Université catholique de Louvain - École polytechnique de Louvain

Étude et comparaison de cellules SRAM *dual-port* sans *assist* dynamique

Jury

Pr. Denis Flandre (promoteur)
Dr. David Bol (assistant encadrant)
Pr. Jean-Didier Legat
Julien DeVos

Mémoire présenté en vue de l'obtention du grade d'ingénieur civil électricien par

Sébastien BERNARD

Juin 2011

Je tiens à remercier mon promoteur, le Professeur Denis Flandre, pour m'avoir donné l'occasion de mener mes recherches dans son département, ainsi que mon assistant encadrant, le Docteur David Bol, pour son aide efficace et sa très grande disponibilité tout au long de l'année et ses conseils éclairés et sa patience durant la correction de ses *libelli*. Je tiens également à remercier M. Julien DeVos et M. Guillaume Polissard pour leurs explications nombreuses et patientes du logiciel Eldo, ainsi qu'Olivier de Caritat pour ses astuces Eldo et pour les nombreux *Questions pour un champion* que nous avons gagnés ensemble. Enfin, je voudrais remercier mes parents pour leur soutien lors de la rédaction de ce mémoire, ainsi que toutes les personnes qui ont lu et corrigé ce texte afin d'améliorer l'orthographe et le style.

Sébastien

Résumé

Une mémoire SRAM à grande vitesse et faible consommation est un élément essentiel de la technologie des circuits intégrés. Par ailleurs, une architecture *dual-port* permet d'améliorer les performances totales d'un circuit numérique. Dans ce travail, des architectures de cellules SRAM *dual-port* sont étudiées afin de les comparer et d'analyser des pistes d'amélioration des performances de l'état de l'art, principalement en terme de consommation et de surface.

Le tableau mémoire contenant les cellules devra atteindre une fréquence de fonctionnement de 1GHz, sous une tension d'alimentation d'1V, avec des transistors de la technologie 32nm FDSOI. Pour commencer l'étude, l'utilisation d'un *buffer* comme système de lecture est motivée pour l'amélioration des performances. Ensuite, plusieurs systèmes d'écriture et de lecture sont comparés selon leurs performances propres. Il en ressort que la cellule composée d'un *latch* appelé 5TPMOS avec un transistor d'accès en écriture SVt combiné avec un *buffer* de lecture à deux transistors de type HVt, présente les plus faibles consommation et surface de silicium tout en garantissant les contraintes de vitesse et de fiabilité. Cependant, une étude plus approfondie montre que cette cellule ne respecte plus la contrainte de vitesse pour une variation trop importante de température et de tension d'alimentation. Afin de résoudre cette limitation, des perspectives d'amélioration sont proposées et brièvement étudiées, notamment une diminution de la tension d'alimentation des cellules.

Abstract

High speed and low power SRAM is a crucial element for numeric circuits. In this connection, a dual-port memory architecture improves the total chip performance. This work studies and compares SRAM cells dedicated to dual-port application in order to compare them and to seek some ways of improvement of the state of the art .

The SRAM memory array will be able to work at 1GHz frequency under a 1V supply, with transistors from 32nm FDSOI technology. To start the study, the use of read buffers is motivated to improve performances. Then, several write and read systems are compared according to their own performances. In conclusion, it appears that the cell composed by a 5TPMOS latch with write access transistor SVt combined with a 2T read buffer presents the lower consumption and aera, and guarantees the constraints of speed and robustness. However, a deeper study shows that this cell does not respect a large variation of temperature and supply voltage. To solve this limitation, perspectives of improvement are proposed and briefly studied, in particular the decrease of cell supply voltage.

Table des matières

Introduction

La mémoire SRAM (*Static Random Acces Memory*) est une mémoire vive associée à un circuit électronique numérique. Elle est fabriquée en technologie CMOS standard, ce qui permet de l'incorporer aisément *on-chip* avec la logique, afin d'augmenter la vitesse d'accès et de réduire la consommation totale du circuit. Elle est la mémoire principale des systèmes embarqués, car elle peut aussi bien contenir les lignes de code et les données du programme, que remplir le rôle de mémoire cache de niveau un ou deux.

Cependant, les prévisions ITRS affirment que la mémoire SRAM est le principal frein à l'amélioration des performances des circuits, alors qu'elle occupera la plus grande partie de la surface des systèmes intégrés sur puce (SoC) et représentera leur principale source de consommation [1].

Une quantité croissante de mémoire embarquée est utilisée dans les circuits intégrés, et par conséquent, la conception de mémoire SRAM à haute densité et basse consommation devient un enjeu critique [2]. Ces dernières années, la demande pour des mémoires SRAM *multi-port* n'a cessé d'augmenter, car elles améliorent les performances globales des circuits [3]. En particulier, la mémoire *dual-port* présente deux ports d'accès, ce qui permet de lire et d'écrire simultanément dans la mémoire. Cette propriété est par exemple fort utile dans les processeurs pipelinés (architecture von Neumann) qui peuvent, grâce à elle, lire l'instruction (étage de FETCH) et écrire le résultat (étage de WRITE-BACK) sur un même cycle d'horloge. Elle est également un avantage dans des applications où plusieurs processeurs sont mis en série pour augmenter le débit, le traitement de signaux (vidéo) par exemple [4]. Dans ces applications, une liste de type FIFO (*First-In First-Out*) est souvent incorporée entre les processeurs pour stocker temporairement les données. Les processeurs écrivent et lisent dans cette FIFO quand ils ont fini de traiter leur donnée, indépendamment des processeurs voisins ; ceci permet de rendre le flot de données plus continu, et d'augmenter le débit moyen.

Dans ce travail, plusieurs cellules de mémoire SRAM seront étudiées en vue de s'intégrer dans une application *dual-port*. De cette étude seront proposées plusieurs pistes d'amélioration potentielles des performances par rapport à l'état de l'art. Au vu des prévisions ITRS, une attention particulière sera mise sur la diminution de la surface de silicium et de la consommation énergétique de la cellule mémoire, ceci dans le but d'augmenter l'autonomie des circuits - pour applications portables - et diminuer leur coût de fabrication et l'impact environnemental associé.

Les applications grand public demandent des fréquences de fonctionnement toujours plus élevées. Par exemple, les processeurs Atom d'Intel et Snapdragon de Qualcomm, présents dans de nombreux *smartphones* et des consoles de jeux portables, ont une fréquence d'horloge de 2GHz et 1.5GHz respectivement. Pour maintenir cette fréquence de fonctionnement avec une consommation relativement faible, des transistors très prometteurs ont été développés par le CEA-Leti à Grenoble. Ils proviennent de la technologie 32nm FDSOI (*Fully Depleted Silicon On Insulator*), dont les performances sont accrues par rapport aux technologies Bulk classiques.

Dans ce travail, le tableau mémoire SRAM *dual-port* devra atteindre une fréquence de fonctionnement de 1GHz, l'ordre de grandeur des applications actuelles, en utilisant les transistors 32nm FDSOI sous leur tension d'alimentation nominale de 1V.

Le premier chapitre explique plus précisément le contexte de ce travail, ainsi que l'état de l'art dans ce domaine. Il présente également des systèmes capables d'améliorer les performances d'une mémoire SRAM, les *assists* statiques. Ils seront un degré de dimensionnement supplémentaire lors de l'étude des cellules *dual-port*.

Le deuxième chapitre définit rigoureusement les critères servant à l'étude et à la comparaison des cellules SRAM *dual-port*.

Le troisième chapitre étudie ces cellules et compare leurs performances définies dans le chapitre précédent. La première section de ce chapitre motive la méthodologie poursuivie dans ce travail. En effet, l'étude du système d'écriture et du système de lecture se fera en deux parties : les systèmes d'écriture dans la deuxième section et les systèmes de lecture dans la troisième section. La dernière section de ce chapitre conclut et réunit les résultats de l'étude. En particulier, il sera montré que la cellule composée d'un *latch* appelé 5TPMOS avec un transistor d'accès en écriture SVt combiné avec un *buffer* de lecture à deux transistors de type HVt, permet d'améliorer les performances des cellules de l'état de l'art en terme de consommation et de surface de silicium, avec un prix à payer sur la vitesse d'écriture (+45%).

Le quatrième chapitre étudie plus en profondeur la cellule retenue dans le chapitre précédent, en particulier face aux variations de fabrication, de tension d'alimentation et de température (*Process Voltage Temperature* ou PVT). Il montre que la cellule proposée n'est pas robuste face aux écarts maximaux spécifiés. Toutes les contraintes environnementales doivent donc *in fine* être prises en compte dans la conception de cellules SRAM.

Le cinquième chapitre explore des perspectives d'amélioration des cellules *dual-port*, en terme de fréquence de fonctionnement et de consommation. En particulier, une piste d'amélioration pourrait également permettre de rendre la cellule proposée robuste face aux variations maximales : une diminution de la tension d'alimentation des cellules.

En résumé, nous avons proposé dans ce travail une méthodologie de comparaison de cellules SRAM *dual-port*, basée sur l'utilisation d'un *buffer* de lecture permettant de séparer l'étude en deux parties distinctes. Nous avons ensuite sélectionné les systèmes d'écriture et de lecture en minimisant la surface et la consommation totale de la cellule. Finalement, la cellule SRAM *dual-port* composée d'un *latch* 5TPMOS à transistor d'accès SVt et d'un *buffer* à deux transistors HVt implémentée dans la technologie 32nm FDSOI permet une fréquence de fonctionnement de 1GHz, consomme typiquement $45,7pW$ de puissance statique et une énergie moyenne par accès de $7.7fJ$, et requiert huit transistors de taille minimale.

Chapitre 1

Etat de l'art des cellules SRAM

Ce chapitre explicite le fonctionnement général d'une mémoire de type SRAM. Dans ce contexte, la première section définit aussi précisément dans quel cadre s'incrit ce travail. L'état de l'art des cellules classiques étant le plus développé, il sera présenté dans la deuxième section. La troisième section présente la cellule *dual-port* actuellement utilisée dans l'état de l'art.
La quatrième section présente et explique les systèmes périphériques, appelés *assists*, utilisés pour améliorer les performances des cellules SRAM. Les *assists* statiques constituent un degré de liberté de l'étude de ce travail.
Enfin, la cinquième section introduit et explique le concept de variabilité dans les circuits électroniques, extrêmement important dans le contexte des mémoires SRAM.

1.1 Fonctionnement d'une mémoire SRAM

Une mémoire SRAM se présente sous la forme d'un tableau avec m lignes et n colonnes (figure 1.1). Chaque ligne est composée de plusieurs cellules, contenant chacune une donnée ou *bit*. Une donnée est une valeur logique binaire, 0 ou 1, représentée physiquement dans la mémoire par la valeur des tensions des noeuds de la cellule. Un ensemble de données représente un mot. Le tableau contient donc tout un ensemble de mots. Le programme contenu dans le processeur interfacé avec la mémoire contient des instructions qui représentent l'écriture d'un mot, le résultat d'un calcul par exemple, ou la lecture d'un mot, pour réutiliser un résultat précédent par exemple. Ces mots sont écrits et lus dans le tableau SRAM.

Lorsque la mémoire n'est pas accédée, la valeur contenue dans chaque cellule ne doit pas varier, quels que soient les niveaux de tension des signaux extérieurs, sous peine de perdre l'information.

Lorsqu'on veut écrire et sauvegarder un mot dans la mémoire, l'adresse du mot est envoyée en entrée du décodeur de ligne et la valeur du mot (*entrée des données* figure 1.1) est envoyée dans le circuit de rafraichissement. Le décodeur a une et une seule sortie à tension haute. Cette sortie est l'entrée d'un *driver* servant à charger une longue ligne d'interconnexion, appelée une *WordLine* (WL). Lorsque la WordLine est à une tension haute, on dit qu'elle est activée. Le circuit de rafraichissement charge les longues lignes d'interconnexion parcourant les colonnes du tableau, appelées *BitLine* (BL), en fonction de la valeur des nouvelles données. Seules les cellules dont la *WordLine* est activée verront leur donnée modifiée par les *BitLines*. Les autres ne changent pas de valeur au cours d'une écriture.

Lorsqu'on veut lire un mot dans la mémoire, l'adresse du mot est à nouveau envoyée en entrée du décodeur de ligne. L'information de la valeur de la donnée se transmet sur les tensions

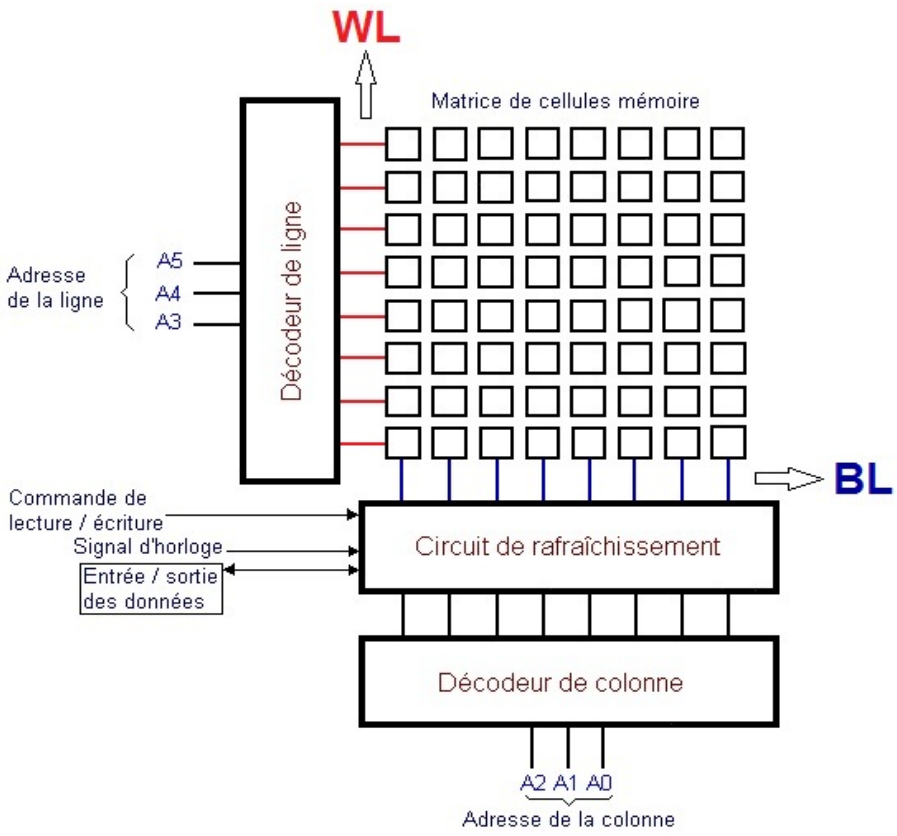


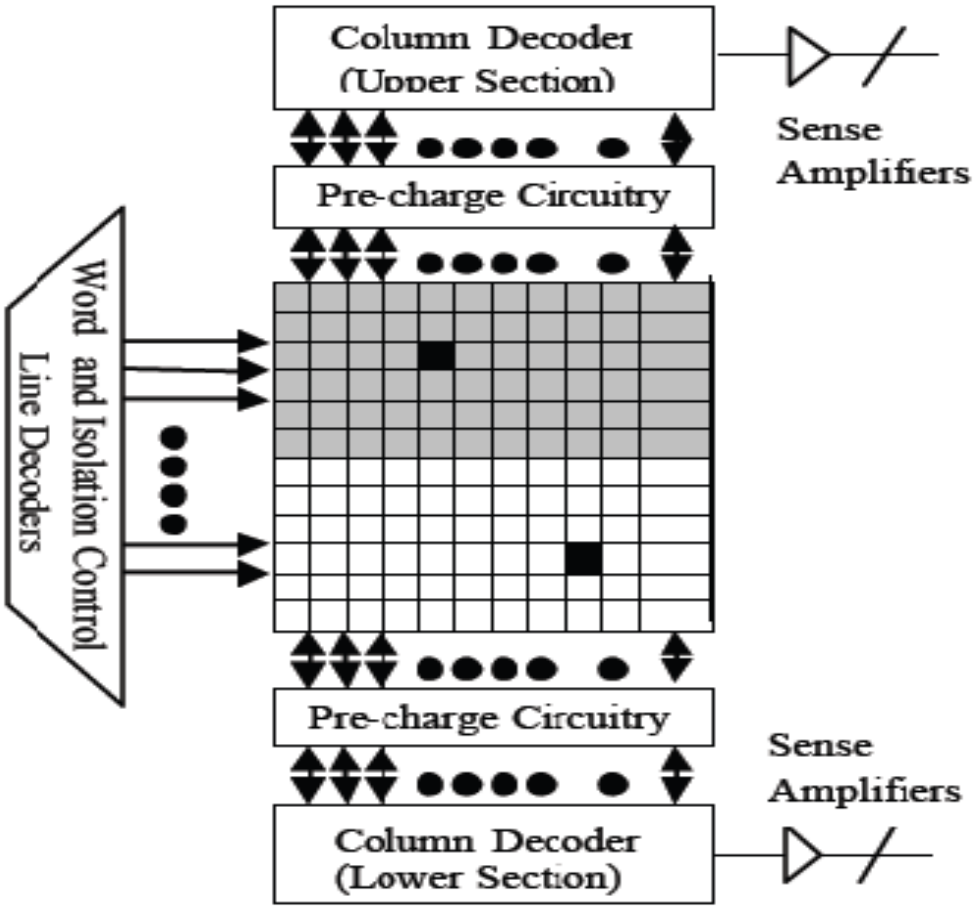
Figure 1.1 – Schéma général d'une mémoire SRAM [5].

des *BitLines*. Cette tension est ensuite récupérée par le circuit de rafraichissement qui renvoie l'information en sortie (*sortie des données*).

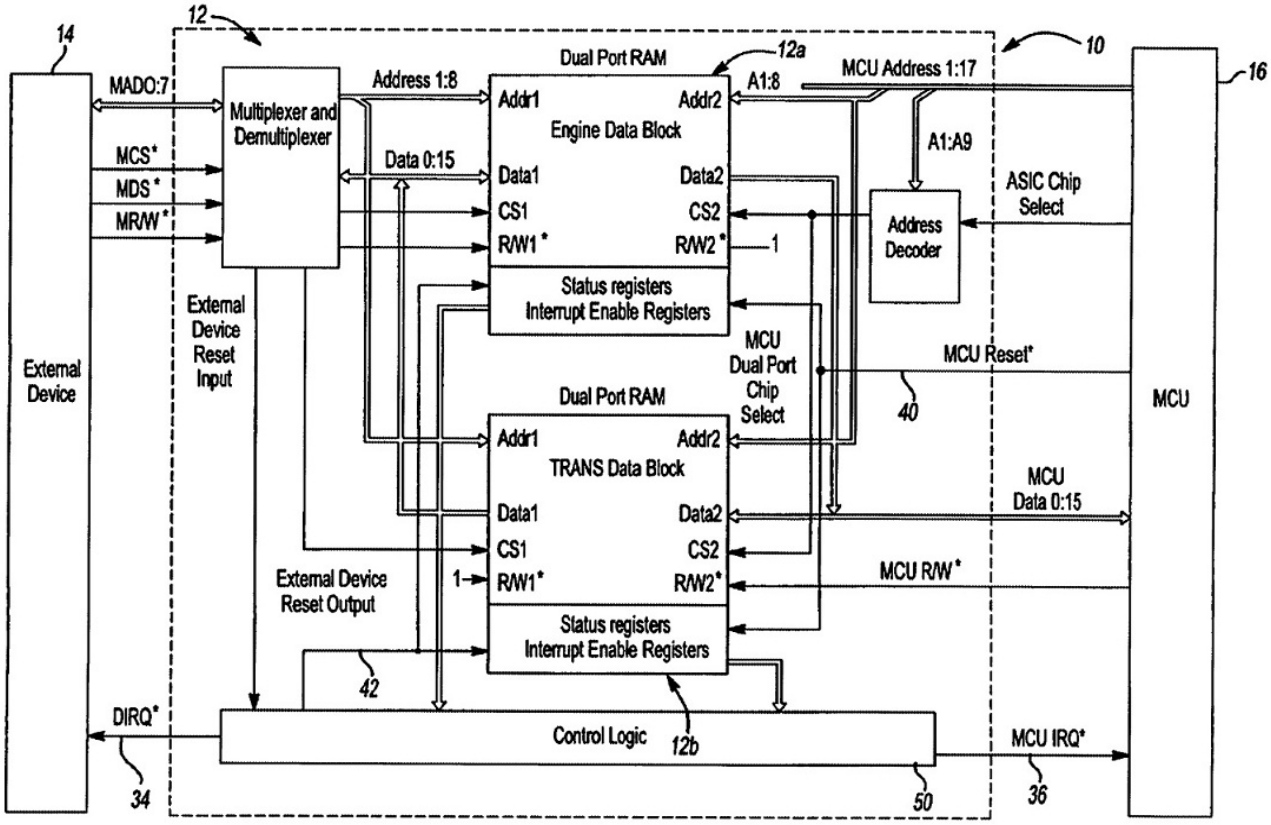
Un signal de commande informe la mémoire si l'opération demandée est une écriture ou une lecture. Dans les deux cas, toutes les opérations présentées dans le paragraphe précédent se déroulent durant un cycle cadencé par le *signal d'horloge*. Lorsque la taille d'un mot est inférieur au nombre de cellules par ligne, l'adresse de la donnée est divisée en deux parties, l'une commandant le décodeur de ligne, l'autre le décodeur de colonne.

Dans une application *dual-port*, on souhaite, rappelons-le, une écriture et une lecture simultanées dans le tableau SRAM. Pour ce faire, les architectures suivantes pourraient être envisagées :

- Avoir physiquement deux systèmes d'entrées-sorties, un pour l'écriture et un pour la lecture [6]. L'avantage est que la valeur suivante à écrire peut être préparée pendant une lecture (SRAM pipelinée). Néanmoins, il est toujours impossible d'écrire et de lire sur un même cycle d'horloge.
- Avoir au moins deux sous-tableaux [7] (figure 1.2(a)). Des périphériques externes ont déjà été proposés [8], pour gérer une écriture et une lecture simultanées avec cette architecture (figure 1.2(b)). Néanmoins, il reste deux désavantages à cette architecture. Premièrement, pour un même nombre de cellules, il y a moins de mots par tableau et plus de périphériques sont nécessaires pour l'activation des BitLines et des WordLines. Deuxièmement, si les mots que l'on souhaite écrire et lire en même temps se trouvent dans le même sous-tableau, le système périphérique devra nécessairement introduire un temps de latence. Ce deuxième désavantage peut être outrepassé lors de la compilation du programme exécuté par le processeur interfacé avec la mémoire, mais pas nécessairement dans tous les cas.



(a) SRAM dédoublée pour permettre l'écriture et la lecture simultanées [7].



(b) Architecture de circuits périphériques pour gestion *dual-port*. L'arbitreur permet de gérer les adresses [8].

Figure 1.2 – Exemple d'architecture de mémoire SRAM *dual-port*.

Dans ce travail, l'objectif est de pouvoir lire et écrire simultanément, quelle que soit l'adresse des mots. Evidemment, un système périphérique (simplifié) devra vérifier que l'on n'écrive et ne lise pas simultanément sur *le même mot*. Cette hypothèse est en fait très peu restrictive.

Dans un système de type FIFO, le système de gestion périphérique gère l'adressage pour avoir le comportement FIFO. En particulier, il informe le(s) processeur(s) si la FIFO est vide ou remplie, auquel cas il interdit respectivement la lecture ou l'écriture. Dans un système *multi-core*, s'il y a écriture et lecture simultanées sur une même adresse, l'architecture globale doit résoudre ce conflit. Soit l'ancienne valeur du mot doit être lue, auquel cas l'écriture se produit à une autre adresse ou bien attend le cycle suivant, soit la nouvelle valeur du mot doit être lue, et l'information peut se transmettre sans passer par le tableau mémoire.

Ce travail se concentrera sur le niveau d'abstraction le plus bas d'une mémoire SRAM, c'est-à-dire la cellule mémoire élémentaire et ses *drivers* de charge des *BitLines* et *WordLines*. Par hypothèse, les périphériques externes gèrent l'accès en écriture et en lecture de sorte qu'il n'y a jamais de conflits d'adressage.

Dans la suite de ce travail, BitLine(s) sera noté BL et WordLine(s), WL.

1.2 Les cellules SRAM classiques

Cette section présente différentes architectures de cellules SRAM classiques de type *single-port*, ainsi que leur fonctionnement général. Même si elles ne sont pas originairement destinées à une application *dual-port*, certaines architectures pourront être légèrement modifiées et adaptées à l'application *dual-port*.

1.2.1 Cellule 6T classique

La cellule classique à six transistors, représentée à la figure 1.3, fonctionne de la façon suivante :

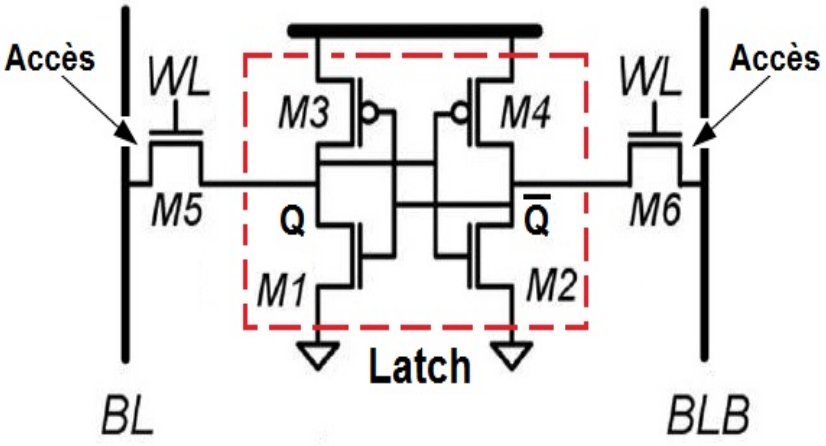


Figure 1.3 – Schéma d'une cellule conventionnelle à six transistors [9].

En rétention, les deux inverseurs montés en tête-bêche (M1 à M4) forment un système de contre-réaction permettant d'avoir deux points stables, ($Q = 0$, $\bar{Q} = 1$) et ($Q = 1$, $\bar{Q} = 0$), représentant respectivement la valeur logique 0 et 1. Ce système est appelé un ***latch*** et les

noeuds Q et $\bar{Q}$, contenant l'information sur la donnée, sont appelés noeuds de rétention. Les transistors d'accès (M5 et M6) sont coupés pour minimiser l'influence d'une variation des BL.

Pendant une écriture, la *BitLine* (BL) est préchargée à la valeur de la donnée que l'on veut écrire, la tension d'alimentation V_{dd} pour la valeur 1, et la masse pour 0. La *BitLine* complémentaire (BLB) est préchargée à l'inverse logique de la donnée, puis la *WordLine* (WL) est activée. Si la valeur précédente de la donnée est identique à la nouvelle, les niveaux de tension des noeuds de rétention ne varient pas, et la WL se coupe à la fin du temps d'écriture pour permettre l'écriture sur d'autres lignes. Si la valeur précédente était l'inverse binaire de la nouvelle donnée, les transistors d'accès chargent ou déchargent les noeuds de rétention Q et $\bar{Q}$, pour faire basculer la cellule d'un point stable à un autre. Ces transistors d'accès doivent pouvoir vaincre la contre-réaction du *latch*. En particulier, il leur faut délivrer un courant plus important que les transistors du *latch* pour pouvoir charger ou décharger les capacités des noeuds de rétention durant le temps d'écriture, et forcer la donnée à écrire. S'ils n'y arrivent pas et que la valeur de la donnée n'a pas changé, il y a échec d'écriture.

Pour la lecture, les *BitLines* sont préchargées à la tension d'alimentation V_{dd}, puis la *WordLine* est activée. Un chemin de décharge se crée alors entre une des deux *BitLines* et la masse, passant par le noeud qui contenait la valeur basse. L'autre BitLine reste à tension haute puisque l'autre noeud de rétention contenait la valeur haute. Les deux BitLines sont les entrées d'un amplificateur opérationnel (appelé *sense-amplifier*) qui aura sa sortie à tension haute ou basse, en fonction de la différence relative de tensions des BitLines. On peut ainsi retrouver la valeur de la donnée à partir de la sortie du *sense-amplifier*. La tension du noeud de rétention contenant la valeur basse monte lors de la décharge de la BitLine. Si elle dépasse un certain seuil, elle fera basculer l'autre inverseur, qui lui-même modifiera la valeur du premier inverseur (contre-réaction du *latch*), et la donnée sera perdue (lecture destructrice). Cette tension milieu est déterminée par la "force" relative entre le transistor d'accès (M5 ou M6) et le transistor NMOS de pied (M1 ou M2) se trouvant sur le chemin de décharge. Plus le transistor d'accès est fort par rapport au NMOS du *latch*, plus la tension sera haute et le risque de perdre la donnée augmentera.

On remarque que le rôle des transistors d'accès est contradictoire en lecture et en écriture. Pour l'écriture ils doivent être suffisamment forts pour changer l'état du *latch*, et inversément pour la lecture. Le compromis communément utilisé est d'avoir un transistor NMOS de pieds avec un courant supérieur à celui du transistor d'accès, lui-même supérieur à celui du PMOS (pour un même quadruplet de tensions à leurs bornes). Ainsi la lecture est robuste, et lors de l'écriture, un des deux noeuds de rétention pourra être déchargé et l'autre se charger grâce à la contre-réaction du *latch*. Ce réglage de courant peut se faire via la largeur des transistors, en utilisant par exemple une largeur pour les transistors de pieds (M1 et M2) deux fois supérieure aux transistors d'accès : $W_{g,access} = W_{g,PMOS} = W_{min}$ et $W_{g,NMOS} = 2W_{min}$ ([10], [11] et [12]).

On remarque directement qu'on ne peut pas lire et écrire en même temps dans le tableau avec ce type de cellule, car les deux BitLines sont utilisées pour les deux opérations.

Pour donner un point de comparaison, le tableau 1.1 montre la densité des SRAM et les marges de bruit (SNM [1]) obtenues pour différents types de cellules de l'industrie.

1. La marge de bruit quantifie la robustesse, la fiabilité de la cellule. Une définition plus précise est donnée en section 2.4.5

Table 1.1 – LP = Low Power, GP = General Purpose, HD = High Density, HP = High Performance.

Technologie CMOS	Surface par cellule [μm^2]	SNM [mV]	V_{dd} [V]
65nm Intel LP [13]	0.68	250	1.1
65nm IBM HP [14]	0.676	NC	1.2
65nm IBM HD [14]	0.54	150	1.2
65nm STM [15]	0.5	240	1.2
32nm IBM GP [16]	0.157	213	0.9
32nm IBM LP [17]	0.157	250	1.1
32nm IBM HP [18]	0.149	320	1.1
32nm Intel HD [19]	0.148	NC	1.1
32nm Intel HD [19]	0.171	NC	1.1
28nm TSMC HD [20]	0.127	220	1
28nm TSMC LP [20]	0.155	160	0.7

1.2.2 Autres architectures

Dans la littérature, plusieurs autres architectures de cellules SRAM ont été proposées pour améliorer certaines caractéristiques de la cellule 6T classique. Certaines ont des systèmes séparés de lecture et d'écriture, d'autres non. Dans cette section sont présentées différentes architectures ainsi que leur fonctionnement générique. De là, il sera plus aisé de déterminer quelles cellules utiliser pour l'application *dual-port*.

Cellule 6T + 2T

La cellule représentée à la figure 1.4 sera appelée cellule 6T+2T dans ce travail (au lieu de 8T [9], pour ne pas confondre avec la cellule *dual-port* étudiée plus loin). Le système de rétention de la donnée est le même que celui de la cellule classique à 6 transistors, et l'écriture se déroule de la même manière. Par contre, la lecture se fait par l'ajout de deux transistors supplémentaires, M7 et M8. Le transistor M7 a sa grille connectée à un noeud de rétention (ici $\bar{Q}$), et le transistor M8 a sa grille connectée à la WL de lecture (RDWL) - différente de celle d'écriture - et son drain connecté à la BL de lecture (RDBL), qui n'est pas utilisée pour l'écriture. Lorsque la ligne n'est pas lue (RDWL = 0V), le chemin entre RDBL et la masse est coupé quelle que soit la valeur de la donnée.

Lorsque les deux transistors sont ouverts, i.e. que la WL est activée *et* que la valeur du noeud Q est haute, la BL est déchargée. Cette dernière est connectée à un *sense-amplifier single-ended* (dont la deuxième entrée est à une tension fixe [9]). La valeur de la donnée est alors transmise en sortie de la même manière que la cellule 6T classique. Si Q représente un 1 logique, le chemin est fermé, la BL ne se décharge pas (idéalement) et le *sense-amplifier* garde sa sortie à une tension haute. Ce système de lecture peut être copié de l'autre côté de la cellule pour avoir une lecture différentielle.

Ce système de lecture composé de deux transistors est appelé un ***buffer* de lecture**. La caractéristique principale de tous les buffers de lecture est que l'information de la donnée de la

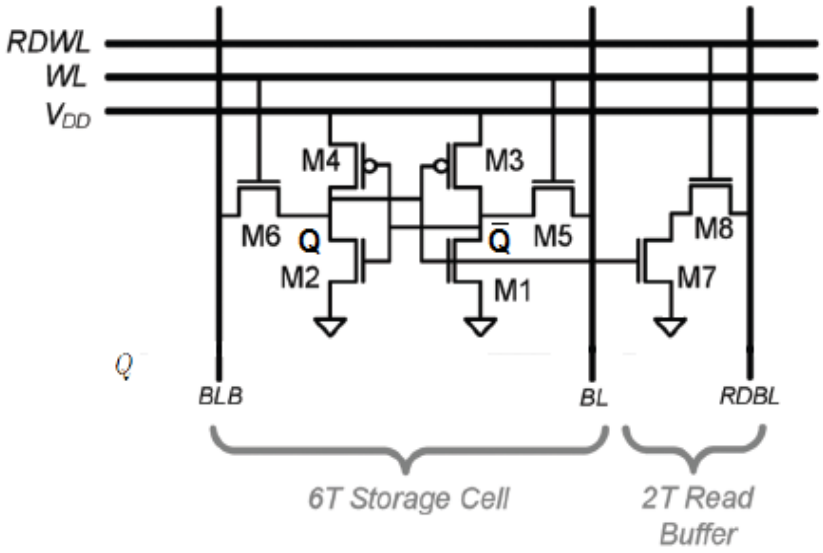


Figure 1.4 – Cellule 6T+2T [9].

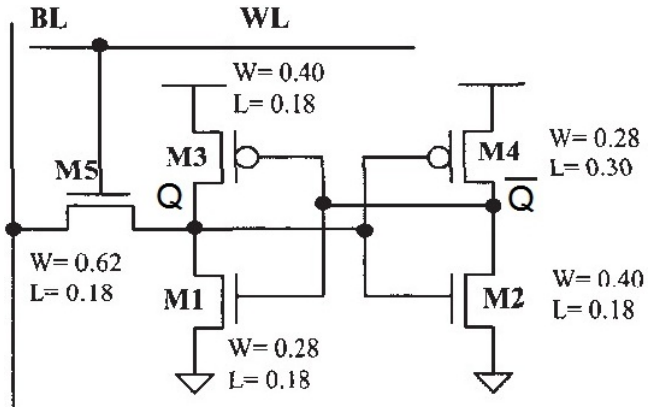


Figure 1.5 – La cellule 5T [10].

cellule est contenue sur la grille d'un ou plusieurs transistors. Ceci permet un "découplage´´ quasiment parfait entre le système de rétention (et d'écriture) et le système de lecture.

Cellule 5T

En supprimant la BL complémentaire (BLB) de la cellule 6T, et le transistor d'accès correspondant, nous obtenons la cellule 5T proposée dans [10].

Cette cellule fonctionne sur le même principe que la cellule 6T. Néanmoins, l'écriture effective d'un 1 logique ne peut se faire que par un dimensionnement approprié des transistors (figure 1.5). La BL doit être préchargée à une tension $V_{pc} < V_{dd}$ pour obtenir une lecture non destructrice et ne pas forcer l'écriture d'un 1 logique. Il faut donc générer dynamiquement une tension supplémentaire dans le circuit. La marge de bruit en lecture ainsi obtenue est inférieure de 50% à celle d'une cellule 6T classique. A la tension d'alimentation de [10], la marge de bruit reste toutefois suffisante pour avoir un tableau fonctionnel et robuste face à la variabilité de fabrication (section 1.5).

Outre le gain en surface de silicium, elle permet aussi une diminution de l'énergie dynamique (une seule BL à charger) et de la puissance statique (un courant de fuite en moins).

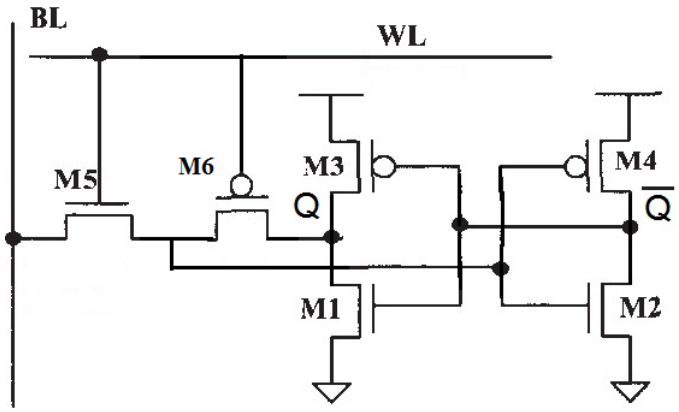


Figure 1.6 – La cellule 5T PMOS (image inspirée de [21] et [10]).

Cellule 5T PMOS

En ajoutant un transistor PMOS à la cellule 5T tel qu'illustré sur la figure 1.6), nous obtenons la cellule 5T PMOS proposée dans [21].

Lorsque la WL est activée, le transistor PMOS est coupé et casse la contre-réaction de l'inverseur contrôlant le noeud Q. Ceci permet au transistor d'accès d'écrire plus facilement un 0 ou d'un 1 logique, car plus aucun autre courant ne vient charger ou décharger le noeud Q.

L'écriture d'un 1 logique reste néanmoins plus délicate que celle d'un 0. En effet, lorsque la BL est à une tension haute et que le transistor d'accès charge le noeud Q, sa tension V_{gs} égale à $V_{WL} - V_Q$ diminue progressivement. C'est donc le mode le plus critique d'un point de vue dynamique.

La lecture ne peut plus se faire par la même BL avec un *sense-amplifier single-ended*. Le transistor PMOS coupant le chemin de décharge, la tension du noeud Q va nécessairement augmenter et faire basculer l'autre inverseur. L'utilisation d'un *buffer* de lecture est indispensable.

Cellule à trigger de Schmitt simplifié

Cette cellule, proposée dans [12], fonctionne exactement de la même manière que la cellule 6T classique, ce qui donne un grand avantage d'un point de vue fabrication/adaptation à la technologie existante.
L'ajout des transistors supplémentaires (figure 1.7) permet d'obtenir un *latch* avec un seuil de basculement beaucoup plus élevé que deux inverseurs montés en tête-bêche. Le but recherché est d'améliorer sensiblement la marge de bruit en lecture, par rapport à la cellule 6T classique.

Cellule de Hobson

La cellule de la figure 1.8(a) a une BL pour effectuer l'opération d'écriture (WB) et une autre BL pour l'opération de lecture (RB). Les WL sont, elles aussi, différentes pour l'écriture et la lecture, ce qui convient parfaitement pour une application *dual-port*.

Le système d'écriture est équivalent à la cellule 5T. Le système de lecture est sensiblement semblable à celui d'une cellule 6T classique (avec un *sense-amplifier single-ended*).

Dans [22], l'auteur explique que la principale difficulté de cette cellule est l'écriture d'un 1 logique au travers d'un seul transistor NMOS. Cette difficulté est contournée dans l'article par une variante (figure 1.8(b)) qui utilise ce qu'on appelle une assistance en écriture (*Write assist*), système qui sera discuté dans la section 1.4.

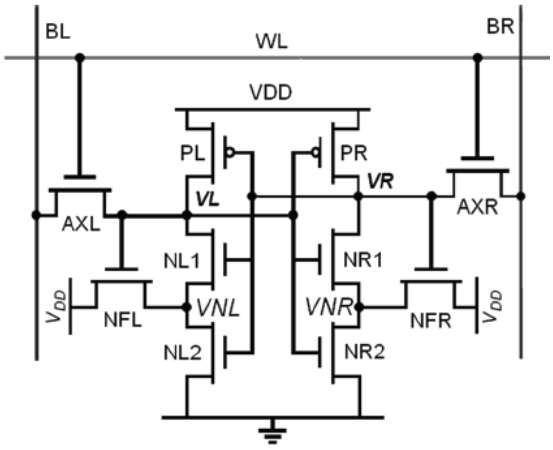
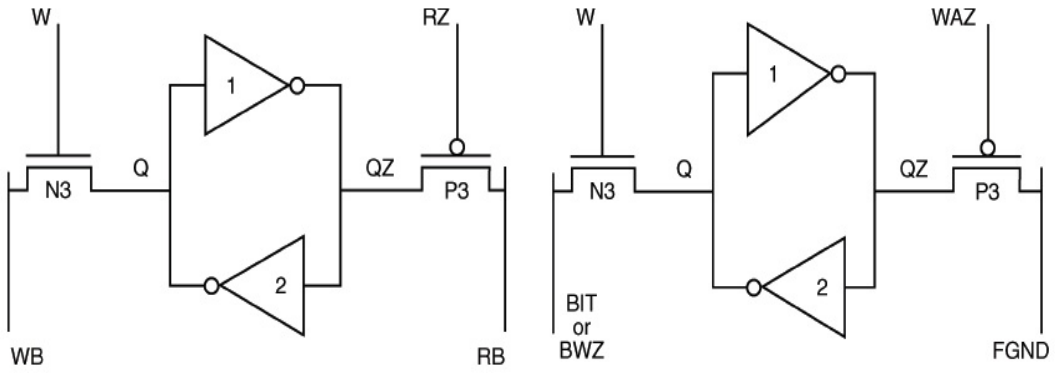


Figure 1.7 – Latch composé de deux triggers de Schmitt simplifiés [12].



(a) Cellule de Hobson avec BL de lecture et d'écriture dédicacées.

(b) Cellule de Hobson avec *Write Assist* pour faciliter l'écriture d'un "1" logique.

Figure 1.8 – Autre architecture de cellule à six transistors [22].

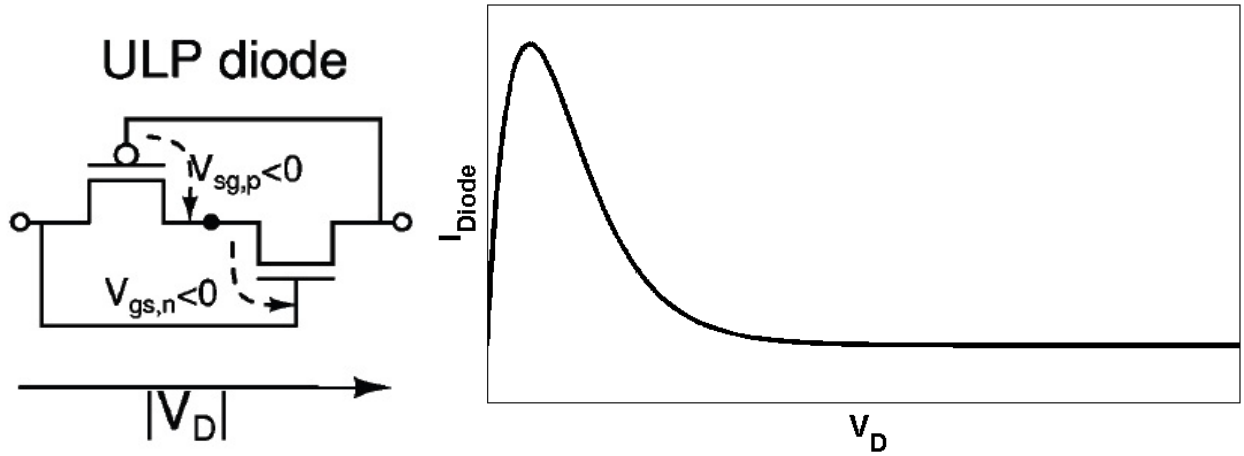
Cellule 7T ULP

C'est le troisième type de latch rencontré dans la littérature, après les inverseurs montés en tête-bêche et le trigger de Schmitt simplifié. Il est basé sur le principe de la diode ULP montée en *reverse* [23]. Deux transistors MOS classiques sont interconnectés pour obtenir une relation courant-tension représentée à la figure 1.9(b).

La mise en série des deux diodes de type ULP permet d'obtenir deux points stables pour le noeud X (figure 1.10(b)). L'ajout d'un transistor d'accès termine la cellule. Notons qu'il n'y a pas de noeud complémentaire dans ce *latch*.

Lors de l'écriture, la WL est activée et le transistor d'accès charge ou décharge ce noeud. Il n'y a qu'une seule BL car un seul noeud contient l'information de la donnée. Le courant des diodes est un courant sous-seuil de transistor, la contre-réaction du *latch* est donc négligeable pour le transistor d'accès M_W. Cela implique également que la lecture ne peut se faire par la BL. L'utilisation d'un *buffer* s'impose. Celui retenu dans [23] est le buffer 2T rencontré précédemment (cellule 6T+2T). De plus, vu les courants de régénération de ce *latch* sous des courants sous-seuil, il peut y avoir un temps relativement long pour que le noeud X atteigne la tension d'alimentation V_{dd} [24].

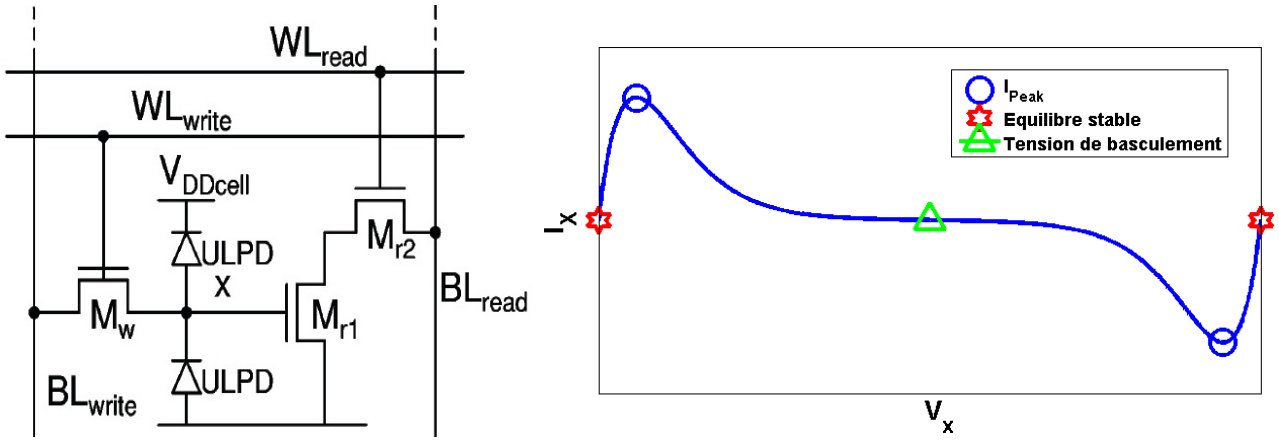
Notons que des variantes des cellules 6T et 6T+2T avec des inverseurs en technologie ULP ont déjà été étudiées dans [24]. Elles sont constituées d'un grand nombre de transistors, ce qui ne permet pas d'atteindre une mémoire SRAM dense.



(a) Schéma de principe de la diode ULP [23].

(b) La relation courant-tension de la diode ULP, caractérisée par une résistance négative et un pic de courant.

Figure 1.9 – La diode ULP en mode *reverse*.



(a) La cellule 7T ULP [23].

(b) La caractéristique courant-tension du *latch* ULP (uniquement les deux diodes)

Figure 1.10 – Le *latch* ULP.

Inverseur *tri-state*

L'architecture représentée à la figure 1.11 est un autre *buffer* de lecture. En effet, l'information de la donnée est contenue sur les grilles des transistors P3 et N5 constituant un inverseur. Les deux autres transistors en parallèle (P4 et N6) sont les transistors d'accès en lecture [4].

En rétention, la WL (RWL) est maintenue à une tension basse, tandis que la WordLine complémentaire (/RWL) est maintenue à la tension haute. Ainsi, les deux transistors d'accès sont coupés et la BL voit une haute impédance à la sortie de l'inverseur.

Pendant une lecture, la WL connectée au NMOS (RWL) est chargée tandis que celle du PMOS (/RWL) est déchargée. L'inverseur peut alors forcer la BL à la valeur qu'il contient, soit V_{dd} soit *gnd*. Cette BL n'est donc pas connectée à un *sense-amplifier*, mais à un inverseur qui transmettra l'information en sortie (signal **ReadOut** sur la figure 1.11).

On peut considérer que ce buffer a deux entrées et une sortie. Les entrées sont les valeurs de $\bar{Q}$ et RWL et la sortie est la tension de la BitLine. Sa table de vérité est alors caractérisée par trois états :

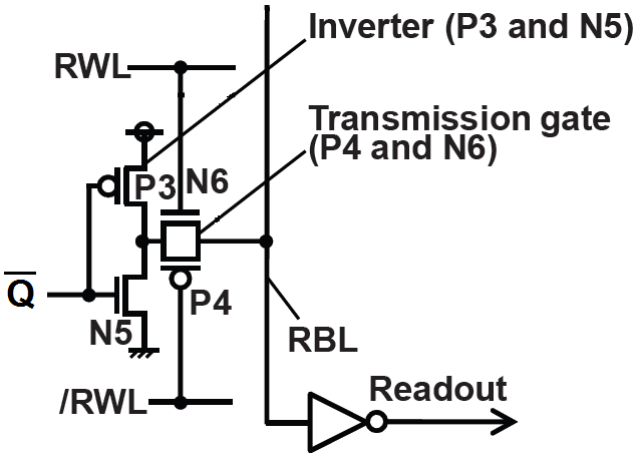


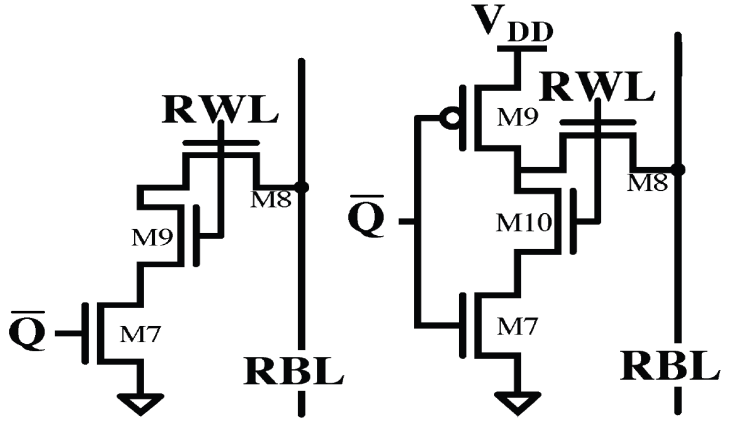
Figure 1.11 – Le buffer nommé inverseur tri-state [4].

Bit	WL	BL
0	1	0
1	1	1
X	0	Z

d'où son nom d'inverseur *tri-state*. Un seul transistor d'accès ne suffit pas que pour faire basculer l'inverseur du signal **ReadOut** suffisamment rapidement, d'où l'utilisation de la logique de passage (*transmission gate*).
L'absence de *sense-amplifier* et multiplexeur en sortie permet une réduction importante de la consommation [21].

Buffers 3T et 4T

Les deux *buffers* de lecture nommés 3T [25] et 4T [26], représentés respectivement aux figures 1.12(a) et 1.12(b), fonctionnent sur le même principe que le *buffer* 2T. En rétention, la WordLine (RWL) est coupée et la BitLine (RBL) voit une haute impédance ; en lecture, la WL est activée et la BL, connectée à un *sense-amplifier single-ended*, se décharge ou non en fonction de la valeur de la donnée.



(a) Buffer nommé 3T [25]. (b) Buffer nommé 4T [26].

Figure 1.12 – Comme dans tout *buffer*, le signal $\bar{Q}$ peut être remplacé par Q.

L'ajout du ou des transistor(s) supplémentaire(s) a pour but de réduire le courant de fuite vu de la BL en rétention. Ceci afin de diminuer la consommation et d'augmenter le nombre maximal de cellules que l'on peut mettre par BL (voir section 2.4.6). Néanmoins, ce faible courant de fuite entraine inexorablement un faible courant de décharge. Ces *buffers* seront donc dédiés à une application SRAM proche de 100MHz [26].

1.3 La cellule SRAM *dual-port* conventionnelle 8T

La cellule 8T *dual-port* conventionnelle a un principe de fonctionnement très semblable à celui de la cellule 6T [27]. Du point de vue architecture, deux transistors d'accès, deux BL et une WL ont été ajoutés (figure 1.13).
Pendant l'écriture, uniquement la WL d'écriture (WLW) est activée et les deux BL d'écriture (BLW et /BLW) sont chargées en fonction de la donnée à écrire.
Les deux BL de lecture (BLR et /BLR) peuvent être maintenues à tout instant à la tension d'alimentation V_{dd}. Elles sont également connectées aux entrées d'un petit amplificateur opérationnel, un *sense-amplifier*. Lors de la lecture, seule la WL de lecture (WLR) est activée et l'une des deux BL de lecture se décharge faisant basculer le *sense-amplifier* qui transmettra la donnée en sortie.

Les contraintes de fiabilité discutées pour la cellule 6T classique restent les mêmes ; elles seront étudiées plus longuement dans la section 3.1.

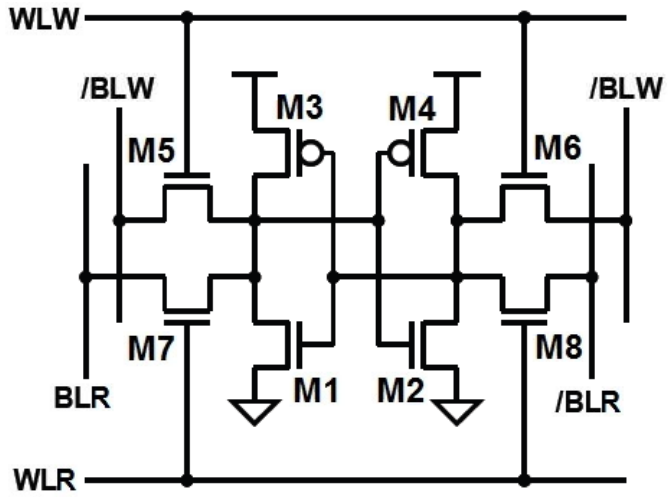


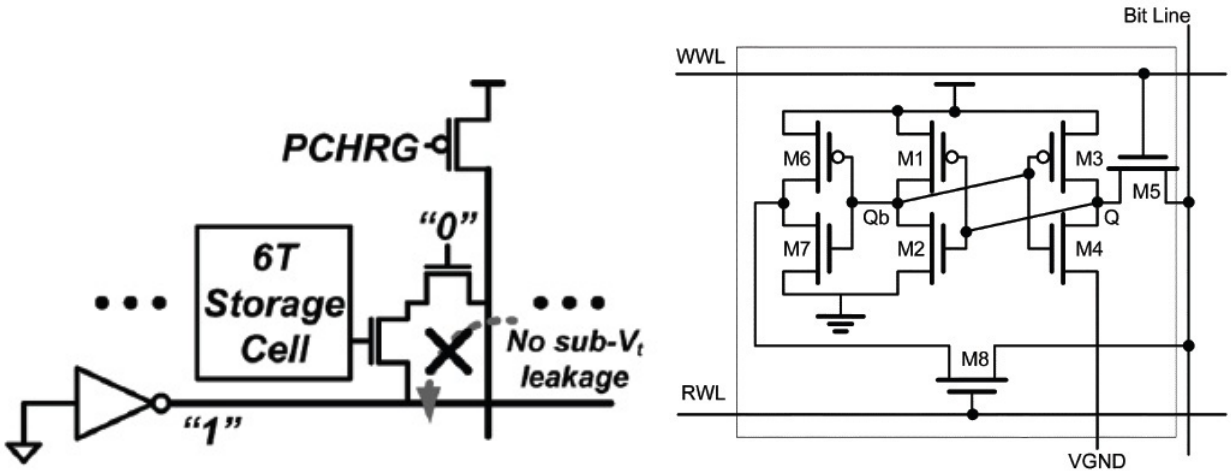
Figure 1.13 – Schéma d'une cellule *dual-port* conventionnelle à huit transistors [27].

1.4 Les *assists*

Les *assists* sont des systèmes qui permettent d'améliorer les performances - en particulier les marges de bruit - des cellules SRAM sans modifier leur topologie, grâce à une polarisation non-standard. Ces polarisations sont appliquées aux connections avec lesquelles la cellule interagit avec le monde extérieur :

- La tension d'alimentation vue par la cellule
- La tension de masse vue par la cellule
- La tension de *WordLine*
- La tension de *BitLine*

Certains *assists* sont statiques, dans le sens où leur effet est dédié au mode de repos d'une cellule mémoire. Par exemple, la valeur de la tension WL en rétention est négative pour réduire le courant de fuite des transistors d'accès [24] ; ou encore, la masse virtuelle du buffer de lecture 2T maintenue à tension haute pour supprimer le courant de fuite (figure 1.14(a)). On peut aussi imaginer une tension d'alimentation statique de la cellule différente de celle des WL et BL.



(a) La masse vue par le buffer 2T est maintenue à V_{dd} pour supprimer le courant de fuite sur la BL [9].

(b) Cellule SRAM à 8 transistors avec polarisation asymétrique de la masse [30]. La topologie est modifiée pour s'adapter à l'*assist*.

Figure 1.14 – Deux exemples d'*assists*

Les *assists* dynamiques sont activés au cours d'une lecture ou d'une écriture. Par exemple, diminuer la tension d'alimentation virtuelle des cellules durant l'écriture permet d'améliorer la marge de bruit [9] - [11]. Un *charge pump* permet d'atteindre temporairement une tension de grille supérieure à celle de l'alimentation, et donc d'augmenter le courant du *buffer* de lecture [11]. La technique du *charge pump* peut aussi être utilisée pour produire une tension de BL négative durant l'écriture, facilitant la décharge du transistor d'accès [28].
Une grande partie de ces *assists* est résumée dans [29].

Allant plus loin, des architectures ont été conçues pour être adaptées à un *assist* particulier. Dans [30], une cellule à une seule BitLine voit son efficacité en écriture accrue, grâce à la masse virtuelle de transistors non contenus dans le *latch* (figure 1.14(b)).

Les assists dynamiques ne seront pas étudiés ici car ils sortent du cadre de ce travail.

1.5 Variabilité

Lors de la fabrication des transistors, il y a toujours un écart inconnu et variable entre la largeur et la longueur de la grille souhaitées et celles effectivement gravées. De même, le dopage des tranches de silicium induit une tension de seuil effective, V_t, différente de celle attendue. Avec la diminution toujours croissante de la taille des transistors, ces variations de paramètres physiques sont non négligeables pour les technologies avancées (< 180nm). On distingue deux types de variabilité : la variabilité globale et locale.

La variabilité globale se fait ressentir à l'échelle d'une tranche de silicium. Les performances globales d'un circuit près du centre de la tranche peuvent être significativement différentes

de celles d'un circuit situé au bord (voir figure 1.15). Par ailleurs, les performances varient également de tranche à tranche.

La variabilité locale a ses effets à l'échelle du transistor. Du fait des dimensions extrêmement petites des transistors actuels, pratiquement à l'échelle de l'atome, la position de chaque atome de dopant a une influence accrue sur la tension de seuil globale du transistor. La figure 1.16 représente la tension de seuil locale rencontrée par un électron dans le canal. On peut imaginer que l'électron se meut dans les vallées de potentiel formées par les atomes dopants. La variabilité locale implique qu'au sein même d'un circuit, deux transistors adjacents fabriqués pour avoir la même tension de seuil peuvent avoir des V_t effectifs significativement différents.

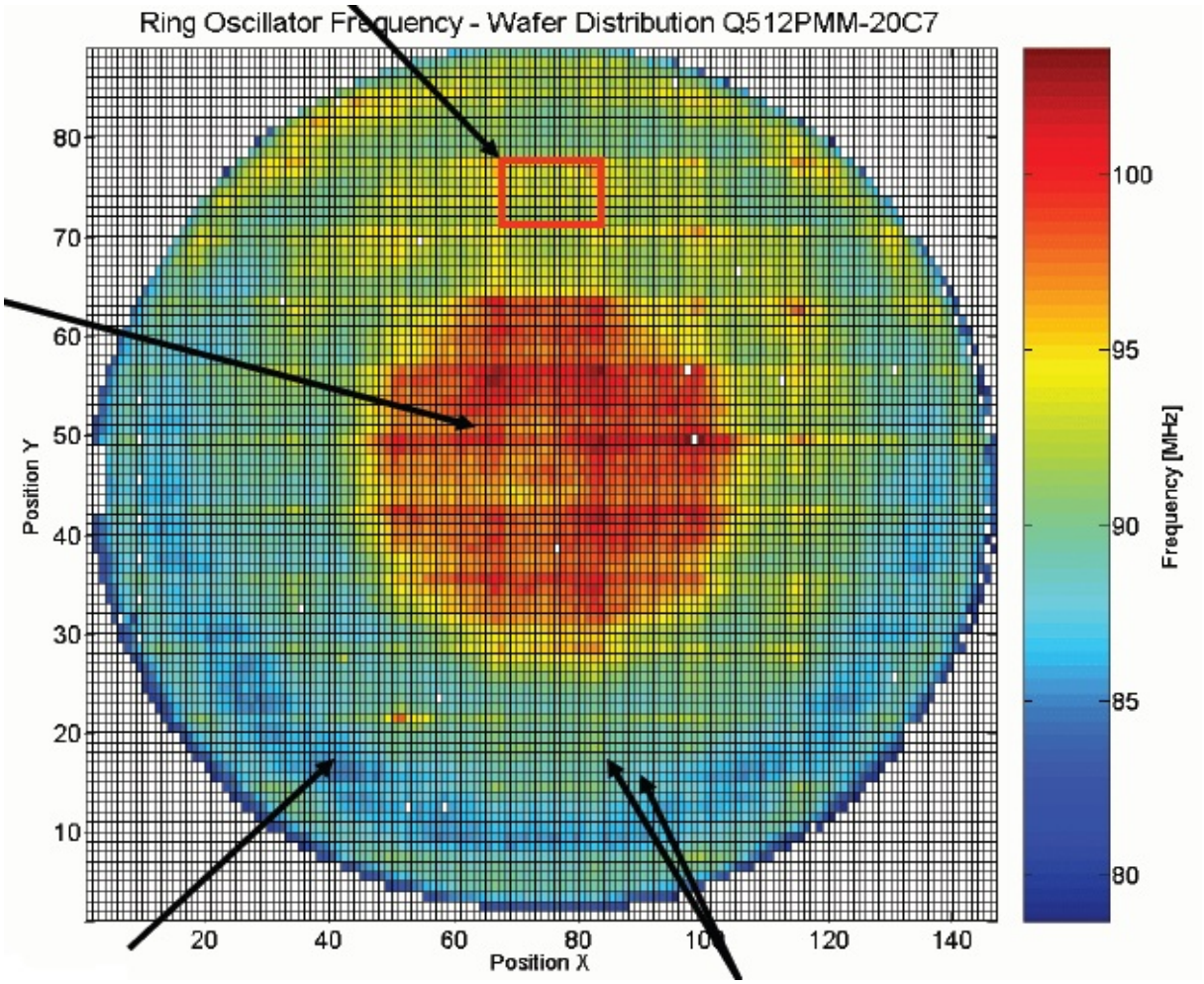


Figure 1.15 – Exemple de variabilité globale : distribution de la fréquence d'oscillation de circuits numériques sur une tranche de silicium [31].

Modélisation

Les concepteurs de circuit doivent donc tenir compte de ces effets pour rendre leurs circuits fonctionnels face à la variabilité. Pour ce faire, ils ont à leur disposition les paramètres numériques rendant compte de la variabilité et une loi statistique modélise la variabilité locale, ou *mismatch*.

Pour tenir compte de la variabilité globale, le ***corner*** dans lequel se situe le circuit est fixé lors des simulations. En fixant le *corner*, les changements de caractéristisques physiques s'appliqueront à tous les transistors du circuit. Pour ne pas simuler dans tous les états possibles de circuit, cinq corners représentant les cas extrêmes sont proposés dans les modèles :

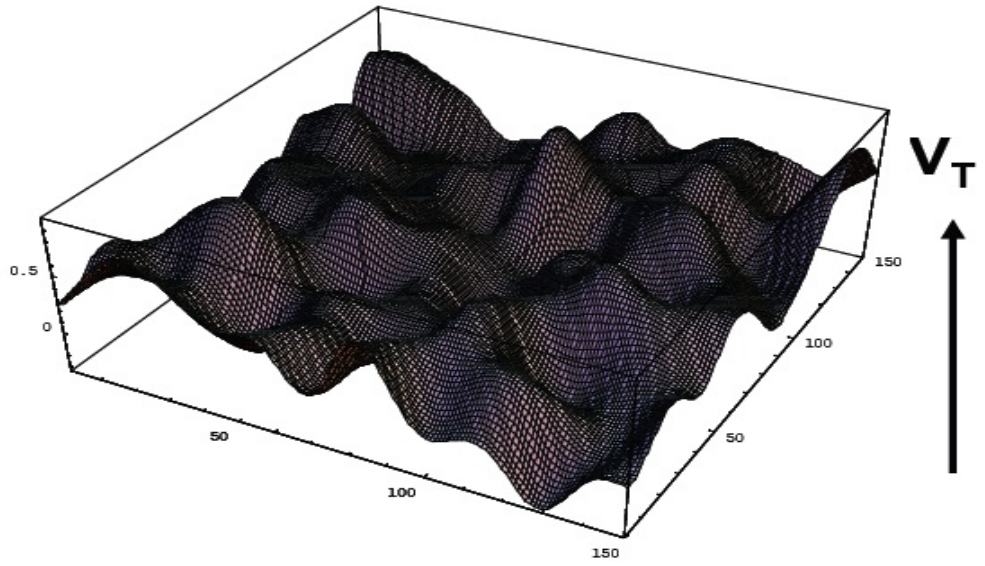


Figure 1.16 – Représentation de la variabilité locale : tension de seuil le long d'un canal de transistors [31].

- le corner Typical-Typical (TT) : tous les paramètres physiques sont à leur valeur nominale, déterminée par la fonderie.
- le corner Fast-Fast (FF) : la valeur des paramètres est fixée de sorte que le courant de drain des transistors vaut le maximum observé.
- le corner Slow-Slow (SS) : la valeur des paramètres est fixée de sorte que le courant de drain des transistors vaut le minimum observé.
- le corner Fast-Slow (FS) : le dopage des transistors PMOS et NMOS se fait à des étapes différentes lors de la fabrication des circuits. Il peut donc y avoir au final une asymétrie entre la tension de seuil moyenne des NMOS et des PMOS. Pour modéliser ceci, ce corner modélise tous les transistors NMOS comme s'ils étaient dans le cas Fast, et tous les transistors PMOS comme s'ils étaient dans le cas Slow.
- le corner Slow-Fast (SF) est l'opposé du précédent.

Après avoir fixé le *corner*, les simulations Monte-Carlo (MC) sont utilisées pour tenir compte de la variabilité locale. Une simulation Monte-Carlo répète la simulation principale un grand nombre de fois. A chaque itération, une nouvelle valeur de paramètres physiques est donnée à chaque transistor. La valeur de la tension de seuil [2] est fixée de manière aléatoire suivant une loi normale, dont la moyenne est donnée par le corner et l'écart-type est donné par la loi

$$\sigma_{V_t} = \frac{A_{Vt}}{\sqrt{W_g L_g}} [31] \tag{1.1}$$

où A_{Vt} est un paramètre dépendant de la technologie, fixée dans le modèle des transistors, et W_g et L_g la largeur et la longueur de grille d'un transistor. Cette loi statistique signifie physiquement que, plus la taille de la grille du transistor est grande, moins la variabilité se fait ressentir. Son caractére gaussien implique que 99% des transistors se situent entre $\pm 3\sigma$ de la moyenne.

Les simulations MC sont utilisées pour connaitre la distribution statistique de paramètres qui dépendent des caractéristiques physiques des transistors. En effet, même avec la loi normale fixée par le modèle extrait de l'industrie, on ne connait pas *a priori* la distribution des

2. Les variations des autres paramètres physiques suivent une autre loi. Néanmoins, la variation de la tension de seuil est le facteur prédominant sur la variation totale du courant de drain du transistor [32].

paramètres qui en dépendent. Des études ont été faites à ce sujet [33]. Pour notre propos, on considère que ces paramètres suivent également une loi normale, dont la moyenne et la déviation standard sont évaluées après un très grand nombre d'itérations MC. Par exemple, grâce à l'estimation de la distribution statistique de la marge de bruit d'une cellule, on peut connaitre la probabilité que celle-ci ne soit pas fonctionnelle. Comme une cellule SRAM est répétée un très grand nombre de fois dans un circuit intégré, cette probabilité doit être très petite. On considère que le rendement d'une mémoire composée de N cellules, est donné par

$$\eta_{SRAM} = \eta_{cell}^{N}$$

où η_{SRAM} est le rendement de fabrication de la mémoire et η_{cell} celui d'une seule cellule.

Dans ce travail, nous allons considérer une stabilité traditionnelle dite 6σ [34] ; ceci signifie que, si la marge de bruit est toujours positive à six écarts-types de la moyenne, alors toutes les cellules du circuit intégré sont fonctionnelles avec une probabilité d'erreur inférieure à *un milliardième*. Malgré cela, sur les milliards de circuits intégrés qui sortent chaque année des fonderies, certaines cellules risquent de ne pas être fonctionnelles. Il faut alors, soit les tester après fabrication, soit inclure de la redondance dans le circuit pour diminuer encore plus drastiquement la probabilité d'avoir un circuit défectueux ("*self-repairs*" SRAM [35])

1.6 Résumé

Après avoir présenté le fonctionnement général d'une mémoire SRAM, le contexte et l'application des cellules SRAM *dual-port* a été expliqué. Un aperçu de l'état de l'art des cellules *single-port* et *dual-port* a été présenté, ainsi que les *assists* utilisés dans les tableaux SRAM pour améliorer leurs performances. Finalement, le concept de variabilité, une non-idéalité intrinsèque des circuits électroniques, a été introduit. Il sera présent à chaque étape du raisonnement dans l'étude des cellules.

Avant d'analyser et de comparer les cellules de l'état de l'art entre elles, les critères de comparaison qui serviront de base à l'étude seront précisément définis.

Chapitre 2

Base de comparaison de cellules SRAM

Afin d'assurer la validité des résultats obtenus, il est critique d'établir une base de comparaison réaliste des cellules SRAM.

Dans ce chapitre, la première section expose les caractéristiques principales des transistors de la technologie considérée (32 nm FDSOI).
Le dimensionnement de la cellule 8T conventionnelle dans cette technologie est motivé dans la deuxième section.
Dans la troisième section, le circuit de test des cellules mémoire sera présenté et ses paramètres justifiés en fonction de la cellule 8T définie précédemment.
Enfin, la quatrième section définit précisément les caractéristiques étudiées et la mesure des performances des cellules.

Pour rappel, l'objectif est d'obtenir un tableau de cellules SRAM *dual-port* robustes face à la variabilité, avec une fréquence de fonctionnement de 1GHz sous une tension d'alimentation de 1V, tout en minimisant la surface de silicium utilisée et la consommation électrique.

2.1 Technologie 32nm FDSOI considérée

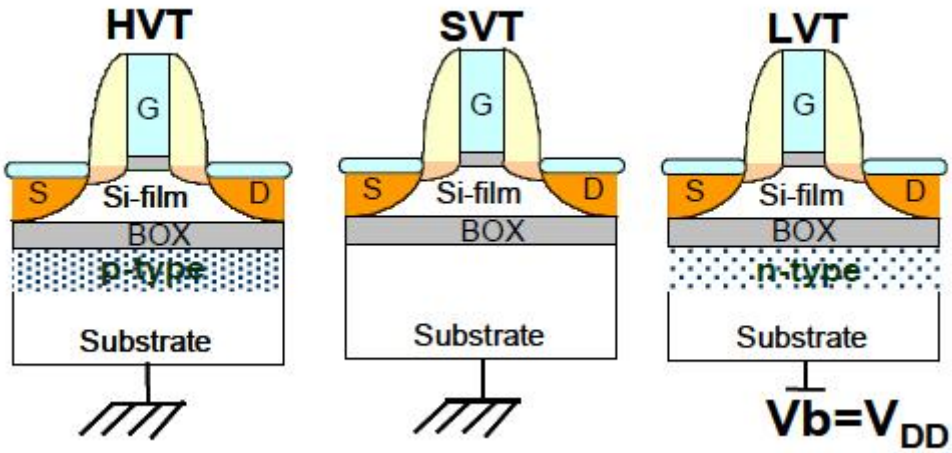


Figure 2.1 – Transistors NMOS de la technologie FDSOI 32nm du CEA-Leti [36]. Trois tensions de seuil obtenues grâce à un dopage et une polarisation sous l'oxyde enterré (BOX) différents.

Table 2.1 – Caractéristiques I_{on} - I_{off} pour les six types de transistors en régime saturé ($V_{ds} = 1V$) et linéaire ($V_{ds} = 0.05V$).

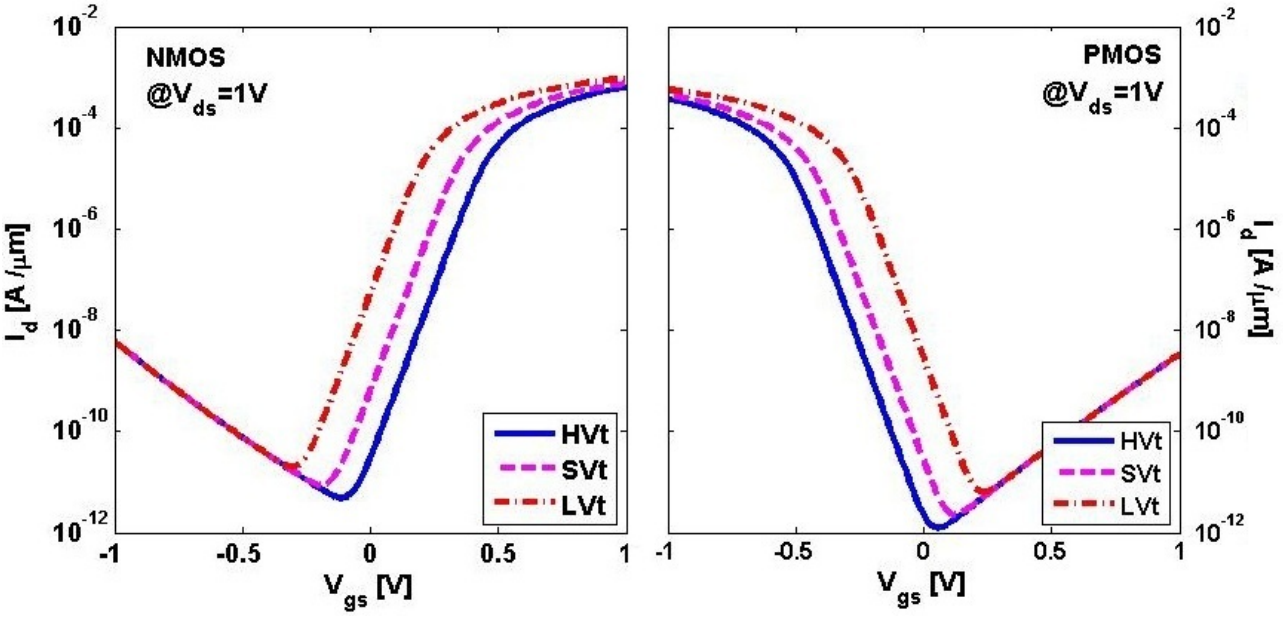
Type de transistor	$I_{on,sature}$ [/μm]	$I_{on,lineaire}$ [/μm]	$I_{off,sature}$ [/μm]
NMOS LVt	949.5 μ A	115.4 μ A	55.4 nA
NMOS SVt	763.1 μ A	110.9 μ A	654.9 pA
NMOS HVt	634.5 μ A	101.6 μ A	31.2 pA
PMOS LVt	568.1 μ A	69.1 μ A	2.7 nA
PMOS SVt	440.3 μ A	65.2 μ A	25.8 pA
PMOS HVt	365.3 μ A	59.2 μ A	2.2 pA

Les transistors de la technologie UTBOX-FDSOI 32nm du CEA-Leti sont représentés à la figure 2.1. Contrairement aux transistors Bulk classiques, les transistors SOI possèdent une couche d'isolant électrique relativement épaisse, appelée oxyde enterré (BOX), entre le canal sous la grille et le substrat. Ceci permet notamment de réduire les capacités parasites et les courants de jonction [37]. Si le film actif de silicium (Si-film) est suffisamment mince, toute la zone contenue entre l'oxyde de grille et l'oxyde enterré est entièrement déplétée. Ce type de transistor est donc appelé *Fully-Depleted Silicon On Insulator* (FDSOI).
Si la couche d'isolant de l'oxyde enterré est très grande par rapport aux dimensions du transistor, l'influence du substrat sur celui-ci est négligeable. Dans cette technologie, l'épaisseur du BOX n'est que de 10nm et les transistors sont appelés *Ultra-Thin BOX* (UTBOX). Ce très mince isolant permet de produire trois tensions de seuil différentes grâce à un dopage et une polarisation différents du substrat sous l'oxyde enterré [36]. La tension de seuil n'est donc pas fixée par le dopage du canal, et celui-ci ne contient pas d'atome dopant. Cette dernière propriété permet de réduire significativement la variabilité par rapport aux transistors Bulk classiques, car le σ_{V_t} dépend de la concentration en atome dopant [38].
Les transistors de cette technologie ont un oxyde de grille épais de 8nm, constitué de matériaux à haute permittivité électrique, dits $high-\kappa$. Ils ont une longueur de grille minimale $L_{g,min}$ de 30nm, et une largeur de grille minimale $W_{g,min}$ de 80nm. Leur tension d'alimentation nominale est de 1V.

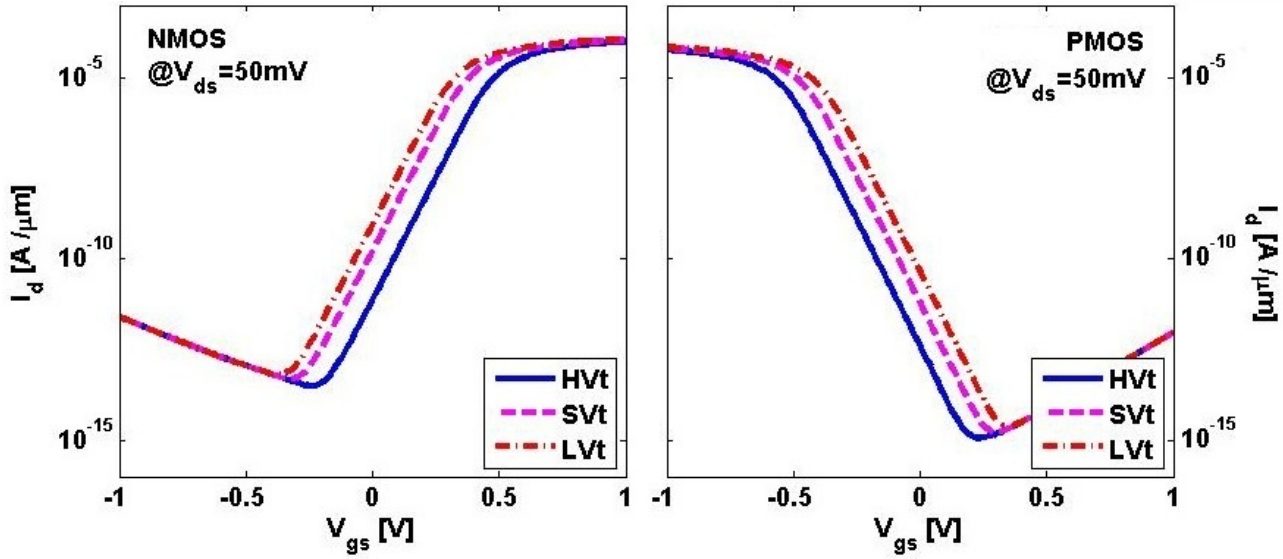
Les transistors fabriqués pour obtenir la plus basse tension de seuil seront appelés LVt (pour *Low-Vt*), ceux ayant la plus grande tension de seuil seront appelés HVt (*High-Vt*), et les transistors à tension de seuil intermédiaire SVt (*Standard-Vt*). Les figures 2.2(a) et 2.2(b) montrent les courbes courant-tension de ces transistors, respectivement en régime saturé et linéaire, pour une largeur de grille de $1\mu m$. Le tableau 2.1 en reprend les principales caractéristiques.

Le tableau 2.2 reprend d'autres caractéristiques extraites de transistors à taille minimale. Les capacités de grille, C_g, et de jonction, C_j, sont présentées pour estimer l'ordre de grandeur de la capacité totale de BL et de WL. Les courants de grille, I_g, sont extraits pour des tensions $V_d = V_s = 0V$, car cette configuration induit une fuite de grille maximale et sera rencontrée plusieurs fois dans la suite de ce travail. La tension de grille vaut 1V et -1V pour le courant de grille *forward* et *reverse* respectivement.

Notons que le courant de grille maximale en *forward* n'est pas négligeable par rapport au courant de drain sous-seuil ; malgré l'utilisation de matériaux $high-\kappa$, le courant parasite de grille reste relativement important.



(a) $V_{ds} = 1$V (régime saturé)



(b) $V_{ds} = 50$mV (régime linéaire).

Figure 2.2 – Courbes $I_d(V_g)$ en échelle logarithmique - *corner* TT, temp. 27°C.

2.2 Définition de la cellule *dual-port* de référence

La cellule de référence qui sera utilisée pour valider le banc de test (ou *testbench*) est la cellule *dual-port* conventionnelle dont tous les transistors ont une tension de seuil standard. Elle sera appelée 8T SVt dans ce travail.

La figure 2.3 montre la marge de bruit de la cellule *dual-port* 8T classique pour les trois modes de fonctionnement, en fonction de la taille des transistors NMOS de pied du *latch*.On remarque que la marge de bruit en lecture SNMR est toujours la plus petite, la lecture est donc l'opération la plus critique. Pour obtenir une marge de bruit suffisante, le rapport de largeur entre les transistors NMOS de pied et les transistors d'accès (appelé *cell ratio*) sera fixé à 2. C'est un rapport usuel pour la cellule classique à six transistors, qui sera à nouveau justifié dans la section 2.4.5.

Table 2.2 – Caractéristiques secondaires pour des transistors de dimensions minimales (Corner TT). Paramètres d'extraction décrits dans la section B.

	C_g (moyen) [fF]	C_j (moyen) [fF]	I_{off} [pA]	I_g *forward* [pA]	I_g *reverse* [pA]
NMOS LVt	0.088	0.06	4431.8	17.7	0.29
NMOS SVt	0.081	0.06	52.3	12.6	0.29
NMOS HVt	0.075	0.06	2.5	6.5	0.29
PMOS LVt	0.085	0.06	217.9	24.1	0.09
PMOS SVt	0.076	0.06	2.1	11.3	0.09
PMOS HVt	0.072	0.06	0.17	3.9	0.09

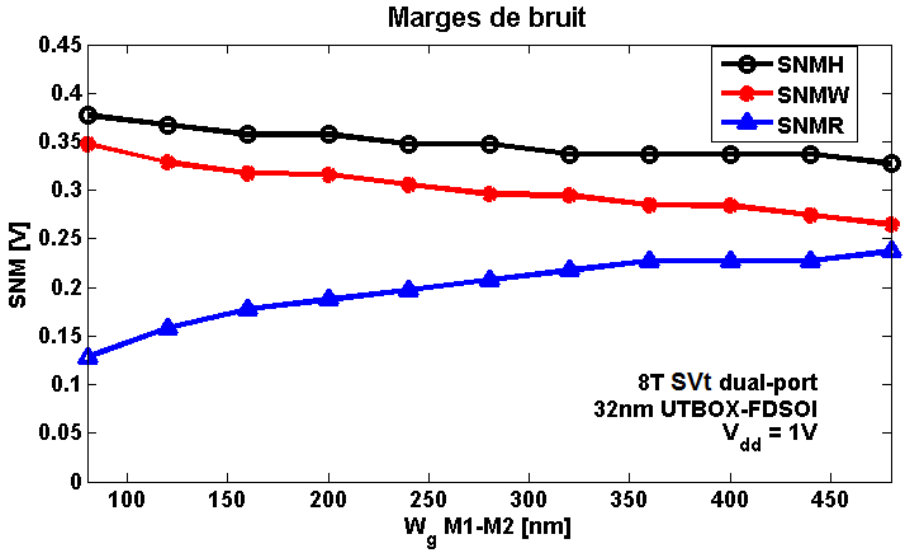


Figure 2.3 – Les marges de bruit de la cellule *dual-port* 8T conventionnelle suivant les trois modes de fonctionnement : en rétention $SNMH$, en écriture $SNMW$ et en lecture $SNMR$ - $W_{g,min}$ pour les six autres transistors, *corner* TT, temp. 27°C.

2.3 *Testbench* de cellule SRAM

Afin d'évaluer les performances absolues, il est nécessaire d'établir un circuit de test le plus réaliste possible.

Les systèmes périphériques (décodeur, mux,...) ne seront pas traités dans ce travail. Le *testbench* représentera uniquement le dernier niveau d'abstraction d'un tableau SRAM de 256x128 cellules (figure 2.4). Ces dimensions sont typiques dans l'étude des tableaux SRAM ([26], [9], [4],...). Dans ce circuit, la cellule de test (*Cell.*) est connectée à ses BL d'écriture (BL_W et $B\bar{L}_W$) et de lecture (BL_R et $B\bar{L}_R$), et à ses WL d'écriture (WL_W) et de lecture (WL_R). Comme expliqué dans la section 1.1, toutes ces grandes lignes d'interconnexion sont connectées à un grand nombre de transistors, et forment une capacité parasite de routage avec le substrat du circuit. Les autres transistors sont modélisés par un transistor unique, et la résistance des vias est négligée. Pour les WL, la grille de ce transistor a des dimensions 128 fois supérieures à la grille totale des transistors d'accès de la cellule étudiée. Les capacités parasites de bord sont négligées, mais la configuration est telle que la capacité de grille est maximale ($V_s = V_d = 0V$). Pour les BL, l'unique transistor a une largeur de grille 256 fois supérieure à celle des transistors d'accès de la cellule étudiée. La capacité de routage formée avec le substrat est modélisée

par une capacité idéale connectée à la ligne d'interconnexion ($C_{WL,rout.}$ et $C_{BL,rout.}$). Leur valeur sera déterminée dans la section 2.3.1. Les BL de lecture sont chargées par un transistor PMOS commandé par le signal de lecture ($ReadEn$). Les autres WL et BL sont chargées et déchargées par un *driver*. Un *driver* est composé de deux inverseurs en série, dont le rapport $\frac{W}{L}$ est croissant. La valeur de ce rapport sera déterminée dans la section 2.3.2.

2.3.1 Les capacités de routage

Les BL et WL sont de grandes lignes de métal parcourant tout le tableau de la mémoire. Un couplage capacitif se crée avec les autres lignes de métal et le substrat du circuit. La capacité induite peut devenir non négligeable par rapport aux capacités de jonction et de grille des transistors. Ces capacités sont extraites du *layout* d'un tableau à 256 lignes et 128 colonnes. Les cellules du tableau sont les 6T classiques en technologie 65nm CMOS Bulk provenant de l'industrie (STMicroelectronics). Au vu du *layout*, nous posons l'hypothèse que seul le couplage capacitif avec le substrat est significatif. Nous approximons la capacité totale par la mise en série de capacités plans idéales, dont les dimensions géométriques et la permittivité relative sont données dans le *layout*. On obtient alors

$$C_{BL,rout.} = 0.79fF$$
$$C_{WL,rout.} = 1fF$$
(technologie 65 nm)

Une capacité plan idéale est donnée par la formule $C = \frac{\epsilon S}{d}$, où ϵ est la permittivité du milieu, S la surface de la capacité et d la distance entre les deux plaques conductrices. Nous posons alors l'hypothèse que *toutes* les dimensions géométriques seront divisées par deux en passant à une technologie 32nm CMOS SOI. Les capacités de routage seront donc celles obtenues en technologie 65nm divisées par deux. Connaissant la capacité de grille C_g et de jonction C_j des transistors (tableau 2.2), nous pouvons retrouver la proportion de la capacité de routage par rapport à la capacité totale de la ligne :

$$\frac{C_{BL,rout.}}{C_{BL}} = \frac{0.395}{256 \cdot 0.06 + 0.395} = 0.025 \tag{2.1}$$
$$\frac{C_{WL,rout.}}{C_{WL}} = \frac{0.5}{128 \cdot 2 \cdot 0.09 + 0.5} = 0.021 \tag{2.2}$$

L'hypothèse de capacités de routage divisées par deux en passant d'une technologie 65nm Bulk à 32nm SOI est pessimiste, car l'oxyde enterré de la technologie SOI diminue fortement la capacité de couplage avec le substrat.

2.3.2 Les drivers

Les dimensions des drivers qui chargent et déchargent les BL et WL déterminent le temps d'écriture et de lecture. Cette section justifie leur choix avec la cellule 8T SVt en banc de test. Elles seront ensuite utilisées pour comparer toutes les cellules étudiées dans ce travail.

Tous les transistors des drivers seront de type SVt, afin de présenter un testbench standard facilement comparable à d'autres technologies. Pour les buffers, les dimensions du premier inverseur seront $\frac{W}{L} = 4$, avec L minimal. Fixer cette dimension, typique en logique CMOS, permet de n'avoir plus qu'un seul degré de liberté. Cette dimension relativement faible par rapport aux

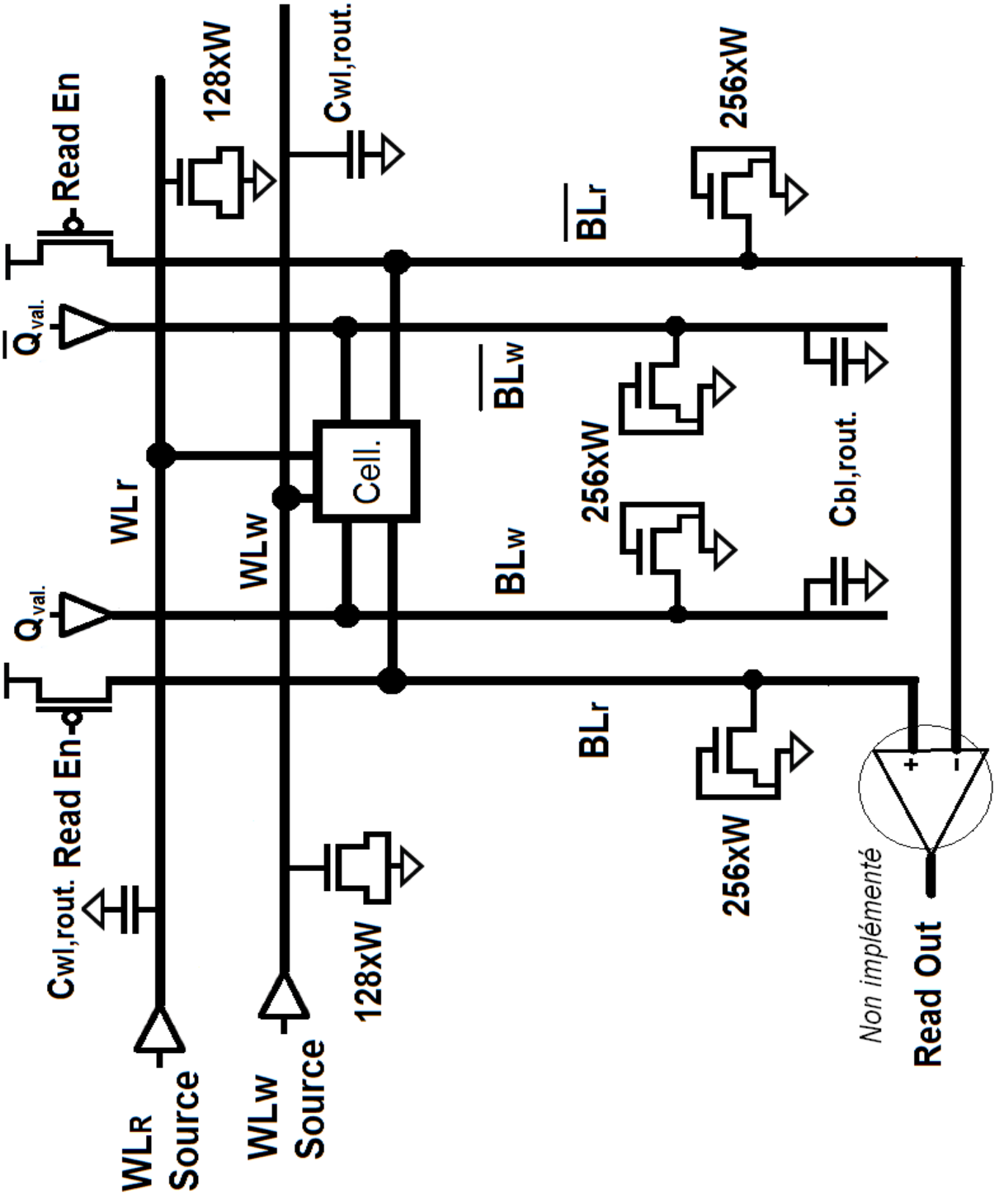


Figure 2.4 – Le testbench pour l'étude et la comparaison des cellules.

dimensions minimales (voir section 2.1) sera justifiée à la fin de cette section. Pour symétriser les inverseurs, le rapport entre les PMOS et NMOS sera de 2 car c'est approximativement le rapport des courants (voir tableau 2.1).

Les PMOS qui préchargent les BL de lecture auront également des dimensions $\frac{W}{L} = 4$, car elles suffisent à recharger la BL avant la fin du temps de lecture (section 2.4.2). Une discussion sur leur dimensionnement est faite dans le chapitre 5.

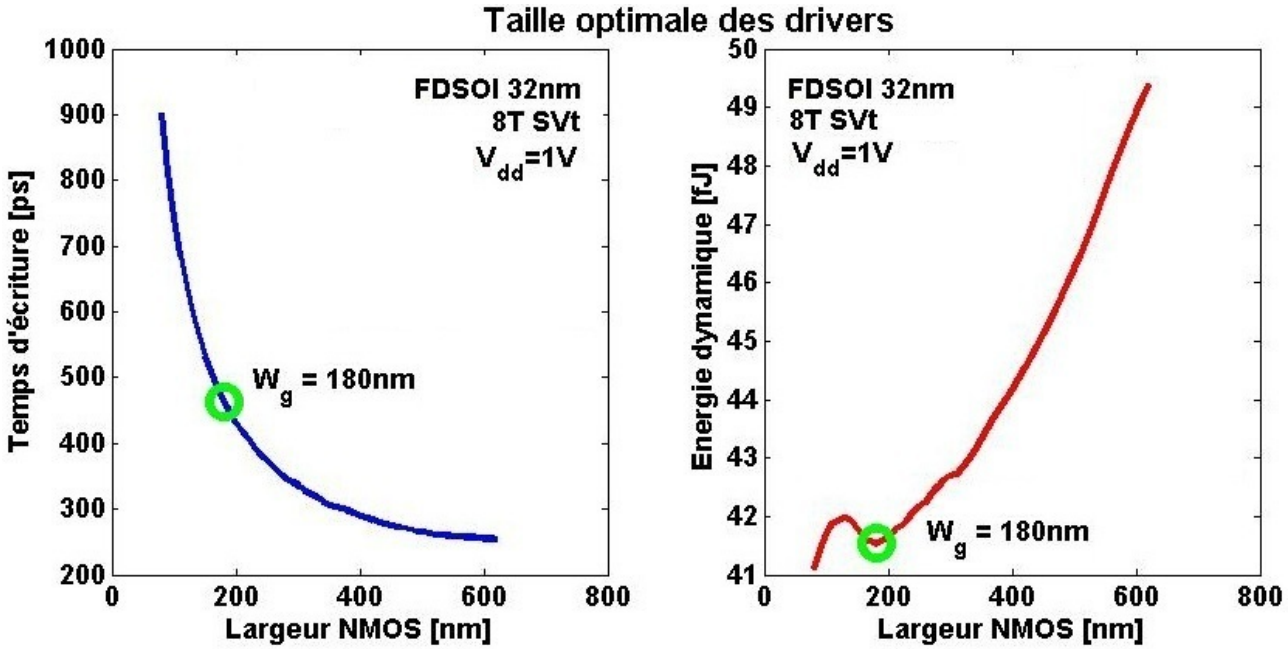


Figure 2.5 – Temps d'accès et énergie dynamique lors d'une écriture en fonction de la taille des drivers (la cellule test est la 8T *dual-port* conventionnelle avec tous les transistors SVt).

La figure 2.5 montre les résultats de simulation obtenus pour le temps d'écriture de la donnée et l'énergie dissipée durant un intervalle de temps fixe[1]. Comme les périphériques externes ne sont pas pris en compte, le temps d'écriture doit être suffisamment inférieur à la période cible (1ns pour une fréquence de fonctionnement de 1GHz). On remarque alors qu'il existe un optimum énergétique qui respecte cette contrainte de temps.

En conclusion, les dimensions des drivers seront :

$L_g = 30nm$
$W_{g,NMOS} = 120nm$ et $W_{g,PMOS} = 240nm$ pour le premier étage des drivers,
$W_{g,NMOS} = 180nm$ et $W_{g,PMOS} = 360nm$ pour le second étage des drivers,
$W_{g,PMOS} = 120nm$ pour le transistor de précharge de la BL de lecture.

L'optimum étant atteint pour un rapport $\frac{W}{L} = 6$, on justifie *a posteriori* le rapport de 4 pour le premier inverseur des *drivers*.

La charge des BL et des WL par les *drivers* est équivalente, en première approximation, à la charge d'une capacité dans un circuit RC. Or, la constante de temps de ce circuit est proportionnelle à la capacité totale, donc au nombre de transistors connectés à la ligne d'interconnexion. Dans le but d'évaluer rapidement et facilement les performances dynamiques dans

1. Définitions exactes dans la section 2.4.1 et 2.4.3

un autre tableau, voici les temps de charge et décharge à 95% de V_{dd} des BL et WL, pour un tableau 256×128 de cellules 8T classiques avec les transistors d'accès SVt :

$$\tau_{BL} = 140ps$$
$$\tau_{WL} = 220ps$$

2.4 Performances de cellule SRAM

Dans cette section, la mesure des performances des cellules - la vitesse, la consommation et la stabilité - sera rigoureusement définie.

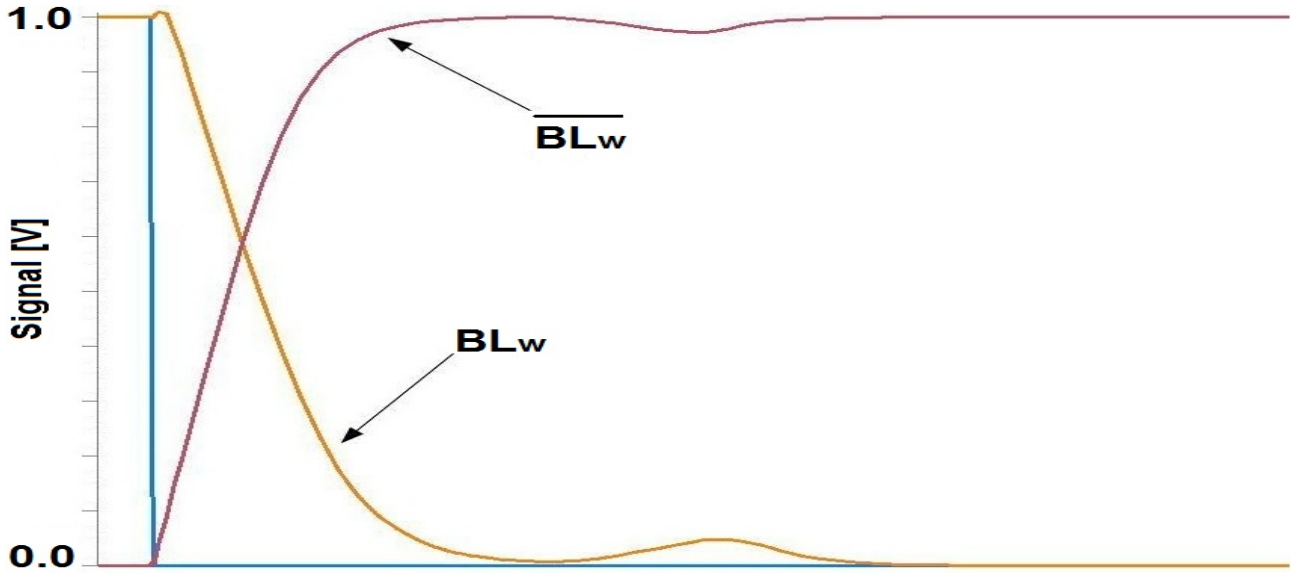
2.4.1 Temps d'écriture

Lors d'une écriture, la tension représentant la valeur de la donnée est modélisée par un *pulse* sur l'entrée du driver de la BL, et cette dernière est préchargée (figure 2.6(a)). Lorsque celle-ci arrive à 95% de sa valeur finale, la WL est activée par un *pulse* à l'entrée du *driver* ($WL_W Source$ figure 2.4). La WL se charge pour permettre l'ouverture des transistors d'accès et l'écriture effective, puis se décharge pour couper les transistors d'accès et permettre une prochaine écriture (figure 2.6(b)).

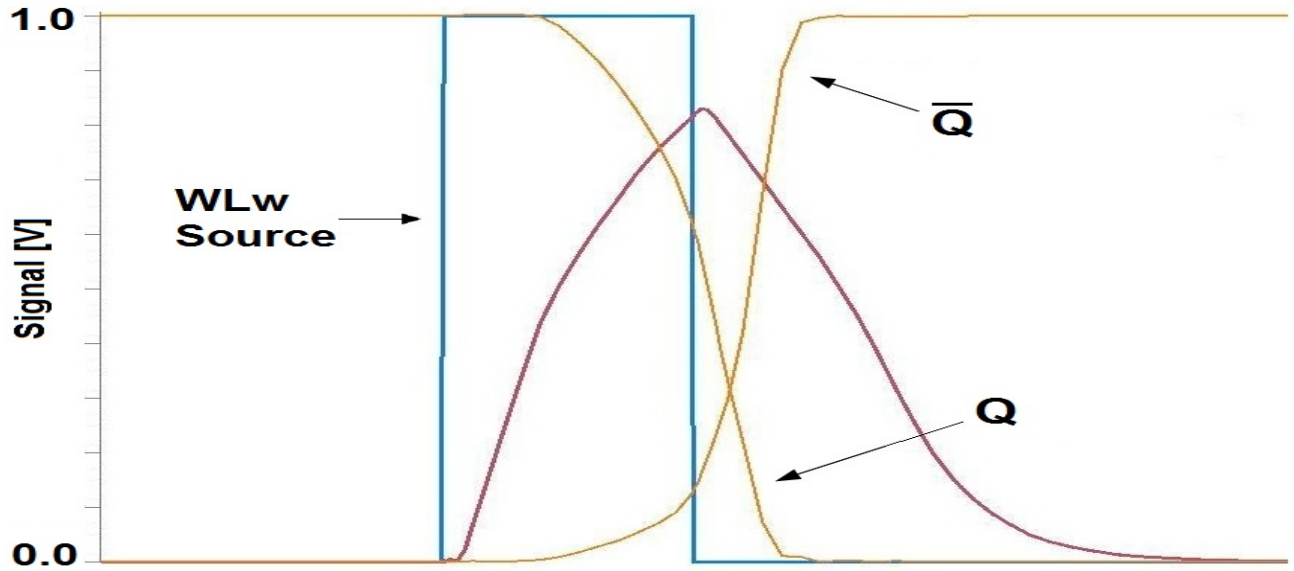
Le temps d'accès en écriture est l'intervalle de temps minimal, défini entre le début du *pulse* sur le driver de la BL et l'instant où la WL a atteint 5% de Vdd, qui permet une écriture effective de la donnée (voir figure 2.6(c)).

Comme on peut le voir, les BL et WL sont chargées de manière séquentielle, comme suggéré dans [25], [39] et [27]. Ceci permet de ne pas dépendre à la fois du nombre de transistors par colonne et par ligne.

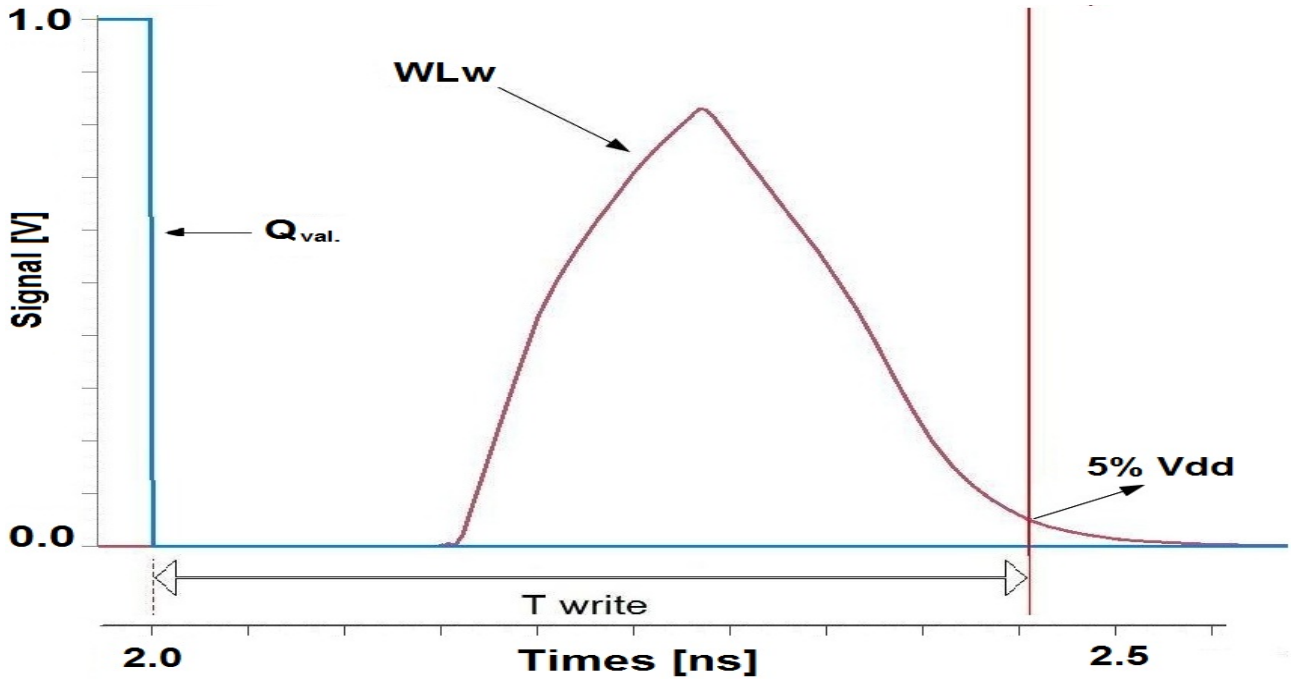
Pour rappel, le temps de propagation de l'arrivée des données à l'entrée de la mémoire jusqu'à la sortie du décodeur n'est pas pris en compte.



(a) Précharge des BitLines.



(b) Les signaux Q et $\bar{Q}$ représentent les noeuds de rétention de la cellule.



(c) Définition du temps d'écriture.

Figure 2.6 – Opération d'écriture.

2.4.2 Temps d'accès en lecture

Lors d'une lecture, la WL est activée par un *pulse* à l'entrée du *driver*, et le transistor PMOS de précharge est coupé par un *pulse* identique sur sa grille.

La WL se charge permettant l'ouverture du transistor d'accès et donc la décharge de la BL, puis se décharge pour permettre une future lecture.

Le temps d'accès en lecture est l'intervalle de temps minimal, défini entre le début du *pulse* sur le driver de la WL et l'instant où celle-ci a atteint 5% de Vdd, qui permet une différence effective de **100mV** entre les deux BLs - ou une chute effective de $100mV$ de la BL lorsqu'il n'y en a qu'une (figure 2.7).

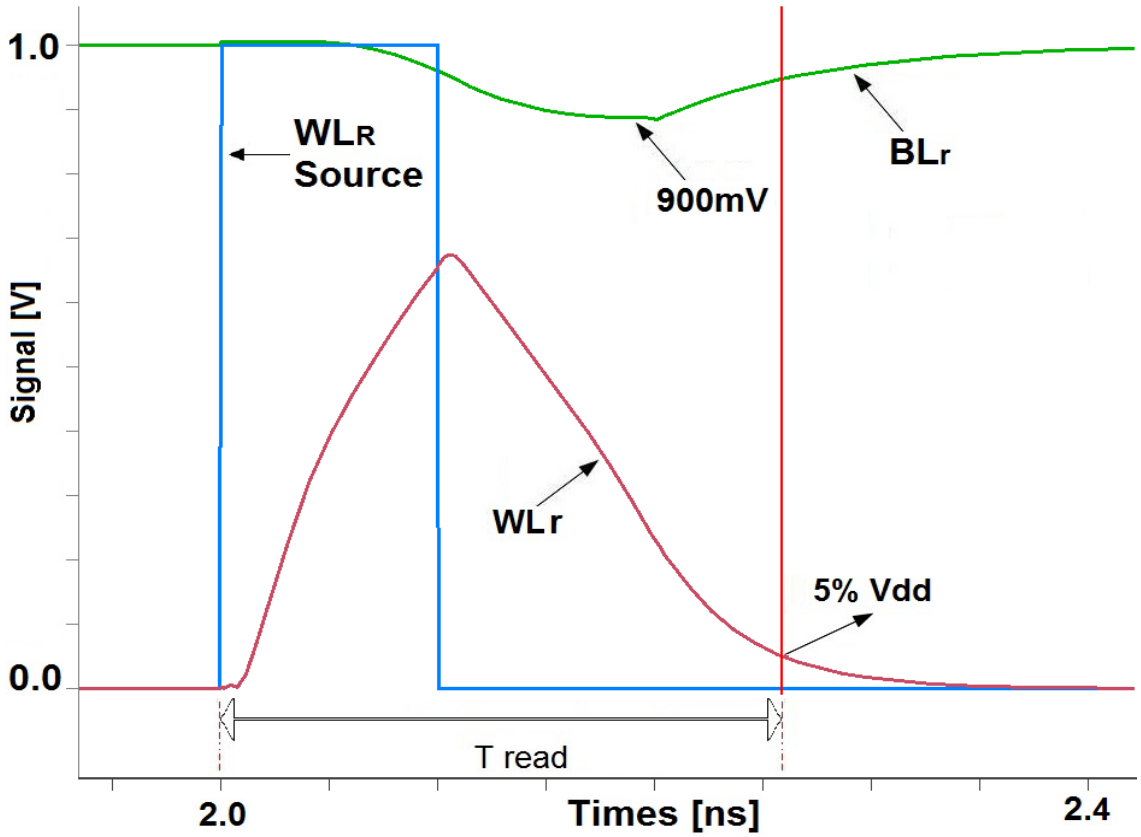


Figure 2.7 – Opération de lecture.

Pour rappel, le temps de propagation de l'arrivée des données à l'entrée de la mémoire jusqu'à la sortie du décodeur, ainsi que le temps de propagation entre l'entrée du *sense-amplifier* et la sortie des données ne sont pas pris en compte.

2.4.3 L'énergie dynamique

L'énergie dynamique par opération est calculée par l'intégration de la puissance débitée de la source d'alimentation sur une période de 1ns, englobant soit une écriture, soit un accès en lecture. On remarque que le courant de fuite intégré sur une période de 1ns est négligeable par rapport à l'énergie de charge des capacités. Donc, en première approximation, l'énergie dynamique sera l'énergie nécessaire à la charge des capacités de BL et WL, ainsi que des cellules accédées.

Les résultats donnés seront l'énergie dynamique moyenne par cellule, dans le but de permettre une transposition aisée des résultats sur un autre tableau mémoire. Energie "moyenne"

signifie que l'énergie sera mesurée lors de l'écriture ou la lecture d'un 1 et d'un 0 logique, pour ensuite être divisée par deux. Pour obtenir la valeur de l'énergie par cellule, il convient de diviser l'énergie fournie à la WL par le nombre de cellules accédées. L'énergie dynamique sera donc donnée par la formule

$$E_{dyn} = V_{dd}\frac{1}{2}\int_0^{t_{op.}} (I_{dd,Cell} + I_{dd,BL} + \frac{I_{dd,WL}}{128})_{op.0} + (I_{dd,Cell} + I_{dd,BL} + \frac{I_{dd,WL}}{128})_{op.1} \qquad (2.3)$$

où I_{dd} est le courant provenant de la source d'alimentation de la cellule, des BL ou des WL, et $op.x$ signifie une opération (écriture ou lecture) avec une donnée binaire x (0 ou 1).
Ce calcul implique quelques imperfections. Dans les simulations, la cellule et les BL contiennent toujours une valeur logique différente de celle de la prochaine écriture. Donc, si le résultat de l'équation 2.3 est multiplié par 128, on considère un taux de commutation de 100%, c'est-à-dire que *toutes* les cellules de la rangée contenaient l'inverse binaire de leur nouvelle valeur. En multipliant par 64, pour représenter un taux de commutation de 50%, le résultat semble plus plausible, mais on introduit alors une erreur sur la contribution énergétique de la WL. Dans l'annexe C, nous montrons que cette erreur reste inférieure à 2%.

2.4.4 Puissance statique

Elle est définie par la puissance moyenne par cellule consommée lorsque aucune opération n'est en cours.

Pour extraire cette valeur, 256 cellules sont connectées sur une BL : 128 contiennent comme donnée un 1 logique, et 128 autres un 0 logique. Les courants issus des alimentations des cellules, des *drivers* et des circuits de précharge, sont ensuite sommés, et cette somme est divisée par 256. Les courants de fuite des drivers décrits dans la section 2.3.2 sont donc pris en compte.

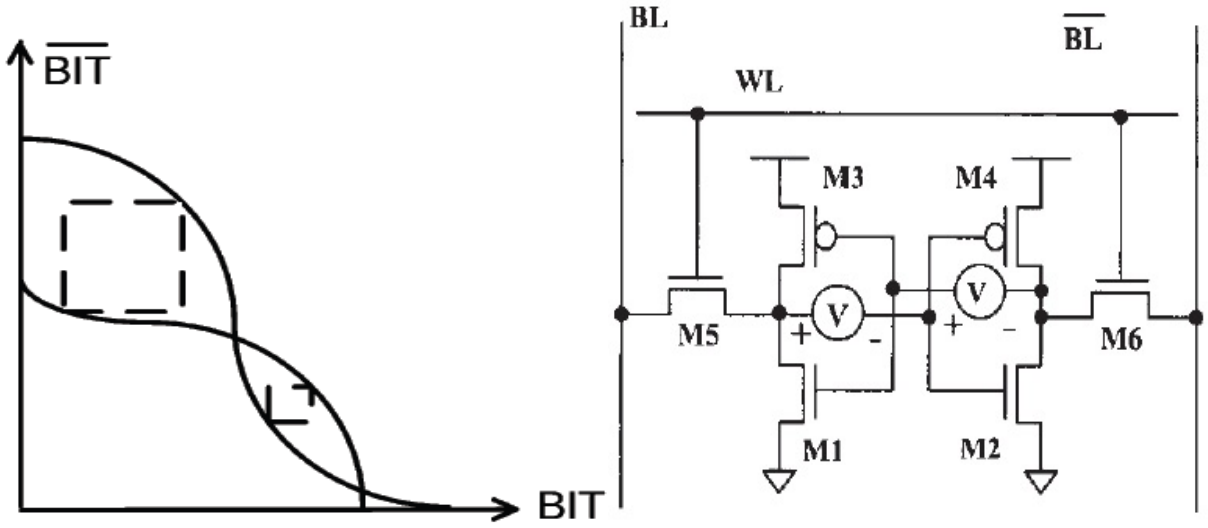
Le courant venant de l'alimentation des drivers des WL n'est pas pris en compte, car les courants de fuite de grilles des transistors d'accès proviendront nécessairement des autres alimentations, puisque V_{WL} est bas. Ces courants sont donc déjà pris en compte dans le résultat final.

2.4.5 Marges de bruit

Le *Static Noise Margin* (ou SNM) est généralement défini par "la tension de bruit minimale qui doit être appliquée au(x) noeud(s) de rétention pour changer l'état de la cellule" [23]. Et ce, pour les trois configurations possibles d'une cellule SRAM : en rétention, en écriture et en lecture.

Cette marge de bruit peut être extraite de deux façons :

- soit en incluant une source de bruit au système V_n (figure 2.8(b)), et en déterminant quel est le niveau de tension minimal faisant basculer l'état de la donnée,
- soit en traçant le graphe dit "papillon" (figure 2.8(a)), et en mesurant la longueur du côté du plus grand carré pouvant être inscrit entre les deux courbes.



(a) Le graphique papillon [40].

(b) Les sources de bruit aux noeuds de rétention de la cellule [10].

Figure 2.8 – Techniques d'extraction du SNM pour les cellules SRAM.

Il apparait alors clairement, que si la marge de bruit est nulle ou négative, l'état de la cellule sera modifié et cette dernière n'est pas fonctionnelle. Une marge de bruit positive ne signifie pas que toutes les cellules du tableau seront fonctionnelles, car il faut tenir compte de la variabilité (section 1.5). Qualitativement, plus le SNM est grand, plus la probabilité qu'une cellule soit inopérante diminue. Mais seules les simulations Monte-Carlo peuvent déterminer s'il est *quantum satis*.

Dans ce travail, nous utiliseraons la méthode d'extraction par source de bruit. Dans le cas des cellules à une seule BL, l'asymétrie du système peut conduire à une marge de bruit différente selon la valeur de la BL. Le pire des cas est alors à prendre en considération.

Les opérations d'écriture et de lecture sont dynamiques. Il faudrait donc, en toute rigueur, mesurer la marge de bruit dynamique, ou *Dynamic Noise Margin* (DNM) [41]. La méthode d'extraction statique sera cependant utilisée pour notre propos, car les marges de bruit alors obtenues seraient supérieures à celles mesurées en dynamique [2], ce qui donne une marge de sécurité supplémentaire.

Pour illustrer nos propos, nous effectuons une simulation Monte-Carlo (400 itérations) de la cellule *dual-port* 8T conventionnelle, avec un cell ratio de 2 et des transistors SVt, dans la technologie FDSOI 32nm considérée dans ce travail. Comme on peut le constater sur la figure 2.9, cette distribution est proche d'une loi normale, dont on peut extraire les deux principales caractéristiques, moyenne et écart-type, par des méthodes statistiques.

En conclusion, pour $W_{NMOS} = 2W_{min}$, la cellule est bien robuste à 6σ face au mismatch, ce qui valide *a posteriori* le choix de la section 2.2. La moyenne de la marge de bruit n'est pas celle obtenue dans le cas nominal, preuve que la dépendance à la tension de seuil n'est pas triviale.

Cette méthodologie sera utilisée dans la suite du travail pour toutes les caractéristiques critiques des cellules, où le pire des cas doit être pris en compte.

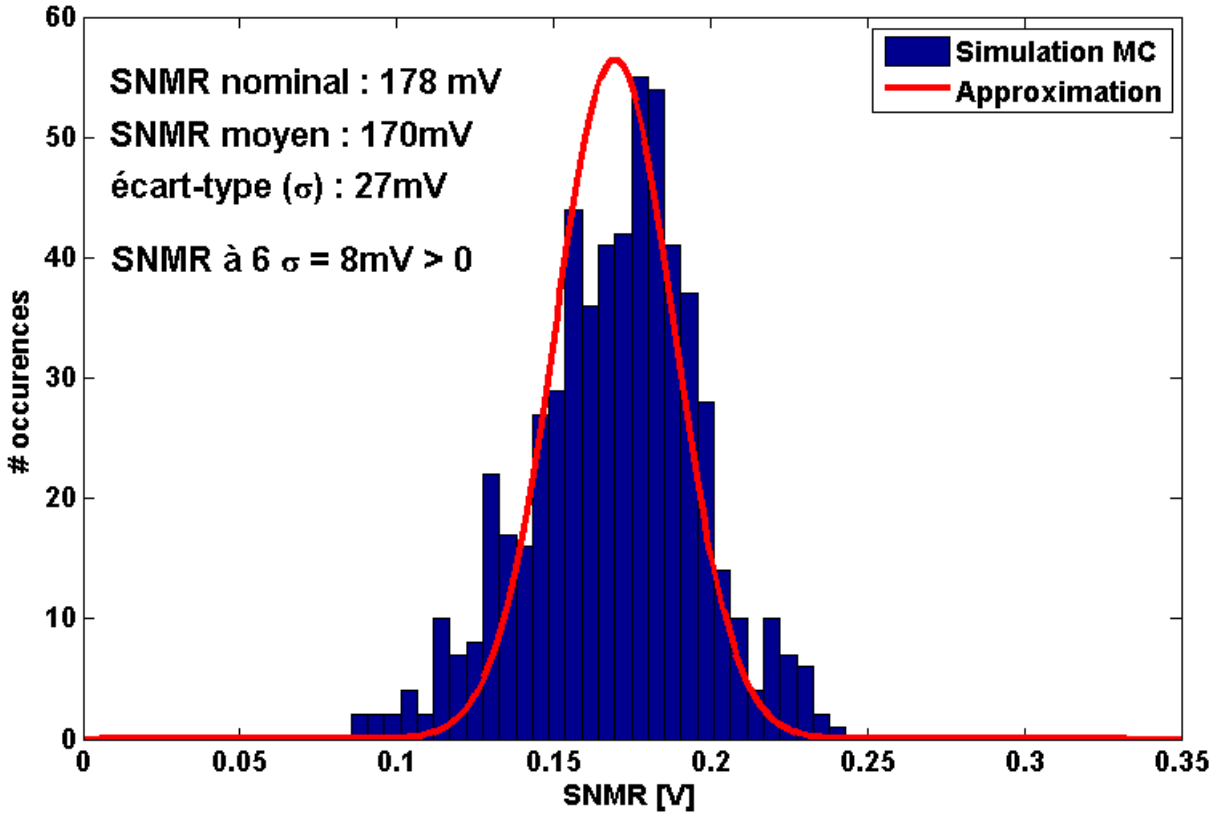


Figure 2.9 – Résultats de simulations Monte-Carlo à 400 itérations pour le $SNMR$ de la cellule 8T SVt définie dans la section 2.2 (*cell ratio* = 2).

2.4.6 Nombre maximal de cellules par BitLine

Quand un très grand nombre de cellules est connecté à une BitLine, le courant de fuite total de toutes les cellules non accédées peut devenir problématique. S'il devient équivalent au courant passant par le transistor d'accès de la cellule accédée I_{read}, il se peut que le tableau n'ait plus le même comportement vis-à-vis des périphériques. Donc, indépendamment des contraintes de vitesse, un nombre infiniment grand de cellules ne peut être connecté aux BL pour des raisons de fiabilité.

Le cas le plus critique à considérer est la lecture d'une cellule ; le nombre maximal de cellules par BL est alors donné par le rapport du courant I_{read} sur le courant I_{off} [9], avec une certaine marge de sécurité.
En effet, si le chemin de décharge est coupé, la tension de la BL devrait idéalement rester à la tension d'alimentation V_{dd}. Néanmoins, comme le transistor PMOS de précharge est coupé, les courants de fuite des cellules de la colonne entrainent une légère diminution ΔV_{BL} de la tension de BL. Il est donc nécessaire de s'assurer que la valeur de ΔV_{BL} reste inférieure à la différence de tension minimale du basculement du *sense-amplifier* connecté à la BL. La somme des courants de fuite doit donc rester inférieure au courant I_{read}, qui lui doit entrainer la BL à la différence de tension de basculement du *sense-amplifier* (figure 2.10).

Plus précisément, le courant I_{read} est le courant statique minimum qui décharge la BL, lorsque sa tension vaut $Vdd - V_{sense\ ampli} = 1V - 0.1V$ dans ce travail. Le courant I_{off} est le courant statique maximal passant par un transistor d'accès fermé, lorsque $V_{BL} = V_{dd} - V_{sense\ ampli} = 0.9V$ dans ce travail.

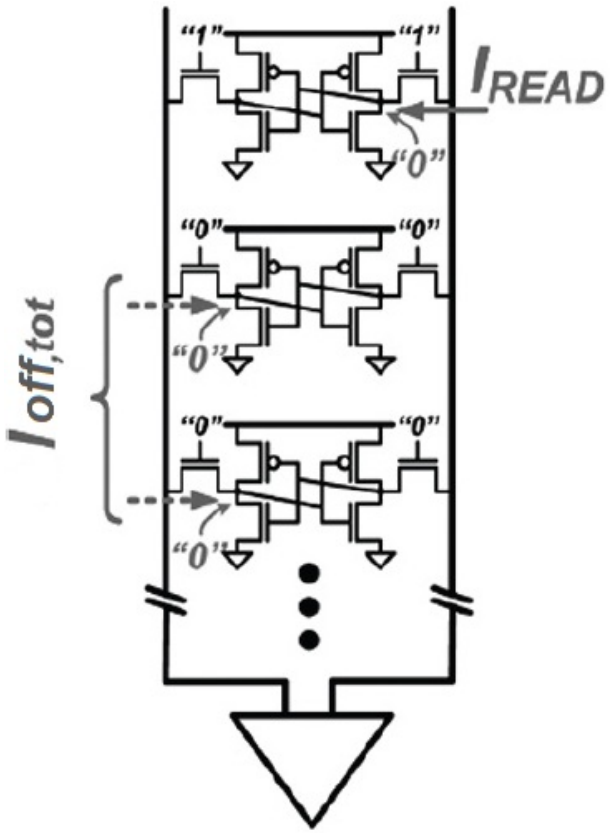


Figure 2.10 – Le rapport I_{read}/I_{off} détermine le nombre maximal de cellules par colonne [9].

Chapitre 3

Comparaison : théorie et simulations

Dans ce chapitre seront étudiées les architectures de cellules qui semblent être adaptées à une application *dual-port*. Le degré de liberté des *assists* statiques est utilisé. La variabilité globale et locale fait partie intégrante de l'étude.

La première section commence par analyser plus en détails les contraintes de dimensionnement de la cellule de l'état de l'art, la cellule 8T conventionnelle. De cette analyse, la démarche d'étude de la suite du chapitre est dégagée : l'utilisation des systèmes de lecture appelés *buffers* permet un découplage parfait entre l'écriture et la lecture, et pourrait améliorer les performances de la cellule *dual-port* de l'état de l'art.

La deuxième section étudie uniquement les *latchs* et systèmes d'écriture, sans aucune mention du système de lecture. Les marges de bruit en rétention et en écriture sont étudiées cellule par cellule, en tenant compte de la variabilité. Ensuite, la vitesse d'écriture et la consommation statique et dynamique sont comparées. La comparaison s'effectue dans le même *corner*, et la variabilité est ensuite prise en compte. Cette section montre finalement que la cellule 5TPMOS à transistors d'accès SVt présente la plus basse consommation tout en respectant les contraintes de vitesse et de fiabilité.

La troisième section étudie uniquement les systèmes de lecture. La vitesse de lecture, la consommation statique et dynamique, ainsi que le nombre maximal de cellules par BL sont étudiés et comparés. Cette section montre enfin que le *buffer* de lecture 2T HVt-HVt présente le meilleur compromis entre consommation, vitesse et rapport I_{read}/I_{off}.

Finalement, la conclusion compare la cellule de l'état de l'art avec la cellule proposée - 5TPMOS avec transistor d'accès SVt et *buffer* 2T HVt-HVt - et montre l'avantage de cette dernière en terme de consommation et de surface de silicium.

Toutes les simulations de ce chapitre sont effectuées à une température de 27°C, sous 1V d'alimentation.

3.1 Motivations

Au vue de la symétrie de la cellule 8T classique (figure 3.1), il y a quatre paires de transistors qui doivent être dimensionnées :
- les transistors NMOS du *latch* (M1 et M2),
- les transistors PMOS du *latch* (M3 et M4),
- les transistors d'accès en écriture (M5 et M6),
- et en lecture (M7 et M8).

Chacune des 4 paires peut avoir 3 tensions de seuil et des dimensions géométriques différentes. Plutôt que de passer en revue les 3^4 possibilités de tensions de seuil, plus le degré de liberté du dimensionnement, attardons-nous un instant sur les degrés de liberté disponibles pour le dimensionnement.

Une augmentation de la largeur de grille W_g des transistors d'accès en lecture augmenterait proportionnellement la capacité des BL et des WL. Mais, pour une taille identique des transistors NMOS de pied (M1 et M2), le courant de décharge n'augmenterait pas de manière proportionnelle (montré en section 3.3.1). Or, le temps de charge ou de décharge Δt d'une capacité C_{tot} est donné, en première approximation, par

$$\Delta t = \frac{C_{tot}\Delta V}{I}$$

où ΔV est la différence de tension sur la capacité produite par le courant I. Les transistors d'accès en lecture doivent donc être de taille minimale, car elle garantit une vitesse maximale pour une surface minimale.

Le temps d'écriture dépend du temps de charge de la BL et de la WL. Augmenter la largeur de grille des transistors d'accès en écriture augmentera également ces temps de charge. Le courant de décharge des noeuds de rétention sera légèrement augmenté, mais sa contribution n'est pas suffisante pour apporter un gain sur le temps total d'écriture (montré en section 3.2.7). Les transistors d'accès en écriture doivent donc également être de taille minimale.

Ces dimensions fixées, analysons les interactions avec les transistors du latch. Pour l'écriture, les transistors PMOS du *latch* M3 et M4 peuvent être les plus "faibles" possible - taille minimale et de type HVt - pour faciliter le travail des transistors d'accès M5 et M6. La tension de seuil, V_t, des transistors d'accès est un compromis entre la consommation et la vitesse d'écriture. Pour la lecture, le dimensionnement se complique. Supposons que les transistors d'accès en lecture M7 et M8 soient les plus faibles possibles : taille minimale et de type HVt (ce qui diminuerait fortement la vitesse de lecture). Les NMOS de pieds M1 et M2 ne peuvent pas être identiques aux transistors d'accès, pour des raisons de fiabilité (SNMR). Soit on augmente leur V_t, ce qui signifie plus de consommation statique, soit on augmente leur W_g, ce qui signifie plus de surface de silicium, et un peu plus de consommation statique. Si on souhaite augmenter la vitesse de lecture, il faut diminuer la tension de seuil des transistors d'accès, et faire de même avec les transistors NMOS du *latch* afin de préserver le SNMR. En augmentant ainsi la force des NMOS du *latch*, on diminue la tension de basculement des inverseurs, et donc la marge de bruit en

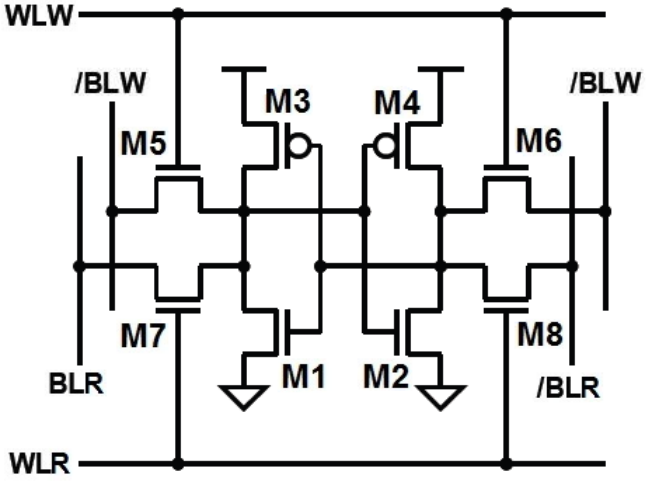


Figure 3.1 – Schéma d'une cellule *dual-port* conventionnelle à huit transistors [27].

lecture ; on pourrait alors augmenter la force du PMOS, ce qui diminuerait la marge de bruit et la vitesse d'écriture et augmenterait la consommation statique.

Cette réflexion théorique sur le dimensionnement montre que toutes les contraintes de la cellule classique à 8 transistors proviennent de son système de lecture. La conception de cette cellule introduit toute une série de compromis, qui amènent bien souvent à sacrifier la consommation au profit de la fiabilité et/ou de la vitesse de fonctionnement. Tous ces problèmes pourraient être contournés si les tensions des noeuds de rétention n'étaient pas sur le chemin de décharge de la BL de lecture. Dans les architectures avec un *buffer* de lecture, l'information de la donnée sauvegardée dans la cellule est contenue sur la grille d'un transistor extérieur au *latch*. Ainsi, pratiquement toutes les contraintes sont contournées.

Ce travail partira donc sur la piste de l'utilisation des *buffers* de lecture, et analysera si les performances des cellules SRAM *dual-port* avec *buffer* sont supérieures ou non à celles des cellules classiques, pour une même vitesse de fonctionnement. A ce stade, on pourrait arguer que l'ajout d'un "système" supplémentaire entrainera nécessairement une surface de silicium par cellule plus importante. Or cette dernière est critique dans une application SRAM. La conclusion de ce chapitre reviendra sur cette question.

3.2 Latch et système d'écriture

Dans cette section, différents *latchs* et systèmes d'écriture proposés dans la littérature seront passés en revue. Pour chacun d'entre eux, les marges de bruit en rétention et en écriture seront d'abord étudiées par simulation. Ceci afin de sélectionner les cellules à rejeter des simulations dynamiques. Pour chaque cellule, quelques réflexions théoriques seront émises sur la vitesse, la consommation et la surface de silicium, afin de mieux interpréter et comparer les résultats de simulation obtenus. Finalement, les cellules seront validées face à la variabilité. Pour rappel, la tension d'alimentation V_{dd} des cellules et du tableau mémoire est de 1V, et la température de simulation de 27°C.

3.2.1 La cellule 6T dédiée à l'écriture

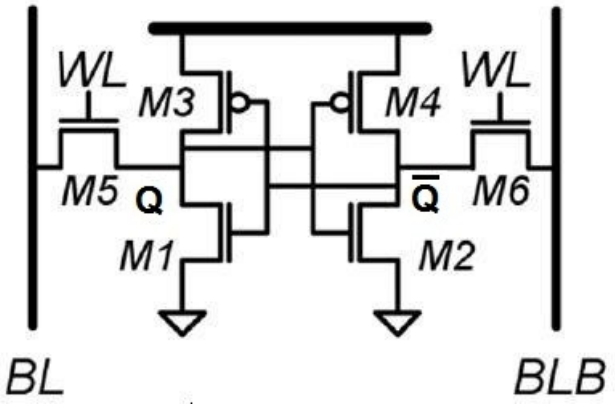


Figure 3.2 – Schéma d'une cellule à six transistors [9].

Cette cellule est architecturalement identique à la cellule 6T classique rencontrée dans la section 1.2.1. Néanmoins, ces deux BL sont ici dédiées uniquement à l'écriture. Ce qui amène à un dimensionnement beaucoup plus simple.

Comme nous l'avons vu dans la section 1.2.1, les transistors d'accès doivent vaincre la contre-réaction du latch, représentée physiquement par le courant I_{on} des inverseurs. Les quatre transistors du *latch* peuvent donc être de taille minimale et de type HVt. De plus, ces paramètres physiques entrainent une consommation minimale pour le *latch* de la cellule. Les deux transistors d'accès auront une taille minimale, pour les mêmes raisons que ceux de la cellule 8T (section 3.1). Et, quelle que soit leur tension de seuil, cette cellule présente une marge de bruit suffisante en rétention et en écriture, et se trouve proche de $V_{dd}/2$ pour la rétention.

Vitesse et consommation

Comme nous l'avons vu dans les sections 2.4.1 et 3.1, pour un *testbench* identique, la vitesse d'écriture des cellules dépend de la capacité de BL, de WL et de l'aptitude du transistor d'accès à forcer le *latch* à changer d'état. Pour cette cellule, la capacité de BL est la capacité de jonction minimale d'un transistor (plus la capacité de routage), et la capacité de WL vaut deux fois la capacité de grille. La vitesse de basculement du *latch* dépend de la tension de seuil des transistors d'accès. Cette dernière sera un compromis entre consommation et vitesse : plus leur tension de seuil est basse, plus vite les transistors d'accès déchargeront le *latch*, mais plus grande sera leur consommation statique. Au sujet de l'énergie dynamique, deux BL sont chargées et déchargées.

Le compromis vitesse/consommation pourrait être contourné par l'utilisation d'un assist statique : la tension de WL négative en rétention. Etudions plus en détail cette possibilité.

Tension de WordLine optimale

Si la tension de WL est négative par rapport à la masse de la cellule et des BL, les transistors d'accès auront dans tous les cas une tension grille-source V_{gs} négative, ce qui diminuera le courant de fuite sous-seuil I_{sub}.
La figure 2.2 nous montre qu'il existe une valeur minimale à ce courant sous-seuil. En effet, en diminuant encore plus la tension de grille, les courants de grille deviennent prédominants, et le courant total de drain augmente. De plus, diminuer la tension de WL augmente l'énergie nécessaire à la charger, c'est-à-dire l'énergie dynamique. Donc, si la valeur de la tension de WL en rétention est trop basse, le circuit consommera au total plus d'énergie qu'il n'en gagne en diminuant le courant sous-seuil des transistors d'accès. Il existe donc une valeur de tension optimale du point de vue énergétique.

Pour déterminer cette tension optimale, quelques hypothèses doivent être posées. Pour chaque transistor d'accès ayant une tension drain-source V_{ds} de 1V, le gain énergétique dû à la réduction du courant sous seuil pour une tension $V_{WL} = -\Delta V$ sur la WordLine est

$$\Delta E_{leak} = V_{dd}(I_{sub}(0) - I_{sub}(-\Delta V))T$$

où T est l'intervalle de temps moyen entre deux accès en écriture. Nous faisons l'hypothèse qu'à la fréquence de fonctionnement du circuit (1GHz), un accès en écriture s'effectue toutes les nanosecondes et que toutes les WL ont la même probabilité d'être accédées. Il en découle $T = 256ns$. Le gain dépend de la relation entre le courant sous-seuil et la tension V_{gs} du transistor d'accès. Or, cette dépendance est hautement sensible à la tension de seuil, donc à la variabilité. Comme un grand nombre de transistors sont connectés à la WL, nous allons considérer que la relation moyenne entre I_{sub} et V_{gs} est donnée par la valeur nominale dans le *corner* considéré.

Table 3.1 – Tension de WL optimale du point de vue énergétique.

Corner	LVt			SVt			HVt	
	ΔV [V]	I_{sub} [pA]	ΔE_{tot} [%]	ΔV [V]	I_{sub} [pA]	ΔE_{tot} [%]	ΔV [V]	I_{sub} [pA]
SF	-0.125	26.8	0.9	0	16	0	0	0.8
SS	-0.14	26.4	0.93	0	21	0	0	1
TT	-0.16	28	0.96	-0.02	24	0.18	0	2.5
FF	-0.18	29	0.98	-0.05	25	0.47	0	5.5
FS	-0.20	29	0.98	-0.06	25	0.58	0	7.8

L'énergie dépensée en surplus lors d'un accès en écriture vaut

$$\Delta E_{Write} = C_{WL}\frac{(V_{dd}+\Delta V)^2}{2} - C_{WL}\frac{V_{dd}^2}{2}$$

où C_{WL} est la capacité totale de la WL. Cette dernière est la somme de la capacité de routage et des capacités de grille de tous les transistors connectés à elle. La capacité de routage étant proportionnelle à la longueur de la WL (section 2.3.1), la capacité totale est proportionnelle au nombre de cellules par WL. Les résultats obtenus ne dépendront donc pas du nombre de cellules par ligne, i.e. des dimensions du tableau.

Si la tension V_{ds} d'un transistor d'accès est nulle, le gain énergétique est nul pour ce transistor. La moitié des transistors d'accès sont considérés avec une tension $V_{ds} = 1V$. Pour les cellules 6T, il y a 256 transistors d'accès connectés à la WL dont 128 avec une tension $V_{ds} = V_{dd}$, et le gain énergétique d'une diminution de la tension de WL en rétention à $-\Delta V$ est

$$\Delta E_{tot} = 128 \cdot (I_{sub}(0) - I_{sub}(-\Delta V))V_{dd}T - C_{WL}\left(\frac{(V_{dd}+\Delta V)^2}{2} - \frac{V_{dd}^2}{2}\right) \tag{3.1}$$

et la tension ΔV optimale est celle qui maximise ΔE_{tot}.

Le tableau 3.1 montre la tension ΔV optimale en fonction de la tension de seuil des transistors d'accès et du *corner* (section 1.5). Comme on peut le voir, les transistors HVt et SVt ne sont pas efficaces d'un point de vue énergétique. Pour que le transistor HVt acquière un gain positif pour une tension ΔV non nulle, il faudrait au moins $T = 5000ns$ dans le corner TT, pour un gain relatif de 12%. Et pour le transistor SVt, il faut au minimum $T = 500ns$ dans le corner SS, pour un gain relatif de 10% (plus de détails en annexe D). Cet *assist* sera donc utilisé, dans le but de sauvegarder de l'énergie, en combinaison avec des transistors LVt uniquement. Notons qu'il y aura toujours un compromis entre vitesse et consommation car, même à la tension de WL optimale, les transistors LVt consomment plus en statique que les SVt et HVt. La figure 3.3 montre le gain relatif des transistors LVt pour les cinq corners, en fonction de ΔV. En fixant une tension de WL à

$$\Delta V_{opt} = -0.125V$$

en rétention, le gain relatif reste proche de 1 dans tous les *corners*. Dans le cas de transistors d'accès SVt et HVt, une tension de WL négative peut avoir de l'intérêt si le tableau mémoire entre en mode veille ; la tension de WL doit alors être fixée pour obtenir un courant de drain minimal I_{min}. Mais cette application sort du cadre de ce travail.

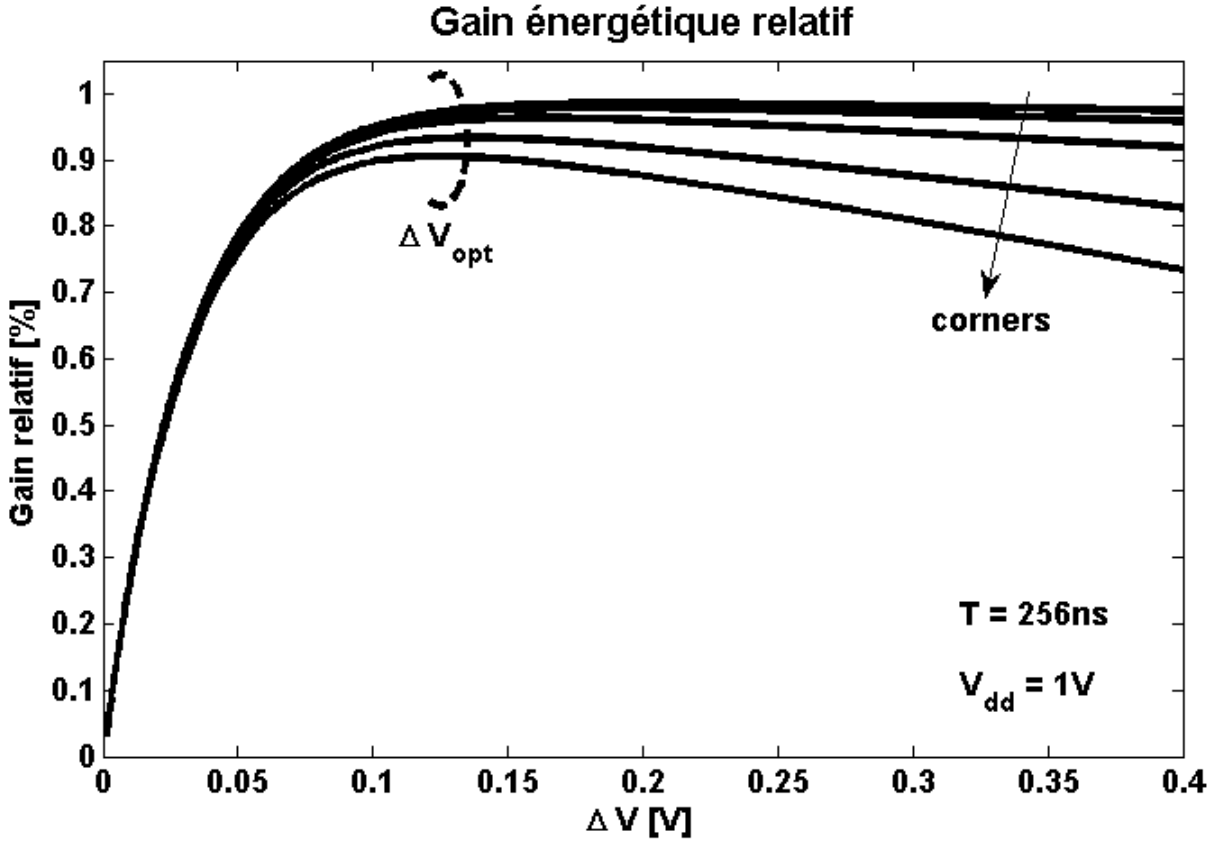


Figure 3.3 – Gain énergétique relatif selon les cinq *corners* des modèles, pour des transistors d'accès LVt.

Durée de vie

Les transistors du *driver* de la WL d'écriture devront supporter une tension plus grande que la tension d'alimentation. Ceci peut influer sur la durée de vie des transistors. La tension d'alimentation nominale étant de 1V pour cette technologie, les transistors du *driver* auront une usure plus rapide que celle des autres transistors du circuit. Une façon de contourner ce problème est de fixer la tension d'alimentation vue par le driver de la WL à $V_{dd} - \Delta V = 0.875V$. L'autre avantage est que le gain énergétique est positif dans tous les cas. Néanmoins, une *quatrième* tension doit pouvoir être générée dans le circuit.

3.2.2 Cellule à trigger de Schmitt simplifié

La principale motivation de l'utilisation de dix transistors est d'augmenter la marge de bruit en lecture. La marge de bruit en rétention, elle, reste aux alentours de $V_{dd}/2$ [12].

Or, l'utilisation d'un *buffer* de lecture permet d'insensibiliser le *latch* au système de lecture, et donc SNMR = SNMH [42]. Cette performance pourrait donc être atteinte en utilisant moins de dix transistors et de surface de silicium.

3.2.3 La cellule 5TPMOS

Comme pour la cellule 6T dédiée à l'écriture, les transistors du *latch* seront de type HVt et de taille minimale. Le transistor PMOS casse la réaction du *latch* pendant l'écriture, ce qui permet d'atteindre un SNMW plus important. L'unique transistor d'accès sera également de

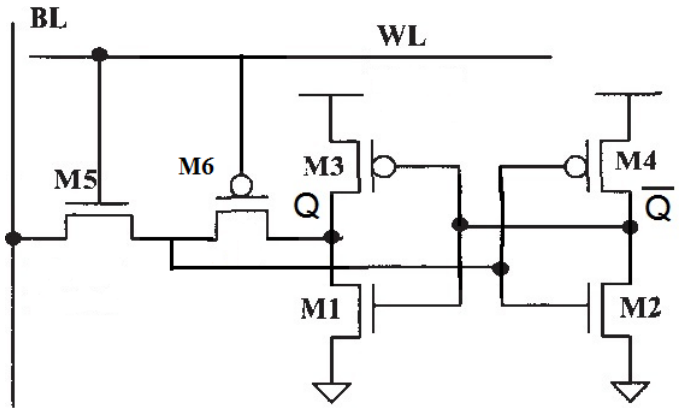


Figure 3.4 – La cellule 5TPMOS [21]

taille minimale, et aura une tension de seuil choisie selon un compromis consommation/vitesse, avec ou sans tension de WL négative pour le transistor LVt.

Vitesse et consommation

La consommation de cette cellule devrait être sensiblement moindre que celle de la cellule 6T. Le transistor PMOS a une même tension de source et de drain en rétention, donc un courant de fuite négligeable. De plus, il n'y a qu'une seule BL à charger et décharger pour écrire dans une cellule, au lieu de deux.

Du point de vue vitesse, il est difficile *a priori* d'identifier la cellule la plus rapide entre la 6T et la 5TPMOS, pour une même tension de seuil des transistors d'accès. Les capacités de BL et de WL sont identiques, avec un petit avantage pour la cellule 5TPMOS car la capacité de grille d'un PMOS est légèrement inférieure à celle d'un NMOS (voir tableau 2.2). Pour la cellule 6T, il y a toujours un transistor d'accès qui a une tension d'*overdrive* ($V_{gs} - V_t$) maximale pour décharger un noeud de rétention, mais il doit d'abord vaincre la réaction du *latch*. Pour la cellule 5TPMOS, la contre-réaction du latch est bloquée et le transistor d'accès doit charger une capacité de jonction en moins. Mais, lors de l'écriture d'un 1 logique, le transistor NMOS d'accès verra sa tension d'*overdrive* progressivement diminuer. Les simulations trancheront et détermineront quel cas de figure est le plus rapide.

Design du transistor PMOS (M6)

La seule contrainte sur ce transistor est qu'il doit présenter un courant de drain négligeable quand il est coupé, ce qui est aisément atteint pour tous les transistors. Une taille minimale permet de ne pas augmenter la capacité de WL, et une tension de seuil HVt le fera passer plus rapidement en régime sous-seuil lors d'une écriture.
La consommation statique de ce transistor provient uniquement du courant de grille, car sa tension drain-source est nulle en rétention. Or, le courant de grille I_g en *forward* est bien minimal pour les transistors PMOS HVt (tableau 2.2). On pourrait penser qu'une tension de seuil plus petite permettrait de régénérer plus vite les tensions du latch à la fin de l'écriture, ce qui diminuerait le temps d'accès. Néanmoins, les simulations dynamiques ont montré que la tension de seuil de ce transistor n'influence en rien la vitesse d'écriture.

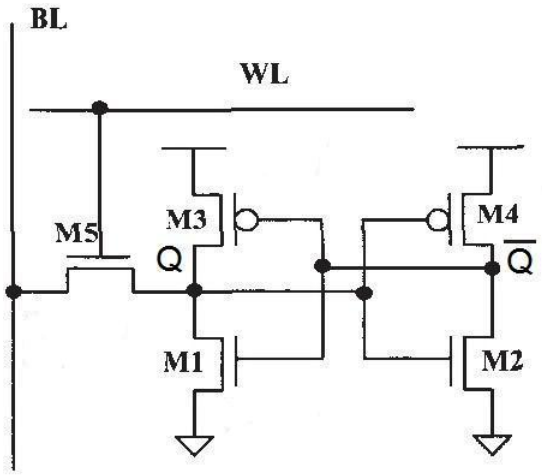


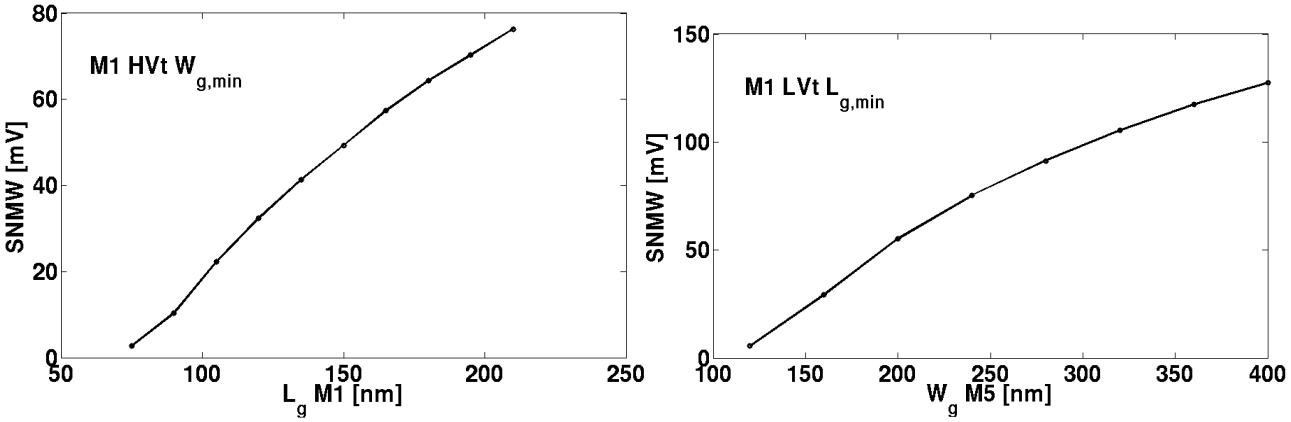
Figure 3.5 – La cellule 5T.

3.2.4 La cellule 5T

Après l'étude des cellules 6T et 5TPMOS, il est naturel de commencer par proposer un *latch* identique. La marge de bruit en rétention sera donc la même. Etudions maintenant la marge de bruit en écriture.

Lors de l'écriture d'un 1 logique se crée un chemin entre la BL, chargée à V_{dd}, et la masse passant par le transistor d'accès et le transistor NMOS du latch (M5-M1 sur figure 3.5). Pour avoir une écriture effective, la tension milieu des deux transistors (celle du noeud Q) doit être suffisante pour faire basculer l'inverseur commandant la valeur de $\bar{Q}$, qui fera alors basculer l'inverseur commandant la valeur de Q. Cette tension milieu est déterminée par la force relative entre les deux transistors composant le chemin de court-circuit. Plus le courant pouvant être débité par M5 est important par rapport à celui de M1, plus la tension sera haute et l'écriture aisée.

Avec un transistor d'accès LVt et le NMOS du *latch* en HVt, tous deux de taille minimale, la tension du noeud Q n'est pas suffisante pour faire basculer le *latch*. Ceci signifie que, quel que soit le temps d'activation de la WL, l'écriture du 1 logique ne se fera jamais.



(a) Pour L_g inférieur à $75nm$, l'écriture ne se produit pas.

(b) Pour W_g inférieur à $120nm$, l'écriture ne se produit pas.

Figure 3.6 – La marge de bruit en écriture en fonction des dimensions des transistors sur le chemin de court-circuit (M1 et M5). La marge de bruit n'est pas suffisante pour des tailles raisonnables de transistors.

Pour pallier ce problème, trois solutions s'offrent à nous :

1. Diminuer la force du transistor NMOS du latch, c'est-à-dire augmenter sa longueur de canal. La figure 3.6(a) montre que, même avec $L_g = 10L_{g,min}$, la marge de bruit en écriture est seulement de 102mV. Pour obtenir une marge de bruit suffisante, la cellule prendrait beaucoup trop de surface de silicium pour notre application SRAM *dual-port* à haute densité.
2. Augmenter la force du transistor d'accès, ce qui est proposé dans [10]. La figure 3.6(b) montre que le SNMW n'atteint que 90mV à $W_g = 4W_{g,min}$. Or, augmenter sa largeur de grille revient à augmenter quasi proportionnellement le temps de charge de la BitLine et de la WordLine d'écriture. La cellule 5TPMOS possède une marge de bruit bien meilleure pour une capacité de BL quatre fois moins grande et une capacité de WL deux fois plus petite, tout en occupant moins de surface de silicium.
3. Diminuer la tension de basculement de l'inverseur M2-M3. Pour ce faire, deux choix s'offrent à nous :
   - soit diminuer la tension d'alimentation du *latch*. Nous y reviendrons dans la section 5.3.
   - soit augmenter la force du NMOS par rapport au PMOS. La figure 3.7 montre que seul un NMOS LVt avec un W_g supérieur à $1.5W_{g,min}$, permet une écriture effective et robuste pour une taille raisonnable.

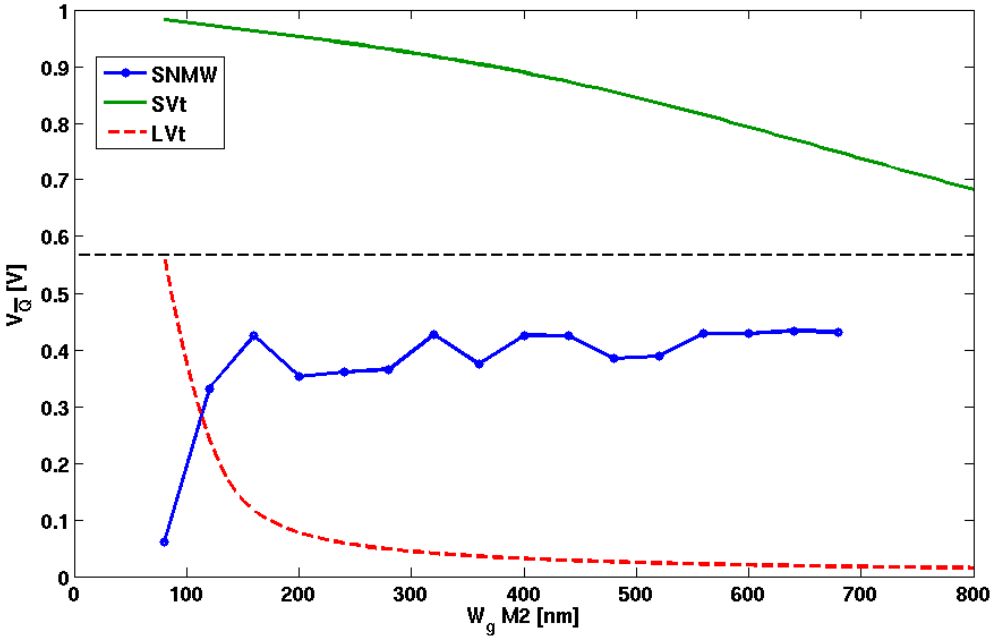


Figure 3.7 – La tension du noeud $\bar{Q}$ lors de l'écriture d'un 1 logique, en fonction de la largeur de grille et de la tension de seuil de **M2**. Le SNMW correspond à la tension obtenue dans le cas LVt. Il appert que seul **M2** LVt avec $W_g = 120nm$ présente une écriture effective et robuste.

En conclusion, une seule configuration permet l'écriture effective et robuste d'un 1 logique face à la variabilité, tout en gardant une surface de silicium acceptable :

- M1 : HVt, $L_g = L_{g,min}$ et $W_g = W_{g,min}$,
- M2 : LVt, $L_g = L_{g,min}$ et $W_g = 1.5W_{g,min}$ (minimum),
- M3 : HVt, $L_g = L_{g,min}$ et $W_g = W_{g,min}$,
- M4 : HVt, $L_g = L_{g,min}$ et $W_g = W_{g,min}$,
- M5 : LVt, $L_g = L_{g,min}$ et $W_g = W_{g,min}$.

L'inverseur M2-M4 ayant été asymétrisé pour permettre l'écriture, la marge de bruit en rétention n'est plus que de 280mV. Pour diminuer la consommation statique, on pourrait choisir d'utiliser un transistor d'accès de type SVt, et d'augmenter la largeur de grille W_g de M2. Des simulations Monte-Carlo d'écriture dynamique ont montré qu'avec $W_{g,M2} = 3W_{g,min}$, l'écriture n'était pas robuste face à la variabilité. Or, cette largeur de grille entraine déjà une consommation plus grande que la cellule proposée.

Vitesse et consommation

Le temps de charge de la WordLine sera approximativement réduit de moitié, ce qui permettra un temps d'écriture plus court. La consommation dynamique sera aussi fortement diminuée par rapport à la cellule 6T, car C_{BL} et C_{WL} sont réduites de moitié (négligeant la capacité de routage).

Cependant, la consommation statique sera beaucoup plus élevée que les *latchs* HVt, à cause de la présence des deux transistors de type LVt. Cette cellule est un compromis extrême entre la vitesse et la consommation.

3.2.5 Cellule de Hobson

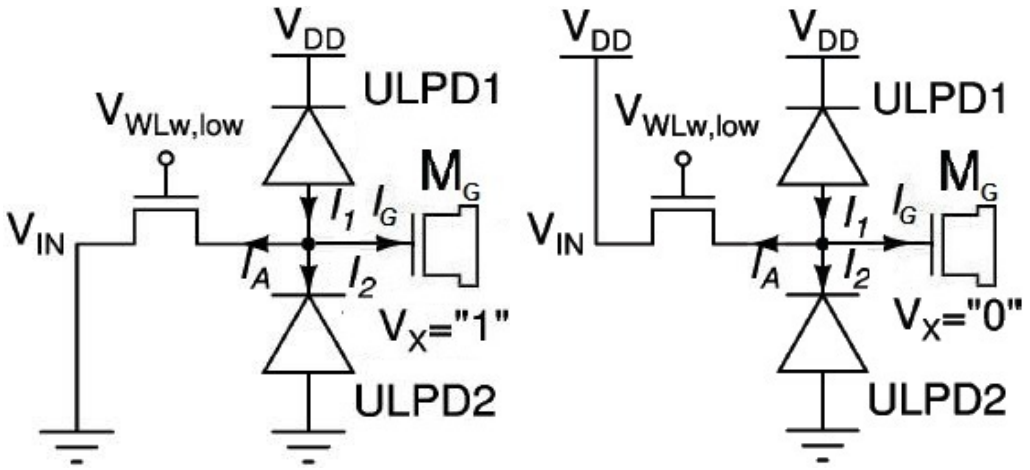
La cellule proposée dans [22] a le même système d'écriture que la cellule 5T. Nous arriverions donc aux mêmes conclusions.

3.2.6 La cellule ULP

On détermine la marge de bruit en rétention en analysant la somme des courants sur le noeud X en fonction de la tension à ce noeud [23]. Posons $I_X = I2 + IA - I1 + Ig$ (figure 3.8(a)). Si I_X est nul, la tension V_X correspondante est un point d'équilibre. Si I_X est positif à la tension V_X, le noeud X se décharge. A contrario, s'il est négatif, le noeud X se charge. Comme illustré à la figure 1.9(b), le *latch* ULP contient deux points d'équilibre stables, et un point d'équilibre instable, appelé tension de basculement dans ce travail en analogie avec les inverseurs montés en tête-bêche. Le SNM est alors déterminé par la plus petite différence entre la tension de basculement, $V_{bascule}$, et un point stable (voir figure 3.8). Ce SNM doit être déterminé quand la tension V_{ds} du transistor du buffer vaut zéro. En effet, pour une certaine tension V_X, le transistor devient beaucoup plus passant que les autres transistors du *buffer* de lecture qui sont coupés. La tension drain-source devient alors pratiquement nulle.

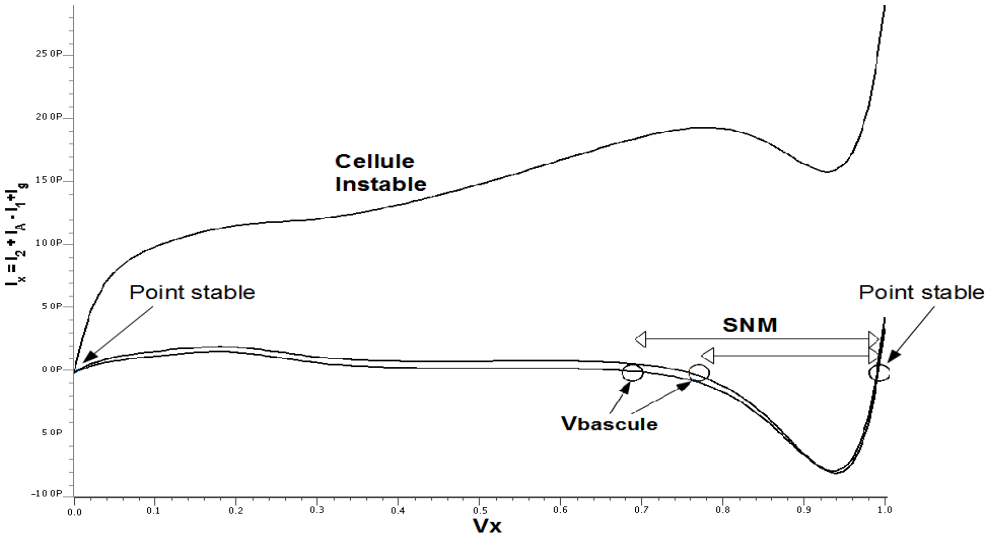
Au vu de la définition I_X, on voit qu'un latch ULP est stable en rétention si les courants de fuite du transistor d'accès et du *buffer* de lecture sont plus petits que le courant maximal I_{Peak} de la diode ULP, qui est également un courant sous seuil.
Comme tous les courants entrant en jeu sont des courants sous-seuil, la marge de bruit dépend très fortement des tensions de seuil de tous les transistors [38]. Il est donc nécessaire de tenir compte de la variabilité à chaque étape du raisonnement.

Le graphique 3.9 montre que le *corner* critique à considérer pour la marge de bruit en rétention est le *corner* FS. En effet dans ce *corner*, le courant de fuite du transistor d'accès NMOS est accentué, tandis que le courant des diodes ULP est moins important car il passe à travers un transistor PMOS. *Pro tempore*, les simulations seront donc toutes faites dans le *corner* FS.



(a) Testbench pour l'extraction le SNMH dans le cas $V_{dd} = 0V$.

(b) Testbench pour l'extraction le SNMH dans le cas $V_{dd} = V_{dd}$.



(c) Exemple de courbe pour I_X. La cellule est aussi instable si la courbe $I_X(V_X)$ est négative sur toute la plage de tension.

Figure 3.8 – Définition de la marge de bruit en rétention pour un latch ULP [23].

Dimensionnement pour la stabilité dans le *corner* FS

Une simulation Monte-Carlo nous montre qu'un latch ULP composé de transistors LVt et un transistor d'accès HVt, tous les cinq de taille minimale, n'est pas stable en rétention dans le *corner* FS, quelque soit le type de *buffer*.

Plusieurs options seront étudiées pour rendre la cellule stable :

1. diminuer le courant sous-seuil du transistor d'accès en rétention
   - en augmentant sa longueur de grille,
   - en diminuant sa tension de grille, donc la tension WordLine, pour atteindre le point de courant de drain minimum (solution exploitée dans [23]).
2. augmenter le courant I_{Peak} en augmentant la taille des transistors des diodes ULP, et en gardant une tension de seuil minimale.
3. diminuer le courant de grille du buffer : utiliser un transistor de dimensions minimales
   - de type HVt,
   - de tension de seuil quelconque en mode *reverse* (voir tableau 2.2).

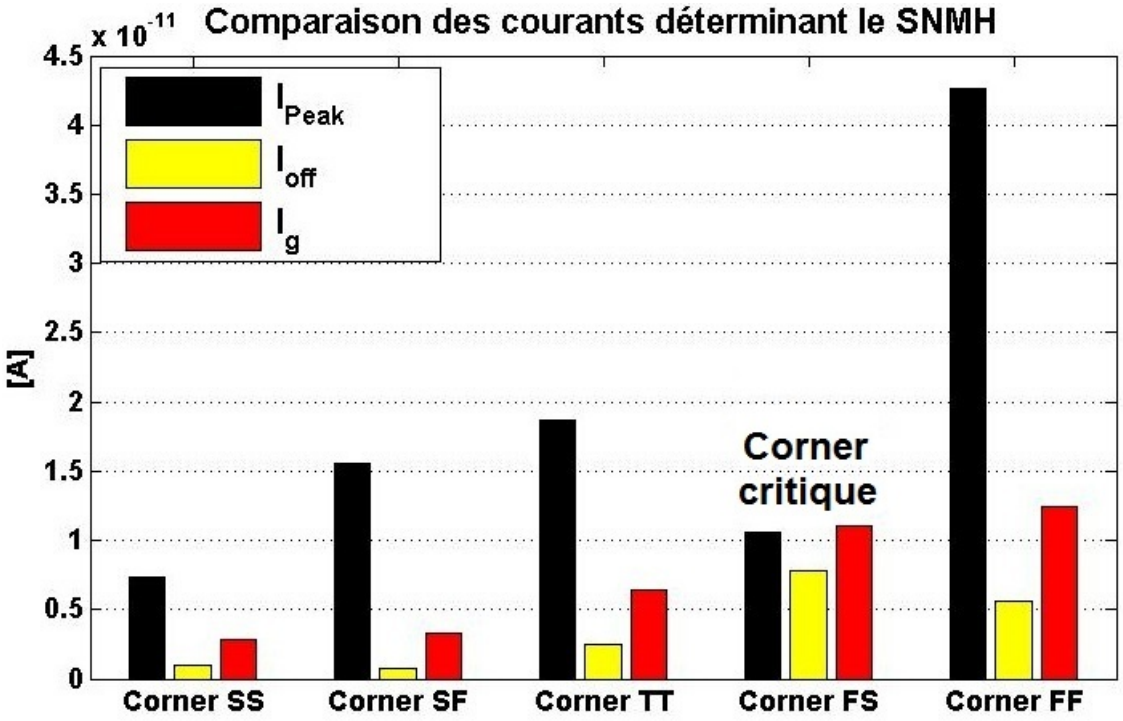
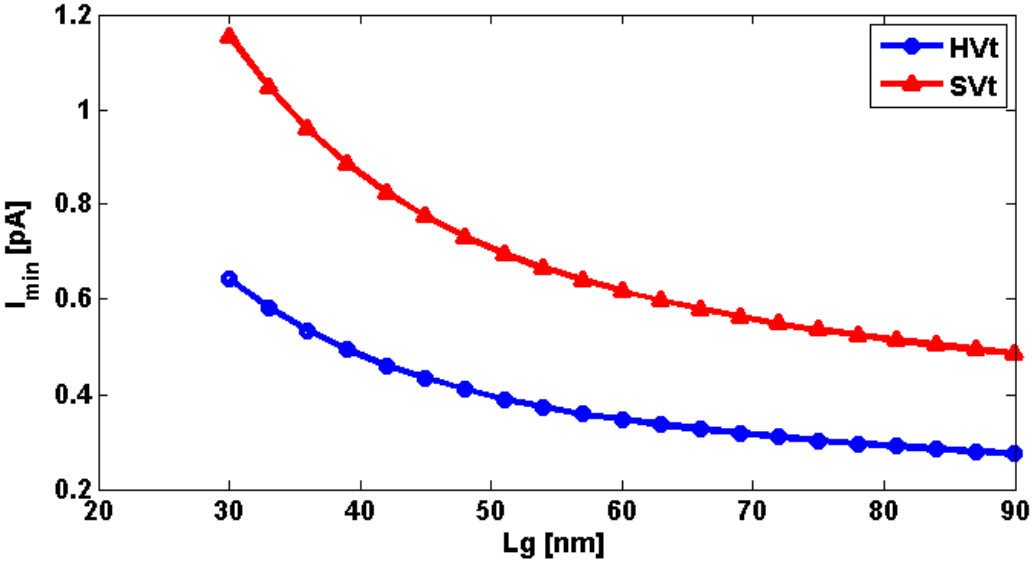


Figure 3.9 – Comparaison du I_{Peak} de la diode ULPD1 avec les courants de fuite de drain et de grille d'un transistor HVt, selon différents *corners*. Le *corner* à considérer pour le SNMH est le *corner* FS.

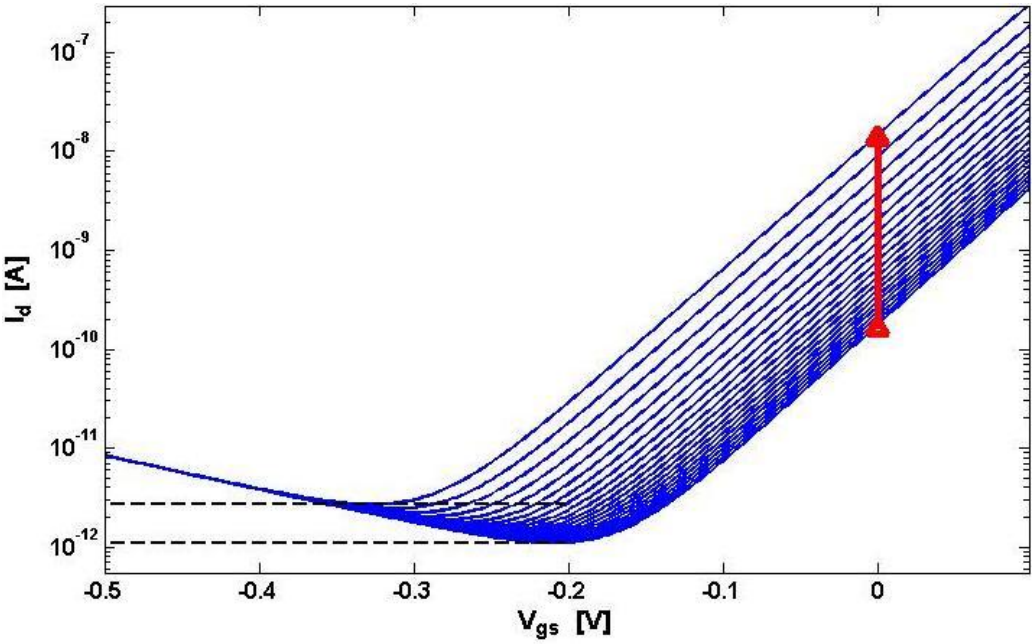
Augmenter la longueur de grille L_g diminue fortement I_{off}, le courant sous-seuil lorsque $V_{gs} = 0V$ ([34]). Malheureusement, le courant minimal obtenable I_{min} ne suit pas cette même dépendance. En effet, comme on le voit sur la figure 3.10, en doublant la longueur de grille, le courant minimal n'est même pas divisé par deux, tandis que le courant I_{off} est divisé par plus de 20. Ceci s'explique par le fait qu'en augmentant la taille du transistor, les courants de grille deviennent plus vite non négligeables. Le point de courant minimal est alors atteint pour une tension V_{gs} plus petite. Doubler la taille du transistor d'accès double le temps de charge de WL et de BL, pour un gain sur I_{min} inférieur à 2. Lorsque nous prenons en compte la variabilité de la tension de seuil, un facteur 2 est marginal. Cette solution est donc à abandonner.

La valeur de la **tension de WL négative en rétention** est fixée dans le but d'atteindre le courant sous-seuil minimal I_{min} (section 2.1), et non de sauvegarder de l'énergie. Comme illustré à la figure 3.11, la variabilité entraine une grande plage de variation des courants sous-seuil. Si la valeur de la tension est fixée en tenant uniquement compte du cas nominal du corner, la variabilité entrainera des courants de drain plus grands que la limite attendue du dimensionnement pour certaines cellules. La valeur de la tension doit alors être choisie afin d'obtenir un courant *minimal dans le pire des cas*. Des simulations Monte-Carlo sont alors utilisées pour déterminer la valeur optimale, comme indiqué à la figure 3.11.

La figure 3.12 montre que, dans le *corner* FS, la largeur de grille des transistors PMOS a la plus grande influence sur le courant I_{Peak} de la diode. La section 2.1 montre que le courant sous-seuil d'un transistor PMOS est d'un ordre de grandeur inférieur à celui d'un NMOS. Il est donc tout à fait logique que le dimensionnement des PMOS ait un rôle plus prononcé sur le courant de la diode que celui des NMOS. Dans le but de maintenir une surface de silicium minimale, seule **la taille des transistors** PMOS sera modifiée, les NMOS restant à une largeur de grille minimale. La diode du haut (ULPD1) a un courant I_{Peak} beaucoup plus grand que celui de la diode du bas. La raison en est que la tension Bulk-Source V_{bs} du transistor PMOS de la diode du haut est plus grande (en valeur absolue) pour des hautes valeurs de V_X que celle



(a) I_{min} atteint à V_{gs} négative en fonction de la longueur de grille. Transistors de type HVt et SVt.



(b) Courbe I_d/V_{gs} pour différentes longueurs de grille (échelle logarithmique). Transistors de type LVt. La flèche rouge montre la variation du courant I_{off}.

Figure 3.10 – Le courant de drain minimal I_{min} en fonction de la longueur de canal L_g.

du transistor PMOS de la diode du bas pour de basses valeurs de V_X. Or, cette polarisation entraine une diminution de la tension de seuil dans cette technologie (section 2.1). Le courant de la diode du bas (ULPD2 figure 3.8(a)) est également important, car il se peut que le courant I_g d'un *buffer* en mode *reverse* soit suffisant pour maintenir le courant I_{SNM} négatif sur toute la plage de tension. Notons que le courant I_{Peak} est plus que quadruplé quand le W_g du PMOS est doublé.

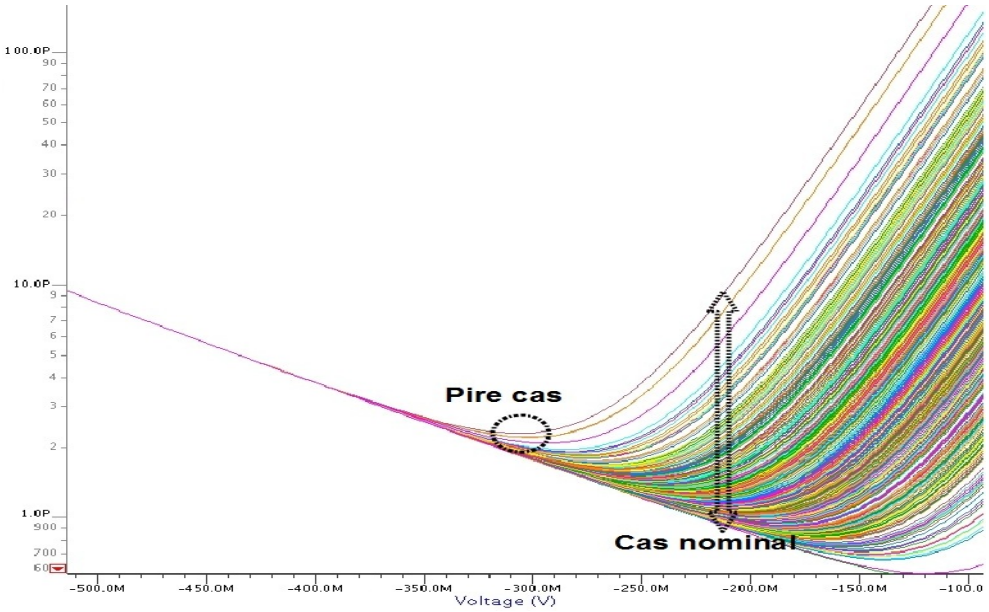


Figure 3.11 – La valeur de la tension de WL doit être choisie afin d'obtenir un courant minimal dans le pire des cas, et non dans le cas nominal du *corner*. Le pire des cas est approximé grâce à une simulation Monte-Carlo.

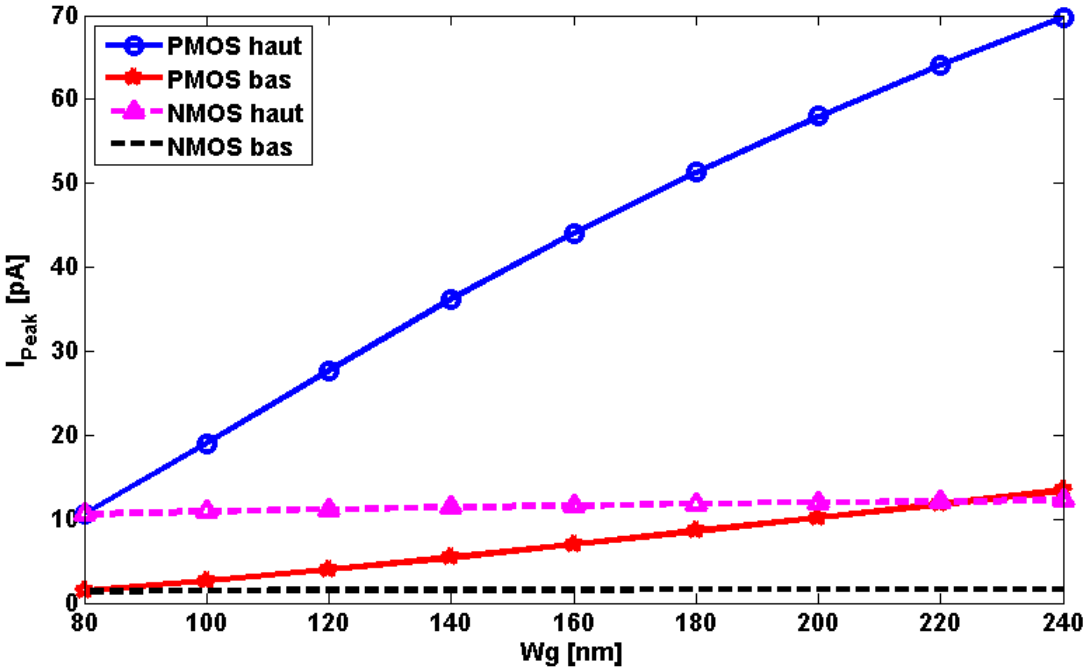


Figure 3.12 – Courant I_{Peak} en fonction des dimensions d'un transistor du *latch*, tandis que les autres transistors sont laissés à taille minimale. Le W_g des transistors PMOS est le plus influent sur le courant I_{Peak}. La diode ULP du haut produit un courant I_{Peak} plus important car la tension V_{bs} de son PMOS est plus grande.

Validation des solutions proposées

Pour déterminer la robustesse de la cellule, deux cas sont à considérer :

- Pour des tensions V_X élevées, le courant I_{Peak} de la diode du haut doit être plus grand que le courant de fuite de drain du transistor d'accès (avec la BL maintenue à la masse) et que le courant de fuite de grille du transistor du *buffer* de lecture, monté en *forward*.
- Pour de faibles tensions V_X, le courant I_{Peak} de la diode du bas doit être plus grand que le courant de fuite de source du transistor d'accès (avec la BL maintenue à 1V) et que le courant de fuite de grille du transistor du *buffer* de lecture, monté en *reverse*.

Dans chaque cas, une simulation Monte-Carlo de 500 itérations nous permettra de comparer les extrema de ces courants.

Le graphique 3.13 montre la valeur minimale obtenue pour I_{Peak} de la diode du haut selon différentes largeurs de grille du transistor PMOS de la diode, et les valeurs maximales obtenues pour I_d et I_g, selon la tension de seuil des transistors. Rappelons que I_{min} est par définition le courant de drain minimal sur toute la plage de tensions V_{gs}, avec $V_d = V_{dd} = 1V$, tandis que I_{off} est le courant de drain obtenu pour $V_{gs} = 0$.

On remarque très clairement que la tension de WL négative en rétention est indispensable. Les courants I_{off} maximaux sont de plusieurs ordres de grandeur supérieurs aux courants I_{Peak} minimaux.

L'utilisation d'un *buffer* de lecture où le transistor NMOS connecté au *latch* a ses tensions de drain et de source nulles en rétention, est également à proscrire dans cette technologie. Le transistor HVt a un courant de grille maximal presque quatre fois plus élevé que le courant I_{Peak} minimal obtenu avec $W_g = 4W_{min}$ pour le transistor PMOS de la diode ULP. Pour garder des dimensions acceptables dans le cadre d'une application SRAM dense, il faut nécessairement utiliser un buffer dont les tensions de source et de drain sont proches de V_{dd} en rétention.

Remarquons que nous comparons ici les cas de figure "extrêmes". En effet, pour dire que la cellule sera instable, il faut que le courant I_g soit maximal **et** que le courant I_{Peak} soit minimal. Ces deux cas de figure ont chacun une chance sur 500 de se produire (le nombre d'itérations de la simulation Monte-Carlo). La probabilité qu'ils se produisent en même temps est donc d'une chance sur 250000. Cette probabilité n'est pas négligeable pour un tableau mémoire SRAM contenant typiquement plusieurs dizaines de milliers de cellules. De plus, vu l'énorme différence entre le courant I_g maximal et le courant I_{Peak} minimal, on rencontrera certainement d'autres configurations intermédiaires où la cellule est instable, et ceci avec un probabilité nettement supérieure.

Le graphique 3.14 montre la valeur minimale obtenue pour I_{Peak} de la diode du bas selon différentes largeurs de grille du transistor PMOS de la diode, et les valeurs maximales obtenues pour I_s et I_g, selon la tension de seuil des transistors. La valeur du courant de grille en mode *reverse* ne varie pratiquement pas lors des simulations Monte-Carlo, autre avantage de ce mode. En fixant la largeur de grille du transistor PMOS de la diode ULP du bas à $2.25W_{g,min}$, la cellule sera robuste face à la variabilité dans tous les cas de figure.

Le graphique 3.15 montre la marge de bruit en rétention à 5σ et 6σ en fonction de la tension de seuil du transistor d'accès, pour différentes valeurs de la largeur de grille du transistor PMOS de la diode du haut (ULPD1).

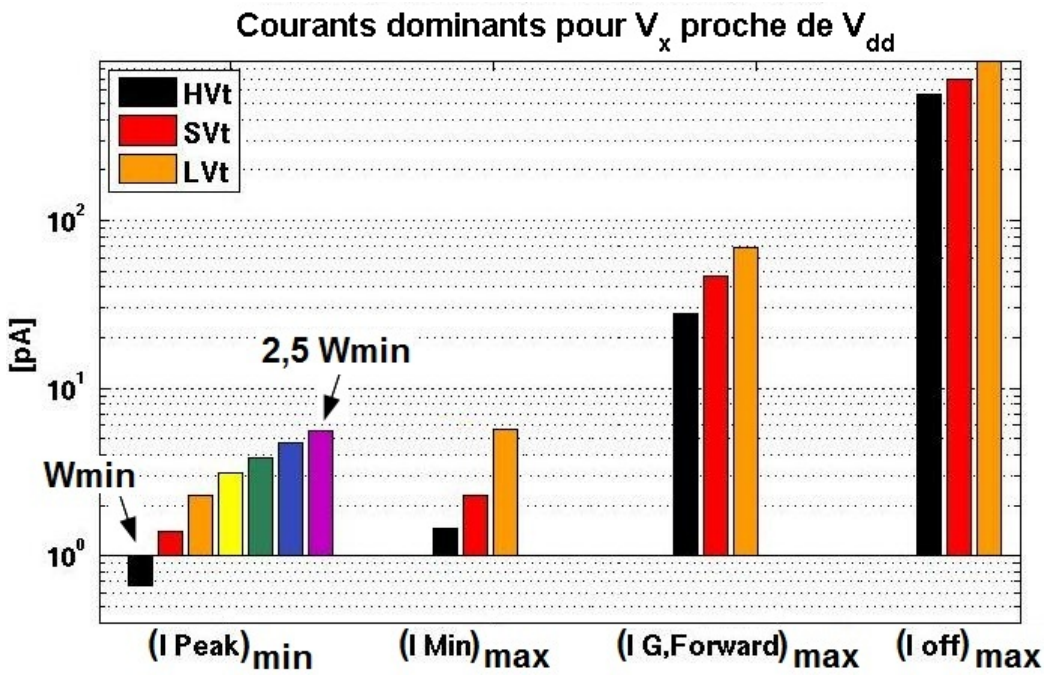


Figure 3.13 – Résultats de simulation Monte-Carlo à 500 itérations. I_{Peak} est le courant maximal de la diode du haut (ULPD1). Les courants I_{off} SVt et LVt ont été mis à l'échelle pour une meilleure vision du graphique.

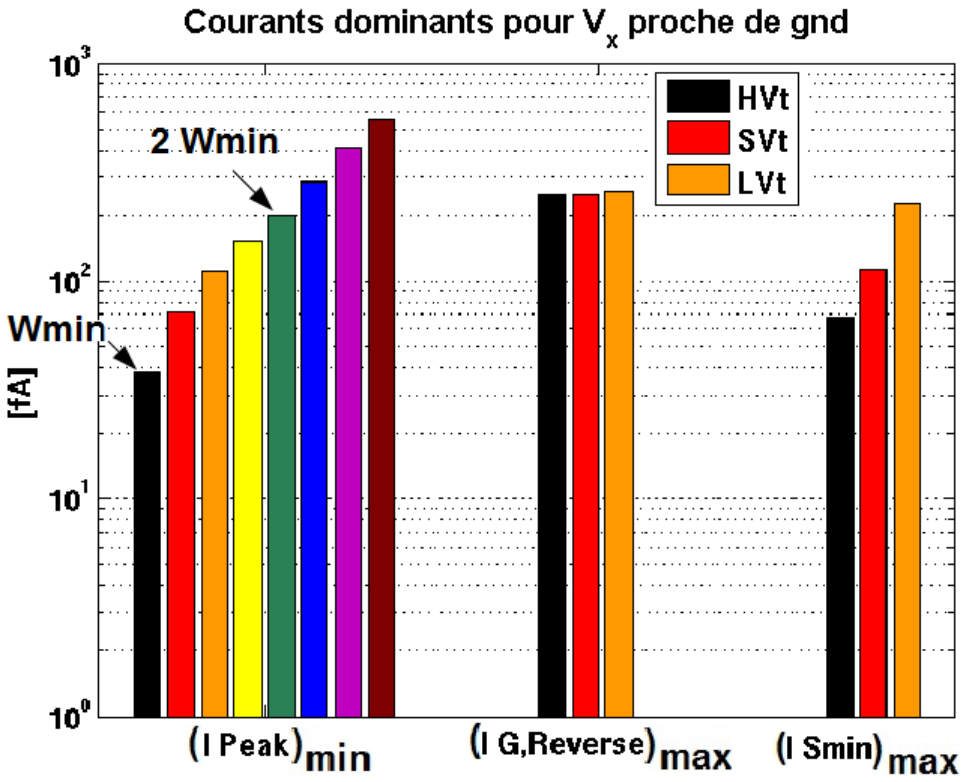


Figure 3.14 – Résultats de simulation Monte-Carlo à 500 itérations. I_{Peak} est le courant maximal de la diode du bas (ULPD2).

En conclusion, pour une stabilité 6σ, les dimensions de la cellule 5T ULP sont
- M_W : HVt, $L_g = L_{g,min}$, $W_g = W_{g,min}$ et $V_{WL} = -0.24V$ en rétention
- PMOS ULPD1 : LVt, $L_g = L_{g,min}$ et $W_g = 1.25W_{g,min}$,
- NMOS ULPD1 : LVt, $L_g = L_{g,min}$ et $W_g = W_{g,min}$,
- PMOS ULPD2 : LVt, $L_g = L_{g,min}$ et $W_g = 2.25W_{g,min}$,
- NMOS ULPD2 : LVt, $L_g = L_{g,min}$ et $W_g = W_{g,min}$.
- Les seules contraintes sur M_G sont sa taille minimale et sa configuration en *reverse* en rétention, $V_S = V_D = V_{dd}$.

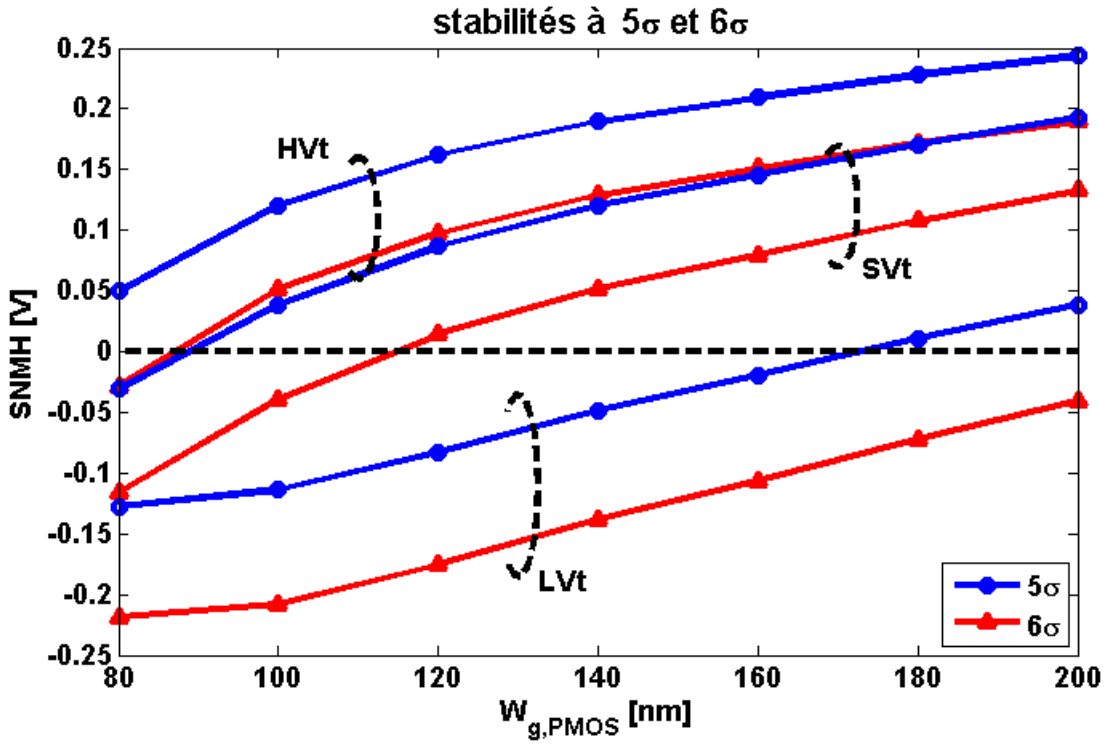


Figure 3.15 – SNMH à 5σ et 6σ en fonction de la largeur de grille du PMOS. Le V_t du transistor d'accès en paramètre.

Si la contrainte de stabilité 6σ est relâchée, nous pouvons diminuer les tailles des transistors. Par exemple, pour une stabilité 5σ, le transistor PMOS de la diode du haut (ULPD1) peut être de taille minimale.

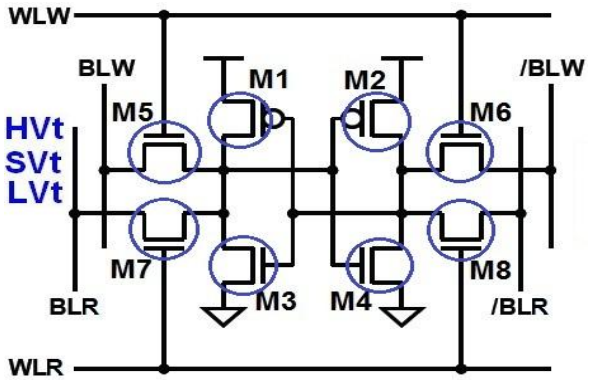
Vitesse et consommation

Ne présentant qu'un seul transistor d'accès, la vitesse de charge de la WL est pratiquement réduite de moitié par rapport aux cellules 6T et 5TPMOS. L'énergie dynamique en sera également réduite.

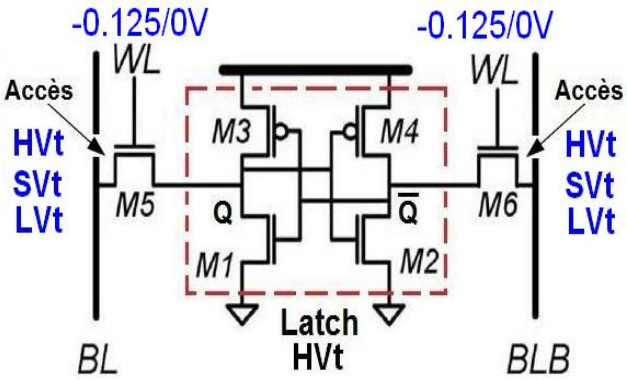
Au vu des très faibles valeurs de courant rencontrées dans cette section, la consommation statique devrait être extrêmement faible. D'où le nom *Ultra Low Power*.

3.2.7 Résultats de simulations et comparaison

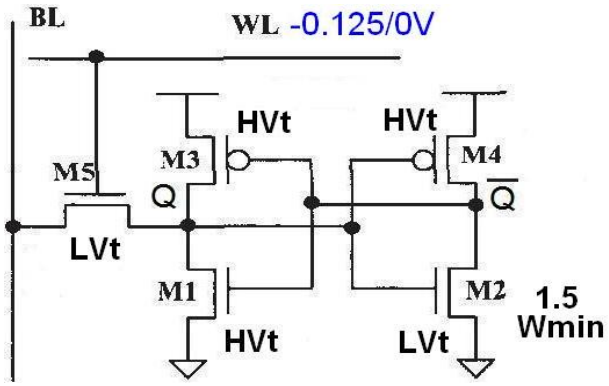
Le tableau 3.2 reprend les principaux résultats de simulations pour toutes les cellules qui présentaient des marges de bruit acceptables. L'architecture de ces dernières est rappelée à la figure 3.16, ainsi que l'explication de la nomenclature.



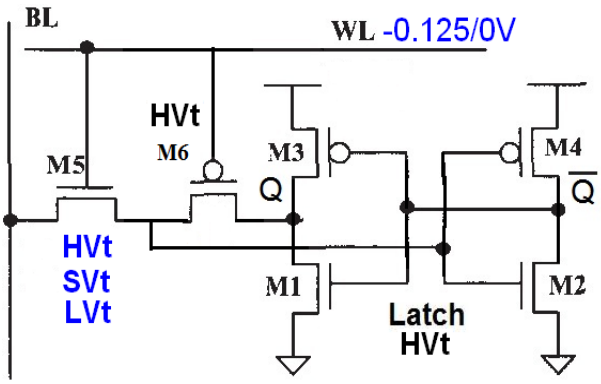
(a) La cellule *dual-port* 8T. Tous les transistors ont la même tension de seuil. *Cell ratio* = 2.



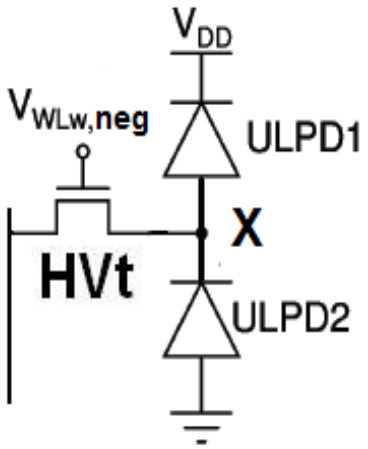
(b) La cellule 6T dédiée à l'écriture.



(c) La cellule 5T. La seule configuration robuste (pour une surface minimale).



(d) La cellule 5TPMOS. Les quatre transistors du *latch* sont HVt et de taille minimale.



(e) La cellule ULP. Dimensions du *latch* à la fin de la section 3.2.6.

Figure 3.16 – *Latchs* et systèmes d'écriture retenus pour les simulations dynamiques. En bleu, les paramètres de simulations.

Table 3.2 – Résultats de simulations. Corner TT, $V_{dd} = 1V$, Température 27°C.

Cellule	T Write [ps]	E Write [fJ]	P stat [pW]	SNMH [mV]	SNMW [mV]
8T HVt	449	15.1	27.3	**404**	-344
8T SVt	455	15.9	239	**359**	-319
8T LVt	469	19.9	22 200	**274**	-283
6T accès HVt	**433**	**14.4**	**17.4**	414	-359
6T accès SVt	409	14.4	67.2	414	-390
6T accès LVt	379	16.4	4340	414	-475
6T accès LVt, V_{WL} nég.	237	14.1	118.5	414	-475
5TPMOS accès HVt	**544**	**7.2**	**15.8**	414	**-456**
5TPMOS accès SVt	**488**	7.1	40.8	414	**-456**
5TPMOS accès LVt	**448**	8.3	2111	414	**-45**6
5TPMOS LVt, V_{WL} nég.	385	7.2	58.5	414	**-45**6
5T	374	9.8	4349	280	-313
5T , V_{WL} nég.	**326**	8.6	2285	280	-313
5T ULP	363	**6.7**	**1**	481	-481

Commençons par détailler ces résultats.

Rappelons que **le temps d'écriture** tient compte de la précharge de la BL. Celui-ci est de 140ps dans le *corner* TT pour 256 cellules connectées à une BL. En retirant ce temps au temps total, nous constatons que les cellules à un seul transistor d'accès sont bien *grosso modo* deux fois plus rapides que celles à deux transistors d'accès. La capacité de WL est donc déterminante pour le temps d'écriture. De ce fait, une tension négative pour la masse des *drivers* de WL augmente fortement le courant de charge, ce qui diminue significativement le temps d'écriture.

La **consommation dynamique** se répartit de la façon suivante (figure 3.17) :

- entre 12 et 18fJ pour la charge de la WL, en fonction du nombre de transistors d'accès et de la valeur maximale atteinte lors de l'écriture, dont 1fJ pour le *driver*. Normalisé par cellule (/128), cela représente moins de 0.15fJ sur le résultat final.
- Charger une BL demande 12fJ, plus 1fJ pour le *driver*. L'énergie pompée à la source d'alimentation étant pratiquement nulle lors de la décharge des capacités, les résultats obtenus proviennent du facteur 1/2 de l'équation 2.3.
- Environ 1fJ est consommé sur la tension d'alimentation des cellules lors de la modification de leur état, sauf pour les cellules de type 5TPMOS qui n'en demandent que 0.45fJ. Le courant provenant des transistors PMOS du *latch* est, soit un courant de court-circuit - quand le transistor d'accès décharge le noeud de rétention -, soit un courant servant à charger la capacité de ce noeud. Or, le transistor PMOS propre à ces cellules bloque le courant pouvant provenir du *latch*. En conclusion, 0.45fJ servent à charger les capacités du *latch*, et une énergie encore plus grande est perdue dans les courants de court-circuit.
- La différence d'énergie restante est due aux courants de courts-circuits, dont la durée varie selon les paramètres des transistors d'accès et du *latch*, et d'autres phénomènes *in situ*.

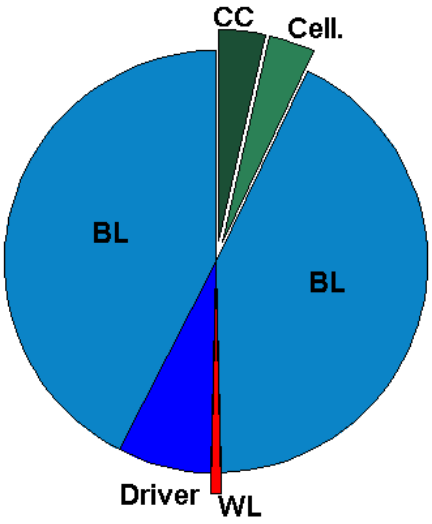


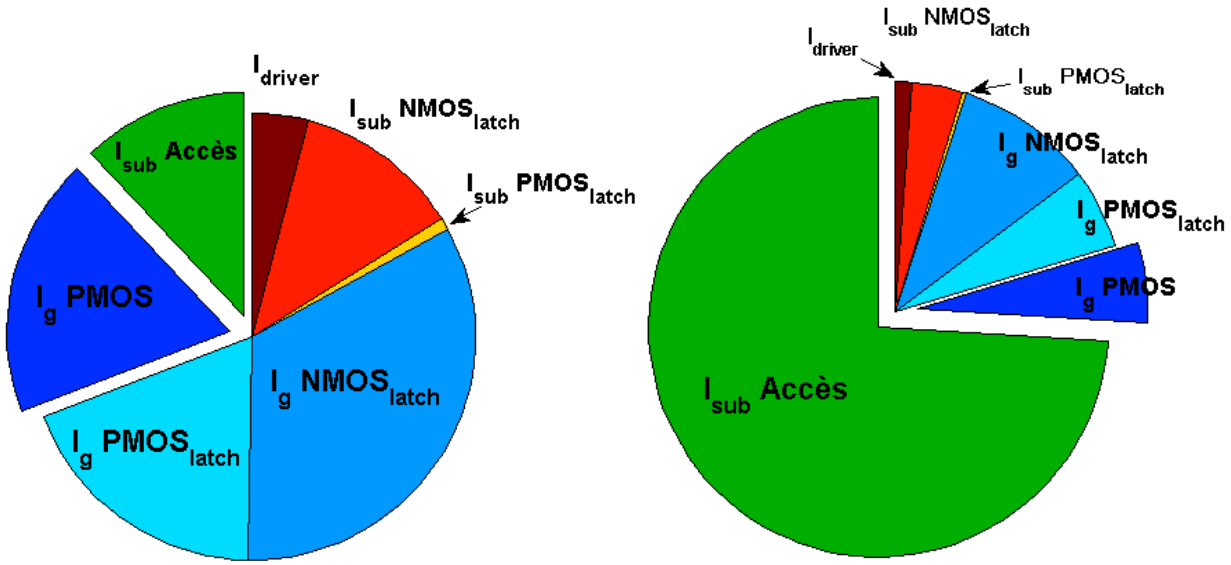
Figure 3.17 – Répartition de l'énergie dynamique moyenne par cellule entre les différentes capacités du circuit.

Notons que les cellules utilisant une tension de WL négative en rétention ont une énergie dynamique inférieure à leur pendant sans assist. Ceci peut paraitre contradictoire puisqu'il faut plus d'énergie pour charger la capacité de WL. Cette différence vient des courants de fuite des transistors d'accès connectés à la BL des cellules non accédées. Ils sont drastiquement diminués du fait de leur tension de grille négative. Comme les transistors de type LVt ont un courant $I_{off} = 4.4nA$ (tableau 2.2), la puissance totale de fuite intégrée sur 1ns vaut environ 1fJ, ce qui explique la différence sur le résultat final.

La **puissance statique** due aux *drivers* de BL est de $256 \times 0.85pW$, ce qui est négligeable dans la plupart des cas, excepté pour la cellule ULP où les *drivers* représentent pratiquement la totalité de la consommation. La consommation statique moyenne due à ces cellules n'est que de $0.15pW$. En se référant au tableau 2.2, il appert que, pour les cellules 8T HVt, 6T accès HVt et 5TPMOS accès HVt, le courant de fuite de grille représente *plus de* 50% de la consommation statique totale ! La figure 3.18 illustre cette répartition pour la cellule 5TPMOS accès HVt. Ce pourcentage passe à 30% pour la cellule 5T accès SVt, et arrive à 15% environ pour les cellules 8T SVt, 6T accès SVt et 5TPMOS accès LVt avec *assist*. Une des solutions pour résoudre cet inconvénient est de diminuer la tension d'alimentation (section 5.3).

Comparaison

En regardant les résultats des cellules 8T conventionnelles du tableau 3.2, nous voyons que pour des dimensions identiques des transistors, le SNMH est plus grand si ces derniers sont de type HVt. L'utilisation d'un *latch* HVt dans la plupart des cellules n'est donc que bénéfique.
Comme attendu, plus la tension de seuil des transistors d'accès est basse, plus le temps d'écriture est court, pour une même architecture de cellules. Néanmoins, le temps d'écriture de la cellule 6T accès HVt est plus court que la cellule 5TPMOS accès LVt. L'écriture d'un 1 logique est donc toujours plus lente, même en cassant la contre-réaction du *latch*.
La consommation dynamique moyenne par cellule est dominée par l'énergie de charge des capacités des BL. Les topologies de cellule à une seule BL ont donc une énergie dynamique pratiquement divisée par deux. La consommation statique, elle, est dominée par la tension de seuil des transistors d'accès. Ainsi, la cellule 6T accès HVt consomme plus en énergie dynamique mais moins en puissance statique que la cellule 5TPMOS accès SVt.



(a) Cellule 5TPMOS accès HVt.

(b) Cellule 5TPMOS accès SVt.

Figure 3.18 – Répartition des courants de fuite dans la puissance statique totale des cellules. Les quartiers détachés sont présents ou non en fonction de la donnée de la cellule et de la tension de BL, à cause de l'asymétrie des cellules à une seule BL.

Dans le but d'aboutir à une cellule à consommation minimale, les cellules seront classées dans l'ordre croissant de consommation, et il sera vérifié, dans cet ordre, si elles tiennent les contraintes de vitesse face à la variabilité.

Il est clair que la cellule ULP est celle qui consomme le moins, suivie par la cellule 5TPMOS accès HVt. Un doute survient alors pour démarquer la cellule 6T accès HVt, ayant la troisième plus faible consommation statique, et les cellules à énergie dynamique presque deux fois moindre (une seule BL).

La consommation totale du tableau de cellules entre deux écritures successives est donnée par

$$E_{tot} = \alpha_{SW} \cdot 128 \cdot E_{W,cell} + 128 \times 255 \cdot P_{stat,cell} \cdot 1ns + 128 \times 256 \cdot P_{stat,cell} \cdot T \qquad (3.2)$$

où T est le temps entre les deux écritures et α_{SW} le taux de commutation des cellules SRAM.

Le tableau 3.3 donne la valeur du temps T nécessaire pour que la 6T accès HVt devienne énergétiquement plus efficace, pour différentes valeurs de α_{SW}. Un nombre négatif ou égal à zéro signifie que la cellule 6T accès HVt est énergétiquement plus efficace que la cellule comparée. La conclusion de ce tableau est que les cellules 5TPMOS accès SVt et 5TPMOS accès LVt avec assist consomment certainement moins que la cellule 6T accès HVt.

Variabilité

Pour déterminer si les cellules respectent la contrainte de vitesse dans tous les cas de figure, il est nécessaire d'effectuer des simulations Monte-Carlo dynamiques dans le *corner* SS. En effet, ce *corner* provoque un courant minimal pour tous les transistors du tableau mémoire. Donc, si toutes les écritures simulées sont achevées en moins de 1ns, nous pouvons conclure que toutes les cellules du tableau supportent une écriture à une fréquence de fonctionnement de 1GHz. Nous allons parcourir toutes les cellules dans l'ordre croissant de consommation, jusqu'à atteindre une cellule parfaitement robuste face à la variabilité.

Table 3.3 – Un nombre nul ou négatif signifie que la cellule 6T accès HVt est plus efficace énergétiquement que l'autre cellule. α_{SW} est le taux de commutation du tableau mémoire.

Cellule	$\alpha_{SW} = 0\%$	$\alpha_{SW} = 1\%$	$\alpha_{SW} = 50\%$	$\alpha_{SW} = 100\%$
5TPMOS accès SVt	0s	8.4 ps	**600 ps**	1202 ps
5TPMOS accès LVt	0s	-1ps	4.7 ps	10.4 ps
5TPMOS accès LVt V_{WL} nég.	0s	4.3	**339 ps**	680 ps
5T	0s	-1ps	1.0 ps	3.1 ps
5T V_{WL} nég.	0s	-1ps	3.9 ps	8.9 ps

1. La cellule ULP. Le temps d'écriture représenté dans le tableau 3.2 est le temps minimal nécessaire pour que le noeud de rétention de la cellule dépasse la tension de basculement (voir le début de la section 3.2.6). Si le transistor d'accès est coupé avant que les tensions des noeuds de rétention atteignent leur valeur nominale, la contre-réaction du *latch* permet d'amener rapidement ces tensions vers un nouvel état stable. Ces courants de régénération sont suffisants, dans la plupart des cellules, pour atteindre cet état dans les quelques picosecondes qui suivent. Néanmoins, le courant de régénération de la cellule ULP étant un courant sous-seuil, il faudra donc un temps non négligeable entre la fin de l'écriture et le moment où le noeud de rétention aura atteint $V_X = V_{dd}$ [24]. De plus, la chute de la tension de WL entraine un couplage capacitif faisant diminuer la tension V_X à la fin de l'écriture (voir figure 3.19).

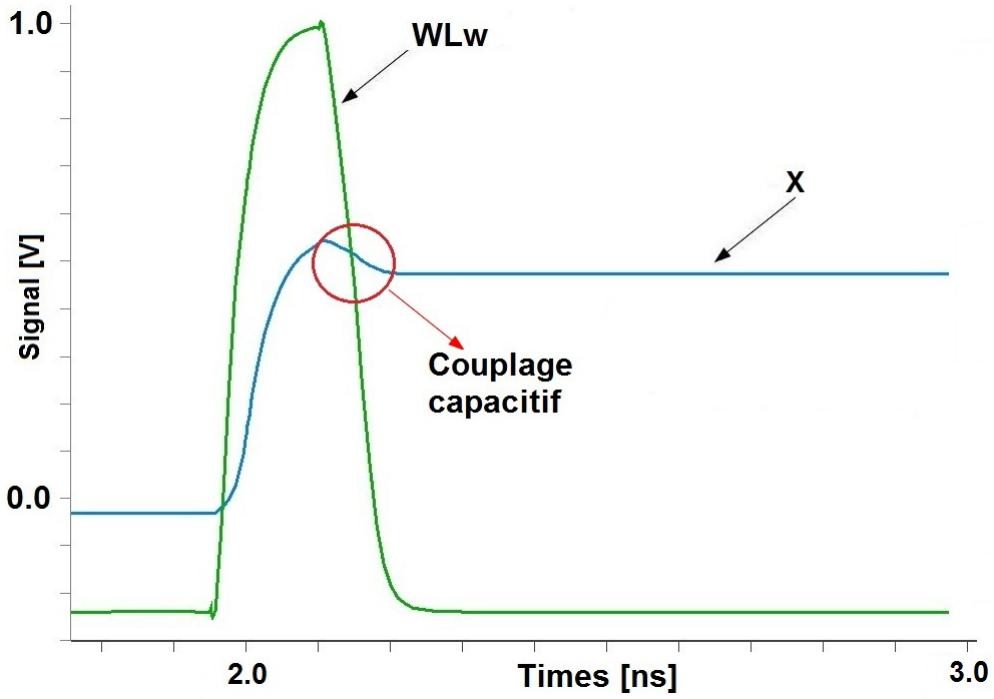


Figure 3.19 – Le couplage capacitif grille-source fait chuter la tension du noeud de rétention de la cellule ULP à la fin de l'écriture. Un temps considérable est nécessaire à la régénération de la valeur de la donnée.

En augmentant le temps d'ouverture de la WL, le transistor d'accès pourra charger le noeud de rétention à une plus grande valeur. Après 1ns d'activation, la tension de ce

noeud a atteint 566mV et son taux de montée est d'environ 0.25mV en 10ns. La cellule mettra donc environ *17360 ns* pour atteindre la tension d'alimentation, et ceci dans le *corner* TT ! Ce qui signifie que lors d'une lecture postérieure, le *buffer* de lecture peut avoir une tension de grille inférieure à 0.6V. Cette tension diminuera énormément son courant de décharge, et donc augmentera substantiellement le temps de lecture. Notons que ce problème se pose quelles que soient les dimensions du tableau mémoire ou des *drivers* de BL et de WL. Une solution envisageable serait de diminuer la tension d'alimentation de la cellule [23]. Nous en reparlerons dans la section 5.3.

2. La cellule 5TPMOS HVt. Lors de simulations Monte-Carlo dans le *corner* SS, le transistor d'accès a souvent une tension de seuil si haute que, lors de l'écriture d'un 1 logique, le temps nécessaire à atteindre la tension de basculement de l'inverseur est bien plus grand que 1ns. De plus, le phénomène de couplage capacitif de la WL se retrouve également ici (voir figure 3.20).

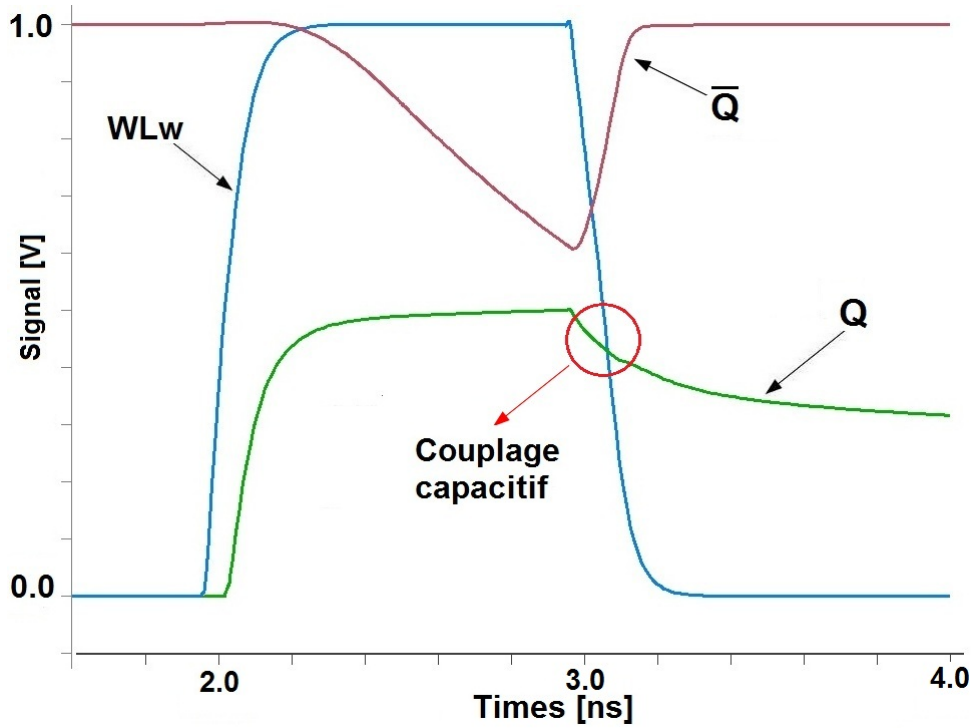


Figure 3.20 – Le couplage capacitif grille-source fait chuter la tension du noeud de rétention de la cellule à la fin de l'écriture.

Par simulation, on détermine qu'avec un temps d'activation de la WL d'environ 2.5ns, l'écriture est enfin effective. Ce qui signifie que, quelles que soient les dimensions du tableau mémoire ou des *drivers* de WL, cette cellule ne convient pas à des applications supérieures à 400MHz. Une solution serait de diminuer la tension de basculement de l'inverseur, mais la taille du NMOS de pied serait alors prohibitive (section 3.2.4). Et, tenant compte du courant de grille, cette cellule aurait alors une consommation statique plus importante que la cellule 5TPMOS à transistor d'accès SVt.

3. La cellule 5TPMOS accès SVt. Une série d'itérations Monte-Carlo dans le *corner* SS ont montré que la tension de seuil du transistor d'accès reste assez basse pour permettre une écriture effective, et que la cellule tient effectivement la contrainte de temps. En considérant une stabilité 6σ sur l'écriture d'un 1 logique, le temps d'écriture est de 700ps plus 150ps de précharge de la BL, pour les définitions du tableau mémoire et du temps d'écriture dans ce travail. Ce qui laisse donc 150ps de battement au décodeur d'entrée pour rester dans la limite de 1ns.

4. La cellule 5TPMOS accès LVt avec tension de WL négative en rétention. Une série d'itérations Monte-Carlo dans le *corner* SS a montré que la tension de seuil du transistor d'accès reste assez basse pour permettre une écriture effective, et que la cellule tient effectivement la contrainte de temps, en laissant beaucoup plus de battements aux systèmes périphériques pour rester dans la limite de 1ns.
5. La cellule 6T accès HVt. Cette cellule respecte très largement la contrainte de vitesse. Pour une application à plus haute fréquence, ou pour d'autres dimensions du tableau mémoire de sorte que les cellules à une seule BL ne peuvent soutenir le rythme, cette cellule serait alors la première à étudier pour préserver une consommation minimale.

Conclusion

Nous avons étudié et comparé plusieurs cellules SRAM selon leurs performances d'écriture et en rétention. Nous avons d'abord dimensionné ces cellules afin d'obtenir des marges de bruits en rétention et en écriture suffisantes. Il est apparu que les cellules dites 5T et ULP demandaient une attention toute particulière pour obtenir des marges de bruit suffisantes, respectivement en lecture et en rétention. Nous avons donc du sélectionner un dimensionnement précis pour ces cellules, contrairement aux cellules dites 6T et 5TPMOS qui gardaient la tension de seuil de leur(s) transistor(s) d'accès en paramètre. En observant que ce paramètre est un compromis entre vitesse et consommation, nous avons cherché à s'affranchir de ce compromis en étudiant un *assist* statique, la tension de WL négative en rétention. Nous avons alors démontré que la polarisation négative de la WL devait avoir une valeur bien précise pour obtenir un gain énergétique. Nous avons ensuite montré que ce gain n'est conséquent qu'avec des transistors d'accès de type LVt.

Ensuite, nous avons comparé les dimensionnements de cellules retenus selon leurs performances dynamiques et leur puissance de fuite statique. Nous avons observé que la cellule 5T est la plus rapide mais que son dimensionnement entraine une consommation statique relativement énorme. Les cellules 6T sont inconditionnellement plus rapides que les cellules 5TPMOS, mais consomment en moyenne deux fois plus d'énergie par écriture. Au sujet de la puissance statique, nous avons d'abord observé que l'*assist* statique utilisé sur la WL diminue significativement la consommation statique et le temps d'écriture, et que le courant de fuite de grille représente une partie très importante de la consommation totale.

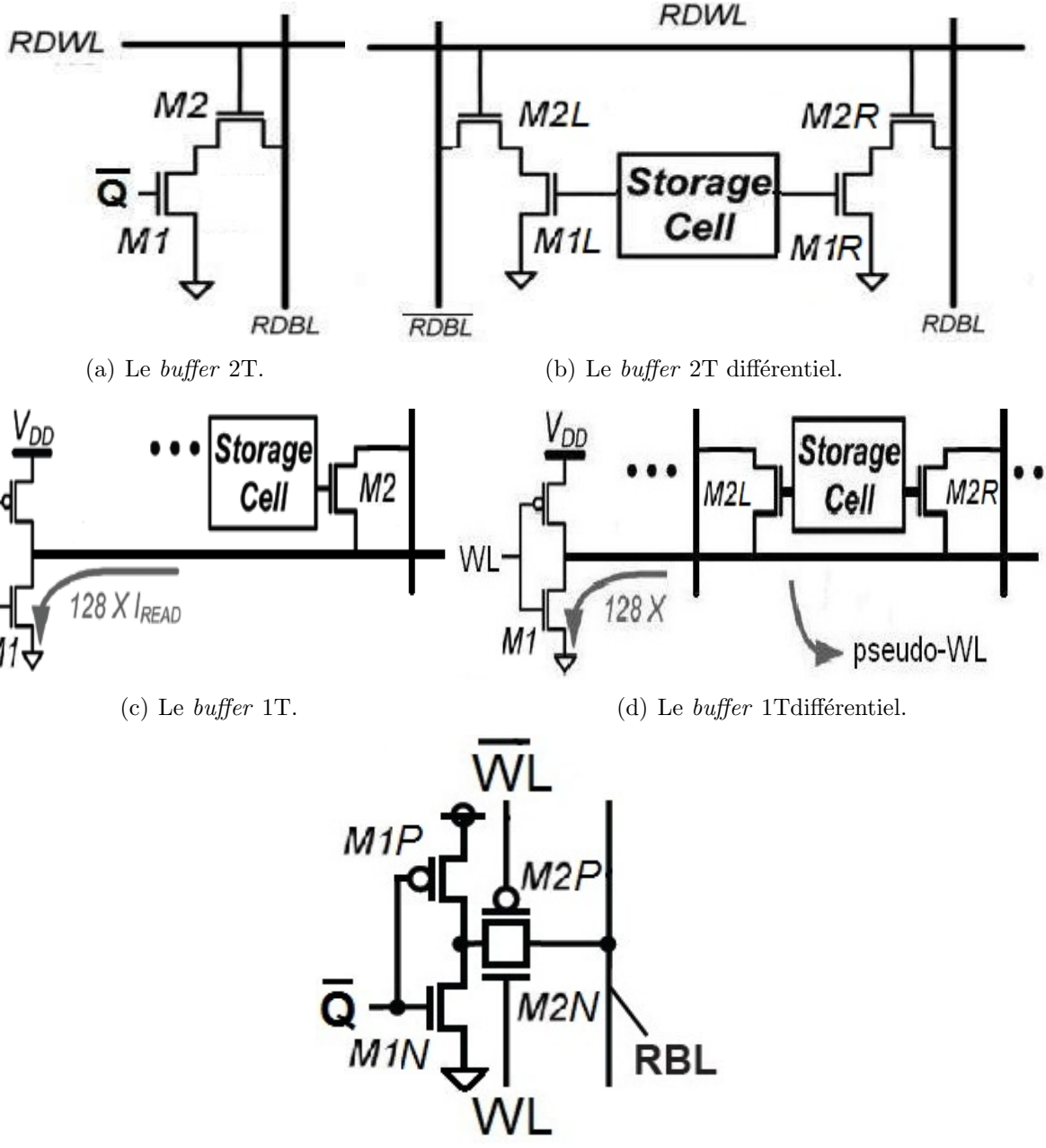
Puis, nous avons classé les cellules selon leur consommation totale, statique et dynamique combinées. Dans l'ordre croissant de consommation, nous avons vérifié grâce à d'autres simulations dynamiques si les cellules respectaient la contrainte de temps du tableau mémoire - 1ns dans ce travail. Il est apparu que les deux cellules qui présentent une consommation minimale (ULP et 5TPMOS accès HVt) ne respectent pas la contrainte de vitesse face à la variabilité de fabrication. Nous avons ensuite validité la fiabilité des autres cellules retenues.

En conclusion, la cellule dite 5TPMOS avec un transistor d'accès de type SVt est le système d'écriture qui possède la plus faible consommation et surface de silicium tout en respectant la contrainte de temps et de fiabilité face à la variabilité.

3.3 *Buffers* de lecture

Après avoir étudié et comparé différents systèmes d'écriture, passons maintenant aux systèmes de lecture retenus dans ce travail : les *buffers* de lecture.

Cinq architectures ont été retenues pour notre application à 1GHz : 1T (différentiel ou à BL unique, figure 3.21(d) et 3.21(c) respectivement), 2T (différentiel ou à BL unique, figure 3.21(b) et 3.21(a) respectivement) et l'inverseur *tri-state* (figure 3.21(e)). Les performances critiques des systèmes de lecture sont la vitesse, la consommation et le rapport I_{read}/I_{off}. Celui-ci donne le nombre maximal de cellules pouvant être connectées à une BL. Notons que les performances de la lecture différentielle conventionnelle de la cellule 8T *dual-port*, sont semblables à celles du *buffer* 2T différentiel.



(a) Le *buffer* 2T.

(b) Le *buffer* 2T différentiel.

(c) Le *buffer* 1T.

(d) Le *buffer* 1Tdifférentiel.

(e) Le *buffer* inverseur tri-state.

Figure 3.21 – Les différentes architectures de *buffer* de lecture étudiées dans ce travail.

Contrairement à la section précédente, les étapes de l'étude se feront caractéristique par ca-

ractéristique. Pour chacune d'elles, les différents *buffers* seront comparés entre eux, selon leurs paramètres de dimensionnement. La vitesse et la puissance statique seront d'abord analysées. Nous verrons qu'un compromis doit être fait entre vitesse et consommation statique pour les *buffers* classiques. Plusieurs pistes seront envisagées pour contourner ce compromis, puis comparées entre elles. Ensuite, l'énergie dynamique sera étudiée selon une méthodologie particulière qui tient compte de la variabilité dans les circuits électroniques. Enfin, la relation entre le rapport I_{read}/I_{off} et les paramètres des transistors sera établie. Une conclusion générale comparera les différents *buffers* de lecture, et déterminera le plus adapté à notre application.

3.3.1 Vitesse - consommation statique

La vitesse de lecture, telle que définie dans ce travail, est la combinaison de deux périodes de temps : la charge de la WL de lecture, et la décharge de la BL. La première dépend essentiellement de la capacité totale de WL, comme nous l'avons vu dans la section 3.2.7. Les architectures demandant deux transistors d'accès par cellule auront donc un temps de charge de la WL environ deux fois plus grand, pour un même *driver* de charge. La décharge de la BL dépend du courant produit par la cellule pour effectuer cette action, le courant I_{read}, et de la capacité totale de BL, principalement la somme des capacités de jonction des transistors.

La consommation statique due au système de lecture provient évidemment des courants de fuite entre l'alimentation et la masse. La BL étant préchargée à la tension d'alimentation V_{dd} au repos, un courant peut se créer entre celle-ci et la masse du *buffer* de lecture. Par définition, un *buffer* de lecture contient la valeur de la donnée sur une grille de transistors. La donnée étant à tension haute dans environ la moitié des cas, les courants de fuite peuvent aussi provenir des courants de grille, qui ne sont pas négligeables comme nous l'avons vu précédemment.

Structure *pull-down* à deux transistors

Que ce soit dans un *buffer* de lecture ou dans la configuration 8T conventionnelle, un minimum de deux transistors sont nécessaires à la décharge (ou non) de la BL de lecture : l'un possédant l'information de la donnée, l'autre l'information d'activation de la WL. La décharge dépendant de ces deux paramètres, les deux transistors doivent être mis en série. Leurs paramètres physiques définissent le courant I_{read} et le courant de fuite lors de la rétention. Le tableau 3.4 montre le courant I_{read} et la puissance statique en fonction des paramètres physiques des transistors.

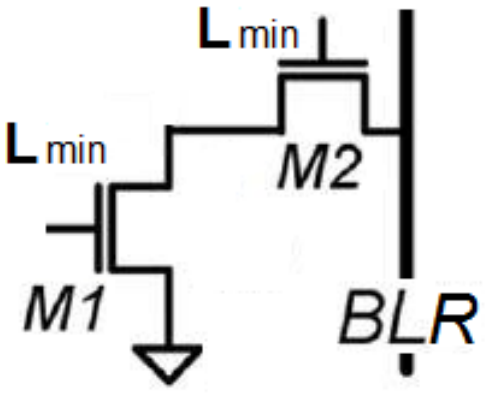


Figure 3.22 – Légende de la première colonne du tableau 3.4.

Table 3.4 – Les paramètres physiques (Largeur de grille et V_{th}) sont représentés à la figure 3.22

$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	I_{read} $[\mu A]$	$P_{stat}[pA]$	$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	I_{read} $[\mu A]$	$P_{stat}[pA]$
1-1 LVt-LVt	45.1	2240	2-1 LVt-LVt	**57.2**	2260
1-1 SVt-LVt	41.2	2220	2-1 SVt-LVt	54.6	2230
1-1 HVt-LVt	37.5	2210	2-1 HVt-LVt	52.3	2210
1-1 LVt-SVt	36.1	50.9	2-1 LVt-SVt	45.2	60.9
1-1 SVt-SVt	33.6	38.9	2-1 SVt-SVt	43.3	46.5
1-1 HVt-SVt	31.4	30.1	2-1 HVt-SVt	41.8	32.9
1-1 LVt-HVt	30	11.4	2-1 LVt-HVt	37.3	18.6
1-1 SVt-HVt	28.2	8.7	2-1 SVt-HVt	35.9	13.8
1-1 HVt-HVt	**26.6**	4.9	2-1 HVt-HVt	34.7	7.6
1-2 SVt-SVt	42.8	61.7	2-2 SVt-SVt	63.1	66.8

Plusieurs conclusions peuvent être tirées du tableau 3.4 :

- Doubler la taille de l'un ou des deux transistors ne double pas le courant de décharge. Par contre, elle double la capacité de jonction ; le transistor connecté à la BL (M2) doit donc garder une taille minimale.
- La tension de seuil du transistor en mode saturé (M1) a la plus grande influence sur le courant de décharge.
- La tension de seuil du transistor connecté à la WL (M2) a la plus grande influence sur la puissance statique. Ceci s'explique par le fait que le transistor connecté au *latch* est passant dans la moitié des cas.
- La différence entre le courant I_{read} de la configuration 2-1 LVt-LVt et de la configuration 1-1 HVt-HVt n'est que d'un facteur 2 dans cette technologie. *A contrario*, la consommation diffère de plusieurs ordres de grandeur. Le compromis entre vitesse et consommation qu'est l'utilisation d'un transistor d'accès LVt, doit être bien pesé.

Pour l'inverseur *tri-state*, trois transistors sont sur le chemin de lecture. Rappelons que le principe de ce *buffer* est d'activer les transistors de passage (M2P - M2N figure 3.21(e)) pour forcer la BL à la valeur de sortie de l'inverseur (M1P-M1N). La tension de la BL doit atteindre au moins $V_{dd}/2$ (dans un sens ou dans l'autre) pour faire basculer l'inverseur de sortie. On voit alors que l'un des deux transistors d'accès va progressivement se couper lors de la charge ou décharge de la BL. Donc, la vitesse de lecture de ce *buffer* a la même dépendance que celle discutée précédemment. Par contre, sa puissance statique dépend de manière similaire de tous les quatre transistors le composant. En effet, les deux transistors contenant l'information de la donnée forment un inverseur, et donc sont eux-mêmes un chemin entre l'alimentation et la masse.

Dans le but de maintenir une surface de silicium acceptable, les dimensions des transistors seront toutes minimales [4].

Table 3.5 – Tension de WL négative en rétention pour les transistors d'accès LVt.

$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	I_{read} $[\mu A]$	$P_{stat}[pA]$	$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	I_{read} $[\mu A]$	$P_{stat}[pA]$
1-1 LVt-LVt	45.1	55.8	2-1 LVt-LVt	57.2	63.4
1-1 SVt-LVt	41.2	36.6	2-1 SVt-LVt	54.6	42.7
1-1 HVt-LVt	37.5	28.0	2-1 HVt-LVt	52.3	30.6

Dans la configuration classique (figure 3.22), le transistor M2 est le principal paramètre du courant I_{read} et de la puissance statique. Sa taille et sa tension de seuil sont donc un compromis entre vitesse et consommation. Quels moyens permettent de s'affranchir de ce compromis ?

1. Utiliser une tension de WL négative en rétention. Comme nous l'avons vu dans la section 3.2.1, cet assist utilisé dans le but de sauvegarder de l'énergie n'est utile qu'avec un transistor de type LVt connecté à la WL. La capacité de WL étant différente dans le cas de la cellule à 6T, la même démarche effectuée dans la section 3.2.1 nous donne une tension de WL optimale de
$$\Delta V_{opt} = -0.145V$$
Le tableau 3.5 met à jour les résultats.
2. Augmenter la longueur de grille diminue drastiquement le courant I_{off} [34] (figure 3.10(b) section 3.2.6). Néanmoins, tant le courant I_{read} que la capacité de WL en seront affectés. La figure 3.23 montre l'évolution du courant I_{read} avec la longueur de grille de M2.

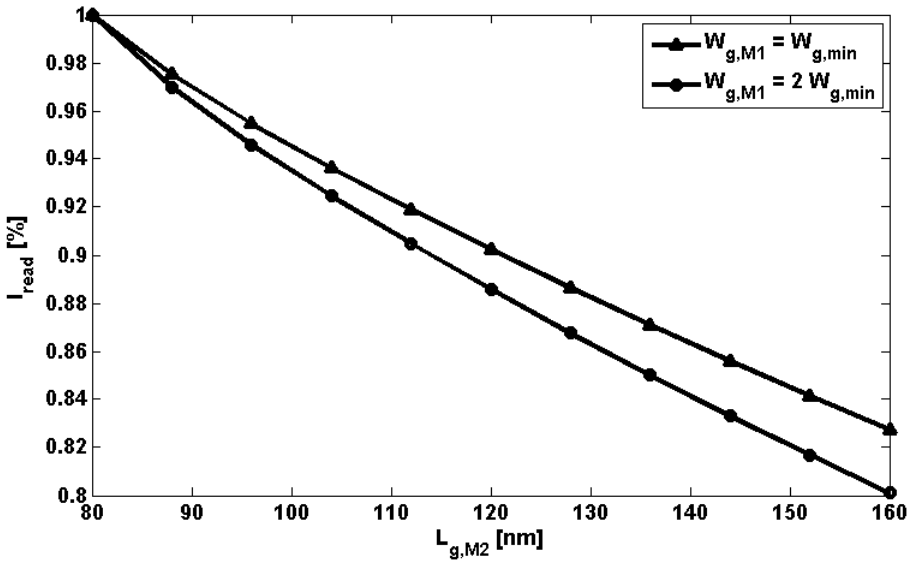


Figure 3.23 – Courant I_{read} relatif en fonction de la longueur de grille du transistor M2.

3. Inverser simplement l'ordre des transistors M1 et M2 (figure 3.22) semble régler définitivement le compromis. En effet, le transistor connecté à la WL ne sera plus en régime saturé lors de la décharge de la BL, et sa tension de seuil pourra donc être maximale pour diminuer la puissance statique tout en minimisant l'impact sur le courant I_{read}. Ceci est faux, car le pire des cas est à considérer pour la décharge de la BL. Cette situation apparait quand toutes les cellules d'une même colonne contiennent une donnée

telle que tous les transistors connectés à la BL sont passants. La capacité de BL à décharger est alors triplée (3 capacités de jonction par cellule) tandis que le courant de décharge I_{read} reste le même. Ceci nous amène alors à une idée inspirée de [9] : le *buffer* dit 1T (figure 3.21(c)).

Buffer 1T

Dans cette architecture, la WL est la grille du transistor en régime linéaire, ainsi que d'un transistor PMOS. La sortie de ce pseudo-inverseur est connectée à la source du transistor en régime saturé, qui contient maintenant l'information sur la donnée. Le transistor PMOS maintient une tension haute quand la ligne n'est pas accédée. Tous les transistors connectés aux cellules sont donc en régime bloqué, même si leur tension de grille est haute. La BL voit alors une haute impédance et une seule capacité par cellule contribue à la capacité totale de BL. Lors d'une lecture, le transistor PMOS se coupe et un chemin de décharge se crée entre la BL et la masse, si le transistor de la cellule est passant.

Pour avoir la même vitesse de décharge, il faut considérer le pire des cas, où tous les transistors connectés à la pseudo-WL sont passants. Donc, le transistor NMOS de l'inverseur devra avoir une taille n (128) fois plus grande que la taille minimale. Pour diminuer cette taille, on peut par exemple utiliser un *assist* dynamique présenté dans [9]. Mais cela sort du cadre de ce travail.

Le design du transistor PMOS est *a priori* non trivial. S'il est trop "fort", son courant risque de devenir non négligeable lors d'une lecture, augmentant alors le temps de décharge. De plus, il augmentera la capacité de WL. S'il est trop faible, il ne rechargera pas assez vite la pseudo-WL, et les transistors de la ligne pourront encore être passants lors de la prochaine lecture. Néanmoins, les simulations ont montré qu'avec un simple transistor SVt de taille minimale, la pseudo-WL atteint une tension haute bien avant 1ns. Sa consommation statique se réduit au courant de fuite de grille, car la tension drain-source V_{ds} est nulle en rétention.

Finalement, une lecture différentielle dans cette architecture s'obtient en ajoutant un deuxième transistor à la cellule, et sa BL correspondante. Ainsi, l'un des deux transistors est nécessairement bloqué et l'autre passant. La lecture se déroule de la même manière que son pendant à BL unique, dans le cas où tous les transistors de la ligne sont passants. Remarquons que cette architecture de lecture différentielle ne requiert que trois transistors par cellule, contrairement à quatre dans le *buffer* 2T différentiel, et un seul transistor est connecté à la WL (si on normalise au nombre de cellules par ligne).

Simulations dynamiques

Les résultats complets sont présentés en annexe A, nous résumons ici les principaux points de comparaison/les principales conclusions de la comparaison.

Pour commencer, il faut noter que, dans notre circuit de test, la chute de $100mV$ sur la BL apparait bien avant que la tension de WL n'arrive à la tension d'alimentation V_{dd}. Cela implique que le courant I_{read} effectif est inférieur à la valeur du tableau 3.4, mais surtout que le temps de charge de la WL est le paramètre principal pour le temps d'accès en lecture. Si les dimensions du tableau ou du *driver* sont telles que la WL se charge très rapidemment par rapport au temps total de lecture, le courant I_{read} deviendra sans doute le paramètre principal.

L'inverseur *tri-state* présente le plus grand temps d'accès en lecture. Comme il n'utilise pas de *sense-amplifier*, il doit charger la BL jusque $V_{dd}/2$. Ses performances relatives dépendent des tensions de seuil des transistors, comme discuté précédemment.

Parmi les autres *buffers*, qui demandent un *sense-amplifier*, le 2T différentiel est le plus lent parce que la capacité de WL est double.

La différence d'architecture entre les *buffers* 2T et 1T provoque une asymétrie des transistors : pour un *buffer* 2T, le transistor en régime saturé (M2) voit sa tension de WL progressivement s'élever au fur et à mesure de la charge de la WL. Par contre, dans le *buffer* 1T, c'est le transistor en régime linéaire (M1) qui voit sa tension V_{gs} progressivement s'élever. Du fait de cette différence, la dépendance du temps d'accès avec les tensions de seuil des transistors du *buffer* est différente entre les deux types de buffer (figure 3.24). Le temps d'accès du *buffer* 2T diminue avec le V_t du transistor d'accès, tandis que le *buffer* 1T suit une évolution plus compliquée. Notons par ailleurs que la différence entre le temps d'accès de la configuration la plus lente et de la plus rapide est seulement de 25%, et que le plus grand temps d'accès respecte largement la contrainte de vitesse, même en tenant compte des périphériques externes. Finalement, le *buffer* 1T différentiel a la même vitesse que son pendant à BL unique.

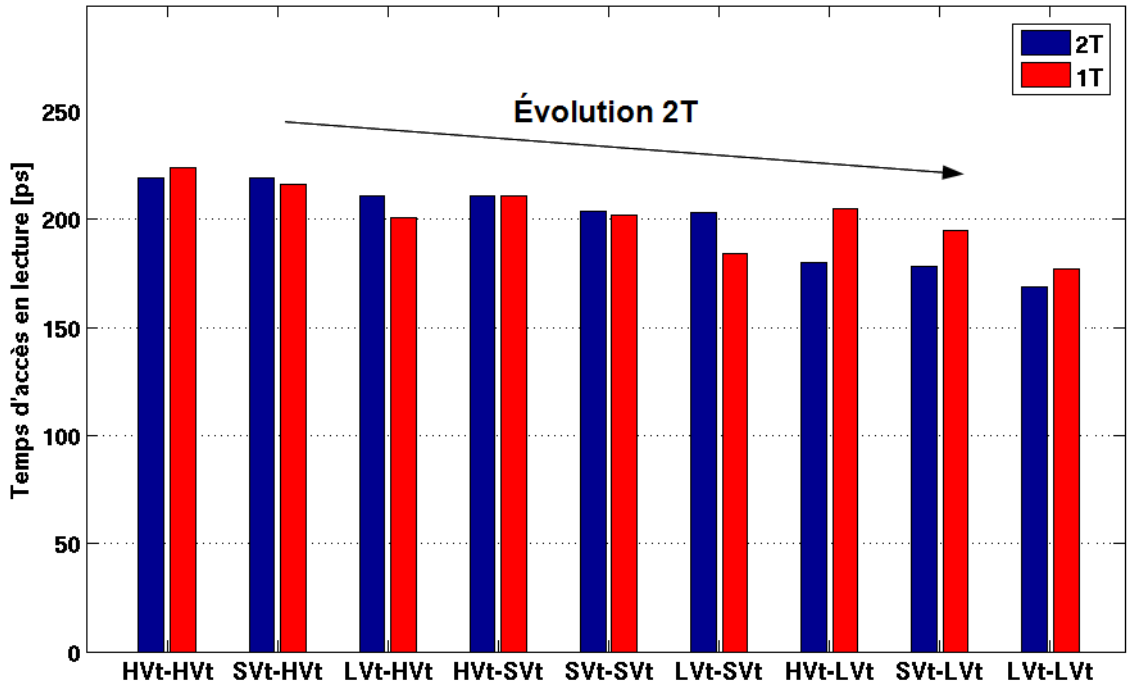


Figure 3.24 – Les temps d'accès en lecture en fonction des caractéristiques physiques des deux transistors du *buffer* (voir figure 3.22)

Pour un même type de transistor connecté à la WL, le *buffer* 1T consomme significativement moins que le *buffer* 2T, sauf pour un transistor de type LVt connecté à la WL. Enfin, la puissance statique dépend fortement du courant de fuite de grille. Dans le *buffer* 2T, le transistor connecté au *latch* est la moitié du temps en mode *forward*, donc à courant de grille non négligeable. Dans un *buffer* 1T, il est en mode *reverse*.

3.3.2 Consommation dynamique

A capacité de WL égales, la différence d'énergie dynamique entre deux *buffers* dépend essentiellement de la chute de tension sur la BL. Dans le cas idéal, cette chute de tension est toujours la même, car nous avons imposé cette condition pour définir le temps d'accès

en lecture. Néanmoins, pour une analyse plus réaliste, il faut tenir compte de la variabilité. En effet, le temps d'activation de la WL est fixé pour que la cellule la plus lente respecte toujours la contrainte de temps. Dès lors, les cellules rapides auront en pratique beaucoup de temps pour décharger la BL. La BL étant nécessairement rechargée à la fin de la lecture (sauf pour l'inverseur *tri-state*), l'énergie dynamique moyenne sera en fait beaucoup plus grande que dans le cas idéal (figure 3.25). Dans [4], l'auteur compare le *buffer* 2T et inverseur *tri-state* en technologie 45nm Bulk, et montre que l'inverseur *tri-state* est globalement plus efficace en terme de consommation. La raison est que l'inverseur *tri-state* ne consomme pas d'énergie lors d'une transition si deux données successives sont identiques. A l'inverse, le *buffer* 2T différentiel obtenait une différence de tension moyenne entre les deux BL de 0.8V à *chaque lecture*, quelles que soient les valeurs relatives des données. Donc, même si l'inverseur demande de charger et décharger complètement la BL, il consomme en moyenne moins que le *buffer* 2T différentiel qui décharge une grande partie de la BL à chaque lecture.

Afin d'effectuer une étude similaire, la méthodologie suivante sera adoptée : le temps d'accès en lecture sera déterminé grâce à une simulation Monte-Carlo dans le *corner* SS. En effet, il est nécessaire de fixer le pire des cas pour rendre toutes les cellules fonctionnelles. Afin d'évaluer la consommation maximale dépensée à chaque lecture, l'énergie dynamique moyenne sera ensuite mesurée dans le *corner* FF. En effet, ce *corner* garantit une décharge maximale de la BL pendant le temps d'accès.

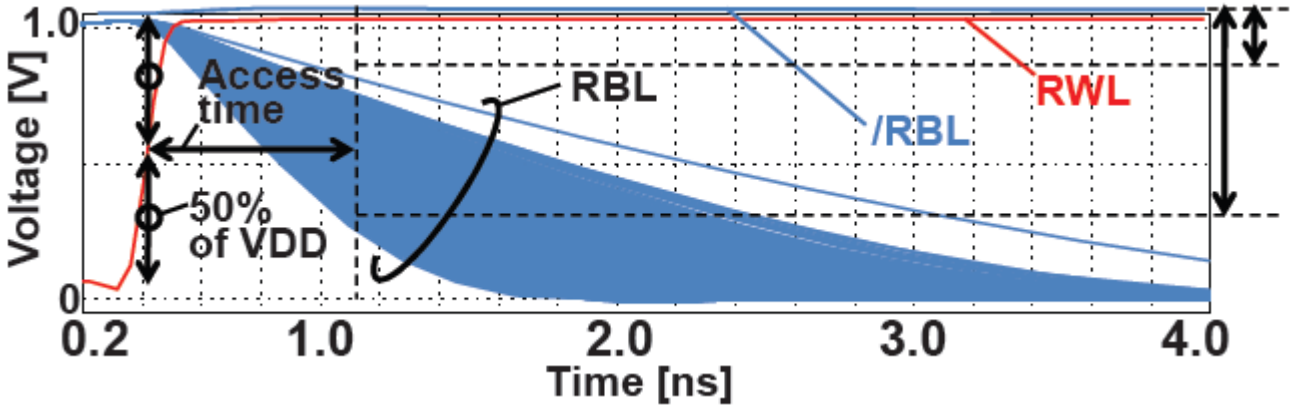


Figure 3.25 – Les cellules plus rapides déchargent énormément la BL durant un temps d'accès en lecture [4].

L'auteur de [4] n'a pas comparé l'inverseur *tri-state* avec les *buffers* de type 1T. *A priori*, on peut espérer que la tension milieu commune à tous les transistors d'une ligne permettra un moyennage du temps d'accès : les transistors des cellules plus rapides verront leur tension V_{gs} diminuer, leur courant de décharge deviendra négligeable et les autres cellules auront donc plus de facilité à décharger leur BL. Si la tension de BL atteint la tension milieu (pseudo-WL sur la figure 3.21(c)), elle ne peut descendre plus bas car le transistor connecté à la cellule commencera à la *charger*. Cette sorte de contre-réaction pourrait permettre de s'approcher du cas idéal, et donc de diminuer l'énergie dynamique totale. En moyenne, la moitié des données d'une ligne sont à 0 et l'autre à 1. Le *buffer* 1T unique aura donc une vitesse effective beaucoup plus grande que dans le pire des cas où, rappelons-le, tous les transistors de la ligne sont ouverts. Cette propriété pourrait alors augmenter l'énergie dynamique, car plus de courant est disponible pour décharger la BL. Néanmoins, le courant maximal ne peut être que celui d'un seul transistor. Or, ce courant maximal est deux voire trois fois plus grand (en fonction des paramètres physiques) que le courant I_{read}, tandis que les BL des cellules contenant la donnée inverse ne seront pratiquement pas déchargées. L'énergie dynamique moyenne pourrait alors globalement diminuer.

Table 3.6 – Comparaison de l'énergie dynamique pour différents types de buffer. L'inverseur *tri-state* consomme la plus grande énergie dynamique dans cette technologie. La variabilité de l'inverseur *tri-state* n'influence pas son énergie dynamique moyenne.

Buffer	Temps d'accès (corner SS)			Energie dynamique (corner FF)		
	μ [ps]	$\frac{\sigma}{\mu}$ [%]	$\mu + 6\sigma$ [ps]	Pire cas [fJ]	Moyenne [fJ]	Meilleur cas [fJ]
1T HVt-LVt	121	6.3	167	5.8	4.5	0.2
2T HVt-LVt	100.5	5.0	130.5	5.1	2.64	0.18
1T diff HVt-LVt	122	6.3	168	5.8	5.8	5.8
2T diff HVt-LVt	138	5.9	188	6.8	6.8	6.8
Inv HVt-LVt	650 ps	-	-	23.2	11.7	0.2

Le tableau 3.6 reprend les résultats de simulation, et est très riche en enseignements. Il apparait clairement que, dans la technologie 32nm FDSOI du Leti, les *buffers* 2T et 1T ont une énergie dynamique inférieure à celle de l'inverseur *tri-state*, même en tenant compte de la variabilité ! [4] explique que dans le traitement d'image, où des données successives sont très souvent identiques, l'avantage de l'inverseur se fait encore plus ressentir. On voit dans le tableau 3.6 que même avec deux données identiques d'affilée sur trois, l'énergie du *buffer* 2T différentiel est toujours plus faible. Ce nombre passe à 8 données identiques d'affilée sur 9 pour le *buffer* 2T à BL unique. De plus, un mécanisme auto-adaptatif pourrait régler le temps d'accès en fonction du *corner*, ce qui diminuerait encore plus l'énergie dynamique des buffers utilisant un *sense-amplifier*. Néanmoins, un *sense-amplifier* est une source importante de consommation d'énergie [21]. Il devrait donc intervenir dans le calcul de l'énergie totale par accès en lecture. Comme ce système n'est pas modélisé dans ce travail, l'inverseur *tri-state* sera abandonné.

Le temps d'accès du *buffer* 1T à BL unique doit être défini dans le cas critique où tous les transistors de la ligne sont passants. L'énergie dynamique moyenne, elle, est définie quand la moitié des transistors sont passants. L'énergie supplémentaire, due à la décharge accentuée sur la moitié des BL, et au courant de court-circuit provenant des cellules non accédées, est plus du double que l'énergie de chaque BL obtenue avec 128 transistors ouverts. C'est pourquoi, au final, l'énergie dynamique moyenne est plus grande que la moyenne arithmétique entre le pire et le meilleur des cas ; mais elle est cependant plus petite que dans le pire des cas.

La variabilité du *buffer* 2T différentiel est proportionnellement plus grande que celui à BL unique. La capacité de WL étant double, le temps de charge par un même *buffer* est approximativement doublé également. Or, le temps d'accès dépend également du courant I_{read} qui décharge la BL. Lors de la montée de la tension de WL, le transistor d'accès devient *progressivement* passant, et le courant qui le traverse est dans un premier temps un courant sous-seuil, dépendant fortement de la variabilité, puis un courant de saturation. Cette période de transition étant plus longue dans le cas différentiel, cela peut expliquer la plus grande variabilité sur le temps d'accès.

Le temps d'accès du *buffer* 2T différentiel est donc proportionnellement plus grand, ce qui entraine une décharge plus importante de la BL dans le cas nominal, et donc une énergie dynamique plus importante que dans le pire des cas du *buffer* 2T à BL unique. Il faut noter que dans le buffer 1T differentiel, la tension de BL se maintient à une certaine valeur (environ 0.8V) du fait des courants de courts-circuits provenant des cellules non-accédées. Il y a un gain énergétique du fait de la décharge moins grande de la BL, et en même temps une perte due aux

courts-circuits. On voit que le gain total est positif dans le cas 1T. De plus, la capacité WL du 1T différentiel est deux fois moins grande que celle du 2T différentiel.

Résumé

L'inverseur *tri-state* présente une énergie dynamique supérieure à celle des autres *buffers*, alors qu'elle devrait être son principal avantage. Ce *buffer* sera donc abandonné pour la suite de ce travail.

L'énergie dynamique moyenne par lecture d'un *buffer* de type 2T est plus faible que celle d'un *buffer* de type 1T dans le cas à BL unique. Pour déterminer lequel des deux consomme le moins, il faut également tenir compte de la puissance statique. Avec le même raisonnement suivi pour les *latchs* (section 3.2.7) :

$$E_{tot} = \alpha_{SW} 128 E_{Read,cell} + 128 \cdot 255 \cdot P_{stat,cell} \cdot 1ns + 128 \cdot 256 \cdot P_{stat,cell} \cdot T$$

Les *buffers* à lecture différentielle consomment plus que les buffers à lecture à BL unique. Dans la configuration différentielle, le *buffer* 1T consomme moins que le *buffer* 2T, que ce soit en énergie dynamique ou statique.

3.3.3 Rapport I_{read}/I_{off} et surface de silicium

Si un grand nombre de cellules peuvent être connectées par BL, les *drivers* et autres systèmes périphériques (*sense-amplifier*,...) seront moins nombreux, pour une même capacité mémoire, ce qui amènera un gain global en surface et en consommation. Néanmoins, ce nombre est limité par les courants de fuite de la BL.

Dans les *buffers* 2T, si la donnée de la cellule activée est telle que le transistor connecté au *latch* est coupé, la tension de la BL devrait idéalement rester à la tension d'alimentation V_{dd}. Néanmoins, comme le transistor PMOS de précharge est coupé, les courants de fuite des cellules de la colonne entrainent une légère diminution ΔV_{BL} de la tension de BL. Le courant de fuite d'un transistor d'accès est maximal si le deuxième transistor du *buffer* est passant. Il est donc nécessaire de s'assurer que la valeur de ΔV_{BL} reste inférieure à la différence de tension minimale qui ferait basculer le *sense-amplifier* connecté à la BL, et ce dans le pire des cas où tous les transistors d'accès connectés à une BL ont un courant de fuite maximal. Pour déterminer si cette contrainte est respectée, le courant I_{read} est comparé au courant I_{off}.

Pour les *buffers* 1T (et l'inverseur *tri-state*), le calcul est le même mais le phénomène est différent. Le problème se pose lorsque la tension de BL ***doit*** chuter. Les transistors NMOS connectés au *latch* des cellules non accédées ont leur tension de source à V_{dd}. Si leur tension de grille est également haute, ils voient leurs tensions V_{ds} ***et*** V_{gs} progressivement augmenter durant la décharge de la BL. Ils injecteront donc un courant sur celle-ci qui aura l'effet inverse souhaité. Il est donc nécessaire de s'assurer que le transistor passant de la cellule activée pourra forcer la BL à la tension de basculement du *sense-amplifier*.Pour déterminer si cette contrainte est respectée, le courant I_{read} est à nouveau comparé au courant I_{off}, tel que défini sur la figure 3.26.

Le courant I_{off} défini pour le buffer 2T diminue avec la tension de BL, tandis que le courant I_{off} défini pour le buffer 1T augmente si la tension de BL diminue. La figure 3.27 compare les deux courants I_{off} en fonction de la tension de la BL. On constate que pour une tension haute de BL, le *buffer* 1T est plus performant que le 2T. La tendance s'inverse à partir d'une certaine

valeur, plus petite que $900mV$. Donc, pour une chute de la BL inférieure ou égale à $100mV$, le *buffer* 1T présente un courant I_{off} plus petit que le *buffer* 2T. La valeur de ce courant dépend de la tension de seuil du transistor connecté à la BL, avec un ou deux ordres de grandeur de différence entre chaque V_t. Notons que l'utilisation d'une WL négative en rétention diminue drastiquement le coruant I_{off}.

Dans les deux cas, le courant I_{off} dépend donc uniquement de la tension de seuil du transistor connecté à la BL ; tandis que le courant I_{read} dépend des deux transistors comme indiqué au tableau 3.4. Le tableau 3.7 montre que le rapport I_{read}/I_{off} diminue avec la tension de seuil du transistor connecté à la BL. Un compromis doit être trouvé entre la vitesse et la surface de silicium. Puisque le rapport I_{read}/I_{off} dépend de la tension de seuil d'un transistor, il dépend forcément du *corner* dans lequel se situe le tableau mémoire. Le tableau 3.8 montre que le rapport I_{read}/I_{off} se dégrade pour les corners à bas V_t des transistors NMOS, et inversément pour les *corners* à haut V_t. Pour être valable dans tous les *corners*, le rapport I_{read}/I_{off} d'une lecture à BL unique doit être calculé différemment. Il faut prendre le courant I_{read} le plus petit parmi tous les *corners* (SF), avec le courant I_{off} le plus grand parmi tous les *corners* (FS) :

$$\frac{I_{read}}{I_{off}}_{single} = \frac{I_{read,min}}{I_{off,max}} = \frac{I_{read,SF}}{I_{off,SS}}$$

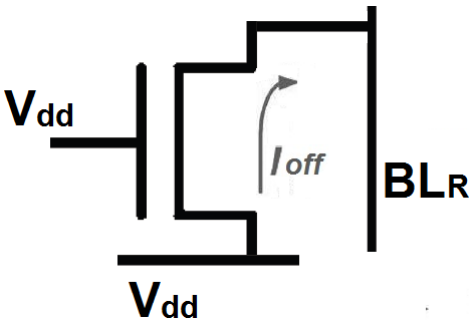


Figure 3.26 – Définition du courant I_{off} pour les *buffers* de type 1T.

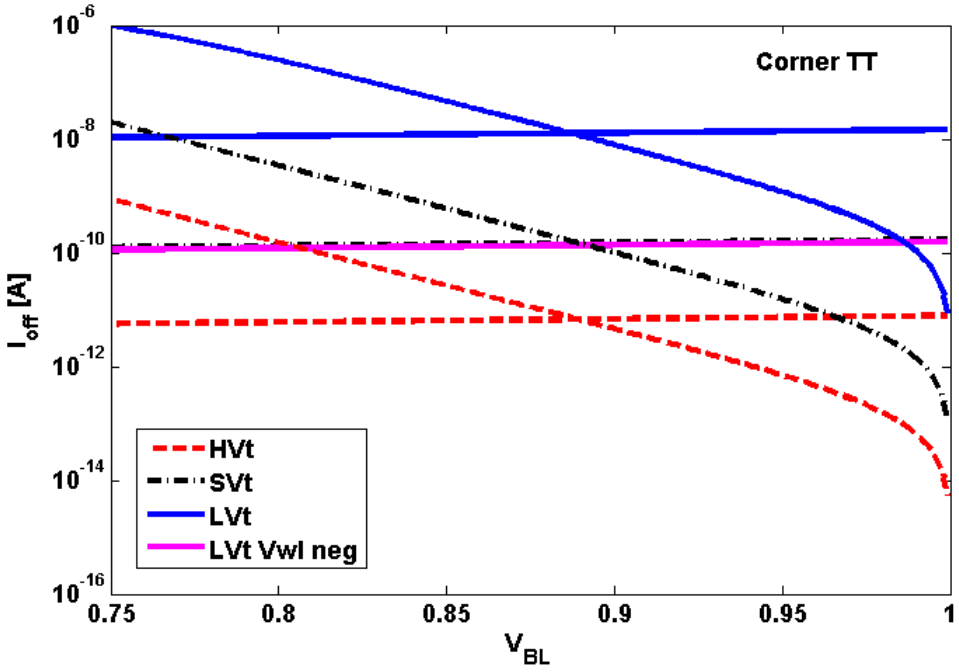


Figure 3.27 – Les différents courants I_{off} en fonction de la tension de BitLine. Pour un même V_t, le courant I_{off} de 1T est plus petit que 2T pour de grandes valeurs de V_{BL}.

Table 3.7 – Comparaison des courants I_{read} et I_{off} pour différents *buffers*. Le rapport I_{read}/I_{off} varie de plusieurs ordres de grandeur selon la tension de seuil des transistors d'accès.

Type de *buffer*	I_{read}	I_{off}	I_{read}/I_{off}
2T 1-1 SVt-**HVt**	28.2 μA	2.5 pA	1.13 10^7
2T 1-1 SVt-**SVt**	33.6 μA	52.4 pA	6.41 10^5
2T 1-1 SVt-**LVt**	41.2 μA	4.43 nA	9.3 10^3
2T 1-1 SVt-LVt V_{WL} nég.	41.2 μA	47.6 pA	8.66 10^5
2T **2**-1 SVt-SVt	43.3 μA	52.4 pA	8.27 10^5

Table 3.8 – Comparaison des courants I_{read} et I_{off} d'un *buffer* pour différents *corners*. Le rapport I_{read}/I_{off} diminue globalement avec la tension de seuil des transistors NMOS.

Buffer 2T SVt-SVt	I_{read}	I_{off}	I_{read}/I_{off}
Corner SF	**27.5** μA	15.9 pA	17.2 10^5
Corner SS	30.3 μA	20.9 pA	14.4 10^5
Corner TT	33.6 μA	52.4 nA	6.4 10^5
Corner FF	37.1 μA	120.6 pA	3.1 10^5
Corner FS	40.8 μA	**171.15** pA	2.4 10^5

Pour une lecture différentielle, même si le temps d'accès est défini dans le *corner* le plus lent, le nombre maximal de cellules par BL doit être pris uniquement dans le *corner* critique.

$$\frac{I_{read}}{I_{off\ diff.}} = \frac{I_{read}}{I_{off\ min}} = \frac{I_{read,FS}}{I_{off,FS}}$$

Du tableau 3.8, on calcule une augmentation de 48% de $I_{read,FS}$ par rapport à $I_{read,SF}$. Donc, un gain de 48% est dû grâce à la lecture différentielle.

Finalement, la tension de seuil du transistor connecté à la BL est un compromis entre la vitesse et la surface de silicium.

3.4 Conclusion

Dans la première section de ce chapitre, nous avons procédé à une étude sommaire de la cellule *dual-port* conventionnelle à 8 transistors. La conclusion fut que beaucoup de compromis étaient nécessaires pour maintenir une marge de bruit en lecture acceptable. L'utilisation d'un *buffer* de lecture semblait être une alternative prometteuse et nous a conduits à une étude des cellules SRAM *dual-port* en deux temps : d'abord une comparaison des systèmes d'écriture, puis une comparaison des systèmes de lecture.

Dans la deuxième section, nous avons analysé chaque système d'écriture. Les cellules 5T et ULP demandent un effort supplémentaire de dimensionnement pour obtenir des marges de bruit suffisantes en rétention et en écriture. Nous avons démontré qu'une polarisation négative de la WL en rétention n'avait un avantage énergétique qu'avec des transistors d'accès de type

LVt et pour une valeur très précise de la tension, définie à partir de la formule 3.1. Enfin, nous avons comparé les systèmes d'écriture. La cellule 5T est la plus rapide mais son dimensionnement entraine une consommation statique énorme. Les cellules 6T sont inconditionnellement plus rapides que les cellules 5TPMOS, mais consomment en moyenne deux fois plus d'énergie par écriture. L'*assist* statique utilisé pour la WL diminue significativement la consommation statique et le temps d'écriture. En tenant compte à la fois de l'énergie dynamique et de la puissance statique, les cellules ont pu être listées en fonction de leur consommation. Les cellules ULP et 5TPMOS accès HVt présentent une consommation minimale mais ne respectent pas la contrainte de vitesse face à la variabilité de fabrication. La cellule 5TPMOS accès SVt est alors le système d'écriture fiable et robuste qui consomme le moins d'énergie.

Dans la troisième section, tous les systèmes de lecture furent comparés performance par performance. Dans un premier temps, il est apparu que la tension de seuil du transistor du *buffer* en régime saturé déterminait la puissance statique et le courant de décharge de la BL. Un compromis devait donc être trouvé entre le temps d'accès en lecture et la consommation statique. Différentes pistes ont été proposées pour contourner ce compromis, dont l'utilisation d'un *buffer* de type 1T. Ensuite, des simulations dynamiques ont montré que les temps d'accès en lecture sont relativement identiques pour une même configuration de transistors, excepté pour le *buffer* 2T différentiel qui paye sa capacité de WL double.
La consommation dynamique a ensuite été étudiée avec une méthodologie particulière. La conclusion est que le *buffer* appelé inverseur *tri-state* présente une consommation dynamique bien supérieure aux autres, alors qu'elle est sa principale performance dans [4]. Ce *buffer* a donc été abandonné pour la suite du travail. Le *buffer* 1T à BL unique consomme plus d'énergie par lecture que le 2T, à cause de ses courants de fuite qui dépendent exponentiellement de la tension de BL. L'ordre s'inverse pour les buffers différentiels, qui consomment en moyenne plus que leur pendant à BL unique.
Dans la troisième partie, nous avons étudié le rapport I_{read}/I_{off} donnant le nombre maximal de cellules par BL. Ce rapport dépend du type de *buffer*, ainsi que de la tension de seuil du transistor connecté à la BL. Les *buffers* 1T permettent un plus grand nombre de cellules par BL pour une sensibilité du *sense-amplifier* inférieure ou égale à $100mV$. Après avoir étudié toutes les performances, on constate que le *buffer* 1T différentiel est plus efficace en terme de vitesse, consommation et surface de silicium que le *buffer* 2T différentiel. Tous *buffers* confondus, le buffer 2T HVt-HVt à BL unique présente la meilleure consommation (statique et dynamique combinées) tout en respectant la contrainte de temps et en offrant un rapport I_{read}/I_{off} plus qu'acceptable.

Après avoir étudié et comparé toutes les performances des systèmes d'écriture et de lecture, il est possible d'établir une comparaison avec la cellule de l'état de l'art. La cellule complète 5TPMOS accès SVt et *buffer* 2T HVt-HVt à BL unique (figure 3.28) présente une puissance statique moyenne de $45.7pW$ et une énergie moyenne par accès de $7.7fJ$, dans des conditions typiques (*corner* TT, à 27°C, sous une tension d'alimentation de 1V). De plus, elle requiert 8 transistors de taille minimale. La cellule *dual-port* conventionnelle à huit transistors de type HVt, présente une puissance statique de $27.3pW$ et une énergie moyenne par accès de $16.4fJ$, sous les mêmes conditions. Elle requiert 6 transistors de taille minmale et 2 transistors de taille double pour garantir sa stabilité. Si un seul des transistors de la cellule 8T est de type SVt, sa consommation statique sera plus grande que la cellule complète 5TPMOS accès SVt et *buffer* 2T HVt-HVt. De plus, si on tient compte de la consommation dynamique comme montré dans les sections 3.2.7 pour l'écriture et 3.3.2 pour la lecture, la cellule 8T a toujours une plus grande consommation en mode actif.

Le système de lecture 5TPMOS accès SVt pèche par son temps d'écriture très grand par rapport aux autres cellules. Si cette cellule ne respecte plus la contrainte du circuit, les cellules 5TPMOS

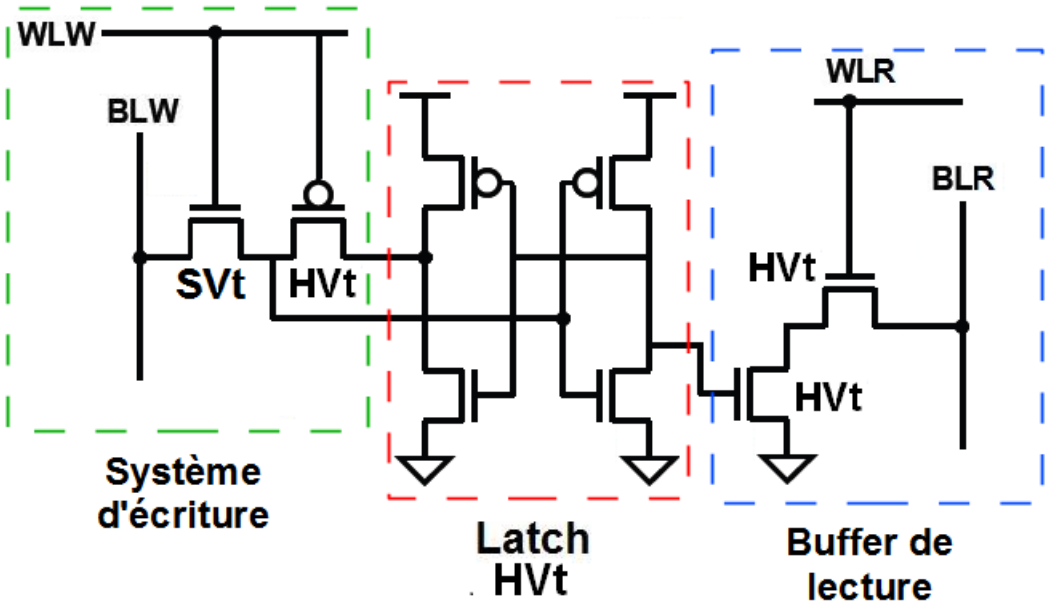


Figure 3.28 – La cellule complète proposée : *latch* 5TPMOS avec transistor d'accès SVt et *buffer* 2T avec deux transistors HVt. Tous les transistors son de taille minimale.

accès LVt et 6T accès HVt pourraient la remplacer tout en gardant une consommation inférieure aux cellules de l'état de l'art. Notons que la cellule 6T accès HVt présente même une vitesse d'écriture plus grande que celle de la cellule 8T HVt.

La vitesse de lecture de la cellule 8T dépend des mêmes paramètres que le système de lecture 2T différentiel, la comparaison peut donc facilement être faite. Malgré sa forte consommation, une lecture différentielle peut être une condition incontournable dans une application. De part son architecture, elle permet une plus grande immunité face au bruit environnant (*crosstalk*, RTS, radiation,...) et permet un meilleur calibrage de la tension de basculement du *sense-amplifier*. Si ce type de lecture s'avère nécessaire, le *buffer* 1T différentiel sera préféré, car sa consommation dynamique et statique est moins importante.

Une dernière observation importante de ce chapitre est que le temps d'écriture est significativement plus élevé que le temps de lecture, tous paramètres et architectures confondus.

Une étude plus approfondie des cellules SRAM doit tenir compte de la variation de tension d'alimentation du circuit, ainsi que de la température ambiante. Le chapitre suivant étudie le comportement des performances face à ses variations, en utilisant la cellule complète 5TPMOS accès SVt et *buffer* 2T HVt-HVt pour les résultats chiffrés.

Table 3.9 – Comparaison de la cellule proposée à une cellule de l'état de l'art.
Corner TT, $V_{dd} = 1V$, Température 27°C.

	T Write [ps]	T Read [ps]	E Write [fJ]	E Read [fJ]	P stat [pW]	SNMH [mV]	SNMW [mV]
8T HVt	449	325	**15.1**	**1.3**	27.3	404	-344
Cellule proposée	488	198	**7.1**	**0.6**	45.7	414	-456

Chapitre 4

Étude PVT

Dans ce chapitre, le comportement des cellules SRAM sera étudié en fonction de la tension d'alimentation, de la température et des variabilités de fabrication (*Process Voltage Temperature* ou PVT). A cause des courants importants pouvant traverser les rails d'alimentation du circuit, les connections du packaging, et les pistes des PCB, une chute résistive et/ou inductive peut induire une variation de la tension d'alimentation vue par le circuit. Une variation typique que les circuits numériques doivent être capables de supporter est de ±10%. Il est donc nécessaire de garantir le fonctionnement de la cellule à 0.9V et 1.1V d'alimentation. Pour garantir une application large public, le tableau mémoire doit également être fonctionnel pour une certaine plage de température. L'intervalle de température pris en charge dans les modèles va de -40°C à 125°C. La variabilité de fabrication quant à elle a déjà été discutée dans la section 1.5.
Pour ne pas entamer une étude inutilement longue, l'étude se focalisera d'abord sur un transistor NMOS de la technologie 32nm FDSOI du Leti, puis sur la cellule complète 5TPMOS accès SVt et *buffer* 2T HVt-HVt. En particulier, les limites en vitesse et marge de bruit seront déterminées, ainsi que les pics extrêmes de consommation de la cellule que l'on peut attendre dans cette technologie.

4.1 Impact sur le transistor FDSOI

Pour mieux comprendre le comportement d'une cellule SRAM complète, il est utile de connaitre les effets de la tension d'alimentation et de la température sur les performances d'un transistor isolé. Le tableau 4.1 montre les performances en fonction de la tension d'alimentation pour un transistor SVt avec $W_g = 1\mu m$. Une variation de 10% de la tension d'alimentation entraine une variation identique d'environ 20% des courants de saturation, mais seulement 10% pour les courants de fuite. Les courants en régime linéaire ont pratiquement une dépendance linéaire avec la tension d'alimentation.

Table 4.1 – Transistor NMOS SVt $W_g = 1\mu m$. Corner TT. Temp = 27°C.

V_{dd}	$I_{on}[\mu A]$ ($V_{ds} = V_{dd}$)	$I_{on}[\mu A]$ ($V_{ds} = 50$mV)	$Ioff[pA]$ ($V_{ds} = V_{dd}$)
0.9V	624 (−18%)	100 (−9%)	577 (−12%)
1V	763 (+0%)	110 (+0%)	655 (+0%)
1.1V	890 (+17%)	118 (+7%)	744 (+13%)

Les figures 4.1 et 4.2 montrent l'évolution de ces mêmes courants avec la température. Comme on peut le voir, le courant I_{on} varie peu dans la gamme de température étudiée (-40°C à 125°C) pour ce modèle de transistor. On observe une diminution de 9% pour les transistors LVt, et seulement 4% et 6% pour les transistors HVt et SVt respectivement. Le courant sous seuil I_{off}, lui, varie très nettement avec la température. Le courant de fuite du transistor HVt est par exemple multiplié par un facteur 113 en passant de 27°C à 125°C.

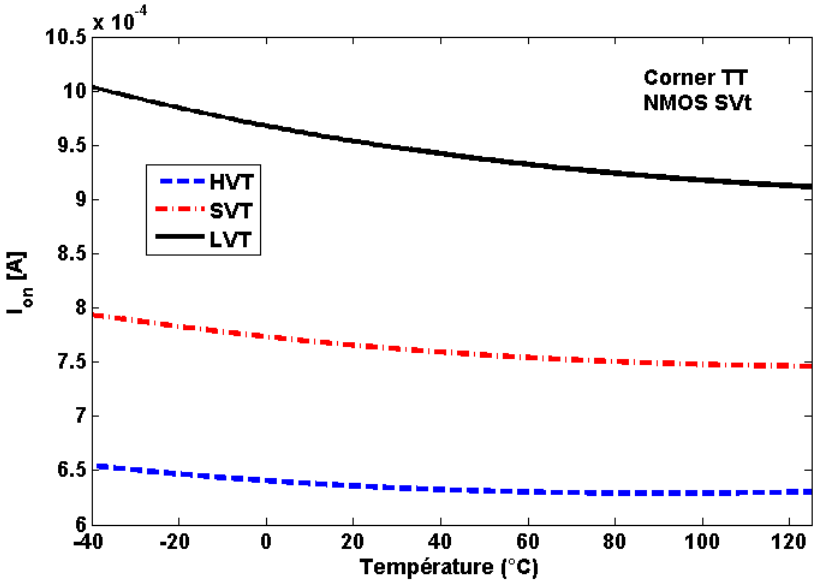


Figure 4.1 – Courant I_{on} par μm de drain en fonction de la température (échelle linéaire).

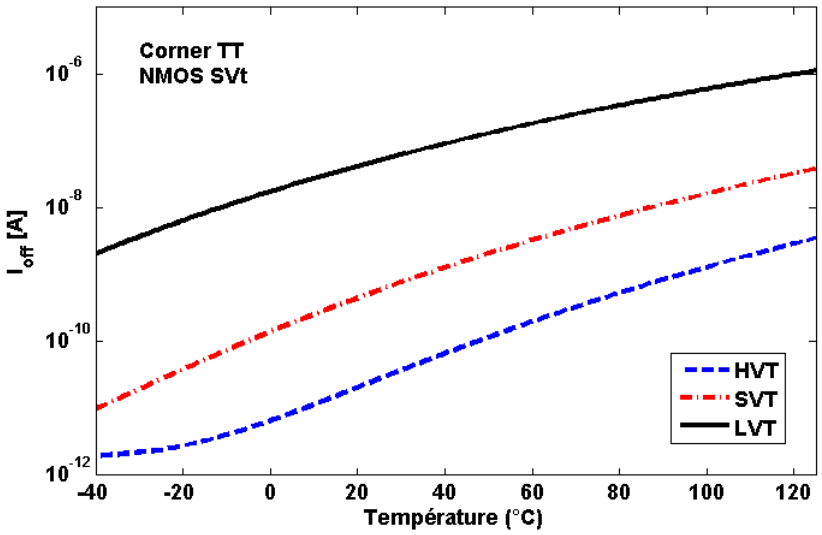


Figure 4.2 – Courant I_{off} par μm de drain en fonction de la température (échelle logarithmique).

4.2 Impact sur la cellule SRAM *dual-port*

Au vu du comportement des transistors face aux variations de leur environnement, les tests suivants semblent les plus pertinents pour une étude plus appronfondie de la cellule SRAM :

1. La limite de stabilité. Pour commencer, le *corner* le plus critique sera déterminé. Ensuite, l'évolution de la marge de bruit avec la température sera illustrée. Finalement, des simu-

lations Monte-Carlo seront effectuées dans le *corner* et la température critique, à 0.9V d'alimentation car une diminution de la tension d'alimentation entraine une diminution de la marge de bruit [32].

2. La consommation statique. La puissance statique maximale est logiquement obtenue dans le *corner* FF, à une tension d'alimentation de 1.1V, pour les températures élevées. L'énergie dynamique dépend essentiellement des charges de capacité (voir les chapitres précédents). Elle suivra alors une évolution quadratique avec la tension d'alimentation, qui a été vérifiée en simulation. Les variations de l'énergie dynamique ne seront donc pas étudiées ici.
3. Le rapport I_{read}/I_{off}. Le courant I_{read} est une combinaison de courant de saturation et linéaire, variant peu avec la température. Le courant I_{off}, par contre, s'élève fortement avec la température. Le pire des cas à considérer est donc les hautes températures dans le *corner* FS (section 3.3.3).
4. La limite en vitesse. Les temps d'accès en écriture et lecture seront simulés dans le *corner* SS à une tension d'alimentation de 0.9V, pour des températures représentatives de la plage considérée. En effet, ces conditions entrainent la plus forte réduction de courant pour les transistors, donc un temps minimal de charge des capacités.

Limite de stabilité

Le *corner* le plus critique pour la marge de bruit en rétention est le *corner* FS (tableau 4.2(a)). En effet, ce *corner* asymétrise le plus les inverseurs composant le *latch*. Le tableau 4.2(b) montre que la marge de bruit se dégrade pour les hautes températures. Les simulations Monte-Carlo ont montré que le *latch*, composé de deux inverseurs montés en tête-bêche avec des transistors HVt de taille minimale, présente une déviation standard pour le SNMH d'environ 25mV. Le *latch* est donc parfaitement robuste à 6σ.

Table 4.2 – Marge de bruit en rétention en fonction du *corner* et de la température du circuit. Le pire des cas à considérer est le *corner* FS dans les hautes températures.

(a) SNMH en fonction du *corner*, à 27°C.

Corner	SNMH [mV]
SS	419
SF	423
TT	414
FS	**386**
FF	404

(b) SNMH en fonction de la température, *corner* FS.

Temperature	SNMH [mV]
-40°C	396
0°C	390
27°C	414
50°C	379
125°C	**358**

Comme nous l'avons vu dans la section 3.2, pour une cellule de type 5TPMOS, la marge de bruit en écriture dépend uniquement de la tension de basculement du *latch*, et le pire des cas à considérer est l'écriture d'un zéro. Nous aurions donc les mêmes comportements que le SNMH, qui dépend lui aussi des caractéristiques des inverseurs du *latch*. De plus, le SNMW est plus grand que le SNMH pour les cellules 5TPMOS (voir tableau 3.2), il est donc moins critique à étudier.

Consommation statique

La consommation statique maximale est atteinte à 1.1V d'alimentation, dans le *corner* FF. Le graphique 4.3 montre l'évolution de la puissance statique en fonction de la température. A 125°C, la puissance est multiplié par un facteur 25 par rapport à la température classique de 27°C. Ce facteur est moindre que pour le transistor NMOS isolé de la section précédente. Ce dernier avait été calculé dans le *corner* TT. La variation des courants de fuite avec la température est moins importante dans le *corner* FF, mais la puissance statique reste plus importante en absolue.

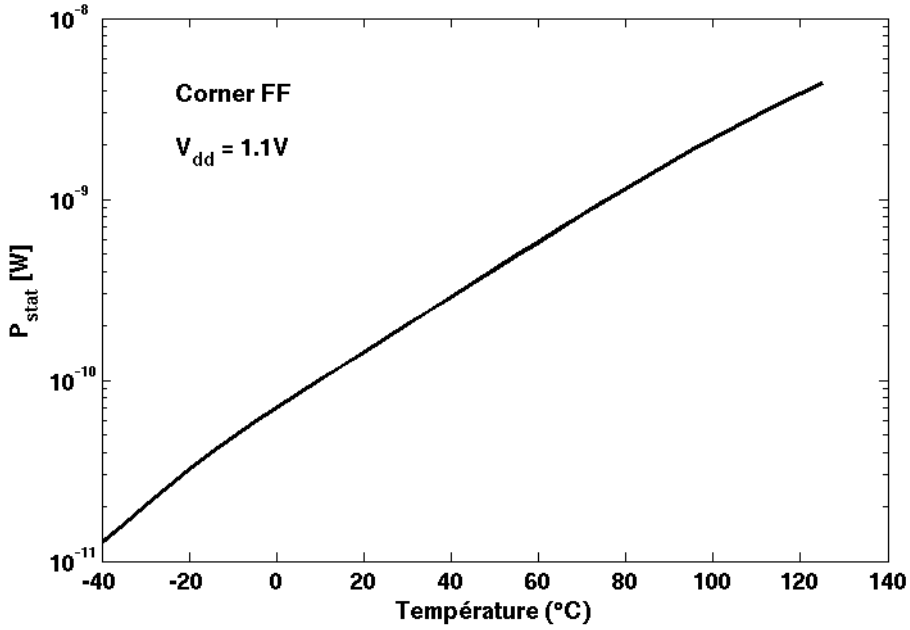


Figure 4.3 – Puissance statique en fonction de la température.

Rapport I_{read}/I_{off}

La section 3.3.3 avait montré que le *corner* à considérer pour un rapport minimal est le *corner* FS. La figure 4.4 représente le rapport I_{read}/I_{off} en fonction de la tension alimentation et de la température. Comme attendu, le rapport chute pour les hautes températures. Il appert également que la chute d'alimentation ne varie pas de manière significative le rapport I_{read}/I_{off}.

En toute rigueur, le courant I_{read} est le courant minimal que peut produire une cellule, et le courant I_{off} la moyenne du courant de fuite que pourraient produire 256 cellules, tous deux déterminés par simulation Monte-Carlo. Pour une tension d'alimentation d'1V à une température de 125°C dans le *corner* FS, la moyenne du courant I_{read} du *buffer* 2T HVt-HVt est de $33\mu A$, pour une déviation standard de $2.3\mu A$. Si nous considérons la valeur critique à 6σ, le courant I_{read} à considérer est de $19.6\mu A$. Le courant I_{off} extrait par simulation Monte-Carlo suit une loi statistique lognormale, et sa moyenne doit être calculée par une méthode statistique différente de celle employée pour une loi gaussienne [38]. La moyenne obtenue alors, pour une tension d'alimentation d'1V à une température de 125°C dans le *corner* FS, est de $0.63nA$. Le rapport I_{read}/I_{off} critique est alors supérieur à 30000, ce qui permet d'affirmer que notre tableau à 256 cellules par BL sera fonctionnel sur toute la plage de température demandée pour les applications grand public.

Remarquons qu'à partir d'un rapport I_{read}/I_{off} nominal d'environ dix millions, nous avons perdu près de trois ordres de grandeur pour le rapport critique à considérer. Or, dans la section

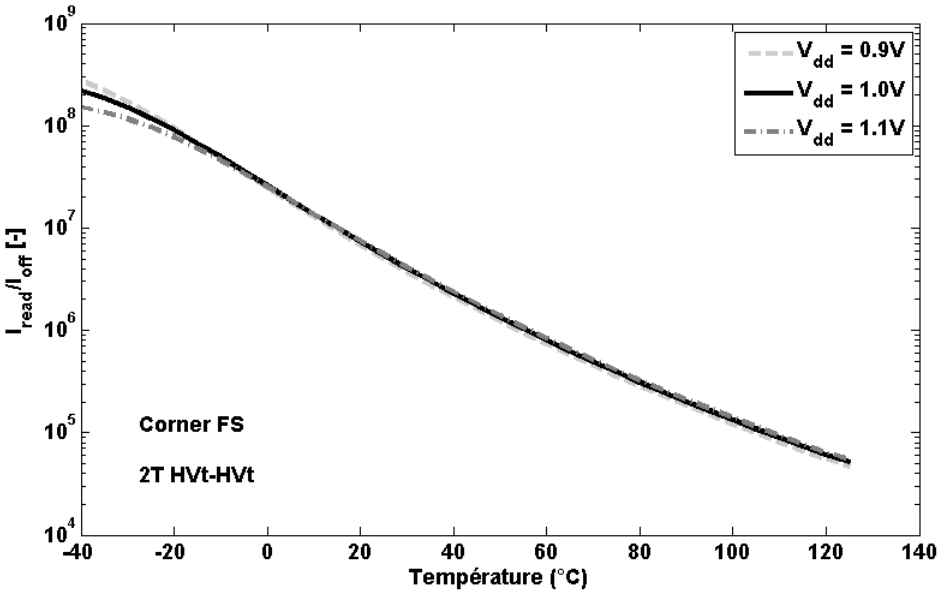


Figure 4.4 – Rapport I_{read}/I_{off} pour le *buffer* 2T HVt-HVT.

3.3.3, nous avions vu qu'un transistor LVt connecté à la BL permettait moins de dix mille cellules par BL, dans le cas typique. Il parait alors peu concevable d'utiliser un transistor LVt sans *assist* pour des applications demandant un fonctionnement à hautes températures.

La limite en vitesse

Pour tester la fonctionnalité de la cellule, des simulations MC sont effectuées dans le *corner* SS à 0.9V d'alimentation. La conclusion est que la cellule 5TPMOS accès SVt n'est pas robuste face à la variabilité. La diminution de la tension d'alimentation a entrainé une diminution de l'écart entre la tension de seuil du transistor d'accès et la tension de basculement du *latch*. Le même problème que la cellule 5TPMOS accès HVt se présente alors (section 3.2.7, figure 3.20). Le transistor d'accès passe en régime sous-seuil et le courant du transistor d'accès n'est plus suffisant que pour charger le noeud de rétention dans le temps imparti. Dans le *corner* SS, il faut un temps d'ouverture de 1.1ns pour avoir une écriture effective et robuste, quel que soit le temps de charge des BL et WL. La cellule n'est donc pas robuste pour une fréquence de 1GHz.

Il est intéressant de noter que l'écriture est effective et robuste face à la variabilité pour des hautes températures ! Pour le comprendre, il faut se rappeler que lors de l'écriture d'un 1 logique, le transistor d'accès voit sa tension grille-source V_{gs} progressivement diminuer. Pour faire basculer l'inverseur du latch, il doit charger le noeud de rétention Q à une tension plus haute que $V_{dd}/2$, car à la fin de l'écriture, le couplage capacitif avec la WL fera chuter cette tension. Or, dans cette configuration, le transistor d'accès est en inversion modérée. La figure 4.5 montre que le courant d'un transistor en inversion modérée, c'est-à-dire proche de sa tension de seuil, augmente avec la température. Donc, dans ce régime, une augmentation de température a l'effet d'une diminution de la tension de seuil. Le pire des cas à considérer pour les systèmes d'écriture à BL unique est l'écriture d'un 1 logique dans les basses températures. Pour les cellules différentielles, comme la 6T, le courant des transistors d'accès reste un courant I_{on}, qui diminue avec la température.

Le système d'écriture 5TPMOS accès SVt est en fait très peu robuste face aux variations de température. Le tableau 4.3 expose sa plage de validité en fonction de la chute de la tension d'alimentation. Il faudrait donc garantir une chute de tension minime et une température

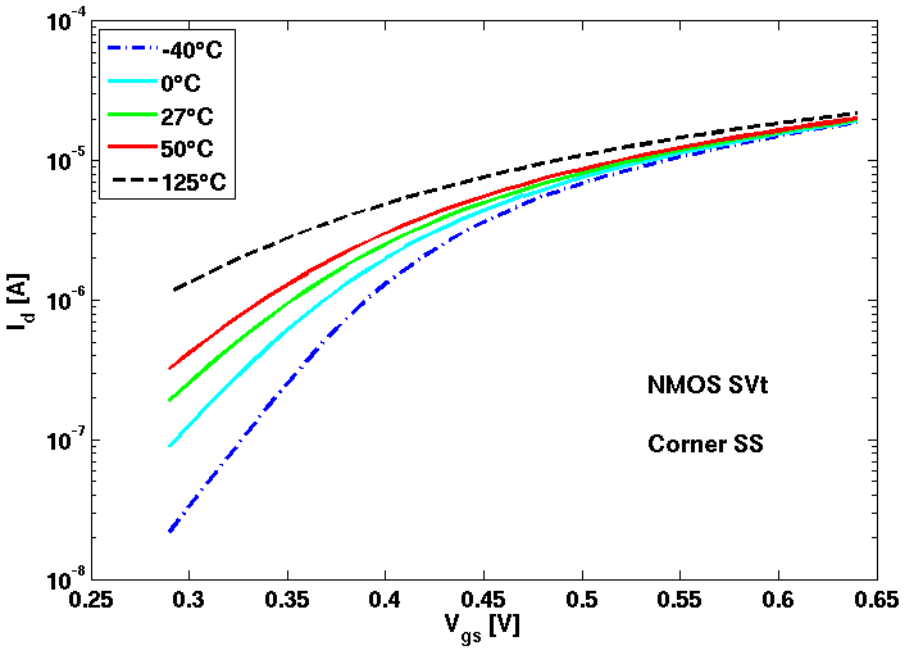


Figure 4.5 – Courbe $I_d(V_g)$ en fonction de la température (échelle logarithmique). Près du seuil, le courant augmente avec la température.

Table 4.3 – Plage de validité des cellules en fonction de la température et de la tension d'alimentation (*Corner* SS).

Température	5TPMOS accès SVt			5TPMOS accès LVt			6T accès HVt		
	0.9V	0.95V	1V	0.9V	0.95V	1V	0.9V	0.95V	1V
-40°C	X	X	X	X	V	V	V	V	V
<-20°C	X	X	**X**	**X**	V	V	V	V	V
<5°C	X	**X**	V	V	V	V	V	V	V
27°C	**X**	V	V	V	V	V	V	V	V
>50°C	V	V	V	V	V	V	V	V	V

minimale de fonctionnement plus élevée que les limites pour application grand-public. Une solution serait d'opter pour le système d'écriture 5TPMOS accès LVt à tension de WL négative en rétention. Cette dernière est robuste pour une chute de 10% d'alimentation, jusqu'à une température de -20°C. Enfin, si l'on souhaite un tableau mémoire robuste face à des variations de 10% de tension d'alimentation jusqu'à une température minimale de -40°C, il faudrait alors passer à la cellule 6T accès HVt.

Le problème majeur des cellules 5TPMOS est que la tension de basculement du *latch* est trop proche de la tension de seuil effective du transistor, dans des conditions particulières. Néanmoins, ce type de cellule est bien adapté aux hautes températures.

La contrainte sur la vitesse de lecture est largement respectée pour le *buffer* 2T HVt-HVt. Ce dernier avait déjà une marge assez conséquente. Le cas le plus lent est obtenu dans le *corner* SS, à 0.9V d'alimentation et, fait *a priori* contre-intuitif, pour les basses températures ! En effet, c'est à -40°C que le temps d'accès en lecture est le plus grand, et vaut 367ps à 6σ. Pendant une lecture, le transistor connecté à la WL voit sa tension V_{gs} progressivement augmenter. Avant

que cette dernière atteigne sa valeur maximale, le transistor était déjà devenu passant et avait commencé à décharger la BL. Nous venons de voir que le courant de drain pour une tension V_{gs} légèrement supérieure au seuil du transistor, diminuait avec la température. On en conclut que le temps d'accès en lecture dépend fortement du régime d'inversion modérée du transistor. Cette propriété avait d'ailleurs déjà été discutée dans la section 3.3.2 pour justifier la variabilité accrue du *buffer* 2T différentielle. Il semble que ce résultat confirme cette conjecture.

4.3 Résumé

Pour garantir la contrainte de stabilité, le *latch* composé de deux inverseurs montés en tête-bêche doit être testé par des simulations Monte-Carlo dans le *corner* FS, à tension d'alimentation minimale (-10%), pour les hautes températures. Le latch 5TPMOS résiste largement à la contrainte 6σ.

Le plus faible rapport I_{read}/I_{off} est atteint dans le *corner* FS, pour une diminution (-10%) de la tension d'alimentation dans les hautes températures. Il permet de maintenir quelques dizaines de milliers de cellules par BL pour un transistor d'accès en lecture HVt.

La puissance statique présente une diminution et une augmentation de plus d'un ordre de grandeur, respectivement à -40°C et 125°C, par rapport à la température nominale de 27°C. La cellule dite 5TPMOS accès SVt et *buffer* 2T HVt-HVt présente une consommation statique de $4nW$ dans le *corner* FF, pour une alimentation de 1.1V à une température de 125°C.

Pour garantir la contrainte de vitesse en écriture, la cellule SRAM doit être simulée dans le *corner* SS pour une diminution de la tension d'alimentation de -10%. Pour les cellules à une seule BL, ce sont les basses températures qui entrainent un comportement critique. La cellule 5TPMOS accès SVt n'est fiable que pour une chute de tension de 5% sur toute la plage de température. La cellule 5TPMOS accès LVt avec V_{WL} négative est fiable pour une chute de 10% mais jusque -20°C uniquement. La cellule 6T accès HVt est fiable dans tous les cas.
La contrainte de vitesse en lecture est largement satisfaite pour le tableau mémoire considéré. C'est encore une fois les basses températures qui entrainent un comportement critique.

En conclusion, le tableau 4.4 résume quel choix de cellule est à privilégier, afin de minimiser la surface et la consommation, en fonction des contraintes de dimensionnement.

Table 4.4 – Résumé des choix de dimensionnement en fonction des contraintes de dimensionnement.

SRAM *dual-port* à 1GHz	Architecture optimale
$V_{alim,min} = 0.95V$, $T_{min} = 5$°C	5TPMOS accès SVt avec 2T HVt-HVt
$V_{alim,min} = 0.9V$, $T_{min} = -20$°C, avec WL nég.	5TPMOS accès LVt avec 2T HVt-HVt
$V_{alim,min} = 0.9V$, $T_{min} = -40$°C	6T accès HVt avec 2T HVt-HVt

Chapitre 5

Perspectives

Dans ce chapitre, nous présentons des idées de conception visant à améliorer les performances des cellules SRAM *dual-port*. Une étude sommaire sera faite, invitant le lecteur à une analyse plus approfondie.

Tout au long du travail, nous avons considéré que le tableau mémoire était en mode actif, avec une écriture et une lecture à chaque cycle d'horloge. Toutefois, dans de nombreuses applications, la mémoire rentre en mode veille pendant un temps relativement long par rapport à sa fréquence de fonctionnement. La première section présente alors un moyen de réduire la consommation statique, applicable préférentiellement en mode veille.

Nous avons vu au chapitre précédent que le système d'écriture de la cellule proposée (5TPMOS accès SVt et *buffer* 2T HVt-HVt) n'était pas robuste face aux variations PVT pour la contrainte en vitesse. Pour pallier ce problème, la deuxième section présente différents moyens d'augmenter la fréquence de fonctionnement des cellules. L'un d'entre eux, la diminution de la tension d'alimentaion des cellules, présente également un avantage pour la consommation et sera étudié plus en profondeur dans la troisième section.

5.1 Réduire la consommation en mode veille

A l'instar de la WL, la tension de BL pourrait être également différente en rétention, toujours dans le but de diminuer la consommation. La finalité des BL d'écriture et de lecture étant très différente, la discussion doit être séparée entre les deux BL.

BL de lecture

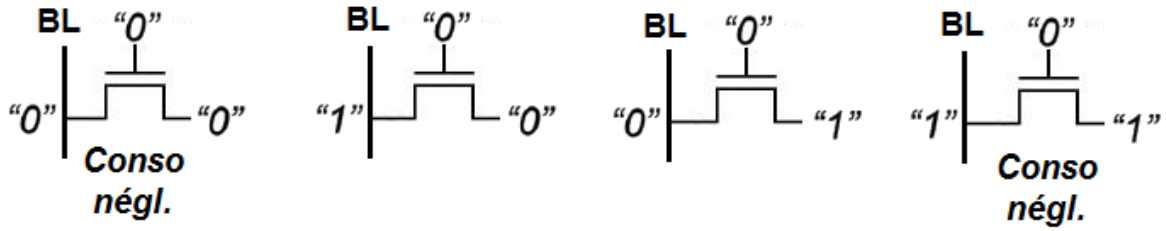
Quand aucune opération de lecture n'est effectuée, le transistor PMOS maintient la BL à la tension d'alimentation. *Ipso facto*, un courant de fuite se crée entre la masse de chaque cellule et la BL. Le tableau 2.2 montre la valeur du courant I_{off} pour un transistor d'accès. Dans un *buffer* de type 2T, le courant total traversant le transistor PMOS est donc 256 fois plus grand. On remarque alors que si le transistor PMOS était coupé en mode veille, son courant de fuite serait de plusieurs ordres de grandeur inférieur à la somme des courants de fuite des cellules. Et donc la consommation moyenne par cellule diminuerait drastiquement. Pour les *buffers* 1T, il est nécessaire de couper les PMOS de précharge des BL, ainsi que le PMOS du pseudo-inverseur connecté à la WL.

Néanmoins, le courant de transistor PMOS étant beaucoup moins important, la tension de BL va progressivement diminuer pour atteindre la tension de masse des cellules. Lors de la mise en mode actif, il faudra donc une certaine énergie et un certain temps pour recharger la capacité de BL. Il est donc nécessaire de

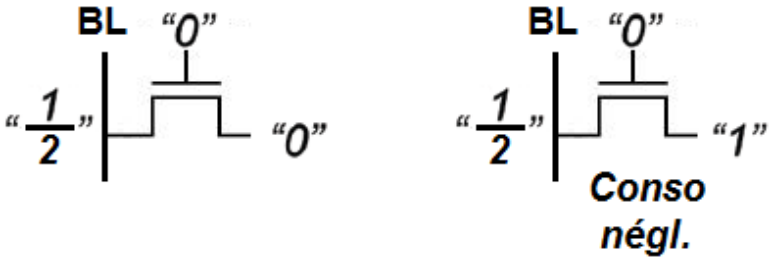
- dimensionner le transistor PMOS pour qu'il puisse recharger la BL dans le délai imparti,
- vérifier que le temps de mise en veille soit suffisant pour avoir un gain énergétique, connaissant le courant de fuite après dimensionnement.

BL d'écriture

Si la tension de BL est à 0 ou 1V, une partie des transistors d'accès aura une tension $V_{ds} = 0V$ donc une consommation négligeable, et l'autre partie $V_{ds} = V_{dd}$ (figure 5.1(a)). En fixant cette tension à $V_{dd}/2$, la tension V_{ds} sera toujours de $V_{dd}/2$, mais une partie des transistors aura une tension V_{gs} négative (figure 5.1(b)). La consommation statique sera donc nécessairement réduite.



(a) Sans *assist*.



(b) Avec *assist*.

Figure 5.1 – Différentes configurations possibles lors de la rétention. Maintenir la tension de BL à $V_{dd}/2$ en veille permet de sauvegarder de l'énergie.

De plus, une polarisation différente de la tension de BL ne modifie pas, en moyenne, la consommation dynamique. La BL tout entière sera chargée pour l'écriture d'un 1 logique, et déchargée pour un 0. Donc, en revenant à chaque fois à $V_{BL} = V_{dd}/2$ à la fin d'une écriture, la consommation dynamique moyenne devrait rester identique. Néanmoins, si la même donnée est écrite plusieurs fois d'affilée, comme dans le traitement du signal par exemple [4], la consommation dynamique risque d'augmenter car la BL reviendrait à chaque fois à $V_{dd}/2$ alors qu'elle aurait pu rester à sa tension précédente.
A la fréquence de fonctionnement cible de ce travail (1GHz), on peut considérer qu'une écriture est effectuée presqu'à chaque cycle d'horloge. Or, la BL est polarisée à chaque écriture car elle est partagée par toutes les cellules de la colonne. De plus, pour réaliser cet *assist*, il sera nécessaire de sélectionner une tension selon le mode de fonctionnement, et des transistors supplémentaires devront alors se trouver entre la tension d'alimentation et la BL, ce qui augmentera le temps d'accès en écriture.

Cet *assist* ne sera donc utilisé que très rarement et entrainera une augmentation du temps de charge de la BL. Il n'a donc pas été traité dans ce travail. Cependant, si la mémoire devait rentrer en mode veille, cette polarisation permettrait de limiter la consommation statique dans ce mode.

5.2 Augmenter la fréquence de fonctionnement

Plusieurs fois au cours de ce travail, l'utilisation d'une polarisation différente de la tension d'alimentation de la cellule a été proposée pour améliorer soit la vitesse d'écriture, soit la marge de bruit en écriture. La section 5.3 y reviendra et, au vu de ses conclusions, il nous semble préférable d'utiliser en premier lieu cet *assist*, si l'on cherche à augmenter la fréquence de fonctionnement du tableau de cellules *dual-port*. Si toutefois, cet *assist* n'est pas applicable dans une application donnée, analysons les possibilités restantes pour augmenter la fréquence de fonctionnement.

Les sections 3.2.7 et 4.2 ont montré que les cellules 5TPMOS accès HVt et SVt nécessitent un temps minimal d'activation de la WL indépendant des temps de charge de BL et WL, lorsque l'on considère le pire des cas. Si ce temps est trop grand pour l'application, il est nécessaire de changer de cellule tout en tentant de minimiser la consommation

La cellule 5TPMOS accès LVt présente la faible consommation dans les systèmes d'écriture restants (section 3.2.7). Néanmoins, elle demande un *assist* et une attention particulière pour le *layout*, car les transistors FDSOI NMOS LVt ont leur tension de substrat arrière polarisée à la tension d'alimentation (section 2.1).

La cellule 6T accès HVt est largement plus rapide que les cellules 5TPMOS. De plus, son temps d'écriture dépend directement des temps de charge de BL et WL qui peuvent être ajustés selon l'application. Elle présente malheureusement une énergie dynamique double, mais deviendrait très intéressante si le circuit passe régulièrement en mode veille.

Le tableau 3.2 montre que le temps de précharge de la BL représente moins de 30% du temps total d'écriture, et nous avons appris dans la section 3.3.1 que le temps de charge de la WL est le paramètre principal pour le temps d'accès en lecture. Il est alors plus bénéfique de commencer par diminuer le temps de charge de la WL pour atteindre plus rapidement la fréquence souhaitée. Pour diminuer la capacité de WL, il faudrait diminuer le nombre de cellules par ligne, c'est-à-dire la taille des mots du tableau mémoire. Or, celle-ci est très souvent fixée par l'application. Le paramètre aisément modifiable pour diminuer le temps de charge de la WL est la taille des *drivers*. L'augmenter conduirait *de facto* à une augmentation de la consommation dynamique et statique, ainsi que de la surface du circuit. On peut aussi imaginer de "découper" la WL en plusieurs morceaux. Tous les morceaux sont accédés en même temps, mais chacun possède son propre *driver*. Chacun de ces petits *drivers* voit une capacité plus petite à charger et le temps de charge s'en trouve donc diminuer. Il convient néanmoins de vérifier quelle solution apporte un minimum de consommation et de surface : avoir un gros *driver* qui respecte la contrainte de temps, ou plusieurs petits *drivers* ?

On peut également diminuer le temps de précharge de la BL en augmentant la taille des *drivers*. A nouveau, cela conduirait *statim* à une augmentation de la consommation dynamique et statique, ainsi que de la surface du circuit. On peut aussi diminuer le nombre de cellules par colonne et ainsi la capacité de BL. La quantité de cellules étant fixée par l'application, il sera alors nécessaire d'ajouter d'autres BL et donc d'autres *drivers*. En fonction de l'application, il convient là encore de vérifier quelle solution apporte un minimum de consommation et de surface : avoir peu de gros *drivers*, ou beaucoup de petits *drivers* ?

5.3 La tension d'alimentation de la cellule

Les lignes d'alimentation des cellules du tableau mémoire sont physiquement séparées de celles des WL et BL. Il est donc plus facile d'y appliquer une tension inférieure à 1V. Ce nouvel *assist* statique présente un grand nombre d'avantages :

1. Augmenter la marge de bruit en écriture. Une diminution de la tension d'alimentation des inverseurs du *latch* entraine nécessairement une diminution de la tension de basculement de ceux-ci. La tension de BL restant à 1V, la tension milieu, apparaissant durant la phase de court-circuit, pourrait être suffisante pour permettre une écriture effective et robuste d'un 1 logique sur la cellule (section 3.2.4).
2. Diminuer le temps d'accès en écriture. Pour les systèmes d'écriture à une seule BL, le transistor d'accès voit sa tension V_{gs} diminuer progressivement lors de l'écriture d'un 1 logique. En diminuant la tension d'alimentation de la cellule, la valeur de la tension de basculement du *latch* est également diminuée. Or, c'est précisement le problème rencontré pour la cellule 5TPMOS accès HVt et SVt (section 3.2.7 et 4.2 respectivement). Dans la section 3.2.7, nous avons aussi vu que le noeud de rétention de la cellule ULP prenait un temps considérable pour atteindre la tension d'alimentation de la cellule après l'écriture d'un 1 logique. Diminuer cette tension raccourcirait ce temps, qui deviendrait pratiquement nul si $V_{dd,Cell} < 1 - V_{t,acces}$.
3. Diminuer la consommation statique. Nous avons vu dans la section 3.2.7 que la consommation statique dépendait principalement du courant de fuite des transistors d'accès, quand ceux-ci étaient de type SVt et LVt. Or, diminuer la tension des noeuds de rétention diminuerait au mieux leur tension V_{gs}, au pire leur tension V_{ds}, et leur courant de fuite dans tous les cas. Pour les cellules à transistors d'accès HVt, la consommation statique dépendait majoritairement des courants de fuite de grille. Or, comme le montre la figure 5.2, le courant de grille diminue exponentiellement avec la tension de grille, que ce soit en mode *forward* ou *reverse*.

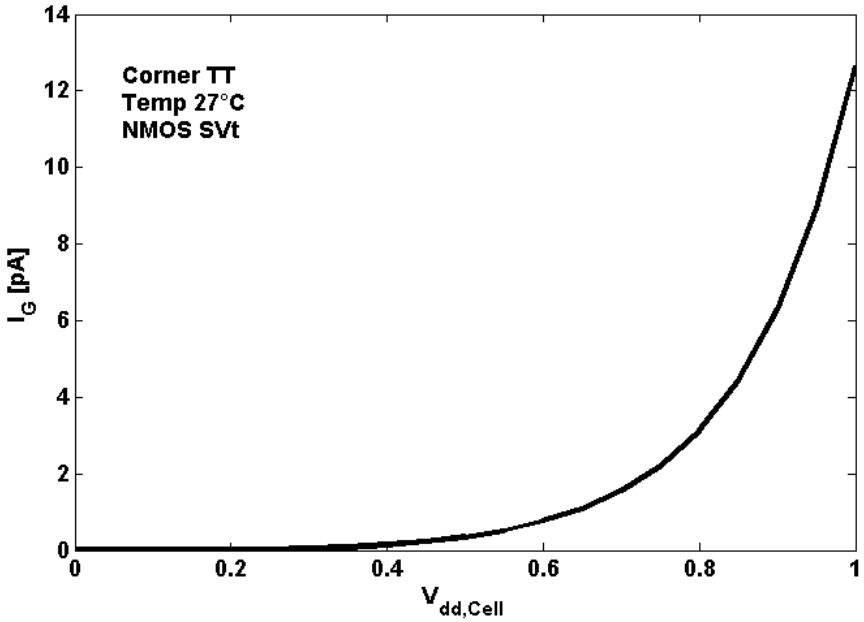


Figure 5.2 – Le courant de fuite de grille diminue exponentiellement avec la tension d'alimentation. Le courant $I_{g,reverse}$, beaucoup plus faible, suit la même dépendance.

4. Diminuer la consommation dynamique. Trivialement, les capacités des noeuds de rétention contiendront moins d'énergie. Le gain serait encore plus important dans le cas du *buffer* de type 1T. En effet, comme vu dans la section 2.4.3, les courants de court-circuit, provoqués par les transistors passants des cellules non accédées, augmentaient fortement la consommation dynamique. Si la tension de la BL reste inférieure à la tension d'alimentation de la cellule, tous les transistors des cellules non accédées auront nécessairement leur tension V_{gs} négative, qu'ils soient passants ou non.

Néanmoins, cet *assist* présente des désavantages au niveau de la marge de bruit en rétention, de la marge de bruit dynamique et de la vitesse de lecture.

La marge de bruit en rétention

Diminuer $V_{dd,Cell}$ diminue la marge de bruit en rétention des *latchs* [32]. La figure 5.3 montre la stabilité de deux inverseurs montés en tête-bêche en fonction de leur tension d'alimentation. Comme on peut le voir, grâce à la fois au dimensionnement (transistors HVt de taille minimale) et à la technologie (FDSOI) ce *latch* reste robuste à 6σ jusqu'à une tension d'alimentation de 0.5V.

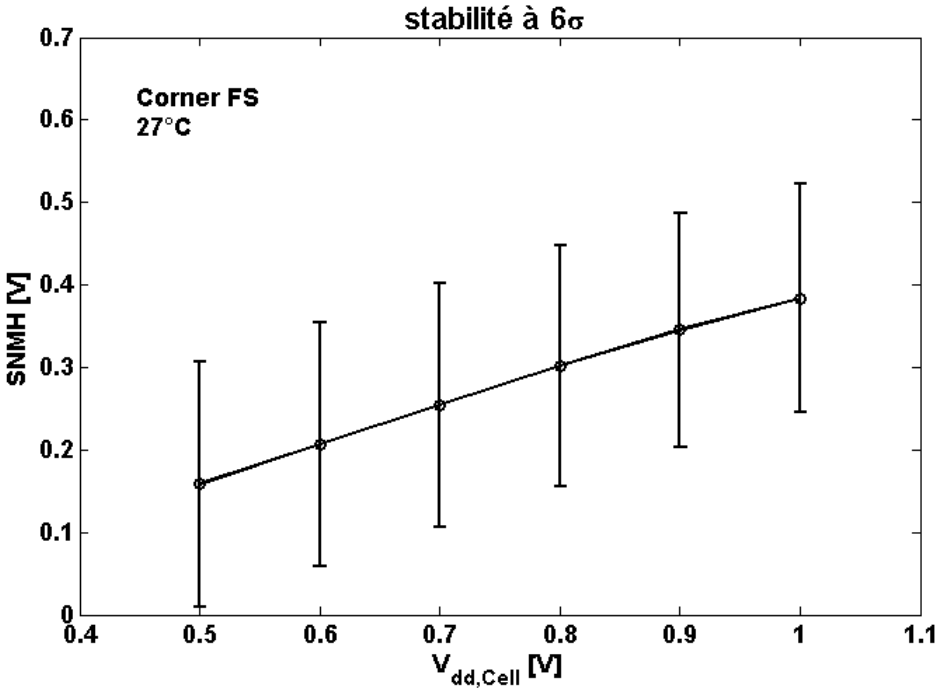


Figure 5.3 – Marge de bruit en rétention en fonction de la tension d'alimentation. Les deux inverseurs montés en tête-bêche, dont les quatre transistors sont de taille minimale et de type HVt, restent robustes à 6σ jusqu'à $0.5V$.

La cellule ULP a demandé un effort particulier pour obtenir une marge de bruit en rétention suffisante dans tous les corners. La figure 5.4 compare les courants intervenant dans la stabilité du *latch* ULP (section 3.2.6) en fonction de la tension d'alimentation de la cellule. La figure ne montre que les valeurs nominales de ces courants, or la section 3.2.6 a montré que la variabilité doit être prise en compte. Néanmoins, plusieurs observations qualitatives peuvent être faites.

Il apparait très clairement que le rapport entre le courant I_{Peak} et les courants de fuite, I_{min} et $I_{g,Buffer}$, grandit avec la diminution de $V_{dd,Cell}$. Ce qui pourrait amener à une relaxation des contraintes et, en suivant la même démarche que la section 3.2.6, une cellule demandant

moins de surface de silicium pourrait être obtenue. En particulier, le courant de fuite de grille diminue fortement avec $V_{dd,Cell}$, car il dépend exponentiellement de la tension de grille (voir précédemment dans la section). Ceci pourrait permettre l'utilisation d'un *buffer* de lecture en mode *forward* pour cette cellule, ce qui n'était pas le cas précédemment (voir section 3.2.6). Et l'utilisation d'un *buffer* en mode *reverse* pourrait être possible avec une diode du bas contenant des transistors de taille minimale.

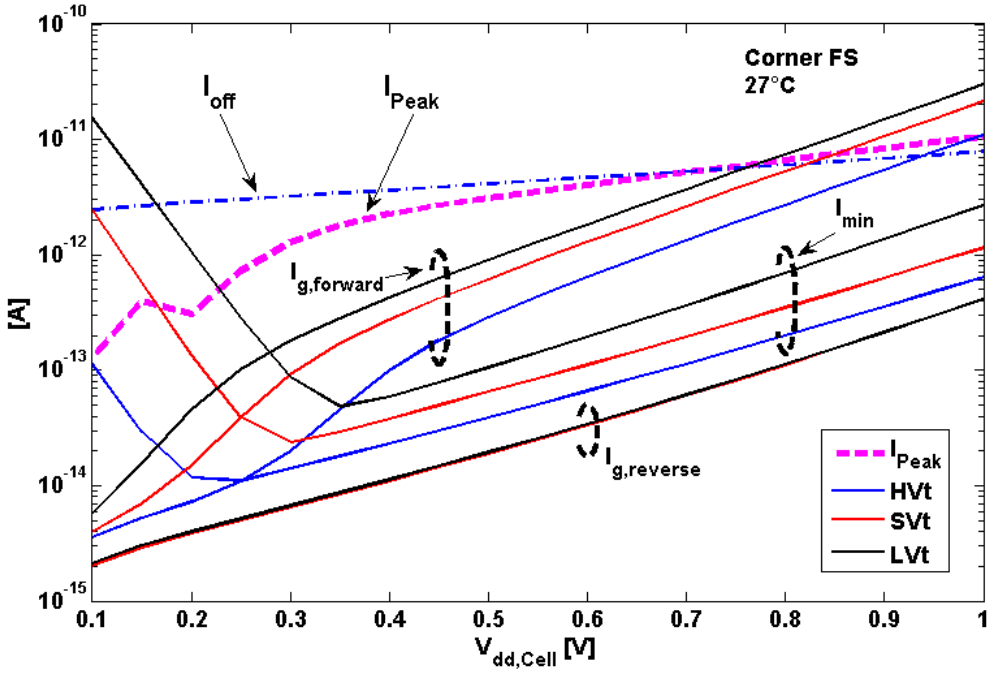


Figure 5.4 – Comparaison entre les valeurs nominales des courants de fuite et du courant I_{Peak} du *latch* ULP, pour différentes tensions d'alimentation. Les valeurs données ici sont les valeurs nominales, sans tenir compte de la variabilité. Le courant I_{Peak} de la diode du bas vaut $1.4pA$, et n'a pas été représenté par souci de clarté.

Néanmoins, on constate que nous ne pouvons pas descendre en dessous de 0.3-0.4V d'alimentation, car les courants minimaux de drain deviennent plus grands que les courants I_{Peak}. On remarque également que l'utilisation d'une tension de WL négative en rétention est indispensable. Le courant sous-seuil du transistor HVt devient plus grand que le courant maximal des diodes ULP, I_{Peak}, quand la tension de seuil diminue. Or, il fallait une taille considérable aux transistors du *latch* pour avoir une cellule fonctionnelle à $V_{dd} = 1V$, en tenant compte de la variabilité (section 3.2.6).

Si l'on souhaite diminuer la tension d'alimentation en dessous de 0.3V, [12] montre que la cellule à trigger de Schmitt simplifié que nous avons brièvement rencontrée dans les sections 1.2.2 et 3.2.2, reste fiable en rétention jusqu'à une tension d'alimentation de $160mV$. Néanmoins, elle nécessite 12 surfaces minimales de silicium [12], ce qui n'est pas optimal pour une application à haute densité. De plus, aucun de ces transistors n'a une tension V_{gs} négative, et il n'est dès lors pas dit que la consommation à 0.16V d'alimentation soit plus faible que celle d'une cellule ULP à 0.3V.

La marge de bruit dynamique

A 1V d'alimentation, quand le transistor d'accès se coupe, le *latch* régénère très rapidement les noeuds de rétention. En diminuant la tension d'alimentation, on diminue le courant pouvant des transistors du latch et donc, la rapidité de sa contre-réaction. Si la tension d'alimentation reste au-dessus de la tension de seuil des transistors HVt ($V_{t,H} \cong 0.5V$), la contre-réaction du *latch* est toujours suffisante pour régénérer les noeuds dans les quelques picosecondes qui suivent la fin de l'écriture.

Néanmoins, si la tension d'alimentation diminue encore, le *latch* prend de plus en plus de temps à ramener les noeuds de rétention à leur valeur nominale. La tension se trouvant alors sur la grille du *buffer* pourrait être plus basse que prévue, et augmenter le temps d'accès maximale en lecture.

Lors de la lecture, l'utilisation d'un *buffer* permettait un découplage et une immunité totale face aux couplages capacitifs induits par la WL ou la BL de lecture. Néanmoins, quand la tension d'alimentation tombe en dessous de la tension de seuil des transistors, un pic de tension commence à se faire ressentir sur les noeuds de rétention. Le temps d'accès en lecture ne sera pas négativement affecté, car la tension sera plus haute que prévue lors de la charge de la WL. Mais, la variation non désirée des noeuds de rétention pourrait provoquer un basculement de la donnée du *latch*.

Finalement, les basses températures sont un problème important pour les performances dynamiques des circuits à très faible tension d'alimentation [43].

Le temps d'accès en lecture

Diminuer la tension d'alimentation de la cellule diminue *de facto* la tension de grille du transistor du *buffer* de lecture. Le temps d'accès en lecture est dominé par le temps de charge de la WL et le courant de décharge de la BL, I_{read} (section 2.4.2). Or, le courant de décharge est fortement dépendant de la tension de grille des deux transistors du chemin de décharge.

En diminuant la tension d'*overdrive* du chemin de lecture, on sortirait du cadre de la comparaison d'une vitesse de lecture à 1V, qui est la base de ce travail. De plus, compte tenu des résultats de la section 2.2, les cellules de l'état de l'art ne peuvent être comparées dans ces conditions, car elles présenteraient un SNMR beaucoup trop petit et ne seraient plus robustes face aux variations PVT (encore un avantage de l'utilisation d'un *buffer* de lecture).

Néanmoins, nous avons vu que le temps d'accès en lecture était largement inférieur au temps d'accès en écriture, pour les systèmes d'écriture à une seule BL. Diminuer la tension d'alimentation pourrait alors faire converger ces temps d'accès vers une même valeur, comme l'illustre schématiquement la figure 5.5. Le tableau 5.1 illustre, sur la cellule 5TPMOS accès SVt et *buffer* 2T HVt-HVt, le gain en performance qu'apporte ce nouvel *assist*.

Table 5.1 – Résultats de simulations. Corner TT, $V_{dd} = 1V$, Température 27°C.

Cellule 5TPMOS - 2T	T Write [ps]	T Read [ps]	E Write [fJ]	E read [fJ]	P stat [pW]	$\frac{I_{read}}{I_{off}}$	$\frac{SNMH}{V_{dd,Cell}}$	$\frac{SNMW}{V_{dd,Cell}}$
$V_{dd,Cell} = 1.0V$	488	218	7.3	0.7	**46**	12.3 10^6	0.41	-0.43
$V_{dd,Cell} = 0.8V$	468	348	7.0	0.7	**28**	9.3 10^6	0.46	-0.49

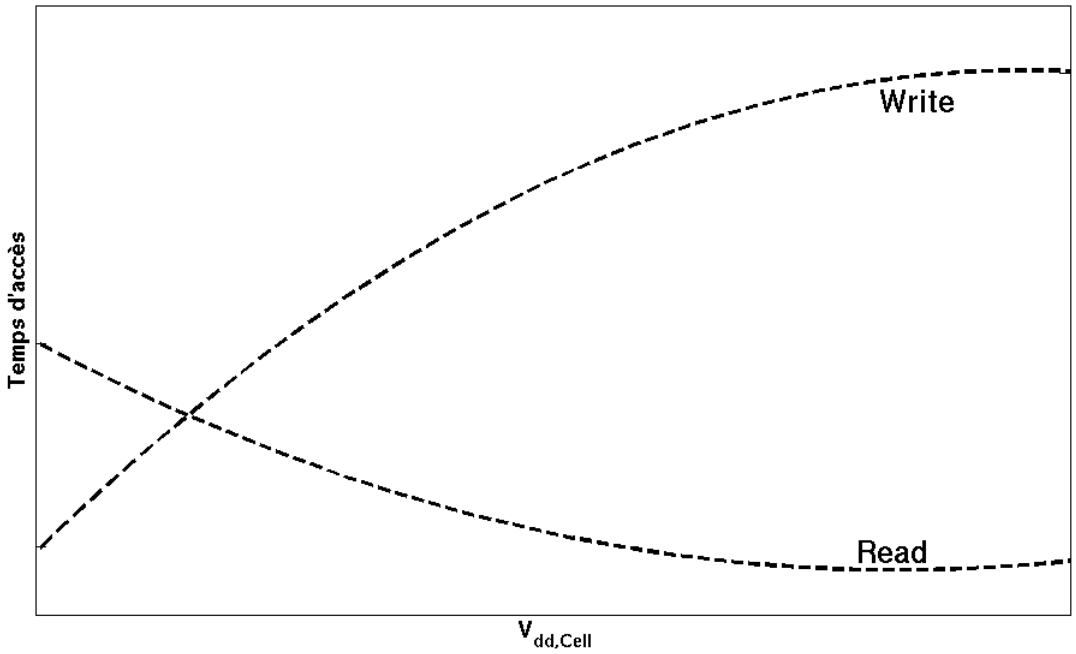


Figure 5.5 – Comportement schématique des temps d'écriture et de lecture en fonction de la tension d'alimentation de la cellule.

Cet exemple motive une démarche différente de celle suivie dans ce travail pour la conception d'une cellule *dual-port* à tension d'alimentation réglable : il paraît judicieux de fixer la tension d'alimentation de la cellule minimale qui permet au système de lecture de respecter la contrainte de temps (et/ou de nombre de cellules par BL). Cette tension fixée, il est fort probable que le système d'écriture sera beaucoup plus facile à dimensionner et verra ses performances s'améliorer.

Plus précisément, il convient de

1. Lister les *buffers* de lecture dans l'ordre croissant de consommation. Le 1T différentiel pourrait devenir le plus efficace, car ses courants de courts-circuits seraient drastiquement diminués si les transistors des cellules non-accédées restent avec une tension V_{gs} négative.
2. Déterminer quelle valeur minimale de tension d'alimentation permet une lecture dans la période cible de fonctionnement. Pour cela, nous avons vu qu'il fallait tenir compte de la variabilité de fabrication (*corner* SS), mais aussi des variations de température (-40°C) et de tension d'alimention ($-10\%V_{dd}$).
3. Lister les systèmes d'écriture dans l'ordre croissant de consommation. Pour ce faire, utiliser la formule 3.2 de la section 3.2.7. Il est possible qu'avec l'amélioration de la marge de bruit en écriture, **la cellule 5T consomme moins** que toutes les autres cellules, ULP exclues.
4. Vérifier les marges de bruit en rétention du *latch* avec la nouvelle tension $V_{dd,Cell}$, *ainsi que le comportement dynamique* de la cellule. Notons que ce schéma est nécessairement fini car nous avons vu que pour $V_{dd,Cell} = 1V$ il est possible d'avoir au moins une cellule robuste face aux variations PVT.
5. Vérifier si un *buffer* plus rapide n'apporterait pas une tension d'alimentation plus faible et un gain de consommation sur l'ensemble de la cellule. En effet, il est probable que les *buffers* qui consomment le moins demandent une plus grande tension d'alimentation pour maintenir la contrainte de vitesse. Or, la tension d'alimentation influence également la consommation du système d'écriture. Le gain sur ce dernier pourrait compenser l'augmentation de consommation due au système de lecture.

Conclusion

Ce travail a étudié et comparé plusieurs systèmes d'écriture et de lecture de cellules SRAM *dual-port* en technologie 32nm FDSOI, destinées à fonctionner à une fréquence de 1GHz sous une tension de 1V.

Dans le premier chapitre, nous avons présenté l'état de l'art dans le domaine *dual-port*, ainsi que des cellules dédiées à une application à accès unique. La connaissance des cellules classiques a permis de sélectionner celles qui pouvaient être utilisées dans une application *dual-port*. L'état de l'art *dual-port* étant très succinct, une comparaison avec les autres cellules fut motivée comme sujet de recherche pour ce travail. Dans la section 1.4, nous avons présenté des systèmes créant une polarisation particulière dans les mémoires SRAM, les *assists*. Nous avons alors motivé l'utilisation de ces systèmes pour améliorer la performance des cellules.

Dans le deuxième chapitre, nous avons défini un circuit de test réaliste afin de pouvoir comparer les cellules entre elles. En particulier, la valeur des capacités de routage a été calculée grâce à des paramètres provenant d'un *layout* de l'industrie, et les dimensions des *drivers* de charge des *BitLines* et *WordLines* ont été fixées afin d'atteindre un optimum énergétique.

Dans le troisième chapitre, nous avons commencé par une brève discussion motivant l'utilisation d'un système de lecture permettant de le découpler du système de rétention : les *buffers*. Ce constat nous a conduits à opérer une étude des cellules *dual-port* en deux temps : d'abord les systèmes d'écriture et de rétention dans la section 3.2 et ensuite les *buffers* de lecture dans la section 3.3. A la fin de ce chapitre, nous étions alors en mesure de comparer une certaine combinaison avec la cellule *dual-port* qui présente un système de lecture couplé au système de rétention.

Dans la deuxième section, nous avons étudié un *assist* statique qui permet de réduire la consommation : une tension de WL négative en rétention. Nous avons montré que cet *assist* n'est efficace que pour des transistors d'accès de type LVt, et pour une valeur définie de la tension.

Grâce aux simulations dynamiques d'écriture, des comparaisons ont pu être faites. Les cellules 5T et ULP sont les plus rapides, car leur capacité de *WordLine* est deux fois plus petite pour un même nombre de cellules. Les cellules 6T sont plus rapides que les cellules 5TPMOS, sauf si celles-ci utilisent une tension de WL négative en rétention.
Il est apparu que certaines cellules consomment plus d'énergie dynamique mais moins de puissance statique que d'autres. Nous avons alors mis au point une méthodologie pour les comparer quantitativement. Nous en avons conclu que certaines cellules 5TPMOS consomment moins d'énergie que la cellule 6T à transistor d'accès HVt. Les simulations ont montré que les deux cellules qui présentent la plus basse consommation, ULP et 5TPMOS accès HVt, n'étaient pas capables de tenir la contrainte de vitesse. La cellule 5TPMOS accès SVt a donc été proposée comme système d'écriture pour l'application *dual-port*.

En étudiant les systèmes de lecture, nous avons d'abord établi une relation entre les performances de vitesse et de consommation statique, et le dimensionnement des transistors. La tension de seuil du transistor se trouvant dans le régime saturé lors de la décharge de la BL, est le paramètre principal influant le courant de décharge. La tension de seuil du transistor connecté à la WL est le paramètre principal influant la consommation statique. Dans les *buffers* 2T, un transistor remplit ces deux rôles, mais pas dans les *buffers* 1T. Il est donc possible d'optimiser à la fois le temps d'accès et la consommation statique. Le compromis entre vitesse et consommation est encore plus marqué pour le *buffer* dit inverseur *tri-state*.
La consommation dynamique a été étudiée selon une méthodologie particulière. L'inverseur *tri-state* a la plus grande consommation dynamique, qui apparaissait pourtant pour son seul avantage. Ce *buffer* a donc été abandonné pour l'étude. Les *buffers* à BL unique consomment en moyenne moins d'énergie par lecture que leur pendant à double BL. Le *buffer* 2T à BL unique consomme moins d'énergie par accès que le *buffer* 1T, à cause des courants de fuite des autres transistors connectés à la BL. Le *buffer* 1T différentiel consomme moins d'énergie par accès que le *buffer* 2T différentiel car la décharge de la BL est moins grande.
Le nombre maximal de cellules par BL dépend de la tension de seuil du transistor connecté à la BL, et peut varier de plusieurs ordres de grandeur en fonction de celle-ci. Le *buffer* dit 2T à deux transistors HVt de taille minimale présente la meilleure consommation (statique et dynamique combinées) tout en respectant la contrainte de vitesse et garantissant un nombre maximal de cellules par BL très acceptable.

A la fin de cette étude, il apparait que l'utilisation d'un *buffer* de lecture permet d'améliorer les performances des cellules SRAM *dual-port*, notamment celles de l'état de l'art.

Dans le quatrième chapitre, nous avons étudié les performances de la cellule complète sélectionnée, *latch* 5TPMOS avec un transistor d'accès en écriture SVt combiné avec un *buffer* de lecture 2T de type HVt-HVt, selon les variations de température et de tension d'alimentation. Il est apparu que le système d'écriture ne pouvait respecter la contrainte de temps pour une diminution de 10% de la tension d'alimentation. Le couplage capacitif avec la *WordLine* d'écriture est le principal frein des performances des cellules 5TPMOS. Fait contre-intuitif, les temps d'écriture et de lecture augmentent pour les basses températures. Ces temps augmentent car le courant est plus faible à basse température pour des transistors en régime proche du seuil. Ce régime est très critique pour les cellules à une seule BL d'écriture et dominant pour les *buffers* de lecture.

Dans le cinquième chapitre, nous avons proposé des changements par rapport au circuit de test afin d'augmenter la fréquence de fonctionnement ou de diminuer la consommation. L'une de ces pistes d'amélioration, la diminution de la tension d'alimentation de la cellule, pourrait permettre à la cellule sélectionnée - 5TPMOS accès SVt et *buffer* 2T HVt-HVt - de devenir robuste face aux variations PVT. Nous présentons un exemple pour illustrer le gain en performance de cet *assist*, suivi d'une nouvelle méthologie de conception de cellules SRAM *dual-port* sans *assist* dynamique.

Bibliographie

[1] International Technology Roadmap for Semiconductor, 2009, chap. “Process integration, devices and structures” & chap. “Emerging research devices”.

[2] M. E. Sinangil, H. Mair, A. P. Chandrakasan, “A 28nm high-density 6T SRAM with optimized peripheral-assist circuits for operation down to 0.6V”, ISSCC, Sess. 14, 14.4, p.260-261, 2011

[3] K. Nii, Y. Tsukamoto, et al., “Synchronous Ultra-High-Density 2RW Dual-Port 8T-SRAM With Circumvention of Simultaneous Common-Row-Access”, JSSC, VOL. 44, NO. 3, MARCH 2009.

[4] H. Noguchi, S. Okumura, Y. Iguchi, H. Fujiwara, Y. Morita, K. Nii, H. Kawaguchi, M. Yoshimoto, “Which is the best dual-port SRAM in 45-nm process technology ? - 8T, 10T single end, and 10T differential -”, IEEE, 2008.

[5] *http ://www.feb-patrimoine.com/nsdat/mediatheque/expos/mam_2008/Borne_MAM/07.html*. Voir aussi *http ://www.freepatentsonline.com/*.

[6] C.-C. Wu, “Reduced size dual-port SRAM cell”, U.S. Pat. Appl. Publ. (2008), US 7359275 B1.

[7] H. Bajwa, X. Chen, “Aera-efficient dual-port memory architecture for multi-core processors”, JSSC, 2007.

[8] R. Manapat, S. K. Koduru “DUAL PORT SRAM”. U.S. Pat. Appl. Publ. (2002), US 6388939 B1.

[9] N. Verma, A.P. Chandrakasan, “A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy”, in IEEE J. Solid-State Circuits, vol. 43, no. 1, pp. 141-149, Jan. 2008.

[10] I. Carlson, S. Andersson, S. Natarajan, A. Alvandpour, “A high density, low leakage, 5T SRAM for embedded caches”, IEEE, 2004.

[11] E. Sinangil, N. Verma, A.P. Chandrakasan, “A reconfigurable 8T ultra-dynamic voltage scalable (UDVS) SRAM in 65nm CMOS”, in IEEE J. Solid-State Circuits, vol. 44, no. 11, pp. 3163-3173, Nov. 2009.

[12] 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.

[13] C.-H. Jan,et al., “A 65nm ultra low power logic platform technology using uni-axial strained silicon transistors”, IEEE, 2005.

[14] A. Steegen , et al., “65nm CMOS technology for low power application”, IEEE, 2005.

[15] F. Arnaud, et al., “Low cost 65nm CMOS platform for low power and general purpose applications”, VLSI, 2004.

[16] F. Arnaud, et al., “32nm general purpose bulk CMOS technology for high performance applications at low voltage”, IEEE, 2009.

[17] X. Chen, et al., "A cost effective 32nm High-K/ metal gate CMOS technology for low power applications with single-metal/gate-first process", VLSI, 2008.

[18] B. Greene, et al., "High performance 32nm SOI CMOS with high-k/metal gate and 0.149 μm^2 SRAM and ultra low-k back end with eleven levels of copper", VLSI, 2009.

[19] C.-H. Jan, et al., "A 32nm SoC platform technology with 2^{nd} generation high-k/metal gate transistors optimized for ultra low power, high performance, and high density product applications", IEDM, 2009.

[20] S.-Y. Wu, et al., "A highly manufacturable 28nm CMOS low power platform technology with fully fonctional 64Mb SRAM unsing dual/tripe gate oxide process", VLSI, 2009.

[21] Y. Jacqmin, "Etude d'architecture de petites mémoires SRAM sans sens-amplifier", travail de fin d'étude, Université Catholique de Louvain, 2010.

[22] 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.

[23] D. Levacq, V. Dessard and D. Flandre, "Low leakage SOI CMOS static memory cell with ultra-low power diode", in IEEE J. Solid-State Circuits, vol. 42, no. 3, pp. 689-702, Mar. 2007.

[24] J. De Vos, "Développement de circuits mémoire à ultra-basse consommation pour applications portables en technologie bulk 130 nm", travail de fin d'étude, Université Catholique de Louvain, 2008.

[25] S. Lin, Y.-B. Kim, F. Lombardi , "Design and analysis of a 32 nm PVT tolerant CMOS SRAM cell for low leakage and high stability", IEEE Transactions on very large scale integration (VLSI) systems, vol. 42, no. , pp. 176-187 Jan 2010.

[26] 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.

[27] K. Nii, et al., "A 45nm single-port and dual-port SRAM family with robust read/write stabilizing circuitry under DVFS environment", IEEE, 2008.

[28] D.P. Wang, et al., "A 45nm dual-port SRAM with write and read capability enhancement at low voltage", IEEE, 2007.

[29] R. W. Mann, et al., "Impact of circuit assist methods on margin and performance in 6T SRAM", in IEEE J. Solid-State Electronics, vol. 54, pp. 1398-1407, 2010.

[30] M.-H. Tu, J.-Y. Lin, M.-C. Tsai, S.-J. Jou, "Single-ended subthreshold SRAM with asymmetrical write/read-assist", IEEE, 2010.

[31] M. Pelgrom, "Different faces of variability", IEEE distinguished lecture, ULC, 2010.

[32] D. Bol, S. Bernard, D. Flandre, "Pre-Silicon 22 nm Compact MOSFET Models for Bulk vs. FD SOI Low-Power Circuit Benchmarks", IEEE SOI Conference, 2011.

[33] V. P.-H. Hu, Y.-S. Wu, M.-L. Fan, P. Su, C.-T. Chuang, "Static noise margin of ultrathin-body SOI subthreshold SRAM cells - an assessment based on analytical solutions of poisson's equation", IEEE TOED, vol. 56, no. 9, 2009.

[34] D. Bol, R. Ambroise, D. Flandre, J-D. Legat, "Channel length upsize for robust and compact subthreshold SRAM", FTFC, 2008.

[35] Bickford, J.P. Rosner, R. Hedberg, E. Yoder, J.W. Barnett, T.S, "SRAM redundancy-silicon area versus number of repairs trade-off", IEEE/SEMI, 2008.

[36] C.Fenouillet-Beranger, et al., "Efficient Multi-VT FDSOI technology with UTBOX for low power circuit design", VLSI, 2010.

[37] D. Flandre, V. Bayot, “Dipositifs électroniques avancés”, notes de cours, Louvain-la-Neuve, 2009.

[38] D. Bol, “Pushing Ultra-Low-Power Digital Circuits into the Nanometer Era”, thèse de doctorat, Université Catholique de Louvain, 2008.

[39] A. Kawasumi, Y. Takeyama, O. Hirabayashi, K. Kushida, Y. Fujimura and T. Yabe, “A Low-Supply-Voltage-Operation SRAM With HCI Trimmed Sense Amplifiers”, IEEE JSSC, vol. 45, no. 11, nov. 2010.

[40] G. Chen, D. Sylvester, D. Blaauw, T. Mudge, “Yield-driven near-threshold SRAM Design”, in IEEE T. VLSI Systems, vol. 18, no. 11, pp. 1590-1597, Nov. 2010.

[41] W. Dong, P. Li, G. M. Huang, “SRAM dynamic stability : theory, variability and analysis”, ICCAD, 2008.

[42] K.-J. Zhang, K. Chen, W.-t. Pan, P.-j. Ma, “A research of dual-port SRAM cell using 8T”, IEEE, 2010.

[43] D. Bol, C. Hocquet, D. Flandre, J.-D. Legat, “The Detrimental Impact of Negative Celsius Temperature on Ultra-Low-Voltage CMOS Logic”, IEEE, 2010.

Annexe A

Résultats complets pour la lecture

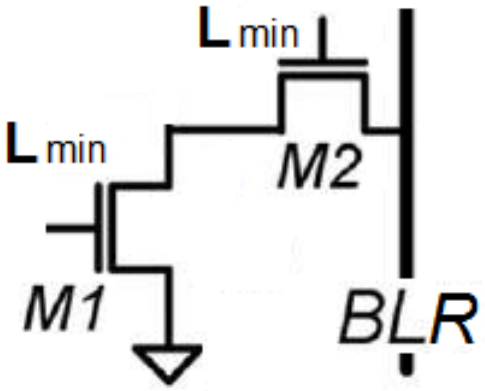


Figure A.1 – Légende de la première colonne des tableaux *ut infra*.

L'énergie dynamique due à la décharge de la BL est idéalement la même pour tous les systèmes de lecture. Comme la capacité et le temps d'activation de la WL varient, de légères variations peuvent apparaitre dans le cas idéal. Néanmoins, ces variations ne changent que la deuxième décimale.

Notons que dans le cas de transistors d'accès LVt, la puissance statique des 256 cellules non accédées intégrée sur 1ns ajoute environ 1fJ sur le résultat. Néanmoins, cette énergie est comprise dans le calcul de la comparaison de l'énergie totale (section 3.2.7).

Table A.1 – Résultats de simulation des performances des *buffers* de lecture 2T à BL unique. Corner TT. Vdd = 1V. Temp = 27°C.

$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	T Read [ps]	E Read [fJ]	P_{stat} [pW]	$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	T Read [ps]	E Read [fJ]	P_{stat} [pW]
1-1 LVt-LVt	169	0.7	2240	2-1 LVt-LVt	157	0.7	2260
1-1 SVt-LVt	175	0.7	2220	2-1 SVt-LVt	160	0.7	2230
1-1 HVt-LVt	178	0.7	2210	2-1 HVt-LVt	168	0.7	2210
1-1 LVt-SVt	203	0.7	50.9	2-1 LVt-SVt	185	0.7	60.9
1-1 SVt-SVt	204	0.7	38.9	2-1 SVt-SVt	187	0.7	46.5
1-1 HVt-SVt	210	0.7	30.1	2-1 HVt-SVt	187	0.7	32.9
1-1 LVt-HVt	211	0.7	11.4	2-1 LVt-HVt	198	0.7	18.6
1-1 SVt-HVt	217	0.7	8.7	2-1 SVt-HVt	198	0.7	13.8
1-1 HVt-HVt	218	0.7	4.9	2-1 HVt-HVt	205	0.7	7.6

Table A.2 – Résultats de simulation des performances des *buffers* de lecture 2T différentiel. Corner TT. Vdd = 1V. Temp = 27°C. Le temps d'accès et l'énergie dynamique sont accrus du fait de la capacité de WL double. La puissance statique est le double de celle du tableau précédent, de par la symétrie du système.

$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	T Read [ps]	E Read [fJ]	P_{stat} [pW]	$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	T Read [ps]	E Read [fJ]	P_{stat} [pW]
1-1 LVt-LVt	235	1.29	4480	2-1 LVt-LVt	218	1.29	4520
1-1 SVt-LVt	239	1.29	4440	2-1 SVt-LVt	239	1.29	4460
1-1 HVt-LVt	249	1.29	4420	2-1 HVt-LVt	229	1.29	4420
1-1 LVt-SVt	315	1.29	101.8	2-1 LVt-SVt	295	1.29	122
1-1 SVt-SVt	323	1.29	77.7	2-1 SVt-SVt	297	1.29	93
1-1 HVt-SVt	332	1.29	60.2	2-1 HVt-SVt	297	1.29	66
1-1 LVt-HVt	358	1.29	22.8	2-1 LVt-HVt	346	1.29	37.2
1-1 SVt-HVt	361	1.29	17.3	2-1 SVt-HVt	347	1.29	27.6
1-1 HVt-HVt	363	1.29	9.8	2-1 HVt-HVt	347	1.29	15.2

Table A.3 – Résultats de simulation des performances des *buffers* de lecture 1T à BL unique et différentiel. Corner TT. Vdd = 1V. Temp = 27°C. Grâce à l'architecture même, les caractéristiques ne varient pratiquement pas entre les deux modes de lecture.

$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	T Read [ps]	E Read [fJ]	P_{stat} [pW]	$\frac{W_{g,M1}}{W_{g,min}} - \frac{W_{g,M2}}{W_{g,min}}$ $Vt_{M1} - Vt_{M2}$	T Read [ps]	E Read [fJ]	P_{stat} [pW]
1-1 LVt-LVt	177	1.3	4370	diff LVt-LVt	177	1.3	4370
1-1 SVt-LVt	195	1.3	53.5	diff SVt-LVt	195	1.3	53.5
1-1 HVt-LVt	205	1.3	4.17	diff HVt-LVt	205	1.3	4.24
1-1 LVt-SVt	184	1.3	4370	diff LVt-SVt	184	1.3	4370
1-1 SVt-SVt	202	1.3	53.5	diff SVt-SVt	202	1.3	53.5
1-1 HVt-SVt	211	1.3	4.17	diff HVt-SVt	211	1.3	4.24
1-1 LVt-HVt	201	1.3	4370	diff LVt-HVt	201	1.3	4370
1-1 SVt-HVt	216	1.3	53.5	diff SVt-HVt	216	1.3	53.5
1-1 HVt-HVt	224	1.3	4.17	diff HVt-HVt	224	1.3	4.24

Table A.4 – Résultats de simulation des performances de l'inverseur *tri-state*. Corner TT. Vdd = 1V. Temp = 27°C. Les tensions de seuil de la première colonne correspondent à celles de l'inverseur et des transistors de passage, respectivement.

$Vt_{M1} - Vt_{M2}$	T Read [ps]	E Read [fJ]	P_{stat} [pW]
LVt-LVt	475	11.7	4640
SVt-LVt	565	11.7	2330
HVt-LVt	630	11.7	2300
LVt-SVt	531	11.7	2370
SVt-SVt	612	11.7	61
HVt-SVt	636	11.7	31.3
LVt-HVt	625	11.7	2340
SVt-HVt	670	11.7	35.1
HVt-HVt	715	11.7	5.39

Annexe B

Extraction des capacités de grille et de jonction

Pour mesurer la capacité de grille, les accès des transistors NMOS et PMOS sont connectés tels que représentés à la figure B.1. Une tension de test augmente de manière linéaire. Pour déterminer la capacité pour chaque tension, on calcule le courant produit par la tension de test divisé par la variation temporelle connue de la tension. La figure B.2 montre que la capacité de grille dépend de la tension de grille. Elle est faible quand le transistor est en régime sous-seuil, augmente progressivement au fur et à mesure que le canal de conduction se forme sous la grille, pour atteindre une valeur limite. La valeur donnée dans le tableau est moyenne parce que la WL doit charger jusque 1V.

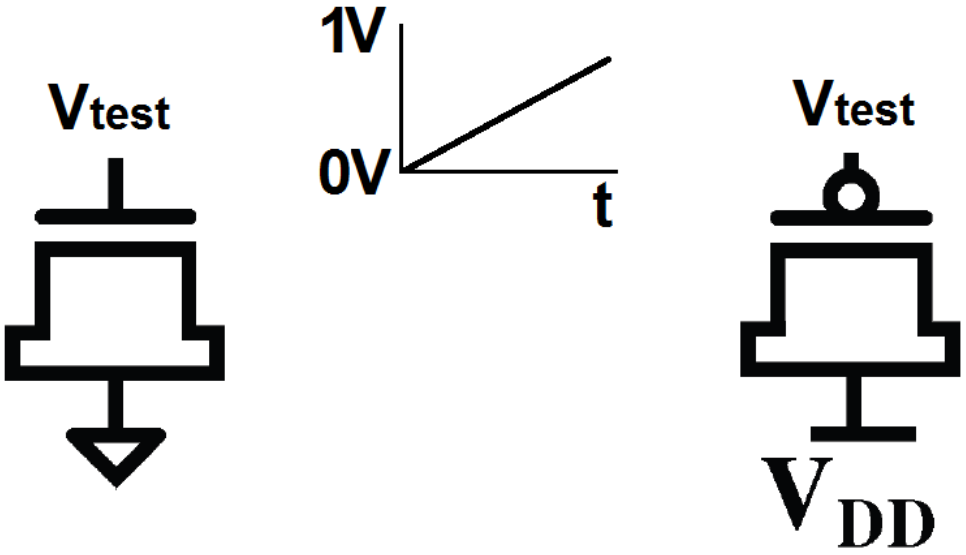


Figure B.1 – *Testbench* d'extraction des capacités de grille.

Pour mesurer la capacité de jonction, les accès sont connectés de sorte que le transistor soit toujours bloqué (figure B.3). Sur la figure B.4, on remarque que la capacité de jonction, elle, ne varie (pratiquement) pas avec la tension de drain (ou de source).

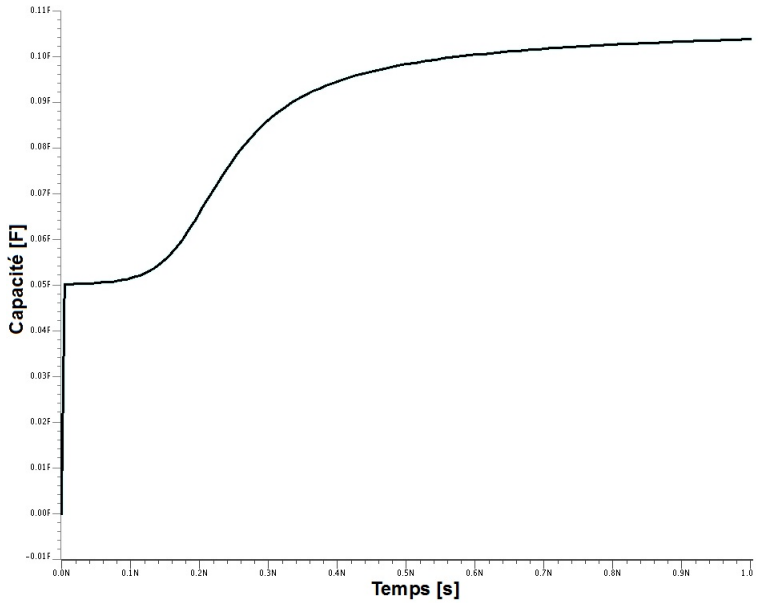


Figure B.2 – La capacité de grille varie en fonction de la tension appliquée à ses bornes. Transistor NMOS SVt sous test.

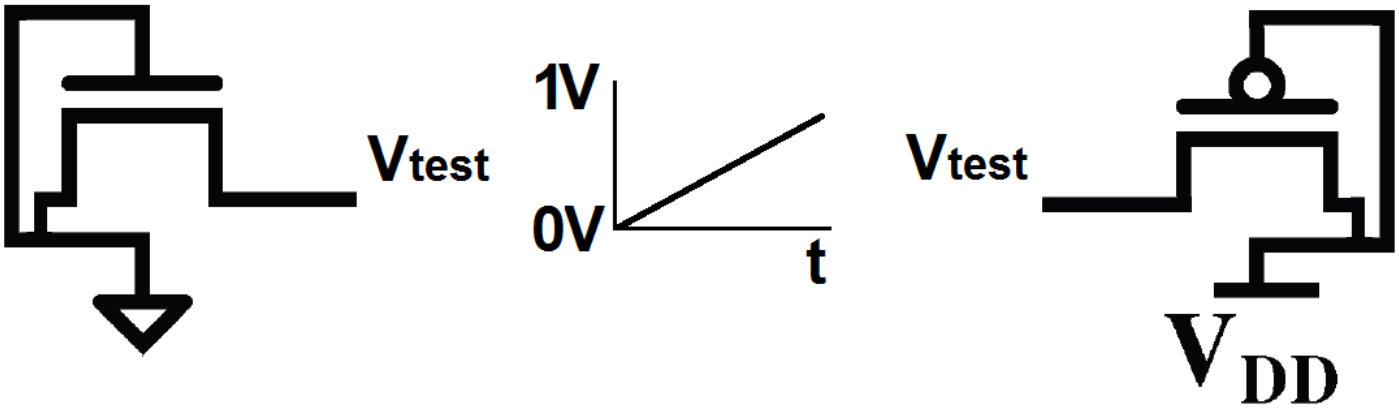


Figure B.3 – *Testbench* d'extraction des capacités de jonction.

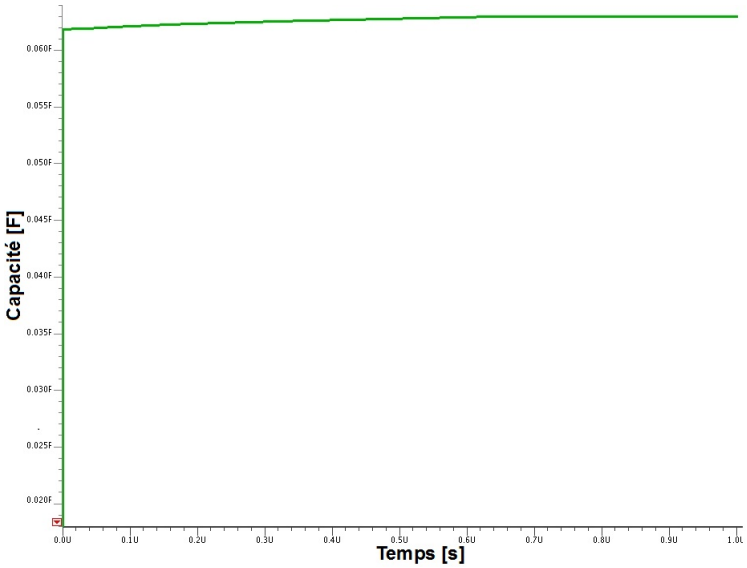


Figure B.4 – La capacité de jonction ne varie pratiquement pas avec la tension appliquée à ses bornes. Transistor PMOS SVt sous test.

Annexe C

Erreur relative du calcul de l'énergie dynamique

Pour rappel, l'énergie dynamique par cellule, que nous avons définie dans la section 2.4.3, est donnée par la formule

$$E_{dyn} = V_{dd} \int_0^{t_{WorR}} \left((I_{dd,Cell} + I_{dd,BL} + \frac{I_{dd,WL}}{128})_{WorR0} + (I_{dd,Cell} + I_{dd,BL} + \frac{I_{dd,WL}}{128})_{WorR1} \right) /2 \tag{C.1}$$

et peut être réécrite sous la forme

$$E_{dyn} = E_{cell} + E_{BL} + \frac{E_{WL}}{128} \tag{C.2}$$

Imaginons maintenant un tableau SRAM fabriqué avec la même technologie, mais avec une longueur de ligne n différente de 128.

Lors d'une écriture, seulement une proportion α_{SW} (le taux de commutation) de cellules changent effectivement d'état sur les n. L'énergie effectivement utilisée est alors

$$E_{dyn,eff} = \alpha_{SW} n E_{cell} + \alpha_{SW} n E_{BL} + n \frac{E_{WL}}{128} \tag{C.3}$$

car E_{WL} est définie pour une ligne de 128 cellules.

L'énergie qui est proposée de calculer est

$$E_{dyn,cal} = \alpha n E_{cell} + \alpha n E_{BL} + \alpha n \frac{E_{WL}}{128} \tag{C.4}$$

L'erreur relative sera alors de

$$\frac{E_{dyn,eff} - E_{dyn,cal}}{E_{dyn,eff}} = \frac{n - \alpha n}{\alpha n} \frac{E_{WL}/128}{E_{cell} + E_{BL} + E_{WL}/128} = 0.01 \frac{1-\alpha}{\alpha} \tag{C.5}$$

selon les valeurs obtenues en section 3.2.7. Si α est très proche de zéro, l'erreur relative explose. Mais, en moyenne, dans un tableau mémoire aux données et entrées aléatoires, α est très proche de 1/2. L'erreur relative est alors de 1%

ὅπερ ἔδει δεῖξαι

Annexe D

Tension de WordLine négative

Cette annexe présente quelques résultats de l'étude de l'*assist* statique qu'est la tension de WL négative en rétention, qui n'étaient pas assez pertinents pour être présentés dans le travail.

La figure D.1 montre clairement que cet *assist* est inefficace avec un transistor HVt.

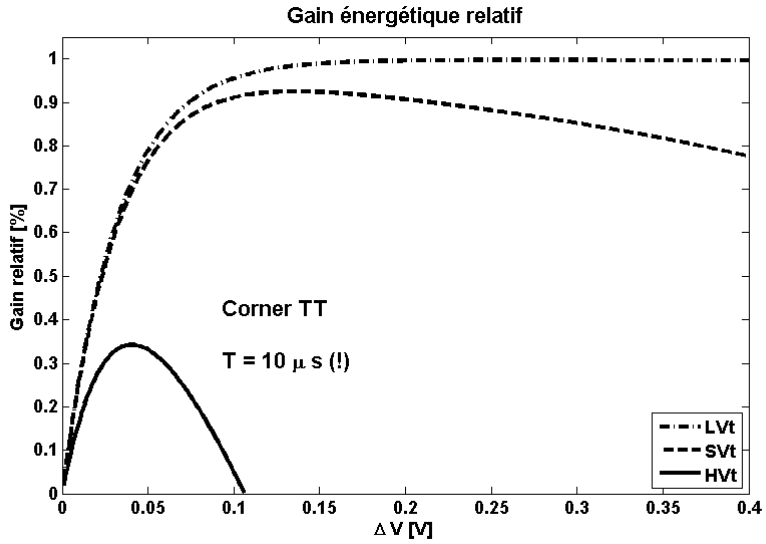


Figure D.1 – Les transistors d'accès HVt ne sont énergétiquement pas efficaces, combinés avec l'*assist* statique.

Dans la section 3.2.1, nous avions fait l'hypothèse, pour fixer le temps moyen entre deux accès T, que les 256 lignes avaient une probabilité égale d'être accédées. Néanmoins, il se peut que cette hypothèse soit peu réaliste. Dans une application FIFO par exemple, les lignes du début de la pile ont beaucoup plus de chance d'être accédées que celles du bas de la pile. La figure D.2 montre que, même dans ces conditions, le transistor LVt reste très efficace énergétiquement. Le transistor SVt par contre reste beaucoup moins efficace, comme montré à la figure D.3.

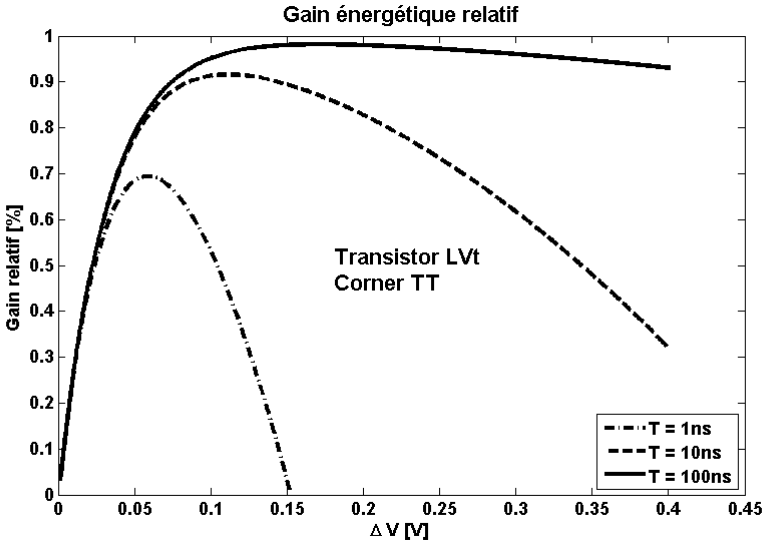


Figure D.2 – Même si la ligne est accédée toutes les $10ns$, le gain relatif reste important pour le transistor LVt.

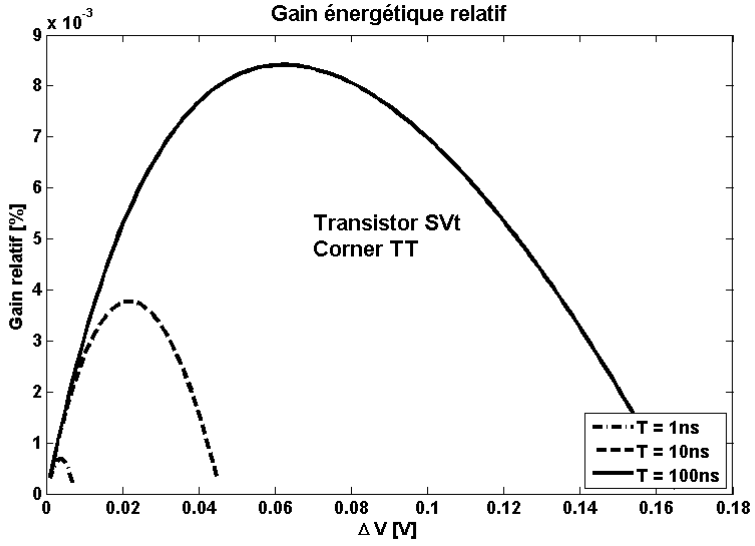


Figure D.3 – Si la ligne est accédée plus souvent que toutes les $256ns$, le gain relatif chute pour le transistor SVt.