à très faible tension.

Ces travaux sont intégrés dans un démonstrateur composé de quatre matrices de 32kb et fabriqué en technologie 65nm.

2 Caractéristiques des mémoires SRAM

Usuellement, les cellules SRAMs industrielles sont symétriques et composées de six transistors : quatre transistors composent la boucle de mémorisation, et deux transistors permettent d'ouvrir ou de fermer l'accès à cette boucle. Sur la Figure 4.1, les trois transistors de gauche forment la partie TRUE (T) tandis que les trois transistors de droite forment la partie FALSE (F). Par construction, la valeur du noeud interne BLTI est toujours égale à l'opposé de la valeur du noeud interne BLFI. Les transistors d'accès, ou passgates en langue anglaise, sont notés PG. Les PMOS permettant de tirer la valeur du noeud interne vers VDD (état logique 1), ou pull-up en langue anglaise, sont notés PU. Les NMOS permettant de tirer la valeur du noeud interne vers GND (état logique 0), ou pull-down en langue anglaise, sont notés PD. Les transistors d'accès sont reliés à la périphérie par les lignes de données, ou bitlines en langue anglaise, notées BL. Ils sont commandés par une ligne de mot, ou wordline en langue anglaise, noté WL.

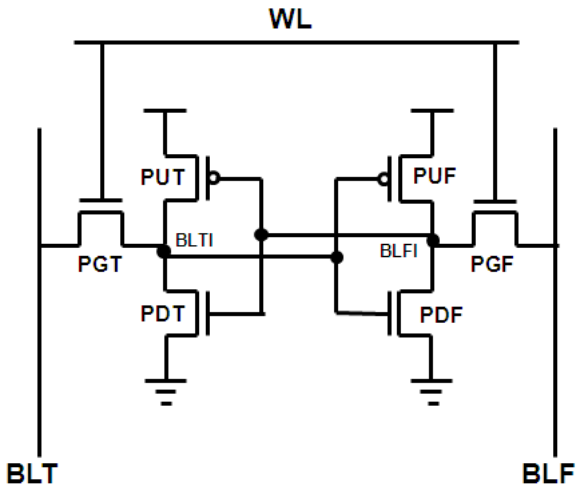

Figure 4.1 – Représentation schématique de la cellule SRAM à six transistors (SRAM 6T).

Les cellules sont disposées en lignes et colonnes. Les cellules sur la même ligne partagent le rail WL d'activation des passgates. Les cellules sur la même colonne partagent le même couple de bitlines. Cette configuration permet la réalisation de trois opérations (Figure 4.2) :

- Le maintien, ou HOLD.
- La lecture, ou READ.
- L' écriture, ou WRITE.

Pour réaliser l'opération HOLD, WL est à GND. PGT et PGF sont bloqués. Les inverseurs maintiennent la valeur des noeuds internes. Usuel-

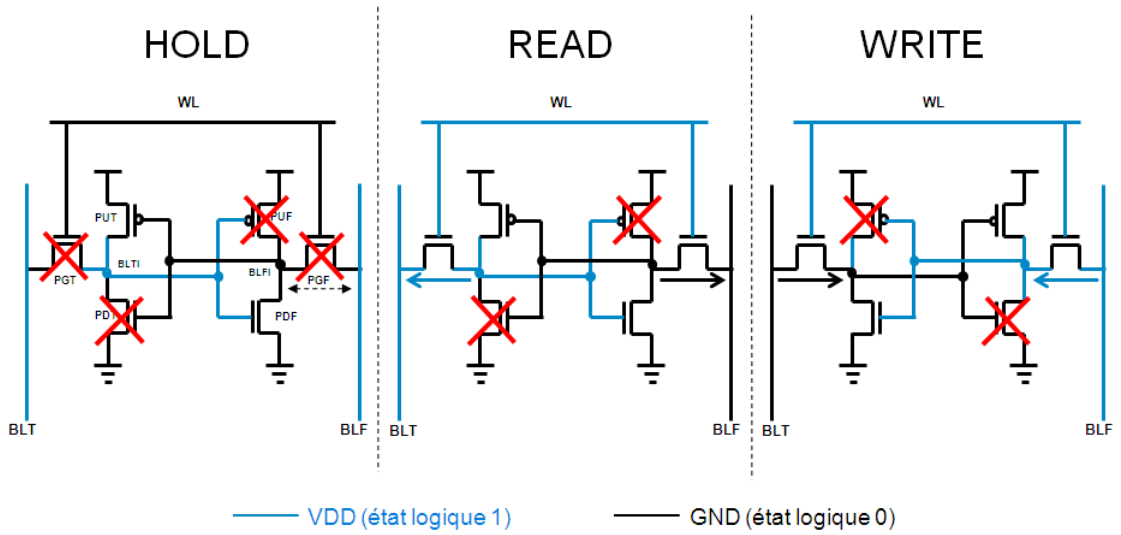

Figure 4.2 – Opérations de maintien (gauche), lecture (centre) et écriture (droite) d'une cellule SRAM 6T.

lement, les bitlines sont préchargées à VDD.
Pour réaliser l'opération READ, WL est à VDD. PGT et PGF sont passants. Les inverseurs maintiennent la valeur des noeuds internes et la communiquent sur les bitlines. Celles-ci étant préchargées à VDD, la bitline connectée au noeud interne possédant la valeur VDD reste stable, tandis que la bitline connectée au noeud interne possédant la valeur GND est tirée vers GND.
Pour réaliser l'opération WRITE, WL est à VDD. PGT et PGF sont passants. Les inverseurs maintiennent la valeur des noeuds internes, tandis que les bitlines sont alimentées en inverse du contenu. Le rapport de courant entre PG et le couple PU/PD doit permettre le basculement de la valeur stockée.

Les cellules voisines du point mémoire accédé sont d'une part affectées par les opérations de lecture et d'écriture, et peuvent d'autre part perturber ces opérations (Figure 4.3).
Les cellules placées sur la même ligne, bien qu'étant en HOLD, ont leurs passgates passantes. Le contenu de la cellule doit être maintenu malgré les perturbations imposées par les bitlines. De plus, une chute de tension est observée sur la bitline reliée au noeud interne possédant la valeur GND, ce qui peut être assimilé à une opération READ non souhaitée.
Les cellules non accédées et placées sur la même colonne, bien qu'étant en HOLD, sont affectées par la polarisation des bitlines lors du WRITE. Le contenue de ces cellules doit rester stable face aux fuites de courant à travers leurs passgates. De plus, lors du READ, les cellules non accédées de la colonne peuvent affecter le contenu lu. A titre d'exemple, dans une colonne de 128 cellules, une cellule est lue, et les autres sont

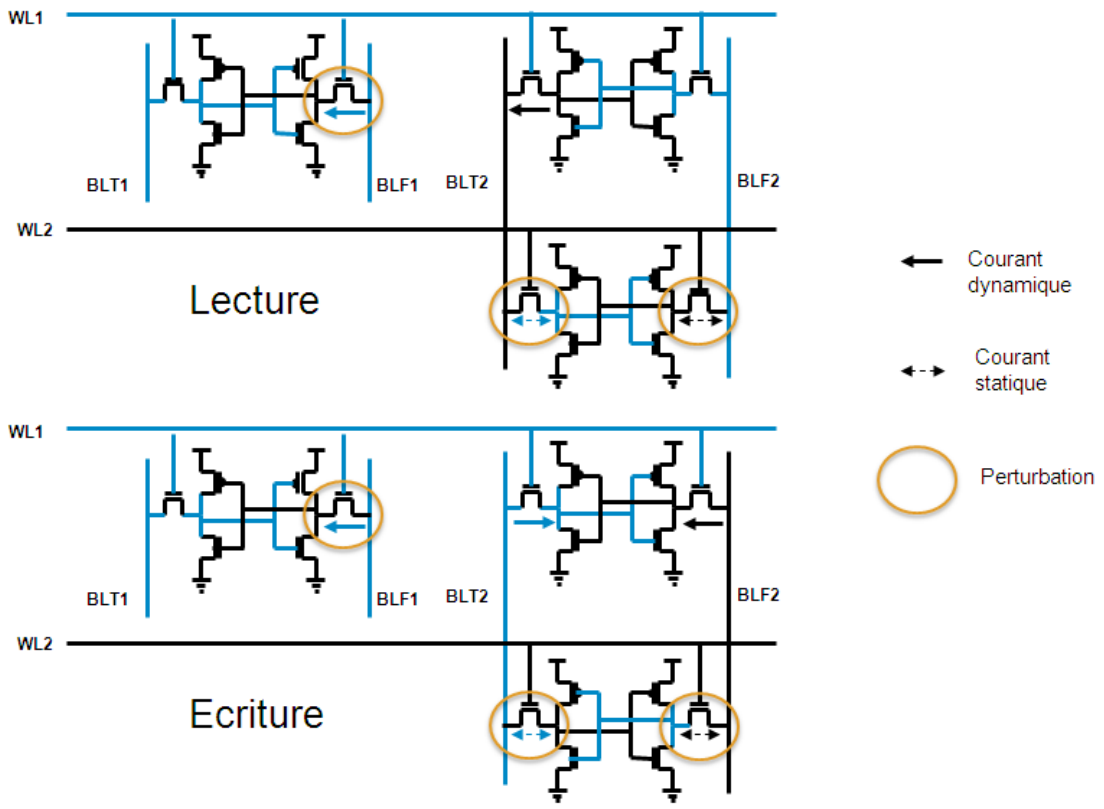

Figure 4.3 – Phénomènes de perturbation au voisinage de la cellule accédée.

en HOLD. Cependant, au pire cas, la cellule lue contient la valeur VDD, et l'ensemble des autres cellules contient la valeur GND. Ainsi, sur la bitline, on lit un courant dynamique VDD, mais également 127 courants statiques parasites GND. Pour lire correctement la valeur VDD, il est nécessaire que le courant dynamique soit au minimum supérieur à 128 fois le courant statique.

On distingue alors au niveau de la cellule les phénomènes de défaillance suivants :

- **En HOLD :** Le noeud interne est perturbé par la fuite des bitlines à travers les passgates.
- **En READ :** Le noeud interne connait une chute ou une hausse de tension lors de l'activation de passgates par les fuites des autres cellules sur les bitlines. De même, la perturbation des bitlines peut provoquer une erreur de lecture.
- **En WRITE :** Le noeud interne résiste au basculement imposé par la polarisation inverse des bitlines.

La capacité de la cellule SRAM à réaliser les opérations HOLD, READ, et WRITE, dépend de sa sensibilité à ces phénomènes. Cette sensibilité

est exprimée par une métrique spécifique.

2.1 Métrique Statique

La métrique statique dédiée aux cellules SRAMs consiste à extraire les marges intrinsèques de celle-ci :

- Sa capacité à maintenir la donnée interne lors du HOLD : c'est la HOLD Static Noise Margin (Hold SNM).
- Sa capacité à maintenir la donnée interne lors du READ : c'est la READ Static Noise Margin (Read SNM).
- Sa capacité à permettre l'écriture lors du WRITE : c'est la WRITE Margin (WM).
- Sa capacité à permettre la lecture correcte lors du READ : c'est le facteur N_{BL}.

SNM

La SNM consiste à mesurer en volt-offset la tension nécessaire sur le noeud interne pour provoquer un basculement **non souhaité** de la valeur stockée dans la boucle de mémorisation.

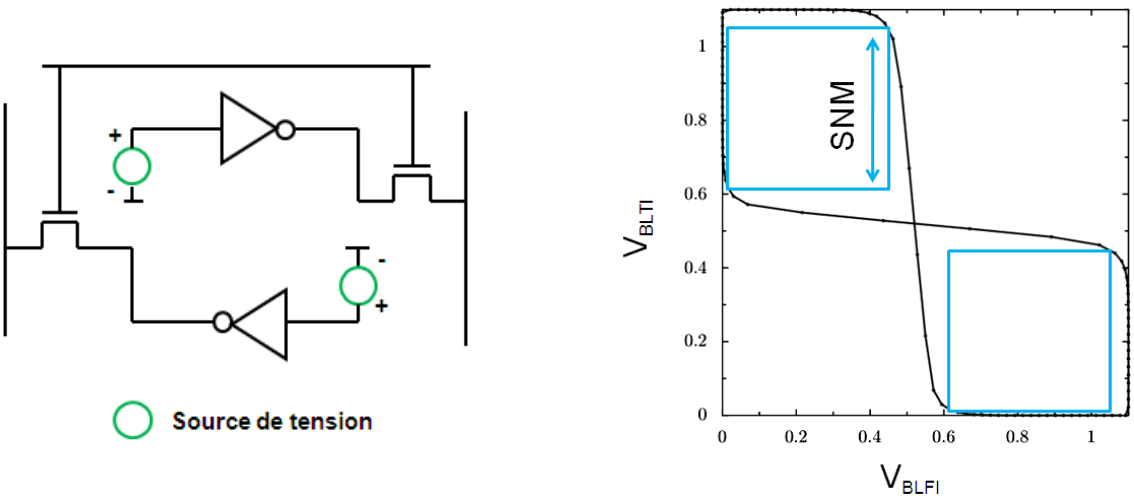

Figure 4.4 – Représentation schématique de la simulation (gauche) et de la figure en papillon (droite) permettant d'extraire la SNM.

Pour mesurer la Hold SNM, la cellule est placée en mode HOLD. Pour mesurer la Read SNM, la cellule est placée en mode READ. Les bitlines sont préchargées à VDD. Les noeuds internes possèdent une valeur logique initiale arbitraire, dans notre cas BLTI = 1 et BLFI = 0. On place une source de tension sur chaque noeud interne que l'on fait varier de

VDD à GND pour BLTI, et de GND à VDD pour BLFI (Figure 4.4). la SNM correspond à la tension de basculement des noeuds. Plus cette tension est élevée, plus la cellule sera stable. La Hold SNM peut cependant se substituer par l'unique analyse de la Read SNM, l'effet du courant dynamique étant supérieur à celui du courant statique.

Il existe une représentation graphique de la SNM, appelée figure en papillon. Elle consiste à tracer la tension de sortie de chaque inverseur de la boucle de mémorisation en fonction de son entrée, et de superposer les deux courbes, ce qui permet d'obtenir une figure en papillon. La longueur du coté du plus grand carré que l'on est capable de dessiner dans chaque boucle correspond à la valeur de la SNM (Figure 4.4).

Cette marge peut être augmentée par la modification du β-ratio de la cellule mémoire. Le β-ratio correspond au rapport de courant entre PD et PG. Plus ce rapport est grand, plus il sera difficile pour les pass-gates d'imposer une perturbation aux noeuds internes. Usuellement, ce rapport est compris entre 1 et 3.

La SNM dépend du dimensionnement, lui même sensible à la variabilité. Par conséquent, la SNM est vérifiée par simulation Monte-Carlo dans le pire cas en termes de point de caractérisation, à savoir le FNSP, puisqu'il accélère les PG et affaiblit les PU de la boucle de mémorisation. La réalisation de 1000 tirages Monte-Carlo autour du pire cas permet une estimation à 6σ de la distribution de la SNM (Figure 4.5).

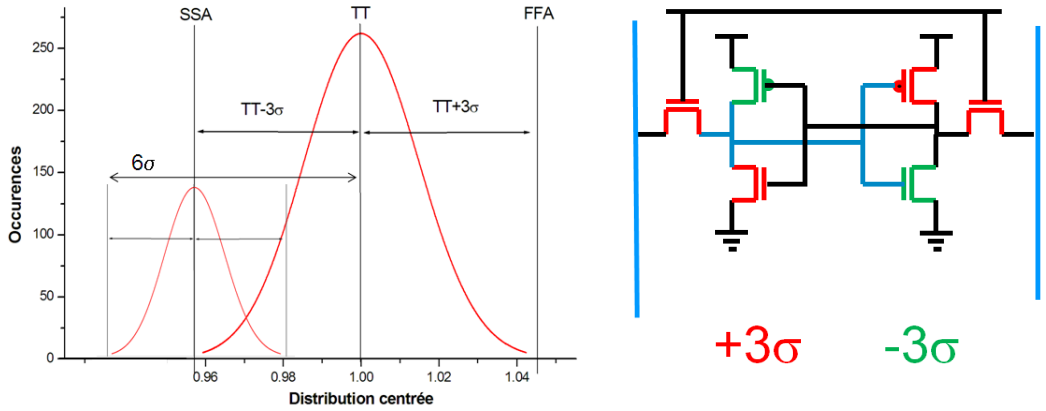

Figure 4.5 – Simulation Monte-Carlo de la SNM à 6σ (gauche), et application de la méthodologie paramétrique (droite).

Il est possible de reproduire le pire cas SNM, à savoir l'extrémité de la distribution à 6σ, sans avoir à réaliser les 1000 tirages Monte-Carlo. En se plaçant dans le point de caractérisation FNSP, on paramétrise

manuellement le σ de la tension de seuil σ_{VT} et de la mobilité σ_{μ} de chaque transistor de la cellule. Le σ est fixé à +3 pour les PG, et à -3 pour les PU et PD (Figure 4.5), +3 correspondant à une augmentation du courant.

Pour reprendre l'exemple introduit en début de chapitre, un défaut de stabilité peut entrainer une perte de la donnée stockée dans la cellule. Cette perte engendre des erreurs sur les opérations d'un microcontrôleur, menant à des résultats erronés en sortie. Les cellules mémoires étant communément utilisées en très grand nombre, le risque d'apparition d'erreurs est grand. Par conséquent, le facteur de mérite minimum de la SNM, correspondant au rapport entre la moyenne de la distribution de la SNM et son écart-type σ, est fixé à 7 par l'industrie.

WM

La Write Margin consiste à mesurer la tension nécessaire sur le noeud interne pour provoquer un basculement **voulu** de la valeur stockée dans la boucle de mémorisation.

Pour mesurer la WM, la cellule est placée en mode WRITE. Les bitlines sont préchargées à VDD. Les noeuds internes possèdent une valeur initiale arbitraire, dans notre cas BLTI = 1 et BLFI = 0. On place une source de tension sur BLT que l'on fait varier de VDD à GND, en maintenant BLF à VDD. La WM correspond à la tension de basculement des noeuds internes (Figure 4.6). Plus cette tension sera haute, moins il faudra d'effort pour écrire dans la cellule.

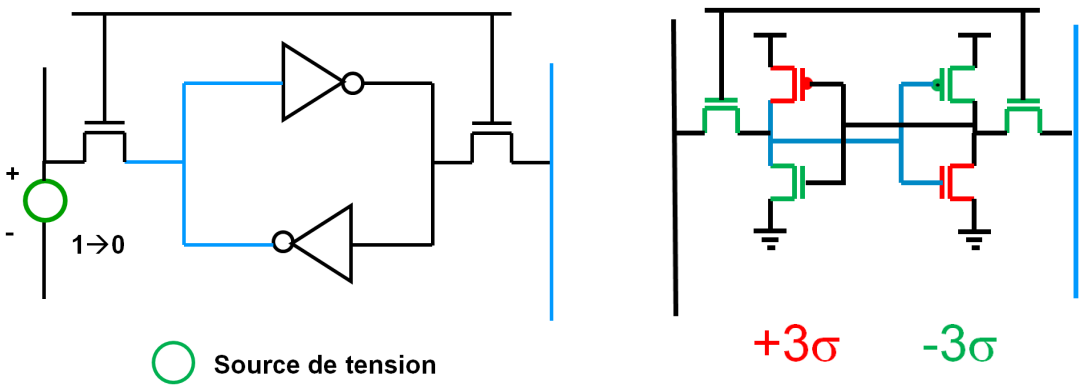

Figure 4.6 – Représentation schématique de la simulation permettant d'extraire la WM (gauche), et application de la méthodologie paramétrique (droite).

Cette tension peut être augmentée par la modification de l'α-ratio de la cellule mémoire. L'α-ratio correspond au rapport de courant entre

PG et PU. Plus ce rapport est grand, plus les passgates auront la capacité d'imposer la valeur des bitlines aux noeuds internes. Usuellement, ce rapport est compris entre 1 et 3.

La WM est vérifiée par simulation Monte-Carlo dans le point de caractérisation le plus défavorable, à savoir le SNFP, puisqu'il affaiblit les PG et accélère les PU de la boucle de mémorisation. La réalisation de 1000 tirages Monte-carlo autour du pire cas permet une estimation à 6σ de la distribution de la WM. Pour obtenir le pire cas de manière paramétrique, on utilise le point de caractérisation SNFP, et on fixe le σ à -3 pour les PG, et à +3 pour les PU et PD (Figure 4.6). Le facteur de mérite minimum défini par l'industrie est de 7.

On observe la symétrie entre SNM et WM. Il s'avère que plus la cellule sera dimensionnée pour être stable, moins elle permettra l'écriture. Ce constat est un fait majeur du submicronique, suite à l'avènement des phénomènes de variabilité avec les technologies 90nm. Le fonctionnement de la cellule correspond donc à un compromis entre stabilité et capacité d'écriture.

Rapport des courants statiques et dynamiques

En supplément de sa stabilité et de sa capacité à être écrite, la cellule doit garantir une lecture correcte de son contenu. Son courant de lecture, correspondant au courant dynamique à travers PG lors de la lecture, doit être supérieur au cumul des courants de fuite à travers la passgate de l'ensemble des cellules de la même colonne. On nomme le facteur N_{BL} le produit α·Lignes, Lignes étant le nombre de lignes, et α étant communément compris entre 5 et 10 (Figure 4.7). Ce facteur garantit la supériorité du courant dynamique sur le courant de fuite.

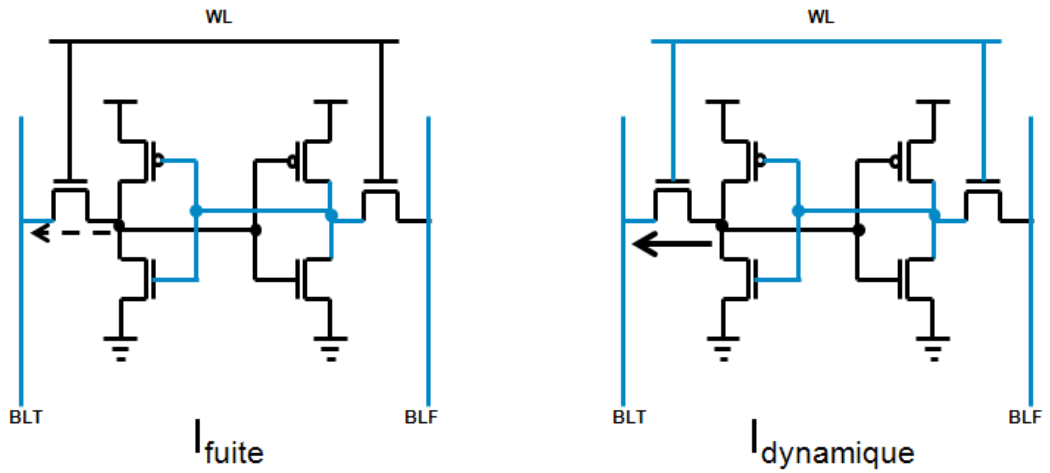

Figure 4.7 – Représentation schématique de la simulation permettant d'extraire le courant de fuite (gauche), et le courant dynamique (droite).

Le facteur N_{BL} est vérifié par une simulation Monte-Carlo dans le point de caractérisation FF, où le courant de fuite à travers les passgates est le plus grand. Pour la simulation du pire cas paramétrique, on se place au point de caractérisation FF, et on fixe le σ à +3 pour les transistors PG.

En augmentant le W des passgates, on améliore leur courant dynamique. Cependant, leur courant de fuite augmente également, de même que la capacité de la bitline. La bitline étant plus capacitive, son temps de charge est plus long, et l'opération de lecture est ralentie. La solution retenue usuellement est la réalisation d'un transistor PG de taille minimale, au détriment de la WM.

2.2 Métrique Dynamique

La simulation statique correspond à une mesure ponctuelle après stabilisation des tensions. Cependant, le temps de cette stabilisation peut être extrêmement long, ce qui est optimiste en regard du fonctionnement réel. La simulation dynamique permet de caractériser la cellule mémoire dans le cadre de la contrainte temporelle du temps de cycle. Elle permet ainsi de vérifier la capacité de la cellule à réaliser les opérations de lecture et d'écriture dans le temps imparti.

On construit alors le stimulus dynamique réalisant les opérations suivantes (Figure 4.8) :

- Le cycle d'horloge est fixé à 100kHz, correspondant à l'horloge pire cas souhaitée.
- La simulation se déroule sur 8 cycles d'horloge : HOLD, WRITE, HOLD, READ, HOLD, WRITE, HOLD, READ.
- Une seule cellule est accédée parmi plusieurs sur une colonne.
- Les cellules d'une même colonne sont initialisées avec le même contenu. Après la première écriture, la cellule accédée contient donc une information en opposition avec celle stockée dans toutes les autres.
- Le temps d'écriture TWRITE correspond au temps mis par la boucle de mémorisation pour basculer.
- Le temps de lecture TREAD correspond au temps mis par la cellule pour tirer la bitline vers GND.
- L'énergie dynamique par opération est extraite.
- Le courant de fuite en HOLD est extrait.

A l'image de la métrique statique, des simulations statistiques Monte-Carlo seront exécutées pour évaluer l'évolution de ces critères en réponse à la variabilité.

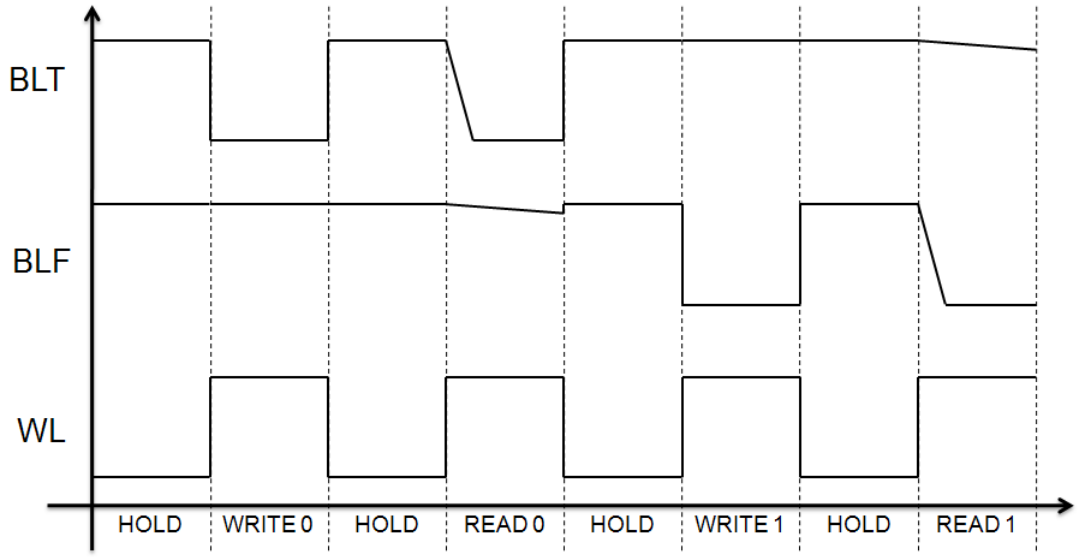

Figure 4.8 – Méthodologie de simulation dynamique des cellules mémoires.

3 Construction de l'environnement de simulation

Le comportement d'une cellule mémoire dépend donc de son environnement. Cependant, une matrice de cellules mémoire permettant une évaluation de ses performances, implique la simulation d'un grand nombre de transistors, ce qui conduit à des temps de simulation importants, notamment pour l'évaluation statistique. Il existe dans l'industrie des méthodologies d'émulation de chemin critique, à l'image du qFO4, adaptées pour les cellules mémoires. Ces chemins sont usuellement construits autour du point mémoire SRAM 6T.

En raison d'une opportunité de fabrication sur silicium en 65nm, les développements suivants seront réalisés dans cette technologie.

3.1 Chemin critique universel

Afin de rendre possible la simulation de différents points mémoires et leur comparaison, un chemin critique universel a été construit. Ce chemin permet d'émuler un point mémoire placé dans le coin supérieur droit d'une matrice. Ce positionnement correspond au pire cas dans la mesure où le temps de propagation sur WL et BL est le plus long.

La construction du chemin critique a été automatisée par l'écriture d'un script. Il a été conçu de telle sorte que la méthodologie soit compatible avec tout type de cellule mémoire et que la taille de la ma-

trice émulée soit paramétrable. Intégrant la génération des chemins critiques, la caractérisation des cellules mémoires en statique comme en dynamique tout en permettant la simulation statistique, et l'extraction de l'ensemble des résultats, il permet de garantir une productivité conforme aux exigences industrielles.

La construction du chemin critique suit les étapes suivantes :

- **Création des WL et BL globales :** Les entrées / sorties WLR, BLTR, et BLFR sont créées.
- **Création des lignes :** WLR est le point de départ d'un réseau RC qui émule le chemin parcouru par la wordline entre WLR et la cellule simulée. Le réseau RC est composé de l'alternance d'une cellule mémoire virtuelle destinée à émuler la capacité d'une cellule, et d'une résistance R correspondant à la connexion métallique entre chaque cellule, répétés suivant le nombre de colonnes de la matrice. Les bitlines des cellules virtuelles sont toujours connectées à VDD, afin de les placer en précharge.
- **Création des colonnes :** BLTR et BLFR sont les points de départ d'un réseau RC qui émule le chemin parcouru par les bitlines entre BLTR/BLFR et la cellule simulée. Le réseau RC est composé de l'alternance d'une cellule mémoire virtuelle destinée à émuler la capacité d'une cellule, et d'une résistance R correspondant à la connexion métallique entre chaque cellule virtuelle, répété suivant le nombre de lignes de la matrice. Les wordlines des cellules sont toujours connectées à GND, afin de les placer en HOLD.

Le chemin critique obtenu est schématisé sur la Figure 4.9. Comme évoqué précédemment, afin de s'adapter à la taille de la mémoire visée, le nombre de cellules en ligne et en colonne est paramétrable. De même, pour s'adapter à l'architecture de la cellule mémoire, le nombre de BL et de WL est paramétrable.

3.2 Stimuli et points de caractérisation

Les stimuli suivant sont définis pour la simulation du chemin critique :

- **SNM :** Extraction de la READ SNM par simulation DC.
- **WM :** Extraction de la WM par simulation DC.
- **DYNAMIC :** Extraction de TREAD, TWRITE, EREAD, EWRITE, et du courant de fuite par simulation transitoire.

Les points de caractérisation suivants sont définis pour la simulation du chemin critique :

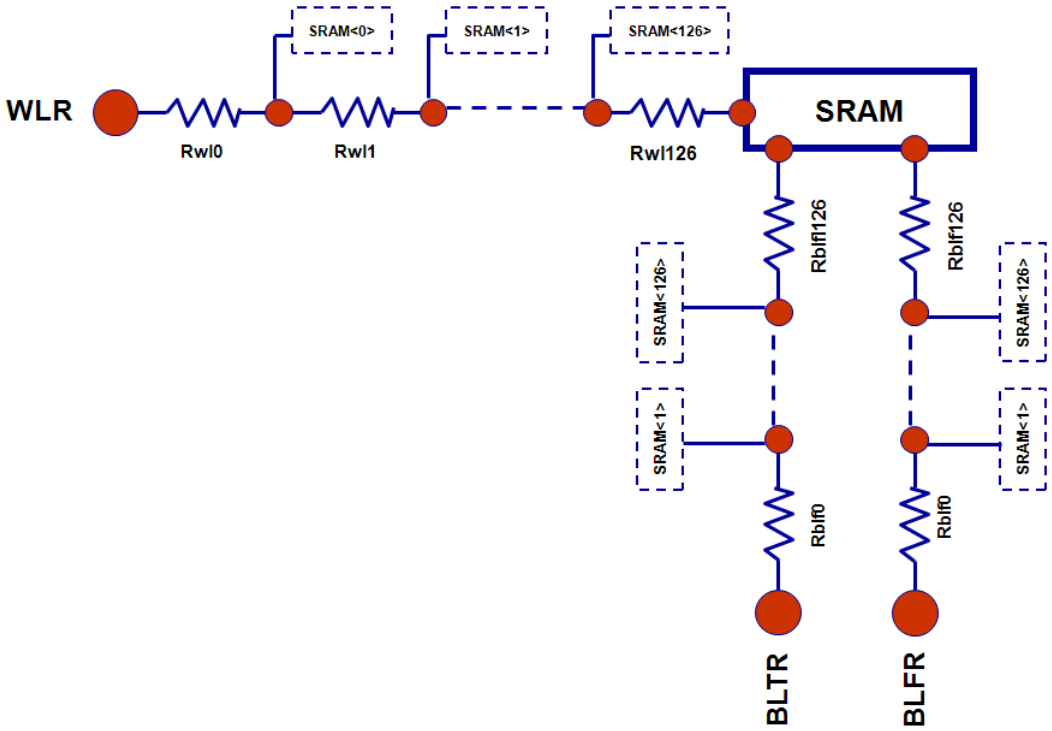

Figure 4.9 – Représentation schématique du chemin critique universel.

- **TT_0.35V_25C & TT_1.2V_25C :** Simulation typique à très faible tension, 25°C, et tension nominale, 25°C.
- **SS_0.35V_0C & SS_1.1V_125C :** Simulation -3σ à très faible tension, 0°C, et tension nominale, 125°C.
- **FF_0.35V_50C & FF_1.3V_m40C :** Simulation +3σ à très faible tension, 50°C, et tension nominale, -40°C.
- **SF_0.35V_0C & SF_1.1V_125C :** Simulation croisée SF à très faible tension et tension nominale.
- **FS_0.35V_0C & FS_1.1V_125C :** Simulation croisée FS à très faible tension et tension nominale.

4 Évolution des caractéristiques

L'objectif est d'observer l'évolution des caractéristiques des points mémoires lorsque la tension d'alimentation est réduite. Cette étude permet de distinguer les avantages offerts par chaque architecture en termes de SNM, de WM, et de comportement dynamique.

4.1 Points mémoires sélectionnés

Pour cette étude, un jeu de quatre cellules mémoire sera simulé :

- **SRAM6T_RBL :** Cellule mémoire industrielle de référence. Utilise les règles de conception des mémoires et les modèles associés.
- **SRAM6T :** Cellule mémoire industrielle de référence portée aux règles de conception standard et utilisant les modèles de transistors standards.
- **SRAM10T_KR :** Cellule mémoire extraite de l'étude bibliographique [71], utilisant deux passgates et deux transistors de pied dont la source est modulée.
- **SRAM10T_ST :** Cellule mémoire extraite de l'étude bibliographique [111], utilisant une structure en Schmitt Trigger.

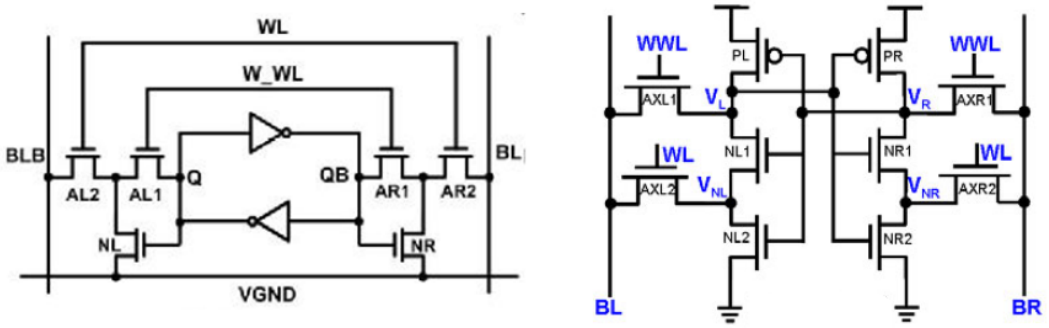

Figure 4.10 – Représentation schématique de la mémoire SRAM10T_KR (gauche) et de la mémoire SRAM10T_ST (droite).

Les deux cellules extraites de l'étude bibliographique (Figure 4.10) sont considérées comme les meilleures candidates pour la conception de mémoires à très faible tension, par leur résultat et leur architecture différentielle. Le dimensionnement des cellules respecte le β-ratio et l'α-ratio de la cellule SRAM6T_RBL. La surface de chaque cellule est résumée dans le Tableau 4.1. Le respect des règles de conception imposées par la plateforme technologique augmente le coût en surface des cellules mémoires, ce qui explique le rapport x1,81 entre les deux cellules SRAM6T. La violation de ces règles, à l'image de la SRAM6T_RBL, est prématurée.

4.2 Simulation et analyse des résultats

La simulation est réalisée en condition typique, à la tension nominale de 1.2V et la tension très faible de 0.35V.

Table 4.1 – Surface des cellules mémoires.

Cellule	$Surface(\mu m^2)$	$S_{CELLULE}$ / S_{6T}
SRAM6T_RBL	0.67	0.55
SRAM6T	1.216	1
SRAM10T_KR	2.244	1.85
SRAM10T_ST	2.179	1.79

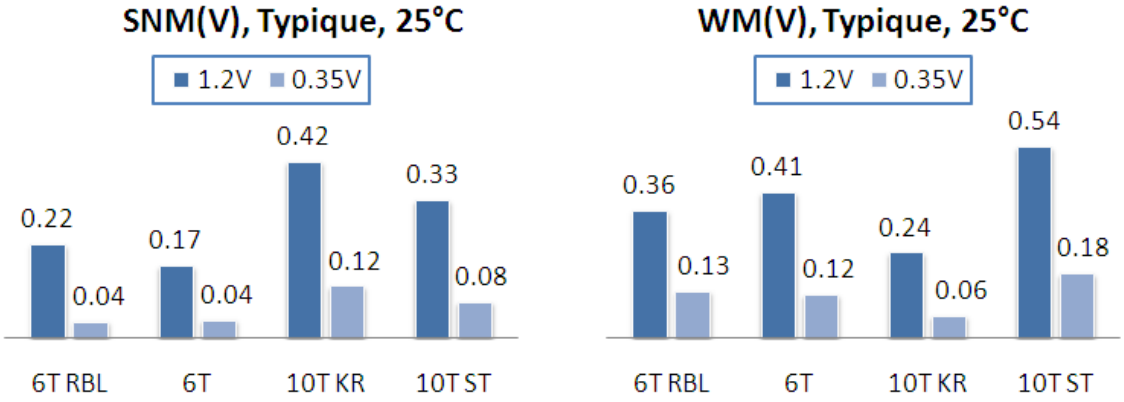

Figure 4.11 – Simulation de la SNM (gauche) et de la WM (droite).

On observe sur l'ensemble des architectures une importante dégradation de la SNM à très faible tension (Figure 4.11). La SNM des cellules 6T est divisée par un facteur x5. La SNM des cellules 10T est divisée par un facteur x4. Ces dernières présentent une SNM plus élevée que les cellules 6T, la cellule SRAM10T_KR présentant la meilleure SNM de ce comparatif.

La WM présente également une importante dégradation, avec un facteur x4 observé sur l'ensemble des cellules. On distingue les effets des architectures SRAM10T_KR et SRAM10T_ST, le premier favorisant la stabilité par l'utilisation de ses deux transistors en série, et le second favorisant l'écriture par son double accès à la bitline.

Alors que le comportement statique des architectures 6T laisse envisager une possible conception à très faible tension, l'analyse dynamique met en évidence leur faiblesse. La cellule SRAM6T_RBL n'est pas fonctionnelle à très faible tension pour un cycle d'horloge de 100kHz. L'autre architecture à six transistors permet d'obtenir le comportement dynamique le plus performant à très faible tension en regard des architectures à dix transistors (Figure 4.12).

En résumé :

- La cellule SRAM10T_KR permet d'obtenir la meilleure SNM.

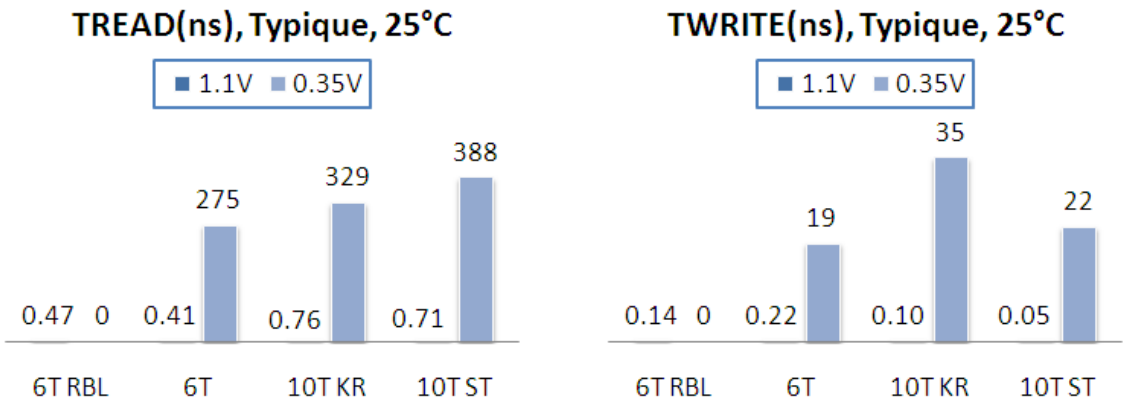

Figure 4.12 – Simulation de TREAD (gauche) et de TWRITE (droite).

- La cellule SRAM10T_ST permet d'obtenir la meilleure WM.
- La cellule SRAM6T permet d'obtenir les meilleures performances dynamiques.
- La cellule SRAM6T_RBL n'est pas fonctionnelle en simulation dynamique.

L'architecture 6T est connue pour être sensible aux variations. On observe sur la Figure 4.13 la comparaison des figures en papillon des deux structures, obtenues après 1000 tirages Monte-Carlo. On note le recouvrement des courbes pour la cellule 6T, correspondant à une SNM nulle.

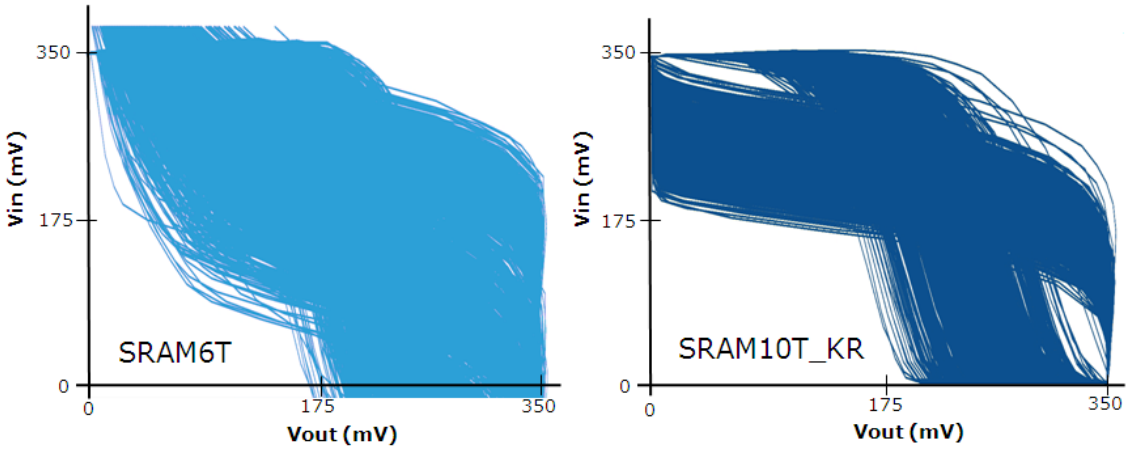

Figure 4.13 – Courbes en papillon des cellules 6T et 10T_KR après 1000 tirages Monte-Carlo.

La stabilité a été citée comme critère principal lors de l'étude bibliographique pour la conception de mémoires à très faible tension. L'architecture de la cellule SRAM10T_KR se distingue par sa stabilité en regard de l'architecture de la cellule SRAM10T_ST, qui est également

pénalisée par son double accès à la BL : si cela permet de favoriser l'écriture, cette solution impose le doublement de la capacité de la BL, pénalisant l'opération de lecture, et limitant la hauteur de la colonne. Enfin, la sensibilité aux variations de l'architecture de la cellule SRAM6T l'écarte de la conception à très faible tension.

Parmi les solutions existantes, la cellule SRAM10T_KR est la meilleure candidate à la lecture de ces résultats.

5 Conception d'un point mémoire

L'objectif est de concevoir un point mémoire fonctionnel à très faible tension mais également à tension nominale. On rappelle que l'architecture SRAM10T_KR utilise une modulation de la source de ses transistors de pied réalisée à partir de transistors DT-MOS, dont la fonctionnalité est limitée aux faibles tensions. De plus, l'écriture est améliorée par une sur-alimentation des grilles des passgates, ce qui implique la génération d'une source de tension additionnelle.

5.1 Architecture

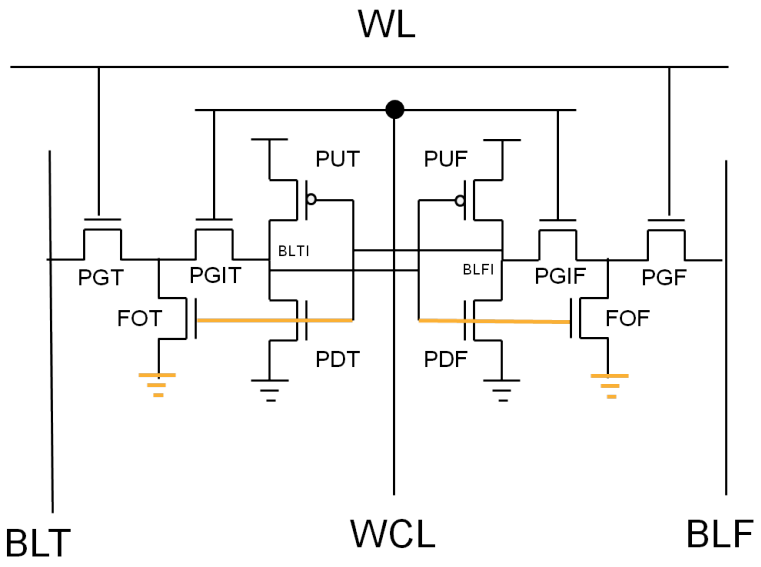

Figure 4.14 – Proposition d'une architecture à 10 transistors, SRAM10T_ULV.

Une cellule à 10 transistors est conçue sur la base de l'architecture 10T_KR (Figure 4.14) avec pour objectif d'être fonctionnelle sans l'utilisation de la modulation des transistors de pied FOT et FOF. Les connections des noeuds internes aux transistors FOT et FOF ont également été

modifiées : la grille du transistor FOT est connectée au noeud interne opposé BLFI, la grille du transistor FOF étant alors connectée au noeud interne BLTI. Ces connexions permettent de supprimer les phénomènes de fuites à travers les passgates internes plutôt que sur les bitlines, afin de préserver la stabilité de la cellule, les fuites sur les bitlines pouvant être réduites par le dimensionnement des transistors de pied.

La comparaison des résultats permet de confirmer la validité de cette architecture face à la structure SRAM10T_KR. Les performances sont maintenues, voir améliorées pour la lecture dynamique (Figure 4.15). De plus, le fonctionnement à tension nominale n'est pas limité.

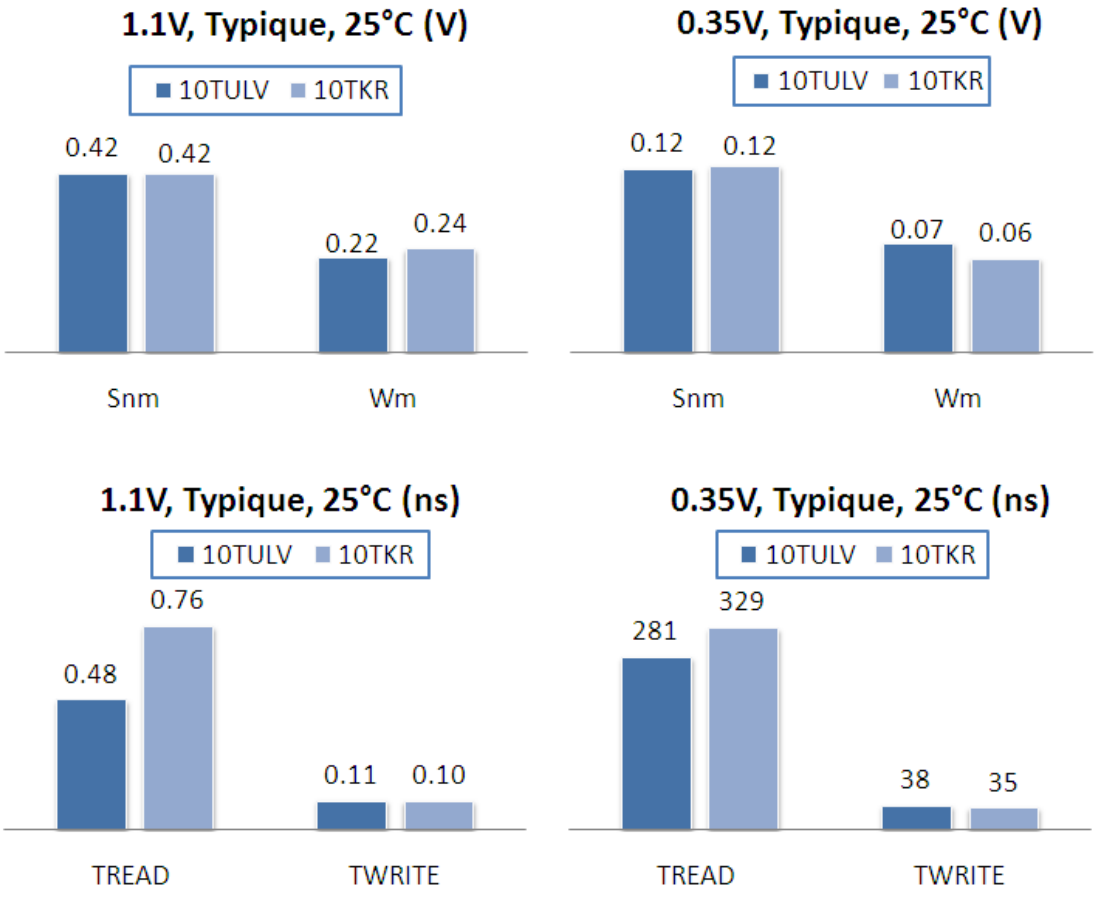

Figure 4.15 – Comparaison des résultats entre les structures SRAM10T_KR et SRAM10T_ULV.

5.2 Dimensionnement de la cellule SRAM10T_ULV

La cellule SRAM10T_ULV a été dimensionnée afin d'améliorer la stabilité, l'écriture, et la sensibilité aux variations. Ce dimensionnement a été réalisé dans un premier temps par l'utilisation d'un outil de dimen-

sionnement automatique nommé WiCkeD [112]. Les contraintes suivantes ont été données à l'outil :

- Une augmentation maximale de W_{PG} fixée à 10% pour limiter l'augmentation de la capacité de la bitline.
- Une variation de la largeur de grille de 0% à 50%.
- Une variation de W de 0% à 50% pour limiter l'impact sur la surface.
- Une SNM supérieure à 20% de VDD.
- Une WM supérieure à 20% de VDD.

Les stimuli développés précédemment sont utilisés. L'objectif est d'obtenir des résultats SNM et WM satisfaisants sur l'ensemble des points de caractérisation TT, FF, SS, SF et FS. Les résultats sont rapportés sur la Figure 4.16. On note une amélioration générale des caractéristiques de la cellule. On observe également une récupération de la fonctionnalité dynamique sur les points de caractérisation SS et SF.

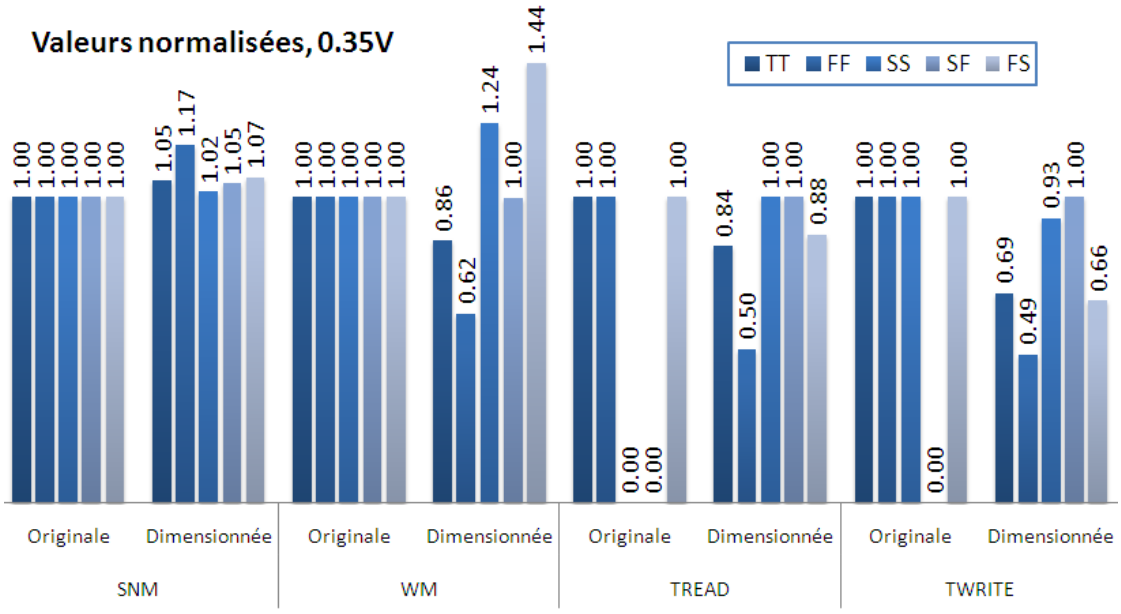

Figure 4.16 – Résultats obtenus sur l'ensemble des points de mesures.

Seule la WM a été dégradée au profit des autres caractéristiques. Le dimensionnement des transistors PU et PD permet le plus de gain en termes de SNM, tandis que la WM est principalement affectée par le dimensionnement des passgates (Figure 4.17).

La limitation observée en termes de WM est liée à la mise en série des transistors d'accès, ce qui réduit considérablement le courant, à l'image des observations faites lors de l'étude des cellules logiques. Pour compenser cette perte de courant, on utilise des passgates PG et

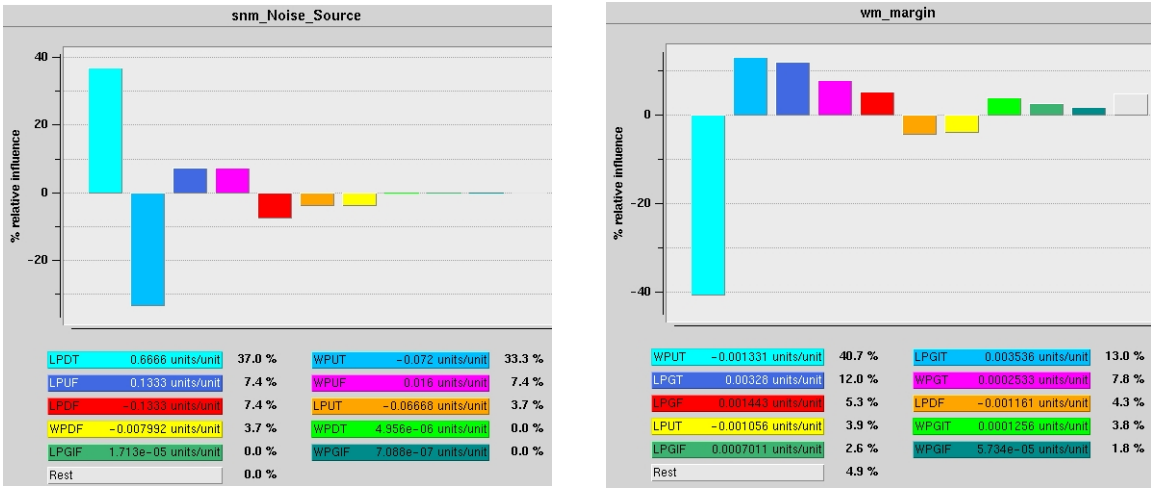

Figure 4.17 – Sensibilité de la SNM (gauche) et de la WM (droite) au dimensionnement des transistors de la cellule.

PGI en LVT. Afin de limiter l'impact sur la consommation, les grilles des passgates sont élargies à 100nm, soit une augmentation de 60%. La SNM est garantie par la fermeture des passagates internes. Après une seconde passe de dimensionnement automatique et quelques ajustements manuels pour limiter la dégradation de la WM, on obtient les résultats affichés en Figure 4.18.

MC 1001 Runs Period = 10000 ns		SRAM10T_ULV					
		TT	FF	SS	SF	FS	MEAN
SNM (mV)	0.35V Centered	96	81	105	100	63	88.8
	1.1V Centered	408	416	386	372	338	384.0
WM (mV)	0.35V Centered	101	111	91	21	74	79.7
	1.1V Centered	241	291	201	111	311	231.2
TREAD (ns)	0.35V Centered	986	143	7694	6113	907	3168.6
	1.1V Centered	1.02	0.65	1.73	1.61	1.25	1.25
	0.35V Fails	0	0	2	0	0	0.4
	1.1V Fails	0	0	0	0	0	0.0
TWRITE (ns)	0.35V Centered	61	12	435	128	383	203.6
	1.1V Centered	0.17	0.09	0.30	0.29	0.19	0.21
	0.35V Fails	0	0	0	1	0	0.2
	1.1V Fails	0	0	0	0	0	0.0

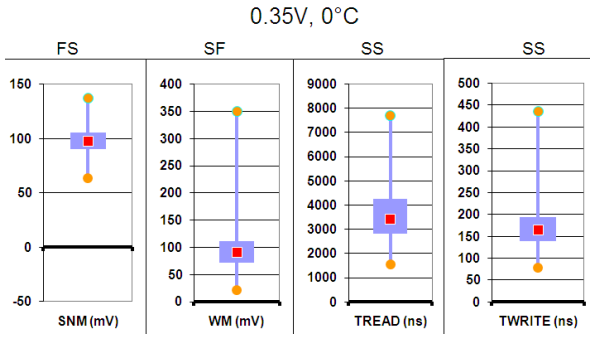

Figure 4.18 – Simulation Monte-Carlo sur l'ensemble des points de caractérisation, avec affichage du moins bon résultat (gauche). La distribution des différentes caractéristiques est affichée en box-plot (droite).

On observe que la WM est la caractéristique la plus faible de cette architecture. Elle reste néanmoins supérieure à 0 dans le pire cas à 6σ, où elle atteint 21mV. La SNM est au pire cas égale à 63mV. Concernant le comportement dynamique, il atteint les 8μs en lecture au pire cas. Au point de caractérisation typique, le résultat est très proche du MHz souhaité dans le cadre de cette thèse. On observe également quelques erreurs parmi les 1000 tirages effectués. Les facteurs de mérite μ/σ de cette architecture ainsi dimensionnée sont reportés sur la Figure 4.19. On note un excellent facteur de mérite pour la SNM, tandis que la WM présente un facteur de mérite inférieur à 3.

MC 1001 Runs Period = 10000 ns		Facteur de mérite µ/σ					
		TT	FF	SS	SF	FS	MEAN
SNM	0.35V Centered	12.7	10.5	14.5	12.8	8.4	11.8
	1.1V Centered	39.4	36.4	40.7	38.7	30.9	37.2
WM	0.35V Centered	2.7	2.9	2.6	3.5	2.1	2.8
	1.1V Centered	16.3	17.8	14.8	8.9	21.4	15.9
TREAD	0.35V Centered	4.1	4.8	3.5	3.3	4.8	4.1
	1.1V Centered	26.6	23.7	30.9	27.0	28.9	27.4
TWRITE	0.35V Centered	4.6	5.3	3.8	4.0	3.1	4.1
	1.1V Centered	33.7	27.1	36.4	25.5	37.2	32.0

Figure 4.19 – Facteurs de mérite pour la cellule SRAM10T ULV.

Concernant le courant statique, on relève une diminution de celui-ci à 0.35V en comparaison de la structure 6T_RBL. Cependant, il est doublé à 1.2V, en raison du nombre de transistors supplémentaires et de l'utilisation du LVT (Figure 4.20). On note également le facteur N_{BL} supérieur à 40 à très faible tension pour une hauteur de 128 lignes.

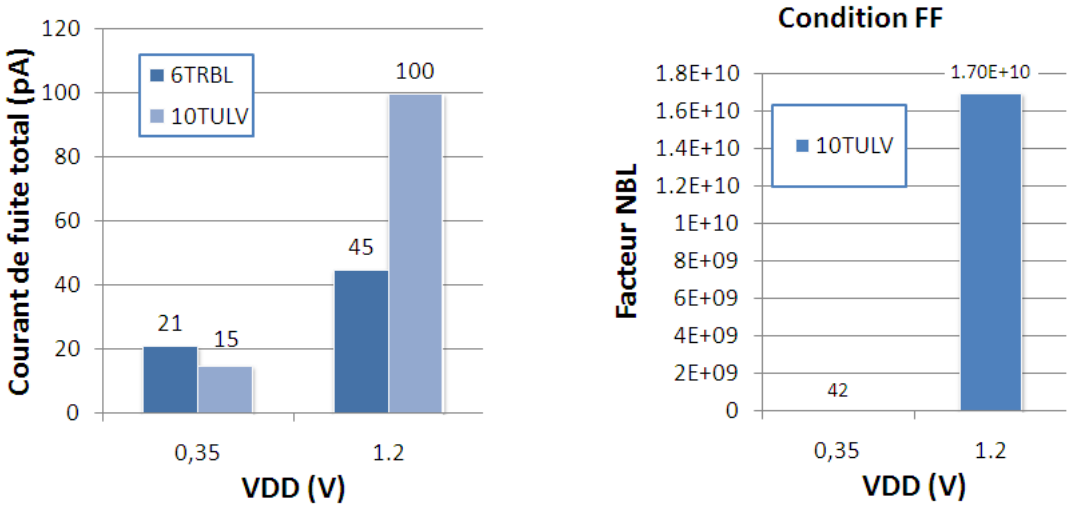

Figure 4.20 – Courant statique total des cellules 6T_RBL et 10T_ULV (gauche), et facteur N_{BL} pour la cellule 10TULV pour une hauteur de 128 lignes.

Ces résultats sont considérés comme les meilleurs pouvant être obtenus avec cette architecture, pour une surface augmentée d'un facteur x2,1 par rapport à la cellule SRAM6T.

6 La périphérie

6.1 Principe

La périphérie désigne l'environnement de la matrice de cellules mémoires permettant la sélection d'une cellule et l'exécution des différentes opérations sur celle-ci. La dimension de cette périphérie doit être réduite au minimum : dans l'industrie, la surface utile d'une mémoire SRAM, sur laquelle est définie le coût, correspond à la matrice des cellules, hors périphérie.

Une mémoire est ainsi habituellement composée des éléments suivants :

- Un Contrôleur.
- Un Décodeur d'adresse.
- Une ou plusieurs matrices, contenant les cellules mémoires.
- Une périphérie de sélection de colonne et d'exécution des opérations.

Le contrôleur est le point d'entrée de la mémoire. Il permet de commander le décodeur ligne, la sélection en colonne, et le type d'opération à effectuer. Le décodeur, en fonction de l'adresse communiquée par le contrôleur, active la WL correspondante. La matrice est décomposée en mots de plusieurs bits par ligne. Le bloc nommé "sélection et opération" est divisé en autant de colonne que la matrice. Lorsque le contrôleur communique une adresse, une ou plusieurs colonnes sont activées et opèrent sur la matrice. Une représentation schématique de l'ensemble de la mémoire est disponible en Figure 4.21.

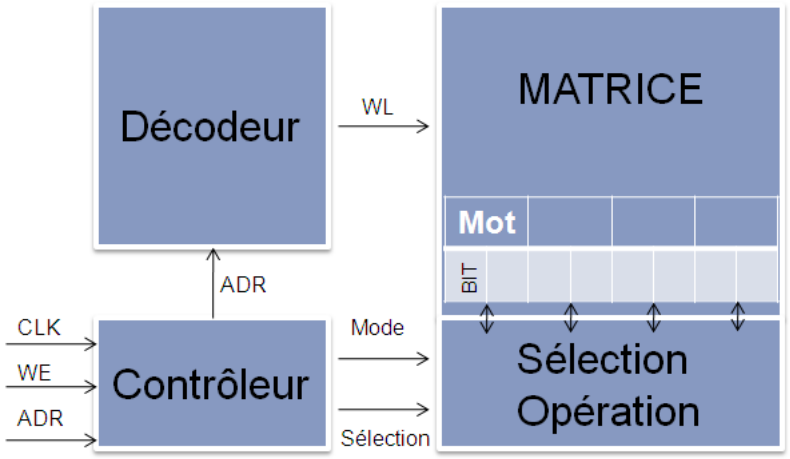

Figure 4.21 – Représentation schématique d'une mémoire SRAM. Le signal "Mode" décrit le type d'opération effectué, WE est l'horloge d'activation de l'écriture, CLK est l'horloge globale, ADR est l'adresse du mot.

Chaque colonne du bloc "sélection et opération" contient les éléments suivants (Figure 4.22) :

- Un circuit de sélection permettant d'activer ou non l'accès à la colonne.
- Un circuit de précharge, destiné à précharger les bitlines à VDD. Cette étape permet de réaliser l'opération HOLD, et de préparer les opérations de lecture et d'écriture.
- Un circuit de lecture permettant la déconnexion des bitlines pour la lecture de la donnée stockée dans la cellule.
- Un circuit d'écriture permettant l'alimentation des bitlines en fonction de la donnée à stocker.

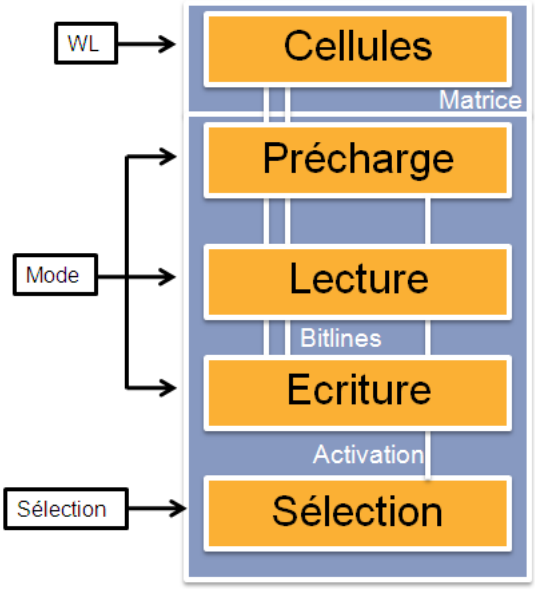

Figure 4.22 – Contenu du bloc de sélection et opération.

6.2 Conception d'une assistance à la lecture

L'opération de lecture est usuellement réalisée par l'utilisation d'un amplificateur différentiel [113]. Cette cellule permet d'enclencher la lecture avant que l'une des bitlines n'ait atteint GND, en détectant une chute de tension ΔVDD sur celle-ci. Cela permet de réduire considérablement le temps de lecture (Figure 4.23).

A très faible tension, cette solution présente une limitation par la faible valeur du ΔVDD. De plus, les amplificateurs différentiels sont séquencés : la lecture est exécutée à un temps T défini par un chemin factice émulant le pire cas en temps de propagation. Ce séquencement rend l'amplificateur particulièrement sensible aux variations à très faible tension. Comme constaté lors de l'étude bibliographique, la

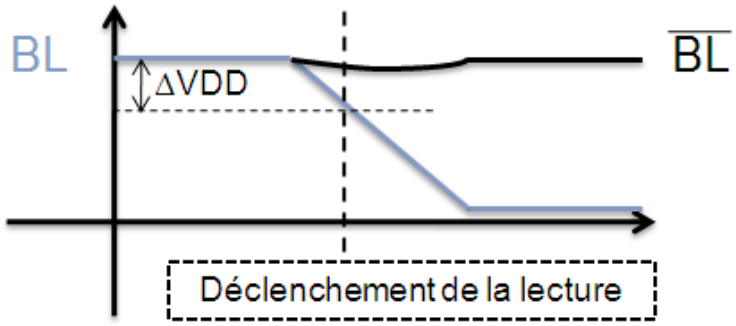

Figure 4.23 – Principe de lecture par l'utilisation d'un amplificateur différentiel.

lecture pleine amplitude est favorisée, ce qui signifie que la lecture n'est réalisée qu'à partir du moment où l'une des bitlines atteint GND.

On note qu'en conséquence de cette lecture pleine amplitude, le temps de lecture relevé lors de l'étude de la cellule est fortement dégradé sur le point de caractérisation SS. On note également que deux tirages parmi mille ne permettent pas l'exécution d'une lecture dans le temps imparti. Afin d'accélérer la lecture à très faible tension, une solution d'assistance a été développée (Figure 4.24).

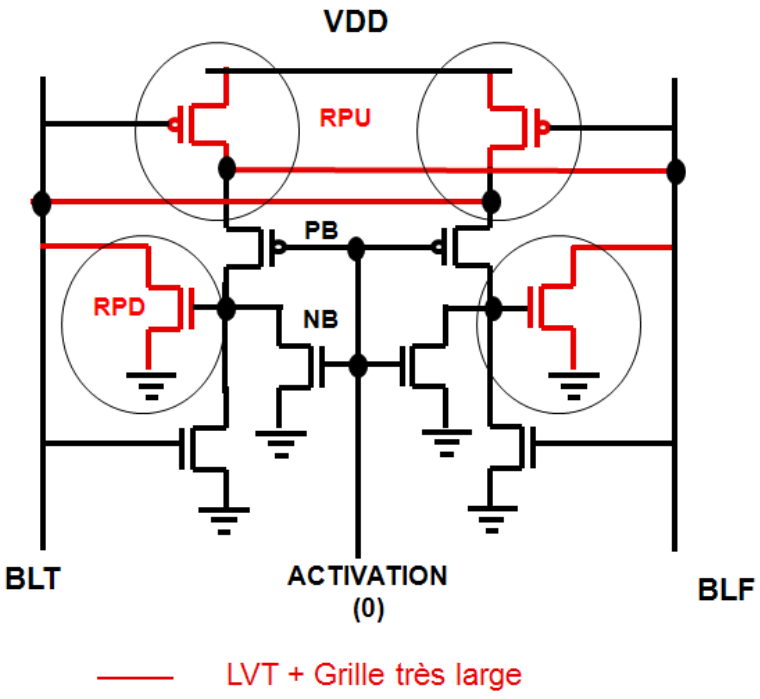

Figure 4.24 – Représentation schématique de la solution d'assistance à la lecture.

Le principe de cette assistance est de dissocier le courant dynamique tirant une bitline vers GND du courant de fuite tirant l'autre bitline vers GND. Une architecture croisée a ainsi été développée. Le courant dynamique va activer le PMOS de tête RPU qui va directement imposer VDD sur l'autre bitline. Cela permet de compenser le courant de fuite provoquant la chute de tension. RPU est un transistor LVT, permettant un courant dynamique élevé, auquel la grille a été élargie de 150% pour minimiser l'impact sur la consommation statique. RPU active également le NMOS de pied RPD, qui permet d'accélérer la chute de la bitline à GND. RPD est en LVT et utilise une grille élargie de 150%. Les transistors de pied et de tête sont directement couplés entre la source et la bitline pour un courant dynamique maximal.

Cette solution peut être désactivée par l'intermédiaire des transistors PB et NB. Elle peut ainsi être synchronisée avec le cycle de lecture. Cette solution ne fonctionne toutefois qu'avec un facteur N_{BL} supérieur à 1 (Figure 4.25).

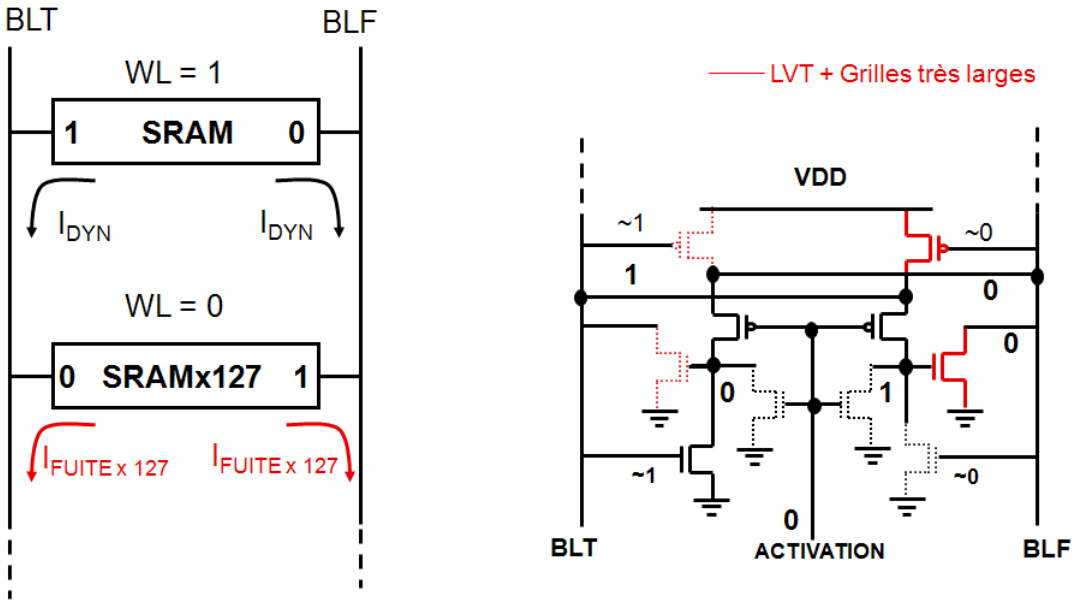

Figure 4.25 – Exemple de fonctionnement de l'assistance à la lecture.

Les résultats obtenus par cette solution sont reportés sur la Figure 4.26. Elle permet de diviser par un facteur x3 le temps de lecture à très faible tension. On observe un gain de 40% à tension nominale. Afin de garantir le fonctionnement de cette cellule sur l'ensemble des points de caractérisation, elle a été intégrée au chemin critique développé précédemment. Les résultats reportés en Figure 4.27 confirme l'efficacité de la solution, avec un temps de lecture pire cas ramené à 4μs, un temps de lecture typique de 600ns, et enfin l'obtention d'une simulation statistique sans erreur.

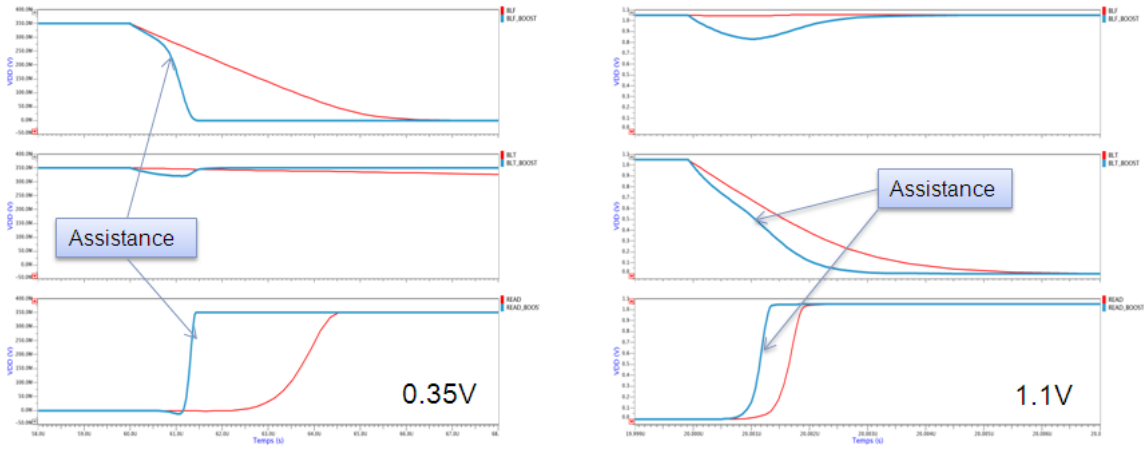

Figure 4.26 – Temps de lecture à 0.35V (gauche) et 1.1V (droite) obtenus avec l'utilisation de l'assistance à la lecture, en condition SS.

MC 1001 Runs Period = 10000 ns		SRAM10T_ULV avec Assistance					
		TT	FF	SS	SF	FS	MEAN
SNM	0.35V Centered	96	81	105	100	63	88.8
	1.1V Centered	408	416	386	372	338	384.0
WM	0.35V Centered	101	111	91	21	74	79.7
	1.1V Centered	241	291	201	111	311	231.2
TREAD	0.35V Centered	623	108	4112	1546	702	1418.1
	1.1V Centered	0.65	0.41	1.06	0.79	0.86	0.76
	0.35V Fails	0	0	0	0	0	0.0
	1.1V Fails	0	0	0	0	0	0.0
TWRITE	0.35V Centered	54	10	393	304	350	222.1
	1.1V Centered	0.17	0.09	0.30	0.30	0.19	0.21
	0.35V Fails	0	0	0	0	0	0.0
	1.1V Fails	0	0	0	0	0	0.0

Figure 4.27 – Simulation Monte-Carlo sur l'ensemble des points de caractérisation du chemin critique avec assistance à la lecture.

En résumé, cette solution permet :

- L'accélération de la lecture du 0.
- La compensation des fuites sur la bitline à VDD par montage croisé.
- Un gain en temps de lecture supérieur à 40%.

6.3 Conception d'une assistance à l'écriture

L'écriture présente une marge réduite à 21mV dans le pire cas. Ce qui signifie qu'il est nécessaire de tirer la bitline en dessous de cette valeur pour que l'écriture s'opère dans la cellule. Pour cela, des buffers à grande sortance sont utilisés. Afin de renforcer leur action et anticiper une prédiction optimiste de cette marge, une solution d'assistance

à l'écriture a été développée(Figure 4.28).

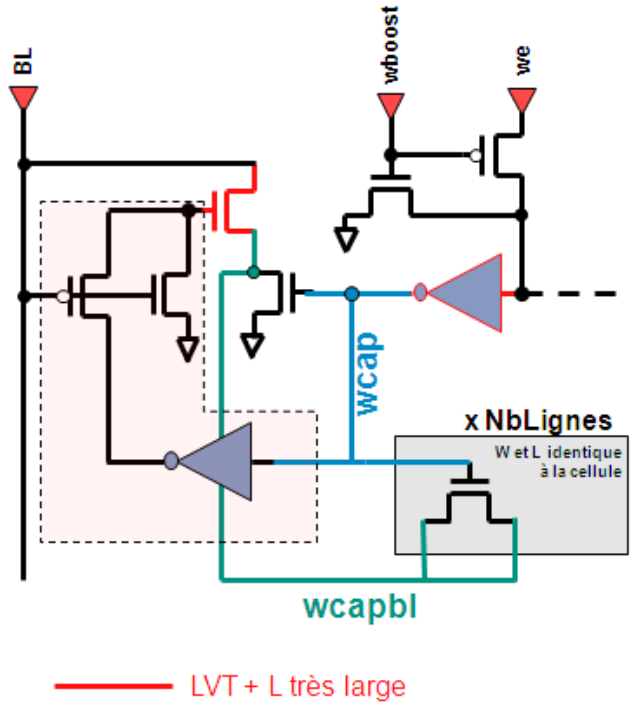

Figure 4.28 – Représentation schématique de la solution d'assistance à l'écriture.

Reprenant le principe d'un brevet STMicroelectronics [114], cette solution consiste à permettre la génération d'une tension négative sur la bitline chargée d'écrire la valeur GND dans la cellule mémoire. La génération d'une telle tension permet de garantir l'écriture, la marge pire cas étant à 21mV. Pour cela, une capacité a été conçue par la mise en parallèle d'un transistor NMOS dont la source et le drain sont connectés pour former une borne de la capacité (wcapbl), tandis que la grille forme l'autre borne (wcap). Ce transistor ayant les mêmes dimensions que PG, et étant mis en parallèle X fois, X étant le nombre de lignes de la matrice, la capacité ainsi créée est équivalente à celle de la bitline. Cette charge peut être intégrée dans la matrice avec un faible coût en surface.

L'assistance fonctionne en deux phases : une phase de précharge et une phase de décharge (Figure 4.29).
Lorsque l'écriture n'est pas activée (we=0), l'inverseur LVT charge la borne wcap à VDD, tandis que wacpbl est chargée à GND.
Lorsque l'écriture (we=1) ainsi que l'assistance (wboost=0) sont activées, wcap est tirée vers GND, ce qui génère une tension -VDD sur la borne wcapbl par effet capacitif. Le circuit de lecture entouré par des

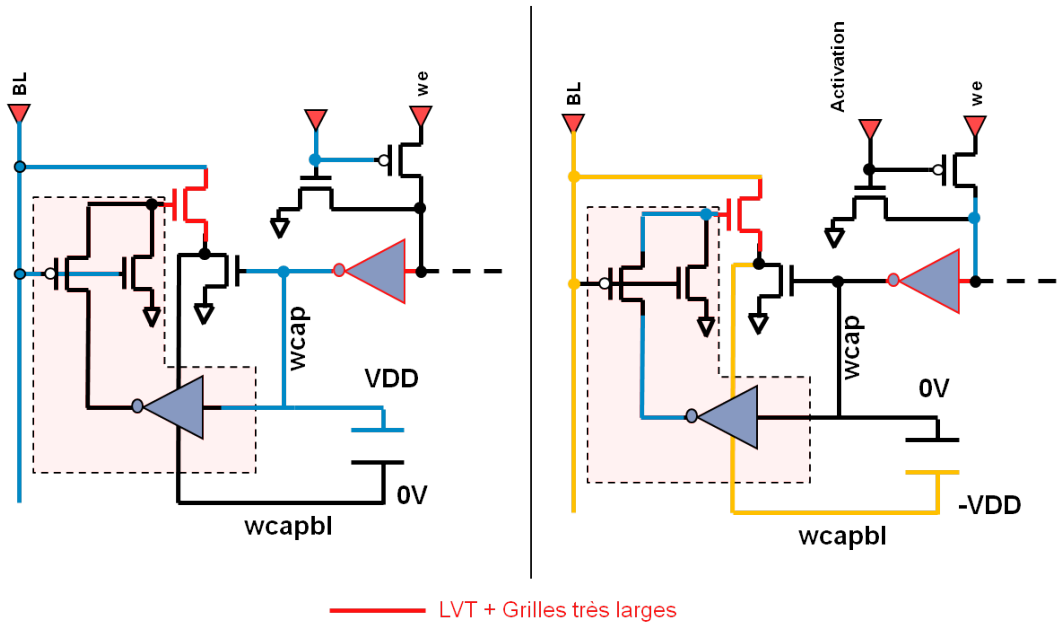

Figure 4.29 – Phases de précharge (gauche) et de charge (droite) de la solution d'assistance à l'écriture.

pointillés sur la Figure 4.28 est chargé de détecter le passage de la bitline à GND : la charge -VDD est alors transmise à la bitline par le transistor NMOS LVT. Il est nécessaire que la bitline ait atteint GND avant de transmettre la charge pour obtenir la plus grande chute de tension.

Cette solution a été implémentée dans le chemin critique afin d'évaluer ses performances. On observe sur la Figure 4.30 qu'elle permet de réduire la tension sur la bitline jusqu'à -60mV à 0.35V et jusqu'à -145mV à tension nominale, dans le point de caractérisation SS. Le temps d'exécution de cette solution est du même ordre de grandeur que l'opération d'écriture.

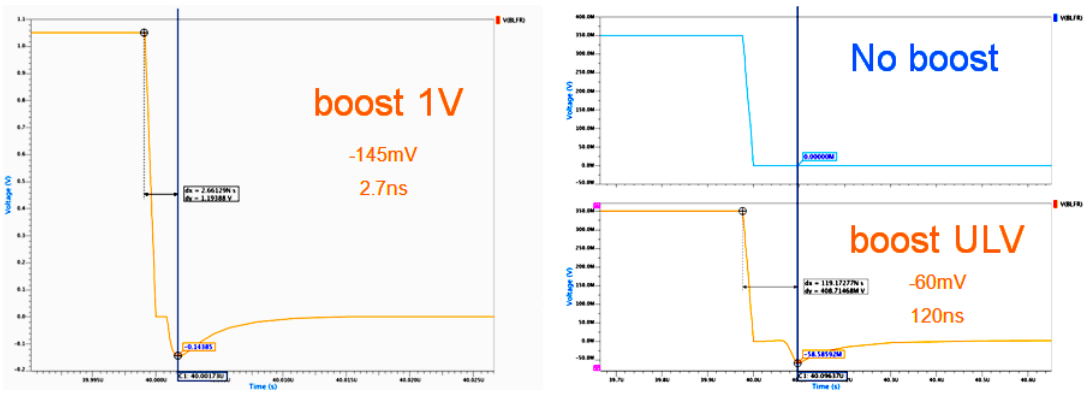

Figure 4.30 – Simulation du chemin critique à tension nominal (gauche) et à très faible tension (droite) avec la solution d'assistance à l'écriture.

7 Réalisation d'un démonstrateur mémoire sur silicium

La cellule mémoire SRAM10T_ULV ainsi que les solutions d'assistance en lecture et écriture ont été intégrées dans un démonstrateur conçu en 65nm (Figure 4.31). L'objectif de ce démonstrateur est de valider le fonctionnement à très faible tension de la mémoire et des solutions d'assistance, et de comparer les résultats en termes de consommation, de vitesse, et de rendement avec l'offre industrielle.

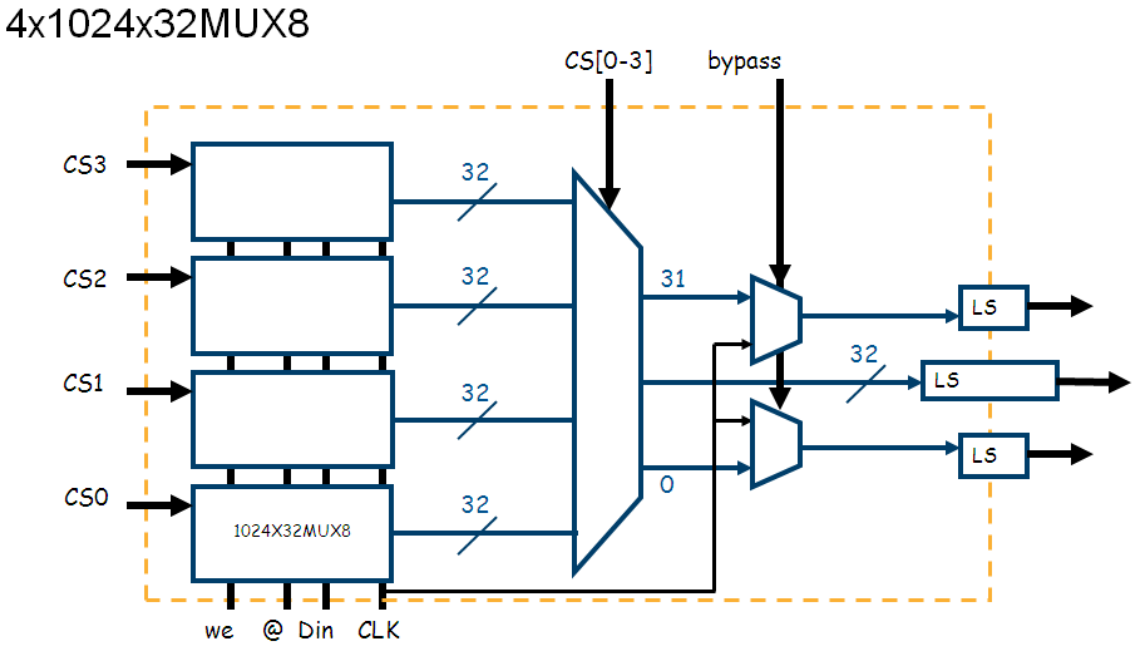

Figure 4.31 – Représentation schématique du démonstrateur mémoire.

7.1 Caractéristiques

Le démonstrateur est composé de quatre mémoires de 32kb. Les entrées et sorties de ces mémoires sont multiplexées. Les mémoires ont été dupliquées quatre fois pour augmenter les données statistiques pouvant être recueillies. Un chemin de caractérisation temporelle a été ajouté. Ce chemin consiste à relier directement l'horloge à une sortie additionnelle. En réalisant la soustraction entre le temps mis par la lecture et le temps mis par ce chemin, on obtient le temps d'accès réel de la mémoire.

Les entrées du démonstrateur, embarqué sur un testchip, proviennent du domaine de tension nominal. Par conséquent, des cellules d'interfaces sont utilisées afin de garantir un rapport cyclique correct. Un

convertisseur de tension a également été placé sur toutes les sorties. Ces cellules ont été développées selon la méthodologie décrite dans le Chapitre 2.

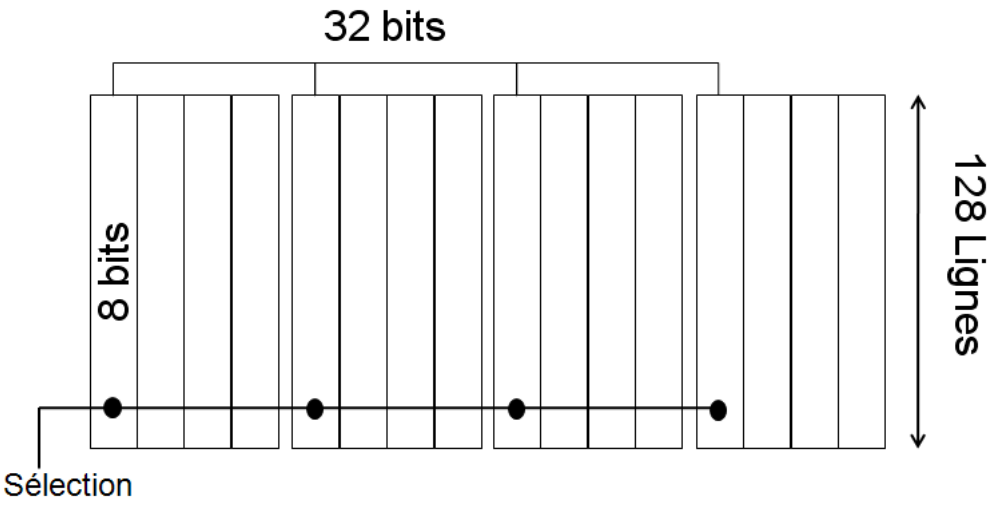

Figure 4.32 – Représentation schématique d'une mémoire.

Ces quatre mémoires présentent les caractéristiques suivantes (Figure 4.32) :

- 128 Lignes.
- 8 mots par lignes.
- Mots de 32bits.
- Solutions d'assistance en lecture et écriture.
- Capacité de la solution d'écriture intégrée dans la matrice.
- Contrôleur avec activation globale des solutions d'assistance.

Le multiplexage permet de séparer les mots en plusieurs groupes de 8bits. Cette opération permet de réduire l'impact local des radiations sur le mot, seule une portion de bit étant alors affectée [71]. Concernant la solution d'assistance en écriture, la colonne de capacités est répétée pour chaque groupe de 8bits.

Les solutions d'assistance ont été synchronisées avec les cycles de lecture et d'écriture. Ainsi, l'assistance en lecture s'opère si :

- L'assistance en lecture est activée au niveau global.
- La précharge est désactivée (Front bas).
- L'horloge d'écriture "we" est désactivée (Front bas).

L'assistance en écriture s'opère si :

- L'assistance en écriture est activée au niveau global.
- La précharge est désactivée (Front bas).
- L'horloge d'écriture "we" est activée(1).
- La cellule d'écriture a fini de tirer la bitline à GND.

Le mécanisme d'écriture (Figure 4.33) intègre une détection de l'état de la bitline. A la suite de la lecture d'un couple 0-1 sur les bitlines, l'écriture est arrêtée, et l'assistance est activé.

Les assistances ne peuvent ainsi être activées simultanément sur le même cycle ou lors d'un cycle de précharge.

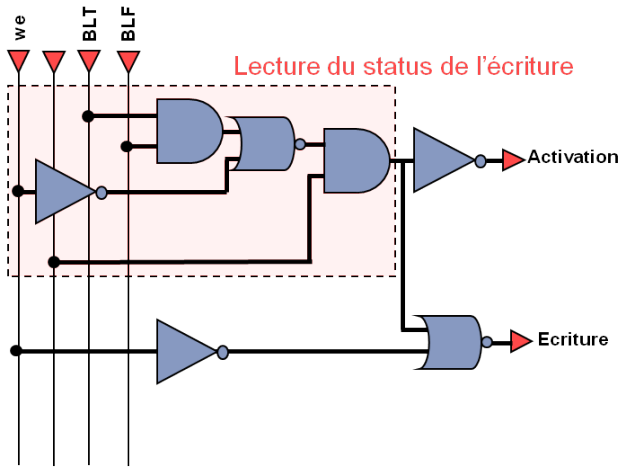

Figure 4.33 – Représentation schématique du mécanisme d'écriture.

7.2 Vérification fonctionnelle et temporelle

Un pattern de test a été construit pour tester et valider la fonctionnalité du démonstrateur. Ce pattern consiste en l'exécution successive d'une l'écriture dans chaque mémoire puis la lecture des données stockées. En voici le détail :

- Horloge fixée à 1MHz.
- Tension fixée à 0,35V et 1,2V.
- Sélection successive des mémoires 0, 1, 2, et 3 (CS).
- Écriture du mot de 32bit 0xAAAAAAAA à l'adresse 0x000 de chaque page.
- Écriture du mot de 32bit 0x55555555 à l'adresse 0x3FF de chaque page.
- Sélection successive des pages 0, 1, 2, et 3.
- Lecture des mots stockés à l'adresse 0x000 de chaque mémoire.

- Lecture des mots stockés à l'adresse 0x3FF de chaque mémoire.
- Activation du chemin de caractérisation temporelle.

Les chronogrammes obtenus pour la simulation à 0,35V sont affichés en Figure 4.34. Ils valident par la simulation la fonctionnalité à très faible tension du démonstrateur.

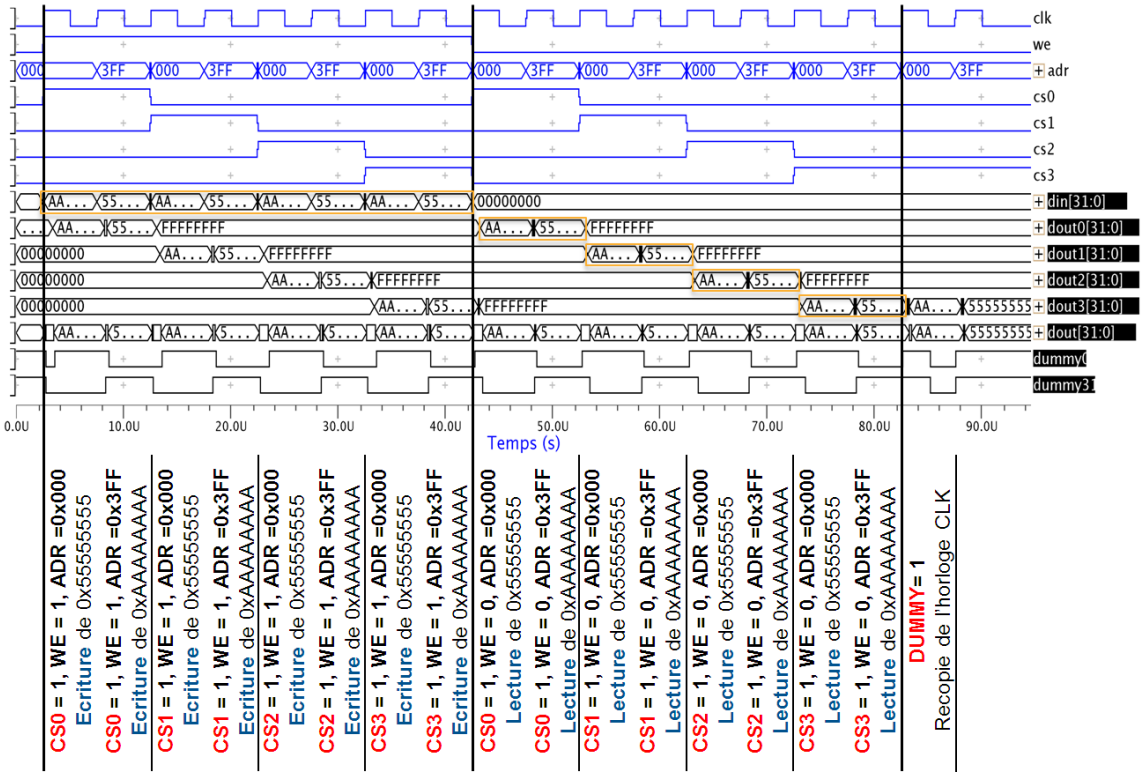

Figure 4.34 – Chronogrammes de la simulation à 0,35V, 25°C, du démonstrateur.

Les caractéristiques suivantes ont été extraites par le simulateur à 0,35V et à 1,2V (Tableau 4.2) :

- **Temps de précharge TPRECH :** Délai entre le front descendant de l'horloge et le passage des bitlines à VDD.
- **Temps de lecture TREAD :** Délai entre le front montant de l'horloge et la lecture des données en sortie.
- **Temps d'écriture TWRITE :** Délai entre le front montant de l'horloge et l'écriture des données dans la cellule.
- **Énergie de lecture EREAD :** Énergie consommée par l'opération de lecture.
- **Énergie d'écriture EWRITE :** Énergie consommée par l'opération d'écriture.
- **Consommation statique ILEAK :** Courant statique sur VDD.

Table 4.2 – Extraction de la vitesse et de la consommation de la mémoire, pour l'écriture et la lecture.

	10T ULV		**ST 6T**	**ST 6T L**
VDD	$0.35V$	$1.2V$	$1.2V$	$1.2V$
V_{MIN} Fonctionnel (V)	0.35		1.05	0.95
Courant Statique A (μA)	2.79	6.66	1.98	0.33
Courant Statique B (nA)	39.98	274.92	*NC*	80.28
EREAD (pJ)	6.95	10.9	18.09	19.38
EWRITE (pJ)	6.96	30.3	20.07	22.29
Temps de Cycle (ns)	1100	1.78	0.67	1.24
Temps d'Accès (ns)	800	0.75	0.45	0.85
Surface (μm^2)	2.62		1.22	

L'utilisation de la très faible tension conduit à une diminution de la vitesse d'un facteur compris entre x500 et x1000, en comparaison de la tension nominale. La consommation statique est réduite d'un facteur x2,4. L'énergie consommée par cycle de lecture est divisée par un facteur x1,6 et l'énergie consommée par cycle d'écriture est divisée par un facteur x4,35.

Ces résultats, obtenus par une simulation SPICE dégradée, sont comparés à deux mémoires industrielles de même taille. Les valeurs sont extraites d'une base de donnée établie par STMicroelectronics à partir de mesures sur silicium. Notons que la tension minimale V_{MIN} de ces mémoires est nettement plus élevée. On constate cependant que si en termes d'énergie par cycle, la mémoire SRAM10T_ULV permet de gagner un facteur x3 en moyenne par opération à très faible tension, la consommation statique reportée lorsque l'on simule le démonstrateur au complet (A) est plus grande que celle des cellules industrielles. On note toutefois qu'une simulation SPICE non dégradée (B) d'une matrice mémoire de 32kb, hors périphérie, annonce une consommation statique à très faible tension réduite d'un facteur x2 comparée à la meilleur offre industrielle en termes de consommation statique. Ces résultats devront cependant être confirmés par des mesures sur silicium.

7.3 Mesures sur silicium et analyse

La mémoire a été testée selon les procédures suivantes :
- EWS : Test des circuits sur silicium, utilisation de pointes.
- PKG : Test des circuits mis en boitier.

Test EWS

Le test a été effectué selon les spécifications suivantes :

- **Température :** 30°C, 55°C.
- **Tension :** 0,3V à 1,2V.
- **Forward Body-Bias (FBB) :** Sans, [VDD-0,15V ; GND+0,15V], [VDD-0.25V ; GND+0,25V].
- **Assistance Lecture (R) et Ecriture (W) :** $R_{OFF}W_{OFF}$, $R_{OFF}W_{ON}$, $R_{ON}W_{OFF}$, $R_{ON}W_{ON}$.
- **Patterns :** Marinescu Checker Board (0101...) et Inverse (1010..).
- **Nombre de pièces :** 307.

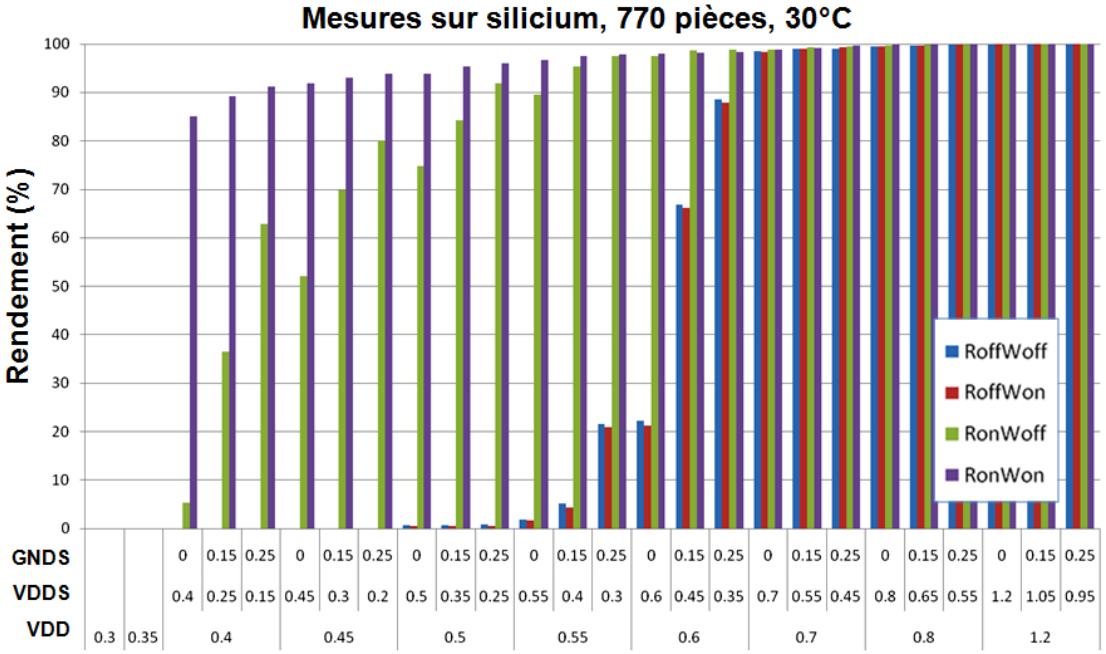

Figure 4.35 – Mesures sur silicium du rendement du démonstrateur à 30°C, en fonction de la tension d'alimentation, du Forward Body-Biasing appliqué, et de l'état d'activation des assistances.

On observe sur la Figure 4.35 l'évolution du rendement en fonction de l'alimentation, du body-biasing appliqué, et des assistances activées. Ces résultats correspondent à la moyenne obtenue sur les quatre matrices, pour l'ensemble des 307 pièces.

Lorsque l'on est dans la configuration $R_{OFF}W_{OFF}$, le rendement chute à partir de 0,6V. Le body-biasing appliqué permet de passer de 20% à 65% pour un FBB de 0,15V, et 88% pour un FBB de 0,25V. En dessous de 0,5V, aucune mémoire n'est fonctionnelle.

Si l'on compare les résultats dans les configurations $R_{OFF}W_{OFF}$ et $R_{ON}W_{OFF}$, on relève une amélioration significative du rendement, notamment à 0,55V où ce dernier passe de moins de 5% à près de 90%. Sachant que la solution d'assistance en lecture ne permet pas de retrouver de la fonctionnalité, mais permet d'accélérer la lecture (c.f. Fi-

gure 4.18a et Figure 4.27), la chute du rendement est liée à une fréquence de test trop élevée et non à une perte de la fonctionnalité. Ainsi, la réduction de la fréquence d'horloge devrait offrir le même résultat en termes de rendement que l'activation de l'assistance en lecture, l'ensemble du décodeur étant entièrement combinatoire. La fréquence de test pour chacune des tensions n'a pas encore été communiquée, exceptée à 0,4V où la fréquence d'horloge a été fixée à 1MHz.

Sur cette même figure, on note lorsque la tension est inférieure à 0,5V une limitation en écriture révélée par la chute de rendement dans la configuration $R_{ON}W_{OFF}$. L'assistance en écriture, en opposition à l'assistance en lecture, permet de retrouver de la fonctionnalité et non d'accélérer l'écriture (c.f. Figure 4.18a et Figure 4.27), ce qui se traduit par une amélioration conséquente du rendement à très faible tension.

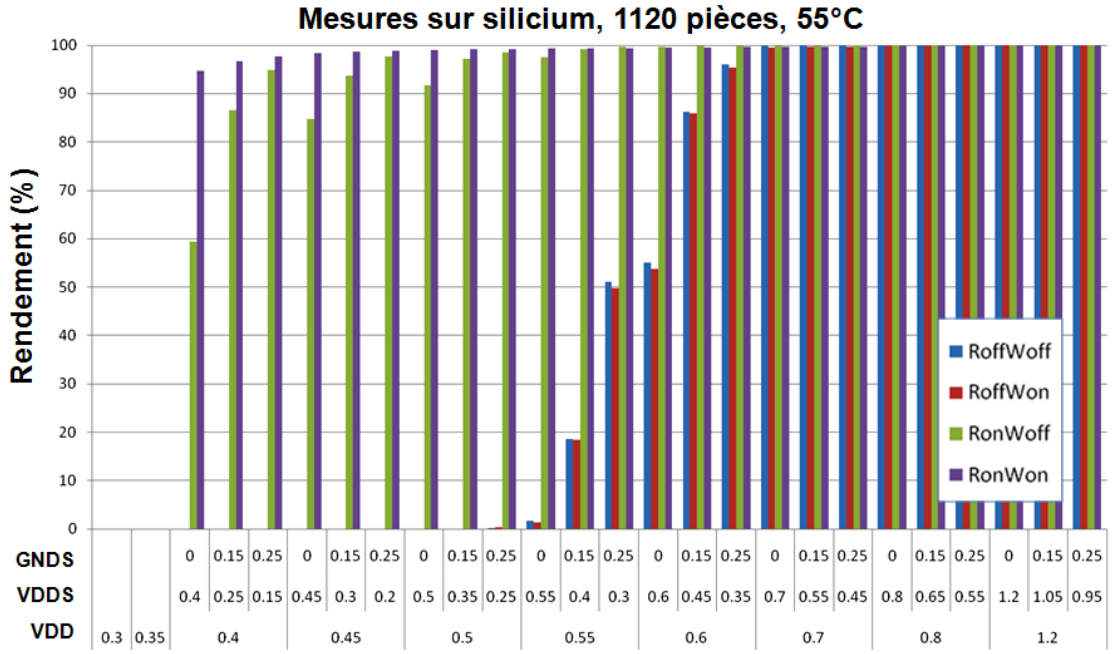

Figure 4.36 – Mesures sur silicium du rendement du démonstrateur à 55°C, en fonction de la tension d'alimentation, du Forward Body-Biasing appliqué, et de l'état d'activation des assistances.

La Figure 4.36 présente les résultats de mesures à 55°C. A cette température, comme présenté lors des précédents travaux sur la conception de circuits numériques, le courant à très faible tension est plus élevé et se traduit par un rendement plus important.

Test PKG

A ce stade, seul un test fonctionnel a été effectué selon les spécifications suivantes :

- **Température :** 30°C.
- **Tension :** 0,35V.
- **Fréquence :** 200kHz.
- **Forward Body-Bias (FBB) :** Sans
- **Assistance Lecture (R) et Ecriture (W) :** $R_{ON}W_{ON}$.
- **Patterns :** Marinescu Checker Board (0101...) et Inverse (1010..).
- **Nombre de pièces :** 24.

Sur les 24 pièces, 23 pièces ont été mesurées fonctionnelles, ce qui représente un rendement de 95%.

Ces résultats confirment la fonctionnalité de la cellule mémoire ainsi que l'efficacité des assistances en écriture et lecture. Les premières mesures de fréquence semblent cohérentes avec les estimations obtenues précédemment par la simulation. Pour compléter ces travaux, les mesures en consommation permettront de valider le gain en consommation offert à très faible tension. Ces résultats ne pourront toutefois être intégré dans les temps.

8 Récapitulatif de la contribution

Ce chapitre a mis en évidence les développements mémoires à très faible tension suivants :

- Rappel de la métrique statique et définition d'une métrique dynamique.
- Développement d'une méthodologie de chemin critique universel automatisée et adaptative.
- Comparatif des cellules mémoires industrielles et bibliographiques.
- Mise en évidence des choix architecturaux favorables au fonctionnement à très faible tension.
- Conception d'un point mémoire fonctionnel à très faible tension et tension nominale.
- Conception de solutions d'assistance en lecture et en écriture.
- Conception d'un démonstrateur de 128kb fabriqué sur silicium 65nm.
- Simulation de la fonctionnalité et estimation des performances.
- Validation sur silicium de la cellule mémoire et des solutions d'assistance.

Ces travaux ont conduit au développement d'une méthodologie permettant la caractérisation de cellules mémoires autre que les structures classiques à six transistors. Par cette méthodologie, les avantages de différentes architectures disponibles dans la littérature ou dans l'industrie ont été mis en évidence, en termes de comporte-

ment statique et dynamique.

Pour combler les limitations de ces solutions et offrir une fonctionnalité de la tension nominale à la très faible tension, une cellule mémoire à dix transistors a été développée et dimensionnée pour tolérer une variabilité de 6σ à 0,35V.

Des solutions d'assistance en lecture et en écriture ont également été développées pour renforcer cette tolérance et améliorer le comportement dynamique en respect des exigences en fréquence fixées pour l'ensemble des travaux de cette thèse.

L'ensemble de ces travaux ont été intégrés dans un démonstrateur fabriqué en technologie 65nm. Les mesures sur silicium ont permis la validation de la fonctionnalité de la cellule et des solutions d'assistance développées, notamment en offrant un rendement élevé.

9 Publications scientifiques

L'étude sur les cellules mémoire a donnée lieu à la publication de deux brevets [115; 116].

Conclusion

10 Contributions

Cette thèse contribue à la réalisation de dispositifs fonctionnant de la tension nominale à la très faible tension dans le respect des exigences industrielles nécessaires à la production et la commercialisation.

Dans un premier temps, les contributions majeures ont été mises en évidence par une étude approfondie de la littérature sur le sujet : la définition des métriques, les phénomènes de variabilité, et les solutions architecturales pour la conception de cellules logiques, séquentielles, et mémoires.

Les développements ont au départ été basés sur la plateforme technologique 40nm. Des mesures sur silicium de transistors élémentaires ont ensuite permis d'évaluer la validité des modèles de simulation à très faible tension. A la suite de cette étude, une méthodologie de simulation et de dimensionnement a été développée. Indépendante du dimensionnement et de la technologie utilisée, elle a permis l'optimisation de cellules logiques en termes de consommation et de vitesse à une tension optimale $V_{OPTIMALE}$ définie à 0,35V, tout en garantissant le maintien des performances à tension nominale. Cette méthodologie offre des résultats prédictifs validés sur silicium.

Des cellules séquentielles ont également été développées et optimisées. Composées de deux Flip-Flops, une rapide et une robuste aux radiations, ces cellules tolèrent une variabilité de 6σ à 0,35V. Pour permettre la communication avec des domaines de tension différents, des cellules d'interfaces d'entrée et de sortie ont été conçues. Ces cellules réalisent une conversion sur la plage de tension 0,35V à 1,2V tout en maintenant un rapport cyclique correct. Leur tolérance à une variabilité de 6σ à 0,35V a été vérifiée. Enfin, des cellules spécifiques aux arbres d'horloge ont été développées. Elles permettent de réduire considérablement la variabilité par l'utilisation combinée de solutions architecturales et technologiques.

Ces cellules ont été intégrées dans des librairies conçues dans le

respect du format industriel en vigueur, puis caractérisées à tension nominale et à très faible tension grâce à la définition de points de caractérisation PVTSC spécifiques à 0,35V. Leur compatibilité avec le flot de conception industriel usuel a été certifié, et des mesures sur silicium ont permis de valider la méthodologie et la fonctionnalité des cellules à très faible tension. Le fonctionnement sur la plage de tension étendue a également été validé.

Les librairies étant certifiées, et la monotonicité fréquentielle entre la tension nominale et la très faible tension étant vérifiée, un circuit numérique a été fabriqué en technologie 40nm en utilisant le flot de conception de circuit industriel usuel. L'analyse des limitations de ce circuit a permis l'adaptation des différents paramètres du flot de conception telles que les marges temporelles ou la construction de l'arbre d'horlorge. Ces adaptations font apparaitre au niveau des outils de conception les limitations observées sur le silicium, permettant alors leurs corrections. L'efficacité du Body-Biasing dans la compensation de la variabilité a également été vérifiée. Enfin, ce circuit confirme par des mesures de courant dynamique le gain énergétique d'un facteur x10 permis par la réduction de la tension d'alimentation.

Pour compléter l'offre, une méthodologie universelle et adaptable, permettant la caractérisation de cellules mémoires autre que les structures classiques à six transistors, a été développée. Elle a permis la distinction des avantages architecturaux d'une sélection de cellules mémoires industrielles et extraites de la littérature. Cette étude a conduit au développement d'une cellule mémoire offrant une fonctionnalité de la tension nominale à la très faible tension, optimisée pour tolérer une variabilité de 6σ à 0,35V. Des solutions d'assistance en lecture et en écriture ont également été développées pour renforcer cette tolérance et améliorer le comportement dynamique en respect des exigences en fréquence fixées pour l'ensemble des travaux de cette thèse.

Cette mémoire et ces solutions ont été intégrées dans un démonstrateur de 128kb fabriqué en technologie 65nm. Comparé à l'offre industrielle, ce démonstrateur affiche, par simulation, une consommation dynamique réduite. Les mesures sur silicium ont permis de démontrer la faisabilité en affichant un rendement supérieur à 90% à très faible tension, ainsi que l'efficacité des solutions d'assistance en lecture et en écriture.

Au final, l'ensemble de ces contributions, compatibles avec les outils et le flot de conception industriels usuels, constitue une plateforme de développement complète. Ces travaux répondent par conséquent à l'objectif fixé au départ de cette thèse, en offrant les méthodologies et les solutions permettant la réalisation de dispositifs à très faible ten-

sion.

11 Perspectives

Les résultats sur silicium obtenus pour les librairies étant conformes aux prédictions, leur commercialisation est envisageable : à termes, elles intègreront l'offre industrielle.

Cette prédictibilité reste à confirmer pour la conception de circuits numériques. Pour y parvenir, les adaptations développées au cours de cette thèse devront être validées sur silicium, c'est pourquoi un nouveau circuit de test est prévu à cet effet. Il peut également être envisagé de remplacer la méthodologie classique à terme, la tension d'alimentation étant réduite au fil des technologies. Notons que des solutions asynchrones peuvent également lever ces limitations liées aux contraintes temporelles.

Concernant le développement mémoire, bien que la simulation apparait comme concluante tout comme les résultats des mesures sur silicium, la commercialisation des solutions nécessitera de compléter les travaux en proposant une mémoire synchrone et non combinatoire. Cela se traduira par l'insertion de latchs et flip-flops sur les entrées / sorties.

Une normalisation de l'offre est également prévue : à ce jour, des librairies sont disponibles en 40nm, le flot de conception adapté est en 40nm, et la cellule mémoire ainsi que les solutions d'assistance sont en 65nm. Si les librairies de cellule ont été portées en 65nm en faisant appel à la méthodologie mise en place, le portage de la cellule mémoire et des solutions d'assistance en 40nm reste à faire.

Pour conclure, la réalisation d'un démonstrateur réalisant une application telle que l'aide à l'audition, à titre d'exemple, et mettant à profit l'ensemble des développements de cette thèse, permettrait de confirmer l'ensemble des travaux et de renforcer la validité industrielle de la plateforme très faible tension. De même, elle offrirait de nouvelles voies d'améliorations au niveau circuit telles que l'optimisation de la longueur des chemins critiques ou la définition de la surface maximale du circuit. Ces nouveaux travaux renforceraient ainsi les performances pouvant être obtenues à très faible tension.

Bibliographie

[1] Kodak. (2005). [Online]. Available : www.kodak.com 1

[2] R. Rotzoll, S. Mohapatra, V. Olariu, M. Wenz, R. andt Grigas, K. Dimmler, O. Shchekin, and A. Dodabalapur, “Radio frequency rectifiers based on organic thin-film transistors,” in *Appl. Phys. Lett*, 2006, p. 88. 1

[3] M. Lysinger, P. Roche, M. Zamanian, F. Jacquet, N. Sahoo, D. McClure, and J. Russell, “A radiation hardened nano-power 8mb sram in 130nm cmos,” in *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, march 2008, pp. 23 –29. 1

[4] P. Urard, “Soc power-reduction techniques,” in *Solid-State Circuits Conference, 2008. ISSCC 2008. Tutorials. IEEE International*, 2008. 1

[5] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, “Low power cmos digital design,” *IEEE Journal of Solid State Circuits*, vol. 27, pp. 473–484, 1995. 1

[6] P. Urard, L. Paumier, V. Heinrich, N. Raina, and N. Chawla, “A 360mw 105mb/s dvb-s2 compliant codec based on 64800b ldpc and bch codes enabling satellite-transmission portable devices,” in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, feb. 2008, pp. 310 –311. 1

[7] R. Swanson and J. Meindl, “Ion-implanted complementary mos transistors in low-voltage circuits,” *Solid-State Circuits, IEEE Journal of*, vol. 7, no. 2, pp. 146 – 153, apr 1972. 2

[8] E. Vittoz and J. Fellrath, “Cmos analog integrated circuits based on weak inversion operations,” *Solid-State Circuits, IEEE Journal of*, vol. 12, no. 3, pp. 224 – 231, jun 1977. 2

[9] K. Roy and R. Krishnammthy, “Design of low voltage cmos circuits,” in *Circuits and Systems, 2001. Tutorial Guide : ISCAS 2001. The IEEE International Symposium on*, 2001, pp. 3.2.1 –3.2.29. 2

[10] J. Haid, R. Weiss, W. Schogler, and M. Manninger, “Design of an energy-aware system-in-package for playing mp3 in wearable computing devices,” in *SOC Conference, 2003. Proceedings. IEEE International [Systems-on-Chip]*, sept. 2003, pp. 35 – 38. 2

[11] A. Lay-Ekuakille, G. Vendramin, and A. Trotta, “Design of an energy harvesting conditioning unit for hearing aids,” in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, aug. 2008, pp. 2310 –2313. 2

[12] Y. Lee, M. Seok, S. Hanson, D. Blaauw, and D. Sylvester, “Standby power reduction techniques for ultra-low power processors,” in *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, sept. 2008, pp. 186 –189. 2

[13] J. Pigott and K. Parmenter. (2009, July) Ultra-low voltage dc-dc converter capability. Freescale. Ultra-Low Voltage DC-DC Converter. 3

[14] E. Yeatman, “Rotating and gyroscopic mems energy scavenging,” in *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*, april 2006, pp. 4 pp. –45. 3

[15] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, “Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits,” *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305 – 327, feb 2003. 6

[16] Z. Xia, G. Du, Y. Song, J. Wang, X. Liu, J. Kang, and R. Han, “Monte carlo simulation of band-to-band tunneling in silicon devices,” *Japanese Journal of Applied Physics*, vol. 46, no. 4B, pp. 2023–2026, 2007. [Online]. Available : http://jjap.ipap.jp/link?JJAP/46/2023/ 7

[17] W.-K. Yeh and J.-W. Chou, “Optimum halo structure for sub-0.1 mu ;m cmosfets,” *Electron Devices, IEEE Transactions on*, vol. 48, no. 10, pp. 2357 –2362, oct 2001. 7

[18] V. Khandelwal and A. Srivastava, “Active mode leakage reduction using fine-grained forward body biasing strategy,” in *Low Power Electronics and Design, 2004. ISLPED ’04. Proceedings of the 2004 International Symposium on*, aug. 2004, pp. 150 – 155. 7, 24

[19] K.-M. Choi, “32nm high k metal gate (hkmg) designs for low power applications,” in *SoC Design Conference, 2008. ISOCC ’08. International*, vol. 01, nov. 2008, pp. I–68 –I–69. 8

[20] X. Yuan, J.-E. Park, J. Wang, E. Zhao, D. Ahlgren, T. Hook, J. Yuan, V. Chan, H. Shang, C.-H. Liang, R. Lindsay, S. Park, and H. Choo, "Gate-induced-drain-leakage current in 45-nm cmos technology," *Device and Materials Reliability, IEEE Transactions on*, vol. 8, no. 3, pp. 501 –508, sept. 2008. 9

[21] T. Skotnicki, G. Merckel, and T. Pedron, "A new punchthrough current model based on the voltage-doping transformation," *Electron Devices, IEEE Transactions on*, vol. 35, no. 7, pp. 1076 –1086, jul 1988. 9

[22] M. Seok, S. Hanson, D. Sylvester, and D. Blaauw, "Analysis and optimization of sleep modes in subthreshold circuit design," in *Design Automation Conference, 2007. DAC '07. 44th ACM/IEEE*, june 2007, pp. 694 –699. 9

[23] S. Hanson, M. Seok, D. Sylvester, and D. Blaauw, "Nanometer device scaling in subthreshold circuits," in *Design Automation Conference, 2007. DAC '07. 44th ACM/IEEE*, june 2007, pp. 700 –705. 10

[24] S. Hanson, B. Zhai, D. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. Das, W. Haensch, E. Nowak, and D. Sylvester, "Ultralow-voltage, minimum-energy cmos," *IBM J. RES. & DEV*, vol. 90, no. 4, p. 469, SEPTEMBER 2006. 12, 20

[25] D. Blaauw and B. Zhai, "Energy efficient design for subthreshold supply voltage operation," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, 0-0 2006, pp. 4 pp. –32. 13

[26] D. Sylvester, D. Blaauw, M. Papaefthymiou, and M. Flynn, "Low power circuits research at the university of michigan," nov 2006. 13, 36

[27] J. Meindl and J. Davis, "The fundamental limit on binary switching energy for terascale integration (tsi)," *Solid-State Circuits, IEEE Journal of*, vol. 35, no. 10, pp. 1515 –1516, oct 2000. 14

[28] S. Nassif, "Process variability at the 65nm node and beyond," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, sept. 2008, pp. 1 –8. 14

[29] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, and M. Quarantelli, "Variation in transistor performance and leakage in nanometer-scale technologies," *Electron Devices, IEEE Transactions on*, vol. 55, no. 1, pp. 131 –144, jan. 2008. 14

[30] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, “Analysis and mitigation of variability in subthreshold design,” in *Low Power Electronics and Design, 2005. ISLPED ’05. Proceedings of the 2005 International Symposium on*, aug. 2005, pp. 20 – 25. 14

[31] N. Verma, J. Kwong, and A. Chandrakasan, “Nanometer mosfet variation in minimum energy subthreshold circuits,” *Electron Devices, IEEE Transactions on*, vol. 55, no. 1, pp. 163 –174, Jan. 2008. 14

[32] S. Hanson, B. Zhai, D. Blaauw, D. Sylvester, A. Bryant, and X. Wang, “Energy optimality and variability in subthreshold design,” *Low Power Electronics and Design, 2006. ISLPED’06. Proceedings of the 2006 International Symposium on*, pp. 363 –365, Oct. 2006. 15

[33] M. Pelgrom, A. Duinmaijer, and A. Welbers, “Matching properties of mos transistors,” vol. 24, no. 5, Oct 1989, pp. 1433 – 1439. 15

[34] N. Jayakumar and S. Khatri, “A variation-tolerant sub-threshold design approach,” in *Design Automation Conference, 2005. Proceedings. 42nd*, june 2005, pp. 716 – 719. 15

[35] K. Roy, J. Kulkarni, and M.-E. Hwang, “Process-tolerant ultralow voltage digital subthreshold design,” in *Silicon Monolithic Integrated Circuits in RF Systems, 2008. SiRF 2008. IEEE Topical Meeting on*, jan. 2008, pp. 42 –45. 15

[36] B. Calhoun, Y. Cao, X. Li, K. Mai, L. Pileggi, R. Rutenbar, and K. Shepard, “Digital circuit design challenges and opportunities in the era of nanoscale cmos,” *Proceedings of the IEEE*, vol. 96, no. 2, pp. 343 –365, feb. 2008. 15

[37] P. Ampadu, “Ultra-low voltage vlsi : are we there yet ?” in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, 0-0 2006, pp. 4 pp. –24. 15

[38] A. J. Martin, “Towards an energy complexity of computation,” *Inf. Process. Lett.*, vol. 77, pp. 181–187, February 2001. [Online]. Available : http://portal.acm.org/citation.cfm?id=375434.375482 15

[39] D. Markovic, V. Stojanovic, B. Nikolic, M. Horowitz, and R. Brodersen, “Methods for true energy-performance optimization,” *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 8, pp. 1282 – 1293, aug. 2004. 15

[40] A. Raychowdhury, S. Mukhopadhyay, and K. Roy, “A feasibility study of subthreshold sram across technology generations,” in

Computer Design : VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on, oct. 2005, pp. 417 – 422. 16

[41] A. Raychowdhury, B. Paul, S. Bhunia, and K. Roy, “Ultralow power computing with sub-threshold leakage : A comparative study of bulk and soi technologies,” in *Design, Automation and Test in Europe, 2006. DATE ’06. Proceedings*, vol. 1, march 2006, pp. 1 –6. 16

[42] J. Ida, K. Tani, M. Ohono, M. Yanagihara, Y. Igarashi, and K. Sakamoto, “Expanding opportunities of ultra low power and harsh applications with fully depleted (fd) soi (invited),” in *SOI Conference, 2009 IEEE International*, oct. 2009, pp. 1 –3. 17

[43] D. Bol, R. Ambroise, D. Flandre, and J.-D. Legat, “Sub-45nm fully-depleted soi cmos subthreshold logic for ultra-low-power applications,” in *SOI Conference, 2008. SOI. IEEE International*, oct. 2008, pp. 57 –58. 17

[44] J.-J. Kim and K. Roy, “Double gate-mosfet subthreshold circuit for ultralow power applications,” *Electron Devices, IEEE Transactions on*, vol. 51, no. 9, pp. 1468 – 1474, sept. 2004. 17

[45] X. Wu, F. Wang, and Y. Xie, “Analysis of subthreshold finfet circuits for ultra-low power design,” in *SOC Conference, 2006 IEEE International*, sept. 2006, pp. 91 –92. 17

[46] D. Kim, Y. Lee, J. Cai, I. Lauer, L. Chang, S. J. Koester, D. Sylvester, and D. Blaauw, “Low power circuit design based on heterojunction tunneling transistors (hetts),” in *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, ser. ISLPED ’09. New York, NY, USA : ACM, 2009, pp. 219–224. [Online]. Available : http://doi.acm.org/10.1145/1594233.1594287 17, 27

[47] D. Bol, D. Flandre, and J.-D. Legat, “Technology flavor selection and adaptive techniques for timing-constrained 45nm subthreshold circuits,” in *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, ser. ISLPED ’09. New York, NY, USA : ACM, 2009, pp. 21–26. [Online]. Available : http://doi.acm.org/10.1145/1594233.1594240 18

[48] B. Graniello, A. Chavan, B. Rodriguez, and E. MacDonald, “Optimized circuit styles for subthreshold logic,” *International Electro Conference*, Oct 2005. 18

[49] H. Soeleman, K. Roy, and B. Paul, “Sub-domino logic : ultra-low power dynamic sub-threshold digital logic,” in *VLSI Design, 2001. Fourteenth International Conference on*, 2001, pp. 211 –214. 18

[50] H. Soeleman, K. Roy, and B. Paul, “Robust subthreshold logic for ultra-low power operation,” *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 9, no. 1, pp. 90 –99, Feb 2001. 18

[51] A. Tajalli, E. Vittoz, Y. Leblebici, and E. Brauer, “Ultra low power subthreshold mos current mode logic circuits using a novel load device concept,” in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, sept. 2007, pp. 304 –307. 19

[52] C.-I. Kim, H. Soeleman, and K. Roy, “Ultra-low-power dlms adaptive filter for hearing aid applications,” *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 11, no. 6, pp. 1058 – 1067, dec. 2003. 19

[53] S. Mukhopadhyay, C. Neau, R. Cakici, A. Agarwal, C. Kim, and K. Roy, “Gate leakage reduction for scaled devices using transistor stacking,” *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 11, no. 4, pp. 716 – 730, aug. 2003. 20

[54] B. Calhoun and A. Wang, A.and Chandrakasan, “Modeling and sizing for minimum energy operation in subthreshold circuits,” *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 9, pp. 1778 – 1786, Sept. 2005. 20

[55] D. Bol, R. Ambroise, D. Flandre, and J.-D. Legat, “Analysis and minimization of practical energy in 45nm subthreshold logic circuits,” Oct. 2008, pp. 294 –300. 20

[56] I. Sutherland, B. Sproull, and D. Harris, *Logical effort : designing fast CMOS circuits*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1999. 20, 36

[57] J. Keane, H. Eom, T.-H. Kim, S. Sapatnekar, and C. Kim, “Subthreshold logical effort : a systematic framework for optimal subthreshold device sizing,” *Design Automation Conference, 2006 43rd ACM/IEEE*, pp. 425 –428, 0-0 2006. 20

[58] B. Calhoun and A. Chandrakasan, “Characterizing and modeling minimum energy operation for subthreshold circuits,” in *Low Power Electronics and Design, 2004. ISLPED ’04. Proceedings of the 2004 International Symposium on*, 2004, pp. 90 – 95. 20

[59] S. Narendra, D. Blaauw, A. Devgan, and F. Najm, “Leakage issues in ic design :trends, estimation and avoidance,” *Proceedings of*

the International Conference on Computer Aided Design, p. 11, 2003. 21

[60] A. Chavan, G. Dukle, B. Graniello, and E. MacDonald, “Robust ultra-low power subthreshold logic flip-flop design for reconfigurable architectures,” in *Reconfigurable Computing and FPGA’s, 2006. ReConFig 2006. IEEE International Conference on*, sept. 2006, pp. 1 –7. 21

[61] J. Seomun, J. Kim, and Y. Shin, “Skewed flip-flop transformation for minimizing leakage in sequential circuits,” in *Design Automation Conference, 2007. DAC ’07. 44th ACM/IEEE*, june 2007, pp. 103 –106. 21

[62] T.-S. Jau, W.-B. Yang, and Y.-L. Lo, “A new dynamic floating input d flip-flop (dfidff) for high speed and ultra low voltage divided-by 4/5 prescaler,” in *Electronics, Circuits and Systems, 2006. ICECS ’06. 13th IEEE International Conference on*, dec. 2006, pp. 902 –905. 21

[63] B. Fu and P. Ampadu, “Comparative analysis of ultra-low voltage flip-flops for energy efficiency,” in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, may 2007, pp. 1173 –1176. 21

[64] H. Alstad and S. Aunet, “Three subthreshold flip-flop cells characterized in 90 nm and 65 nm cmos technology,” in *Design and Diagnostics of Electronic Circuits and Systems, 2008. DDECS 2008. 11th IEEE Workshop on*, april 2008, pp. 1 –4. 21

[65] T.-H. Kim, J. Keane, H. Eom, and C. Kim, “Utilizing reverse short-channel effect for optimal subthreshold circuit design,” *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 15, no. 7, pp. 821 –829, july 2007. 22

[66] Y. Morita, H. Fujiwara, H. Noguchi, Y. Iguchi, K. Nii, H. Kawaguchi, and M. Yoshimoto, “An area-conscious low-voltage-oriented 8t-sram design under dvs environment,” in *VLSI Circuits, 2007 IEEE Symposium on*, june 2007, pp. 256 –257. 22

[67] N. Verma and A. Chandrakasan, “A 256 kb 65 nm 8t subthreshold sram employing sense-amplifier redundancy,” *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 1, pp. 141 –149, jan. 2008. 22

[68] F. Moradi, D. Wisland, S. Aunet, H. Mahmoodi, and T. V. Cao, “65nm sub-threshold 11t-sram for ultra low voltage applications,” in *SOC Conference, 2008 IEEE International*, sept. 2008, pp. 113 –118. 22

[69] B. Calhoun and A. Chandrakasan, “A 256-kb 65-nm sub-threshold sram design for ultra-low-voltage operation,” *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 3, pp. 680 –688, march 2007. 22, 27

[70] J. Kulkarni, K. Kim, and K. Roy, “A 160 mv robust schmitt trigger based subthreshold sram,” *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 10, pp. 2303 –2313, oct. 2007. 22, 27

[71] I. J. Chang, J.-J. Kim, S. Park, and K. Roy, “A 32kb 10t subthreshold sram array with bit-interleaving and differential read scheme in 90nm cmos,” in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, feb. 2008, pp. 388 –622. 22, 105, 121

[72] K. Itoh, M. Horiguchi, and T. Kawahara, “Ultra-low voltage nano-scale embedded rams,” in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, 0-0 2006, pp. 4 pp. –28. 22

[73] J. Lee, Y. J. Lee, and Y. B. Kim, “Sram word-oriented redundancy methodology using built in self-repair,” in *SOC Conference, 2004. Proceedings. IEEE International*, sept. 2004, pp. 219 – 222. 22

[74] H. Shao and C.-Y. Tsui, “A robust, input voltage adaptive and low energy consumption level converter for sub-threshold logic,” in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, sept. 2007, pp. 312 –315. 23

[75] R. K. Ik Joon Chang, Jae-Joon Kim, “Robust level converter design for sub-threshold logic,” *Low Power Electronics and Design, 2006. ISLPED’06. Proceedings of the 2006 International Symposium on*, pp. 14 –19, Oct. 2006. 23

[76] K. von Arnim, E. Borinski, P. Seegebrecht, H. Fiedler, R. Brederlow, R. Thewes, J. Berthold, and C. Pacha, “Efficiency of body biasing in 90-nm cmos for low-power digital circuits,” *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 7, pp. 1549 – 1556, july 2005. 24

[77] C. Neau and K. Roy, “Optimal body bias selection for leakage improvement and process compensation over different technology generations,” in *Low Power Electronics and Design, 2003. ISLPED ’03. Proceedings of the 2003 International Symposium on*, aug. 2003, pp. 116 – 121. 24

[78] C. Kim and K. Roy, “Dynamic vth scaling scheme for active leakage power reduction,” in *Design, Automation and Test in Europe Conference and Exhibition, 2002. Proceedings*, 2002, pp. 163 – 167. 24

[79] J. Tschanz, S. Narendra, R. Nair, and V. De, “Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors,” *Solid-State Circuits, IEEE Journal of*, vol. 38, no. 5, pp. 826 – 829, may 2003. 24

[80] H. Ananthan, C. Kim, and K. Roy, “Larger-than-vdd forward body bias in sub-0.5v nanoscale cmos,” in *Low Power Electronics and Design, 2004. ISLPED ’04. Proceedings of the 2004 International Symposium on*, aug. 2004, pp. 8 – 13. 24

[81] G. l. Giuseppe De Vita, “A voltage regulator for subthreshold logic with low sensitivity to temperature and process variations,” *ISSCC*, 2007. 25

[82] M. Meijer, F. Pessolano, and J. Pineda De Gyvez, “Technology exploration for adaptive power and frequency scaling in 90nm cmos,” in *Low Power Electronics and Design, 2004. ISLPED ’04. Proceedings of the 2004 International Symposium on*, aug. 2004, pp. 14 – 19. 25

[83] B. Calhoun and A. Chandrakasan, “Ultra-dynamic voltage scaling (udvs) using sub-threshold operation and local voltage dithering,” *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 1, pp. 238 – 245, jan. 2006. 25

[84] A. Wang and A. Chandrakasan, “A 180-mv subthreshold fft processor using a minimum energy design methodology,” *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 1, pp. 310 – 319, Jan. 2005. 25

[85] M.-E. Hwang, A. Raychowdhury, K. Kim, and K. Roy, “A 85mv 40nw process-tolerant subthreshold 8x8 fir filter in 130nm technology,” in *VLSI Circuits, 2007 IEEE Symposium on*, june 2007, pp. 154 –155. 25

[86] J.-S. Wang, J.-S. Chen, Y.-M. Wang, and C. Yeh, “A 230mv-to-500mv 375khz-to-16mhz 32b risc core in 0.18 μm cmos,” in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, feb. 2007, pp. 294 –604. 25

[87] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw, “Performance and variability optimization strategies in a sub-200mv, 3.5pj/inst, 11nw subthreshold processor,” in *VLSI Circuits, 2007 IEEE Symposium on*, june 2007, pp. 152 –153. 25

[88] R. Jain, P. Guttal, and D. Parent, “6 bit decimation filter in sub-threshold region,” in *University/Government/Industry Microelectronics Symposium, 2006 16th Biennial*, june 2006, pp. 215 –219. 25

[89] B. Paul, H. Soeleman, and K. Roy, “An 8x8 sub-threshold digital cmos carry save array multiplier,” in *Solid-State Circuits Conference, 2001. ESSCIRC 2001. Proceedings of the 27th European*, sept. 2001, pp. 377 – 380. 25

[90] D. Wolpert and P. Ampadu, “An ultra-low voltage 200 mhz 0.6 pj add-compare-select unit in 180 nm cmos,” in *Circuits and Systems, 2006. MWSCAS ’06. 49th IEEE International Midwest Symposium on*, vol. 1, aug. 2006, pp. 32 –35. 25

[91] J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann, and A. Chandrakasan, “A 65nm sub-vt microcontroller with integrated sram and switched-capacitor dc-dc converter,” in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, feb. 2008, pp. 318 –616. 25

[92] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, “A sub-200mv 6t sram in 0.13 μm cmos,” in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, feb. 2007, pp. 332 –606. 27

[93] L. Haixia, L. Weimin, T. Jianping, L. Shijin, and C. Lei, “Design of a low power radiation hardened 256k sram,” in *Solid-State and Integrated Circuit Technology, 2006. ICSICT ’06. 8th International Conference on*, oct. 2006, pp. 1646 –1648. 27

[94] A. Bhavnagarwala, S. Kosonocky, Y. Chan, K. Stawiasz, U. Srinivasan, S. Kowalczyk, and M. Ziegler, “A sub-600mv, fluctuation tolerant 65nm cmos sram array with dynamic cell biasing,” in *VLSI Circuits, 2007 IEEE Symposium on*, june 2007, pp. 78 –79. 27

[95] Y. Morita, H. Fujiwara, H. Noguchi, K. Kawakami, J. Miyakoshi, S. Mikami, K. Nii, H. Kawaguchi, and M. Yoshimoto, “A vth-variation-tolerant sram with 0.3-v minimum operation voltage for memory-rich soc under dvs environment,” in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 0-0 2006, pp. 13 –14. 27

[96] M. Yamaoka and T. Kawahara, “Operating-margin-improved sram with column-at-a-time body-bias control technique,” in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, sept. 2007, pp. 396 –399. 27

[97] J. Chen, L. Clark, and T.-H. Chen, “An ultra-low-power memory with a subthreshold power supply voltage,” *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 10, pp. 2344 –2353, oct. 2006. 27

[98] A. Agarwal, N. Banerjee, S. Hsu, R. Krishnamurthy, and K. Roy, “A 200mv to 1.2v, 4.4mhz to 6.3ghz, 48x42b 1r/1w programmable register file in 65nm cmos,” in *Solid State Circuits Conference, 2007. ESSCIRC 2007. 33rd European*, sept. 2007, pp. 316 –319. 27

[99] T. Masson and R. Ferrant, “Memory insensitive to disturbances,” U.S. Patent 5 570 313, Oct. 29, 1996. 57

[100] S. Clerc, J.-P. Schoelkopff, F. Firmin, and F. Abouzeid, “Single phase clock and low power dynamic flip-flop,” FR Patent 0 957 053, Oct. 9, 2009. 57, 68

[101] F. Abouzeid, F. Firmin, and S. Clerc, “Method of electronic logic synthesis,” US Patent 12 853 627, Aug. 10, 2010. 68

[102] F. Abouzeid, S. Clerc, M. Renaudin, and G. Sicard, “Design solutions for ultra-low voltage,” *7ème Journées Faible Tension Faible Consommation*, May 2008. 68

[103] F. Abouzeid, S. Clerc, M. Renaudin, and G. Sicard, “45nm ultra-low voltage design,” *European Solid-State Circuits Conference*, September 2008. 69

[104] F. Abouzeid, S. Clerc, M. Renaudin, and G. Sicard, “Ultra-low voltage from 65nm to 32nm,” *8ème Journées Faible Tension Faible Consommation*, June 2009. 69

[105] F. Abouzeid, S. Clerc, F. Firmin, M. Renaudin, and G. Sicard, “A 45nm cmos 0.35v-optimized standard cell library for ultra-low power applications.” in *ISLPED*. ACM, August 2009, pp. 225–230. 69

[106] M. Darnell, “Error control coding : Fundamentals and applications,” *Communications, Radar and Signal Processing, IEE Proceedings F*, vol. 132, no. 1, p. 68, February 1985. 75

[107] M. Keating, D. Flynn, R. Aitken, A. Gibbons, and K. Shi, “Multi-voltage design,” in *Low Power Methodology Manual*. Springer US, 2007, pp. 21–31. 87

[108] A. Chattopadhyay and Z. Zilic, “High speed asynchronous structures for inter-clock domain communication,” in *Electronics, Circuits and Systems, 2002. 9th International Conference on*, vol. 2, 2002, pp. 517 – 520 vol.2. 88

[109] A. Martin and M. Nystrom, “Asynchronous techniques for system-on-chip design,” *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1089 –1120, june 2006. 88

[110] S. Clerc, F. Abouzeid, V. Heinrich, A. Jain, A. Veggetti, D. Crippa, P. Roche, and G. Sicard, “A 40nm cmos, 1.27nj, 330mv, 600khz, bose chaudhuri hocquenghem 252 bits frame decoder,” in *ICICDT*. IEEE, jun. 2010, pp. 78 –81. 90

[111] J. Kulkarni, K. Kim, S. Park, and K. Roy, “Process variation tolerant sram array for ultra low voltage applications,” in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, june 2008, pp. 108 –113. 105

[112] MunEDA. Wicked. [Online]. Available : http://www.muneda.com/ 110

[113] B. Wicht, *Current Sense Amplifiers : for Embedded SRAM in High-Performance System-on-a-Chip Designs*. Springer, July 2003. 114

[114] C. Dray, F. Jacquet, and S. Barasinski, “Sram memory device with improved write operation and method thereof,” US Patent 7 751 229, July 6, 2010. 118

[115] F. Abouzeid and S. Clerc, “Ultra-low voltage 10t sram,” FR Patent 1 051 043, Feb. 15, 2010. 128

[116] F. Abouzeid, S. Clerc, and P. Roche, “Boost circuit for memory device,” FR Patent 1 057 945, Sept. 30, 2010. 128

Bibliographie de l’auteur

Publications

F. Abouzeid, S. Clerc, M. Renaudin, and G. Sicard, "Design solutions for ultra-low voltage," 7ème Journées Faible Tension Faible Consommation, May 2008.

F. Abouzeid, S. Clerc, M. Renaudin, and G. Sicard, "45nm ultra-low voltage design," European Solid-State Circuits Conference, September 2008.

F. Abouzeid, S. Clerc, M. Renaudin, and G. Sicard, "Conception de circtuis intégrés à très faible tension," JNRDM, May 2009.

F. Abouzeid, S. Clerc, M. Renaudin, and G. Sicard, "Ultra-low voltage from 65nm to 32nm," 8ème Journées Faible Tension Faible Consommation, June 2009.

F. Abouzeid, S. Clerc, F. Firmin, M. Renaudin, and G. Sicard, "A 45nm cmos 0.35v-optimized standard cell library for ultra-low power applications." in ISLPED. ACM, August 2009, pp. 225230.

S. Clerc, F. Abouzeid, V. Heinrich, A. Jain, A. Veggetti, D. Crippa, P. Roche, and G. Sicard, "A 40nm cmos, 1.27nj, 330mv, 600khz, bose chaudhuri hocquenghem 252 bits frame decoder," in ICICDT. IEEE, jun. 2010, pp. 78 81.

F. Abouzeid, S. Clerc, F. Firmin, M. Renaudin, and G. Sicard, "40nm cmos 0.35v-optimized standard cell libraries for ultra-low power applications." in TODAES. ACM, 2010 (en révision).

Brevets

S. Clerc, J.-P. Schoelkopff, F. Firmin, and F. Abouzeid, "Single phase clock and low power dynamic flip-flop," FR Patent 0 957 053, Oct. 9, 2009.

F. Abouzeid, F. Firmin, and S. Clerc, "Method of electronic logic synthesis," US Patent 12 853 627, Aug. 10, 2010.

F. Abouzeid and S. Clerc, "Ultra-low voltage 10t sram," FR Patent 1 051 043, Feb. 15, 2010.

F. Abouzeid, S. Clerc, and P. Roche, "Boost circuit for memory device," FR Patent 1 057 945, Sept. 30, 2010.

Index

Résumé

L'alimentation des circuits à très faible tension, permettant une efficacité énergétique multipliée par 10, répond aux contraintes des applications mobiles, au prix d'une variabilité accrue limitant la prédiction des résultats et nécessitant des efforts et méthodologies de conception spécifiques. Cette thèse associe la conception à très faible tension aux exigences industrielles, et présente le développement de cellules digitales optimisées pour la très faible tension, par une méthodologie indépendante de la technologie. Ces cellules, validées par des mesures sur silicium en technologie CMOS 40nm, ont conduit à la fabrication d'un circuit numérique, dont le test met en évidence les adaptations permettant d'améliorer le rendement. Enfin, une cellule mémoire a été conçue et optimisée à très faible tension, ainsi que des solutions d'assistance en lecture et en écriture pour renforcer la tolérance à la variabilité. Un démonstrateur 128kb est fabriqué en 65nm pour valider ces développements.

Mot-clefs

Sous-le-seuil, Énergie, CMOS, SRAM, circuit, logique, variabilité

Title

Subthreshold architecture and digital circuits study in submicronic CMOS technology

Abstract

Ultra-low voltage enables to answer the limitations of the wearable mobile applications with an energy efficiency improved by a factor x10, at the price of an increased transistor variability limiting the predictability of the results.
In respect with the industrial requirements, this thesis presents the development of logical cells optimized at ultra-low voltage, using a technology independent methodology. These cells, certified then validated by silicon measurements in 40nm, led to the design of a digital circuit, fabricated on silicon, which analysis highlighted the adaptations needed to enhance the yield and the predictability of the results. At last, a memory cell was developed and optimized at ultra-low voltage. Read and write assist solutions were conceived in order to reinforce the tolerance to variability. A 128kb memory demonstrator was then fabricated in 65nm to validate these developments.

Keywords

Subthreshold, Energy, CMOS, SRAM, Circuit, Logic, Variability

Adrr : Laboratoire TIMA, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex France
ISBN : 978-2-84813-160-3