Conception de SRAM en 32nm : Exposé des limitations physiques et des solutions étudiées

Matthieu Nongaillard[(a,b)], Richard Ferrant[(c)], Bruno Allard[(a)]
[(a)]Laboratoire AMPERE, 21 avenue Jean Capelle - 69621 Villeurbanne cedex - France
[(b)]NXP, 850 Rue Jean Monnet - 38926 Crolles Cedex - France
[(c)]STMicroelectronics, 850 Rue Jean Monnet - 38926 Crolles Cedex - France

Email : matthieu.nongaillard@insa-lyon.fr

Résumé

Ce papier présente les difficultés de fonctionnement des cellules SRAM dans les technologies les plus avancées. Il expose le fonctionnement d'une cellule ainsi que les différents phénomènes physiques qui limitent de plus en plus ce fonctionnement. Des solutions, au niveau conception sont proposées pour repousser ces effets parasites et améliorer les marges de fonctionnement.

1. Introduction

Depuis maintenant plus de trente ans, la loi de Moore, qui est une loi de réduction des coûts, a toujours été suivie avec succès ; la réduction de la taille des transistors a toujours permis de respecter cette « loi ». Toutefois, en raison des limites physiques soulevées par la réduction des échelles, il est difficile aujourd'hui, de suivre cette tendance. En particulier, l'industrie du semiconducteur ne peut plus se contenter de diminuer la taille des transistors d'une génération à l'autre, il devient également nécessaire de trouver de nouveaux leviers d'action. A chaque nouveau noeud technologique, de nouvelles limites qui n'étaient que peu influentes, deviennent prépondérantes, par exemple la dispersion des performances devient la principale difficulté pour mener à bien une nouvelle conception.

Dans cet article, nous présentons d'abord les principales limites rencontrées, et ensuite, une méthode d'approche pour repousser ces limites.

2. Présentation de la SRAM

En raison de leur niveau de performance, les mémoires statiques à accès aléatoire (SRAM) sont largement utilisées pour l'élaboration des circuits intégrés (certains circuits comme les microprocesseurs comprennent jusqu'à 70% de SRAM).

La cellule SRAM la plus utilisée est une cellule à 6 transistors (on parle alors de 6T-SRAM). Le point mémoire consiste en 2 inverseurs montés tête-bêche et de 2transistors d'accès (Fig.1a). La lecture et l'écriture se font avec 2 signaux de commande : la paire de *bit lines* (BL & BL/) et la *word line* (WL). Pour écrire dans une cellule mémoire, on force la valeur dans la cellule sélectionnée : la *word line* est activée, la paire de *bit line* est forcée à des valeurs CMOS saturés et transfère la donnée à l'intérieur de la cellule. Pour la lecture, la paire de *bit line* est laissée flottante et l'ouverture de la *word line* transfert l'information de la cellule sur la paire de *bit line*.

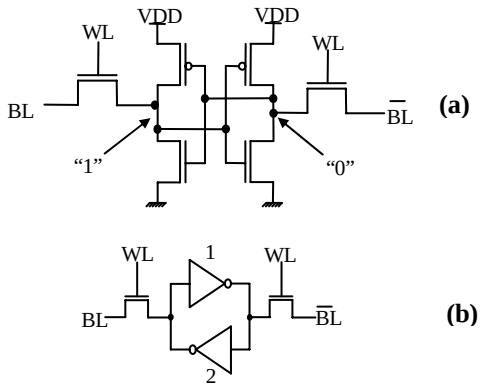

Figure 1 : (a) représentation électrique d'une mémoire 6T-SRAM et (b) vue schématique correspondante

3. Problèmes des SRAM pour les dernières générations technologiques

3.1 La marge statique au bruit

Une des caractéristiques principales de la cellule est sa marge statique au bruit souvent appelée par son acronyme anglais SNM (*Static Noise Margin*). La SNM permet de visualiser la stabilité d'un point mémoire, c'est-à-dire, sa capacité à conserver une information et de résister aux perturbations. On peut visualiser la SNM en traçant la caractéristique Vout (Vin) de l'inverseur 1 puis la caractéristique Vin (Vout) de l'inverseur 2. La SNM est représentée par la diagonale du plus grand carré inscrit dans l'une des deux boucles de la courbe en papillon associé au point mémoire [1]. C'est l'opération

de lecture (et non d'écriture) d'un point mémoire qui est la plus sensible en terme de conservation de l'information. C'est pourquoi la courbe de SNM est prise dans ces conditions (transistors d'accès passants) et on parle alors de *read SNM*.

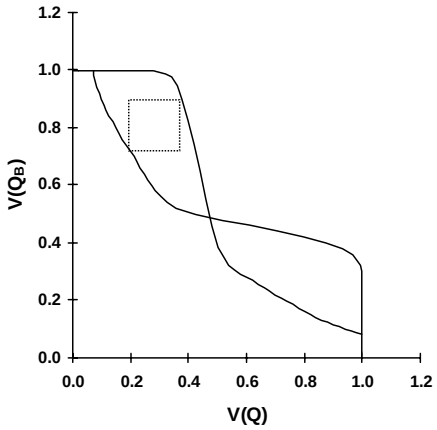

Figure 2 : SNM d'un point mémoire sur une courbe en papillon

La SNM des cellules SRAM des dernières générations technologiques est affectée par beaucoup de phénomènes parasites : désappariement des transistors, courants de fuite, diminution de la fenêtre de fonctionnement (à cause de la réduction des dimensions des transistors et la tension d'alimentation VDD) et la taille de la matrice par rapport aux bus.

3.2 Phénomènes parasites

3.2.1 Désappariement

Lorsqu'une série de circuits est réalisée, on peut mesurer des variations de caractéristiques sur plusieurs niveaux : lot à lot, plaquette à plaquette, puce à puce : ce sont des variations globales. Les variations qui nous intéressent le plus sont les variations locales qui correspondent à des différences entre les transistors d'une même cellule.

La plupart du temps en technologie CMOS, les transistors fonctionnent par paires et les légères variations dues aux procédés de fabrication modifient les caractéristiques des transistors les uns par rapport aux autres. La principale cause du désappariement est la fluctuation aléatoire du nombre de dopant (RDF : *Random Dopants Fluctuations*) [2]. La tension de seuil étant directement proportionnel au nombre de dopants dans le canal du transistor, leur variation aléatoire modifie les tensions de seuil des transistors (V_T).

$$V_{th\infty} = V_{FB} + 2\phi_F + \frac{1}{C_{ox\_eot}}\sqrt{2\varepsilon_{Si} q N_B (2\phi_F)} \quad (1)$$

où V_{FB} est la tension de bande plate, ϕ_F le potentiel de surface en inversion forte, C_{OX} la capacité de l'oxyde, ε_{Si} la permittivité du Silicium, q la charge et N_B le nombre de dopant.

Une des conséquences de ce désappariement, représentée figure 3, est la réduction de la zone de stabilité des points mémoires [3].

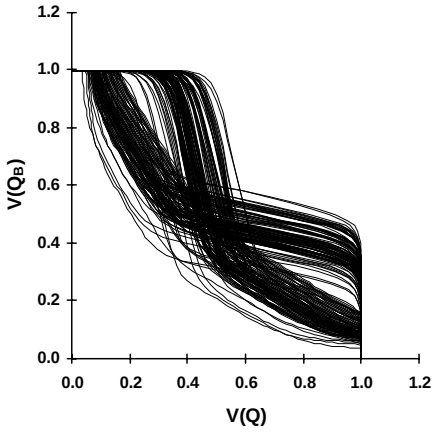

Figure 3 : SNM d'un point mémoire avec prise en compte du désappariement

La déviation standard σ_{VT} sur la tension de seuil est donnée par la relation:

$$\sigma_{VT} = AV_T * \sqrt{\frac{1}{W * L}} \quad (2)$$

où AV_T est une donnée intrinsèque à la technologie utilisée, W et L respectivement les largeurs et longueurs de la grille du transistor. Pour limiter l'action du désappariement sur les tensions de seuil, on ne peut agir que sur 2 facteurs. Nos leviers d'action sont le AV_T ou le W*L : on a le choix entre une action design en agissant sur les dimensions des transistors, ou une action sur les matériaux pour modifier le AV_T. Seulement, pour qu'une action sur le design diminue la disparité des tensions de seuil, on devrait augmenter les tailles des transistors. Or pour être en adéquation avec la loi de Moore, on ne doit pas arrêter de diminuer les dimensions des transistors : pour réduire le désappariement des transistors, il ne reste plus qu'a agir sur le AV_T des matériaux. **Il est très difficile, de par la physique du semiconducteur, de trouver un compromis entre la réduction des dimensions et le désappariement** [4]. L'objectif de l'étude sera en particulier de trouver une technologie permettant de réduire le AV_T. Les possibilités privilégiées étant le FDSOI (Full Depleted Silicon On Insulator) et le Dual Gate CMOS.

Le niveau de qualité acceptable pour une matrice de points mémoires est d'avoir au maximum 100 puces de 10Megabits défectueuses pour 1 millions produites, soit une probabilité de 10^{-11} défaillances par cellule. On peut traduire ce chiffre en équivalent de sigma de la loi gaussienne, c'est-à-dire environ 7σ (Fig.5). En d'autres termes, toutes les cellules ayant une déviation par rapport à la moyenne, inférieure à 7σ seront fonctionnelles.

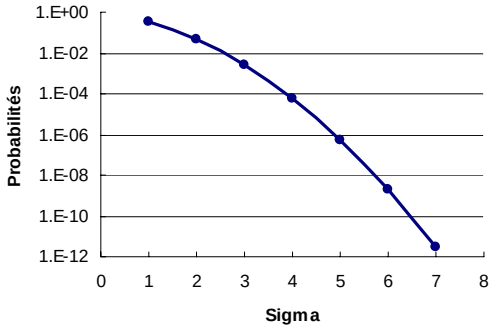

Figure 5 : Probabilités d'évènement en fonction de la loi gaussienne

3.2.2 Courant de fuites

La consommation d'une mémoire SRAM est proportionnelle au nombre de cellules qui la composent et au courant de fuite des transistors. Or, ce dernier reste à peu près identique de génération en génération alors que le nombre de cellules double à chaque nouveau nœud technologique. La consommation augmente donc en conséquence et devient donc critique pour l'élaboration des derniers circuits. Par ailleurs, à cause du désappariement des caractéristiques des transistors, certaines cellules peuvent générer des courants de fuite suffisants pour faire basculer l'état d'un point mémoire. Ces deux phénomènes parasites nous imposent d'avoir des courants de fuite inférieurs à 1pA pour la génération 45nm.

3.2.3 Réduction des dimensions

La figure 4 représente l'évolution de la zone de fonctionnement d'une cellule mémoire.

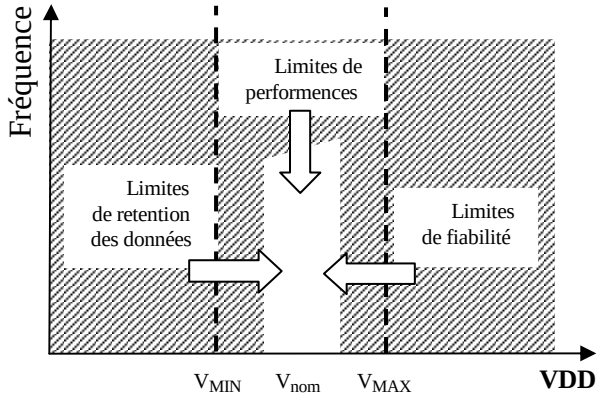

Figure 4 : Fenêtre de fonctionnement d'un point mémoire

A chaque diminution des dimensions des transistors et donc des cellules mémoires, les contraintes sur les paramètres des transistors sont de plus en plus restrictives et la fenêtre de fonctionnement de la cellule mémoire se réduit. Plus cette fenêtre diminue, plus il est difficile d'arriver à faire fonctionner correctement une matrice de points mémoires. On réduit également la marge de flexibilité que l'on pouvait avoir pour les solutions design de la matrice mémoire.

3.2.4 Problèmes dynamiques

Avec la réduction des échelles, certains problèmes prennent de plus en plus d'importance. C'est le cas des délais de propagation des signaux dans les bus de connexions. En effet, la taille des mémoires est de plus en plus importante (plus de 10 Mbits) et donc, les bus de connexion sont eux aussi plus longs. Par conséquent, à cause de la réduction des dimensions, certains phénomènes prennent de l'importance et modifient le comportement des délais de propagation : ces derniers ne sont plus linéaires.

$$\tau = R * C \quad (3)$$

Dans l'équation (3), ce n'est plus la composante capacitive qui pose problème, mais la composante résistive : la résistivité des matériaux augmente quand les sections diminuent. Les délais augmentent alors de façon exponentielle. Plus le noeud technologie est petit, plus la distance critique devient courte : on passe de 250µm en 90nm à 150µm en 65nm, puis à 70µm en 45nm et environ à 40µm en 32nm [5]. Un des avantages de la SRAM par rapport aux autres mémoires est sa rapidité. Or, de grands délais d'accès l'handicaperaient beaucoup. Pour une cellule mémoire de surface 0.374µm² en technologie 45nm, il est possible de placer 128 cellules mémoire avant de rencontrer de réels problèmes de délais.

4. Solutions Etudiées

On peut poser la problématique sous plusieurs angles : réduire l'influence du désappariement dans les cellules, ou augmenter le rendement pour atteindre une qualité acceptable. Nous présentons ici quelques éléments de la seconde approche qui consiste à augmenter le rendement des points mémoires fonctionnels en récupérant des cellules défectueuses.

4.1 Amplificateur de récupération de SNM négatives

Une large majorité des erreurs de lecture sont dues à des points mémoire qui ont une SNM négative. Une cellule qui a une SNM négative, est une cellule dont les lobes de la courbe en papillon ne se coupent qu'une fois: il n'y qu'une zone de stabilité. Lors de la lecture d'une cellule de ce type, l'information est systématiquement perdue si elle n'est pas initialement en zone stable. Pour rendre fonctionnel ces points mémoire, on utilise un amplificateur de lecture spécifique : avant que la cellule ne perde son information, l'amplificateur de lecture va lire, puis réinscrire l'information dans la cellule [6].

Le temps de rétention de l'information d'une telle cellule est de l'ordre de quelques centaines de

picosecondes. En utilisant cette méthode, on récupère environ 2σ de SNM négative.

4.2 Code correcteur d'erreur

L'utilisation d'un code correcteur d'erreurs est une bonne méthode pour augmenter le rendement de cellules fonctionnelles en contrepartie d'une augmentation de la superficie de la mémoire ainsi que du temps de lecture. En effet, plus ce code est performant, plus la superficie qu'il occupe sera grande : il est adapté aux matrices mémoires de grandes tailles (plusieurs Mbits).

L'action d'un code correcteur d'erreur de type hamming a été étudié sur une mémoire SRAM [7], et l'ordre du code correcteur d'erreur (c'est-à-dire, le nombre d'erreurs corrigées par mot) va déterminer le nombre d'erreurs qui peuvent être corrigées et détectées ainsi que le surcoût en superficie que cela va engendrer. L'utilisation d'un code correcteur d'erreur d'ordre 2, permet de diviser les contraintes par deux pour un coût supplémentaire en superficie inférieur à 12% (pour des mots de 128bits) (Tab.1).

Longueur d'un mot	Extra ECC (Double correction)	< à 100ppm / 8Meg Sans ECC	< à 100ppm / 8Meg Avec ECC
128	15 (11.7%)	7σ	4.4σ

Table 1. Etude statistique de l'action d'un ECC sur le rendement d'une mémoire SRAM.

5. Conclusion

Nous avons montrés que le fonctionnement des cellules SRAM se dégrade lorsque les technologies deviennent plus agressives. Dans ce papier nous avons proposé des solutions de conception qui peuvent sensiblement réduire cette dégradation. Même si elles permettent de continuer à suivre la « loi » de Moore, à terme, ces solutions ne seront pas suffisantes et devront être complétées par d'avantages d'efforts, tant au niveau conception que procédés de fabrication.

Références

[1] E.Seevinck, Frans J List and J.Lohstroh, *"Static-Noise Margin Analysis of MOS SRAM Cells"*, IEEE Journal of Solid-State Circuits, Vol. Sc-22, No. 5, October 1987.

[2] B.Cheng, S.Roy, Groy, A.Brown and A.Asenov, *"Impact of Random Dopant Fluctuation on Bulk CMOS 6T SRAM Scaling"*, ESSDERC September 2006

[3] A.Bhavnagarwala, X.Tang and JD.Meindi, *"The Impact of Intrincic Device Fluctuations on CMOS SRAM Cell Stability"*, IEEE Journal of Solid-State Circuits, Vol. 36, No. 4, April 2001.

[4] MJ.Pelgrom, Duinjaijer, Welbers, *"Matching propieties of MOS transistors"*, IEEE Journal of Solid-State Circuits, Vol. 24, No. 5, 1989

[5] M.Sellier et al, soumis à la conférence VLSI 2007.

[6] R.Ferrant et al, soumis à la conférence ICMTD 2007.

[7] C.Maufront, Crolles 2 Alliance, rapport interne.