Etude de la technologie SOI partiellement désertée à très basse tension pour minimiser l'énergie dissipée et application à des opérateurs de calcul.

Alexandre Valentian

TELECOM PARIS
école nationale supérieure des télécommunications

École Doctorale d'Informatique, Télécommunications et Électronique de Paris

Thèse

présentée pour obtenir le grade de Docteur
de l'École Nationale Supérieure des Télécommunications

Spécialité : **Électronique et Communications**

Alexandre VALENTIAN

Etude de la technologie SOI partiellement désertée à très basse tension pour minimiser l'énergie dissipée et application à des opérateurs de calcul

Soutenue le 17 mai 2005 devant le jury composé de :

Michel Ciazynski	Président
Marc Renaudin	Rapporteurs
Jean-Didier Legat	
Marc Belleville	Examinateurs
Jean-François Naviner	
Amara Amara	Directeur de thèse

Remerciements

Ce manuscrit représente l'aboutissement de plusieurs années de travail de recherche dans le domaine de la très basse tension en technologie SOI, effectuées à l'ISEP en collaboration avec le CEA-Leti et STMicroelectronics. Bien que cette thèse porte mon nom, elle est véritablement le fruit de la coopération de plusieurs personnes, qui m'ont encadré, supporté, guidé tout au long de ces années. Ce sont ces personnes que je tiens ici à remercier.

Mon directeur de thèse Amara Amara et le professeur Andrei Vladimirescu ont été mes principaux guides au cours de cette aventure. Je retiens du professeur Amara ses grandes qualités d'encadrement et plus généralement ses qualités humaines. Il a été une personne essentielle pour mon initiation et mon introduction au monde de la recherche, il s'est investi dans mon travail comme peu le font et a toujours été derrière moi dans les moments difficiles. Le professeur Vladimirescu a également été essentiel dans le déroulement de ce travail, par sa disponibilité et les nombreuses discussions que nous avons eu. Sa très bonne connaissance des outils de simulation électrique, de la modélisation des transistors et de la conception des circuits analogiques a réellement élargi mes compétences dans ces domaines.

J'ai également eu la chance de disposer d'excellents conseillers dans le secteur industriel, le docteur Marc Belleville au CEA-Leti et le docteur Philippe Flatresse de STMicrolectronics. Je connais Marc Belleville depuis de nombreuses années, puisqu'il a été mon maître de stage. C'est quelqu'un d'extrêmement compétent et qui dispose d'une bonne expérience dans des domaines variés, avec qui il est toujours très enrichissant d'avoir une conversation. J'ai connu Philippe Flatresse plus récemment, au cours de notre collaboration avec STMicroelectronics. Philippe Flatresse est d'un contact étonnamment facile et a été disponible dès le début. Il m'a permis de découvrir le domaine de la conception de bibliothèques de cellules standards, ainsi que les contraintes introduites par l'industrie dans un travail de recherche, où il faut que ça aille vite, par opposition au monde académique, où l'on souhaite atteindre la solution optimale.

Ces années de recherche n'auraient pas été couronnées de succès sans l'implication de toute l'équipe du laboratoire d'électronique de l'ISEP, qui je dois l'avouer, a été sollicitée surtout lorsque des difficultés se présentaient. Je tiens à remercier notamment Thomas Ea qui m'a aidé à l'écriture de codes vhdl et à la définition de l'architecture du circuit de traitement d'images par ondelettes ; pour sa patience lorsque je lui demandais à maintes reprises de relancer les licences de différents logiciels ; et pour sa relecture très rigoureuse du manuscrit. Je remercie également Frédéric Amiel, pour ses conseils et ses enseignements sur les architectures numériques et la représentation signée. Sans oublier les échanges très nombreux et très importants que j'ai eu avec mes co-thésards, Razvan Ionita et surtout Olivier Thomas. Nous avons commencé ensemble, mûri ensemble et terminé « presque » ensemble.

Mes remerciements ne sauraient être complets sans exprimer ma gratitude aux professeurs Jean-Didier Legat et Marc Renaudin, mes rapporteurs, pour leurs commentaires et remarques très enrichissants sur mon travail, ainsi que leurs nombreuses questions issues d'une lecture très attentive de ce manuscrit. Je souhaite de même remercier le professeur Michel Ciazynski, le président de mon jury et directeur de l'ISEP, sans qui tout ce travail n'aurait pu être possible.

Enfin, je dois une dette certaine à ma femme Sandrine et à mes parents, qui ont été mes plus fidèles supporteurs et qui m'ont aidé à traverser cette période. Ma femme m'a été tout particulièrement indispensable, que ce soit pour sa confiance indéfectible en ma réussite, ou pour la correction des fautes d'orthographe et la mise en page du manuscrit.

J'oublie très certainement des personnes, mais celles-ci se reconnaîtront et doivent être assurées de ma gratitude.

Abrégé

L'évolution des technologies des semi-conducteurs vers des géométries de plus en plus fines permet un accroissement des performances et des fonctionnalités par puce mais s'accompagne simultanément d'une augmentation de la puissance dissipée. Alors que les utilisateurs sont de plus en plus friands d'applications portables, la conception de circuits intégrés doit désormais prendre en compte le budget de puissance alloué. Il est donc essentiel de développer des circuits microélectroniques très basse puissance. La réduction de la tension d'alimentation V_{DD} s'avère une approche très intéressante puisque cela permet de réduire la puissance dynamique quadratiquement et la puissance statique des courants de fuite exponentiellement. L'utilisation de tensions d'alimentation très basses (ULV) a été explorée à Stanford dès 1990 en utilisant une technologie spéciale, dont les transistors possèdent des tensions de seuil proches de zéro volt. Cependant, bien que réduire fortement la tension d'alimentation soit une méthode efficace pour diminuer la consommation, elle ne peut pas être appliquée arbitrairement car cela affecte négativement les performances, le délai dans les portes augmentant exponentiellement lorsque V_{DD} devient inférieur à la tension de seuil. Il faut donc trouver un compromis entre vitesse et consommation. Du point de vue technologique, la technologie SOI-PD (Silicium sur Isolant Partiellement Désertée) s'avère très intéressante en ULV : elle présente des performances entre 25% et 30% supérieures à celles obtenues en CMOS à substrat massif. La technologique SOI permet donc de diminuer la consommation des circuits intégrés à fréquence de fonctionnement égale.

Pour mieux appréhender le comportement des transistors SOI opérés en inversion faible, un modèle analytique et physique simple a tout d'abord été développé. La consommation d'un circuit dépendant fortement du style logique employé, plusieurs styles ont été comparés et celui présentant le meilleur produit puissance-délai a été choisi pour réaliser une bibliothèque de cellules standards. La problématique de la propagation de données sur des interconnexions longues, alors que les transistors fournissent peu de courant, a été abordée : un nouveau circuit de transmission en mode courant a été proposé. Enfin, un circuit de traitement d'image par paquets d'ondelettes a été développé et synthétisé grâce à la bibliothèque précédente.

Sommaire

Liste des figures

Liste des tableaux

1 Introduction

L'industrie microélectronique connaît depuis de nombreuses années une évolution exponentielle des fonctionnalités offertes par les microprocesseurs, prédite ou plutôt guidée par la loi de Moore [Moor65], qui prévoit un doublement du nombre de transistors par circuit intégré tous les deux ans. Cette évolution a été rendue possible par la miniaturisation sans cesse renouvelée des composants élémentaires des circuits microélectroniques, aujourd'hui les transistors MOS à effet de champ. Les améliorations parallèles dans le domaine des architectures, avec l'utilisation de cœurs RISC, du pipeline ou de mémoires caches, ont conduit à une augmentation des performances d'un facteur 1,5 à 2 tous les ans aux cours des décennies 80 et 90. Les seules contraintes étaient jusqu'alors le délai et la surface. L'apparition du premier microprocesseur dissipant plusieurs dizaines de Watts au début des années 1990 fut une prise de conscience pour l'industrie dans son ensemble [Dobb92]. La puissance consommée doit être limitée pour des raisons de dissipation de chaleur et de fiabilité. Par la suite, la croissance du marché des applications portables, assistants personnels (PDA), téléphones cellulaires, récepteurs GPS, a accrue la contrainte sur l'énergie dissipée, le succès de ces applications dépendant de l'autonomie des batteries.

Il est donc essentiel de développer des circuits microélectroniques très basse puissance. La réduction de la tension d'alimentation V_{DD} s'avère une approche très intéressante puisque cela permet de réduire la puissance dynamique quadratiquement et la puissance statique, due aux courants de fuite, exponentiellement. L'utilisation de tensions d'alimentation très basses (ULV) a été explorée à Stanford en 1990 [Burr91] et reprise depuis [Hass01]. Cependant, bien que réduire fortement la tension d'alimentation soit une méthode efficace pour diminuer la consommation, elle ne peut pas être appliquée arbitrairement car cela affecte négativement les performances, le délai dans les portes augmentant exponentiellement lorsque V_{DD} devient inférieur à la tension de seuil. Il faut donc trouver un compromis entre vitesse et consommation. Pour les applications ne nécessitant pas des performances élevées, inférieures à 100MHz, travailler en ULV permet de réduire l'énergie dissipée.

Du point de vue technologique, la technologie SOI-PD (Silicium sur Isolant Partiellement Désertée) s'avère très intéressante en ULV. Elle a pour avantages, par rapport au CMOS à substrat massif (*bulk*), la modulation dynamique de la tension de seuil, des capacités de jonction réduites d'un ordre de grandeur et l'absence de l'effet source suiveur dans l'empilement de transistors. Elle présente des performances entre 25% et 30% supérieures à celles obtenues en CMOS *bulk* [Shah99]. La technologique SOI permet donc de diminuer la consommation des circuits intégrés à fréquence de fonctionnement égale.

L'objectif de cette thèse est de développer des circuits arithmétiques fonctionnant à très basse tension. La technologie utilisée est la technologie SOI-PD 0,13µm de STMicroelectronics. Le but recherché est de réduire au maximum la consommation d'énergie. Le chapitre 2 commence par la problématique de la dissipation d'énergie. Les différentes sources de consommation et les techniques existantes pour les réduire sont détaillées. La technologie SOI est introduite dans le chapitre 3 : après avoir présenté ses avantages et ses inconvénients face au CMOS *bulk*, le choix de la valeur de la tension d'alimentation est explicité. En ULV, les transistors opérant entre les modes d'inversion faible et d'inversion modérée, il est primordial de bien modéliser les courants de conduction sous le seuil. Le chapitre 4 décrit donc un modèle analytique et physique simple de caractérisation du courant sous le seuil. Le choix du style logique est très important concernant la dissipation d'énergie mais aussi concernant la robustesse des circuits réalisés en ULV : le chapitre 5 compare différents styles logiques et différents circuits séquentiels et détaille la bibliothèque de cellules standards qui a été développée, en partant d'une bibliothèque de

STMicroelectronics. Dans le chapitre 6, le problème de la transmission d'informations sur les interconnexions longues, dû à la faiblesse des courants fournis par les transistors, est abordé. Deux modes de transmission sont comparés : le mode tension et le mode courant. Pour le dimensionnement des inverseurs en mode tension, un programme Matlab a été conçu, prenant en compte la variation du rapport optimal des mobilités en fonction de la largeur des transistors et les discontinuités introduites par les contacts de substrat. Pour le mode courant, un circuit émetteur-récepteur a été développé. Le chapitre 7 décrit quant à lui un circuit de traitement d'image par ondelette pour une application de reconnaissance par l'iris de l'œil. Ce circuit a été synthétisé grâce à la bibliothèque basse tension qui a été développée. Les résultats obtenus dans la technologie SOI sont comparés à la technologie silicium sur substrat massif. Enfin, le chapitre 8 présente la conclusion.

2 Notions de base

L'augmentation de la densité d'intégration des transistors et la demande toujours plus forte en systèmes portatifs alimentés par batterie ont fait de la réduction de la puissance dissipée un des objectifs prioritaires lors de la conception de circuits intégrés. Bien que des circuits très économes en énergie ont depuis longtemps été nécessaires – montres, calculatrices solaires –, la puissance dissipée par les circuits microélectroniques est désormais devenue un facteur limitatif dans nombre d'applications, facteur qui doit être considéré très tôt dans la phase de conception.

La dissipation de puissance peut s'interpréter de deux manières différentes. Dans le domaine des systèmes à très hautes performances, la puissance consommée se traduit par la quantité de chaleur à évacuer. Les transitions rapides des circuits créent en effet des points chauds qu'il convient de refroidir efficacement par le boîtier. La limitation vient donc ici, dans un premier temps, de la capacité à extraire suffisamment de quantité de chaleur et, dans un deuxième temps, du coût que cela induit.

Dans le domaine des applications portables, la puissance consommée se traduit par la durée de vie de la batterie. Ce qui est important n'est pas tant de diminuer les pics de courant que de réduire la consommation moyenne. Bien évidemment, ces deux domaines ne sont pas exclusifs, c'est-à-dire que l'on peut avoir à réduire la puissance consommée à la fois pour des

problèmes thermiques et de durée de vie des batteries, le marché étant demandeur de mobilité accrue accompagnée de toujours plus de fonctionnalités, telles que la voix, la vidéo ou encore les jeux.

Dans ce chapitre, nous allons expliciter plus en détails les besoins de développer des circuits basse puissance, donner les différentes sources de dissipation de puissance puis les tendances des technologies à venir et les solutions apportées.

2.1 Problématique

La puissance consommée par les circuits microélectroniques a émergé aujourd'hui comme étant une des contraintes les plus importantes lors de la conception de ces circuits. L'accroissement des densités d'intégration grâce à des procédés lithographiques de plus en plus fins et l'accroissement simultané de la fréquence de fonctionnement se traduisent par une augmentation très forte de la puissance dissipée, augmentation qui est exponentielle ainsi que le montre la Figure 1 : les valeurs données concernent les processeurs Intel, du 386 au Pentium [Inte98]. La puissance dissipée par ce type de processeur dépasse désormais la centaine de Watts.

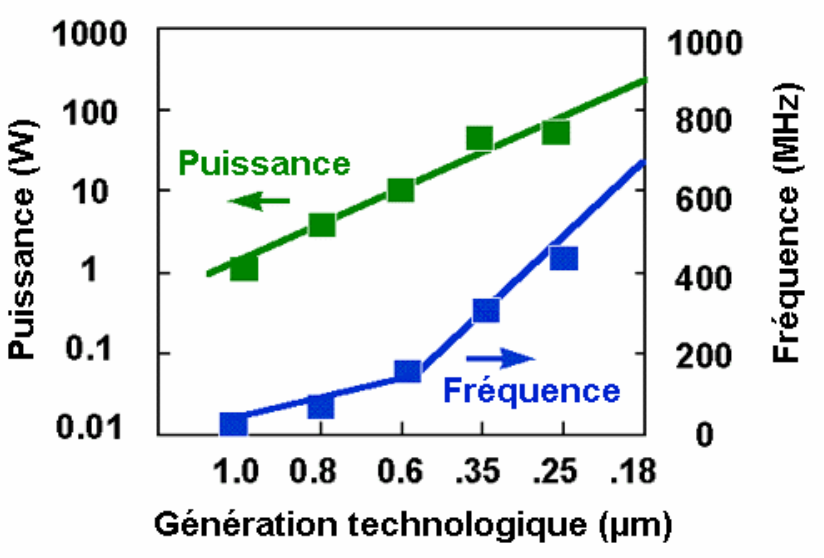


Figure 1 Evolution de la puissance consommée et de la fréquence de fonctionnement des processeurs Intel, en fonction de la génération technologique [Inte98].

Les prévisions ITRS – *International Technology Roadmap for Semiconductors* – mises à jour en 2004 montrent une évolution linéaire plutôt qu'exponentielle de la puissance dissipée

par les microprocesseurs à usage général : cette évolution, représentée dans la Figure 2, est même limitée aux alentours de 200W à partir du nœud technologique hp65[1] jusqu'au nœud technologique hp22, c'est-à-dire entre 2008 et 2018. Cette évolution n'est bien évidemment pas naturelle mais représente la limite maximale autorisée pour des problèmes de refroidissement. En ce qui concerne les applications portables, les contraintes sont encore plus fortes puisque les prévisions ITRS limitent la puissance consommée à moins de 3W.

La dissipation de puissance étant un objectif de conception à part entière, nous allons dans la suite en voir les enjeux puis détailler les différentes sources de consommation et les solutions qui peuvent être apportées.

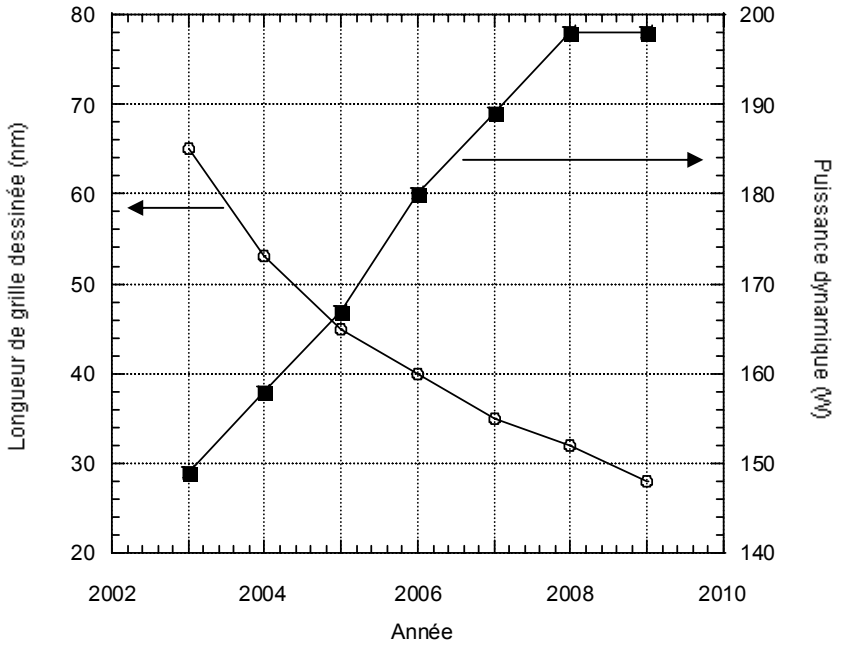


Figure 2 Prévisions ITRS'04 concernant la puissance dissipée par les microprocesseurs à usage général.

2.1.1 Dissipation de chaleur

Comme indiqué dans la Figure 1, les processeurs actuels présentent une consommation moyenne maximale approchant voire dépassant les 100W, la consommation instantanée pouvant être encore plus élevée. Le courant consommé dans un circuit microélectronique est exclusivement transformé en chaleur, suivant la loi d'Ohm. A moins que cette chaleur ne soit

[1] Pour diminuer les confusions, les données ITRS se basent désormais sur des nœuds technologiques correspondant aux motifs géométriques les plus petits des fondeurs, qui sont actuellement obtenus pour les cellules DRAM. Le nœud technologique hp65 signifie donc un *half-pitch* ou demi pas de 65nm pour le métal des cellules DRAM.

évacuée, le circuit intégré s'échauffe, ce qui dégrade les propriétés électriques du circuit et peut aller jusqu'à détruire le composant.

Le coût engendré par les systèmes de refroidissement est le suivant [Vara02] : lorsque la consommation est inférieure à 1W, un simple boîtier en plastique peut être utilisé, au coût d'environ 1 centime de dollar par broche. De 1 à 2W, des dissipateurs de chaleur doivent être ajoutés au boîtier, faisant augmenter le prix à 2 ou 3 centimes. Entre 2 et 10W cependant, un boîtier céramique est substitué au boîtier plastique : ces dispositifs coûtent entre 5 et 10 centimes de dollar par broche. Au-delà de 10W, des systèmes de ventilation doivent être ajoutés.

La dissipation de chaleur des processeurs actuels a atteint des proportions telles, de l'ordre de plusieurs dizaines de watts/cm², que des systèmes de refroidissement très complexes et donc très coûteux doivent être utilisés. La production de ces systèmes, qui se fait en très grands volumes, implique qu'un gain limité sur la quantité de chaleur à évacuer se traduit par d'importantes économies pour le fabricant. Réduire la puissance consommée est donc nécessaire pour des raisons de coût et de durée de vie du produit.

2.1.2 Systèmes portables

La demande du marché pour des applications portables ne se dément pas et s'accompagne en plus d'une volonté d'avoir des systèmes autonomes pendant une longue durée : le système ne doit pas seulement être déplaçable, il doit être nomade. Les utilisateurs veulent pouvoir utiliser leurs ordinateurs portables, leurs outils de communication ou multimédia sans avoir à se relier à une source d'énergie externe. La durée de vie des batteries est donc un facteur déterminant dans le processus d'achat. Dans le même temps, les contraintes de la portabilité imposent des restrictions sévères sur la taille, le poids et donc l'énergie stockée. Augmenter la capacité de stockage est un des moyens de s'attaquer au problème.

Les batteries utilisées ont fait des progrès ces dernières années [Buch03] : la batterie conventionnelle nickel-cadmium permet de stocker entre 45 et 80 Wh/kg. Celle-ci a été remplacée par la batterie nickel-métal-hybride, offrant entre 60 et 120Wh/kg de densité d'énergie. Désormais, la technologie qui s'impose est la technologie lithium-ion qui a une densité comprise entre 110 et 160Wh/kg. De nouvelles technologies sont en cours de développement : on peut citer la technologie lithium-polymère ou encore zinc-air. Cependant, les gains en capacité de stockage sont limités et ne peuvent pas suivre l'augmentation exponentielle de la consommation des processeurs : le développement de systèmes basse

consommation est nécessaire pour diminuer l'énergie demandée aux batteries et augmenter l'autonomie des systèmes portables.

2.1.3 Fiabilité

De nombreux problèmes de fiabilité des circuits microélectroniques sont directement liés à leur consommation. Le premier d'entre eux est le phénomène d'électromigration, durant lequel des atomes des lignes d'alimentation peuvent migrer, aboutissant à des courts-circuits ou des sectionnements des lignes de métal. L'électromigration des atomes est directement liée à l'intensité du courant qui traverse les lignes : une consommation électrique trop importante peut provoquer un vieillissement prématuré du circuit. De plus, nous l'avons vu plus haut, toute la consommation d'énergie est transformée en chaleur : les circuits ont donc tendance à fonctionner à des températures élevées. Ces températures élevées accélèrent d'autres mécanismes de vieillissement des circuits tels que rupture des jonctions, claquage de l'oxyde de grille ou encore endommagement du boîtier d'encapsulation.

Au-delà de ces phénomènes physiques, une forte consommation de courant peut avoir des conséquences sur le comportement électrique des systèmes microélectroniques. Un courant important transporté par le réseau d'alimentation va provoquer une chute de potentiel dans les lignes d'alimentation, due à leur résistance non nulle : la tension d'alimentation fournie à différentes parties du système n'est pas la même. Certaines parties du circuit vont fonctionner à une tension inférieure à la tension nominale pour laquelle elles ont été conçues. Ce problème est connu sous le nom de « chute *iR* ». De plus, de forts courants tirés par une partie du circuit vont provoquer des bruits transitoires dans le réseau d'alimentation, qui nécessiteront des capacités encombrantes pour être atténués – mais pas totalement éliminés. L'inductance des lignes est également un facteur à considérer : des variations brusques du courant demandé vont engendrer des variations en tension, qui peuvent affecter le fonctionnement d'autres parties du circuit. Ce phénomène est connu sous le nom de « rebond d'alimentation » ou problème « *Ldi/dt* ».

Tous ces problèmes nécessitent bien évidemment des outils adaptés pour pouvoir être gérés : outil de dimensionnement des lignes d'alimentation, outil de vérification électrique, … Les technologies futures, présentant des motifs géométriques plus petits, un nombre plus important de transistors et une fréquence de fonctionnement plus élevée, ne vont faire qu'aggraver ces phénomènes : la diminution de la consommation de courant aide à les atténuer.

2.1.4 Puissance et énergie

Avant de commencer à expliciter les différentes sources de dissipation d'énergie, il est important de faire la distinction entre les deux termes clé liés à la consommation de courant, que sont la puissance et l'énergie. L'énergie représente la quantité de travail dépensée pour réaliser une opération logique et s'exprime en Joules. La puissance représente la vitesse à laquelle cette quantité d'énergie est dissipée et a pour unité le Watt, ou Joules/sec. L'objectif de réduire la puissance dissipée par un circuit est donc différent de celui consistant à en réduire l'énergie.

La puissance est essentiellement un problème pour des questions de dissipation de chaleur. Si trop de joules sont convertis en quantité de chaleur pendant un temps très court, les systèmes de dissipation de chaleur n'arriveront pas à l'évacuer entièrement : une élévation de la température et une dégradation thermique en résulteront, comme nous l'avons vu précédemment. A l'inverse, si la même quantité de chaleur est relâchée en un temps plus long, le dispositif d'évacuation de la chaleur maintiendra la température à un niveau acceptable. Dans les deux cas, l'énergie dissipée est la même.

La réduction d'énergie porte sur le nombre total de Joules consommés et non pas sur la vitesse à laquelle ils le sont. Le facteur de qualité que l'on considère lors de la conception d'un circuit est donc l'« énergie par opération » : cette métrique représente l'efficacité réelle d'un circuit à effectuer une opération. Minimiser l'énergie consommée revient à augmenter l'autonomie d'un système alimenté par batterie.

L'inconvénient de ne considérer que le seul paramètre énergie est qu'il n'y a pas de notion de temps, et donc de performance d'un processeur. Le produit énergie $\times$ délai est alors quelques fois utilisé, le terme « délai » étant le temps nécessaire à l'exécution d'une opération. De manière plus générale, on peut considérer comme facteur de mérite l'expression $énergie^x \times délai^y$, x étant d'autant plus élevé que l'on veut favoriser l'énergie dissipée, et y la performance.

2.2 Les sources de dissipation d'énergie

Nous allons prendre la logique CMOS conventionnelle comme exemple, les sources de dissipations d'énergie étant les mêmes pour d'autres styles logiques. La puissance

consommée par une porte logique est composée de trois termes : les puissances dynamique, de court-circuit et de courants de fuite. La Figure 3 montre un inverseur chargé par une capacité C_L et soumis à un créneau en entrée. Les chemins des différentes sources de courant sont illustrés par des flèches, dont l'épaisseur des pointillés indique l'importance dans la consommation totale.

La première source de dissipation d'énergie est, de part son importance, la consommation dynamique : elle est due au changement d'état d'une porte, qui implique un chargement ou un déchargement de la capacité de sortie. Vient ensuite l'énergie consommée par les courants de fuite qui est de plus en plus importante du fait de l'abaissement des tensions de seuil des transistors. Enfin, lors d'une commutation lente de l'entrée, les deux transistors NMOS et PMOS peuvent être simultanément passant, créant un chemin de conduction direct entre l'alimentation et la masse : c'est le courant de court-circuit. Il convient de rajouter à ces sources celle due à un courant statique, qui peut être présent dans des circuits spécialisés tels des circuits analogiques ou dans des styles logiques, telle la logique pseudo-NMOS. Nous allons dans la suite détailler ces différentes sources de consommation d'énergie.

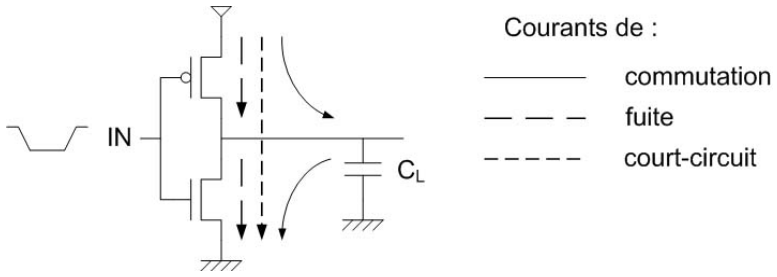


Figure 3 Illustration des principales sources de dissipation d'énergie à l'aide d'un inverseur.

2.2.1 Puissance de commutation

La puissance dissipée lors de la commutation des nœuds d'un circuit est la source principale de consommation dynamique et est bien souvent confondue avec le terme « puissance dynamique ». L'énergie n'est consommée que lors de la charge de la capacité de sortie, c'est-à-dire pour une transition haut vers bas de la tension d'entrée. La quantité de charge stockée dans la capacité est $C_L V_{DD}$ et l'énergie fournie par l'alimentation pour la charger est $C_L V_{DD}^2$. Cette quantité d'énergie est dissipée par le transistor NMOS lorsque la tension de sortie est ramenée à la masse.

L'expression de la puissance dynamique est la suivante [Veen98]:

$$P = a \cdot f \cdot C_{TOT} \cdot V_{DD}^2,$$

Équation 1

où a est l'activité de la porte, f la fréquence de fonctionnement, V_{DD} la tension d'alimentation et C_{TOT} la capacité totale de sortie, qui est composée de la capacité de charge C_L, de la capacité de sortie intrinsèque de la porte et de la capacité parasite due à l'interconnexion. La puissance dissipée dépend directement de l'activité de la porte et varie quadratiquement avec la tension d'alimentation V_{DD}.

2.2.2 Puissance de court-circuit

Lors du changement de l'état logique à l'entrée d'une porte, les transistors des arbres N et P peuvent être conducteurs simultanément, provoquant un courant de court-circuit entre l'alimentation et la masse. Cela se produit lorsque l'entrée varie lentement devant la sortie. L'expression de cette puissance de court-circuit pour un inverseur non chargé est la suivante [Veen98]:

$$P = \frac{\beta}{2} \cdot (V_{DD} - 2V_T)^3 \cdot \frac{\tau}{T},$$

Équation 2

avec β le facteur de gain, V_{DD} la tension d'alimentation, V_T la tension de seuil – considérée égale pour les transistors NMOS et PMOS –, τ la pente du signal d'entrée et T sa période. Bien évidemment, cette expression n'est valable que pour une tension d'alimentation supérieure à la somme des tensions de seuil des transistors NMOS et PMOS. Pour une tension d'alimentation inférieure, le courant de court-circuit est négligeable. La règle de conception, lorsque l'on considère le courant de court-circuit, consiste à avoir des temps de montée – ou de descente – à l'entrée de la porte, moins de deux fois supérieurs à ceux de la sortie, en dimensionnant correctement les transistors. La puissance de court-circuit peut ainsi être limitée à 10% de la consommation dynamique, devenant un effet du deuxième ordre [Chan95].

2.2.3 Courants de fuite

Un transistor n'est pas un interrupteur idéal : non seulement sa résistance n'est pas nulle lorsqu'il est passant, mais elle n'est pas non plus infinie lorsqu'il est bloqué. Il existe un courant I_{OFF} lorsque le transistor est dans un état bloqué. Ce courant est fonction du profil de

dopage et des dimensions physique et effective du canal, de la profondeur des jonctions drain/source, de la tension de seuil V_T, de la tension d'alimentation V_{DD} et de la température. Un transistor est affecté par huit courants de fuite différents [Kesh97] qui sont illustrés Figure 4 :

- le courant de conduction sous le seuil I_{STH},
- le courant dû à l'abaissement de la barrière de potentiel par le drain I_{DIBL},
- le courant de fuite du drain induit par la grille I_{GIDL},
- le courant de fuite de la jonction p-n du drain polarisée en inverse I_R,
- le courant tunnel à travers l'oxyde de grille I_{OX},
- le courant de grille dû à l'injection de porteurs chauds I_{II},
- le courant de perforation I_{PT},
- le courant de surface du canal dû à un effet de canal étroit.

Il faut noter que le courant tunnel à travers l'oxyde de grille I_{OX} ne se manifeste que lorsqu'un potentiel non nul est appliqué sur la grille, c'est-à-dire lorsque le transistor est passant. Quant au courant de grille dû à l'injection de porteurs chauds I_{II}, il traduit un vieillissement du transistor, à la suite de l'introduction d'électrons et de trous dans l'oxyde.

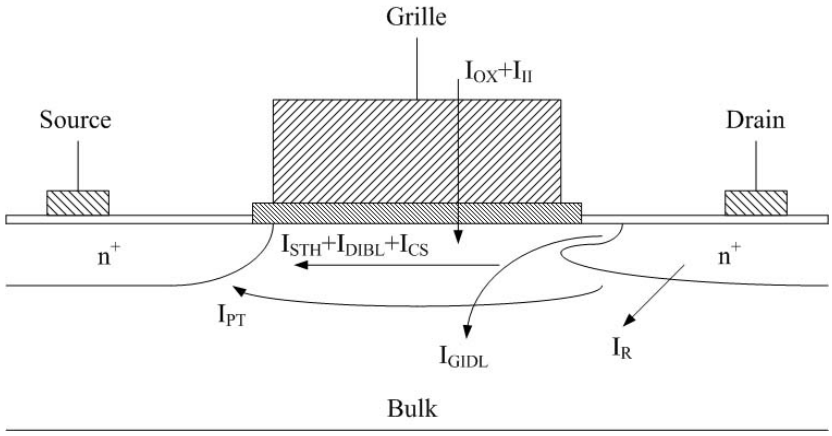


Figure 4 Illustration des différents courants de fuite présents dans un transistor fortement sous-micronique.

Les courants de fuite dominants qui composent le courant I_{OFF} sont :

- le courant sous-seuil I_{STH},
- le courant de polarisation inverse de la jonction p-n au niveau du drain I_R,
- le courant I_{DIBL},
- et le courant I_{GIDL}.

Courant de fuite du drain induit par la grille (GIDL)

Le courant GIDL trouve son origine au niveau du chevauchement du drain par la grille : il est dû à un effet tunnel de bande à bande et dépend fortement du champ électrique transverse et du profil de dopage de la jonction. Il se manifeste pour des polarisations de grille négatives et des valeurs élevées de V_{DS}, comme on peut le voir Figure 5.

Abaissement de la barrière de potentiel par le drain (DIBL)

Le DIBL se produit lorsqu'un potentiel élevé est appliqué au drain : la région de déplétion du drain interagit avec la source près de la surface, abaissant la barrière de potentiel. La source introduit alors plus de porteurs dans le canal sans variation du potentiel de grille. L'effet DIBL se manifeste d'autant plus que la tension V_{DS} est élevée et la longueur effective L_{eff} du transistor courte : il est proportionnel à V_{DS}/L_{eff}^2 [Veen98]. L'effet DIBL abaisse la tension de seuil du transistor mais ne modifie pas la pente sous le seuil : le DIBL peut être mesuré comme la variation du courant I_{DS} pour une variation de la tension V_{DS}, à tension V_{GS} constante. L'effet DIBL est illustré Figure 5 : il déplace la courbe vers le haut et la gauche lorsque la tension V_{DS} augmente.

Courant de polarisation inverse de la jonction p-n

Le courant de polarisation inverse I_R a deux composantes principales : la première est la diffusion de porteurs minoritaires près du bord de la région de déplétion, la deuxième provient de la génération de paires électrons-trous dans la région de déplétion. Le courant de fuite de la jonction en inverse dépend de la surface de la jonction et de la concentration du dopage.

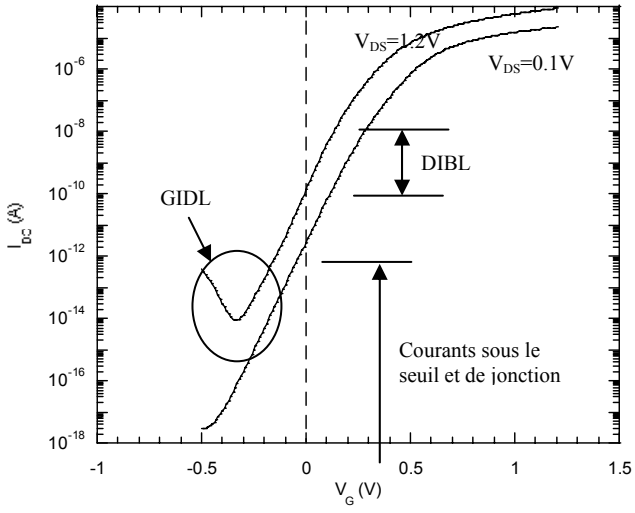


Figure 5 Caractéristique I_D (V_{GS}) d'un transistor NMOS montrant les composants principaux du courant I_{OFF} que sont le courant sous le seuil, le courant inverse de la diode, le DIBL et le GIDL dans la technologie SOI 0,13µm.

Courant de conduction sous le seuil

Le courant de conduction sous le seuil ou courant en inversion modéré est le courant entre la source et le drain qui a lieu lorsque la tension V_{GS} est nulle. C'est un courant de porteurs minoritaires le long de la surface du canal : il est fonction de la tension de seuil V_T et de la pente sous le seuil S et a pour expression [Hori93] :

$$I_{fuite} = W \cdot \frac{I_0}{W_0} \cdot 10^{-\frac{V_T}{S}},$$

Équation 3

où I_0/W_0 représente la référence de densité de courant. Ce courant de fuite varie exponentiellement avec la tension de seuil comme indiqué Figure 6 : cette figure représente le courant I_{DS} en fonction de la tension de grille V_{GS} d'un transistor NMOS pour deux polarisations de substrat, V_{BS}=0V et V_{BS}=0,6V, et pour une tension V_{DS}=1,2V. Les simulations Eldo[2] ont été réalisées à l'aide du modèle SOI 0.13µm de STMicroelectronics et montrent qu'une variation de 150mV de la tension de seuil entraîne une modification d'un facteur 103 du courant de fuite, ce qui est cohérent avec une pente sous le seuil d'environ 80mV/décade.

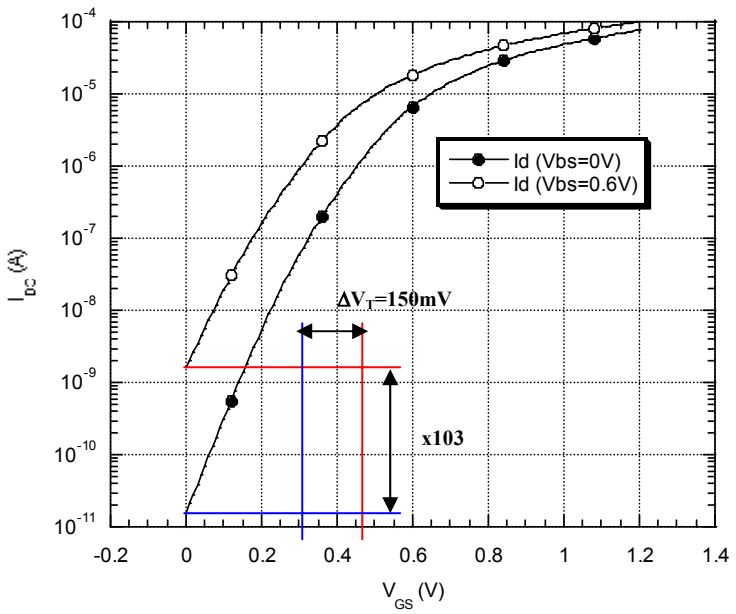


Figure 6 Influence de la tension de seuil sur les courants de fuite d'un transistor NMOS.

[2] Eldo est le simulateur électrique de Mentor Graphics

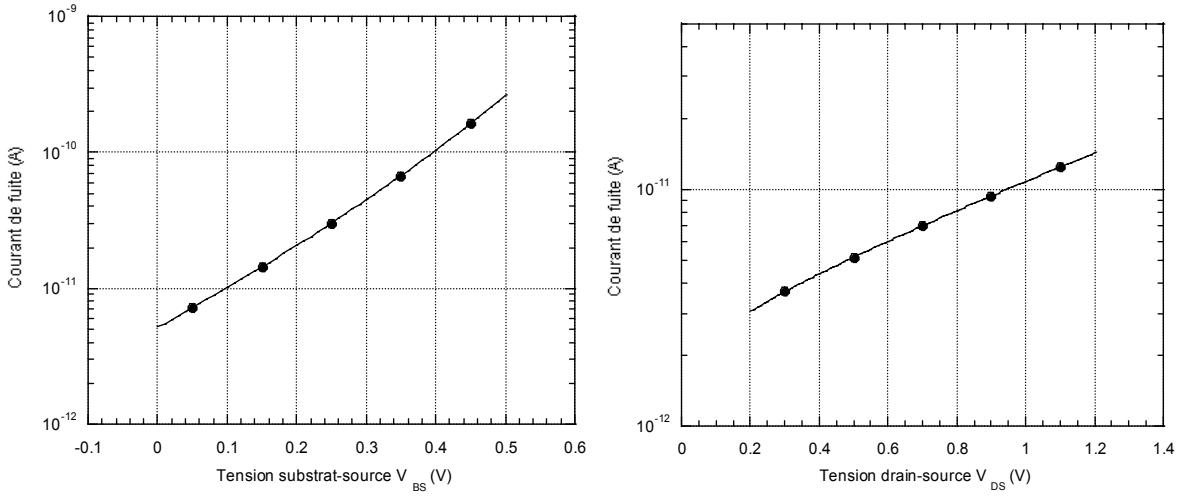
La tension de seuil d'un transistor dépend de la tension substrat-source V_{BS} selon l'expression suivante :

$$V_T = V_{T0} + \gamma\left(\sqrt{2\psi_F - V_{BS}} - \sqrt{2\psi_F}\right)$$

Équation 4

où le paramètre γ représente le coefficient d'effet de substrat et le paramètre ψ_F le potentiel de Fermi dans le substrat. Mais la tension de seuil d'un transistor dépend aussi de la tension drain-source V_{DS}. Deux effets entrent en jeu. Le premier est la « rétroaction statique du drain ». La zone de déplétion sous la grille est influencée par le potentiel de canal, qui varie du drain à la source, et est donc influencée par la tension V_{DS}. Un potentiel de drain plus élevé va augmenter la zone de déplétion, augmenter le nombre de porteurs minoritaires et réduire la barrière que le potentiel de grille doit surmonter pour créer une couche d'inversion. L'augmentation de la tension V_{DS} a donc pour conséquence de réduire la tension de seuil V_T. Le deuxième effet est un effet que l'on a vu précédemment, l' «abaissement de la barrière de potentiel par le drain » ou DIBL.

La Figure 7 montre l'évolution du courant de fuite d'un transistor en fonction de sa tension substrat-source V_{BS}, Figure 7a, et de sa tension drain-source V_{DS}, Figure 7b. L'échelle des ordonnées étant logarithmique dans les deux cas, on remarque que le courant de conduction sous le seuil varie exponentiellement tant avec la polarisation du substrat que la tension drain-source, qui est proportionnelle à la tension d'alimentation V_{DD}.



(a) **(b)**

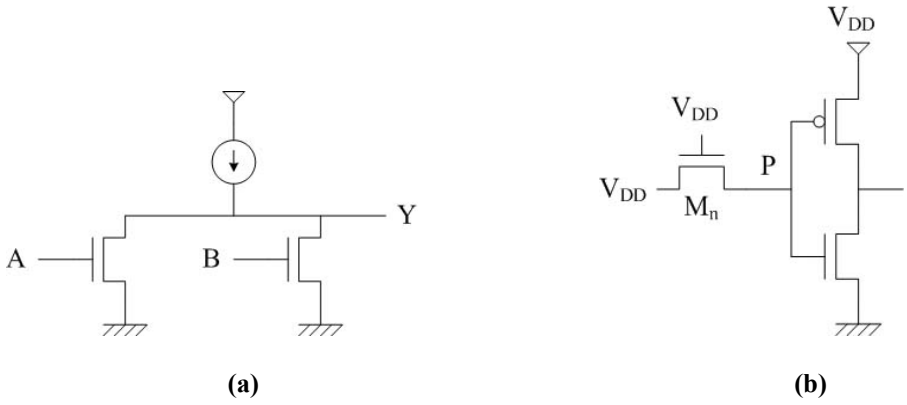
Figure 7 Simulations Eldo de l'évolution du courant de fuite d'un transistor en fonction (a) de la polarisation du substrat et (b) de la tension drain-source.

2.2.4 Courant statique

Les exemples de ce type de courant peuvent être variés. Les circuits qui dissipent ce type de puissance contiennent un chemin de conduction de l'alimentation à la masse qui peut conduire du courant même lorsque le circuit est dans un état stable, non transitoire. Un exemple de ce type de circuit est le style logique NMOS ou pseudo-NMOS, dans lequel respectivement une résistance ou un transistor PMOS toujours passant servent à fournir le niveau logique '1' en sortie de la porte. La Figure 8-a présente une porte NOR à deux entrées de style NMOS. Cette porte génère un courant statique lorsque ses entrées *A* ou *B* sont au niveau logique '1', et cela même en l'absence d'activité. Puisque la vitesse de la porte dépend de la magnitude du courant fournie par la source de courant, ce style logique se révèle très consommateur d'énergie. Cela explique en grande partie pourquoi la technologie CMOS a été très largement adoptée.

Un autre exemple est la présence de niveaux logiques dégradés en sortie de portes à transistors de passage qui induisent une consommation statique dans l'étage suivant, comme montré Figure 8-b. Le transistor M_n ne passant pas correctement le niveau logique '1', le nœud interne *P* voit son potentiel s'élever à seulement $V_{DD} - V_{TN}$. C'est suffisant pour que la sortie de l'inverseur soit à 0V, mais cela induit un courant de court-circuit statique puisque le transistor PMOS de l'inverseur n'est pas correctement bloqué. La consommation statique peut représenter une part importante de la consommation totale si la porte est longtemps au repos et dans ce cas, un restaurateur de niveau faible devant le transistor Mn doit être introduit pour amener le potentiel du nœud *P* à V_{DD}.

Le courant statique étant évité dans les blocs logiques, c'est bien souvent un courant de polarisation présent dans les blocs analogiques.



(a) **(b)**

Figure 8 (a) Porte NOR à deux entrées en logique NMOS; (b) Transistor de passage présentant un niveau logique dégradé en entrée d'un inverseur.

Dans le chapitre suivant, nous allons présenter les différentes méthodes existantes pour réduire la consommation d'énergie des circuits microélectroniques.

2.3 Techniques de réduction de la puissance dissipée

Nous l'avons vu précédemment, la puissance dissipée par les microprocesseurs augmente de façon exponentielle et doit être limitée pour des raisons de coût, d'autonomie ou encore de fiabilité. La puissance dynamique est encore aujourd'hui la source principale de dissipation d'énergie des circuits microélectroniques. Les différents paramètres qui entrent dans sa formulation a, f, C_L, V_{DD}, sont autant de pistes pour réduire la puissance consommée : c'est l'objet de la première partie. Avec les technologies fortement sous-microniques, la puissance statique due aux courants de fuite devient une source de plus en plus importante de consommation : les différentes techniques existantes pour réduire les courants de conduction sous le seuil sont détaillées dans la deuxième partie.

2.3.1 Technique de réduction de la puissance dynamique

Nous n'allons considérer ici que la puissance de commutation puisque la puissance de court-circuit peut, comme indiqué plus haut, être limitée à 10% de la consommation dynamique totale par un dimensionnement correct des transistors. D'après l'Équation 1, la puissance de commutation dépend de l'activité du circuit – équivalente à la probabilité de transition –, de la fréquence de fonctionnement, de la valeur de la charge commutée et de la tension d'alimentation. Pour réduire la puissance dissipée, on peut réduire chaque terme séparément sans affecter les autres ou bien trouver un compromis entre les différents paramètres.

Réduction de l'activité

L'activité peut être réduite à différents niveaux, du plus élevé au plus bas : niveau système, niveau architecture, niveau logique ou RTL, niveau circuit et niveau technologie. D'une manière générale, la puissance dissipée peut être diminuée à tous les niveaux d'abstraction précédents, les gains les plus importants étant réalisés aux niveaux les plus élevés. Ce qui nous intéresse ici sont les niveaux circuit et architecture. On peut néanmoins brièvement citer au niveau système l'utilisation de logiciels basse consommation, qui réduisent l'activité en optimisant notamment les boucles avec des techniques telles que le fusionnement de boucles

ou le déroulement de boucles. De la même manière, on peut utiliser différents codes pour représenter les données et choisir celui qui présente le moins d'activité. La réduction du nombre d'opérations permet des gains très intéressants au niveau de l'énergie consommée, au-delà même de la simple réduction d'activité. En effet, si l'on réduit le nombre d'opérations à effectuer d'un facteur *p*, on peut alors réduire la tension d'alimentation globalement d'un facteur *p*, tout en gardant les mêmes performances, puisque le temps de propagation d'un circuit CMOS est approximativement égal à [Saku90] :

$$T = \beta C_L \frac{V_{DD}}{(V_{DD} - V_{TH})^{\alpha}}$$

Équation 5

La puissance dynamique étant proportionnelle à V_{DD}^2, cela permet une réduction d'énergie d'un facteur p^3.

Au niveau circuit, la technique consiste à réduire l'activité inutile. Considérons tout d'abord les transitions parasites – *glitch*. Cela se produit lorsque les entrées d'une porte logique n'arrivent pas toutes au même moment : la sortie de la porte va alors présenter des variations avant d'arriver dans son état stable. Pour illustrer le phénomène, la Figure 9-a présente une structure cascadée de deux portes Nand : lorsque les entrées ABC passent de 100 à 111, le délai introduit par la première porte Nand provoque une fausse transition de la deuxième porte. En effet, le nœud interne Int varie après l'entrée C et le produit $C \times Int$ doit de nouveau être évalué par la deuxième porte. Pour réduire les fausses transitions, il faut égaliser les délais des différents chemins, notamment pour les nœuds fortement chargés, si possible en introduisant des inverseurs. Il faut de préférence éviter les structures cascadées, qui introduisent différents retards : par exemple, la fonction $((A \times B) \times C) \times D$ peut s'écrire $(A \times B) \times (C \times D)$, formulation qui possède l'avantage d'avoir des délais équilibrés.

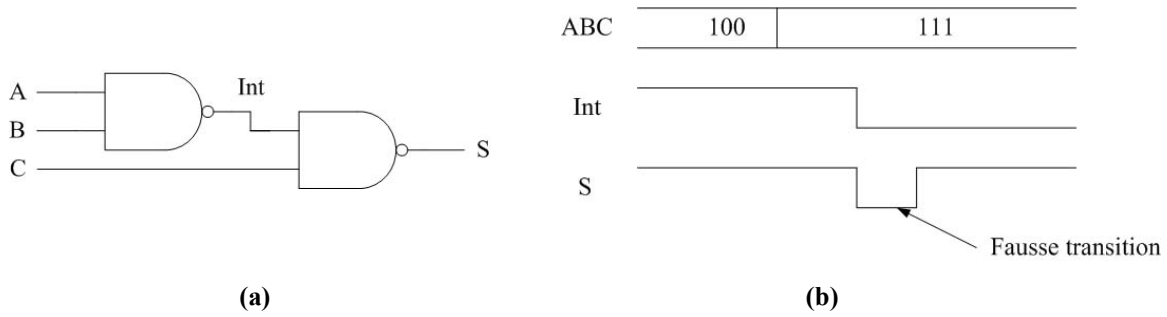


Figure 9 Illustration des fausses transitions dans les portes logiques.

Au niveau architecture, la propagation des transitions parasites peut être évitée en pipelinant le chemin de données. Un pipeline est réalisé en segmentant les unités fonctionnelles en plusieurs étages et en insérant des latchs ou des bascules entre les étages. Un exemple de pipeline à l'aide de bascules est donné Figure 10.

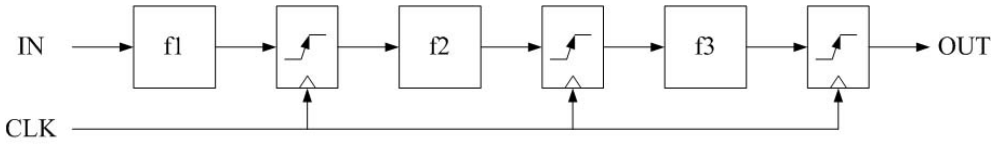


Figure 10 Pipeline d'un chemin de données à l'aide de bascules.

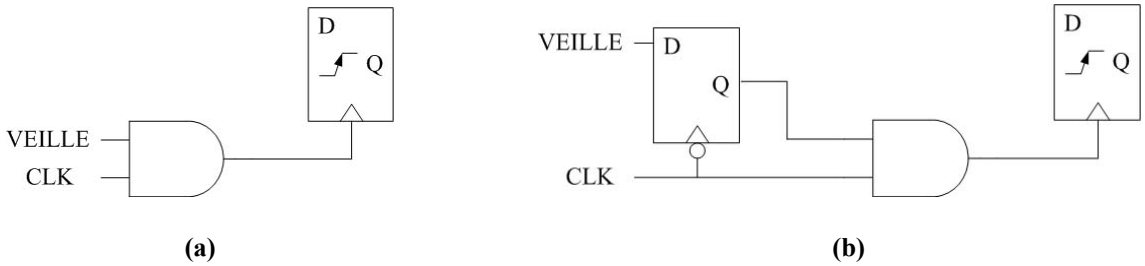
Une telle architecture n'est néanmoins pas exempte d'inconvénients. On peut relever :

- une surface accrue due à la présence des bascules,
- une consommation dynamique plus élevée dans l'arbre d'horloge,
- la latence introduite par la profondeur du pipeline.

Isolation d'horloge

Dans les systèmes synchrones, qui représentent la quasi-totalité des circuits actuels, la technique pour réduire l'activité inutile, en dehors des transitions parasites, consiste à stopper localement les horloges des blocs qui sont en veille ou qui doivent seulement maintenir une information : cette technique est appelée *clock gating* ou *gated clock* en anglais [Beni94], [Gunt00]. Activer sélectivement des horloges locales est très important pour économiser l'énergie, notamment dans les processeurs à usage général, car les différents systèmes qui les composent ne sont activés qu'une fraction du temps.

Le principe de l'isolation d'horloge est d'associer à chaque élément séquentiel un bloc de contrôle CG qui inhibe le signal d'horloge lorsque la condition VEILLE est vraie. Le moyen le plus simple pour l'implémenter dans un circuit à base de bascules est montré Figure 11-a : il s'agit d'une simple porte AND. Cette implémentation est cependant sensible aux fausses transitions présentes sur le signal de veille. Pour éviter cela, un latch est introduit pour mémoriser la valeur du signal de VEILLE, tel que montré sur la Figure 11-b. Lorsque l'horloge est à l'état haut, les fausses transitions sont filtrées par le latch ; lorsque l'horloge est à l'état bas, elles sont filtrées par la porte AND.



(a) **(b)**

Figure 11 Implémentation du bloc de contrôle CG : (a) circuit le plus simple; (b) circuit permettant de filtrer les fausses transitions.

La technique de l'isolation d'horloge s'avère très efficace dans le cas où des unités fonctionnelles importantes ne sont pas utilisées pendant une fraction significative du temps de calcul. [Hash02] a montré une réduction de 70% de l'énergie dissipée par un décodeur vidéo MPEG4 en utilisant massivement cette technique : 90% des bascules sont isolées par un bloc CG.

Réduction de la capacité commutée

La réduction de la capacité commutée peut se faire à tous les niveaux d'abstraction. Au niveau architecture, celle-ci est obtenue en augmentant la localité des données. Un processeur consomme une partie de son énergie à lire ou à écrire des instructions et des données et donc à les déplacer de la mémoire vers les unités de calcul et inversement. Si l'information peut être stockée plus près de l'endroit où elle est nécessaire, l'énergie dépensée à la transférer est réduite grâce à des longueurs de bus plus courtes. La localité des données est également utilisée pour améliorer les performances d'un processeur [Henn96] : cette technique n'est efficace que si les données présentent une localité temporelle suffisante, c'est-à-dire si un ensemble de données est utilisé fréquemment.

Au niveau circuit, on peut réduire la capacité commutée en redimensionnant les transistors. La largeur des transistors dans les portes logiques est adaptée à la sortance. La charge d'une porte CMOS est composée de trois paramètres :

- la capacité de grille de l'étage suivant,
- la capacité parasite de sortie, due aux diffusions de drain des transistors,
- et la capacité parasite de l'interconnexion.

Les deux premiers paramètres sont proportionnels à la largeur des transistors. Au premier ordre, le délai d'une porte CMOS peut donc s'écrire :

$$Délai = (C_{grille} + C_{diffusion} + C_{interconnexion}) / I$$

$$\propto (C_{grille} + C_{diffusion}) / largeur + C_{interconnexion} / largeur$$

Équation 6

Si la charge en sortie dépend peu de la capacité de l'interconnexion, la réduction de la largeur des transistors aura un effet bénéfique sur la consommation sans trop affecter le délai, d'après l'Équation 6.

Au niveau technologie, les capacités peuvent être réduites par des géométries plus fines, des capacités de jonction plus petites et des interconnexions optimisées. D'une génération technologique à la suivante, des gains importants peuvent ainsi être réalisés. Le Tableau 1 montre les effets du dimensionnement de la technologie sur la puissance consommée par un processeur à usage général, le PowerPC 604 [Stor94]. A fréquence et tension d'alimentation égales, le PowerPC 604 consomme 37% de moins en technologie 0,15µm par rapport à la technologie 0,25µm, ce qui indique que les capacités commutées ont été au moins réduites du même ordre de grandeur. En effet, la puissance due aux courants de court-circuit peut être considérée inférieure à 10% de la puissance dynamique dans les deux cas, si les transistors ont été correctement dimensionnés, et les courants de fuite sont supérieurs dans la technologie 0,15µm par rapport à la technologie 0.25µm, du fait de tensions de seuil inférieures.

Tableau 1 Diminution de la puissance dissipée par le PowerPC 604 par la réduction de la géométrie de la technologie.

L (µm)	0,25	0,15
L_{eff} (µm)	0,15	0,10
V_{DD} (V)	1,8	1,5
Surface (mm²)	20	7,5
Horloge (MHz)	225	330
Puissance (W)	2,35	1,5
@100MHz et V_{DD}=1.5V		
Puissance (W)	0,72	0,45

La réduction des paramètres géométriques des technologies permet également de réaliser un gain d'échelle : l'augmentation des densités d'intégration autorise la réalisation de systèmes sur puce, en intégrant des circuits auparavant montés en composants discrets. Le nombre total de puces dans le produit final est réduit, et par là même, sa consommation d'énergie.

Réduction de la tension d'alimentation

La réduction de la tension d'alimentation offre un gain quadratique de la puissance dissipée. C'est donc un moyen très efficace de réduire la consommation de courant, qui néanmoins impacte défavorablement les performances des portes : le temps de propagation d'une porte en fonction de la tension d'alimentation a été donné dans l'Équation 5. Pour compenser l'accroissement du délai, la tension de seuil des transistors est diminuée, ce qui a pour conséquence une augmentation de l'énergie dissipée par les courants de fuite. Il faut alors trouver un compromis entre la réduction de la puissance dynamique et l'augmentation de la puissance statique. Les tensions d'alimentation et de seuil optimales sont obtenues lorsque les puissances statique et dynamique sont égales [Hass01] : on obtient alors le point d'énergie minimum. Néanmoins, baisser de manière trop importante la tension de seuil V_T ne va pas sans poser de sérieux problèmes de robustesse : la variation normale de V_T durant le processus de fabrication et avec des paramètres environnementaux, tels que la température, peut alors représenter une part importante de la valeur désirée de V_T [Gonz97]. Cette variation importante de V_T rend délicate l'opération des circuits et peut entraîner une consommation statique importante.

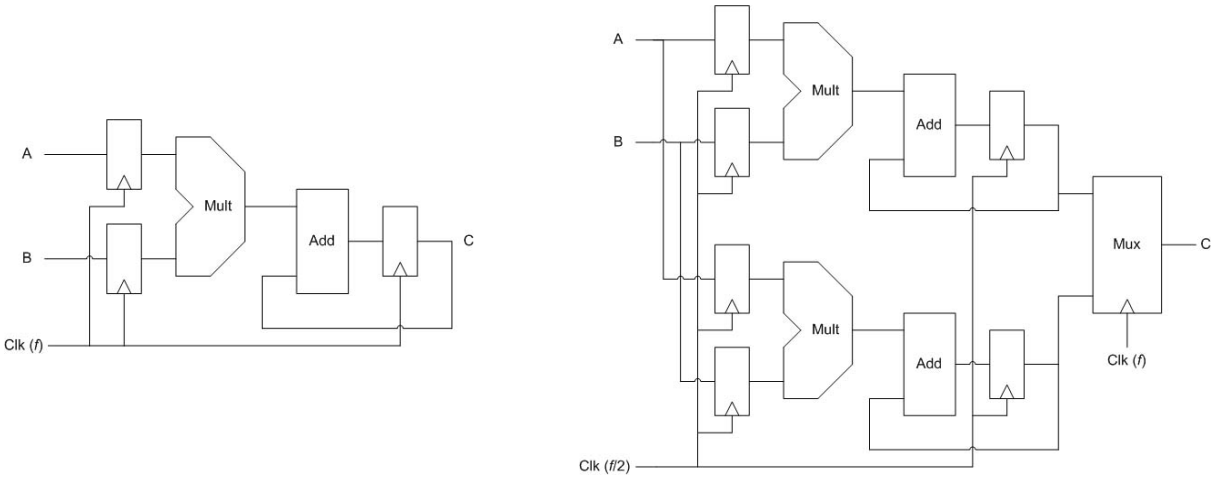
Une autre façon de baisser la tension d'alimentation V_{DD} sans réduire de façon importante la tension de seuil consiste à faire varier dynamiquement la tension V_{DD}, pour l'adapter aux performances requises [Mack90]. La tension d'alimentation est diminuée lorsque la charge demandée au processeur n'est pas très élevée et augmentée lorsque la charge est importante. De la même façon, la fréquence de fonctionnement du circuit est adaptée aux besoins : en fonction de la charge demandée, une fréquence d'horloge est choisie puis la tension V_{DD} minimale est générée. Des exemples de ce type de processeur se trouvent dans [Tran04] et [Inte04].

Réduction de la fréquence d'horloge

La réduction de la fréquence de fonctionnement offre une réduction linéaire de la puissance dissipée d'après l'Équation 1. Cependant, baisser la fréquence de fonctionnement n'a aucun

effet sur l'énergie consommée mais diminue uniquement la puissance dissipée. La baisse de la fréquence d'horloge est bien souvent la conséquence d'une baisse de la tension d'alimentation, qui comme nous l'avons vu précédemment, s'accompagne d'une diminution des performances. Il y a alors deux architectures possibles pour compenser la diminution du flot de données – *throughput* : la mise en parallèle des unités de calcul et le pipeline.

La parallélisation consiste à faire travailler simultanément deux circuits plus lents. La Figure 12 montre une application de la parallélisation à un chemin de données réalisant une opération de multiplication-accumulation. Le schéma de gauche représente le chemin de données simple, avec pour fonction : $C = C + A \cdot B$. Dans le schéma de droite, une fréquence *f*/2 est utilisée pour les registres d'entrée des multiplicateurs. Le flot de données à la fréquence *f* est récupéré après le multiplexeur.



(a) **(b)**

Figure 12 (a) Chemin de données simple ; (b) architecture parallèle permettant de compenser une perte de performances due à une diminution de la tension d'alimentation.

Le pipeline, présenté Figure 10, peut être vu également comme une mise en parallèle des unités de calcul où, au lieu de multiplier le nombre de circuits, les unités sont partitionnées dans le temps, permettant à plusieurs calculs d'avoir lieu simultanément sur la même ressource matérielle.

2.3.2 Techniques de réduction des courants de fuite

Traditionnellement, la tension d'alimentation V_{DD} est réduite d'une génération technologique à l'autre pour des raisons de mise à l'échelle de différents paramètres géométriques et électriques, ceci afin de respecter la tension de claquage de l'oxyde de grille. Pour garder de bonnes performances, la tension de seuil V_T est également diminuée, un

rapport V_{DD}/V_T sensiblement égal à quatre étant conservé. Les courants de fuite ou courants de conduction sous le seuil sont ainsi fortement augmentés, comme indiqué par l'Équation 3. La Figure 6 montre une pente sous le seuil du courant de drain de 80mV par décade, ce qui signifie qu'une variation de 80mV de la tension de seuil va entraîner une variation des courants de fuite d'un ordre de grandeur. Avec la réduction technologique, ces courants ne pourront plus être négligés. Différentes techniques pour les réduire ont donc été proposées :

VTCMOS (Variable Threshold CMOS)

Cette technique consiste à faire varier dynamiquement la tension de seuil de tous les transistors en contrôlant le potentiel du substrat face arrière [Kuro96]. Lorsqu'une partie du circuit est au repos, une polarisation négative est utilisée pour rendre la tension substrat-source négative et augmenter la tension de seuil. A l'inverse, lorsque le circuit est actif, la polarisation redevient positive. L'inconvénient de cette méthode est de devoir générer la tension de contrôle, dont l'amplitude de variation se situe hors de la plage $[0,V_{DD}]$. De plus, si cette technique est facilement utilisable dans une technologie CMOS massif triple well, elle n'est pas exploitable en technologie SOI, où il faudrait contrôler les substrats flottants de tous les transistors.

SRBB (Selective Reverse Body Biasing of PD-SOI transistors)

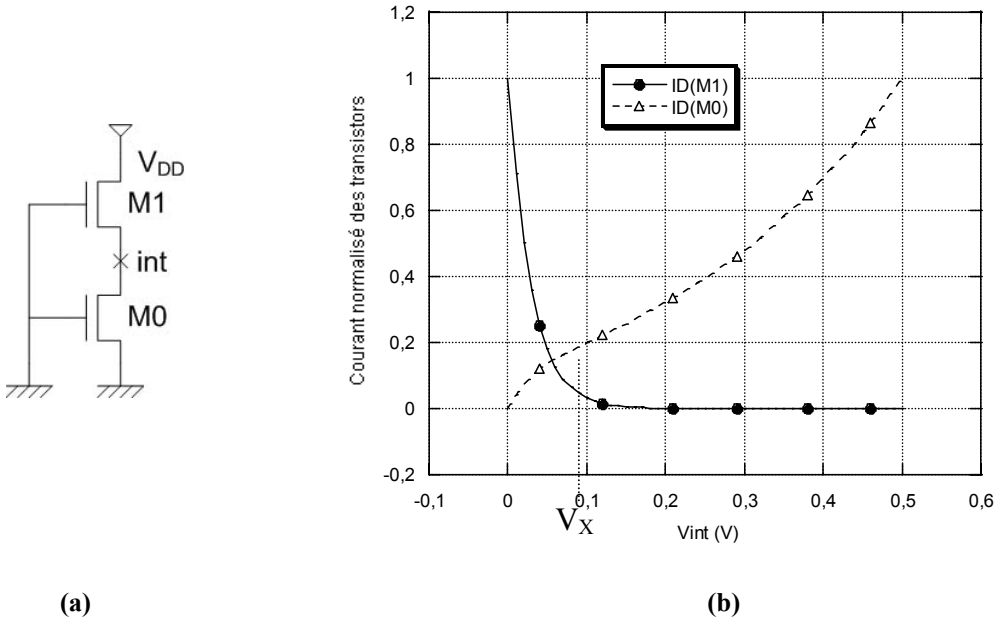
Cette méthode est similaire à la précédente et est adaptée à la technologie SOI [Casu01]. L'idée de départ est la suivante : pour diminuer la consommation statique d'un circuit, il n'est pas nécessaire de contrôler la tension de seuil de tous les transistors. Pour savoir quels transistors doivent être contrôlés, il faut étudier les courants de fuite de chaque cellule en fonction de la valeur des entrées. L'inconvénient de ce schéma est qu'il faut connaître à l'avance quelles sont les valeurs des entrées lorsque le circuit est au repos. Il est également affecté par les mêmes limitations que la technique précédente, à savoir qu'il faut générer une tension négative et que le nombre de contacts nécessaires peut être élevé.

Empilement de transistors

Cette technique tire partie du fait que deux transistors bloqués, en série, vont avoir un courant de conduction sous le seuil bien plus faible qu'un seul transistor bloqué [Nare01]. Un transistor est donc dédoublé et les deux sont placés en série – voir Figure 13-a. La Figure 13-b montre les courants de fuite des deux transistors en fonction du potentiel du nœud interne *int.* La réduction des courants de fuite est due à l'effet source suiveur : le potentiel de la source du transistor M1 augmente ce qui diminue exponentiellement son courant de drain. Dans le

même temps, la tension V_{DS} du transistor M0 augmente, ainsi que le courant de fuite. La tension d'équilibre V_X est atteinte lorsque les deux courants précédents s'égalisent, comme indiqué Figure 13-b.

Cette approche permet de réduire modérément les courants de fuite en utilisant une technologie standard mais présente l'inconvénient de nécessiter de nouveaux outils de synthèse, à même d'identifier les transistors pouvant être dédoublés sans entraîner de pénalité en délai.



(a) **(b)**

Figure 13 (a) Empilement de deux transistors, (b) Courants de fuite des transistors M0 et M1 en fonction du potentiel du noeud interne.

MTCMOS (Multiple Threshold CMOS)

Dans la méthode précédente, les transistors placés en série pour gagner en empilement sont toujours commandés par les entrées des portes considérées. Dans le schéma MTCMOS, le transistor placé en série est commandé par un signal de veille. Il s'agit d'un transistor à haut V_T, pour autoriser de faibles courants de fuite lorsqu'il est bloqué, alors que les transistors constituant la porte sont à bas V_T [Muto95]. La technique MTCMOS nécessite un dimensionnement correct de ce transistor de veille : il doit être suffisamment gros pour ne pas dégrader les performances de la porte mais pas trop pour des raisons de surface et de consommation dynamique lors de la commutation entre modes. Néanmoins, il ralentira considérablement la porte lorsque la tension d'alimentation V_{DD} est faible.

Mélange de techniques

Ces différentes techniques peuvent être combinées. Par exemple, dans [Das03], les auteurs utilisent deux transistors de veille en série, c'est-à-dire à la fois la technique MTCMOS et l'empilement de transistors. L'avantage cité est une réduction encore plus importante des courants de fuite, associée à très peu de pénalités au niveau des performances.

2.4 Les circuits très basse tension

L'utilisation d'une tension d'alimentation très basse, de l'ordre de quelques centaines de millivolts, permet de réaliser des circuits dissipant très peu d'énergie. Cette approche a émergé au début des années 1990. Des circuits UBT – Ultra Basse Tension – ont été développés dans la technologie CMOS à substrat massif et le SOI. Nous allons voir que les techniques employées sont différentes dans les deux cas.

2.4.1 L'UBT en technologie CMOS à substrat massif

La technologie CMOS UBT a été développée à Stanford en 1990 [Burr91]. Pour garder de bonnes performances à des tensions d'alimentation très basses, les tensions de seuil V_T des transistors sont considérablement réduites, devenant même négatives. Il faut alors contrôler dynamiquement le potentiel de substrat des transistors afin d'ajuster leur tension de seuil et ceci pour deux raisons : d'une part, afin de maîtriser les variations de V_T dues au procédé de fabrication et à la température, car compte tenu de la très faible valeur de la tension de seuil, les performances sont fortement dépendantes de cette dernière, et d'autre part, afin de diminuer les courants de fuite [Burr94]. La Figure 14 montre une vue simplifiée de transistors NMOS et PMOS et les prises caisson ajoutées pour contrôler leur tension de seuil en polarisant le substrat.

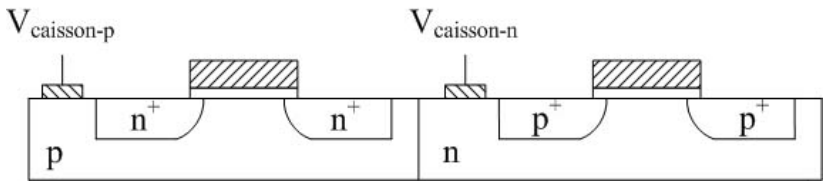


Figure 14 Schéma de transistors NMOS et PMOS montrant les prises des caissons P et N nécessaires pour ajuster les tensions de seuil V_{TN} et V_{TP}.

L'inconvénient de cette technique est de devoir distribuer des lignes d'alimentation supplémentaires, étant donné que les potentiels de substrat sont différents des valeurs de l'alimentation. La surface occupée par le circuit est alors augmentée à cause du placement moins efficace des cellules standards lors de la synthèse.

Cependant, grâce à cette technologie, les circuits CMOS peuvent fonctionner jusqu'à des tensions d'alimentation très basses : Burr et Shott donnent l'exemple d'un circuit d'encodage/décodage opérant sans erreur à 125mV [Burr94]. Les principales caractéristiques de la technologie UBT sont résumées dans le Tableau 2.

Tableau 2 Caractéristiques de la technologie UBT de Stanford, CMOS 0.5µm [Burr94].

	NMOS	PMOS
V_{T0}	-120mV	20mV
V_{BS}	[0 : -9] V	[0 : 4,5] V

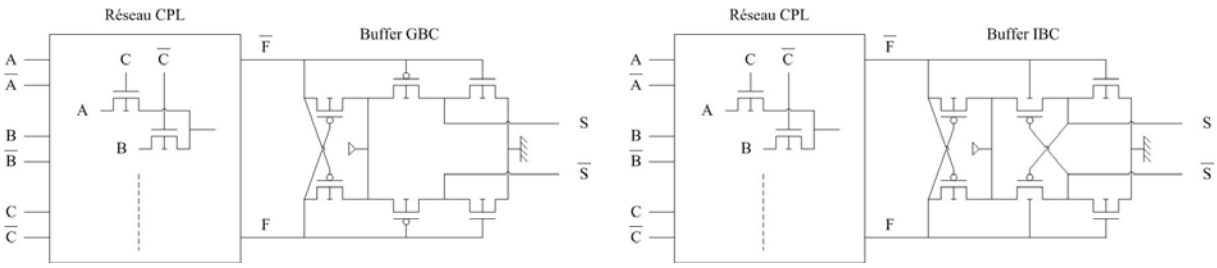
Cette technique consistant à utiliser des transistors ayant une tension de seuil proche de 0V a été développée par la suite : on peut notamment citer [Hass01]. Néanmoins, dans les technologies fortement sous-microniques actuelles, les variations de la longueur de grille sont très importantes et le faible niveau de dopage employé dans cette approche ne permet plus de compenser leurs effets en polarisant le substrat : cet inconvénient majeur a été démontré par Heer et Al. pour la technologie 0,13µm avec des tensions de seuil de 150mV [Heer04]. Cette approche n'est donc plus viable.

2.4.2 L'UBT en technologie SOI

La particularité de la technologie SOI étant que le substrat de chaque transistor est isolé, il n'est plus possible de polariser simplement toutes les zones de silicium actives des transistors de manière à remonter leur tension de seuil V_T (nous allons présenter plus en détail cette technologie dans le chapitre suivant). Les circuits UBT en SOI n'utilisent donc pas de transistors à V_T nul, mais exploitent à la place une structure particulière du SOI, le transistor DTMOS. L'idée, contrairement au CMOS à substrat massif, n'est donc pas de relever la tension de seuil, mais de l'abaisser dynamiquement par le potentiel présent sur la grille.

Dans [Fuse96], les auteurs réalisent leurs circuits en logique CPL et utilisent exclusivement des transistors DTMOS. L'avantage cité est que les niveaux logiques en sortie

des arbres NMOS sont moins dégradés puisque la tension de seuil des transistors DTMOS est basse lorsqu'ils sont passants : elle est égale à 0,17V pour une polarisation de substrat de 0,5V alors qu'elle vaut 0,4V pour une polarisation de substrat nulle. Les auteurs proposent deux schémas pour la bufferisation en sortie des arbres, appelés GBC – *Gate-Body Connected* – et IBC – *Input-Body Connected*. Ces deux structures sont présentées Figure 15.



(a) **(b)**

Figure 15 (a) Schéma GBC; (b) schéma IBC.

Le schéma GBC est le schéma classique de la bufferisation de la logique CPL : il est constitué de deux inverseurs et d'un latch de type P. Le latch permet de restaurer le niveau logique '1' tandis que les inverseurs assurent le découplage entrée/sortie. Le schéma IBC permet d'améliorer plusieurs points en utilisant une structure différentielle au niveau des buffers de sortie. Premièrement, la taille du réseau CPL peut être réduite puisque celui-ci n'attaque plus que deux transistors PMOS au lieu de deux transistors PMOS et deux transistors NMOS. Les auteurs en déduisent donc qu'ils peuvent diviser par deux la taille des transistors du réseau CPL : ils ne semblent pas avoir pris en compte la capacité introduite par les contacts de substrat, qui est loin d'être négligeable. Deuxièmement, la vitesse de basculement de la sortie est améliorée puisque la tension de seuil du transistor PMOS décroît avant même que la sortie ne bouge.

Dans [Dous96], les auteurs utilisent une combinaison de transistor à bas V_T et body flottant et de transistors à haut V_T montés en DTMOS en technologie SOI totalement désertée. Le transistor DTMOS sert uniquement à réduire la consommation statique due aux courants de fuite en coupant l'alimentation. Un schéma simplifié est donné Figure 16

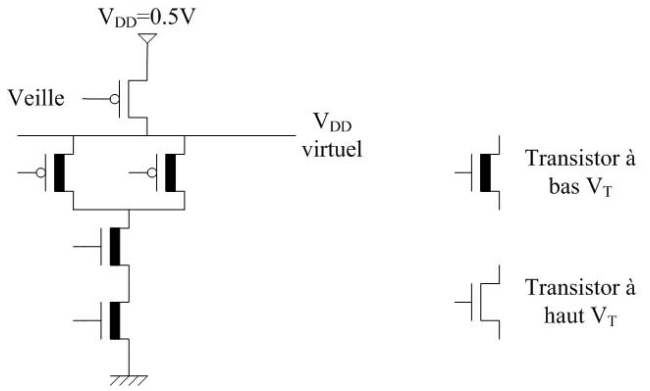


Figure 16 Schéma du circuit MTCMOS utilisant des transistors à body flottant et des transistors DTMOS.

Dans [Hiro98], les auteurs réalisent un multiplexeur/démultiplexeur 8 bits en utilisant une matrice de transistors DTMOS pour faciliter la conception et la migration technologique. A la tension d'alimentation de 0,5V, les résultats obtenus sont :

- 320MHz et 2mW pour le multiplexeur,
- 380MHz et 1,4mW pour le démultiplexeur.

La structure de base de la matrice de portes est représentée Figure 17-a et un exemple de réalisation de porte Nand est donné Figure 17-b. Comme on peut le constater, la place occupée par le contact du body est réduite et la forme de la grille n'est pas tellement différente de ce qu'elle serait pour un transistor à substrat flottant. La capacité parasite introduite par le contact devait être limitée dans cette technologie. Les auteurs indiquent que la surface occupée par cette matrice de transistors DTMOS est à peine supérieure à celle occupée par une matrice de transistors à body fixe.

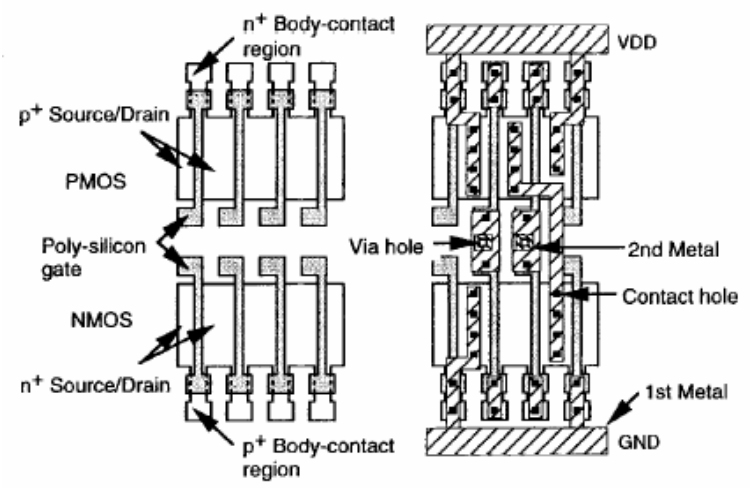


Figure 17 (a) Structure de matrice de portes de base; (b) Exemple de réalisation de porte Nand.

Dans [Soel00], les auteurs introduisent ce qu'ils appellent la logique DTMOS sous le seuil, qui revient en fait à n'utiliser que des transistors DTMOS comme dans [Dous96]. Les avantages sont un meilleur produit puissance*délai et une moindre variabilité du délai en fonction des paramètres environnementaux.

2.5 Conclusion

La consommation de puissance des circuits microélectroniques est devenu un problème de premier ordre pour des raisons de fiabilité, de portabilité et de dissipation de chaleur. Après avoir détaillé les différentes sources de dissipation d'énergie et identifié les deux principales, à savoir la puissance dynamique et la puissance statique des courants de fuite, nous avons explicité de manière exhaustive les techniques permettant de les réduire. En ce sens, l'utilisation de circuits très basse tension est intéressante, lorsque les performances demandées ne sont pas élevées. En silicium massif, l'idée retenue est d'utiliser des transistors ayant des tensions de seuil proches de zéro, afin de compenser la baisse de performance. Cependant, cette approche technologique n'est plus viable pour des technologies fortement sous-microniques car elle ne permet plus de compenser les variations de la longueur de grille. En technologie SOI (technologie que nous allons présenter dans la suite), l'approche retenue est d'utiliser des tensions de seuil standards, puisque qu'il n'est pas possible de contacter les substrats de tous les transistors. Les auteurs présentent la logique CPL comme un style basse consommation, de part sa faible capacité d'entrée. Ils n'hésitent pas à faire un usage intensif du transistor DTMOS, celui-ci augmentant grandement les performances au détriment de la consommation dynamique. Notre choix va se porter sur cette deuxième approche, à savoir l'utilisation d'une technologie SOI standard, en faisant une utilisation raisonnée des transistors DTMOS, après avoir étudié de quelle manière en tirer avantage, puisque leur utilisation exclusive ne se justifie évidemment pas de part leur capacité de grille importante.

3 La technologie SOI

Les avancées technologiques ont par le passé été guidées par des contraintes de délai et de surface et ont accessoirement contribué à réduire la puissance consommée par les circuits intégrés. Un facteur α est utilisé pour la mise à l'échelle des différents paramètres géométriques ou électriques lors du passage d'une génération technologique à l'autre, tels que l'épaisseur de grille, la longueur effective du transistor, la largeur des pistes d'interconnexion, le dopage, la tension d'alimentation, … Ce terme α, d'une valeur comprise entre 1,25 et 1,33, permet une augmentation des performances et de la densité respectivement d'un facteur α et α^2 et une diminution de la puissance dissipée d'un facteur $1/\alpha^2$. Les diminutions successives du motif minimal des procédés lithographiques réduisent les capacités des transistors, et donc la valeur des capacités à charger et à décharger. Aujourd'hui, l'amélioration de la consommation est un objectif à part entière, comme nous l'avons vu au Chapitre 2. Des technologies spécifiquement basse consommation doivent être développées : la technologie SOI – « *Silicon On Insulator* » ou « silicium sur isolant » – est, à cet aspect, très intéressante.

3.1 Le transistor MOS SOI

Comme son nom l'indique, la technologie SOI se caractérise par la présence d'un oxyde enterré. En fonction de l'épaisseur de la couche de silicium actif présente au dessus, dans laquelle les zones de diffusion sont implantées, deux technologies se distinguent : celle appelée « totalement désertée » pour une mince couche de silicium, de l'ordre de 20nm, et celle appelée « partiellement désertée » pour une couche plus épaisse, de l'ordre de 100nm [Oklo02]. Bien que la technologie SOI totalement désertée présente de meilleures caractéristiques I-V, notamment une meilleure pente sous le seuil, elle a l'inconvénient de nécessiter un contrôle très précis de l'épaisseur de la couche de silicium : celle-ci étant totalement désertée, le potentiel sous la grille et donc la tension de seuil sont directement fonction de son épaisseur. Cela implique une contrainte importante sur le procédé de fabrication, qui ne peut pas être aujourd'hui respectée avec un rendement suffisant. De plus, l'épaisseur de silicium actif doit être mise à l'échelle d'une génération technologique à l'autre et les coûts associés au développement de cette technologie ne peuvent dès lors pas être amortis sur plusieurs générations. Enfin, la mince couche de silicium induit des résistances d'accès élevées au niveau du drain et de la source. Ces trois raisons expliquent pourquoi la technologie SOI totalement désertée n'est pas utilisée commercialement.

La technologie SOI partiellement désertée, à l'inverse, ne nécessite pas de couche très mince de silicium, ce qui réduit ses coûts de fabrication : c'est une technologie aujourd'hui mature, utilisée notamment par IBM [Reed04]. La structure d'un transistor MOS SOI est donnée Figure 18. Du fait de l'épaisseur de la couche de silicium superficiel, il existe une zone de silicium qui ne peut être totalement désertée de porteurs mobiles par la grille : cette région est appelée body du transistor, c'est le quatrième terminal du transistor MOS à effet de champ. Il peut être laissé flottant ou connecté à une tension extérieure, son potentiel déterminant la tension de seuil du transistor, comme montré dans l'Équation 4. Lorsque le body est laissé flottant, son potentiel est déterminé par différents phénomènes de charge et de décharge, qui sont :

- le courant d'ionisation par impact,
- le courant de fuite des jonctions PN polarisées en inverse,
- les couplages capacitifs,
- le courant de jonction polarisée en direct,
- la génération thermique de paires électron trou,

- le phénomène de recombinaison.

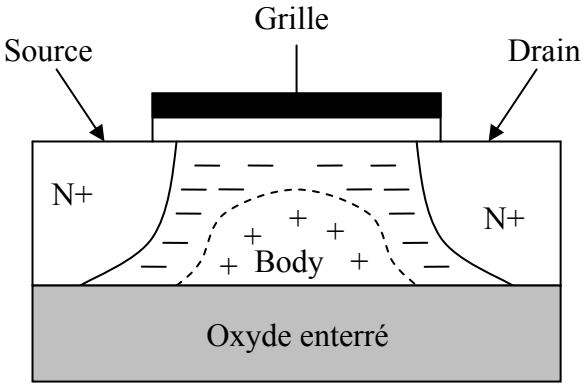


Figure 18 Structure d'un transistor NMOS SOI.

Le courant d'ionisation par impact apparaît aux fortes polarisations de drain. C'est un phénomène qui se produit dans la zone de pincement du canal : les porteurs minoritaires ayant atteint leur vitesse de saturation – des électrons pour le cas d'un transistor NMOS – sont attirés par le fort champ électrique du drain et frappent les atomes de silicium près de l'oxyde de grille, créant des paires électron trou. Parmi celles qui ne se recombinent pas, les électrons sont évacués par le drain et les trous se dirigent vers le body, zone de moindre potentiel, augmentant ainsi la tension V_{BS}.

Le courant de fuite de la jonction PN est le courant qui passe à travers la diode polarisée en inverse. Il provient de trois mécanismes :

- la recombinaison de paires électron trou dans la zone de charge d'espace,
- la déformation du gradient de dopage par des défauts ou impuretés, passant d'un type N à un type P,
- le franchissement de la barrière de potentiel par des porteurs possédant suffisamment d'énergie.

Il a pour expression l'équation classique du courant d'une diode :

$$I = I_0 \cdot \left(e^{qV/kT} - 1 \right),$$

Équation 7

avec I_0 le courant de génération de la diode, V la tension aux bornes de la jonction et T la température. Le courant de fuite dépend exponentiellement de la tension aux bornes de la jonction et de la température.

En transitoire, le body est couplé de façon capacitive à la grille et au drain. Le couplage de la grille s'effectue par la capacité d'oxyde de silicium et la capacité de déplétion et a lieu tant que le canal n'est pas créé : lorsque ce dernier existe, il n'y a plus de couplage grille body car toute variation de quantité de charge sur la grille est compensée par une variation de quantité de charge dans le canal d'inversion, et non plus dans le body. Le couplage avec le drain s'effectue par la capacité de zone de charge d'espace. Sa valeur est fonction de la polarisation de la jonction puisque la largeur de la zone désertée fixe l'épaisseur de la capacité. Le couplage avec le drain à tension nominale est plus important que celui avec la grille, comme montré Figure 19. Cette simulation a été réalisée sur un inverseur à l'aide du modèle SOI 0,13µm de STMicroelectronics.

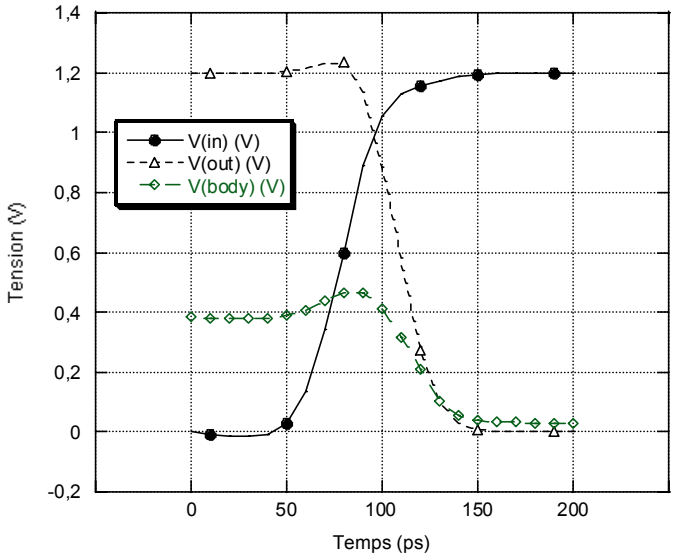


Figure 19 Couplage capacitif du body d'un transistor NMOS en fonction des variations des tensions de grille et de drain.

La polarisation de la jonction PN en direct se produit lorsque la tension à ses bornes dépasse 0,6V : c'est généralement la jonction body source qui est concernée. La diode peut être rendue passante soit par le courant d'ionisation par impact qui charge le body, soit par couplage capacitif. Il y a alors déclenchement du bipolaire parasite horizontal, NPN pour un transistor NMOS et PNP pour un transistor PMOS.

Pour résumer, il existe trois chemins de charge et deux chemins de décharge du body. Les mécanismes de charge sont les courants de fuite des jonctions polarisées en inverse, phénomène qui est de l'ordre de la milliseconde ; à tension élevée, l'ionisation par impact

ajoute également des charges dans le body. Les deux chemins de décharge sont les jonctions polarisées en direct, phénomène qui dure quelques dizaines de microsecondes. La valeur statique du potentiel du body dépend de l'équilibre entre ces différents phénomènes de charge et de décharge.

3.2 Les avantages du SOI face au silicium à substrat massif

De part la couche d'oxyde enterré qui isole la zone active du transistor, la technologie SOI possède des propriétés intéressantes par rapport au silicium à substrat massif. Globalement, les transistors SOI apportent 25% à 30% de performances en plus par rapport à leurs homologues à substrat massif [Shah99], [Alle99]. Cette amélioration des performances est due à trois paramètres.

3.2.1 Elimination des capacités de jonction

Le principal avantage de la technologie SOI est la réduction des capacités de jonction des transistors. Grâce à la présence de l'oxyde enterré, les diffusions de drain et de source sont limitées à la couche de silicium superficielle : le bas des diffusions touche l'oxyde enterré, qui possède une permittivité relative de 3,8 contre 11,6 pour le silicium. La composante de surface des capacités de jonction est fortement réduite ; seule demeure la composante de périmètre, les capacités de jonction ne dépendant plus que de la capacité latérale body source ou body drain. Elles sont réduites d'un ordre de grandeur par rapport au silicium massif, ce qui diminue la capacité totale d'un circuit de l'ordre de 20% [Shah99], et améliore les performances et la consommation dynamique.

3.2.2 Tension de seuil plus basse

Les tensions de seuil des transistors SOI sont dynamiquement plus basses que celles des transistors à substrat massif grâce à la polarisation positive du substrat : lors d'une transition de la tension de grille, mettant le transistor d'un état bloqué dans un état passant, le couplage capacitif grille body augmente de 10% à 15% le courant de drain.
De plus, l'affaiblissement de la tension de seuil – V_T *roll-off* – associé aux canaux courts s'atténue avec l'augmentation de la polarisation du substrat : cela permet d'abaisser le V_T lors du procédé de fabrication. Ainsi, les performances des transistors à canal long sont améliorées

tout en gardant des courants de fuite et une marge au bruit acceptables pour les transistors à canal court.

3.2.3 Effet source suiveur

En MOS *bulk*, les transistors montés en série présentent un effet de substrat : la chute de tension à travers un transistor va élever le potentiel de la source de celui placé au-dessus – dans le cas de transistors de type N. Puisque le substrat est relié à la masse, la tension V_{BS} de ce transistor devient négative, ce qui augmente sa tension de seuil et diminue les performances de la porte. En SOI, le potentiel du body flottant se situe entre ceux de la source et du drain, donc la tension V_{BS} est positive – hors effet transitoire dû au couplage capacitif : c'est l'effet source suiveur. Les portes à empilement de transistors ne sont pas ralenties en SOI comme elles le sont en CMOS à substrat massif, ce qui permet d'ajouter plus de logique par étage.

Le gain en performances du SOI au niveau transistor ne se retrouve pas intégralement au niveau système. Le délai d'un circuit microélectronique est composé du délai dans les portes et du délai dans les interconnexions, or le SOI n'améliore quasiment pas ce dernier : un très léger gain est obtenu pour le premier niveau de métal, puisque la capacité de couplage métal/substrat est réduite par la présence de l'oxyde enterré. Cependant, les niveaux de métal plus élevés sont isolés du substrat par le métal 1 et le délai dans les interconnexions se situe essentiellement aux niveaux supérieurs. Néanmoins, les performances d'un circuit étant fixées par le chemin critique, le gain global obtenu s'approche du gain au niveau transistor, car le délai du chemin critique est composé, pour une plus grande part, du délai dans les portes, que le reste du circuit.

3.2.4 Résistance accrue aux radiations

Grâce à la couche d'oxyde enterré, le SOI est plus résistant aux radiations : la technologie SOI a commencé à être utilisée plus de deux décennies auparavant dans des applications spatiales ou militaires car elle présente une meilleure immunité aux particules alpha et aux radiations cosmiques – protons ou neutrons de grande énergie. La zone sensible en SOI est limitée à la zone active du transistor : une particule alpha ou des rayons cosmiques vont générer des paires électron-trou tout au long de leur trajectoire mais n'auront plus d'effet une fois passés dans l'oxyde enterré ou en dessous. L'épaisseur de silicium superficiel étant

beaucoup plus faible que l'épaisseur d'un substrat de silicium massif, le nombre de charges générées est beaucoup plus faible en SOI. Ces charges sont évacuées par les jonctions polarisées en inverse.

3.3 Les inconvénients du SOI

3.3.1 Effet d'histoire

A l'inverse, le SOI présente des inconvénients face à la technologie MOS à substrat massif, au premier rang desquels un potentiel de substrat flottant. Comme indiqué précédemment, ce potentiel est déterminé par les couplages capacitifs et, au repos, par les courants de fuite des diodes formées par les jonctions body/source et body/drain. L'état de charge du body dépend alors des transitions passées, du taux d'activité, de la fréquence de fonctionnement, de la pente en entrée, … C'est ce qu'on appelle l'effet d'histoire. Pour le caractériser, il convient d'introduire les notions suivantes : première transition, deuxième transition et état permanent. Prenons l'exemple d'un inverseur. La Figure 20 montre deux signaux d'entrée, l'un partant d'un équilibre statique à 0V, l'autre d'un équilibre statique à V_{DD}. Pour chaque signal, après un long équilibre statique, on peut définir une première transition et une deuxième transition, pour un passage de '0' vers '1' ou de '1' vers '0'. A cause d'équilibres statiques différents du potentiel du body, les première et deuxième transitions dans une même direction vont provoquer des temps de propagation différents. Ainsi, l'effet d'histoire fait varier le délai des portes.

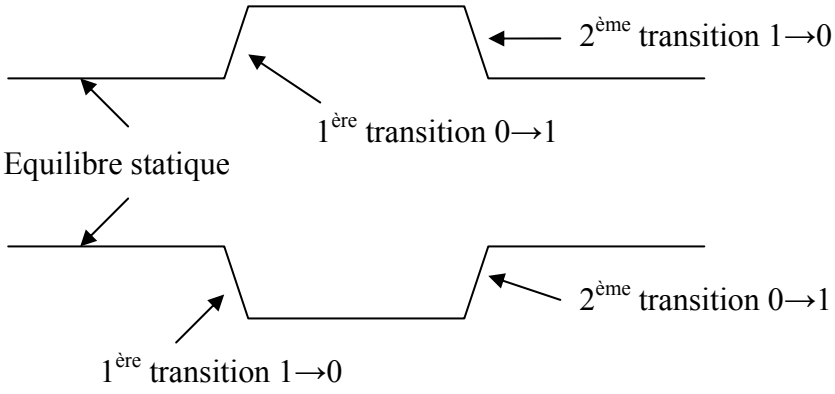


Figure 20 Définitions de la première transition et de la deuxième transition.

3.3.2 Variabilité du délai

Les variations du délai d'un circuit sont problématiques, notamment pour les systèmes synchrones haute performance, car :

- la variabilité[3] dépend de la composition du chemin considéré,
- la variabilité n'est pas uniforme : deux chemins identiques connaîtront des variations de délai différentes,
- la marge allouée, au niveau de la synchronisation, pour compenser les variations possibles du délai, affecte négativement les performances.

En CMOS *bulk*, les sources de variation du délai sont d'ordre temporel et d'ordre spatial. Les variations d'ordre temporel sont :

- la charge ou la décharge d'états précédents,
- les effets d'électrons chauds,
- la charge résiduelle de jonction,
- l'électromigration

Les variations d'ordre spatial sont :

- la variation des paramètres technologiques tels que la tension de seuil V_T et la longueur effective des transistors,
- la variation de la largeur des lignes d'interconnexion,
- la variation de la tension dans la matrice d'alimentation du circuit,
- les sources de bruit telles que le bruit d'alimentation, le couplage capacitif, le partage de charge,
- la synchronisation du signal d'horloge à travers le circuit,
- la variation de la température.

La variabilité d'origine spatiale représente la majeure partie de la variabilité totale. En SOI, il faut rajouter l'effet d'histoire vu précédemment et l'auto-échauffement des transistors puisque l'oxyde de silicium est un mauvais dissipateur thermique. La Figure 21 montre la distribution de la variabilité du délai d'un circuit en technologie SOI [Bern00]. La variabilité temporelle due à l'effet d'histoire est du même ordre de grandeur que la variabilité spatiale due aux

[3] La variabilité est l'écart d'une série d'observations ou de mesures à une mesure de tendance centrale. L'indice de variabilité le plus fréquemment employé est l'écart type.

couplages capacitifs ou aux variations des paramètres du procédé de fabrication. Le phénomène de l'effet d'histoire introduit par le body flottant des transistors SOI ne représente donc pas un problème majeur.

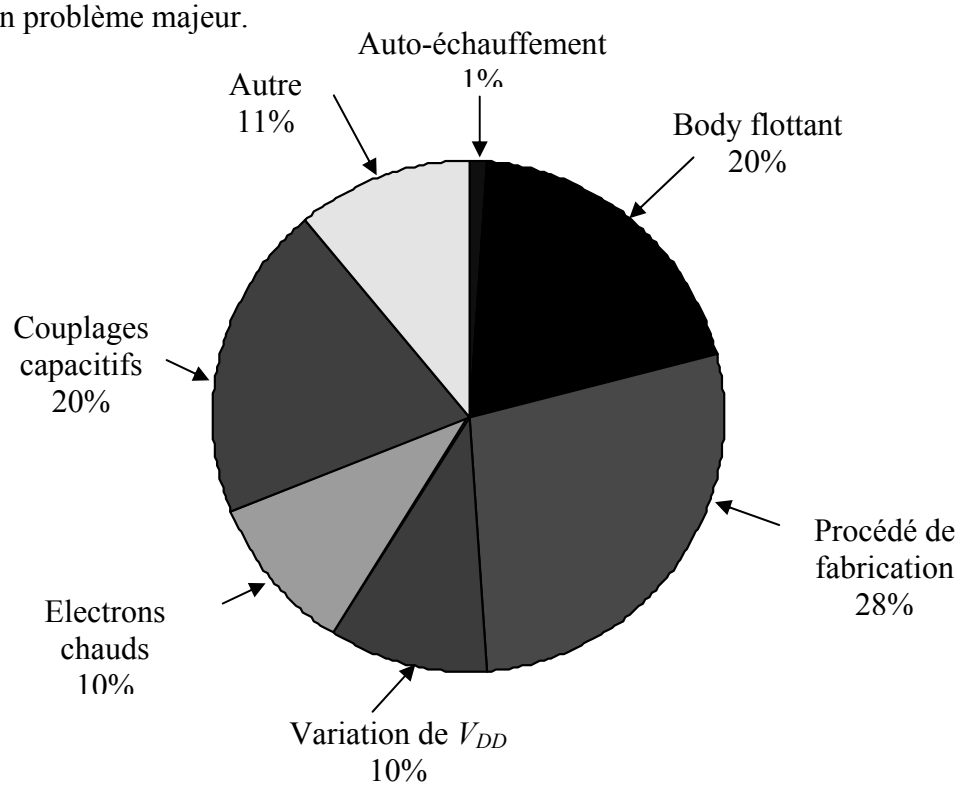


Figure 21 Distribution de la variabilité du délai d'un circuit microélectronique en SOI

3.3.3 Coût

Le coût des substrats SOI est souvent cité pour justifier la réticence de l'industrie à utiliser cette technologie. Il faut en effet partir d'un substrat en silicium massif et lui faire subir plusieurs étapes pour obtenir un substrat SOI. Dans le cas de la technique la plus répandue, à savoir le procédé SmartCut de SOITEC, ces étapes sont respectivement : l'oxydation thermique, l'implantation de protons, le collage moléculaire, le découpage, le recuit et le polissage. Un substrat SOI coûte donc nécessairement plus qu'un substrat silicium. Il est très difficile d'obtenir des informations sur ce surcoût – ce doit être une donnée commerciale jalousement gardée : la seule valeur disponible date de l'année 2002, durant laquelle un substrat SOI de 200mm de diamètre était vendu 320€ contre 80€ pour son homologue silicium. Aujourd'hui, cet écart a forcément diminué, grâce à des volumes de production en hausse, même s'il n'est pas possible de le quantifier. Il faut noter qu'une partie de ce surcoût peut-être compensée par un layout plus compact en SOI et donc une surface réduite.

3.4 Pourquoi le SOI partiellement déserté maintenant ?

La motivation pour migrer vers la technologie SOI est tout d'abord due au fait que ce n'est qu'une évolution de la technologie CMOS *bulk*. Cela permet de garder les mêmes procédés technologiques, mis à part le *wafer*, et donc de diminuer fortement les coûts de développement. Utiliser les mêmes matériaux, les mêmes jeux d'outils et les mêmes procédés permet également de bien maîtriser, à l'avance, la variabilité spatiale, identique à celle du CMOS à substrat massif.

La diminution de la dissipation de puissance est désormais un objectif à part entière et le SOI possède pour cela des avantages considérables, tels que la réduction des capacités de jonction et des performances accrues pour la même tension d'alimentation V_{DD}, permettant de baisser cette dernière tout en gardant les mêmes performances qu'en CMOS *bulk*. Finalement, la réduction des dimensions géométriques, d'une génération à l'autre, conduit à une diminution des capacités et donc à une sensibilité plus élevée aux radiations ; grâce à la présence de l'oxyde enterré, le SOI permet de contrer cette évolution.

Bien évidemment, pour choisir de changer de technologie et justifier les investissements nécessaires, il faut que cette technologie reste prometteuse pour les futures générations lithographiques. Actuellement, en CMOS sur silicium massif, les bénéfices obtenus en réduisant le motif minimal diminuent de génération en génération, à cause du rapport V_{DD}/V_T qui ne peut être maintenu constant : la tension de seuil ne peut pas être abaissée aussi agressivement que la tension d'alimentation, de manière à garder les courants de fuite à un niveau acceptable. En SOI, la polarisation positive du substrat pour des tensions V_{DS} élevées oblige à augmenter la tension de seuil, ce qui laisserait à penser que cette technologie est tout autant limitée que le *bulk* au niveau de la mise à l'échelle de la tension de seuil V_T par rapport à la tension d'alimentation. Cependant, pour des tensions de drain plus faibles, l'effet d'ionisation par impact diminue fortement, ce qui abaisse la valeur du potentiel du body, et permet de réduire plus agressivement la tension de seuil. Ainsi, en évoluant vers des technologies de plus en plus sous-microniques, le SOI devrait connaître une mise à l'échelle de V_T par rapport à V_{DD} plus favorable, grâce à des courants de fuite chutant plus rapidement, et donc un gain en performances plus constant.

La technologie SOI est aujourd'hui de plus en plus employée, tant dans le domaine des applications portables que dans celui des processeurs haute performance, comme on peut le

voir dans le Tableau 3. Avec la baisse du coût des *wafers*, conséquence de la hausse des volumes, le SOI tend à devenir un concurrent très sérieux pour le CMOS à substrat massif.

Tableau 3 Circuits réalisés en SOI et relevés dans ISSCC.

Type	Technologie	Performance	V_{DD}	Compagnie	Année
Multiplicateur 16 b	0,3µm	200MHz	0,5V	Toshiba	1996
DRAM 16Mb	0,5µm	46ns	1V	Mitsubishi	1997
Power PC 32 b	0,25µm	580MHz	2V	IBM	1999
Power PC 64 b	0,2µm	550MHz	1,8V	IBM	1999
Power PC 64 b	0,18µm	660MHz	1,5V	IBM	2000
ALU	0,18µm	380ns	1,5V	Intel	2001
Additionneur 32 b	0,08µm	1ns	1,3V	Fujitsu	2001
PA-RISC 64 b	0,18µm	1GHz	1,5V	HP	2001
SPARC64	0,13µm	1.3GHz	-	Fujitsu	2003
Microcontrôleur	0,10µm	400MHz	0,8V	Mitsubishi	2003

3.5 Choix d'une très basse tension d'alimentation

3.5.1 Définition de la valeur de la tension d'alimentation

Nous allons tout d'abord commencer par définir la notion de facteur de qualité, que nous allons utiliser par la suite. Le facteur de qualité choisi est le produit puissance*délai, car il représente l'énergie dissipée par un circuit, ce qui est primordial dans le cas d'applications portables. En faisant ce choix, on s'oriente vers des circuits plutôt lents, car on favorise autant la minimisation de la puissance dissipée que celle du délai. Pour des circuits haute performance, comme par exemple des serveurs, le facteur de qualité considéré est le produit énergie*délai, voire énergie*délai².

Nous l'avons vu précédemment, il existe plusieurs façons de réduire la dissipation de puissance d'un circuit. La puissance dynamique représente encore aujourd'hui la plus grande partie de sa consommation : pour des applications ne nécessitant pas des performances élevées, le moyen le plus efficace de réduire cette consommation est d'abaisser la tension d'alimentation V_{DD}, comme indiqué Équation 1, la puissance dynamique dépendant quadratiquement de V_{DD}. Bien évidemment, puisque dans le même temps, la tension de seuil V_T n'est pas modifiée, cette opération diminue les performances des transistors. Il existe un

optimum pour lequel la baisse de la puissance dissipée compense le plus l'augmentation du délai. Les Figure 22-a et b montrent l'évolution du produit puissance*délai d'un inverseur avec une sortance de 4 en fonction de la tension d'alimentation V_{DD}, respectivement pour des transistors isocontact (dont le body est relié à la source) et des transistors à body flottant. Dans chaque cas, deux types de transistors ont été considérés : des transistors *High Speed* (HS) à basse tension de seuil, et des transistors *Low Leakage* (LL) à tension de seuil élevée.

Pour chaque courbe, on remarque que l'énergie dissipée chute rapidement avec la baisse de la tension d'alimentation V_{DD}. Ceci est dû au fait que la puissance dynamique diminue quadratiquement. L'énergie dissipée arrive alors à un minimum avant de repartir à la hausse : lorsque l'on approche de la tension de seuil, le délai des transistors augmente exponentiellement, puisque l'on passe d'un courant de saturation à un courant de fuite. L'optimum pour les transistors isocontact et à body flottant ne se situe donc pas à la tension d'alimentation nominale : il se situe autour de 0.7V pour les transistors *Low Leakage* et à 0.5V pour les transistors *High Speed*. On remarque également que l'inverseur en HS possède une énergie jusqu'à 43% plus faible que l'inverseur en LL pour des transistors isocontact et jusqu'à 24% plus faible pour des transistors à body flottant. La fréquence de fonctionnement étant très faible (1 MHz), on peut en conclure que les courants de fuite plus élevés en technologie HS ne posent pas un problème majeur.

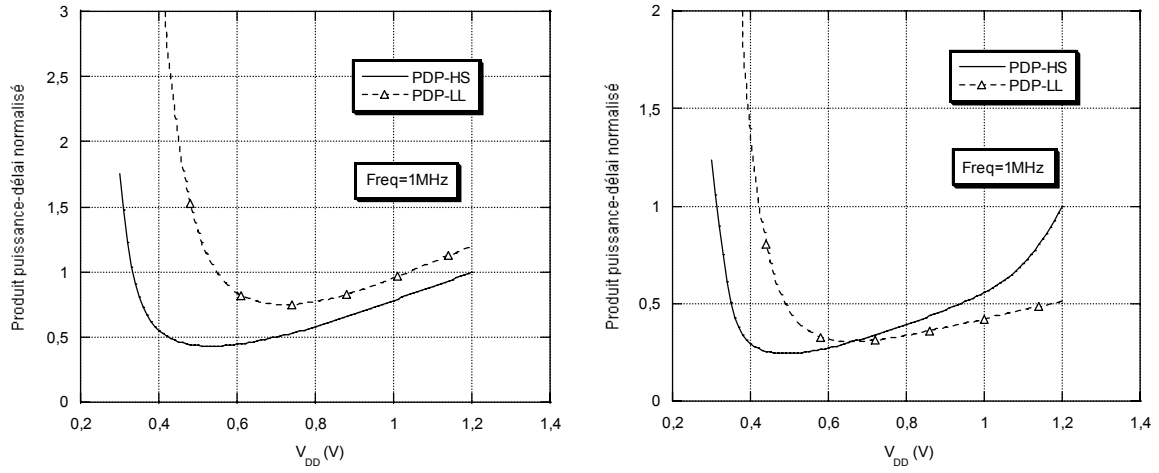


Figure 22 Produit puissance*délai d'un inverseur doté d'un *fanout* de 4 en fonction de la tension d'alimentation V_{DD}, pour deux tensions de seuil différentes, haute (LL : Low Leakage) et basse (HS : High Speed) : (a) transistors isocontact, (b) transistors à body flottant.

Nous faisons donc le choix de travailler avec une tension d'alimentation égale à 0.5V, qui minimise l'énergie dissipée, en utilisant des transistors HS. Hormis la perte de performances, les avantages en sont multiples, comme nous allons le voir dans la suite.

3.5.2 Avantages d'un tel choix

Le premier avantage évident est la forte diminution de la puissance dissipée. Comme nous l'avons vu, la puissance dynamique chute de manière quadratique. Mais ce n'est pas la seule. La puissance statique des courants de fuite diminue également exponentiellement, comme cela a été montré Figure 7-a (se reporter au Chapitre 2, paragraphe 2.3 pour plus de détail). Cela est dû essentiellement au fait que la tension de seuil d'un transistor dépend de la tension V_{DS}, à cause des effets de rétroaction statique du drain et d'abaissement de la barrière de potentiel au niveau de la source. Enfin, la puissance de court-circuit statique, lors du changement d'état d'une porte, est quasiment nulle, puisque la somme des tensions de seuil des transistors NMOS et PMOS est inférieure à la tension d'alimentation. Ainsi, les transistors de type opposé ne peuvent être simultanément passants.

En travaillant avec une tension d'alimentation inférieure à la tension de seuil d'une diode, on peut utiliser une structure particulière au SOI, le transistor DTMOS – *Dynamic Threshold MOS* [Assa94]. Il s'agit d'un transistor dont le body est relié à la grille : le schéma est donné Figure 23.

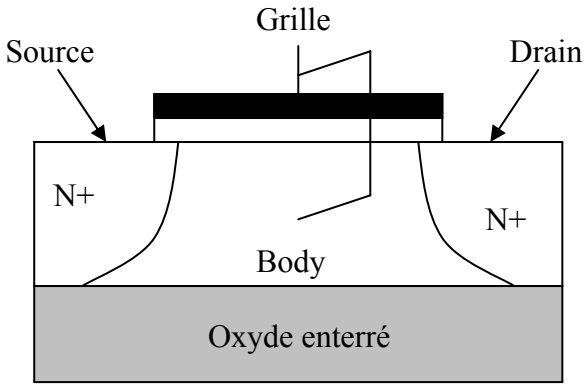


Figure 23 Coupe d'un transistor DTMOS : le body est relié à la grille

Le transistor DTMOS possède le double avantage d'un courant de saturation élevé lorsqu'il est passant et d'un courant de fuite faible lorsqu'il est bloqué. Dans le cas où la tension sur la grille est amenée à la tension d'alimentation V_{DD}, le potentiel du substrat augmente

également, ce qui a pour effet de diminuer la tension de seuil, d'après l'Équation 4. La zone de déplétion augmentant, le potentiel requis sur la grille pour créer le canal d'inversion est plus faible. La tension de seuil ainsi obtenue est plus basse que celle de transistors à body flottant, dont le potentiel du substrat aurait été augmenté par un couplage capacitif avec la grille. A l'inverse, lorsque la tension de grille est amenée à 0, la tension de seuil du transistor augmente ce qui limite le courant de fuite. La Figure 24 montre cet avantage : le courant I_{ON} du transistor de type N monté en DTMOS est égal à celui du transistor dont le body est maintenu à 0.5Vet le courant de fuite I_{OFF} du DTMOS est aussi faible que celui du transistor dont le body est maintenu à 0V. La pente sous le seuil d'un transistor DTMOS est bien meilleure que celle d'un transistor à body flottant, illustré en bleu dans le graphique : elle vaut ici 63mV/décade contre 78mV/décade respectivement, ce qui la rapproche de la pente idéale de 60mV/décade [Assa94].

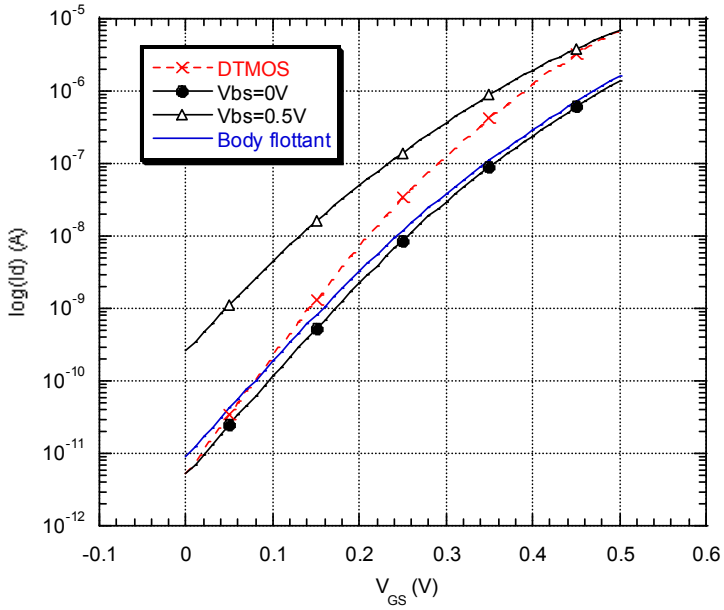


Figure 24 Comparaison des courants de drain d'un transistor NMOS dans différentes configurations : DTMOS, V_{BS}=0V, V_{BS}=V_{DD} et body flottant.

Le transistor DTMOS n'est néanmoins pas exempt d'inconvénient, le premier étant une capacité de grille plus importante, à cause du contact de body. Un transistor DTMOS de type NMOS en technologie 0,13µm de STMicroelectronics est montré Figure 25 : un chemin de conduction de type P est créé entre la zone active du transistor sous la grille et le contact. La forme en ailette de la grille est utilisée pour diminuer sa capacité.

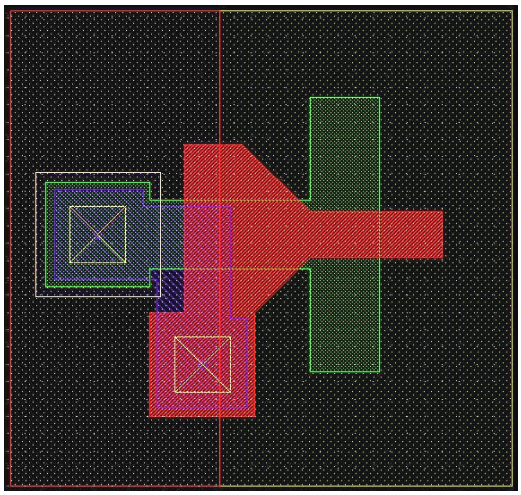


Figure 25 Layout d'un transistor DTMOS montrant le contact de body.

La Figure 26 montre la capacité de grille d'un transistor à body flottant et celle d'un transistor DTMOS à 1 ou 2 contacts en fonction de la largeur de la grille – pour avoir un bon contrôle du canal par la grille, il faut ajouter un deuxième contact si la largeur du transistor dépasse 2.5µm ; au-delà de 5µm, il faut placer un autre transistor DTMOS en parallèle. Pour des largeurs de grille très faibles, la capacité introduite par le contact de body peut doubler la capacité de grille.

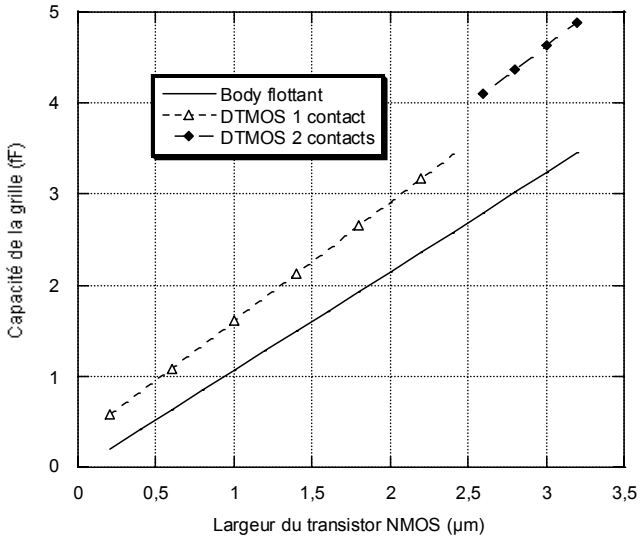


Figure 26 Comparaison des capacités de grille d'un transistor à body flottant et d'un transistor DTMOS, avec 1 ou 2 contacts selon la largeur.

Pour déterminer s'il est opportun d'utiliser des transistors DTMOS, nous introduisons la métrique suivante :

$$\frac{I_{ON}(DT)}{I_{ON}(FB)} \Big/ \frac{C_G(DT)}{C_G(FB)}$$

Équation 8

Si le résultat obtenu est supérieur à 1, alors il est plus intéressant d'utiliser des transistors DTMOS par rapport à des transistors à body flottant car les premiers vont introduire moins de capacité de grille sur l'étage précédent qu'ils ne vont fournir de courant additionnel à l'étage suivant. A l'opposé, si le résultat est inférieur à 1, les transistors DTMOS doivent être évités car ils ne vont faire que ralentir le chemin dans lequel ils ont été placés. Cette métrique a été calculée pour les transistors NMOS et PMOS en technologie SOI HS et LL. Les résultats sont présentés Figure 27. Cela implique que :

- en technologie SOI HS, les transistors DTMOS n'amélioreront la vitesse que s'ils sont suffisamment larges, c'est-à-dire qu'ils doivent être utilisés pour attaquer des charges élevées ;
- en technologie SOI LL, les transistors DTMOS vont apporter un gain quelque soit leur taille.

Ces conclusions peuvent être expliquées par le fait que les bénéfices apportés par les transistors DTMOS dépendent de la variabilité de leur tension de seuil V_T. La modulation de V_T est fonction du niveau de dopage et par conséquent, dans le cas de la technologie HS, le niveau de dopage plus faible limite les variations de V_T. Le gain obtenu sur le courant I_{ON} par l'utilisation d'un transistor DTMOS n'est pas assez important par rapport à l'augmentation de la capacité de grille pour des petites tailles. Il ne faut pas en conclure que l'utilisation de transistors DTMOS n'est pas intéressante en technologie HS, il faut seulement limiter leur emploi aux capacités importantes.

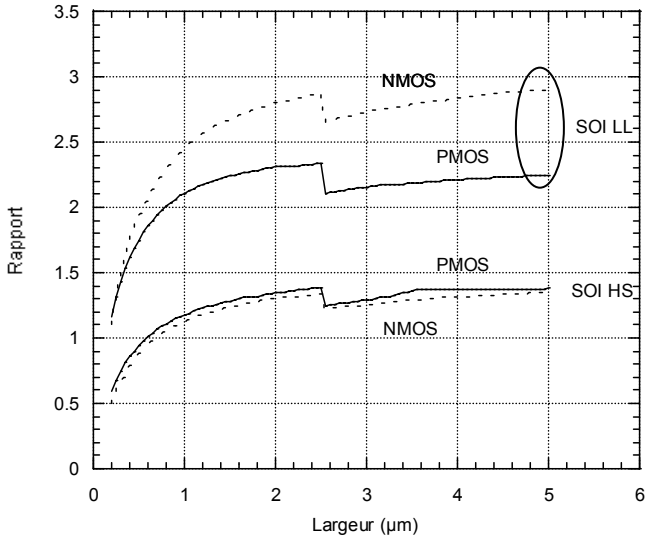


Figure 27 Le rapport défini dans l'Équation 8 est calculé pour les technologies HS et LL en fonction de la taille des transistors. Les cassures observées sur les courbes sont dues au passage de 1 à 2 contacts de body.

3.6 Conclusion

La technologie SOI est une technologie qui se prête très bien aux applications basse consommation. Elle présente en effet des capacités de jonction plus réduites qu'en silicium massif, ce qui diminue la consommation dynamique. De plus, la modulation dynamique de la tension de seuil et la suppression de l'effet source-suiveur inverse permettent d'améliorer les performances et donc de diminuer la tension d'alimentation. Nous avons vu que la tension d'alimentation optimale pour minimiser l'énergie dissipée se situe autour de 0.5V, et qu'il faut pour cela utiliser des transistors haute performance (HS) : ils présentent en effet une énergie dissipée plus faible que les transistors à bas courant de fuite (LL). En opérant à cette tension d'alimentation, nous pouvons utiliser des transistors DTMOS, dont la tension de seuil peut être contrôlée précisément, au détriment d'une capacité de grille en hausse. Ces transistors permettent d'avoir simultanément un courant I_{ON} plus important et un courant I_{OFF} plus réduit. Une métrique a été proposée pour déterminer si l'utilisation d'un transistor DTMOS est intéressante – c'est-à-dire si le gain en délai procuré est supérieur au retard causé dans l'étage précédent par une capacité de grille en hausse. Cette métrique a montré que les transistors DTMOS doivent être de taille importante en technologie HS pour apporter un bénéfice, et donc attaquer une charge importante, alors qu'en technologie LL, des transistors

DTMOS de petite taille peuvent être utilisés, grâce à la plus grande variabilité de leur tension de seuil.

4 Modèle sous seuil d'évaluation de la technologie SOI

Comme nous l'avons vu précédemment, la dissipation d'énergie optimale d'un circuit est obtenue pour une tension d'alimentation très basse. Bien évidemment, cela s'effectue au détriment des performances. Cependant, certaines applications ne nécessitent pas de hautes performances, telles que les systèmes auditifs, les cartes sans contact ou encore les capteurs tirant leur énergie de l'environnement – appelée *energy scavenging*. Utiliser les transistors en inversion modérée ou même sous la tension de seuil s'avère dans ce cas très attractif. [Swan72] a montré la robustesse de la caractéristique de transfert d'un inverseur opéré à très faible tension, jusqu'à 100mV. Pour estimer le comportement des transistors opérés en inversion faible, un modèle d'évaluation est nécessaire. Des modèles tels que EKV [Enz95] et le modèle MOSFET trans-régional basse puissance [Aust98] ont été développés pour la conception et la simulation de circuits analogiques ou à signaux mixtes basse tension. L'inconvénient de ces modèles est que l'expression de leur courant de drain est relativement complexe et ne permet pas de faire des calculs « à la main ». D'autres modèles ont été développés de manière à simplifier l'équation du courant de drain, tel le modèle MOSFET physique à loi puissance alpha [Bowm99] – dérivé du précédent modèle MOSFET à loi

puissance alpha [Saku90] qui ne décrivait pas les caractéristiques sous seuil. L'inconvénient de ce modèle est qu'il ne prend pas en compte les effets de canal court.

Nous avons donc développé un modèle simple mais précis, basé sur des paramètres physiques, qui décrit les caractéristiques sous seuil des transistors en prenant la dépendance de la tension body-source et les effets de canal court. Grâce à ce modèle, il est possible de dériver le temps de propagation d'un inverseur, en tenant compte de la pente en entrée et du potentiel de body des transistors SOI : les résultats sont appliqués à une structure d'oscillateur en anneau.

4.1 Modèle sous seuil analytique

Le courant de drain I_{DS} dans la région de conduction sous le seuil dépend exponentiellement de la tension grille-source V_{GS} [Sze81]. Cette dépendance s'exprime en fixant une référence de densité de courant $d_0=I_0/W_0$ mesurée à $V_{GS}=V_{T0}$ et du facteur de pente sous le seuil S :

$$I_{DS}(V_{GS}) = W \cdot \frac{I_0}{W_0} \cdot 10^{\frac{|V_{GS}|-|V_{T0}|}{S}},$$

Équation 9

où W représente la largeur de la grille du transistor. La constante V_{T0} représente la tension V_{GS} nécessaire pour qu'un transistor de largeur W_0 fournisse un courant I_0 – voir Figure 28. Elle n'a donc aucune relation avec la tension de seuil V_T de forte inversion du transistor. Le terme S représente la pente sous le seuil, qui est fonction de la capacité de déplétion C'_D et de la capacité d'oxyde de grille C'_{OX}, par unité de surface :

$$S = \frac{K \cdot T}{q} \cdot \ln(10) \cdot \left(1 + \frac{C'_D}{C'_{OX}}\right)$$

Équation 10

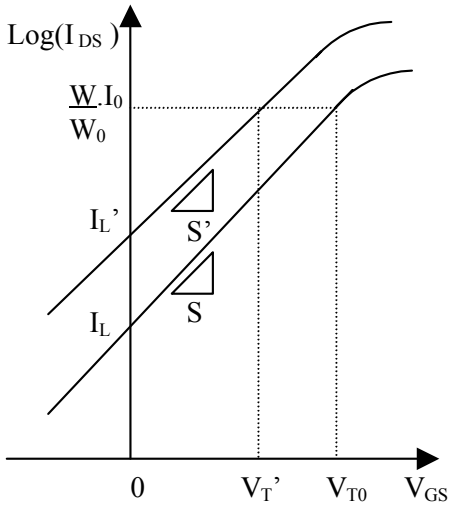


Figure 28 Caractéristiques sous le seuil d'un transistor NMOS.

Pour une tension grille-source nulle, le courant de fuite a pour valeur :

$$I_L = W \cdot d_0 \cdot 10^{-\frac{|V_{T0}|}{S}}$$

Équation 11

Comme illustré Figure 28, le courant de conduction sous le seuil augmente exponentiellement lorsque la tension de seuil – définissant la référence de courant – varie de V_{T0} à V_T'. En SOI, la variation de la tension de seuil est due à l'effet de body, qui n'est pas inclue dans les équations précédentes.

4.1.1 Dépendance de la tension de body

La variation de la tension body-source va avoir un impact à la fois sur la valeur de la tension de seuil et sur celle de la pente sous le seuil. Considérons dans un premier temps la tension de seuil, le passage de celle-ci de V_{T0} à V_T' peut s'exprimer par une variation ΔV_T :

$$V_T' = V_{T0} + \Delta V_T$$

Équation 12

L'expression de ΔV_T est similaire à celle concernant la tension de seuil en forte inversion :

$$\Delta V_T = \gamma\left(\sqrt{2 \cdot \Psi_F - (1-\alpha) \cdot V_{BS}} - \sqrt{2 \cdot \Psi_F}\right)$$

Équation 13

Le paramètre α a été introduit pour mieux décrire la dépendance du la tension body-source. Le terme γ représente le facteur de body :

$$\gamma = \frac{\sqrt{2 \cdot q \cdot N_A \cdot \varepsilon_0 \cdot \varepsilon_{SI}}}{C_{OX}'}$$

Équation 14

et Ψ_F le potentiel de quasi-Fermi dans le body :

$$\Psi_F = \frac{K \cdot T}{q} \cdot \ln\left(\frac{N_A}{n_i}\right)$$

Équation 15

La pente sous le seuil, quant à elle, dépend de la capacité de déplétion, comme montré Équation 10. Cette capacité étant fonction de la variation de la quantité de charge de déplétion Q_D', elle dépend directement du potentiel de surface Ψ_S :

$$C_D' = \frac{\delta Q_D'}{\delta \Psi_S} \qquad \text{avec :} \qquad Q_D' = \sqrt{2 \cdot q \cdot N_{!B} \cdot \varepsilon_0 \cdot \varepsilon_{si} \cdot \left(\Psi_S - V_{BS}\right)}$$

$$\text{donc } C_D' = \sqrt{\frac{q \cdot N_B \cdot \varepsilon_0 \cdot \varepsilon_{si}}{2 \cdot \left(2 \cdot \Psi_F - V_{BS}\right)}}$$

Équation 16

4.1.2 Dépendance de la tension de drain

Nous allons à présent introduire les caractéristiques $I_{DS}(V_{DS})$. Comme illustré Figure 29, le courant de drain a la même forme qu'en forte inversion : il existe une zone linéaire et une zone saturée. Puisque que le transistor ne fonctionne pas en régime linéaire ni en régime saturé, nous avons appelé ces régions respectivement pseudo-linéaire et pseudo-saturée. La tension V_{DS} qui distingue ces régions est appelée tension de pseudo-saturation et est à peu près égale à 100mV. Les modèles publiés [Enz95], [Aust98], décrivent bien la partie pseudo-linéaire mais pas celle pseudo-saturée : la Figure 29 montre l'erreur introduite par les modèles publiés – en pointillés – et le comportement sous le seuil réel – en trait continu.

Dans la zone pseudo-linéaire, le courant de drain dépend exponentiellement de la tension V_{DS} :

$$I_{DS}(V_{DS}) = W \cdot d_0 \cdot 10^{\frac{|V_{GS}|-|V_T'|}{S}} \cdot \left(1 - e^{-\frac{m \cdot V_{DS}}{V_{th}}}\right)$$

Équation 17

où m est un paramètre d'ajustement proche de 1 et V_{th} la tension thermique.

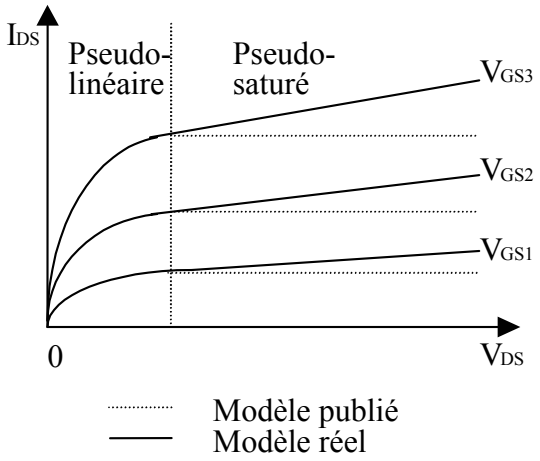


Figure 29 Caractéristiques I_{DS} (V_{DS}) d'un transistor NMOS.

Dans la région de pseudo-saturation, la dépendance du courant de drain devient linéaire :

$$I_{DS}(V_{DS}) = W \cdot d_0 \cdot 10^{\frac{|V_{GS}|-|V_T'|}{S}} \cdot (a + \lambda \cdot V_{DS})$$

Équation 18

En combinant l'Équation 17 et l'Équation 18, on obtient l'expression complète du modèle de comportement sous seuil d'un transistor SOI, en fonction des tensions V_{GS}, V_{BS} et V_{DS} :

$$I_{DS} = W \cdot d_0 \cdot 10^{\frac{|V_{GS}|-|V_T'|}{S}} \cdot \left(1 - e^{-\frac{m \cdot V_{DS}}{V_{th}}}\right) \cdot (a + \lambda \cdot V_{DS})$$

Équation 19

Nous avons négligé l'effet DIBL, c'est-à-dire l'impact de la tension de drain sur la tension de seuil du transistor car cet effet se manifeste plus pour des longueurs de canal effectives courtes et des tensions de drain élevées : le DIBL est en effet proportionnel à V_{DS}/L_{eff}^2 [Veen98]. Or nous avons développé ce modèle pour une technologie SOI 0.25µm utilisée à une tension d'alimentation inférieure à 500mV. Cet effet devra être pris en compte dans la

modélisation de technologies fortement sous-microniques, de motif minimal inférieur ou égal à 0.13µm. Pour résumer, le modèle est décrit à l'aide de cinq paramètres :

- un paramètre de procédé :

  la concentration N_A, qui donne γ et Ψ_F

- quatre paramètres d'ajustement :

  α, qui fixe la variation de la tension de seuil,

  m, pour mieux décrire la région pseudo-linéaire du courant $I_{DS}(V_{DS})$,

  a et γ, extraits de la courbe $I_{DS}(V_{DS})$, qui représentent respectivement l'ordonnée à l'origine et la pente de la région pseudo-saturée.

Nous pouvons désormais comparer notre modèle à des simulations Eldo et au modèle MOSFET physique à loi puissance alpha [Bowm99]. Comme on peut le constater Figure 30, le modèle physique à loi puissance alpha ne décrit pas correctement les effets de canal court des transistors opérant sous le seuil : il ne modélise pas la région de pseudo-saturation.

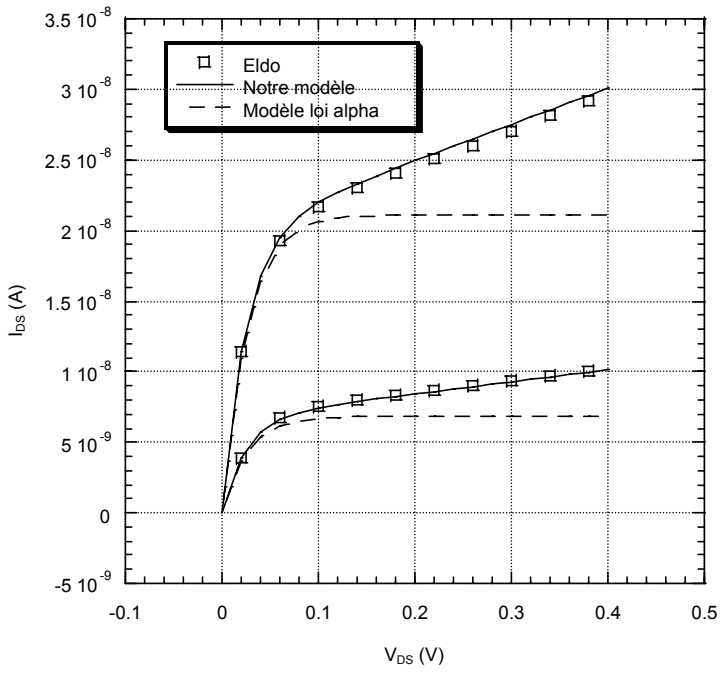
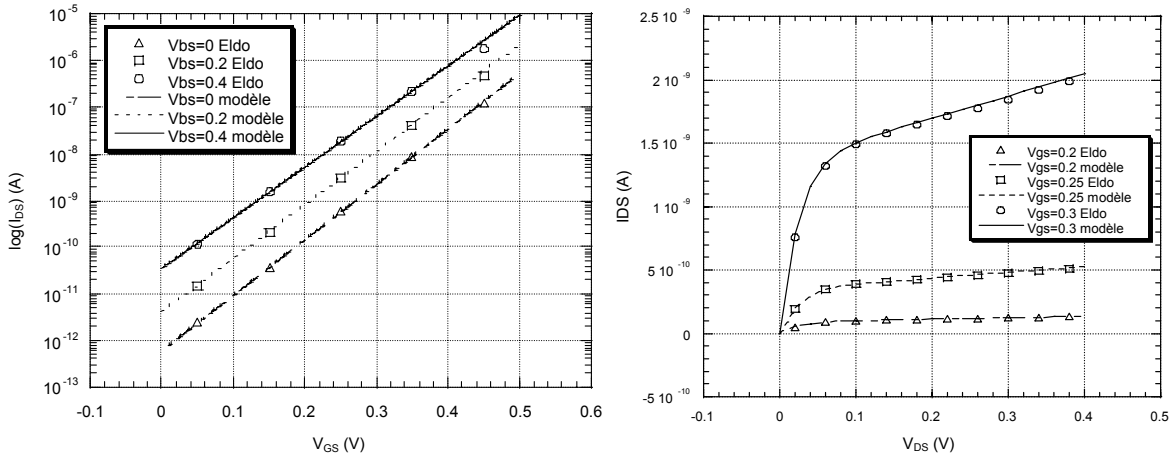


Figure 30 Comparaison des caractéristiques $I_{DS}(V_{DS})$ de notre modèle, de simulations Eldo et du modèle physique à loi puissance alpha.

La Figure 31 montre plus en détails les comparaisons de notre modèle sous-seuil analytique avec les simulations Eldo : la Figure 31-a affiche les courbes I_{DS} (V_{GS}) pour différentes polarisations de substrat, tandis que la Figure 31-b affiche les caractéristiques I_{DS} (V_{DS}) pour différentes tensions de grille.

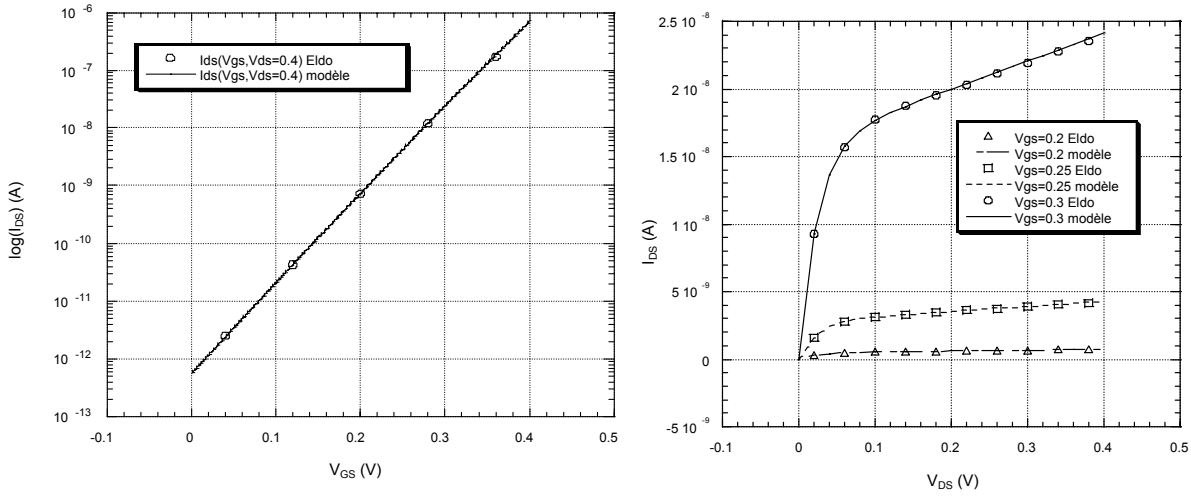
Grâce à ce modèle analytique, on peut facilement modéliser un transistor DTMOS – dont le body est relié à la grille – en égalisant les tensions V_{BS} et V_{GS} : le terme V_{GS} est introduit dans les équations décrivant l'effet de body, à savoir les variations de la tension de seuil et de la pente sous le seuil. Les résultats obtenus en procédant de la sorte sont très proches des simulations Eldo comme montré Figure 32-a et Figure 32-b.

Finalement, nous comparons notre modèle avec des mesures effectuées sur silicium : celles-ci concernent la caractéristique I_{DS} (V_{GS}) sur cinq décades pour une tension drain-source égale à 0.1V.



(a) **(b)**

Figure 31 Comparaison de notre modèle analytique avec des simulations Eldo pour (a) I_{DS} (V_{GS}) et (b) I_{DS} (V_{DS})



(a) **(b)**

Figure 32 Comparaison du modèle de DTMOS avec des simulations Eldo pour (a) I_{DS} (V_{GS}) et (b) I_{DS} (V_{DS}).

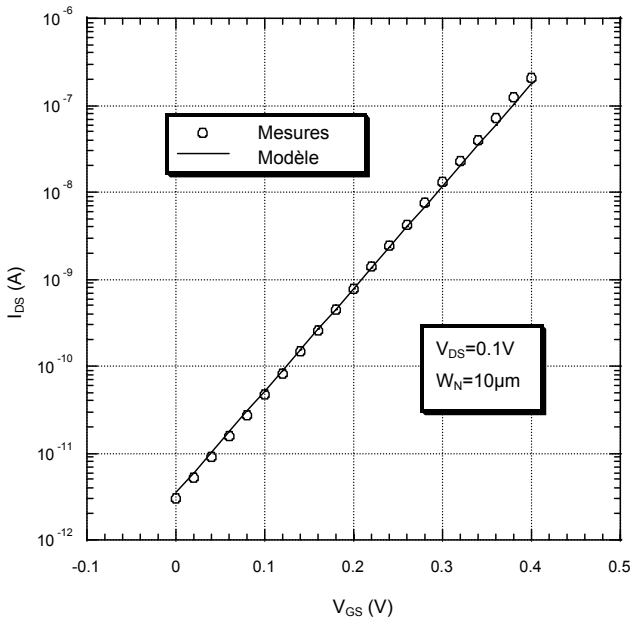


Figure 33 Comparaison de la caractéristique I_{DS} (V_{GS}) entre notre modèle analytique et des mesures silicium pour V_{DS}=0.1V.

4.2 Application à un inverseur

Le modèle ayant des expressions de courant de drain relativement simples, nous allons pouvoir l'utiliser de manière à décrire le comportement d'un inverseur en très basse tension. Dans un premier temps, nous allons étudier l'inverseur dans le domaine statique, c'est-à-dire sa caractéristique de transfert. Puis nous allons poursuivre l'analyse en dynamique et dériver les équations du temps de propagation d'un inverseur en fonction de la charge en sortie et de la pente en entrée pour deux cas : une pente en entrée rapide devant celle de sortie et une pente en entrée lente. Finalement, nous allons appliquer ces formules à l'étude d'un oscillateur en anneaux.

4.2.1 Analyse statique

La Figure 34 montre la caractéristique de transfert d'un inverseur pour différentes polarisations de substrat des transistors NMOS et PMOS. Les courbes en trait noir continu sont extraites de simulations Eldo, tandis que celles en trait rouge pointillé sont obtenues à l'aide du modèle analytique et de Matlab. Le point intéressant d'une caractéristique de transfert est la tension de seuil logique V_M, définie comme étant la tension en entrée pour

laquelle la tension en sortie vaut $V_{DD}/2$. L'expression de V_M peut être obtenue en égalisant les courants des transistors NMOS et PMOS et en résolvant la tension V_{GS} des transistors – voir Annexe 1.

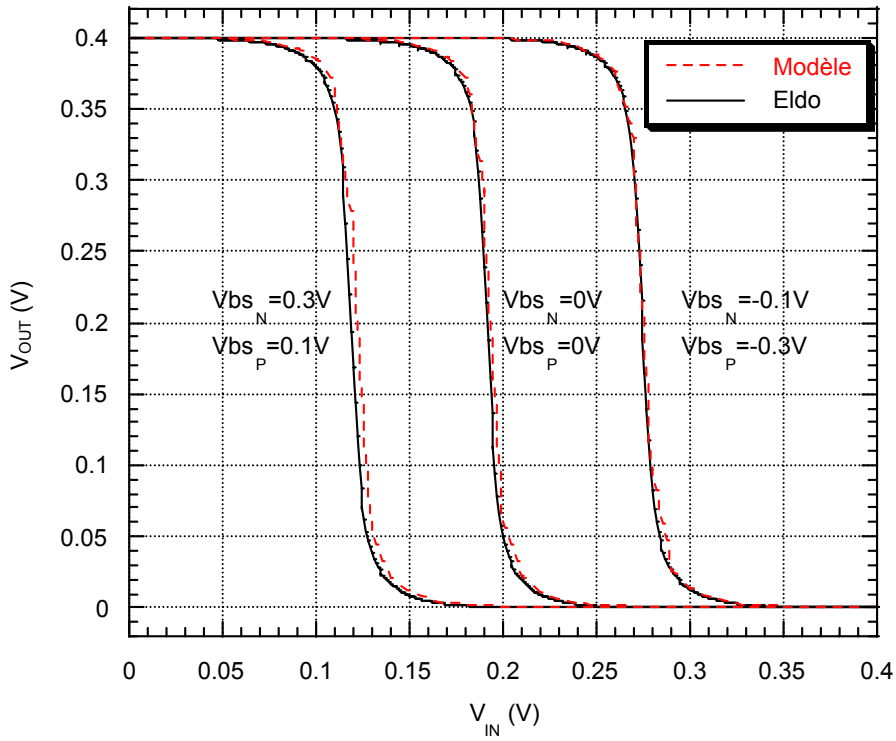


Figure 34 Comparaison des caractéristiques de transfert d'un inverseur entre le modèle et des simulations Eldo pour différentes polarisations de substrat des transistors NMOS et PMOS.

La tension de seuil logique a ainsi pour expression :

$$V_M = S_{np}\left[\frac{V_{Tn}}{S_n} + \frac{V_{DD} - |V_{Tp}|}{S_p}\right] + \log_{10}\left(\frac{d_{op}}{d_{on}} \cdot \frac{a + \lambda_p \frac{V_{DD}}{2}}{a + \lambda_n \frac{V_{DD}}{2}} \cdot \frac{W_p}{W_n}\right),$$

Équation 20

où les indices n et p correspondent aux transistors NMOS et PMOS respectivement. On peut alors calculer la tension de seuil logique en fonction de la tension V_{BS} des transistors, ce qui est représenté Figure 35. L'erreur entre les valeurs simulées et calculées est inférieure à 3,5% pour des valeurs élevées de V_{BS}.

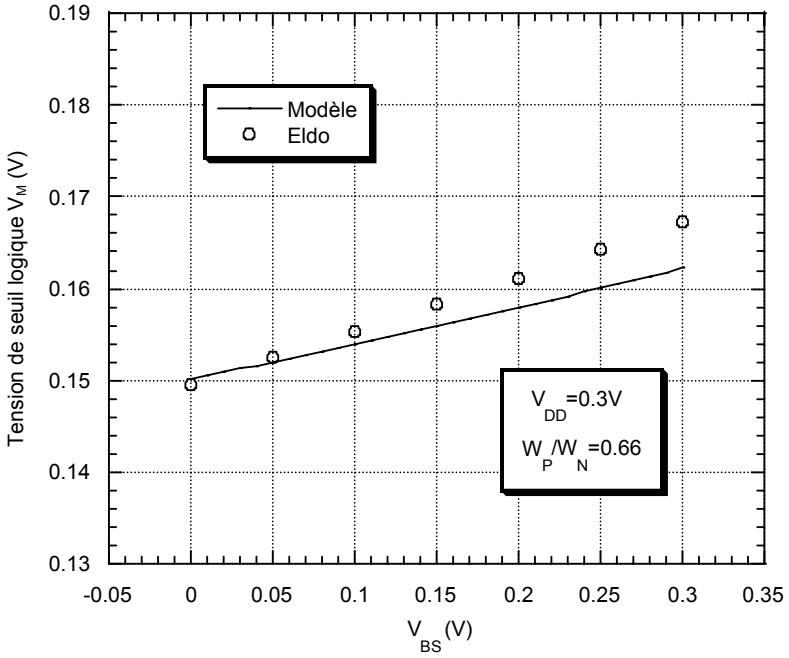


Figure 35 Evolution de la tension de seuil logique V_M en fonction de la tension V_{BS}, égale en valeur absolue pour les deux transistors.

4.2.2 Analyse dynamique

Dans cette partie, nous allons étudier le temps de propagation d'un inverseur, tp_{HL} et tp_{LH}, qui est le temps séparant le passage de la tension d'entrée par $V_{DD}/2$ du passage de la sortie par $V_{DD}/2$, respectivement dans le cas de la décharge et de la charge de la capacité de sortie. Le temps de propagation d'un inverseur dépendant de la pente du signal en entrée, considérons tout d'abord le cas d'une transition rapide. Nous pouvons alors faire l'hypothèse que le signal d'entrée varie plus rapidement que la tension de sortie, ce qui signifie que la tension d'entrée aura atteint sa valeur maximale avant que la sortie n'ait eu le temps de varier. Dans le cas de la décharge de la capacité de sortie, le transistor PMOS peut être considéré bloqué lors de la transition tandis que le transistor NMOS est saturé. D'après l'Annexe 2, on obtient pour expression :

$$tp_{HL} = \frac{C_{tot}}{I_{SS_n}} \left[\frac{1}{\lambda_n \left(1 - e^{-\frac{V_{sat_n}}{V_{th}}} \right)} \cdot \log \left(\frac{a + \lambda_n V_{DD}}{a + \lambda_n \cdot 0.5 \cdot V_{DD}} \right) \right]$$

Équation 21

Dans le cas de la charge de la sortie, le raisonnement est similaire : cette fois-ci, c'est le transistor NMOS qui est considéré bloqué tandis que le PMOS est saturé. L'Annexe 3 donne l'expression suivante :

$$tp_{LH} = \frac{C_{tot}}{I_{SS_p}} \left[\frac{1}{\lambda_p \left(1 - e^{\frac{V_{sat\,p}}{V_{th}}} \right)} \cdot \log \left(\frac{a + \lambda_p \cdot 0.5 \cdot V_{DD}}{a + \lambda_p V_{DD}} \right) \right]$$

Équation 22

Dans l'Équation 21 et l'Équation 22, la capacité de sortie C_{tot} est une constante : elle ne présente pas de variation en tension, alors que c'est généralement inclus dans les modèles de capacité Eldo. Nous faisons néanmoins cette approximation pour la raison suivante : la capacité de sortie est composée de quatre termes, représentés graphiquement dans la Figure 36, qui sont la capacité de recouvrement grille-drain C_{gd}, la capacité de jonction drain-body C_{db}, la capacité d'entrée de l'étage logique suivant C_{L} et enfin la capacité d'interconnexion. Cependant, les deux premières capacités présentent de faibles variations puisque les variations en tension sont petites. Ce raisonnement est également valable pour la capacité d'entrée de l'étage suivant : la capacité de grille est relativement constante puisque, le canal d'inversion n'étant jamais créé, on a toujours en série la capacité d'oxyde de grille et la capacité de déplétion. On peut donc, au premier ordre, faire l'approximation que la capacité C_{tot} est constante.

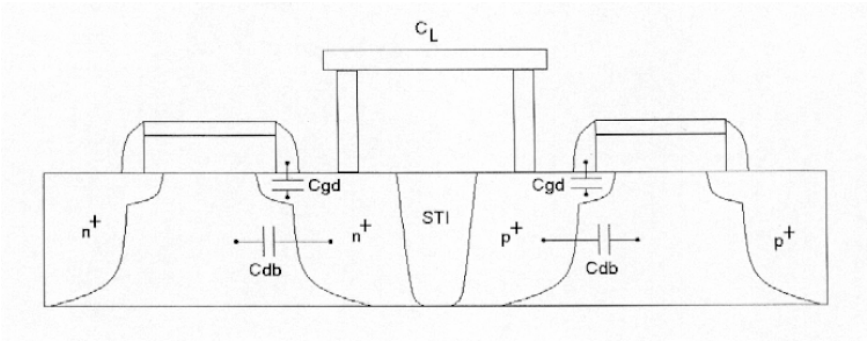


Figure 36 Différents termes de la capacité de sortie d'un inverseur.

La comparaison des formules précédentes avec des simulations Eldo est donnée Figure 37. L'utilité de ce type de modèle pour des circuits SOI est démontrée Figure 38 où le temps de propagation est donné en fonction de la polarisation des substrats des transistors.

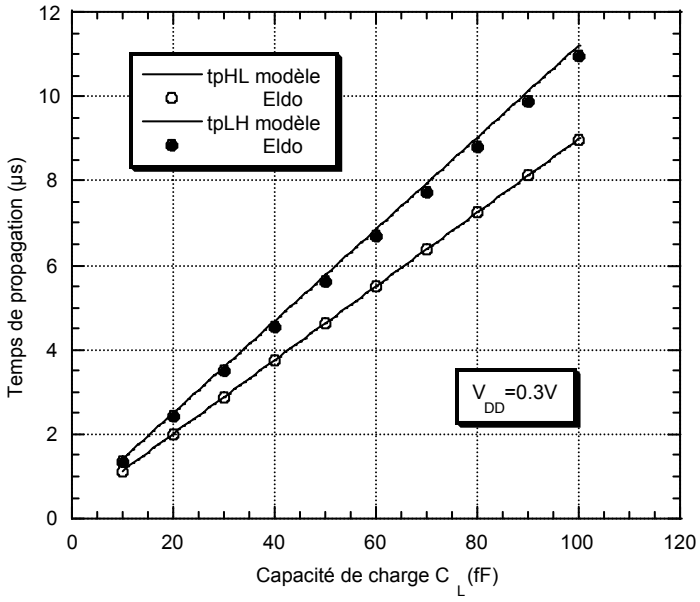


Figure 37 Temps de propagation tp_{LH} et tp_{HL} d'un inverseur pour différentes capacités de charge.

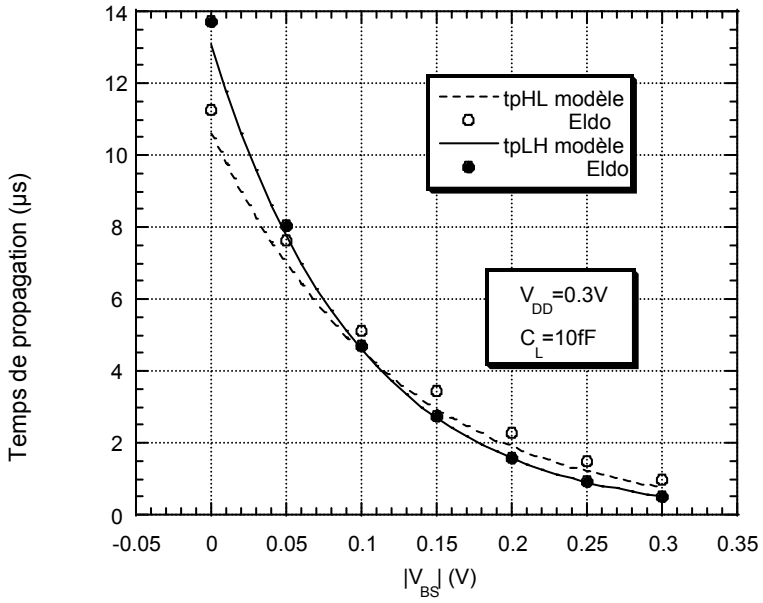


Figure 38 Temps de propagation d'un inverseur en fonction de la polarisation des substrats des transistors.

Considérons maintenant le cas où l'entrée varie dans le même intervalle de temps que la sortie, ou plus lentement. Dans ce cas, le courant de drain du transistor saturé ne dépendra plus seulement de la tension V_{DS} mais aussi de la tension V_{GS} : le courant de charge ou de décharge sera fonction de la tension de sortie et du temps de transition de l'entrée. Nous pouvons ici aussi négliger le courant du transistor opposé, grâce aux caractéristiques sous seuil des transistors. Prenons le cas de la décharge de la sortie : la Figure 39 montre la pente en entrée, la tension de sortie – échelle de gauche – et les courants des transistors NMOS et PMOS – échelle de droite. Le courant du transistor PMOS est de l'ordre du nA, soit au moins deux ordres de grandeur plus petit que celui du transistor NMOS. Cela s'explique pour les raisons suivantes : dans la partie I, la tension V_{GS} du PMOS est très petite, donc son courant de drain qui en dépend exponentiellement est lui aussi très petit ; dans la partie II, c'est la tension V_{DS} qui cette fois est inférieure à la tension de saturation où là encore, le courant de drain devient exponentiellement petit. Le courant du transistor PMOS peut donc être négligé.

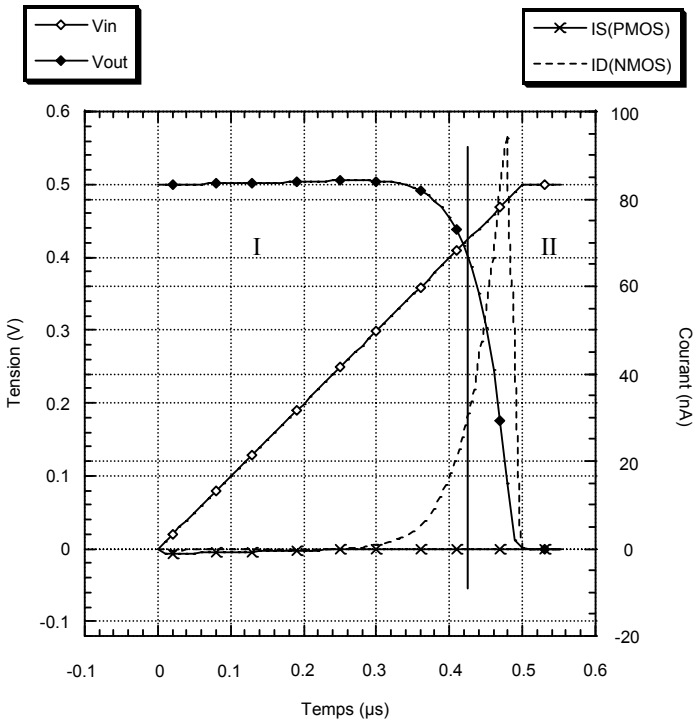


Figure 39 Courants des transistors NMOS et PMOS pour une entrée variant lentement, dans le cas d'une décharge de la capacité de sortie.

Pour obtenir le délai, il faut considérer l'équation différentielle suivante :

$$I_D(t, V_{OUT}) = C \cdot \frac{dV_{OUT}}{dt}$$

Équation 23

Après intégration, détaillée dans l'Annexe 4, on obtient :

$$tp_{HL} = \frac{S_n}{k} \cdot \log_{10}\left[\frac{k \cdot C \cdot \log(10)}{A_n \cdot \lambda_n \cdot S_n} \cdot \log\left(\frac{a_n + \lambda_n V_{DD}/2}{a_n + \lambda_n V_{DD}}\right) + 1\right] - \frac{V_{DD}}{2k},$$

Équation 24

où A_n est une constante et k la pente du signal d'entrée. Dans le cas de la charge de la capacité de sortie, l'expression est la même – voir l'Annexe 5 :

$$tp_{LH} = -\frac{S_p}{k} \cdot \log_{10}\left[-\frac{k \cdot C \cdot \log(10)}{A_p \cdot \lambda_p \cdot S_p} \cdot \log\left(\frac{a_p - \lambda_p V_{DD}/2}{a_p - \lambda_p V_{DD}}\right) + 1\right] - \frac{V_{DD}}{2|k|}$$

Équation 25

Nous comparons les résultats obtenus dans la Figure 40.

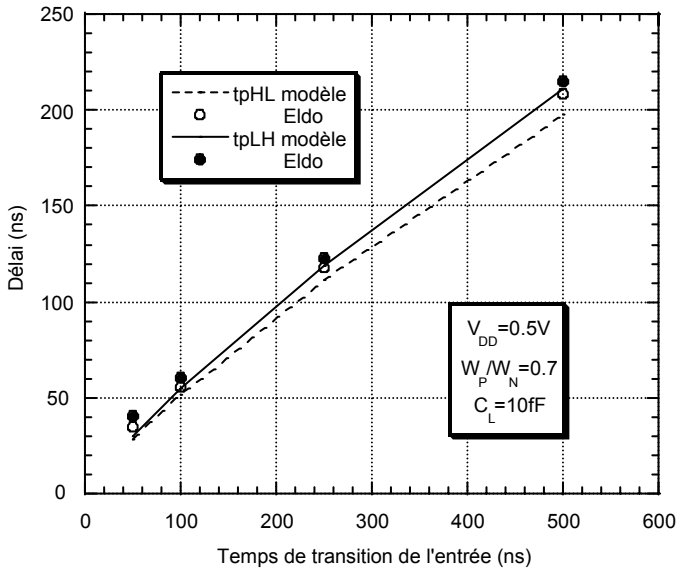


Figure 40 Comparaison des valeurs estimées et des valeurs simulées du temps de propagation d'un inverseur en fonction de la pente en entrée.

4.2.3 Oscillateur en anneau

Grâce aux formules développées précédemment, nous pouvons faire l'étude d'un oscillateur en anneau. Pour cela, il nous suffit d'additionner les temps de propagation tp_{LH} et tp_{HL} obtenus dans les Annexes 2 et 3, de manière à obtenir le temps séparant les transitions des sorties des inverseurs *i* et *i+1* par $V_{DD}/2$. La chaîne d'inverseurs est composée de dix inverseurs et d'une porte Nand. Les valeurs estimées par le modèle et celles simulées par Eldo sont montrées Figure 41 : les performances intrinsèques de la technologie SOI varient exponentiellement avec la tension d'alimentation V_{DD}.

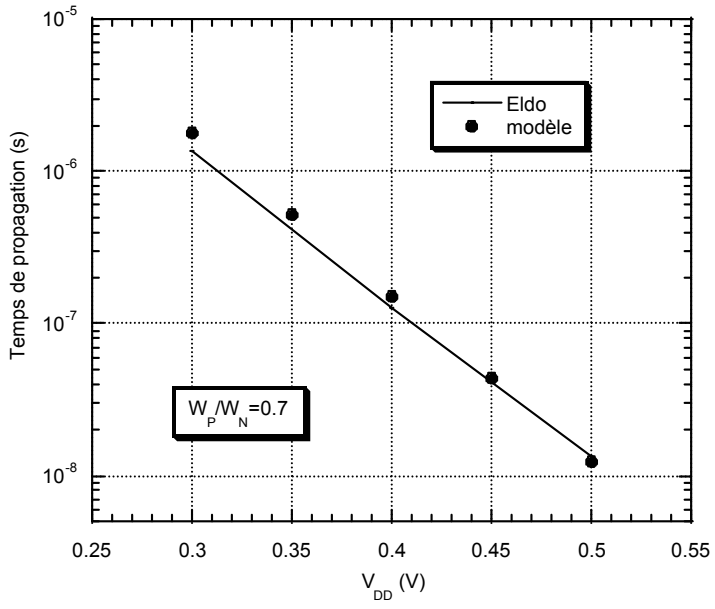


Figure 41 Performances intrinsèques de la technologie SOI en fonction de la tension d'alimentation V_{DD}.

4.3 Conclusion

La technologie SOI permet de réduire la consommation d'énergie des circuits microélectroniques en supprimant quasiment les capacités de jonction des transistors et donc en réduisant la capacité totale de ces circuits. De plus, elle s'avère très intéressante pour travailler à basse tension par rapport à la technologie CMOS sur silicium massif grâce à la modulation dynamique de la tension de seuil V_T et à l'élimination de l'effet source suiveur dans l'empilement de transistors, deux phénomènes qui permettent de gagner en vitesse et qui

sont d'autant plus bénéfiques que la tension d'alimentation V_{DD} est faible, c'est-à-dire la différence V_{DD}-V_T réduite.

Pour minimiser l'énergie dissipée et augmenter l'autonomie d'applications portables, nous faisons le choix de fixer la tension d'alimentation à 0,5V. Nous préférons les transistors HS aux transistors LL, qui permettent de gagner en performances tout en n'augmentant pas trop la consommation puisque la puissance statique due aux courants de fuite est petite devant la puissance dynamique. Nous avons dès lors développé un modèle physique simple pour estimer le comportement des transistors SOI opérés en faible inversion : ce modèle permet de calculer l'expression du temps de propagation de portes logiques et a été implémenté sous Matlab.

5 Circuits combinatoires et séquentiels

Le problème de la réduction de la puissance dissipée par les circuits microélectroniques peut être considéré à différents niveaux, du niveau le plus abstrait au plus détaillé, comme nous l'avons vu dans le chapitre 2. Au niveau circuit, le choix du style logique permet des gains potentiels intéressants. En effet, tous les paramètres importants concernant la dissipation d'énergie tels que le taux d'activité ou la valeur des capacités commutées sont directement influencés par le style logique.

Bien évidemment, le choix d'un style logique ne se résume pas à ces seules considérations : il faut également prendre en compte la vitesse et la surface des portes logiques réalisées de même que leur robustesse à diverses conditions de fonctionnement et leur immunité au bruit. Dans le cas de l'utilisation d'outils de synthèse logique pour réaliser les fonctions souhaitées, la facilité d'utilisation et la compatibilité avec les circuits environnants sont aussi des paramètres importants.

Dans une première partie, nous allons étudier les différentes familles de circuits combinatoires, en nous concentrant essentiellement sur les circuits statiques. Les contraintes pour une basse consommation seront brièvement rappelées et celles concernant la synthèse logique seront exposées. Puis, nous allons comparer plusieurs bascules afin de déterminer

laquelle possède le meilleur facteur de qualité. Pour finir, nous présenterons la bibliothèque de cellules standards qui a été développée.

5.1 Circuits combinatoires

5.1.1 Contraintes sur les styles logiques

Comme indiqué précédemment, réduire la tension nominale est le moyen le plus efficace de diminuer l'énergie consommée : c'est la raison pour laquelle nous travaillons avec une tension d'alimentation V_{DD} égale à 0.5V. Puisque la tension V_{DD} est inférieure à la somme des tensions de seuil des transistors NMOS et PMOS, il n'y a pas de courant de court-circuit : la source dominante de dissipation d'énergie en technologie CMOS est alors la puissance dynamique – voir Équation 1. Celle-ci dépend de la tension d'alimentation V_{DD}, de la fréquence d'horloge *f*, de l'activité *a* d'un nœud, de sa capacité C_{TOT} et du nombre de nœuds. On peut agir sur chacun de ces paramètres pour réduire la puissance consommée, et les différentes méthodes existantes ont été présentées au Chapitre 2. Cependant, dans le choix du style logique que l'on va faire, deux paramètres sont fixés :

- la tension d'alimentation V_{DD} est ici égale à 0.5V,
- la fréquence *f* est souvent considérée constante, au niveau circuit, pour respecter des contraintes de performance : la réduction de la fréquence d'horloge se décide au niveau architecture.

Tous les autres paramètres sont influencés par le style logique ; on peut alors énumérer les conditions suivantes pour une basse consommation :

- *Réduction des capacités commutées* : les capacités des nœuds C_L, dues aux lignes d'interconnexion et aux capacités de grille et de diffusion des transistors, doivent être réduites au minimum. Pour cela, la longueur et le nombre de connexions à l'intérieur d'une même cellule et donc sa surface doivent être minimaux. Une partie importante de la capacité d'un circuit logique étant due aux capacités des transistors, la réduction de la taille des transistors est un moyen très efficace de diminuer la valeur des capacités commutées : les transistors doivent être de taille minimale sur les chemins non critiques [Roge96] et ne doivent être augmentés que si leur capacité de charge est dominée par le facteur extrinsèque [Raba96]. Le style logique doit donc être robuste

quant au dimensionnement des transistors, c'est-à-dire garantir un fonctionnement correct avec des transistors minimaux.

- *Réduction du nombre de nœuds* : le nombre de nœuds est directement influencé par le style logique choisi. Il dépend bien évidemment du nombre de transistors mais aussi de la complexité de la logique : les structures différentielles présentent plus de nœuds fortement capacitifs.
- *Réduction du taux d'activité* : le taux d'activité peut être modifié principalement aux niveaux système et architecture. Au niveau circuit, on peut réduire le taux d'activité en choisissant une logique statique plutôt qu'une logique dynamique. En effet, dans le cas d'une porte statique, la sortie est stable tant que l'entrée ne varie pas : hormis le courant de conduction sous le seuil, l'énergie est dissipée dans les seules portes qui changent d'état. Dans le cas des logiques dynamiques, l'énergie est dissipée pendant et immédiatement après une variation de l'horloge : une porte peut consommer du courant même lorsque ses entrées ne varient pas ; le taux d'activité est donc plus élevé.

L'utilisation d'une tension d'alimentation très faible, proche des tensions de seuil des transistors, apporte une contrainte supplémentaire : la marge aux bruit des portes s'en trouve réduite et le style logique choisi doit être robuste pour garantir un fonctionnement correct.

Voyons maintenant les contraintes imposées par l'utilisation d'outils de synthèse logique. Les portes doivent présenter les caractéristiques électriques suivantes :

- *sortance* : les portes doivent avoir de bonnes capacités de charge – sortance – de manière à fonctionner correctement dans n'importe quelle configuration.
- *découplage entrée/sortie* : dans le cas où la sortie n'est pas découplée de l'entrée, les portes mises les unes à la suite des autres forment un réseau d'interrupteur. Les transistors n'étant pas des interrupteurs idéaux, puisque leur résistance série n'est pas nulle, le délai dans cette chaîne augmente quadratiquement, selon le délai d'Elmore [Elmo48]. Il faut alors limiter le nombre de portes en série en introduisant des inverseurs. Cependant, lorsqu'on utilise des outils de synthèse logique, on veut pouvoir placer les portes dans n'importe quel ordre : une inversion logique par étage est alors nécessaire. En conséquence, les entrées et sorties de chaque porte doivent être découplées pour satisfaire à la contrainte de la synthèse logique.

- *niveaux logiques corrects* : l'amplitude des signaux en sortie des portes doit être égale à la tension d'alimentation. Dans le cas des logiques à transistor de passage, des restaurateurs de niveau sont nécessaires pour assurer un fonctionnement fiable.

Comme nous allons le voir dans la suite, chaque style logique a ses avantages et ses inconvénients. Aucun ne permet d'optimiser à la fois la vitesse, la consommation et la surface, mais les contraintes concernant la basse tension et la basse consommation que nous avons énoncées précédemment vont nous permettre d'éliminer certains styles.

5.1.2 Styles logiques

Il nous faut d'abord faire la distinction entre la logique statique et la logique dynamique, qui ont des caractéristiques très différentes au niveau de leur dissipation d'énergie.

Logique dynamique

Le fonctionnement de la logique dynamique est basé sur le stockage d'une charge sur un nœud et la charge – ou décharge – conditionnelle de ce nœud en fonction des entrées. Une porte en logique dynamique est constituée d'un réseau de transistors NMOS ou PMOS pour réaliser la fonction logique et de deux transistors, de type opposé à celui de l'arbre logique, pour les phases de précharge et d'évaluation. Elle présente l'avantage d'avoir une capacité d'entrée ainsi qu'une capacité parasite de sortie faibles et donc des temps d'évaluation courts. Par contre, elle est sensible à des effets parasites tels que les courants de fuite – qui nécessitent de rafraîchir périodiquement le nœud de stockage et imposent donc une fréquence d'horloge minimale –, le partage de charges avec les transistors de l'arbre d'évaluation et les couplages capacitifs avec le signal d'horloge, dont les fronts sont très raides. La logique dynamique sacrifie donc la marge au bruit pour les performances.

La consommation d'une porte dynamique est importante : son taux d'activité est plus élevé que celui d'une porte statique puisque sa sortie peut varier à chaque coup d'horloge, même en l'absence de transitions sur ses entrées. De plus, la présence de nombreux transistors de précharge et d'évaluation induit une charge considérable sur l'arbre d'horloge, qui doit propager un signal dont l'activité est de 1. La logique dynamique n'est donc pas une logique basse consommation mais une logique haute performance [Chan95] et ne sera pas considérée par la suite.

Logique CMOS conventionnelle

Les portes en logique CMOS conventionnelle sont réalisées avec un arbre logique NMOS relié à la masse et un arbre complémentaire PMOS relié à l'alimentation V_{DD}. Grâce aux transistors complémentaires NMOS et PMOS, un seul arbre est passant quelque soit l'entrée – hormis pendant une transition. Il existe donc toujours une connexion directe entre la sortie et la masse ou V_{DD}. L'amplitude de variation de la tension de sortie est égale à la tension d'alimentation et l'impédance de sortie des portes est faible, ce qui résulte en de bonnes marges aux bruits. Les niveaux logiques en sortie ne dépendant pas des tailles des différents transistors, les transistors peuvent être minimaux. C'est un style logique appelé « sans ratio », par opposition aux « logiques à ratio », pour lesquelles les niveaux logiques sont déterminés par les dimensions relatives des transistors composant le circuit.

L'impédance d'entrée des portes réalisées en CMOS est idéalement infinie, puisque l'oxyde de grille d'un transistor est un isolant – on voit bien apparaître des courants tunnel pour des épaisseurs de grille faibles, mais ils restent pour l'instant négligeables devant les autres sources de courant. En statique, une porte CMOS peut donc attaquer un nombre très élevé d'autres portes, c'est-à-dire avoir une sortance très grande et rester fonctionnelle. Le nombre de portes en sortie doit néanmoins être limité car le délai se dégrade avec la sortance.

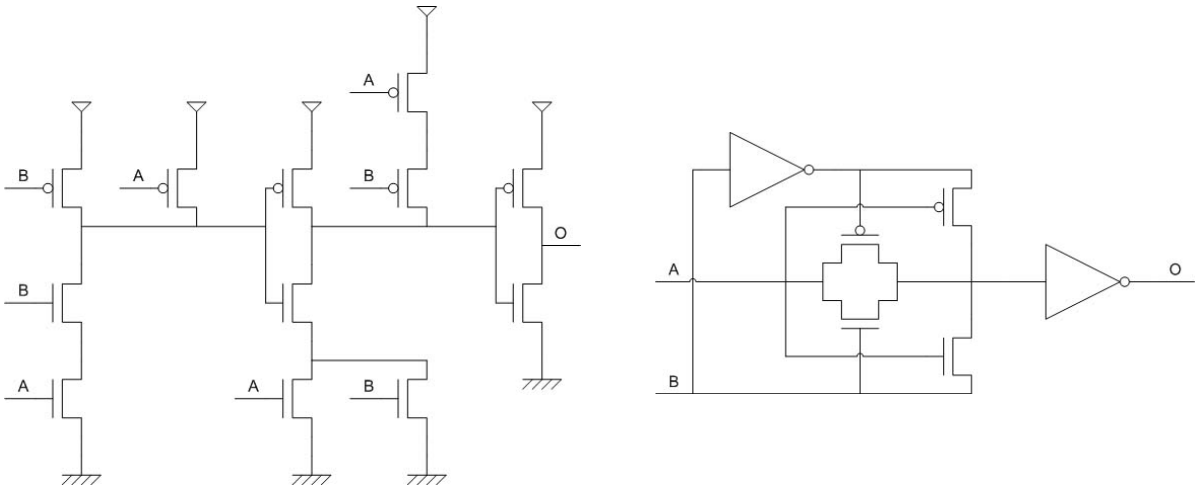
La robustesse d'un inverseur CMOS face à une réduction de la tension d'alimentation a été montrée dans [Swan72] : la caractéristique de transfert reste correcte jusqu'à une tension d'alimentation égale à 4kT/q, soit 100mV. Selon [Veen98], la marge au bruit d'un inverseur est supérieure ou égale à $0.42*V_{DD}$.

Des portes monotoniques simples, telles que des NAND/NOR ou AOI/OAI, peuvent être réalisées efficacement avec peu de transistors, résultant en une surface et une consommation restreintes. De plus, l'inversion du signal à chaque étage se faisant naturellement, il n'est pas nécessaire d'introduire un inverseur en sortie, ce qui améliore le délai. Les portes non monotoniques, telles que les XOR et les multiplexeurs, requièrent des implémentations plus complexes mais restent efficaces.

Les signaux d'entrée étant seulement reliés à des grilles, l'utilisation et la caractérisation des portes s'en trouvent facilitées, comme nous le verrons dans la suite. La conception de ces portes est également très directe, grâce aux paires de transistors NMOS et PMOS. Elles remplissent donc les contraintes de la facilité d'utilisation et de versatilité imposées par la synthèse logique.

L'inconvénient majeur de la logique CMOS conventionnelle est la présence des transistors PMOS : leurs porteurs majoritaires possédant une mobilité moindre que les électrons des transistors NMOS, il faut augmenter leur taille pour équilibrer les caractéristiques de transfert et les temps de propagation des portes. Les transistors PMOS deviennent prohibitivement gros dans le cas de portes à entrance élevée, telles les portes NOR, où ils sont placés en série. Cependant, l'entrance est généralement limitée à 3 et la réorganisation des fonctions logiques aide à pallier ce problème. Par exemple, l'équation $A+B+C+D$ peut s'écrire $\overline{(\overline{A+B})\cdot(\overline{C+D})}$: le délai introduit par la porte NAND2, mise à la place de l'inverseur, est largement compensé par le délai gagné dans les portes NOR2, puisque ce dernier diminue quadratiquement avec la sortance. De plus, en technologie sous-micronique, le rapport des mobilités tend à diminuer puisque les porteurs des transistors NMOS et PMOS atteignent leur vitesse de saturation : les meilleures performances des portes sont obtenues pour un rapport W_P/W_N égal à 1.5. Le problème posé par les transistors PMOS tend donc à s'amoindrir et peut dans tous les cas être géré par une réorganisation des fonctions logiques.

Des portes de passage, qui sont constituées d'un transistor NMOS et d'un transistor PMOS en parallèle, peuvent être utilisées pour améliorer l'efficacité des portes non monotoniques : ce style logique est appelé CMOS+. La Figure 42 montre une porte XOR réalisée en CMOS et en CMOS+ : la porte de passage permet d'avoir un circuit contenant moins de transistors.



(a) **(b)**

Figure 42 Porte XOR à deux entrées réalisée en : (a) CMOS et (b) CMOS+.

Comme nous l'avons vu, l'inconvénient majeur de la logique CMOS conventionnelle est la présence de gros transistors PMOS, même si ce problème tend à s'atténuer avec les nouvelles technologies. Des styles logiques ont donc été introduits dans le passé pour n'utiliser que des arbres NMOS dans la réalisation des fonctions logiques voulues.

Logiques à ratio

La technique pour réduire la complexité des circuits CMOS – c'est-à-dire réduire le nombre de transistors PMOS – consiste à n'utiliser que le réseau NMOS et à remplacer le réseau PMOS par une charge. Comme montré Figure 43, la charge peut être soit un élément passif, telle une résistance, soit un élément actif, tel un transistor. La première technologie ayant existé est celle utilisant une charge résistive ; cependant, la relation linéaire du courant avec la tension aux bornes de la résistance fait que le courant de charge chute rapidement lorsque la tension de sortie augmente. Le transistor NMOS à déplétion a été la première réponse pour trouver une meilleure charge – une charge qui se rapproche d'une source de courant, c'est-à-dire dont le courant fourni dépend peu de la tension à ses bornes. Finalement, la logique pseudo-NMOS est apparue avec le transistor complémentaire PMOS et s'est révélée encore meilleure que l'approche précédente, car la tension de grille ne chute pas lorsque la sortie augmente et le transistor n'est pas affecté par l'effet de substrat.

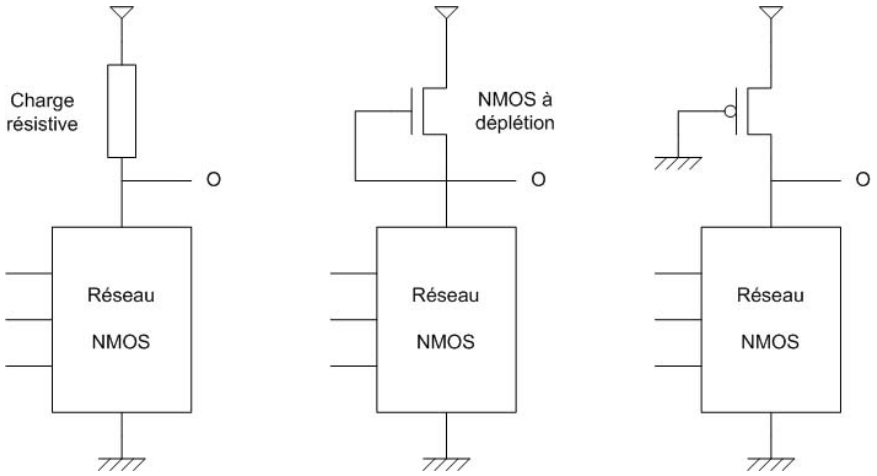


Figure 43 Différentes portes logiques à ratio : (a) charge résistive, (b) transistor NMOS à déplétion, (c) pseudo-NMOS.

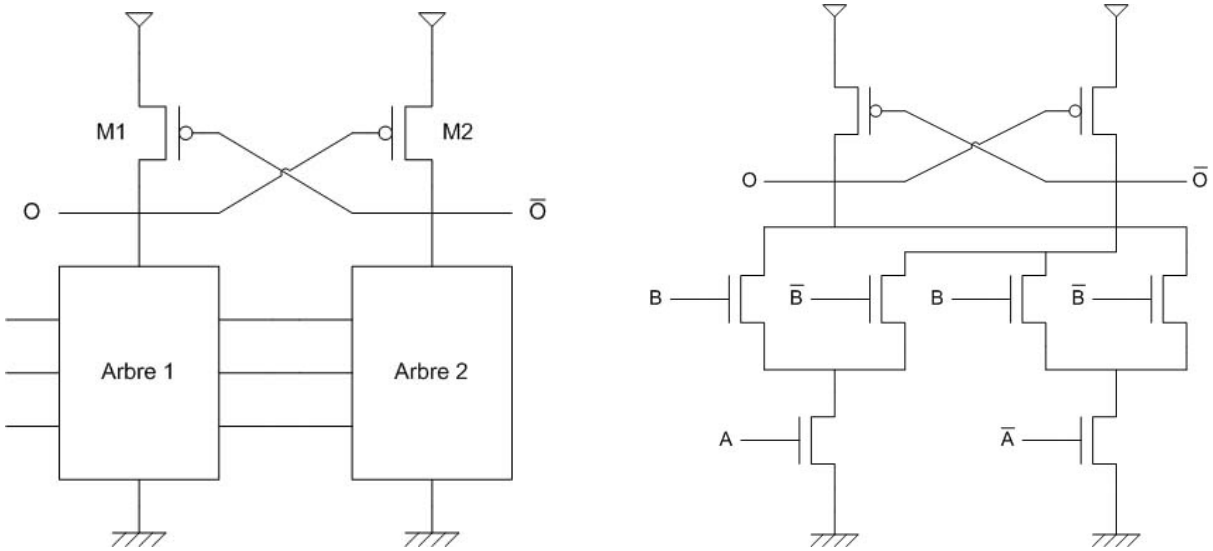
Il est évident que ces logiques ou technologies ne sont plus utilisées aujourd'hui car leur consommation, due au court-circuit statique lorsque l'arbre NMOS est passant, est bien trop considérable : elles ont depuis été avantageusement remplacées par la technologie CMOS.

Néanmoins, ce principe consistant à n'utiliser que des arbres NMOS peut être modifié de manière à ne plus dissiper de courant statique : cela a donné naissance à la logique *Differential Cascode Voltage Switch Logic* (DCVSL).

Logique DCVSL

Le principe de cette logique est d'utiliser deux arbres NMOS complémentaires [Hell84]. La Figure 44-a montre le principe de base de la porte ; les réseaux 1 et 2 implémentent la fonction logique voulue et son inverse. Le fonctionnement de la porte est le suivant : considérons que, en fonction des entrées, l'arbre 1 est passant et l'arbre 2 est bloqué. La sortie O est tirée à la masse, ce qui active le transistor M2 et amène la sortie $\overline{O}$ à la tension V_{DD}. Le transistor de charge M1 est alors coupé. Il n'y a alors pas de courant de court-circuit statique puisque seuls l'arbre 1 et le transistor M2 sont passants.

La logique DCVSL a l'avantage en vitesse de la logique pseudo-NMOS, les capacités parasites en sortie étant diminuées. Cependant, la consommation dynamique est élevée car il faut charger une des deux sorties à chaque changement d'état de la porte. La surface occupée est accrue par la présence des deux réseaux, ce qui est néanmoins en partie compensé par le fait que de la logique peut être partagée entre les deux arbres : cela est illustré à l'aide de la porte XOR à deux entrées dans la Figure 44-b.



(a) **(b)**

Figure 44 (a) Principe de base d'une porte DCVSL ; (b) Porte XOR à deux entrées.

Jusqu'à présent, nous n'avons vu que des logiques dont les entrées des portes sont reliées à des grilles : les logiques à transistors de passage permettent d'utiliser les transistors comme des interrupteurs et de réduire l'impédance d'entrée des portes.

Logiques à transistors de passage

Les logiques à transistors de passage utilisent les transistors comme des interrupteurs [Radh85]. A la différence de la logique CMOS conventionnelle, où les entrées ne sont connectées qu'aux grilles des transistors, les entrées des portes à transistors de passage peuvent être reliées aux sources de ces derniers – dont la capacité est réduite de 20% par rapport à la capacité de grille. La représentation schématique d'une telle porte est donnée Figure 45-a. Puisque le réseau n'est pas relié aux sources d'alimentation, la présence d'un inverseur en sortie est nécessaire : c'est la contrainte du découplage entrée/sortie précédemment énoncée.

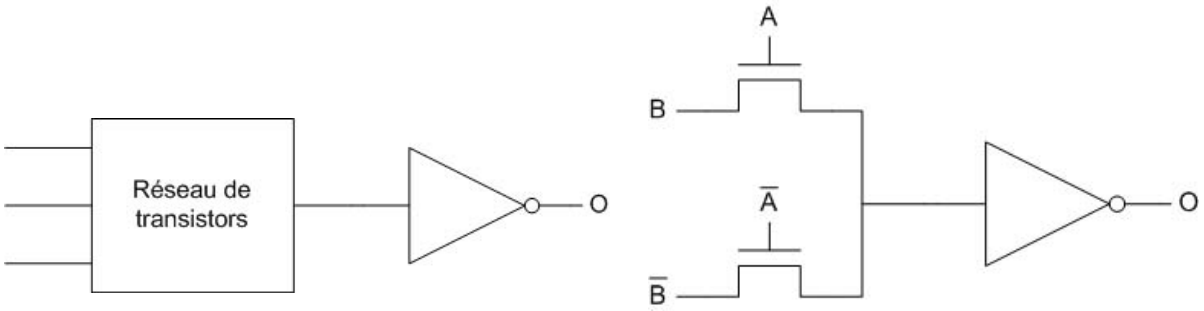


Figure 45 (a) Principe de base d'une porte à transistors de passage ; (b) Porte XOR à deux entrées.

Une porte XOR à deux entrées est représentée Figure 45-b : l'utilisation de seulement deux transistors permet de réaliser la fonction logique $A \cdot B + \overline{A} \cdot \overline{B}$, ce qui est très intéressant au niveau de la surface occupée et de la valeur des capacités à charger. L'inconvénient de cette logique est la chute de potentiel V_{DD}-V_T à travers les transistors, ce qui dégrade les niveaux logiques. L'inverseur est aussi là, dans une certaine mesure, pour restaurer les niveaux ; cependant, le niveau logique '1' faible entraîne un courant de court-circuit statique dans l'inverseur de sortie. Pour remédier à cela, un restaurateur de niveau doit être utilisé.

Logique à transistors de passage et restaurateurs de niveau [Yano96]

Un restaurateur de niveau est un transistor PMOS relié à la tension d'alimentation V_{DD} et utilisé dans une boucle de rétroaction. Le schéma de la Figure 45-b est modifié tel que montré dans la Figure 46. La sortie de l'inverseur commande la grille du transistor, le bloquant

lorsque le réseau NMOS passe le niveau logique ‘0’ et l’activant lorsque la sortie du réseau approche du niveau logique ‘1’. Le dimensionnement du restaurateur de niveau doit être considéré avec précaution : en effet, bien qu’il ne s’agisse pas d’une logique à ratio puisqu’il n’y a jamais de chemin statique allant de l’alimentation V_{DD} à la masse, un conflit existe néanmoins entre le transistor PMOS et les transistors NMOS du réseau lors d’un changement d’état de la porte. Pour que le réseau NMOS puisse écrire un ‘0’ logique sur le nœud interne situé devant l’inverseur, c’est-à-dire que la tension de ce nœud chute suffisamment pour faire basculer l’inverseur, il faut que le transistor PMOS soit sensiblement de la même taille que les transistors NMOS [Raba96]. Le transistor PMOS présente l’inconvénient de ralentir la porte lors d’une transition bas vers haut de la sortie mais l’avantage de l’accélérer dans l’autre sens.

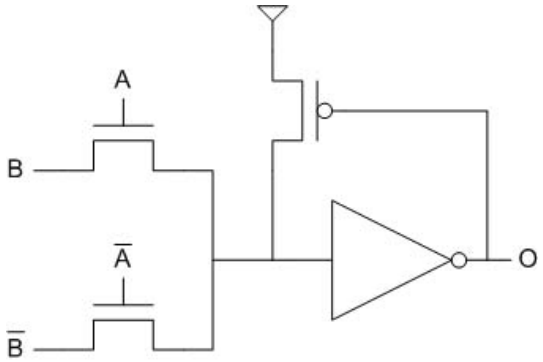


Figure 46 Porte XOR à transistors de passage et restaurateur de niveau.

Le problème majeur de ce style logique reste néanmoins la chute de potentiel à travers les transistors NMOS pour des applications très basse tension. La structure de restauration du niveau logique ne fonctionne plus pour une tension d’alimentation $V_{DD}<V_{TN}+|V_{TP}|$, car la sortie de l’inverseur ne peut plus basculer au niveau logique ‘0’ et activer le transistor PMOS. Pour respecter la contrainte d’une très faible tension d’alimentation, les seules logiques à base de transistor de passage qui fonctionnent sont donc les logiques à base de porte de passage ou les logiques différentielles. Nous allons les détailler dans la suite.

Logique à portes de passage

L’élément de base est une porte de transmission, composée d’un transistor NMOS et d’un transistor PMOS en parallèle. Le transistor NMOS va passer correctement le niveau logique ‘0’ et le transistor PMOS le niveau logique ‘1’, de sorte qu’il n’y a pas plus de chute de potentiel due à la tension de seuil. Une porte de passage est bidirectionnelle, c’est-à-dire qu’elle peut transmettre une information indifféremment dans un sens ou l’autre, sur activation du signal de contrôle. La Figure 47 montre un exemple de circuit, une porte XOR à

deux entrées. Lorsque l'entrée B vaut '1', les deux transistors, dont les grilles sont reliées à A, se comportent en inverseur, chacun ayant à passer la bonne valeur logique : la fonction réalisée est $\overline{A} \cdot B$. Lorsque B vaut 0, ils sont bloqués et la porte de passage est active : la sortie est alors égale à $A \cdot \overline{B}$.

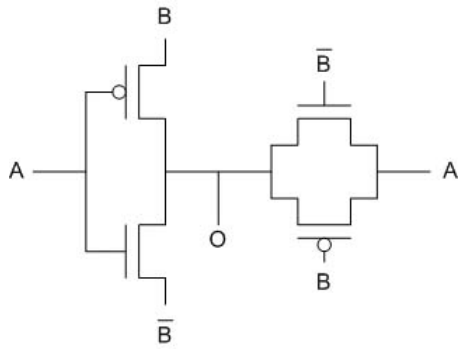


Figure 47 Porte XOR deux entrées à porte de passage.

Pour ne pas avoir à générer les entrées complémentées dans chaque porte, les portes de passage peuvent être utilisées dans deux arbres complémentaires : ce style logique est appelé D*ouble Pass-transistor Logic* (DPL) [Suzu93]. La Figure 48 donne l'exemple d'une porte XOR. Néanmoins, la logique DPL réalise une structure différentielle relativement inefficace, comportant beaucoup de transistors PMOS et beaucoup de nœuds ce qui augmente le délai, la surface et la consommation. Elle n'est pas compétitive face à la logique différentielle CPL que nous allons voir dans la suite.

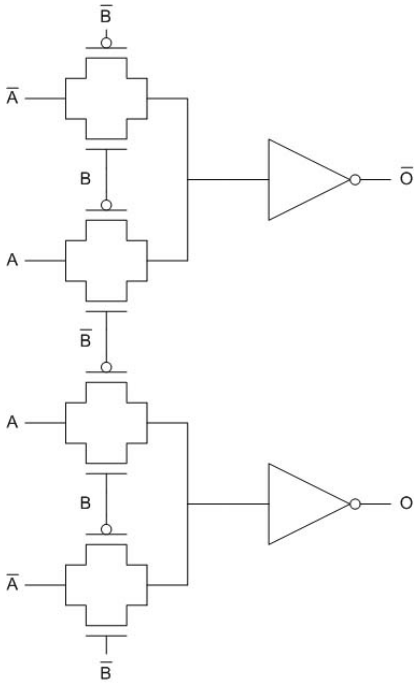


Figure 48 Porte XOR à deux entrées en logique DPL.

Logique à transistors de passage complémentaires

La logique *Complementary Pass-Transistor Logic* (CPL) utilise deux arbres complémentaires de transistors de passage NMOS et des restaurateurs de niveau – voir l'exemple d'une porte XOR dans la Figure 49. Les arbres complémentaires sont utilisés pour générer des sorties complémentaires et ainsi éviter d'avoir à inverser chaque signal d'entrée. Les avantages de la logique CPL sont une faible capacité d'entrée, qui diminue la consommation dynamique, une bonne capacité de charge grâce aux inverseurs de sortie et une structure différentielle rétro-bouclée, les deux derniers points contribuant à améliorer le délai. A l'inverse, la structure différentielle induit de forts courants de court-circuit, à cause des deux restaurateurs de niveau. De plus, le nombre de nœuds, le nombre d'interconnexion et la surface occupée sont plus élevés à cause des signaux complémentaires.

La logique CPL permet de réaliser efficacement des portes complexes telles que des portes XOR, des additionneurs un bit ou des multiplexeurs. A l'inverse, elle n'est pas compétitive par rapport au CMOS conventionnel pour des portes monotoniques simples, comme des NAND ou des NOR.

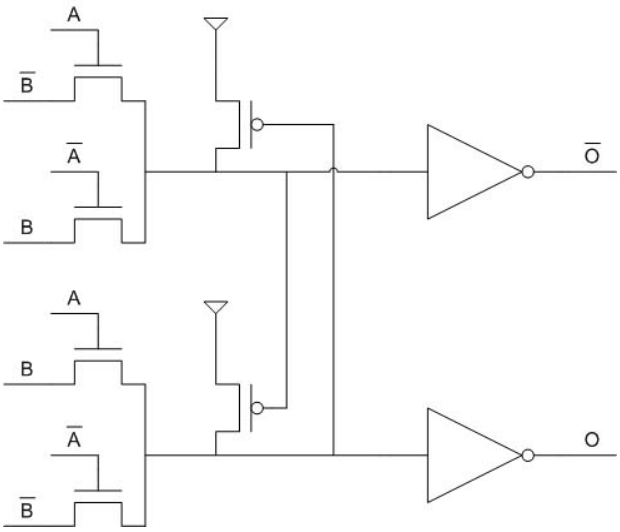


Figure 49 Porte XOR à deux entrées en logique CPL.

5.1.3 Comparaison des styles logiques

Styles logiques retenus

De tous les styles logiques présentés plus haut, qui ne constituent pas une liste exhaustive mais représentent les grandes familles logiques pouvant être considérées, nous n'allons comparer que ceux qui respectent les contraintes résultant de l'utilisation d'une basse tension d'alimentation et d'outils de synthèse logique.

Les logiques à ratio sont bien évidemment à proscrire car elles présentent une forte consommation statique et une marge au bruit très faible. La logique DCVSL en est une évolution directe : elle présente une bien meilleure charge – absence de courants de court-circuit – grâce à sa structure différentielle. La capacité d'entrée de ces portes est diminuée du fait de la seule présence de l'arbre NMOS pour réaliser la fonction logique : la sortance de la porte précédente est par conséquent réduite, améliorant les performances de ce style logique. Cependant, le conflit existant, lors d'une transition, entre les transistors de charge et les arbres d'évaluation anéantit cet avantage en vitesse. La consommation dynamique de cette logique est élevée à cause de la structure différentielle : il faut charger une des deux sorties à chaque transition de la porte, ce qui augmente l'activité. La logique DCVSL n'est pas compétitive au niveau de l'énergie dissipée par rapport au CMOS conventionnel et se montre équivalente au niveau du délai [Chu87] : on va donc en ce point abandonner ce style logique.

Les logiques à base de transistors de passage s'avèrent beaucoup plus intéressantes, notamment en technologie SOI, pour laquelle les capacités de jonctions sont réduites d'un facteur 10 par rapport au silicium passif. Utiliser les transistors comme des multiplexeurs permet donc de réaliser des portes ayant une faible capacité d'entrée. Cependant, il y a plusieurs limitations, comme nous l'avons vu :

- des inverseurs doivent être introduits pour découpler les sorties des entrées, ce qui diminue l'avantage en vitesse inhérent à ce style logique,
- la chute de potentiel V_{DD}-V_{TN} à travers un transistor NMOS détruit la marge au bruit et engendre une consommation statique dans l'inverseur de sortie : il faut donc utiliser, soit des portes de passage, soit des restaurateurs de niveau.

Les portes de passage présentent l'inconvénient de nécessiter des signaux complémentaires. Pour ne pas avoir à inverser les entrées de chaque porte, un style différentiel doit être utilisé : c'est le cas de la logique DPL. Ce style logique présente une très bonne marge au bruit,

puisque les niveaux logiques ‘0’ et ‘1’ sont correctement transmis. Les portes ont également une bonne capacité de charge grâce aux inverseurs de sortie. Néanmoins, la logique DPL réalise une structure différentielle relativement inefficace, puisqu’elle nécessite autant de transistors PMOS que de transistors NMOS. Il y a donc un nombre important de nœuds présentant des charges élevées. La nécessité d’inverser les signaux de sorties pour découpler les sorties des entrées ralentit un peu plus ces portes. La logique différentielle à base de portes de passage n’est donc pas compétitive face aux styles logiques suivants et n’est pas considérée dans la comparaison.

La logique CPL représente la seule alternative valable – à base de transistors de passage – à la logique CMOS conventionnelle, présentant de faibles capacités d’entrée, une bonne sortance grâce aux inverseurs de sortie et permettant de réaliser efficacement des portes non monotoniques. La faible impédance d’entrée du CPL est naturellement améliorée en technologie SOI. L’inconvénient de cette logique est sa marge au bruit réduite – présence de restaurateurs de niveaux – et sa structure différentielle qui, tout en étant censée améliorer le délai, augmente la puissance dynamique.

Finalement, la logique qui semble la mieux adaptée à de faibles tensions d’alimentation est la logique CMOS conventionnelle. Celle-ci est robuste, tant au niveau de la marge au bruit que du dimensionnement des transistors. Elle réalise une inversion logique à chaque étage ce qui évite le recours à un inverseur en sortie et améliore le délai. Les portes simples sont implémentées avec peu de transistors ce qui diminue la surface et la consommation. De plus, les portes plus complexes peuvent être réalisées à l’aide de portes de passage – CMOS+. La logique CMOS+ peut être vue comme une logique asymétrique, à base de portes de passage, par opposition au DPL, qui est différentiel.

Nous allons comparer dans la suite les logiques les plus appropriées à une utilisation basse tension et basse énergie, à savoir les styles CMOS, CMOS+ et CPL.

Arrangement des portes pour une comparaison équitable

Des conditions de simulation réalistes doivent être réunies pour comparer les différents styles logiques. En particulier, il ne faut pas commander les entrées d’une porte directement par le simulateur : les pentes que ce dernier introduit sont idéales et ne varient pas en fonction de l’impédance d’entrée du circuit. On avantagerait ainsi les logiques ayant une forte impédance d’entrée. Pour la même raison, il faut charger la sortie de la porte analysée à l’aide d’une ou plusieurs autres portes identiques et non pas avec des capacités de valeur fixe. La

configuration utilisée pour la comparaison des styles logiques est montrée Figure 50 : la porte analysée est ici la porte I2. Elle est commandée par les sorties des portes I0 et I1, qui génèrent les signaux A et B, et doit charger deux portes, I4 et I5. La porte I3 sert à introduire une sortance correcte sur les sorties de I0 et I1. Une sortance de deux est choisie pour avoir un cas réaliste. Des capacités d'interconnexion de 5fF sont ajoutées. On peut ainsi, pour chaque style, prendre en compte l'influence de l'impédance d'entrée ou encore la capacité de charge.

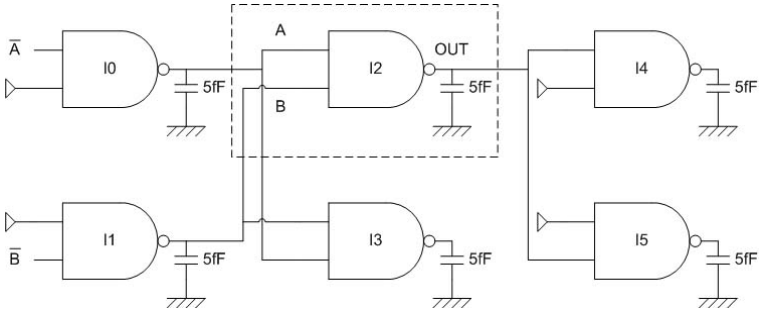


Figure 50 Arrangement des portes pour avoir des conditions de simulation réalistes.

Résultats

Les portes sont simulées en présentant en entrée toutes les transitions possibles, soit 2^{2N} combinaisons, avec N le nombre d'entrées. Le délai de la porte est choisi comme étant le délai le plus élevé. La puissance dissipée mesurée est la puissance moyenne sur toutes les transitions. Pour comparer équitablement les trois styles, plusieurs types de portes sont analysés : des portes monotoniques telles que NAND2 et NOR3, plutôt favorables au CMOS et des structures de type multiplexeur, telles que XOR2 et MUX4, qui sont à l'inverse avantageuses pour le CPL. Une porte OR6 est considérée pour évaluer l'impact de l'empilement de transistors et de la technique de réorganisation logique. De même, un additionneur un bit est ajouté à la liste des portes car il est souvent utilisé dans la comparaison de styles logiques, étant un élément essentiel dans nombre de fonctions arithmétiques, notamment les additionneurs finaux – Ripple Carry, Carry Bypass, Carry Select – et les multiplicateurs.

Les tailles des transistors sont dimensionnées de manière à obtenir le plus petit facteur de qualité, i.e. le produit puissance délai. Les résultats concernant le délai, la puissance et l'énergie dissipée sont présentés dans le Tableau 4. Comme on peut le remarquer, le style CMOS+ n'est pas utilisé pour les trois premières portes : en effet, ce n'est qu'une évolution du CMOS en lui associant des portes de passage. Or les portes monotoniques simples ne peuvent pas être améliorées en utilisant des portes de passage, c'est-à-dire posséder moins de

transistors. En revanche, la structure de portes de type multiplexeur peut être simplifiée grâce à la logique CMOS+.

Tableau 4 Comparaison des styles logiques CMOS, CMOS+ et CPL.

Fonction réalisée	Style logique	Délai (ns)	Puissance (nW)	PDP (aJ)
NAND2	CMOS	0,69	62,5	43,1
	CPL	0,78	111,1	86,7
NOR3	CMOS	0,98	31,6	31
	CPL	0,99	55,7	55,1
OR6	CMOS	1,52	7,5	11,4
	CMOS-reorg	1,09	14,5	15,8
	CPL	1,95	18,7	36,5
	CPL-reorg	1,59	29,8	47,4
XOR2	CMOS	0,76	78,5	59,7
	CMOS+	0,82	72,4	59,4
	CPL	0,79	121,2	95,7
MUX4	CMOS	1,11	108,5	120,4
	CMOS+	0,94	107,1	100,7
	CPL	0,87	148,2	128,9
FA	CMOS	1,22	152,7	186,3
	CMOS+	1,24	156,8	194,4
	CPL	1,2	269,2	323

Pour les portes NAND2 et NOR3, la logique CMOS a une dissipation de puissance quasiment divisée par un facteur deux par rapport à la logique CPL, ce qui est dû à la structure différentielle de cette dernière. Le délai est également meilleur pour le CMOS car la logique CPL possède une faible marge au bruit et le conflit qui en résulte, lors d'une transition, entre les restaurateurs de niveau et les arbres d'évaluation NMOS, ralentit ces portes.

On retrouve sensiblement les mêmes rapports pour la porte OR6. On considère, en plus, ici la réorganisation logique consistant à remplacer cette porte, où il y a beaucoup – trop – de transistors empilés, par deux portes NOR3 en parallèle et une NAND2 : c'est ce qui est appelé

CMOS-reorg et CPL-reorg dans la colonne des styles logiques. On constate, pour la logique CMOS comme pour la CPL, que le délai est amélioré de l'ordre de 25% grâce à la diminution de l'empilement ; cependant, la puissance dynamique est presque doublée à cause de l'augmentation du taux d'activité, due à la mise en parallèle des deux portes NOR3 : une entrée seulement sur trois peut désormais faire changer la sortie, contre une sur six auparavant. Le produit puissance délai est donc moins bon suite à la réorganisation logique, mais celle-ci demeure souhaitable pour améliorer les fronts montant et descendant du signal de sortie – même si, grâce à la faible tension d'alimentation V_{DD} utilisée, on ne doit pas s'inquiéter des courants de court-circuit statique induits dans la porte suivante.

Pour la porte XOR2, pourtant défavorable à la logique CMOS, le délai est quasiment identique, avec un très léger avantage au CMOS. La puissance dynamique, quant à elle, est 50% plus élevée pour le CPL. Le style logique CMOS+ n'améliore pas l'énergie dissipée par rapport au CMOS, car le délai et la puissance dynamique sont sensiblement identiques. En revanche, le CMOS+ est avantageux pour le multiplexeur MUX4 : l'énergie dissipée est diminuée de 22% et 17% par rapport au CPL et CMOS respectivement, la porte en CPL étant la plus rapide.

Finalement, concernant l'additionneur final, c'est la logique CMOS qui est la plus intéressante car elle est plus rapide et consomme moins que le CPL et le CMOS+.

5.1.4 Conclusion

La logique CMOS est supérieure dans presque tous les points à la logique CPL. Elle présente une meilleure marge au bruit, possède un découplage entrée/sortie inhérent et consomme dynamiquement moins quelque soit le type de porte, dans un facteur allant jusqu'à deux. Son temps de propagation est presque toujours inférieur, sauf dans un cas très avantageux pour le CPL – le multiplexeur. De plus, son intégration dans des outils de synthèse logique est aisée et cette logique se caractérise facilement. L'ajout de portes de passage – logique CMOS+ – ne s'est avéré intéressant que dans le cas du multiplexeur. Pour la porte XOR2, et donc pour l'additionneur 1 bit, le produit puissance délai n'est pas amélioré.

La logique CMOS conventionnelle a donc encore de beaux jours devant elle : les logiques « exotiques » qui permettaient des gains en vitesse et/ou en consommation à des tensions d'alimentation élevées, tel que le CPL, ne fonctionnent plus ou voient leur gain anéanti. Le

principal avantage du CMOS reste finalement sa robustesse, ce qui se vérifiera d'autant plus dans les technologies à venir, où la sur-polarisation de grille, la différence V_{DD}-V_T, diminue.

5.2 Circuits séquentiels

La plupart des circuits VLSI sont des circuits synchrones : un signal d'horloge synchronise le flot de données et réduit la complexité du système. De plus, le fait d'introduire des registres entre des unités fonctionnelles (Unité Arithmétique et Logique, …) permet de tester facilement le circuit après sa fabrication : les registres peuvent être chargés et lus depuis l'extérieur. Les circuits mémoire sont donc des éléments très importants d'une bibliothèque de cellules standards.

Avant d'étudier des cellules séquentielles, il faut déterminer le type d'horloge que l'on va utiliser, à savoir une horloge monophase ou multiphase. La Figure 51 montre une machine à états finis générique, à horloge monophase. L'élément mémorisant de type D peut-être soit un latch, soit une bascule, c'est-à-dire être actif sur niveau ou actif sur front. Dans le cas du latch, le problème de la transparence de l'élément mémorisant impose une contrainte importante sur la largeur de la période de transparence de l'horloge. En effet, cette période doit être suffisamment courte pour que la donnée n'ait pas le temps de retraverser le bloc logique et suffisamment longue pour que le latch ait le temps de mémoriser la donnée. Pour réduire la contrainte au niveau du temps et ainsi simplifier la conception de circuits VLSI, des bascules sont utilisées en lieu et place des latchs : la donnée en entrée est mémorisée sur un front de l'horloge et ne peut pas traverser le bloc logique plus d'une fois dans une demie période d'horloge.

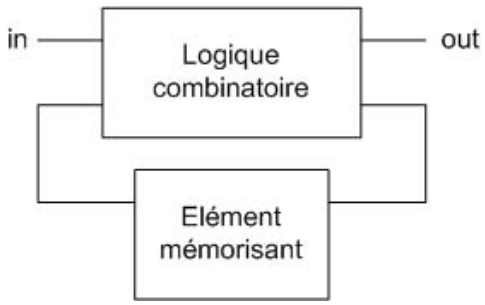


Figure 51 Machine à états finis générique.

La Figure 52 illustre un système biphase à horloges non recouvrantes. Dans ce cas, l'utilisation de latch est aisée puisqu'il n'y a pas de contrainte de temps : un des latchs est transparent pendant que l'autre mémorise. Une autre façon d'exploiter une horloge biphase est d'employer une bascule dont la partie maître est commandée par une horloge et la partie esclave par une autre. La difficulté avec l'approche biphase consiste à s'assurer que les horloges ne se recouvrent pas, ce qui rendrait transparents la bascule ou les deux latchs, simultanément.

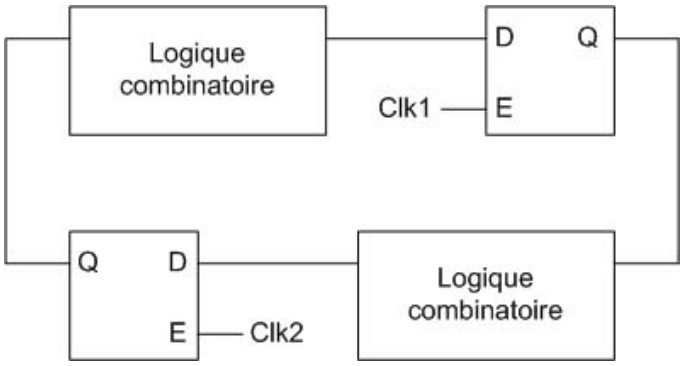


Figure 52 Structure biphase à horloges non recouvrantes et latchs.

Dans l'analyse des éléments mémoire qui suit, nous allons considérer uniquement le cas de l'horloge monophase, qui est beaucoup plus simple à implémenter en utilisant des outils de synthèse et évite le routage de deux signaux d'horloge.

5.2.1 Analyse

Nous allons ici rappeler les définitions des paramètres temporels d'une bascule tels qu'ils ont été définis par Unger et Tan dans [Unge86]. Ces paramètres sont au nombre de trois :

- délai *Clk*-*Q* : temps de propagation entre le terminal de l'horloge *Clk* et le terminal de sortie *Q*, en considérant que l'entrée *D* est restée stable suffisamment longtemps avant le front d'horloge ;
- temps de *setup* : temps minimum pendant lequel l'entrée *D* doit rester stable avant un front de l'horloge, de manière à assurer que la sortie *Q* deviendra égale à la nouvelle valeur de *D* ;
- temps de *hold* : temps minimum pendant lequel l'entrée *D* doit rester constante après un front de l'horloge, de manière à assurer que la sortie *Q* restera stable.

Pour mesurer les performances d'une bascule, le seul paramètre *Clk*-*Q* n'est pas suffisant. En effet, il ne tient pas compte du temps de *setup* nécessaire pour que la sortie acquière la bonne

valeur. La nouvelle donnée doit pouvoir être évaluée aussi près que possible d'un front d'horloge pour exploiter au maximum la période d'horloge. Ainsi, le temps de *setup* doit être comptabilisé dans le délai total de l'élément mémoire, si bien que l'on choisit le délai *D-Q* pour en évaluer les performances.

Le facteur de qualité considéré est le produit puissance délai. Pour extraire les paramètres de dissipation de puissance intéressants, le montage de la Figure 53 est utilisé dans l'étude des bascules. Ce montage assure des conditions de simulation assez proches de ce que l'on peut trouver dans un circuit. Les inverseurs en entrée permettent d'avoir des signaux d'horloge et de donnée réalistes, alors qu'ils sont issus, à l'origine, de sources de tension idéales. La capacité de sortie simule la sortance des étages suivants et est équivalente à la capacité de grille de vingt inverseurs de taille minimale.

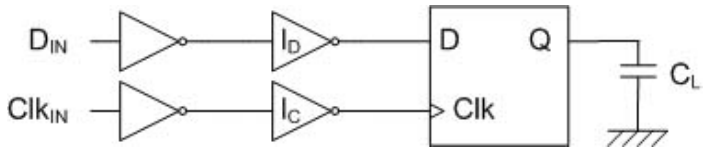


Figure 53 Montage utilisé pour obtenir une simulation réaliste et extraire les différents paramètres de consommation d'une bascule.

Trois sources de dissipation de puissance peuvent être isolées : la puissance consommée sur l'arbre d'horloge, celle consommée par la donnée et enfin la puissance interne de la bascule. Ces trois sources sont mesurées de manière à avoir une idée générale de la dissipation de puissance à l'intérieur et autour de la bascule :

- l'énergie consommée localement par l'horloge est mesurée grâce à l'inverseur I_C. Pour obtenir une valeur plus exacte, on retire la consommation de ce même inverseur à vide, de manière à soustraire la consommation introduite par la charge et la décharge de la capacité de sortie parasite de l'inverseur ;
- l'énergie consommée par la donnée est mesurée par l'inverseur I_D. De la même manière que précédemment, on retire sa consommation à vide ;
- la consommation interne de la bascule est calculée en mesurant le courant provenant de l'alimentation et en y soustrayant le courant consommé par la capacité de sortie. La puissance consommée par la charge de la capacité de sortie étant du même ordre de grandeur que la puissance intrinsèque d'une bascule, la différence de consommation entre les bascules serait atténuée si on la prenait en considération. Pour cela, on retire

la moitié de la valeur absolue du courant passant dans la capacité de sortie, puisqu'il est égal, au signe près, lors de la charge et lors de la décharge.

La consommation d'un circuit dépend du vecteur que l'on applique en entrée et notamment de son taux d'activité α – qui représente le nombre moyen de transitions de la donnée par période d'horloge. Un taux d'activité de 0.5, avec autant de transitions vers le haut que vers le bas, reflète la consommation interne moyenne d'une bascule. Un taux de 1 va entraîner une consommation maximale tandis qu'un taux de 0 reflètera la consommation minimale. Des exemples de ces différents vecteurs sont illustrés Figure 54.

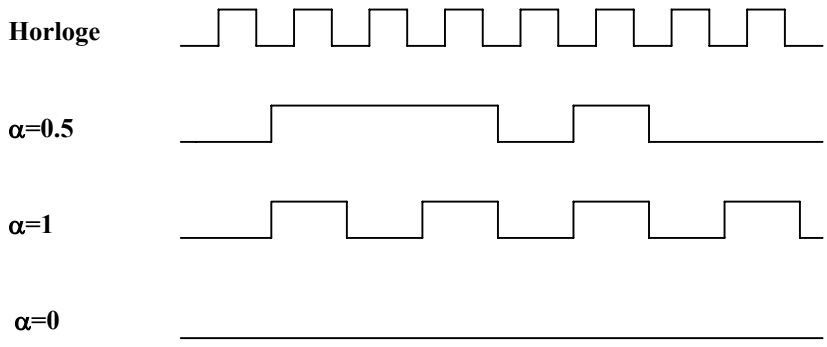


Figure 54 Les différents vecteurs utilisés en entrée.

5.2.2 Comparaison

Puisque les applications visées sont basse consommation, seules des bascules maître-esclave statiques sont considérées : les bascules semi-dynamiques ou hybrides[4] sont réservées aux applications haute performance [Stoj99]. En outre, les structures différentielles préchargées consomment à chaque cycle d'horloge, même lorsque la donnée en entrée ne varie pas.

Nous allons, dans la suite, comparer trois latchs maître-esclave : leur dimensionnement a été effectué avec le produit puissance délai pour objectif. La première bascule présentée est une modification du latch dynamique C²MOS : elle est représentée Figure 55. Les principaux avantages de cette structure sont :

- une faible charge introduite sur l'arbre d'horloge grâce à la bufferisation interne,

[4] Les bascules hybrides sont en fait des latchs dont la période de transparence est très réduite : cette période de transparence est définie par le front de l'horloge.

- une bonne robustesse à la variation de la pente du signal d'horloge,
- des inverseurs de rétroaction commandés par l'horloge qui assurent un fonctionnement statique.

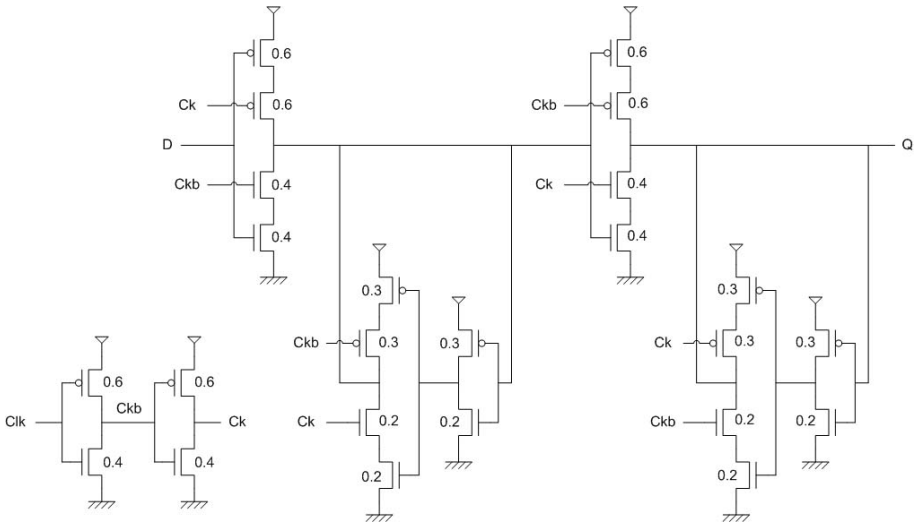


Figure 55 Bascule maître-esclave mC²MOS : les tailles des transistors sont indiquées en micromètres.

La deuxième bascule est celle du Power PC [Gero94]. C'est une des structures classiques les plus rapides car elle est basée sur des portes de passages et a donc un chemin direct très court, comme on peut le voir Figure 56. De part la charge élevée que cette bascule introduit sur l'arbre d'horloge et le fait de devoir générer deux horloges complémentaires, une bufferisation interne est là aussi préférée.

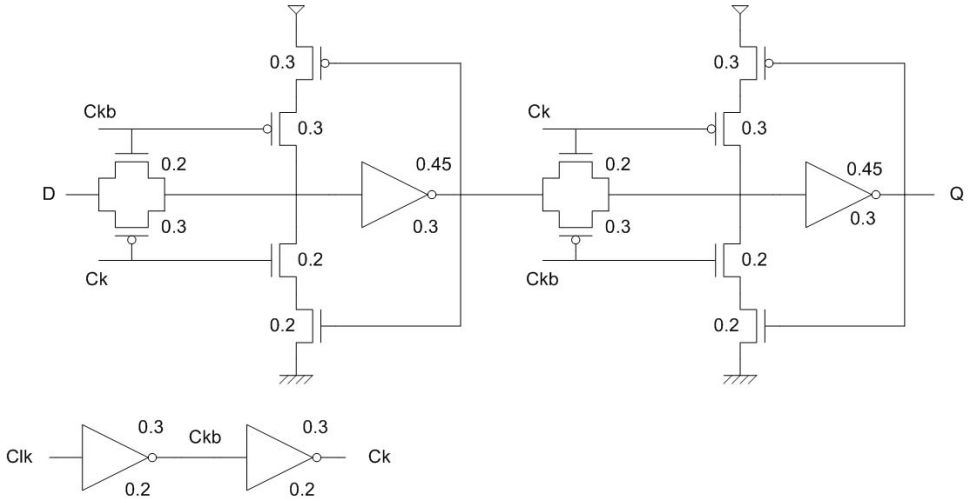


Figure 56 Bascule maître-esclave du PowerPC.

Enfin, la dernière bascule maître-esclave est basée sur des latchs TSPC (*True Single Phase Clocking*) [Yuan89] et est illustrée Figure 57. Elle est composée de quatre inverseurs en série et dissipe donc une faible puissance interne. Cependant, il faut faire attention que la pente du signal d'horloge ne soit pas trop lente pour éviter des problèmes de transparence. De plus, le stockage de la donnée est dynamique puisqu'il n'y a pas de boucle de rétroaction : la donnée peut alors être perdue si la fréquence de fonctionnement est trop faible.

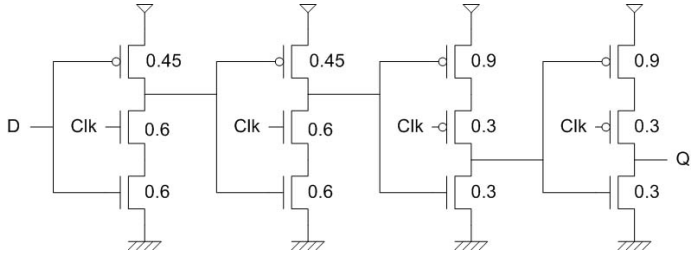


Figure 57 Bascule maître-esclave TSPC (*True Single Phase Clocking*).

Le Tableau 5 donne les résultats des simulations concernant les performances des trois bascules. Le taux d'activité de 0,5 présenté Figure 54 est utilisé et la fréquence d'horloge est de 100MHz. On se rend bien compte ici que le paramètre à considérer pour établir la performance d'une bascule est le délai D-Q et non pas le délai Clk-Q. En effet, si l'on regarde simplement le délai Clk-Q, la bascule TSPC est la plus rapide. Or cette dernière est affectée par un temps de *setup* très important, devenant ainsi l'élément mémorisant le plus lent. La bascule la plus rapide est celle du Power PC, grâce aux portes de passage offrant un chemin direct court. Au niveau de la dissipation de puissance, la bascule mC²MOS est désavantagée par la bufferisation interne du signal d'horloge et par le nombre important de transistors (24) qui la composent, et ainsi présente la puissance la plus élevée. La bascule TSPC étant celle qui contient le moins de transistors, c'est tout logiquement celle qui dissipe le moins d'énergie. Néanmoins, au niveau de l'énergie dissipée, c'est-à-dire du produit puissance délai, la bascule du Power PC s'avère la plus intéressante, juste devant la bascule TSPC et la bascule mC²MOS.

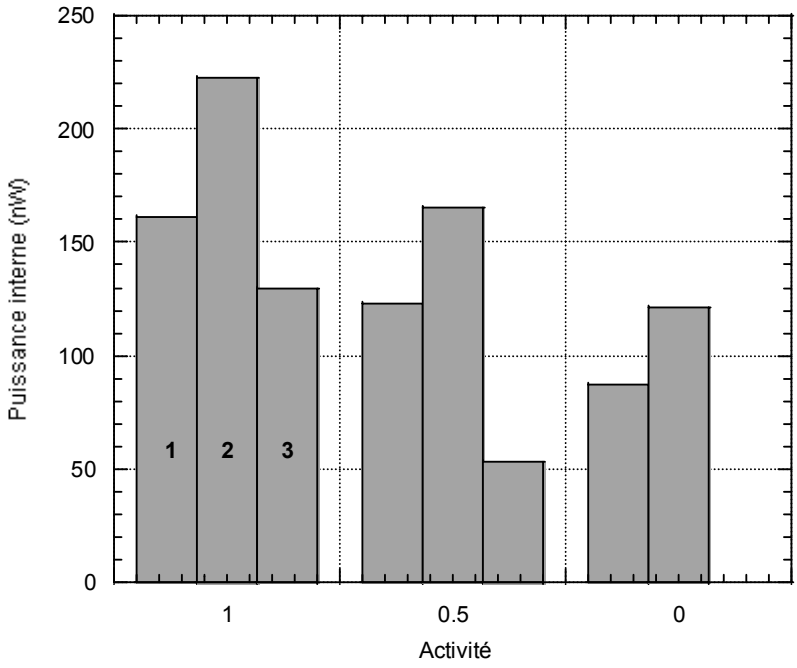
Tableau 5 Paramètres temporels, de puissance et PDP des bascules.

Bascule	mC²MOS	Power PC	TSPC
Clk-Q ll (ps)	679	584	206
Clk-Q hh (ps)	620	741	501
D-Q ll minimum (ps)	961	783	376
D-Q hh minimum (ps)	975	819	1009
Temps de setup optimal (ps)	270	70	510
Puissance horloge (nW)	36,3	18,4	67,2
Puissance donnée (nW)	6,4	12,4	7,5
Puissance interne (nW)	165	122,8	53,1
Puissance totale (nW)	207,7	153,6	127,8
PDP (fJ)	0,2	0,125	0,13

La Figure 58 montre la puissance interne dissipée par les bascules en fonction du taux d'activité de la donnée. On constate que dans le cas des bascules mC²MOS et PowerPC, la bufferisation interne du signal d'horloge occasionne une consommation de puissance interne élevée en l'absence d'activité de la donnée : elle représente la moitié de leur consommation pour un taux d'activité de 1. Néanmoins, c'est seulement de la puissance qui est retirée de l'arbre d'horloge.

La dissipation de puissance de la bascule TSPC a été fixée arbitrairement à 0 pour un taux d'activité nul car il n'y a pas d'activité interne contrairement aux cas précédents : le résultat obtenu par simulation est légèrement négatif (-0.8nW) et est dû aux couplages capacitifs par les grilles des transistors attaqués par le signal d'horloge qui injectent plus de courant dans

l'alimentation qui n'en est consommé. Ce résultat négatif, qui suggérerait que la bascule génère de l'énergie, n'est pas considéré comme significatif et est fixé à 0 – tout au plus, la consommation de la bascule est due aux courants de fuite.



1 : Power PC 2 : mC²MOS 3 : TSPC

Figure 58 Puissance interne des bascules en fonction du taux d'activité.

Nous choisissons la bascule PowerPC car c'est celle qui a le meilleur produit puissance-délai. La bascule TSPC est très proche à ce niveau-là, cependant elle présente des délais très déséquilibrés selon le sens de transition de la sortie et a l'inconvénient de mémoriser la donnée de façon dynamique et donc d'être beaucoup plus sensible aux courants de fuite.

5.3 La bibliothèque de cellules standards

Pour faciliter la conception de circuits complexes, une bibliothèque de cellules standards est utilisée. Celle-ci est basée sur la bibliothèque HD9GPLL de STMicroelectronics qui contient plus de trois cents cellules. C'est une bibliothèque de style logique CMOS et en technologie BULK 0.13µm avec des transistors LL. La première partie du travail consista à la caractériser en technologie SOI et à basse tension et en considérant des transistors HS.

5.3.1 Caractérisation de la bibliothèque

Une bibliothèque de cellules standards contient le délai, la pente des signaux de sortie et la puissance dissipée, pour chaque porte, en fonction des conditions de caractérisation. Celles-ci comprennent :

- la tension d'alimentation,
- la température,
- le type de procédé des transistors : typique, pire cas, meilleur cas ;
- la gamme des capacités de charge, variant en fonction des capacités de sortance des portes,
- et la gamme de pentes d'entrée.

Généralement, une bibliothèque est caractérisée en fonction de deux conditions, pire cas et meilleur cas. La condition de pire cas sert, lors de la synthèse d'un circuit, à maximiser le temps de *setup*, pour s'assurer que la donnée arrivera suffisamment tôt à l'entrée d'un bascule quelques soient les conditions. La condition de meilleur cas sert à repérer les problèmes de temps de *hold* qui en résultent, la donnée suivante arrivant trop tôt pour que la précédente soit correctement mémorisée. Nous avons également caractérisé la bibliothèque pour le cas typique, de manière à pouvoir correctement comparer les technologies BULK et SOI.

Par exemple, pour caractériser la technologie SOI HS en typique, nous avons considéré les valeurs suivantes :

- condition centrale : Procédé = typique
  Tension = 0.5V
  Température = 25°C
- liste des pentes : 0,0056 ; 0,1016 ; 0,2616 ; 0,5336 ; 1,0696 ; 1,5 ; 2 ; 2,5 ; 3 ; 3,5 ; 4 ; 4,5
- liste des charges (pour X0) : 0,004 ; 0,01 ; 0,02 ; 0,03 ; 0,04
  (pour X8) : 0,004 ; 0,036 ; 0,104 ; 0,278 ; 0,596 ; 1,28

Les choix de sortance sont conformes à la bibliothèque standard de référence. Concernant la liste de pente, les pentes lentes – de 3ns à 4,5ns – ont été rajoutées pour tenir compte des performances moindres des portes.

Caractérisation en délai

Les seuils utilisés pour mesurer les pentes sont de 10% et 90% de la tension d'alimentation V_{DD}. Pour mesurer un temps de propagation, la largeur d'une impulsion ou encore les paramètres temporelles d'une bascule, les seuils utilisés sont :

- 40% de V_{DD} pour un signal montant,
- 60 % de V_{DD} pour un signal descendant.

Les pentes et délais obtenus sont placés dans un tableau à deux dimensions, en fonction de la gamme de pentes du signal d'entrée et de la gamme de capacités de charge en sortie.

Paramètres temporels des bascules

Aux paramètres définis par Unger et Tan, qui ont été explicités dans le paragraphe 5.2.1, s'ajoute la mesure de la largeur d'impulsion minimale du signal d'horloge requise pour mémoriser une valeur correcte. Les temps de *setup* et de *hold* sont calculés pour une charge fixe, généralement la valeur moyenne, et toute la gamme de pente en entrée. Le délai Clk-Q est mesuré pour les deux gammes de charge et de pente. Quant à la largeur d'impulsion minimale, elle est mesurée pour un seul couple pente-charge, généralement la pente la plus raide et une charge moyenne.

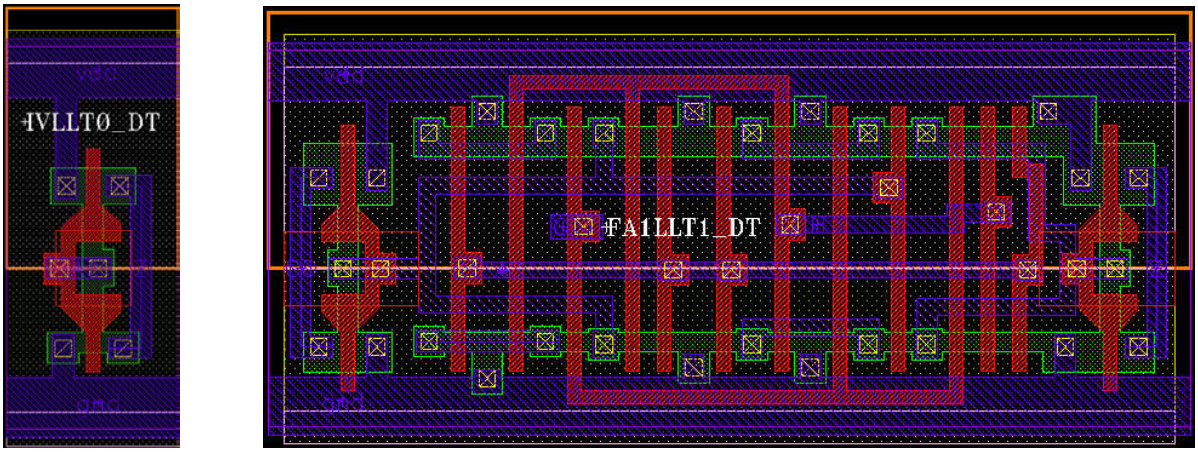
Caractérisation en puissance

Deux types de puissance sont donnés dans la bibliothèque : la puissance statique due aux courants de fuite et la puissance dynamique. La puissance statique est mesurée pour toutes les combinaisons possibles en entrée, en additionnant le courant provenant de l'alimentation et de toutes les entrées. La puissance dynamique est la puissance dissipée à l'intérieur de la bascule lorsqu'une entrée varie : elle provient du courant de charge des nœuds internes et des courants de fuite, puisqu'il n'y a pas de courant de court-circuit à la tension d'alimentation à laquelle nous travaillons. Nous excluons de cette mesure le courant de charge de la capacité de sortie. La puissance interne est mesurée pour les deux gammes de pente et de charge.

5.3.2 Contenu de la bibliothèque

La bibliothèque utilisée à 0,5V est un sous-ensemble de la bibliothèque HD9GPLL originelle et ne contient plus que 150 cellules environ. Les autres cellules, notamment toutes les bascules, ne passent pas la caractérisation à 0,5V : leur fonctionnalité n'est plus assurée à très basse tension. Différents types de bascules et de latchs, basés sur la bascule du PowerPC ont donc été rajoutés – voir l'Annexe 6. De plus, de manière à pouvoir tirer profit de l'usage

des transistors DTMOS, des portes contenant ce type de transistor ont été développées : ce sont des inverseurs, des buffers, des additionneurs 1bit et les différentes bascules et latchs. Les transistors DTMOS ont été rajoutés dans les inverseurs de sortie, puisque comme nous l'avons vu dans le paragraphe 3.5.2, les charges à l'intérieur d'une cellule ne sont pas suffisamment importantes pour en tirer avantage. Les cellules ont été dimensionnées avec comme objectif, la minimisation du produit puissance délai. Dans ce sens, un rapport de mobilité de 1,5 a été retenu, car cette valeur permet de diminuer les capacités de grille des portes. Pour toutes les portes introduites, au moins trois tailles différentes ont été considérées, l'inverseur et le buffer en comptant six. La Figure 59 montre les layouts d'un inverseur DTMOS et d'un additionneur 1 bit avec inverseur de sortie DTMOS.



(a) **(b)**

Figure 59 Layouts de portes incluant des transistors DTMOS : (a) un inverseur de taille 0, (b) un additionneur 1 bit de taille 1.

5.4 Conclusion

Dans une première partie, la comparaison de différents styles logiques statiques a conclu à la supériorité de la logique CMOS sur presque tous les points : meilleure marge au bruit, délai et consommation inférieurs et facilité d'emploi et de caractérisation pour synthétiser des circuits complexes. Dans la suite, nous avons recherché une bascule présentant un bon facteur de qualité et nous avons choisi la bascule du PowerPC, à base de portes de passage. Nous avons donc à partir là, développé une bibliothèque de cellules standards, basée sur la

bibliothèque HD9GPLL de STMicroelectronics et adaptée à une utilisation basse tension. Des transistors DTMOS ont été introduits afin de pouvoir estimer leur impact par la suite.

6 Etude de la propagation des signaux sur les interconnexions longues

Les technologies sous-microniques avancées, 0,18µm ou moins, ont permis d'augmenter considérablement la densité des transistors en diminuant la taille du motif minimal. Ces technologies ont notamment permis d'ouvrir un nouveau domaine dans l'intégration des composants, les systèmes sur puce ou SOC (System-On-Chip). La longueur et le nombre des interconnexions globales (bus, horloge, …) ont donc augmenté : les valeurs des capacités des interconnexions sont aujourd'hui plusieurs ordres de grandeur plus élevées que les capacités intrinsèques des portes CMOS.

Cette tendance s'observe également au niveau des interconnexions plus courtes, à l'intérieur d'un même bloc: les fils ont une importance croissante dans la conception d'un circuit sous-micronique car le délai se déplace de plus en plus des portes vers les fils. La Figure 60 montre que le délai des interconnexions dépasse celui des portes dès les nœuds technologiques 0,18µm en aluminium et 0,13µm en cuivre. Dans la prochaine génération, les fils représenteront 75% du délai total [Lev].

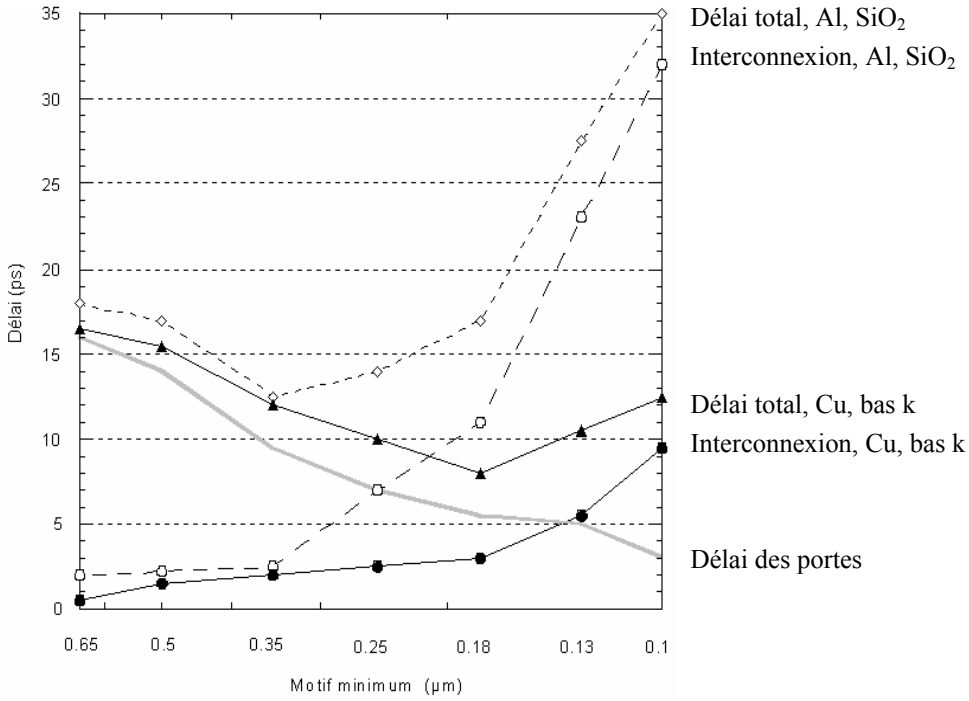


Figure 60 Comparaison des délais intrinsèques des portes et des fils en aluminium Al et en cuivre Cu pour différentes générations technologiques [Lev].

Ce problème est atténué lorsque l'on travaille en très basse tension, car les performances des portes sont plus basses qu'à tension nominale, tandis que le temps de propagation dans les fils ne dépend pas de la tension d'alimentation. Néanmoins, pour les interconnexions longues, il va falloir utiliser de gros transistors pour charger les lignes rapidement car le courant fourni est faible. Le nombre d'étages d'inverseurs attaquant les interconnexions va être plus élevé qu'à tension d'alimentation nominale, pénalisant la consommation.

Il faut donc développer des techniques permettant de garder des performances correctes tout en diminuant la dissipation d'énergie. Dans la suite, nous allons brièvement étudier les propriétés intrinsèques d'une ligne d'interconnexion et le temps de propagation d'un signal en tension et en courant. Puis, nous allons approfondir les deux modes possibles de transmission d'une donnée sur un fil : le mode tension et le mode courant. Pour le mode tension, nous allons expliciter quel type de bufferisation est optimal et dériver une méthode de dimensionnement. Pour le mode courant, nous allons introduire les bases de la lecture en courant et détailler les circuits existants. De là, nous allons développer un circuit de lecture fonctionnant en très basse tension. Enfin, nous allons établir des règles de conception en fonction de la longueur de la ligne.

6.1 Etude d'une ligne d'interconnexion

6.1.1 Modélisation

Les lignes d'interconnexion servent à relier deux portes entre elles, en propageant un signal en mode courant ou en mode tension. En technologie CMOS, on peut représenter ces signaux comme des rampes, puisqu'ils sont carrés. Pour déterminer quel type d'analyse est suffisant, il faut connaître la bande passante du signal – tout signal carré peut se décomposer en une somme de sinusoïdes et a donc une bande passante limitée. [Dhae92] donne une expression largement utilisée pour calculer la bande passante maximale d'une onde:

$$f_{max} = 0.35/T_r$$

Équation 26

où T_r est le temps de montée du signal. La bande passante est donc de quelques gigahertz puisque dans nos applications, le temps de montée est de l'ordre de la nanoseconde. L'analyse de l'interconnexion peut alors se faire dans le domaine temporel : il n'est pas nécessaire de recourir à l'équation du télégraphe et encore moins aux équations de Maxwell car il n'y a pas d'effet de peau à ces fréquences basses.

Les lignes sont souvent modélisées par des éléments RC. Pour savoir s'il faut introduire le paramètre de l'inductance L, [Isma99] donne les limites inférieure et supérieure de la longueur d de la ligne nécessaires pour considérer l'inductance, en fonction du temps de montée T_r et des paramètres R, L, C par unité de longueur :

$$\frac{T_r}{2\sqrt{LC}} < d < \frac{2}{R}\sqrt{\frac{L}{C}}$$

Équation 27

Si la longueur de la ligne est en dehors de ces bornes, il indique que l'on peut négliger l'inductance. Cependant, l'Équation 27 n'inclut ni l'impédance de sortie de l'émetteur ni l'impédance d'entrée du récepteur : ces bornes sont fausses dans le cas d'une ligne chargée par une source résistive et terminée par une charge capacitive, comme montré dans [Dhao02], ce qui représente une transmission en mode tension. Puisque l'inductance des lignes a beaucoup augmenté depuis la métallisation cuivre, nous décidons d'introduire ce paramètre par défaut dans la modélisation des interconnexions.

Une ligne peut être représentée par un seul élément RLC (modèle L) ou par une série d'éléments distribués (modèles π et T). Les circuits équivalents π et T sont présentés Figure 61.

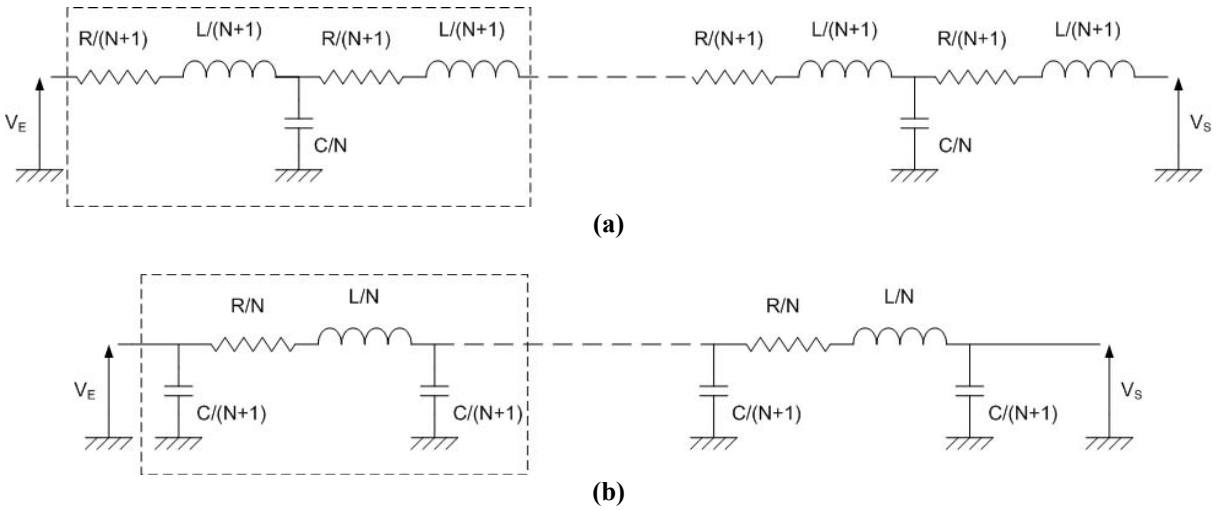


Figure 61 Circuits équivalents d'une ligne RLC : (a) représente le modèle π et (b) le modèle T.

Dans [Saku83], Sakurai montre que T_3 et π_3 modélisent mieux une ligne RC que le modèle L. Les modèles T et π sont différents pour des valeurs N assez petites mais identiques pour de grandes valeurs. Dans la suite, nous utilisons le modèle π, puisqu'il faut bien en choisir un – et qu'il sera équivalent au modèle T pour la valeur de N que nous allons considérer.

La précision de la modélisation d'une ligne RLC augmente avec la valeur de N, mais une trop grande valeur pénalise la durée des simulations électriques. Pour déterminer la valeur de N adéquate, la réponse en fréquence du modèle π est présentée Figure 62 pour différentes valeurs de N : 1, 3, 5 et 100. Comme nous pouvons le constater, il y a un phénomène de saturation de la précision pour des valeurs élevées de N. La valeur N=5 est donc suffisante pour modéliser correctement une ligne de transmission, puisqu'elle donne des résultats très proches de N égal à 100.

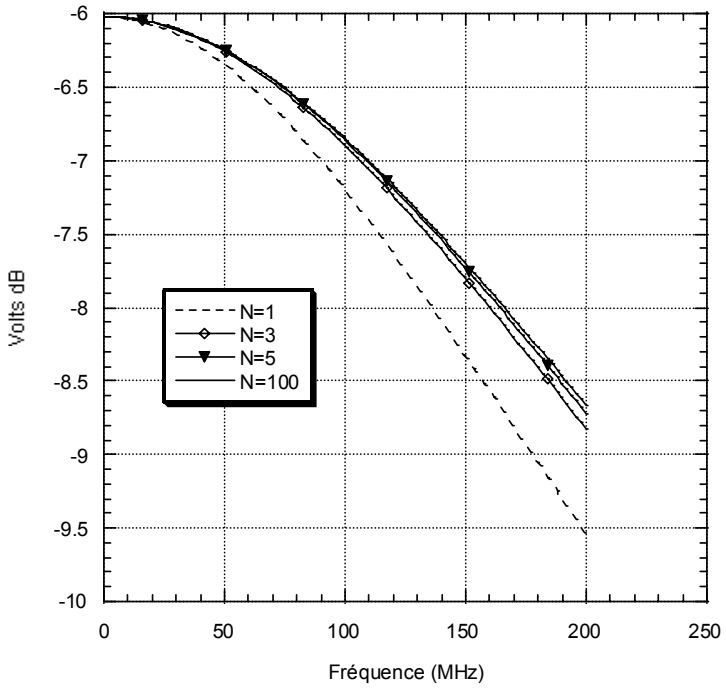


Figure 62 Réponse en fréquence du modèle π pour N=1, N=3, N=5 et N=100. La longueur de la ligne est de 10mm avec les paramètres suivants : C=180fF/mm, R=100Ω/mm, L=1nH/mm.

6.1.2 Etude du délai intrinsèque

Comme expliqué précédemment, le calcul est effectué dans le domaine temporel. Le signal d'entrée est considéré comme étant une rampe, ce qui s'approche correctement des signaux CMOS. Cependant, comme la rampe ne peut ici être tronquée comme dans la réalité, l'analyse est statique (en mode DC) et ne peut inclure d'effets transitoires. L'expression générale du délai intrinsèque est la suivante :

$$\delta t = \frac{L_T + \frac{R_T C_T}{2}\left[R_B + \frac{R_T}{3} + R_L\left(1 + 2\frac{R_B}{R_T}\right)\right]}{R_T + R_B + R_L}$$

Équation 28

où R_T est la résistance totale de la ligne, C_T sa capacité totale, L_T son inductance totale, R_B la résistance de la source en entrée et R_L la résistance en sortie de la ligne. Pour une transmission en mode tension, la résistance R_L est très grande.

L'expression du délai se réduit à :

$$\delta t_v = \frac{R_T C_T}{2}\left(1 + 2\frac{R_B}{R_T}\right)$$

Équation 29

Pour une transmission en mode courant, la résistance en sortie de la ligne est très petite. Le temps de propagation vaut alors :

$$\delta t_i = \frac{L_T + \frac{R_T C_T}{2}\left(R_B + \frac{R_T}{3}\right)}{R_T + R_B}$$

Équation 30

Si l'on considère une ligne de transmission de 10mm de long, avec les paramètres suivants : R_T=1kΩ, C_T=1,8pF, LT=10nH et une source avec une résistance d'entrée R_B=500Ω, on obtient :

$$\delta t_v = 1.8ns\,,$$

et

$$\delta t_i = 0.6ns\,.$$

La transmission en mode courant est intrinsèquement plus rapide qu'en mode tension.

6.2 Analyse du mode tension

6.2.1 La transmission en mode tension

L'étude des inverseurs attaquant des interconnexions est ancienne. A l'origine, les chaînes d'inverseur étaient utilisées mais la technique qui a émergé au fil des ans avec beaucoup de succès est celle de l'insertion de répéteurs. De très nombreuses méthodes sont données dans la littérature quant à leur dimensionnement et leur placement optimaux [Gine90], [Lill96], [Alpe98]. L'intérêt des répéteurs est de couper la ligne en plusieurs morceaux plus petits. En effet, en reprenant l'Équation 29 et en négligeant le délai introduit par la résistance de l'émetteur, le délai vaut :

$$\delta t_v = \frac{R_T C_T}{2} = \frac{R_l C_l}{2} l^2$$

Équation 31

avec R_l et C_l respectivement la résistance et la capacité par unité de longueur et *l* la longueur de la ligne. Le délai dépend donc quadratiquement de la longueur. En insérant des répéteurs, le délai devient linéairement dépendant de *l* – plus le retard introduit par les répéteurs.

Cependant, aussi intéressante que soit cette technique, on a indiqué précédemment que le délai d'un inverseur, en très basse tension, n'est pas petit devant le délai de la ligne. La Figure 63 montre la comparaison des temps de propagation d'un inverseur – avec une sortance de quatre (FO4) en fonction de la tension d'alimentation V_{DD} – et d'une ligne de transmission – en fonction de sa longueur *l*. L'abscisse inférieure indique la tension V_{DD}, l'abscisse supérieur la longueur *l*. Le temps de commutation de l'inverseur augmente exponentiellement avec la diminution de V_{DD} : à 0,5V, il est égal au délai intrinsèque d'une ligne de 8,5mm de long. Il n'est donc pas intéressant en très basse tension de couper la ligne en plusieurs morceaux : le délai serait perdu dans les inverseurs. On ne peut pas ici utiliser la technique très répandue de l'insertion de répéteurs, il faut utiliser une chaîne d'inverseurs à la place.

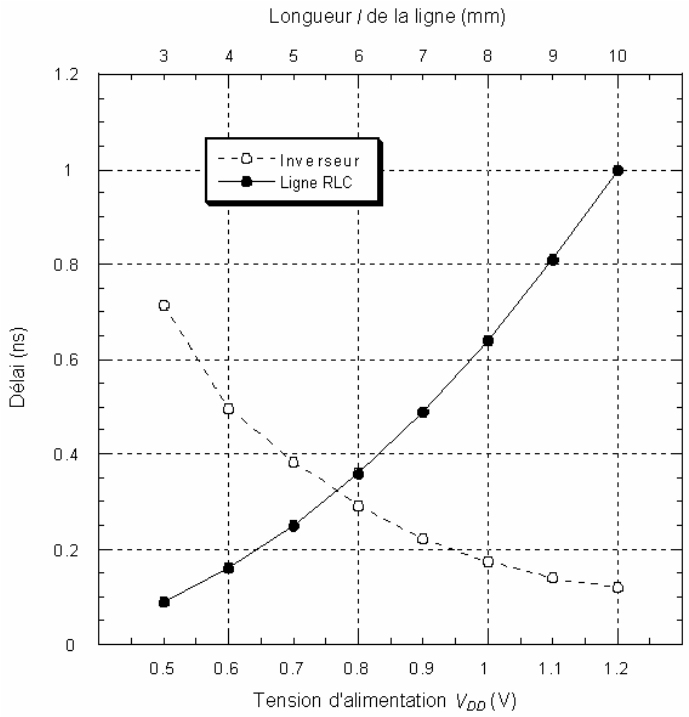


Figure 63 Comparaison du délai d'un inverseur avec un fanout de 4 en fonction de V_{DD} et du délai d'une ligne RLC en fonction de l.

6.2.2 La chaîne d'inverseurs

La première méthode d'optimisation d'une chaîne d'inverseurs a été présentée par Lin et Linholm en 1975 [Lin75]. Elle consistait à considérer que chaque inverseur est β fois plus grand que l'inverseur précédant et à trouver la valeur β optimale. Lin et Linholm ont trouvé que pour une chaîne de N inverseurs, le délai optimal est obtenu lorsque le rapport du courant de sortie divisé par la capacité de sortie est constant pour chaque étage. Jaeger [Jaeg75] a montré par le calcul que ce rapport optimal est la valeur *e*. Le nombre N d'inverseurs est alors :

$$N = \frac{\ln Y}{\ln \beta},$$

Équation 32

où $Y=C_L/C_y$, avec C_L la capacité de charge et C_Y la capacité de grille d'un inverseur minimum. Cependant, les optimisations précédentes ne prenaient pas en compte la valeur de la capacité intrinsèque de sortie d'un inverseur – due essentiellement aux diffusions de drain et aux recouvrements grille-drain. La méthode d'optimisation a donc été améliorée par la suite dans [Kanu83], [Li90] et [Hede94] en adoptant un modèle de capacité de sortie incluant la capacité de sortie intrinsèque et la capacité de grille de l'étage suivant ; ce modèle est appelé "capacité divisée" – *split-capacitor* – et est présenté Figure 64.

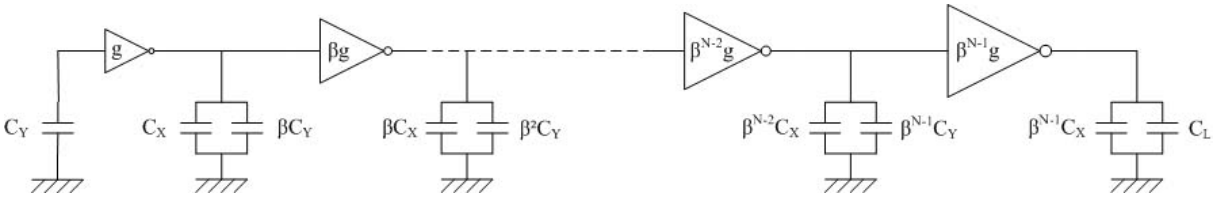


Figure 64 Chaîne d'inverseurs avec un rapport β fixe et le modèle de capacité divisée.

La capacité de sortie des inverseurs est séparée en deux parties : la capacité de sortie intrinsèque C_x et la capacité de grille de l'étage suivant C_y. Chaque étage est plus gros que le précédant d'un facteur β. Les expressions de la conductance, des capacités et de la constante de temps du $k^{ème}$ étage sont alors :

$$g_k = \beta^k g\,,$$

$$C_{x,k} = \beta^k C_x\,,$$

$$C_{y,k} = \beta^{k+1} C_y\,,$$

$$\tau_k = \frac{\beta^k C_x + \beta^{k+1} C_y}{\beta^k g} = \frac{C_x + \beta \cdot C_y}{g}$$

Équation 33

avec g, C_x et C_y respectivement la conductance, la capacité d'entrée et la capacité de sortie intrinsèque d'un inverseur minimal. Le délai total de la chaîne est donné par :

$$\tau_{tot} = \sum_{k=0}^{N-1} \tau_k = N \frac{C_x + \beta C_y}{g}$$

Équation 34

En considérant l'Équation 32, le délai total vaut :

$$\tau_{tot} = \frac{\ln Y}{g} \cdot \frac{1}{\ln \beta} \cdot \left(C_x + \beta \cdot C_y \right)$$

Équation 35

En dérivant l'équation précédente par rapport à β, on obtient le rapport β optimal en résolvant l'expression suivante :

$$\ln \beta \cdot \left(1 + \frac{C_x}{C_y} \cdot \frac{\ln \beta}{\beta} \right) - 1 = 0$$

Équation 36

Bien qu'on ne puisse pas trouver une solution simple, l'Équation 35 est une fonction convexe et n'admet qu'une seule solution. Cette fonction a été tracée Figure 65 pour un rapport C_x/C_y égal à 1 : le délai minimum est pour β=3,6.

En prenant en compte la capacité de sortie intrinsèque, le facteur β optimal n'est donc plus fixe – valeur *e* précédente – mais dépend du rapport C_x/C_y.

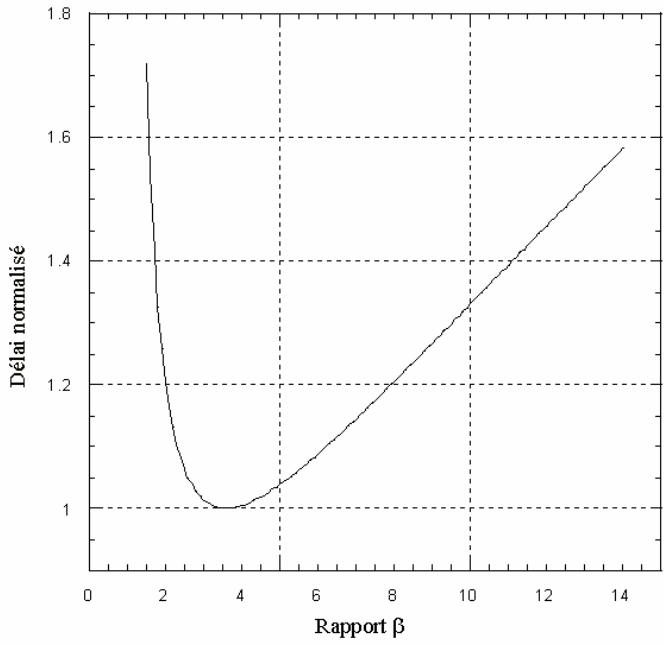


Figure 65 Délai normalisé en fonction du rapport β pour Cx/Cy=1.

[Shoj88] a montré qu'un facteur β fixe entre les étages est la configuration qui présente le délai le plus petit, car le délai de chaque étage est constant. Cependant, cette configuration n'est plus optimale lorsque des contraintes autres que le délai sont considérées, par exemple le produit puissance-délai ou encore la surface occupée. [Vemu91] est le premier à avoir introduit un facteur β variable en fonction de la position de l'inverseur dans la chaîne : son évolution suit une loi géométrique. Le rapport du $k^{ème}$ inverseur vaut :

$$\beta_k = \beta^{\frac{k(k+1)}{2}}$$

Équation 37

Le gain en consommation et en surface est intuitif : en permettant au facteur β d'augmenter avec le numéro de l'étage, le nombre d'inverseurs nécessaires pour attaquer une charge C_L est plus petit, d'où des valeurs de capacités commutées et une surface occupée moindres. Puisqu'en basse puissance, nous voulons optimiser le produit puissance-délai, une chaîne d'inverseurs optimale, dans notre cas, possédera un rapport β variable d'un étage à l'autre si les contraintes de délai ne sont pas fortes. Dans la suite, nous allons détailler la méthode d'optimisation de la chaîne d'inverseurs. Pour simplifier l'analyse, nous allons commencer en étudiant le cas de la seule contrainte en délai.

6.2.3 Optimisation de la chaîne d'inverseurs en délai

Les chaînes d'inverseurs considérées ici sont réalisées uniquement avec des transistors DTMOS puisque ceux-ci présentent un bien meilleur courant I_{ON}, et sont par conséquent bien adaptés pour attaquer une grosse charge. Optimiser une chaîne d'inverseur en technologie sous-micronique n'est pas trivial. En effet, si l'on garde le rapport W_P/W_N constant d'un étage à l'autre, le temps de propagation dans la chaîne sera différent en fonction que l'on passe un '0' logique ou un '1' logique : il faut prendre en compte les effets de canal étroit qui impliquent d'une part que le rapport des mobilités des transistors NMOS et PMOS n'est pas constant – voir Figure 66 – , mais également qu'on ne peut plus écrire simplement la conductance d'un transistor en fonction de sa largeur. On ne peut donc pas dériver comme plus haut un facteur β_i optimal. La première partie de mon travail d'optimisation consiste donc à prendre en compte les effets de canal étroit mais aussi les non-linéarités, en fonction de la largeur, des capacités d'entrée et de sortie dues aux contacts de body. En effet, la capacité introduite par un contact n'est pas négligeable : elle est de 0,35fF alors que la capacité de grille vaut en moyenne 1,6fF/µm.

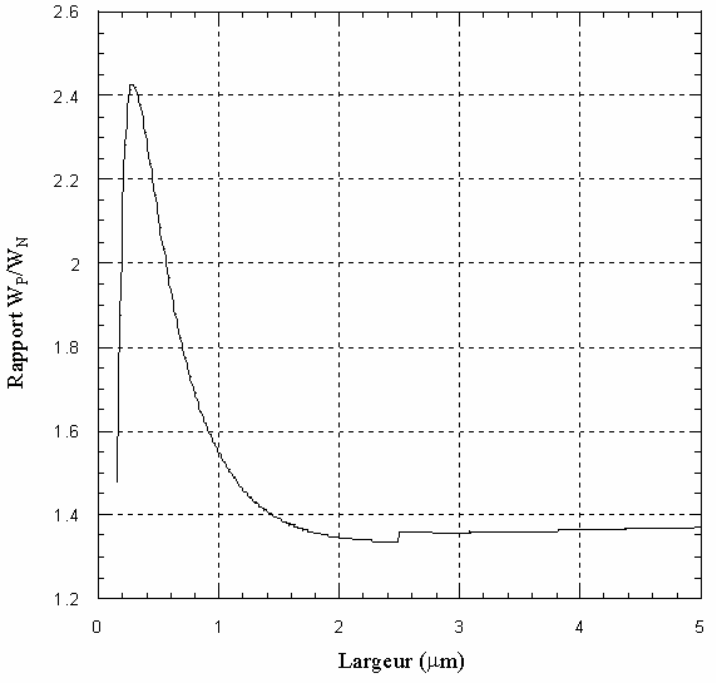


Figure 66 Rapport des mobilités en fonction de W (µm).

La variation du rapport des mobilités et l'introduction des capacités dues aux contacts de body ont pour effet de modifier la capacité de charge d'un étage à l'autre ainsi que le courant I_{ON}. Le rapport capacité de charge sur courant fourni n'est pas le même pour tous les étages,

donc le délai n'est pas constant dans chaque étage et la configuration obtenue n'est pas optimale. Pour optimiser la chaîne, nous introduisons le paramètre K, qui doit être maintenu constant :

$$K_i = \frac{C_X^i + C_Y^{i+1}}{I_i}$$

Équation 38

Ce paramètre représente le temps de montée ou de descente d'un inverseur en seconde/Volt. En effet, le rapport K a pour expression :

$$K = \frac{\tau}{V_{DD}}$$

Équation 39

La difficulté avec le modèle de la capacité divisée vient du fait que les capacités d'entrée et de sortie des transistors varient avec les tensions d'entrée et de sortie. De plus, il existe un effet Miller dû à la capacité mutuelle entre l'entrée et la sortie. Il est donc difficile de déterminer le rapport C_x/C_y. Pour déterminer de manière précise ce rapport, on effectue des simulations Eldo de la manière suivante [Hede94]. Deux chaînes d'inverseurs, idéalement infinies, sont réalisées avec deux facteurs β différents, l'une avec le facteur β_1, l'autre avec le facteur β_2. On considère alors le temps de propagation d'un inverseur au milieu de la chaîne – pour avoir des conditions réalistes au niveau de la pente du signal d'entrée et de la charge en sortie. On obtient les valeurs t_{p1} et t_{p2} pour les deux inverseurs et en résolvant le système d'inconnues à l'aide de l'Équation 33, on a :

$$\frac{C_x}{C_y} = \frac{\beta_1 t_{p2} - \beta_2 t_{p1}}{t_{p1} - t_{p2}}$$

Équation 40

Dans notre cas, la chaîne d'inverseurs sera réalisée avec des transistors DTMOS. Avec la méthode ci-dessus, on obtient un rapport de 1.08. La capacité de sortie intrinsèque d'un inverseur est donc légèrement plus grande que la capacité de grille.

Avant de développer plus avant la méthode d'optimisation, il nous faut déterminer de manière relativement précise le délai dans un étage de la chaîne. Il semble assez évident que

le temps de propagation d'un inverseur va dépendre de sa taille *w*, de la charge en sortie et du temps de montée ou de descente de l'entrée *k* – que l'on considère identiques. Le temps de propagation de l'inverseur *i* dans la chaîne vaut donc :

$$tp_i = \frac{C_X^i + C_Y^{i+1}}{C \cdot I_i} + D \cdot k_i$$

Équation 41

avec *C* et *D* deux paramètres dépendants de la technologie. L'Équation 41 n'est pas valable dans tous les cas, par exemple si le temps de montée *k* est beaucoup plus grand que le temps de propagation *tp*. La Figure 67 montre la limite de validité de cette formule : elle reste vraie au dessus du trait noir continu gras, ce qui représente fort heureusement la totalité des cas dans une chaîne d'inverseurs, puisque la taille des inverseurs va croissante.

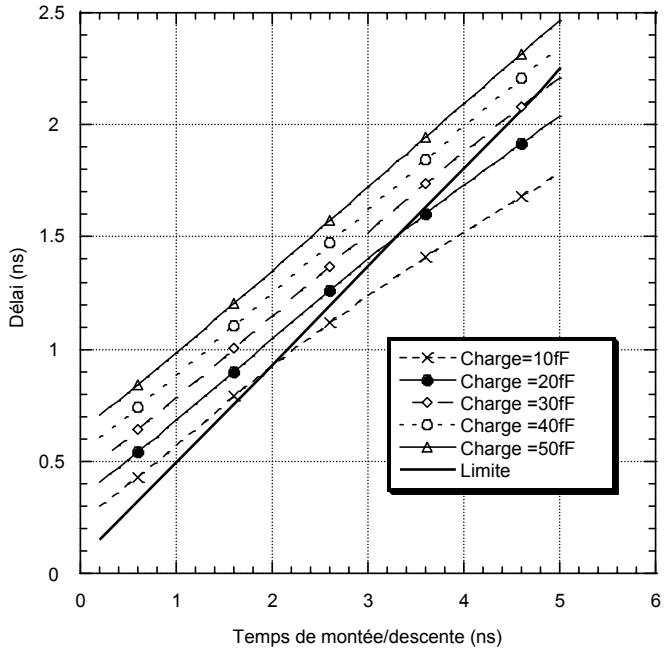


Figure 67 Délai d'un inverseur en fonction du temps de montée en entrée *k*, pour différentes valeurs de capacité de charge.

Il nous faut maintenant exprimer le courant *I*. Sa variation en fonction de la capacité de charge et de *k* est donnée Figure 68. Dans cette figure est reportée la limite définie précédemment, symbolisée par la ligne noire : la région qui nous intéresse se situe au-dessus.

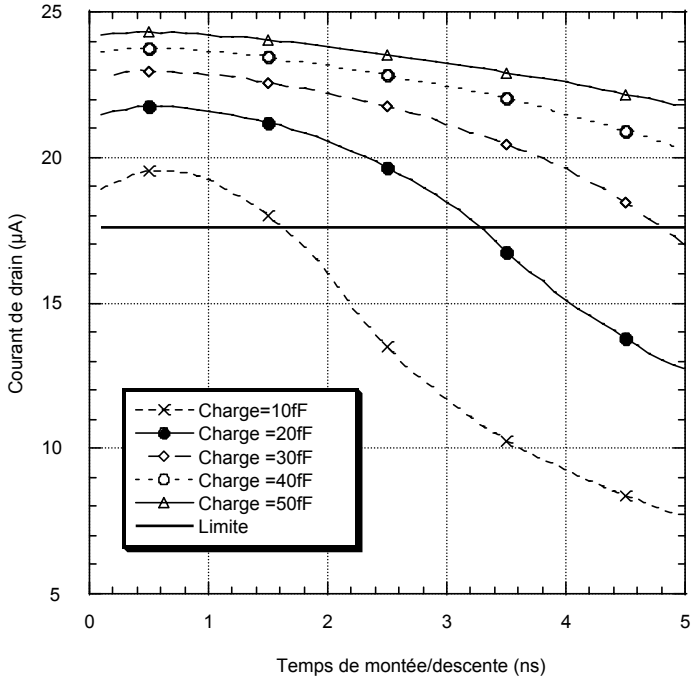


Figure 68 Valeur du courant de sortie d'un inverseur en fonction du temps de montée *k*, pour différentes capacités de charge.

On en déduit l'expression du courant suivante :

$$I_i = A \cdot w_i \cdot \left(1 - \frac{C_Y^i}{C_X^i + C_Y^{i+1}}\right) \cdot \left(1 - \frac{k_i}{B}\right)$$

Équation 42

avec *A* et *B* deux paramètres empiriques.

De la même manière, on définit l'expression du temps de montée de l'inverseur *i* par :

$$k_{i+1} = \frac{C_X^i + C_Y^{i+1}}{E \cdot I_i}$$

Équation 43

E étant là encore un paramètre empirique. On l'appelle k_{i+1} car c'est le temps de montée en entrée de l'inverseur *i+1*. La Figure 69 montre la chaîne d'inverseurs que l'on veut optimiser. L'inverseur en pointillés représente l'inverseur qui aurait une capacité de grille de valeur C_L. Le paramètre à optimiser est le paramètre β, qui vaut C_Y^2/C_Y^1.

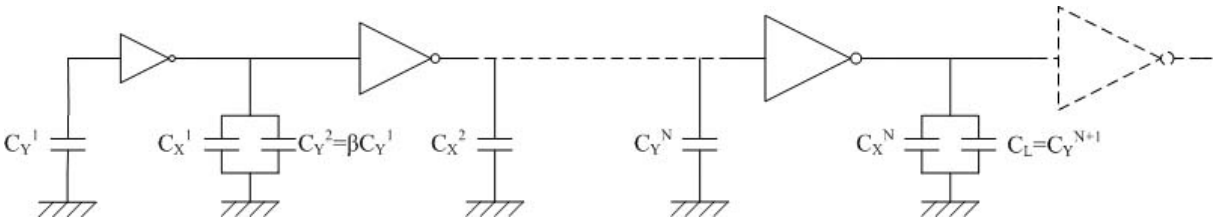


Figure 69 Chaîne d'inverseurs à optimiser en délai.

Un programme Matlab a été écrit pour l'optimisation de la chaîne ci-dessus. La structure générale de l'algorithme est précisé sur la voir Figure 70. On calcule le nombre d'étages N nécessaire à l'aide de l'équation qui suit :

$$N = \frac{\ln\left(\frac{C_L}{C_Y^1}\right)}{\ln\left(\beta_{opt}\right)}$$

Équation 44

avec C_Y^1 la capacité de grille d'un inverseur minimum et β_{opt} le rapport théorique optimal entre les étages, déduit de l'Équation 36 avec $C_X/C_Y=1{,}08$. On sélectionne une valeur β en partant de la valeur 1; on calcule alors les tailles des transistors de chaque étage pour que le rapport K reste identique et on incrémente β jusqu'à ce que C_Y^{N+1} soit très proche de C_L. On obtient alors la configuration optimale, à savoir un temps de propagation égal pour tous les étages. Le facteur β est incrémenté de la façon suivante pour améliorer le temps de calcul :

$$\beta = \beta + \frac{C_L - C_Y^{N+1}}{2 \cdot C_L}$$

Équation 45

Le pas est très grand au début et diminue à mesure que l'on approche de la bonne valeur : l'algorithme converge rapidement vers la solution. La taille des transistors de chaque inverseur est déduit en fonction de la capacité de grille voulue et du nombre de contacts nécessaires à partir d'une table, dont les valeurs sont extraites de la Figure 66. Un exemple d'application est donné dans le Tableau 6 : le programme donne la taille des transistors NMOS et PMOS de chaque étage en fonction de la capacité de charge C_L, pour avoir le délai optimal. Nous pouvons remarquer que le nombre d'étages augmente assez peu avec la capacité de la ligne à charger.

Tableau 6 Tailles des transistors de chaque étage de la chaîne d'inverseurs en fonction de la capacité de charge C_L.

	Taille des transistors (µm)	C_L=100fF	C_L=500fF	C_L=1000fF	C_L=2000fF
Etage 1	W_N	0,2	0,2	0,2	0,2
	W_P	0,46	0,46	0,46	0,46
Etage 2	W_N	1,57	1,66	1,51	1,66
	W_P	2,18	2,29	2,11	2,28
Etage 3	W_N	7,99	8,87	7,46	8,8
	W_P	10,96	11,94	10,24	11,8
Etage 4	W_N		46,15	36,79	45,5
	W_P		62,89	50,05	62
Etage 5	W_N				239,95
	W_P				326,36

Le schéma général du programme est présenté Figure 70.

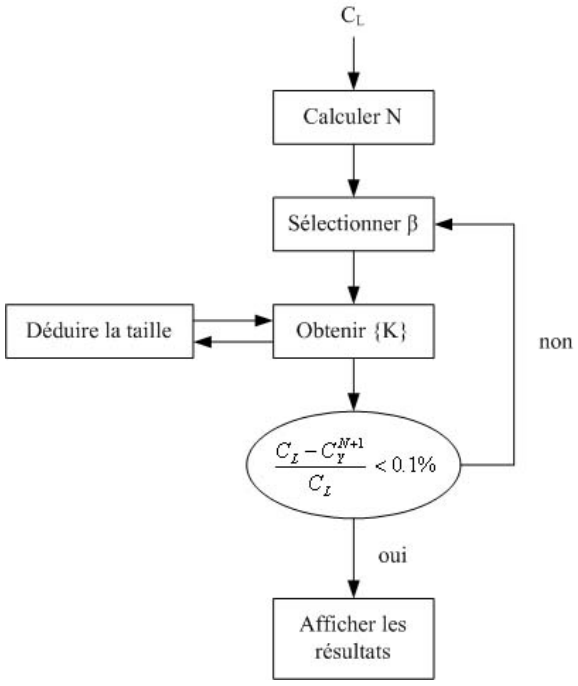


Figure 70 Schéma général de l'algorithme d'optimisation du délai.

6.2.4 Optimisation de la puissance dissipée sous contrainte de délai

Après avoir considéré le cas de l'optimisation en délai, nous allons nous intéresser ici à l'optimisation de la puissance. Plutôt que de minimiser le produit puissance délai, nous allons optimiser la puissance dissipée sous contrainte de délai, c'est-à-dire avoir la puissance la plus faible possible pour un temps de propagation donné. En effet, pour les circuits à consommation limitée, il faut pouvoir diminuer la puissance consommée tout en gardant une certaine puissance de calcul : cela représente un cas plus général que de simplement minimiser le produit puissance délai car il est ainsi possible de décrire toute la courbe puissance=f(délai).

Dans [Dhar91], Dhar et Franklin montrent que l'on peut utiliser le théorème de Khun-Tucker pour résoudre un problème d'optimisation : ce théorème donne les conditions nécessaires et suffisantes pour trouver le minimum d'un programme convexe. Le problème d'optimisation est le suivant : il faut trouver l'ensemble {w} tel que la fonction puissance *p(w)* est minimisée et que la contrainte en délai *d(w)* est respectée, c'est-à-dire :

$$d(w) = \sum_{i=1}^{N} \tau_i - \tau_{\max} \leq 0$$

Équation 46

Dhar et Franklin montrent que la solution optimale est donnée par :

$$\frac{w_{i+1}}{w_i} = \frac{w_i}{w_{i-1}} + \mu w_i$$

$$d(w) = 0$$

Équation 47

Contrairement à la solution donnée pour l'obtention du délai minimal, l'optimum est ici un facteur β qui augmente d'un facteur μw_i à mesure que l'on se déplace dans la chaîne d'inverseurs, ce qui rejoint le résultat de [Vemu91]. L'algorithme est donc modifié de la manière suivante :

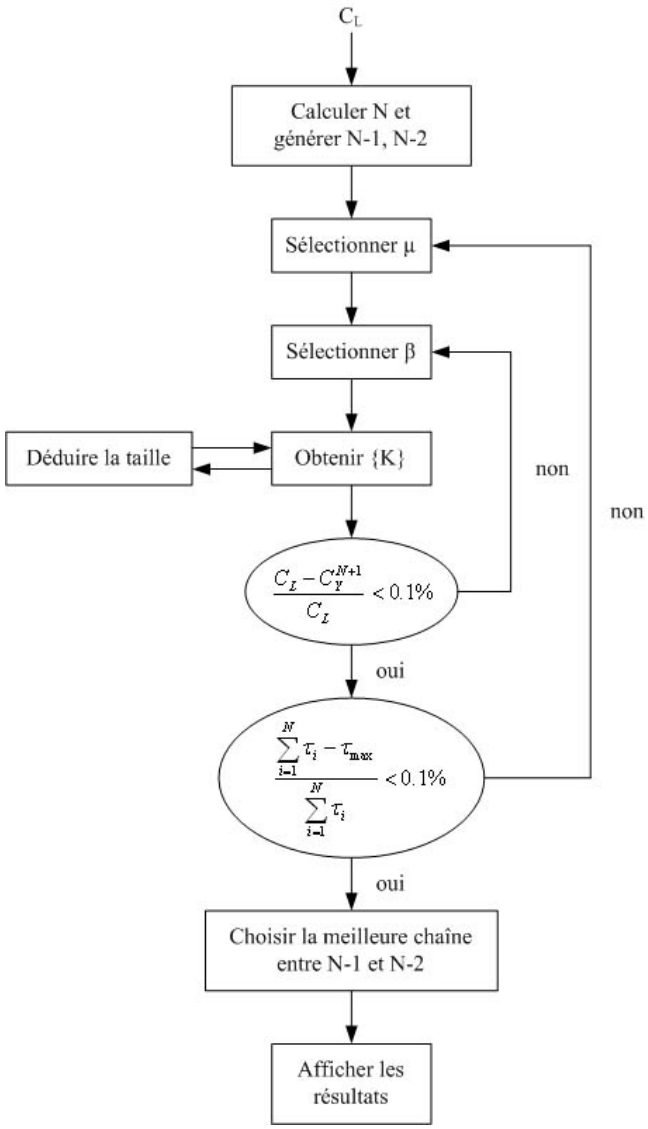


Figure 71 Schéma général de l'algorithme d'optimisation de la puissance consommée sous contrainte de délai.

Nous sélectionnons μ en partant de 0 – qui représente le cas où le délai est le plus petit – et nous augmentons sa valeur jusqu'à ce que le temps de propagation coïncide avec la contrainte. L'incrémentation de μ se fait comme suit :

$$\mu = \mu + \frac{\sum_{i=1}^{N} \tau_i - \tau_{\max}}{\sum_{i=1}^{N} \tau_i} \cdot 10^{-11}$$

Équation 48

Comme précédemment, le pas est grand au début pour converger rapidement vers la solution et diminue à son approche pour augmenter la précision. Le nombre d'étages N est calculé de la même manière que pour le premier algorithme; il représente donc le nombre d'étages nécessaire pour avoir un délai minimum. La solution optimale à notre problème d'optimisation en aura moins et comme on ne sait pas à priori sa valeur, on fait une recherche exhaustive sur N-1 et N-2.

En faisant varier la contrainte sur le délai, on peut obtenir la courbe puissance-délai, qui est présentée Figure 72. Comme on peut le noter, il est très intéressant de travailler dans la partie gauche de la courbe, là où la pente est très élevée, car le gain en consommation est très important pour une diminution minime de la performance. Nous pouvons ainsi constater que si la contrainte sur le délai est relâchée de 4% par rapport au délai minimum, la consommation se trouve réduite de 50%.

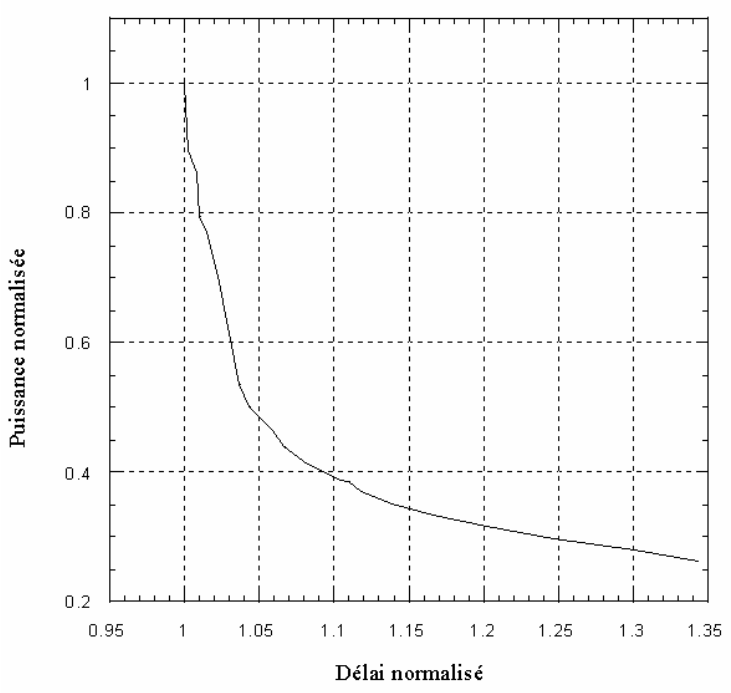


Figure 72 Courbe représentant la puissance dissipée par une chaîne d'inverseurs en fonction de la contrainte sur le délai.

6.3 Analyse du mode courant

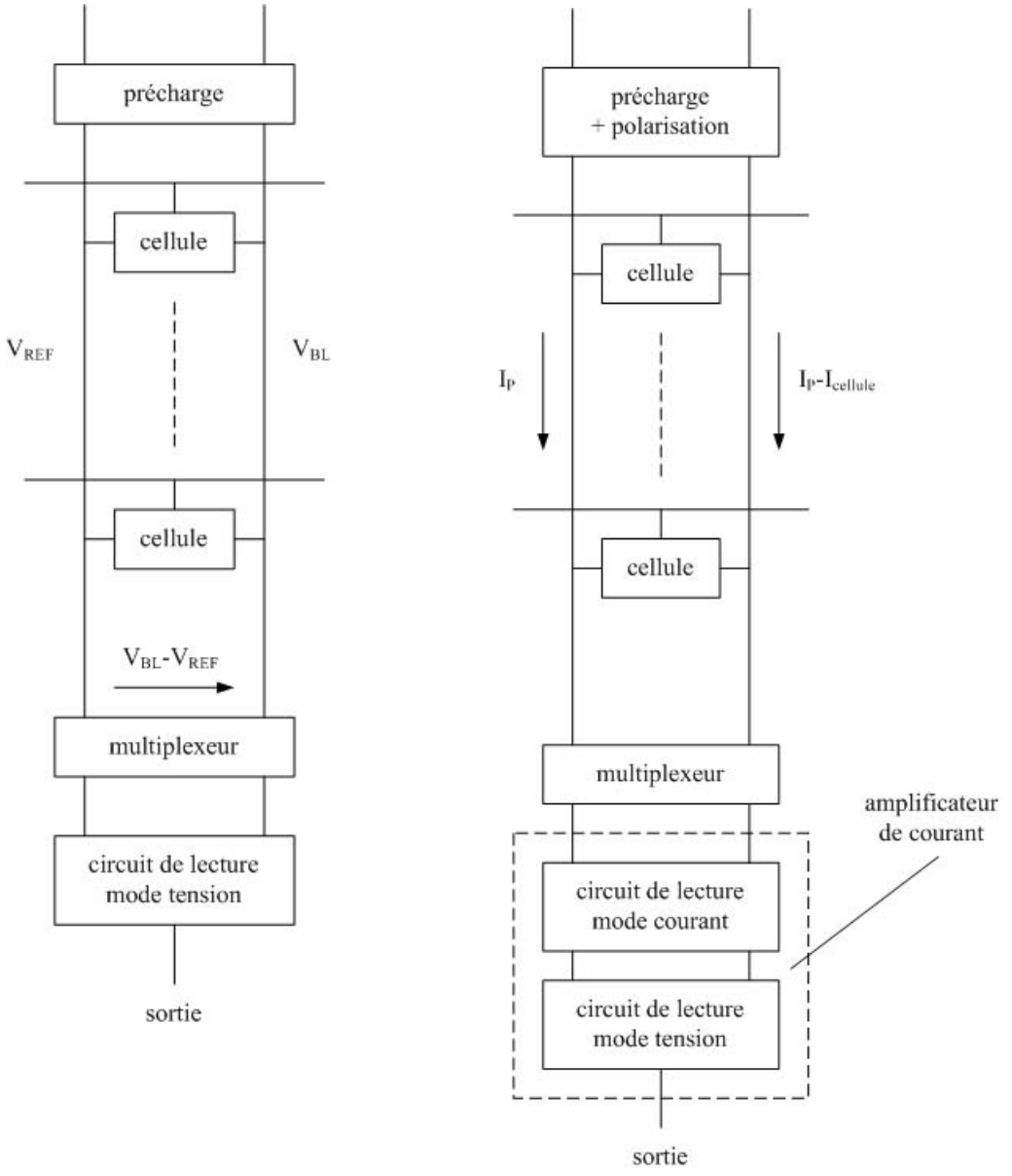
Nous allons ici expliquer les bases de la lecture en mode courant, puis détailler les problèmes de conception rencontrés en très basse tension. Pour les surmonter, nous allons concevoir étape par étape un circuit émetteur-récepteur en mode courant. Mais tout d'abord, nous allons faire une brève présentation des mémoires SRAM et de leurs circuits de lecture : le problème de transmission de données sur des interconnexions hautement capacitives est en effet présent depuis longtemps dans les mémoires. Il est donc très intéressant d'analyser les solutions proposées et de voir si nous pouvons les transposer à la transmission de données sur des interconnexions globales; il ne s'agit pas ici de faire une étude exhaustive de tous les types d'amplificateurs de lecture, cela a déjà été réalisé par Shibata en 1996 [Shib96], qui a le premier classé les circuits selon plusieurs types. Cette classification a été récemment reprise par Wicht [Wich03].

6.3.1 Mémoires SRAM

L'étude des mémoires SRAM (Static Random Access Memory) se révèle intéressante dans le cas présent. En effet, leur architecture consiste en une matrice de points mémoire, classiquement des inverseurs rétro-couplés, que l'on peut accéder en sélectionnant une ligne de mots et en lisant des lignes de bits : chaque point mémoire doit charger une ou deux lignes de bits (dans le cas différentiel, qui est le plus répandu). Or, les points mémoire doivent être les plus petits possibles, pour des raisons évidentes de surface, et ne délivrent que peu de courant : la charge des lignes de bits, hautement capacitives, se fait lentement. Des circuits de lecture ont donc été développés pour accélérer la lecture des données.

Les amplificateurs de lecture les plus couramment utilisés ont été les amplificateurs de lecture en mode tension. La Figure 73-a montre le principe de la lecture en tension. Avant la lecture, les transistors de charge amènent la ligne de bit à un potentiel prédéfini V_{REF} proche de V_{DD}. Pendant la lecture, ces transistors de charge sont désactivés; du courant est alors tiré par la cellule du côté où le "0" est mémorisé. Ce courant i_p décharge la ligne de bit correspondante, tandis que la ligne de bit complémentaire reste au potentiel de précharge V_{REF}, puisque que le côté de la cellule où est mémorisé le "1" logique ne consomme pas de courant. La différence de potentiel entre les deux lignes de bit V_{BL}-V_{REF} est détectée par un amplificateur de lecture en tension, qui est un comparateur de tension dont la sortie restitue des niveaux logiques CMOS. L'inconvénient de la lecture en tension est sa lenteur, car la

différence de potentiel doit être suffisamment importante pour le circuit de lecture : c'est la décharge de la ligne de bit par le point mémoire qui fixe les performances.



(a) **(b)**

Figure 73 Principe de fonctionnement de la lecture : (a) en mode tension et (b) en mode courant.

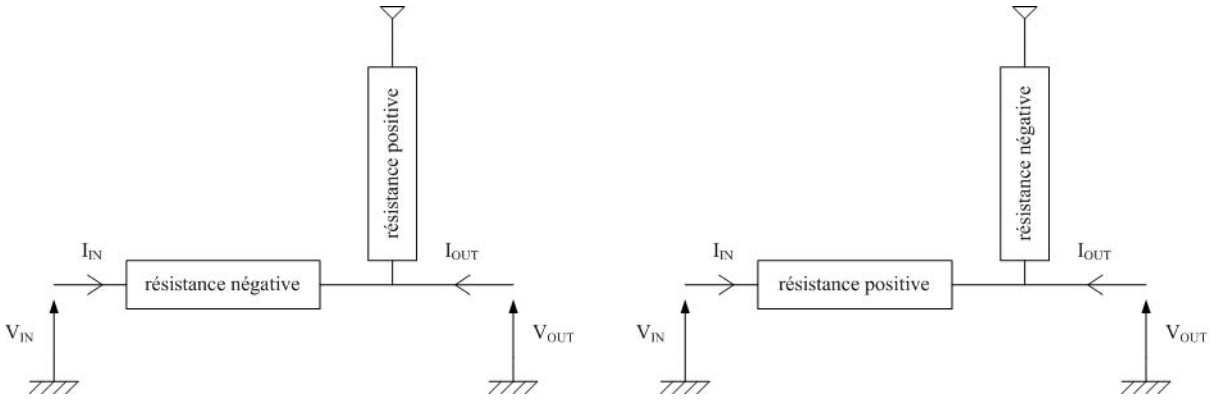
Les amplificateurs de lecture en mode courant sont de plus en plus utilisés pour surmonter ces problèmes de délai en lecture, malgré les difficultés d'industrialisation. Dans [Seev91], Seevinck et al. montrent que le mode courant est intrinsèquement plus rapide que le mode tension, résultat qui a été présenté dans le paragraphe 6.1.2. Le principe de la lecture en courant est détaillé Figure 73-b. Cette fois, ce n'est plus une différence de tension que l'on cherche à lire mais une différence de courant. Classiquement, un circuit de lecture en courant

est inséré avant le comparateur de tension pour générer la différence de potentiel auparavant présente sur les lignes de bits complémentaires. Comme pour le mode tension, les lignes sont préchargées à un potentiel V_{REF} proche de V_{DD}. Pendant la lecture, un courant de polarisation I_P est envoyé dans les lignes de bit. Ce courant I_P sert à polariser le circuit de lecture afin qu'il présente une impédance d'entrée faible : la résistance série doit être minimale pour limiter les variations de potentiel sur les lignes. Le courant I_P, égal dans les deux lignes de bit, est tiré par la cellule mémoire du côté où le "0" est mémorisé ; il reste constant dans la ligne complémentaire car la cellule ne consomme pas de courant du côté du "1" logique. Cette différence de courant I_P-I_{cell} est lue par le circuit de lecture, qui génère une différence de tension proportionnelle en sortie : l'ensemble circuit de lecture en mode courant et circuit de lecture en mode tension constitue l'amplificateur de courant.

Nous allons expliquer plus en détail le fonctionnement des amplificateurs en mode courant en partant des bases de la lecture en courant.

6.3.2 Les bases de la lecture en mode courant

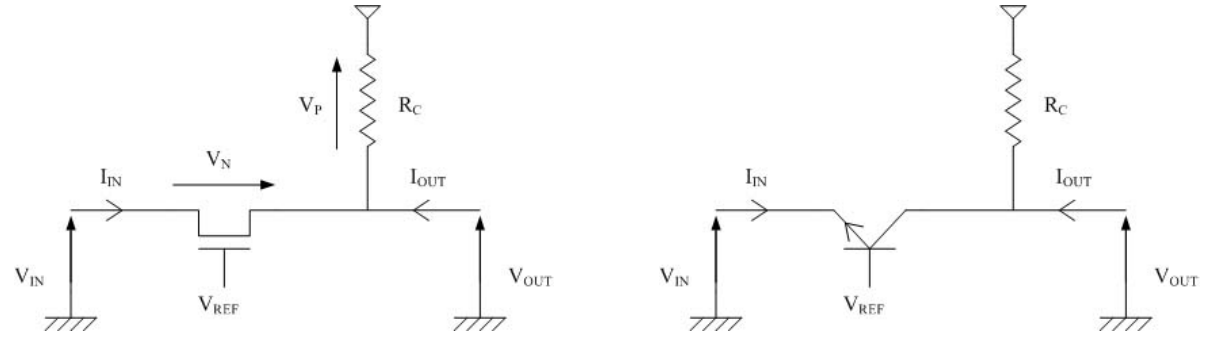
La fonction de base d'un amplificateur de courant est la conversion courant-tension : le courant présent en entrée est transformé en une tension en sortie par l'intermédiaire d'une résistance. Or, le circuit de lecture doit avoir une impédance d'entrée faible pour limiter les variations de tension sur l'interconnexion. En effet, le temps de lecture consiste en la somme du temps pour transmettre la donnée sur la ligne et du temps pour la détection de cette donnée par l'amplificateur : le temps pour transmettre la donnée peut être élevé si une grosse variation d'amplitude est nécessaire, c'est-à-dire si la résistance totale – résistance série de la ligne et impédance d'entrée – est importante. Pour diminuer l'impédance d'entrée, un circuit présentant une résistance interne négative est indispensable pour compenser la résistance positive utilisée pour la lecture. Deux exemples de circuit avec une résistance positive et une résistance négative sont présentés Figure 74.



(a) **(b)**

Figure 74 Principe de la lecture en courant.

Le schéma Figure 74-a peut être réalisé en utilisant un simple transistor MOSFET monté en grille commune [Seno93], comme illustré Figure 75-a. Ce montage est dérivé des amplificateurs de courant à base de transistor bipolaire (Figure 75-b).



(a) **(b)**

Figure 75 Amplificateurs de courant : (a) MOS et (b) bipolaire.

Le potentiel de référence V_{REF} sert à mettre le transistor dans la région de saturation, il est donc ici égal à V_{DD}. Le schéma équivalent petit signal, montré Figure 76, permet de calculer l'impédance d'entrée :

$$Z_{IN} = \frac{V_{IN}}{i_{IN}} = \frac{V_{GS}}{I_{IN}}$$

Équation 49

car la tension sur la grille est égale à V_{DD}. On obtient :

$$Z_{IN} = \frac{1 + g_d R_C}{g_m + g_d}$$

Équation 50

avec g_m la transconductance, g_d la conductance et R_C la résistance de charge. Puisqu'en saturation, la conductance g_d est faible, on a alors :

$$Z_{IN} = 1/g_m$$

Équation 51

L'impédance d'entrée est proche de zéro. La résistance du transistor NMOS est la suivante :

$$R_{NMOS} = Z_{IN} - R_C$$

Équation 52

ce qui montre que l'impédance introduite par le transistor est négative. Améliorer la vitesse de ce circuit revient à augmenter la transconductance g_m puisque cela fait diminuer Z_{IN}. Cependant, un compromis doit être trouvé car pour augmenter g_m, il faut soit augmenter le courant de polarisation au détriment de la consommation, soit augmenter la taille du transistor au détriment de la surface.

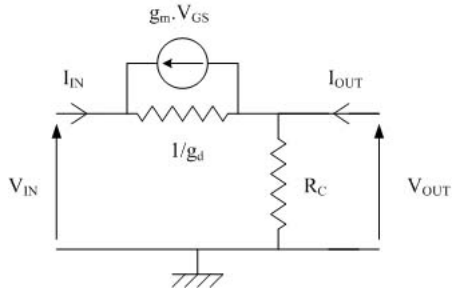


Figure 76 Schéma équivalent petit signal de l'amplificateur de type A.

Ces circuits de lecture, appelés de type A [Shib96], ne sont pas efficaces en tant qu'amplificateurs de courant car leur impédance d'entrée et leur vitesse dépendent de la transconductance du MOSFET. Celle-ci est bien plus faible que la transconductance d'un bipolaire, ce qui a rendu difficile la conception de circuits de lecture comme indiqué dans [Seno93]. C'est encore plus délicat en très basse tension car la transconductance des transistors MOS est beaucoup plus faible. En SOI, elle peut être améliorée en utilisant un DTMOS, qui présente un meilleur rapport I_{ON}/I_{OFF} que les transistors à body flottant. La transconductance d'un DTMOS demeure néanmoins plus de deux fois plus faible que celle des transistors MOS à tension V_{DD} nominale (voir Figure 77).

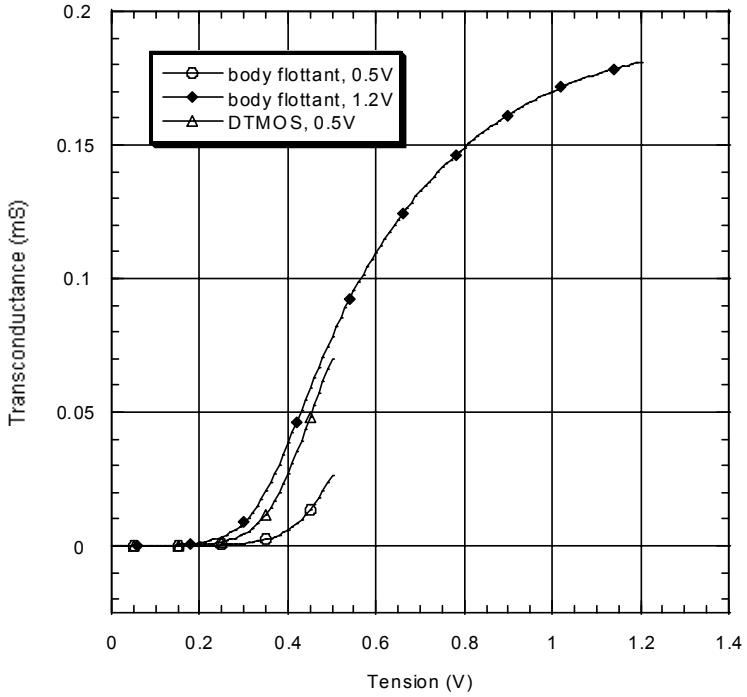
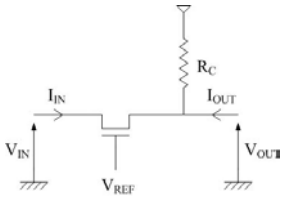
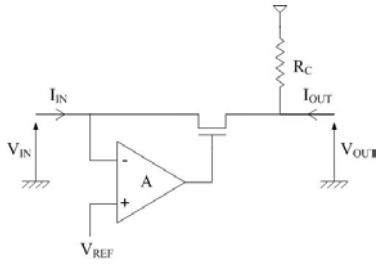
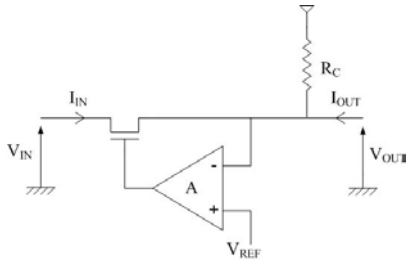


Figure 77 Transconductance d'un transistor NMOS à body flottant à 0.5V et 1.2V et monté en DTMOS à 0.5V en fonction de la tension appliquée sur la grille.

Il faut donc améliorer la transconductance du transistor grâce à un amplificateur additionnel. Cet amplificateur sert à contrôler le potentiel de référence V_{REF} de la grille du MOSFET, auparavant maintenue à un potentiel fixe. Il existe deux familles d'amplificateurs : les amplificateurs à rétroaction de l'entrée et les amplificateurs à rétroaction de la sortie. Les circuits de lecture ont donc été classés selon trois types par Shibata [Shib96], qui sont présentés dans le Tableau 7. Le type A est simplement le circuit vu précédemment avec le transistor MOSFET à grille commune. Le type B a une rétroaction de l'entrée par l'intermédiaire d'un amplificateur de gain A. Le type C a une rétroaction de la sortie, également par l'intermédiaire d'un amplificateur de gain A. Z_{TR} représente la trans-impédance du montage (dV_{OUT}/dI_{IN}).

Tableau 7 Les principaux types d'amplificateurs de courant.

	Type A	Type B	Type C
			
Z_{IN}	$\frac{1}{g_{m0}}$	$\frac{1}{(1+A)\cdot g_{m0}}$	$\frac{1-A\cdot R_C\cdot g_{m0}}{g_{m0}}$
Z_{TR}	R_C	R_C	R_C

Pour les circuits de type B et C, la transconductance effective est augmentée de la manière suivante : lorsque le potentiel de la source du transistor MOS augmente, le potentiel de sa grille diminue. Ainsi, la tension V_{GS} varie sur de plus grandes amplitudes. Pour le circuit de lecture de type B, on obtient par analyse petits signaux l'impédance d'entrée suivante :

$$Z_{IN} = \frac{1}{(1+A)\cdot(g_m + \frac{g_d}{1+A})}$$

Équation 53

En négligeant la conductance devant la transconductance, l'Équation 53 se simplifie en :

$$Z_{IN} = \frac{1}{(1+A)\cdot g_m}$$

Équation 54

Par analogie avec l'Équation 51, on obtient la transconductance effective suivante :

$$g_{me} = (1+A)\cdot g_m$$

Équation 55

La transconductance effective du transistor MOS est augmentée 1+A fois.

Pour le circuit de lecture de type C, on a un gain en boucle dû à la rétroaction de la sortie. La tension de sortie V_{OUT} a pour valeur initiale :

$$V_{OUT} = R_C \cdot g_{m0} \cdot V_{IN}$$

Équation 56

avec g_{m0} la transconductance intrinsèque du transistor. Après la rétroaction, la tension de sortie prend comme valeur :

$$V_{OUT} = R_C \cdot g_{me} \cdot V_{IN}$$

Équation 57

avec g_{me} la transconductance effective :

$$g_{me} = (1 + A \cdot R_C \cdot g_{me}) \cdot g_{m0}$$

Équation 58

Une plus grande transconductance peut être obtenue grâce à la rétroaction de la sortie par rapport à une rétroaction de l'entrée. Néanmoins, la réponse à un échelon en courant est dégradée par rapport au circuit de lecture de type B car l'amplificateur de gain A est inclus dans le chemin critique de l'amplificateur de lecture.

Les circuits de type C étant essentiellement des circuits différentiels, nous choisissons de partir d'un circuit de lecture de type B pour développer le circuit de transmission en mode courant, de manière à éviter d'avoir à utiliser deux lignes d'interconnexions parallèles pour transmettre une seule donnée. Avant de commencer, nous allons voir dans le paragraphe suivant les circuits qui ont été proposés récemment, ainsi que leurs avantages et inconvénients.

6.3.3 Etat de l'art

En 2001, Maheshwari et Burleson [Mahe01] ont proposé un circuit de lecture différentiel et préchargé appelé « *modified clamped bit-line sense amplifier* » basé sur un circuit de lecture utilisé dans les mémoires DRAM [Blal92], qui est montré Figure 78. C'est un circuit de type C car les sorties sont utilisées pour la rétroaction. Le fonctionnement de ce circuit est le suivant. Initialement, les sorties sont égales à $V_{DD}/2$ grâce à l'activation du transistor d'égalisation : les inverseurs rétro-couplés sont maintenus dans une région où leur gain est très élevé. Des signaux d'entrée différentiels sont utilisés : du courant est envoyé dans une des

entrées et tiré de l'autre. Lorsque le temps écoulé est suffisamment important pour que la différence de courant soit significative, le transistor d'égalisation est coupé. La sortie correspondant à l'entrée depuis laquelle le courant est retiré bascule alors au niveau logique '0' tandis que la sortie complémentaire rejoint le niveau logique '1'. Les avantages de ce circuit sont :

- sa vitesse grâce à la précharge qui maintient les inverseurs dans un état de gain en tension élevé,
- sa robustesse au bruit de mode commun grâce aux entrées différentielles.

Ses inconvénients sont :

- une surface occupée importante du fait de la présence de deux interconnexions,
- la nécessité d'utiliser un signal de synchronisation en tension entre l'émetteur et le récepteur, signal qui doit être en mode tension.

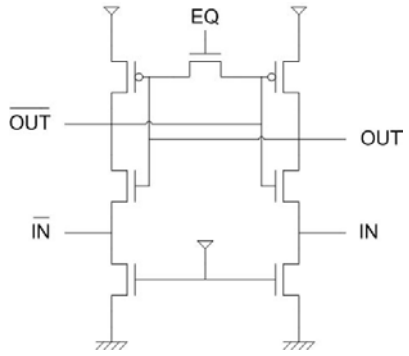


Figure 78 Circuit de lecture de type C appelé « *Modified Clamped Bit-line Sense Amplifier* ».

En 2002, Katoch [Kato02] a proposé un schéma de transmission en mode courant basé sur la propagation d'impulsions – voir Figure 79. A l'entrée de la ligne, un générateur d'impulsions transforme un changement de niveau logique en un train d'impulsions, qui est utilisé pour commander une source de courant. Le courant injecté dans la ligne est recopié au niveau du récepteur par un miroir de courant : ainsi, la variation en tension a lieu sur un nœud interne faiblement capacitif plutôt que sur l'interconnexion. Lorsque la variation en tension est suffisamment importante, le latch est verrouillé au niveau logique '1' puis est réinitialisé à '0' par une ligne à retard composée d'une chaîne de trois inverseurs. Le train d'impulsions en tension obtenu est reconverti en niveaux logiques par le régénérateur de niveau. L'avantage de ce circuit est qu'il est statique, c'est-à-dire qu'il ne nécessite pas de signal de

synchronisation avec l'émetteur. Son inconvénient, qui n'est pas des moindres, est que la largeur des impulsions est un paramètre de conception très important, qui dépend de la capacité de la ligne : une impulsion trop courte ne fournirait pas suffisamment de courant pour charger le nœud interne et faire basculer la sortie du latch tandis qu'une impulsion trop longue déchargerait entièrement la ligne d'interconnexion.

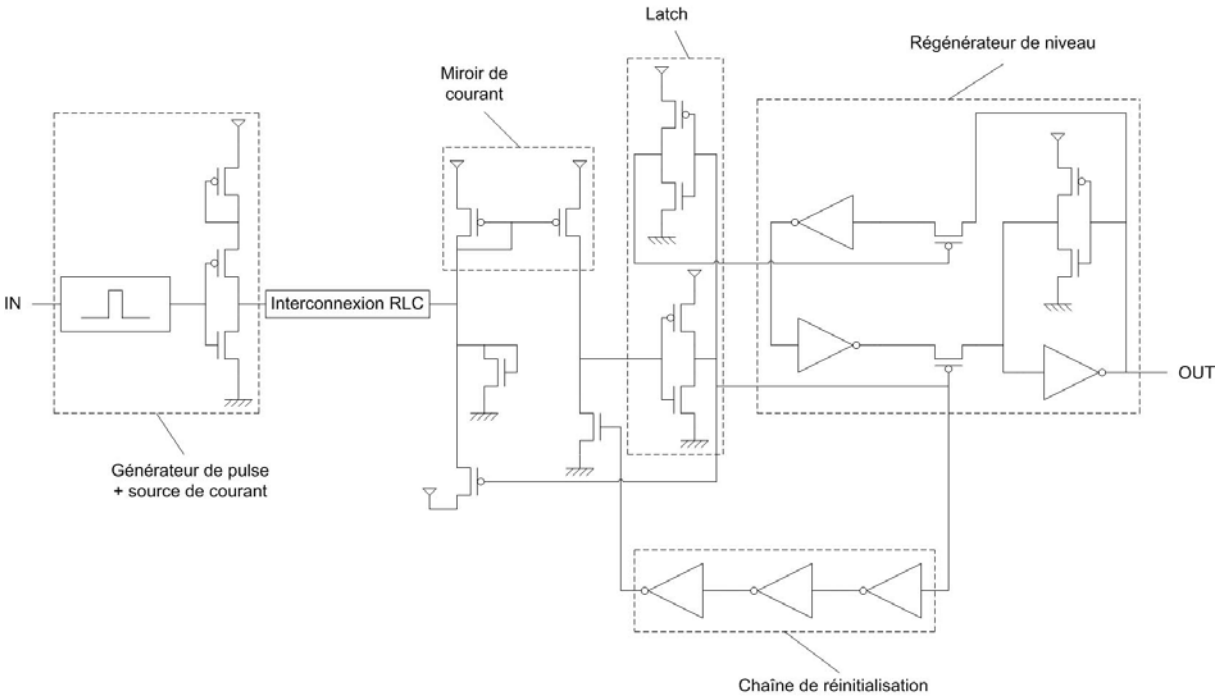


Figure 79 Circuit de transmission en mode courant basé sur la propagation d'impulsions.

Ce que nous voulons développer est un circuit de transmission en mode courant qui ne nécessite pas de signal de synchronisation en mode tension, car dans ce cas inverse, le problème n'est qu'à moitié résolu. Le circuit doit donc être statique et son fonctionnement ne doit pas dépendre de paramètres de conception variables comme dans le cas du schéma présenté ci-dessus : nous nous attacherons par conséquent à développer un circuit émetteur-récepteur auto-régulé.

6.3.4 La transmission en mode courant

Puisque nous avons fait le choix de ne pas utiliser de signal de synchronisation en mode tension, nous ne disposons pas de signaux de précharge et de lecture. Nous ne pouvons donc pas utiliser de circuits de lecture dynamiques puisque nous ne savons pas à quel moment la donnée va arriver : nous ne considèrerons que des circuits de lecture statiques. L'inconvénient de cette contrainte est la relative lenteur des amplificateurs de courant statiques par rapport

aux amplificateurs dynamiques. En effet, les circuits de lecture dynamiques sont préchargés dans un état où ils présentent un gain élevé et commutent rapidement vers un état stable lors de la phase de lecture. Nous voulons également nous affranchir de l'utilisation de deux interconnexions pour transmettre une donnée : les circuits développés dans la suite ne devront pas être des circuits différentiels. La Figure 80 illustre le concept de la transmission en mode courant que nous allons utiliser.

Figure 80 Schéma de la transmission en mode courant.

L'élément de base de ce schéma de transmission est le circuit de lecture. Nous faisons le choix d'un circuit de type B, c'est-à-dire possédant une rétroaction de l'entrée, puisque les circuits de type C sont généralement des circuits différentiels. Le circuit de lecture proposé, appelé LI²-CSA pour 'Low Input Impedance Current Sense Amplifier', possède un amplificateur de type inverseur, de gain proche de 1 – voir Figure 81.

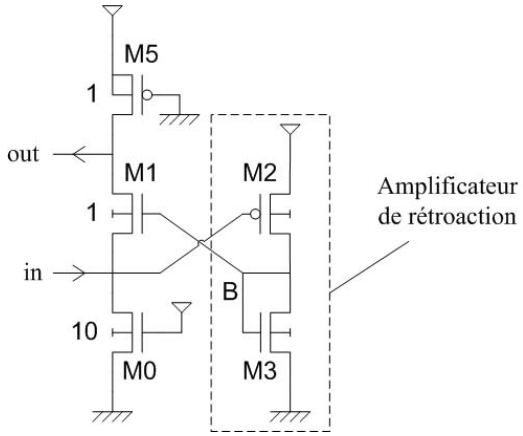


Figure 81 Amplificateur de type B proposé.

La source de courant est un simple inverseur CMOS. L'entrée du circuit de lecture est appelée *in* et la sortie *out*, pour des raisons de simplicité. Le transistor M0 génère le courant de polarisation I_P. Le transistor M5 représente la résistance aux bornes de laquelle la lecture du courant est effectuée : c'est la résistance R_C illustrée dans le Tableau 7. Le circuit dans le cadre en pointillés est l'amplificateur additionnel : son rôle est de générer une tension de polarisation pour le transistor M1 qui soit en opposition de phase avec le signal d'entrée. Les

transistors M2 et M3, avec M3 monté en diode, gardent le transistor M1 dans un état passant : ils sont de type DTMOS pour améliorer le gain de la rétroaction – proche de 1.

Le fonctionnement du montage est le suivant. Quand l'entrée de l'inverseur vaut zéro (voir Figure 80), du courant passe dans l'interconnexion vers l'entrée du circuit de lecture, augmentant la tension de la source du transistor M1. Grâce à l'amplificateur de rétroaction, la tension du nœud B commence à décroître, abaissant encore la tension V_{GS} du transistor M1 et augmentant la tension de sortie. Puisque le courant de drain d'un transistor varie exponentiellement avec la tension grille-source, la sortie *out* présente un gain en tension important : les valeurs d'excursion maximales de la tension de sortie sont proches des niveaux logiques CMOS. Réciproquement, quand l'inverseur tire du courant de l'entrée *in* du circuit de lecture, le potentiel de la source du transistor M1 commence à décroître, entraînant des effets inverses à ceux décrits précédemment. Les valeurs des tensions V_{IN}, V_{OUT} et V_B sont montrées Figure 82.

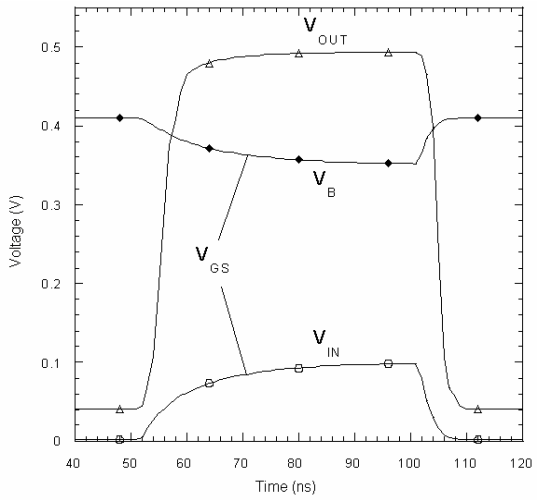


Figure 82 Tensions simulées aux noeuds in, out et B.

Le rôle du transistor M0 est de limiter l'amplitude des variations en tension sur l'interconnexion. Cependant, il limite la vitesse de charge de la ligne en tirant à la masse tout le courant provenant de l'inverseur ; or V_{IN} est la quantité qui contrôle le courant i_{OUT} – c'est ici V_{IN} mais cela peut être n'importe quelle autre tension interne, par exemple la tension de sortie. Cela a pour conséquence que la charge de la ligne est plus lente que la décharge, d'où des inégalités au niveau des temps de propagation. Pour pallier ce problème, le circuit de

lecture de type B est modifié pour présenter une impédance d'entrée variable – voir Figure 83 – et est appelé VI²-CSA, pour 'Variable Input Impedance Current Sense Amplifier'.

Ce circuit exploite les caractéristiques des transistors opérés en très basse tension. En effet, lorsque l'entrée du circuit de lecture vaut zéro, sa sortie vaut aussi zéro et le transistor M0 est bloqué : l'impédance d'entrée du circuit est donc importante et nous ne sommes plus dans le cas idéal de la transmission en courant. Néanmoins, il ne faut pas oublier que la quantité de contrôle est V_{IN} et qu'il faut de toute façon une variation en tension à l'entrée. L'avantage est ici une charge plus rapide de l'interconnexion puisque le courant provenant de l'inverseur n'est plus tiré à la masse par le transistor M0. Pour une faible variation de la tension de l'interconnexion, le transistor M0 est activé et entre dans un état de saturation puisque, comme nous l'avons vu dans le paragraphe 4.1, la tension de drain pour laquelle les transistors saturent en ULV est à peu près égale à 100mV. La charge de la ligne est dès lors limitée.

Pour comparer les deux circuits précédents, nous mettons sur l'entrée D de l'inverseur un signal de rapport cyclique ½ et nous observons le temps de propagation de ce signal à la montée (tp_{HH}) et à la descente (tp_{LL}) en fonction de sa fréquence. Comme on peut le voir Figure 84, les temps de propagation sont beaucoup plus stables avec le circuit VI²-CSA.

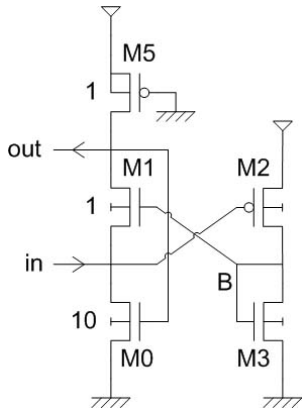


Figure 83 Circuit de lecture modifié avec impédance d'entrée variable VI²-CSA.

Cependant, l'inconvénient de ces deux circuits est leur importante consommation statique, due au court-circuit réalisé. Dans les deux cas, l'amplitude de la charge de l'interconnexion est limitée par le transistor M0 qui fait chuter à la masse tout le courant provenant de l'inverseur. Il nous faut donc trouver un moyen d'autolimiter cette charge. C'est dans ce sens que le montage émetteur-récepteur suivant a été introduit.

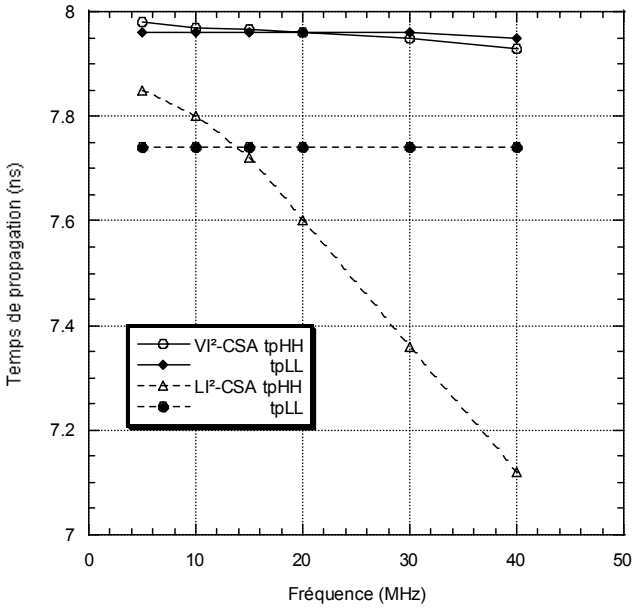


Figure 84 Temps de propagation du circuit LI²SA et du circuit VI²SA en fonction de la fréquence du signal.

6.3.5 Circuit émetteur-récepteur ARCS

L'idée de ce circuit, appelé ARCS pour 'Auto-Regulated Current Sensing scheme', est d'arrêter la source de courant dès que la charge de la ligne a atteint un niveau suffisant pour permettre à l'amplificateur de courant en sortie de l'interconnexion de lire la bonne valeur. Il est composé d'une partie émettrice et d'une partie réceptrice (voir Figure 85). L'élément de base est le circuit de lecture en courant précédent, sans le transistor M0 qui n'est plus nécessaire.

Le circuit compris dans le cadre en pointillés est l'émetteur. Il est constitué d'un circuit de lecture appelé S1 et d'une source de courant, commandée par S1. Le fonctionnement est le suivant : lorsque l'entrée D_{IN} passe à '0', le transistor Md1 est activé puisque la tension au point *P* vaut 0. Du courant passe alors dans l'interconnexion, ce qui augmente légèrement son potentiel. La sortie du circuit de lecture S1 initialement à '0' passe à '1' lorsque l'amplitude de la tension sur la ligne est suffisante, ce qui désactive le transistor de charge Md1 par l'intermédiaire du signal C et des transistors M0 et M1. Le délai introduit par l'inverseur I1 et les transistors M0 et M1 assure que le circuit de lecture S2 en sortie de l'interconnexion sera en mesure de lire la bonne valeur; en effet, celui-ci compense le délai introduit par la ligne de transmission (qui implique que S2 va commuter après S1) et garantit que l'amplitude en sortie de la ligne sera suffisante.

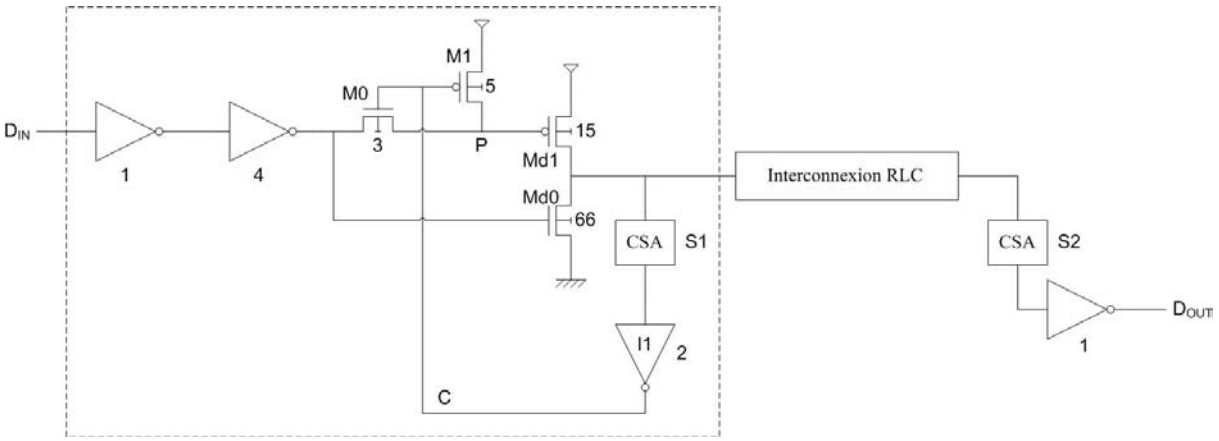


Figure 85 Schéma du circuit émetteur-récepteur ARCS.

Pour évaluer la sensibilité du circuit aux différentes sources de bruit, nous utilisons la méthode d'analyse du pire cas du rapport signal/bruit présentée dans [Dall98] pour mesurer la robustesse du circuit émetteur-récepteur, à l'aide de simulations Monte Carlo. Les sources de bruit sont classées dans deux catégories : les sources de bruit proportionnelles à l'amplitude du signal et les sources de bruit indépendantes :

$$V_N = K_N \cdot V_S + V_{IN}$$

Équation 59

$K_N.V_S$ représente les sources de bruit qui sont proportionnelles à l'amplitude du signal (V_S) comme le couplage capacitif entre deux interconnexions adjacentes, et les bruits d'alimentation introduits par le signal V_S. Le terme V_{IN} représente quant à lui les sources de bruit indépendantes, dues aux disparités des paramètres de la technologie telle la tension de seuil, ou aux variations de la tension d'alimentation indépendantes du signal V_S. Le Tableau 8 résume ces différents paramètres de bruit, dont la description est donnée par la suite.

Le coefficient de couplage capacitif K_C est égal au rapport de la capacité de couplage latéral sur la capacité intrinsèque de l'interconnexion. Pour des lignes de 10 mm de long, espacées de 2µm et de 2pF de capacité, K_C vaut 0,2. Le coefficient d'atténuation des bruits de couplage Atn_C est lié à la capacité du buffer, chargeant ou déchargeant la ligne, à compenser un bruit capacitif. Pour un inverseur CMOS, il est égal à 0,2; dans le cas du circuit ARCS, puisque la source de courant n'est pas toujours activée, nous le fixons à 1, c'est-à-dire que nous considérons le pire cas dans lequel il n'y a pas d'atténuation.

Le bruit d'alimentation K_{PS} introduit par la commutation du signal V_S est estimé à 5% pour une transmission monofil et à 1% pour une transmission différentielle. Le pire cas pour le paramètre K_N est alors :

$$K_N = Atn_C \cdot K_C + K_{PS}$$

Équation 60

La sensibilité *Rx_S* et le décalage en entrée *Rx_O* du récepteur dépendent des variations des paramètres de procédé et de la tension d'alimentation.

Le bruit d'alimentation *PS* indépendant du signal V_S est limité à 5% de V_{DD} pour une matrice d'alimentation bien conçue. Le coefficient d'atténuation du bruit d'alimentation Atn_{PS} est défini comme étant le rapport de la variation de la tension de seuil du récepteur sur la variation de la tension d'alimentation.

Le décalage émetteur-récepteur *Tx_O* est dû aux variations des paramètres technologiques et de la tension d'alimentation V_{DD} entre l'émetteur et le récepteur. Le pire cas pour le paramètre V_{IN} est :

$$V_{IN} = Rx\_O + Rx\_S + Atn_{PS} \cdot PS + Tx\_O$$

Équation 61

Le rapport signal/bruit SNR prend pour valeur :

$$SNR = \frac{V_S}{2 \cdot V_N}$$

Équation 62

Tableau 8 Les différentes sources de bruit.

K_N	K_C	Coefficient de couplage capacitif.
	Atn_C	Coefficient d'atténuation du couplage capacitif.
	K_{PS}	Bruit d'alimentation dû à la commutation de V_S.
	Pire cas :	$K_N=Atn_C.K_C+K_{PS}$
V_{IN}	*Rx_O*	Décalage d'entrée du récepteur.
	Rx_S	Sensibilité du récepteur.
	PS	Bruits d'alimentation indépendants : 5% de V_{DD}
	Atn_{PS}	Atténuation du bruit d'alimentation.
	Tx_O	Décalage émetteur-récepteur.
	Pire cas :	$V_{IN}=Rx\_O+Rx\_S+Atn_{PS}.PS+Tx\_O$

Ces différents paramètres de bruit sont calculés pour le circuit émetteur-récepteur et comparés à ceux d'un inverseur générant des signaux d'amplitude maximale. Les résultats sont donnés dans le Tableau 9. On note que le circuit en mode courant a un rapport signal/bruit 4 fois moins bon qu'un inverseur, ce qui n'est pas étonnant du fait du très mauvais coefficient d'atténuation que nous avons choisi : celui-ci vaut 1 pour le circuit ARCS contre 0.2 pour l'inverseur ce qui rend le circuit en mode courant beaucoup plus sensible aux bruits de couplage. De plus, l'amplitude du signal V_S est cinq fois plus faible que pour l'inverseur qui génère des signaux dont la tension est comprise entre 0 et V_{DD} : le rapport signal/bruit s'en trouve donc dégradé. Cependant, le circuit mode courant est beaucoup moins sensible aux bruits d'alimentation grâce à un coefficient d'atténuation Atn_{PS} de 0,048 contre 0,52. De plus, sa sensibilité aux variations des paramètres technologiques est équivalente à celle de l'inverseur : *Rx_O* à 0,032V contre 0,028V et *Rx_S* à 0,014V contre 0,013V.
Le rapport signal/bruit est uniquement dégradé du fait de couplages capacitifs. La règle de conception, pour avoir une transmission fiable, est donc d'espacer les interconnexions puisque le coefficient de couplage est linéairement proportionnel à la distance. Pour avoir le même rapport signal/bruit qu'un inverseur, il faut mettre quatre fois plus d'espace entre deux lignes de transmission : ça n'est pas très pénalisant pour des signaux globaux puisque le routage s'effectue avec des niveaux de métal élevés (métal 5, 6 ou 7) qui ont en général une densité de surface assez faible.

Tableau 9 Valeurs des différents paramètres de bruit pour le circuit émetteur-récepteur et l'inverseur.

Circuit	V_S (V)	K_C	Atn_C	K_{PS} (V)	*Rx_O* (V)	*Rx_S* (V)	*PS* (V)	Atn_{PS}	*Tx_O* (V)	*SNR*
Emetteur-récepteur	0,1	0,2	1	0,005	0,032	0,014	0,025	0,048	0,03	0,51
Inverseur CMOS	0,5	0,2	0,2	0,025	0,028	0,013	0,025	0,52	0,038	2,01

6.3.6 Comparaison courant-tension

Nous comparons les performances et la consommation des circuits VI²-CSA et ARCS à celles d'une chaîne d'inverseurs en fonction de la longueur de ligne à attaquer. La chaîne d'inverseurs et les circuits de lecture en courant ont été dimensionnés avec le produit

puissance-délai comme objectif d'optimisation. Le signal d'entrée a un rapport cyclique de ½. Le délai normalisé des trois circuits est donné Figure 86.

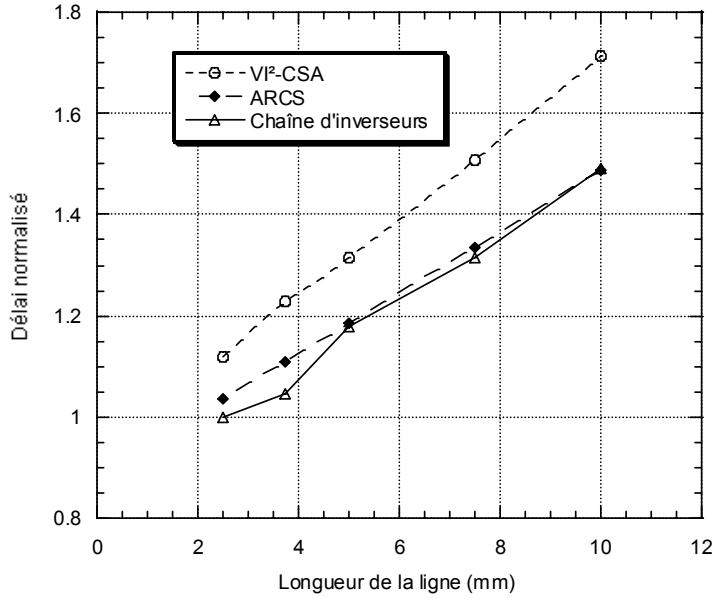


Figure 86 Comparaison des délais normalisés des circuits VI²-CSA, ARCS et de la chaîne d'inverseurs en fonction de la longueur de la ligne de transmission.

Comme on peut le voir, les trois délais augmentent linéairement avec la longueur de la ligne et donc avec sa charge : c'est assez intuitif pour la chaîne d'inverseurs, mais ça l'est moins pour les amplificateurs de courant. En effet, les circuits de lecture en mode courant présents dans les SRAMs ont notamment pour avantage d'avoir un temps de réponse qui dépend très peu de la capacité de charge de la ligne de bit – voir par exemple [Seev91]. Or dans notre cas, comme indiqué précédemment, les circuits ne peuvent pas être préchargés et donc ne sont pas dans un état où ils présentent un gain élevé : la réponse de l'amplificateur de courant dépend beaucoup de la quantité de contrôle de la rétroaction, qui est V_{IN} – puisque ce sont des circuits de type B. Cela implique donc d'avoir une charge de la ligne de transmission, qui même si elle demeure très limitée, dépend de sa longueur.

La chaîne d'inverseurs est le circuit le plus rapide ce qui semble encore une fois contre intuitif. Cependant, cela s'explique pour trois raisons :

- en ULV, le délai est dans les portes et pas dans les fils : bien que la transmission d'une donnée soit intrinsèquement plus rapide en mode courant qu'en mode tension, comme nous l'avons vu au paragraphe 6.1.2, ce gain se trouve atténué ;

- les circuits de lecture sont statiques et ne peuvent pas être maintenus dans un état de gain en tension élevé, ce qui a pour conséquence qu'ils sont moins rapides que des circuits de lecture dynamiques ;
- enfin, nous avons choisi d'optimiser le produit puissance-délai des différents circuits pour avoir une base de comparaison correcte : puisque les amplificateurs de courant consomment beaucoup moins à haute fréquence pour des charges élevées – comme nous allons le voir par la suite –, il reste une marge d'amélioration pour les circuits en mode courant.

La comparaison de la consommation des circuits a été effectuée pour des fréquences allant de 1MHz à 80 MHz et des longueurs de ligne valant 2,5mm, 5mm, 7,5mm et 10mm : voir respectivement les Figure 87-a, -b, -c et -d. Le circuit VI²-CSA a une consommation relativement constante avec la fréquence, ce qui est dû au courant de court-circuit pendant une demi-période. La chaîne d'inverseurs et le circuit ARCS ont une dissipation de puissance qui varie linéairement avec la fréquence. Nous pouvons constater que la chaîne d'inverseurs a une dissipation purement dynamique tandis que le circuit ARCS possède :

- une composante statique illustrée à fréquence faible,
- une composante dynamique, représentée par la pente de la courbe, qui est plus petite que celle de la chaîne d'inverseurs.

Pour des lignes de distances inférieures à 2,5mm, la chaîne d'inverseurs aura toujours un avantage au niveau de la consommation par rapport au circuit ARCS quelque soit la fréquence. Par contre, pour une ligne de 5mm, on constate que le circuit ARCS consomme moins d'énergie dès la fréquence de 20MHz et que le gain peut atteindre 30% pour des fréquences élevées. Pour 7,5mm, le circuit ARCS consomme moins dès 10MHz et le gain s'élève jusqu'à 45%. Pour 10mm, le gain est obtenu dès 5MHz et culmine à 55%.

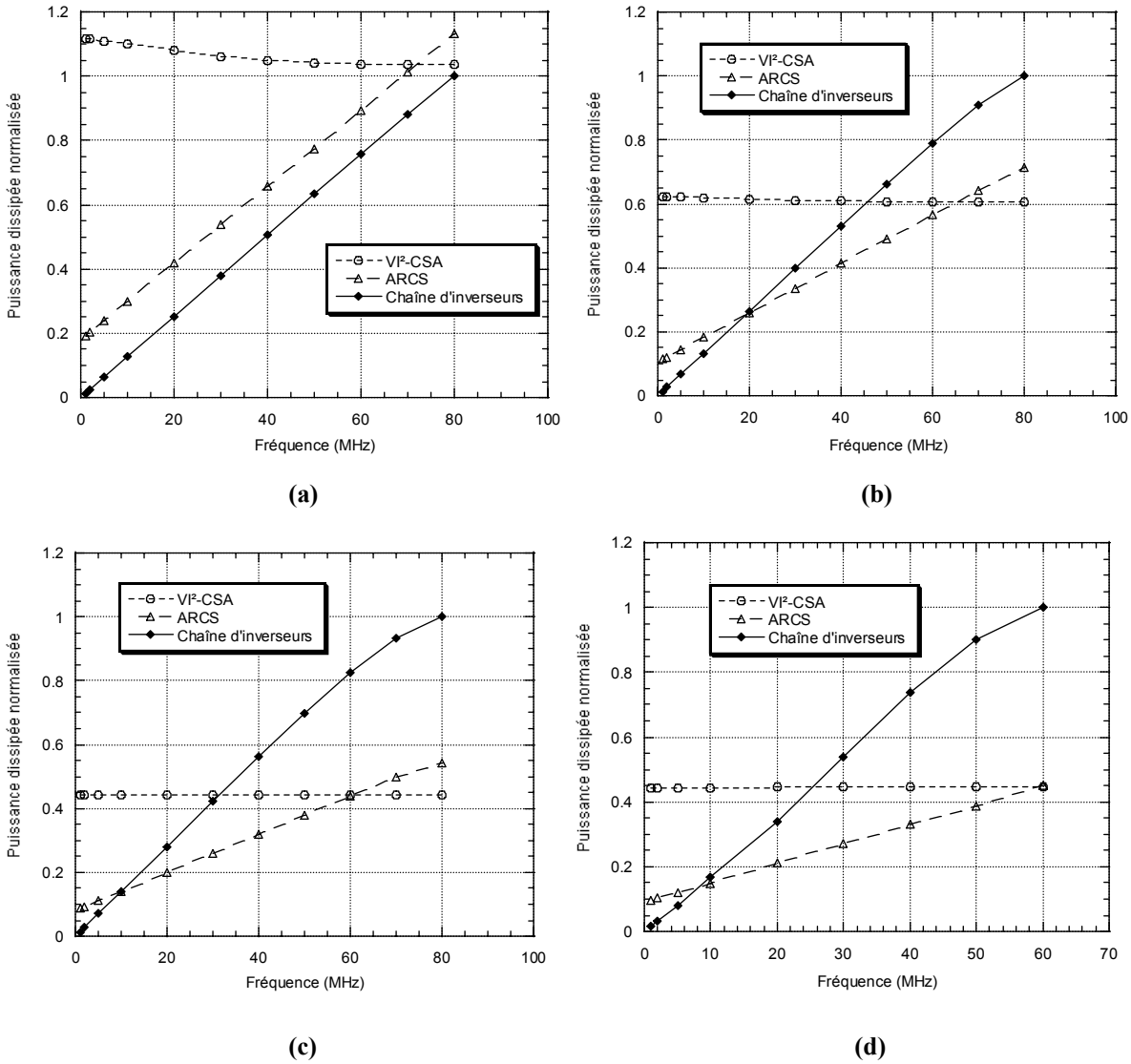


Figure 87 Comparaison de la puissance dissipée des circuits VI²-CSA, ARCS et de la chaîne d'inverseurs pour des longueurs de ligne de : (a) 2.5mm, (b) 5mm, (c) 7.5mm et (d) 10mm.

On peut donc définir un espace de conception pour décider dans quel cas choisir une transmission en mode courant ou une transmission en mode tension. Cet espace de conception est représenté Figure 88. Il est évident que la transmission en mode courant sera intéressante dans le cas de lignes hautement capacitives de part sa composante dynamique réduite alors que la transmission en mode tension doit être utilisée pour des lignes courtes et des taux d'activité faibles.

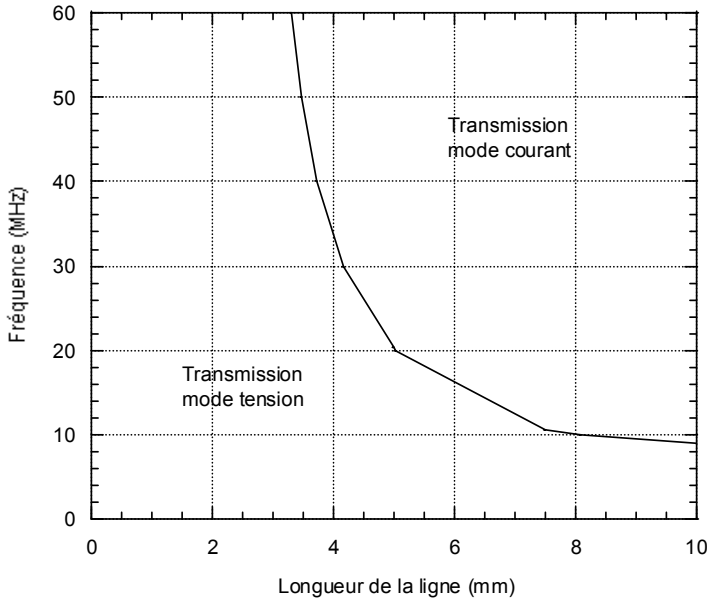


Figure 88 Espace de conception en fonction de la longueur de l'interconnexion et de la fréquence du signal transmis.

6.4 Conclusion

L'étude des lignes de transmission a été effectuée. Il a été montré que leur modélisation ne nécessite que peu d'éléments RLC. L'étude de la transmission en mode tension a révélé que, en très basse tension, la chaîne d'inverseur est une configuration plus avantageuse que l'insertion de répéteurs, puisque le délai se trouve dans les portes et pas dans les fils. Les cas de l'optimisation d'une chaîne d'inverseurs en délai et sous contrainte de puissance ont été étudiés : pour cela, un modèle du temps de propagation d'un inverseur a été proposé, puis un programme a été développé prenant en compte les effets de canal étroit sur les caractéristiques I-V et les discontinuités dues aux contacts de body.

L'étude des amplificateurs de courant dans leur généralité a été réalisée et leur utilisation dans les mémoires SRAMs expliquée. Deux circuits de lecture ont été proposés : le circuit VI²-CSA et le circuit ARCS. Ce dernier est celui qui présente les meilleures caractéristiques performance-consommation. Comparée au mode tension, on note que la configuration ARCS est surtout intéressante pour de longues lignes de transmission et pour des fréquences élevées. Dans ce cas, le gain en puissance dissipée peut atteindre 55% à 80MHz, pour un temps de propagation égal.

7 Application au traitement d'images par ondelettes

Dans ce chapitre, nous allons nous intéresser au développement d'un circuit de traitement d'image utilisant la transformée en ondelettes. Ce travail s'inscrit dans le cadre du projet IRISEP. Le but du projet IRISEP est le développement d'un système d'identification par l'iris, qui soit fiable et d'un coût raisonnable, et puisse concurrencer les systèmes commerciaux existants. L'algorithme développé à l'ISEP [Rydg04-a], [Rydg04-b] est basé sur l'extraction de signatures à l'aide de paquets d'ondelettes. Nous allons dans une première partie faire un bref historique du traitement du signal et expliquer les limitations de la transformée de Fourier qui ont conduit au développement de la transformée en ondelettes. Il ne s'agit nullement de s'attaquer à la partie théorique ni aux choix qui ont été arrêtés concernant le type d'ondelette, ce qui serait hors propos, mais d'expliquer simplement ce qu'est une ondelette et permettre d'introduire les particularités de l'analyse par ondelette choisie pour extraire la signature de l'iris.

Dans un deuxième temps, nous détaillerons l'architecture utilisée pour implémenter la transformée en ondelette.

7.1 Historique du traitement du signal

7.1.1 La transformée de Fourier et ses dérivées

La transformée de Fourier

Lorsque l'on s'intéresse au domaine du traitement du signal, la transformée de Fourier est une référence majeure. En 1807, le mathématicien français Joseph Fourier indique que tout signal périodique peut être décomposé en une somme infinie de sinus et de cosinus de fréquences diverses, dont on fait varier d'une part les amplitudes en les multipliant par des coefficients, et d'autre part les phases, de manière à ce qu'elles s'additionnent ou se compensent. La transformée de Fourier a pour expression mathématique :

$$X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-i\omega t} dt$$

Équation 63

L'information sur le temps étant présente, la transformée inverse est possible. Cependant, cette information est cachée dans les phases et est en pratique impossible à extraire. Comme indiqué par l'équation précédente, il est nécessaire de connaître tout l'historique d'un signal pour pouvoir calculer sa transformée de Fourier : il faut alors choisir de l'analyser soit en fréquence, soit dans le temps, les deux analyses n'étant pas possible simultanément. Cela constitue la limitation majeure de la transformée de Fourier, car celle-ci ne permet pas d'étudier des signaux non-stationnaires, c'est-à-dire qui varient brusquement. Or, la plupart des signaux réels sont non-stationnaires : on peut citer l'exemple de la voix. Dans le domaine des images, la distinction entre signaux stationnaires et non-stationnaires est également importante : un signal qui se répète sera perçu comme une texture ou bien le fond de l'image, tandis qu'un signal qui ne se répète pas sera identifié comme étant un objet.

La transformée de Fourier à fenêtre

Pour palier le manque d'information temporelle d'une transformée de Fourier, la méthode d'analyse à base de transformée de Fourier à fenêtre est apparue. L'idée est non plus d'étudier un signal dans sa globalité mais de l'analyser par morceaux. Cette méthode revient à effectuer une analyse spectrale : on applique au signal une fenêtre qui sert de masque, à l'intérieur de laquelle le signal est considéré localement stationnaire, puis on déplace cette fenêtre le long du signal afin de l'analyser entièrement. Gabor reprendra cette méthode dans les années 40.

La fenêtre est alors représentée par une fonction gaussienne *g(x)*, et deux réels *a* et *b*, représentant respectivement la largeur de la fenêtre et sa translation :

$$g_{a,b}(t) = e^{iat} g(t-b)$$

Équation 64

Il existe toujours une limitation à cette méthode d'analyse : la taille de la fenêtre choisie. Si celle-ci est trop petite, l'information contenant la totalité du signal, représentée par les basses fréquences, est perdue. A l'inverse, si celle-ci est trop grande, l'information contenue dans les hautes fréquences, c'est-à-dire les détails, est noyée dans la moyenne. Il faut donc judicieusement choisir la taille de la fenêtre pour pouvoir décomposer un signal en une combinaison temps-fréquence correcte.

Néanmoins, certaines informations peuvent demeurer cachées, par exemple un signal variant brutalement au milieu d'une zone de faibles variations. De plus, il n'est pas possible d'avoir à la fois une vue d'ensemble et une vue détaillée du signal analysé. En décomposant un signal sous forme de fonctions qui ne sont plus des sinusoïdes pures, il est possible de condenser l'information dans les domaines temporel et fréquentiel en effectuant une analyse multi-résolution.

7.1.2 La transformée en ondelettes

L'analyse multi-résolution

Plutôt que de conserver une enveloppe fixe dans laquelle le nombre d'oscillations varie, l'utilisation des ondelettes permet de conserver un nombre constant d'oscillations dans une enveloppe que l'on peut contracter et dilater à volonté : c'est le principe de l'analyse multi-résolution. La Figure 89 détaille ce principe : le signal d'origine est décomposé en sous-signaux de différentes résolutions à l'issue de cinq étapes. La première étape transforme le signal *f(x)* en un signal de détail $W_2^1 f(x)$ et un signal d'approximation $S_2^1 f(x)$. Ce signal est à son tour décomposé en un détail et une approximation $W_2^2 f(x)$ et $S_2^2 f(x)$ et ainsi de suite jusqu'à obtenir $W_2^5 f(x)$ et $S_2^5 f(x)$. L'analyse multi-résolution permet de rapprocher les analyses temporelle et fréquentielle (ou spatiale et fréquentielle pour une image). Elle exploite l'idée que tout signal peut être construit par raffinements successifs, c'est-à-dire par l'ajout de détails en passant d'une résolution à l'autre.

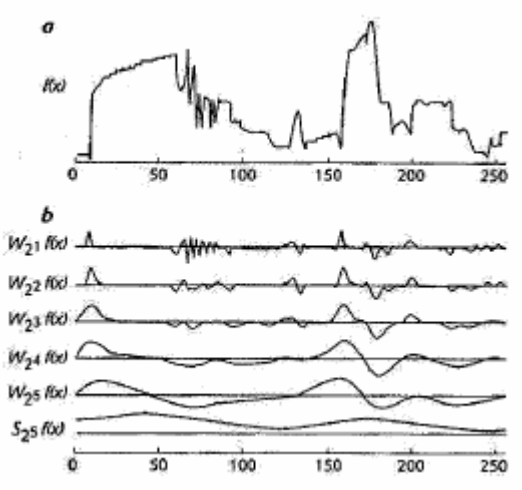


Figure 89 Transformée en ondelettes d'un signal montrant les fonctions d'approximation successives et la fonction de détail [Hubb95].

La transformée en ondelettes continue

L'analyse multi-résolution est intimement liée à la transformée en ondelettes. C'est en fait une manière de décrire la transformée en ondelettes : celle-ci consiste en effet à obtenir l'approximation d'un signal en le projetant sur un espace d'approximation $\psi(x)$ [Mall89]. Bien évidemment, pour ne pas perdre d'informations, il faut aussi projeter le signal sur un espace de détail $\varphi(x)$. La transformée en ondelettes utilise des translations et des dilatations (fonctions d'expansions) d'une fonction fixe appelée ondelette mère ψ. Dans le cas de la transformée en ondelettes continue, les paramètres de dilatation a et de translation b varient de manière continue :

$$\Psi_{a,b}(x) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-b}{a}\right)$$

Équation 65

La transformée en ondelettes d'un signal $f(x)$ produit une fonction de deux variables (le temps et l'échelle d'analyse du signal) $W(a,b)$ représentant la projection du signal f sur la base d'ondelettes $\psi_{a,b}$:

$$W(a,b) = \frac{1}{\sqrt{|a|}}\int f(x)\psi_{a,b}(x)dx$$

Équation 66

La transformée en ondelettes discrète

Puisque la translation de l'ondelette est continue, l'information obtenue est infiniment redondante. Pour diminuer cette redondance, la transformée en ondelettes discrète est utilisée : les translation et dilatations s'effectuent alors selon des valeurs discrètes. L'ondelette présentée dans l'Équation 65 est modifiée de la manière suivante :

$$\Psi_{m,n}(x) = \frac{1}{\sqrt{a_0^m}} \psi\left(\frac{1}{a_0^m} - nb_0 \right)$$

Équation 67

Les paramètres de translation et de dilatation discrétisés sont définis par :

$a = a_0^m$ et $b = nb_0 a_0^m$, avec $a_0 > 1$ et $b_0 > 0$, des entiers relatifs fixés [Cohe92].

Les ondelettes orthogonales

Cependant, il existe encore de la redondance. Pour la supprimer totalement, il faut utiliser des ondelettes dites orthogonales. L'orthogonalité signifie que l'information capturée par une ondelette est totalement décorrélée de celle capturée par une autre. L'utilisation d'ondelettes orthogonales va de soi pour la compression d'images, car cela permet de ne garder que l'information nécessaire et d'assurer la réversibilité.

Deux ondelettes mères orthogonales permettent d'effectuer une analyse multi-résolution orthogonale : les espaces de détail et d'approximation sont alors orthogonaux, c'est-à-dire que la projection des vecteurs de la base de l'un des espaces sur l'autre donne zéro. Chaque ondelette et sa fonction d'échelle associée sont également orthogonales. Une famille d'expansion très populaire a été crée par Ingrid Daubechies. Une autre fonction d'expansion également fréquemment utilisée est celle de Haar. Ces deux ondelettes et leur fonction d'échelle associée sont illustrées ci-dessous.

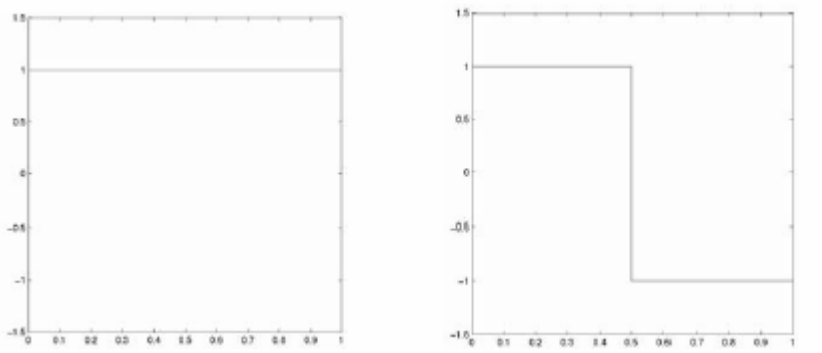


Figure 90 La fonction d'échelle de Haar (gauche) et l'ondelette (droite).

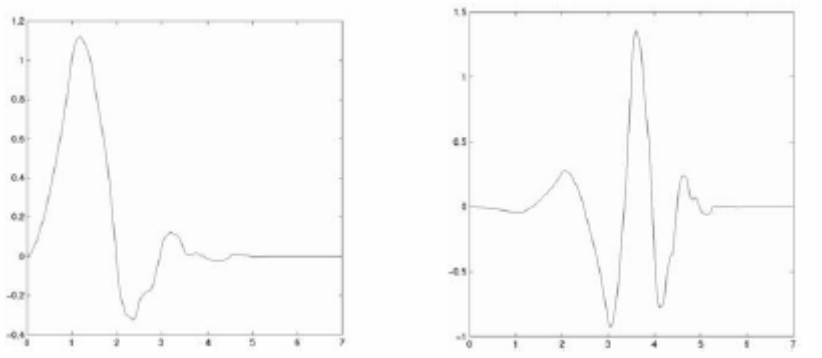


Figure 91 La fonction d'échelle Daubechies 4 (gauche) et l'ondelette associée (droite).

Les bancs de filtre

Une des raisons du succès de la transformée en ondelettes est son implémentation matérielle efficace. La méthode la plus répandue pour implémenter une transformée en ondelettes est l'utilisation de bancs de filtre sous forme d'une structure pyramidale, technique qui provient des travaux de Mallat [Mall89]. Deux filtres FIR calculent à chaque niveau les coefficients d'ondelette (les détails) et les coefficients d'échelle (l'approximation) : ils possèdent alors des caractéristiques respectivement passe-haut et passe-bas. Pour garder le même nombre d'échantillons en sortie et en entrée, les produits de convolution issus des filtres sont sous-échantillonnés par un facteur deux. Seule la sortie du filtre passe-bas, c'est-à-dire l'approximation, est de nouveau traitée par les deux filtres. Cette structure pyramidale est illustrée Figure 92 pour un filtrage 1D.

La transformée inverse peut être obtenue en faisant tourner l'algorithme à l'envers et en utilisant une autre paire de filtres FIR. Les quatre filtres (deux pour la décomposition et deux pour la reconstruction) associés au sous-échantillonnage forment un banc de filtres conjugués en quadrature.

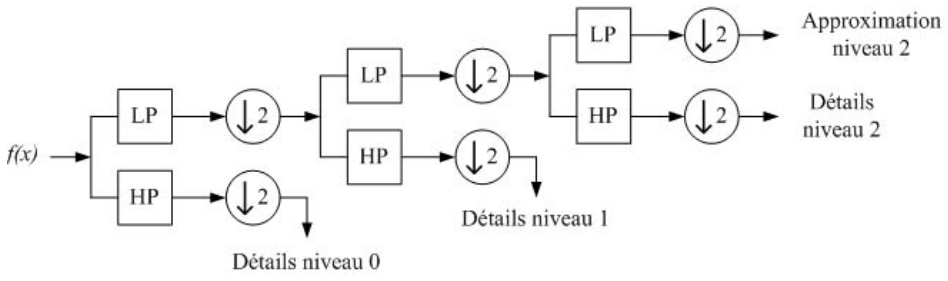


Figure 92 Transformée en ondelettes discrète 1D sur trois niveau à l'aide d'un banc de filtres.

Jusqu'à présent, seule la transformée en ondelettes sur une dimension a été considérée. Pour l'appliquer à un signal en deux dimensions – une image –, la transformée est calculée en deux

étapes. La première étape consiste à effectuer la transformée 1D précédente sur les lignes de l'image, ce qui produit deux nouvelles images issues des filtres passe-bas et passe-haut. La deuxième étape consiste à appliquer la transformée sur les colonnes : on obtient alors quatre images différentes, qui correspondent aux coefficients d'ondelette et de fonction d'échelle à un niveau donné. Ces quatre images représentent l'approximation, les détails horizontaux, les détails verticaux et les détails diagonaux de l'image d'origine. L'approximation provient du filtrage passe-bas dans les deux directions et est une image basse résolution de l'image d'origine. Les images contenant les détails subissent un filtrage passe-haut dans au moins une direction. La structure du banc de filtres utilisé pour la transformée en ondelettes discrète 2D est montrée Figure 93.

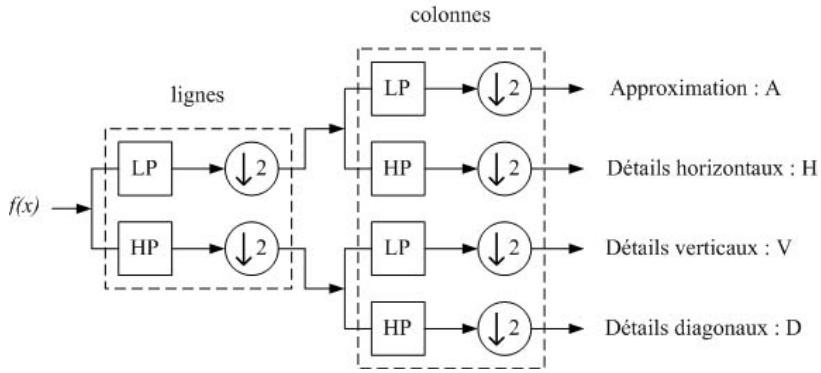


Figure 93 Banc de filtres pour la transformée en ondelettes discrète 2D.

Les coefficients obtenus sont rangés comme indiqué dans la Figure 94. Seuls les coefficients d'approximation sont transformés au niveau suivant, ce qui signifie que la partie basse fréquence de l'image est divisée en parties de plus en plus petites.

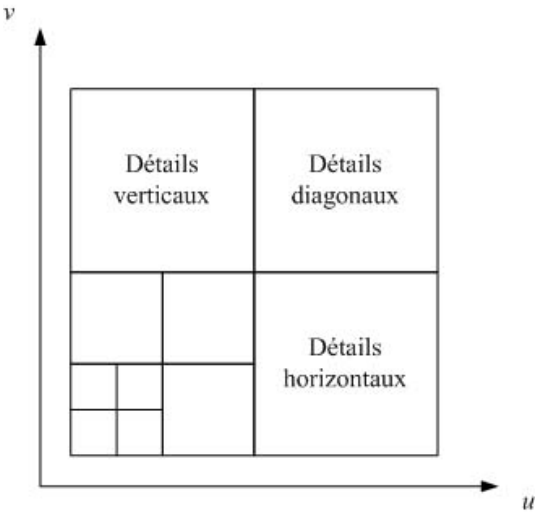


Figure 94 Plan de fréquence *(u,v)* après une transformée en ondelettes discrète 2D sur trois niveaux.

7.2 Application à la reconnaissance par l'iris

Pour pouvoir reconnaître un iris, il faut en extraire une signature et la comparer à une base de données de signatures de référence, en utilisant une mesure de distance telle que la distance de Hamming. Pour cela, l'iris est d'abord localisé et isolé du reste de l'œil puis la texture est transformée des coordonnées cartésiennes en coordonnées polaires afin d'être déroulée facilement. On obtient alors une image carrée ou rectangulaire qui ne contient que l'iris. La signature est extraite à partir des propriétés particulières de chaque iris et en fonction des transformations appliquées à l'image.

7.2.1 La signature extraite d'une transformée en ondelettes

Une transformée en ondelettes sur un nombre n de niveaux est appliquée à l'image, où n dépend de la résolution de l'image. La moyenne et l'écart de chaque sous-image obtenue sont alors extraits pour former la signature. Dans [Zhu00], Zhu *et al.* utilisent cette méthode pour la reconnaissance de l'iris en choisissant l'ondelette Daubechies 4 et en effectuant cinq niveaux de transformation. Ils excluent la sous image contenant les détails les plus fins car celle-ci est fortement bruitée, et obtiennent par conséquent 26 caractéristiques. La signature de référence la plus proche est extraite en mesurant la distance euclidienne :

$$WED(k) = \sum_{i=1}^{N} \frac{\left(f_i - f_i^{(k)}\right)^2}{\left(\delta_i^{(k)}\right)^2}$$

Équation 68

avec f_i la i^{ème} caractéristique de l'iris à reconnaître, et $f_i^{(k)}$ et $\delta_i^{(k)}$ respectivement la valeur moyenne et la déviation standard de l'iris de référence. L'indice k pour lequel la distance euclidienne est minimisée identifie la personne k.

Cette méthode exploite les caractéristiques globales de l'iris et offre donc l'avantage de l'invariance en rotation : un iris tourné présente, une fois déroulé, des caractéristiques décalées qui ne modifient pas les statistiques globales. A l'inverse, les détails ne sont pas exploités, diminuant le nombre de facteurs discriminants pour distinguer différentes signatures. Erik Rydgren [Rydg04-a] a montré que de meilleurs résultats pouvaient être obtenus en considérant également les moyennes et hautes fréquences contenues dans une image.

7.2.2 La signature extraite des sous-images de paquets d'ondelettes

Lorsque l'on veut séparer plus précisément les parties moyenne et haute fréquence d'une image, il faut effectuer la transformée en ondelettes discrète sur une ou plusieurs des sous-images obtenues au premier niveau de transformation : c'est ce qui est appelé transformée par paquets d'ondelettes. Une telle transformée est souvent représentée sous la forme d'un arbre, comme montré dans la Figure 95. Dans cet exemple, toutes les sous-images obtenues au premier niveau ont été transformées dans le deuxième. Les extrémités de l'arbre sont notées de gauche à droite.

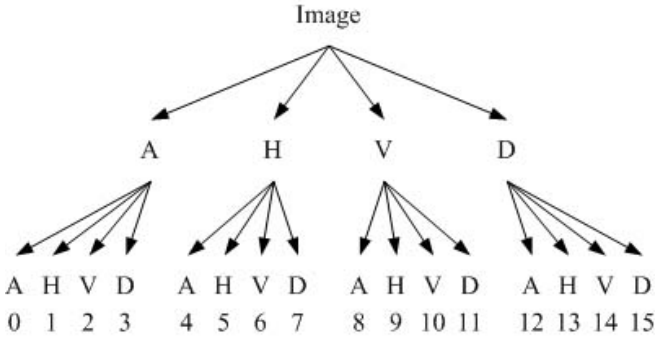


Figure 95 Arbre de paquets d'ondelettes après une décomposition sur deux niveaux.

La transformée par paquets d'ondelettes permet de diviser plus finement tout ou partie du plan de fréquence de l'image. Les coefficients d'ondelette obtenus sont rangés comme indiqué dans la Figure 96.

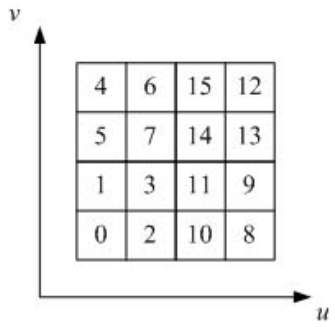


Figure 96 Plan de fréquence *(u,v)* après une décomposition par paquets d'ondelettes sur deux niveaux.

Dans [Rydg04-b], cette transformée en ondelettes donne les meilleurs résultats, concernant la discrimination des signatures, associée au filtre biorthogonal 1,3. Le filtre biorthogonal 1,3 possède 6 coefficients pour les filtres passe-bas et passe-haut :

- LP = {-0,0884 ; 0,0884 ; 0,7071 ; 0,7071 ; 0,0884 ; -0,0884}
- HP = {0 ; 0 ; -0,7071 ; 0,7071 ; 0 ; 0}

Dans la suite, nous allons voir comment implanter la transformée par paquets d'ondelettes.

7.3 Implémentation matérielle

Nous avons choisi un codage en virgule fixe pour simplifier les calculs. Pour un nombre de passes limité à 5 ou 6 (nombre de transformations appliquées à l'ondelette), ce qui est suffisant pour notre application, la précision des calculs nécessite de coder ces derniers sur 16 bits. Comme vu précédemment, une transformée en ondelettes peut être implémentée sous la forme d'un banc de filtre. L'opération de base consiste à effectuer le produit de convolution entre les coefficients du filtre et les pixels de l'image. Pour cela, nous choisissons une structure MAC (multiplication-accumulation), illustrée Figure 97 car les performances demandées au système ne sont pas très élevées. En effet, il ne s'agit pas de traiter un flux vidéo mais une seule image en un temps acceptable pour un être humain, c'est-à-dire quelques dizaines voire quelques centaines de millisecondes. Il n'est donc pas nécessaire d'effectuer les multiplications en parallèle.

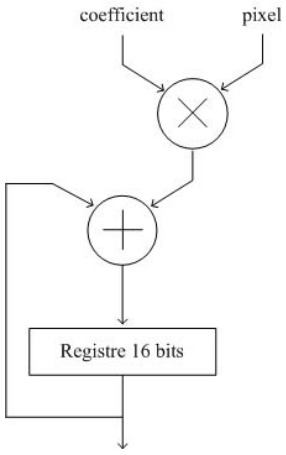


Figure 97 Structure MAC (multiplication-accumulation).

Nous allons, tout d'abord, détailler l'architecture d'un multiplieur et préciser celle qui sera retenue. Puis, nous introduirons différents types d'additionneurs et nous effectuerons leur comparaison après synthèse pour déterminer lequel possède le meilleur produit puissance délai. Nous verrons enfin les différents éléments constitutifs du bloc de traitement d'image.

7.3.1 Le multiplieur

Au niveau le plus bas, la multiplication d'un multiplicande X par un multiplicateur Y peut être vue comme une suite de décalages et d'additions. Pour une multiplication non signée, cela consiste à additionner *n* copies de X décalées, *n* étant la taille du multiplicateur Y. Une multiplication est divisée en trois parties : la génération des produits partiels, leur réduction et enfin le calcul de la somme finale. Une telle procédure est représentée Figure 98 pour une multiplication 6x6.

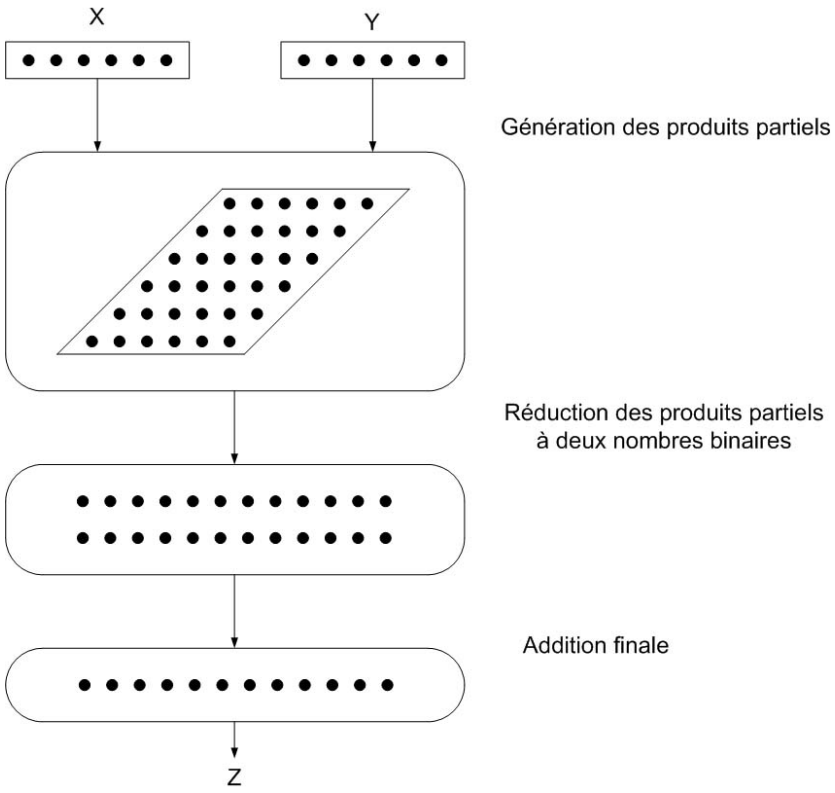


Figure 98 Etapes d'une multiplication.

Génération des produits partiels

La première étape d'une multiplication est la génération des *n* copies décalées du multiplicande X. La valeur du multiplicateur à la position de poids *i* définit si la copie est additionnée ou non : si Y_i vaut zéro, le multiplicande décalé *i* fois n'est pas additionné, dans le cas contraire, il l'est. Cela requiert donc n-1 additions : il est à noter qu'une structure de type « carry save » est utilisée, c'est-à-dire que les retenues d'une ligne sont propagées sur la ligne suivante.

La technique pour générer les produits partiels dépend de la représentation des nombres : dans le cas présent, une simple porte AND est suffisante. Néanmoins, une multiplication à base plus élevée permet de réduire le nombre de produits partiels. Dans le cas de nombres signés, la représentation du multiplicateur par le code de Booth à base 4 est couramment utilisée. Celle-ci est basée sur le fait que lorsqu'un bit du multiplicateur vaut zéro, une ligne entière d'additionneurs est inutile, les bits en entrée étant directement propagés en sortie. Bien évidemment, cette ligne ne peut pas être supprimée car il n'est pas possible de connaître à l'avance quel bit sera à zéro. Le code de Booth permet de tirer profit de la présence de zéros pour diviser par deux le nombre de produits partiels. Pour encoder le multiplicateur, les valeurs de trois bits consécutifs sont considérées : le résultat de l'encodage est donné dans le Tableau 10. La multiplication par deux est générée par un décalage d'un bit vers la gauche et le changement de signe est réalisé en prenant le complément à deux du multiplicande X. La génération des produits partiels est donc obtenue à l'aide d'un multiplexeur qui permet de sélectionner au choix X_i, $\overline{X_i}$, X_{i+1}, $\overline{X_{i+1}}$ et '0', l'indice *i* étant compris entre 0 et n-1.

Tableau 10 Codage de Booth à base 4.

Y_{i+1}	Y_i	Y_{i-1}	B_i
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	2
1	0	0	-2
1	0	1	-1
1	1	0	-1
1	1	1	0

L'avantage du code de Booth est une nette diminution du nombre de produits partiels à additionner et donc du délai et de la puissance dissipée par le multiplieur. Globalement, chaque niveau d'encodage de Booth divise par deux le nombre de produits partiels. Néanmoins, la présence de la logique nécessaire pour encoder le multiplicateur et décoder le multiplicande augmente la consommation. En pratique, le codage de Booth est limité à un voire deux niveaux, mais rarement plus. Le schéma du codage de Booth est montré Figure 99.

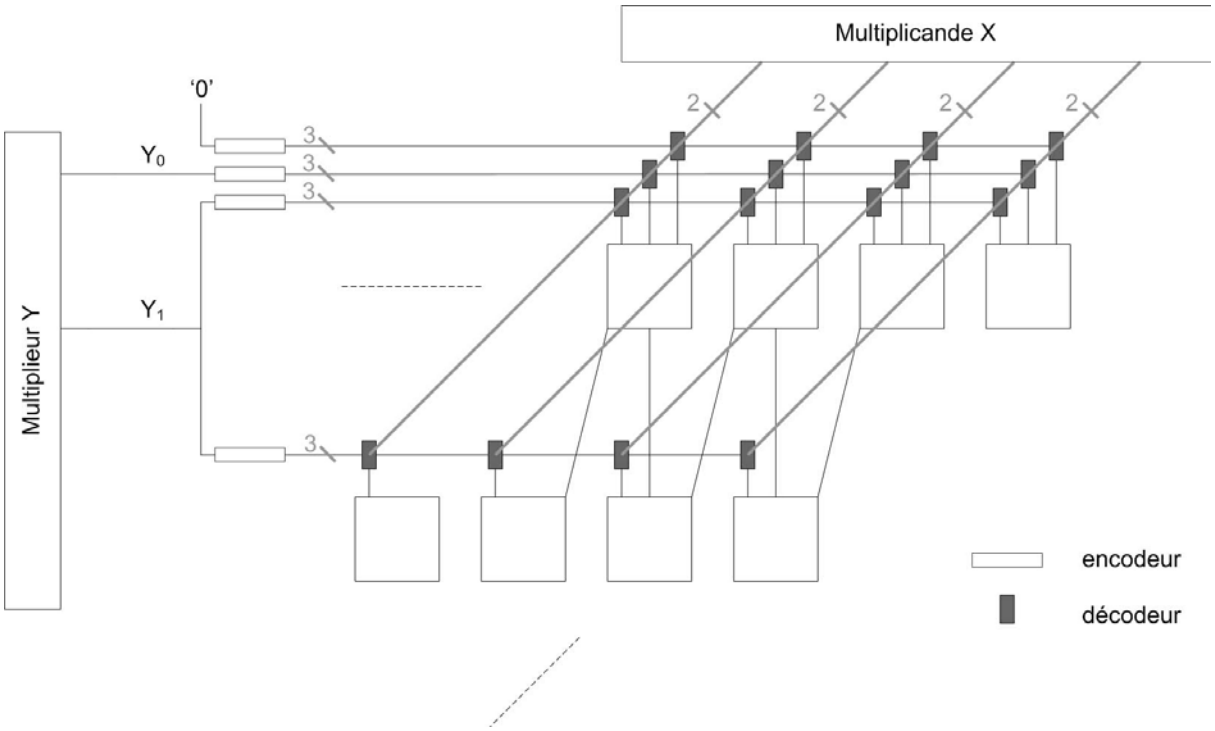


Figure 99 Schéma montrant les encodeurs et les décodeurs de Booth utilisés pour la réduction du nombre de produits partiels.

Puisque nous utilisons des nombres signés, nous choisissons d'utiliser le code de Booth pour diminuer le nombre de produits partiels et ainsi réduire l'énergie dissipée par le multiplieur.

Réduction des produits partiels

La réduction des produits partiels s'effectue en additionnant les bits d'une même colonne. Comme indiqué précédemment, une structure de type « carry save » est utilisée, c'est-à-dire que les retenues des additions sont propagées aux lignes suivantes. Cela permet de rendre indépendantes les additions entre elles, sur une même ligne, et d'éviter que le délai ne soit de la forme $O(n^2)$. Grâce à une addition « carry save », nous pouvons réduire trois lignes de bits en deux lignes (somme et retenue) dans le délai d'un additionneur 1 bit. Un nombre n de lignes peut ainsi être réduit à deux.

La structure la plus simple consiste à additionner ligne à ligne les produits partiels. Un exemple portant sur une multiplication 6x6 est donnée Figure 100. Pour commencer, nous additionnons trois lignes de bits, en utilisant des additionneurs et des demi additionneurs comme montré Figure 100-b. Les trois lignes de produits partiels sont réduites à deux lignes de bits. La somme et la retenue obtenues sont alors additionnées à la ligne de bits suivante.

Après *n*-2 additions « carry save », les produits partiels sont réduits à deux lignes de bits. L'avantage de cette structure est qu'elle est très régulière, avec des fils de faible longueur : les mêmes additionneurs sont dupliqués suivant la largeur et la hauteur du multiplieur. Le délai de cette structure est de la forme *O(n)*. Cette relative lenteur est due au fait que les additions sont effectuées en série.

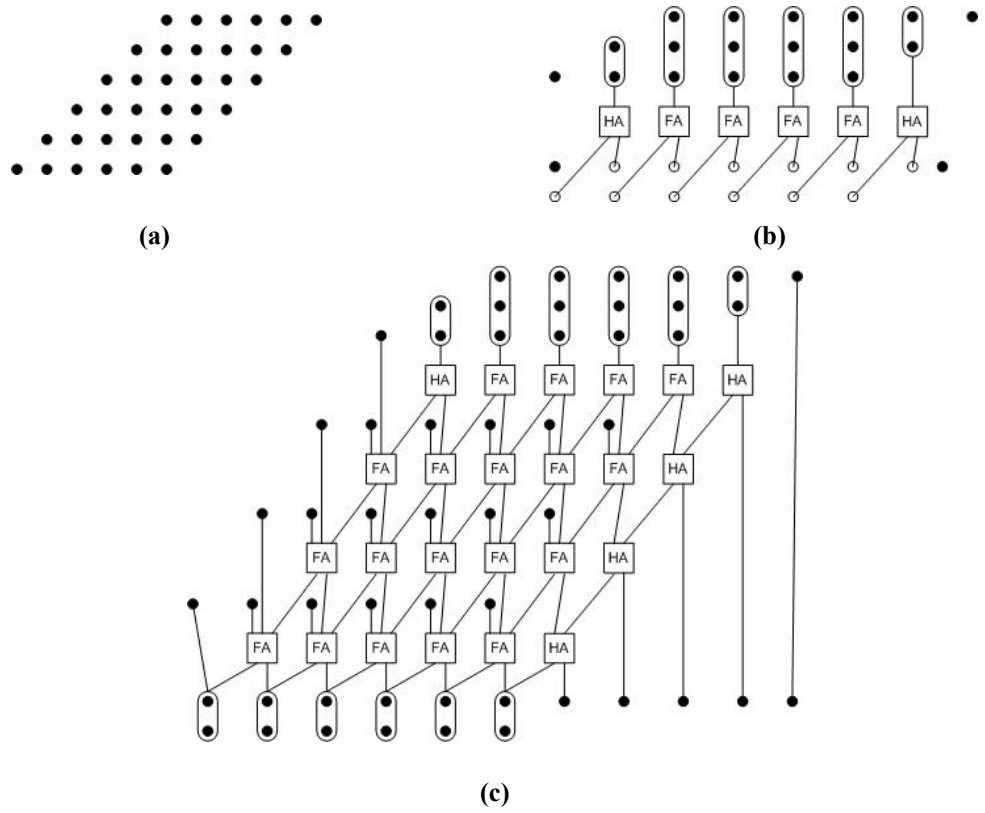


Figure 100 Réduction des produits partiels ligne à ligne : (a) la matrice initiale de produits partiels, (b) la réduction de trois lignes de bits par une rangée d'additionneurs « carry save », (c) la structure d'additionneurs permettant de réduire les six lignes de produits partiels à deux lignes de bits.

Il est possible de paralléliser les opérations en utilisant un arbre de Wallace [Wall64]. Deux additions « carry save » sont effectuées simultanément, ce qui réduit beaucoup plus rapidement le nombre de produits partiels. La Figure 101 montre un exemple d'arbre de Wallace pour une multiplication 6x6 réalisé en suivant l'algorithme de Dadda [Dadd65], qui minimise le nombre d'opérateurs de réduction. Dans cet exemple, seul le premier objectif de réduction nécessite l'utilisation d'additionneurs « carry save » en parallèle.

L'avantage des arbres de Wallace est que le délai est proportionnel à $O(\log_{3/2} n)$. Les inconvénients, par rapport à la structure séquentielle précédente, sont :

- leur irrégularité qui introduit des fils plus longs,

- l'addition finale de taille $2n - log_{3/2}n$ contre n précédemment.

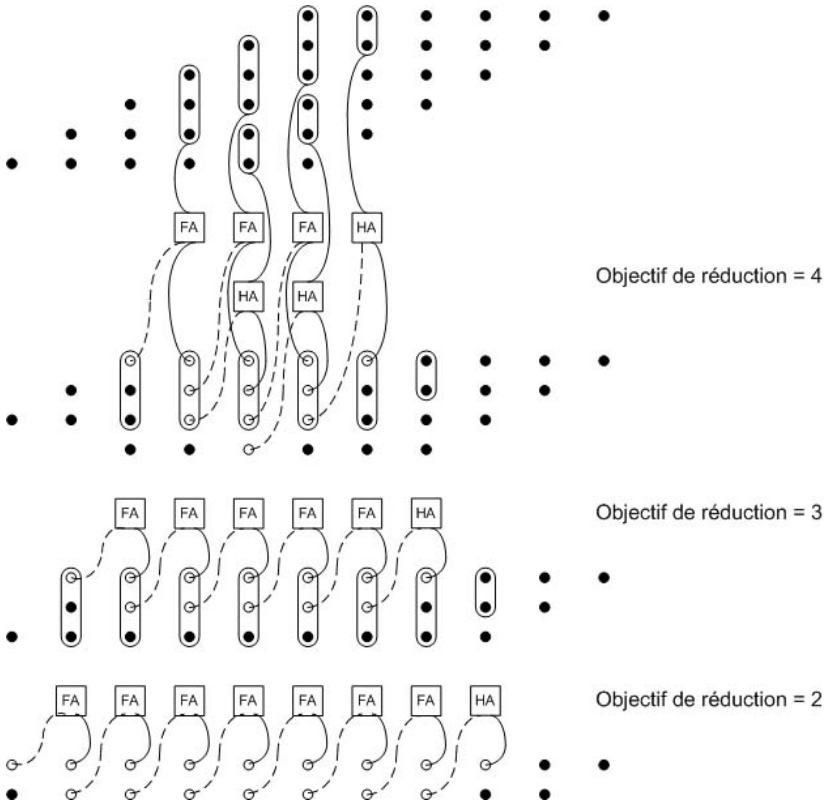


Figure 101 Arbre de Wallace généré suivant l'algorithme de Dadda, garantissant le minimum d'opérateurs de réduction.

Néanmoins, nous choisissons ce type de réduction de produits partiels pour l'avantage qu'il présente en délai sans nuire à la consommation : la diminution du nombre d'opérateurs, et donc de l'activité, compense en effet l'augmentation des capacités d'interconnexion. Meier *et al.* [Meie96] suggèrent un avantage en énergie de 10% pour l'arbre de Wallace par rapport à la structure de réduction ligne à ligne. Un arbre de Wallace est très souvent utilisé en combinaison avec le codage de Booth pour réaliser des multiplieurs haute performance et basse consommation [Laks02] [Liao02] [Frie97].

Addition finale

La méthode couramment utilisée pour diminuer le délai d'une multiplication consiste à accélérer l'addition finale. Comme nous allons le voir dans la suite, de nombreux compromis existent entre vitesse et consommation selon le type d'additionneur considéré.

7.3.2 L'additionneur

Le choix de l'additionneur est important, non seulement pour respecter des contraintes de délai, mais également pour participer à la réduction de la puissance dissipée. Il existe de nombreux types d'additionneurs, de l'additionneur *ripple carry* dont le délai est linéaire avec le nombre de bits *n*, aux additionneurs parallèles permettant de s'affranchir en partie de la cascade de la retenue, avec des délais en $\sqrt{n}$ ou en $\log(n)$. Nous allons nous intéresser dans la suite à quatre additionneurs.

Additionneur *Ripple Carry*

Cet additionneur est le plus simple que l'on puisse imaginer. Les additionneurs 1 bit sont placés en série, la retenue sortante de l'un étant reliée à la retenue entrante du suivant. Le délai de pire cas survient lorsque la retenue, générée au bit de poids faible, se propage jusqu'au bit de poids fort : le délai est proportionnel au nombre de bits de l'additionneur. Puisque cette structure est la plus simple, c'est aussi une des plus économe en énergie : la valeur de la capacité totale est limitée par le faible nombre de portes ; cependant, le nombre de fausses transitions peut être élevé à cause de la propagation de la retenue.

Même si cela a peu d'intérêt de l'utiliser pour effectuer l'addition finale, l'additionneur *ripple carry* se retrouve dans des sous blocs d'additionneurs plus rapides, par exemple *carry bypass* ou *carry select*. Par conséquent, les paramètres de vitesse et de consommation de l'additionneur *ripple carry* influencent la performance globale de nombre d'additionneurs. La structure de cet additionneur est montrée Figure 102.

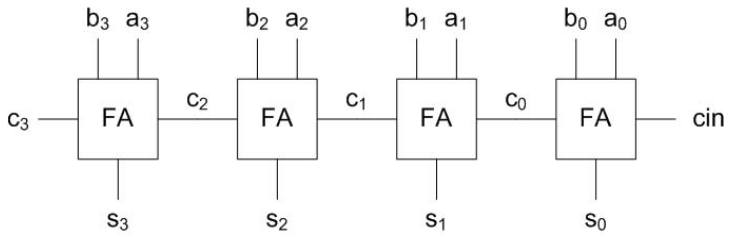


Figure 102 Additionneur *Ripple Carry* 4 bit

Additionneur *Carry Select*

L'additionneur *carry select* est un additionneur parallèle divisé en sous blocs. Pour chaque sous bloc, deux additionneurs *ripple carry* sont implémentés, un pour effectuer le calcul en considérant que la retenue en entrée vaut '0', l'autre en considérant qu'elle vaut '1'. Le résultat est alors choisi à l'aide d'un multiplexeur, lorsque la valeur de la retenue en entrée est

connue. Le schéma de principe est montré Figure 103a. L'optimisation de l'additionneur *carry select* par rapport au *ripple carry* réside dans le fait que la retenue est propagée de sous bloc en sous bloc et non plus d'additionneur à additionneur. Le délai s'en trouve réduit, au détriment de la dissipation de puissance, qui augmente à cause d'un nombre de portes et d'un taux d'activité en hausse.

Un moyen d'améliorer ce schéma d'additionneur est de remarquer que le temps introduit par le multiplexeur de l'étage précédent peut être utilisé pour ajouter un additionneur supplémentaire dans l'étage suivant. Ainsi, la taille des sous blocs n'est plus constante mais augmente de 1 à chaque étage – voir Figure 103c. Le délai de cet additionneur n'augmente plus linéairement avec le nombre de bits mais augmente en racine carrée.

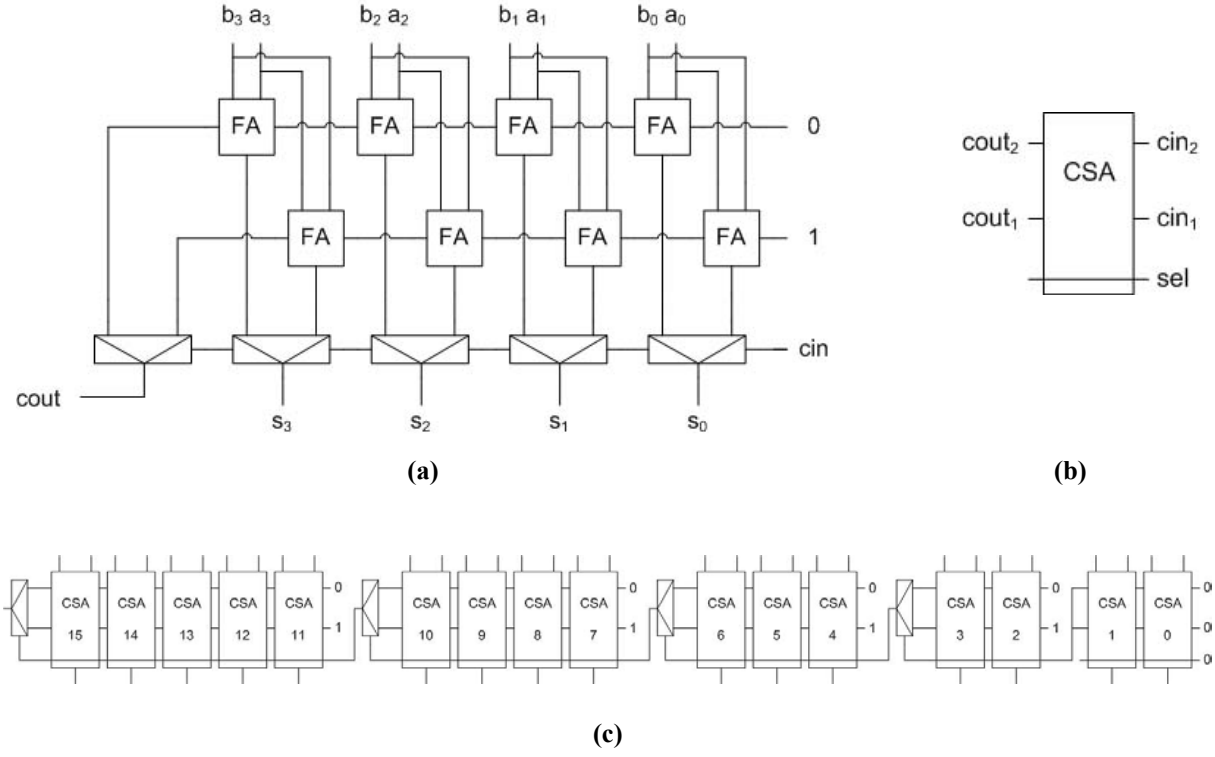


Figure 103 Additionneur *Carry Select* : (a) Structure d'un bloc de 4 bits, (b) Symbole d'un bloc 1 bit, (c) Amélioration de l'additionneur pour avoir un délai variant en racine carrée en fonction du nombre de bits.

Additionneur de Kogge et Stone

Pour obtenir un additionneur plus rapide que le précédent, il faut s'affranchir un peu plus de la propagation de la retenue et en confier le calcul à des cellules spécialisées : c'est le cas

des additionneurs logarithmiques. Pour cela, on introduit les signaux intermédiaires G (Générer) et P (Propager). Leur expression est la suivante :

$$G_i = A_i \cdot B_i,\ P_i = A_i \oplus B_i$$

Équation 69

Lorsque G=1, une retenue est générée quelque soit la retenue entrante. Lorsque P=1, la retenue entrante est propagée à la sortie. La somme et la retenue sortante de l'étage *i* peuvent alors être calculées de la manière suivante :

$$Cout_i = G_i + P_i \cdot Cin_i$$

$$S_i = P_i \oplus Cin_i$$

Équation 70

Le schéma de l'additionneur de Kogge et Stone [Kogg73] est donné Figure 104. Cinq types de cellules sont nécessaires ; elles sont représentées par des carrés blanc et noir, des cercles blanc et noir et un triangle blanc :

- le carré noir réalise les fonctions logiques définies dans l'Équation 69,
- le cercle noir réalise les fonctions logiques suivantes :

$$G_{j+1,i} = \overline{\left(G_{j,i} + P_{j,i}\right) \cdot G_{j,i-k}},\ P_{j+1,i} = \overline{P_{j,i} \cdot P_{j,i-k}}$$

- le cercle blanc :

$$G_{j+1,i} = \overline{\left(G_{j,i} \cdot P_{j,i}\right) + G_{j,i-k}},\ P_{j+1,i} = \overline{P_{j,i} + P_{j,i-k}}$$

- le carré blanc :

$$S_i = P_{0,i} \,\overline{\oplus}\, G_{\max,i-1}$$

- enfin, le triangle blanc représente un inverseur.

Les cellules représentées par les cercles noir et blanc sont appelées cellules de Brent et Kung (BK). Cet additionneur possède les avantages de présenter une sortance constante de deux à chaque étage et une profondeur logique réduite mais occupe une surface importante et utilise beaucoup de cellules BK.

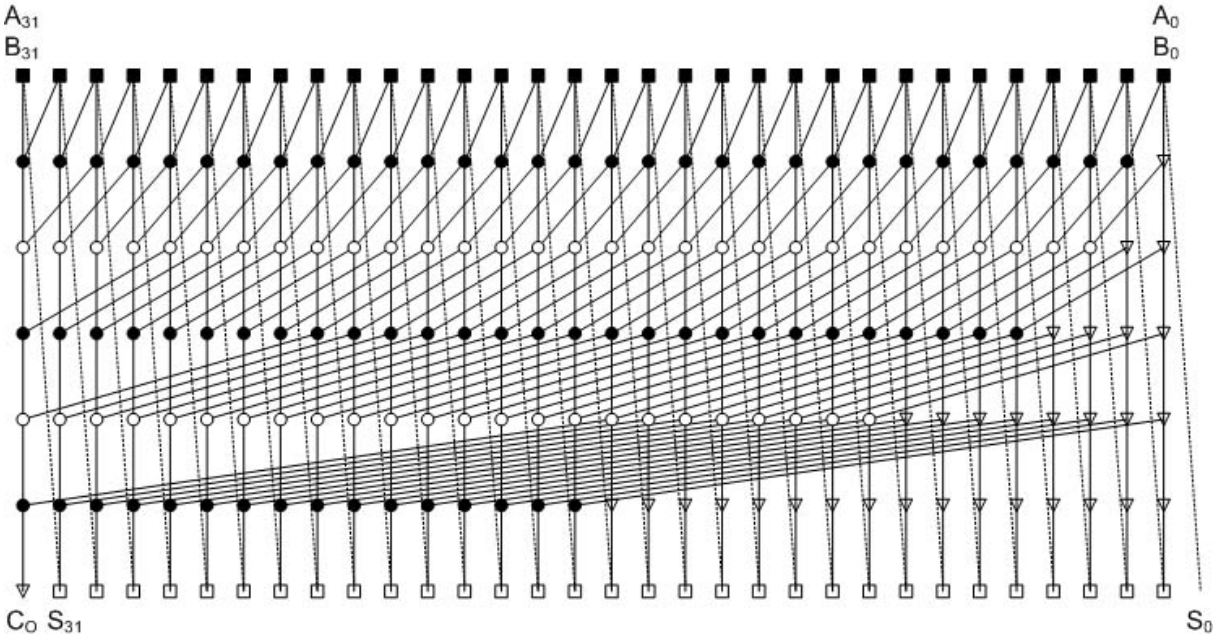


Figure 104 Additionneur de Kogge et Stone.

Additionneur de Han et Carlson

L'additionneur de Han et Carlson [Han87] représente un meilleur compromis délai-surface par rapport à l'additionneur précédent : un niveau logique est ajouté en échange d'un gain important au niveau de la surface occupée. Comme on peut le voir Figure 105, il n'y a qu'une seule cellule BK par paire de bits quelque soit la ligne considérée. Cet additionneur peut être condensé en largeur d'un facteur 2, ce qui réduit la capacité des fils horizontaux. Comme précédemment, les cercles noirs représentent une cellule BK dont la fonction logique est :

$$G_{j+1,i} = G_{j,i} + P_{j,i} \cdot G_{j,i-k},\ P_{j+1,i} = P_{j,i} \cdot P_{j,i-k}$$

Équation 71

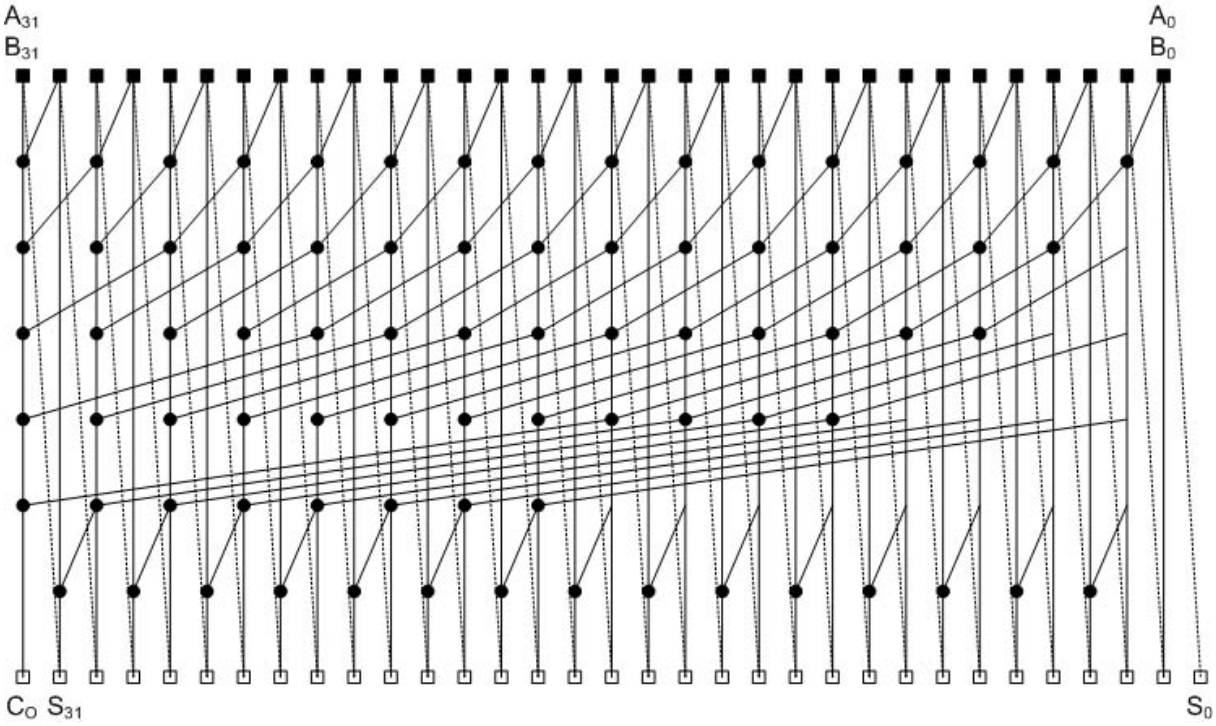


Figure 105 Additionneur de Han et Carlson.

Comparaison

Les quatre additionneurs précédents ont été comparés synthétisés à l'aide de la bibliothèque de cellules standards présentée dans le paragraphe 5.3 et comparés en délai et en puissance. Le flot utilisé est présenté dans la Figure 106 ; l'outil Design Compiler de Synopsys est utilisé pour synthétiser les circuits en les contraignant en délai et en surface. Le fichier Verilog obtenu est simulé à l'aide de VCS pour générer le fichier d'activité puis est introduit dans Prime Power pour estimer la consommation du circuit à l'aide des tables de puissance dynamique et de puissance statique de la bibliothèque.

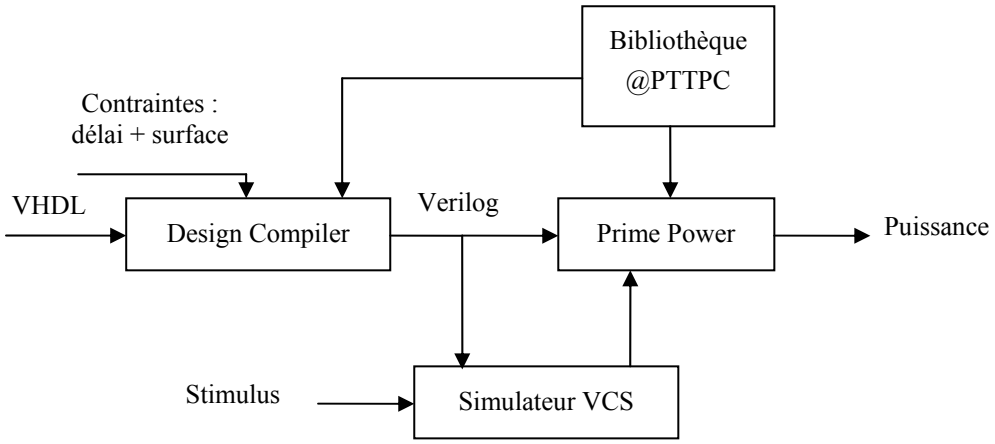


Figure 106 Flot de synthèse pour un point de fonctionnement PTTPC donné (Procédé, Tension, Température, Pente, Capacité).

Les résultats obtenus sur les additionneurs sont présentés Figure 107.

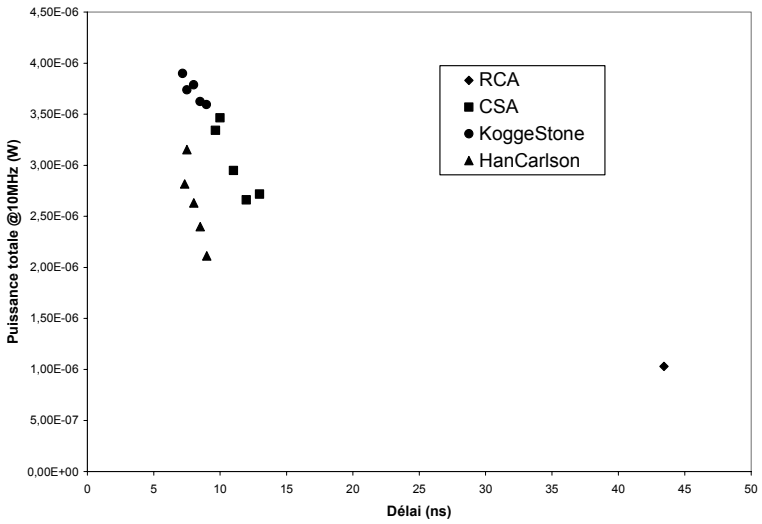


Figure 107 Comparaison du délai et de la puissance dissipée par les additionneurs 32 bits.

L'additionneur de Kogge et Stone est le plus rapide car c'est celui qui présente le moins de profondeur logique mais c'est également le plus consommateur d'énergie puisqu'il contient un très grand nombre de portes. L'additionneur de Han et Carlson présente des performances à peine inférieures tout en consommant 30% de puissance en moins. L'additionneur *Carry Select* en temps racine carrée est 30% plus lent que les additionneurs logarithmiques tout en ayant une puissance dissipée comprise entre celles de ces deux additionneurs.

Pour les trois additionneurs précédents, plusieurs points sont présentés, qui illustrent chacun une contrainte en délai différente. Comme on peut le remarquer dans la Figure 107, l'additionneur *Ripple Carry* n'est caractérisé que pour un seul point. La raison est la suivante : si nous laissons l'outil de synthèse optimiser cet additionneur, la liste d'interconnexions obtenue ne ressemble plus du tout à celle d'un additionneur *Ripple Carry* et le délai approche celui de l'additionneur *Carry Select*. Il faut donc mettre des attributs 'set_dont_touch' sur les cellules demi-additionneur et additionneur 1 bit pour empêcher leur optimisation et la perte de la structure particulière du *Ripple Carry*. Ainsi, nous n'obtenons qu'un seul couple délai-surface.

Au final, l'additionneur de Han et Carlson est celui qui présente le meilleur produit puissance délai. C'est donc cet additionneur qui est utilisé dans la suite.

7.3.3 Architecture du bloc de transformée par paquets d'ondelettes

L'architecture générale du bloc de transformée par paquets d'ondelettes est présentée dans la Figure 108. La taille des images à transformer est de 512x512 pixels, codés sur 16 bits. Le bloc est composé de sept éléments différents : un bus de données est utilisé pour en relier plusieurs entre eux.

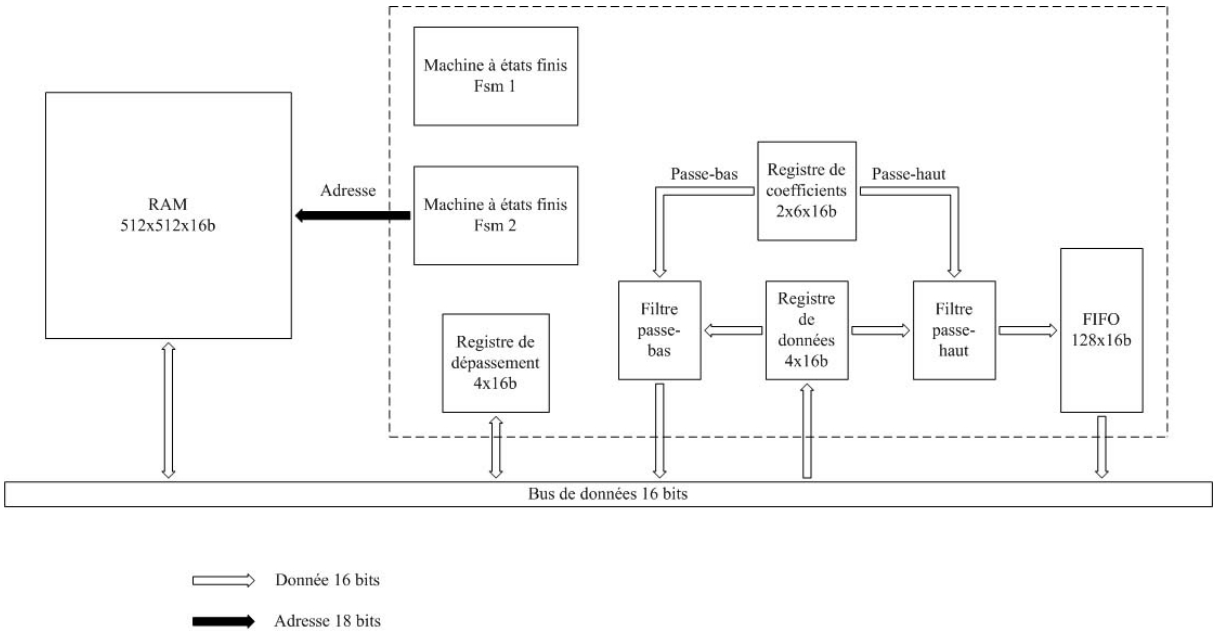


Figure 108 Architecture générale du circuit de transformée par paquets d'ondelettes : les différents signaux de contrôle ont été omis pour éviter de surcharger la figure.

Les filtres

Il s'agit des structures MAC définies dans la Figure 97. Le multiplieur utilise un codage de Booth et un arbre de Wallace ; son additionneur final et l'additionneur 16 bits sont de type Han-Carlson. La sortie du multiplieur, sur 31 bits, est tronquée à 16 bits : les bits de position 15 à 30 sont utilisés pour l'addition. Le registre est mis à zéro avant un nouveau calcul. Après six multiplications et cinq additions, la valeur est disponible en sortie.

La sortie du filtre passe-bas est connectée sur le bus de données : en effet, la valeur obtenue peut être directement stockée dans la mémoire RAM puisque, à l'endroit où celle-ci va être écrite, le pixel a déjà été lu. Le résultat issu du filtre passe-haut doit, lui, être écrit dans la deuxième moitié de la ligne (ou de la colonne) comme indiqué dans la Figure 94. Par conséquent, lorsque les filtres sont appliqués au début d'une ligne (ou d'une colonne), le produit de convolution issu du filtre passe-haut ne peut pas être écrit immédiatement dans la mémoire RAM, car il écraserait un pixel non lu : il est donc stocké temporairement dans une mémoire FIFO (*First In First Out*).

Registre de données

Pour pouvoir sous-échantillonner le signal d'entrée par deux, les filtres sont décalés de deux pixels à chaque nouveau calcul. Ainsi, quatre pixels sont communs à deux produits de convolution consécutifs : plutôt que de relire ces quatre pixels dans la mémoire RAM, ce qui prendrait beaucoup de temps, ceux-ci sont stockés dans un registre de données. Au début du calcul sur une ligne ou sur une colonne, six pixels doivent être lus depuis la mémoire, puis seulement deux pour le calcul suivant.

Registre de coefficients

Comme son nom l'indique, il s'agit du registre dans lequel sont stockés les coefficients des filtres. Il y a donc en fait deux registres de six mots de 16 bits. Les sorties des deux registres sont reliées aux filtres passe-haut et passe-bas pour fournir les coefficients correspondant au pixel lu. Nous n'avons pas utilisé de mémoire ROM pour pouvoir modifier les valeurs des coefficients.

Mémoire FIFO

Cette mémoire sert à stocker temporairement les résultats des calculs issus du filtre passe-haut, le temps que les emplacements où ils doivent être écrits en mémoire se libèrent. Sa taille est limitée à 128 mots de 16 bits pour une image de 512x512 pixels, car les premiers produits

de convolution peuvent être écrits lorsque les extrémités des filtres arrivent à la moitié de l'image : à la suite du sous-échantillonnage par deux, seuls 128 coefficients haute fréquence ont été calculés. Les données de la FIFO sont écrites en mémoire aussi vite que des données y sont lues, c'est-à-dire à raison de deux écritures pour deux lectures de pixels.

Registre de dépassement

Pour pouvoir effectuer tous les produits de convolution sur une ligne ou sur une colonne, il manque (taille du filtre – 2) pixels, soit dans notre cas, quatre valeurs. Lorsqu'ils arrivent en bout de ligne (ou de colonne), les filtres dépassent. Pour pouvoir gérer cet effet de bord, la technique du repliement de l'image est utilisée : les pixels du côté gauche (ou en bas) de l'image sont utilisés pour remplacer ceux manquant du côté droit (ou en haut). Bien évidemment, les pixels que nous voulons utiliser ont depuis été effacés par les coefficients issus du filtre passe-bas. C'est la raison pour laquelle ils sont préalablement sauvés dans le registre de dépassement : lors du premier produit de convolution, lorsque six pixels sont lus depuis la mémoire, les quatre premiers sont écrits dans ce registre.

Machines à états finis

Pour simplifier la partie contrôle, nous avons divisé le problème en deux et donc écrit deux machines à états finis. La première gère les paramètres suivants :

- direction de la transformée (ligne ou colonne),
- taille de l'image à traiter (et donc longueur des lignes ou des colonnes, en fonction du niveau de la transformée en cours),
- établissement des valeurs des décalages en x et en y pour l'adressage mémoire, en fonction de la sous-image qui doit être transformée (les décalages valent zéro pour le premier niveau de transformation).

Tous ces paramètres sont transmis à la deuxième machine à états finis. Son rôle est de contrôler la transformée en ondelettes en fonction de la direction, de la taille de l'image et des décalages spécifiés : elle gère l'adressage mémoire et génère tous les signaux de contrôle nécessaires pour commander les cinq blocs que nous avons détaillés précédemment. Les différents signaux de contrôle sont :

- remise à zéro des registres des filtres, du registre de données et du registre de dépassement,
- lecture ou écriture dans la mémoire RAM,

- lecture ou écriture dans la mémoire FIFO,
- lecture ou écriture du registre de dépassement,
- écriture dans le registre de données,
- lecture des valeurs correspondantes depuis le registre de données et le registre de coefficients.

Lorsque les calculs sur un bloc de lignes ou de colonnes sont terminés, la deuxième machine à états finis met un drapeau à '1' pour le signaler à la première, et ainsi autoriser une nouvelle transformation.

Un diagramme simplifié du fonctionnement de la deuxième machine à états finis est donné dans la Figure 109 : cela représente les différents états décrits lors de la transformée en ondelettes d'une seule ligne (ou d'une seule colonne). Un indice i a été introduit pour servir de condition de passage entre les états : celui-ci indique la position du dernier coefficient des filtres. Il est à noter que les multiplications et additions, non représentées, s'effectuent en parallèle de la lecture des pixels, que ce soit depuis la mémoire RAM, depuis le registre de données ou depuis le registre de dépassement.

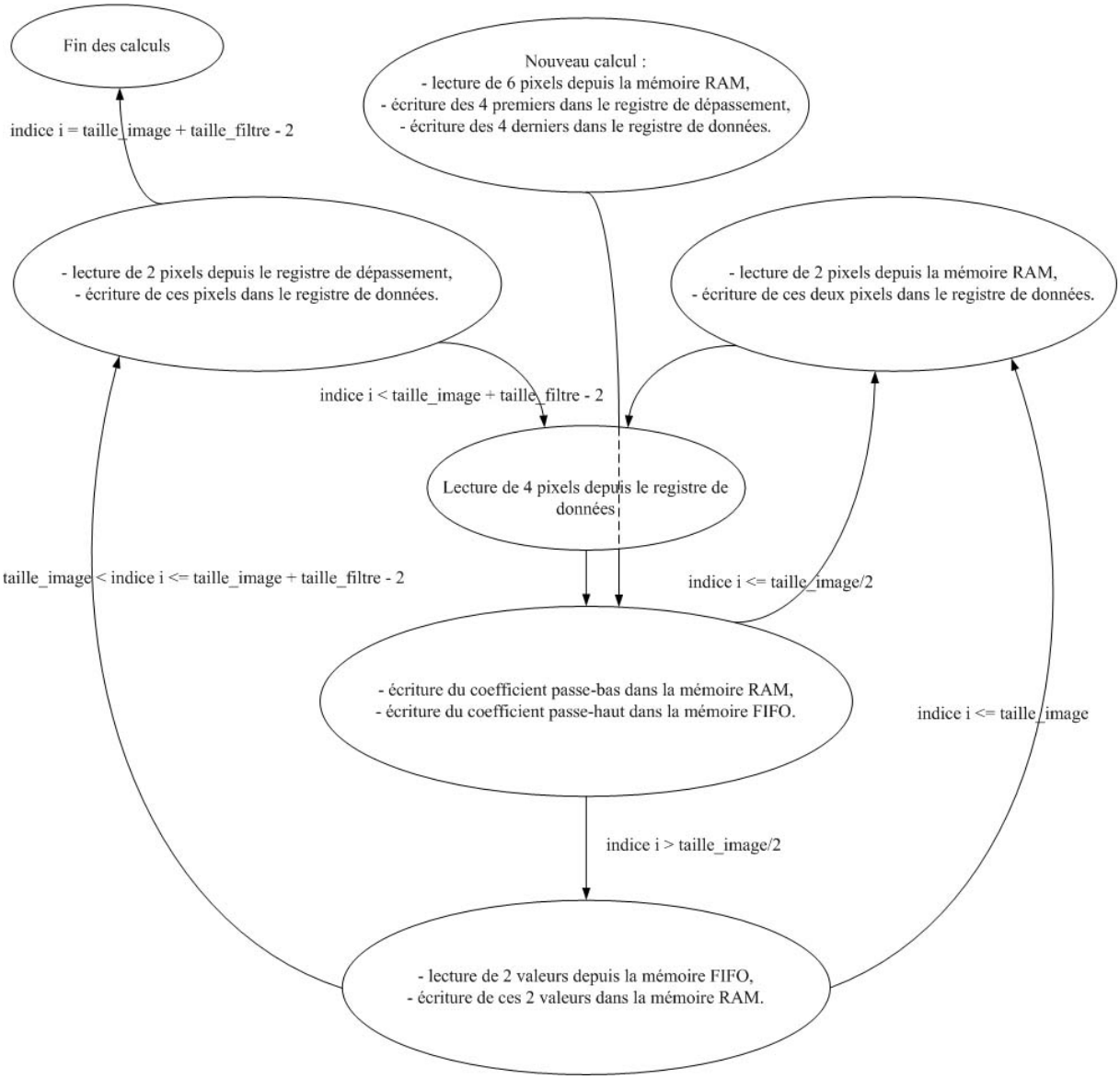
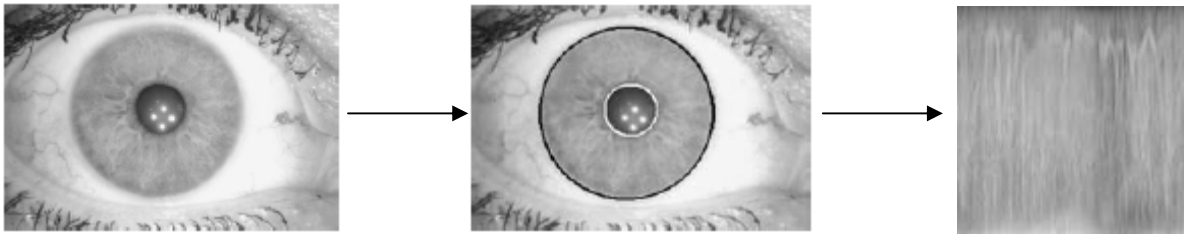


Figure 109 Les différents états de la deuxième machine à états finis.

7.3.4 Résultats de synthèse

Le bloc de traitement d'image par ondelette a été synthétisé dans trois technologies différentes : SOI *High-Speed* (HS), BULK *High-Speed* et BULK *Low-Leakage* (LL), en typique, à la tension d'alimentation de 0,5V et à la température de 25°C. Nous faisons le choix de considérer des transistors en typique plutôt qu'en pire cas, car c'est généralement dans le cas nominal que des technologies sont comparées. Les bascules qui ont été développées précédemment pour la bibliothèque basse tension en SOI sont réutilisées pour les bibliothèques BULK et ont donc été caractérisées à l'aide de modèles Eldo BULK. Le flot présenté dans la Figure 106 a été utilisé pour évaluer ces trois technologies. Pour mesurer la

puissance dissipée, une image d'iris déroulé a été considérée : celle-ci est de taille 128x128 pixels et est donnée dans la Figure 110-c. Il faut noter que compte tenu de la taille du fichier d'activité généré par le simulateur, la mesure de puissance n'a pas pu être effectuée sur toute l'image mais seulement sur une partie : pour une transformée en ondelettes sur l'image entière, le fichier d'activité fait 4Go et l'outil PrimePower de Synopsys refuse de l'ouvrir. Néanmoins, le temps de simulation est suffisamment long – supérieur à 10000 périodes d'horloge – pour que les résultats de puissance obtenus soient réalistes.



(a) **(b)** **(c)**

Figure 110 (a) Image d'un œil avant traitement, (b) localisation de l'iris, (c) image d'iris déroulé de taille 128x128 utilisée pour la comparaison de la puissance dissipée.

Les résultats de synthèse sont donnés dans le Tableau 11. La technologie SOI HS se situe, au niveau des performances, entre les technologies BULK HS et BULK LL, avec une fréquence de fonctionnement maximum du bloc de transformée en ondelettes de 40MHz contre respectivement 100MHz et 21MHz. Une telle différence de fréquence entre SOI HS et BULK HS peut étonner au premier abord mais elle s'explique très simplement par les valeurs différentes des tensions de seuil de ces deux technologies. En effet, celles-ci ont été optimisées pour la tension d'alimentation nominale de 1,2V. Le critère d'optimisation choisi par STMicroelectronics est celui des courants de fuite : ainsi, en SOI, compte tenu des effets de body importants à cette tension d'alimentation – qui en moyenne abaissent la tension de seuil V_T et donc augmentent les courants de fuite –, la tension de seuil doit être augmentée au niveau technologique, par des niveaux de dopage plus importants. Lorsque nous travaillons à très basse tension, les effets de body sont considérablement diminués et la tension de seuil du SOI HS devient alors supérieure à celle du BULK HS. Le SOI HS est donc plus lent : il faudrait pouvoir optimiser cette technologie pour la très basse tension afin d'effectuer une comparaison équitable au niveau des performances entre ces technologies.

Tableau 11 Comparaison des technologies SOI HS, BULK HS et BULK LL en typique, pour une tension d'alimentation de 0.5V et une température de 25°C.

	Fréquence (MHz)	Ptot @ f_{max} (mW)	Pdyn/MHz (µW)	Pfuites (µW) % Ptot	Surface
SOI HS	40	0,40	9,53	22 (5,45%)	151135
BULK HS	100	1,20	11,39	59,2 (4,94%)	151377
BULK LL	21	0,27	12,07	11,2 (4,23%)	199602

Pour pouvoir comparer les trois circuits obtenus par synthèse, nous prenons comme facteur de qualité la puissance dynamique par MHz, ce qui revient à effectuer le produit de la puissance dissipée et du délai. Nous pouvons alors constater que la technologie SOI est celle qui présente la plus faible énergie dissipée, inférieure de 16% au BULK HS et de 21% au BULK LL. En ce qui concerne les courants de fuite, ils sont directement proportionnels aux différentes tensions d'alimentation et sont donc les plus élevés pour le BULK HS et les plus faibles pour le BULK LL. Dans tous les cas, la puissance des courants de fuite, à la fréquence de fonctionnement maximale, est limitée à 5% de la puissance totale, ce qui confirme que l'introduction de techniques de limitation des courants de fuite n'est pas nécessaire en technologie 0,13µm.

Un point intéressant à noter, lorsque nous comparons les résultats obtenus, est celui de la surface occupée dans les trois technologies. En effet, celle-ci est identique en SOI HS et en BULK HS mais est plus de 30% supérieure en BULK LL. Cela s'explique par l'obligation que l'outil de synthèse Design Compiler a de respecter les règles de DRC (*Design Rule Check*) notamment max_transition. Les portes en BULK LL étant plus lentes, l'outil de synthèse doit insérer plus de buffers et d'inverseurs pour accélérer les fronts montant et descendant, augmentant ainsi la surface.

Une précision reste à apporter concernant l'avantage d'utiliser des transistors DTMOS. Comme indiqué dans la partie décrivant la bibliothèque de cellules standards, nous avons introduit des portes possédant des transistors DTMOS dans leurs inverseurs de sortie. Ces portes devraient donc être avantageuses dans le cas de grosses charges. Nous avons donc comparé deux synthèses en SOI HS, une en incluant ces portes, l'autre sans, en utilisant la commande *set_dont_use* de Synopsys. Les résultats obtenus, tant en vitesse qu'en surface ou en consommation, sont quasiment identiques : la version sans DTMOS possède un trop faible

avantage en vitesse sur la version avec DTMOS pour être jugé significatif. Il n'est donc pas possible d'estimer à ce niveau l'avantage apporté par l'utilisation de transistors DTMOS.

7.3.5 Conclusion

Dans cette partie, nous avons vu l'implémentation d'un circuit de transformée d'image par ondelettes. Compte tenu de son application, la reconnaissance de personnes par l'iris, et donc du faible niveau de performances demandé, nous avons choisi une structure de type MAC pour la réalisation des filtres, les calculs ne nécessitant pas d'être effectués en parallèle. Pour le multiplicateur, nous avons opté pour un codage de Booth puisque nous utilisons des nombres signés, et un arbre de Wallace pour la réduction des produits partiels. Après avoir comparé différents types d'additionneurs 16 bits, nous avons choisi l'additionneur de type Han-Carlson pour son meilleur produit puissance-délai et nous l'employons pour l'addition finale dans le multiplicateur et pour l'addition des produits $coefficient \times pixel$ obtenus.

Les différentes parties constitutives du circuit de traitement d'image sont détaillées : les particularités de la transformée en ondelette imposent de sous-échantillonner par deux le signal d'entrée, de faire du repliement d'image ou encore de ranger de manière bien précise les coefficients obtenus dans le plan de fréquence. Finalement, nous donnons une représentation simplifiée des états de la machine à états qui commande la majeure partie des signaux de contrôle et l'adressage mémoire.

Les fichiers source VHDL sont synthétisés dans trois technologies : SOI HS, BULK HS et BULK LL. La comparaison des résultats obtenus montre que le circuit synthétisé en technologie SOI est celui qui présente le meilleur facteur de qualité, c'est-à-dire la plus faible puissance dissipée par MHz. A la fréquence de fonctionnement maximale de 40MHz, ce circuit est capable d'effectuer la transformée en ondelettes d'une image de taille 128x128 pixels sur un niveau en 3,2ms.

8 Conclusion

La puissance dissipée par les circuits microélectroniques représente une contrainte de plus en plus forte lors de leur conception, que ce soit pour des raisons de coût, notamment celui du système de refroidissement, des raisons de fiabilité ou bien des raisons d'autonomie dans le cas de systèmes portables. Bien évidemment, ce problème n'est pas nouveau puisqu'il est apparu avec les premières montres à quartz. Seulement aujourd'hui, il concerne un spectre très large d'applications. Pour pallier ce problème, la réduction de la tension d'alimentation s'avère très efficace dans le cas où les performances attendues ne sont pas trop élevées. Cela permet en effet de réduire quadratiquement la puissance dynamique et exponentiellement la puissance due aux courants de fuite. Il faut néanmoins trouver un compromis entre d'un côté la consommation et de l'autre la vitesse, cette dernière chutant rapidement lorsque la tension d'alimentation devient inférieure à la tension de seuil.

Pour mieux appréhender le comportement des transistors SOI opérés en très basse tension, nous avons développé un modèle physique et analytique simple représentant le courant fourni par des transistors en inversion modérée et prenant en compte la tension de substrat V_{BS} et les effets de canal court. Ce modèle décrit correctement des transistors à substrat flottant et des transistors DTMOS et permet de calculer le temps de propagation de portes simples et différents paramètres tels que la tension de seuil logique d'un inverseur.

Nous nous sommes, dans une deuxième partie, intéressés au problème posé par la propagation de données sur des interconnexions longues. Il est en effet connu que, dans les technologies fortement sous-microniques actuelles, le délai est de plus en plus présent dans les fils car il tend à augmenter d'une génération technologique à l'autre alors que, dans le même temps, le délai dans les portes diminue. Intuitivement, nous pourrions penser que ce problème est aggravé à très basse tension puisque le courant fourni par les transistors est très faible. C'est en fait l'inverse qui se produit, les portes étant plus lentes. Cependant, pour charger des interconnexions longues, il va falloir utiliser de très gros transistors et donc beaucoup d'étages de bufferisation, pénalisant par la même la consommation. Nous nous sommes donc attachés à réduire la puissance dissipée lors de la propagation de données. Nous avons comparé deux modes de transmission, le mode tension et le mode courant. Pour le mode courant, nous avons développé un circuit émetteur-récepteur basé sur un circuit de lecture de type B. Il ressort de cette comparaison que le mode courant est très avantageux sur des interconnexions longues et des taux d'activité élevés, réduisant la puissance dissipée jusqu'à un facteur 2, mais qu'il doit être proscrit dans le cas d'interconnexions courtes et de taux d'activité faible.

Dans la suite, nous avons réalisé une bibliothèque de cellules standards, adaptée à la basse tension, en partant d'une bibliothèque de STMicroelectronics. Cette bibliothèque est basée sur le style logique CMOS, qui après comparaison avec d'autres styles logiques, s'avère être celui qui possède le meilleur facteur de qualité et la meilleure marge au bruit. Les éléments séquentiels sont basés sur la bascule du PowerPC, qui est constituée de portes de passage et est très efficace en basse tension.

Enfin, nous avons développé un circuit de traitement d'image par paquet d'ondelette, dans le cadre du projet IRISEP, destiné à la reconnaissance par l'iris de l'œil. Ce circuit a été synthétisé grâce à la bibliothèque précédente. Il présente de bonnes performances avec une fréquence de fonctionnement maximale de 40MHz, permettant de traiter une image de 128x128 pixels sur un niveau de transformée en 3,2ms. En outre, il présente une puissance dissipée de seulement 9,5µW/MHz. En cela, l'utilisation de la technologie SOI s'avère très intéressante puisque cette valeur est 16% inférieure à ce que nous pouvons obtenir en technologie silicium à substrat massif, à la même tension d'alimentation.

Pour conclure, l'axe de recherche développé ici, à savoir l'utilisation d'une technologie haute performance (SOI *High-Speed*) à basse tension d'alimentation, s'avère être une solution

d'avenir, à la condition, pour les technologies futures, d'introduire des techniques de réduction des courants de fuite.

Les points à améliorer

Le circuit émetteur-récepteur que nous avons développé a été fondu par STMicroelectronics. Cependant, les tests effectués n'ont pas permis de valider son fonctionnement. Le problème provient du convertisseur de niveau utilisé en sortie, dont le rôle est de transformer un signal de 0,5V d'amplitude en un signal de 1,2V d'amplitude. Celui-ci a été dimensionné avec les premières cartes modèle et refuse obstinément de passer d'un état à l'autre. Le même convertisseur a été utilisé en sortie d'oscillateurs en anneaux placés sur le même *wafer*, et nous n'avons visualisé aucune oscillation en sortie de ces oscillateurs.

L'autre point concerne le circuit de transformée en ondelettes, dont la seule synthèse logique a été réalisée. Toutes les vues géométriques, nécessaires à l'outil de placement/routage, n'ont pas pu être générées depuis notre bibliothèque de cellules standards, notamment la localisation des plots d'entrée/sortie. Ainsi, il n'a pas été possible de réaliser le layout. Pour effectuer l'analyse de puissance, un modèle statistique des capacités de fils (*wire load model*) a été utilisé.

Les axes de recherche qui restent à explorer

Il faudrait étudier plus avant la raison pour laquelle l'utilisation de transistors DTMOS n'apporte pas de gain significatif sur un circuit comportant autant de portes que le circuit de transformée en ondelettes, et donc beaucoup de nœuds fortement capacitifs.

De la même manière, il serait intéressant de pouvoir estimer l'avantage retiré d'une bibliothèque à double tension de seuil, même si celui-ci doit être limité compte tenu du faible pourcentage de la puissance des courants de fuite dans la puissance totale.

Bibliographie

[Alle99] D. Allen et al., "A 0.2µm 1.8V SOI 550MHz 64b PowerPC Microprocessor with Copper Interconnects," *International Solid-State Circuits Conference*, pp. 438-439, February 1999

[Alpe98] C. J. Alpert, A. Devgan and S. T. Quay, "Buffer Insertion for Noise and Delay Optimization," *DAC*, pp. 362-367, 1998

[Assa94] F. Assaderaghi, et al., "A Dynamic Threshold Voltage MOSFET (DTMOS) for Ultra Low-Voltage Operation," *IEDM 1994*

[Aust98] B. Austin, K. Bowman, X. Tang and J. D. Meindl, "A Low-Power Transregional MOSFET Model for Complete Power-Delay Analysis of CMOS Gigascale Integration (GSI)," in *Proc. 11th Annual IEEE Int. ASIC Conf.*, pp. 125-129, September 1998

[Beni94] L. Benini, et al., "Saving Power by Synthesizing gated clocks for sequential circuits," vol. 11, no 4, pp. 32-41, 1994

[Bern00] Kerry Bernstein and Norman J. Rohrer, *SOI Circuit Design Concepts*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 2000

[Blal92] Travis Blalock and Richard Jaeger, "A High-Speed Sensing Scheme for 1T Dynamic RAM's Utilizing the Clamped Bit-Line Sense Amplifier," *IEEE Journal of Solid-State Circuits,* vol. 27, No. 4*,* April 1992

[Bowm99]K. Bowman, B. Austin, J. Eble, X. Tang and J. D. Meindl, "A Physical Alpha-Power Law MOSFET Model," *IEEE Journal of Solid-States Circuits*, vol. 34, October 1999

[Buch03] Isidor Buchmann, "What's the best battery?," www.batteryuniversity.com/partone-3.htm

[Burr91] J. Burr and M. Peterson, "Ultra Low Power CMOS Technology," *NASA VLSI Design Symposium*, pp. 4.2.1-4.2.13, 1991

[Burr94] J. Burr and J. Shott, "A 200mV Self-Testing Encoder/Decoder Using Stanford Ultra-Low-Power CMOS," *International Solid-State Circuits Conference*, pp. 84-85, February 1994

[Casu01] M. R. Casu, et al., "Synthesis of Low-Leakage PD-SOI Circuits with Body-Biasing," *ISLPED 2001*, pp. 287-290

[Chan95] A. Chandrakasan and R. Brodersen, *Low Power Digital CMOS Design*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1995

[Chu87] K. M. Chu and D. L. Pulfrey, "A comparison of CMOS circuit techniques: Differential Cascode voltage switch logic versus conventional logic," *IEEE Journal of Solid-State Circuits,* vol. 22, pp. 528 - 532, August 1987

[Cohe92] A. Cohen, *Ondelettes et Traitement Numérique du Signal*, Masson, 1992

[Dadd65] Dadda, L.: "Some schemes for parallel multipliers," *Alta Freq*. , pp. 349–356, 1965

[Dall98] W. Dally and J. Poulton, *Digital Systems Engineering*, Cambridge U.K.: Cambridge Univ. Press, 1998

[Das03] Koushik Das and Richard Brown, "Ultra Low-Leakage Power Strategies for Sub-1V VLSI: Novel Circuit Styles and Design Methodologies for Partially Depleted Silicon-On-Insulator (PD-SOI) CMOS Technology," *Proc. of the 16th International Conference on VLSI Design*, 2003

[Dhae92] T. Dhaene and D. Zutter, "Selection of Lumped Element Models for Coupled Lossy Transmission Lines," *IEEE Trans. Computer-Aided Design*, vol. 11, pp. 805-815, July 1992

[Dhao02] I. Ben Dhaou, "Low-Power Design Techniques for Deep Submicron Technology with Application to Wireless Transceiver Design," *PhD Dissertation*, Royal Institute of Technology, Sweden, 2002

[Dhar91] S. Dhar and M. Franklin, "Optimum Buffer Circuits for Driving Long Uniform Lines," *IEEE Journal of Solid-State Circuits*, vol. 26, no. 1, pp. 32-40, January 1991

[Dobb92] D. W. Dobberpuhl et al., "A 200MHz 64b Dual Issue CMOS Microprocessor, " *IEEE Journal of Solid-State Circuits*, vol. 27, no. 11, pp. 1555-1567, November 1992

[Dous96] T. Douseki et al., "A 0.5V SIMOX-MTCMOS Circuit with 200ps Logic Gate," *International Solid-State Circuits Conference*, pp. 84-85, 1996

[Elmo48] W. C. Elmore, "The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers," *Journal of Applied Physics*, Vol. 19, pp. 55-63, January 1948

[Enz95] C. C. Enz, F. Krummenacher and E. A. Vittoz, "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications," *Special Issue Analog Integrated Circuits and Signal Processing J. Low-Voltage and Low-Power Design*, vol. 8, pp. 83-114, July 1995

[Frie97] R. Frier, "Minimizing Energy Dissipation in High-Speed Multipliers," *International Symposium on Low Power Electronics and Design*, pp. 214-219, August 1997

[Fuse96] T. Fuse et al., "0.5V SOI CMOS Pass-Gate Logic," *International Solid-State Circuits Conference*, pp. 88-89, 1996

[Gero94] G. Gerosa et al., "A 2.2W, 80MHz Superscalar RISC Microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 1440-1452, December 1994

[Gine90] L. P. van Ginneken, "Buffer Placement in Distributed RC-tree Networks for Minimal Elmore Delay," in *Proc. Int. Symp. on Circuits and Systems*, pp. 865-868, 1990

[Gonz97] R. Gonzalez, B. Gordon and M Horowitz, "Supply and Threshold Voltage Scaling for Low Power CMOS," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1210-1216, August 1997

[Gosw99] Jaideva Goswami and Andrew Chan, *Fundamentals of Wavelets: Theory, Algorithms and Applications*, Wiley Series in Microwave and Optical Engineering, Wiley Interscience, 1999

[Gunt00] S. H. Gunther, F. Binns, D. Carmean and J. C. Hall, "Managing the impact of increasing microprocessor power consumption," in *Proc. Intel Technology Journal*, March 2000

[Han87] T. Han and D.A. Carlson, "Fast Area-Efficient VLSI Adders," *Proceedings of the 8th IEEE Symposium on Computer Arithmetic*, 1987, pp. 49-56

[Hash02] T. Hashimoto, et al., "A 27-MHz/54-MHz 11-mW MPEG-4 Video Decoder LSI for Mobile Applications," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 11, pp 1574-1581, November 2002

[Hass01] K. Hass, J. Venbrux and P. Bhatia, "Logic Design Considerations for 0.5-Volt CMOS," *Conf. on Advanced Research in VLSI*, March 2001

[Hede94] N. Hedenstierna and K. O. Jeppson, "Comments on 'The Optimum CMOS Tapered Buffer Problem'," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 2, pp. 155-159, February 1994

[Heer04] J. Heer et al., Ultra Low Power special session, Date 2004 Paris

[Hell84] L. G. Heller, W. R. Griffin, J. W. Davis, and N. G. Thoma, "Cascode voltage switch logic: A differential CMOS logic family," *IEEE International Solid-State Circuits Conference,* vol. XXVII, pp. 16 - 17, February 1984

[Henn96] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, Morgan Kaufmann, San Francisco, CA, 1996

[Hori93] M. Horiguchi, T. Sakata and K. Itoh, "Switched Source-Impedance CMOS Circuit for Low Standby Subthreshold Current Giga-Scale LSI's," *IEEE Journal of Solid-States Circuits*, vol. 28, November 1993

[Hubb95] B. Hubbard, *Ondes et ondelettes*, Collection Sciences d'Avenir, Pour la science, Paris, 1995

[Inte98] Intel Technology Journal Q3'98

|Inte04] http://www.intel.com/design/intelxscale/

[Isma99] Y. Ismail, E. Friedman and J. Neves, "Figures of Merit to Characterize the Importance of On-Chip Inductance," *IEEE Trans. VLSI Syst.*, vol. 7, pp. 442-449, December 1999

[ITRS04] http://www.itrs.net/Common/2004Update/2004_000_ORTC.pdf

[Jaeg75] R. C. Jaeger, "Comments on 'An Optimized Output Stage for MOS Integrated Circuits'," *IEEE Journal of Solid-State Circuits*, vol. SC-10, no. 3, pp. 185-186, June 1975

[Kanu83] A. Kanuma, "CMOS Circuit Optimization," *Solid-State Electron.*, vol. 26, no. 1, pp. 47-58, 1983

[Kato02] Atul Katoch, Evert Seevinck and Harry Veendrick, "Fast Signal Propagation for Point to Point On-Chip Long Interconnects using Current Sensing," *European Solid-State Circuits Conference*, September 2002

[Kesh97] A. Keshavarzi, K. Roy and C. Hawkins, "Intrinsic Leakage in Low Power Deep Submicron ICs, " *International Test Conference*, p. 146, November 1997

[Kogg73] P. Kogge and H. Stone, "A Parallel Algorithm for the Efficient Solution of a General Class of Recurrent Solutions," *IEEE Trans. Computers*, vol. C-22, no. 8, pp. 786-793, August 1973

[Kuro96] Tadahiro Kuroda, et al., "A 0.9V, 150MHz, 10mW, 4mm², 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage (VT) Scheme, " *IEEE Journal of Solid-State Circuits*, pp. 1770-1779, November 1996

[Laks02] Lakshmanan, M. Othman and M. Ali, "High performance parallel multiplier using Wallace-Booth algorithm," *IEEE International Conference on Semiconductor Electronics*, pp. 433-436, December 2002.

[Lev] Lavi Lev and Ping Chao, "Down to the Wire: Requirements for Nanometer Design Implementation," Cadence White Papers, http://www.cadence.com/whitepapers/4064_NanometerWP_fnlv2.pdf

[Li90] N. C. Li, G. L. Haviland and A. A. Tuszynski, "CMOS Tapered Buffer," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 4, pp. 1005-1008, August 1990

[Liao02] M.-J. Liao, C.-F. Su, C.-Y. Chang and A. Wu, "A Carry-Select-Adder Optimization Technique for High-Performance Booth-Encoded Wallace-Tree Multipliers," *IEEE International Symposium on Circuits and Systems*, vol. 1, pp. I-81 – I.84, May 2002

[Lill96] J. Lillis, C.-K. Cheng and T.-T. Lin, "Optimal Wire Sizing and Buffer Insertion for Low Power and a Generalized Delay Model," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 3, pp. 437-447, March 1996

[Lin75] H. C. Lin and L. W. Linholm, "An Optimized Output Stage for MOS Integrated Circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-10, no. 2, pp. 106-109, April 1975

[Mack90] P. Macken, et al., "A Voltage Reduction Technique for Digital Systems," *International Solid-States Circuits Conference*, pp. 238-239, February 1990

[Mahe01] Atul Maheshwari and Wayne Burleson, "Current Sensing for Global Interconnects, Secondary Design Issues: Analysis and Solutions," *IEEE International Workshop on Power and Timing Modeling, Optimization and Simulation*, September 2001

[Mall89] Stéphane Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 11, no.7, 1989

[Meie96] P. Meier, R. Rutenbar and R. Carley, "Exploring Multiplier Architecture and Layout for Low-Power," *Custom Integrated Circuits Conference*, May 1996

[Moor65] Gordon E. Moore, "Cramming More Components Onto Integrated Circuits," *Electronics*, vol. 38, April 1965

[Muto95] Shin'ichiro Mutoh, et al., "1-V Power Supply High-Speed Digital Circuit Technology with Multi-Threshold Voltage CMOS," *IEEE Journal of Solid-State* Circuits, vol. 30, no. 8, pp. 847-854, August 1995

[Nare01] Siva Narendra, et al., "Scaling of Stack Effect and its Application for Leakage Reduction," *ISLPED 2001*, pp. 195-200

[Oklo02] Vojin G. Oklobdzija, *The Computer Engineering Handbook*, pp. 2-54, CRC Press, 2002

[Pigu97] C. Piguet, et al., "Low-Power Design of 8-bit Embedded CoolRISC Microcontroller Cores," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 7, pp. 1067-1078, July 1997

[Raba96] Jan M. Rabaey, *Digital Integrated Circuits: A Design Perspective*, Prentice Hall, 1996

[Radh85] D. Radhakrishnan, S. R. Whitaker, and G. K. Maki, "Formal design procedures for pass transistor switching circuits," *IEEE Journal of Solid-State Circuits,* vol. 20, pp. 531 - 536, April 1985

[Reed04] http://www.reedelectronics.com/electronicnews/article/CA381507?stt=000&industryid=2

[Roge96] R. Rogenmoser, H. Kaeslin and N. Felber, "The Impact of Transistor Sizing on Power Efficiency in Submicron CMOS Circuits," *Proc. 22nd European Solid-State Circuits Conference*, pp. 124-127, Neuchâtel, Switzerland, September 1996

[Rydg04-a] Erik Rydgren, "Iris Recognition Using Wavelet Analysis," *Master of Science Thesis*, ISEP, Paris, 2004

[Rydg04-b] Erik Rydgren, Thomas Ea, Frederic Amiel, Florence Rossant and Amara Amara, "Iris Features Extraction Using Wavelet Packets," *IEEE International Conference on Image Processing*, Singapore, October 2004

[Seev91] Evert Seevinck, *Senior Member*, *IEEE*, Petrus J. van Beers, and Hans Ontrop, "Current-Mode Techniques for High-Speed VLSI Circuits with Application to Current Sense Amplifier for CMOS SRAM's," *IEEE Journal of Solid-States Circuits*, vol. 26, no. 4, April 1991

[Saku83] T. Sakurai, "Approximation of Wiring Delay in MOSFET LSI," *IEEE Journal of Solid-State Circuits*, vol. SC-18, pp. 418-426, August 1983

[Saku90] T. Sakurai and R. Newton, "Alpha-Power Law MOSFET Model and Its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-States Circuits*, Vol. 25, April 1990

[Shah99] Ghavam Shahidi, et al., "Partially-Depleted SOI Technology for Digital Logic," *IEEE International Solid-State Circuits Conference*, 1999

[Shib96] Nobutaro Shibata, "Current Sense Amplifiers for Low-Voltage Memories," *IEICE Trans. Electron.*, vol. E79-C, no. 8, August 1996

[Shoj88] M. Shoji, *CMOS Digital Circuit Technology*, Englewood Cliffs, Prentice-Hall, 1988

[Soel00] H. Soeleman et al., "Robust Ultra-Low Power Sub-Threshold DTMOS Logic," *International Symposium on Low Power Electronics and Design*, 2000

[Stoj99] V. Stojanovic and V. Oklobdzija, "Comparative Analysis of Master-Slave Latches and Flip-Flops for High Performance and Low-Power Systems," *IEEE Journal of Solid-States Circuits*, vol. 34, no.4, pp. 536-548, April 1999

[Stor94] J. M. C. Stork, "Technology Leverage for Ultra-Low Power Information Systems," *IEEE Symposium on Low Power Electronics*, Tech. Dig., pp. 52-55, October 1994

[Suzu93] M. Suzuki, N. Ohkubo, T. Shinbo, T. Yamanaka, A. Shimizu, K. Sasaki, and Y. Nakagome, "A 1.5-ns 32-b CMOS ALU in double pass-transistor logic," *IEEE Journal of Solid-State Circuits,* vol. 28, pp. 1145 - 1151, November 1993

[Swan72] R. M. Swanson and J. D. Meindl, "Ion-implanted Complementary MOS Transistors in Low-Voltage Circuits," *IEEE Journal of Solid-States Circuits*, vol. SC-7, pp. 146-153, April 1972

[Sze81] S. M. Sze, *Physics of Semiconductor Devices*, 2nd edition New-York: Wiley 1981

[Tran04] http://www.transmeta.com/efficeon/efficeon_tm8800.html

[Unge86] S. H. Unger and C. Tan, "Clocking Schemes for High-Speed Digital Systems," *IEEE Trans. Comput.*, vol. C-35, pp. 880-895, October 1986

[Vara02] Hemmige Varadarajan, et al., "Low-Power Design Issues," in *The Computer Engineering Handbook*, Vojin G. Oklobdzija, CRC Press, 2002

[Veen98] Harry Veendrick, *Deep-Submicron CMOS ICs*, Kluwer Academic Publishers, Deventer, the Netherlands, 1998

[Vemu91] S. R. Vemuru and A. R. Thorbjornsen, "Variable-Taper CMOS Buffers," *IEEE Journal of Solid-State Circuits*, vol. 26, pp. 1265-1269, September 1991

[Wall64] C. S. Wallace, "Suggestions for a Fast Multiplier," *IEE Trans. Electron. Computers*, EC-13, pp. 14-17, 1964

[Wich03] Bernhard Wicht, *Current Sense Amplifiers for Embedded SRAM in High-Performance System-on-a-Chip Designs*, Springer Verlag, Heidelberg, 2003

[Yano96] K. Yano, Y. Sasaki, K. Rikino, and K. Seki, "Top-down pass-transistor logic design," *IEEE Journal of Solid-State Circuits,* vol. 31, pp. 792 - 803, June 1996

[Yuan89] J. Yuan and C. Svensson, "High-speed CMOS circuit technique," *IEEE Journal of Solid-State Circuits,* vol. 24, pp. 62 - 70, February 1989

[Zhu00] Y. Zhu, T. Tan and Y. Wang, "Biometric Personal Identification Based on Iris Patterns," *International Conference on Pattern Recognition*, Barcelona, Spain, 2000

Publications associées à ce travail

Journal :

"Modeling Subthreshold SOI Logic for Static Timing Analysis," Alexandre Valentian, Olivier Thomas, Andreï Vladimirescu, Amara Amara, *IEEE Transactions on VLSI*, June 2004

Publications :

"A 130nm Partially Depleted SOI Technology Menu for Low-Power Applications," Nicolas L'Hostis, Alexandre Valentian, Philippe Flatresse, Amara Amara, *IEEE Northeast Workshop on Circuits and Systems*, Quebec, CANADA, 2005

"Ultra-Low-Voltage Robust Design Issues in Deep-Submicron CMOS," Andreï Vladimirescu, Yu Cao, Olivier Thomas, Alexandre Valentian, Razvan Ionita, Jean Rabaey, Amara Amara, *IEEE Northeast Workshop on Circuits and Systems*, Montreal, CANADA, June 2004.

"On-Chip Signaling for Ultra Low-Voltage 0.13µm CMOS SOI Technology," Alexandre Valentian, Amara Amara, *IEEE Northeast Workshop on Circuits and Systems*, Montreal, Canada, June 2004

"Modélisation du délai d'une porte CMOS SOI en faible inversion," Alexandre Valentian, Olivier Thomas, Amara Amara, Andreï Vladimirescu, *Journées d'Etudes Faible Tension Faible Consommation*, Paris, Mai 2003

"An Accurate Estimation Model for Subthreshold CMOS SOI Logic," Olivier Thomas, Alexandre Valentian, Andreï Vladimirescu, Amara Amara, *IEEE European Solid-State Circuit Conference*, Firenze, Italy, September 2002.

Annexes

Annexe 1 : Calcul de la tension de seuil logique

La tension de seuil logique est la tension présente en entrée pour laquelle la sortie est à $V_{DD}/2$. La tension drain-source V_{DS} des transistors NMOS et PMOS étant supérieure à la tension de saturation sous le seuil – définie dans la Figure 29 – , ils sont tous les deux dans la région de saturation : on peut alors négliger l'exponentielle dans l'expression de leur courant de drain. La tension de seuil logique est obtenue en égalisant les courants des transistors :

$$W_n \cdot d_{on} \cdot 10^{\left(\frac{V_M - V_{Tn}}{S_n}\right)} \cdot \left(a + \lambda_n \frac{V_{DD}}{2}\right) = W_p \cdot d_{op} \cdot 10^{\left(\frac{V_{dd} - V_M - |V_{tp}|}{S_p}\right)} \cdot \left(a - \lambda_p \frac{V_{DD}}{2}\right)$$

Équation 72

On obtient alors :

$$V_M = \frac{S_n \cdot S_p}{S_n + S_p}\left[\frac{V_{tn}}{S_n} + \frac{V_{DD} - |V_{tp}|}{S_p} + \log_{10}\left(\frac{d_{op}}{d_{on}} \cdot \frac{a - \lambda_p \frac{V_{DD}}{2}}{a + \lambda_n \frac{V_{DD}}{2}} \cdot \frac{W_p}{W_n}\right)\right]$$

Équation 73

Annexe 2 : Calcul du temps de propagation tp_{HL} pour une entrée rapide

On considère ici le cas de la décharge de la capacité de sortie. Puisque le signal d'entrée varie rapidement, les effets du transistor PMOS peuvent être négligés car celui-ci se trouve dans un état bloqué. Le temps de propagation est défini comme étant le temps séparant les transitions de l'entrée et de la sortie par $V_{DD}/2$; il faut donc intégrer entre V_{DD} et $V_{DD}/2$ l'équation suivante :

$$i_D(V_{OUT}) = C_{tot} \cdot \frac{dV_{OUT}}{dt}$$

Équation 74

Le transistor NMOS se trouvant dans l'état saturé, l'expression de son courant de drain est la suivante :

$$I_D = I_{SS_n} \cdot \left(1 - e^{-\frac{V_{sat_n}}{V_{th}}}\right) \cdot (a + \lambda_n V_{out})$$

Équation 75

Le temps de propagation vaut donc :

$$tp_{HL} = \frac{C_{tot}}{I_{SS_n}} \left[\frac{1}{\lambda_n \left(1 - e^{-\frac{V_{sat_n}}{V_{th}}}\right)} \cdot \log\left(\frac{a + \lambda_n V_{DD}}{a + \lambda_n \cdot 0.5 \cdot V_{DD}}\right) \right]$$

Équation 76

Annexe 3 : Calcul du temps de propagation tp_{LH} pour une entrée rapide

De la même manière que précédemment, il faut intégrer le courant du transistor PMOS. L'expression du temps de propagation est donnée par :

$$tp_{LH} = C_{tot} \int_{-V_{DD}}^{-0.5V_{DD}} \frac{dV_{out}}{I_D(V_{out})}$$

$$tp_{LH} = \int_{-V_{DD}}^{-0.5V_{DD}} \frac{dV_{out}}{I_{SS_p}\left(1 - e^{\frac{V_{sat_p}}{V_{th}}}\right)(a + \lambda_p V_{out})}$$

Équation 77

On obtient alors :

$$tp_{LH} = \frac{C_{tot}}{I_{SS_p}} \left[\frac{1}{\lambda_p \left(1 - e^{\frac{V_{sat_p}}{V_{th}}}\right)} \cdot \log\left(\frac{a + \lambda_p \cdot 0.5 \cdot V_{DD}}{a + \lambda_p V_{DD}} \right) \right]$$

Équation 78

Annexe 4 : Calcul du temps de propagation tp_{HL} pour une entrée lente

Comme expliqué précédemment à l'aide de la Figure 39, une approximation au premier ordre valide consiste à négliger le courant du transistor PMOS : celui-ci est en effet plusieurs ordres de grandeur plus petit que le courant de décharge du transistor NMOS dû aux propriétés sous le seuil des transistors. Il faut donc intégrer le courant du transistor NMOS par rapport à V_{GS} et à V_{DS}, soit respectivement par rapport au temps t et à la tension de sortie V_{OUT}. L'équation différentielle à intégrer est :

$$I_D(t, V_{OUT}) = C \cdot \frac{dV_{OUT}}{dt}$$

L'intégration se fait entre les bornes $[0,t]$ et $[V_{DD}, V_{DD}/2]$:

$$A_n \int_0^t 10^{\frac{kt}{S_n}} dt = C \int_{V_{DD}}^{V_{DD}/2} \frac{dV_{OUT}}{a_n + \lambda_n V_{OUT}}$$

Équation 79

où k est la pente du signal d'entrée et $A_n = W \cdot d_{on} \cdot 10^{-\frac{VT_{WN}}{S_n}} \cdot \left(1 - e^{-m\frac{V_{sat_n}}{V_{th}}}\right)$. Après intégration, on a :

$$\frac{A_n \cdot S_n}{k \cdot \log(10)} \left(10^{\frac{kt}{S_n}} - 1\right) = \frac{C}{\lambda_n} \log\left(\frac{a_n + \lambda_n V_{DD}/2}{a_n + \lambda_n V_{DD}}\right)$$

ce qui donne :

$$t = \frac{S_n}{k} \log_{10}\left[\frac{k \cdot C \cdot \log(10)}{A_n \cdot \lambda_n \cdot S_n} \log\left(\frac{a_n + \lambda_n V_{DD}/2}{a_n + \lambda_n \cdot V_{DD}}\right) + 1\right]$$

Équation 80

Le temps t obtenu représente le temps séparant les transitions de l'entrée par 0 et de la sortie par $V_{DD}/2$. Pour avoir le temps de propagation tp_{HL}, il faut soustraire le temps mis par l'entrée pour atteindre la tension $V_{DD}/2$:

$$tp_{HL} = \frac{S_n}{k}\log_{10}\left[\frac{k\cdot C\cdot\log(10)}{A_n\cdot\lambda_n\cdot S_n}\log\left(\frac{a_n+\lambda_n V_{DD}/2}{a_n+\lambda_n\cdot V_{DD}}\right)+1\right]-\frac{V_{DD}}{2k}$$

Équation 81

Annexe 5 : Calcul du temps de propagation tp_{LH} pour une entrée lente

Comme dans l'Annexe 4, on néglige le courant de court-circuit provenant du transistor NMOS. On intègre le courant de saturation du transistor PMOS :

$$A_p \int_0^t 10^{\frac{kt}{S_p}} dt = C \int_{-V_{DD}}^{-V_{DD}/2} \frac{dV_{OUT}}{a_p + \lambda_p V_{OUT}}$$

Équation 82

avec $A_p = W_p \cdot d_{op} \cdot 10^{-\frac{|VT_{WP}|}{S_p}} \cdot \left(1 - e^{-m\frac{V_{sat_p}}{V_{th}}}\right)$.

On obtient :

$$t = -\frac{S_p}{k} \log_{10}\left[-\frac{k \cdot C \cdot \log(10)}{A_p \cdot \lambda_p \cdot S_p} \log\left(\frac{a_p - \lambda_p V_{DD}/2}{a_p - \lambda_p V_{DD}} \right) + 1 \right]$$

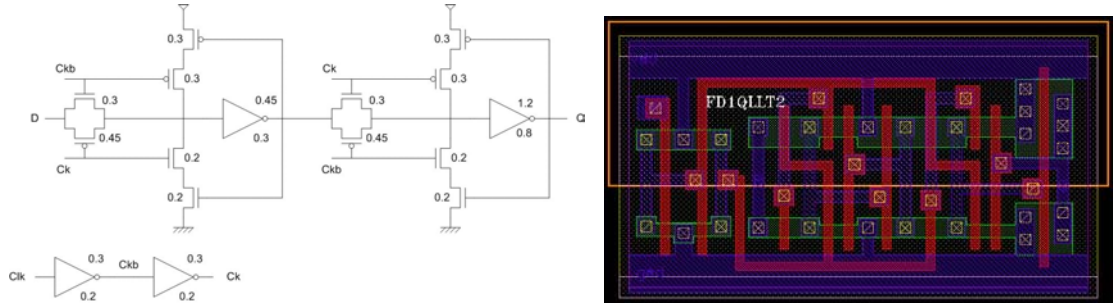
Il faut là aussi prendre en compte le temps mis par l'entrée pour atteindre la tension $V_{DD}/2$. L'expression du temps de propagation est alors :

$$tp_{LH} = -\frac{S_p}{k} \log_{10}\left[-\frac{k \cdot C \cdot \log(10)}{A_p \cdot \lambda_p \cdot S_p} \log\left(\frac{a_p - \lambda_p V_{DD}/2}{a_p - \lambda_p V_{DD}} \right) + 1 \right] - \frac{V_{DD}}{2|k|}$$

Équation 83

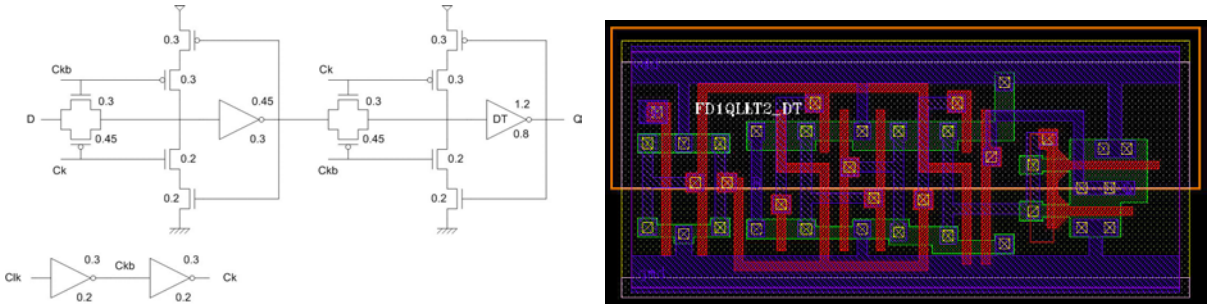
Annexe 6 : Les éléments séquentiels introduits dans la bibliothèque

Dans cette annexe, nous avons reproduit les éléments séquentiels que nous avons développés pour la bibliothèque. Ceux-ci ont été dimensionnés avec le produit puissance-délai pour objectif : les tailles des transistors sont exprimées en micromètres.
La Figure 111 représente une bascule D de taille deux. Une variante avec inverseur de sortie DTMOS est donnée dans la Figure 112.



(a) **(b)**

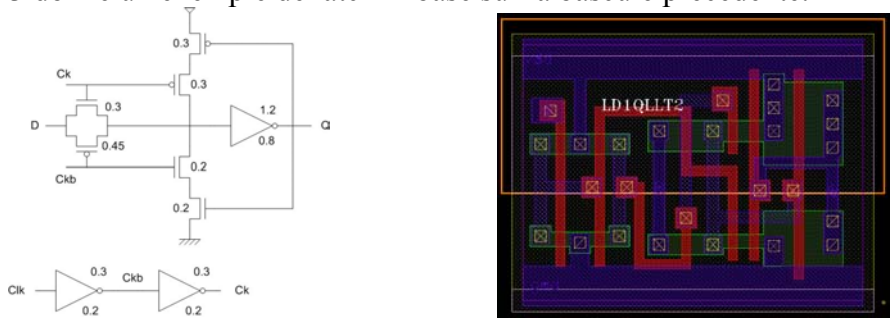
Figure 111 (a) Schéma et (b) layout de la bascule D.



(a) **(b)**

Figure 112 (a) Schéma et (b) layout de la bascule D avec inverseur de sortie de type DTMOS.

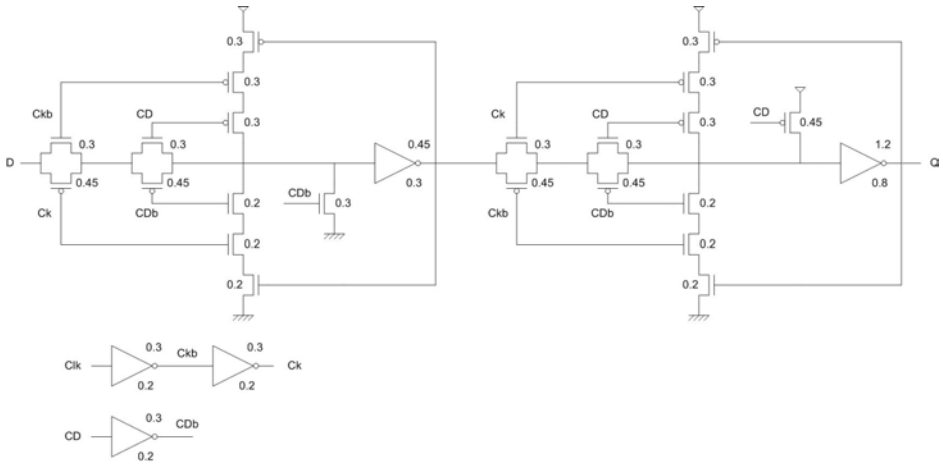
La Figure 113 donne un exemple de latch D basé sur la bascule précédente.



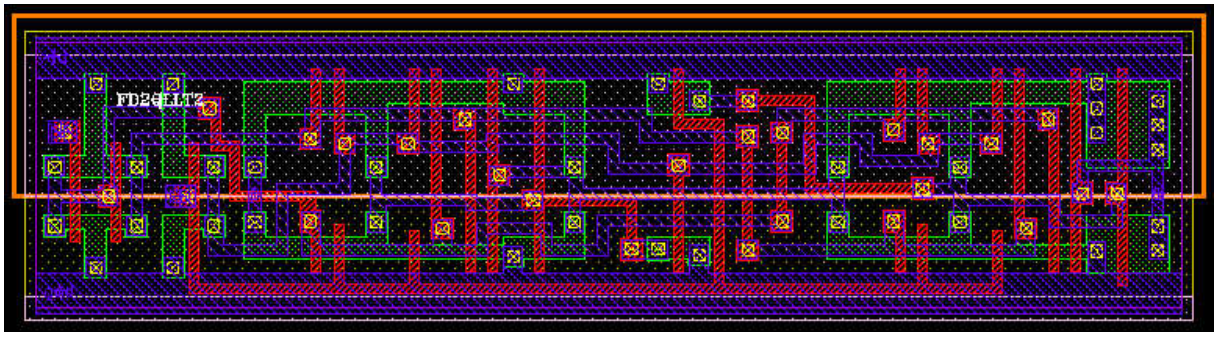
(a) **(b)**

Figure 113 (a) Schéma et (b) layout du latch D.

La Figure 114 représente la bascule D avec commande de remise à zéro asynchrone.



(a)



(b)

Figure 114 (a) Schéma et (b) layout de la bascule D avec remise à zéro asynchrone (commance CD).

Tous les éléments séquentiels précédents existent en plusieurs tailles et en variante DTMOS, qui n'ont pas été montrées ici.