UNIVERSITE CATHOLIQUE DE LOUVAIN

Ecole Polytechnique de Louvain

Département d'Electricité

Développement de circuits mémoire à ultra-basse consommation pour applications portables en technologie bulk 130 nm

Promoteur : Professeur D. Flandre

Mémoire présenté en vue de l'obtention du grade d'ingénieur civil en électromécanique section mécatronique par

Julien De Vos

Louvain-la-Neuve

Année académique 2007-2008

Je tiens à remercier mon promoteur, le Professeur Denis Flandre pour ses remarques et ses conseils avisés tout au long de ce travail ainsi que M. David Bol pour son aide précieuse et sa grande disponibilité pour répondre à mes nombreuses questions. Je tiens également à les remercier tous les deux de m'avoir offert l'occasion de rédiger un article sur ce travail. J'adresse également mes remerciements à Mme Brigitte Dupont pour son aide lorsque certains problèmes techniques ont été rencontrés sur les plateformes SUN de l'Université. Enfin, je remercie toutes les personnes qui m'ont soutenu durant cette année et qui ont lu ce texte en vue d'en améliorer la qualité.

Résumé

L'évolution des technologies CMOS vise une réduction de la taille des transistors. S'il est ainsi possible d'améliorer la densité d'intégration et les performances de ces dispositifs, les courants de fuite deviennent significatifs. Ceux-ci ont un impact direct sur la durée de vie des batteries des nombreuses applications portables que l'on rencontre actuellement sur le marché. Ce problème se pose surtout dans le cas des mémoires SRAMs qui occupent une place significative des circuits sur puce. En effet lors des mises en veille, il n'est pas possible de couper l'alimentation de ces dernières sans perdre l'information stockée. Par ailleurs, la variabilité introduite lors de la fabrication rend difficile l'obtention de SRAMs dont la robustesse est garantie.

Dans ce contexte, une nouvelle architecture de cellule SRAM a été développée. Elle est basée sur une nouvelle technique CMOS permettant de réduire les pertes statiques dues aux courants de fuite. L'assemblage de deux MOSFETs permet de réaliser un dispositif que nous appellerons « transistor ULP ». Les courants de fuite de celui-ci sont réduits de plusieurs ordres de grandeur grâce à l'application de tensions de polarisation négatives auto-induites entre grille et source. A l'aide de deux transistors ULP, il est alors possible de réaliser des inverseurs dont la courbe de transfert DC possède une hystérèse. Ils permettent de réaliser différentes cellules mémoires. Parmi celles-ci une cellule comportant 12 transistors apporte une solution au besoin d'obtenir une mémoire possédant à la fois de très faibles pertes statiques et une grande stabilité tout en maintenant de bonnes performances dynamiques. En effet, l'hystérèse présente sur la courbe caractéristique de l'inverseur ULP permet d'amener la marge de bruit de la cellule au-delà de $V_{DD}/2$, tandis qu'un buffer composé de deux transistors garantit un faible temps de lecture. Des simulations ont été réalisées en technologie bulk 130 nm. Les pertes statiques de la cellule ne sont que de 13 pW. Cela représente une réduction d'un facteur 55 par rapport à une cellule 6 transistors conventionnelle. De manière plus particulière, une étude de la robustesse face aux variations environnementales et technologiques de la cellule 12 transistors a été réalisée et montre de bons résultats. Des essais réalisés à l'aide d'un modèle prédictif 45 nm ont prouvé la fonctionnalité de la cellule sous les technologies futures, à condition de prendre certaines mesures pour contrer l'impact des courants de grille.

Abstract

CMOS technology evolution targets a reduction of transistor size. Yet it does allow to improve the integration density and the devices performance, leakage currents start to become important. These have an impact on the battery life time of mobile applications that can now be found on the market. This problem is a primary concern for SRAM memories that occupy a large area of systems-on-chip. Indeed during stand-by operation it is not possible to shut their power down without losing stored data. Moreover process variability makes it difficult to get SRAMs with guarantied robustness.

In this context a new SRAM architecture has been developed. It is based on a new CMOS technique that allows to cut off static leakage. Combination of two MOSFETs can lead to a device called « ULP transistor ». Its leakage currents are reduced by orders of magnitude thanks to negative gate to source self-biasing. Inverters can be build, based on ULP transistors. These inverters show hysteresis in their DC transfer curve. They can be used to build memory cells. One of these cells counts 12 transistors and gives a solution to the need of getting a memory with low static power and great stability, while keeping good dynamic performances. Indeed the hysteresis of the ULP inverter allow a static noise margin higher than $V_{DD}/2$ and a buffer made of two transistors guaranties a small read time. Simulations have been made in the 130 nm bulk technology. The static power consumption is only 13 pW. This represents a reduction by a factor 55 as compared to a 6-transistor conventional cell. More specifically, a study of the robustness of the 12-transistor cell to environmental and process variations has been done, exhibiting good results. Finally, trials with a predictive 45 nm models prove the functionality of the cell for technologies to come, provided that steps are taken to mitigate gate leakage.

Table des matières

1. Introduction

Ces dernières années ont vu la diminution continue de la taille des transistors sur les circuits intégrés CMOS. Cette évolution permet d'atteindre des densités d'intégration et des performances de plus en plus élevées, mais elle s'accompagne également d'une augmentation importante des courants de fuite. Ainsi, une des préoccupations principales des concepteurs de circuits intégrés, notamment pour produits portables, est devenue la réduction de la puissance dissipée.

Le laboratoire de microélectronique (DICE) de l'Ecole Polytechnique de Louvain a déposé récemment des brevets concernant l'utilisation d'un dispositif permettant de réduire significativement les pertes statiques dues aux courants de fuite des transistors des circuits CMOS. Une thèse liée à ces recherches, menée par D. Levacq, a abouti à l'élaboration d'une mémoire statique à accès aléatoire (SRAM) basée sur des diodes à « ultra basse consommation » (ULP). D'autres études actuellement en cours, et toujours fondées sur ces brevets, ont pour objectif le développement de circuits logiques basés sur l'utilisation de transistors ULP.

Par ailleurs, les grands acteurs du monde des semi-conducteurs se réunissent régulièrement pour définir les grands défis à venir des prochaines années. Cette rencontre mène à l'élaboration de l'« *International Technology Roadmap for Semiconductors* » (ITRS). Ce document permet de définir les principaux objectifs que devra rencontrer l'industrie de silicium afin de garantir l'évolution constante de ce domaine. D'après l'édition 2007 [1], des obstacles seront rencontrés dans le maintien des performances futures des SRAMs. En effet, avec la réduction des dimensions des transistors, outre les défis propres aux procédés de fabrication, il sera difficile de garantir la stabilité de la cellule. Il ne sera pas non plus aisé de limiter les courants de fuite. Pourtant, l'ITRS rappelle également que les SRAMs sont capitales pour garantir les performances de nombreux systèmes étant donné qu'elles sont utilisées sur les circuits à cadence élevée. Pour comprendre cette préoccupation, il faut savoir que les SRAMs représentent une partie importante de la surface totale et de la puissance consommée par de nombreux systèmes sur puce (SoC) [2]. De plus, contrairement aux autres blocs logiques, il est impossible de déconnecter l'alimentation des SRAMs afin de réduire leur courant de fuite durant les périodes de mise en veille ou "standby". La consommation statique devient donc une part significative des pertes de ces dispositifs.

C'est dans ce cadre rempli de défis que j'ai choisi de réaliser mon travail de fin d'études. En effet, si l'on excepte l'étude de D. Levacq, un regard critique sur les recherches en cours sur le

développement de SRAMs montre qu'il n'existe actuellement pas de réelles solutions pour maintenir des performances dynamiques élevées tout en réduisant la consommation statique de la cellule et en améliorant sa stabilité. C'est pourquoi, une nouvelle architecture de cellule, basée sur l'utilisation d'inverseurs à ultra faible consommation (ULP), a été élaborée. Elle permet d'apporter une solution valable à ces deux problèmes si l'on se permet d'occuper une plus grande surface de silicium.

Après avoir étudié les performances de différentes cellules SRAM, il ressort qu'une cellule composée de 12 transistors permet de répondre aux différents besoins énoncés ci-dessus. Cette cellule est basée sur l'utilisation de transistors ULP. Leurs courants de fuite sont réduits de plusieurs ordres de grandeur grâce à l'application auto-induite d'une tension V_{GS} négative. Deux de ces transistors peuvent être assemblés pour former un inverseur ULP. Celui-ci montre une réduction de la puissance statique de 3 ordres de grandeur, alors que son délai est significativement accru. La caractéristique DC de cet élément possède une hystérèse qu'il sera possible de mettre en valeur afin d'améliorer la stabilité d'un latch réalisé à l'aide de deux inverseurs ULP. La cellule ULP 12 transistors est élaborée sur base d'un tel latch. En outre, deux transistors permettent de réaliser une opération de lecture. De cette manière, la stabilité de la cellule n'est pas perturbée lors de cette opération, comme cela est le cas pour une cellule à 6 transistors conventionnelle, et les performances dynamiques sont maintenues malgré le délai élevé de l'inverseur ULP. Ainsi, la cellule possède une marge de bruit proche de 0.75 V_{DD}, et sa consommation statique ne s'élève qu'à 13 pW, ce qui représente une réduction par un facteur 55 des pertes statiques de la cellule 6 transistors conventionnelle. Une analyse réalisée à l'aide d'un modèle prédictif montre que pour les technologies futures, la cellule sera sensible à l'accroissement des courants de grille. Cependant, une série des mesures adéquates permet de maintenir ses fonctionnalités tout en préservant une très bonne stabilité.

Afin de pouvoir développer cette étude, il a été nécessaire d'apprendre à utiliser les différents outils du chercheur en microélectronique. C'est ainsi que certains logiciels de simulations tels qu'ELDO ont été largement mis à contribution afin de pouvoir analyser les performances des cellules réalisées à l'aide de modèles industriels de type BSIM3. Par ailleurs, pour évaluer la surface occupée par la cellule, le logiciel Cadence a été utilisé dans le but de réaliser certains layouts. Enfin, il ne faudrait pas oublier de mentionner le regard critique qu'il a fallu porter sur un nombre conséquent de publications actuellement parues dans le monde des mémoires. Cette mise en situation s'est soldée par la publication d'un article portant sur ce travail [3] et par la participation à une parution liée au développement de portes logiques à faible consommation [4].

La première partie de ce travail introduit le fonctionnement de cellule mémoire classique réalisée à l'aide de 6 transistors. Les différentes opérations réalisées sur cette cellule y sont décrites ainsi que l'architecture générale d'une mémoire. Ensuite, un tour d'horizon des techniques actuellement mises en œuvre pour réduire la consommation des cellules SRAMs a été effectué. Il permet de mettre en évidence le fait qu'il n'existe pas à ce jour de solutions disponibles pour les applications requérant de bonnes performances dynamiques et une grande stabilité.

Dans la deuxième partie, une nouvelle architecture CMOS ULP est présentée. En regroupant deux MOSFETs ensemble, nous verrons qu'il est possible de créer une nouvelle structure possédant les caractéristiques d'un transistor dont les courants de fuite sont réduits de plusieurs ordres de grandeur. L'inverseur ULP sera également présenté.

Dans la troisième partie, différentes architectures basées sur l'utilisation de transistors ULP seront étudiées. Cette analyse portera sur le fonctionnement dynamique de ces cellules ainsi que sur

leur stabilité en rétention et durant les opérations de lecture et d'écriture. A cet effet, certaines méthodes permettant de caractériser cette stabilité devront être adaptées.

La quatrième partie tend à montrer la robustesse d'une cellule à 12 transistors face à différentes familles de perturbations. En effet, il faut pouvoir garantir le bon fonctionnement de la cellule sous une certaine plage de température et malgré une réduction de la tension d'alimentation. Par ailleurs, durant la fabrication, les caractéristiques des transistors peuvent présenter un écart par rapport aux prédictions, localement ou sur l'ensemble de la puce. Il faut alors garantir l'opérabilité de la cellule malgré ces variations.

Enfin, les dernières parties de ce travail présenteront un exemple de layout d'une cellule à 12 transistors et différentes pistes de solutions pour améliorer ses performances. La cellule sera également simulée à l'aide de modèle prédictif de type BSIM4 afin de prouver sa fonctionnalité dans les technologies futures.

2. Les mémoires SRAM

2.1. Introduction

L'évolution de ces dernières années tend à montrer que les mémoires SRAM occupent une surface de plus en plus importante dans les circuits sur puces et microprocesseurs, si bien que, comme nous pouvons l'observer à la figure 2.1, dans certains microprocesseurs pas moins de 30 % de l'aire disponible est occupée par celles-ci [5] [6]. Par ailleurs, à chaque nouveau nœud technologique les courants de fuite des circuits CMOS gagnent en importance. Cette situation mène à la constatation que la principale source de dissipation de puissance des SRAMs, et par extension des microprocesseurs et SoC, est liée aux pertes statiques [5] [7]. Les applications portables se multipliant, il devient dès lors primordial pour la durée de vie de leur batterie d'apporter certaines solutions au niveau circuit pour restreindre la consommation de l'électronique embarquée.

Avant de présenter notre contribution, ce chapitre vise à présenter les principales caractéristiques d'une mémoire SRAM, et poursuit en donnant un état de l'art des différentes solutions proposées dans la littérature pour réduire les courants consommés de celles-ci.

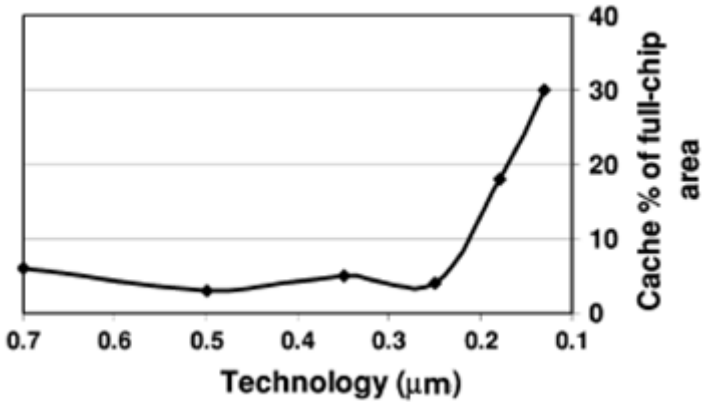


Fig. 2.1 : Evolution du pourcentage de l'aire occupée par les SRAMs en fonction du nœud technologique sur les circuits VLSI (figure reprise de [5]).

2.2. Fonctionnement d'une SRAM

2.2.1. Contexte

Nous allons ici tenter de cerner les domaines d'applications des SRAMs. Pour ce faire, il est utile d'en comprendre les propriétés et de réaliser une comparaison avec d'autres types de mémoires.

Par opposition à de nombreuses mémoires telles que ROM, EPROM…, les DRAMs et SRAMs sont volatiles. Cela signifie que lorsqu'elles ne sont plus alimentées, leurs données sont perdues. Par ailleurs, les SRAMs sont dites à accès aléatoire (RAM pour « random access memory »). Cette dénomination signifie que l'on peut accéder indépendamment à n'importe quelle cellule de la mémoire, que ce soit en lecture ou en écriture.

Pour fonctionner, les SRAMs stockent une valeur grâce à deux inverseurs bouclés sur eux-mêmes. Si l'on ajoute deux transistors d'accès, on peut dénombrer, à la figure 2.2 (a), 6 transistors par cellule mémoire (on parle donc de cellule 6T). Elles occupent ainsi une aire plus importante que les

DRAMs, qui sont représentées à la Fig. 2.2 (b). Celles-ci mémorisent une valeur binaire sur une capacité par l'intermédiaire d'un transistor d'accès. Il faut noter que petit à petit, les courants de fuite de ce transistor vont décharger la capacité et que, contrairement aux SRAMs, il est ici nécessaire de réaliser périodiquement une opération de rafraichissement. Un autre inconvénient des DRAMs est que le temps d'accès en lecture ou en écriture est plus élevé que dans le cas de la SRAM. Par temps d'accès, il faut comprendre le temps compris entre l'activation de la cellule mémoire et la disponibilité de la valeur en sortie (dans le cas d'une lecture), ou l'écriture de la valeur mémorisée (dans le cas d'une écriture). Le tableau 2.1 permet de comparer qualitativement ces deux types de mémoires.

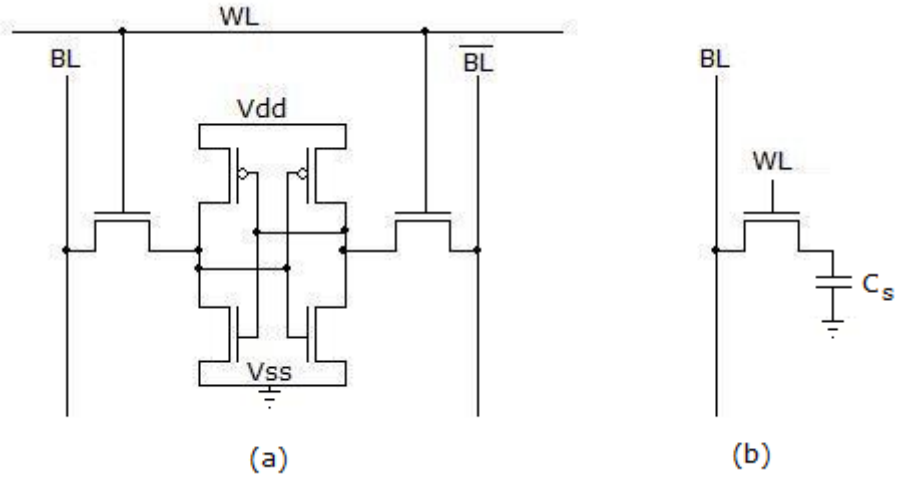


Fig. 2.2 (a) Cellule SRAM 6 transistors, (B) Cellule DRAM 1 transistor.

	densité	vitesse	cout	rafraichissement	volatile
SRAM	faible	rapide	cher	Non	Oui
DRAM	élevée	moyenne	moyen	Nécessaire	Oui

Tableau 2.1 : Comparaison qualitative des mémoires SRAM et DRAM.

Nous pouvons comprendre grâce à cette comparaison que la mémoire SRAM sera généralement utilisée dans de plus petits tableaux que la DRAM étant donné son coût plus élevé. Par contre, sa vitesse d'accès sera son principal atout. C'est pourquoi elle sert fréquemment de mémoire cache aux processeurs. Son rôle est alors de contenir une série de variables auxquelles on accède fréquemment. Lorsque le contenu d'une variable doit être connu, le processeur examinera d'abord si elle est disponible dans la cache avant d'inspecter les mémoires plus lentes d'accès.

2.2.2. Opérations générales sur une cellule SRAM

La figure 2.3 représente une mémoire SRAM typique. Dans ce cas-ci, il existe 2^{m+n} cellules réparties en m lignes et n colonnes. Lors de chaque opération, un signal d'adresse de taille m+n permettra de sélectionner une wordline et une paire de bitlines à l'aide des décodeurs d'adresse. L'intersection de la colonne et de la ligne ainsi sélectionnées définit l'unique cellule sur laquelle une opération de lecture ou d'écriture est réalisée[1]. Les circuits périphériques permettent de faire

[1] En réalité, plusieurs accès sont réalisés simultanément sur des cellules différentes de manière à écrire ou lire un mot binaire. Cependant, en vue de simplifier les explications, nous supposerons dans ce travail de fin d'études que l'accès se fait toujours sur une seule cellule.

fonctionner la mémoire SRAM, c'est par le biais de ceux-ci que les valeurs mémorisées sont transmises à l'extérieur, qu'une horloge est fournie à la mémoire, etc.

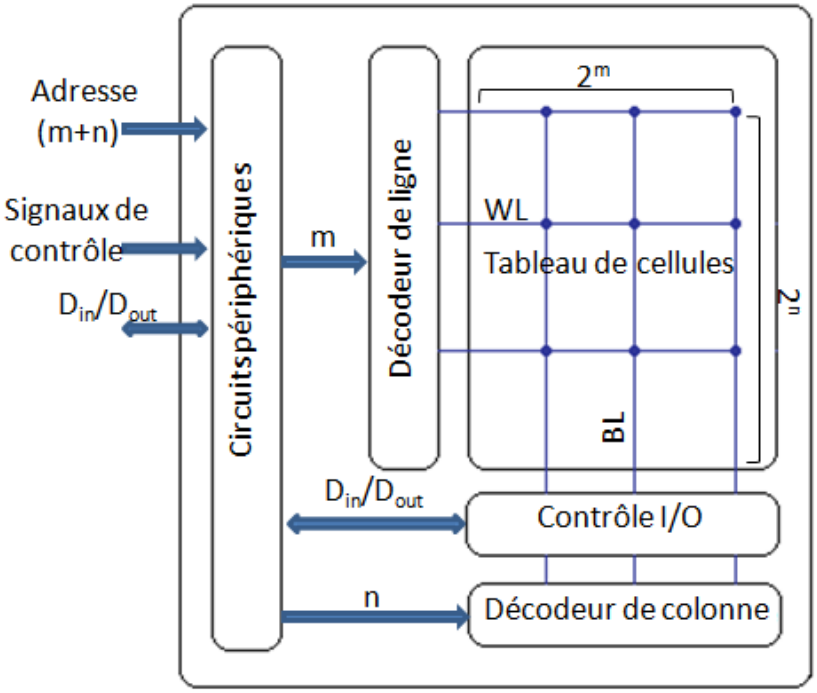


Fig. 2.3 Topologie d'une mémoire SRAM.

Il faut noter que sur cette représentation, il n'existe qu'un seul banc mémoire. Cependant, une pratique courante consiste à diviser les wordlines en plusieurs sections. La mémoire est alors divisée en plusieurs bancs comme illustré à la figure 2.4, typiquement de 2 à 8 [8]. Cette astuce permet de gagner en termes de délais et de puissance active grâce à une réduction de la capacité de wordline et de cellules activées simultanément. Cependant, si le nombre de bancs devient excessif, l'accroissement de délai lié aux nouveaux circuits de prédécodage peut devenir le plus important.

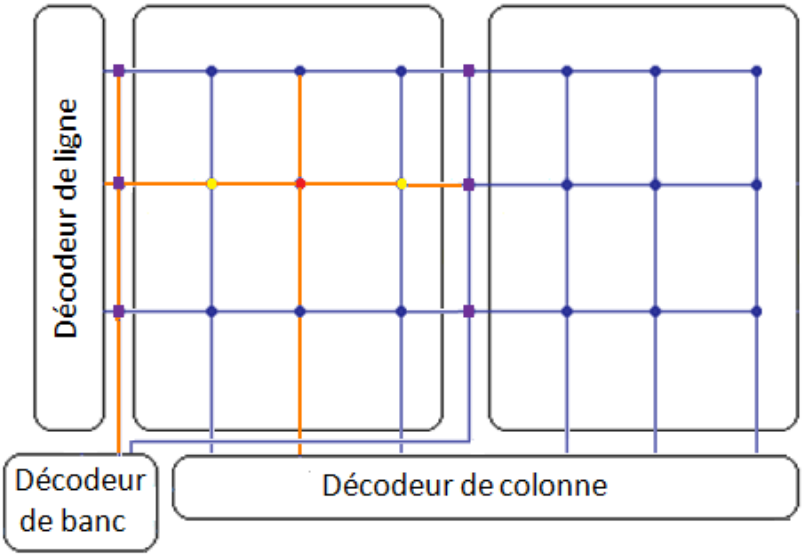


Fig. 2.4 La division de la SRAM en bancs de données permet d'améliorer les performances.

2.2.3. Opérations de lecture et écriture sur une SRAM [9]

Les opérations d'écritures et de lectures sont similaires à quelques petits détails près. Ces opérations sont détaillées à la figure 2.5. Nous supposons ici que l'état de la cellule est tel que la tension au nœud N1 soit haute et celle de N2 faible. Les NMOS M1 et M2 seront appelés les « drivers », les PMOS M3 et M4 « load ». Dans le cas d'une lecture, les bitlines sont tout d'abord préchargées à V_{DD} grâce aux charges « Z ». Si à la figure 2.5 cette opération est automatique à travers les deux charges passives, il peut arriver qu'un signal de précharge soit envoyé à la grille de deux PMOS qui seront coupés par la suite pour réduire les pertes actives. Ensuite, la wordline active les deux NMOS d'accès M5 et M6. $\overline{BL}$ est alors déchargée à travers M6 étant donné que le nœud N2 est à l'état bas. Cependant la capacité parasite de $\overline{BL}$ n'est pas négligeable, et cette décharge prend un certain temps. Afin de réduire les temps d'accès et de minimiser les pertes actives, dès qu'une tension suffisante s'est développée entre BL et $\overline{BL}$, les switchs de colonne sont activées et la donnée est envoyée vers le sense-amplifier. Une fois que celui-ci aura traité l'information, la wordline est désactivée ce qui met fin au cycle de lecture. Il faut noter que la tension du nœud mémorisant l'état bas est dégradée lors de cette opération. En effet, les transistors d'accès ont initialement une tension de drain égale à V_{DD}. M6 tendra donc à remonter la tension du nœud N2. Suite à cette perturbation, la « load » M3 est affaiblie par rapport au « driver » M1 et la tension au nœud N1 est également dégradée. Si la perturbation sur N2 est trop importante, l'état du latch pourrait basculer. C'est pourquoi les NMOS des inverseurs possèdent une largeur de grille environ deux fois plus élevée que celle des transistors d'accès. Il s'agit du ratio β de la mémoire.

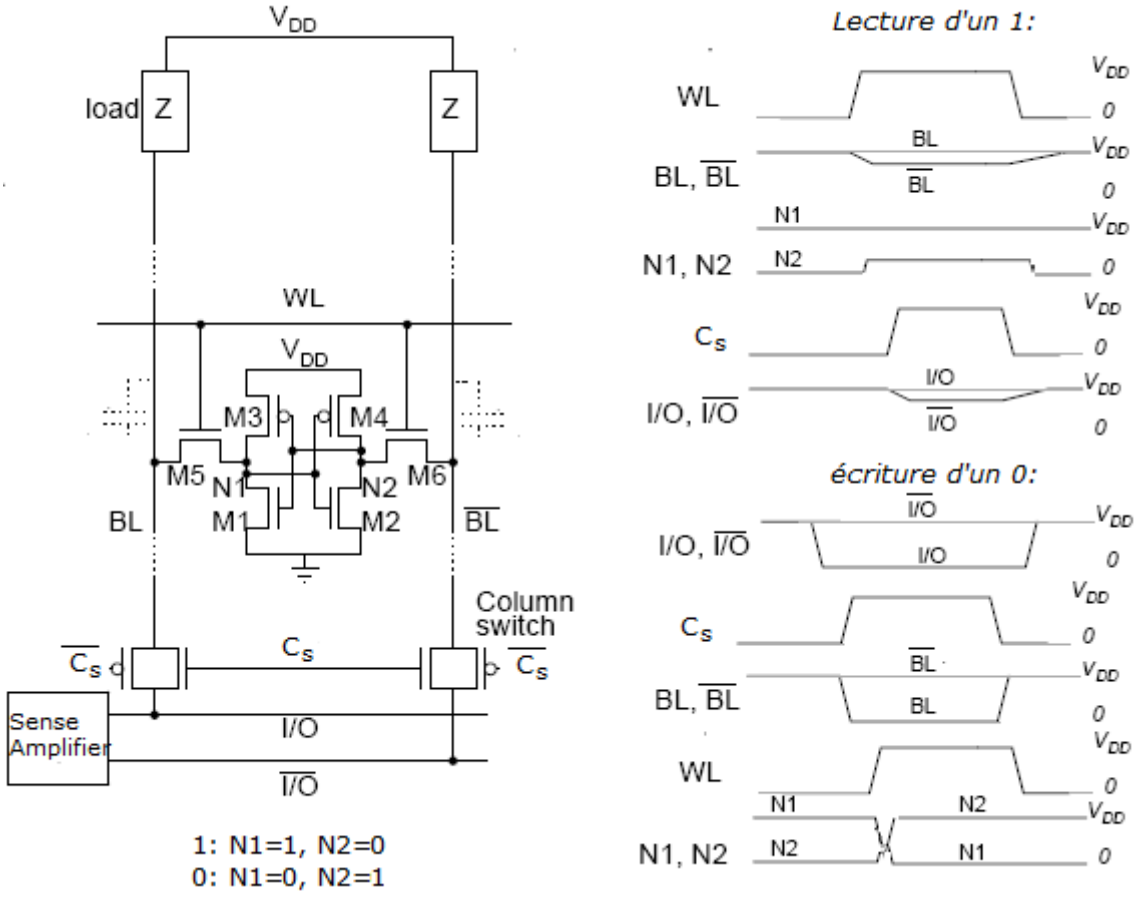


Fig.2.5 Opérations de lecture et d'écriture sur une SRAM

Lors d'une opération d'écriture, la valeur à écrire sur la mémoire est d'abord transmise aux bitlines à travers les switchs de colonnes. Ensuite, la wordline est activée. Le PMOS M3 forme alors un diviseur de tension avec le transistor d'accès M5 et la tension au nœud s'établit à un niveau intermédiaire. Si elle est inférieure au seuil de basculement de l'inverseur M2-M4, l'écriture est réussie. Pour que l'écriture soit effective, il est cependant nécessaire que le transistor d'accès M5 soit suffisamment fort devant le PMOS de l'inverseur M3 pour pouvoir transmettre un 0. Les PMOS possèdent donc toujours une largeur de grille minimale.

2.3. Etat de l'art

2.3.1. Introduction

Dans cette section, nous allons examiner les différentes techniques actuellement utilisées pour réduire la consommation des cellules SRAM. En effet, comme nous l'avons déjà mentionné, ces dernières années ont été marquées par l'augmentation de leurs consommations de puissance avec l'avancée des nœuds technologiques. Avant de réaliser ce tour d'horizon, il faut remarquer que plusieurs attributs permettent de caractériser l'efficacité d'une cellule donnée :

- la vitesse d'accès pour les opérations de lecture et d'écriture,
- la stabilité de la cellule, la fiabilité pour les opérations de lecture et d'écriture et sa sensibilité face aux perturbations,
- la taille de la cellule mémoire, sa densité d'intégration,
- la consommation en fonctionnement, que l'on peut séparer en consommation active (énergie dépensée pour réaliser une opération de lecture et d'écriture), ou statique (énergie nécessaire à la mémorisation d'une valeur).

En fonction de l'application ciblée, un des critères prendra plus d'importance que les autres. Ainsi, pour un pacemaker, ou toute une série d'applications portables, le critère prépondérant est la faible consommation de manière à valoriser la durée de vie de la batterie. Par contre, pour d'autres domaines, la vitesse d'accès sera de première importance, de manière à faire fonctionner la SRAM à une cadence supérieure au GHz [10] [11]. On peut encore aisément imaginer qu'une application spatiale accordera beaucoup d'importance à la fiabilité de la cellule.

Dans le cadre de ce TFE, nous allons nous pencher plus particulièrement sur les applications privilégiant une faible consommation de fonctionnement. Contrairement aux autres circuits logiques, il est indispensable de maintenir la tension d'alimentation de la cellule lors des périodes d'inactivité, afin de ne pas perdre l'information stockée par la mémoire. Des solutions particulières aux mémoires SRAM doivent donc être développées, d'autant plus que la consommation des mémoires SRAM représente une fraction non négligeable de la puissance dissipée par les mémoires cache des microprocesseurs [6] [12] [13]. Il est donc du premier intérêt, pour certaines applications à faible consommation, de réduire leurs courants de fuite.

Les techniques présentées à ce jour dans la littérature pour réduire ces fuites sont ici passées en revue. Le point commun à toutes ces solutions est qu'un compromis sera réalisé et que, pour réduire la consommation, un sacrifice est réalisé en termes de stabilité, de vitesse d'accès ou encore de

densité d'intégration. Si ces solutions ont chacune leur originalité, il est cependant possible de les classifier en quatre grandes catégories.

- La première catégorie vise une économie de l'énergie active liée aux opérations de lecture et d'écriture.
- La deuxième catégorie tente, en conservant l'architecture de la cellule 6T, de réduire les pertes par l'application de niveaux de polarisations en tension adéquats ou par l'utilisation de transistors à tension de seuil élevée.
- La troisième garde toujours une architecture 6T, mais met en place un mode de fonctionnement de veille, ou stand-by.
- Enfin, la quatrième catégorie regroupe de nouvelles architectures proposées pour la cellule.

Notons que la première catégorie se distingue des trois autres qui s'efforcent de réduire la consommation statique de la cellule elle-même. Evidemment, certaines propositions peuvent être hybrides et appartenir à plusieurs groupes.

2.3.2. Mécanisme de réduction de l'énergie active

Lors d'une opération de lecture et d'écriture, plusieurs éléments de la mémoire provoquent une consommation d'énergie, comme illustré à la figure 2.6. On parlera dans ce cas de pertes actives, car elles ont lieu uniquement lorsqu'une action est entreprise sur la mémoire. Il existe principalement trois sources de telles dissipations. Tout d'abord, la charge et la décharge de lignes hautement capacitives telles que wordlines et bitlines provoquent une dépense d'énergie proportionnelle à $C*V^2$ (1). Par ailleurs, lors de l'activation d'une wordline, toutes les bitlines sont déchargées par un courant de colonne suite à l'activation des transistors d'accès (2). Enfin, le courant de fonctionnement du sense-amplifier et des circuits périphériques lors d'une lecture est une dernière source de dissipation (3).

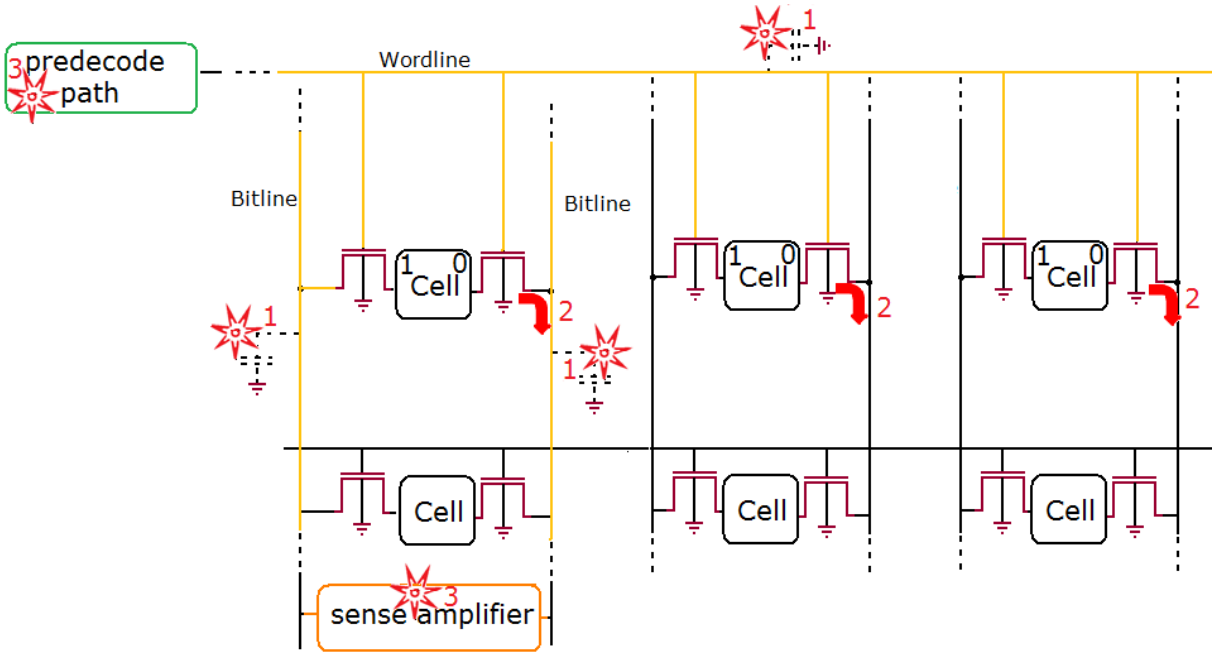


Fig. 2.6 : Sources de pertes actives d'une SRAM.

Le courant de fuite par colonne est proportionnel au nombre de bitlines connectées à une même wordline. Une sous-division de la wordline en plusieurs parties permet donc de réduire ce courant par un facteur égal au nombre de subdivisions [13]. En outre, d'après [14], tant que ce nombre n'excède

pas une certaine valeur, le délai global d'activation de wordline est réduit grâce à la réduction de sa capacité. Il est encore possible de réduire le courant de colonne par l'utilisation de charge d'impédance variable [14]. Dans ce cas, les PMOS de charge sont coupés lors des opérations d'écriture de manière à donner aux bitlines une impédance élevée. Il est possible d'étendre cette technique aux opérations d'écriture. La wordline ne doit alors être activée que le temps nécessaire au développement d'une tension suffisante pour être lue par le sense-amplifier de manière à ne pas dépenser une énergie inutile dans la décharge complète d'une bitline hautement capacitive [15]. Pour pouvoir créer ce pulse, il est nécessaire de connaitre les délais de fonctionnement de la cellule SRAM. La solution classique pour représenter un délai consiste à réaliser une chaine d'inverseur. Cependant, suite à la variabilité de procédés de fabrication entrainant une dispersion de la tension de seuil des transistors et à la dépendance de ce délai vis-à-vis des conditions de fonctionnement, une simple chaîne d'inverseur peut ne pas être satisfaisante. Une autre technique proposée dans [15] consiste à obtenir ce délai par une technique de réplique de la bitline.

La charge et la décharge de lignes à forte capacité sont également deux composantes importantes de la consommation active. A nouveau, une subdivision de la wordline permet de réduire significativement sa capacité [14] [16]. La même technique peut également être appliquée aux bitlines, notamment pour les opérations d'écritures [17]. De manière générale, un effort particulier a été mis dans la réduction du swing de tension appliqué au chemin de prédécodage d'adresse [16] ou aux bitlines lors d'une écriture [16] [17] [18]. En effet, le cas de l'écriture est critique dans le sens où la technique généralement utilisée consiste à charger une bitline à V_{DD} et à décharger complètement l'autre bitline, ce qui implique une dépense d'énergie proportionnelle à $2*C_{BL}*V_{DD}^2$, où C_{BL} représente la capacité de bitline. Afin de réduire leur swing de tension, trois techniques ont été proposées [17].

- La première consiste à ne précharger les bitlines qu'à $V_{DD}/2$ [16]. Cette technique possède cependant le désavantage de dégrader la stabilité de la cellule lors d'une lecture.

- La deuxième utilise un sense-amplifier local et une subdivision des bitlines [17]. Seul un faible swing est appliqué à la bitline hautement capacitive et à la sous-bitline sélectionnée. Par la suite, le sense-amplifier local amplifie le signal avant l'opération d'écriture. Cette technique permet de ne pas dégrader les marges de bruit et est schématisée à la figure 2.7(b).

- Enfin, la dernière proposition consiste à briser la boucle de contre-réaction lors d'une opération d'écriture, comme illustré à la figure 2.7(a). A cet effet, [18] et [19] proposent de contrôler la tension de masse de la cellule. En écriture, celle-ci est laissée flottante ce qui désactive la boucle de contre-réaction. Dès lors, seule une faible différence de 100 mV entre les deux bitlines est suffisante pour que lors de la reconnexion de la tension de masse la mémoire restaure la valeur à mémoriser, la cellule agissant alors comme une sorte de sense-amplifier.

Pour ces deux premières sources de pertes actives, il faut garder à l'esprit que les gains que permettent d'obtenir les mesures présentées sont en réalité liée à la taille de la mémoire utilisée. En effet, la capacité parasite des lignes, ou le nombre de bitlines raccordées à une wordline seront d'autant plus importants que le tableau des cellules mémoires est conséquent. Pour des mémoires de plus petite taille, il faut se poser la question de l'utilité de ces mesures par rapport aux complexifications qu'elles entrainent en termes de leur mise en œuvre. Ainsi, il est probable que la

solution présentée dans [17] et représentée à la figure 2.7 (a) entraine un surcout en termes de surface occupée et de complexité de fonctionnement de la mémoire.

La dernière source de dissipation active est le courant utilisé par les sense-amplifiers lorsqu'ils sont activés. Une des solutions envisagées dans [14] consiste ici à n'activer un sense-amplifier de type miroir de courant que le temps nécessaire à une amplification du signal suffisante, avant de restaurer entièrement les niveaux de tension par un latch. L'avantage est ici que le sense-amplifier de type miroir de courant ne nécessite qu'une différence de bitline faible pour pouvoir amplifier le signal, tandis que le sense-amplifier de type latch ne consomme que très peu de courant [20].

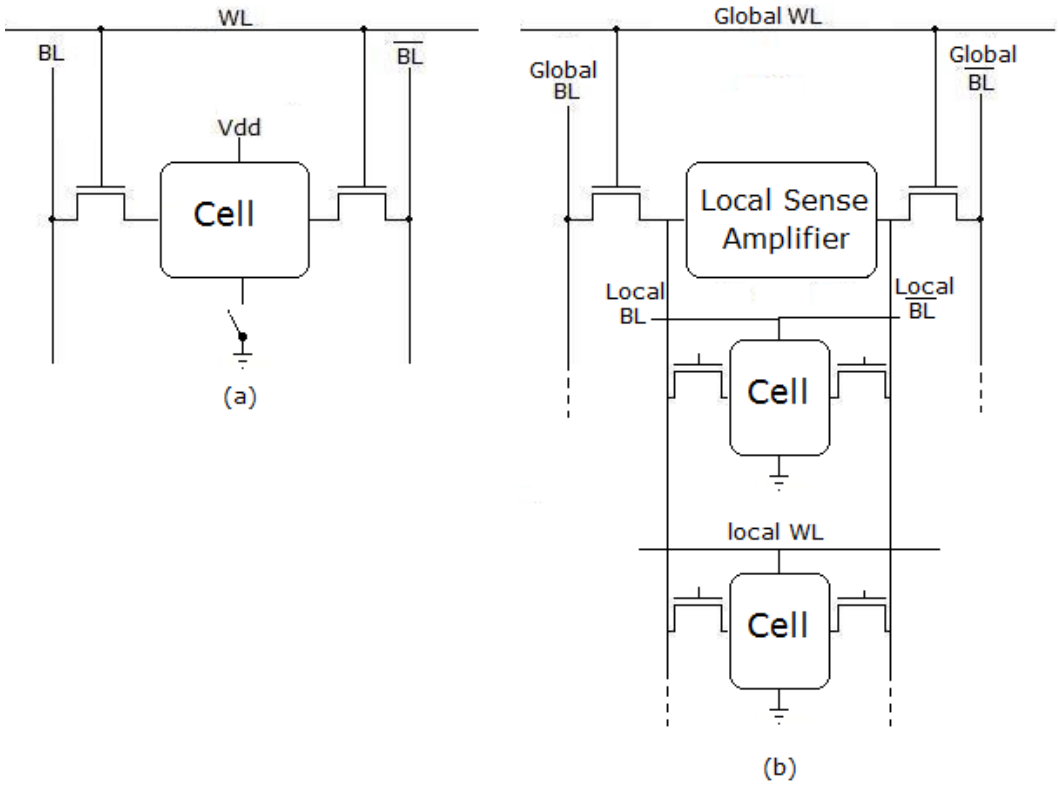


Fig. 2.7 : (a) La tension de référence de la cellule est débranchée lors d'une écriture de manière à briser la boucle de contre réaction. (b) Lors d'une écriture, un sense-amplifier local amplifie les tensions de bitlines globales.

2.3.3. Cellule 6T sans mode stand-by

L'utilisation de transistors à tension de seuil élevée permet de réduire significativement les courants de fuite d'une cellule SRAM. Néanmoins, l'utilisation de transistors plus lents se paye en termes de performances [16]. C'est ainsi qu'un certain nombre de SRAMs actuellement développées sont de type MT-CMOS (multi-thresold CMOS) ou VT-CMOS (variable threshold CMOS) [16] [21]. Les transistors de la cellule mémoire proprement dite possèdent une tension de seuil élevée, pour minimiser les fuites de la cellule, tandis que les circuits périphériques sont réalisés à l'aide de transistors à tension de seuil plus faible et donc plus rapides. Néanmoins, cette approche possède ses limites. En effet, l'utilisation de transistors à tension de seuil élevée augmente significativement le temps d'accès en lecture, suite à une diminution du courant de lecture qu'est capable de développer la

cellule mémoire. De plus, la stabilité de la cellule est réduite lors de la réduction de la tension d'alimentation. [20]

Pour pallier ce problème, certains auteurs ont proposé une combinaison adéquate de composants à faible ou haute tension de seuil dans la cellule mémoire elle-même. Ceci permet d'atteindre à la fois un faible courant de fuite tout en maintenant un courant de lecture élevé. Une étude sur les différentes combinaisons envisageables a été réalisée dans [22]. D'après cette étude, le meilleur compromis entre performances, réduction des fuites et surface de silicium utilisée est obtenu en utilisant des composants à tension de seuil faible pour les transistors d'accès, et une tension de seuil élevée pour les quatre autres transistors (Fig. 2.8a). Pour obtenir une marge de bruit[2] suffisante, il faut choisir une largeur de grille minimale pour les transistors d'accès et accroitre celle des drivers NMOS. Cette solution permet d'après les auteurs de réduire les pertes statiques de 20% par rapport à une cellule classique. Cependant, le prix à payer ici est une surface occupée plus importante. En effet, l'utilisation de transistors de types différents au sein d'une même cellule rend le layout moins dense étant donné qu'il ne sera plus possible d'imbriquer des transistors dont le dopage du canal sera différent. Les auteurs estiment que la cellule occupe une surface 30% plus importante que la cellule classique.

Une autre solution pour maintenir une grande vitesse de lecture consiste à élever la tension de wordline au-delà de V_{DD}. Pour minimiser le surplus de la consommation engendré, [21] propose de n'élever la tension de wordline que durant les premiers instants de la lecture, avant l'activation du sense-amplifier (Fig. 2.8b). Une fois que le sense-amplifier traite les données, la tension de wordline est ramenée à V_{DD}. Le défi consiste dans ce cas à pouvoir créer des pulses synchronisés, afin d'envoyer les données des bitlines au sense-amplifier au bon moment. Cette difficulté est particulièrement présente dans les systèmes sous batterie étant donné que les délais évoluent notamment avec la tension de batterie [21]. Notons également que la stabilité de la cellule est dégradée ce qui limite la tension maximale applicable à la wordline [21]. Evidemment, cette proposition ne permet pas à proprement parler de réduire les pertes statiques, mais uniquement limiter leur augmentation si un boost de wordline est nécessaire.

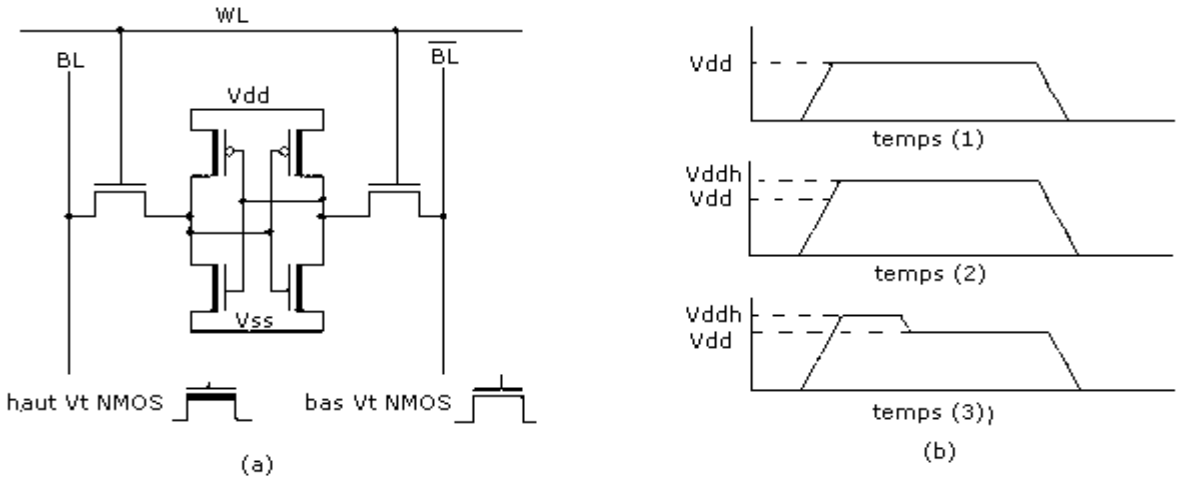


Fig. 2.8 : (a) Cellule SRAM à double tension de seuil, un faible V_T aux transistors d'accès et un haut V_T aux transistors du latch fournit le meilleur compromis entre performances et réduction des fuites [22]. (b) Un boost bref de la tension de wordline (3) permet d'obtenir un temps de lecture plus court (1) tout en minimisant les pertes occasionnées par les techniques de boost conventionnelles (2) [21].

[2] La notion de marge de bruit sera abordée à la section 4.8.

Cependant, de nombreux auteurs utilisent une autre voie pour réduire les fuites de la cellule. Il s'agit cette fois d'utiliser un niveau de polarisation adéquat aux différents accès. La principale ligne de conduite consiste à réduire le plus drastiquement possible la tension de rail à rail de la cellule tout en maintenant une stabilité suffisante [23] [24]. Certains auteurs appliquent également une tension de wordline négative en rétention [25]. Cette technique peut se retrouver aussi bien pour une cellule 6T standard que pour une autre architecture et permet de réduire le courant de fuite à travers les transistors d'accès de deux ordres de grandeur [20]. En effet, le V_{GS} de ces transistors devient alors négatif ce qui a pour effet de couper le courant sous seuil.

La difficulté lorsque l'on réduit la tension d'alimentation de la cellule est de garantir sa stabilité alors que la variabilité des procédés de fabrication s'accroit à chaque nouveau nœud technologique. L'existence d'une tension d'alimentation minimale pour garantir le bon fonctionnement de la cellule a ainsi été mise en évidence dans [6] et [7]. En plus des difficultés liées à la variabilité déjà citées précédemment, d'autres limitations sont imposées par le comportement sous seuil des transistors. En effet, dans cette zone de fonctionnement, où la tension d'alimentation est inférieure à la tension de seuil des transistors, [7] et [23] nous rappellent que la dépendance des courants par rapport à V_T et V_{GS} est exponentielle. Pour pallier ce phénomène, [23] suggère d'augmenter la longueur de canal des transistors (Fig. 2.9 (a)). Cela permet non seulement de réduire l'impact de la variabilité, mais également de l'effet DIBL[3] (Drain Induced Barrier Lowering). Les simulations réalisées grâce à un modèle prédictif 45nm montrent qu'une dépense de surface de silicium de 10 % supérieure à celle d'une cellule classique permet de réduire la tension d'alimentation jusqu'à 400 mV et le courant de rétention de la cellule par un facteur 20 tout en maintenant une bonne stabilité et de bonnes performances dynamiques.

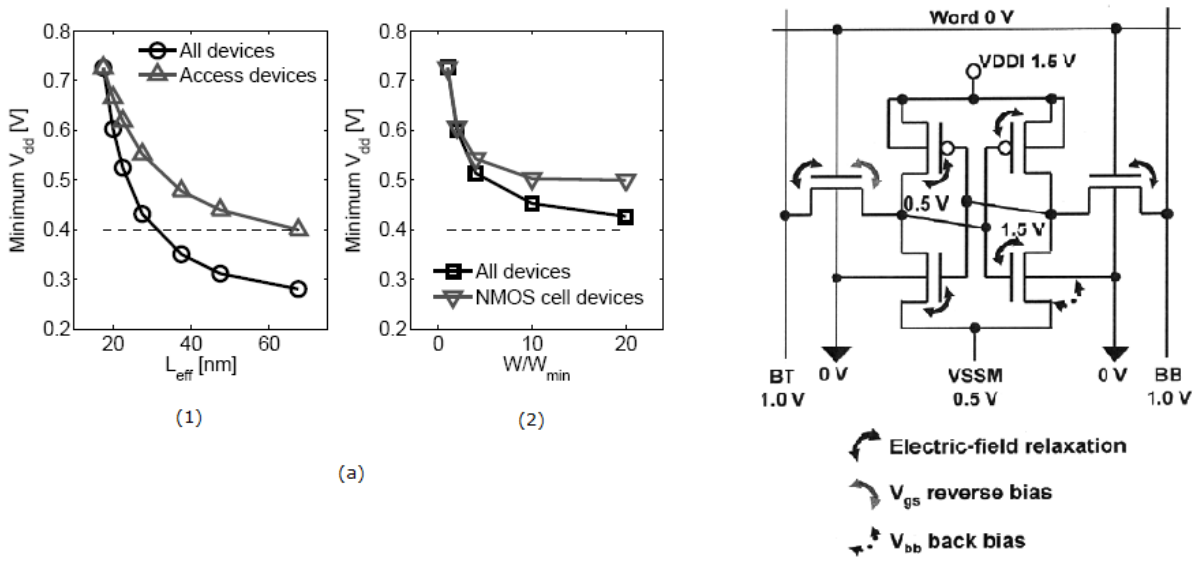


Fig. 2.9 : (a) Une augmentation de la longueur effective du canal de tous les transistors (courbes avec les ronds) permet le bon fonctionnement de la cellule sous une certaine tension d'alimentation (1). Pour obtenir un effet similaire, l'accroissement de la largeur de grille devrait être déraisonnable (figure reprise de [23]). (b) L'application de niveaux de polarisation adéquats permet de réduire plusieurs composantes des courants de fuite de la cellule (figure reprise de [24]).

[3] L'effet DIBL sera présenté à la section 7.2.2.

Une autre proposition consiste à ajuster de manière adéquate l'ensemble des niveaux de polarisation de la cellule [24]. L'application des niveaux de tension représentés à la fig. 2.9 (b) permet de réduire le courant statique de la cellule de plus de 80% d'après les auteurs de [24] par rapport à un schéma de polarisation standard. L'utilisation d'une tension de masse plus élevée pour la cellule que pour les wordlines et bitlines permet l'application de V_{GS} négatives aux transistors d'accès et donc de réduire leur courant sous seuil. L'écart entre tensions de grille et de source ou de drain est également réduit à certains endroits ce qui permet de réduire les courants de grille et ceux induits par effet GIDL (Grid Induced Drain Lowering) du circuit. Enfin, les polarisations de substrat des drivers NMOS permettent de réduire leur courant sous seuil. En effet, leur V_{BS} est négatif ce qui fait jouer l'effet substrat en faveur d'une réduction du courant de fuite.

2.3.4. Cellule 6T avec mode stand-by

Une autre technique consiste à faire entrer la SRAM en mode stand-by (ou veille) lorsque l'on n'y accède pas durant de longues périodes. Contrairement aux autres circuits logiques, il n'est cependant pas possible de couper totalement la tension d'alimentation de la SRAM, car les données mémorisées seraient perdues ! La solution la plus largement utilisée consiste à réduire la tension de rail à rail de la cellule à un niveau garantissant la rétention des données. Deux possibilités sont alors ouvertes : réduire la tension d'alimentation [6] ou augmenter la valeur de la tension de masse [5].

Une étude plus approfondie sur les effets d'une modification de la polarisation de la cellule durant la mise en stand-by a été réalisée dans [25]. Les résultats tendent à montrer qu'il est plus efficace de relever la tension de masse de la cellule que d'abaisser sa tension d'alimentation. En effet, une augmentation de la tension de référence a non seulement pour effet de réduire le V_{DS} des transistors de la cellule, mais induit également un V_{GS} négatif aux transistors d'accès. En outre, l'effet substrat permet encore de réduire le courant sous seuil et celui induit par effet GIDL (Gate induced Drain Leakage). Ces observations sont confirmées par [5] et [24].

L'apport de [5] consiste à ajuster le rehaussement de la tension de masse. En effet d'après ses auteurs, une technique utilisée habituellement est d'utiliser un PMOS monté en diode lorsque l'on désire entrer en veille [26], ce qui fixe la tension de masse de la cellule aux alentours de la tension de seuil du PMOS, comme illustré à la figure 2.10 (c). Si la cellule doit sortir du mode veille, des NMOS permettent de restaurer la tension de masse. Cependant, dans le cas des cellules fonctionnant sous faible tension d'alimentation et pour des technologies inférieures aux 100 nm, la sensibilité de la cellule aux variations de procédé est importante. La stabilité doit donc être traitée avec attention et imposer une valeur arbitraire à la tension de masse peut poser problème. La solution proposée consiste à rehausser la tension de masse secteur par secteur sur la cellule. Le circuit d'ajustement est constitué de NMOS de forces différentes (Fig. 2.10). En fonction de la valeur de tension désirée, ceux-ci sont sélectionnés ou non, ce qui permet d'ajuster finement la tension appliquée. Les mesures sur un circuit expérimental montrent une réduction des fuites lors de la mise en stand-by de 60 % à 80 % par rapport à leur cellule classique de référence.

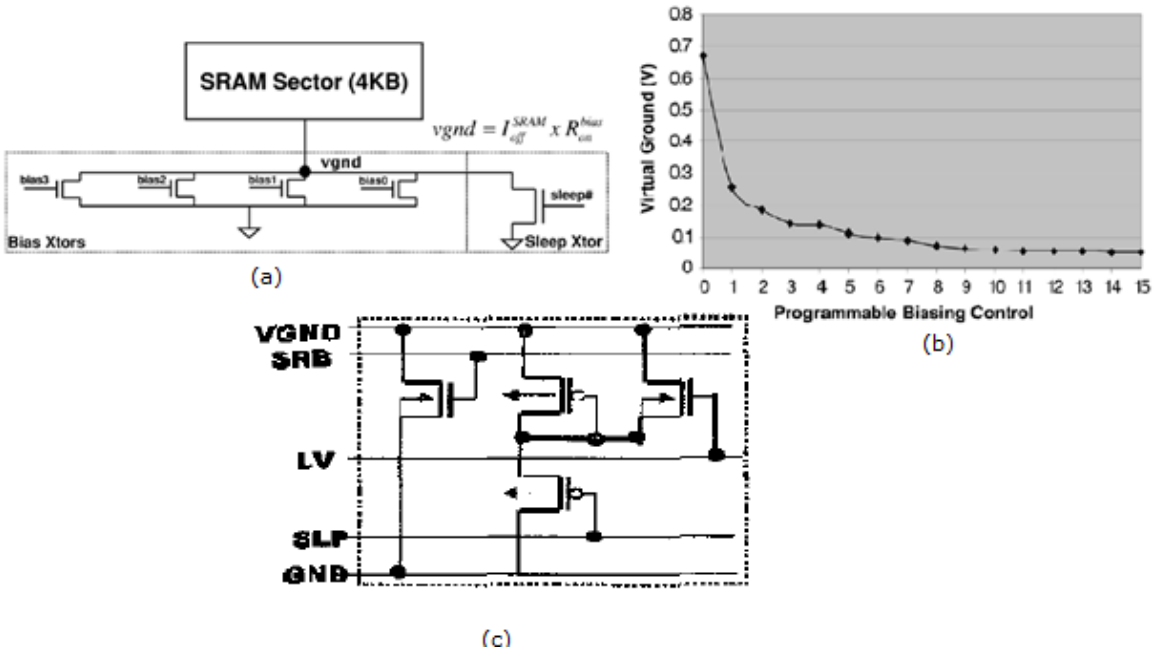


Fig. 2.10 : Le contrôle de la tension de masse (b) est réalisé à l'aide de NMOS de forces différentes (a) (figures reprises de [5]). (c) Des PMOS montés en diode permettent de créer une masse virtuelle. Les NMOS ramènent la tension de masse à 0 lorsque l'on quitte le mode veille (figure reprise de [26]).

L'inconvénient de la mise en veille de la mémoire est que le retour au mode de fonctionnement actif prend un certain temps et consomme une certaine quantité d'énergie [5]. Cette méthodologie n'est donc efficace que pour les mémoires où les périodes d'inactivité sont suffisamment longues. En outre, en dehors du mode de veille, aucune réduction des fuites n'est observée.

2.3.5. Nouvelles architectures de cellule

Encore une fois, le potentiel d'une réduction de la tension d'alimentation de la cellule jusque des valeurs inférieures à la tension de seuil des transistors est un des principaux moteurs de développement de nouveau type de cellule. En effet, une diminution de V_{DD} permet de réduire la puissance dynamique quadratiquement. Le bénéfice de la réduction de V_{DD} s'étend encore dans les nouveaux nœuds technologiques plus sensibles aux effets de canal court grâce à la réduction de l'effet DIBL (drain-induced barrier lowering), car une plus faible tension V_{DS} est appliquée aux transistors [27].

Cependant, certaines difficultés rendent la cellule 6T inopérante sous une trop faible tension d'alimentation. En effet, il devient impossible de maintenir une bonne stabilité lors des opérations de lecture ou d'écriture [7] [27] [28]. Une autre difficulté se retrouve dans la réduction du rapport I_{ON}/I_{OFF} des transistors d'accès à faible tension d'alimentation, ce qui provoque une interaction entre éléments passants et bloqués [27]. Enfin, la déviation de la tension de seuil due aux variations de procédés forme une dernière barrière. En effet, il faut rappeler qu'en régime sous seuil, le courant dépend exponentiellement de V_T [7] [28]. Une série de nouvelles cellules ont ainsi vu le jour pour réduire le risque d'erreur à tension d'alimentation donnée. L'intérêt de celles-ci réside dans le fait que la réduction des fuites est réalisée en permanence et pas uniquement lors de la mise en veille du

dispositif [27].

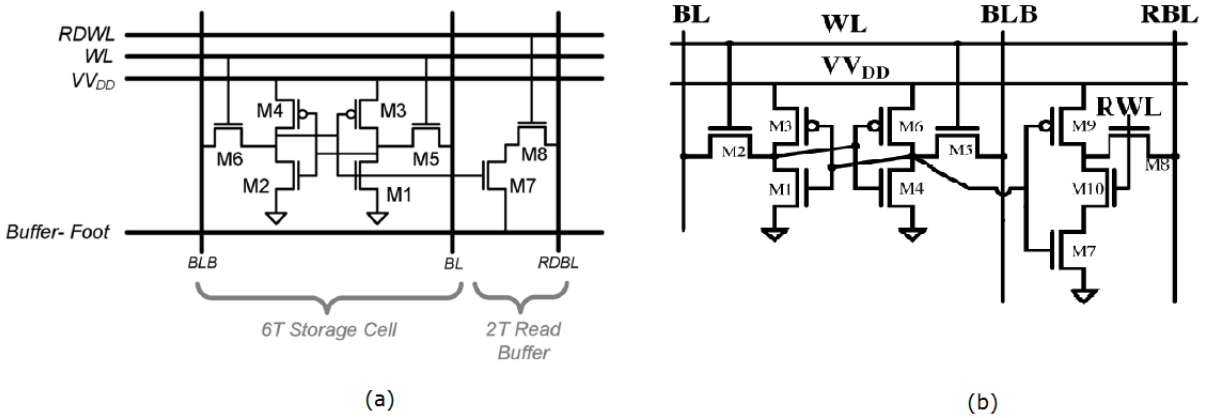


Fig.2.11 : (a) Cellule SRAM 8T, les transistors M7 et M8 assurent l'opération de lecture sans dégrader la marge de bruit (figure reprise de [27]). (b) Cellule SRAM 10T, les transistors M7 à M10 assurent l'opération de lecture sans dégradation de la marge de bruit. Un grand nombre de cellules peut être connecté à une même bitline sans mesure supplémentaire (figure reprise de [7]).

Dans [27], l'observation est faite que le premier obstacle à une réduction de la tension d'alimentation de la mémoire est la dégradation de la marge de bruit lors d'une opération de lecture. La solution consiste alors à ajouter deux transistors supplémentaires afin de réaliser l'opération de lecture. On peut voir à la figure 2.11 (a) que lorsque la wordline de lecture est activée, en fonction de l'état de la mémoire, M7 est soit passant, soit bloqué et peut ou non décharger la bitline de lecture. Cette solution nécessite l'ajout d'une wordline et d'une bitline et donc une modification de l'architecture globale de la SRAM. Le nœud mémoire est isolé du circuit de lecture à travers la grille de M7 et la marge de bruit n'est donc plus dégradée lors de cette opération. Malheureusement la dégradation du rapport I_{ON}/I_{OFF} des transistors d'accès en lecture peut entrainer une mauvaise interprétation de la valeur lue. En effet, le courant de fuite de l'ensemble des cellules auxquelles on n'accède pas et reliées à la même bitline peut être suffisant pour décharger la bitline de lecture. La solution apportée consiste à relever la tension de source des transistors M7 des cellules inactives, afin de supprimer leur courant de fuite. Afin de garantir le résultat de l'opération d'écriture malgré la variabilité des tensions de seuil des transistors, la tension d'alimentation de la cellule est réduite lors d'une écriture, de manière à affaiblir les charges PMOS. Un prototype valide la fonctionnalité de la cellule pour un V_{DD} descendant jusque 350 mV ce qui permet de réduire les pertes statiques par un facteur 20 en comparaison d'une cellule 6 transistors alimentée sous 1V. Cependant, la surface occupée est accrue de 30% par rapport à cette cellule de référence [27].

Une alternative à cette solution est détaillée dans [7]. Cette fois-ci, 10 transistors sont utilisés. Nous pouvons observer à la Fig. 2.11(b) que quatre transistors formant un buffer sont maintenant dédiés à l'opération de lecture. Ici, le problème de la dégradation du courant I_{ON}/ I_{OFF} des transistors d'accès est pris en charge grâce au transistor M10 dont le rôle est de couper le courant de fuite du buffer. Le courant de fuite est également limité grâce au PMOS M9. En effet, selon les auteurs, pour les technologies inférieures aux 100 nm, le courant de fuite des PMOS est supérieur à celui de NMOS. Dès lors, au repos, M9 maintient son drain proche de V_{DD}. Comme la bitline de lecture est également préchargée à V_{DD}, cela garantit un V_{DS} minimal aux bornes du transistor d'accès. Cette solution offre ainsi la possibilité de relier un grand nombre de cellules (par rapport à une architecture 6T) à la même bitline et donc de réduire le nombre de sense-amplifiers requis. Si cette cellule amène les pertes à un même niveau que la solution présentée précédemment dans [27], elle permet de relier un nombre virtuellement illimité de cellules aux mêmes bitlines sans montrer de problèmes à la lecture. Toutefois,

une estimation de la surface utilisée montre qu'elle occupe une surface supplémentaire de 60%, par rapport à une cellule 6 transistors conventionnelle, ce qui est peut-être trop important au vu de ses apports.

Afin de réduire plus avant la tension d'alimentation de la cellule, [28] propose une nouvelle solution. Le principe de la cellule repose sur le remplacement des inverseurs d'un latch classique par des triggers de Schmidt. Ceux-ci induisent une variation du seuil de basculement du latch en fonction du sens de transition des données. Afin d'éviter un nombre trop important de transistors sur la cellule, seul le NMOS de l'inverseur possède un tel mécanisme. La cellule est plus robuste lors d'un accès en lecture. En effet, si nous supposons à la fig. 2.12(a) que VL=1 et VR=0, le risque lors d'un accès en lecture est que l'élévation de VR ne vienne faire basculer l'inverseur de gauche. Or, l'action de NFL est de remonter la tension VNL. Dès lors, NL1 requiert une tension plus élevée pour basculer et le danger d'instabilité lors d'une lecture est écarté. Un prototype prouve la fonctionnalité de la cellule pour une tension d'alimentation de 160 mV. Par rapport à une cellule standard dont la tension d'alimentation est posée à 400 mV, de manière à présenter la même probabilité d'erreur, cette solution permet de réduire les pertes statiques de 20%, selon les auteurs. Notons qu'ici l'architecture globale de la mémoire peut être la même que dans le cas de cellule 6T. En effet, les opérations de lectures et d'écritures sont réalisées à l'aide des mêmes bitlines différentielles et l'accès à la cellule est également réalisé à l'aide d'une seule wordline.

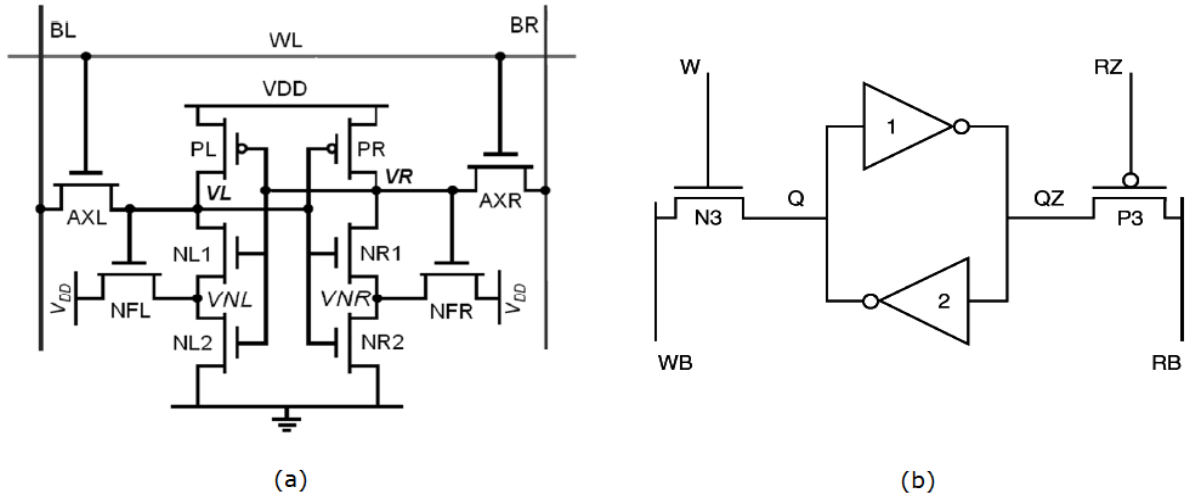


Fig. 2.12 : (a) Cellule SRAM 10T. L'action de NFL et de NFR permet d'augmenter le seuil de basculement haut des inverseurs et donc d'accroitre la stabilité en lecture (figure reprise de [28]). (b) La lecture et l'écriture sont réalisées sur des lignes séparées. La prédominance de « 1 » mémorisés permet de réduire les fuites à travers les transistors d'accès (figure reprise de [12]).

Les solutions présentées ci-dessus se basent sur une réduction de la tension d'alimentation de la cellule. Il faut cependant garder à l'esprit qu'une conséquence directe de cette stratégie réside dans l'augmentation du temps nécessaire à une opération de lecture étant donné que les transistors sollicités pour une opération de lecture devront opérer avec une tension d'overdrive réduite. Dans certains cas ils devront même opérer en régime sous seuil. Dès lors, ces solutions permettent le plus souvent une cadence de la mémoire de l'ordre de 10 kHz ce qui est bien éloigné des 100 MHz des mémoires plus rapides.

Il faut mentionner l'existence d'autres types de cellules ne se basant pas forcément sur une réduction de la tension d'alimentation. Dans [12], il s'agit de garder une structure proche de celle de la

cellule 6 transistors mais en séparant les bitlines de lecture et d'écriture comme représenté à la fig. 2.12 (b). La solution apportée ici permet d'après les auteurs de réduire la consommation statique de 60 % environ par rapport à une cellule conventionnelle tout en maintenant les performances dynamiques de la cellule. La difficulté consiste essentiellement à pouvoir écrire la valeur 1 sur la cellule étant donné que le NMOS d'accès ne peut totalement transmettre une tension haute. Pour contourner le problème, il suffit d'activer le PMOS d'accès en lecture relié à une bitline déchargée. L'écriture d'un « 1 » logique est ainsi facilitée, car le NMOS de l'inverseur de droite est affaibli. Ceci revient en quelque sorte à briser la boucle de contre-réaction du latch, ce qui est effectivement utilisé dans un certain nombre de cellules pour faciliter l'opération d'écriture [7] [29]. La réduction des fuites est liée à l'observation que pour certaines mémoires caches, la valeur 0 est beaucoup plus présente que la valeur 1. Si cette logique est inversée, la cellule mémorisée permet alors de couper les fuites des deux transistors d'accès. Cependant, cette solution ne reste valable que dans les cas où cette hypothèse reste vérifiée.

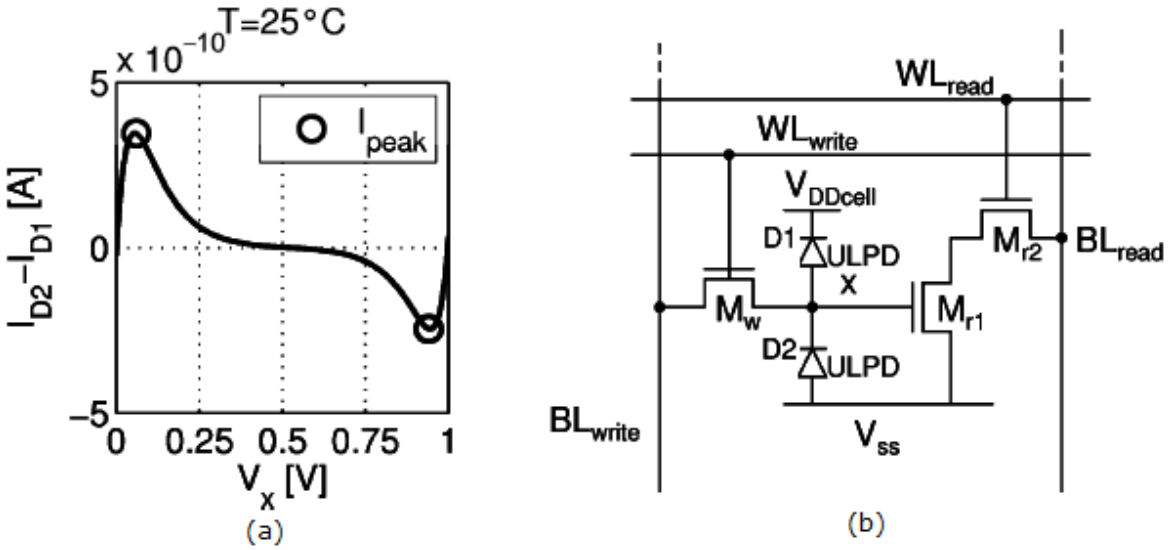


Fig. 2.13 : (a) L'évolution des courants dans la cellule en fonction de la tension du nœud de mémorisation confère un effet mémoire au montage diode. (b) Cellule 7T réalisée à l'aide de deux transistors ULP montés en diode. (Figures reprises de [20]).

Une dernière proposition se base sur l'utilisation de diodes ULP [20]. Il s'agit ici d'utiliser deux transistors ULP tels que ceux présentés à la section 3.4 montés en diodes, et placés en série. Une analyse sur les courants de la cellule démontre qu'un effet mémoire est ainsi obtenu tout en réduisant drastiquement les courants de fuite de la cellule. Par exemple, comme observé à la figure 2.13 (a), lorsque la tension de la capacité parasite du nœud mémoire s'éloigne de V_{DD}, le courant dans la diode du haut domine et charge cette capacité jusque V_{DD}. Un buffer de lecture est utilisé afin de ne pas détruire la valeur mémorisée. Cette cellule comporte donc 7 transistors au total comme illustré à la figure 2.13 (b), et permet d'obtenir à la fois une bonne stabilité, un fonctionnement à température élevée et une bonne vitesse d'accès. Malheureusement, les couplages capacitifs détériorent la valeur mémorisée juste après une écriture. Cette valeur n'est restaurée qu'après une période de quelques µs, car les diodes ULP ne peuvent délivrer un fort courant, durant laquelle la consommation de la cellule est légèrement accrue par une composante active. Par ailleurs, d'après les auteurs, sans l'utilisation de matériaux « high-k »[4], les courants de grille dans des technologies plus avancées risquent de détériorer

[4] Les matériaux « high-k » sont en fait utilisés pour remplacer l'oxyde de Silicium comme diélectrique de grille pour les transistors MOS. Nous parlerons plus en détail de leur intérêt à la section 7.7.2.

la fonctionnalité de la cellule. Toutefois, il faut constater que ce même genre de problème se pose pour nombre des architectures présentées ci-dessus et que le passage aux matériaux « high-k » semble assuré dans les années à venir.

2.4. Conclusion

Le tableau 2.2 permet de faire une synthèse des caractéristiques de certaines SRAMs présentées. A titre de comparaison, les résultats d'une cellule 6 transistors à hautes performances dynamiques sont également fournis. Afin de pouvoir effectuer une analyse objective, les différentes cellules ont été simulées sous ELDO avec des modèles de transistors de la même technologie bulk 130nm[5]. Ceci explique les différences qu'il est possible de retrouver entre les résultats présentés sur ce tableau et ceux annoncés par les auteurs. Toutefois, dans tous les cas, la transposition des résultats en technologie bulk 130 nm semble être cohérente avec ce qui est présenté dans la littérature. L'écart le plus important concerne le courant statique présenté par la cellule de Osada [24]. Cet écart s'explique par le fait que les auteurs utilisent un type de transistor possédant une tension de seuil très élevée. Toutefois, les auteurs annoncent également une réduction de pertes statique de 80 % par rapport à leur cellule de référence. Ce rapport est bien du même ordre de grandeur que celui que nous obtenons dans ce tableau. La dénomination I_{stat} correspond au courant consommé par la cellule lorsque celle-ci est dans un état de rétention.

	V_{DD} [mV]	Transistors	I_{stat}	$I_{stat\ 6T}/\ I_{stat}$	I_{READ}	$I_{READ}/\ I_{stat}$
6T bas V_T	1	6	15.9 nA	0.04	51.7 µA	$3.2\ 10^3$
6T haut V_T	1	6	710 pA	1	38.5 µA	$54.2\ 10^3$
8T faible V_{DD}[27]	400	8	420 pA	1.7	0. 86 µA	$2.0\ 10^3$
Kulkarny [28]	400	10	352 pA	2.0	0.57 µA	$1.6\ 10^3$
Osada [24]	1	6	84 pA	8.5	12.9 µA	$153.5\ 10^3$
Levacq [20]	1	7	5.8 pA	122.4	30.6 µA	$5\ 276\ 10^3$

Tableau 2.2 : performances de certaines cellules utilisant des techniques « low power ». Les cellules ont été simulées en technologie bulk 130nm sous une température de 25 °C.

Ce tableau montre que la piste d'une diminution de la tension d'alimentation ne permet pas une réduction des pertes statique exceptionnelle, même si les auteurs affirment qu'il est possible de réduire la valeur de V_{DD} à des valeurs aussi faibles que 160 mV [28]. Actuellement, une seule cellule permet de réduire significativement les courants de fuite. Il s'agit de la cellule à 7 transistors basée sur l'utilisation de diodes ULP présentée par D. Levacq [20].

Il faut cependant considérer certains facteurs de mérites propres à l'application visée. Dans le cadre de ce travail, nous nous efforcerons de développer une cellule permettant de réduire de manière importante la consommation statique tout en maintenant de hautes performances dynamiques. C'est pourquoi nous accorderons une certaine importance au rapport $I_{READ}/\ I_{stat}$. En effet, ce rapport donne une image du temps nécessaire à une opération de lecture et de la réduction des pertes statiques de la cellule. Il aurait toutefois également été possible de viser par exemple un domaine « low cost, low power » où cette fois la surface de silicium consommée devient plus importante que le courant de lecture. Ce rapport $I_{READ}/\ I_{stat}$ nous montre encore une fois que pour ce type d'application, d'autres pistes qu'une réduction de la tension d'alimentation doivent être apportées. A ce titre, la cellule de D. Levacq se démarque sans conteste du lot de cellules déjà étudié.

[5] Une caractérisation des performances de ces transistors est réalisée à la section suivante.

Il semble que la cellule présentée dans [20] apporte déjà une très bonne solution au problème posé. Cependant, dans ce travail, nous rechercherons également à développer une cellule avec une stabilité accrue. Nous verrons qu'il est possible d'obtenir des résultats nettement supérieurs à ce que la littérature actuelle propose. Cette notion de stabilité sera abordée plus en détail à la section 4.8. Notons qu'elle est de première importance alors que l'ITRS 2007 annonce qu'il sera difficile de garantir le bon fonctionnement des SRAMs rapides pour les technologies futures [1].

3. Le transistor et l'inverseur ULP

3.1. Introduction

Ces dernières années, la loi de Moore a gouverné la réduction de la taille des transistors. Cette diminution de la longueur de grille des transistors permet d'économiser de la surface sur les circuits intégrés et d'accroitre le courant qu'ils sont capables de délivrer. Néanmoins, ceci s'accompagne d'une augmentation des courants de fuite, notamment du courant sous seuil et du courant de grille. Pourtant, s'il est vrai qu'une partie du marché demande des circuits toujours plus rapides et « performants », certaines applications nécessitent quant à elles le développement de circuits à très faible consommation d'énergie.

C'est dans cette optique qu'un nouveau mécanisme de réduction des courants de fuite a été élaboré par l'UCL [8] [20] [30]. Une diode à ultra basse consommation (diode ULP) et ses applications ont été étudiées [8] [20]. Sur base du même principe de réduction des courants de fuite, il a été ensuite possible de mettre au point des transistors à ultra basse consommation (transistor ULP) sur base desquelles un inverseur et certains circuits logiques ont été élaborés [30]. Ces éléments logiques serviront de base à l'élaboration d'une cellule SRAM ULP détaillée dans ce travail.

La seconde partie de cette section abordera la description des modèles de transistors utilisés. Ensuite, la troisième partie permettra de faire un bref rappel sur les composantes du courant de fuite d'un transistor. La quatrième partie introduira le fonctionnement du transistor ULP et la cinquième le fonctionnement de l'inverseur ULP.

3.2. Modèle des transistors utilisés

Nous utilisons deux modèles industriels BSIM3 de transistors bulk du nœud technologique 0.13 µm. Un de ces transistors est dit « high speed » (HS), l'autre « low leakage » (LL). Comme nous le verrons par la suite, pour fonctionner correctement, les transistors ULP doivent être composés de transistors présentant un courant de fuite relativement élevé. C'est pourquoi ils seront réalisés en utilisant les transistors « HS ». Les comparaisons de performances avec une technologie CMOS classique seront effectuées avec des transistors « LL ». Ceci permet d'obtenir une base objective de comparaison, il s'agit d'analyser les performances des transistors ULP par rapport à une technique classique de réduction des fuites.

Afin de mettre en évidence les caractéristiques de ces modèles, les courbes I_D-V_{GS} ont été tracées pour un PMOS et un NMOS. La largeur des grilles est de 1µm. La tension de substrat des NMOS est connectée à la tension la plus faible du circuit, celle des PMOS à la tension la plus élevée. Cette convention a été prise tout au long de ce travail. Ainsi, afin d'alléger les schémas, les connexions de substrats respectant cette convention ne seront pas systématiquement indiquées sur les schémas.

- Nous obtenons pour le modèle « LL » les résultats suivants :

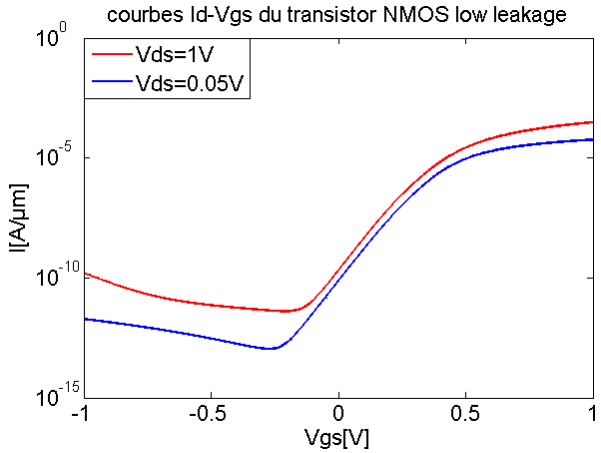


Fig. 3.1 : Courbe I_D-V_{GS} du NMOS LL pour V_D=0.05V et 1V. V_S=V_B=0V, W=1µm, L=0.13µm.

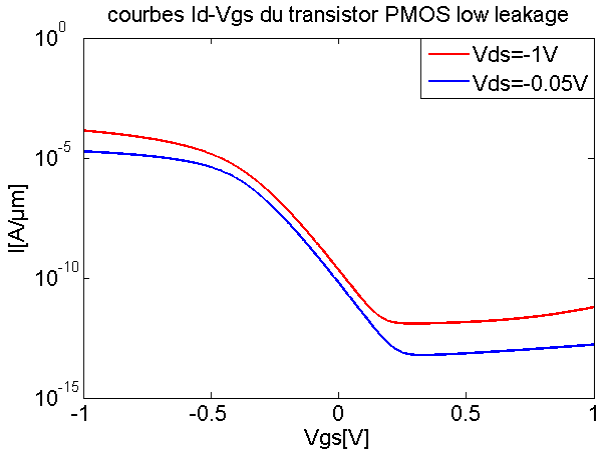


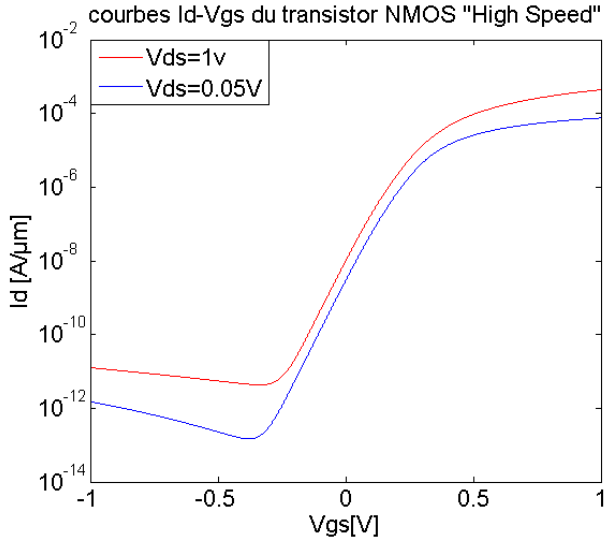
Fig 3.2 : Courbe I_D-V_{GS} du PMOS LL pour V_D=-0.05V et -1V. V_S=V_B=0V, W=1µm, L=0.13µm.

$\|V_{DS}\|$ [V]	I_{ON} NMOS [A/µm]	I_{ON} PMOS [A/µm]	I_{OFF} NMOS [A/µm]	I_{OFF} PMOS [A/µm]
1	3.1E-04	1.4E-04	2.0e-10	2.1e-10
0.05	5.7E-05	1.9E-05	7.3e-11	6.2e-11

Tableau 3.1 : courants I_{ON} et I_{OFF} des transistors low leakage. Le courant est donné pour un $\|V_{GS}\|$=1 ou 0V et pour $\|V_{DS}\|$ = 1V et 0.05V.

La tension de seuil de ces transistors s'élève à 450 mV pour le NMOS et à 400 mV pour le PMOS, d'après le fondeur ayant développé ces modèles.

Le modèle « HS » possède quant à lui les caractéristiques suivantes :



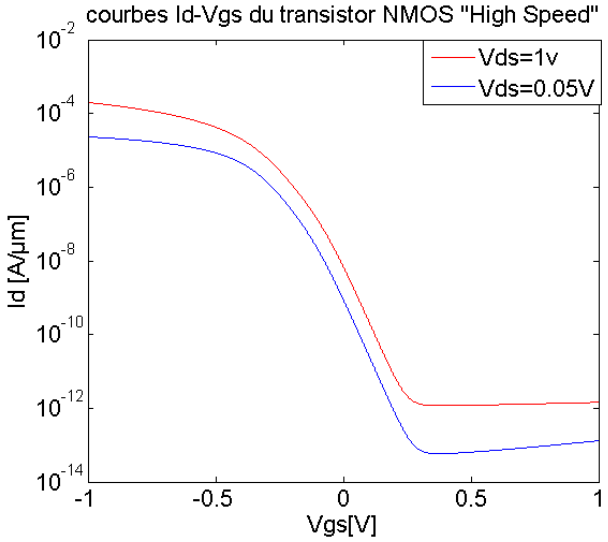


Fig 3.3 : A gauche, courbe I_D-V_{GS} du NMOS HS, à droite courbe I_D-V_{GS} du PMOS HS, pour $\|V_D\|$=0.05V et 1V. V_S=V_B=0V, W=1µm, L=0.13µm.

$\|V_{DS}\|$ [V]	I_{ON} NMOS [A/µm]	I_{ON} PMOS [A/µm]	I_{OFF} NMOS [A/µm]	I_{OFF} PMOS [A/µm]
1	4.4e-04	2.0E-04	1.0e-08	6.9e-09
0.05	7.5e-05	2.3e-05	2.9e-09	8.9e-10

Tableau 3.2 : courants I_{ON} et I_{OFF} des transistors high speed. Le courant est donné pour un $\|V_{GS}\|$=1 ou 0V et pour $\|V_{DS}\|$ = 1V et 0.05V.

Cette fois-ci, la tension de seuil de ces transistors s'élève à 340 mV pour le NMOS et à 300 mV pour le PMOS.

Plus de détails sur l'évolution des courants I_{ON} et I_{OFF} avec la température sont disponibles à l'annexe 1.

3.3. Les courants de fuite d'un transistor

Nous pouvons distinguer trois principales sources de courant de fuite dans un transistor :

- courant sous seuil,
- courant de grille,
- courant de jonction.

La figure 3.4 illustre ces courants. Le courant sous le seuil représente le courant s'écoulant entre source et drain alors que le transistor se trouve à l'état bloqué. Le courant de grille représente les électrons s'échappant du canal vers la grille par effet tunnel. Enfin, le courant de jonction est le courant s'écoulant du drain vers le substrat à travers la jonction PN présente entre ces deux zones. Notons que ce dernier peut être pratiquement annulé à température ambiante par l'utilisation de substrat de type SOI (silicon on insulator), où une couche d'oxyde de silicium permet d'isoler les zones actives de la connexion du substrat.

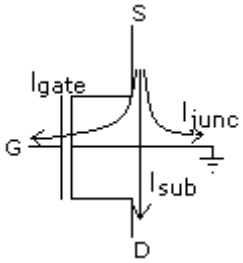


Fig. 3.4 : les sources fuites d'un transistor.

Dans le cas d'un transistor MOS, le courant sous seuil est la principale source de fuite lorsque la tension V_{GS} est nulle. Certaines équations permettent d'en obtenir une représentation précise [8] [31]. Lorsque la longueur de transistor diminue, l'effet DIBL (drain induced barrier lowering) influence significativement son importance. Cet effet représente la diminution de contrôle de la tension de grille sur le canal au profit de la tension de drain. Il se traduit par une réduction de la tension de seuil du transistor lorsque la tension de drain augmente.

En outre, la diminution de la longueur de canal requiert également la réduction de l'épaisseur de la capacité d'oxyde de grille. Cela est nécessaire pour pouvoir maintenir une capacité de grille suffisante pour garantir un bon contrôle du canal. Toutefois lorsque cette épaisseur atteint l'ordre du nanomètre, l'effet tunnel permet à certains électrons présents dans le canal de traverser l'oxyde de grille. Cette source de fuite prend donc de l'importance dans les nouveaux nœuds technologiques. L'apparition de matériaux « high-k » permettra de contourner ce problème. Il s'agit de matériaux possédant une constante diélectrique supérieure à celle de l'oxyde de silicium et donc pour lesquels une épaisseur d'oxyde moins fine serait requise pour obtenir la même capacité de grille. Leur introduction a été prévue par l'ITRS 2007 comme étant une des innovations majeures pour 2008 [1].

Les différents courants de fuite d'un transistor ont été simulés pour le modèle « LL ». Leur représentation est donnée à la figure 3.5. Le courant sous seuil est bien ici la contribution majeure au courant de fuite du transistor.

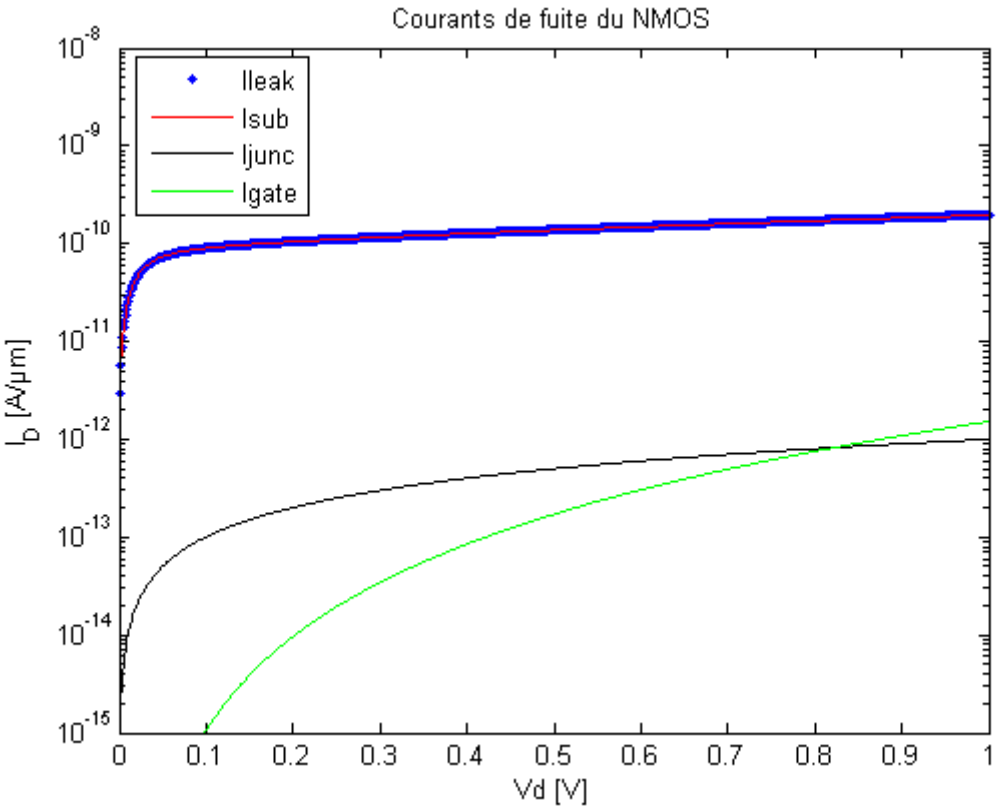


Fig. 3.5: Le courant de fuite d'un transistor LL et ses composantes. V_{DS}=1V, V_{BS}=0V, V_{GS}=0V, W=1µm, L=0.13µm, T=25 °C.

3.4. Le transistor ULP

3.4.1. La structure du transistor ULP

Les transistors ULP sont représentés sur la figure 3.6. Ils sont constitués d'un NMOS et d'un PMOS. Nous analyserons ici le fonctionnement du transistor ULP de type N en détail. Des raisonnements semblables peuvent être effectués pour le transistor ULP de type P. Le transistor ULP de type N est constitué d'un transistor NMOS classique empilé sur un PMOS. Ce montage permettra de réduire le courant de fuite du transistor ULP bien plus efficacement que ne l'aurait fait un simple empilement de deux NMOS, grâce à la création d'une boucle de contre-réaction. Pour des raisons de symétrie, la largeur du PMOS est égale à celle du NMOS multiplié par l'inverse du rapport de mobilité des porteurs. Dans la suite de ces explications, les lettres G, S et D majuscules désigneront les accès des transistors ULP, tandis que les lettres g, s, d minuscules ceux des transistors sous-jacents.

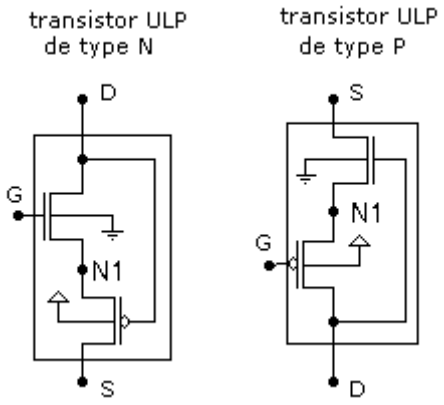


Fig 3.6 : les transistors ULP.

3.4.2. Mécanisme de réduction des courants de fuite

Dans le cas des transistors ULP, les courants de fuite sont réduits grâce à l'application automatique d'une tension V_{gs} négative sur le NMOS et le PMOS du transistor ULP. Ceci permettra de diminuer significativement le courant sous le seuil, qui comme nous avons pu l'examiner à la figure 3.5, est la principale source de fuite d'un transistor MOS. La tension de source de ces deux transistors est celle du nœud N1 (fig. 3.6). Rappelons que la largeur des transistors est choisie de manière à compenser la différence de mobilité des porteurs et que le substrat du NMOS est connecté à la masse et celui du PMOS à V_{DD}. Ainsi, lorsque le transistor est bloqué (V_G=0V), la tension du nœud 1 doit valoir $V_{DS}/2$ par symétrie. On en déduit que les tensions V_{gs} et V_{sg} du PMOS et du NMOS valent $-V_{DD}/2$, ce qui réduit significativement le courant sous le seuil.

L'expression du courant sous le seuil d'un NMOS est donnée par [8] [31] :

$$I_{subt,n} = I_{Sn} \cdot e^{\frac{Vbs\cdot(n-1)}{n\cdot U_T}} \cdot e^{\frac{Vgs}{n\cdot U_T}} \cdot e^{\frac{\eta\cdot Vds}{n\cdot U_T}} \cdot (1 - e^{\frac{-Vds}{n\cdot U_T}}) \qquad (2.1)$$

Avec I_{Sn} un courant de référence qui correspond graphiquement à l'extrapolation du courant de drain pour V_{GS}=V_{BS}=0 avec un V_{DS} suffisant pour obtenir la saturation du transistor tout en restant suffisamment faible pour que l'effet DIBL n'intervienne pas [8]. Le facteur n est le coefficient d'effet de substrat. U_T représente la tension thermique et η est le coefficient DIBL. Dans le cas du modèle « HS » utilisé, la valeur de I_{Sn} est de 1.47 nA, celle de n est de 1.4 est celle du coefficient DIBL η est de 0.09.

Il est possible de simplifier cette expression (2.1) pour obtenir le courant sous le seuil d'un transistor NMOS (2.2) dont les tensions V_{GS} et V_{BS} sont nulles. En définissant le courant sous seuil du transistor ULP de type N comme étant celui de son NMOS, son expression peut également être évaluée (2.3)

$$I_{subt,n} = I_{Sn} \cdot e^{\frac{\eta\cdot Vds}{n\cdot U_T}} \cdot (1 - e^{\frac{-Vds}{n\cdot U_T}}) \qquad (2.2)$$

$$I_{subt,NULP} = I_{Sn} \cdot e^{\frac{-V_{DS}\cdot(n-1)}{2\cdot n\cdot U_T}} \cdot e^{\frac{-V_{DS}}{2\cdot n\cdot U_T}} \cdot e^{\frac{\eta\cdot V_{DS}}{2\cdot n\cdot U_T}} \cdot (1 - e^{\frac{-V_{DS}}{2\cdot n\cdot U_T}}) \qquad (2.3)$$

Avec dans (2.3) V_{DS} égale à la tension entre drain et source du transistor ULP.

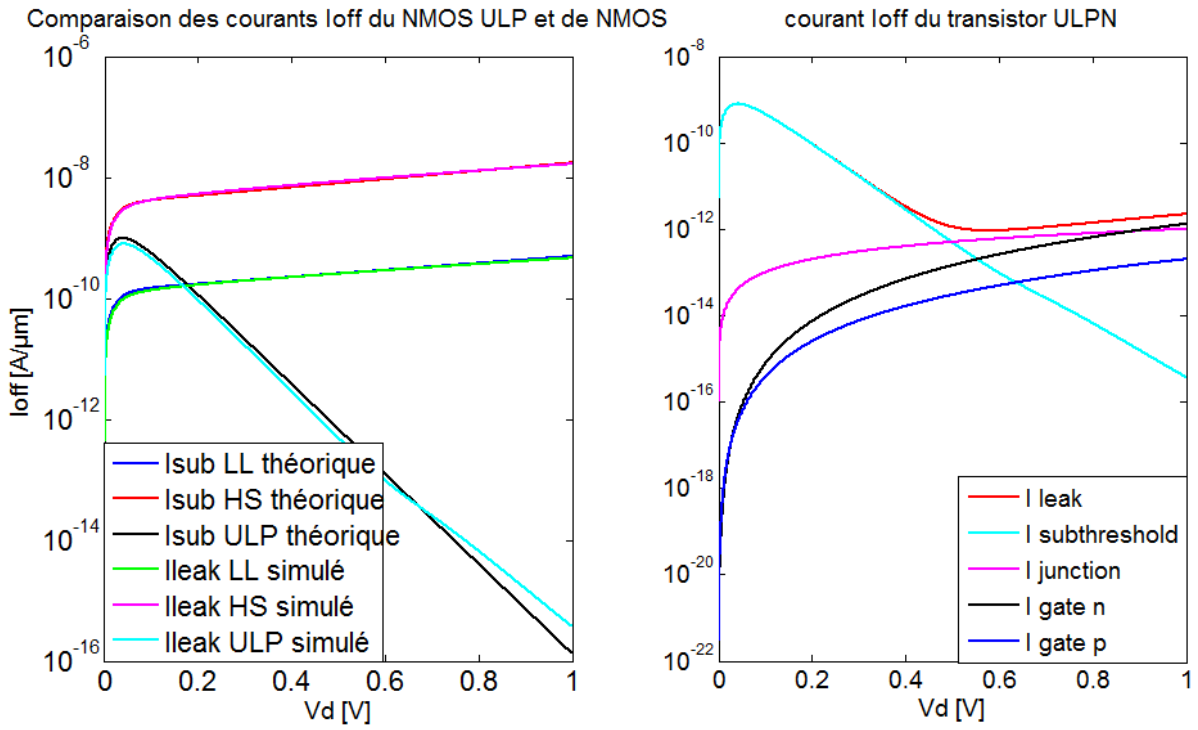


Fig. 3.7 : A gauche comparaison du courant de fuite théorique et simulé du transistor ULP, du NMOS LL et du NMOS HS. A droite, simulation des différentes composantes du courant de fuite du transistor ULP. W_N=1 µm, W_P=2.4µm, L=0.13µm, T=25°C.

La figure 3.7 reprend à gauche l'évolution du courant sous seuil du transistor ULP de type N en fonction de V_D. Elle est comparée à celle des NMOS HS et NMOS LL. Les courbes théoriques données par les expressions (2.1) à (2.3) sont également tracées. La correspondance entre prévision théorique et résultat de simulation est bonne dans le cas des NMOS classiques. Pour le transistor ULP, le modèle théorique reste correct pour des tensions V_D inférieures à 0.6V. Au-delà, nous pouvons observer un écart entre la pente théorique et simulée. Cet écart est lié à la non-symétrie des transistors constituant le transistor ULP de type N. En effet, leurs tensions de seuil ne sont pas exactement égales. En outre, pour cette valeur de V_D, le courant de jonction du PMOS commence à devenir non négligeable et son impact n'est pas pris en compte par notre modèle théorique

Le graphe de droite de la figure 3.7 trace l'évolution du courant de fuite du transistor ULP de type N et de ses différentes composantes. Dans un premier temps, le courant sous seuil augmente à cause de l'augmentation de la tension V_{ds} du NMOS et du PMOS. Ensuite, lorsque V_D continue d'augmenter, la tension V_{gs} des transistors devient de plus en plus négative ce qui va fortement réduire le courant sous seuil. Ce courant va finir par devenir négligeable devant le courant de jonction qui va alors déterminer la valeur du courant de fuite pour de plus grands V_D. Pour V_D proche de 1V, le courant de grille du NMOS devient la composante majoritaire du courant de fuite.

Notons enfin la présence d'une valeur de tension d'alimentation permettant de minimiser les courants de fuite du transistor ULP. Cette valeur se situe aux alentours de V_{DD}=0.6V. L'utilisation de substrat SOI et de diélectrique « High-k » permettra de réduire significativement les courants de fuite

du transistor ULP. En effet, après une certaine valeur de V_{DD}, les courants de jonction et de grille sont les composantes essentielles du courant de fuite du transistor ULP. Ceux-ci seraient tous deux réduits significativement par l'utilisation combinée de ces techniques.

3.4.3. Caractéristique I_D-V_{GS} du transistor ULP

La figure 3.8 trace les courants I_{ON} des transistors ULP pour différentes valeurs de la tension d'alimentation. Nous pouvons remarquer immédiatement qu'ils sont également bien plus faibles que le courant de fonctionnement d'un transistor classique. Pour comprendre ceci, examinons la courbe correspondant à une tension d'alimentation de 0.6V.

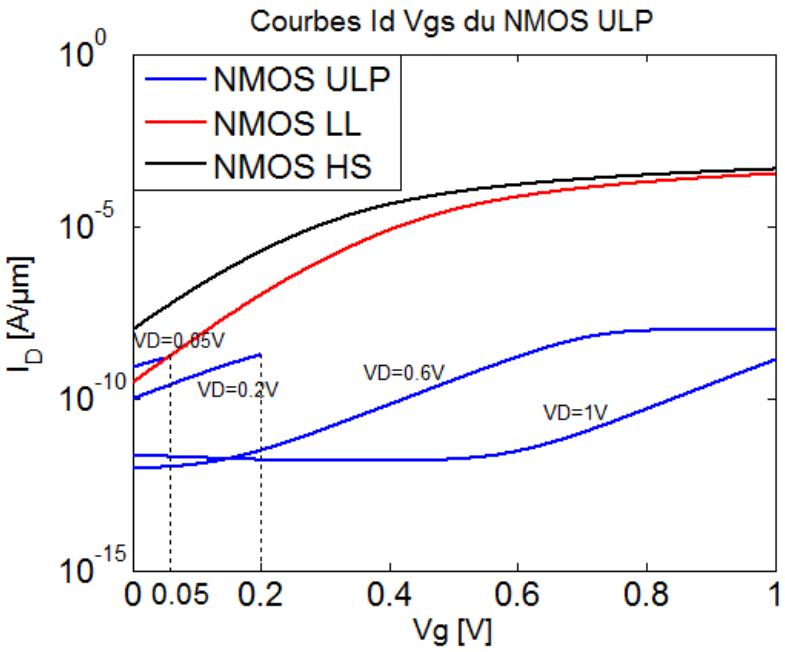


Fig. 3.8 : En bleu, courbes I_D V_{GS} du NMOS ULP pour V_{DS}=0.05, 0.2, 0.6 et 1V, en rouge courbe I_D V_{GS} du NMOS HS pour V_{DS}=1V, en noir du NMOS LL pour V_{DS}=1V. V_{BS}=0V. W_N=1µm, W_P=2.4µm, L=0.13µm, T=25°C.

- Dans un premier temps, l'augmentation de V_{GS} accroit progressivement la valeur du courant sous seuil. Cet effet n'est pas visible, car comme on a pu le voir à la fig. 3.7, ce courant sous seuil est bien inférieur aux courants de jonction et de grille. La valeur du courant I_{OFF} reste donc « bloquée » aux alentours du pA, ce qui correspond à la somme des courants de jonction et de grille que nous avons pu observer à la fig. 3.7 pour V_D = 0.6V. Pour les différentes courbes présentées, le courant I_{OFF} est bien le plus faible lorsque V_D vaut 0.6V, comme mentionné précédemment.
- Lorsque V_{GS} continue d'augmenter, le courant sous seuil devient plus important que le courant de jonction et de grille. Le courant I_{ON} commence à s'élever. Le NMOS de la figure 3.6 voit son impédance équivalente diminuer par rapport à celle du PMOS lorsque V_{GS} augmente, ceci entraine une augmentation de la tension de source de deux transistors (nœud N1 fig. 3.6). Le transistor PMOS possède donc une tension V_{sg} proche de zéro et reste en régime de faible inversion.

- Ensuite, V_{GS} dépasse la valeur V_D. Le transistor NMOS quitte progressivement le régime de faible inversion, pour entrer dans le régime triode lorsque V_{GS} dépasse V_D+V_{TN}. L'impédance du NMOS est alors insignifiante devant celle du PMOS et nous pouvons donc remplacer le NMOS par un court circuit. Nous obtenons alors le schéma équivalent d'un PMOS monté en diode bloquée. Le courant I_{ON} reste donc limité par le courant de fuite du PMOS. Le courant observé vaut alors environ 10 nA, ce qui est bien du même ordre de grandeur que le courant de fuite indiqué au tableau 3.2 pour le PMOS « HS ». Notons enfin que, pratiquement, ce régime n'est pas utilisé dans les circuits logiques, car la tension de grille ne peut y dépasser la tension d'alimentation.

3.5. L'inverseur ULP

Maintenant que nous avons mieux compris le fonctionnement des transistors ULP, nous pouvons les associer pour former un inverseur ULP dont le schéma est présenté à la figure 3.9. Le PMOS P1 et le NMOS N2 sont les transistors qui limitent le courant de l'inverseur. Le rapport de leur largeur a donc été choisi pour compenser la différence de mobilité des porteurs afin de garantir la symétrie entre le temps de montée et de descente. Les autres transistors possèdent une largeur de grille minimale de 0.15 µm.

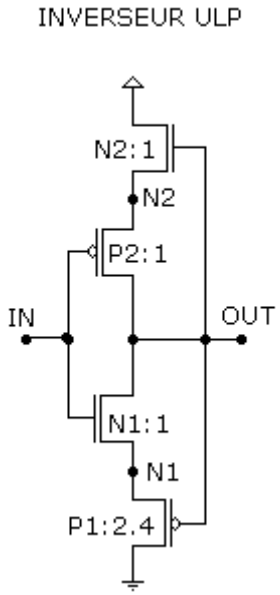


Fig. 3.9 : l'inverseur ULP : $W_{N1, N2, P2}$ = 0.15 µm, W_{P1} = 0.4 µm, V_{BNMOS}=0V, V_{BPMOS}=V_{DD}.

3.5.1. Courbe caractéristique de L'inverseur ULP

La courbe de transfert de l'inverseur est donnée à la figure 3.10. La tension d'alimentation a été choisie égale à 0.6V, ce qui correspond à la tension optimale des transistors du point de vue des courants de fuite.

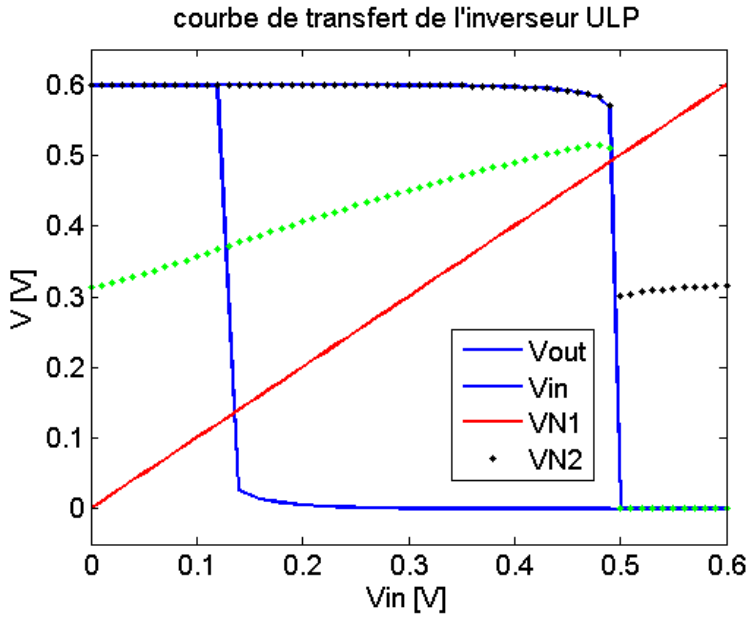


Fig. 3.10 : courbe de transfert de l'inverseur ULP. W_N=W_{P2}=0.15µm, W_{P1}=0.4µm V_{DD}=0.6V T=25°C.

Pour comprendre ce graphe, nous allons nous intéresser à l'évolution des signaux lors d'un flanc montant de la tension d'entrée. Une analyse similaire peut être réalisée lors d'un flanc descendant.

- Dans un premier temps, V_{IN} = 0. Le transistor P1 possède un grand V_{gs} et donc une faible impédance équivalente. Nous pouvons l'approximer par un court-circuit. La tension V_{OUT} est donc proche de V_{N2}. Il s'en suit que N2 observe un V_{gs} nul. Par ailleurs, le NMOS ULP est bloqué, la tension V_{gs} de N1 et de P1 est donc proche de $-V_{DD}/2$, ils possèdent dès lors une impédance équivalente bien plus élevée que celle de N2. Ainsi, les tensions V_{N2} et V_{OUT} sont proches de V_{DD}. La tension V_{N1} est proche de $V_{DD}/2$ comme nous l'avons déjà vu précédemment (section 3.3).
- Lorsque V_{IN} augmente, la symétrie de V_{gs} des transistors N1 et P1 implique que V_{N1} évolue de manière à être égale à $(V_{IN}+V_{OUT})/2$. N1 et P1 continuent d'observer un V_{gs} négatif. V_{OUT} reste donc proche de V_{DD}.
- Petit à petit, V_{IN} s'approche de V_{N1}. La tension V_{gs} des transistors N1 et P1 est de moins en moins négative. Leur impédance équivalente s'approche donc de celle de N2, ce qui réduit V_{OUT}. La diminution de V_{OUT} tend à réduire encore la tension V_{gs} de N1 et P1 ce qui diminue encore leur impédance équivalente. Une réaction positive est donc enclenchée. V_{OUT} finit par basculer sous V_{N1}, N1 quitte alors le régime de faible inversion et V_{OUT} se décharge à V_{N1}, ce qui augmente le V_{gs} de P1 qui va alors également accélérer la décharge de V_{OUT} par une autre boucle de réaction positive. Le PMOS ULP se bloque alors, P2 et N2 possèdent tous deux un V_{gs} négatif. Tous ces phénomènes se déroulent en cascade et provoquent une transition très abrupte de V_{OUT} par rapport à ce qui est observé dans le cas d'un inverseur conventionnel.

3.5.2. Évaluation des performances

Les performances de l'inverseur ont été évaluées grâce à des simulations ELDO. Pour les résultats présentés ci-dessous, un fan-out de 4 inverseurs ULP a été choisi. Une tension d'alimentation de 0.5V

a été appliquée à la cellule. Les temps de montée et de descente des signaux sont conditionnés par le délai d'une autre porte logique similaire. Afin de pouvoir tenir compte de ce point, d'autres inverseurs ULP placés en cascade ont servi de buffer d'entrée. Une comparaison est réalisée avec d'autres types d'inverseurs dont la définition est donnée dans le paragraphe ci-dessous. Les valeurs de comparaison sont empruntées à [30] qui obtient des résultats similaires pour les performances de l'inverseur ULP. Le courant I_{ON} représente le pic de courant observé dans la courbe caractéristique DC de l'inverseur, comme illustré à la figure 3.11. Au plus, celui-ci sera élevé, au plus le délai de l'inverseur sera faible. Cette valeur n'est donnée que pour l'inverseur ULP et l'inverseur classique, la valeur d'intérêt étant ici plutôt le délai de l'inverseur.

type	V_{DD} [V]	I_{stat}	I_{ON}	Délai	Puissance à 1kHz	Puissance à 10 kHz
ULP	0.5	2.7 pA	0.4 nA	2.5 µs	1.4 pW	3.2 pW
Standard	0.5	2.6 nA	14.5 µA	0.16 ns	1.3 nW	1.3 nW
High V_T	0.5	133 pA	-	0.45 ns	67 pW	69.6 pW
High V_T	0.2	88 pA	-	43 ns	17.6 pW	18 pW
High V_T RBB 1.2V	0.5	5.5 pA	-	4.6 ns	8.1 pW	10.5 pW

Tableau 3.3 : comparaison des performances de l'inverseur ULP, les valeurs présentées aux trois dernières lignes proviennent de [30].

L'inverseur standard présente un courant statique bien plus élevé que celui de l'inverseur ULP (par 3 ordres de grandeur). Par contre, l'inverseur ULP possède un délai plus élevé de 5 ordres de grandeur. Toutefois, les applications visées par les transistors ULP prennent comme facteur de mérite une faible puissance consommée et non pas de hautes performances dynamiques. Le rapport de la puissance dissipée par l'inverseur standard sur celle dissipée par l'inverseur ULP à 1kHz vaut 1000 et 400 à 10kHz. Par 1 ou 10 kHz, nous entendons une variation de l'entrée de l'inverseur à ces fréquences. Notons que pour l'inverseur standard, la puissance consommée ne varie pas dans cette gamme de fréquences, car elle est alors dominée par la consommation statique du circuit ce qui n'est pas le cas de l'inverseur ULP.

L'inverseur ULP doit être comparé aux autres techniques de réduction de puissance consommée. Une première méthode consiste à utiliser des transistors à tension de seuil élevée (inverseur « High V_T ». La quatrième ligne du tableau permet d'obtenir une comparaison grâce à l'utilisation de transistor à haute tension de seuil pour le même nœud technologique. Le rapport de la puissance consommée reste ici élevé (48 à 1 kHz et 22 à 10 kHz).

Une autre technique consiste à utiliser une faible tension d'alimentation de manière à faire fonctionner les transistors en régime sous le seuil. Les résultats donnés ici sont obtenus pour une tension d'alimentation de 0.2V ce qui est un choix très agressif pour le procédé 0.13 µm. Toutefois, même pour cette valeur, le rapport de puissance consommée s'élève toujours à 13 à 1Khz et à 6 à 10 kHz.

Le dernier mécanisme de réduction de la puissance consommée consiste à appliquer une tension de substrat négative (inverseur « High V_T RBB »). Cette technique permet d'obtenir une forte réduction de la puissance consommée. Cette fois la réduction de la consommation de l'inverseur ULP s'élève à 82.7 % à 1kHZ et à 69.5 % à 10kHz, ce qui reste appréciable. Si elle permet d'obtenir une consommation qui se rapproche de celle de l'inverseur ULP, l'application d'une tension négative sur

le substrat possède certains inconvénients. En effet, il est entre autres nécessaire de réaliser un sous circuit supplémentaire capable de générer cette tension de substrat.

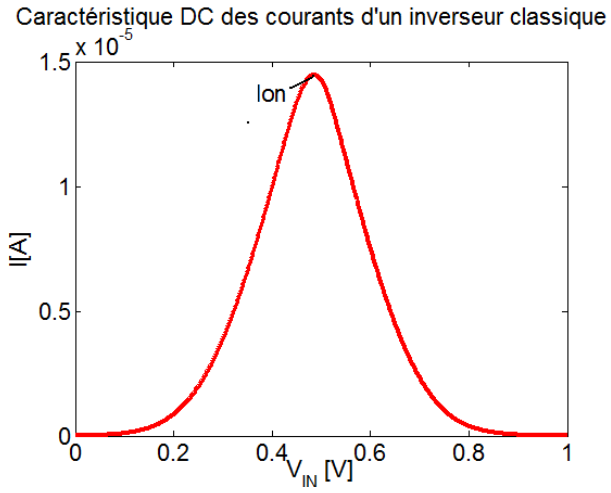


Fig. 3.11 : courbe caractéristique DC de l'évolution du courant dans un inverseur CMOS conventionnel.T=25°C.

3.6. Conclusion

L'utilisation conjointe d'un transistor NMOS et d'un transistor PMOS permet de réaliser un transistor ULP. Les courants de fuite de ce transistor sont réduits de plusieurs ordres de grandeur grâce à une forte réduction du courant sous seuil. Ce phénomène s'obtient lorsque le transistor ULP est bloqué par le développement d'une tension V_{gs} négative sur ses transistors constitutifs.

La combinaison de deux transistors ULP permet de créer un nouveau type d'inverseur. Lorsqu'il est inactif, celui-ci est traversé par un très faible courant statique, ce qui réduit alors fortement la puissance dissipée. Néanmoins, le délai reste plus élevé que celui d'autres inverseurs à basse consommation. Son utilisation semble donc réservée aux applications basse fréquence requérant une faible consommation.

4. Les mémoires SRAM ULP

4.1. Introduction

Différents schémas de cellules SRAM sont étudiés dans cette section. Cette analyse nous permettra de sélectionner l'architecture de cellule ULP optimale par rapport aux critères que nous avons développés à la section 2.4. Nous nous focalisons ici sur le fonctionnement de la cellule elle-même et non sur celui de la mémoire SRAM dans son ensemble. Les principales opérations visées sont donc l'écriture d'une valeur dans une cellule, la rétention de cette valeur et sa lecture. Afin de pouvoir quantifier les performances de chaque architecture, des simulations ELDO ont été réalisées.

Dans un premier temps, nous analyserons les cellules classiques à 6 et à 8 transistors. Les résultats obtenus nous serviront de base de comparaisons pour évaluer les performances des cellules ULP. Ensuite, nous aborderons différents choix d'architectures ULP. Les notions de stabilités face aux sources de bruits extérieures seront discutées en fin de section. Les simulations présentées ici ont été réalisées sous une température de 25°C. Une étude du comportement d'une des cellules présentée sous d'autres températures sera réalisée à la section 5. Rappelons que les connexions de substrat sont placées à la tension la plus faible du circuit dans le cas des NMOS et à la tension la plus élevée pour les PMOS.

4.2. La cellule 6 transistors

4.2.1. Présentation de la cellule

Les opérations de lecture et d'écriture sur une cellule standard ont déjà été présentées à la section 2.2.3, nous ne reviendrons plus sur ce point ici. Un schéma de ce type de cellule est rappelé à la figure 4.1. Notons que nous faisons ici l'hypothèse qu'un sense-amplifier est disponible par paire de bitlines. Certaines mémoires SRAM n'en possèdent qu'un nombre égal à la quantité de cellules auxquelles on accède simultanément, cet élément étant raccordé aux datalines qui sont elles-mêmes reliées aux bitlines à travers les transistors « bitline switch ». Dans la plupart des cellules modernes à faible consommation, il n'est cependant plus possible de ne disposer que d'un seul sense-amplifier. En effet, la réduction de la tension d'alimentation dégrade le rapport I_{ON}/ I_{OFF} des transistors d'accès de chaque cellule [23]. Il s'en suit que la somme des courants de fuite des cellules inactives peut atteindre le même ordre de grandeur que le courant I_{ON} de l'élément auquel on accède si un trop grand nombre de cellules sont connectées au même sense-amplifier. Lors d'une lecture, il devient dès lors difficile de discriminer un « 1 » d'un 0 étant donné que les bitlines se déchargent de la même manière, quelle que soit la valeur de l'élément mémorisé [27] [23]. Afin d'assurer la fonctionnalité de la mémoire [27] suggère de limiter le nombre de cellules raccordées aux mêmes bitlines de manière à ce que l'ensemble des courants de fuite des cellules inactives reste inférieur par un facteur 10 au courant de lecture.

Avant toutes choses, précisons que l'architecture et le dimensionnement de la cellule présentée sont basés sur [9] en tenant compte des suggestions de [8] en ce qui concerne la modélisation des bitlines.

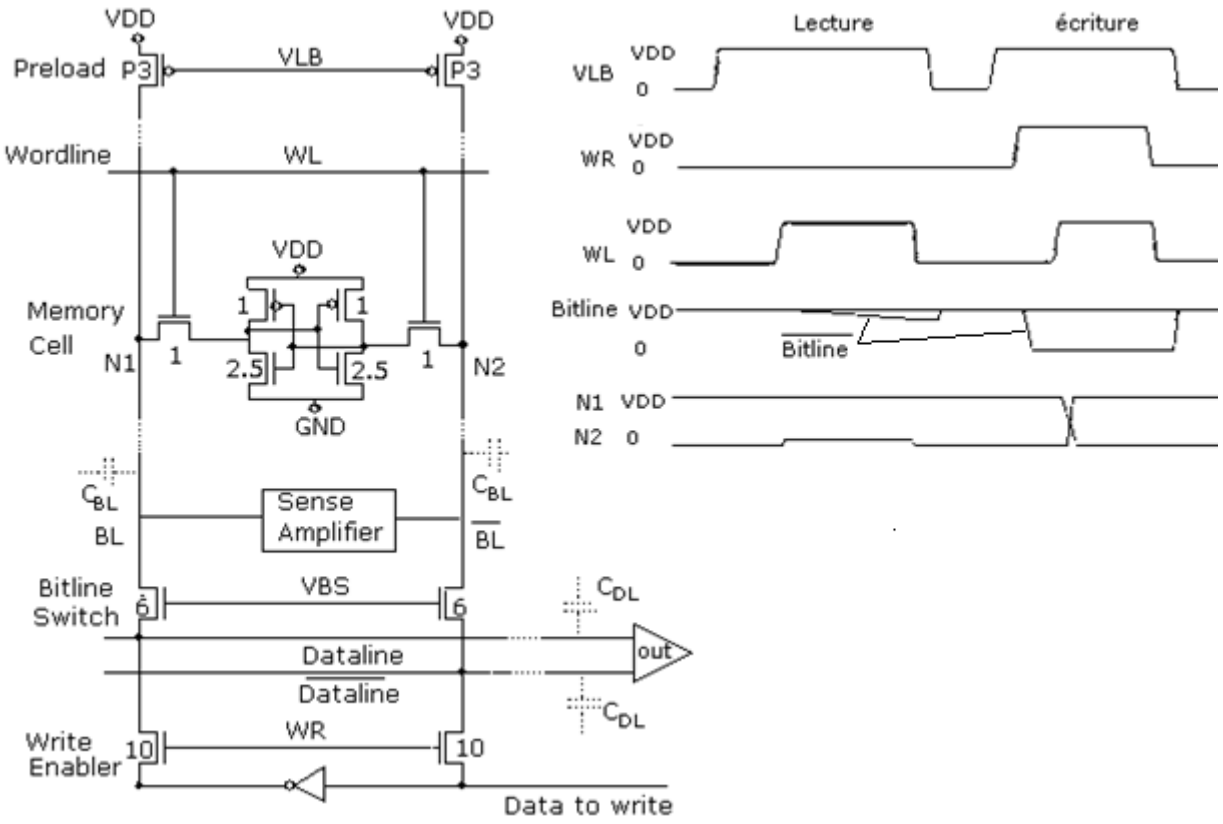


Fig. 4.1 : Schéma de l'architecture de la mémoire SRAM simulée et diagramme temporel des différents signaux appliqués. Les rapports W/L des transistors sont indiqués sur le schéma. Le sense-amplifier n'a pas été inclus à la simulation.

4.2.2. Simulations de la cellule

L'architecture présentée à la figure 4.1 a d'abord été simulée sur ELDO sous une tension d'alimentation de 1V. Les effets d'une réduction de la tension d'alimentation ont ensuite été observés grâce à un deuxième essai sous 0,5V. Nous avons choisi une mémoire de 256 cellules par bitline. Ainsi, la valeur approximée de capacité de bitlines est de 150 fF d'après les valeurs fournies dans [8]. Pour garantir le succès d'une opération de lecture nous avons considéré, comme suggéré dans [8], que les transistors d'accès devaient développer une différence de tension de 100 mV entre les bitlines pour que l'information puisse être traitée par le sense-amplifier.

- V_{DD}=1V

Le but de cette simulation est de valider l'architecture proposée ci-dessus et de recueillir une base de données pouvant servir à la comparaison avec les architectures ULP. Lors d'une lecture, les simulations montrent à la figure 4.2 que 0.39ns sont nécessaire au développement d'une tension de 100mV entre les deux bitlines. Lors d'une écriture, la wordline doit être activée durant 40ps pour faire basculer l'état de la cellule.

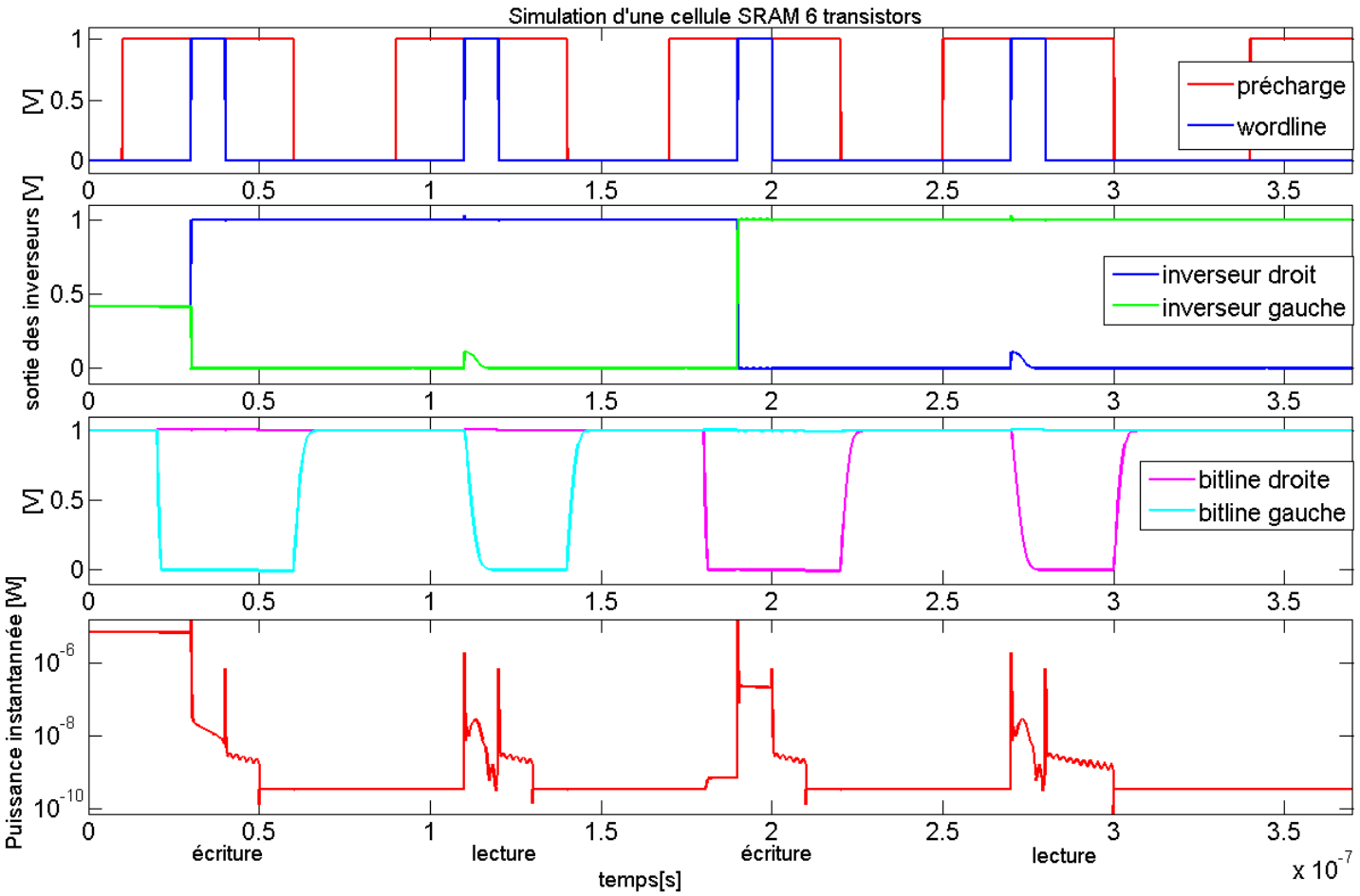


Fig. 4.2 : Simulation ELDO d'une cellule SRAM 6T classique. V_{DD}=1V. Deux opérations de lecture et d'écriture sont alternativement réalisées.

Sur le graphe du haut, lorsque le niveau de la tension précharge est bas, les bitlines sont chargées à V_{DD}. L'autre courbe indique l'évolution de la tension de wordline. Le deuxième graphe indique la valeur mémorisée par les inverseurs du point mémoire. On peut remarquer lors d'une lecture la petite perturbation sur le zéro logique. La perturbation s'élève à 110mV. Celle-ci est due au phénomène expliqué dans la section 2.2.3. Un bon rapport de taille entre les transistors NMOS de l'inverseur et d'accès permet de garder un contrôle sur ce phénomène. Le troisième graphe montre l'évolution des tensions de bitlines. Rappelons que d'après [8], lors d'une lecture, une différence de tension de 100mV est une valeur suffisante pour pouvoir être traitée par le sense-amplifier. Les tensions de bitlines sont bien restaurées à V_{DD} lors des précharges. Enfin, la puissance instantanée dissipée dans la cellule mémoire est affichée sur le dernier graphe. Lorsque la cellule n'est pas sollicitée, celle-ci vaut 0.71 nW en tenant compte des pertes dues aux transistors d'accès (celles-ci sont de 0.37nW).

- V_{DD}=0.5V

Cette deuxième simulation nous permet de visualiser l'effet d'une réduction de la tension d'alimentation sur le fonctionnement de la mémoire. Cette pratique est courante dans les circuits à faible consommation et ces résultats nous seront donc utiles pour l'évaluation des performances des cellules ULP. Les résultats de la simulation sont donnés à la figure 4.3.

Nous pouvons d'abord remarquer sur le dernier graphe que la consommation statique de la cellule a été réduite par environ un facteur 2 (consommation de l'ordre de 0.38 nW). Toutefois, la vitesse de

la lecture est également affectée. En effet, la tension d'overdrive des transistors est fortement diminuée ainsi que le courant qu'ils peuvent développer. Cette fois-ci pour pouvoir faire apparaitre une tension de 100mv entre les deux bitlines, les simulations montrent que les transistors d'accès doivent être activés par la wordline pendant 4.0 ns, c'est-à-dire 10 fois plus longtemps que pour une tension d'alimentation de 1 V.

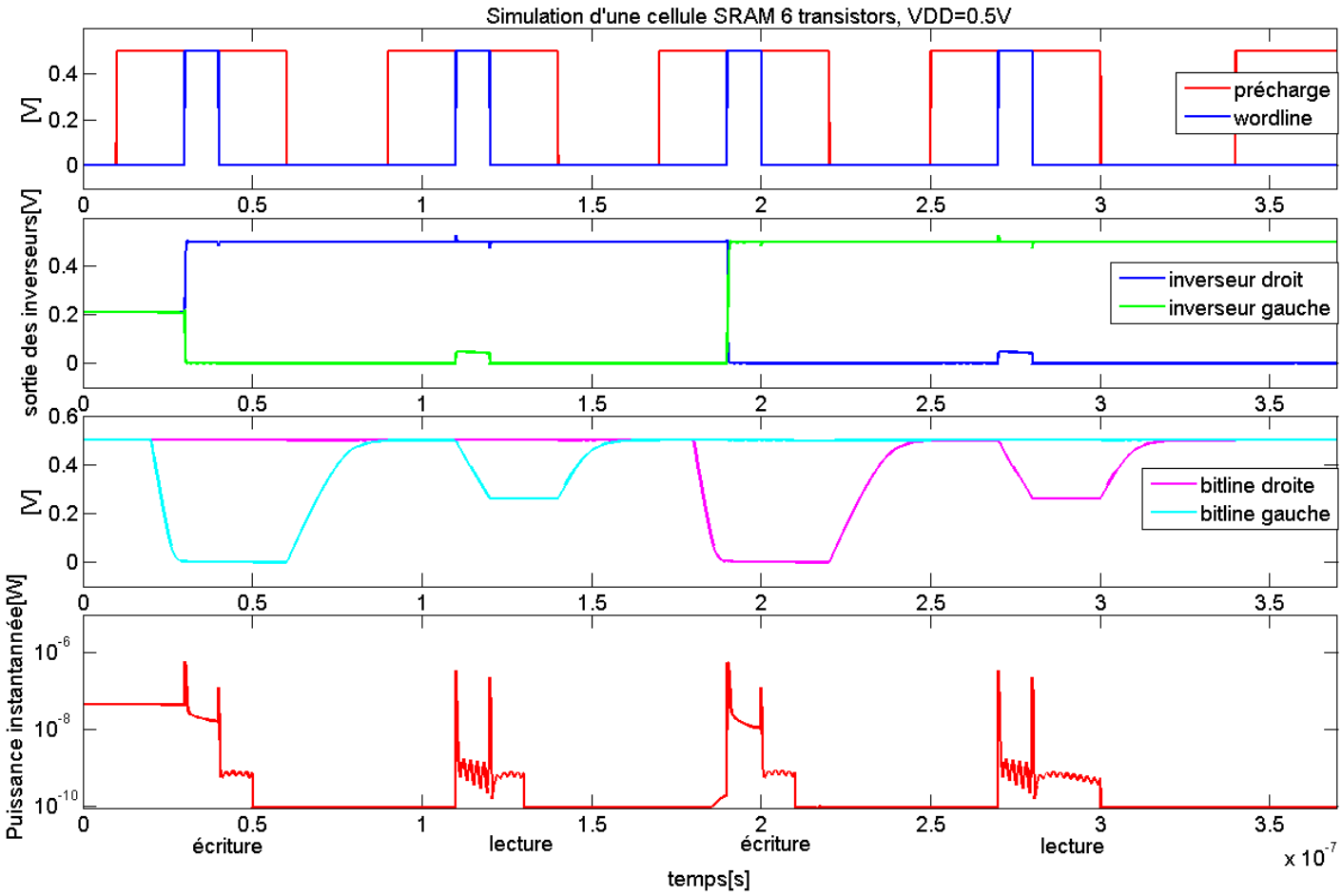


Fig. 4.3 : Simulation ELDO d'une cellule SRAM 6T classique. V_{DD}=0.5v, deux opérations de lecture et d'écriture sont alternativement réalisées.

4.3. La cellule 8 transistors

4.3.1. Présentation de la cellule

Une des principales limitations de la cellule à 6 transistors concerne sa sensibilité lors des opérations de lecture, comme nous le verrons à la section 4.8.1. Une alternative consiste à sacrifier un peu de surface et à ajouter un buffer de lecture via deux transistors supplémentaires [7] [27]. Une telle cellule est schématisée à la figure 4.4. Cette fois-ci, les nœuds de mémorisation sont complètement isolés du circuit de lecture et aucune dégradation des tensions mémorisées n'est observée lors de l'activation de la wordline de lecture. Il n'est donc plus nécessaire que les NMOS des inverseurs

soient plus forts que les transistors d'accès. Il est dès lors possible de prendre la taille minimale pour tous les transistors (ici L=0.13 µm, W=0.15µm).

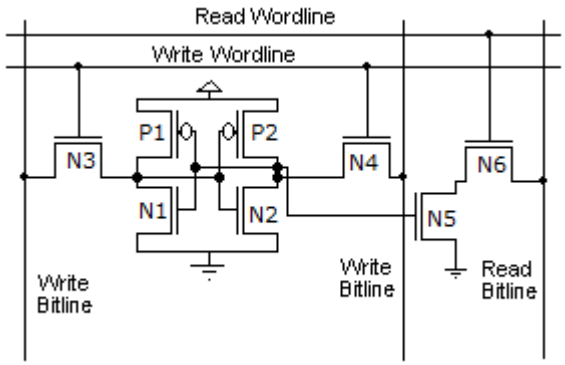


Fig. 4.4 : Schéma d'une cellule 8 transistors. Les transistors N5 et N6 servent de buffer de lecture.

L'opération d'écriture est identique à celle de la cellule à 6 transistors. Lors d'une lecture, la bitline de lecture est préchargée à V_{DD}. Ensuite, la wordline de lecture est activée. Un chemin peut donc être créé au travers de N5 et de N6 entre la tension de masse et la bitline de lecture. En fonction de la valeur mémorisée, ces transistors déchargeront la bitline ou la laisseront chargée à V_{DD}. La lecture n'est donc plus réalisée par comparaison du niveau de deux bitlines (lecture différentielle), mais en récupérant l'information contenue sur une seule bitline.

4.3.2. Simulation de la cellule

Une simulation de la cellule a été réalisée pour une tension d'alimentation de 1 V. Les résultats de cette simulation sont présentés à l'annexe 2. Cette fois-ci, une durée de 0.50ns est nécessaire pour faire chuter la tension de bitline de 100mV. Il pourrait sembler surprenant que cette opération soit plus longue que dans le cas de la cellule à 6 transistors, alors que le phénomène de dégradation de niveau mémorisé n'est plus présent. Ceci s'explique par le fait que cette fois-ci tous les transistors sont de taille minimale, contrairement à la cellule 6 transistors. Une comparaison peut être faite en augmentant la longueur de grille des transistors N5 et N6 de la figure 4.4 de manière à ce qu'ils occupent la même surface que les transistors N1 et N2 de la figure 4.1. Cette fois, le temps de lecture ne s'élève plus qu'à 0.34ns, ce qui est bien 0.05ns plus rapide que dans le cas de la cellule 6 transistors. Lors d'une écriture, le temps nécessaire au basculement de l'état des inverseurs reste sensiblement le même. Il vaut cette fois-ci 37ps. Ceci s'explique par le fait que la cellule est isolée du circuit de lecture par la grille de N5, et que le circuit de lecture n'a donc que très peu d'influence sur l'écriture. La consommation de la cellule s'élève maintenant à 0.95nW en moyenne. Celle-ci évolue en fonction de la valeur stockée par la cellule. En effet, les courants de fuite du buffer de lecture dépendent de la tension de grille de N5 et donc de la valeur mémorisée par la cellule.

4.4. La cellule ULP 10 transistors

4.4.1. Présentation de la cellule

L'architecture de la mémoire est semblable à celle de la cellule à 6 transistors si ce n'est que les inverseurs CMOS ont été remplacés par des inverseurs ULP. Les transistors d'accès représentent maintenant la principale source de perte de la cellule. Pour pouvoir bénéficier pleinement de

l'utilisation de la logique ULP, une tension de wordline négative leur est appliquée en rétention. Cela permet de réduire les fuites travers des transistors d'accès de deux ordres de grandeur grâce à l'application de tension V_{GS} négative. Une représentation de cette architecture est fournie à la figure 4.5. La connexion de substrat des NMOS est reliée à la masse, celle des PMOS à V_{DD}. Les inverseurs ULP sont réalisés à l'aide de transistors « HS », car ceux-ci possèdent un courant de fuite suffisant pour le bon fonctionnement des inverseurs ULP. Les transistors d'accès et autres transistors périphériques à la cellule sont de type « LL » de manière à minimiser les pertes statiques.

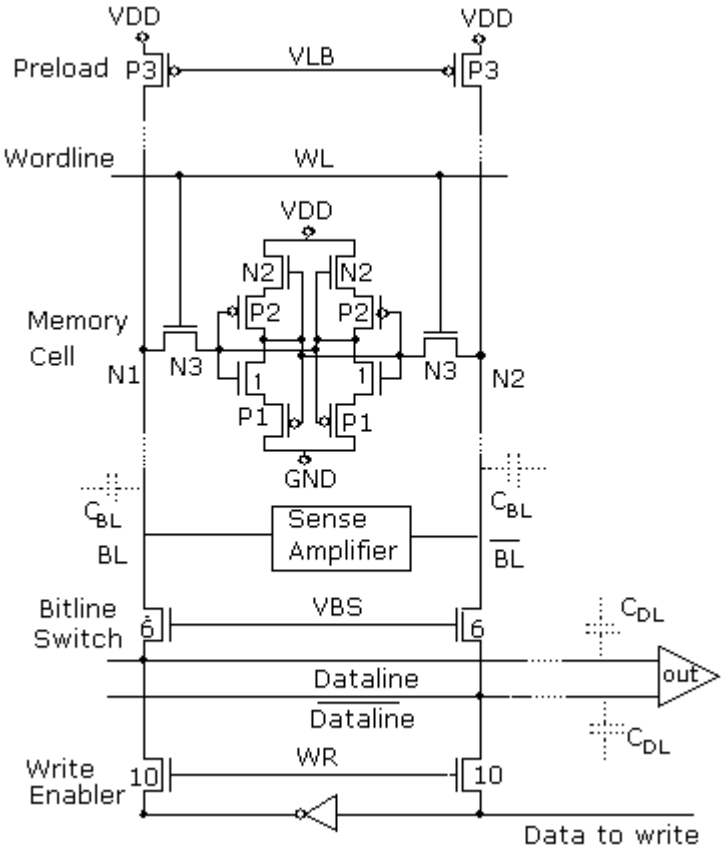


Fig 4.5 : Schéma de la cellule SRAM ULP 10 transistors. V_{BN}=0, V_{BP}=V_{DD}. Le sense-amplifier n'a pas été simulé. Lorsque fixés, les rapports W/L sont indiqués sur le schéma.

Cette fois-ci, la cellule sera plus lente, surtout pour la lecture de données à cause du faible courant qu'est capable de fournir un transistor ULP. Mais les différentes opérations de lecture et d'écriture peuvent être réalisées dans la même séquence que pour la cellule SRAM standard.

Le point délicat est ici le dimensionnement du transistor d'accès. En effet, nous avions vu dans le cas de la cellule SRAM classique, que pour éviter la perturbation de la valeur mémorisée, il était nécessaire de choisir un transistor NMOS plus « fort » pour les inverseurs de la mémoire que pour les portes d'accès. Ce n'est manifestement pas possible dans le cas où l'on utilise des transistors ULP pour réaliser ces inverseurs. En effet, lorsqu'ils sont passants, l'impédance équivalente de ces transistors revient à celle d'un transistor bloqué. Comme nous pourrons l'observer dans les simulations, cela pose un double problème. En effet, d'une part, la robustesse de la cellule se trouve fortement dégradée. D'autre part, cela engendre également un temps de lecture encore fortement accru.

Plusieurs configurations ont donc été testées dans le but de mettre en évidence la portée d'une modification du dimensionnement de certains transistors. Le tableau 4.1 donne les différentes dimensions des transistors de la mémoire représentés à la figure 4.5. Un test a également été réalisé en préchargeant les bitlines à $V_{DD}/2$ au lieu de V_{DD}.

	N2		**N3**		**P1**		**P2**		**P3**	
	W[µm]	L[µm]	W[µm]	L[µm]	W[µm]	L[µm]	W[µm]	L[µm]	W[µm]	L[µm]
Configuration n°1	0.15	0.13	**0.15**	**0.13**	**0.15**	**0.13**	0.15	0.13	**0.75**	**0.13**
Configuration n°2	0.15	0.13	**0.15**	**0.13**	**0.75**	**0.13**	0.15	0.13	**0.75**	**0.13**
Configuration n°3	0.15	0.13	**0.15**	**0.65**	**0.15**	**0.13**	0.15	0.13	**0.15**	**0.13**
Configuration n°4	0.15	0.13	**0.15**	**0.65**	**0.75**	**0.13**	0.15	0.13	**0.15**	**0.13**
Configuration n°5	**030**	**0.13**	0.15	0.65	0.75	0.13	**0.45**	**0.13**	0.15	0.13

Tableau 4.1: les différentes configurations du dimensionnement de la mémoire simulée.

4.4.2. Simulations de la cellule

Les simulations ont été réalisées sous une tension d'alimentation de 1 V. Les résultats sont donnés à la figure 4.6 pour la configuration nº 4 (tableau 4.1). Les graphes obtenus dans les autres cas sont disponibles en annexe 2. Cette configuration a donné les meilleurs résultats au prix de plus d'espace occupé. Le temps nécessaire à une opération de lecture est de 7.85 µs. Une comparaison plus détaillée des résultats obtenus pour les différents dimensionnements est réalisée plus loin.

Le premier graphe indique comme précédemment les périodes de précharge et d'activation de wordline. Remarquons l'application d'une tension de wordline négative en rétention. Le deuxième graphe représente le niveau de sortie des inverseurs. On observe une forte dégradation du zéro lors d'une opération de lecture (1). Ceci est dû à la faiblesse du NMOS ULP de l'inverseur par rapport au transistor d'accès. Une solution à ce problème qui n'a pas été examinée ici pourrait consister à réduire la tension de wordline de manière à réduire la force du transistor d'accès. Cependant, même si le point critique est le courant des transistors ULP, cela aurait pour conséquence d'accroitre le temps de lecture.

Dans l'état de rétention (2), la sortie de l'inverseur mémorisant l'état bas n'est pas exactement égale à zéro (3 mV). Il faut ici mettre en cause l'impédance équivalente élevée du transistor NMOS ULP de l'inverseur. Le diviseur résistif, illustré à la figure 4.7, qu'il forme avec les transistors d'accès et de précharge lorsqu'ils sont bloqués implique ce petit écart. Dans notre cas, l'application d'une tension de wordline négative permet d'augmenter fortement l'impédance équivalente du transistor d'accès. Ceci permet de limiter ce phénomène.

Enfin, (3) permet de voir ce qui se passe lors d'une opération d'écriture sur une autre cellule de la même bitline. A cet effet, nous avons volontairement retardé le moment d'activation des transistors d'accès par la wordline alors qu'une donnée à écrire est présente sur les bitlines. A nouveau, l'application d'une tension de wordline négative permet d'isoler le nœud mémoire. Il s'ensuit que la tension mémorisée n'est pas dégradée.

Des mesures avec une tension de wordline nulle ont également été réalisées. Dans ce cas de figure, l'élévation du zéro en rétention (2) est de 21 mV et durant une écriture sur une autre cellule (3), la tension du nœud mémoire à l'état haut chute à 925mV. Enfin, le quatrième graphe nous donne l'évolution de la puissance dissipée par le point mémoire. L'utilisation des inverseurs ULP permet de réduire significativement sa valeur à 11.7 pW, une réduction par un facteur 60 de la consommation d'une cellule classique. Notons la présence de petites perturbations sur le niveau de tension des nœuds de mémorisation liée à l'utilisation des transistors d'accès (4). Une discussion sera réalisée à ce sujet à la section 4.5.2, où ce phénomène sera mieux visualisé.

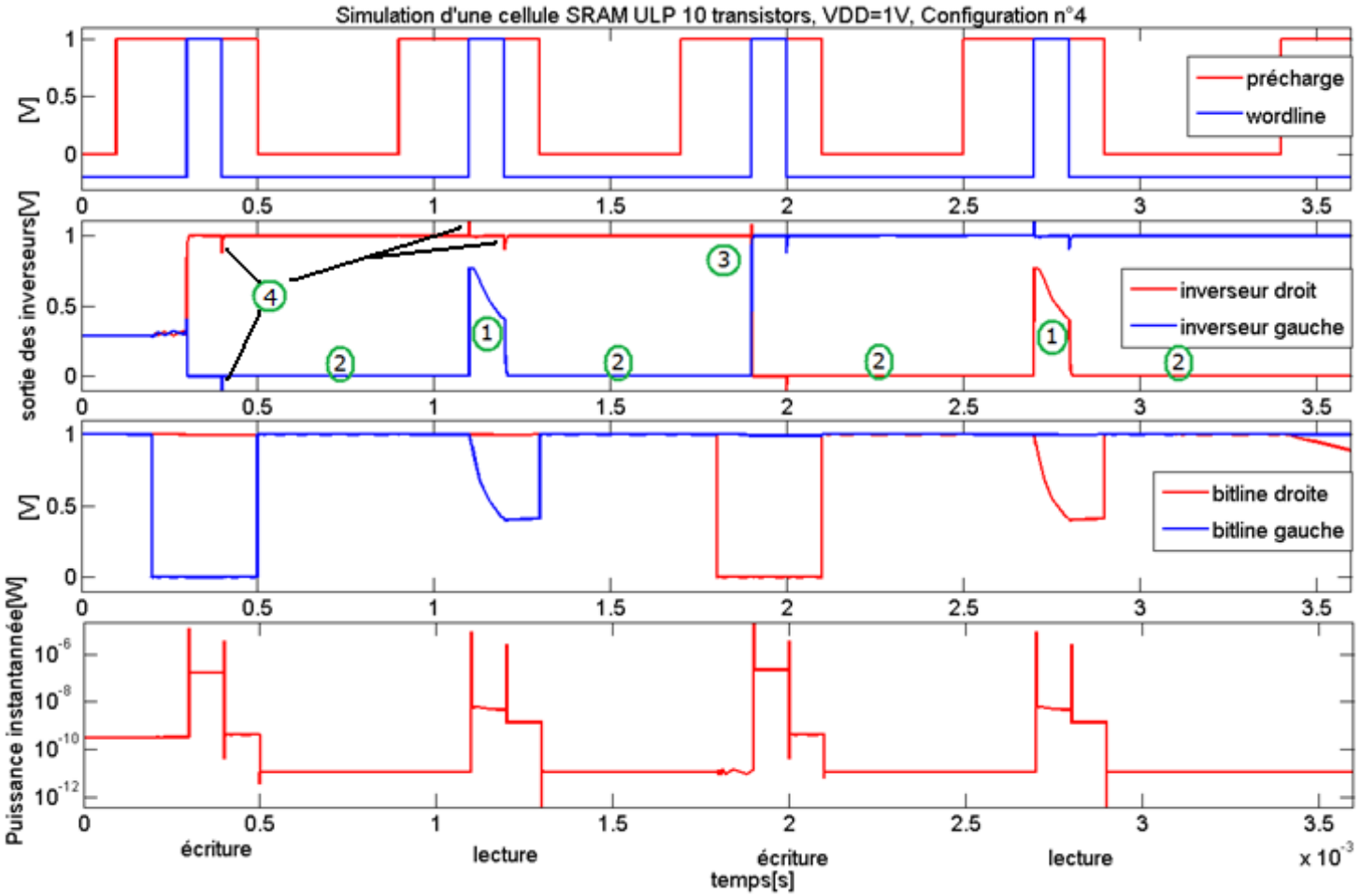


Fig. 4.6 : Simulation ELDO d'une cellule SRAM ULP 10 transistors. V_{DD}=1V. Deux opérations d'écriture et deux opérations de lecture sont alternativement réalisées.

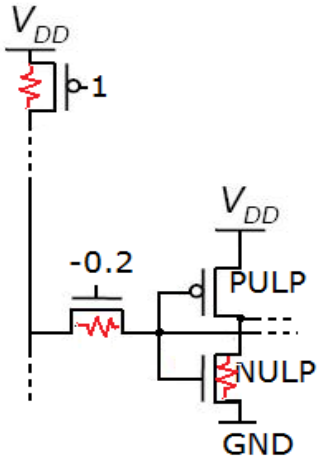


Fig. 4.7 : Représentation du diviseur résistif formé par la cellule en mode rétention.

Le tableau 4.2 reprend les performances observées pour les différents dimensionnements de la cellule.

	Temps de lecture [µs]	Niveau bas à la lecture* [mV]	Niveau bas en rétention* [mV]	Consommation statique [pw]
Configuration n°1	54.6	847	17	11.17
Configuration n°2	7.63	800	3	11.21
Configuration n°3	56.2	820	16	11.5
Configuration n°4	7.85	765	3	11.7
Configuration n°4, V_{RETWL}=0*	19	765	21	-

*Tableau 4.2 : performances mesurées lors de la simulation pour les différentes configurations de dimensionnement de la mémoire. *la deuxième colonne donne la dégradation de la tension au nœud mémorisant l'état bas en début de lecture, la troisième colonne de l'état haut en rétention. La dernière ligne correspond aux mesures réalisées en appliquant une tension de wordline dont le niveau bas est de 0V au lieu de -0.2 V.*

On peut remarquer que le fait d'accroitre la largeur de grille du PMOS du transistor ULP de type N (configurations 2 et 4) permet une nette amélioration des performances. L'effet est essentiellement d'accroitre le courant que sont capables de délivrer les transistors ULP de type N. Ceci garantit donc un temps de lecture significativement plus court. Un autre effet est d'accroitre la capacité présente au nœud de mémorisation. Ceci permet de réduire l'amplitude de la dégradation de niveau mémorisé bas lors d'une lecture, car les charges redistribuées lors de l'activation de la wordline seront réparties sur une plus grande capacité.

En mode de rétention, le dimensionnement du transistor d'accès (configurations 3 et 4) perd de son attrait si l'on s'autorise à appliquer une tension de wordline négative. En effet, dans ce cas, l'isolation du nœud mémoire en rétention est bien supérieure. Par contre, lors d'un accès en lecture, l'augmentation de la longueur de transistor d'accès permet d'opposer une résistance au phénomène de redistribution des charges entre capacités liées au point mémoire et capacités de bitlines et datalines. L'élévation de tension du nœud de mémorisation est donc réduite, mais le temps de lecture a

légèrement augmenté. Cependant, seul un dimensionnement excessif de la longueur de grille du transistor d'accès permettrait de réduire significativement ce phénomène.

La réduction de la largeur de grille du transistor de précharge (configurations 1 et 2) devrait permettre de réduire son courant de fuite au sacrifice du temps de précharge. Dès lors, le temps de lecture devrait être amélioré étant donné que l'isolation de la bitline par rapport à son alimentation est accrue. Il serait donc plus aisé pour le point mémoire de décharger la capacité parasite de bitline. Cependant, seul un léger effet a pu être observé.

Le temps de lecture est amélioré par l'application d'une tension de wordline négative. En effet, en rétention, les niveaux logiques au sein de la cellule sont meilleurs dans ce cas de figure. Ceci permet d'améliorer les performances des transistors d'accès lors d'une lecture.

Le tableau 4.3 reprend une synthèse des dépendances de certaines caractéristiques du circuit mémoire par rapport à ces différents dimensionnements.

	Temps de lecture	**Niveau bas à la lecture**	**Niveau bas en rétention**	**Espace occupé**
Augmenter L du transistor d'accès	± 0	+	± 0	_
Diminuer W du transistor de précharge	0	0	0	0
Augmenter W du PMOS du NMOS_ULP	+	+	+	_
Tension de wordline négative en rétention	+	0	+	0

Tableau 4.3 : synthèse des effets des modifications du dimensionnement des différents transistors du circuit.

- Précharge à $V_{DD}/2$.

Enfin, la cellule a été simulée dans le cas d'une précharge des bitlines et datalines à $V_{DD}/2$. L'objectif est ici de voir s'il est possible de réduire le temps de lecture notamment en réduisant les conséquences du phénomène de redistribution des charges au nœud de mémorisation. La configuration n° 5 du dimensionnement de la cellule (voir tableau 4.1) a été appliquée. Comme nous pouvons le voir sur le troisième graphe de la figure 4.8, la dégradation des états haut et bas des nœuds de mémorisation est telle que la lecture ne peut être effectuée dans un temps plus rapide, il est même détérioré (32 µs) et la robustesse de la cellule à toute perturbation est très faible. Cette solution ne semble donc pas pouvoir être retenue.

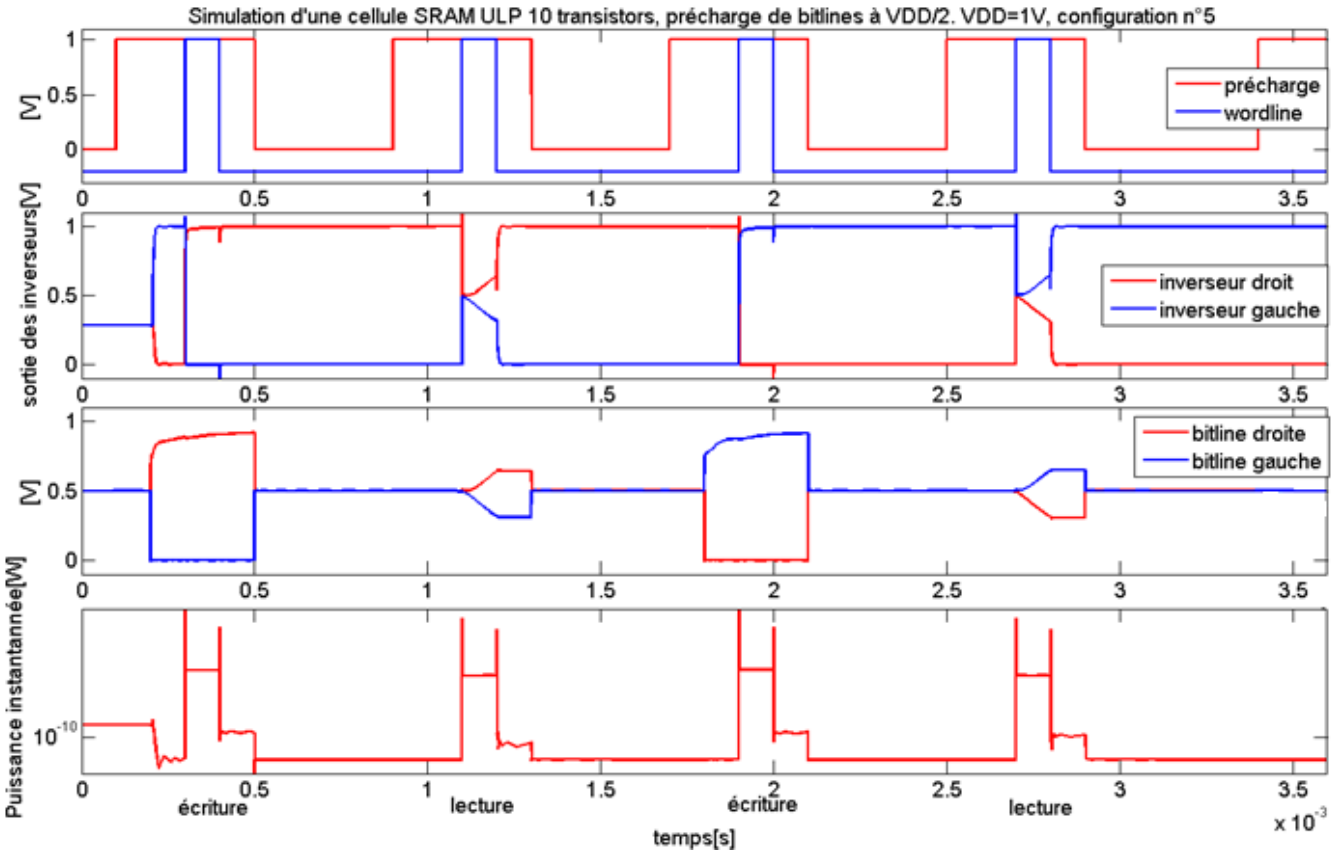


Fig. 4.8 : Simulation ELDO d'une cellule SRAM ULP 10 transistors. V_{DD}=1V, précharge des bitlines V_{DD}/2. Deux opérations d'écritures et deux opérations de lectures sont alternativement réalisées.

4.5. La cellule ULP 12 transistors

4.5.1. Présentation de la cellule

Comme pour la cellule étudiée précédemment, des transistors « HS » permettent de réaliser le point mémoire grâce à un latch ULP. Les transistors d'accès sont de type « LL » afin de minimiser leur courant de fuite.

A nouveau, deux inverseurs ULP montés en tête-bêche forment le cœur de la cellule mémoire (fig. 4.9). L'accès en écriture au latch ULP est réalisé grâce à deux transistors d'accès N7 et N8, ce qui permet de s'affranchir du délai élevé de l'inverseur ULP lors d'une écriture. Ici, deux transistors supplémentaires permettent de réaliser l'opération de lecture sur une bitline séparée, comme pour la cellule 8 transistors classique. Ceci permet également de ne plus dépendre du délai élevé de l'inverseur ULP pour l'opération de lecture. La cadence de la SRAM est donc comparable à celle de la cellule 8 transistors. De manière à tirer pleinement profit du mécanisme ULP, le courant de fuite des transistors d'accès est limité en rétention par l'application d'une tension négative de 200mV sur les wordlines de lecture et d'écriture.

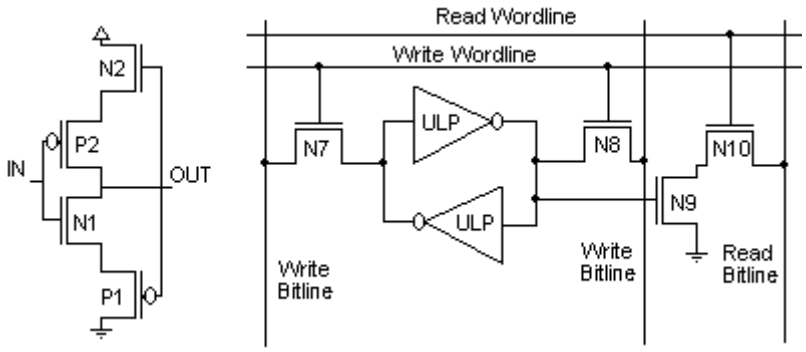


Fig 4.9 : Schéma de la cellule SRAM 12T. Un buffer supplémentaire permet de réaliser l'opération de lecture. $V_{BN}=0$, $V_{BP}=V_{DD}$.

Etant donné que le nœud mémoire est isolé des transistors d'accès en écriture par la grille de N9, il est maintenant possible de garder la taille de tous les transistors minimale (w=0.15µm, L=0.13µm). Les transistors d'accès N7 et N9 doivent être dimensionnés uniquement pour garantir le succès d'une opération d'écriture. Pour cela ils doivent être suffisamment forts pour vaincre la partie « pull up » de l'inverseur ULP. Toutefois, ils peuvent être choisis de taille minimale, car ils resteront dans tous les cas bien plus forts que les transistors ULP de type P.

4.5.2. Simulations de la cellule

La cellule a été simulée sous ELDO sous une tension d'alimentation de 1V. La figure 4.10 reprend les résultats de la simulation. La première constatation est que l'ajout de transistors supplémentaires permet d'atteindre des vitesses de lecture et d'écriture de l'ordre de grandeur de ceux d'une mémoire SRAM classique. En effet, le temps de lecture n'est que de 0.5ns. Le temps d'écriture s'élève à 1.37ns. Pour expliquer ce temps d'écriture un peu plus élevé, il faut se souvenir de l'hystérèse présente dans la caractéristique de l'inverseur ULP. Le NMOS d'accès devant transmettre une valeur haute devra amener le nœud de mémorisation proche de V_{DD} pour dépasser le seuil de basculement haut. Or, les NMOS d'accès ne peuvent transmettre une valeur plus élevée que V_{DD}-V_{TN}[6] et travaillent en fin d'écriture avec une tension d'overdrive réduite. Dans le cas de la cellule ULP 10 transistors, un temps d'écriture de cet ordre de grandeur ne pose pas de problèmes étant donné que la cadence de la mémoire est imposée par l'opération de lecture. Ceci n'est plus le cas pour la cellule ULP 12 transistors.

Lors d'une lecture, il faut à ce titre considérer le cas le plus défavorable où une donnée est lue immédiatement après avoir été écrite. En effet, nous venons de voir que les NMOS d'accès en écriture, N7 et N9, ne peuvent pas totalement transmettre une valeur haute au nœud de mémorisation. Le latch ULP va donc devoir régénérer cette tension ce qui, étant donné le délai de l'inverseur ULP, est relativement long (2.4µs). Il est donc possible que lors d'un accès en lecture, la tension V_{GS} du NMOS N9 ne soit pas maximale, ce qui ralentit légèrement la lecture. Dans le cas d'une lecture suivant immédiatement une écriture, le temps de lecture s'élève à 0.75ns.

La consommation statique de la cellule s'élève cette fois à 13pW.

[6] Il est utile de rappeler ici qu'une des possibilités pour réduire le temps nécessaire à une opération de lecture de n'importe quelle cellule consiste à élever la tension de wordline au-delà de la tension d'alimentation du latch [21]. Toutefois dans certains cas, cette solution pourra être incompatible avec les conditions de stabilité (nous discuterons de la notion de stabilité plus loin dans cette section). Une piste similaire, consistant à réduire la tension d'alimentation du latch en laissant celle des wordlines inchangée, a été développée plus loin dans ce travail à la section 7.1.

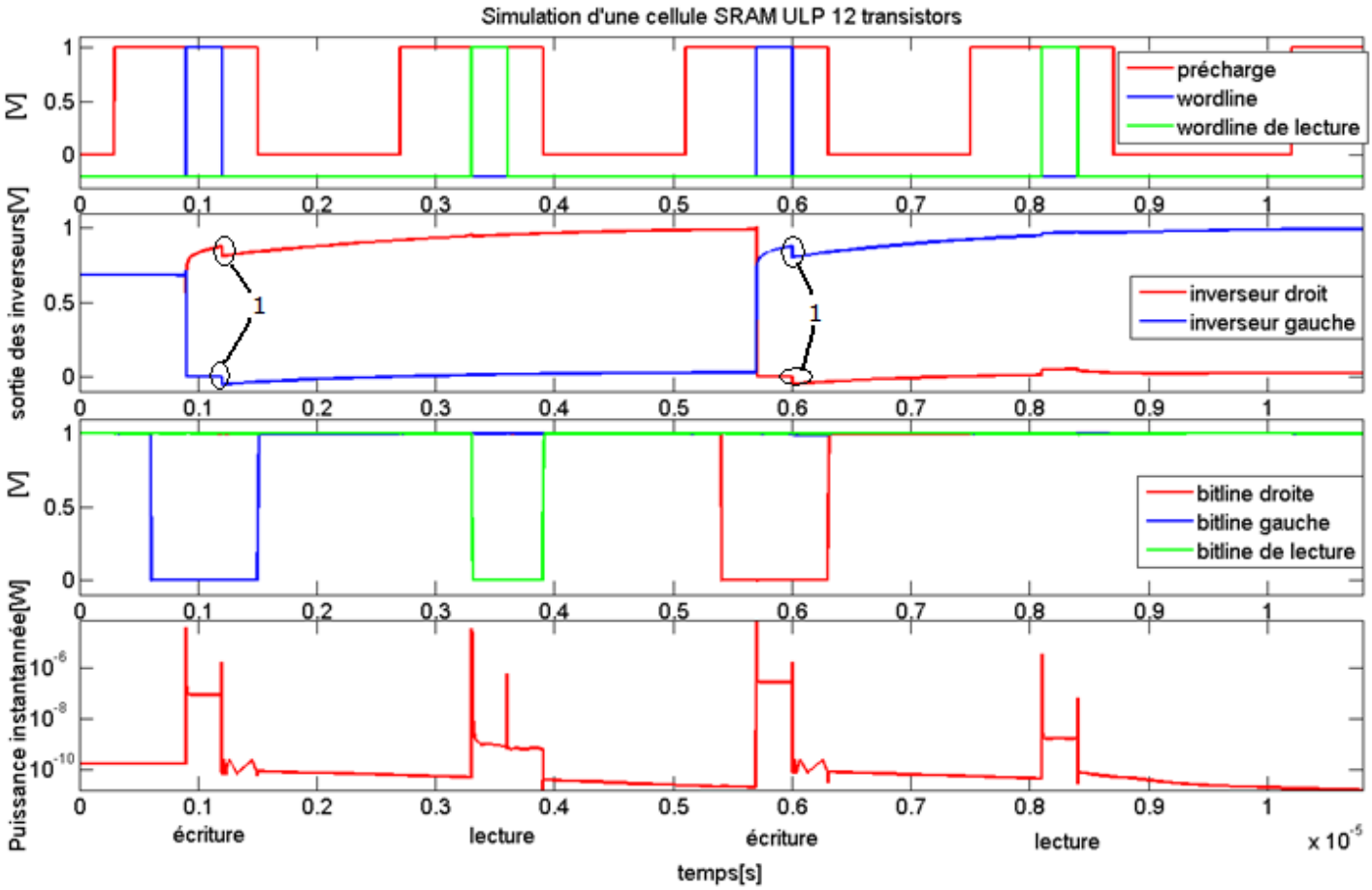


Fig 4.10: Simulation ELDO d'une cellule SRAM ULP 12 transistors. V_{DD}=1V. Deux opérations d'écriture et deux opérations de lecture sont alternativement réalisées.

Notons enfin la présence d'une perturbation de niveaux de tensions des nœuds de mémorisation suite à une opération d'écriture. Ce phénomène était déjà présent pour la cellule à 10 transistors, mais est plus visible au graphe 4.10 grâce au choix de l'échelle de temps. Cette perturbation est liée d'une part au phénomène de couplage capacitif et d'autre part à celui d'injection de charges.

Il existe un certain couplage capacitif entre la capacité d'overlap des transistors d'accès et la capacité du nœud de mémorisation. Lors d'un abaissement de la tension de wordline, cela entraine une variation de la tension mémorisée :

$$\Delta V_{node} = -\frac{C_{overlap}}{C_{overlap} + C_{node}} \cdot (V_{WL\,on} - V_{WL\,off})$$

Où C_{node} comprend les capacités parasites des inverseurs ULP, la capacité de jonction du transistor d'accès en écriture et les capacités de grille et d'overlap de N9. $C_{overlap}$ représente la capacité d'overlap du transistor d'accès en écriture comprise entre sa grille et le nœud de mémorisation.

Par ailleurs, lorsque les transistors d'accès sont coupés, les charges présentes dans le canal doivent être évacuées. Il s'en suit une injection de charges négatives au nœud de mémorisation. Ce phénomène est décrit par [32] qui donne une évaluation de la quantité de charges injectées de cette manière :

$$Q_{inj} = 0.5 \cdot C_{ox} \cdot (V_{WL\,on} - V_{th\,access} - V_{node})$$

Il est donc possible d'en déduire la modification du niveau de tension du nœud de mémorisation :

$$\Delta V_{node} = -\frac{Q_{inv}}{C_{node} + C_{overlap}}$$

Ces deux phénomènes ensemble entrainent une variation de la tension mémorisée d'environ 60 mV, ce qui contribue quelque peu au temps nécessaire aux inverseurs ULP pour restaurer le 1 logique suite à une écriture.

4.6. La cellule ULP 8 transistors

4.6.1. Présentation de la cellule

A nouveau, les transistors « high speed » permettent de réaliser le point mémoire et les transistors d'accès sont de type « low leakage ». Cependant, cette fois-ci nous utilisons un latch ULP d'une nouvelle architecture présentée à la figure 4.11 (a). Cette cellule est en réalité une variante de celle réalisée dans [8]. Afin de minimiser les fuites à travers les transistors d'accès, la tension WLR est de -200 mV lorsqu'aucune écriture n'est réalisée.

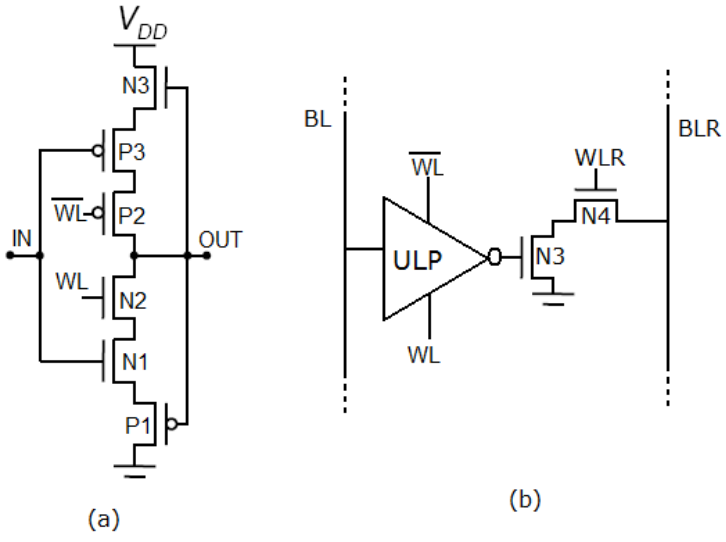


Fig. 4.11 : (a) Schémas d'un nouveau latch ULP. (b) Architecture de la cellule ULP 8 transistors. $V_{BN}=0$, $V_{BP}=V_{DD}$.

Avant d'examiner la structure du point mémoire dans son ensemble, il est utile de comprendre le fonctionnement de ce nouveau latch.

Deux configurations principales doivent être envisagées. En effet, il est possible de représenter ce bloc par deux schémas équivalents différents en fonction de la valeur de WL fig.4.12.

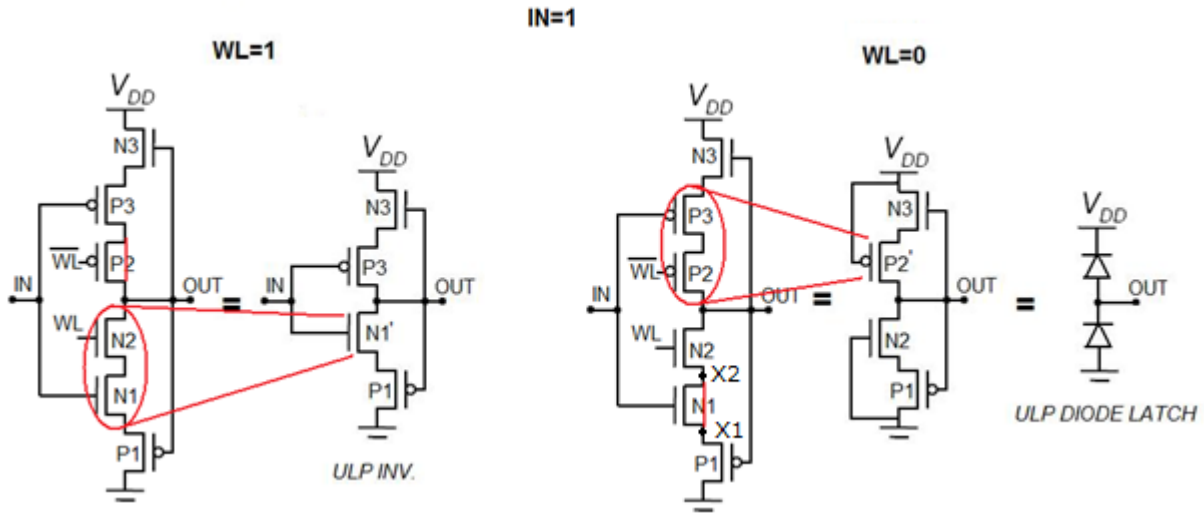


Fig. 4.12 : Schémas équivalents du latch ULP pour IN=1 : à gauche lorsque WL=1, à droite lorsque WL=0. Un raisonnement similaire peut être réalisée lorsque IN=0.

- Lorsque WL=1 et IN=1, le transistor P2 possède un large V_{SG} et donc une faible impédance équivalente par rapport à P3 et N3. Il est donc raisonnable de le remplacer par un court-circuit dans le schéma équivalent du bloc. En effet, ceux-ci forment un transistor ULP à l'état bloqué et possèdent respectivement une tension V_{sg} et V_{gs} proche de $V_{DD}/2$. Les transistors N2 et N1 possèdent tous deux un large V_{gs} et peuvent être remplacés par un seul NMOS équivalent N1'. Ces deux modifications apportées, nous retrouvons le schéma de l'inverseur ULP. La tension OUT s'équilibrera donc à l'inverse de la tension IN. Un raisonnement similaire peut-être effectué dans le cas où IN vaut 0.

- Lorsque WL=0 et IN=1, N2 doit posséder un V_{gs} négatif. En effet, les tensions des nœuds X1 et X2 doivent être comprises entre 0 et VDD, en fonction de la valeur de la tension du nœud OUT. N1 observe quant à lui, de par la valeur du noeud IN, un V_{gs} positif. Dès lors, le transistor N1 possède une impédance équivalente très faible devant celle de N2 et peut donc être modélisé par un court-circuit dans le schéma équivalent du bloc. Par ailleurs, P3 et P2 sont tous deux bloqués. Leur V_{sg} évoluera de la même manière en fonction de la tension du nœud de sortie OUT. Ils peuvent donc être remplacés par un seul transistor PMOS équivalent P2'. Un raisonnement similaire peut-être réalisé lorsque IN est au niveau bas. Le schéma du latch se simplifie donc à celui de deux transistors ULP montés en diodes.

Comme expliqué dans [8], l'assemblage de deux diodes ULP montées en séries permet d'obtenir un effet de mémorisation de la tension de sortie. Cela sera examiné plus en détail au point 4.6.2. Nous pouvons donc utiliser ce nouveau type de latch ULP pour réaliser une mémoire. Lorsque nous désirerons écrire une valeur, la configuration de l'inverseur ULP sera utilisée, ensuite cette valeur sera mémorisée par la configuration des deux diodes (appelée « latch diodes » par la suite). Les transistors N2 et P2 vont donc en quelque sorte faire office de transistors d'accès en écriture activés par la tension de leur grille WL.

4.6.2. Le latch diode

La figure 4.13 reprend le schéma du latch diode. La capacité C_{node} comprend l'ensemble des capacités parasites présentes au nœud de sortie. Pour comprendre comment ce montage peut mémoriser une valeur, il faut examiner plus en détail l'évolution des différents courants en présence.

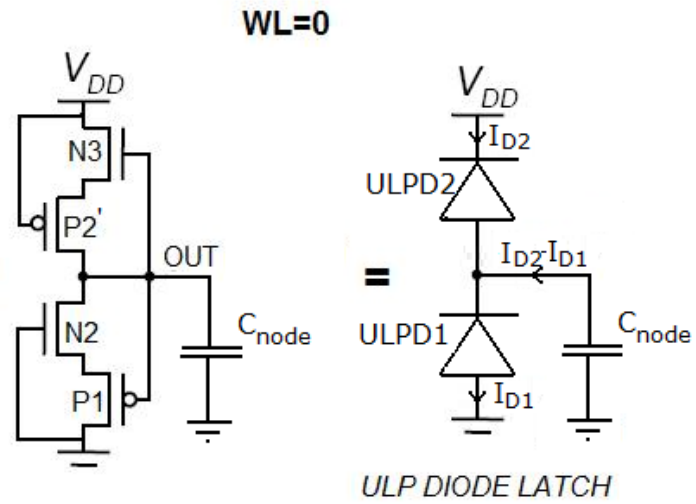


Fig 4.13 : représentation du latch diode.

Lorsque la tension de sortie est inférieure à une certaine valeur, le courant I_{D1} domine le courant I_{D2}. Dès lors, la capacité C_{node} est déchargée petit à petit. Par contre, si la tension OUT est supérieure à une autre valeur, le courant I_{D2} est plus important que le courant I_{D1} et la capacité C_{node} est cette fois-ci chargée à V_{DD}. Ainsi, si une tension quelconque est mémorisée, sa valeur sera régénérée en cas de perturbation d'amplitude limitée. Il existe donc trois états d'équilibres possible pour le montage, soit OUT vaut 0 ou V_{DD} et dans ce cas nous sommes dans un état stable, où OUT prend une valeur intermédiaire pour lequel I_{D1}=I_{D2}, ce qui correspond à un état instable.

La vitesse avec laquelle la tension OUT atteint un équilibre stable dépend de l'amplitude de la courbe I_{D2}-I_{D1}. La figure 4.14 permet d'observer la courbe caractéristique d'une diode ULP. Une comparaison est réalisée avec un NMOS de type « HS » monté en diode. Nous pouvons noter à première vue que pour une polarisation positive, le courant dans la diode ULP est proche de celui du NMOS monté en diode. Néanmoins, cette zone de fonctionnement ne nous intéresse pas vraiment, car la diode ULP est surtout utilisée dans la région de polarisation négative. Son intérêt est ici évident, car pour une tension de polarisation supérieure à – V_{DD}/2, son courant de fuite est réduit de 5 ordres de grandeur par rapport à celui du montage NMOS. Lorsque la tension de polarisation part de zéro, le courant de fuite de la diode ULP augmente d'abord rapidement suite à l'augmentation de la tension V_{DS}. Par la suite, sa tension V_{GS} devient de plus en plus négative ce qui permet de couper le courant de fuite qui diminue de plus en plus. Enfin, un minimum est atteint lorsque le courant de jonction devient plus important que le courant sous seuil du transistor. Ce courant augmente quant à lui avec la tension de polarisation.

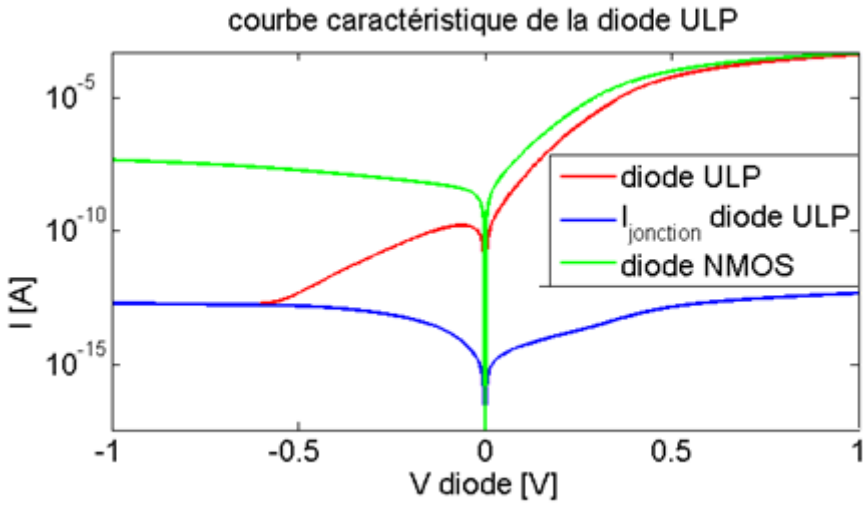


Fig. 4.14 : courbes caractéristiques de la diode ULP, comparées à celle d'un NMOS « HS » monté en diode (W_N=1µm, W_P=2.5µm, $L_{N,P}$=0.13µm, Vb_{NMOS}=-1V, $Vb_{NMOSULP}$=0V, $Vb_{PMOSULP}$=1V).

Il est intéressant de noter que lorsque la température augmente, le courant sous seuil des transistors augmente également ce qui accroit la valeur des courants de régénération de la diode. Nous obtenons alors une cellule mémoire dont la capacité de restauration augmente avec la température. La même remarque peut-être effectuée au sujet du latch ULP formé de deux inverseurs bouclés sur eux-mêmes que nous avons utilisés pour les cellules ULP à 10 et 12 transistors.

La figure 4.15 permet d'observer l'évolution des courants dans un latch diode. Lorsque OUT est proche de 0, le courant ID1 est plus grand que le courant ID2. La courbe rouge représentant le courant dans ULPD1 possède l'amplitude la plus importante. A l'inverse, lorsque la tension de sortie est proche de 1, la courbe bleue associée au courant dans la diode ULPD1 domine.

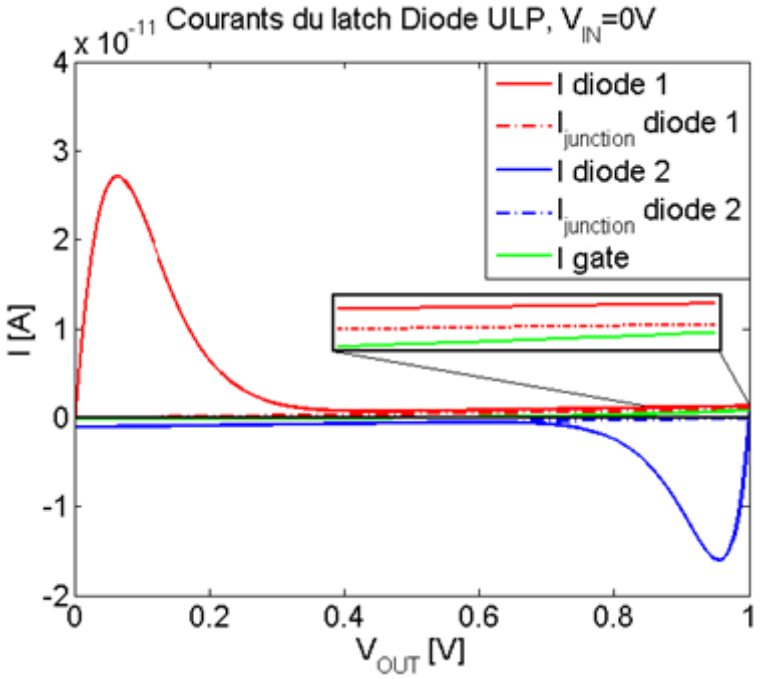


Fig. 4.15 : décomposition des courants de la diode ULP. W_N=0.15µm, W_P=0.45µm, $L_{N,P}$=0.13µm, V_{IN}=0 V_{BN}=0V, V_{BP}=1V.

4.6.3. Simulation de la cellule

La cellule a été simulée sous ELDO sous une tension d'alimentation de 1V (fig. 4.16). L'organisation générale de la mémoire reste la même que celles présentées dans les sections précédentes si ce n'est l'utilisation d'une seule bitline d'écriture suite aux modifications de l'architecture de la cellule.

Comme nous pouvions nous y attendre, le temps d'écriture est élevé (5µs) à cause du délai de l'inverseur ULP. En effet, l'utilisation d'une seule bitline nous oblige à attendre que l'inverseur ULP fasse basculer l'état de la cellule vers sa valeur d'équilibre. Par contre, le circuit de lecture est indépendant du latch ULP. Il est du même type que celui de la cellule 8 transistors classique ou de la cellule ULP 12 transistors. Le temps de lecture est donc semblable à celui d'une cellule 8 transistors et ne vaut que 0.50 ns. La consommation statique de la cellule ne vaut que 6.2 pW. Cette cellule ne possède qu'un seul chemin entre la tension d'alimentation et la masse, elle permet donc d'obtenir le courant de rétention le plus faible.

Il faut également remarquer que la cellule subit toute une série d'effets de type couplage capacitif comme décrit à la section 4.5.2. Ceux-ci ont lieux lorsque les tensions de wordline ou de bitline d'écriture changent de valeur. En effet, certains des transistors de la cellule ont leur grille directement raccordée à ces lignes. Toutefois, ces phénomènes ne possèdent pas une amplitude suffisante pour déstabiliser l'état de la cellule.

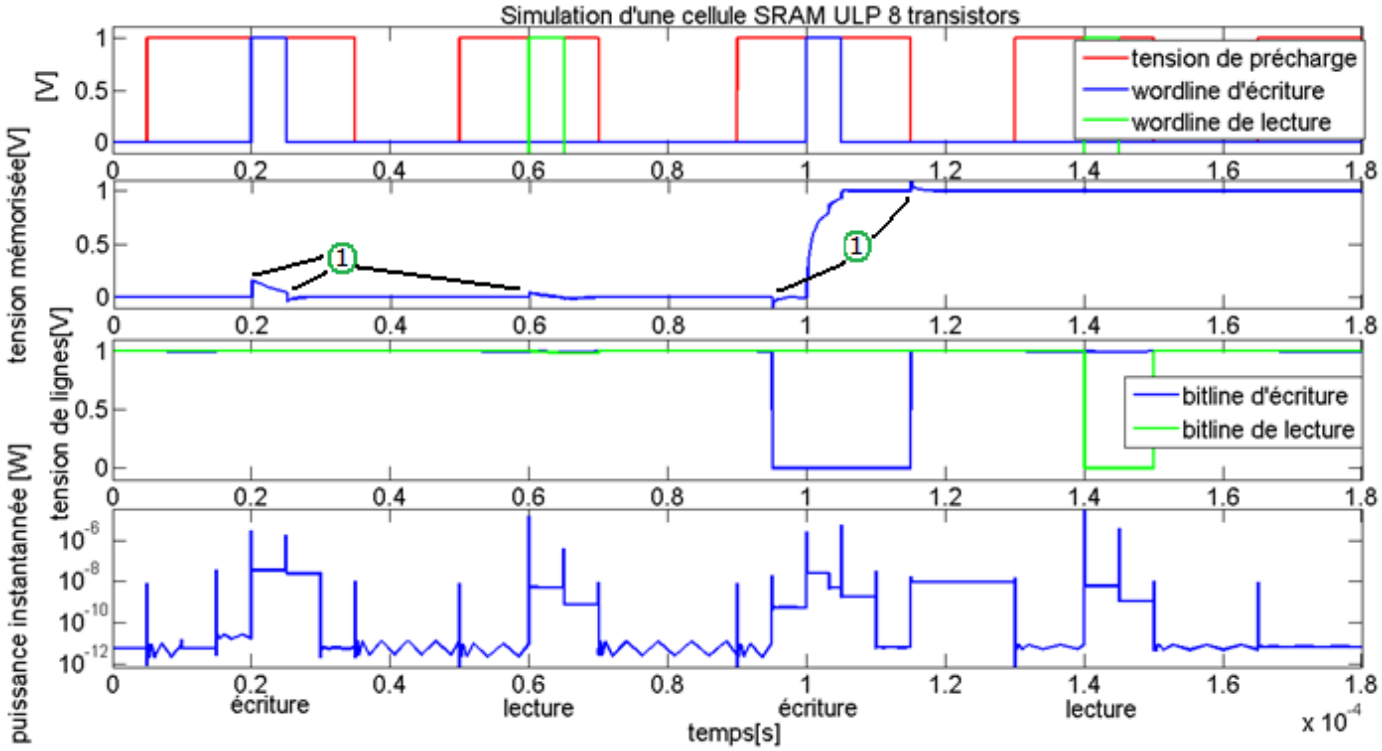


Fig 4.16: Simulation ELDO d'une cellule SRAM ULP 8. V_{DD}=1V. Deux opérations d'écriture et de lecture sont alternativement réalisées.

4.7. Synthèse

Le tableau 4.4 reprend l'ensemble des performances de chaque cellule. Rappelons que ces résultats ont été extraits à l'aide de simulation ELDO réalisée à l'aide de modèle BSIM3 industriels sous une température de 25°C. Le dimensionnement des transistors des différentes cellules correspond aux valeurs présentées dans les paragraphes précédents. Le temps de lecture indiqué ici correspond au temps nécessaire au développement d'une différence de tension de 100 mV entre les deux bitlines ou le cas échéant à une chute de tension de 100 mV sur l'unique bitline de lecture. Enfin, les bitlines ont toujours été préchargées à l'aide de PMOS avant une opération de lecture, ces mêmes PMOS étant coupés lors de la lecture proprement dite, de manière à isoler les bitlines de l'alimentation.

Si la cellule 8T classique ne semble ici pas posséder d'avantages par rapport à la cellule 6T, ceux-ci seront mis en valeur dans les sections suivantes.

	$T_{lecture}$	$T_{écriture}$	$P_{statique}$	$I_{lecture}$	$I_{lecture}/I_{stat}$	$T_{accès\ moyen}$
6T	0.39 ns	25 ps	0.71 nW	38.5 µA	$54.4 \cdot 10^3$	215 ps
$6T_{VDD=0.5}$	4.0 ns	205 ps	0.38 nW	3.75 µA	$9.8 \cdot 10^3$	2.10 ns
8T	0.50 ns	37 ps	0.95 nW	30.0 µA	$31.6 \cdot 10^3$	250 ps
ULP 10T	7.85 µs	4.76 ns	11.7 pW	1.91 nA	163.2	3.93 µs
ULP 12T	0.75 ns	1.37 ns	13 pW	20.0 µA	$1.54 \cdot 10^6$	1.23 ns
ULP 8T	0.50 ns	7.5 µs	6.2pW	30.0 µA	$4.84 \cdot 10^6$	3.75 µs

Tableau 4.4 : comparaison des performances des différentes cellules étudiées. V_{DD}=1V sauf si le contraire est mentionné. T= 25 °C.

Outre le rapport $I_{lecture}/I_{stat}$ dont nous avons déjà parlé, le temps d'accès moyen est également donné. Il correspond à la moyenne arithmétique des temps d'accès en lecture et en écriture. Il complète donc l'information sur les performances dynamiques des différentes cellules.

L'utilisation d'inverseurs ULP pour la réalisation d'une cellule SRAM permet effectivement de réduire la consommation du circuit de deux ordres de grandeur. L'utilisation d'un buffer de lecture garantit un temps de lecture élevé. Il assure alors aux cellules ULP un rapport $I_{lecture}/ I_{stat}$ très élevé. La cellule ULP 12 transistors permet également de garder un temps d'accès faible en écriture et en lecture au sacrifice d'une plus grande surface consommée. Remarquons également que la réduction de la tension d'alimentation de la cellule 6T ne permet d'améliorer aucun des deux facteurs de mérites $I_{lecture}/I_{stat}$ et $T_{accès\ moyen}$, contrairement aux mémoires ULP. Cependant par rapport à ces deux critères, la cellule ULP 10 transistors est décevante. Evidemment, il existe d'autres applications où la surface consommée a un plus grand impact, notamment pour des raisons de prix de fabrication, et où les performances dynamiques ne sont pas de première importance. Il faut alors pouvoir reconsidérer l'intérêt de la cellule ULP 10 transistors, et le bien-fondé des autres cellules. Il pourrait également à ce moment être intéressant de sacrifier la rapidité de la cellule pour en réduire la tension d'alimentation à $V_{DD}/2$. Notons enfin que la cellule ULP 8 transistors présentée ici ne possède pas de performances équivalentes à la cellule ULP 7 transistors analysée dans [8]. Par ailleurs, comme nous le verrons dans la suite de cette section, sa marge de bruit n'est pas non plus améliorée.

4.8. Les marges de bruit

En fonctionnement, chaque cellule de la mémoire SRAM sera soumise à quelques perturbations. Leurs sources sont variées, il peut s'agir de radiations liées à l'exposition du circuit aux rayons cosmiques, à une particule α [8]... Il faut également garantir que le circuit puisse fonctionner malgré les variations de procédés de fabrication. Par exemple, un dopage non uniforme du substrat pourrait modifier la tension de seuil du transistor, une grille plus longue, diminuer son courant.

Il est donc nécessaire d'établir à quel point la cellule développée sera robuste face à ce « bruit ». Plusieurs critères permettent ainsi de la caractériser en mode rétention, lors d'une opération de lecture ou encore lors d'une écriture.

4.8.1. La marge de bruit statique en rétention et écriture

La marge de bruit statique (SNM) est définie par la tension de bruit minimale qui doit être appliquée aux deux nœuds de mémorisation pour basculer l'état de la cellule [33] [34]. Elle permet d'obtenir une image de la stabilité de la cellule face aux perturbations statiques, en mode de rétention ou lors d'un accès en lecture. Sa définition est illustrée à la figure 4.17. D'après [34], sa valeur doit être suffisante pour garantir le bon fonctionnement de la cellule malgré des perturbations dynamiques supplémentaires.

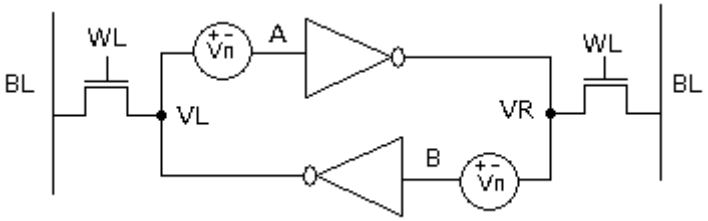


Fig. 4.17 : La marge de bruit est définie par la tension minimale de V_N nécessaire pour faire basculer l'état de la cellule.

La SNM peut être déduite de la courbe caractéristique des inverseurs de la cellule mémoire. Elle correspond à la longueur du plus grand carré qu'il est possible de tracer entre ces deux courbes [33] [34] formant ainsi un diagramme appelé « diagramme papillon ». Lors d'une lecture, les deux transistors d'accès sont activés et les deux bitlines sont préchargées à V_{DD}. Dans le cas du mode de rétention, nous considérons le cas où la tension des bitlines est opposée à celle des nœuds de mémorisation. Ceci correspond au cas le plus défavorable où une valeur opposée à celle mémorisée est écrite sur une autre cellule de la même colonne. La figure 4.18 illustre ces deux configurations pour une cellule à 6 transistors. Notons que l'effet d'une source du bruit statique V_N correspond à un décalage des courbes caractéristiques. Lorsque V_N vaut la SNM de la cellule, les deux courbes ne se croisent plus que ponctuellement [33] [34].

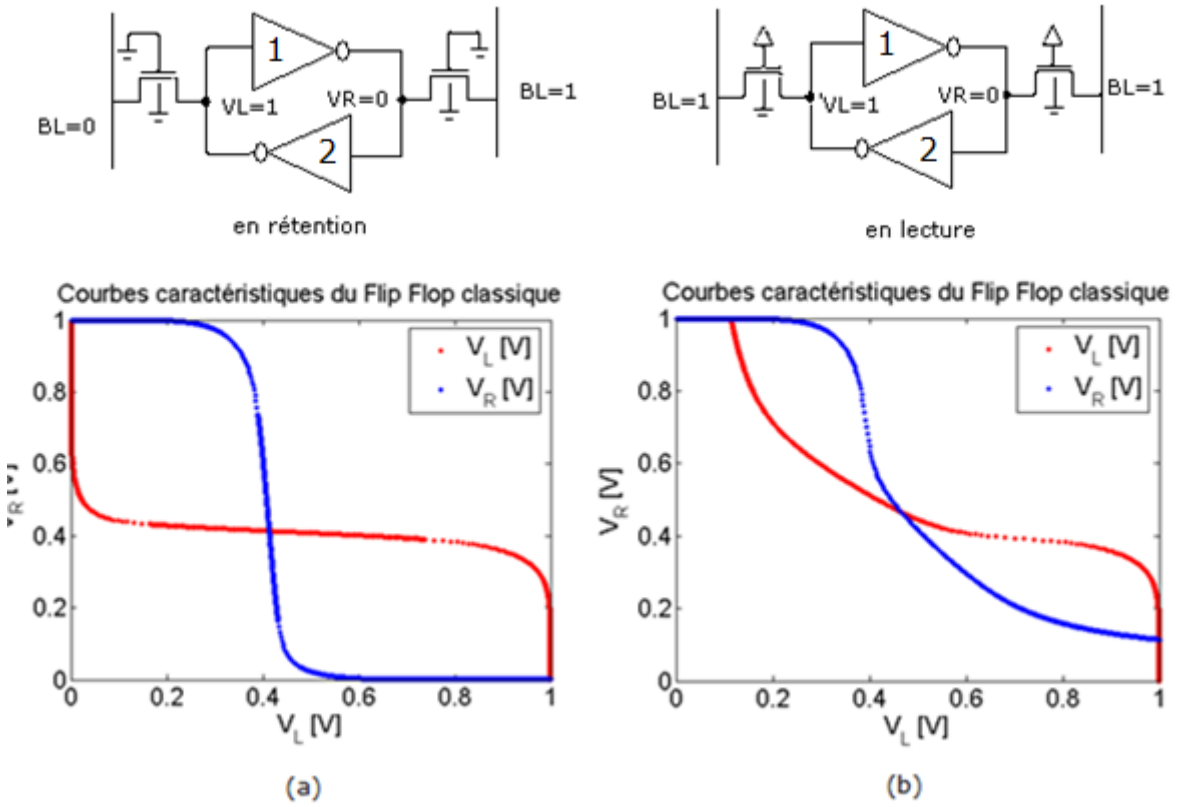


Fig. 4.18 : courbes permettant de calculer la SNM dans le cas du flip flop classique, (a)en rétention et (b) lors d'une lecture.

Dans le cas d'une cellule SRAM classique, lors d'une écriture, le nœud mémorisant le « 0 » voit sa tension se dégrader. Ce phénomène est lié à la présence d'un diviseur résistif entre le transistor d'accès et le NMOS de l'inverseur. Cela rend la cellule plus sensible lors d'un accès en lecture.

Pour la cellule ULP, il faut prendre en compte l'hystérèse présente dans la courbe caractéristique de l'inverseur. Si une source de bruit quelconque tend à perturber l'état de la cellule, elle devra augmenter la tension au nœud de mémorisation à l'état bas et réduire celle du nœud à l'état haut. C'est pourquoi une des courbes du diagramme papillon correspond à la courbe caractéristique de l'inverseur ULP pour une pente montante de son entrée et l'autre pour une pente descendante de son entrée. On peut observer à la figure 4.19 qu'il est maintenant possible de tracer deux carrés de tailles différentes dont un seul permet de déduire la SNM. Comme nous l'avons vu au paragraphe précédent, l'application d'une tension de bruit V_N correspond à un décalage des deux courbes. Les sources de bruit vont donc tendre à faciliter le basculement des inverseurs. Si on se ramène à la figure 4.17, on comprend que V_N va avoir pour effet de réduire la tension d'entrée de l'inverseur 1 et que donc la tension V_L nécessaire à faire basculer cet inverseur est plus proche de V_{DD}. V_N tend donc soit à monter la courbe caractéristique de l'inverseur pour lequel nous avons appliqué une pente descendante, ou à décaler l'autre courbe sur la gauche. L'état de la cellule sera sur le point de basculer lorsqu'elles ne se croiseront plus qu'en un point unique. La marge de bruit correspond donc au plus grand carré qu'il est possible d'inscrire dans la plus grande des deux boucles.

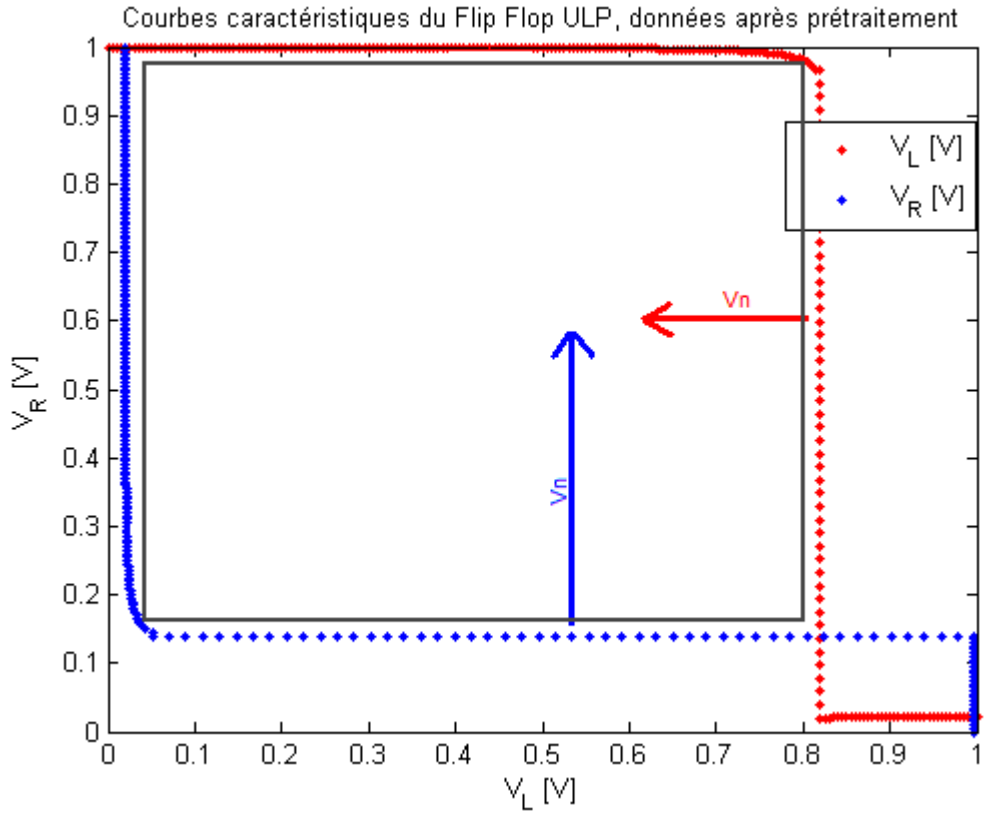


Fig. 4.19 : calcul de la SNM de la cellule ULP. V_L mémorise l'état haut, V_R l'état bas (cf fig. 13). La SNM est définie par la longueur du plus grand carré qu'il est possible d'inscrire dans le rectangle de droite.

Des simulations sous ELDO ont permis d'obtenir les SNM pour les différentes cellules ULP, ainsi que pour une cellule 6 transistors classique ou 8 transistors. Dans le cas du latch ULP, l'hystérèse présente dans la caractéristique des inverseurs ULP entraine un accroissement de la robustesse de la cellule face au bruit, et permet même d'obtenir des marges de bruit supérieures à $V_{DD}/2$. Cependant, une architecture « classique » de la mémoire (cellule ULP à 10 transistors) rend la cellule très sensible lors d'un accès en lecture. En effet, les transistors ULP possèdent une impédance équivalente beaucoup plus élevée que les transistors d'accès. Ainsi, lors d'un accès en lecture la tension de nœud mémorisant le « 0 » est bien plus dégradée que dans le cas d'une cellule conventionnelle. Un dimensionnement judicieux des différents transistors permet d'améliorer la SNM de la cellule. Une possibilité peut être d'augmenter la longueur des transistors d'accès tout en augmentant la largeur des PMOS P1 des inverseurs ULP (fig. 3.9), nous avons vu à la section 3.4.3 que ce sont ces transistors qui limitent le courant du transistor ULP de type N. Il est également possible de contourner le problème grâce à l'ajout d'un buffer de lecture. C'est ce qui est réalisé dans la cadre de la cellule ULP à 12 transistors vue à la section 4.5. Pour cette cellule, la SNM reste donc identique en rétention et en lecture.

Le tableau 4.5 permet de comparer la SNM des cellules ULP à celles des cellules à 6 transistors et à 8 transistors. Notons que pour la cellule à 6 transistors, des résultats similaires peuvent être trouvés dans la littérature [8] [28] [35]. Une réduction de la tension d'alimentation de la cellule classique détériore fortement la SNM en lecture. En dessous de 0.75V, la dégradation de la tension de nœud mémoire à l'état bas devient trop importante. Une constatation semblable est réalisée par [7]. Ceci peut s'expliquer par exemple par l'accroissement de l'impact de la dissymétrie des tensions de seuil du NMOS et du PMOS en régime sous le seuil.

La marge de bruit de la cellule ULP à 8T doit être déterminée par une autre procédure. Bien que celle-ci ne soit expliquée qu'à la section 4.8.3, sa valeur est indiquée ici pour la comparaison.

Type	V_{DD} [V]	W (P1) [µm]	L (N3) [µm]	SNM rétention [mV]	SNM lecture [mV]
6T	1	-*	-*	353	170
6T	0.75	-*	-*	293	53
8T	1	-**	-**	355	355
10T ULP	1	0.75	0.65	786	45
12T ULP	1	0.15	0.13	715	715
12T ULP	0.75	0.15	0.13	501	501
12T ULP	0.5	0.15	0.13	285	285
8T ULP	1	0.15	0.13	280	280

*Tableau 4.5 : comparaison des SNM de la cellule 6T SRAM et des cellules ULP. * pour la cellule 6 transistors,WNdriver/WNaccess=2, les autres transistors possèdent une taille minimale (W=0.15µm, L=0.13 µm). ** pour la cellule 8 transistors classique, tous les transistors sont de taille minimale.*

Dans le cas des cellules ULP à 10 et 12 transistors, les simulations montrent que l'hystérèse présente dans la caractéristique de l'inverseur permet d'améliorer la SNM de plus de 100% lorsque l'on ne réalise pas une opération de lecture sur la cellule. Malheureusement, lors d'une telle opération, une forte dégradation de la marge de bruit statique est observée pour la cellule à 10 transistors. Un redimensionnement de la cellule ne permet d'élever la marge de bruit qu'à environ 25% de celle de la cellule à 6T. Notons que les cellules ULP ne sont que peu affectées par une réduction de V_{DD}, la marge de bruit restant supérieure $V_{DD}/2$.

4.8.2. La marge de bruit en écriture

Ce critère vise à garantir qu'une opération d'écriture permettra de modifier l'état de la cellule. Il s'agit ici essentiellement de vérifier que les inverseurs d'accès seront suffisamment forts pour vaincre la boucle de contre-réaction du Flip Flop. Pour faciliter l'écriture, dans une cellule classique, les PMOS des inverseurs sont de taille minimale de manière à être plus faibles que les transistors d'accès.

Une mesure de la facilité avec laquelle il est possible d'écrire sur une cellule est déduite de la notion de marge de bruit. Il faut que la cellule ne soit pas trop robuste face à écriture, de manière à ce que la valeur mémorisée soit modifiée. Si l'on se réfère à la notion de marge de bruit statique étudiée à la section 3.1, il faut cette fois que la SNM soit négative [7] [23] [27]. La figure 4.20 illustre la manière avec laquelle les courbes caractéristiques de la cellule sont obtenues. Les bitlines possèdent une valeur opposée aux niveaux de tensions présents aux nœuds de mémorisation. La wordline est activée de manière à réaliser l'opération d'écriture. Cependant, des sources de bruit V_N perturbent les entrées des inverseurs. Ces sources permettent de simuler la force de la boucle de contre-réaction. En effet, lorsque V_N est négatif, les entrées des inverseurs seront dans un état opposé à celui que les bitlines tentent d'imposer. Les sources V_N s'opposent à la tentative d'écriture comme la boucle de contre-réaction du latch le ferait. Au fur et à mesure que V_N augmente et devient positif, les inverseurs voient leur entrée évoluer dans le sens des valeurs des bitlines. V_N facilite de plus en plus l'écriture en provoquant le basculement des inverseurs.

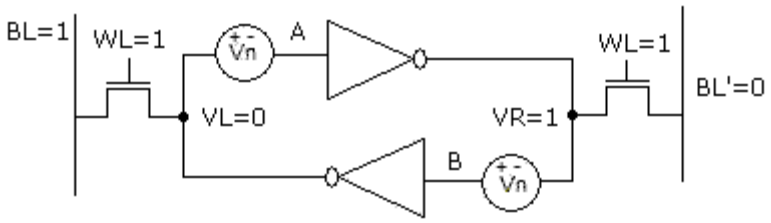


Fig. 4.20 : Schéma du circuit permettant de déduire la marge de bruit en écriture.

Pour que l'opération d'écriture soit possible, il faut que l'effet des tensions de bitlines soit plus fort que la boucle de contre-réaction. Ainsi, le basculement de la cellule doit avoir lieu pour des valeurs de V_N négatives. La valeur de V_N pour laquelle les tensions mémorisées V_L et V_R se croisent (fig. 4.20) définit la marge de bruit en écriture. Notons que cette valeur porte le nom de marge de bruit par analogie avec le critère de marge de bruit statique (section 4.8.1). Cependant, la marge de bruit en écriture n'est pas du tout déterminée par la même méthode. Lors du calcul de la marge de bruit statique en rétention ou écriture, aucune source de bruit V_N n'a été appliquée au circuit. La figure 4.21 donne les résultats d'une simulation de la marge de bruit en écriture pour une cellule 6 transistors.

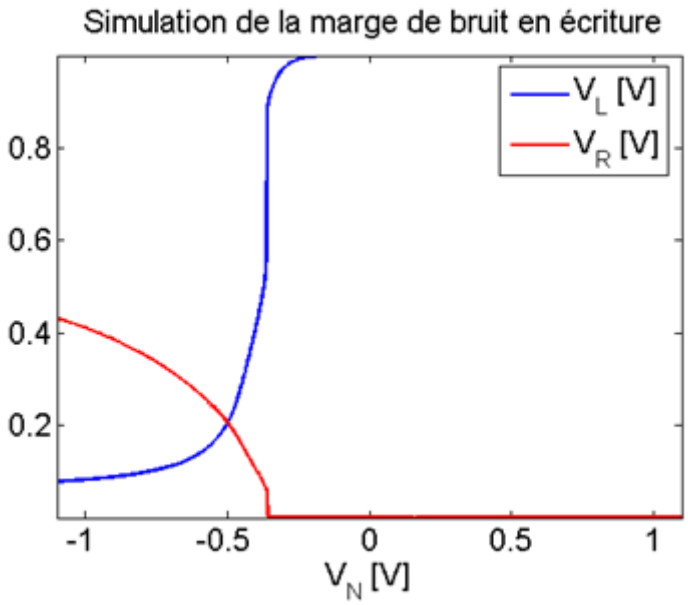


Fig. 4.21 : calcul de la SNM en écriture de la cellule 6 transistors. La SNM est définie par la valeur de V_N pour laquelle les deux courbes se croisent. Une valeur négative indique que l'opération d'écriture est réalisée avec succès.

Dans le cas de la cellule ULP, aucun croisement des courbes ne peut être observé, comme illustré à la figure 4.22a. En effet, l'impédance équivalente d'un transistor ULP reste très élevée même lorsqu'il est actif. Dès lors, seulement une très faible variation de tension peut-être observée sur le transistor d'accès et la tension du nœud de mémorisation reste très proche de celle de la bitline correspondante. Nous pouvons remarquer toutefois le basculement de l'inverseur de gauche au point 1 et de celui de droite au point 2. Le basculement de la courbe bleue est nettement plus visible. Cela s'explique par le fait que le NMOS d'accès ne peut pas entièrement transmettre un 1. Par contre une fois que l'inverseur bascule, son PMOS peut tirer la tension mémorisée jusque V_{DD}. Toutefois, il est difficile de définir de manière précise la marge de bruit en écriture. Nous pouvons considérer que l'écriture sera réussie une fois que les deux inverseurs auront basculé. La solution consiste alors à

penser en termes de courants. Considérons un inverseur dont l'entrée est montante. Nous pouvons considérer que le basculement aura lieu une fois que le courant s'écoule majoritairement par le NMOS de l'inverseur plutôt que par le PMOS (point 4 à la figure 4.22b). Ceci nous amène donc à définir un nouveau critère pour déterminer la marge de bruit en écriture. L'opération sera considérée comme achevée une fois que les deux inverseurs auront basculé.

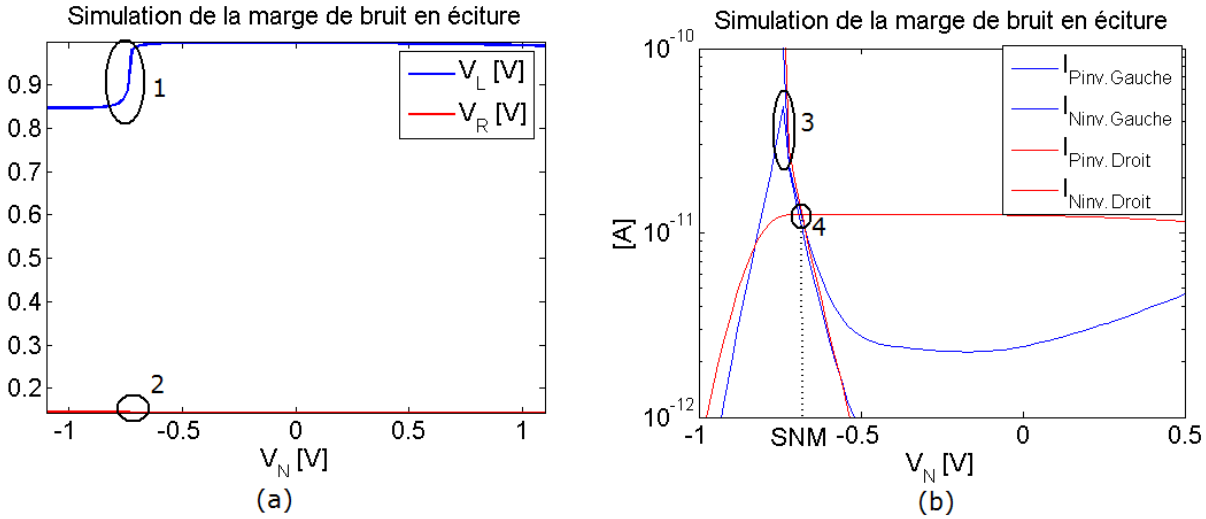


Fig. 4.22 : (a) calcul de la SNM en écriture de pour la cellule ULP. Les courbes ne se croisent pas. (b) La SNM est définie par la valeur de V_N pour laquelle les courbes de courant d'un même inverseur se croisent. Une valeur négative indique que l'opération d'écriture est bien réalisée.

Le pic de courant observé à la figure 4.22b (zone 3) peut sembler étrange. Il correspond tout simplement au courant circulant dans le PMOS de l'inverseur alors que sa sortie n'est pas encore ramenée à V_{DD}. Une fois que l'état de l'inverseur est totalement basculé, la sortie de l'inverseur s'approche de la tension d'alimentation ce qui réduit le V_{DS} du PMOS. Son courant chute alors comme nous pouvons l'observer sur le graphe.

Le tableau 4.6 reprend les valeurs de la marge de bruit en écriture pour les différentes cellules étudiées. Le cas de la cellule ULP à 8 transistors sera traité dans la section suivante. Tous les transistors sont de taille minimale (W=0.15 µm, L=0.13µm), excepté les NMOS des inverseurs de la cellule 6 transistors (W=0.30 µm), la cellule ULP 10 transistors se trouve dans la configuration n°4 décrite à la section 4.4. Nous pouvons remarquer que la cellule ULP possède une marge de bruit en écriture meilleure que les cellules classiques. Cela est lié au fait que les transistors ULP sont très faibles par rapport aux transistors d'accès et donc que ceux-ci n'éprouvent que très peu de difficultés à vaincre la boucle de contre-réaction du latch.

Type	W (P1) [µm]	L (N3) [µm]	SNM écriture [mV]
6T	-*	-*	-499
8T	-**	-**	-611
10T ULP	0.75	0.65	-656
12T ULP	0.15	0.13	-660

*Tableau 4.6 : comparaison des SNM en écriture de la cellule 6T et des cellules ULP. * pour la cellule 6 transistors,WNdriver/WNaccess=2, les autres transistors possèdent une taille minimale (W=0.15µm, L=0.13 µm). ** pour la cellule 8 transistors classique, tous les transistors sont de taille minimale.*

4.8.3. La marge de bruit de la cellule ULP 8T

La cellule ULP à 8 transistors, dont le schéma est rappelé à la figure 4.23a, n'est pas réalisée à l'aide de deux inverseurs bouclés sur eux-mêmes. Ainsi, il n'est pas possible d'étudier la stabilité d'un tel montage face aux perturbations par les méthodes décrites précédemment.

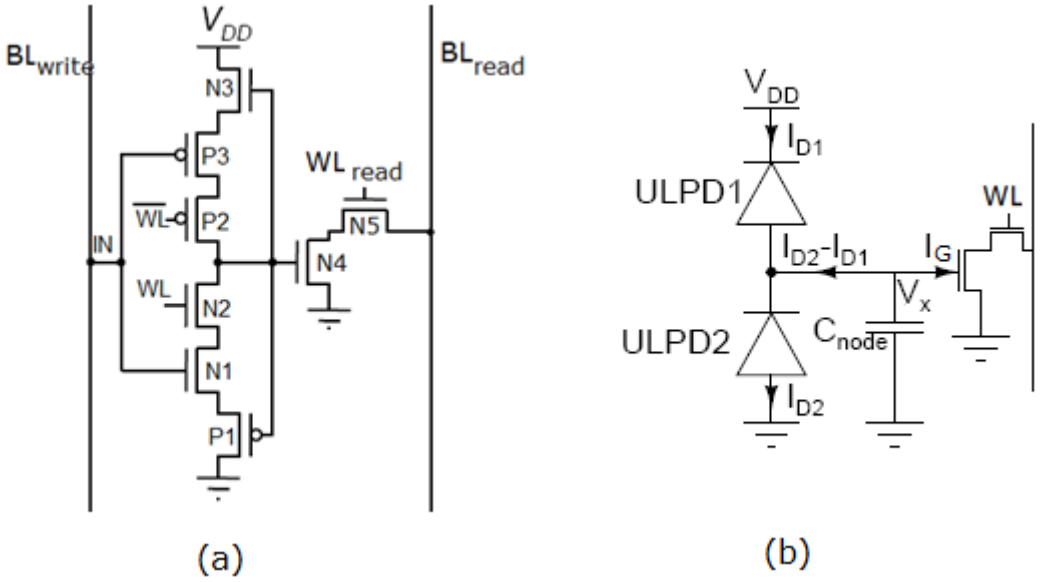


Fig. 4.23 : (a) Schéma d'une cellule ULP à 8 transistors. (b) Schéma équivalent du latch en rétention.

Dans ce cas-ci, la marge de bruit peut-être déduite de la courbe caractéristique de la diode ULP. Une décomposition de l'évolution des courants présents dans le latch est donnée à la figure 4.15. Lorsque la cellule mémorise un zéro, elle mémorisera correctement cette valeur tant que la condition suivante est vérifiée : $I_{D2} > I_{D1}-I_G$. Cependant lorsque la tension mémorisée est proche de zéro, le courant de grille peut être négligé devant I_{D1} et I_{D2}. La condition se ramène donc à $I_{D1}>I_{D2}$. Lorsque la cellule mémorise un « 1 », la condition devient $I_{D1}>I_{D2}+I_G$. Dans ce cas, le courant de grille n'est plus négligeable. En effet, la tension V_{GS} du transistor d'accès équivaut alors à la tension d'alimentation de la cellule. Ce cas est donc plus sensible.

Cette analyse doit tenir compte d'autres effets. Premièrement, la tension V_{IN} présente sur la bitline a une influence sur la courbe caractéristique de la diode. En effet, nous avons vu qu'en fonction de la valeur de V_{IN}, les deux transistors P2 et N2 ne se comportent pas de la même manière. V_{IN} introduit donc une dissymétrie dans le circuit qui se reporte sur la courbe caractéristique (fig. 4.24). En outre, les courants de jonctions peuvent entrainer l'apparition de zéro supplémentaires sur la courbe caractéristique du latch. En effet, lorsque V_{OUT} s'approche de $V_{DD}/2$, la diode « passante » voit son courant de régénération diminuer, et l'autre diode est toujours bloquée. Le courant de jonction de cette diode devient alors la principale composante du courant de fuite du latch. Ce courant peut même finir par égaliser, voire dépasser celui de la diode active lorsque celle-ci est sur le point de se couper. Ce phénomène restreint significativement les marges du bruit du montage. Il est toutefois possible de s'en affranchir grâce à l'utilisation de la technologie SOI comme cela est fait dans [8]. Dans ce cas, seul le courant de grille reste à redouter. Notons que lors du passage à des technologies supérieures, ce courant de grille prendra de l'importance suite à la diminution de l'épaisseur de l'oxyde de grille. L'utilisation de matériaux « High-K » pourra apporter une solution [36]. Notons enfin qu'une élévation de la température de fonctionnement résout le problème du courant de jonction. En effet, le courant sous seuil augmente alors suffisamment pour que l'influence du courant de grille ou de jonction soit négligeable. La marge de bruit tend alors vers $V_{DD}/2$.

En fin de compte, quatre possibilités doivent être envisagées pour le calcul de la marge de bruit, en fonction de la valeur de la bitline d'écriture et de la valeur mémorisée. Le pire cas correspond à une marge de bruit statique de 280mV. Cette valeur est nettement inférieure à celle obtenue pour les cellules ULP à 10 ou 12 transistors (voir tableau 4.5).

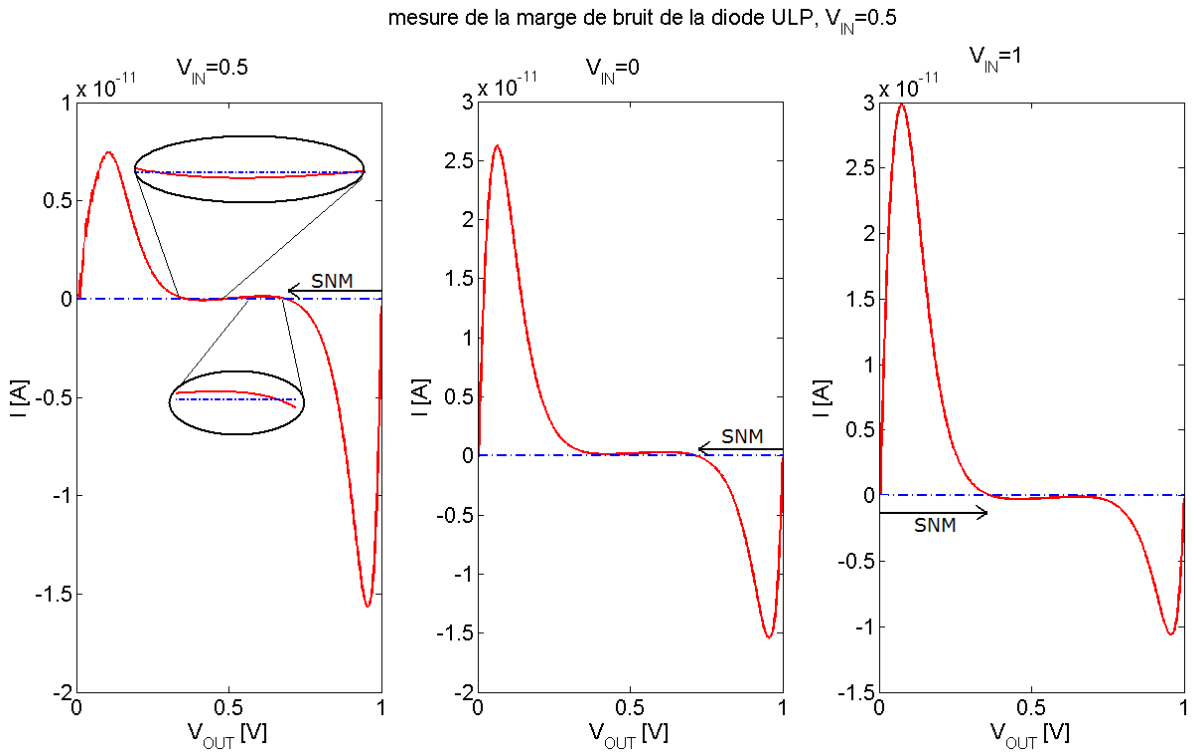


Fig. 4.24 : courbe caractéristique de latch diode. W_N=0.15µm, W_P=0.45µm, $L_{N,P}$=0.13µm, VIN=0, V_{BN}=0V, V_{BP}=1V.

4.9. Conclusion

Différentes SRAMs réalisées à l'aide de transistors ULP ont été étudiées. Leurs performances ont été comparées à celles de cellules 6 et 8 transistors. Sur base de facteurs de mérites choisis préalablement, la cellule ULP 12 transistors se distingue en offrant un grand rapport $I_{lecture}/I_{stat}$ et une vitesse d'accès élevée en échange d'un sacrifice en termes de surface occupée. Cette cellule présente des pertes statiques 50 fois plus petites que celle d'une cellule 6 transistors conventionnelle.

Par la suite, la robustesse des cellules ULP face aux perturbations statiques et la stabilité d'écriture ont été étudiées. Certaines méthodes d'évaluation de ces paramètres ont dû être adaptées à la technologie ULP. A nouveau la cellule ULP 12 transistors se démarque des autres cellules ULP. Cette cellule permet d'obtenir une SNM supérieure à $V_{DD}/2$, ce qu'il n'est pas possible d'atteindre avec une cellule réalisée à partir d'inverseur CMOS classique. Sa marge de bruit en écriture est également remarquable.

Dans la suite de ce travail, nous allons approfondir l'étude de cette cellule. En particulier sa tolérance aux variations de fabrication, son niveau de polarisation en tension optimal et son comportement sous des températures plus élevées

5. Sensibilité aux variations

5.1. Introduction

Les résultats obtenus à la quatrième section ne tiennent pas compte des dispersions liées aux fluctuations du procédé de fabrication. Celles-ci se traduisent essentiellement par une variation de la tension de seuil des transistors et de la longueur de grille [6]. De plus, l'environnement sous lequel la mémoire sera amenée à fonctionner peut également avoir une influence sur les performances. Une étude de la robustesse de la cellule ULP à 12 transistors par rapport à ces fluctuations est réalisée dans cette section. Il faut à ce titre considérer un défaut commun à l'ensemble de la surface de la tranche de silicium (« global corner »), mais aussi une variation d'un transistor à l'autre (« local variation ») [37]. Ces analyses seront complétées par des simulations réalisées sous diverses températures et le cas d'une réduction de la tension d'alimentation sera également envisagé.

5.2. Variation globale

Nous allons étudier ici la sensibilité de la cellule face à une perturbation présente sur l'ensemble de la surface de silicium. Cette perturbation se répercute directement sur les performances des transistors en présence. Il faut à ce titre considérer quatre cas extrêmes appelés « global corners » :

- Une augmentation de la tension de seuil des NMOS et de celle des PMOS (SS),
- Une augmentation de la tension de seuil des NMOS et une diminution de celle des PMOS (SF),
- Une diminution de la tension de seuil des NMOS et une augmentation de celle des PMOS (FS),
- Une diminution de la tension de seuil des NMOS et de celle des PMOS (FF).

Les résultats seront comparés au cas de figure standard (TT) et à l'effet d'une réduction de la tension d'alimentation de 100 mV, pouvant résulter par exemple d'une baisse de tension de la batterie du dispositif.

La figure 5.1 permet d'observer l'évolution des marges de bruit en écriture et en lecture. Il faut pouvoir garantir que celles-ci restent suffisantes dans tous les cas envisagés. Rappelons que la marge de bruit en rétention est ici équivalente à celle en lecture étant donné que la cellule est isolée du circuit de lecture à travers la grille d'un transistor.

Le cas de figure le plus contraignant en ce qui concerne l'écriture est celui d'une réduction de la tension d'alimentation. Néanmoins, la marge de bruit reste largement négative, même sous des températures élevées. Elle s'élève à environ 400 mV pour une température de 75 °C et à environ 300 mV pour une température de 125 °C, ce qui garantit clairement le succès de l'opération. En dehors d'une réduction de l'alimentation, la combinaison de NMOS et PMOS plus rapides est la plus contraignante. Cela se comprend par le fait que le mécanisme de contre réaction du latch, que doivent vaincre les transistors d'accès, est alors renforcé. Remarquons que la cellule 8 transistors classiques possède une marge de bruit en écriture beaucoup plus stable avec la température. Cela s'explique par le fait que dans la cellule ULP, les transistors composants le latch ULP fonctionnent en régime sous le seuil. Dans cette zone de fonctionnement, les courants varient exponentiellement avec la température[7]. Dès lors, les transistors d'accès, fonctionnant en régime normal et donc dont le courant diminue légèrement avec la température, ont de plus en plus de difficultés à vaincre le mécanisme de contre-

[7] Ce phénomène est observable à l'annexe 4.

réaction du latch lorsque la température augmente. Dans le cas de la cellule 8 transistors classique, les MOS des inverseurs fonctionnent en régime normal, ce qui rend la cellule moins sensible à une augmentation de la température.

Lors d'une lecture, la combinaison de NMOS rapides et de PMOS plus lents se révèle être la plus critique. Cette configuration accentue en fait le déséquilibre présent dans la courbe caractéristique de l'inverseur. En effet, pour préserver la surface de silicium, les PMOS constitutifs du latch ont été choisis de taille minimale. Il existe donc un déséquilibre entre le transistor ULP relié à l'alimentation, dont le courant I_{ON} est limité par un NMOS et celui relié à la tension de masse où I_{ON} est limité par un PMOS. Cependant, la marge de bruit reste supérieure à $V_{DD}/2$ pour la gamme de température étudiée et est deux fois plus importante que celle de la cellule 8 transistors classique.

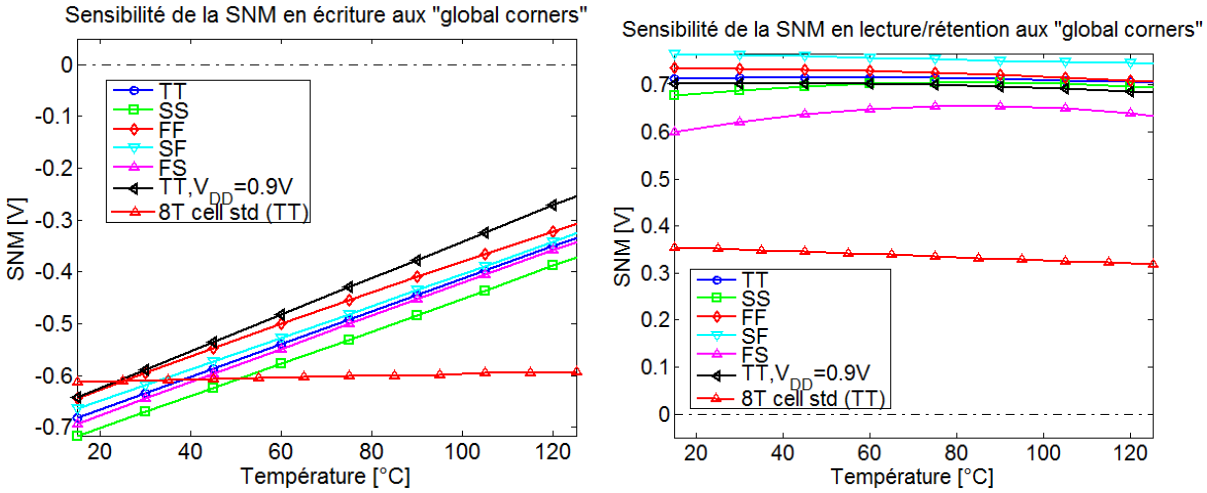


Fig. 5.1 : Marges de bruit de la cellule ULP 12 transistors sous les différents corners. A gauche en écriture, à droite en lecture. Les valeurs sont normalisées à V_{DD}=1V. Une comparaison est réalisée avec la cellule 8 transistors classique.

L'impact de ces variations sur les performances dynamiques de la cellule a également été étudié. La figure 5.2 représente l'évolution du courant développé par les deux transistors de lecture de la cellule sous différentes températures. A nouveau, une réduction de la tension d'alimentation a été envisagée. Il ressort que hormis la réduction de l'alimentation de la cellule, des transistors NMOS plus lents (corners SS et SF) sont les plus pénalisants pour le temps de lecture, indépendamment de la situation des PMOS. Ceci est lié au fait que le circuit de lecture ne comporte que des NMOS. Ainsi, une réduction de leurs performances dynamiques a pour premier effet de réduire le courant développé par ces deux transistors lors d'une opération de lecture. Par ailleurs, il faut noter que le courant de lecture de la cellule ULP augmente avec la température. Pour une cellule 8 transistors standard, l'effet inverse est observé. En effet, les MOSFETs subissent une dégradation de leur courant I_{ON} lorsque la température augmente[8]. Par opposition, le courant sous seuil augmente quant à lui avec la température. Ceci entraine un plus grand courant ON des inverseurs ULP à température élevée (fonctionnement sous-seuil), qui restaurent donc plus rapidement le 1 transmis par le transistor d'accès et améliorent donc la tension de grille du buffer de lecture.

[8] Voir annexe 4.

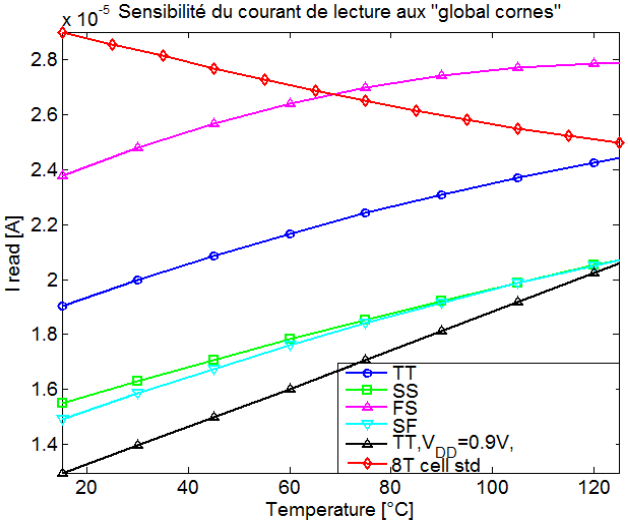


Fig. 5.2 : Evolution du courant de lecture de la cellule ULP 12 transistors en fonction de la température sous les différents corners. Une comparaison est réalisée avec la cellule 8 transistors classique.

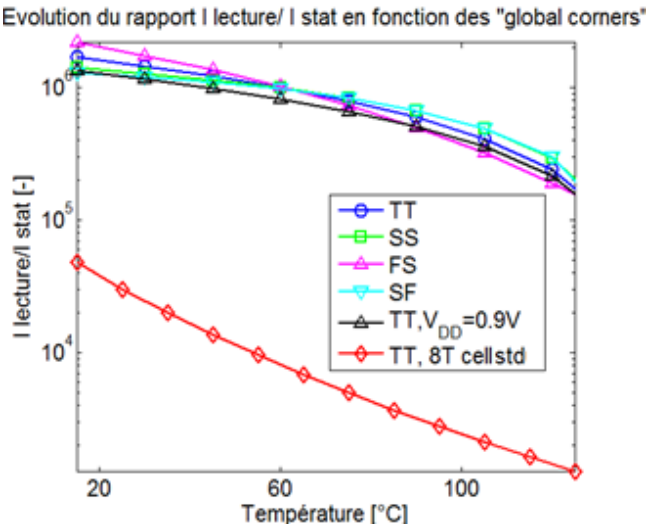


Fig. 5.3 : Evolution du rapport entre courant de lecture et courant statique de la cellule ULP 12 transistors en fonction de la température sous différents corners. Une comparaison est réalisée avec la cellule 8 transistors classique.

Outre la robustesse de la cellule face aux dispersions, le rapport entre courants I_{ON} et I_{OFF} des transistors d'accès peut poser des problèmes de fonctionnalités [27]. De manière plus particulière, comme nous l'avons déjà mentionné, il peut être difficile de discriminer un 1 d'un 0 lors d'une lecture. En effet, si le rapport I_{ON}/ I_{OFF} des transistors de lecture est trop faible, l'ensemble des courants I_{OFF} des transistors d'accès raccordés à la même bitline peut égaliser le courant I_{ON} des transistors d'accès de la cellule lue. Dès lors, la bitline se décharge à une vitesse similaire que la valeur mémorisée soit 1 ou 0. Par ailleurs, un grand nombre de cellules à faible consommation réduisent les pertes en fonctionnant en régime sous seuil [7] [28]. Cette technique possède l'inconvénient de restreindre significativement la fréquence de fonctionnement de la cellule.

C'est pourquoi nous avons choisi d'étudier ici le rapport entre courant de lecture et courant statique de la cellule. Ce courant comprend le courant statique de la paire d'inverseurs ULP et les courants de fuite des transistors d'accès. L'évolution de ce rapport en fonction de la température est représentée à la figure 5.3. Un grand rapport signifie que la cellule permet de réduire significativement la consommation de la cellule tout en maintenant une fréquence d'opération élevée. Comme I_{stat} contient également les courants de fuite des transistors de lecture, cela indique également qu'un nombre élevé de cellules pourra être raccordé aux mêmes bitlines. En effet, cela garantit que les courants I_{OFF} des transistors d'accès inactifs ne viendront pas perturber l'opération de lecture. Les courbes montrent que l'utilisation d'inverseurs ULP permet d'améliorer significativement ce rapport en comparaison à une cellule 8T réalisée à l'aide d'inverseurs classiques.

5.3. Variations locales

Nous venons d'étudier la sensibilité de la cellule aux perturbations globales. La variabilité locale est cependant plus critique alors que la taille des transistors devient de plus en plus petite [37]. Il s'agit ici d'évaluer l'impact d'une modification des performances d'un transistor à l'autre au sein d'une même cellule. Il n'est donc plus possible de ne considérer qu'un nombre fini de possibilités. La solution consiste alors à réaliser un nombre suffisamment élevé de simulations pour pouvoir réaliser une analyse statistique sur base de celles-ci. Nous parlerons alors de simulation de type Monte-Carlo. Etant donné l'investissement en temps nécessaire pour ces analyses, nous n'étudierons ici que la

sensibilité des marges de bruits face aux variations locales. La figure 5.4 permet donc d'observer les résultats d'une simulation de Monte-Carlo des marges de bruits. A chaque fois, 10 000 simulations ont été effectuées. Ce nombre élevé est nécessaire pour recueillir suffisamment de données éloignées de la moyenne. En effet, il se peut que la loi de distribution de ces données marginales soit différente de celle des résultats plus standards. Une comparaison est réalisée avec la cellule 8T afin d'évaluer le bien-fondé de la cellule ULP 12 transistors.

Si les deux cellules se comportent de manières fortement semblables en écriture, le cas de la lecture/rétention mérite notre attention. En effet, malgré une marge de bruit nettement supérieure, la cellule ULP est plus sensible aux variations locales et les résultats sont plus dispersés que pour la cellule 8T. Cela est dû à la structure même de l'inverseur ULP qui place les transistors en régime sous seuil où la sensibilité aux variations est amplifiée [38]. Il s'en suit que dans le cas d'une variation respectivement 4σ et 6σ (reprenant respectivement 99.997% et 99.99985% des données) la marge de bruit de la cellule ULP ne s'élève plus qu'à 280 mV et 205 mV en lecture. Des valeurs comparables sont obtenues pour la cellule 8 transistors. Ces valeurs sont suffisantes pour garantir la fonctionnalité de la cellule et indiquent qu'elle sera toujours opérationnelle sous de plus faibles tensions d'alimentation. De plus, la loi de variation des paramètres, fournie pour les modèles BSIM3 utilisés, prise en compte pour obtenir ces résultats ne tient pas compte de la promiscuité des transistors sur la tranche de silicium. Or, il existe une forte corrélation des paramètres de transistors proches comme le montre [39]. En effet, il n'est pas réaliste de considérer que deux transistors éloignés de moins d'un micromètre voient pour l'un une tension de seuil rehaussée de 100 mV ($3\sigma_{*}V_T$) et pour l'autre abaissée de 100 mV. Ceci entraine que la variabilité des marges de bruit obtenue est ici fortement surestimée et les cas que nous devons considérer sont donc bien plus proches des valeurs moyennes. Les SNM moyennes valent quant à elle 700 mV pour la cellule ULP 12 transistors contre seulement 355 mV pour la cellule 8 transistors conventionnelle[9].

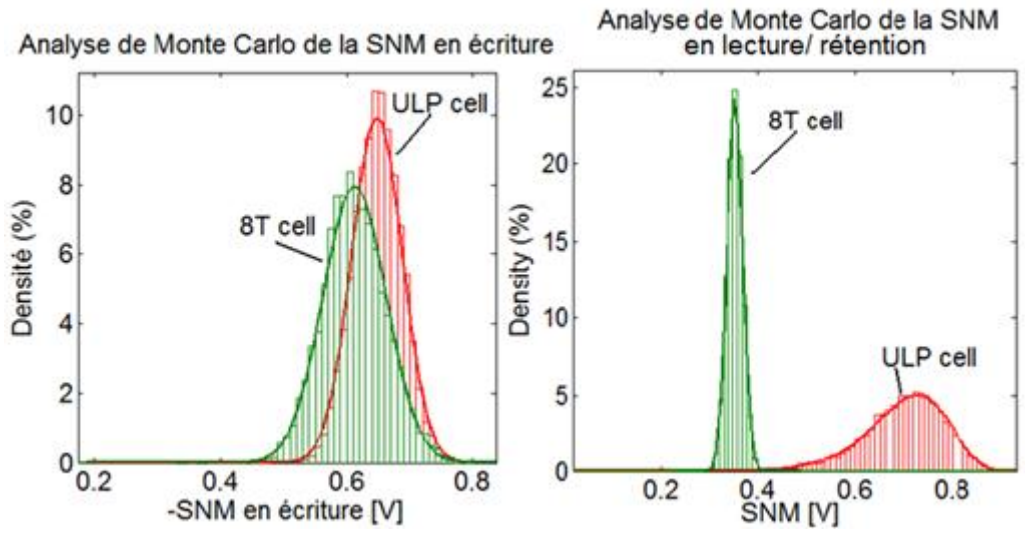


Fig. 5.4: Simulations de Monte-Carlo (10k) des marges de bruits des cellules 8 transistors classique et 12 transistors ULP. A gauche en écriture, à droite en lecture. σ_{VT}=34mV.

Notons également que dans le cas d'une opération d'écriture, la cellule ULP est moins sensible aux variations locales que la cellule 8T. Cela traduit le fait que malgré le mode de fonctionnement

[9] Pour de plus amples informations sur les lois de distributions permettant d'obtenir ces valeurs, veuillez vous référer à l'annexe 4.

sous seuil des inverseurs ULP, les transistors d'accès restent beaucoup plus forts. Ils n'éprouvent dès lors aucune difficulté à transmettre une valeur au nœud de mémorisation malgré l'hystérèse présente sur la courbe caractéristique de l'inverseur.

5.4. Conclusion

Des simulations ont été réalisées afin d'étudier la sensibilité de la cellule ULP 12 transistors face aux variations technologiques. Une étude de la sensibilité des performances par rapport à la température et à une réduction de la tension d'alimentation a également été effectuée.

Les résultats montrent que la cellule est robuste face aux variations globales à l'ensemble de la surface de silicium. A cause de son fonctionnement sous seuil, la cellule est plus sensible aux variations locales, mais conserve une excellente marge de bruit même en considérant une variation 6σ. Les fonctionnalités de la cellule sont maintenues jusque des températures supérieures à 125 °C.

Le rapport du courant $I_{lecture}/I_{stat}$ montre que la cellule permet d'obtenir une réduction des pertes statiques importantes tout en garantissant une cadence de fonctionnement élevée. En outre, il montre qu'il sera possible de raccorder un nombre élevé de cellules à la même paire de bitlines.

6. Exemple de layout

La surface occupée par un circuit aura un impact direct sur son cout de fabrication. Ce point devient très critique dans le cas des mémoires où des milliers de cellules doivent être juxtaposées. Il n'est donc pas étonnant que des règles de design spécifiques soient proposées par les fondeurs pour la réalisation d'une SRAM. Ces règles permettent par exemple un plus petit espacement minimal entre deux zones actives.

Afin d'évaluer la surface nécessaire à la réalisation d'une cellule ULP 12 transistors, un layout a été réalisé. Ce layout respecte les règles de designs du fondeur « ST » pour la technologie bulk 130 nm. Deux cellules classiques à 6 et 8 transistors ont également été dessinées pour la comparaison.

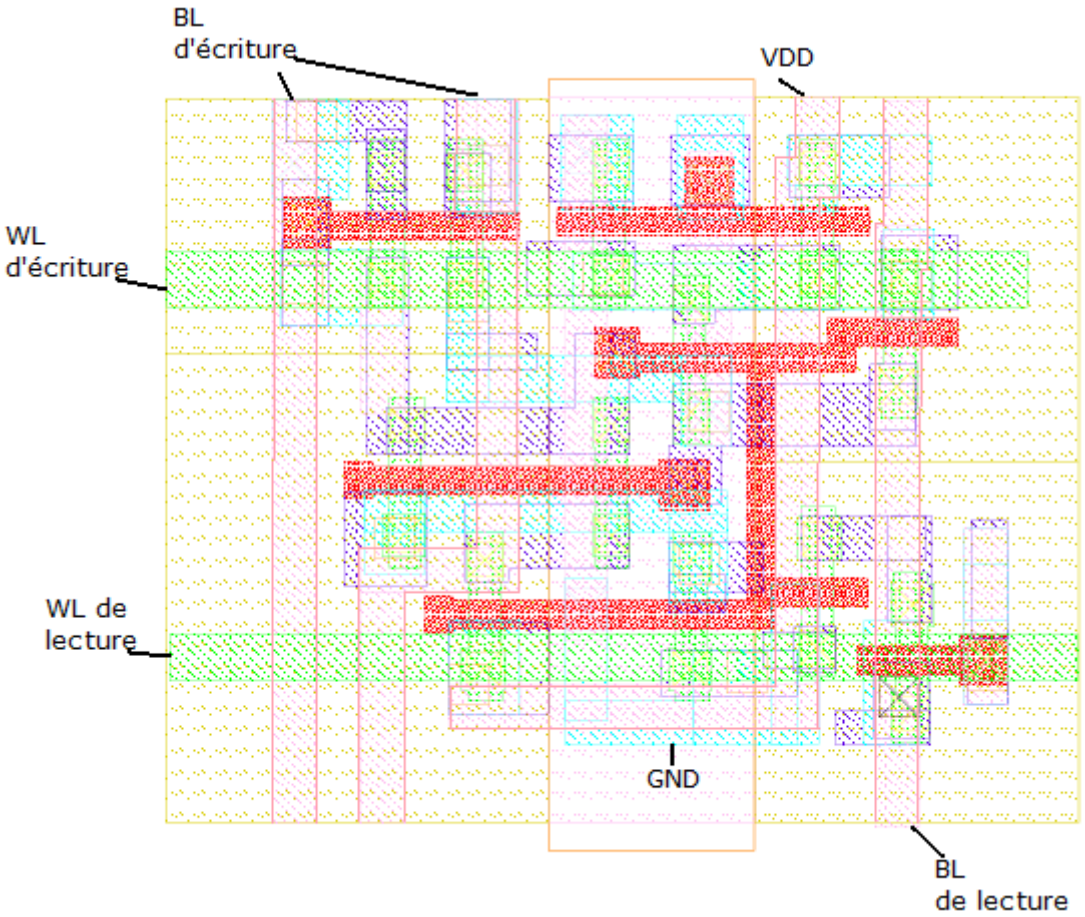


Fig. 6.1 : Layout d'une cellule ULP 12 transistors en technologie bulk 130 nm. Les dimensions de la cellule sont de 3.65 µm sur 3.15 µm, pour une surface totale occupée de 11.50 µm^2.

La cellule ULP 12 transistors, représentée à la figure 6.1, occupe une surface 2,6 fois plus importante que la cellule 6 transistors classiques, dont un exemple de layout est donné à la figure 6.2 (a). Ce rapport élevé est dû à l'utilisation de deux types de transistors différents au sein de la cellule.

En effet, il est alors nécessaire de respecter certaines « design rules » empêchant de définir une zone active trop proche des bordures d'une zone de dopage ce qui rend le layout un peu moins dense. Par rapport à une cellule classique 8 transistors, pour les mêmes raisons, la surface occupée reste 1,9 fois plus importante. Un exemple de layout de cette cellule est représenté à la figure 6.2 (b).

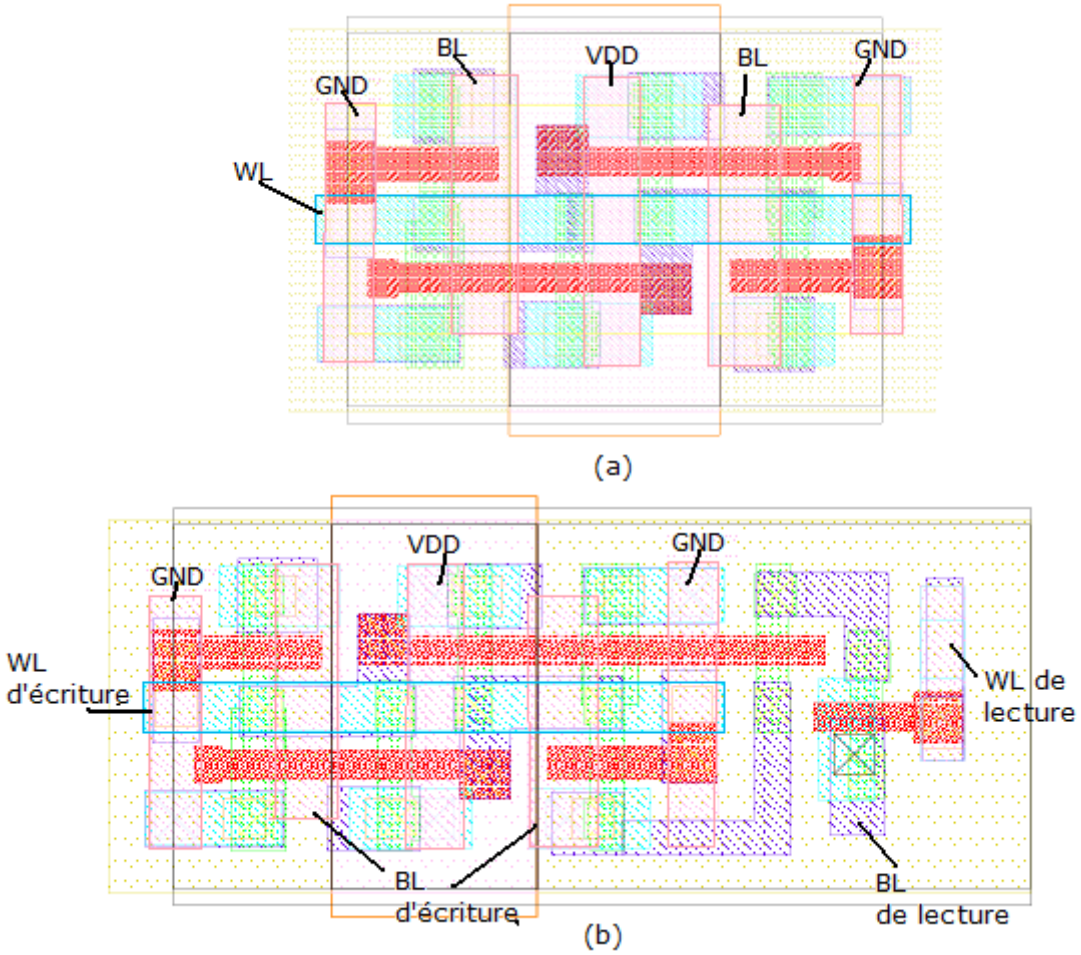


Fig. 6.2 : (a) layout d'une cellule classique 6 transistors en technologie bulk 130 nm. Les dimensions de la cellule sont de 2.86 µm sur 1.53 µm, pour une surface totale occupée de 4.36 µm^2. (b) layout d'une cellule classique 8 transistors en technologie bulk 130 nm. Les dimensions de la cellule sont de 4.00 µm sur 1.53 µm, pour une surface totale occupée de 6.12 µm^2.

7. Perspectives

7.1. Modification des niveaux de polarisation

Dans les sections précédentes, nous avons vu qu'afin de réduire les courants de fuite à travers les transistors d'accès, il est utile d'appliquer une tension de wordline négative en rétention. Cependant, il existe d'autres possibilités pour atteindre cet objectif. Il est même possible de modifier les tensions de polarisation de la cellule pour réduire plus avant les pertes statiques.

Pour atteindre cet objectif, nous allons envisager deux mesures séparées : une réduction de la tension d'alimentation des deux inverseurs ULP ou une augmentation de leur tension de masse. Nous rappelons ici le schéma de la cellule ULP à 12 transistors à la figure 7.1 pour faciliter les explications.

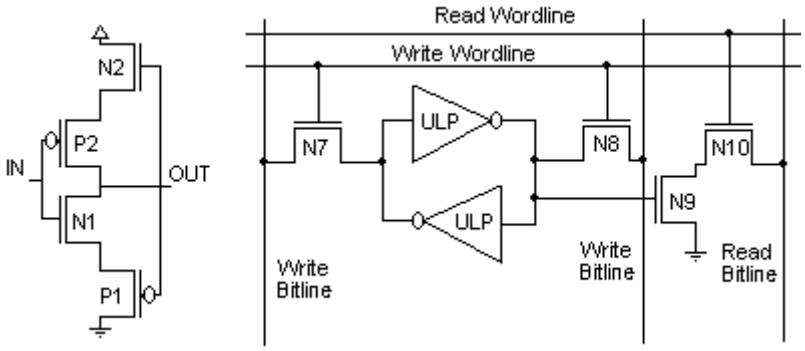


Fig 7.1 : Schéma de la cellule SRAM ULP 12 transistors. $V_{BN}=0$, $V_{BP}=V_{DD}$.

Réduction de la tension d'alimentation

Comme les pertes statiques dépendent fortement de V_{DD}, sa réduction permet facilement de réduire les pertes statiques. Il pourrait sembler que cette mesure aura pour effet d'augmenter le temps de lecture. En effet, la tension appliquée à la grille de N9 est restreinte par cette action et il en est donc de même de sa tension d'overdrive. Cependant, il faut se rappeler que les transistors d'accès en écriture N7 et N8 ne peuvent transmettre une valeur haute à la cellule qu'à leur tension de seuil près. Par la suite, les inverseurs ULP doivent régénérer ce niveau logique, ce qui prend quelques microsecondes. Si l'on veut évaluer le temps de lecture de la cellule, il faut donc considérer le cas le plus critique où les inverseurs n'auront pas encore régénéré le 1 logique.

Partant de cette constatation, on peut remarquer qu'il n'est pas pénalisant pour le temps de lecture de réduire la tension V_{DD} des inverseurs d'une valeur inférieure ou égale à la tension de seuil des transistors d'accès (soit environ 200 mV). Au contraire, en plus de réduire les pertes statiques nous diminuons alors également les pertes actives de la cellule en simplifiant le travail que devra fournir l'inverseur ULP lors d'une écriture.

Élévation de la tension de référence de la cellule

Une augmentation de la tension de référence de la cellule permet au même titre que la réduction de V_{DD} de réduire les pertes statiques. En outre, cela a d'autres impacts sur la cellule tels que l'application d'une tension V_{GS} négative aux transistors d'accès. Il n'est donc plus nécessaire d'imposer une tension de wordline négative en rétention. De plus, le niveau de polarisation du substrat par rapport à la source de certains transistors a été modifié. Les fuites des transistors d'accès sont alors encore réduites grâce à l'évolution de V_{BS}. Si l'on suppose une élévation de 200 mV de cette tension, la figure 7.2 illustre ces deux phénomènes.

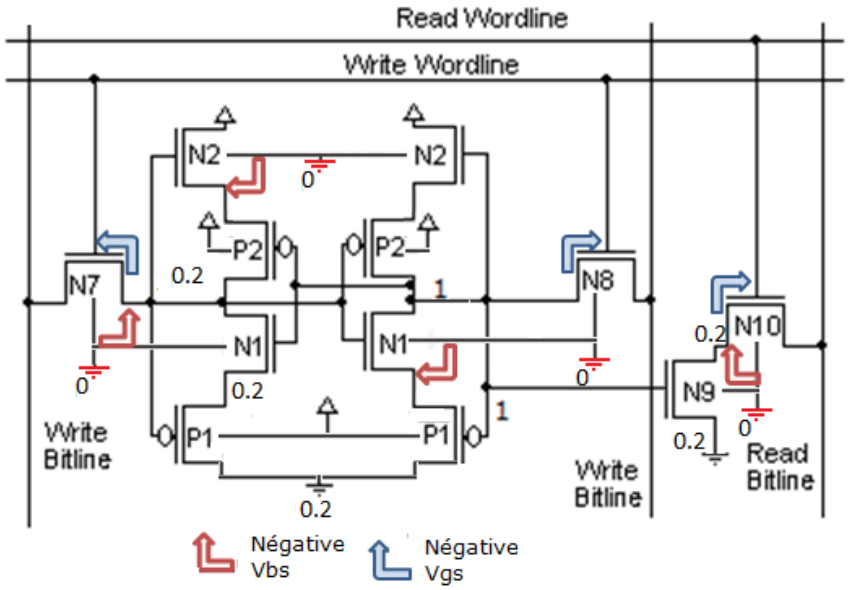


Fig. 7.2: Modifications des niveaux de polarisation de la cellule ULP 12 dues à l'augmentation de la tension de référence de la cellule de 200 mV.

Toutefois, lors d'une écriture sur une autre cellule reliée aux mêmes bitlines, un des transistors d'accès ne possède plus une tension V_{GS} négative si le niveau de la bitline est abaissé à 0V. Si ce transistor est relié au nœud de mémorisation chargé à V_{DD}, son courant de fuite augmente alors sensiblement. Cette perturbation peut déstabiliser l'état de la cellule. Pour éviter ce phénomène, lors d'une écriture, la bitline ne doit pas être déchargée totalement, mais seulement à 200 mV.

Des simulations ont été réalisées afin de quantifier l'action d'une modification des niveaux de polarisation. Les résultats de ces simulations sont repris au tableau 7.1. Une variation de 200 mV a été appliquée aux à V_{DDcell} et $V_{GNDcell}$.

	$T_{lecture}$	$T_{écriture}$	$P_{statique}$	$T_{restauration}$
Cas standard	0.75 ns	1.37 ns	13 pW	2.4 µs
↘ V_{DDcell}	0.77 ns	0.15 ns	9.86 pW	7.18 ns
↗ $V_{GNDcell}$	1.95 ns	1.34 ns	10.0 pW	1.3 µs
↗ $V_{GNDcell}$ ↘ V_{DDcell}	1.97 ns	0.15 ns	7.7 pW	7.15 ns

Tableau 7.1: résultats de l'augmentation de la tension de masse de la cellule de 200 mV, de la réduction de la tension d'alimentation de la cellule de 200 mV et d'une combinaison de ces deux mesures. T=25 °C.

Il ressort de ces simulations que l'action la plus efficace consiste à réduire la tension d'alimentation de la cellule. En effet, cela permet de réduire les pertes statiques de 25% tout en maintenant les performances dynamiques de la cellule. De plus, les inverseurs ne doivent plus restaurer le 1 logique après une opération d'écriture. Notons que le temps d'écriture est significativement amélioré grâce au fait que les transistors d'accès peuvent plus facilement transmettre une valeur haute au nœud de mémorisation. Cela revient en quelque sorte, si l'on prend le point de vue de la tension de wordline à utiliser une technique de « boosted wordline » telle que proposé dans [21].

Si l'augmentation de la tension de référence permet d'obtenir une réduction des pertes du même ordre de grandeur, le temps de lecture lui est dégradé. Cela est lié au fait que la tension d'overdrive du transistor N9 (fig. 7.1) est réduite de 200mV. En effet, afin de maintenir une tension V_{GS} négative au transistor d'accès en lecture N10, il est nécessaire de relever également la tension de source de N9. Notons que le temps de restauration de la valeur mémorisée est ici plus court. En effet, la tension de wordline d'écriture n'est plus ramenée à une valeur négative suite à un accès. Cela mène à une réduction des effets de couplages capacitifs consécutifs à une opération d'écriture décrits à la section 4.5.2.

Lorsque l'application le nécessite, il est possible de couper plus avant les pertes statiques en combinant l'augmentation de $V_{GNDcell}$ et la réduction V_{DDcell}. Ceci permet une diminution des pertes par un facteur proche de 2. Cependant, il faut alors faire avec les inconvénients liés à l'accroissement de $V_{GNDcell}$ en termes de temps de lecture

7.2. Résultats en technologie avancée (45 nm)

Une étude a été réalisée afin d'établir la viabilité de la cellule dans les technologies future. Pour ce faire, la cellule a été simulée avec un modèle prédictif 45nm de type BSIM4 [40]. La question de la stabilité de la cellule se pose alors. Peut-on garantir que les effets de canal court ne vont pas perturber la cellule au point de lui faire perdre les données mémorisées ? D'autant plus que pour de si petites longueurs de grille, les variations de procédés sont plus difficiles à contrôler ce qui accroit significativement la variabilité des caractéristiques des transistors.

Dans l'objectif de répondre à ces questions, nous décrirons tout d'abord les principaux effets de canal court afin de bien cerner les différents phénomènes en jeux sous les technologies inférieures aux 100 nm. Ensuite, nous passerons à la simulation proprement dite de la cellule étudiée.

7.2.1. Caractéristiques des transistors utilisés

Avant toute chose, les caractéristiques des transistors utilisés sont présentées ci-dessous. Comme nous l'avons déjà mentionné, les simulations ont été réalisées à l'aide de modèles prédictifs 45 nm. Malheureusement, ces modèles ne sont pas calibrés correctement en ce qui concerne les courants de jonctions. Cependant, pour cette technologie et notre application, comme nous le verrons par la suite, les problèmes se posent surtout par rapport aux courants de grille, qui sont quant à eux bel et bien repris par ces modèles. A nouveau, une combinaison de transistors de type « high speed » et de type « low leakage » a été choisie. Les transistors « high speed » permettent de réaliser les inverseurs ULP tandis que les transistors « Low Leakage » correspondent aux transistors d'accès. Les oxydes de grille sont réalisés à l'aide de matériaux « high-k » afin de réduire les courants de grille. Une discussion sera réalisée à ce sujet au paragraphe suivant.

- Voici les caractéristiques du modèle « LL » :

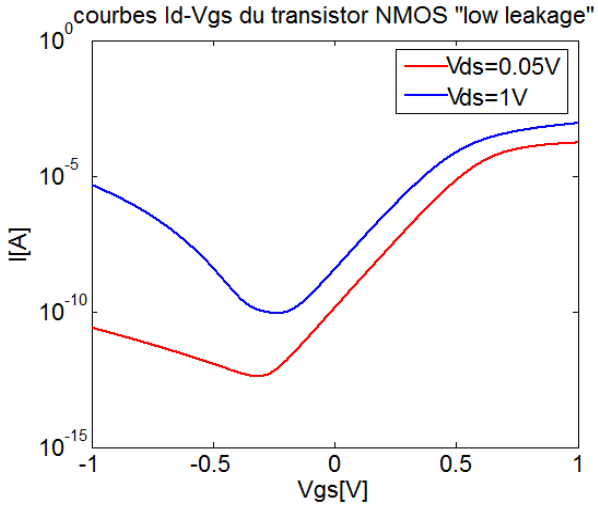


Fig. 7.1 : Courbe I_D-V_{GS} du NMOS LL pour V_D=0.05V et 1V. V_S=V_B=0V, W=1µm, L=45nm, T=25°C.

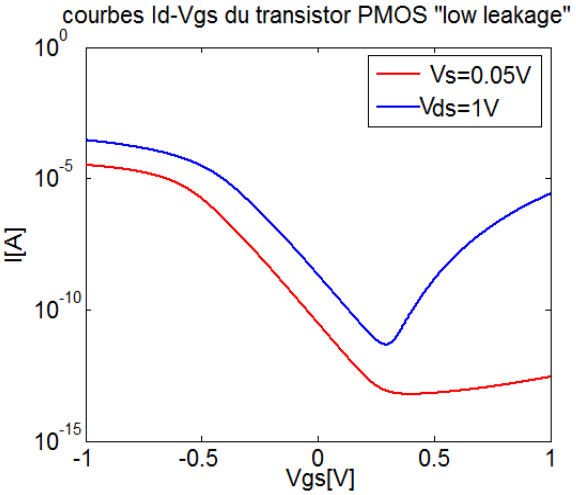


Fig 7.2 : Courbe I_D-V_{GS} du PMOS LL pour V_D=-0.05V et -1V. V_S=V_B=0V, W=1µm, L=45nm, T=25°C.

Une extraction à partir de la simulation de la tension de seuil de ces transistors montre qu'elle s'élève à 456 mV pour le NMOS et à 348 mV pour le PMOS pour une température de 25 °C et une tension V_{DS} de 1V.

$\|V_{DS}\|$ [V]	I_{ON} NMOS [A/µm]	I_{ON} PMOS [A/µm]	I_{OFF} NMOS [A/µm]	I_{OFF} PMOS [A/µm]
1	8.7E-04	2.9E-04	3.8e-9	2.2e-9
0.05	1.7E-04	3.3E-05	1.4e-10	3.2e-11

Tableau 7.1 : courants I_{ON} et I_{OFF} des transistors low leakage. Le courant est donné pour un $\|V_{GS}\|$=1 ou 0V et pour $\|V_{DS}\|$ = 1V et 0.05V, T=25°C.

Le modèle « HS » possède quant à lui les caractéristiques suivantes :

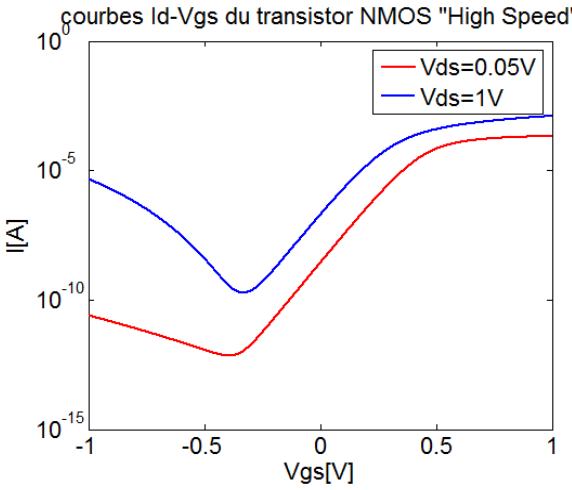


Fig. 7.3 : Courbe I_D-V_{GS} du NMOS HS pour V_D=0.05V et 1V. V_S=V_B=0V, W=1µm, L=45nm, T=25°C.

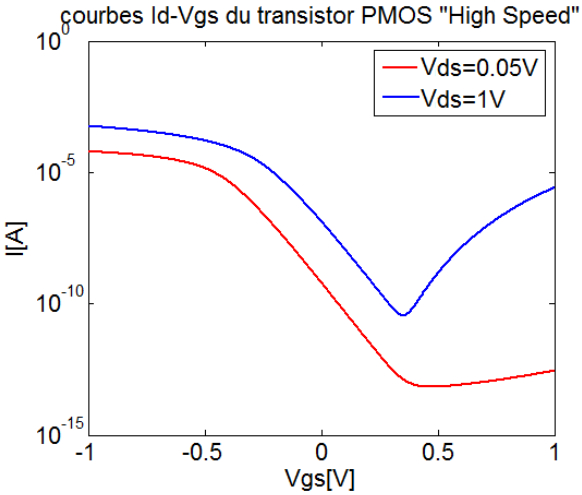


Fig 7.4 : Courbe I_D-V_{GS} du PMOS HS pour V_D=-0.05V et -1V. V_S=V_B=0V, W=1µm, L=45nm, T=25°C.

Une extraction à partir de la simulation de la tension de seuil de ces transistors montre qu'elle s'élève à 272 mV pour le NMOS et à 184 mV pour le PMOS pour une température de 25 °C et une tension V_{DS} de 1V.

$\|V_{DS}\|$ [V]	I_{ON} NMOS [A/µm]	I_{ON} PMOS [A/µm]	I_{OFF} NMOS [A/µm]	I_{OFF} PMOS [A/µm]
1	1.3e-3	5.8E-04	2.0e-7	1.4e-07
0.05	2.1e-04	6.2e-05	2.8e-09	6.4e-10

Tableau 7.2 : courants I_{ON} et I_{OFF} des transistors high speed. Le courant est donné pour un $|V_{GS}|$=1 ou 0V et pour $|V_{DS}|$ = 1V et 0.05V, T=25°C.

Nous pouvons remarquer sur ces graphes que les courants de fuite des transistors sont beaucoup plus importants lorsque la tension V_{DS} vaut 1V. Comme nous le verrons dans la section suivante, ceci est dû aux courants de grille, qui ici sont d'autant plus importants que la tension V_{GS} est grande.

7.2.2. Les effets de canal court

Alors que l'on réduit les dimensions caractéristiques des transistors, il faudrait diminuer la valeur de la tension d'alimentation des circuits pour maintenir une évolution constante des champs électriques au sein des dispositifs CMOS [41]. L'augmentation de ces champs mène à une détérioration de certaines caractéristiques des transistors. C'est ce que l'on appelle les effets de canal court.

Nous décrirons ici uniquement les impacts sur les courants de fuite des effets DIBL et de courants de grille. Cependant, d'autres difficultés peuvent également être mises en évidence telle que l'augmentation des résistances d'accès par exemple [1] [42].

Les courants de grille

Alors que les dimensions des transistors diminuent, il faut pouvoir garantir le maintien du contrôle de la grille sur le canal. Il est donc nécessaire d'accroitre la capacité de grille C_{OX} en diminuant l'épaisseur de l'oxyde. Cependant, lorsque celle-ci atteint l'ordre du nanomètre, les charges présentes dans le canal peuvent traverser la grille par effet tunnel [42]. Il s'agit d'un phénomène décrit par la physique quantique où une particule a une probabilité non nulle de franchir une barrière de potentiel sans posséder l'énergie requise à ce passage. D'après [43], ce phénomène limitera le choix du SiO_2 comme oxyde de grille lorsque son épaisseur devrait devenir inférieure aux 2 nm. Toutefois, plus récemment l'ITRS 2007 a prévu l'introduction de matériaux « high-k » pour cette année 2008. Il s'agit de matériaux possédant une constante diélectrique plus élevée que celle du SiO_2. A capacité identique, l'épaisseur de l'oxyde peut alors être augmentée, ce qui permet de réduire significativement les courants de grille. Il est donc possible de définir la notion d'épaisseur d'oxyde équivalente (EOT). Pour une capacité et un matériau de constante diélectrique K_{highK} donnés, celle-ci correspond à l'épaisseur de SiO_2 requise à l'obtention de cette même capacité.

$$EOT = T_{oxyde} * {K_{SiO_2}}/{K_{high\ k}}$$

Afin de mieux cerner l'ampleur de ce phénomène, le tableau 7.3, extrait de l'ITRS 2007, donne une prévision de ces courants pour les prochaines années.

Year of Production	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*
EOT: Equivalent Oxide Thickness [2]						
Extended planar bulk (Å)	11	9	7.5	6.5	5.5	5
$J_{g,limit}$: *Maximum gate leakage current density [5]*						
Extended Planar Bulk (A/cm^2)	8.00E+02	9.09E+02	1.00E+03	1.11E+03	1.25E+03	1.43E+03

Tableau 7.3 : Prévision de l'évolution des courants de grille et de l'épaisseur d'oxyde équivalente pour les prochaines années. Les cases surlignées en jaune et en rouge correspondent à l'utilisation de matériaux « high-k ». Tableau issu de [1]

La société Intel a choisi d'introduire les matériaux « high-k » dans la réalisation de processeurs pour le nœud technologique 45 nm[10][44].

L'effet DIBL (Drain Induced Barrier Lowering)

L'effet DIBL apparait lorsque les zones de déplétion entre source et substrat et entre drain et substrat atteignent des dimensions similaires à la longueur de grille du transistor [41] [42]. On peut alors observer un certain contrôle du drain sur le canal. En effet, normalement seule la grille du transistor peut ou non permettre aux charges de traverser le canal. Dans le cas d'un NMOS au repos, les jonctions PN entre zones actives et substrat sont bloquées. Il existe une barrière de potentielle empêchant les charges de traverser la jonction comme illustré à la figure 7.5.

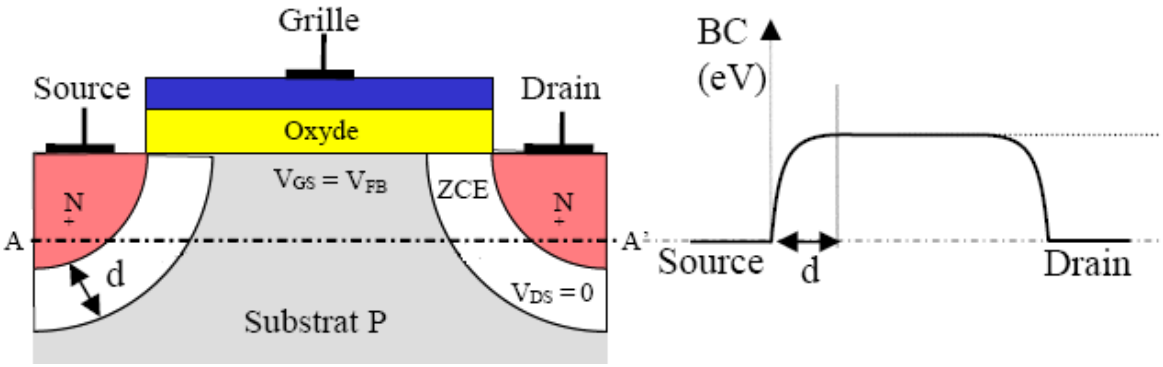


Fig. 7.5 : L'effet DIBL apparait lorsque les zones de charge d'espace (ou zone de déplétion) atteignent des dimensions similaires à la longueur de grille. Figure issue de [41].

Si la tension de drain augmente, la zone de déplétion entre drain et substrat s'étend et peut rejoindre celle de la source. Ceci entraine une réduction de la barrière de potentielle comme indiqué à la figure 7.6. Il est alors possible de comprendre la conséquence de ce phénomène en parlant d'une réduction de la tension de seuil V_T du transistor correspondant à cette baisse de la barrière de potentiel et proportionnelle à la tension de drain. Le contrôle du canal n'est donc plus uniquement dépendant de la tension de grille.

[10] Ce choix a été décrit par Gordon Moore lui-même, peut-être exagérément, comme « l'évolution la plus marquante pour les transistors depuis quarante ans » [44].

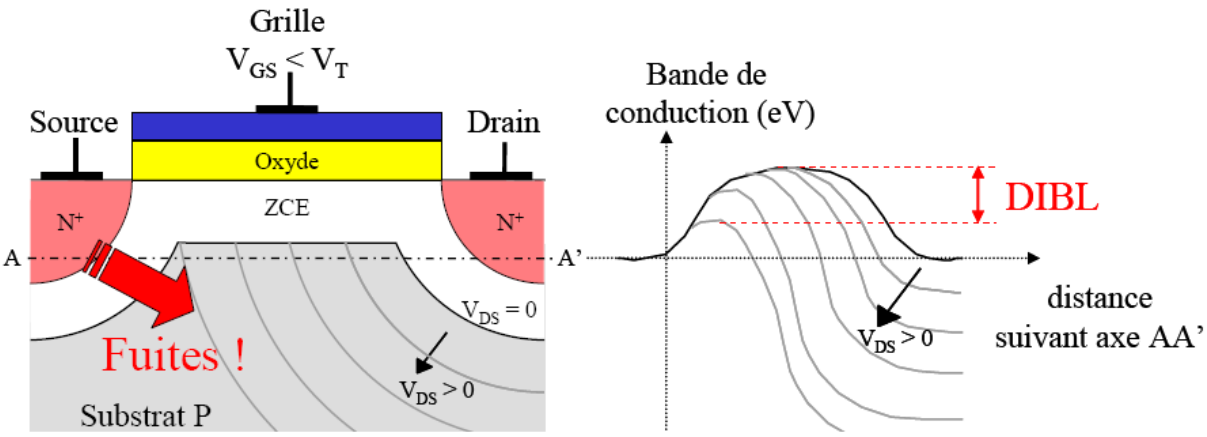


Fig. 7.6 : Lorsque les zones de déplétions se rejoignent sous l'influence de la tension de drain, il apparait une baisse de la barrière de potentielle entre source et substrat. Ceci peut-être apparenté à une réduction de la tension de seuil du transistor. Figure issue de [41]

Illustration

Il peut être intéressant de visualiser à l'aide d'une simulation l'impact de ces différents effets sur un transistor du modèle considéré et sur les transistors ULP. A cet effet, la figure 7.7 permet de visualiser l'importance des effets de canal court.

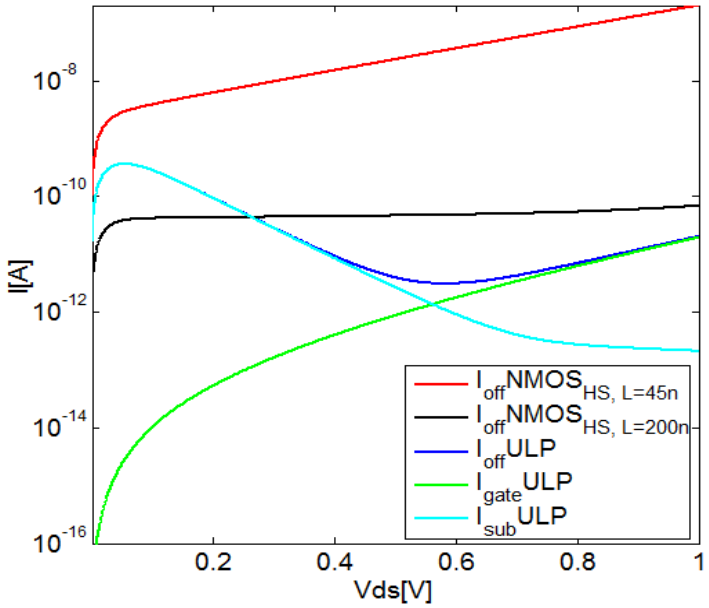


Fig. 7.7 : Evolution des courants I_{OFF} du NMOS HS pour W=1µm, L= 45 nm et L= 200nm, et de l'inverseur ULP pour W_N=1µm et W_P=45nm. V_G=0V, T=25°C.

Tout d'abord, la différence de la pente du courant OFF par rapport à V_{DS} entre une longueur de grille de 45 nm et 200 nm permet de visualiser l'effet DIBL. Pour la courbe rouge, correspondant à une longueur de grille de seulement 45 nm, on peut observer une nette dépendance du courant de fuite

par rapport à la tension de drain. Si on augmente la longueur de la grille (la courbe noire), cette dépendance disparait presque, ce qui montre bien qu'il s'agit là de l'effet DIBL.

Ensuite, nous pouvons remarquer que les courants de fuite du transistor ULP sont nettement influencés par les courants de grille pour des tensions de drain supérieure à 600 mV. Pour limiter l'impact de ceux-ci, il est donc préférable de choisir une tension d'alimentation de 600 mV. Les modèles utilisés font usage de matériaux high-k. C'est ce qui permet de limiter dans une certaine mesure les courants de grille, qui pour le transistor « HS » restent négligeables devant le courant sous-seuil.

7.2.3. Simulation de la cellule ULP 12 transistors

Il s'agit tout d'abord de dimensionner la cellule pour assurer sa fonctionnalité sous cette technologie 45 nm. Pour ce faire, des analyses Monte-Carlo (1k) ont été réalisées afin d'étudier la marge de bruit en lecture et en écriture de la cellule. Prendre une taille minimale pour tous les transistors de la cellule ne permet plus cette fois-ci de garantir une stabilité 6σ. En outre, une nette dégradation de la marge de bruit en lecture/rétention est observée. En effet, celle-ci n'est plus supérieure à $V_{DD}/2$, mais est plutôt proche de 400 mV pour V_{DD}=1V.

Pour comprendre la diminution des marges de bruit, il faut examiner plus en détail les différents courants de fuite de ces transistors. Ceux-ci sont représentés à la figure 7.7. Il ressort que cette fois il faut jouer avec les courants de grille qui ont pris de l'importance comme expliqué à la section 7.2.2. L'ensemble de ceux-ci est représenté à la figure 7.8 sur une cellule ULP 12 transistors. Sur cette figure, on peut observer que les niveaux logiques seront détériorés si les courants de grille sont trop importants devant le courant de chaque inverseur. Or, le courant de régénération des inverseurs est limité au courant sous seuil de P1 et N2 comme nous l'avons vu à la section 3.5. La courbe caractéristique de l'inverseur ULP en technologie 45 nm est en effet dégradée. Nous pouvons observer à la figure 7.9 une forte détérioration du 0 logique et dans une moindre mesure du 1 logique. Ce phénomène est d'autant plus important que V_{IN} augmente et accroit donc les courants de grille. Ceci est la cause de la perte de marge de bruit en lecture/écriture.

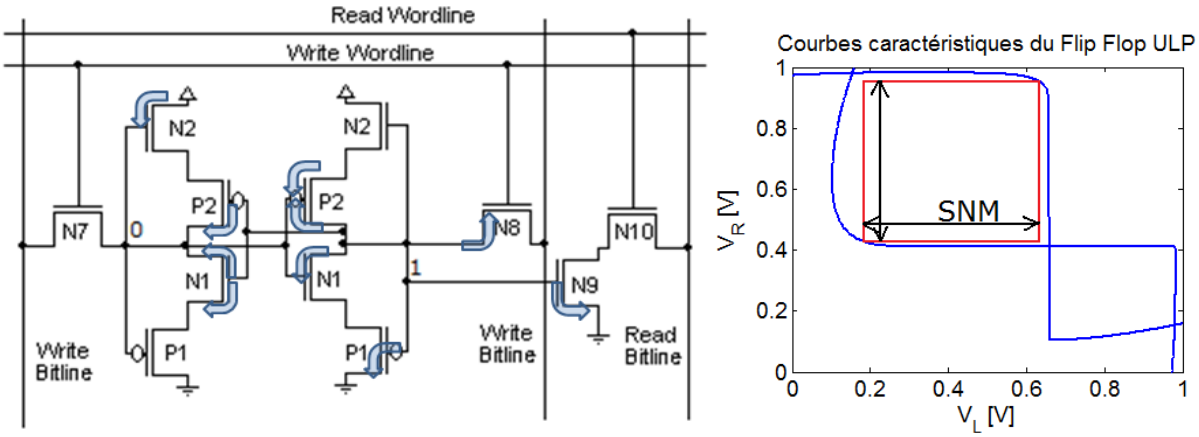


Fig. 7.8:A gauche, les courants de grille de la cellule ULP 12 transistors. A droite, les courbes permettant de déterminer la marge de bruit en rétention/lecture. On observe une dégradation de la marge de bruit suite à la forme des courbes caractéristique du latch ULP.

Pour contourner ce problème, un redimensionnement de l'inverseur est possible afin d'accroitre son courant I_{ON}. Pour cela, il suffit d'augmenter la largeur de P1 et N2. Une autre solution déjà mentionnée consiste à réduire la tension d'alimentation à 600 mV. Cependant, dans ce cas de figure la différence entre

seuils de basculement haut et bas est réduite. Ceci entraine également une perte de marge de bruit, c'est pourquoi cette solution n'a pas été retenue.

La figure 7.9 nous montre également que les seuils de basculement de l'inverseur sont plus proches de $V_{DD}/2$ qu'en technologie bulk 130 nm. La raison de ceci est liée à l'effet DIBL. En effet, nous avons vu que lorsque la tension de drain est élevée, celui-ci accroit le courant dans le canal. Dès lors lorsque la sortie d'un inverseur ULP est haute, le courant du transistor ULP de type N sera plus important par effet DIBL. Lorsque l'entrée de cet inverseur suivra un flanc montant, l'effet DIBL va donc aider son transistor ULP de type N à décharger le nœud de sortie et mènera plus rapidement au basculement. Un même raisonnement peut être effectué pour le cas d'un flan descendant. C'est ce qui réduit l'hystérèse de l'inverseur.

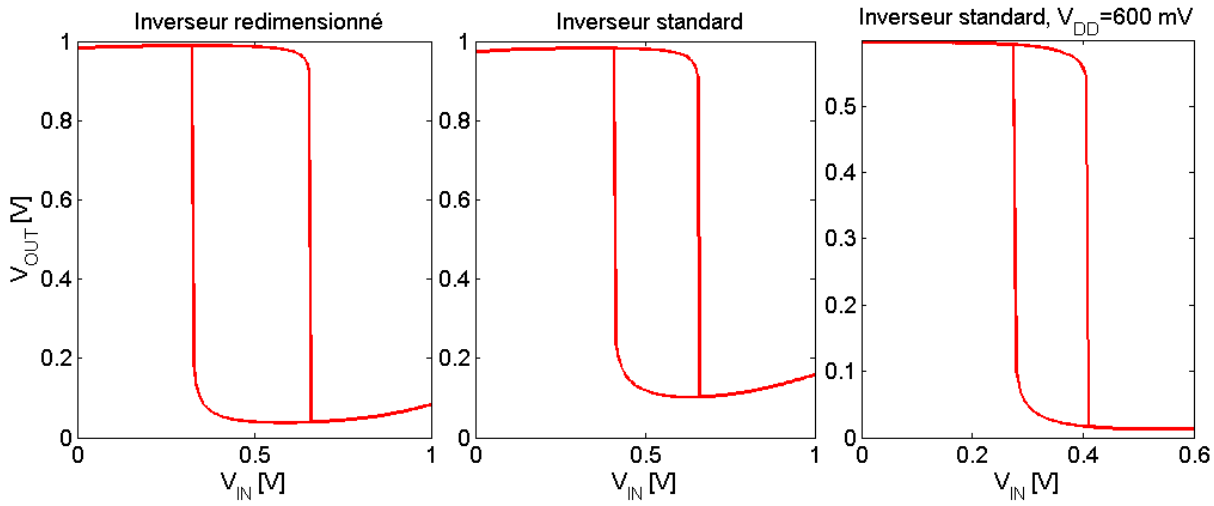


Fig. 7.9: Courbes caractéristiques de l'inverseur ULP (W= 80 nm, L=45 nm). La figure de gauche correspond au cas d'un redimensionnement du PMOS relié à la masse (W_{footer}=240 nm, L_{footer}= 45nm) et du NMOS relié à V_{DD} (W_{header}=120 nm, L_{header}= 45 nm),V_{DD}=1V. La figure du milieu correspond aux cas ou tous les transistors possèdent le même dimensionnement (W= 80 nm, L= 45 nm), V_{DD}=1V. Il en est de même pour la figure de droite, V_{DD}=600 mV. T=25°C.

Si un redimensionnement de la cellule réduit l'impact du courant de grille, malheureusement les procédés de fabrication entrainent une plus grande variabilité que pour la technologie 130 nm. Pour garantir une marge de bruit suffisante, il faut donc sacrifier un peu de surface de silicium. En effet, la variabilité des propriétés des transistors peut être réduite en augmentant la largeur de l'ensemble des transistors de la cellule de 20 nm (W=80 nm au lieu de 60 nm)[11]. Les simulations de Monte-Carlo montrent alors que la cellule possède une marge de bruit en lecture/rétention de 220 mV et de -160 mV en écriture pour une variabilité 6σ. Les moyennes des résultats obtenus s'élèvent à 465 mV en lecture/rétention et à -360 mV en écriture. Ces résultats restent très bons et montrent qu'il sera toujours possible d'utiliser cette cellule dans les technologies futures.

[11] [37] montre que σ_{VT} est proportionnel à $\frac{1}{\sqrt{WL}}$.

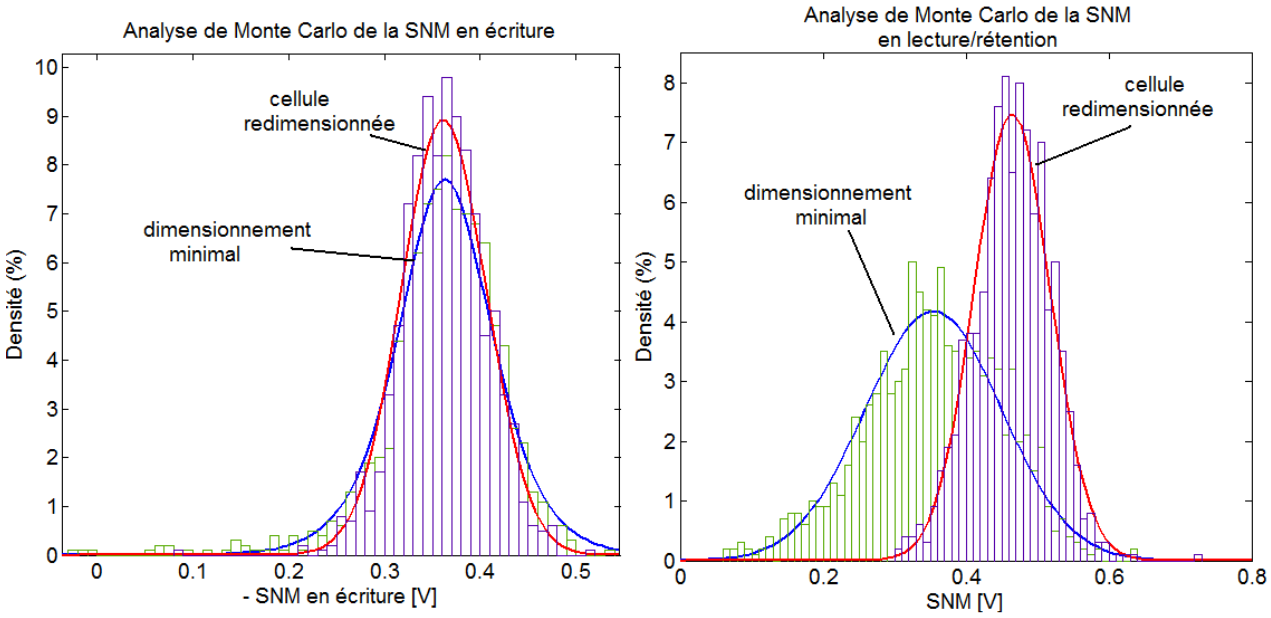


Fig. 7.4: Simulations de Monte-Carlo (1k) des marges de bruits de la cellule 12 transistors ULP pour un dimensionnement minimal de tous les transistors (W=60 nm, L=45 nm) et pour un redimensionnement de la cellule (W= 80nm sauf W_{P1}= 240nm et W_{N2}=120 nm, L=45 nm,). A gauche en écriture, à droite en lecture. =34mV pour des transistors de taille minimale (W=60 nm, L= 45 nm).

8. Conclusion

Ces dernières années ont vu une réduction continue de la taille des transistors présents dans les circuits sur puces. Si cette évolution permet d'atteindre des densités d'intégrations et des performances dynamiques supérieures, elle entraine toutefois une augmentation des pertes statiques des transistors. Par ailleurs, les SRAMs occupent une partie importante de nombre de circuits tels que les microprocesseurs. Comme il n'est pas possible de couper l'alimentation de ces mémoires lors d'une mise en veille, sous peine de perdre les informations mémorisées, leur consommation de puissance statique est une partie prépondérante des pertes de tels systèmes. Il est donc du premier intérêt d'arriver à développer une solution permettant de mitiger les courants de fuite de chaque cellule composant la mémoire. Par ailleurs, l'ITRS 2007 annonce que, dans les années à venir, il faudra s'attendre à rencontrer certaines difficultés pour contrer la perte de stabilité des SRAMs, suite à l'impact croissant de la variabilité des performances des transistors due aux fluctuations lors de la fabrication.

A cet effet, des montages basés sur l'assemblage d'un NMOS et d'un PMOS, et appelés transistors ULP permettent de réduire les courants de fuite de plusieurs ordres de grandeur par rapport à un MOSFET classique. Le mécanisme des réductions des fuites est basé sur l'application d'une tension V_{GS} négative auto-induite. De cette manière, le courant sous seuil est coupé lorsque la tension entre drain et source du transistor ULP augmente. Il est alors possible de réaliser un inverseur basé sur deux de ces transistors ULP. Cet inverseur possède des pertes statiques réduites de trois ordres de grandeur par rapport à celles de l'inverseur CMOS. Il faut également noter la présence d'une hystérèse sur la courbe caractéristique de ce montage qui pourra être mise en valeur pour accroitre la stabilité d'un latch ULP. Malheureusement, il y a un prix à payer en termes de performance dynamique. En effet, le délai de ce type d'inverseur est accru de 5 ordres de grandeur par rapport à l'inverseur CMOS.

Différentes cellules SRAM ont été élaborées sur base de ces éléments ULP. Si on considère comme facteurs de mérites les performances dynamiques et la réduction des pertes statiques, une cellule composée de 12 transistors se révèle être le meilleur compromis. Cependant, une dépense importante doit être réalisée en termes de surface de silicium occupée. Cette cellule utilise comme élément de mémorisation un latch ULP composé de deux inverseurs ULP. Un buffer de lecture composé de deux NMOS permet de maintenir un temps de lecture indépendant du délai élevé de l'inverseur ULP. La cellule mémoire montre une réduction des pertes statiques par un facteur 50 par rapport à une cellule 6 transistors conventionnelle. Par ailleurs, l'hystérèse des inverseurs ULP permet d'obtenir une marge de bruit supérieure à $V_{DD}/2$ impossible à atteindre à l'aide d'une architecture conventionnelle. La robustesse de la cellule a également été prouvée par rapport aux variations technologiques ou environnementales (température, tension d'alimentation…).

Enfin des essais ont montré que pour les technologies à venir, la cellule subira l'impact de l'accroissement des courants de grille. En effet, ces courants deviennent une composante importante des pertes statiques des transistors ULP. Certaines mesures doivent donc être prises pour maintenir les fonctionnalités de l'inverseur ULP et donc de la cellule ULP à 12 transistors. Il s'agit essentiellement de se baser sur des matériaux « high-k » comme oxyde de grille et de redimensionner la largeur des transistors en présence.

Finalement, il serait possible de continuer sur le chemin qui a commencé à se dégager. Si la cellule proposée présente une très faible consommation statique, il reste à réaliser l'ensemble des circuits périphériques. Ceux-ci devraient permettre quant à eux une réduction des pertes dynamiques de la mémoire. Une fois ces designs validés, il resterait alors à envoyer un prototype en fabrication

afin de valider expérimentalement la SRAM. Une autre piste prometteuse est liée à certains designs envoyés en fabrication auprès du fondeur OKI (ceux-ci sont disponibles à l'annexe 5). Il s'agissait de vérifier expérimentalement les marges de bruit de la cellule sur un prototype. Celui-ci a été réalisé à l'aide de transistors intrinsèques. L'utilisation de ces transistors semble prometteuse dans le cadre de l'utilisation de la technologie ULP, permettant un courant I_{ON} supérieur tout en maintenant les pertes statiques à un niveau très faible. Une piste à suivre…

9. Bibliographie

[1] The International Technology Roadmap for Semiconductors, 2007 edition.

[2] N. Kim, K. Flautner, D. Blaauw, T. Mudge, "Circuit and microarchitectural techniques for reducing cache leakage power", *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 2, pp. 167-184, Feb. 2004.

[3] J. De Vos, D. Bol, D. Flandre, "Cellule SRAM 12 transistors à ultra faible courant de fuite", *Proc. Journée d'étude faible tension faible consommation (FTFT)*, 8ed, pp. 111-115, Mai 2008.

[4] D. Bol, J. De Vos, R. Ambroise, et al., "Building Ultra-Low-Power High-TemperatureDigital Circuits in Standard High-Performance SOI Technology", *Solid-State Electronics , 10p., In Press.*

[5] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, M. Bohr, "SRAM Design on 65-nm CMOS Technology With Dynamic Sleep Transistor for Leakage Reduction", *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 895-901, Apr. 2005.

[6] H. Qin, Y. Cao, D. Markovic et al., "Standby supply voltage minimization for deep sub-micron SRAM", *Microelectronics Journal*, vol.36, no. 1, pp. 789-800, Mar. 2005.

[7] B. H. Calhoun, A. P. Chandrakasan, "A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation", *IEEE Journal of Solid-State Circuits*, vol. 42, no. 3, pp. 680-688, Mar. 2007.

[8] D. Levacq, "Low Leakage SOI CMOS Circuits Based on the Ultra-Low Power Diode Concept", *thèse de doctorat, Université Catholique de Louvain*, 2006.

[9] M. Ken, "Digital Integrated Circuit Design", pp438-450, Oxford University Press, 2000.

[10] Y. Wang, H. J. Ahn, U. Bhattacharya et al., "Design in 65 nm Ultra-Low-Power CMOS Technology With Integrated Leakage Reduction for Mobile Applications", *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 172-179, Jan. 2008.

[11] J. Pille, C. Adams, T. Christensen et al., "Implementation of the Cell Broadband Engine™ in 65 nm SOI Technology Featuring Dual Power Supply SRAM Arrays Supporting 6 GHz at 1.3 V", *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 163-171, Jan. 2008.

[12] R. F. Hobson, "A new Single-Ended SRAM Cell With Write-Assist", *IEEE Transactions on very large scale integration (VLSI) systems*, vol. 15, no. 2, pp. 173-181 Feb 2007 .

[13] B. Bhaumik, P. Pradlan, G.S. Visweswaran, et al., "A Low Power 256 KB SRAM Design", *Proceedings of the 12th International Conference on VLSI Design*, pp. 67-70, Jan. 1999

[14] K. Itoh, K. Sasaki, Y. Nakagome, "Trends in Low-Power RAM Circuit Technologies", *Proc. IEEE* vol. 83, no. 4, pp. 524-543, Apr. 1995.

[15] B. S. Amrutur, M. A. Horowitz, " A Replica Technique for Wordline and Sense Control in Low-Power SRAM's", *IEEE Journal of Solid-State Circuits*, vol. 33, no. 8, pp. 1208-1219, Aug. 1998.

[16] K. W. Mai, T. Mori, B. S. Amrutur, R. Ho, B. Wilburn, M.A. Horowitz, I. Fukushi, T. Izawa, S. Mitarai, "Low-Power SRAM Design Using Half-Swing Pulse-Mode Techniques", *IEEE Journal of Solid-State Circuits*, vol. 33, no. 11, pp. 1659-1671, Nov. 1998.

[17] B.D. Yang, L.S. Kim, "A Low-Power SRAM Using Hierarchical Bit Line and Local Sense Amplifiers" , *IEEE Journal of Solid-State Circuits*, vol. 40, no. 6, pp. 1366-1376, Jun. 2005.

[18] H. Mizuno, T. Nagano, "Driving Source-line Cell Architecture for sub 1-V High Speed Low-Power Applications", *IEEE Journal of Solid-State Circuits*, vol. 31, no. 4, pp. 552-557, Apr. 1996.

[19] K. Kanda, H. Sadaaki, T. Sakurai, "90% Write Power-Saving SRAM Using Sense-Amplifying Memory Cell", *IEEE Journal of Solid-State Circuits*, vol. 39, no. 6, pp. 927-953, Jun. 2004.

[20] D Levacq, V. Dessard, D. Flandre, "Low Leakage SOI CMOS Static Memory Cell With Ultra-Low Power Diode", *IEEE Journal of Solid-State Circuits*, vol. 42, no. 3, pp. 689-702, Mar. 2007.

[21] H. Morimura, N. Shibita, "A Step-Down Boosted-Wordline Scheme for 1-V Battery-Operated fast SRAM's", *IEEE Journal of Solid-State Circuits*, vol. 33, no. 8, pp. 1220-1227, Aug. 1998.

[22] F. Hamzaoglu, Y. Ye, A. Keshavarzi et al., "Analysis of Dual-V_T SRAM Cells With Full-Swing Single-Ended Bit Line Sensing for On-Chip Cache", *IEEE Transactions on very large scale integration (VLSI) systems*, vol. 10, no. 2, pp. 91-95 Apr. 2002 .

[23] D. Bol, R. Ambroise, D. Flandre, J.D. Legat, "Channel Length Upsize for Robust and Compact Subthreshold SRAM", *Proc. Journée d'étude faible tension faible consommation (FTFT)*, 8ed, pp. 117-120, Mai 2008.

[24] K. Osada, Y. Saitoh, E. Ibe, K. Ishibashi, "16.7-fA/Cell Tunnel-Leakage-Suppressed 16-Mb SRAM for Handling Cosmic-Ray-Induced Multierrors", *IEEE Journal of Solid-State Circuits*, vol. 38, no. 11, pp. 1952-1957, Nov. 2003.

[25] S. Romanovsky, A. Achyuthan, S. Natarajan, W. Leung, "Leakage Reduction techniques in a 0.13um SRAM Cell", *VLSI Design, Proceedings. 17th International Conference on*, 2004.

[26] A.J. Bhavnagarwala, S.V. Kosonocky, et all, " A pico-joule class, 1 GHz, 32 KByte/spl times/64 b DSP SRAM with self reverse bias", *Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2003.

[27] N. Verma, A. P. Chandrakasan, "A 256-kb 65-nm 8T Sub-threshold SRAM Employing Sense-Amplifier Redundancy", *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 141-149, Jan. 2008.

[28] J. P. Kulkarni, K. Kim, K. Roy, "A 160 mV Robust Schmitt Trigger Based Subthreshold SRAM", *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2303-2312, OCT. 2007.

[29] J. Chen, L. T. Clark, T.H. Chen, , "An Ultra-Low-Power Memory With a Subthreshold Power Supply Voltage", *IEEE Journal of Solid-State Circuits*, vol. 41, no. 10, pp. 2344-2353, Oct. 2006.

[30] D. Bol et al., "Building Ultra-Low-Power Low-Frequency Digital Circuits with High-Speed Devices", *in Proc. of the 14th IEEE International Conference on Circuits, Electronics and Systems, (ICECS'07)*, Marrakech, Dec. 2007.

[31] E. Vittoz, J. Fellrath, "CMOS analog integrated circuits based on weak inversion operation", *IEEE journal of solid state circuits*, vol. SC12 (3), pp224-231, 1977.

[32] J. Krupar, R. Srowik, J. Schreiter, et all, "Minimizing charge injection errors in high-precision, high-speed SC-circuits", *Symp. On Circuits and Systems ISCAS 2001*, vol. 1, pp. 727-730, May 2001.

[33] J.Lohstroh, E. Seevinck, J. DE GROOT, "Worst-case Static Noise Margin Criteria for Logic Circuits and their Mathematical Equivalence", *IEEE journal of solid state circuits*, vol. SC18 (6), pp803-807, 1983.

[34] E. Seevinck, F. J. List, J.Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells", *IEEE journal of solid state circuits*, vol. SC22 (5), pp748-754, 1987.

[35] Thomas, O., Amara, A, "Ultra low voltage design considerations of SOI SRAM memory cells", *Circuits and Systems*, pp: 4094 – 4097, Vol. 4, 2005

[36] D. Bol, D. Flandre, "ULP diode and latch in 45nm technology", technical report

[37] A. Asenov, A.R. Brown, J.H. Davies, S. Kaya and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs", *in IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1837-1852, Sep. 2003.

[38] B. Zhai, S. Hanson, D. Blauw and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", Proc. IEEE/ACM Int. Symp.Low-Power Electron. Des., 2005, pp. 20-25.

[39] P. Friedberg, Y. Cao, J. Cain, et all, "Modeling Within-Die Spatial Correlation Effets for Process-Design Co-Optimization", *Symp. On Quality of Electronic Design, ISQED*, 6p., 2005.

[40] W. Zhao, Y. Cao, "New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration", *Symp. On Quality of Electronic Design, ISQED*, 6p., Mar 2006.

[41] J. Saint-Martin, "Etude par simulation Monte Carlo d'architectures de MOSFET ultracourts à grille multiple sur SOI", *thèse de doctorat*, Université Paris XI Orsay, 2005.

[42] Michael Stockinger, “Optimization of Ultra-Low-Power CMOS Transistors”, *thèse de doctorat,* Université de Wien, 2000 (site internet “http://www.iue.tuwien.ac.at/phd/stockinger/index.php”, visité le 9/05/2008).

[43] S. Thompson, P. Packan, M. Bohr, “MOS Scaling: Transistor Challenges for the 21st Century”, *Intel Technology Journal*, 1998.

[44] Intel, site officiel, “http://www.intel.com/corporate/techtrends/emea/fra/45nm/index.htm ” et “ http://www.intel.com/technology/architecture-silicon/45nm-core2/demo/index.htm?iid=tech _arch_45nm+body_demo”, visités le 9/05/2008.

10. Annexe 1 : dépendance des modèles BSIM3 utilisés à la température

Les figures ci-dessous donnent l'évolution des courants de fuite des transistors du modèle BSIM 3 utilisé en fonction de la température. Remarquons que les courants de fuite augmentent exponentiellement avec la température alors que les courants ON diminuent. Cela entraine donc une dégradation du rapport I_{ON}/I_{OFF} pour des températures élevées.

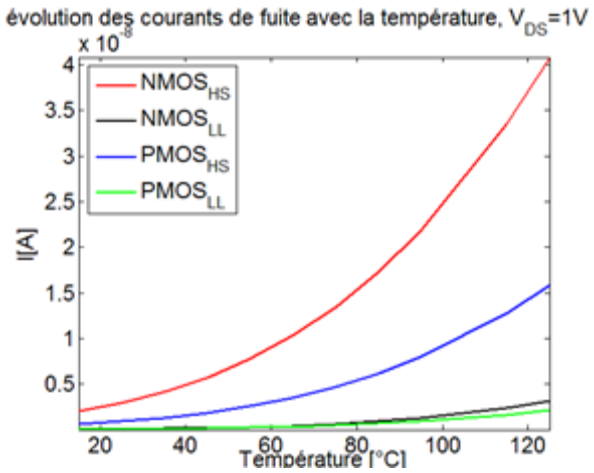


Fig. 10.1 : Evolution des courants de fuite des transistors HS et LL pour V_D=1V, V_S=V_B=0V, W=1µm, L=0.13µm.

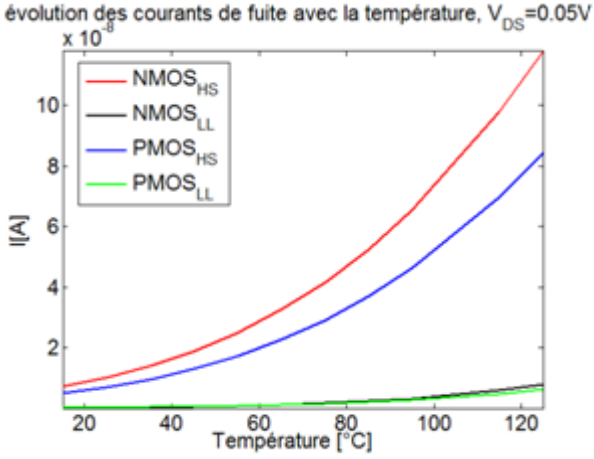


Fig. 10.2: Evolution des courants de fuite des transistors HS et LL pour V_D=0.05V, V_S=V_B=0V, W=1µm, L=0.13µm.

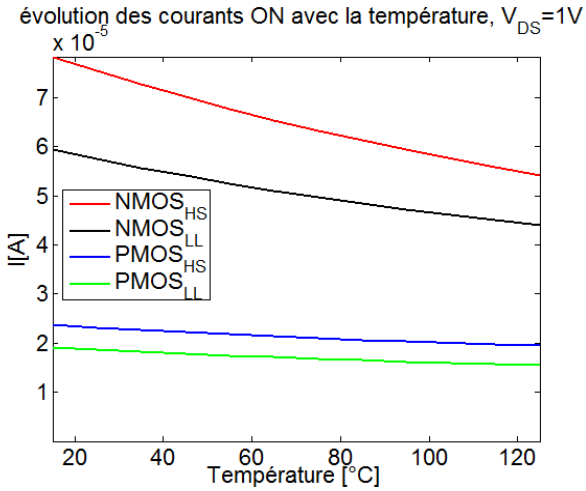


Fig. 10.3: Evolution des courants ON des transistors HS et LL pour V_D=1V, V_S=V_B=0V, W=1µm, L=0.13µm.

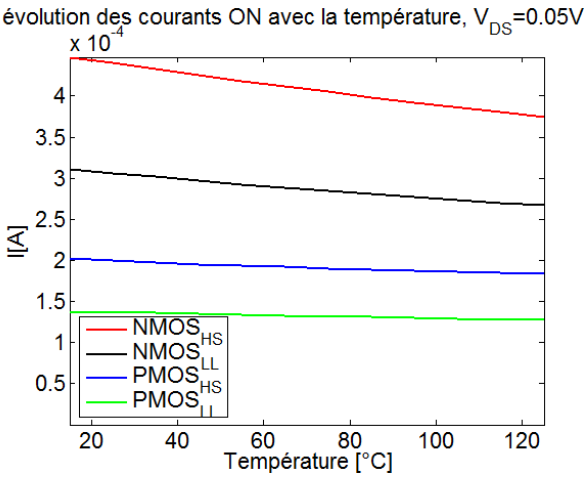


Fig. 10.4: Evolution des courants ON des transistors HS et LL pour V_D=0.05V, V_S=V_B=0V, W=1µm, L=0.13µm.

11. Annexe 2 : Simulation de cellules SRAM

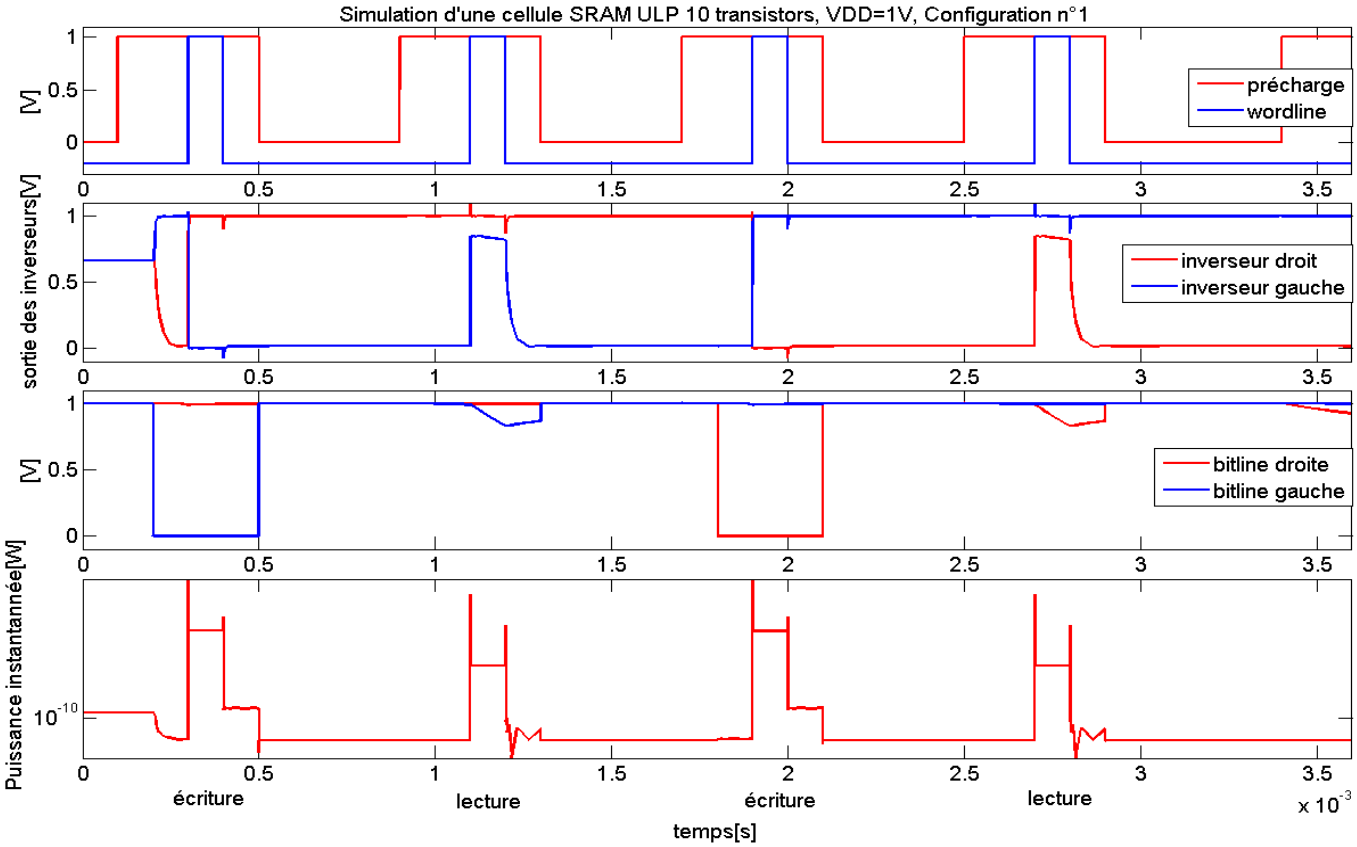


Fig 11.1: Simulation ELDO d'une cellule SRAM ULP 10 transistors pour la configuration n°1.

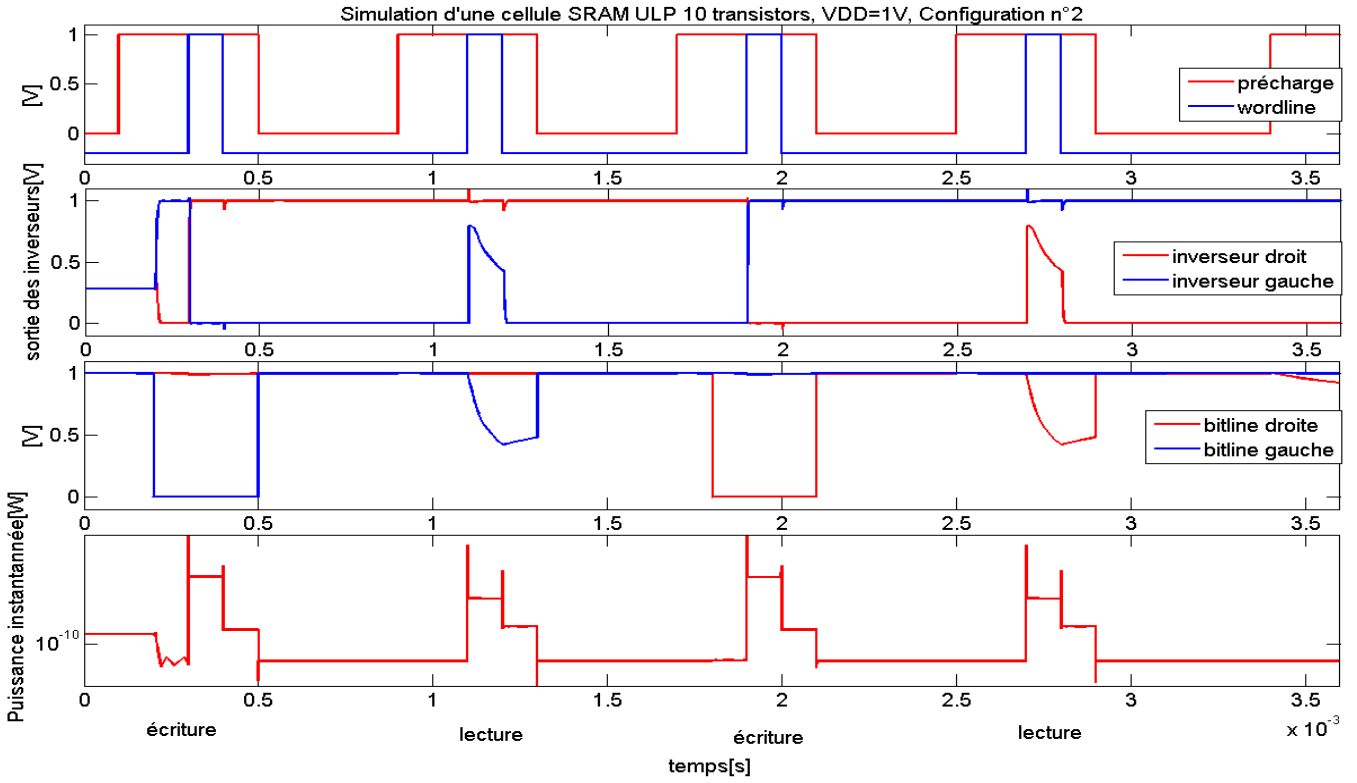


Fig 11.2 : Simulation ELDO d'une cellule SRAM ULP 10 transistors pour la configuration n°2.

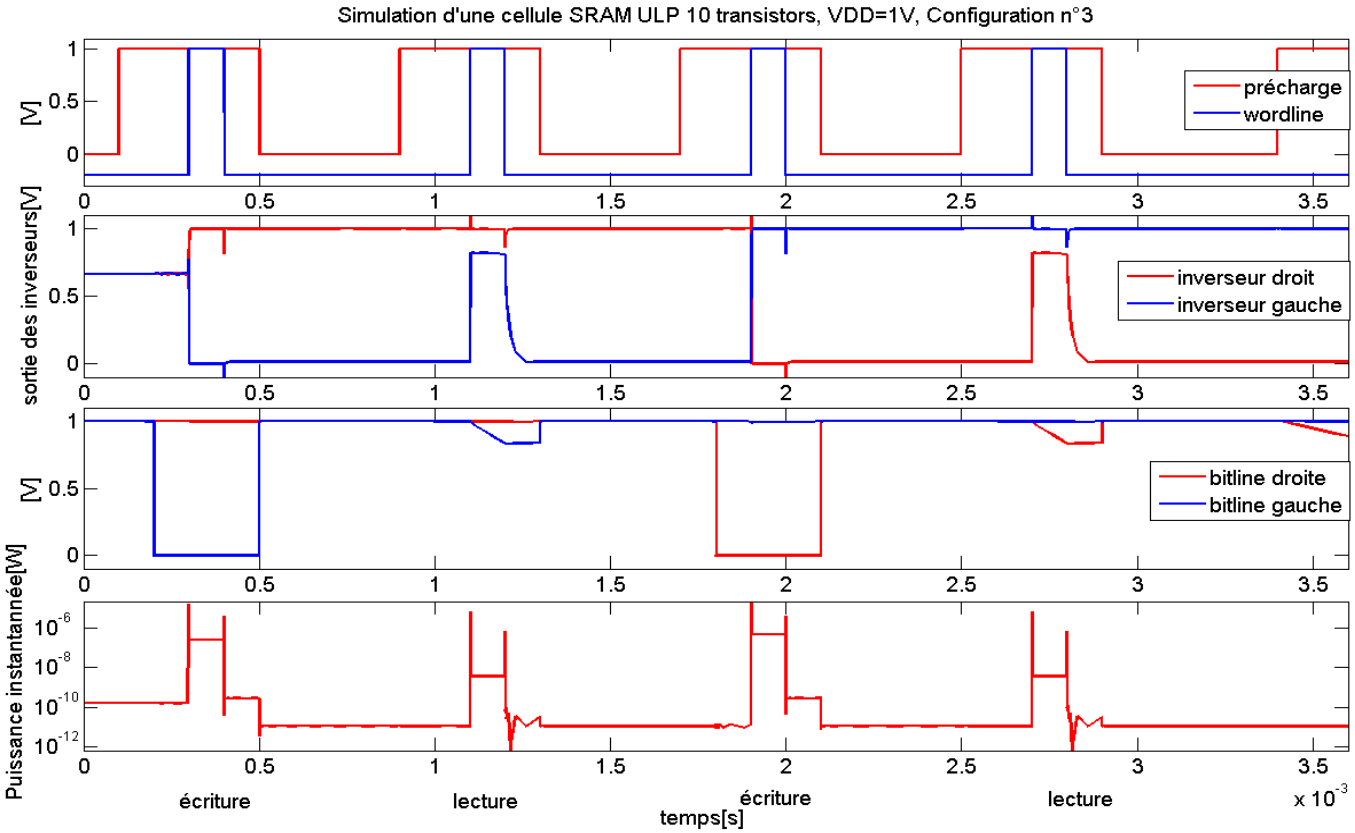


Fig 11.3 : Simulation ELDO d'une cellule SRAM ULP 10 transistors pour la configuration n°3.

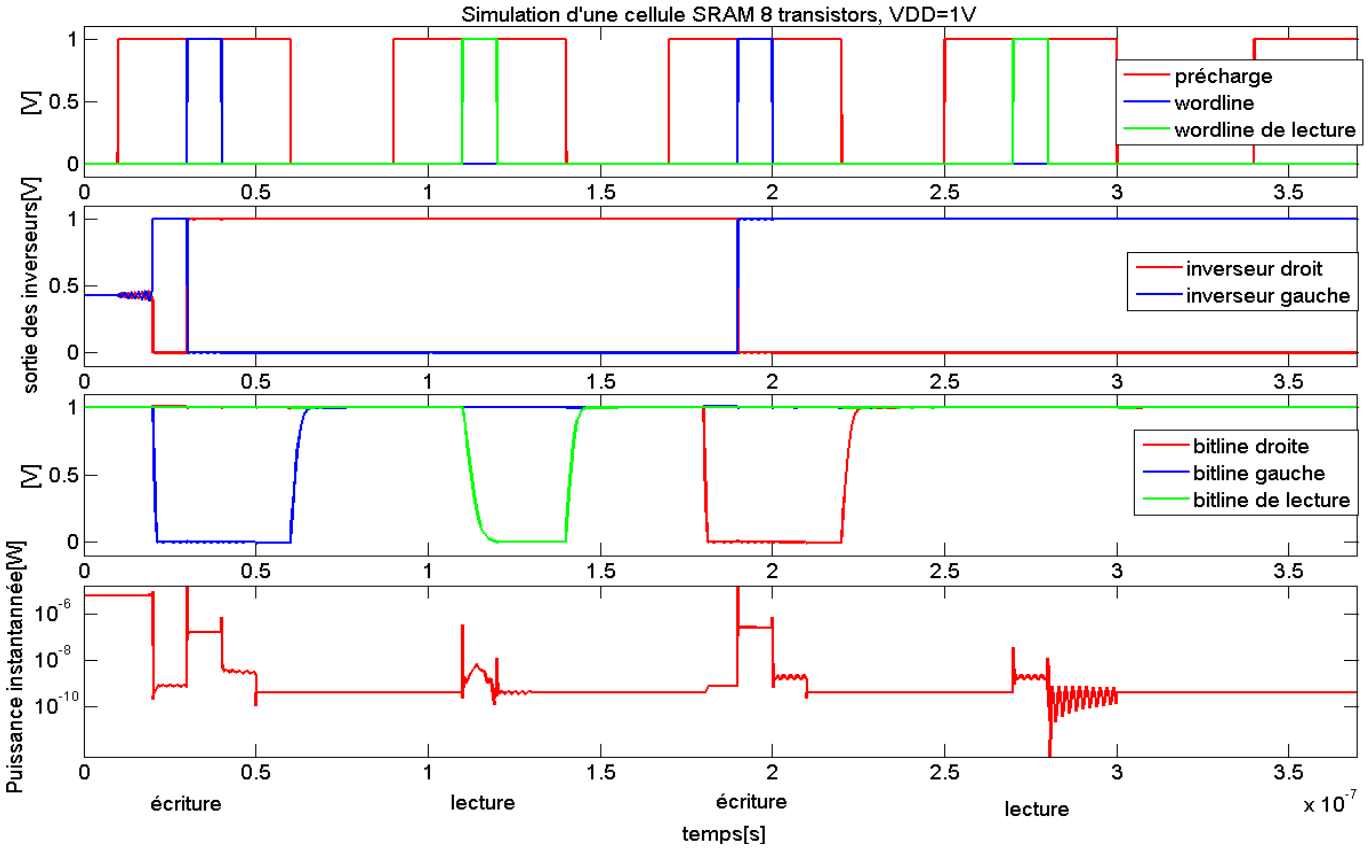
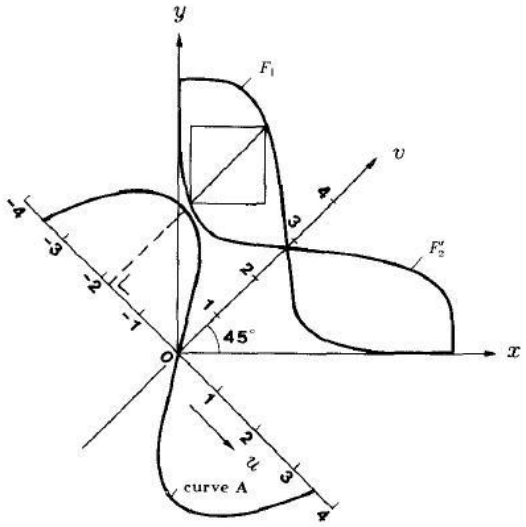


Fig. 11.4: Simulation ELDO d'une cellule SRAM 8 transistors. V_{DD}=1. Deux opérations d'écritures et deux opérations de lectures sont alternativement réalisées.

12. Annexe 3: le calcul de la marge de bruit

Une méthode pour déterminer le plus grand carré inscriptible dans le graphe en papillon est proposée dans [33]. Elle consiste à effectuer une rotation de 45 ° du graphe et de calculer la longueur des diagonales des carrés inscriptibles en faisant tout simplement la différence des ordonnées des deux courbes ainsi obtenues (fig. 12.1).



SNM estimation based on "maximum squares" in a 45° rotated coordinate system.

Fig. 12.1 : méthode de calcul de la SNM, tiré de [33].

Cette technique a été appliquée aux résultats de la simulation. Toutefois, la courbe caractéristique de l'inverseur ULP est très abrupte lors du basculement. Ainsi, il existe un réel trou dans les données lors du changement de repère qui ne peut être évité que par un échantillonnage excessif lors des simulations. La solution apportée à ce problème a été d'implémenter un petit algorithme « bouche-trou » qui va compléter les « données manquantes » avant la rotation. Une mesure de l'erreur d'estimation peut également être déduite. Il s'agit d'une courbe proportionnelle à la différence d'abscisse des deux ordonnées permettant de calculer une diagonale. Il s'agit évidemment d'ajuster la précision de l'algorithme « bouche-trou » de manière à limiter cette erreur lors du calcul de la diagonale du plus grand carré inscriptible. La figure 12.2 donne un exemple de l'action de ces différents algorithmes.

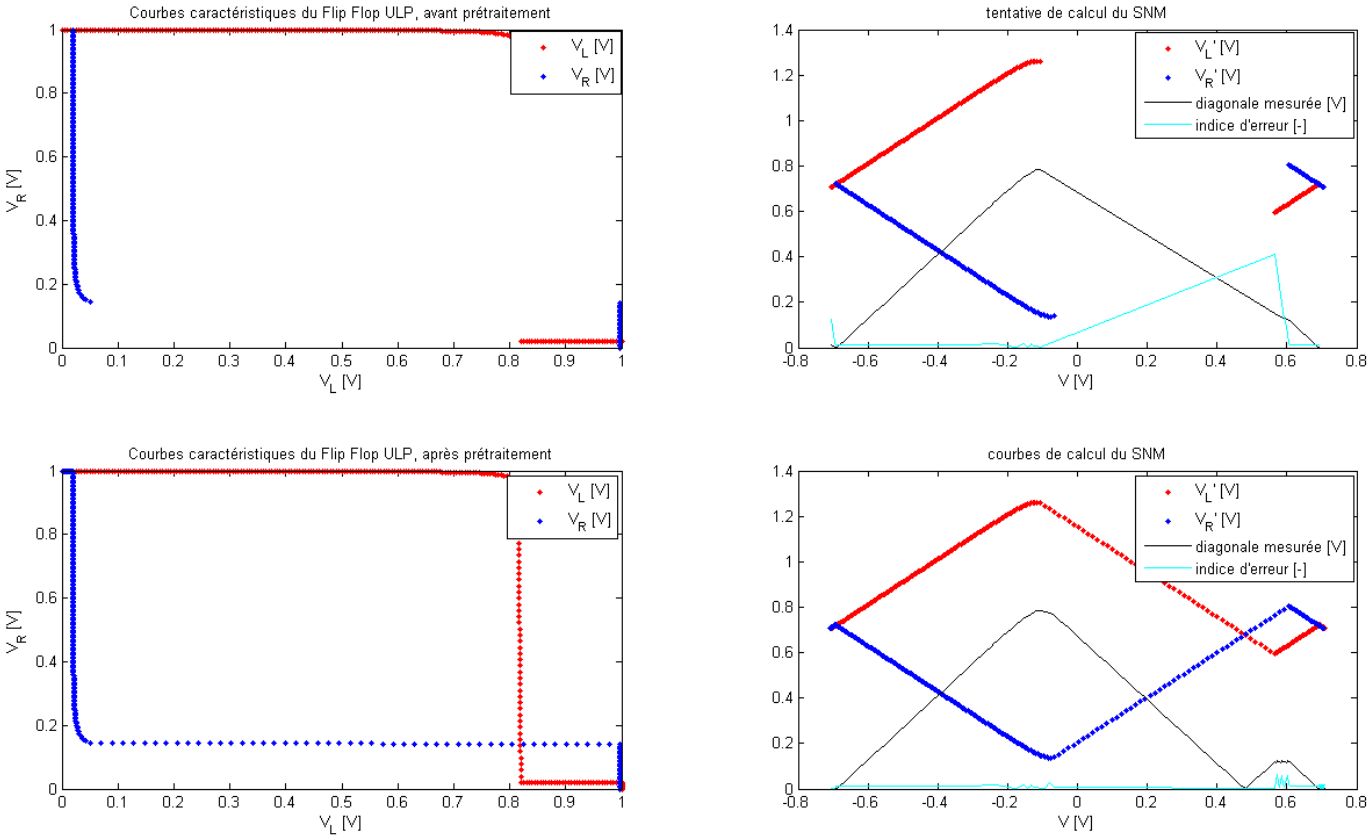


Fig. 12.2 : résultat de l'algorithme de calcul de la SNM. En haut à gauche, données de la simulation, en haut à droite tentative du calcul de la SNM sans prétraitement. En bas à gauche données traitées et en bas à droite, calcul de la SNM avec le traitement de données. Exemple pour le cas d'une cellule ULP en rétention pour W P1=0.7µm, L N3=0.65µm.

13. Annexe 4 : complément de probabilité

La loi de distribution qui permet le mieux de modéliser les valeurs obtenues en rétention est une loi de Weibull. Cependant, celle-ci surestime la probabilité d'occurrence pour les valeurs extrêmes, comme illustré à la figure 13.1. Il est possible qu'il ne s'agisse pas d'une surestimation, mais tout simplement que le nombre de données reçues pour ces probabilités d'occurrence ne soit pas suffisant pour être représentatif. Toutefois, il ne faut pas oublier que les modèles utilisés ne tiennent pas compte de la corrélation existante entre les paramètres de deux transistors très proches. Pourtant, une telle corrélation a été démontrée [39].

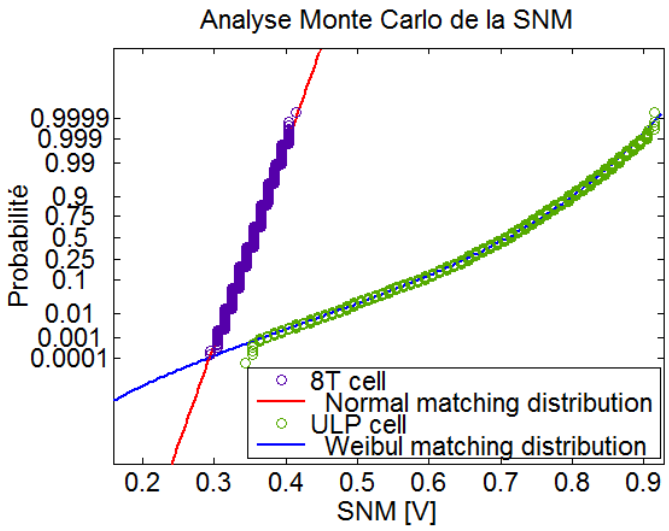


Fig. 13.1 : Simulations de Monte-Carlo (10k) des marges de bruits des cellules 8 transistors et 12 transistors ULP. Graphe des probabilités d'occurrence. σ_{VT}=34mV.

Il peut sembler étrange, alors que la tension de seuil des transistors suit une loi normale, que la marge de bruit de la cellule suive une loi de Weibull. Pour comprendre ceci il faut se souvenir que la cellule ULP 12 transistors fonctionne en régime sous le seuil. A ce titre, les courants délivrés par les transistors varient exponentiellement avec la tension de seuil. Comme ces courants influencent fortement la marge de bruit, on peut comprendre que celle-ci ne suive plus une loi normale.

Les paramètres des lois normales et de Weibull sont donnés ci-dessous à titre indicatif.

- Pour la loi de Weibull, la densité de probabilité est donnée par :

$$f(x;k,\lambda) = \left(k/\lambda\right)\left(x/\lambda\right)^{(k-1)} e^{-(x/\lambda)^k}.$$

Où k et λ sont des paramètres de la loi de distribution. Ceux-ci doivent être strictement positifs. Notons la présence d'un terme dépendant exponentiellement de x. L'utilisation de cette loi est donc bien justifiée de par la dépendance exponentielle des courants par rapport à la tension de seuil des transistors en régime sous le seuil.

L'espérance et la variance sont données par :

$$\mu = \lambda\Gamma\left(1+\frac{1}{k}\right), \quad \sigma^2 = \lambda^2\Gamma\left(1+\frac{2}{k}\right) - \mu^2.$$

- Pour la loi normale, la densité de probabilité est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5(\frac{x-\mu}{\sigma})^2}.$$

14. Annexe 5 : design OKI

Des structures permettant de tester les performances de la cellule ULP 12 transistors ont été envoyées en fabrication. Celles-ci devaient permettre de réaliser des mesures concernant les pertes statiques de la cellule, sa marge de bruit en écriture et en lecture. Les transistors utilisés pour réaliser la cellule sont de type intrinsèque. Ceci offre certains avantages qu'un travail ultérieur pourrait tenter de quantifier et de mettre à profit pour optimiser la cellule ULP 12 transistors. Par exemple, le rapport I_{ON}/I_{OFF} des transistors ULP réalisés à l'aide de tels éléments est très élevé.

Tout d'abord, voici une description des transistors utilisés pour réaliser la cellule. Les transistors d'accès sont toujours réalisés à l'aide d'éléments « low leakage » tandis que les inverseurs ULP mettent à profit les avantages des transistors intrinsèques. Il s'agit d'une technologie SOI 150 nm.

- Voici les caractéristiques du modèle « LL » :

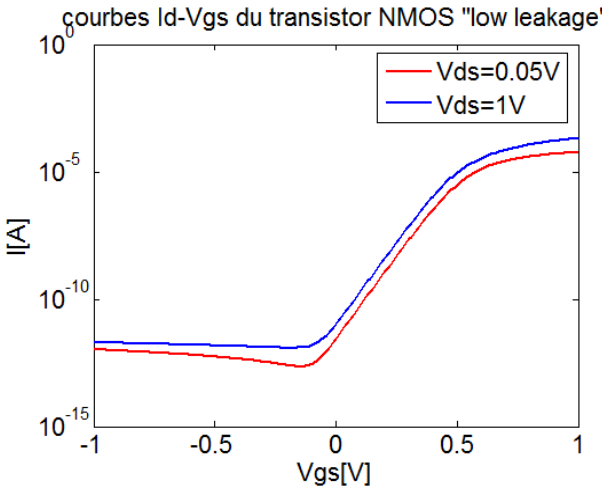


Fig. 14.1 : Courbe I_D-V_{GS} du NMOS LL pour V_D=0.05V et 1V. V_S=V_B=0V, W=1µm, L=0.15µm, T=25°C.

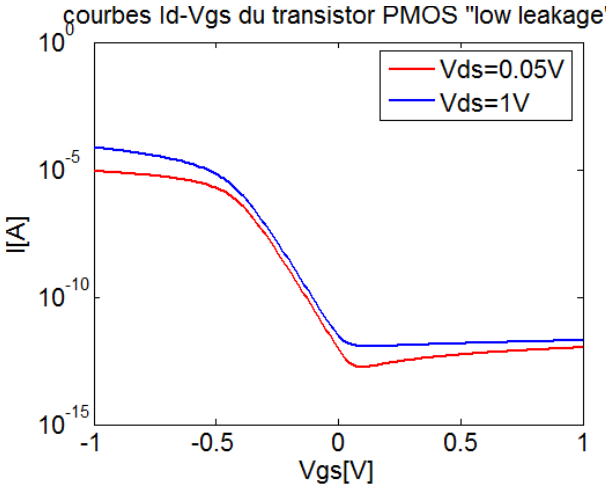


Fig 14.2 : Courbe I_D-V_{GS} du PMOS LL pour V_D=-0.05V et -1V. V_S=V_B=0V, W=1µm, L=0.15µm, T=25°C.

$\|V_{DS}\|$ [V]	I_{ON} NMOS [A/µm]	I_{ON} PMOS [A/µm]	I_{OFF} NMOS [A/µm]	I_{OFF} PMOS [A/µm]
1	2.1E-4	2.9E-4	8.0E-12	3.1E-12
0.05	5.9E-5	8.5E-6	2.1E-12	9.3E-13

Tableau 14.1 : courants I_{ON} et I_{OFF} des transistors low leakage. Le courant est donné pour un $\|V_{GS}\|$=1 ou 0V et pour $\|V_{DS}\|$ = 1V et 0.05V, T=25°C.

- Le modèle intrinsèque possède quant à lui les caractéristiques suivantes :

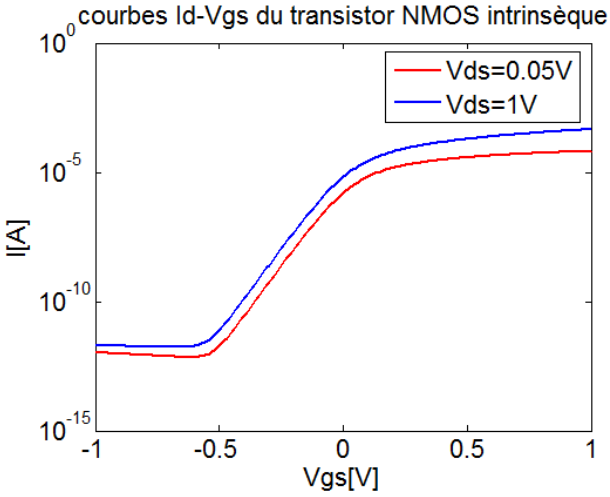


Fig. 14.3 : Courbe I_D-V_{GS} du NMOS intrinsèque pour V_D=0.05V et 1V. V_S=V_B=0V, W=1µm, L=0.15µm, T=25°C.

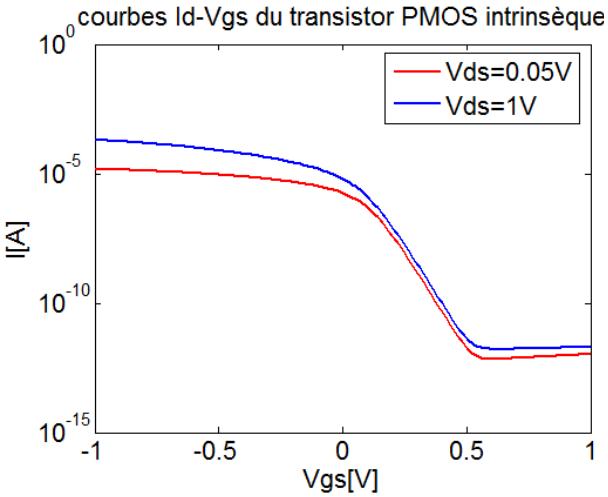


Fig 14.4 : Courbe I_D-V_{GS} du PMOS intrinsèque pour V_D=-0.05V et -1V. V_S=V_B=0V, W=1µm, L=0.15µm, T=25°C.

$\|V_{DS}\|$ [V]	I_{ON} NMOS [A/µm]	I_{ON} PMOS [A/µm]	I_{OFF} NMOS [A/µm] (V_{GS}=0V)	I_{OFF} PMOS [A/µm] (V_{GS}=0V)	I_{OFF} NMOS [A/µm] (V_{GS}=-0.5V)	I_{OFF} PMOS [A/µm] (V_{GS}=0.5V)
1	4.6E-4	2.0E-4	6.7E-6	6.4E-6	7.8E-12	3.1E-12
0.05	6.6E-5	1.5E-5	1.6E-6	1.8E-6	1.9E-12	1.9E-12

Tableau 14.2 : courants I_{ON} et I_{OFF} des transistors intrinsèques. Le courant est donné pour un $\|V_{GS}\|$=1 ou 0V et pour $\|V_{DS}\|$ = 1V et 0.05V, T=25°C.

La courbe I_D-V_{GS} du transistor ULP réalisé à l'aide de transistors intrinsèques pour une tension V_{DS} de 1V est présentée à la figure 14.5. La courbe obtenue pour un NMOS LL est ajoutée pour la comparaison. Nous pouvons remarquer que le rapport entre I_{ON} et I_{OFF} est proche de 10^7. Celui-ci est donc bien plus élevé que ce que nous obtenions précédemment en technologie bulk 130 nm (ce rapport s'élevait environ à 10^4).

Différents layouts ont été réalisés. Ceux-ci sont représentés aux figures 14.6 à 14.9. Ils permettent d'évaluer le bon fonctionnement de la cellule en lecture et en écriture, ainsi que ses pertes statiques (figure 14.7), la marge de bruit en rétention et en lecture (fig 14.6), ou en écriture (fig. 14.8). Enfin, une vue d'ensemble avec les pads d'accès est donnée (fig. 14.9). Afin de limiter la surface de silicium utilisée, lorsque cela est possible, les pads d'accès servent à plusieurs structures de tests différentes (par exemple, le même pad fournit la tension de masse de toutes les cellules).

Pour l'évaluation des marges de bruit, afin de ne pas rendre les structures de tests trop sensibles aux variations de procédé, la largeur des transistors a été augmentée à 1 µm. La longueur de grille reste minimale (0.15µm).

En ce qui concerne la structure de test permettant de déterminer la consommation statique de la cellule, les transistors sont tous de taille minimale. Malheureusement, une erreur s'est glissée dans ce layout. Dès lors, la tension de wordline de lecture ne peut être appliquée correctement. Toutefois, il

est toujours possible de laisser le circuit de lecture flottant et d'utiliser la cellule comme s'il s'agissait d'une cellule ULP 10 transistors.

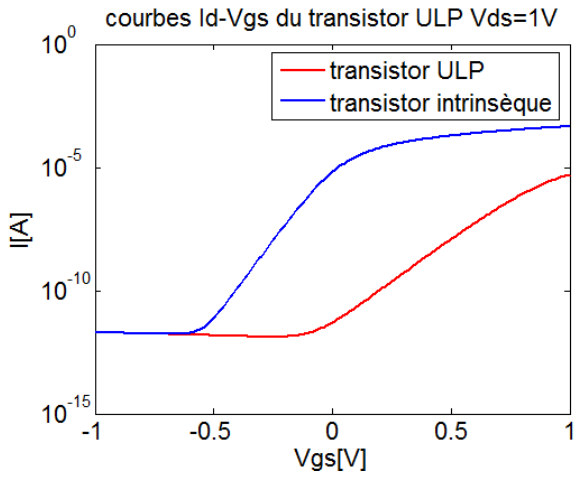


Fig. 14.5 : Courbe I_D-V_{GS} du transistor ULP de type N et du NMOS LL pour V_D= 1V. V_S=V_B=0V, W_N=1µm, W_P=2.4µm, L=0.15µm, T=25°C.

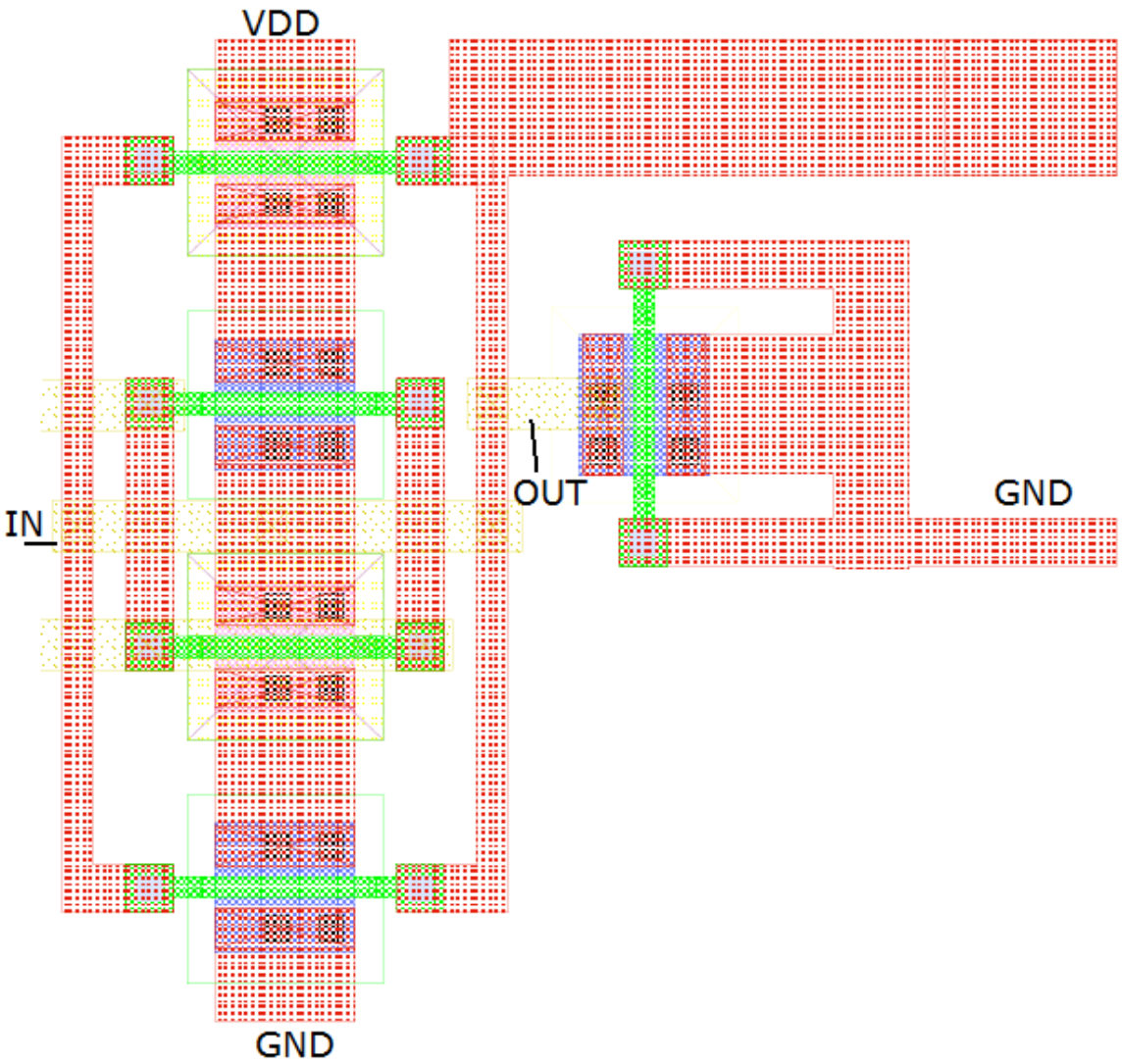


Fig. 14.7 : Layout de la structure de test permettant d'évaluer la SNM en lecture et rétention.

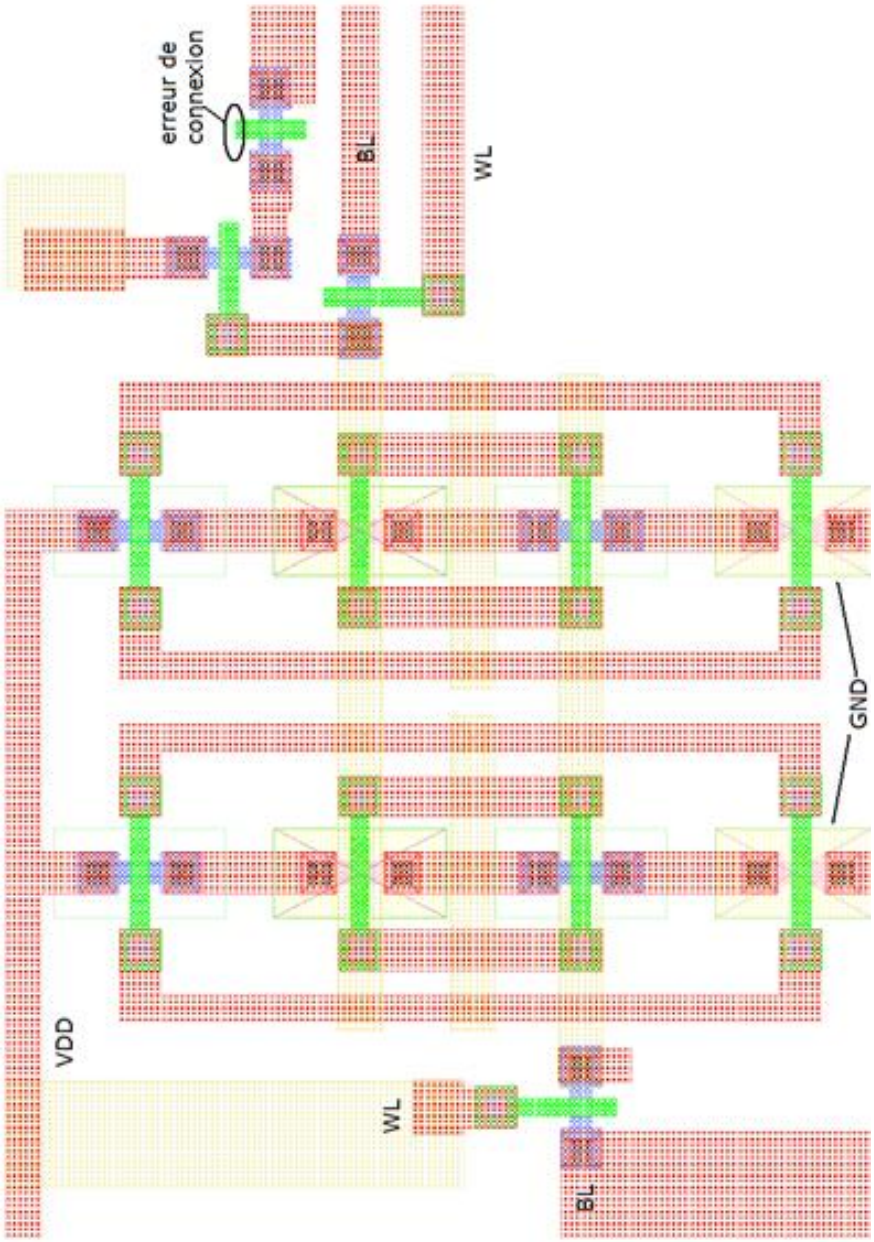


Fig. 14.6 : Layout de la structure de test permettant d'évaluer les pertes statiques de la cellule ainsi que de réaliser des opérations de lecture et d'écriture. Une erreur de connexion rend le buffer de lecture inopérant. La cellule peut être utilisée comme une cellule ULP 10 transistors.

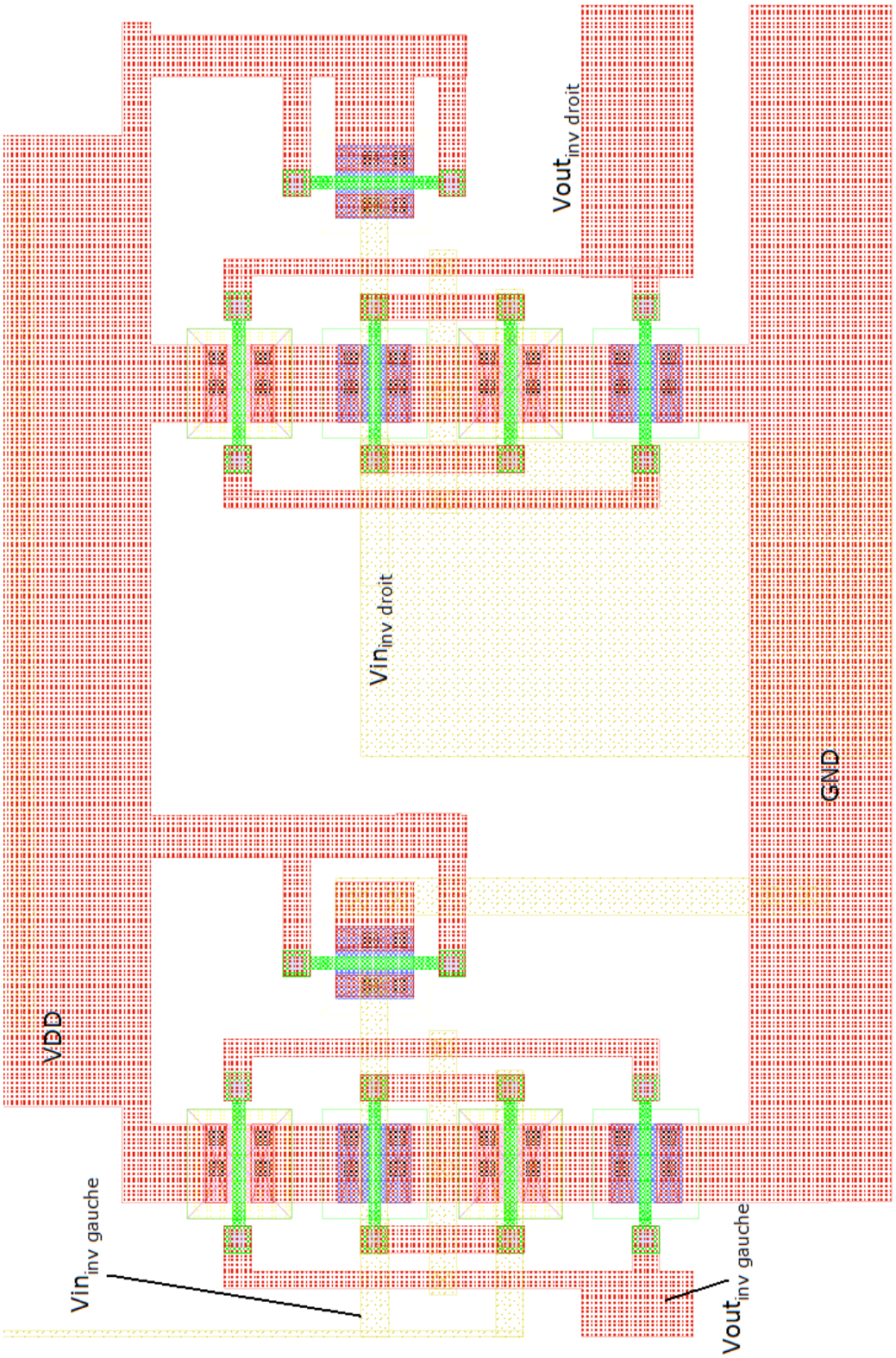


Fig. 14.8 : Layout de la structure de test permettant d'évaluer la SNM en écriture. Des sources de tension externes doivent relier les deux inverseurs ensemble en prenant le rôle des sources de bruit.

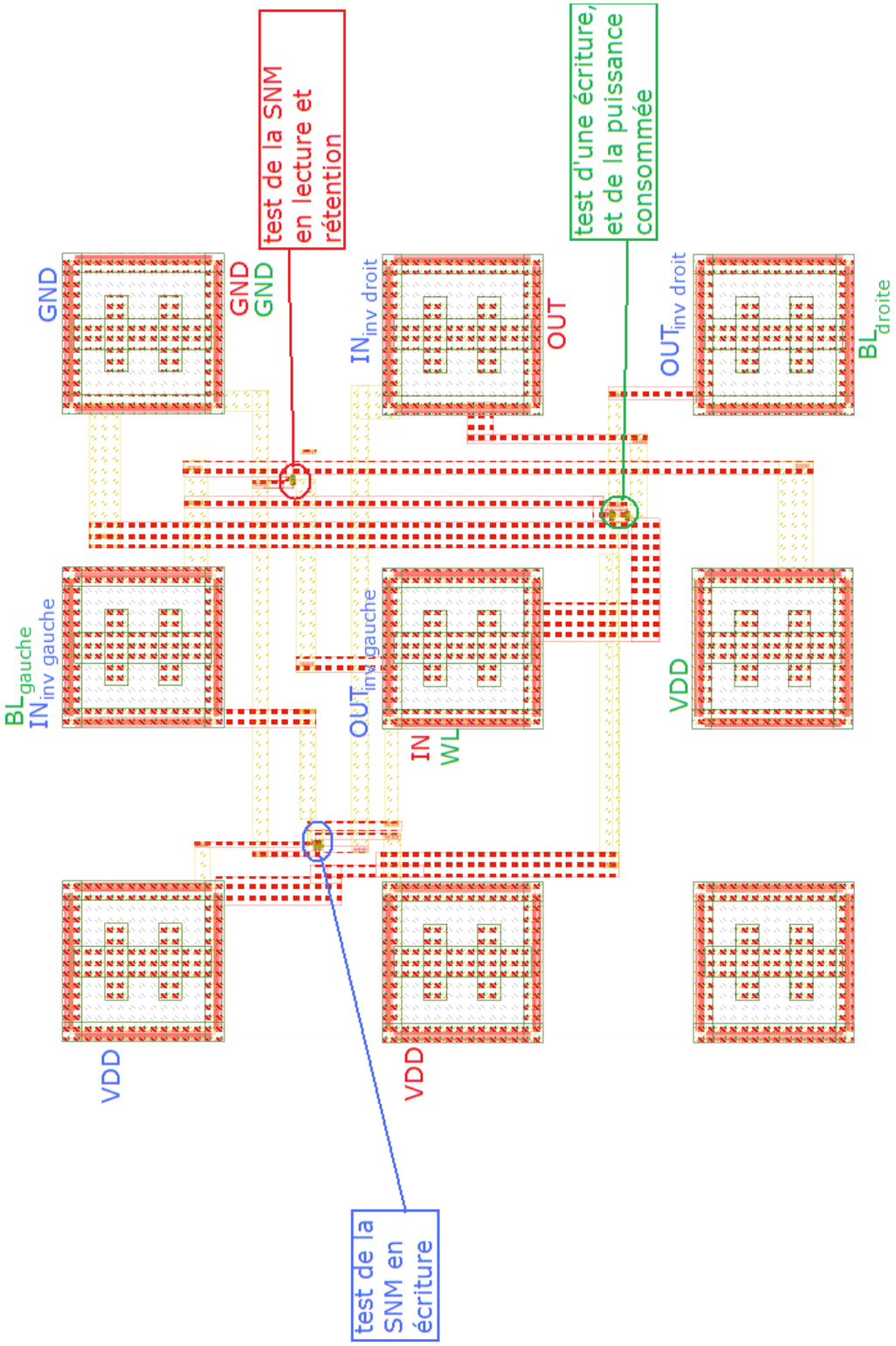


Fig. 14.9 : Vue d'ensemble des structures de tests, les tensions à appliquer aux différents pads sont indiquées.

15. Annexe 6 : publications

Cellule SRAM 12 transistors à ultra faible courant de fuite [3]

Julien De Vos, David Bol, Denis Flandre

Présentation réalisée dans le cadre des 7emes journées faible tension, faible consommation.

Building Ultra-Low-Power High-Temperature Digital Circuits in Low-Voltage Deep Submicron SOI Technology [4]

David Bol, Julien De Vos, Denis Flandre and Jean-Didier Legat

Abstract

For ultra-low-power applications, digital integrated circuits may operate at low frequency to reduce dynamic power consumption. At high temperature, the power consumption of such circuits is completely dominated by static power dissipation due to leakage current. In this contribution, we propose a new logic style, namely Ultra-Low-Power (ULP) logic style which achieves negative Vgs self-biasing, to benefit from the small area and low dynamic power of low-voltage deep-submicron SOI technologies while keeping ultra-low leakage, even at high temperature. In 0.13-µm partially-depleted SOI technology, the static power consumption at 200◦C is reduced by nearly 3 orders of magnitude at the expense of increased delay.