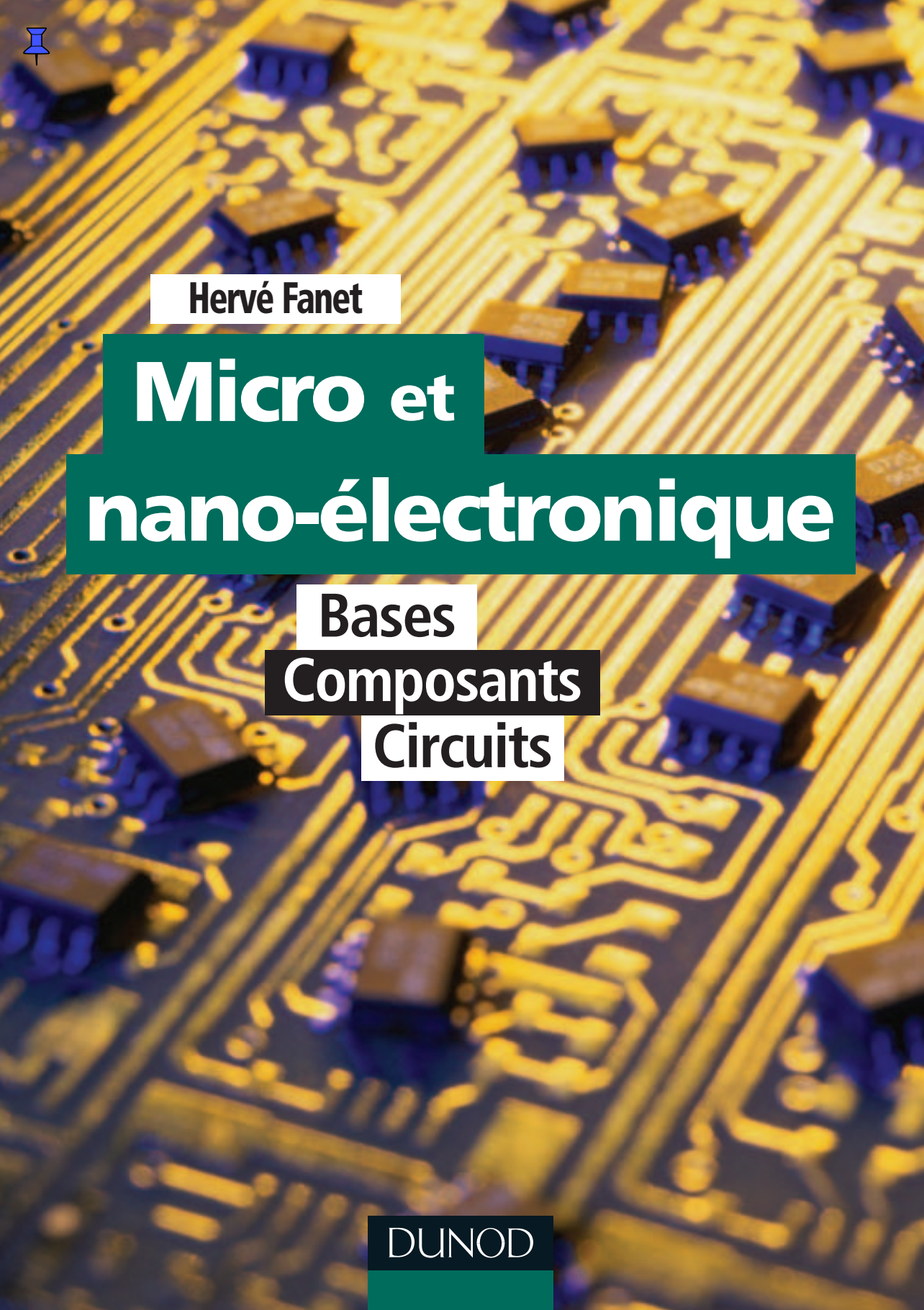


Hervé Fanet

Micro et nano-électronique

Bases
Composants
Circuits

DUNOD

Technologie électronique

Hervet Fanet

Micro et nano-électronique

Bases • Composants • Circuits

DUNOD

ISBN 2 10 049141 5

Table des matières

Introduction

La technologie micro-électronique progresse très rapidement depuis 1950. Les progrès réalisés contribuent à augmenter de manière significative la part de l'électronique dans les produits industriels. Cette évolution amène à faire chuter le coût d'une fonction électronique donnée puisque de plus en plus de transistors sont intégrés dans un même circuit. Cette évolution va-t-elle se poursuivre dans les années futures ? Quelles sont les limites physiques à la miniaturisation poussée ? De nouvelles technologies sont-elles susceptibles de remplacer le transistor ? Cet ouvrage tente de répondre à ces questions en étudiant à la fois les aspects technologiques et architecturaux des systèmes électroniques intégrés.

La première partie de cet ouvrage (chapitres 1 et 2) est un résumé des méthodes électriques et physiques nécessaires à la compréhension du fonctionnement des composants électroniques. Les lois de l'électromagnétisme sont le plus souvent utilisées mais aussi les principes de base de la physique des semi-conducteurs. Les transistors ont aujourd'hui des dimensions inférieures au micron si bien que les lois de la mécanique quantique sont nécessaires pour comprendre certains mécanismes. Les effets tunnel et balistique sont les plus représentatifs.

La deuxième partie (chapitres 3, 4, 5 et 6) est une description assez détaillée du transistor MOS car ce dispositif est la brique de base avec laquelle sont fabriqués aujourd'hui la plupart des circuits électroniques intégrés. Le fonctionnement du transistor bipolaire est décrit de manière succincte car il est de moins en moins utilisé. Une description simplifiée des procédés de fabrication de l'industrie micro-électronique est également donnée au chapitre 6.

La troisième partie (chapitres 7, 8 et 9) est consacrée à la description de l'intégration des fonctions de base de l'électronique, fonctions analogiques et fonctions numériques. Le chapitre 9 décrit les principales caractéristiques des circuits intégrés complexes, processeurs et mémoires.

La dernière partie de l'ouvrage (chapitres 10, 11 et 12) introduit les composants nanométriques et les architectures associées. Des composants nouveaux comme les nanofils et les nanotubes, les dispositifs à électron unique et les molécules fonctionnalisées sont décrits ainsi que leurs perspectives d'utilisation dans les systèmes de traitement de l'information.

En résumé, cet ouvrage tente de donner au lecteur une vision globale de l'électronique intégrée en traitant à la fois les aspects physiques, technologiques et architecturaux.

Chapitre 1

De la micro à la nano-électronique

La part de l'électronique est croissante dans les produits de grande consommation, et cette évolution se fait sans faire croître le prix des produits de manière significative. Les raisons à cela sont le remarquable niveau d'automatisation atteint dans la fabrication des composants électroniques et la réduction continuelle de leurs tailles. Cette évolution va-t-elle continuer de manière inexorable, et de nouveaux dispositifs vont-ils apparaître dans les années futures ?

Répondre à ces questions est le sens général de cet ouvrage.

1.1 La part de l'électronique dans les produits industriels

La part prise par l'électronique dans les produits industriels ne fait que croître d'année en année et apporte le plus souvent aux produits un facteur de différentiation fort. Pensons à l'apport de l'électronique dans l'automobile, la photographie, les activités de gestion des entreprises, la communication entre personnes…

Les produits sont très divers et pourtant trois fonctions principales sont réalisées par les systèmes électroniques dans les produits :

- transporter des données d'un point à un autre ;
- effectuer des calculs à la demande de l'utilisateur ;
- opérer un contrôle-commande.

Trois exemples aident à comprendre cette classification. Un système de télévision permet de transmettre une image du studio d'enregistrement à l'écran du téléspectateur. L'écolier qui effectue une opération à l'aide d'une calculette utilise l'électronique comme technique de calcul. Les circuits dans une machine à laver le linge reçoivent des données et commandent des actions (chauffage, ouverture de vannes, démarrage de moteur).

Toutes ces fonctions, et c'est la raison du succès de l'électronique, peuvent se réaliser avec un seul composant : le transistor. Le transistor est un dispositif qui permet de commander un courant à l'aide d'une tension appliquée. Il est analogue au robinet qui permet de régler le flux d'un liquide. Quand l'opération est continue, on parle d'électronique analogique. Quand l'opération est du tout ou rien, on parle d'électronique numérique. Le courant passe ou ne passe pas. L'électronique numérique qui ne connaît que deux états est donc bien adaptée au système binaire de numération. Notons qu'une électronique à plusieurs états serait peut-être mieux adaptée à un autre système de numération.

Le transistor n'est pas le premier composant utilisé pour faire des calculs. Les premiers ont été les dispositifs mécaniques qui ont par exemple permis à Pascal de construire une machine à calculer. Les métiers à tisser réalisaient avec des systèmes mécaniques des opérations de contrôle-commande sophistiquées. Ils ont d'ailleurs inspiré l'inventeur Charles Babbage qui a imaginé le premier ordinateur faisant usage d'un programme. Cet ordinateur n'a cependant pas pu être réalisé par l'inventeur. Le relais électromécanique s'est imposé comme un composant de choix dans les applications électriques. Il a également été utilisé comme une sorte de transistor en mode logique. Enfin, le tube à vide a été inventé puis utilisé aussi bien en analogique qu'en logique. Le premier ordinateur appelé ENIAC comportait des milliers de tubes de type triode pour effectuer quelques calculs. Ensuite, le transistor est apparu et a remplacé tous les dispositifs précédents. Le transistor est particulièrement intéressant car il est de taille réduite et travaille avec des tensions électriques faibles.

La *figure 1.1* montre la place de l'électronique dans tous les domaines de la vie quotidienne : au travail, à la maison, dans les transports, dans un centre de soins, dans un lieu de loisirs.

La deuxième invention déterminante est celle du circuit intégré. Un système électronique se réalise en interconnectant des milliers et souvent des millions de transistors entre eux. L'idée du circuit intégré est de réaliser tous les transistors dans un même morceau de matériau et de réaliser les interconnexions également dans cet élément appelé puce. Cette idée est attribuée à Jack Kilby en 1958. Elle permet véritablement de réaliser des systèmes électroniques à faible coût puisque les opérations de fabrication des transistors et des interconnexions peuvent être automatisées. Ajoutons à cela qu'il est possible de fabriquer quelques milliers de circuits identiques en même temps, et on comprend facilement pourquoi des objets aussi complexes peuvent être aussi bon marché. La fabrication collective des circuits intégrés est représentée de manière simplifiée *figure 1.2*.

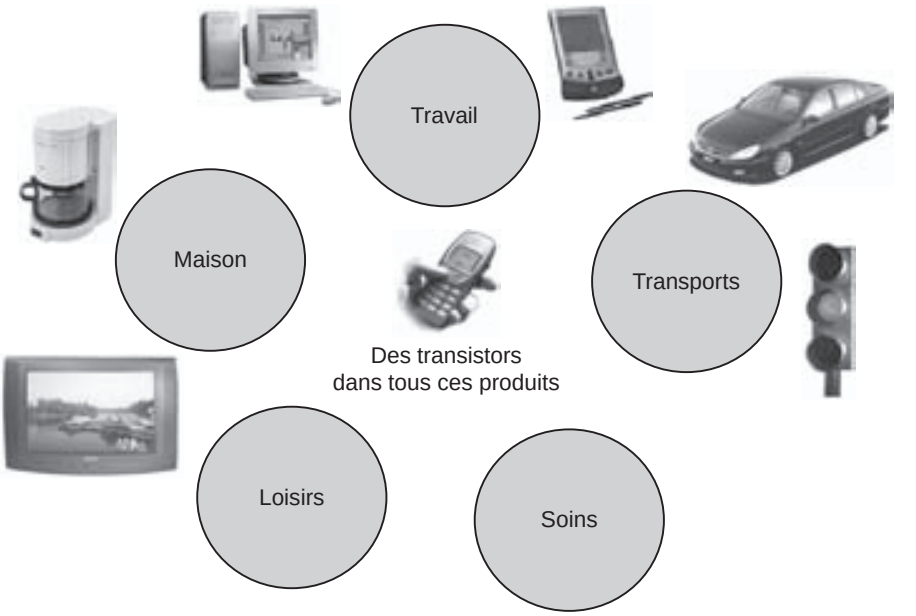


Figure 1.1 – Des transistors dans tous ces produits.

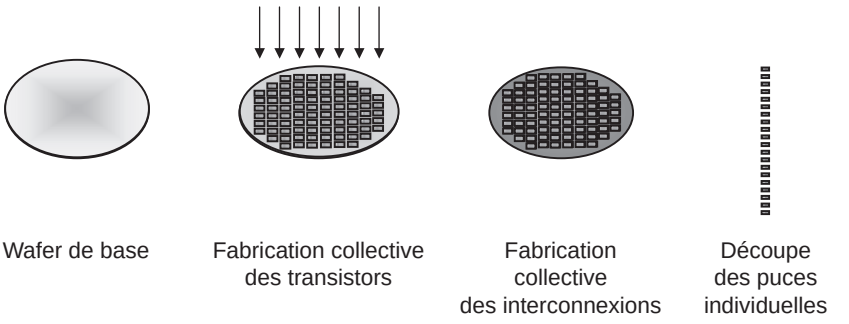


Figure 1.2 – Fabrication collective des circuits intégrés.

Le matériau de base se présente sous forme d'un disque de semi-conducteur. Nous expliquerons par la suite pourquoi le matériau de base n'est ni un isolant ni un conducteur. Le silicium s'est imposé comme le meilleur choix. Il est en effet abondant comme matière première puisque fait à partir de sable. D'autre part, son oxyde naturel, le dioxyde de silicium est stable.

Le disque de silicium appelé wafer est alors traité collectivement. Tous les circuits et tous les transistors des circuits sont réalisés en même temps. Il faut plusieurs étapes dans la fabrication qui sera décrite en détail dans le chapitre 6. Ces étapes sont toutefois mises en œuvre sur tous les transistors à la fois. Les interconnexions entre transistors sont réalisées de la même manière.

Pour être complet, il faut ajouter que quelques composants supplémentaires doivent également être fabriqués collectivement car ils sont nécessaires au fonctionnement des circuits électroniques. Ce sont les résistances, les condensateurs et les selfs. Leur nombre est cependant faible en proportion.

On comprend alors que le coût de fabrication d'un transistor diminue quand la taille du transistor diminue et quand la taille du wafer augmente. Le coût d'une opération sur un wafer est en effet relativement constant. Cette évolution est manifeste sur la *figure 1.3* qui représente l'évolution du coût de production d'un transistor en micro-électronique.

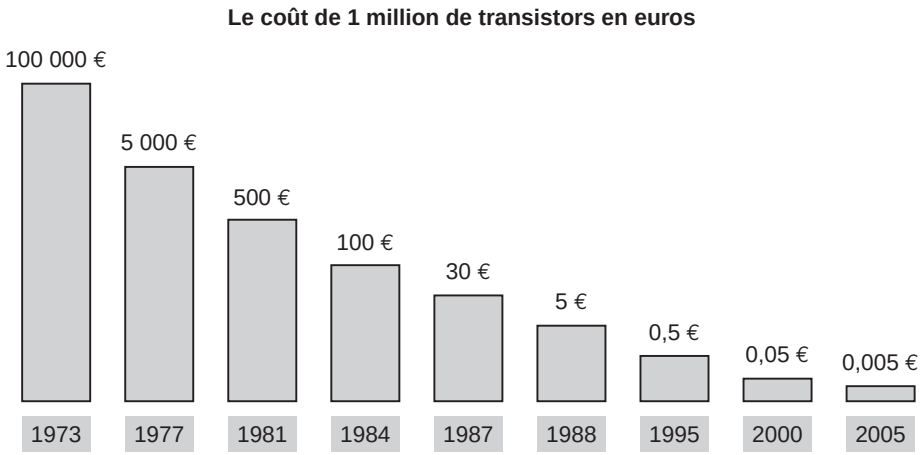


Figure 1.3 – Diminution du coût d'un transistor.

Cette décroissance exponentielle du coût de production est appelée loi de Moore.

1.2 La miniaturisation et la loi de Moore

La diminution du coût d'un transistor ou d'une fonction est donc basée sur la réduction de taille du transistor. Gordon Moore, ingénieur chez Intel, a énoncé cette évolution de la manière suivante : le nombre de transistors intégrés sur une puce double tous les 18 mois. Cette déclaration n'est pas une loi mais une simple observation et rien ne peut réellement expliquer la période de 18 mois. La *figure 1.4* montre la différence entre la règle de Moore et ce qui s'est véritablement passé dans l'industrie micro-électronique.

La loi de Moore s'est finalement appliquée avec une constante de temps plus courte que celle imaginée dans les prévisions initiales. Le nœud de la technologie λ est défini comme la demi-distance la plus petite entre deux lignes conductrices.

La *figure 1.5* représente un transistor dans un circuit intégré.

Le transistor MOS est formé de deux zones conductrices appelées source et drain. Une électrode appelée grille est placée au-dessus du dispositif et contrôle le courant qui circule de la source vers le drain. La distance L entre source et drain est très faible et voisine du nœud de la technologie. Elle est appelée longueur du canal. La largeur du dispositif W est petite mais supérieure à la longueur, de deux à quelques centaines de fois λ. Elle est choisie pour procurer au transistor des propriétés électriques satisfaisantes et est donc plus élevée que la valeur minimale autorisée par la technologie.

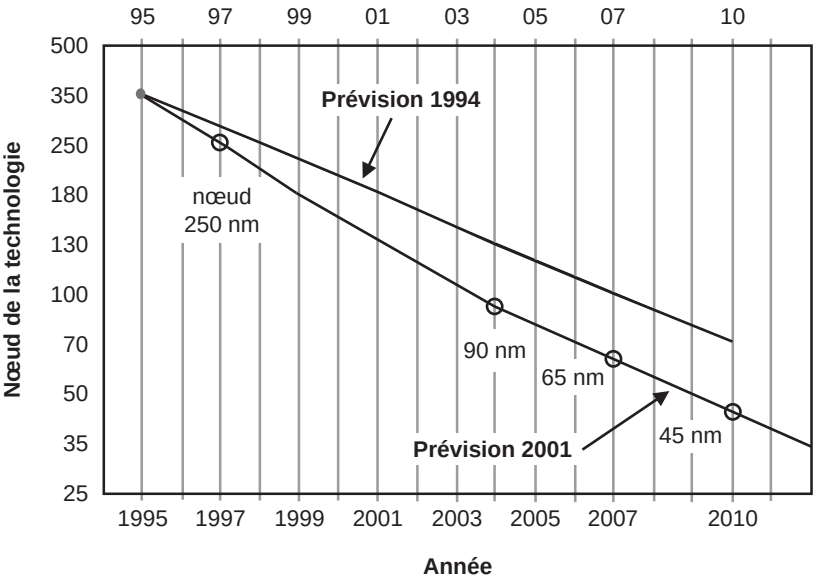


Figure 1.4 – La loi de Moore.

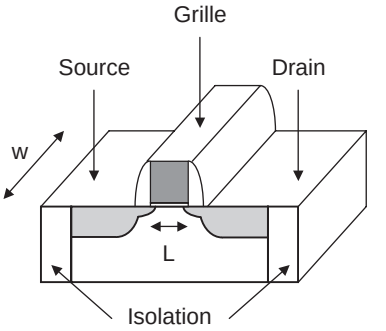


Figure 1.5 – Le transistor MOS.

Pour réaliser des fonctions numériques, une valeur de 2 λ peut suffire. Pour réaliser des fonctions analogiques, la largeur doit être augmentée. La dimension λ est liée aux possibilités de la lithographie, c'est-à-dire la technique optique qui permet de fabriquer les transistors comme il sera expliqué dans le chapitre 6.

De nombreux transistors de longueur minimale sont réalisés sur le wafer. La largeur du transistor peut être choisie selon la fonction à réaliser. Il est également possible de fabriquer des transistors plus longs que le transistor minimum pour réaliser des fonctions de type analogique. Les transistors sont ensuite reliés entre eux pour former un circuit donné. La largeur minimale d'une connexion est également voisine du nœud de la technologie. Le facteur de réduction doit aussi s'appliquer aux interconnexions car sinon la taille de la puce ne serait pas réduite proportionnellement.

On peut s'interroger sur la pérennité de la loi de Moore. Les difficultés techniques augmentent en effet au fur et à mesure que les tailles diminuent. Des difficultés de rendement de fabrication apparaissent par exemple à partir du nœud 90 nm. Cependant, jusqu'au nœud 90 nm, la loi de Moore a toujours été appliquée et tout laisse à penser que les difficultés seront surmontées jusqu'au nœud 22 nm.

Pour terminer cette courte introduction, il faut évoquer le coût des équipements de l'industrie micro-électronique et le coût des masques fabriqués à chaque fois que l'on désire produire un circuit nouveau. Le coût des équipements est principalement le coût des machines d'insolation, de dépôt et de gravure qui permettent la fabrication collective. Toutes ces opérations seront détaillées dans le chapitre 6. Il faut y ajouter le coût lié à l'ultra propreté. Une poussière de 1 micron de diamètre est plus large que la partie active du transistor et plus large qu'une connexion. Les poussières doivent donc être éliminées ce qui explique la nécessité de réaliser la fabrication dans des salles sans poussière appelées salles blanches. En fait, il y a toujours des poussières résiduelles que les dispositifs de filtrage de l'air ne peuvent éliminer et on définit plusieurs classes de propreté en fonction du niveau recherché. En définitive, le coût total des installations est très élevé ce qui explique des regroupements industriels et introduit une notion de taille critique dans l'industrie micro-électronique.

Pour fabriquer un circuit intégré, il faut réaliser un jeu de masques particulier. Les masques sont des plaques de quartz sur lesquelles sont gravés avec une précision extrême les motifs servant à la réalisation du circuit. La résolution de gravure de ces motifs est inférieure au micron ce qui explique les coûts très élevés. La *figure 1.6* montre comment le coût d'un jeu de masques augmente au fur et à mesure que la technologie s'affine.

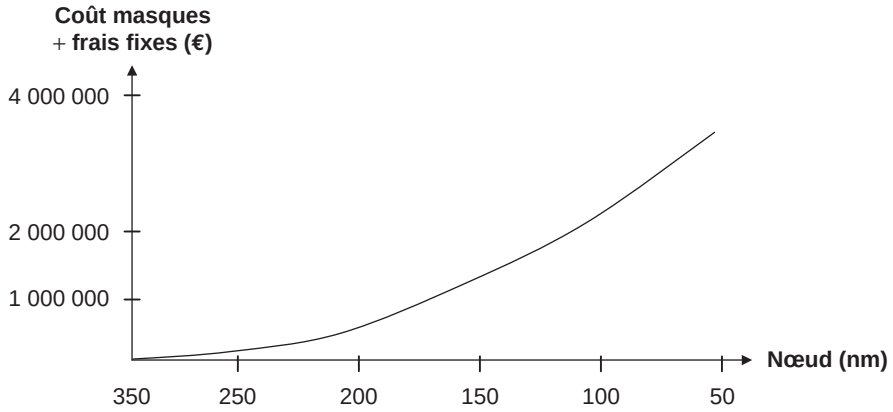


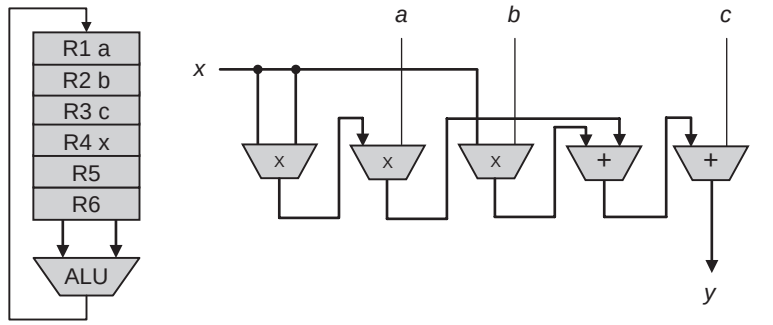
Figure 1.6 – Le coût de fabrication d'un circuit.

Cette courbe met en relief un véritable problème de fond pour les futures générations de circuits. Le coût représenté sur la *figure 1.6* est à répartir sur les circuits réalisés. Dans une technologie 50 nm par exemple, si 100 000 circuits sont fondus, le coût par circuit des frais fixes et du jeu de masques est alors de 40 € ce qui est considérable pour un circuit intégré. Les nouvelles technologies micro-électroniques sont donc économiquement rentables pour des quantités importantes de circuits identiques, en pratique plusieurs millions. Cela a des conséquences très importantes pour l'architecture des circuits comme il sera vu dans le chapitre 11.

1.3 La convergence du matériel et du logiciel

L'évolution de l'électronique numérique amène à réaliser des fonctions de type microprocesseur dans la majorité des circuits intégrés. Une fonction numérique pourra donc se réaliser sur un circuit intégré de deux manières : par un logiciel ou par un composant matériel spécifique. Pour illustrer cette évolution majeure de l'électronique intégrée prenons l'exemple simple du calcul de la fonction du deuxième degré.

$$y = ax^2 + bx + c$$



Mul R4 R4 R5
Mul R4 R2 R6
Add R6 R3 R6
Mul R5 R1 R5
Add R5 R6 R5

Solution logicielle **Solution matérielle**

Figure 1.7 – Équivalence matérielle-logicielle.

Dans la solution logicielle, les diverses opérations sont faites par une unité arithmétique et logique (ALU) qui modifie le contenu des registres R1 à R6. L'opération (Mul R4 R4 R5) veut dire que le contenu de R4 est multiplié par le contenu de R4 et que le résultat est rangé dans R5. On définit de même une opération addition. Les calculs s'effectuent séquentiellement en utilisant les mêmes blocs matériels. Ils sont concentrés dans l'espace et répartis dans le temps.

Dans la solution matérielle, les opérations sont effectuées en parallèle pour la plupart par des opérateurs addition et multiplication. Les calculs sont répartis dans l'espace et concentrés dans le temps. Cette solution est plus rapide mais consomme plus de silicium.

D'un point de vue logique, les deux techniques sont équivalentes et on peut changer du logiciel en matériel et du matériel en logiciel. Ce principe est à la base de la conception des systèmes intégrés complexes et est appelé co-design.

Si on envisage une application complexe, par exemple la mise en œuvre de protocoles de communication ou une série de traitements d'images, la première chose à faire est de décomposer l'ensemble de l'application en tâches élémentaires et de choisir quelles tâches seront exécutées sous forme d'un programme se déroulant sur un processeur et quelles tâches seront exécutées par des blocs spéci-

fiques. Il faut ensuite programmer les différentes tâches logicielles, réaliser les blocs logiques pour les tâches de type matériel et vérifier la cohérence de l'ensemble.

Pour pouvoir éventuellement traiter différentes applications avec le même circuit, il est intéressant d'identifier dans les tâches logicielles celles qui ne dépendent pas de l'application mais des blocs matériels du circuit. Ces briques logicielles sont appelées HDS, *Hardware Dependent Software*. Le but de la méthodologie de conception des systèmes complexes est d'automatiser au maximum les opérations. C'est la seule manière pour augmenter la productivité du design et pour réduire le nombre d'erreurs. Le sujet est complexe et donne encore lieu à une intense activité de recherche.

1.4 Miniaturisation du transistor et composants alternatifs

Deux questions se posent pour l'avenir de la micro-électronique :

- Quel est le plus petit transistor possible ?
- Le transistor sera-t-il remplacé par d'autres composants ?

Ces deux questions seront traitées en détail dans les chapitres 10 et 12 mais il est possible de donner dans cette introduction quelques idées générales sur le sujet.

Le fait de réduire les dimensions du transistor en dessous du micron amène à réduire la charge électrique Q dans le canal. Cette charge mesure l'état du dispositif. Quand elle est nulle, le transistor est non passant, et quand elle est non nulle, il est conducteur. Le transistor de capacité électrique C ne pourra fonctionner à une température T donnée que si l'énergie associée à cette charge soit Q^2/C est supérieure aux fluctuations thermiques qui sont de l'ordre de k_BT. Elle doit même être très supérieure car le choix entre les deux états possibles doit pouvoir se faire avec une probabilité d'erreur infiniment faible. La réduction de la taille du transistor conduit à une diminution de la charge de conduction et donc à une diminution de l'énergie associée.

Pensons à un circuit intégré comportant des millions de transistors commutant à fréquence élevée. La réduction de la taille du transistor va donc entraîner l'apparition d'erreurs logiques sauf si les dispositifs travaillent à très basse température. Le refroidissement des dispositifs est très difficilement envisageable, il faut donc se tourner vers des solutions basées sur de nouvelles architectures tolérantes aux fautes. De nombreux travaux ont été menés dans ce domaine sans toutefois aboutir à des solutions réellement applicables. Il ne faut pas trop augmenter le nombre de transistors et donc la surface de silicium.

La réduction de la taille du transistor conduit également à des difficultés dans les procédés de fabrication en particulier dans l'usage des techniques lithographiques. Le transistor est fabriqué à partir d'un masque fait à une échelle élevée. Une optique de qualité permet d'effectuer cette réduction. Cette méthode est dite *top-down* et se complexifie au fur et à mesure que les dimensions se réduisent. Une autre méthode serait de fabriquer par voie chimique des dispositifs ayant les dimensions souhaitées. Cette méthode est dite *bottom-up*. Il faut alors résoudre deux problèmes :

- fonctionnaliser chaque dispositif élémentaire, c'est-à-dire lui donner une propriété électrique de conduction ou de non conduction ;
- faire passer l'information d'un dispositif à un autre pour assurer une fonction globale.

Quelques dispositifs nouveaux sont étudiés dans les laboratoires, comme les nanotubes de carbone ou les nanofils de semi-conducteur. Des molécules présentant des propriétés électriques particulières sont également des candidats possibles. Enfin, l'informatique quantique se propose de traiter l'information au niveau même des atomes individuels. Ces thèmes seront étudiés plus en détail dans les chapitres 11 et 12 de cet ouvrage.

Chapitre 2

Les principes physiques de base

Le but de ce chapitre est de rappeler les principes physiques indispensables à la compréhension du fonctionnement des composants de la micro-électronique. Ces notions de base sont issues de deux domaines principaux : l'électromagnétisme et la physique des solides. La physique statistique est parfois nécessaire pour expliquer le comportement d'un nombre fini d'éléments.

Dans la micro-électronique conventionnelle, le fonctionnement des composants s'explique principalement par la loi de Coulomb à la condition de savoir exprimer les densités d'électrons et de trous à partir de règles issues de la physique des solides. On en arrive alors à écrire un modèle électrique du composant exprimant la relation entre courant et tension. Quand plusieurs composants sont interconnectés, il suffit d'écrire les lois de Kirchhoff pour comprendre le fonctionnement de l'ensemble.

Passons maintenant aux nanodispositifs. La notion de courant doit être revue, ce qui conduit à oublier la loi d'Ohm pour expliquer la conduction et à prendre en compte les propriétés de l'onde associée à un électron. Il ne doit plus être considéré comme une simple particule au sens de la mécanique de Newton. Les effets tunnel sont les plus caractéristiques de cette importante différence.

Quelques éléments seront également donnés pour expliquer les méthodes de calcul des circuits électroniques en statique et en dynamique.

2.1 Les forces électriques et le potentiel

Le fonctionnement des composants électroniques est fondamentalement l'étude du transport des électrons dans les dispositifs. La notion de trou sera introduite ultérieurement mais elle est une simple commodité de représentation. La conduction est toujours assurée par le déplacement d'électrons. La force à prendre en compte est la force électrique entre un électron en mouvement et les autres charges du dispositif à savoir les noyaux des atomes chargés et les autres électrons. Cette force est la force de Coulomb.

2.1.1 *Champ électrique et potentiel*

Si on considère dans le vide une particule fixe ayant une charge donnée q au milieu d'un ensemble de particules de charges Q_i et situées à la distance r_i de la particule considérée, la force qui s'exerce sur cette particule est donnée par la formule de Coulomb.

$$\mathbf{F} = q \cdot \sum_i \frac{Q_i}{4\pi\varepsilon_0} \frac{1}{r_i^2} \mathbf{u_i} \tag{2.1}$$

La *figure 2.1* illustre la formule et représente les vecteurs unitaires $\mathbf{u_i}$ qui indiquent la direction de la force associée à chaque particule chargée. La composante est attractive quand les charges q et Q_i sont de signes opposés et répulsive quand elles sont de même signe. La constante ε_0 est la permittivité du vide et égale à $8{,}85 \times 10^{-12}$ dans le système MKSA. Dans un autre milieu, elle a une valeur plus élevée égale à $\varepsilon_r \varepsilon_0$. Pour le silicium, la valeur de ε_r est égale à 11,7.

Quand le milieu considéré n'est pas le vide, il suffit de faire la différence entre les charges Q_i étrangères au milieu considéré et les charges des atomes du milieu lui même pour en arriver à cette notion de permittivité relative. Cette représentation est beaucoup plus simple que de pendre en compte l'effet de toutes les charges. Elle sera expliquée dans le paragraphe 2.1.4.

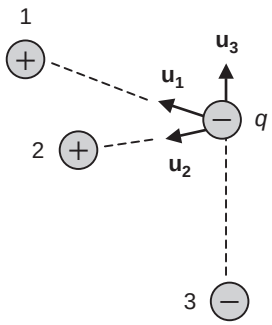


Figure 2.1 – La force de Coulomb.

Pour calculer la force qui s'exerce sur une charge, il est très commode d'introduire le champ électrique **E** et le potentiel électrique. Le champ électrique est simplement la force qui s'exerce sur une charge unitaire. La force qui s'exerce sur une charge quelconque de valeur q est donc :

$$\mathbf{F} = q\,\mathbf{E}$$

La forme particulière de la force de Coulomb est telle qu'il est possible pour une répartition donnée de charges Q_i de définir en tout point de l'espace une fonction $V(x, y, z)$ permettant de calculer simplement le champ électrique. Cette fonction est le potentiel électrique. La relation entre champ et potentiel est en tout point :

$$E_x = -\frac{\partial V}{\partial x}$$

$$E_y = -\frac{\partial V}{\partial y}$$

$$E_z = -\frac{\partial V}{\partial z}$$

On écrit en notation compacte :

$$\mathbf{E} = -\mathbf{grad}\, V$$

Remarquons que cette définition du potentiel n'est possible qu'à cause de la forme mathématique particulière de la force électrique. On dit que la force dérive d'un potentiel et on prouve assez facilement que dans ce cas le travail de la force le long d'un chemin ne dépend que des points de départ et d'arrivée et non pas du chemin choisi. On montre également que le potentiel en un point et créé par un ensemble de charges s'exprime simplement par la relation suivante :

$$V = \frac{1}{4\pi\varepsilon}\sum_i \frac{Q_i}{r_i} \tag{2.2}$$

Dans cette relation, r_i est la distance entre la charge i et le point considéré. Quand les charges sont en grand nombre, on passe de la somme à une intégrale en considérant cette fois la densité de charge ρ par unité de volume :

$$V(x, y, z) = \frac{1}{4\pi\varepsilon}\int \frac{\rho(x', y', z')}{r}\, \mathrm{d}x'\, \mathrm{d}y'\, \mathrm{d}z'$$

Dans cette relation, la distance r s'exprime par :

$$r = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}$$

Si on représente les points dans l'espace correspondant à une même valeur du potentiel, on obtient une surface dite équipotentielle. Cette surface est une sphère quand il y a une charge ponctuelle unique dans l'espace. Elle est de forme plus complexe quand il y a une distribution de charge quelconque. En général, on représente les surfaces pour des valeurs réparties uniformément. On obtient alors les équipotentielles du problème. On peut monter que par définition même du potentiel, le champ électrique est normal aux surfaces équipotentielles. Cette démonstration est laissée au lecteur. Le plus simple est de raisonner par l'absurde en supposant que le champ a une valeur non nulle en plus de sa composante normale et en exprimant que la variation du potentiel est nulle quand on se déplace sur la surface équipotentielle.

2.1.2 *Le théorème de Gauss*

Il est maintenant possible de rappeler un des théorèmes les plus importants de l'électromagnétisme, à savoir le théorème de Gauss. Ce théorème établit la relation entre le flux du champ électrique sur une surface fermée et la charge contenue dans la surface. La *figure 2.2* représente une surface quelconque et un cas particulier, celui d'une sphère comportant une charge en son centre. Le flux

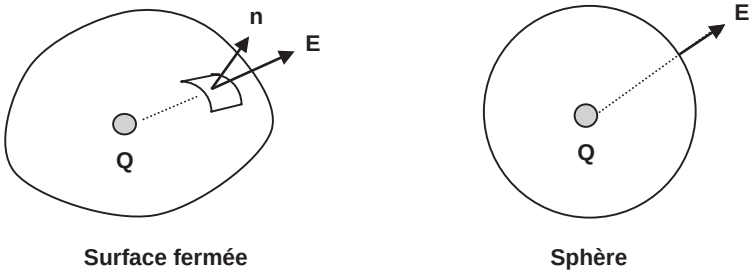


Figure 2.2 – Le théorème de Gauss.

d'un vecteur **E** le long d'une surface est l'intégrale sur toute la surface du produit scalaire de ce vecteur par le vecteur unitaire **n** normal à la surface. Il est noté :

$$\Phi = \int_S \mathbf{E} \cdot \mathbf{n}\, ds$$

Si on considère dans une première étape un volume V délimité par une surface fermée S et comportant une seule charge Q ponctuelle à l'intérieur, il est assez facile de montrer avec la valeur particulière du champ électrique que :

$$\Phi = \int_S \mathbf{E} \cdot \mathbf{n}\, ds = \frac{Q}{\varepsilon_0}$$

Le cas particulier de la sphère avec une charge en son centre permet une vérification très simple de cette formule. Il est également facile de montrer que si la charge est à l'extérieur du volume le flux est nul, et que finalement le flux est relié à la somme des charges présentes dans le volume par :

$$\Phi = \int_S \mathbf{E} \cdot \mathbf{n}\, ds = \frac{\sum_i Q_i}{\varepsilon_0}$$

Quand le milieu n'est plus le vide mais un milieu de permittivité donnée, on généralise simplement la formule.

$$\Phi = \int_S \mathbf{E} \cdot \mathbf{n}\, ds = \frac{1}{\varepsilon} \sum_i Q_i$$

La formule se généralise à une distribution continue de charge et,

$$\Phi = \int_S \mathbf{E} \cdot \mathbf{n}\, ds = \frac{1}{\varepsilon} \int_V \rho\, dx\, dy\, dz \tag{2.3}$$

Quand le volume est petit, on obtient une formule limite qui vient du fait que le rapport entre le flux et le volume associé à la surface considérée tend vers une limite finie appelée la divergence du vecteur.

$$\mathrm{div}\mathbf{E} = \lim_{V \to 0} \frac{\int_S \mathbf{E} \cdot \mathbf{n}\, ds}{V}$$

On montre que la divergence d'un vecteur se calcule par :

$$\text{div}\mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}$$

On écrit donc :

$$\text{div}\mathbf{E} = \frac{\rho}{\varepsilon} \qquad \textbf{(2.4)}$$

Cette relation locale qui est une conséquence de la loi de Coulomb permet de relier le champ et la densité locale de charge. Elle est connue comme une des équations de Maxwell et est la relation la plus utile à connaître pour expliquer le fonctionnement des dispositifs électroniques.

Elle s'exprime sous une autre forme en remplaçant le champ par son expression en fonction du potentiel. On obtient alors l'équation dite de Laplace.

$$\text{div}(-\textbf{grad}\,V) = \frac{\rho}{\varepsilon}$$

soit,

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} + \frac{\rho}{\varepsilon} = 0 \qquad (2.5)$$

Cette équation permet de calculer le potentiel électrique, la densité de charge étant connue. En pratique, il est nécessaire de prendre en compte des conditions dites aux limites qui sont généralement la donnée de valeurs particulières du potentiel en des endroits déterminés de l'espace appelés électrodes.

2.1.3 *Exemple du condensateur plan*

Un exemple est indiqué *figure 2.3*. C'est celui du condensateur plan. On considère deux électrodes conductrices idéales portées à deux potentiels 0 V et 10 V à l'aide de générateurs extérieurs et on se propose de calculer le potentiel et le champ électrique en tout point de l'espace. On prouve également que le potentiel est nul loin des électrodes. Cela est lié à la formule en $1/r$ du potentiel. Pour simplifier le problème du condensateur plan, on supposera qu'il est placé dans une boîte conductrice de grande taille et portée au potentiel 0 V.

La résolution de l'équation peut se faire numériquement mais il est assez facile de prévoir le résultat approché à partir de règles de bon sens. Les conducteurs étant supposés parfaits, ils sont donc équipotentiels. Pensons à la loi d'Ohm pour comprendre cette hypothèse. Les charges du système sont nécessairement situées à la surface des électrodes. Pour le montrer, on peut appliquer le théorème de Gauss. Quand les tensions sont appliquées sur les électrodes, les charges se déplacent puis se « bloquent » à la surface des électrodes.

Si on se place au milieu du dispositif, on comprend en tenant compte des propriétés de symétrie du système, que les équipotentielles sont distribuées régulièrement et qu'elles sont parallèles aux électrodes. La *figure 2.3* représente les équipotentielles 5 V et 8 V comme exemples. À l'extérieur du dispositif, on peut de même prévoir la position des équipotentielles en intégrant les symétries du système. Si le problème se limitait à calculer le potentiel entre deux plaques conductrices infinies, il serait facile de conclure que les équipotentielles sont des plans parallèles aux électrodes et réparties uniformément. Dans le cas plus concret représenté *figure 2.3*, on considère les deux autres conden-

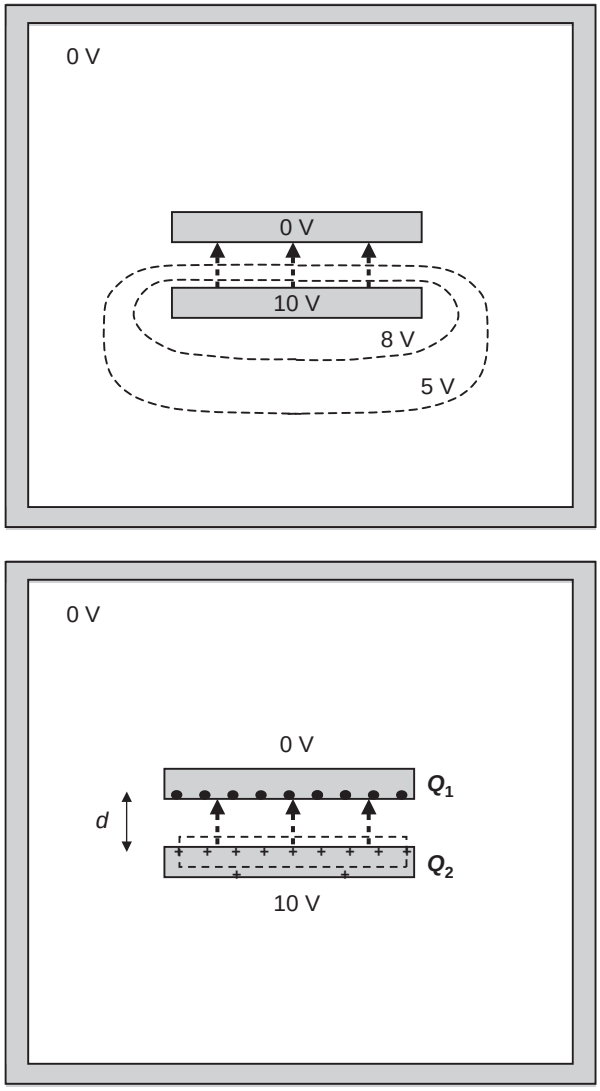


Figure 2.3 – Exemple de résolution de l'équation de Laplace.

sateurs formés avec les parois de la boîte extérieure et on raisonne par continuité pour les régions intermédiaires.

Il est maintenant possible de relier les charges formées sur les surfaces des électrodes et les potentiels. Appliquons pour cela le théorème de Gauss à la surface fermée indiquée sur la figure. On suppose que le champ est constant le long de la surface. Les considérations de symétries permettent de le comprendre. Le champ est perpendiculaire aux électrodes. En effet, par définition même du potentiel, le champ électrique est toujours perpendiculaire aux surfaces équipotentielles. Imaginons le système en trois dimensions, les électrodes ont une surface A. Le champ étant nul dans l'électrode, on obtient :

$$E\,A = \frac{Q_2}{\varepsilon}$$

Cette relation s'écrit également :

$$E = \frac{\sigma}{\varepsilon}$$

Elle relie champ et densité de charge σ par unité de surface au niveau d'une électrode métallique. Cette formule établie sur un exemple est en fait généralisable.

Le champ étant constant et d étant la distance entre les électrodes, on peut écrire par définition du potentiel :

$$V_2 - V_1 = E \cdot d$$

On en déduit donc :

$$Q_2 = \varepsilon \frac{A}{d} (V_2 - V_1) \tag{2.6}$$

Dans l'exemple, les potentiels sont 10 V et 0 V. Cette relation est la formule du condensateur plan. Il est facile de montrer par un calcul du même type mais dans l'autre électrode que la charge Q_1 s'exprime par :

$$Q_1 = -\varepsilon \frac{A}{d} (V_2 - V_1)$$

Les deux charges sont égales et de signes opposés. On en déduit facilement, toujours en appliquant le théorème de Gauss, que la charge située à la surface externe de l'électrode portée à 10 V est nulle. En réalité, elle est de très faible valeur quand on résout le problème sans approximation, c'est-à-dire quand on abandonne l'hypothèse d'une boîte de très grande taille.

2.1.4 *Les champs électriques dans la matière*

Il faut revenir un peu sur la permittivité diélectrique relative introduite de manière assez brutale dans le premier paragraphe. Prenons l'exemple d'un condensateur plan. Si il y a le vide entre les armatures, la relation entre la charge stockée sur une armature et la tension aux bornes s'écrit :

$$Q_2 = \varepsilon_0 \frac{A}{d} (V_2 - V_1)$$

On dit que le condensateur a une capacité C donnée par :

$$C = \varepsilon_0 \frac{A}{d}$$

Si le milieu entre les électrodes n'est plus le vide mais un matériau solide ou liquide, la capacité devient :

$$C = \varepsilon \frac{A}{d}$$

Elle est plus élevée car la permittivité est plus élevée que celle du vide. Pour une même tension, le condensateur stocke plus de charge.

Pour comprendre l'origine de ce phénomène, on suppose que le milieu est constitué d'atomes ou de molécules entre les armatures. On suppose qu'il peut se décrire par un ensemble de « dipôles »,

c'est-à-dire d'objets présentant une répartition de charge non symétrique dans l'espace quand un champ est appliqué. La molécule ou l'atome présente une distribution de charge variant de manière continue dans son volume, négative par endroits et positive ailleurs.

En l'absence de champ électrique, cette distribution peut être nulle en moyenne. Dans ce cas, les molécules sont dites non polaires. La distribution peut être non nulle et montrer une région négative et une région positive. Les molécules sont alors polaires. Quand on applique un champ électrique, les électrons des atomes et des molécules se déplacent et la dissymétrie de la densité de charge augmente. Elle augmente quand les molécules sont polaires et elle apparaît quand les molécules sont non polaires. On dit que la matière se polarise.

La *figure 2.4* représente le condensateur non polarisé, le condensateur polarisé et le condensateur polarisé équivalent modélisé par sa constante diélectrique équivalente. Les molécules du milieu diélectrique sont représentées par des « dipôles » rigides présentant une dissymétrie de charge. Dans le cas des atomes, il faut imaginer une déformation des orbitales atomiques.

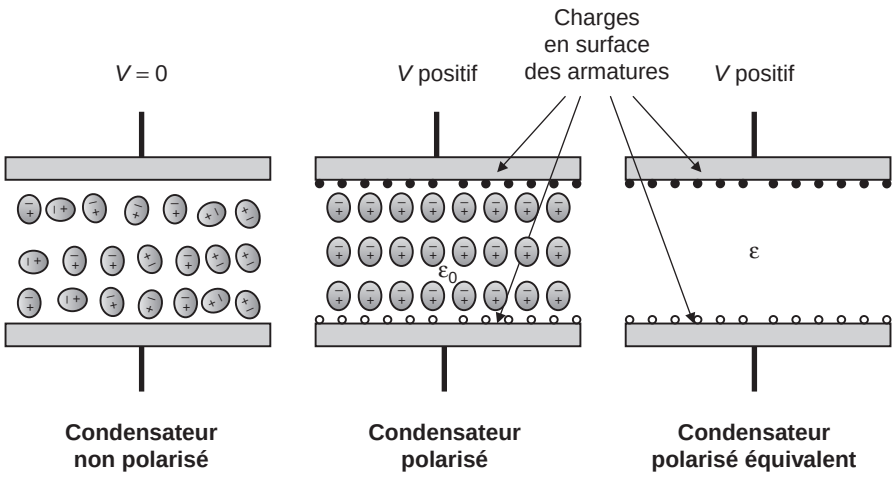


Figure 2.4 – Polarisation de la matière.

Sans champ appliqué, les molécules ou les atomes s'orientent de manière aléatoire. Quand un champ est appliqué les molécules et les atomes s'orientent, les parties négatives étant attirées par l'armature positive. La dissymétrie augmente avec le champ. Si on considère que la molécule est équivalente à deux charges opposées de même valeur Q et séparées d'une distance a, on peut définir le moment du dipôle p par le produit $Q \cdot a$, et on admet que cette valeur est proportionnelle au champ électrique appliqué.

$$\mathbf{p} = \alpha \mathbf{E}$$

Le coefficient α est la polarisabilité de la molécule ou de l'atome. Il vaut environ $10^{-40} C\ m^2/V$ pour les atomes usuels. Les moments dipolaires permanents quand ils existent sont en général beaucoup plus élevés que les moments induits. Dans ce cas, le champ électrique appliqué n'augmente pas le moment de chaque molécule ou atome mais oriente les « dipôles » et crée donc un moment moyen important.

Revenons maintenant au condensateur donné en exemple. On peut considérer que le champ à l'intérieur résulte de la superposition de deux systèmes de charges. Le premier est le condensateur dans le vide. Le second est la tranche de diélectrique placée entre les deux électrodes flottantes. Pour comprendre ce qui se passe, on peut imaginer la séquence suivante. Dans un premier temps, on charge dans le vide le condensateur formé par les deux plaques en appliquant une tension V_0, ce qui a pour effet de créer des charges Q_0 et $-Q_0$ sur les plaques. Dans un deuxième temps, on isole les plaques puis on intercale la tranche de diélectrique. La charge stockée n'a pas changé mais la différence de potentiel entre les plaques se modifie. Le champ résultant est la somme du champ créé par les charges Q_0 et le champ créé par les dipôles du diélectrique.

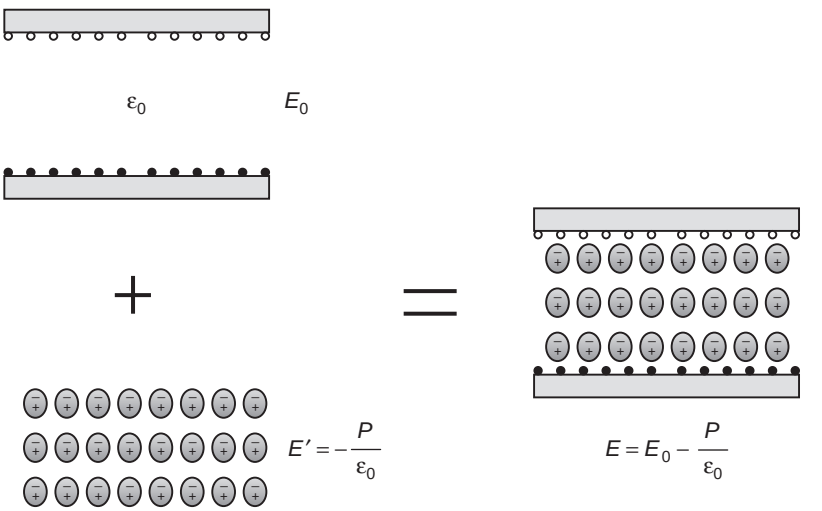


Figure 2.5 – Permittivité relative.

La valeur du champ créé par les dipôles n'est pas tout à fait évidente. Pour trouver cette valeur du champ, il faut considérer la densité de moments dipolaires par unité de volume. C'est le produit du nombre de dipôles par unité de volume N multiplié par le moment dipolaire $\mathbf{p}$ d'un dipôle. Elle est notée $\mathbf{P}$. La densité de charge par unité de surface au niveau des électrodes est alors P. Pour trouver cette relation simple, il faut considérer une colonne de dipôles entre les deux armatures et tenir compte du fait que les charges des dipôles en série s'annulent deux à deux. Le champ est donc $-P/\varepsilon_0$. Le signe moins s'explique facilement en tenant compte des signes des charges. Au total, le champ dans le dispositif est donc :

$$\mathbf{E} = \mathbf{E}_0 - \frac{\mathbf{P}}{\varepsilon_0}$$

On en déduit le calcul de la capacité du condensateur contenant la tranche de diélectrique :

$$C = \frac{C_0 V_0}{V} = \frac{C_0 V_0}{\left(E_0 - \frac{P}{\varepsilon_0}\right) d} = C_0 \frac{E_0}{E_0 - \frac{P}{\varepsilon_0}}$$

En exprimant cette valeur en fonction du champ dans le dispositif.

$$C = C_0 \frac{E + \frac{P}{\varepsilon_0}}{E} = C_0\left(1 + \frac{P}{\varepsilon_0 E}\right)$$

La capacité a donc augmenté de valeur par rapport à celle du condensateur dans le vide et tout se passe comme si la permittivité s'était modifiée pour prendre la valeur :

$$\varepsilon = \varepsilon_0(1 + \chi_r)$$

avec,

$$\chi_r = \frac{P}{E} = N\alpha$$

Ce résultat, établi sur l'exemple du condensateur plan, a une valeur générale et justifie l'utilisation de la permittivité relative dans les matériaux.

2.1.5 *Introduction de l'induction magnétique*

Pour terminer ces rappels importants, il reste à commenter le domaine de validité de ces diverses lois dérivant de la formule de Coulomb. On peut par exemple se demander si ces lois établies pour des charges fixes restent valables quand les charges sont en mouvement.

Une hypothèse forte est de supposer que la charge est invariante par transformation relativiste. Cela veut dire que dans un repère en mouvement uniforme à vitesse **v** par rapport au repère initial, la mesure de la charge ne change pas. Pour mesurer le champ électrique créé par une charge Q en mouvement dans un repère donné, on place une charge d'essai q fixe dans le repère choisi et on mesure la force qui s'exerce sur cette charge. L'angle formé entre la droite reliant les deux charges et la vitesse **v** a une influence sur la valeur du champ mesuré.

Il est donc plus naturel pour mesurer la charge Q de l'entourer par une surface sphérique et de calculer l'intégrale du champ. Le théorème de Gauss, supposé valable, permet de calculer la charge. Nous allons donc faire deux hypothèses fondamentales :

- le théorème de Gauss reste vrai pour des charges en mouvement ;
- la charge a la même valeur que dans le régime statique.

Cette dernière hypothèse est très forte. Rappelons que les distances et la masse ne sont pas invariantes quand on change de repère. Ce résultat se généralise : la charge totale d'un ensemble de particules qui se déplacent ne dépend pas de leurs mouvements.

Le champ électrique contrairement à la charge dépend du repère choisi. Revenons à l'exemple du condensateur plan. Il est supposé fixe dans le repère R ainsi que les charges sur les armatures.

Dans le repère fixe par rapport aux charges, on calcule le champ par le théorème de Gauss et on trouve le champ en fonction de la densité de charge σ par unité de surface.

$$E_z = \frac{\sigma}{\varepsilon_0}$$

Dans le repère R' en translation uniforme selon l'axe des x de la droite vers la gauche et à la vitesse v par rapport au repère R, l'expression est différente. En effet, si on prend les mêmes particules chargées donc les mêmes charges en fonction du principe d'invariance de la charge, elles occupent

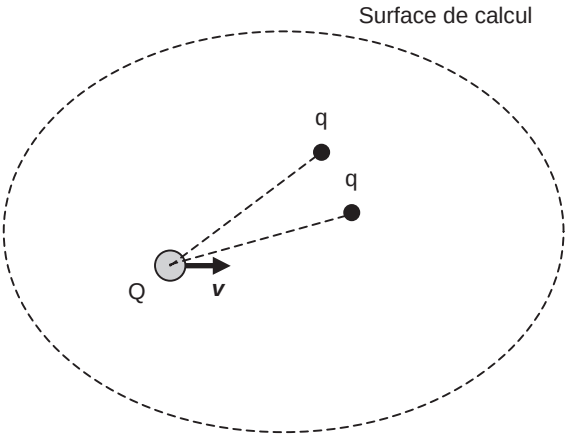


Figure 2.6 – Mesure de la charge.

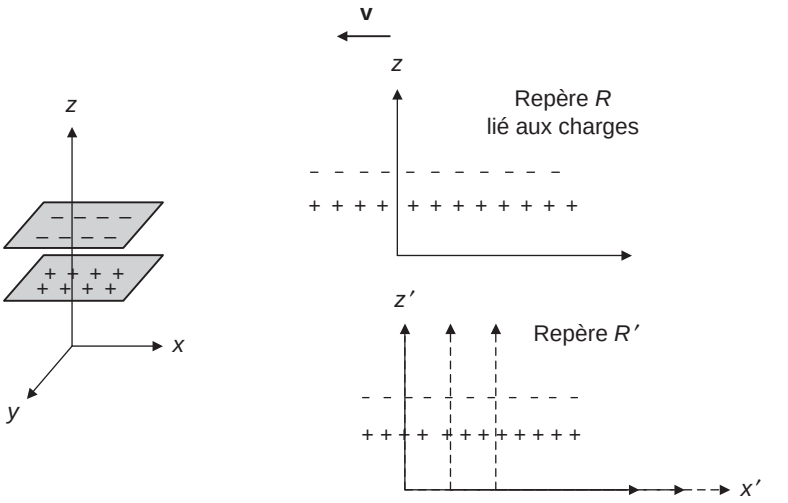


Figure 2.7 – Le champ dans un repère en mouvement.

une surface différente vue de *R′* en raison du principe de contraction des longueurs. En fait, il n'y a que la dimension selon *x* qui se contracte. On obtient alors :

$$E_z' = \frac{\sigma}{\varepsilon_0}\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$$

Ce simple calcul montre que la loi de Coulomb ne s'applique plus dans le repère *R′* dans lequel les charges sont mobiles. Dans cette expression *c* est la vitesse de la lumière dans le vide. La grandeur *c* est une constante universelle.

Nous allons maintenant introduire une autre grandeur nécessaire pour calculer la force qui s'applique sur une charge en mouvement. Cette nouvelle grandeur est l'induction magnétique. Pour l'introduire, il est possible de prendre un exemple simple comme il est fait dans l'ouvrage figurant en référence **[1]** dans la bibliographie.

On considère deux fils chargés avec des charges positives et négatives de mêmes densités λ par unité de longueur, qui défilent de gauche à droite et de droite à gauche à la vitesse v_0. La *figure 2.8* illustre cet exemple. Dans le repère R, les charges défilent avec des vitesses v_0 et $-v_0$ mais une charge d'essai fixe dans ce repère voit une densité de charge globale nulle et donc un champ électrique nul. On suppose maintenant que la charge d'essai se déplace à une vitesse v dans le repère R. Plaçons nous dans le repère R' lié à la charge d'essai en mouvement.

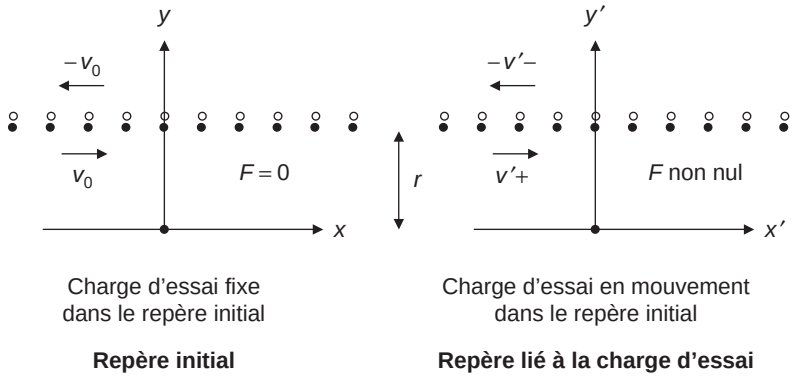


Figure 2.8 – Force sur une charge en mouvement.

Pour calculer l'effet relativiste de la vitesse sur les longueurs et donc sur les densités de charges, il faut calculer les vitesses dans R' des charges positives et négatives. Elles ne sont pas égales et en conséquence les densités de charge par unité de longueur sont différentes.

Les vitesses dans R' des charges positives et négatives sont respectivement :

$$v'_+ = \frac{v_0 - v}{1 - \frac{v_0 v}{c^2}} \qquad v'_- = \frac{v_0 + v}{1 + \frac{v_0 v}{c^2}}$$

On en déduit les densités linéaires de charge dans le repère R'.

$$\lambda'_+ = \frac{\lambda}{\sqrt{1 - \frac{v'^2_+}{c^2}}} \sqrt{1 - \frac{v_0^2}{c^2}} \qquad \lambda'_- = \frac{\lambda}{\sqrt{1 - \frac{v'^2_-}{c^2}}} \sqrt{1 - \frac{v_0^2}{c^2}}$$

La force qui s'exerce sur la charge d'essai dans le repère lié à cette charge se calcule alors à l'aide du théorème de Gauss.

$$F' = q \frac{\lambda'_+ - \lambda'_-}{2 \pi \varepsilon_0 r}$$

On trouve après un calcul un peu long :

$$F' = \frac{q \lambda v v_0}{\pi \varepsilon_0 r c^2} \frac{1}{\sqrt{1 - \frac{v_0^2}{c^2}}}$$

Dans cette expression valable dans le repère R', la densité λ et la vitesse v de la charge d'essai sont calculées dans le repère R. La densité linéaire de charge s'exprime dans le repère R en fonction du courant circulant dans le fil.

$$I = 2 \lambda v_0$$

On obtient donc dans le repère R en appliquant le facteur de correction relativiste.

$$F = \frac{q v I}{2 \pi \varepsilon_0 r c^2}$$

On constate l'effet total du mouvement exprimé par le produit du courant dans le fil par la vitesse de la charge d'essai. En résumé, la charge fixe est soumise à une force nulle tandis que la charge en mouvement est soumise à une force proportionnelle à sa vitesse et au courant dans le fil d'induction. Cet effet relativiste se manifeste pour une vitesse de la charge d'essai très inférieure à la vitesse de la lumière.

Les principes de base qui permettent d'établir ce résultat fondamental sont l'invariance de la charge, les équations relativistes dans les changements de repères et le théorème de Gauss. Le fait que des effets relativistes se manifestent dans l'électromagnétisme classique peut surprendre. Il faut cependant avoir en mémoire le principe de conservation de la charge dans la nature qui conduit à placer en général dans une unité de volume autant de charges positives que de charges négatives. La charge de ce volume vue à une distance importante est donc quasi-nulle ce qui veut dire que ce sont seulement les variations et en particulier les variations dues aux vitesses qui seront observées à distance. La formule précédente peut s'écrire également sous la forme suivante :

$$F = q v B$$

avec

$$B = \frac{I}{2 \pi \varepsilon_0 c^2} \frac{1}{r}$$

Il suffit de poser

$$\varepsilon_0 \mu_0 c^2 = 1$$

On retrouve alors la forme bien connue de l'induction magnétique due à un fil conducteur :

$$B = \frac{\mu_0}{2 \pi} \frac{I}{r}$$

Les constantes ont les valeurs suivantes :

$$\mu_0 = 4 \pi \cdot 10^{-7} \text{ MKSA} \qquad \varepsilon_0 = 8{,}854 \cdot 10^{-12} \text{ F} \cdot \text{m} \qquad c = 2{,}998 \cdot 10^8 \text{ m/s}$$

On peut généraliser assez facilement ce résultat et obtenir la forme la plus générale de la force qui s'exerce sur une charge en mouvement en présence d'autres charges en mouvement :

$$\mathbf{F} = q\,\mathbf{E} + q\,(\mathbf{v} \wedge \mathbf{B}) \tag{2.6}$$

Le champ électrique **E** est calculé comme la force s'exerçant sur une charge d'essai unitaire supposée fixe. Il se calcule à l'aide de la loi de Coulomb. L'induction magnétique **B** est introduite pour calculer l'autre force qui dépend de la vitesse **v**. Dans cette relation, le symbole $\wedge$ représente le produit vectoriel.

Il faut alors savoir calculer l'induction magnétique. Elle est liée aux charges en mouvement du système. Dans le cas simple d'un système de charges en mouvement réduit à un fil traversé par un courant d'intensité I dans le vide, le module de l'induction magnétique à la distance r du fil est donné par la relation :

$$B = \frac{1}{2\,\pi\,\varepsilon_0\,c^2}\,\frac{I}{r}$$

La direction est indiquée sur la *figure 2.9*. Le champ est dans le plan normal au fil et tangent au cercle indiqué sur la figure. Le sens du champ est relié au sens du courant dans le fil par la règle dite du tire-bouchon. Le tire-bouchon tourne dans le sens indiqué par le champ magnétique et progresse dans le sens du courant.

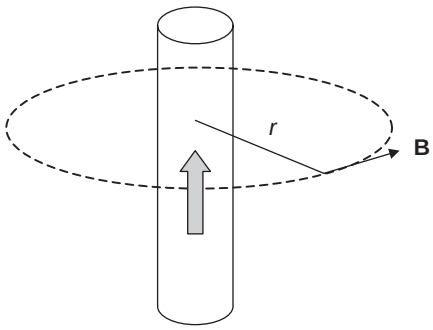


Figure 2.9 – Le champ magnétique créé par un fil.

Dans le cas plus général, l'induction magnétique se calcule à partir d'une autre grandeur **A** appelée potentiel vecteur, par la relation :

$$\mathbf{B} = \mathbf{rot}\,\mathbf{A} \tag{2.7}$$

Le potentiel vecteur se calcule à partir de toutes les densités de courant présentes $\mathbf{J}(x', y', z')$ par la relation :

$$\mathbf{A}(x, y, z) = \frac{\mu_0}{4\,\pi} \int \frac{\mathbf{J}(x', y', z')}{r}\,\mathrm{d}x'\,\mathrm{d}y'\,\mathrm{d}z' \tag{2.8}$$

Dans cette formule, r est la distance du point où on calcule le potentiel vecteur à la source de courant prise en compte. Rappelons que la densité de courant est le courant par unité de surface quand on considère une surface infinitésimale en un point donné de l'espace.

$$r = \sqrt{(x' - x)^2 + (y' - y)^2 + (z' - z)^2}$$

Le rotationnel est défini par :

$$(\mathbf{rot}\,\mathbf{A})_x = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}$$

$$(\mathbf{rot}\,\mathbf{A})_y = \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}$$

$$(\mathbf{rot}\,\mathbf{A})_z = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}$$

On montre assez facilement à partir de ces relations que div **B** est nulle, ce qui veut dire que le flux de l'induction magnétique à travers une surface fermée est nul. Quand le milieu n'est plus le vide, la permittivité μ_0 est remplacée par μ. L'introduction de la perméabilité relative est équivalente à celle de la permittivité relative. Ce point ne sera pas traité dans ce chapitre.

En définitive, champ électrique et induction magnétique sont étroitement liés et sont les manifestations d'un même objet physique, le tenseur électromagnétique. Considérons les composantes du champ et de l'induction dans un repère R donné puis les mêmes composantes dans un repère en translation uniforme v le long de l'axe des x. La vitesse v est mesurée dans R. On montre que les relations de transformation sont les suivantes :

$$E'_x = E \qquad B'_x = B_x$$

$$E'_y = \gamma(E_y - \beta\, c\, B_z) \qquad B'_y = \gamma\left(B_y + \beta\,\frac{E_z}{c}\right)$$

$$E'_z = \gamma(E_z + \beta\, c\, B_y) \qquad B'_z = \gamma\left(B_z - \beta\,\frac{E_y}{c}\right)$$

avec,

$$\beta = \frac{v}{c} \quad \text{et} \quad \gamma = \frac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}$$

Ces formules de transformation montrent la symétrie des quantités **E** et $c\mathbf{B}$. Elles peuvent se simplifier quand la vitesse v est très inférieure à c.

$$\mathbf{E}' = \mathbf{E} + \mathbf{v} \wedge \mathbf{B}$$

$$\mathbf{B}' = \mathbf{B} - \frac{\mathbf{v}}{c} \wedge \frac{\mathbf{E}}{c}$$

2.2 Courant et courant de déplacement

Le courant est une grandeur fondamentale dans l'étude des circuits électroniques qui peut sembler triviale dans une première approche. Deux notions plus complexes doivent cependant être expliquées en détail : le courant de déplacement et le courant au sens de la mécanique quantique.

Le courant électrique est le flux d'un ensemble de charges à travers une surface. La *figure 2.10* illustre cette définition et permet de calculer la quantité de charge qui traverse une surface infinitésimale donnée par unité de temps en fonction de la vitesse et de la densité de ces charges au niveau de la surface.

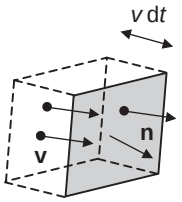


Figure 2.10 – Le courant à travers une surface.

On obtient facilement en considérant les charges qui se sont déplacées pendant un intervalle de temps dt :

$$\mathrm{d}Q = \rho\, \mathbf{v} \cdot \mathbf{n}\, \mathrm{d}t\, \mathrm{d}s$$

La densité de courant est définie par :

$$\mathbf{J} = \rho\, \mathbf{v} \tag{2.9}$$

Le courant traversant la surface A s'écrit donc :

$$I = \int_A \mathbf{J} \cdot \mathbf{n}\, \mathrm{d}s$$

Quand plusieurs types de charges sont à considérer en un même point, il y a plusieurs types de densités et plusieurs vitesses. Il suffit de sommer sur les différents types pour obtenir le courant total.

$$I = \sum_i \int_A \mathbf{J}_i \cdot \mathbf{n}\, \mathrm{d}s$$

Si on considère maintenant une surface fermée S entourant un petit élément de volume, le calcul du flux total peut se faire comme dans le cas du champ électrique :

$$\phi = \int_S \mathbf{J} \cdot \mathbf{n}\, \mathrm{d}s = \int_V \mathrm{div}\, \mathbf{J}\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z$$

Ce flux est la variation de la charge contenue dans le volume par unité de temps. Cette variation de charge dans l'élément de volume considéré s'écrit également en appliquant le principe de conservation de la charge :

$$\Delta Q = -\int_V \frac{\partial \rho}{\partial t}\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z$$

On écrit donc :

$$\mathrm{div}\, \mathbf{J} + \frac{\partial \rho}{\partial t} = 0$$

Exprimons maintenant la densité de charge en fonction du champ comme il a été vu précédemment.

$$\mathrm{div}\, \mathbf{E} = \frac{\rho}{\varepsilon}$$

alors,

$$\text{div } \mathbf{J} + \text{div } \varepsilon \frac{\partial \mathbf{E}}{\partial t} = 0$$

Le vecteur $\mathbf{J} + \varepsilon \frac{\partial \mathbf{E}}{\partial t}$ a donc une divergence nulle.

Dans ce cas, son flux à travers une surface fermée est nul, ce qui exprime la conservation de ce flux comme le montre la *figure 2.11*. Le terme dépendant de la dérivée du champ par rapport au temps est appelé courant de déplacement. Il faut donc bien avoir à l'esprit quand on parle de conservation du courant que ce qui se conserve n'est pas le courant dû au mouvement des charges mais la somme de ce courant et du courant de déplacement.

Illustrons cette propriété en examinant la conduction dans un fil cylindrique comme le montre la *figure 2.11*.

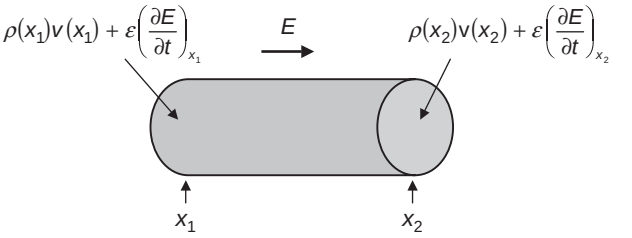


Figure 2.11 – La conservation du courant.

Par raison de symétrie, le champ est dirigé dans l'axe du cylindre. Les charges se déplacent parallèlement au cylindre. Le flux du courant total à travers le cylindre, courant de conduction + courant de déplacement, est non nul uniquement au niveau des deux faces du cylindre. On suppose aussi que la densité de courant et le champ électrique sont constants sur chaque face. La conservation du flux conduit alors à écrire :

$$\rho(x_1)v(x_1) + \varepsilon\left(\frac{\partial E}{\partial t}\right)_{x_1} = \rho(x_2)v(x_2) + \varepsilon\left(\frac{\partial E}{\partial t}\right)_{x_2}$$

Le courant en x_2 varie donc en fonction du courant en x_1 et des variations locales du champ électrique. Si on admet que le champ électrique se propage dans le fil à la vitesse $c/\sqrt{\varepsilon_r}$ on comprend la conservation du courant en régime dynamique. En régime établi ou continu, le courant de conduction se conserve le long du fil.

La conservation du courant ne s'explique pas par le déplacement des charges entre les deux points considérés. Si on applique la conservation du courant à un fil ayant une longueur d'un mètre, les charges en mouvement à l'une des extrémités n'atteindront jamais l'autre extrémité. Elles se recombineront dans le fil. Et pourtant, les courants aux deux extrémités sont égaux car le courant mesuré en un point est en quelque sorte informé par le champ de la valeur à l'autre extrémité. Le courant de déplacement, introduit par Maxwell, était initialement considéré comme un déplacement de charges réelles dans le vide appelé Ether. Aujourd'hui ce courant est considéré comme purement virtuel.

L'ensemble de ces résultats peut être résumé dans les quatre équations de Maxwell invariantes par transformation relativiste. Remarquons que seule la première de ces équations est vraiment nouvelle. La seconde est liée à la conservation du courant. La troisième est l'expression du théorème de Gauss. La quatrième a été introduite dans le paragraphe 2.1.5.

$$\begin{aligned} \mathbf{rot}\,\mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \mathbf{rot}\,\mathbf{B} &= \varepsilon\,\mu\,\frac{\partial \mathbf{E}}{\partial t} + \mu\,\mathbf{J} \\ \mathrm{div}\,\mathbf{E} &= \frac{\rho}{\varepsilon} \\ \mathrm{div}\,\mathbf{B} &= 0 \end{aligned} \qquad (2.10)$$

On y ajoutera les deux relations :

$$\mathrm{div}\,\mathbf{J} + \frac{\partial \rho}{\partial t} = 0$$

$$\varepsilon_0\,\mu_0\,c^2 = 1$$

2.3 La fonction d'onde et le courant à l'échelle des atomes

2.3.1 *Des particules au courant*

Si nous nous plaçons à l'échelle quantique, que devient alors la notion de courant ? L'objectif des paragraphes suivants est de montrer comment on passe de la description d'un ensemble de particules à la notion de courant. Le chemin est assez long pour y arriver mais il permet d'introduire toutes les notions physiques utiles à la compréhension de la micro-électronique.

La progression du raisonnement est illustrée *figure 2.12*.

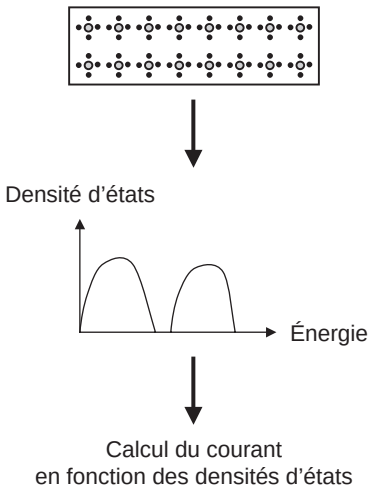


Figure 2.12 – De l'interaction entre particules au courant électrique.

2.3.2 *Représentation des grandeurs physiques en mécanique quantique*

Comme il est expliqué dans les ouvrages de mécanique quantique, un système physique de *N* particules ne se représente pas par un ensemble de *N* points matériels en mouvement dont on connaît les positions et les vitesses. Il se représente par une fonction d'état ψ dépendant des 3*N* coordonnées associées à ces *N* particules. Cette fonction est la fonction d'onde du système. C'est un nombre complexe caractérisé par son module et sa phase.

Plaçons-nous dans le cas d'une seule particule dans un monde à une dimension. La fonction d'onde dépend généralement du temps et est notée $\psi(x,t)$. Pour faire le lien entre la fonction d'onde et la représentation classique, il suffit de savoir que le module de la fonction d'onde est la densité de probabilité de présence de la particule. En d'autres termes, la probabilité de trouver la particule en x à dx près est $\psi^*(x,t)\psi(x,t)\,\mathrm{d}x$. Dans cette relation $\psi^*(x,t)$ est le complexe conjugué de la fonction d'onde.

Ces notions se généralisent facilement pour une particule dans un monde à trois dimensions puis pour *N* particules. Cette définition de la fonction d'onde a pour conséquence immédiate que son intégrale sur tout l'espace est égale à 1. Il faut bien faire attention au fait que dans cette représentation la variable d'espace x n'est pas la position de la particule mais la variable continue repérant un point quelconque de l'espace. La position de la particule est distribuée dans tout l'espace en fonction du module de sa fonction d'onde. Le concept de paquet d'ondes sera introduit ultérieurement, il permet alors faire le lien entre la représentation ondulatoire que nous venons de décrire et la notion classique de la particule dont on peut à tout moment définir une position précise.

En mécanique quantique, les grandeurs physiques habituelles (vitesse, énergie cinétique, position) ne se représentent plus par des fonctions dépendant des coordonnées et des vitesses des *N* particules mais par des opérateurs agissant sur la fonction d'onde. Le *tableau 2.1* montre la correspondance entre les grandeurs physiques en mécanique classique et les opérateurs associés en mécanique quantique dans le cas simple d'un système à une particule et à une dimension.

Tableau 2.1

Grandeur physique	Expression en mécanique classique	Expression de l'opérateur en mécanique quantique
Position	x_0	$x \cdot \psi(x)$
Impulsion	$m v_0$	$\frac{\hbar}{i}\frac{\partial \psi}{\partial x}$
Énergie	E_0	$i\hbar\frac{\partial \psi}{\partial t}$

Les expressions en mécanique classique sont indicées car elles sont relatives à des valeurs données. La position de la particule est par exemple à l'abscisse 1,2 mètre. En mécanique quantique, elles sont des opérateurs agissant sur la fonction d'onde. La position est par exemple définie par la multiplication de la fonction d'onde par la variable x. Elle transforme une fonction d'onde en une autre fonction d'onde. La constante $\hbar$ qui intervient dans ces formules de conversion est la constante caractéristique de la physique quantique. C'est la constante de Planck. Sa valeur est :

$$\hbar = 1{,}054 \times 10^{-34}\ \mathrm{J \cdot s}$$

Il est alors possible de passer du mode quantique au monde macroscopique en introduisant la notion de valeur moyenne d'un opérateur. La grandeur physique A dans le monde macroscopique est la moyenne au sens mathématique de l'opérateur associé.

$$\langle A \rangle = \int \psi^* A \psi \, \mathrm{d}x$$

Dans le cas d'un système à une particule et à une dimension, la position sera par exemple :

$$\langle x \rangle = \int \psi^* x \psi \, \mathrm{d}x$$

L'impulsion sera :

$$\langle p \rangle = \int \psi^* \frac{\hbar}{i} \frac{\mathrm{d}\psi}{\mathrm{d}x} \, \mathrm{d}x$$

L'énergie sera :

$$\langle E \rangle = \int \psi^* i\hbar \frac{\partial \psi}{\partial t} \, \mathrm{d}x$$

Ces définitions se généralisent facilement quand on passe à trois dimensions.

2.3.3 *L'équation de Schrödinger*

Il reste à savoir comment on calcule cette fonction d'onde. La solution est donnée par la résolution de l'équation fondamentale de la mécanique quantique à savoir l'équation de Schrödinger. À partir de l'Hamiltonien de la mécanique classique, somme de l'énergie cinétique et de l'énergie potentielle, on forme l'opérateur Hamiltonien H en remplaçant les variables algébriques classiques par les opérateurs selon le tableau de correspondance précédent. L'équation de Schrödinger s'écrit alors :

$$i\hbar \frac{\partial \psi}{\partial t} = H\psi \tag{2.11}$$

Elle permet de calculer la fonction d'onde et ensuite de calculer les valeurs moyennes des autres opérateurs.

Un exemple simple peut préciser ces notions assez différentes des notions habituelles. C'est celui d'un électron libre dans l'espace. L'Hamiltonien est réduit à l'énergie cinétique.

$$H = \frac{p^2}{2m}$$

L'opérateur Hamiltonien en mécanique quantique s'écrit donc de la manière suivante, en appliquant deux fois de suite la dérivation.

$$H\psi = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2}$$

Pour simplifier le problème, on se limite en général à la recherche des solutions stationnaires c'est-à-dire des états ayant une énergie bien déterminée E_0. Dans ce cas, l'action de l'opérateur H se limite à multiplier la fonction d'onde par E_0. On dit que la valeur E_0 est une valeur propre de l'opérateur H et que dans ce cas la fonction d'onde est une fonction propre de l'opérateur considéré. Cela n'est pas évident et constitue un des postulats de la mécanique quantique.

2.3.4 *Ensemble complet d'observables qui commutent*

Reprenons de manière générale les notions d'opérateur et de fonction propre de l'opérateur. Les opérateurs en général ne commutent pas contrairement aux fonctions ce qui veut dire que l'effet de AB sur une fonction d'onde est en général différent de BA. Le lecteur pourra facilement le vérifier avec les opérateurs position et impulsion définis précédemment.

Un opérateur A étant donné, on admet comme un des postulats de la mécanique quantique que les valeurs possibles de la grandeur physique associée sont les valeurs propres de l'opérateur, c'est-à-dire les solutions de l'équation :

$$A\psi = a\psi$$

Dans cette équation ψ est une fonction d'onde possible du système physique considéré (une ou plusieurs particules), A indique une action sur cette fonction (par exemple une dérivation par rapport à une coordonnée mais bien d'autres sont possibles) et a est une valeur numérique réelle.

La solution n'est en général pas unique et on obtient un ensemble de valeurs possibles pour a. Pour chaque valeur de a, il y a souvent plusieurs fonctions solutions notées ψ_{na}. On dit alors que la valeur propre a est dégénérée. L'indice n identifie les différentes fonctions correspondant à la même valeur a.

Les valeurs propres pour avoir un sens physique doivent être réelles et non pas imaginaires, ce qui impose à l'opérateur la propriété mathématique d'hermiticité. On dit aussi que l'opérateur est une observable. On prouve dans ce cas que l'opérateur doit être tel que :

$$\int \psi^* (A\psi)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z = \int (A\psi)^* \psi\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z$$

Quand deux opérateurs physiques commutent, on peut montrer qu'ils ont alors en commun un ensemble de fonctions propres. On montre également qu'il est toujours possible de choisir ces fonctions de telle manière que deux fonctions propres correspondant à deux valeurs propres différentes soient orthonormales. Dans le cas d'un système à une particule, cette condition s'écrit :

$$\int \psi_a^*(\mathbf{r})\, \psi_b(\mathbf{r})\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z = 0$$

On en arrive alors à la notion de système complet d'opérateurs physiques qui commutent.

Pour étudier un système physique, il est très intéressant d'identifier un ensemble d'opérateurs physiques A, B, C… tous commutant entre eux et possédant un ensemble **unique** de fonctions propres communes. L'Hamiltonien en fait généralement partie mais d'autres opérateurs sont ajoutés pour obtenir cette propriété. Dans ce cas, à chaque fonction propre est associé de manière unique un ensemble de nombres, les valeurs propres des divers opérateurs, permettant d'identifier de manière précise l'état du système. Ces valeurs sont appelées **nombres quantiques**. Les opérateurs de symétrie et l'Hamiltonien sont souvent choisis pour construire ce système. Ces notions difficiles mais importantes sont résumées dans les équations ci-dessous illustrant l'exemple d'un système complet de trois observables qui commutent.

$$\begin{aligned} H\psi(x,y,z,t) &= E\psi(x,y,z,t) \\ A\psi(x,y,z,t) &= a\psi(x,y,z,t) \\ B\psi(x,y,z,t) &= b\psi(x,y,z,t) \end{aligned}$$

Une solution est donc notée $\psi_{E,a,b}(x,y,z,t)$.

La mécanique quantique appliquée aux solides s'intéresse à un ensemble d'opérateurs pour décrire un système et rarement à un opérateur unique. Dans le cas des cristaux, on s'intéresse à

l'opérateur Hamiltonien mais aussi à l'opérateur de translation qui transforme $\psi(\mathbf{r})$ en $\psi(\mathbf{r}+\mathbf{T})$, car le système est invariant par des translations particulières correspondant à la symétrie par translation du cristal. On prend aussi en compte l'opérateur de parité qui transforme $\psi(\mathbf{r})$ en $\psi(-\mathbf{r})$.

2.3.5 *Cas de l'électron seul dans l'espace*

Un autre postulat de la mécanique quantique est l'affirmation suivante : si la valeur de la grandeur physique a une valeur donnée avec une probabilité de 1, alors nécessairement la fonction d'onde est une fonction propre de l'opérateur correspondant à cette valeur propre. Reprenons le cas de l'opérateur Hamiltonien de l'électron libre. On obtient, si la probabilité est égale à 1 pour que l'énergie ait la valeur E_0 :

$$i\hbar\frac{\partial\psi}{\partial t} = E_0\psi$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = E_0\psi$$

On peut alors décomposer la fonction en une partie dépendant du temps et une partie dépendant de la position :

$$\psi(x,t) = \psi_0(x)f(t)$$

On obtient,

$$\frac{f'(t)}{f(t)} = \frac{E_0}{i\hbar}$$

soit,

$$f(t) = \alpha\exp^{-i\frac{E_0 t}{\hbar}}$$

Ensuite, on calcule

$$\psi_0(x) = \beta\exp^{\frac{i}{\hbar}\sqrt{2mE_0}\,x}$$

En résumé, la fonction d'onde s'écrit :

$$\psi(x,t) = \gamma\exp^{\frac{i}{\hbar}(\sqrt{2mE_0}\,x - E_0 t)}$$

On peut donner une autre expression de la fonction d'onde en posant :

$$k_0 = \frac{\sqrt{2mE_0}}{\hbar}$$

$$E_0 = \hbar\omega_0$$

Donc,

$$\psi(x,t) = \gamma\exp^{i(k_0 x - \omega_0 t)}$$

On reconnaît l'équation d'une onde plane qui se propage avec un vecteur d'onde k_0 à une fréquence $\omega_0/2\pi$. La longueur d'onde associée est $2\pi/k_0$.

La constante de proportionnalité peut s'exprimer en fonction de la longueur L de l'espace à une dimension, en remarquant que l'intégrale du module sur cette distance est l'unité.

$$\gamma = \frac{1}{\sqrt{L}}$$

Il est maintenant possible de calculer les valeurs moyennes de la position, de l'impulsion et de l'énergie. Ce sont les grandeurs physiques classiques associées à l'onde. On trouve dans ce cas simple en intégrant de $-L/2$ à $L/2$:

$$\begin{aligned} \langle x \rangle &= 0 \\ \langle p \rangle &= \hbar k_0 \\ \langle E \rangle &= E_0 \end{aligned}$$

Ces relations montrent la correspondance entre les termes de la fonction d'onde et les grandeurs énergie et impulsion de la particule. Ces relations sont en fait plus générales et sont les relations fondamentales de passage du monde quantique (k et ω) au monde classique (E et p).

$$\begin{aligned} p &= \hbar k \\ E &= \hbar \omega \end{aligned} \tag{2.12}$$

2.3.6 *Le courant en mécanique quantique*

Il est maintenant possible de définir le courant en mécanique quantique. Le principe de conservation de la norme de la fonction d'onde amène à définir le courant associé à une fonction d'onde par :

$$j = \mathrm{Re}\left(e\, \psi^{\star} \frac{\hbar}{i\, m} \frac{\partial}{\partial x} \psi \right) \tag{2.13}$$

On généralise en trois dimensions en exprimant le courant par le gradient. C'est donc un vecteur défini en tout point de l'espace.

On peut également définir l'opérateur courant en un point $\mathbf{r_0}$ de l'espace par la relation :

$$j(\mathbf{r}_0) = \frac{e}{2\, m} (\mathbf{p}\, \delta(\mathbf{r} - \mathbf{r}_0) + \delta(\mathbf{r} - \mathbf{r}_0)\mathbf{p}) \tag{2.14}$$

Dans cette expression, $\mathbf{p}$ est l'opérateur impulsion et $\delta(\mathbf{r} - \mathbf{r}_0)$ est un opérateur défini par :

$$\delta(\mathbf{r} - \mathbf{r}_0)\psi(\mathbf{r}) = \psi(\mathbf{r}_0)$$

La valeur définie en 2.13 n'est autre que la valeur moyenne de cet opérateur comme le lecteur pourra le démontrer facilement. Cet exemple illustre également la différence entre fonction et opérateur.

Il est enfin nécessaire de définir un opérateur courant pour un système dans lequel il n'est plus possible de déterminer un état donné mais seulement un mélange d'états possibles. Ces états sont définis par leurs fonctions d'onde $\psi_m(\mathbf{r}, t)$ et des probabilités p_m associées. Il ne faut pas confondre ces probabilités avec celles qui permettent de mesurer une valeur donnée d'un opérateur parmi toutes les valeurs possibles. On définit alors l'opérateur densité $\rho(\mathbf{r})$ par :

$$\rho(\mathbf{r})\psi(\mathbf{r}, t) = \sum_m p_m [\int \psi_m^{\star}(\mathbf{r}, t)\psi(\mathbf{r}, t)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z] \cdot \psi_m(\mathbf{r}, t)$$

L'opérateur densité de courant est alors dans ce cas :

$$\mathbf{j} = \frac{e}{2m}[\rho(\mathbf{r}) \cdot \mathbf{p} + \mathbf{p} \cdot \rho(\mathbf{r})] \tag{2.15}$$

2.4 Les électrons dans les solides et les niveaux énergétiques

2.4.1 *Les électrons dans les solides*

Le but de ce paragraphe est de montrer que, dans un solide, les électrons n'ont pas des énergies quelconques mais des énergies en nombre fini et réparties dans des bandes autorisées. La répartition des énergies selon des bandes et non pas de manière continue permet d'expliquer que certains solides sont conducteurs et que d'autres sont isolants. Cette propriété ne peut pas être établie par une théorie classique et il faut faire appel à la mécanique quantique.

Il faut également expliquer que les états possibles des électrons sont eux aussi quantifiés, c'est-à-dire en nombre fini. Ce nombre est grand mais non infini. Enfin, il est nécessaire de tenir compte du principe de Pauli qui établit que deux électrons ne peuvent être dans le même état quantique. Ce principe fondamental permet de comprendre de nombreuses propriétés des solides, en particulier les propriétés électriques et optiques.

Considérons un solide formé d'atomes placés avec une certaine régularité. Quand celle ci est parfaite, les atomes sont placés à des distances fixes les uns des autres, on dit alors que la structure est cristalline. Quand cette régularité est vérifiée sur des distances assez faibles, on dit que la structure est polycristalline. Quand la répartition est totalement irrégulière, on dit que la structure est amorphe. La *figure 2.13* montre ces trois états dans un espace à deux dimensions pour simplifier la représentation.

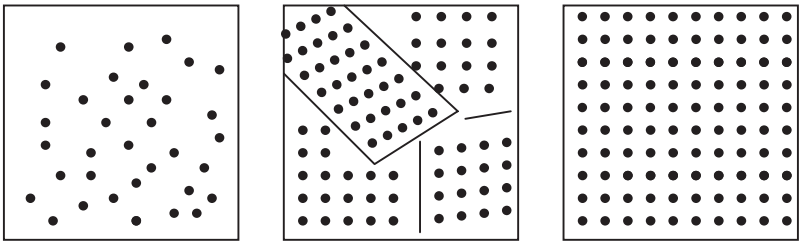


Figure 2.13 – Les trois états d'un solide.

Les électrons de chaque atome peuvent se diviser en deux familles, les électrons fortement liés au noyau par la force électrique et répartis sur les couches complètement remplies et les électrons faiblement liés au noyau et situés sur les couches incomplètes. On les appelle électrons de valence. Le solide peut donc se représenter comme un ensemble d'ions fixes formés par les noyaux et leurs électrons liés et d'un gaz d'électrons de valence faiblement liés. Les atomes de silicium comportent par exemple quatre électrons de valence.

Calculer la fonction d'onde du solide est une opération d'une énorme complexité et inaccessible même avec des outils de calcul très puissants. Il faut faire un grand nombre d'approximations. La

première est de considérer les ions comme fixes puisqu'ils sont beaucoup plus lourds que les électrons. La seconde approximation est de considérer que la somme de toutes les interactions électriques (électron-ion et électron-électron) est équivalente à un potentiel électrique moyen agissant sur les électrons de valence.

On considère alors qu'un électron de valence quelconque est soumis à un potentiel qui ne dépend que de la variable position dans l'espace. C'est la même fonction pour tous les électrons de valence. Dans un cristal, ce potentiel est une fonction périodique. La périodicité du motif permet de comprendre que le potentiel présente la même propriété. Au niveau des atomes, les électrons de valence sont dans le potentiel électrique de l'ion positif privé des électrons de valence et ce potentiel présente un maximum. L'énergie potentielle y est alors minimale puisqu'elle est le produit de la charge négative de l'électron par le potentiel. L'électron exprimera alors une forte envie de rester dans cette position d'énergie minimale.

La modélisation du problème par le potentiel moyen en simplifie considérablement la résolution. La fonction d'onde du système est dans le cas général fonction des N électrons de valence présents dans le solide, les ions étant supposés immobiles. L'hypothèse du potentiel moyen permet de supposer que chaque électron de valence est soumis au même potentiel et le problème ne dépend plus que des trois variables d'espace de la fonction d'onde de l'électron considéré et du temps. Il faut remarquer que cette approche du problème conduit à « délocaliser » les électrons de valence. Puisque tous les électrons sont soumis au même potentiel, il devient impossible d'affecter un électron à un atome particulier.

Quand nous serons amenés à considérer un électron localisé dans l'espace, par exemple un électron créé par ionisation en un point donné, il sera nécessaire de former un paquet d'ondes pour avoir une représentation correcte de la situation. Ce paquet d'ondes est alors formé comme une combinaison linéaire d'états propres de l'énergie.

Dans l'hypothèse du potentiel moyen, la fonction d'onde de l'électron de valence obéit à l'équation de Schrödinger dans laquelle $V(\mathbf{r})$ est le potentiel moyen.

$$i\hbar \cdot \frac{\partial}{\partial t}\psi(\mathbf{r},t) = -\frac{\hbar^2}{2m} \cdot \Delta\psi(\mathbf{r},t) + V(\mathbf{r}) \cdot \psi(\mathbf{r},t) \tag{2.16}$$

Dans cette équation m est la masse de l'électron et Δ est l'opérateur de Laplace égal à la somme des dérivées partielles du second ordre.

$$\Delta\psi(\mathbf{r},t) = \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2}$$

On s'intéresse aux fonctions d'onde correspondant à une énergie E donnée de l'électron. On dit alors que l'état est stationnaire. Dans ce cas, l'équation de Schrödinger se simplifie :

$$i\hbar \cdot \frac{\partial}{\partial t}\psi(\mathbf{r},t) = E \cdot \psi(\mathbf{r},t)$$

La solution s'écrit alors :

$$\psi(\mathbf{r},t) = \psi(\mathbf{r},0) \cdot e^{-\frac{iEt}{\hbar}}$$

La fonction d'onde est donc le produit d'une fonction dépendant de la position et d'une onde plane. La probabilité de présence de l'électron en un point donné est alors indépendante du temps. Si on reporte cette valeur dans l'équation générale, on obtient, en notant $\psi(\mathbf{r},0) = \psi(\mathbf{r})$:

$$-\frac{\hbar^2}{2\,m}\cdot\Delta\psi(\mathbf{r})+V(\mathbf{r})\cdot\psi(\mathbf{r})\;=\;E\cdot\psi(\mathbf{r})$$

Cette équation, indépendante du temps, permet de calculer la fonction d'onde des électrons de valence dans un solide.

Il est maintenant nécessaire de tenir compte de la structure cristalline du solide. On peut montrer qu'il est possible de définir pour un cristal donné un vecteur **T** tel que la structure du solide soit invariante dans une translation de ce vecteur. La *figure 2.14* représente un cristal dans l'espace réel à trois dimensions. Le même motif se répète périodiquement selon les trois directions avec des périodes définies par les vecteurs $\mathbf{a_1}$, $\mathbf{a_2}$ et $\mathbf{a_3}$. Il est important de faire la différence entre les points qui représentent le réseau périodique et les atomes eux mêmes. Dans l'exemple représenté, il y a deux atomes par maille élémentaire. Le vecteur **T** est une combinaison linéaire de ces trois vecteurs.

$$\mathbf{T}\;=\;p\cdot\mathbf{a}_1+q\cdot\mathbf{a}_2+r\cdot\mathbf{a}_3$$

Dans cette relation, les nombres p, q et r sont des entiers relatifs, éventuellement nuls. La figure représente deux atomes par maille élémentaire mais il peut y avoir un ou plusieurs atomes dans cette maille élémentaire.

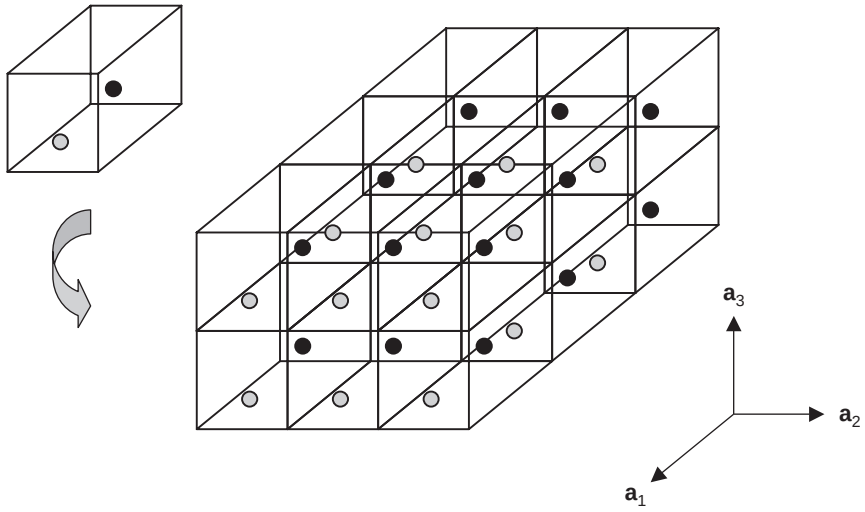


Figure 2.14 – La périodicité du réseau.

Le potentiel moyen doit alors satisfaire à la condition :

$$V(\mathbf{r})\;=\;V(\mathbf{r}+\mathbf{T})$$

Un théorème classique de la physique des solides, le théorème de Bloch, montre alors que la fonction d'onde des électrons de valence s'écrit :

$$\psi(\mathbf{r})\;=\;\mathrm{e}^{i\,\mathbf{k}\cdot\mathbf{r}}\cdot u_{\mathbf{k}}(\mathbf{r}) \qquad (2.17)$$

Dans cette formule, la fonction $u_{\mathbf{k}}(\mathbf{r})$ est périodique et de période **T**. À chaque valeur de **k** sont associées éventuellement plusieurs fonctions, par exemple deux fonctions pour les deux états pos-

sibles du spin. Cette forme de la fonction d'onde explique que la probabilité de présence de l'électron a la périodicité du réseau ce qui est physiquement évident. Ce n'est pas le cas de la fonction d'onde qui elle n'a pas la périodicité du réseau. Dans la résolution de l'équation de Schrödinger, on obtient une relation entre l'énergie E et la valeur de $\mathbf{k}$. Cette relation est appelée relation de dispersion. Pour expliquer ce théorème et la signification du paramètre $\mathbf{k}$, il est nécessaire de détailler la résolution de l'équation de Schrödinger dans un solide.

2.4.2 *Quelques cas simples*

Dans un premier temps, nous allons étudier le cas le plus simple possible, celui de N électrons dans un solide à une dimension en négligeant le potentiel moyen. L'équation de Schrödinger s'écrit :

$$-\frac{\hbar^2}{2\,m}\cdot\frac{\mathrm{d}^2\psi(x)}{\mathrm{d}x^2} = E\cdot\psi(x)$$

Les solutions de cette équation sont des combinaisons linéaires d'exponentielles e^{ikx}. La relation entre le paramètre k et l'énergie est obtenue en remplaçant la fonction d'onde par sa valeur dans l'équation de Schrödinger.

$$E = \frac{\hbar^2\,k^2}{2\,m}$$

Pour compléter l'analyse, il faut tenir compte du fait que le solide est fini de longueur L. Une solution naturelle serait d'écrire que la fonction d'onde est nulle à l'extérieur du solide puisque l'électron est confiné dans la matière. Par continuité on pourrait alors écrire :

$$\psi(0) = \psi(L) = 0$$

En fait, une autre condition sera choisie.

$$\psi(0) = \psi(L)$$

Cette condition appelée règle de Born-von Kármán est moins restrictive. Elle peut se justifier de manière rigoureuse en faisant appel à un postulat de la mécanique quantique qui impose à tout opérateur d'être hermitique. Cette règle est appliquée à l'opérateur impulsion. La fonction d'onde s'écrit alors :

$$\psi(x) = \frac{1}{\sqrt{L}}\cdot\mathrm{e}^{i\,k\,x}$$

La condition de Born-von Kármán s'écrit :

$$\psi(L) = \frac{1}{\sqrt{L}}$$

soit,

$$k\,L = 2\,\pi\,n$$

n étant un nombre entier positif ou négatif. L'énergie de l'électron est donc :

$$E_n = \frac{\hbar^2}{2\,m}\cdot\left(\frac{n}{2\,L}\right)^2$$

Cette analyse simple montre que l'énergie de l'électron ne peut prendre que des valeurs discrètes et que le paramètre k appelé vecteur d'onde est lui aussi quantifié.

À chaque valeur de n correspondent deux états associés aux deux valeurs du spin. Si N électrons sont disponibles dans le solide, les états se remplissent par énergies croissantes. C'est une conséquence du principe de Pauli qui interdit à deux électrons d'être dans le même état. Si les électrons pouvaient occuper le même état ils se placeraient tous dans l'état d'énergie minimale. Le principe de remplissage des états conduit donc à atteindre une valeur maximale E_{F0} donnée par la relation :

$$E_{F0} = \frac{\hbar^2}{2\,m} \cdot \left(\frac{N}{4\,L}\right)^2$$

Si N/L est égal à 0,4 électron par angström, on trouve E_{F0} égal à 1 eV.

La *figure 2.15* représente l'énergie d'un électron de valence en fonction du vecteur d'onde dans le cas d'un solide à une dimension de longueur L et de maille élémentaire a.

Quand k varie de $-\pi/a$ à $+\pi/a$, il y a donc $(2\pi/a)/(2\pi/L)$ états possibles – en fait, le double, à cause des deux spins et finalement $2\,L/a$ états. Il y a en réalité une grande quantité de valeurs possibles pour k et la figure en représente beaucoup moins que dans la réalité.

Si, par maille élémentaire, il y a quatre électrons de valence alors il y a $4\,L/a$ électrons à placer. Deux zones d'extension $2\pi/a$ sont nécessaires. Ce résultat ne serait plus vrai avec trois électrons de valence par maille. Les valeurs de k multiples de π/a délimitent des zones qui nous ont servi uniquement à dénombrer les états. Ces zones jouent un rôle beaucoup plus important comme il le sera étudié dans le paragraphe suivant.

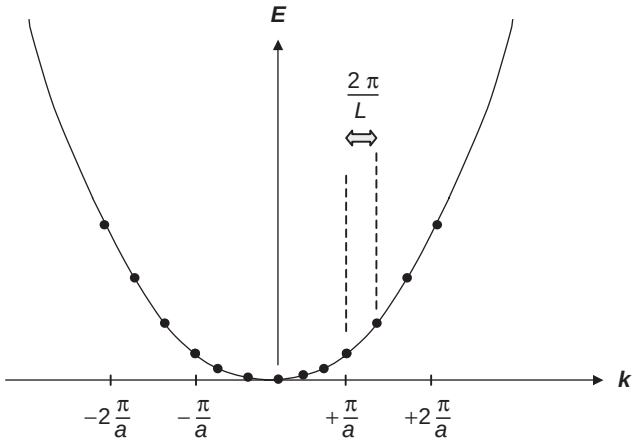


Figure 2.15 – Électrons libres dans un système à une dimension.

La généralisation à trois dimensions ne pose pas de problème. Les fonctions d'onde s'écrivent pour un cube de côté L.

$$\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{V}} \cdot e^{i\mathbf{k}\mathbf{r}}$$

Le vecteur **k** ne peut prendre que des valeurs discrètes :

$$k_x = 0, \pm\frac{2\pi}{L}, \pm\frac{4\pi}{L}, \ldots$$

$$k_y = 0, \pm\frac{2\pi}{L}, \pm\frac{4\pi}{L}, \ldots$$

$$k_z = 0, \pm\frac{2\pi}{L}, \pm\frac{4\pi}{L}, \ldots$$

L'énergie est donc elle aussi quantifiée :

$$E = \frac{\hbar^2 k^2}{2m}$$

Si N électrons sont disponibles, ils remplissent les états par ordre d'énergies croissantes car deux électrons ne peuvent avoir le même état. L'énergie atteint ainsi la valeur maximale E_{F0} appelée énergie de Fermi. Les états d'énergie constante correspondent à des vecteurs **k** localisés sur la surface d'une sphère. La valeur maximale du module de **k** est notée k_F. Comme deux états de spins différents correspondent à un volume $(2\pi/L)^3$ dans l'espace des **k**, on peut écrire :

$$2 \cdot \frac{\frac{4}{3}\pi k_F^3}{\left(\frac{2\pi}{L}\right)^3} = N$$

Cette relation fixe l'énergie de Fermi du gaz d'électrons. On remplace k par la valeur particulière k_F. Le calcul conduit à des valeurs de quelques eV (7 eV pour le cuivre et 5,5 eV pour l'or).

2.4.3 *Apparition des bandes*

Le cas d'un solide à une dimension avec introduction du potentiel du cristal donne lieu au modèle de Krönig-Penney. Il explique l'essentiel des propriétés des solides.

On suppose le potentiel moyen égal à une valeur constante V_0 au voisinage des ions et nul ailleurs. La distance entre atomes est a, et la zone d'action du potentiel s'étend sur une distance b. Ce modèle

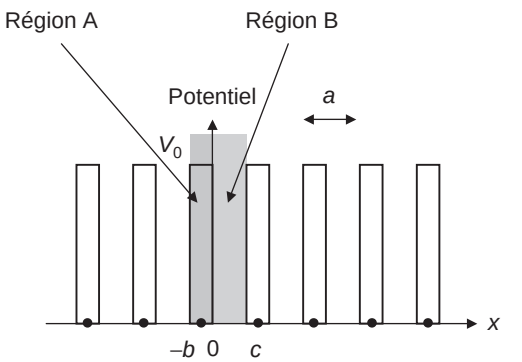


Figure 2.16 – Cristal à une dimension

simple intègre le fait que les électrons se localisent préférentiellement au niveau des atomes. En cherchant les solutions dans ce modèle, nous verrons que certaines valeurs de l'énergie ne sont pas possibles ce qui introduit la notion de bande.

L'équation de Schrödinger s'écrit alors au voisinage des atomes, soit dans la région A :

$$\frac{\partial^2 \psi(x)}{\partial x^2} + \frac{2m}{\hbar^2}(E + eV_0)\,\psi(x) = 0$$

Elle s'écrit dans les régions situées ente les atomes, par exemple la région B :

$$\frac{\partial^2 \psi(x)}{\partial x^2} + \frac{2m}{\hbar^2} E\,\psi(x) = 0$$

Les deux solutions dans les deux régions A et B sont des combinaisons de deux fonctions exponentielles. Dans la région A,

$$\psi(x) = A \cdot e^{ik_1 x} + B \cdot e^{-ik_1 x}$$

avec,

$$k_1 = \sqrt{\frac{2m}{\hbar^2}(E + eV_0)}$$

Dans la région B, on écrit :

$$\psi(x) = C \cdot e^{ik_2 x} + D \cdot e^{-ik_2 x}$$

avec,

$$k_2 = \sqrt{\frac{2m}{\hbar^2} E}$$

On obtient donc un système de solutions à quatre paramètres. En écrivant la continuité de la fonction d'onde et de sa dérivée en x égal à zéro et en x égal à c, on obtient quatre équations entre ces paramètres. Pour écrire la continuité de la fonction d'onde et de sa dérivée en x égal à c, il faut utiliser le fait que la fonction d'onde obéit à la relation :

$$\psi(x + a) = e^{ika} \cdot \psi(x)$$

Cette relation n'a rien d'évident. Elle peut se déduire de la périodicité du potentiel. De manière plus générale, elle exprime que la fonction d'onde cherchée est aussi une fonction propre de l'opérateur translation. En référence aux généralités écrites sur la recherche d'un ensemble complet d'observables qui commutent, on cherche ici des fonctions propres communes à l'opérateur Hamiltonien et à l'opérateur de translation qui exprime la symétrie du réseau. Dans cette relation k est un paramètre inconnu qui ne doit pas être confondu avec les paramètres k_1 et k_2 des fonctions d'onde solutions.

En $x = 0$, on écrit donc :

$$A + B = C + D$$
$$iA \cdot k_1 - iB \cdot k_1 = iC \cdot k_2 - iD \cdot k_2$$

En $x = c$, on écrit de même en tenant compte de la règle de translation énoncée précédemment :

$$C \cdot e^{ik_2 c} + D \cdot e^{-ik_2 c} = e^{ika} \cdot (A \cdot e^{-ik_1 b} + B \cdot e^{ik_1 b})$$
$$ik_2 C \cdot e^{ik_2 c} - ik_2 D \cdot e^{-ik_2 c} = e^{ika} \cdot (ik_1 A \cdot e^{-ik_1 b} - ik_1 B \cdot e^{ik_1 b})$$

Le déterminant de ce système doit être nul pour qu'il y ait des solutions non nulles. Un calcul assez fastidieux permet d'obtenir la condition :

$$P \frac{\sin(\varepsilon a)}{\varepsilon a} + \cos(\varepsilon a) = \cos(k a) \qquad (2.18)$$

Dans cette relation de dispersion, les variables sont définies par :

$$P = \frac{e m a V_0 b}{\hbar^2} \quad \text{et} \quad \varepsilon = \frac{2m}{\hbar^2} \cdot E$$

On a également supposé dans ce calcul que b était petit. La fonction correspondant au premier membre de l'équation est représentée *figure 2.17*.

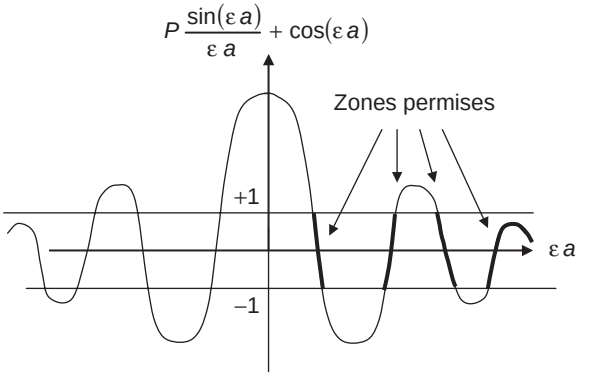


Figure 2.17 – Résolution de l'équation de dispersion.

Comme la fonction de la *figure 2.16* est égale à un cosinus elle est nécessairement comprise entre −1 et +1 ce qui permet de définir les zones possibles pour la grandeur $a\varepsilon$. Les valeurs négatives de l'énergie n'ont pas de signification physique. Toutes les énergies ne sont pas autorisées et les énergies possibles se répartissent sous forme de bandes. On introduit à partir de cet exemple ce résultat fondamental en physique des solides. Il a en fait une portée générale.

Si nous choisissons une valeur de l'énergie, la relation de dispersion permet de calculer le paramètre k à une constante additive $n \cdot 2\pi/a$ près puisque seul son cosinus est fixé. Il est alors appelé quasi-vecteur d'onde à cause de cette indétermination. Il est ensuite possible de calculer la fonction d'onde en tout point du cristal à partir de son expression dans les régions A et B. Il suffit d'appliquer :

$$\psi(x + pa) = e^{ikpa} \cdot \psi(x)$$

Le paramètre k étant choisi, on écrit la fonction d'onde sous la forme :

$$\psi(x) = e^{ikx} \cdot u_k(x)$$

Il est facile de montrer que la fonction $u_k(x)$ est périodique et de période a. On retrouve l'expression du théorème de Bloch. Le choix du paramètre k peut être effectué de telle sorte que sa valeur, pour une énergie donnée, soit la plus proche possible de la valeur du vecteur d'onde k de l'électron libre ayant la même énergie. Le choix de ce paramètre est un point délicat de la théorie des solides.

En fait et contrairement au cas des électrons libres, le paramètre k n'est pas une valeur propre de l'opérateur impulsion car l'opérateur de Hamilton d'un solide ne commute pas avec l'opérateur impulsion. L'opérateur qui commute avec l'Hamiltonien est l'opérateur translation dont l'action est définie par :

$$T_{na}\psi(x) = \psi(x + na)$$

La périodicité du potentiel moyen entraîne que cet opérateur commute avec l'Hamiltonien à la condition que la translation soit un multiple entier de la maille élémentaire. Le paramètre k est alors une valeur propre de cet opérateur.

$$T_{na}\psi_k(x) = \psi_k(x + na) = \exp^{ikna}\psi_k(x)$$

Pour éviter de compter les états plusieurs fois, on se restreint à la valeur de k dans la zone comprise entre $-\pi/a$ et π/a. Cette zone est appelée zone de Brillouin. Les fonctions d'onde propres ne sont pas des ondes planes mais sont exprimées par les équations données au début de cette analyse. Des discontinuités apparaissent pour les multiples de $2\pi/a$, comme il sera expliqué dans la suite. L'apparition de ces discontinuités n'est pas surprenante si on pense à la diffraction des ondes dans les cristaux.

Quand P augmente, les zones autorisées sont plus limitées. Les deux cas extrêmes, P nul et P infini sont intéressants. Quand P est nul, la relation de dispersion devient :

$$\cos(\varepsilon a) = \cos(k a)$$

soit :

$$\varepsilon a = k a + 2\pi n$$

On obtient alors pour n égal à zéro, ce qui correspond à un choix particulier de k :

$$E = -\frac{\hbar^2 k^2}{2m}$$

Remarquons une nouvelle fois que le paramètre k est défini à un multiple entier de $2\pi/a$ près. Le choix n'est donc pas unique. Le choix n égal à 0 permet de faire correspondre le paramètre k au vecteur d'onde de l'électron supposé libre mais un autre choix serait possible.

Quand P est infini, la relation de dispersion s'écrit puisque le deuxième membre est fini :

$$\sin(\varepsilon a) = 0$$

$$E = \frac{\pi^2 \hbar^2 n^2}{2 m a^2}$$

Les valeurs de l'énergie sont discrètes et sont celles d'une particule bloquée dans un puits de potentiel. Quand P a une valeur intermédiaire, des zones interdites apparaissent comme le montre la *figure 2.18*.

La *figure 2.18* représente l'évolution des énergies possibles quand le paramètre P varie de 0 à l'infini ou, ce qui est équivalent, quand l'effet du réseau cristallin est de plus en plus intense. On comprend ainsi l'origine des bandes d'énergie dans les solides. Cette répartition des énergies permises a une

importance extrême pour toutes les propriétés électriques et optiques des solides. Les résultats ont été établis dans le cas particulier d'un solide à une dimension. Les conclusions restent valables pour les solides réels à trois dimensions. Les calculs sont cependant d'une grande complexité.

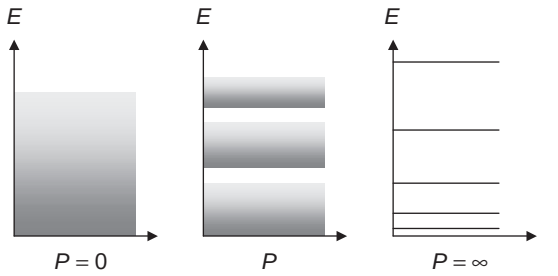


Figure 2.18 – Les différents régimes du cristal à une dimension.

La périodicité du graphe représentant l'énergie en fonction du pseudo-vecteur d'onde invite à limiter la représentation dans la région s'étendant de π/a de part et d'autre de l'origine. Toutes les autres parties des courbes s'en déduisent par translation d'un multiple de $2\pi/a$. Cette zone est appelée zone de Brillouin. La *figure 2.19*, page suivante, montre comment il est possible, par des translations de k d'un multiple entier de $2\pi/a$, de ramener toutes les bandes dans la zone de Brillouin. Cette opération revient à choisir le vecteur k comme il a été vu précédemment parmi toutes les valeurs possibles. Toutes les fonctions de Bloch peuvent être obtenues par des vecteurs k de la zone de Brillouin ainsi construite.

La *figure 2.19* montre les deux cas : électrons libres dans un cristal et électrons soumis à un potentiel périodique. Dans les deux cas, tous les états seront obtenus en décalant le vecteur k d'autant de fois $2\pi/a$ qu'il le faut pour que le nouveau vecteur k soit dans la zone dite de Brillouin. En réalité, la représentation 2.19(c) est la bonne représentation physique et les points représentés sur le schéma pour une énergie donnée E_0 et décalés d'un multiple de $2\pi/a$ représentent le même état de la fonction d'onde, fonction propre de l'Hamiltonien pour la valeur E_0 de l'énergie.

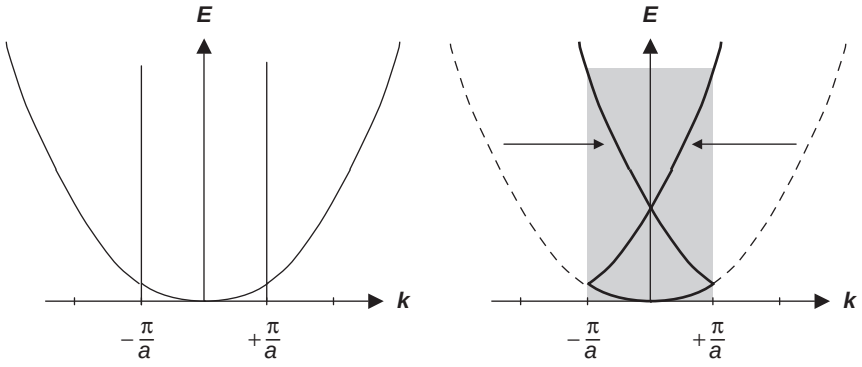
2.4.4 *Le cristal à trois dimensions*

Dans un solide cristallin, le même motif se répète un grand nombre de fois. On peut alors définir un maillage de l'espace obtenu à partir de trois vecteurs $\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}$ comme il a été vu dans le paragraphe 2.4.1. Dans le cas le plus général, ces vecteurs sont quelconques. La périodicité du réseau s'exprime par :

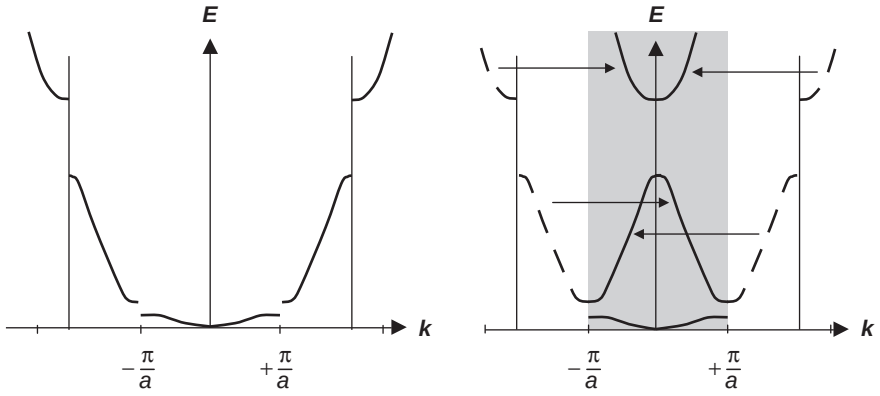
$$f(\mathbf{r}) = (\mathbf{r} + m \cdot \mathbf{a}_1 + n \cdot \mathbf{a}_2 + p \cdot \mathbf{a}_3) = f(\mathbf{r} + \mathbf{T})$$

Dans cette relation, la fonction f est par exemple le potentiel. Le réseau du silicium est représenté à titre d'exemple *figure 2.20*. Il est dit de type diamant. Il est formé par deux réseaux cubiques face centrée décalés. L'un d'entre eux est représenté par des sphères grises sur la figure et l'autre par des sphères blanches. Toutes les sphères représentent des atomes de silicium.

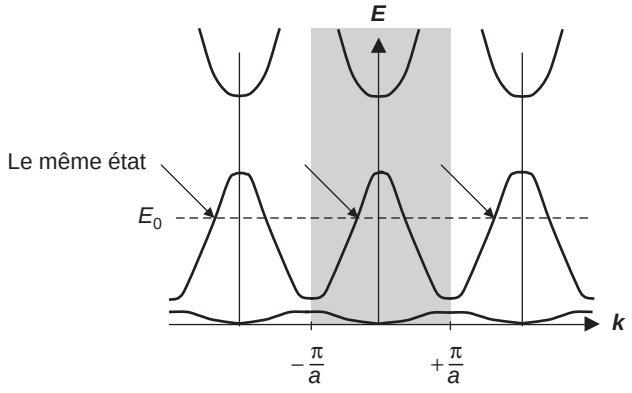
Il est maintenant possible d'exprimer la périodicité du réseau en écrivant les fonctions périodiques sous forme d'une série de Fourier à trois dimensions. Cette décomposition introduira naturellement le réseau réciproque. La fonction périodique f, potentiel ou autre grandeur physique liée à la périodicité du réseau, s'exprime par :



a) Électrons libres dans un cristal



b) Électrons soumis à un potentiel périodique



c) Électrons dans la représentation périodique

Figure 2.19 – Construction de la zone de Brillouin.

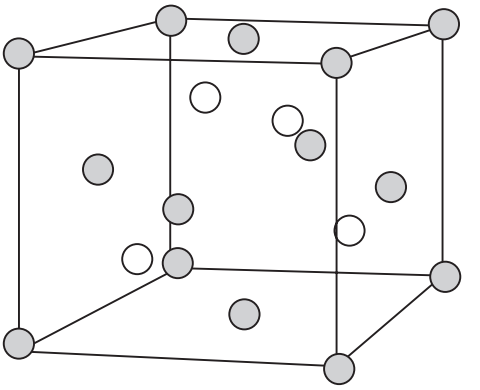


Figure 2.20 – La maille élémentaire du silicium.

$$f(\mathbf{r}) = \sum_{\mathbf{G}} C_{\mathbf{G}} \cdot e^{i\,\mathbf{G}\cdot\mathbf{r}}$$

La somme est faite sur tous les vecteurs **G** tels que **G**.**T** soit égal à $2\pi n$, n étant un entier relatif. Dans cette relation, **T** est un vecteur du réseau direct tel que défini au début de ce paragraphe. Il est alors facile de montrer que, moyennant cette condition, la fonction f est bien périodique et que les vecteurs **G** sont obtenus comme des combinaisons entières de trois vecteurs **A1**, **A2**, **A3** appelés vecteurs de base du réseau réciproque.

$$\mathbf{G} = h \cdot \mathbf{A}_1 + p \cdot \mathbf{A}_2 + q \cdot \mathbf{A}_3$$

Dans cette relation, h, p, q sont des entiers relatifs. Les vecteurs de base du réseau réciproque vérifient :

$$\mathbf{A}_1 \cdot \mathbf{a}_j = 2\pi \cdot \delta_{ij}$$

Le symbole $\ddot{a}_{ij}$ est nul dans tous les cas sauf quand i et j sont égaux. Dans ce cas, il est égal à l'unité. Les vecteurs $\mathbf{A}_i$ se construisent à partir des vecteurs $\mathbf{a}_i$. Ils sont normaux aux plans formés par deux vecteurs du réseau direct et leur module satisfait l'équation de définition. On montre par exemple que le réseau réciproque d'un réseau cubique face centrée est un réseau cubique centré.

Nous allons maintenant généraliser les résultats obtenus dans le paragraphe 2.4.2. Nous y avons appris qu'il suffisait de faire varier le vecteur k entre les valeurs $-\pi/a$ et $-\pi/a$ pour décrire tous les états du système. De la même manière, dans l'espace à trois dimensions, il suffit d'étudier les états pour les valeurs de **k** dans la zone de Brillouin. Elle est obtenue dans le réseau réciproque à partir de l'origine en construisant tous les plans médiateurs des droites reliant l'origine aux nœuds voisins de l'espace réciproque. Tous les vecteurs de l'espace réciproque peuvent alors se construire comme la somme d'un vecteur de la zone de Brillouin et d'un vecteur de translation de type **G** défini précédemment.

La *figure 2.21* représente la zone de Brillouin d'un cristal à deux dimensions. Elle est délimitée par les lignes médiatrices entre les sommets du réseau réciproque. Les nœuds du réseau réciproques sont indiqués par des points noirs. Ce cas est simple car les vecteurs $\mathbf{A}_1$ et $\mathbf{A}_2$ sont perpendiculaires. Ils pourraient ne pas l'être.

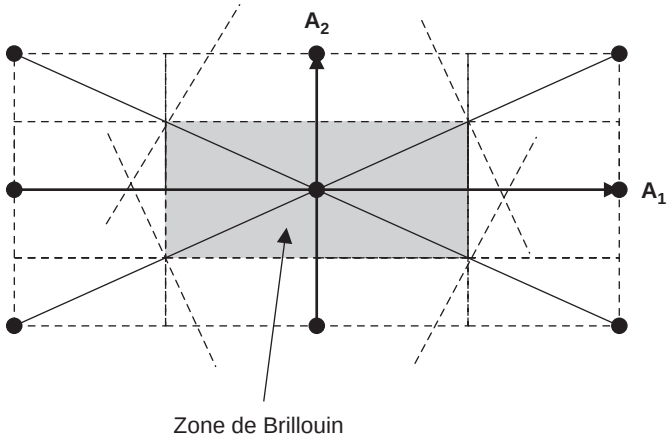


Figure 2.21 – Zone de Brillouin dans un cristal à deux dimensions.

Revenons maintenant aux fonctions d'onde des électrons de valence. Pour une énergie donnée, le cas à une dimension nous a montré que plusieurs valeurs de k étaient possibles à $n \cdot 2\pi/a$ près. Ce résultat se généralise à trois dimensions et les valeurs possibles de $\mathbf{k}$ correspondant à une énergie donnée sont définies à un vecteur $\mathbf{G}$ près du réseau réciproque. Rappelons que le vecteur $\mathbf{k}$ est le paramètre de la fonction de Bloch exprimée par :

$$\psi(\mathbf{r}) = e^{i\mathbf{kr}} \cdot u_{n\mathbf{k}}(\mathbf{r})$$

Rappelons que le vecteur $\mathbf{k}$ est également une valeur propre de l'opérateur de translation qui commute avec l'Hamiltonien.

$$\psi(\mathbf{r} + \mathbf{T}) = e^{i\,\mathbf{kT}} \cdot \psi(\mathbf{r})$$

La représentation des états dans la zone de Brillouin a conduit à ramener dans cette zone des états qui étaient à l'extérieur dans la représentation initiale ou étendue ce qui explique l'apparition de plusieurs bandes et l'indice n. On note alors la fonction de Bloch $\psi_{n\mathbf{k}}(\mathbf{r})$ expression dans laquelle l'indice n indique la bande choisie. La *figure 2.19* aide à comprendre cette manière de compter les états.

On suppose que le solide considéré est un cube dont les côtés sont L_x, L_y et L_z. Si on exprime maintenant que la fonction d'onde prend la même valeur quand x varie de L_x, quand y varie de L_y, quand z varie de L_z, on obtient avec n, p, q entiers relatifs :

$$k_x = \frac{2\pi \cdot n}{L_x}$$

$$k_y = \frac{2\pi \cdot p}{L_y}$$

$$k_z = \frac{2\pi \cdot q}{L_z}$$

Dans un espace à une dimension, quand k_x varie de Δk, il y a $\Delta k \cdot L_x/2\pi$ valeurs possibles correspondant chacune à un état. Ce résultat se transpose facilement à trois dimensions. Le nombre d'états par unité de volume de l'espace des k est alors :

$$n_e = \frac{V}{(2\pi)^3}$$

Le nombre d'états occupés est directement lié au nombre d'atomes par unité de volume et au nombre d'électrons de valence. Ce sont les électrons de valence qui remplissent les états possibles. Dans le cas à une dimension, N atomes conduisent à définir dans la zone de Brillouin un nombre d'états par bande égal à :

$$n_B = \frac{2\frac{\pi}{a}}{2\frac{\pi}{L}} = N$$

Il faut doubler ce nombre pour tenir compte des deux états de spin possibles. Il y a donc $2N$ états possibles par bande dans un matériau comportant N atomes. Ce résultat se généralise à un cristal réel. Chaque maille élémentaire fournit une valeur indépendante de k à chaque bande d'énergie, ce qui correspond à deux états en tenant compte des deux spins.

Si le réseau comporte un atome par maille avec un seul électron de valence par atome, alors il est possible de remplir une bande mais seulement à moitié. Si deux atomes sont présents par maille, toujours avec un électron non lié par atome, une bande peut être totalement remplie. Si l'atome comporte trois électrons de valence et si la maille élémentaire est formée d'un atome, deux bandes seront également remplies mais la seconde sera incomplète ce qui confère au matériau des propriétés de conduction électrique.

Par bande, les états se remplissent comme il est indiqué *figure 2.22* dans un cristal à deux dimensions. Ils remplissent l'intérieur d'un disque en prenant d'abord les états correspondant aux énergies les plus basses. La frontière entre les états occupés et non occupés est un cercle appelé cercle de Fermi.

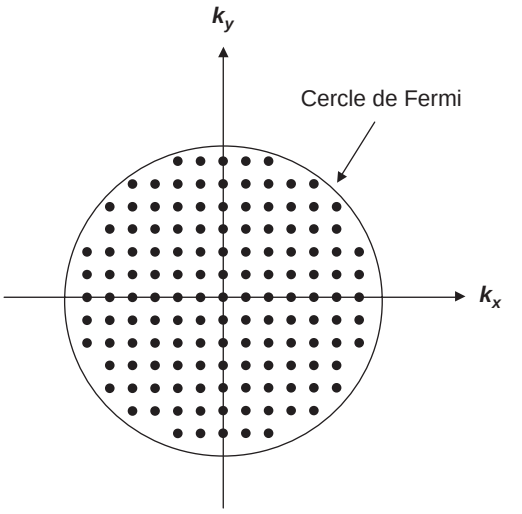


Figure 2.22 – La surface de Fermi dans un cristal à deux dimensions.

Pour simplifier l'analyse, on considère que les électrons sont libres et que l'énergie des électrons est du type :

$$E = -\frac{\hbar^2 k^2}{2m}$$

Dans un cristal à trois dimensions, le raisonnement est le même mais les électrons remplissent alors une sphère appelée sphère de Fermi et les surfaces iso-énergétiques sont des surfaces sphériques. La *figure 2.23* montre la surface de Fermi pour le sodium et le cuivre. En fait, à cause du potentiel du cristal, l'énergie des électrons n'est pas simplement donnée par la formule $E = \hbar^2 k^2 / 2m$, ce qui explique que la surface de Fermi n'est pas une simple sphère.

Figure 2.23 – Surfaces de Fermi dans un cristal à trois dimensions.

Le vecteur d'onde k_F correspondant à la valeur de l'énergie la plus élevée atteinte E_{F0} est donc donné par la relation :

$$2 \cdot \frac{4}{3}\pi k_F^3 \cdot \frac{V}{(2\pi)^3} = N$$

Dans cette relation, N est le nombre d'électrons de valence dans le volume considéré. On en déduit la valeur du vecteur d'onde et la valeur de l'énergie correspondante appelée énergie de Fermi. Elle est, par exemple, de 7 eV pour le cuivre. Quand $\mathbf{k}$ varie de $\mathrm{d}k_x\,\mathrm{d}k_y\,\mathrm{d}k_z$, le nombre d'états varie de $2V/(2\pi)^3 \cdot \mathrm{d}k_x\,\mathrm{d}k_y\,\mathrm{d}k_z$.

Nous allons maintenant définir une notion très importante pour le fonctionnement des composants, la densité d'états. Elle est notée ρ. Le nombre d'états possibles par unité de volume pour les électrons de valence, quand l'énergie varie entre E et $E + \mathrm{d}E$, est égal à $\rho(E)\,\mathrm{d}E$. On écrit alors :

$$\rho(E)\,\mathrm{d}E = \frac{2}{(2\pi)^3} \cdot 4\pi k^2\,\mathrm{d}k$$

Dans le cas simple des électrons libres dans le cristal, l'énergie est purement cinétique et on peut donc écrire :

$$\mathrm{d}E = \hbar \cdot \sqrt{\frac{2E}{m}} \cdot \mathrm{d}k$$

On obtient alors :

$$\rho(E) = \frac{\sqrt{2}(m)^{3/2}}{\pi^2 \hbar^3} \cdot \sqrt{E} \qquad (2.19)$$

2.4.5 *La conduction et la notion de trou*

La conduction est le phénomène de base des composants électroniques. Elle sera étudiée dans trois modes de représentation selon la nature des dispositifs et selon leurs dimensions :

- représentation classique des charges en mouvement ;
- représentation semi-classique pour les solides cristallins ;
- représentation quantique pour les systèmes de très petites tailles ou désordonnés.

Dans la représentation classique, la conduction est le mouvement de charges ponctuelles, électrons ou ions. Les équations à appliquer sont celles de la dynamique de Newton et les équations de Maxwell pour calculer les forces électriques. La notion de trou est totalement étrangère à cette description. On introduira deux composantes pour le courant : le courant de conduction et le courant de déplacement.

Dans la représentation semi-classique utilisée dans l'étude des semi-conducteurs, le courant est défini comme la somme du courant de conduction et du courant de déplacement mais les charges contribuant véritablement à ce mécanisme ne sont plus qu'une petite partie des électrons du solide. Ce sont les électrons de valence dont l'énergie est en haut de la bande de valence ou en bas de la bande de conduction. Dans les systèmes de très petite taille, le courant est défini uniquement par sa représentation quantique et de nouveaux concepts seront alors nécessaires pour expliquer la conduction.

Comment expliquer l'origine du courant dans un solide ? Dans un premier temps, nous reprenons le cas du solide à une dimension car la représentation des phénomènes y est plus simple. Le réseau de courbes reliant l'énergie E et le vecteur d'onde k dans la zone de Brillouin est appelé diagramme de bandes. Il est caractéristique du solide étudié. La *figure 2.24* illustre deux cas possibles. Pour simplifier la figure, trois bandes sont représentées. Il peut y en avoir plus en fonction du nombre d'électrons de valence disponibles.

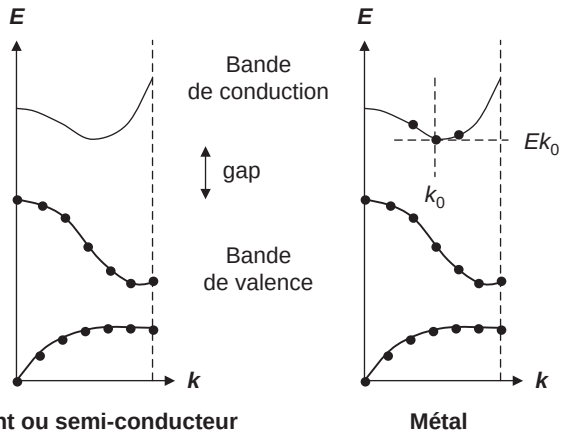


Figure 2.24 – Conduction dans un semi-conducteur ou un métal.

Dans le premier cas, illustré *figure 2.24*, la bande de valence est complètement remplie. La seule solution pour qu'un électron change d'énergie est qu'il passe dans la bande située au-dessus, appelée bande de conduction. En effet, deux électrons ne peuvent être dans le même état quantique puisque

les électrons sont des Fermions. L'intervalle séparant les deux bandes et noté E_g est appelé le gap. Il est de quelques eV dans les solides. Quand sa valeur est limitée, de l'ordre de 1 eV, le solide est un semi-conducteur car une conduction réduite est quand même possible. Quand sa valeur est plus importante, toute conduction est impossible car le changement d'énergie de l'électron est trop élevé. Le solide est alors un isolant.

Dans le deuxième cas, illustré *figure 2.24*, les états sont occupés par énergies croissantes, mais la bande supérieure appelée bande de conduction est incomplètement remplie. Rappelons que les valeurs possibles de k sont en nombre fini et que les valeurs correspondantes de E sont également en nombre fini. Dans ce cas, un électron de la bande de conduction peut changer légèrement d'énergie suite à une sollicitation extérieure. Nous verrons par la suite que la conduction électrique est alors possible. Le solide est dans ce cas un métal.

Dans un solide à une dimension, la représentation de la *figure 2.24* est naturelle à partir de la relation entre énergie et vecteur k. Pensons au modèle introduit dans le paragraphe 2.4.3 pour comprendre l'origine de cette représentation. Dans un solide à trois dimensions, la représentation est plus complexe. Il faut alors se placer dans une direction donnée de l'espace des vecteurs d'onde et faire varier le module de **k** noté k dans cette direction. Il est nécessaire d'inspecter toutes les directions possibles. En pratique, seules quelques directions privilégiées sont indiquées dans la littérature et correspondent aux nœuds du réseau réciproque.

Pour établir la relation entre la structure de bandes et la conduction électrique, il est nécessaire de revenir sur la notion de paquet d'ondes. La mécanique quantique postule que si ψ_1 et ψ_2 sont des états possibles d'un système, toute combinaison linéaire de ces états est également un état possible. On peut généraliser à une intégrale. Il est donc possible de construire un état en formant un paquet d'ondes de fonctions de Bloch pour décrire un électron localisé dans l'espace.

$$\psi(x,t) = \int_{-\infty}^{+\infty} f(k) \cdot u_{nk}(x) \cdot \mathrm{e}^{i\left(k \cdot x - \frac{E(k) \cdot t}{\hbar}\right)} \mathrm{d}k$$

On suppose que la fonction f présente des valeurs significatives uniquement dans un domaine proche de la valeur k_0. Le terme exponentiel oscille et la sommation conduit généralement à des interférences destructives. Le seul cas contraire est quand la phase de l'exponentielle reste constante quand k varie. Typiquement elle n'effectue alors qu'une oscillation ou moins autour de k_0. On peut alors écrire que la dérivée de la phase de l'exponentielle est nulle pour cette valeur de k.

$$x = \frac{t}{\hbar} \cdot \left(\frac{\mathrm{d}E}{\mathrm{d}k}\right)_{k = k_0}$$

Cette équation identifie les régions de l'espace dans lesquelles la fonction d'onde de l'électron n'est pas nulle. Tout se passe comme si l'électron se déplaçait avec une vitesse égale à :

$$v_g = \frac{1}{\hbar} \cdot \left(\frac{\mathrm{d}E}{\mathrm{d}k}\right)_{k = k_0}$$

Cette relation se généralise facilement :

$$\mathbf{v_g} = \frac{1}{\hbar} \mathbf{grad}\, E_{\mathrm{k} = \mathrm{k}_0} \tag{2.20}$$

Cette vitesse est la vitesse de groupe du paquet d'ondes. C'est également la vitesse de la particule classique associée à l'électron. Une manière plus rigoureuse d'établir cette relation fait appel à la

notion de valeur moyenne d'une grandeur physique en mécanique quantique. La valeur moyenne de l'opérateur vitesse égal à *p/m* est d'après les résultats du paragraphe 2.3 :

$$\langle v \rangle = \int \psi^{\star} \frac{\hbar}{i\,m} \frac{\mathrm{d}\psi}{\mathrm{d}x}\,\mathrm{d}x$$

La fonction d'onde choisie pour faire ce calcul est une fonction de Bloch caractérisée par un vecteur $\mathbf{k_0}$ et un indice de bande *n*. On peut alors calculer cette valeur de la vitesse en fonction de l'énergie. On obtient également la formule 2.20.

Si nous exprimons maintenant la relation entre *E* et *k*, appelée relation de dispersion, il est possible de calculer la vitesse de groupe. On examine le cas d'un électron situé dans la bande de conduction dans un métal ou un semi-conducteur et on se place au voisinage du minimum de la fonction énergie atteint pour *k* égal à k_0. Ce minimum est éventuellement atteint pour *k* nul mais ce n'est pas toujours le cas. La *figure 2.24* aide à comprendre les développements suivants. La fonction énergie peut s'écrire en première approximation :

$$E(k) \approx E(k_0) + (k - k_0)^2 \cdot A$$

On suppose donc que la courbe est voisine d'une parabole. Si on définit la masse effective par :

$$m_e^{\star} = \frac{\hbar^2}{2\,A}$$

On obtient :

$$v_g = \frac{\hbar}{m_e^{\star}}(k - k_0)$$

Nous comprendrons plus loin pourquoi cette grandeur est appelée masse.

Appliquons maintenant un champ électrique de valeur *E*. Un certain nombre d'électrons vont changer d'état. Pour les bandes pleines, il n'y a aucune possibilité puisque deux électrons ne peuvent occuper le même état en fonction du principe de Pauli. La conduction est donc possible uniquement pour des bandes incomplètement remplies. C'est par exemple le cas des électrons en bas de la bande de conduction, nombreux dans un métal et plus rares dans un semi-conducteur. Le travail de la force électrique est égal à la variation d'énergie de l'électron. Nous nous plaçons ici dans une description semi-classique de la conduction et nous utilisons la relation 2.20.

$$\mathrm{d}E = \hbar \cdot v_g \cdot \mathrm{d}k = -\mathrm{e} \cdot E \cdot v_g \cdot \mathrm{d}t$$

Comme,

$$\frac{\mathrm{d}v_g}{\mathrm{d}t} = \frac{\hbar}{m_e^{\star}} \cdot \frac{\mathrm{d}k}{\mathrm{d}t}$$

On obtient :

$$-eE = m_e^{\star} \cdot \frac{\mathrm{d}v_g}{\mathrm{d}t}$$

Cette équation est l'équation fondamentale de la dynamique à la condition de remplacer la masse par la masse effective. On comprend donc la notation « masse effective » introduite précédemment.

Considérons maintenant dans un semi-conducteur un électron situé en haut de la bande de valence et supposons que quelques états ont été libérés par passage d'électrons dans la bande de conduction. Le paragraphe suivant montre que ce processus est possible par agitation thermique. Ce sont bien évidemment les électrons de valence ayant les énergies les plus élevées donc en haut de la bande de valence qui peuvent le plus facilement passer dans la bande de conduction. La relation de dispersion sera exprimée de la même manière par une approximation parabolique. La différence avec le cas précédent est la suivante : le coefficient du terme quadratique est maintenant négatif étant donné la forme de la courbe. La *figure 2.24* en donne une illustration.

$$E(k) \approx E(k_0) - (k - k_0)^2 \cdot B$$

Un électron en haut de la bande de valence peut alors changer d'état puisque quelques états ont été libérés. Si on définit la masse effective de l'électron par :

$$m_e^\star = \frac{\hbar^2}{2B}$$

On peut écrire :

$$eE = m_e^\star \cdot \frac{\mathrm{d}v_g}{\mathrm{d}t}$$

Tout se passe comme si l'électron avait une masse équivalente négative ou une charge positive. On choisit généralement cette dernière possibilité et on représente la conduction en haut de la bande de valence par des particules équivalentes appelées trous puisqu'elles correspondent à l'absence d'électrons. Les trous ont donc une charge positive opposée à celle de l'électron et une masse effective donnée par la courbure de la courbe de dispersion. **Ce n'est qu'un artifice de représentation et la conduction est dans tous les cas assurée par la modification du vecteur d'onde des électrons du solide.** Cette représentation est cependant très commode et sera utilisée dans toute la suite de l'ouvrage. Tout se passe comme si le trou avait une masse positive mais différente de sa masse mécanique et une charge positive. Le trou se déplace alors dans le sens du champ électrique. Cette représentation est d'un usage courant pour comprendre le fonctionnement des composants.

Il est facile de reprendre le raisonnement précédent dans le cas réel d'un solide à trois dimensions. La représentation de la relation de dispersion est plus complexe. On représente en général l'énergie en fonction du module du vecteur **k** en choisissant pour **k** une direction privilégiée dans l'espace réciproque. Alors, pour une valeur donnée du module de **k**, il y a une énergie correspondante. Il y a en général plusieurs états possibles pour cette valeur et cette direction de **k**, au moins deux pour les deux valeurs de spin. La *figure 2.25* représente les courbes de dispersion pour le silicium. Pour donner une idée de la représentation à trois dimensions, les courbes sont représentées dans deux directions du réseau réciproque, les directions (111) et (100).

Pour obtenir une représentation complète il faudrait représenter l'énergie pour toutes les directions de **k**.

Il est instructif sur cet exemple de compter les électrons de valence par bande. Ce décompte doit intégrer les cas de dégénérescence, c'est-à-dire les cas correspondant à plusieurs états possibles pour une même valeur de **k** dans une bande donnée. L'origine des dégénérescences est un phénomène complexe non traité dans cette introduction.

Pour une valeur particulière k_0, on trouve dans le silicium un état deux fois dégénéré puis un état deux fois dégénéré et enfin un état quatre fois dégénéré. Au total, il y a donc 8 états possibles dans la zone de Brillouin. Il y a également 8 électrons de valence par maille puisqu'il y a deux atomes de

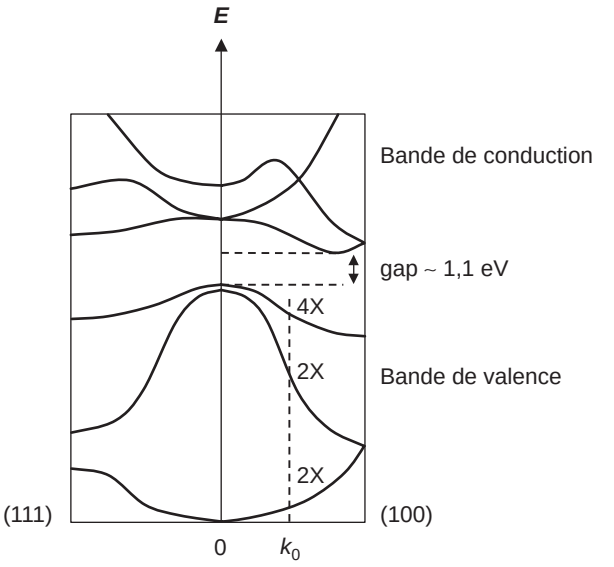


Figure 2.25 – Représentation à trois dimensions pour le silicium.

silicium par maille élémentaire. La bande de valence est donc pleine et la bande de conduction vide à température nulle. La valeur du gap est relativement faible, le silicium a donc bien le comportement d'un semi-conducteur.

La *figure 2.25* illustre une propriété particulière du silicium. La valeur minimale de l'énergie dans la bande de conduction ne correspond plus à **k** égal à zéro. Une transition ne peut donc se faire en conservant la valeur de **k**. Il faut faire intervenir une troisième particule pour conserver le vecteur d'onde dans l'interaction. Cette troisième particule est un phonon, pseudo particule associée aux vibrations du réseau. Le semi-conducteur est alors dit à gap indirect. Le lecteur est invité à consulter les ouvrages de physique du solide pour en savoir plus sur ce sujet (références **[3]** et **[4]**).

Dans un solide à trois dimensions, l'énergie au voisinage du minimum de la bande de conduction ou du maximum de la bande de valence s'exprime par :

$$E(\mathbf{k}) \approx E(\mathbf{k_0}) + \sum_{ij} \left(\frac{\partial^2 E(\mathbf{k})}{\partial k_i \partial k_j} \right)_{\mathbf{k} = \mathbf{k_0}} (k_i - k_{0i})(k_j - k_{0j})$$

Les indices *i* et *j* expriment les trois coordonnées. On définit ainsi le tenseur des masses effectives comme dans le cas à une dimension. Les masses effectives peuvent également se définir dans une direction particulière du vecteur **k**. Elles sont différentes de la masse de l'électron et les valeurs sont en général plus faibles.

2.4.6 *Les densités d'états*

Dans ce paragraphe, nous allons calculer le nombre d'états possibles pour les électrons de conduction et de valence dans le cas d'un semi-conducteur. La connaissance du nombre d'états dans une bande d'énergie donnée permettra d'en déduire le nombre d'électrons disponibles pour la conduction électrique et donc la conductivité du matériau. Rappelons que les électrons de conduction sont unique-

ment ceux qui sont situés en bas de la bande de conduction ou en haut de la bande de valence. À température nulle, il n'y en a pas.

Pour calculer le nombre d'états nous ferons quelques hypothèses simplificatrices. La première est de considérer les électrons dans des états proches des états d'électrons libres. Toutefois, on supposera que la masse des électrons et des trous est la masse effective pour tenir compte de l'effet du cristal. L'énergie s'écrit pour un électron de la bande de conduction :

$$E(k) = E_C + \frac{\hbar^2 k^2}{2 m_e^*}$$

Les surfaces d'énergie constante sont donc des surfaces sphériques si on suppose la masse effective isotrope dans l'espace des $\boldsymbol{k}$. Les états sont en nombre fini comme il a été vu dans le paragraphe 2.2.3. Nous y avons vu qu'il suffit de faire varier le vecteur $\boldsymbol{k}$ dans la zone de Brillouin pour obtenir tous les états. La densité d'états, c'est-à-dire le nombre de valeurs possibles de $\boldsymbol{k}$ dans un élément de volume de l'espace réciproque $\mathrm{d}k_x\,\mathrm{d}k_y\,\mathrm{d}k_z$ est dans une bande donnée :

$$\mathrm{d}n = \frac{V}{(2\pi)^3} \cdot \mathrm{d}k_x\,\mathrm{d}k_y\,\mathrm{d}k_z$$

Il faut multiplier par 2 pour tenir compte des deux spins possibles de l'électron. Les états vont être occupés en commençant par ceux d'énergies les plus basses. Les surfaces d'égale énergie dans notre modèle simplifié sont des sphères. Le nombre d'états compris entre E et $E + \mathrm{d}E$ est donc :

$$V\,\rho(E) \cdot \mathrm{d}E = 2 \cdot 4\pi k^2\,\mathrm{d}k \cdot \frac{V}{(2\pi)^3}$$

Comme :

$$E = E_C + \frac{\hbar^2 k^2}{2 m_e^*}$$

On obtient par unité de volume :

$$\rho(E) = \frac{\sqrt{2}\, m_e^{*\frac{3}{2}}}{\pi^2 \hbar^3} \cdot \sqrt{E - E_C} \qquad (2.21)$$

Le même raisonnement peut se faire dans une autre bande, on trouve une relation équivalente à la masse effective près, qui prend une autre valeur.

Nous allons maintenant étudier l'effet de la température. À basse température, la bande de conduction d'un semi-conducteur est vide et la bande de valence est pleine. Quand la température augmente, quelques électrons peuvent passer de la bande de valence à la bande de conduction ce qui autorise une conduction électrique. Ce phénomène se décrit en introduisant la fonction de Fermi qui exprime la probabilité qu'un électron occupe un état d'énergie E. Cette probabilité $f(E)$ s'écrit en fonction de la température T et d'un paramètre E_F appelé abusivement niveau de Fermi. Ce paramètre ne doit pas être confondu avec l'énergie de Fermi E_{F0}, étudiée précédemment, même si ces deux valeurs sont assez proches.

$$f(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{k_B T}\right)} \qquad (2.22)$$

La fonction de Fermi est représentée *figure 2.26* pour T égal à zéro et pour deux températures différentes de zéro, T_2 étant la plus élevée. En fait, le paramètre E_F est le potentiel chimique du gaz d'électrons comme il sera expliqué paragraphe 2.5. On calcule alors le nombre d'électrons par unité de volume dans la bande de conduction :

$$n = \int_{E_C}^{\infty} \rho(E) \cdot f(E) \cdot \mathrm{d}E$$

Un calcul simple, en supposant le terme exponentiel très supérieur à l'unité, conduit à :

$$n = N_e \cdot \exp\left(\frac{E_F - E_C}{k_B T}\right) \tag{2.23}$$

avec,

$$N_e = 2 \cdot \left(\frac{m_e^* k_B T}{2 \pi \hbar^2}\right)^{3/2} \tag{2.24}$$

L'ordre de grandeur de N_e est de 10^{19} par cm^3 à température ambiante.

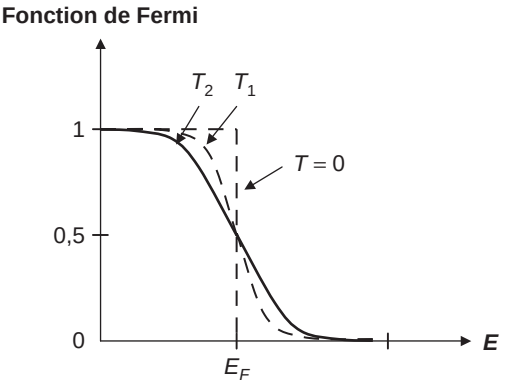


Figure 2.26 – Distribution de Fermi.

Le même raisonnement peut s'appliquer aux trous de la bande de valence. On obtient :

$$p = N_h \cdot \exp\left(-\frac{E_F - E_V}{k_B T}\right) \tag{2.25}$$

avec

$$N_h = 2 \cdot \left(\frac{m_h^* k_B T}{2 \pi \hbar^2}\right)^{3/2} \tag{2.26}$$

Dans ces formules, le choix de la masse effective n'est pas toujours évident puisqu'il dépend de la branche choisie dans une bande et de la direction d'observation dans l'espace des **k**.

Dans un semi-conducteur pur, le nombre de trous de la bande de valence est égal au nombre d'électrons de la bande de conduction puisqu'un trou est en fait créé par le passage d'un électron de la bande de valence à la bande de conduction. On en déduit donc $n_e = n = n_i$ et en conséquence :

$$E_i = \frac{E_C + E_V}{2} + \frac{3}{4} k_B T \cdot \ln\left(\frac{m_h^*}{m_e^*}\right)$$

On simplifie souvent cette formule en négligeant le deuxième terme. On dit alors que le niveau de Fermi d'un semi-conducteur pur est au milieu de la bande interdite. La concentration intrinsèque n_i s'exprime de la manière suivante :

$$n_i^2 = N_e N_h \exp -\left(\frac{E_C - E_V}{k_B T}\right)$$

Il est alors possible d'exprimer les concentrations en fonction de l'énergie E_i.

$$n = n_i \cdot \exp\left(\frac{E_F - E_i}{k_B T}\right) \qquad (2.27)$$

$$p = n_i \cdot \exp\left(\frac{E_i - E_F}{k_B T}\right)$$

Ces deux relations seront souvent utilisées dans la suite de cet ouvrage.

2.4.7 *Le dopage des semi-conducteurs*

Dans certains cas, il est nécessaire de créer des électrons de conduction ou des trous en quantité supérieure à celle donnée par l'agitation thermique. Une technique consiste à insérer dans le semi-conducteur des éléments dits dopants susceptibles de fournir des charges.

La *figure 2.27* représente à deux dimensions l'effet de l'introduction d'un atome dopant dans le silicium. Le silicium possède quatre électrons de valence et, par liaison covalente, chaque atome de silicium est lié à quatre atomes voisins. Quand un atome est introduit comportant cinq électrons de valence, c'est par exemple le cas d'un atome de phosphore, le cinquième électron de valence se trouve peu lié dans le cristal comme le monte la *figure 2.26* et est disponible pour la conduction électrique. L'atome est dit donneur. La traduction de cet effet dans la représentation des niveaux d'énergie est l'apparition d'un niveau possible pour cet électron proche du niveau inférieur de la bande de conduction. Un écart de 50 meV est une valeur typique.

De manière analogue, si on introduit un atome comportant trois électrons de valence, comme il est représenté en 2.27, une des liaisons covalente sera rompue dans le semi-conducteur et de ce fait un transfert d'électron sera probable de la bande de valence vers ce site pour compléter la liaison. Ce comportement est donc celui d'un trou et un niveau énergétique apparaît à proximité du haut de la bande de valence. L'atome est alors dit accepteur.

Un semi-conducteur dopé avec des impuretés pentavalentes est dit de type *n*. Si on appelle *n* la densité d'électrons dans la bande de conduction, *p* la densité de trous dans la bande de valence et N_D la densité de donneurs, on peut écrire pour assurer la conservation de la charge :

$$n = N_D + p$$

Dans la pratique, la densité de trous est très inférieure à la densité de dopants. On obtient donc :

$$n = N_D$$

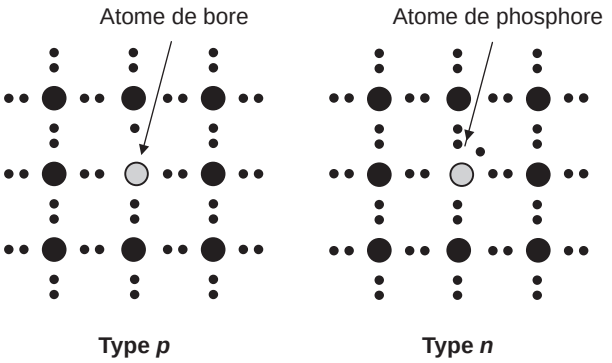


Figure 2.27 – Dopage de type p et de type n.

Cette relation indique que les électrons de conduction sont en première approximation fournis par les dopants.

De même, pour un semi-conducteur de type *p*, on obtient :

$$p = N_A$$

Pour terminer ce paragraphe, nous allons préciser la position du potentiel chimique pour les semi-conducteurs dopés. Pour cela, il suffit d'appliquer les relations précédentes en tenant compte des relations générales 2.24 et 2.26 données précédemment. On obtient alors pour un semi-conducteur de type *n* :

$$E_C - E_F = k_B T \cdot \ln\left(\frac{N_e}{N_D}\right)$$

Le niveau de Fermi du semi-conducteur de type *n* est donc très proche du niveau de la bande de conduction. Pour un semi-conducteur de type *p*, on obtient de même :

$$E_F - E_V = k_B T \cdot \ln\left(\frac{N_h}{N_A}\right)$$

Pour un semi-conducteur de type *p*, le niveau de Fermi est très proche du niveau de la bande de valence car k_BT est égal à 26 mV à température ambiante. La *figure 2.28* résume les cas possibles.

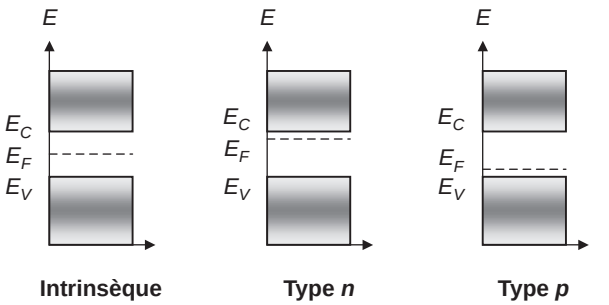


Figure 2.28 – Semi-conducteurs dopés et niveaux de Fermi.

2.4.8 *Le courant dans un semi-conducteur*

Nous avons, dans les paragraphes précédents, calculé les densités d'électrons et de trous dans les solides. Il est maintenant possible de calculer les courants. Le courant électrique dans un solide est composé de deux termes : un courant de conduction dû au déplacement des électrons et des trous sous l'effet du champ électrique et un courant de diffusion.

Examinons tout d'abord le courant de conduction. Le paragraphe 2.4.4 a montré que l'équation du mouvement de l'électron de la bande de conduction pouvait s'écrire :

$$-e\,\mathbf{E} = \hbar \cdot \frac{\mathrm{d}\mathbf{k}}{\mathrm{d}t}$$

Il en est de même pour les trous de la bande de valence. Cette équation conduit à un déplacement général des vecteurs d'onde et à un déplacement de la sphère de Fermi. Si ce modèle était complet, il conduirait à une augmentation constante et illimitée du vecteur d'onde, ce qui est contraire à l'expérience. En fait, les électrons et les trous sont soumis dans un solide à diverses collisions : collisions avec les impuretés, collisions avec les vibrations du réseau, collisions avec les défauts du cristal. L'ensemble de ces phénomènes est équivalent à une force de rappel qui s'ajoute à la force électrique et l'équation de transport s'écrit :

$$-e\,\mathbf{E} - \frac{\hbar(\mathbf{k}-\mathbf{k_0})}{\tau} = \hbar\,\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{k}-\mathbf{k_0})$$

Dans cette équation, on exprime la variation du vecteur d'onde par rapport à sa valeur initiale et on fait intervenir une constante τ, homogène à un temps. La solution permanente est donc :

$$\mathbf{k}-\mathbf{k_0} = -\frac{e\tau}{\hbar}\cdot\mathbf{E}$$

Si on exprime la relation entre la vitesse de groupe et le vecteur d'onde, on écrit pour un électron libre :

$$\mathbf{v}-\mathbf{v_0} = -\frac{e\tau}{m}\cdot\mathbf{E}$$

Pour un électron ou un trou dans un solide, il suffit de remplacer la masse par la masse effective. Il faut également noter que cette vitesse est la vitesse de dérive ($\mathbf{v} - \mathbf{v_0}$) et non pas la vitesse correspondant à l'agitation thermique représentée par le module de $\mathbf{v_0}$. La *figure 2.29* représente la modification du vecteur d'onde et le déplacement de la sphère de Fermi suite à l'application d'un champ électrique.

Il est maintenant possible d'exprimer la densité de courant associée à n électrons par unité de volume :

$$\mathbf{J_e} = \frac{n\,e^2\tau}{m_e^\star}\cdot\mathbf{E}$$

On reconnaît bien évidemment la loi d'Ohm. Le terme $e\,\tau / m_e^\star$ est appelé mobilité et noté μ, sa valeur varie de 100 cm^2/V/s à 30 000 cm^2/V/s quand on passe de matériaux de mauvaise qualité à des matériaux de bonne qualité. Une formule équivalente peut être établie pour les trous. Les valeurs de la mobilité sont plus élevées pour les électrons que pour les trous. Elle vaut 1 500 cm^2/V/s pour les électrons dans le silicium et 450 cm^2/V/s pour les trous dans le silicium.

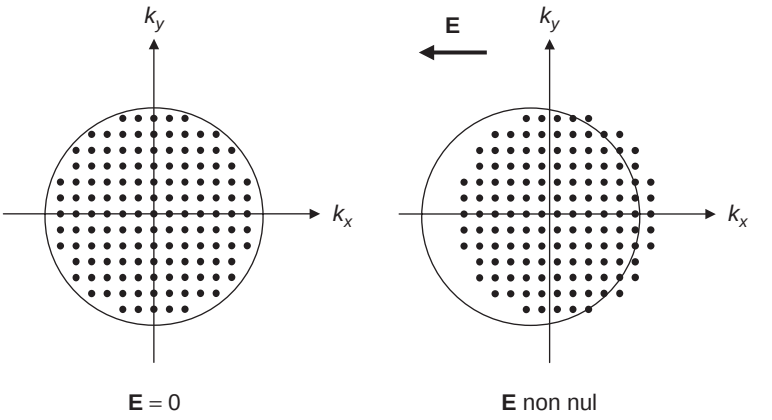


Figure 2.29 – Modification du vecteur d'onde et courant.

Le courant de conduction n'est pas seul responsable de la conduction électrique. Il s'y ajoute un courant dit de diffusion quand deux régions présentent des concentrations en porteurs différentes. Les porteurs, électrons ou trous, diffusent de la région de forte concentration vers la région de faible concentration et cela proportionnellement à la différence des concentrations. Le courant de diffusion s'écrit donc pour les électrons :

$$\mathbf{J_e} = e D_n \cdot \mathbf{grad}\, n$$

Il s'écrit pour les trous :

$$\mathbf{J_h} = -e D_p \cdot \mathbf{grad}\, p$$

Le courant total est la somme du courant de conduction et du courant de diffusion. Les signes s'expliquent en examinant les différents cas possibles : électrons et trous. On obtient donc pour les électrons :

$$\mathbf{J_e} = \frac{n e^2 \tau}{m_e^*} \cdot \mathbf{E} + e D_n \cdot \mathbf{grad}\, n \qquad (2.28)$$

$$\mathbf{J_e} = \mu_n \cdot \mathbf{E} + e D_n \cdot \mathbf{grad}\, n$$

On obtient de même pour les trous :

$$\mathbf{J_h} = \frac{p e^2 \tau}{m_h^*} \cdot \mathbf{E} - e D_p \cdot \mathbf{grad}\, p \qquad (2.29)$$

$$\mathbf{J_h} = \mu_p \cdot \mathbf{E} - e D_p \cdot \mathbf{grad}\, p$$

Le courant total est la somme des deux contributions. Il est possible d'obtenir une relation entre mobilité et coefficient de diffusion. Cette démonstration due à Einstein n'est pas présentée dans cet ouvrage.

$$D_n = \frac{k_B T}{\mu_n} \qquad (2.30)$$

Que devient le courant à l'échelle quantique ? Il faut alors revenir au courant d'une charge élémentaire dans un état caractérisé par son impulsion p. la définition classique du courant est :

$$J = e\frac{p}{m}$$

On définit donc par la règle de correspondance l'opérateur courant dont on peut calculer la valeur moyenne :

$$\langle J \rangle = \frac{e}{m}\int \psi^{\star}\,\frac{\hbar}{i}\frac{\partial}{\partial x}\psi\,\mathrm{d}x$$

Si on applique cette relation à un état donné d'un système unidimensionnel, la fonction d'onde est :

$$\psi(x) = \frac{1}{\sqrt{L}}\exp^{ikx}$$

On obtient alors :

$$\langle J \rangle = \frac{e}{m}\hbar\, k \tag{2.31}$$

2.5 Ensemble de particules : potentiel chimique et niveau de Fermi

Le potentiel chimique est une notion fondamentale dans le fonctionnement des dispositifs solides. Le potentiel chimique est intimement lié aux tensions électriques appliquées aux bornes des composants électroniques et est un paramètre significatif expliquant le fonctionnement de ces composants. Pour en comprendre la signification, il est nécessaire de revenir à quelques notions de base de la thermodynamique statistique

On considère donc N électrons susceptibles d'occuper des états énergétiques caractérisés par leur énergie E_i et on suppose que pour chaque valeur E_i il y a g_i états différents. On dit que l'état est g_i fois dégénéré. Pensons par exemple au cas des électrons libres dans l'espace à trois dimensions. Toutes les valeurs de k sur une sphère donnée correspondent à des états de même énergie. La dégénérescence est donc un nombre élevé. Le problème que nous avons à résoudre est de déterminer combien d'électrons occupent les états d'énergie E_i.

Ces nombres sont notés n_i. Pour résoudre ce problème, nous appliquerons le principe suivant. La thermodynamique statistique postule que les états se remplissent de telle manière que le nombre de combinaisons possibles conduisant à la même énergie totale soit maximum. On suppose que l'énergie totale est donnée ainsi que le nombre total d'électrons. On suppose également que les électrons sont indiscernables. Le nombre de façons de placer n_i électrons indiscernables parmi g_i états possibles est donné par la relation :

$$combinaisons = \frac{g_i!}{n_i!(g_i - n_i)!}$$

Si maintenant on prend en compte toutes les énergies possibles, on obtient un nombre de combinaisons possibles égal à :

$$\Omega = \prod \frac{g_i!}{n_i!(g_i - n_i)!}$$

On cherche alors à maximiser ce nombre en choisissant les valeurs de n_i particulières. Cette optimisation se fait sous les contraintes suivantes :

$$\sum_i n_i = N$$

$$\sum_i n_i E_i = E$$

On appliquera alors la méthode des multiplicateurs de Lagrange qui permet d'optimiser une fonction sous contraintes. Pour simplifier le calcul, on optimisera $\ln \Omega$ et non Ω, ce qui est équivalent. Si on appelle α et β les deux multiplicateurs, on cherche donc à maximiser la fonction :

$$\Theta = \sum_i (\ln g_i! - \ln n_i! - \ln (g_i - n_i)!) + \alpha\left(N - \sum_i n_i\right) + \beta\left(E - \sum_i E_i\right)$$

Pour aller plus loin dans le calcul, il faut appliquer la formule de Stirling qui est valable uniquement pour des valeurs élevées de n.

$$\ln n! \approx n \ln n - n$$

Quand on dérive la fonction Θ par rapport aux variables du problème, c'est-à-dire les quantités d'électrons n_i, on obtient :

$$-\ln n_i + \ln (g_i - n_i) - \beta E_i - \alpha = 0$$

On obtient donc pour tous les i du problème :

$$\ln \frac{g_i - n_i}{n_i} = \alpha + \beta E_i$$

On en déduit alors :

$$n_i = \frac{g_i}{1 + \exp^{\alpha + \beta E_1}}$$

Il faut maintenant relier ces résultats aux grandeurs thermodynamiques macroscopiques : la température, le potentiel chimique et l'entropie.

La thermodynamique nous apprend que l'énergie d'un ensemble de N particules dans un volume V est une fonction de V, N et S. La variable S, entropie du système, est reliée au nombre d'états possibles par la relation suivante :

$$S = k_B \ln \Omega$$

La constante k_B est la constante de Boltzmann. C'est une grandeur fondamentale de la physique qui relie un nombre d'états à une grandeur physique énergétique. Sa valeur est la suivante :

$$k_B = 1{,}380664 \times 20^{-23}\ \mathrm{J/K}$$

La variation d'énergie peut alors s'écrire :

$$\mathrm{d}E = \frac{\partial E}{\partial S}\mathrm{d}S + \frac{\partial E}{\partial N}\mathrm{d}N + \frac{\partial E}{\partial V}\mathrm{d}V$$

On définit alors à partir des dérivées partielles les grandeurs macroscopiques : température, potentiel chimique et pression.

$$T = \frac{\partial E}{\partial S}$$

$$E_F = \frac{\partial E}{\partial N}$$

$$p = -\frac{\partial E}{\partial V}$$

Plaçons-nous dans le cas où le volume est constant alors :

$$\frac{\mathrm{d}S}{k_B} = \frac{\mathrm{d}E}{k_B T} - \frac{E_F}{k_B T}\mathrm{d}N$$

On peut également calculer dS à partir de la définition de l'entropie et des résultats de l'optimisation. On obtient alors :

$$\frac{\mathrm{d}S}{k_B} = \beta\,\mathrm{d}E + \alpha\,\mathrm{d}N$$

On en déduit donc :

$$\beta = \frac{1}{k_B T}$$

$$\alpha = -\frac{E_F}{k_B T}$$

Finalement, le nombre d'électrons par niveau d'énergie est :

$$n_i = \frac{g_i}{1 + \exp\dfrac{E_i - E_F}{k_B T}}$$

On trouve ainsi de manière rigoureuse l'expression indiquée dans le paragraphe précédent. La *figure 2.30* illustre le mécanisme de remplissage des états en fonction de la température.

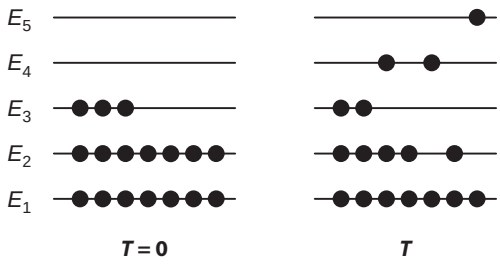


Figure 2.30 – Remplissage des états en fonction de la température.

Cette démonstration permet également d'établir une propriété fondamentale liant potentiel chimique et tension appliquée, qui sera largement utilisée par la suite. Considérons deux volumes contenant $N1$ et $N2$ électrons dans des volumes fixes, portés à même température et caractérisés par leurs potentiels chimiques E_{F1} et E_{F2}. Faisons passer adiabatiquement, c'est-à-dire avec une variation d'entropie nulle, un électron de 1 vers 2. Pour cela, on applique une tension $V_1 - V_2$ entre 1 et 2 comme le montre la *figure 2.31*.

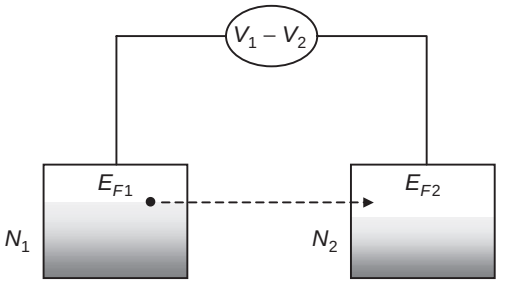


Figure 2.31 – Potentiel chimique et différence de potentiel.

L'énergie totale du système a varié de $e(V_2 - V_1)$, énergie apportée par l'extérieur. La variation d'énergie s'exprime aussi par :

$$\Delta E = T_1 \Delta S_1 + T_2 \Delta S_2 - p_1 \Delta V_1 - p_2 \Delta V_2 + E_{F2} - E_{F1}$$

Comme l'opération est adiabatique et comme les volumes sont fixes, il reste :

$$E_{F2} - E_{F1} = -e(V_2 - V_1) \tag{2.32}$$

Cette relation permet de calculer les différences de potentiels chimiques entre deux solides échangeant des charges et soumis à une différence de potentiel.

Pour terminer cet important paragraphe, il est possible d'étudier le cas limite correspondant à une température nulle. Dans ce cas, la fonction de Fermi tend vers une fonction échelon. Seuls les états inférieurs au potentiel chimique sont occupés, les états d'énergies supérieures sont strictement vides comme cela est représenté sur la *figure 2.30*. Dans ce cas et seulement dans ce cas, énergie de Fermi et potentiel chimique ont la même valeur.

Dans le cas d'un métal, ces deux grandeurs peuvent être liées par la formule approchée quand $k_B T$ est petit devant E_{F0}.

$$E_F = E_{F0}\left(1 - \frac{\pi^2}{12}\left(\frac{k_B T}{E_{F0}}\right)^2\right)$$

2.6 L'effet tunnel : un effet quantique à prendre en compte

L'effet tunnel est un phénomène de nature quantique très important dans la micro et nano-électronique. Il est à prendre en compte dans les technologies au-delà du 90 nm car il est responsable du courant de fuite de grille des transistors de nouvelle génération. Pour le comprendre, il faut penser à la nature ondulatoire d'un électron ou d'un trou. La fonction d'onde a une certaine extension

spatiale. Dans des dispositifs de très faible dimension, typiquement plus petite que l'extension de la fonction d'onde, le module de la fonction d'onde peut avoir une valeur non nulle dans des zones ou la particule au sens classique du terme ne peut être présente. Sommé sur un grand nombre d'électrons, cet effet peut alors conduire à un courant non nul.

Considérons un électron dans un monde à une dimension et soumis à un potentiel tel que celui représenté *figure 2.32*.

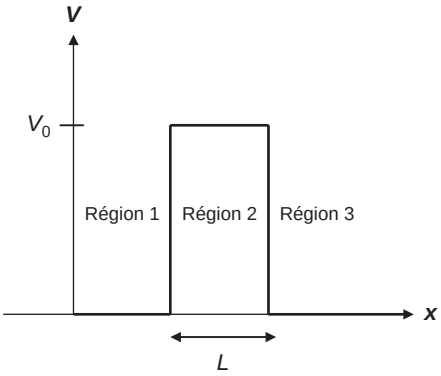


Figure 2.32 – Modèle simple pour l'effet tunnel.

L'espace est divisé en trois régions 1, 2 et 3 en allant de la gauche vers la droite. Le potentiel est nul dans les régions 1 et 3 et égal à V_0 dans la région 2. L'équation de Schrödinger s'écrit alors :

$$i\hbar \frac{\partial \psi(x,t)}{\partial t} = \left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right) \psi(x,t)$$

On s'intéresse aux états stationnaires d'énergie donnée E et on pose :

$$E = \frac{\hbar^2}{2m} \varepsilon \qquad V(x) = \frac{\hbar^2}{2m} U(x)$$

L'équation s'écrit alors selon les régions :

$$\text{Région 1 :} \quad \frac{\partial^2 \psi}{\partial x^2} + \varepsilon \psi = 0$$

$$\text{Région 2 :} \quad \frac{\partial^2 \psi}{\partial x^2} + (\varepsilon - U_0) \psi = 0$$

$$\text{Région 3 :} \quad \frac{\partial^2 \psi}{\partial x^2} + \varepsilon \psi = 0$$

On cherche les solutions sous la forme suivante :

$$\text{Région 1 :} \quad \psi(x) = S \cdot \exp^{i\sqrt{\varepsilon}\, x}$$

$$\text{Région 3 :} \quad \psi(x) = \exp^{-i\sqrt{\varepsilon}\, x} + R \cdot \exp^{i\sqrt{\varepsilon}\, x}$$

$$\text{Région 2 :} \quad \psi(x) = A \cdot \exp^{\eta x} + B \cdot \exp^{-\eta x}$$

avec $\eta = \sqrt{U_0 - \varepsilon}$ et $V_0 = \frac{\hbar^2}{2m} U_0$.

La forme de ces fonctions n'est pas tout à fait évidente et suppose que l'on s'intéresse uniquement à une onde venant des x positifs et se propageant vers les x négatifs. On a également supposé que la barrière était plus haute que l'énergie considérée mais le problème pourrait se résoudre également dans le cas contraire. Cette hypothèse correspond à la définition de l'effet tunnel.

Pour résoudre le système d'équations, il faut déterminer les 5 coefficients. On écrit la continuité de la fonction d'onde aux deux interfaces, ce qui donne deux équations puis la continuité de la dérivée, ce qui donne deux autres équations et enfin on normalise la fonction d'onde, ce qui donne une cinquième équation. Le calcul complet n'est pas détaillé et seul le résultat est donné pour T qui mesure la fraction de l'onde qui a réussi à traverser la barrière.

$$T = |S|^2 = \frac{4\varepsilon(U_0 - \varepsilon)}{4\varepsilon(U_0 - \varepsilon) + U_0^2 \, sh^2 \eta L} \tag{2.33}$$

Quand le produit ηL tend vers 0, l'onde est presque transmise en totalité à travers la barrière. Ce cas correspond à une épaisseur de barrière très faible. On comprend donc que l'effet tunnel se manifeste dans des dispositifs de dimensions très faibles. Une application numérique incluant une différence d'énergie de 1 eV conduit à une valeur de L de l'ordre du nm. Cette valeur est approximativement l'épaisseur attendue des oxydes de grille pour les futures générations de transistors.

2.7 Les densités d'états dans les systèmes nanométriques

Les technologies actuelles permettent de réaliser des dispositifs qui étaient jusqu'à maintenant des cas d'école à savoir les systèmes à deux, une ou zéro dimension. Un plan d'atomes isolé est un exemple de système à deux dimensions obtenu par exemple par un plan de graphène. Un nanofil ou un nanotube sont des exemples de systèmes à une dimension. Une dimension veut dire que l'électron a un seul degré de liberté. Sa fonction d'onde est confinée dans les dimensions transverses du fil ce qui conduit à des valeurs quantifiées discrètes d'une partie de son énergie.

2.7.1 *Système à 0 dimension*

Une boîte quantique est un système à zéro dimension. On peut le considérer comme la limite d'un puits de potentiel quand la dimension tend vers zéro. Les états énergétiques sont alors discrets et dégénérés comme il est représenté *figure 2.33*. Le calcul des énergies et des états a été présenté paragraphe 2.4.3.

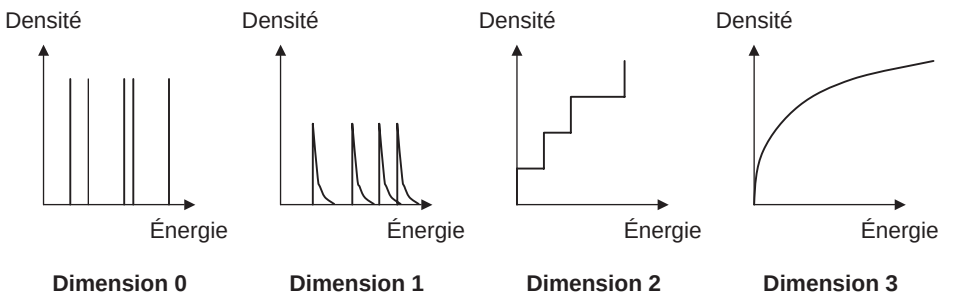


Figure 2.33 – États et densités d'états pour des systèmes de dimensions différentes.

2.7.2 *Système à une dimension*

Ce cas a été traité en introduction dans le paragraphe 2.4.2. Si on se place dans le cas simple des électrons libres, on peut écrire pour la partie « libre » de l'énergie :

$$E = \frac{\hbar^2 k^2}{2m}$$

Les conditions de Born-von Kármán conduisent à quantifier les valeurs possibles de k.

$$k = 2\pi \frac{n}{L}$$

Le nombre d'états possibles pour une variation ΔE est donc, étant donné les deux spins possibles :

$$\Delta n = 2 \frac{\Delta k}{2\pi} L = \frac{1}{\pi} \frac{\Delta k}{\Delta E} L$$

La densité d'états par unité de longueur est donc :

$$\rho(E) = \frac{1}{\pi} \frac{1}{\left(\frac{dE}{dk}\right)}$$

Soit en exprimant la dérivée en fonction de l'énergie :

$$\rho(E) = \frac{1}{\pi} \frac{1}{\hbar} \sqrt{\frac{2m}{E}}$$

En fait, les électrons ne sont pas libres et une partie de l'énergie est l'énergie de confinement. Elle est dans ce cas mesurée par deux nombres quantiques n_1 et n_2. La relation entre l'énergie et le vecteur d'onde n'est plus la relation :

$$E = \frac{\hbar^2 k^2}{2m}$$

Elle s'écrit :

$$E - E_{n_1 n_2} = \frac{\hbar^2 k^2}{2m}$$

La densité d'état s'écrit alors :

$$\rho(E) = \frac{1}{\pi} \frac{1}{\hbar} \sqrt{2m} \sum_{n_1 n_2} \frac{1}{\sqrt{E - E_{n_1 n_2}}} \gamma(E - E_{n_1 n_2})$$

La fonction $\gamma(E - E_{n1n2})$ est nulle quand l'énergie est inférieure à E_{n1n2} et égale à un quand elle est supérieure.

La *figure 2.34* montre des mesures de densité effectuées sur des nanotubes de carbone.

Un exemple est donné *figure 2.34* pour des nanotubes de carbone. Ces dispositifs sont décrits dans le chapitre 12.

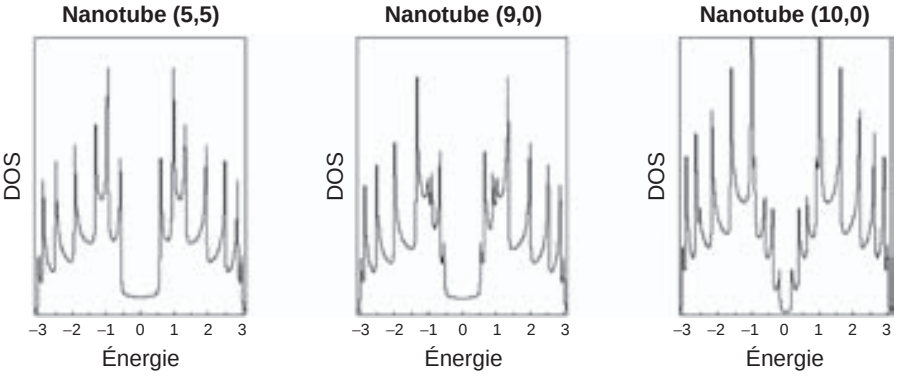


Figure 2.34 – Densités d'états des nanotubes de carbone.

2.7.3 *Systèmes à deux dimensions*

Dans ce cas, l'électron a deux degrés de liberté et l'énergie est confinée selon une seule dimension ce qui introduit des termes E_i. La partie non confinée de l'énergie d'un électron s'écrit alors :

$$E = \frac{\hbar^2 \mathbf{k}^2}{2m}$$

$$\mathbf{k}^2 = k_x^2 + k_y^2$$

Les composantes k_x et k_y sont quantifiées en fonction de la surface considérée.

$$k_x = \pm \frac{2\pi}{L_x} n_x$$

$$k_y = \pm \frac{2\pi}{L_y} n_y$$

Le nombre d'états possibles quand k_x varie de Δk_x et quand k_y varie de Δk_y est donc :

$$\Delta n = \Delta k_x \frac{L_x}{2\pi} \Delta k_y \frac{L_y}{2\pi}$$

Soit par unité de surface et en tenant compte des deux états possibles de spins,

$$\Delta n = \frac{2}{(2\pi)^2} \Delta k_x \Delta k_y$$

Si on considère maintenant l'ensemble des électrons répartis dans le cercle de Fermi, ils se répartissent sur les états possibles en remplissant en priorité les états les moins énergétiques. Une augmentation d'énergie maximale possible Δk correspond à une augmentation du nombre d'états dans la couronne de rayon k et d'épaisseur Δk. Cette couronne a pour surface $2\pi k \Delta k$.

Le nombre d'états y est donc :

$$\Delta n = 2 \frac{2\pi k \Delta k}{(2\pi)^2}$$

Comme,

$$E = \frac{\hbar^2 k^2}{2m}$$

On obtient :

$$\Delta n = \frac{m}{\pi \hbar^2} \Delta E$$

La densité d'états est donc constante en fonction de l'énergie. En fait, l'énergie totale est la somme de cette énergie et de l'énergie de confinement.

$$E = E_n + \frac{\hbar^2}{2m} k^2$$

On obtient donc pour la densité d'états la relation :

$$\rho(E) = \frac{m}{\pi \hbar^2} \sum_n \gamma(E - E_n)$$

Elle a donc une forme d'escalier comme le montre la *figure 2.33*.

Les systèmes de dimension 3 ont été largement étudiés dans les paragraphes précédents.

2.8 Les méthodes de calcul des composants et des circuits

2.8.1 *Conservation du courant et additivité du potentiel*

Ce sont les deux règles de base à appliquer. Pour en comprendre l'application, prenons l'exemple d'un système formé par la mise en contact électrique d'un condensateur plan, d'une résistance en forme de barreau, les deux étant reliés par des conducteurs parfaits. La *figure 2.35* représente une coupe effectuée dans ce système à trois dimensions.

On suppose que le champ électrique est nul à l'extérieur des composants, ce qui n'est pas tout à fait vrai dans la réalité. On suppose également que les conducteurs parfaits représentés en gris clair sont équipotentiels. Par définition même du potentiel on peut écrire :

$$V_0 = V_1 + (V_0 - V_1)$$

Cette relation triviale a pour seul intérêt de relier la tension aux bornes du condensateur à la tension présente dans le conducteur parfait. On peut écrire également entre les armatures du condensateur plan :

$$E = \frac{V_1 - V_0}{d}$$

Dans cette relation, d est l'espace entre les électrodes. Le barreau résistif présente une résistance R dont la valeur est donnée par la relation classique :

$$R = \rho \frac{L}{S}$$

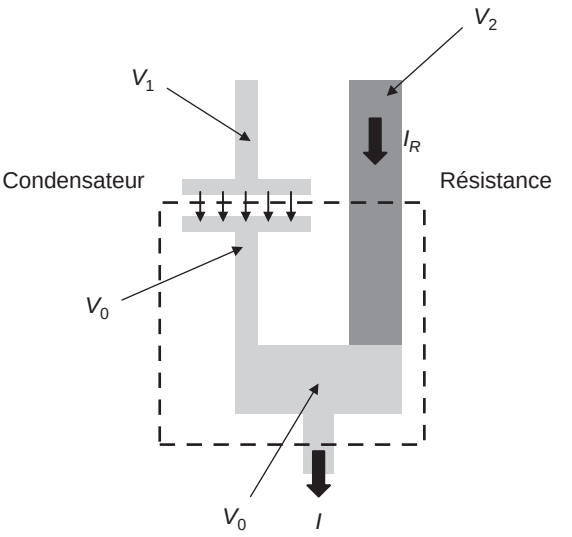


Figure 2.35 – Exemple de circuit.

Dans cette relation, L est la longueur du barreau, S sa section et ρ est la résistivité du barreau conducteur. On écrit alors en appelant I_R le courant traversant la résistance :

$$V_2 - V_0 = R \cdot I_R$$

La seule difficulté de cette formule est le choix des signes. Elle est vraie quand le sens positif choisi pour le courant correspond à un flux de charges positives. Autrement dit, le courant se déplace du potentiel le plus positif vers le potentiel le moins positif. Les électrons se déplacent donc dans le sens inverse du sens positif choisi.

Nous pouvons maintenant appliquer la loi de conservation du courant au sens de l'électromagnétisme en reprenant le résultat fondamental du paragraphe 2.2. La surface fermée choisie est indiquée en pointillés sur la *figure 2.34*. Elle passe entre les armatures du condensateur et à travers la résistance et les conducteurs. Le flux du courant total, courant de conduction et courant de déplacement, est nul à travers cette surface.

$$\varepsilon A \frac{\mathrm{d}E}{\mathrm{d}t} + I_R = I$$

Les signes intègrent le sens du flux, entrant ou sortant. Les relations précédentes permettent d'écrire :

$$\varepsilon A \frac{1}{d} \frac{\mathrm{d}(V_1 - V_0)}{\mathrm{d}t} + \frac{V_2 - V_0}{R} = I$$

On reconnaît la valeur de la capacité C du condensateur.

$$C \frac{\mathrm{d}(V_1 - V_0)}{\mathrm{d}t} + \frac{V_2 - V_0}{R} = I$$

Cette équation est l'équation de base du système. Supposons maintenant que cet ensemble soit relié à une source de tension comme le montre la *figure 2.36*.

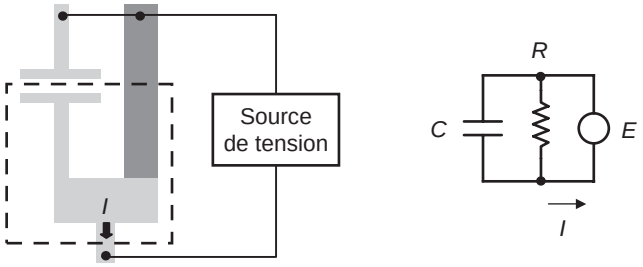


Figure 2.36 – Schéma électrique équivalent.

Dans ce cas, les tensions V_1 et V_2 étant égales, on écrit :

$$C\frac{\mathrm{d}E}{\mathrm{d}t} + \frac{E}{R} = I$$

Cette équation permet donc de calculer le courant si la tension est donnée.
Prenons un autre exemple simple, le condensateur et la résistance sont cette fois en série comme le montre la *figure 2.37*.

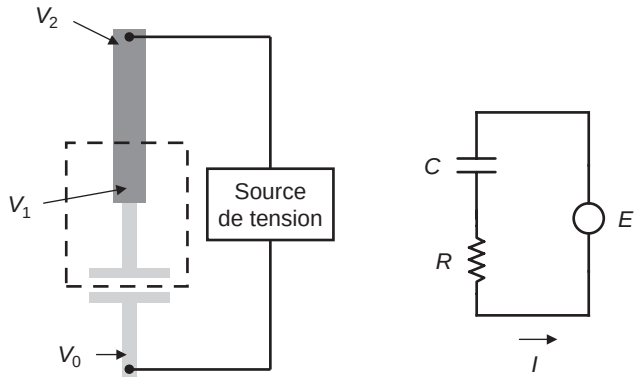


Figure 2.37 – Circuit série.

La *figure 2.37* indique comment on applique la règle d'addition des potentiels et la règle de conservation du courant pour écrire :

$$E = V_2 - V_1 + V_1 - V_0$$

$$A\frac{\varepsilon}{d}\frac{\mathrm{d}(V_1 - V_0)}{\mathrm{d}t} = I$$

$$V_2 - V_1 = RI$$

Au total,

$$\frac{\mathrm{d}E}{\mathrm{d}t} = R\frac{\mathrm{d}I}{\mathrm{d}t} + \frac{I}{C}$$

De manière plus générale, dans un circuit complexe comme le montre la *figure 2.38*, on appliquera la règle de conservation du courant aux nœuds du circuit et on écrira les équations de chaque branche. Nœuds et branches sont définis sur la figure. On peut montrer que l'ensemble de ces équations suffit pour calculer tensions et courants dans chaque branche du circuit total.

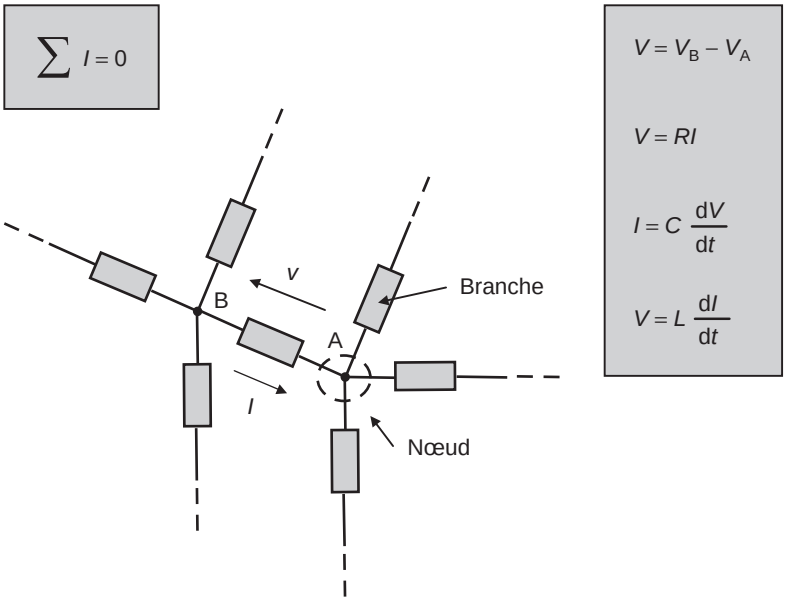


Figure 2.38 – Le calcul des circuits électriques.

La *figure 2.38* résume la méthode de calcul des circuits.

- La somme des courants entrants est nulle en chaque nœud.
- Dans chaque branche, on écrit l'équation courant-tension.

Quand on écrit que la somme des courants entrant dans un nœud est nulle, on applique en fait la conservation du flux du courant total à une surface fermée entourant le nœud. La définition de cette surface, indiquée en pointillés sur la *figure 2.38*, conduit à localiser le nœud considéré. Insistons sur le fait que tension et courant sont, par pure convention, orientés de manière opposée pour pouvoir écrire les équations indiquées avec des signes positifs.

On écrit « $V = RI$ » et non pas « $V = -RI$ ». Les valeurs obtenues dans la résolution du système d'équations, tensions et courants, peuvent être positives ou négatives.

Une équation supplémentaire a été introduite : c'est la relation courant-tension aux bornes d'une self. Un enroulement de fil conducteur ou même une simple connexion manifeste un comportement inductif dû au champ magnétique créé par le passage d'un courant. Cet effet conduit à l'apparition d'une tension induite aux bornes mesurée par la formule $L\,dI/dt$.

2.8.2 *Les méthodes de calcul et la transformée de Laplace*

Le paragraphe précédent a donné les règles de mise en équation du problème. Ce paragraphe donne un résumé des méthodes de résolution. Prenons l'exemple du premier circuit étudié, la mise en parallèle d'un condensateur et d'une résistance.

$$C\frac{\mathrm{d}E}{\mathrm{d}t} + \frac{E}{R} = I$$

Si la tension est donnée, la résolution est immédiate. Dans le cas inverse, c'est-à-dire quand le courant est donné, le problème est plus difficile et on obtient une équation différentielle du premier ordre.

La solution en régime établi, c'est-à-dire ne dépendant pas du temps est triviale si on considère une tension constante E_0. Les dérivées en fonction du temps étant nulles, il reste :

$$\frac{E_0}{R} = I$$

Supposons maintenant que le courant varie brusquement de 0 à I_0. La solution générale de l'équation différentielle est supposée, pour t positif, de la forme :

$$E(t) = A(t)\exp^{-\frac{t}{RC}}$$

On remplace E par sa valeur dans l'équation différentielle ce qui permet de calculer $A(t)$.

$$\frac{\mathrm{d}A}{\mathrm{d}t} = \frac{I_0}{C}\exp^{\frac{t}{RC}}$$

On en déduit $A(t)$ à une constante près.

$$A(t) = RI_0\exp^{\frac{t}{RC}} + A_0$$

La tension $E(t)$ est donc :

$$E(t) = RI_0 + A_0\exp^{-\frac{t}{RC}}$$

Pour déterminer la constante A_0, il suffit de raisonner par continuité en $t = 0$. La tension ne peut varier que de manière continue aux bornes du condensateur. Dans le cas inverse, il y aurait un courant infini. Elle est donc nulle en $t = 0^+$ puisqu'elle était nulle pour les temps négatifs. La tension s'écrit alors :

$$E(t) = RI_0 - RI_0\exp^{-\frac{t}{RC}}$$

Les deux courbes, courant et tension, sont tracées *figure 2.39*.

On peut définir le temps de montée de l'impulsion de tension en prenant les valeurs à 10 % et à 90 % de la valeur maximale et en mesurant l'intervalle de temps correspondant. On obtient alors :

$$t_r = 2{,}2\,RC$$

La résolution des équations différentielles ne pose pas de problème dans les cas simples. Dans les cas plus compliqués, on préfère la méthode de Laplace qui permet de transformer les équations différentielles en équations algébriques. Le principe en est le suivant.

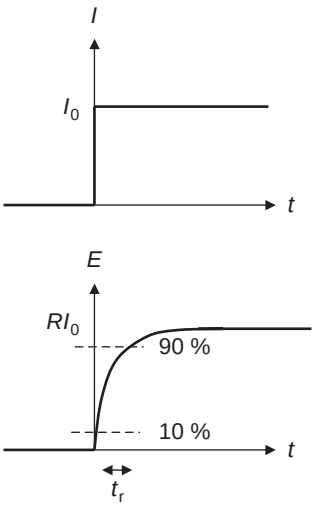


Figure 2.39 – Courant et tension.

On fait correspondre à une fonction du temps $E(t)$, nulle pour les temps négatifs, une fonction de la variable complexe s. Cette fonction sera notée $E(s)$.

$$E(s) = \int_0^\infty E(t)\exp^{-st}\,\mathrm{d}t \tag{2.34}$$

Insistons sur le fait que la fonction d'origine est nulle pour les temps négatifs. Si ce n'était pas le cas, des fonctions différentes pourraient avoir la même transformée de Laplace ce qui poserait des problèmes dans l'utilisation de la méthode. L'intérêt de cette transformation est la simplicité de transformation des fonctions dérivée et intégrale. On peut facilement montrer que si $E(t)$ se transforme en $E(s)$ alors :

$$\frac{\mathrm{d}E}{\mathrm{d}t} \rightarrow sE(s) - E(0^+)$$

$$\int_0^t E(t')\mathrm{d}t' \rightarrow \frac{E(s)}{s}$$

Le symbole $E(0^+)$ exprime la valeur de la fonction réelle E pour une valeur légèrement positive. L'intérêt de cette méthode peut être mis en évidence en étudiant à nouveau le cas du condensateur en parallèle avec la résistance.

$$\frac{\mathrm{d}E}{\mathrm{d}t} + \frac{E}{RC} = \frac{I}{C}$$

La transformée de cette équation s'effectue en tenant compte de la linéarité de la transformation et des résultats précédents :

$$sE(s) + \frac{E(s)}{RC} = \frac{I_0}{Cs}$$

Nous avons également utilisé le fait que $E(0^+)$ est nul par continuité et que la transformée de la fonction en forme de marche représentant le courant est I_0/s. Cette dernière propriété se démontre facilement à partir de la définition de la transformée. On calcule alors la valeur algébrique de la tension :

$$E(s) = RI_0 \frac{1}{s} \cdot \frac{1}{1 + \frac{s}{RC}}$$

Il faut maintenant inverser cette relation pour trouver la fonction du temps correspondante. Une méthode souvent utilisée quand la fonction n'est pas dans une table de correspondance est de la décomposer en somme de termes plus simples. Dans ce cas, on écrit :

$$E(s) = RI_0 \left(\frac{1}{s} - \frac{1}{1 + \frac{s}{RC}} \right)$$

Il suffit de lire dans une table que les inverses des deux termes sont respectivement $\gamma(t)$ et $\gamma(t)\exp^{-\frac{t}{RC}}$ pour en déduire :

$$E(t) = RI_0 \left(\gamma(t) - \gamma(t)\exp^{-\frac{t}{RC}} \right)$$

Rappelons que la fonction $\gamma(t)$ est nulle pour les temps négatifs et égale à 1 pour les temps positifs. On retrouve bien le résultat du paragraphe précédent.

Dans les cas plus complexes, la méthode est la même. On écrit en chaque nœud que la somme des courants est nulle. Les relations courant-tension sont écrites dans chaque branche en notation de Laplace, comme il est indiqué dans le *tableau 2.2*.

Tableau 2.2

Type de branche	Relation
Résistive	$V(s) = R\,I(s)$
Capacitive	$I(s) = CsV(s) - V(0^+)$
Inductive	$V(s) = LsI(s) - I(0^+)$

2.8.3 *La fonction de transfert d'un circuit*

Dans l'étude d'un circuit, il est souvent utile d'exprimer la grandeur de sortie en fonction de la grandeur d'entrée. La grandeur de sortie, en général une tension, est une des inconnues du problème général posé en écrivant en chaque nœud du circuit la conservation du courant et en écrivant dans chaque branche la relation entre le courant et la tension. Quand on fait usage de la méthode de Laplace, le résultat final est donné sous forme d'une fraction rationnelle de la variable s.

$$\frac{V_S(s)}{V_E(s)} = \frac{a_n s^n + a_{n-1} s^{n-1} + \ldots + a_1 s + a_0}{b_m s^m + b_{m-1} s^{m-1} + \ldots + b_1 s + b_0}$$

Cette fraction peut se décomposer de la manière suivante.

$$\frac{V_S(s)}{V_E(s)} = \frac{a_n(s+z_1)(s+z_2)\ldots(s+z_n)}{b_m(s+p_1)(s+p_2)\ldots(s+p_m)}$$

Les termes z du numérateur sont les zéros et les termes p du dénominateur sont les pôles. Ils sont réels ou complexes.

Quand les termes z_i et p_i sont réels, on reconnaît les comportements dérivateurs et intégrateurs ou passe-haut et passe-bas de la théorie des circuits électriques. Le remplacement de s par la variable fréquentielle $i\omega$ permet de passer de la représentation temporelle à la représentation fréquentielle. Quand les termes p_i sont imaginaires, des comportements oscillatoires de la variable de sortie sont attendus. Toutes ces notions sont développées de manière beaucoup plus complète dans des ouvrages spécialisés (référence **[5]**).

La connaissance des pôles et des zéros permet une compréhension très complète des propriétés d'un circuit et sera largement mise en œuvre dans la suite de cet ouvrage. Une approximation souvent utilisée est celle du pôle dominant. Si par exemple, le dénominateur de la fonction de transfert est un polynôme du second degré, il s'écrit :

$$D(s) = as^2 + bs + c$$

soit,

$$D(s) = a(s+s_1)(s+s_2)$$

Supposons maintenant que l'un des pôles soit très supérieur à l'autre, s_1 par exemple.

$$as_1s_2 = c$$
$$as_1 = b$$

On en déduit les valeurs des deux pôles en fonction des coefficients du polynôme.

$$s_1 = \frac{b}{a} \quad \text{et} \quad s_2 = \frac{c}{b}$$

Chapitre 3

La jonction *pn* et la structure Métal-Isolant-Semi-conducteur

Ces deux dispositifs de base sont les briques avec lesquelles il est possible de construire tous les composants et systèmes de la micro-électronique aujourd'hui. Une parfaite compréhension des mécanismes mis en jeu dans ces dispositifs est donc indispensable.

L'objectif de ce chapitre est de comprendre à partir de la description géométrique du dispositif l'évolution des variables internes (potentiel, densité de charge, densité de courant) quand il est soumis à l'effet de potentiels externes appliqués sur ses électrodes. Les notions de base sont rappelées dans le chapitre 2. Les plus utiles sont cependant les notions de base en électromagnétisme et les relations reliant les densités de porteurs et les potentiels. Le potentiel chimique souvent appelé de manière abusive énergie de Fermi est une grandeur fondamentale dans la détermination des densités de charge. Il est directement relié au potentiel extérieur appliqué, comme cela est expliqué dans le chapitre 2.

La mécanique quantique et son application à la physique des solides ne sont pas strictement indispensables à la compréhension du fonctionnement électrique de ces dispositifs, et l'électromagnétisme classique explique la plupart des effets. Il faut cependant faire usage d'un peu de physique des solides pour exprimer la relation entre la densité des porteurs et la tension appliquée. La physique des solides est également nécessaire pour expliquer le concept de « trou ». De plus, des effets nouveaux (l'effet tunnel par exemple) liés aux faibles dimensions ne peuvent être compris si l'on ne se soucie pas de la nature ondulatoire des électrons.

La modélisation de la structure MOS s'appuie sur l'analyse de Y.P. Tsividis donnée référence [7].

3.1 La jonction *pn* ou *np*

La jonction *pn* est un dispositif de base. Son fonctionnement est décrit en détail dans de nombreux ouvrages et dans ce paragraphe nous ne ferons que donner les éléments les plus importants.

On suppose donc que deux blocs de semi-conducteurs sont mis en contact, le premier de type *p* et le second de type *n*. Ce contact doit être pris au sens métallurgique du terme et n'est pas un simple contact physique. Le semi-conducteur de type *p* peut être considéré comme un ensemble d'atomes neutres de silicium comportant en plus quelques dopants accepteurs. Ils sont considérés comme des ions fixes chargés négativement et sont associés à des trous non localisés et mobiles dans le solide. Le semi-conducteur de type *n* est formé d'atomes neutres et comporte quelques ions donneurs considérés comme des charges positives fixes associées à des électrons non localisés et mobiles. La *figure 3.1* illustre les deux cas. Par souci de clarté, on représente uniquement les dopants mais il y a beaucoup plus d'atomes de silicium. Que se passe-t-il donc quand les deux blocs sont mis en contact ?

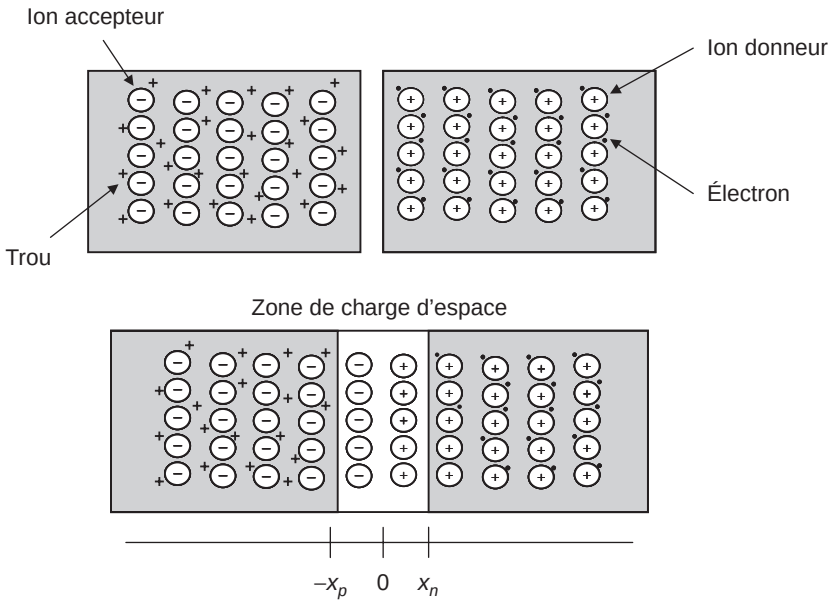


Figure 3.1 – La jonction pn.

Quand les deux blocs sont en contact, les électrons diffusent du côté *n* où ils sont le plus nombreux vers le côté *p* et ils se recombinent avec les trous. Les dopants privés de leurs électrons forment alors une région chargée positivement côté *n*. De même, les trous nombreux dans la région *p* diffusent dans la région *n* et se recombinent avec les électrons. Les dopants de la région *p* privés de leurs trous forment une région chargée négativement. Ces deux régions chargées constituent la zone de charge d'espace de la jonction. Un champ électrique règne donc dans cette zone et varie selon la loi de Poisson donnée dans le chapitre 2. Il est dirigé de la charge positive vers la charge négative et donc de la région *n* vers la région *p*. Il s'oppose au passage des électrons de la région *n* vers la région *p* et donc au courant de diffusion d'électrons. Il s'instaure alors un régime d'équilibre dans lequel le

courant de diffusion des électrons compense le courant de dérive des électrons. Ces notions sont détaillées dans le chapitre 2. Il en est de même pour les trous et le courant total traversant la jonction est nul. Ce résultat n'est pas étonnant car sinon un courant circulerait sans apport d'énergie, ce qui est contraire au bon sens physique. Il est maintenant possible de calculer la différence de potentiel qui apparaît de part et d'autre de la jonction.

Pour simplifier le calcul, on suppose la jonction abrupte. En réalité, la jonction n'est pas abrupte étant donné le mode de fabrication. La densité de charge positive passe brusquement à zéro au-delà de la distance x_n et la densité de charge négative passe brusquement à zéro au-delà de x_p. Pour calculer la variation du potentiel, il suffit d'appliquer la relation de Poisson dans les deux parties de la zone de charge d'espace.

$$\varepsilon_s \frac{\mathrm{d}^2 V}{\mathrm{d}x^2} - eN_A = 0$$

$$\varepsilon_s \frac{\mathrm{d}^2 V}{\mathrm{d}x^2} + eN_D = 0$$

Dans ces relations, N_A et N_D sont respectivement les densités d'accepteurs et de donneurs. On admet pour simplifier que la densité de charge est égale à la densité de dopants. On constate sur ces formules que le champ varie d'autant plus vite que le dopage est élevé. Il faut maintenant faire appel à un peu de physique des solides, rappelée dans le chapitre 2, pour exprimer les densités n et p d'électrons et de trous.

$$n(x) = n_i \exp^{\frac{E_F - E_i(x)}{k_B T}} \tag{3.1}$$

$$p(x) = n_i \exp^{\frac{E_i(x) - E_F}{k_B T}} \tag{3.2}$$

On suppose le système en équilibre ce qui veut dire que le potentiel chimique E_F est constant en fonction de la position. Cette hypothèse n'est plus vraie quand la jonction est soumise à une différence de potentiel extérieure. Le chapitre 2 clarifie ces notions importantes. L'énergie E_i est comprise entre les énergies de valence et de conduction. Elle est placée environ au milieu du gap. De manière plus précise, elle est donnée par la relation :

$$E_i = \frac{1}{2}\left(E_C + E_V + k_B T \ln \frac{m_v^*}{m_c^*} \right)$$

La constante n_i est caractéristique du semi-conducteur. Sa valeur pour le silicium est 0,015 par μ^3 à température ambiante. Le dernier terme de la formule dépend des masses effectives et est en général négligeable. Si on applique les relations 3.1 et 3.2 aux extrémités du dispositif, on peut écrire :

$$N_D = n_i \exp^{\frac{E_F - E_{in}}{k_B T}}$$

$$N_A = n_i \exp^{\frac{E_{ip} - E_F}{k_B T}}$$

On en déduit donc,

$$N_A N_D = n_i^2 \exp^{\frac{E_{ip} - E_{in}}{k_B T}}$$

Comme la différence d'énergie est la différence de potentiel à la charge près, la différence de potentiel aux bornes de la jonction est donc :

$$\phi_F = \frac{k_B T}{e} \ln \frac{N_A N_D}{n_i^2} \tag{3.3}$$

Le terme $k_B T/e$ vaut 26 mV à température ambiante. La différence de potentiel calculée précédemment n'est cependant pas mesurable. Pour s'en convaincre, il faut tenir compte des différences de potentiel qui apparaissent entre les pointes de mesure et le semi-conducteur comme il sera expliqué dans le paragraphe 3.3. Appliquons maintenant une différence de potentiel entre les deux blocs de semi-conducteurs comme il est représenté *figure 3.2*.

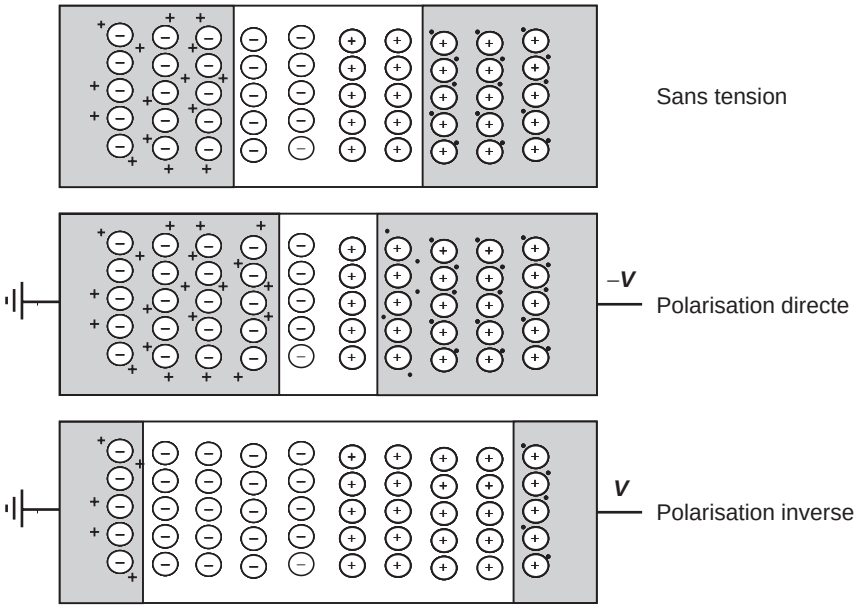


Figure 3.2 – Jonction polarisée.

Dans un premier temps, une tension négative est appliquée entre la région *n* et la région *p*. Les électrons de la partie *n* sont repoussés vers la zone centrale de même que les trous de la partie *p*. La zone de charge d'espace diminue alors. Le champ étant faible, il s'oppose peu au passage des porteurs majoritaires. En résumé, le courant de diffusion est important et un courant total de forte valeur traverse la jonction. Elle est dite polarisée en direct.

Inversement, appliquons une tension positive entre la partie *n* et la partie *p*. Les électrons sont alors repoussés de la zone centrale ainsi que les trous. La zone de charge d'espace s'élargit. Le champ électrique s'oppose au passage des porteurs majoritaires à travers la jonction et le courant traversant la jonction est très faible. La jonction est dite polarisée en inverse. La jonction peut être également vue comme un condensateur plan dont l'épaisseur est la largeur de la zone de charge d'espace. Dans ce cas, la largeur est importante et la capacité est donc faible.

Le calcul du courant traversant une jonction est décrit dans de nombreux ouvrages consacrés aux semi-conducteurs. Le résultat est présenté sans démonstration. Il s'exprime par une relation valable en mode direct et en mode inverse à condition de respecter la convention de signe pour la tension. La tension V est positive quand elle est appliquée entre la partie p et la partie n. Le courant I traversant la jonction est alors :

$$I = I_S\left(\exp^{\frac{V}{\phi_t}} - 1\right) \tag{3.4}$$

Le terme ϕ_t est égal à k_BT/e soit 26 mV à température ambiante. Le terme I_S est le courant de saturation inverse et sa valeur est :

$$I_S = eA\left(\frac{D_pN_A}{L_p} + \frac{D_nN_D}{L_n}\right) \tag{3.5}$$

Dans cette relation, les grandeurs D et L sont respectivement les coefficients de diffusion et les longueurs de diffusion. Elles sont définies pour les trous et les électrons. A est la section de la jonction. Les longueurs de diffusion sont définies par les relations :

$$L_n = \sqrt{D_n\tau_n}$$

$$L_p = \sqrt{D_p\tau_p}$$

Dans ces relations, les paramètres τ sont les durées de vie des porteurs dans le matériau. Les coefficients de diffusion ont été définis dans le chapitre 2.

3.2 Les potentiels de contact

Le problème des contacts entre matériaux différents est d'une grande importance dans les dispositifs micro-électroniques bien qu'étant souvent passé sous silence. Il est possible d'en donner une représentation simple en définissant le potentiel de contact. Si on considère deux matériaux différents 1 et 2 mis en contact, isolants ou semi-conducteurs, il y a diffusion des porteurs des régions de forte concentration vers les régions de faible concentration tout comme dans la jonction *pn*. Il se forme alors de la même manière une zone de charge d'espace et une différence de potentiel apparaît naturellement sans tension extérieure appliquée. On notera cette tension V_{12} égale à $V_1 - V_2$.

Quand le matériau 2 est du silicium intrinsèque, le potentiel V_2 est noté V_I. La différence de potentiel entre le matériau 1 et le silicium intrinsèque est notée ϕ. Quand le matériau 1 est un semi-conducteur elle est appelée potentiel de Fermi et est notée $-\phi$. Le potentiel de Fermi est défini comme l'inverse du potentiel de contact. Cette définition n'a pas d'autre justification que la tradition des notations de la micro-électronique.

$$\phi_F = -(V_1 - V_I)$$

L'étude de la jonction *pn* a montré que si le matériau 1 est un semi-conducteur de type p dopé avec une concentration N_A d'atomes accepteurs alors,

$$V_1 - V_I = -\phi_t \ln\frac{N_A}{n_i}$$

Les trous diffusant de la région *p* vers la région intrinsèque, la différence de potentiel entre la région *p* et la région intrinsèque est bien négative. Dans le cas d'un semi-conducteur de type *n*, on obtient :

$$V_1 - V_I = \phi_t \ln \frac{N_D}{n_i}$$

Si maintenant on met en contact deux matériaux quelconques et si on veut calculer le potentiel de contact en fonction des potentiels de contact de chacun d'eux relativement au silicium intrinsèque, il suffit de considérer l'assemblage de la *figure 3.3*.

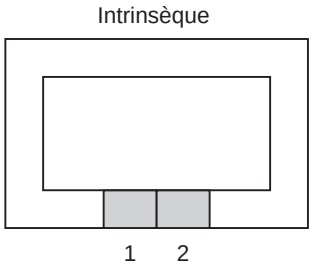


Figure 3.3 – Calcul du potentiel de contact entre deux matériaux.

Pour calculer le potentiel de contact entre les matériaux 1 et 2, il suffit d'écrire :

$$V_1 - V_2 = V_1 - V_I + V_I - V_2$$
$$V_1 - V_2 = \phi_1 - \phi_2$$

De même, si on considère plusieurs matériaux en série comme dans la *figure 3.4*, on peut écrire :

$$V_1 - V_n = V_1 - V_2 + V_2 - V_3 + \ldots + V_{n-1} - V_n$$

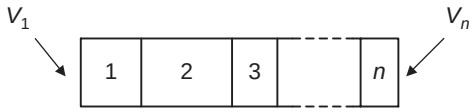


Figure 3.4 – Matériaux en série.

Pour terminer ce paragraphe, il reste à donner quelques valeurs de potentiel de contact relativement au silicium intrinsèque pour différents métaux et pour du silicium dopé : *tableau 3.1*.

On notera $\phi_F = \phi_t \ln \frac{N_A}{n_i}$ dans le cas d'un semi-conducteur dopé *p* et $\phi_F = -\phi_t \ln \frac{N_D}{n_i}$ dans le cas d'un semi-conducteur dopé *n*.

Tableau 3.1.

Matériau	Potentiel de contact
Argent	−0,4 V
Or	−0,3 V
Cuivre	0 V
Nickel	0,15 V
Aluminium	0,6 V
Silicium *p*	$-\phi_t \ln N_A/n_i$
Silicium *n*	$\phi_t \ln N_D/n_i$
Silicium intrinsèque	0 V

3.3 La structure Métal-Oxyde-Semiconducteur

3.3.1 *Description phénoménologique*

Dans le chapitre 1, nous avons présenté un dispositif dans lequel un courant peut être commandé par une tension appliquée non pas directement sur le matériau mais à travers une mince couche d'oxyde. Rappelons les choix de ce matériau. Si le dispositif était isolant, le courant ne pourrait circuler. Si le matériau était un métal, le champ électrique ne pourrait pénétrer à l'intérieur et dans ce cas l'action de commande par la tension serait impossible. Le choix se porte alors naturellement vers les semi-conducteurs.

La première étape est de comprendre comment une tension appliquée à travers un oxyde peut contrôler la charge stockée. C'est l'objectif de ce paragraphe. Le chapitre 4 montrera ensuite comment ce modèle évolue quand deux réservoirs de charge sont placés de part et d'autre de ce dispositif et comment un courant peut circuler de l'un à l'autre.

Le dispositif de base est donc un empilement de trois matériaux : un semi-conducteur dopé de type *p* ou *n*, un isolant de faible épaisseur (quelques nanomètres pour les technologies les plus avancées) et une couche métallique appelée grille. La technologie des circuits intégrés a conduit pendant de longues années à choisir du silicium fortement dopé à la place du métal pour des raisons qui seront expliquées ultérieurement mais le comportement est équivalent à celui d'un métal. Il faut ajouter à cela une couche métallique en face arrière pour prendre le contact électrique, en général de l'aluminium. Ce dispositif est représenté *figure 3.5*.

Quelques commentaires sont nécessaires pour expliquer cette figure. Le silicium est de type *p* mais un dispositif équivalent peut être imaginé avec du silicium de type *n*. Le dopage du silicium est d'environ 10^{17} dopants par cm^3 soit 10^5 par μ^3. L'épaisseur de l'oxyde est d'environ 2 nm pour les technologies actuelles soit 2/1 000 de micron ce qui représente quelques couches atomiques. L'épaisseur de silicium est très importante et ne joue aucun rôle dans le fonctionnement électrique. La dimension longitudinale que nous noterons *x* est supposée très supérieure à la dimension transverse de la zone chargée correspondant à l'épaisseur du dispositif. La troisième dimension, non représentée, est également très supérieure à l'épaisseur. En pratique, un tel dispositif s'étend sur quelques microns de côté.

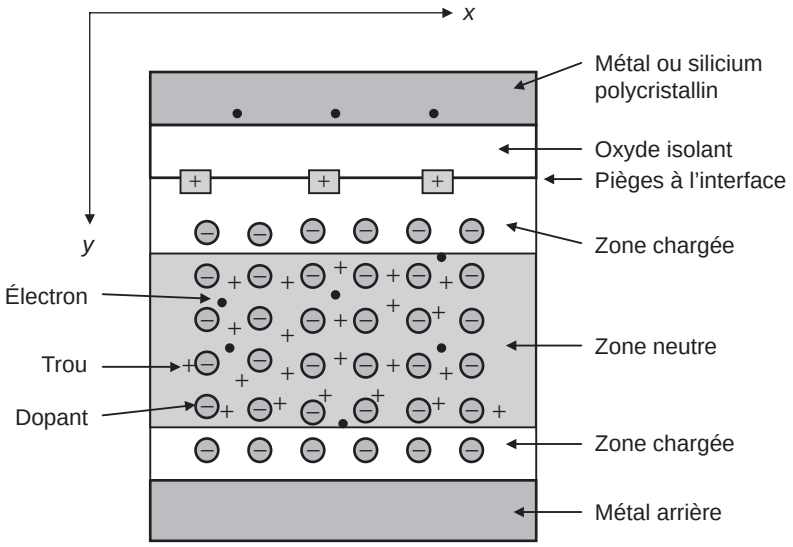


Figure 3.5 – Le dispositif Métal-Oxyde-Semiconducteur.

À l'interface entre le semi-conducteur et l'oxyde, apparaît une fine couche de charges inhérente au processus de fabrication. Ces charges sont soit des impuretés soit des atomes de silicium ayant des liaisons manquantes. La densité est de l'ordre de 0,001 fC par μ^2 et leur charge est souvent positive. Le semi-conducteur de base est formé d'un grand nombre d'atomes de silicium non représentés sur la figure et neutres. Un faible nombre de dopants (des atomes de Bore pour le silicium *p*) sont ajoutés pendant la fabrication du dispositif. Ces atomes sont fixes, chargés négativement et associés chacun à un trou mobile symbolisé par une croix.

Quelques paires électron-trou peuvent être créées thermiquement ce qui explique la présence de quelques électrons dans le silicium, représentés par des billes noires. Pour trouver le nombre de ces électrons, il suffit d'appliquer :

$$nN_A = n_i^2$$

Comme n_i vaut $0{,}015/\mu^3$ à température ambiante, on trouve pour N_A égal à $10^5/\mu^3$:

$$n = 2 \text{ par mm}^3$$

Étudions maintenant le comportement de ce dispositif sans tension extérieure appliquée. Des échanges de porteurs ont lieu entre les différentes régions dans lesquelles les concentrations sont différentes. La zone de silicium en contact avec l'oxyde est concernée à cause des charges présentes en surface qui créent un champ. Il en est de même pour la zone en contact avec le métal en face arrière car les électrons peuvent diffuser du métal vers le silicium. Comme dans la jonction *pn*, les mouvements de charge par diffusion conduisent à la création d'un champ électrique qui s'oppose à ce mouvement et les courants de diffusion et de dérive se compensent. Le courant total traversant le dispositif est nul mais des échanges ont lieu dans les deux zones blanchies sur le dessin dans lesquelles le champ électrique n'est pas nul. Dans la zone centrale grisée, la densité locale de charge est nulle et le champ électrique également ce qui veut dire que le potentiel est constant.

Détaillons un peu les potentiels formés aux contacts en considérant cette fois un système complet incluant des électrodes capables d'appliquer une tension extérieure comme le montre la *figure 3.6*. On suppose que les électrodes sont fabriquées avec le même métal. On appellera V_{bulk} le potentiel dans la zone neutre du dispositif. Il sera pris comme référence.

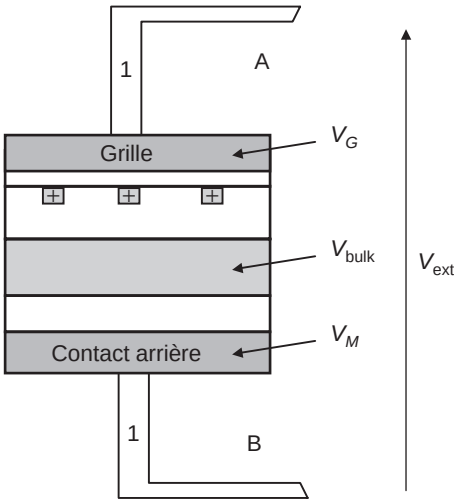


Figure 3.6 – Les potentiels de contact.

Dans une première étape, exprimons la différence de potentiel extérieure V_{ext} en fonction des différents potentiels présents dans le dispositif.

$$V_{\text{ext}} = V_A - V_G + V_G - V_{\text{bulk}} + V_{\text{bulk}} - V_M + V_M - V_B$$

Exprimons les potentiels par rapport au silicium intrinsèque comme il a été vu dans le paragraphe 3.2.

$$V_{\text{ext}} = \phi_1 - \phi_G + V_G - V_{\text{bulk}} + \phi_{\text{si}} - \phi_M + \phi_m - \phi_1$$

Les potentiels ϕ figurant dans cette formule sont les potentiels de contact par rapport à l'intrinsèque et sont donnés dans les tables. Ils sont notés ϕ_1 pour le métal de l'électrode, ϕ_{si} pour le semi-conducteur, ϕ_M pour le métal face arrière et ϕ_G pour le matériau de grille. Il reste :

$$V_{\text{ext}} = -\phi_G + V_G - V_{\text{bulk}} + \phi_{\text{si}}$$

On notera alors :

$$\phi_{MS} = \phi_{\text{si}} - \phi_G$$

Cette notation est traditionnellement utilisée dans le monde des semi-conducteurs. Elle reste valable quand la grille n'est pas un métal mais du silicium polycristallin fortement dopé.

On obtient donc finalement :

$$V_{\text{ext}} = V_G - V_{\text{bulk}} + \phi_{MS} \tag{3.6}$$

La tension présente dans le dispositif V_{bulk} dépend de la tension appliquée et des potentiels de Fermi du semi-conducteur et du matériau de grille et absolument pas des autres matériaux utilisés dans le dispositif, contact arrière et matériaux des fils de contact. Il est utile de donner quelques valeurs numériques pour ϕ_{MS} en appliquant les formules du tableau du paragraphe 3.2.

Pour du silicium dopé *p* de dopage 10^5 par μ^3 en contact avec de l'aluminium, on trouve :

$$\phi_{si} = -0{,}4\ \text{V}$$

$$\phi_G = 0{,}6\ \text{V}$$

$$\phi_{MS} = -0{,}4\ \text{V} - 0{,}6\ \text{V} = -1\ \text{V}$$

Si le métal de grille est remplacé par du silicium polycristallin fortement dopé, son potentiel de Fermi est environ 0,56 V, la valeur de ϕ_{MS} est alors peu différente :

$$\phi_{MS} = -0{,}4\ \text{V} - 0{,}56\ \text{V} = -0{,}96\ \text{V}$$

Si le silicium de base est dopé *n* on peut calculer les potentiels de la même manière mais le potentiel de contact du semi-conducteur est dans ce cas de signe positif.

Appliquons maintenant une tension nulle entre grille et contact arrière comme le montre la *figure 3.7*. Les trous de la zone *p* sont repoussés vers l'intérieur par les ions en surface et une zone de charge d'espace chargée négativement apparaît. De même, une zone chargée négativement apparaît au niveau du contact arrière car des électrons du métal diffusent et se recombinent aux trous.

Si on augmente encore la tension sur la grille, la zone de charge d'espace au niveau de l'oxyde s'étend davantage. Celle créée au niveau du contact arrière reste inchangée car le potentiel de la grille n'a pas d'effet dans cette région. La *figure 3.7* ne respecte pas les échelles car l'épaisseur de silicium neutre, représenté par la zone grisée, est en fait beaucoup plus importante que les zones de charge d'espace. Il y a un facteur mille environ.

Quand la tension augmente encore, elle finit par être assez importante pour attirer les quelques électrons présents dans le dispositif qui vont alors s'accumuler à l'interface oxyde-semi-conducteur. Cet effet est appelé l'inversion du semi-conducteur puisqu'il se forme une couche de porteurs minoritaires. C'est le phénomène de base du fonctionnement du transistor. La *figure 3.7* illustre la formation de la couche d'inversion.

Quand la tension augmente encore, cette couche d'électrons forme une sorte d'écran électrostatique entre la grille et le semi-conducteur et l'excédent de tension est quasiment appliqué intégralement de part et d'autre de l'oxyde. Au fur et à mesure que la charge négative augmente dans le silicium, il se forme sur la grille une charge positive de la même manière que les deux armatures d'un condensateur se chargent avec des polarités opposées.

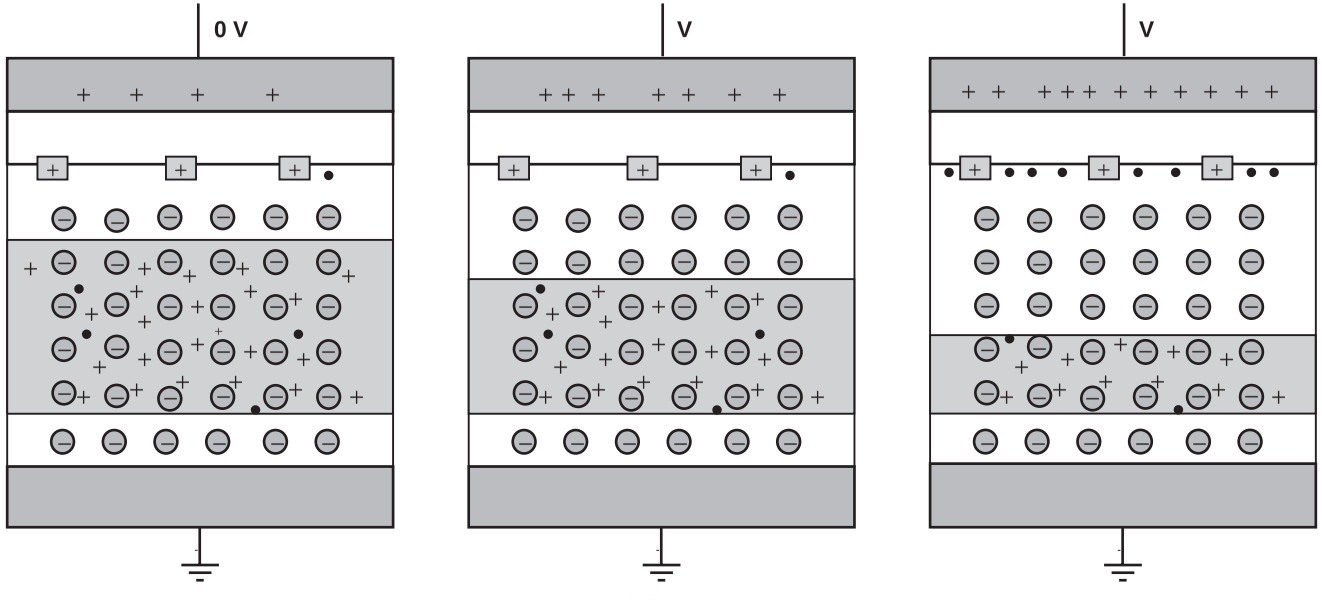
Il est possible de donner la profondeur de la zone de charge d'espace en dessous de l'oxyde de grille. Il suffit d'appliquer la loi de Poisson dans le silicium.

$$\varepsilon_s \frac{\mathrm{d}E}{\mathrm{d}y} = -eN_A$$

La densité est supposée égale à celle des dopants en suivant une approximation très classique.

En fonction du potentiel, on obtient :

$$\varepsilon_s \frac{\mathrm{d}^2E}{\mathrm{d}y^2} = eN_A$$



a) Tension nulle

Une zone de charge d'espace de faible épaisseur se forme

b) Tension légèrement positive

Une zone de charge d'espace de faible épaisseur se forme

c) Tension positive

Une couche d'inversion apparaît

Figure 3.7 – Formation de la couche d'inversion.

On en déduit facilement la profondeur de la zone dans laquelle règne un champ électrique en fonction de la différence de potentiel ΔV entre la surface et la partie profonde du silicium dans laquelle le champ est nul.

$$y_b = \sqrt{\frac{2\,\varepsilon_s \Delta V}{eN_A}}$$

Quelques valeurs typiques donnent des ordres de grandeur.

$$\Delta V = 1\,V$$
$$\varepsilon_s = 0{,}104\,fF/\mu$$
$$e = 1{,}6\,10^{-4}\,fC$$
$$N_A = 10^5/\mu^3$$

On calcule alors :

$$l \approx 0{,}1 \text{ micron}$$

3.3.2 *Le modèle électrique*

L'objectif de ce paragraphe est d'exprimer la charge d'inversion en fonction de la tension appliquée. En effet, les propriétés de conduction des MOSFET dépendent de la valeur de cette charge d'inversion responsable de la conduction. Le dispositif de base est représenté *figure 3.8*. La profondeur est notée *y* et l'origine est prise à l'interface oxyde-semi-conducteur. Le potentiel représenté est la différence de potentiel entre un point à une profondeur *y* et un point en profondeur dans le matériau, région dans laquelle le champ est nul. Cette différence de potentiel est représentée sur la *figure 3.8*.

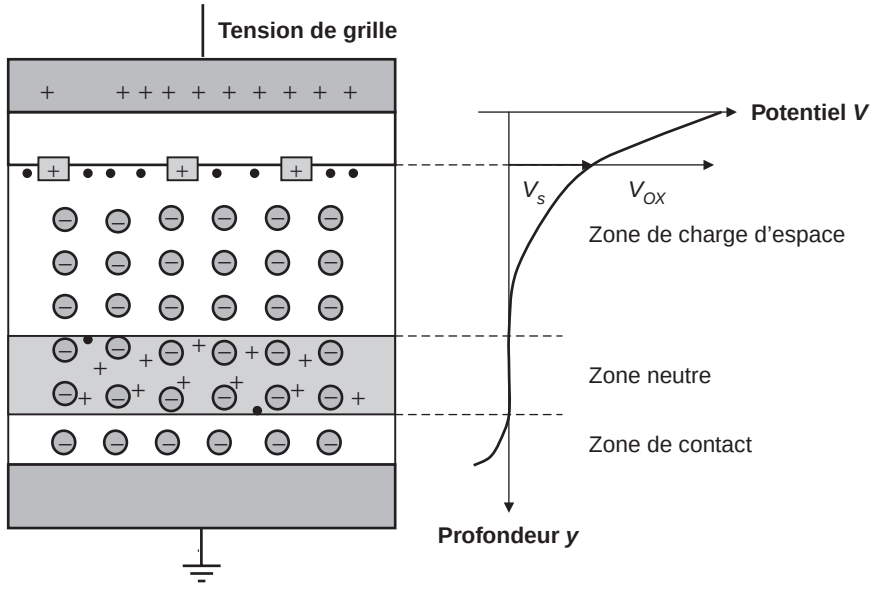


Figure 3.8 – Le modèle électrique du dispositif MOS.

On suppose que le système est en équilibre thermodynamique. Le potentiel chimique du gaz d'électrons est alors égal à celui des trous. Cette hypothèse ne sera plus valable dans le transistor qui échange des charges avec l'extérieur et il sera nécessaire de définir des pseudo-potentiels chimiques différents pour les électrons et les trous. On part alors des relations classiques données dans le chapitre 2 pour exprimer les densités d'électrons et de trous.

$$n(y) = n_i \exp^{\frac{E_F - E_i(y)}{k_B T}}$$

$$p(y) = n_i \exp^{\frac{E_i(y) - E_F}{k_B T}}$$

L'énergie E_i est environ au milieu du gap et dépend de la profondeur à cause du potentiel électrique qui n'est pas constant dans le dispositif. Le potentiel chimique est lui constant par définition de l'équilibre thermodynamique. La valeur de n_i est à température ambiante de 0,015 par μ^3. On peut exprimer les valeurs de la concentration d'électrons n_s à l'interface oxyde-semi-conducteur et également la concentration n_b d'électrons en profondeur. En profondeur ou dans le bulk veut dire dans la région grisée sur la figure. Dans cette région, le champ est nul.

$$n_s = n_i \exp^{\frac{E_F - E_{is}}{k_B T}}$$

$$n_b = n_i \exp^{\frac{E_F - E_{ib}}{k_B T}}$$

De même, les densités de trous p_s à l'interface et en profondeur p_b sont :

$$p_s = n_i \exp^{\frac{E_{is} - E_F}{k_B T}}$$

$$p_b = n_i \exp^{\frac{E_{ib} - E_F}{k_B T}}$$

De plus, il est admis que la densité de trous en profondeur est :

$$p_b = N_A$$

On en déduit alors facilement :

$$\frac{n_b}{N_A} = \exp^{\frac{2(E_F - E_{ib})}{k_B T}}$$

$$\frac{n_s}{n_b} = \exp^{\frac{E_{ib} - E_{is}}{k_B T}}$$

Si on appelle V_S la différence entre le potentiel à l'interface et le potentiel en profondeur alors :

$$E_{is} - E_{ib} = -eV_s$$

Définissons maintenant le potentiel de Fermi ϕ_F du substrat par la relation :

$$\phi_F = \frac{E_F - E_{ib}}{-e}$$

On rappelle que la quantité $k_B T/e$ est notée ϕ_t.
La densité d'électrons à l'interface s'écrit alors :

$$n_s = N_A \exp^{\frac{V_s - 2\phi_F}{\phi_t}} \tag{3.8}$$

Il reste à exprimer le potentiel de Fermi en fonction du dopage et la relation est établie entre la concentration des charges d'inversion et le potentiel de surface. La relation exprimant la concentration de trous dans le substrat permet d'écrire :

$$N_A = n_i \exp^{\frac{\phi_F}{\phi_t}}$$

On en déduit :

$$\phi_F = \phi_t \ln \frac{N_A}{n_i}$$

Le potentiel de Fermi est positif pour un semi-conducteur de type *p*, il est négatif pour un semi-conducteur de type *n* et donné par la relation :

$$\phi_F = -\phi_t \ln \frac{N_D}{n_i}$$

On retrouve bien les valeurs données dans l'étude des potentiels de contact.
En pratique, le potentiel de Fermi varie assez peu avec le dopage étant donné la relation logarithmique. Il est compris entre 0,2 V et 0,5 V. La *figure 3.9* indique les valeurs possibles.

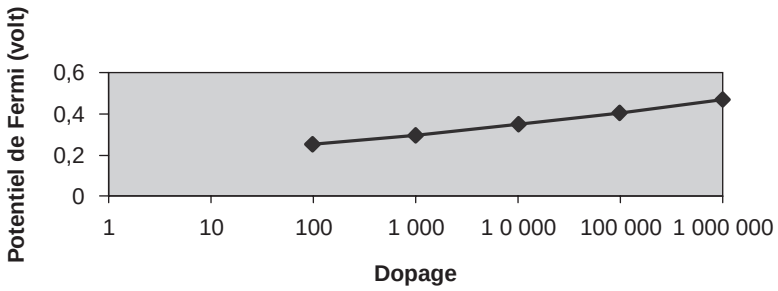


Figure 3.9 – Variation du potentiel de Fermi.

Les expressions précédentes font appel aux résultats de la physique des solides établis dans le chapitre 2. Maintenant nous ne ferons plus usage que de l'électromagnétisme pour expliquer le fonctionnement des composants classiques de la micro-électronique. Une exception cependant avec la prise en compte des effets tunnel.
Pour exprimer la densité des électrons d'inversion en fonction du potentiel appliqué il est nécessaire de considérer les différentes charges présentes dans le système en s'appuyant sur la *figure 3.10*.
Les différents types de charge sont les suivants :

- Les charges positives Q_G accumulées à la surface de la grille à l'interface avec l'oxyde.

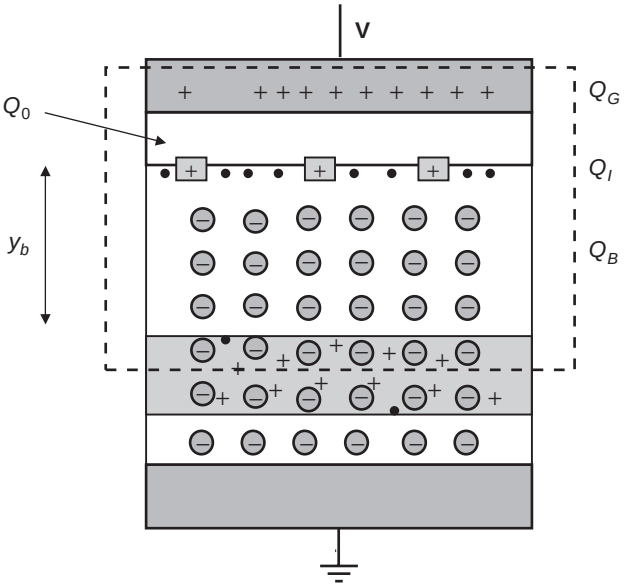


Figure 3.10 – Les charges du problème.

L'électromagnétisme montre qu'il n'y a pas de charge dans les métaux. Dans le cas d'une grille en silicium polycristallin fortement dopé, on fera la même hypothèse.

- Les électrons minoritaires d'inversion Q_I.
  Ils forment une couche à l'interface entre le semi-conducteur et l'oxyde et l'épaisseur de cette couche est infiniment mince.
- Les charges positives Q_0, fixes, à l'interface entre l'oxyde et le semi-conducteur.
- Les ions dopants privés de leurs trous Q_B dans la zone de charge d'espace.
  Elle s'étend sur une profondeur y_b de moins d'un micron.
- La densité des électrons mobiles dans le semi-conducteur notée $n(y)$.
  Les électrons de la couche d'inversion correspondent à $n(0)$.
- La densité des trous mobiles notée $p(y)$.

On ne prend pas en compte la charge au niveau du contact arrière car on exprime le potentiel entre la grille et l'intérieur du silicium et non pas entre la grille et le contact arrière. Cette charge apparaîtrait comme une constante dans les développements qui vont suivre.

Pour éviter toute confusion, reprenons l'expression 3.6 exprimant le potentiel appliqué.

$$V_{\text{ext}} = V_G - V_{\text{bulk}} + \phi_{MS}$$

Dans la suite du calcul, les potentiels sont mesurés relativement au potentiel en profondeur dans le silicium, appelé V_{bulk}, comme il est indiqué sur la *figure 3.9*. On notera en particulier :

$$V_s = V(0) - V_{\text{bulk}}$$
$$V_{\text{ox}} = V_G - V(0)$$

On écrit donc :

$$V_{\text{ext}} = V_G - V_{\text{bulk}} + \phi_{MS} = V_{\text{ox}} + V_s + \phi_{MS}$$

La résolution de ce système suppose connus la densité des charges de surface Q_0, le dopage du semi-conducteur N_A et la tension appliquée à la grille V_{ext}.
Les inconnues du problème sont les charges (Q_I, Q_B, Q_G), les densités de charge ($n(y)$ et $p(y)$), les potentiels ($V(y)$, V_s, V_{ox}) ainsi que la profondeur de la zone de charge d'espace y_b.
Les densités d'électrons n_b et de trous p_b sont également inconnues. Il y a au total 11 inconnues. Il faut donc écrire 11 équations indépendantes pour résoudre le problème. Elles seront notées de 3.9 à 3.19. La première est déjà établie.

$$V_{\text{ext}} = V_{\text{ox}} + V_s + \phi_{MS} \tag{3.9}$$

Les considérations précédentes ont montré que ϕ_{MS} était donné par des tables.
La deuxième équation exprime la densité de charge $n(y)$. Le calcul est équivalent à celui effectué précédemment pour calculer la densité de charge à l'interface.

$$\frac{n(y)}{n_b} = \exp^{\frac{V(y)}{\phi_t}} \tag{3.10}$$

De même,

$$\frac{p(y)}{p_b} = \exp^{-\frac{V(y)}{\phi_t}} \tag{3.11}$$

Dans la zone neutre,

$$p_b - n_b = N_A \tag{3.12}$$

De plus,

$$n_b p_b = n_i^2 \tag{3.13}$$

La charge sur la grille est reliée à la différence de potentiel V_{ox} par la relation des condensateurs plans. On raisonnera plus facilement sur les charges par unité de surface notées prime pour ne pas les confondre avec les charges globales. La charge de la grille par unité de surface s'écrit donc :

$$Q'_G = C'_{\text{ox}} V_{\text{ox}} \tag{3.14}$$

La valeur de C'_{ox}, capacité par unité de surface, est donnée par la relation classique des condensateurs plans et ne dépend que des dimensions du dispositif et de la constante diélectrique de l'oxyde, toutes grandeurs connues.
Il est possible d'écrire la loi de Poisson reliant densité et potentiel.

$$\frac{\mathrm{d}^2 V}{\mathrm{d}y^2} + e\left(\frac{p(y) - n(y) - N_A}{\varepsilon_s}\right) = 0 \tag{3.15}$$

Il est également possible de calculer la charge d'inversion par unité de surface en fonction de la densité d'électrons, il suffit d'intégrer dans toute la zone de charge d'espace :

$$Q'_I = \int_0^{y_b} -e\, n(y)\, \mathrm{d}y \tag{3.16}$$

La profondeur de la zone de charge d'espace est la solution de l'équation :

$$\left(\frac{\mathrm{d}V}{\mathrm{d}y}\right)_{y=y_b} = 0 \tag{3.17}$$

La charge Q'_B par unité de surface peut s'écrire :

$$Q'_B = \int_0^{y_b} -eN_A \mathrm{d}y \tag{3.18}$$

Enfin, il est possible d'écrire une règle de conservation de la charge. Cette règle est toujours délicate à appliquer car il n'est pas évident d'identifier les charges qui réellement se conservent. La manière la plus sûre est d'appliquer le théorème de Gauss en trouvant une surface au travers de laquelle le flux du champ est nul. La surface identifiée *figure 3.10* est intéressante car le champ y est nul sur toutes les faces ou parallèle aux faces considérées.

On peut donc écrire :

$$Q'_G + Q'_0 + Q'_I + Q'_B = 0 \tag{3.19}$$

On a négligé dans cette formule la charge des trous restant dans la zone de charge d'espace. On suppose qu'ils sont tous repoussés ou recombinés.

La résolution de ce système est assez complexe du point de vue calcul et seul le résultat est donné. Le détail figure en annexe de ce chapitre. En supposant la densité d'électrons négligeable dans le silicium profond ($p_b = N_A$), on trouve alors :

$$\left(\frac{\mathrm{d}V}{\mathrm{d}y}\right)^2 = \frac{2eN_A}{\varepsilon_s}\left[\phi_t \exp^{-\frac{V}{\phi_t}} + V - \phi_t + \exp^{-2\frac{\phi_F}{\phi_t}}\left(\phi_t \exp^{\frac{V}{\phi_t}} - V - \phi_t\right)\right] \tag{3.20}$$

Cette équation ne peut se résoudre analytiquement qu'en effectuant d'autres approximations.

3.3.3 *Le régime de forte inversion*

Le principe du calcul est de calculer le champ électrique à l'interface soit pour *y* égal à 0. L'équation générale établie précédemment et le théorème de Gauss permettront alors de calculer facilement la charge d'inversion. La *figure 3.11* illustre l'application du théorème de Gauss au calcul du champ à l'interface.

La surface délimitée sur la figure comprend la charge d'inversion mais pas les charges positives à l'interface oxyde-semi-conducteur. Elle passe dans la partie neutre du dispositif. Le flux à travers cette surface est donc égal à la somme des charges à l'intérieur ce qui s'écrit :

$$\left(-\frac{\mathrm{d}V}{\mathrm{d}y}\right)_{y=0} = \frac{Q'_I + Q'_B}{\varepsilon_s}$$

Reprenons maintenant l'équation générale en l'appliquant à *y* égal à zéro et en supposant que le potentiel V_s est très supérieur à ϕ_t soit 26 mV à température ambiante.

$$\left(\frac{\mathrm{d}V}{\mathrm{d}y}\right)^2_{y=0} = \frac{2eN_A}{\varepsilon_s}\left[\phi_t \exp^{-\frac{V_s}{\phi_t}} + V_s - \phi_t + \exp^{-2\frac{\phi_F}{\phi_t}}\left(\phi_t \exp^{\frac{V_s}{\phi_t}} - V_s - \phi_t\right)\right]$$

De plus, le quotient ϕ_F/ϕ_t est largement supérieur à l'unité comme le montre la *figure 3.9*. La formule se simplifie, il reste :

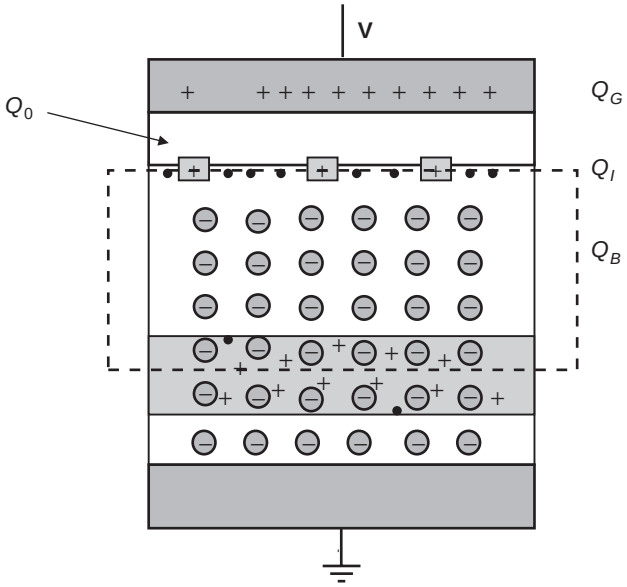


Figure 3.11 – Le calcul du champ à l'interface.

$$\left(\frac{\mathrm{d}V}{\mathrm{d}y}\right)^2_{y=0} = \frac{2eN_A}{\varepsilon_s}\left[V_s + \phi_t \exp^{\frac{V_s - 2\phi_F}{\phi_t}}\right]$$

On en déduit donc :

$$\left(\frac{\mathrm{d}V}{\mathrm{d}y}\right)_{y=0} = \pm\left(\frac{2eN_A}{\varepsilon_s}\left[V_s + \phi_t \exp^{\frac{V_s - 2\phi_F}{\phi_t}}\right]\right)^{1/2}$$

et ensuite,

$$Q'_I + Q'_B = \pm\varepsilon_s\sqrt{\frac{2eN_A}{\varepsilon_s}\left(V_s + \phi_t \exp^{\frac{V_s - 2\phi_F}{\phi_t}}\right)}$$

On choisira pour des raisons physiques la valeur négative. La charge qui nous intéresse est uniquement Q_I, il faut donc maintenant calculer Q_B. Pour cela, il suffit de supposer que la charge d'inversion est infiniment mince. La charge Q_B est alors le produit de la densité de dopant par le volume de la zone de charge d'espace de profondeur y_b. Ce calcul s'effectue simplement en écrivant l'équation de Poisson dans le semi-conducteur.

$$\varepsilon_s \frac{\mathrm{d}^2V}{\mathrm{d}y^2} = eN_A$$

On en déduit que V varie avec le carré de y,

$$\varepsilon_s V = \frac{eN_A}{2}(y - y_b)^2$$

On en déduit la valeur de y_b en fonction de V_s :

$$\varepsilon_s V_s = eN_A \frac{y_b^2}{2}$$

Le calcul de Q'_B est alors simple ainsi que celui de Q'_I.

$$Q'_B = -\sqrt{2eN_A\varepsilon_s}\sqrt{V_s}$$

La charge d'inversion est donc :

$$Q'_I = -\sqrt{2eN_A\varepsilon_s}\left[\sqrt{V_s + \phi_t \exp^{\frac{V_s - 2\phi_F}{\phi_t}}} - \sqrt{V_s}\right] \qquad (3.21)$$

Cette relation est fondamentale et mérite d'être un peu commentée. La charge d'inversion est proportionnelle à la surface du dispositif et au dopage. Elle est nulle quand le potentiel de surface est faible devant le potentiel de Fermi puis augmente exponentiellement au-delà de cette valeur.

Rappelons que le potentiel de surface n'est pas le potentiel appliqué sur la grille. Il reste encore à établir la relation entre ces deux potentiels. Pour cela, il faut revenir aux équations générales du dispositif complétées par la relation établie précédemment. Le potentiel extérieur appliqué V_{ext} sera noté V_G.

$$Q'_G + Q'_0 + Q'_I + Q'_B = 0$$

$$V_G = V_{\text{ox}} + V_s + \phi_{MS}$$

$$Q'_G = C'_{\text{ox}} V_{\text{ox}}$$

$$Q'_I + Q'_B = -\varepsilon_s \sqrt{\frac{2eN_A}{\varepsilon_s}\left(V_s + \phi_t \exp^{\frac{V_s - 2\phi_F}{\phi_t}}\right)}$$

Ce système de quatre équations permet de calculer les charges et potentiels inconnus : $Q'_I + Q'_B$, Q'_G, V_s et V_{ox}. On en déduit facilement :

$$V_G = \phi_{MS} - \frac{Q'_0}{C'_{\text{ox}}} + V_s + \frac{1}{C'_{\text{ox}}}\sqrt{2eN_A\varepsilon_s}\sqrt{V_s + \phi_t \exp^{\frac{V_s - 2\phi_t}{\phi_t}}} \qquad (3.21)$$

Cette relation permet de calculer le potentiel de surface en fonction de la tension extérieure appliquée entre grille et contact arrière. Il n'y a pas de solution analytique. L'hypothèse que nous allons faire maintenant est uniquement valable en forte inversion, c'est-à-dire pour des tensions appliquées suffisamment élevées. On suppose que le potentiel de surface est constant. Cette hypothèse semble peu fondée. Il suffit cependant d'examiner les résultats fournis par la résolution numérique de l'équation donnant le potentiel de surface pour s'apercevoir qu'il varie assez peu quand la tension appliquée sur la grille augmente. La *figure 3.12* illustre ce comportement.

Tout se passe comme si, au-delà d'une certaine valeur, toute la tension était appliquée aux bornes de la couche d'oxyde. Si on considère l'écran électrique formé par la couche d'inversion, ce phénomène apparaît peu surprenant. La suite du calcul consiste à prendre une valeur bien choisie pour cette valeur constante. Elle est notée ϕ_B. Certains auteurs choisissent alors $2\,\phi_F$ en se basant sur le terme exponentiel de la formule. D'autres choisiront une valeur légèrement supérieure en prenant par exemple $2\,\phi_F + 6\,\phi_t$.

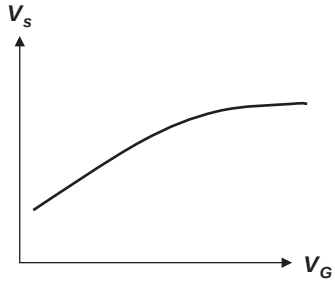


Figure 3.12 – Variation du potentiel de surface.

Reprenons maintenant certaines des équations générales du système pour calculer la charge d'inversion.

$$Q'_G + Q'_0 + Q'_I + Q'_B = 0$$
$$V_G = V_{ox} + V_s + \phi_{MS}$$
$$Q'_G = C'_{ox} V_{ox}$$
$$Q'_B = -\sqrt{2eN_A\varepsilon_s}\sqrt{V_s}$$

À partir de ces quatre équations et en prenant une valeur donnée du potentiel de surface, on obtient facilement :

$$Q'_I = -C'_{ox}(V_G - V_T) \tag{3.23}$$

Dans cette formule le paramètre V_T est appelé tension de seuil. Il est défini par :

$$V_T = \phi_{MS} - \frac{Q'_0}{C'_{ox}} + \phi_B + \frac{1}{C'_{ox}}\sqrt{2eN_A\varepsilon_s}\sqrt{\phi_B} \tag{3.24}$$

Ces deux formules sont fondamentales dans l'étude des MOS. Elles quantifient la formation de la couche d'inversion. La tension de seuil dépend du paramètre ϕ_{MS} variant lui-même avec la nature du matériau de grille, des charges de surface et d'un terme en racine de la tension. Ce dernier terme est appelé effet « body » dans la littérature. Le coefficient $\sqrt{2eN_A\varepsilon_s}/C'_{ox}$ est appelé coefficient γ.

Pour donner à cette formule un sens physique, on peut dire que le potentiel doit franchir un certain nombre de barrières pour créer une charge d'inversion : le potentiel de contact ϕ_{MS}, puis la barrière des charges de surface, puis la barrière du silicium et enfin la barrière de la charge de la zone de charge d'espace. Le lecteur comprendra cependant que la valeur de la tension de seuil n'est pas évidente comme certains ouvrages pourraient parfois le laisser supposer.

Pour terminer cette description du dispositif MOS, il est possible de donner quelques valeurs numériques des grandeurs physiques pour différentes technologies : *tableau 3.2.*

Les valeurs de ce tableau sont soit données par la technologie soit calculées par les formules des paragraphes précédents. C'est le cas de la profondeur de la zone de charge d'espace, du coefficient γ et de la tension de seuil. La valeur de la capacité de l'oxyde par unité de surface est prise à 0,035 fF par micron d'épaisseur pour du dioxyde de silicium. On constate une diminution de la tension de seuil quand la technologie est plus fine, ce qui est un phénomène très important dans l'industrie du semi-conducteur.

Tableau 3.2

	0,80 micron	0,18 micron	0,045 micron
Épaisseur de l'oxyde en nm	14	4	1,5
C'_{ox} en fF/µ³	2	4	25
N_A en atomes/µ³	10^{14}	10^{15}	10^{16}
Zone de charge d'espace en micron	0,8	0,10	0,05
Charges de surface en fF/µ³	0,1	0,08	0,001
ϕ_B en V	0,75	0,80	0,85
γ en $V^{1/2}$	0,4	0,2	0,15
Tension de seuil en V	0,26	0,18	0,13

3.3.4 *Le régime de faible inversion*

Dans le régime de faible inversion, la tension appliquée sur la grille est plus faible mais le potentiel de surface est encore supposé très supérieur à ϕ_t. Par contre, il est inférieur à $2\phi_F$. Les approximations de la formule générale ne sont plus valables. Ce régime était exceptionnel dans les anciennes technologies mais, pour les technologies avancées et en particulier pour les circuits à faible consommation, ce régime de fonctionnement est relativement standard.

Reprenons l'équation générale 3.21 en supposant cette fois que le terme $\phi_t \exp^{\frac{V_s - 2\phi_F}{\phi_t}}$ noté alors ξ est largement inférieur à V_s. L'équation s'écrit, rappelons-le :

$$Q'_I = -\sqrt{2eN_A\varepsilon_s}\left[\sqrt{V_s + \phi_t \exp^{\frac{V_s - 2\phi_t}{\phi_t}}} - \sqrt{V_s}\right]$$

soit,

$$Q'_I = -\sqrt{2eN_A\varepsilon_s}[\sqrt{V_s + \xi} - \sqrt{V_s}]$$

Un développement limité de la racine carrée permet d'écrire :

$$Q'_I = -\sqrt{2eN_A\varepsilon_s}\,\frac{\xi}{2\sqrt{V_s}}$$

Reprenons alors les équations du système.

$$Q'_G + Q'_0 + Q'_I + Q'_B = 0$$
$$V_G = V_{OX} + V_s + \phi_{MS}$$
$$Q'_G = C'_{OX} V_{OX}$$
$$Q'_B = -\sqrt{2eN_A\varepsilon_s}\sqrt{V_s}$$
$$Q'_I = -\sqrt{2eN_A\varepsilon_s}\,\frac{\xi}{2\sqrt{V_s}}$$

Pour simplifier le calcul, on néglige la charge d'inversion Q'_I devant la charge Q'_B ce qui permet d'écrire à partir des quatre premières équations :

$$V_G = \phi_{MS} - \frac{Q'_0}{C'_{OX}} + V_s + \gamma\sqrt{V_s}$$

Cette équation du deuxième degré en $\sqrt{V_s}$ permet de calculer le potentiel de surface en fonction du potentiel extérieur appliqué. Une seule des racines a un sens physique. On peut ensuite calculer la charge d'inversion avec cette valeur.

Il est cependant possible de simplifier encore en considérant que le potentiel de surface a une valeur donnée égale à $1{,}5\ \phi_F$ dans le terme en racine carrée. C'est une valeur intermédiaire entre 0 et $2\ \phi_F$, valeur à partir de laquelle commence le régime d'inversion forte. On exprime alors la pente de la fonction donnant la valeur de V_G au voisinage de $1{,}5\ \phi_F$ par :

$$n_0 = \frac{\mathrm{d}V_G}{\mathrm{d}V_s} = 1 + \frac{\gamma}{2\sqrt{1{,}5\ \phi_F}}$$

On peut alors calculer le potentiel de surface en fonction de la tension de grille en faisant une approximation linéaire.

$$V_s = \frac{1}{n_0}\left(V_G - \phi_{MS} + \frac{Q'_0}{C'_{OX}}\right)$$

Appelons V_X la valeur de la tension de grille pour laquelle le potentiel de surface est égal à $1{,}5\ \phi_F$.

$$V_s - 1{,}5\ \phi_F = \frac{V_G - V_X}{n_0}$$

$$V_X = \phi_{MS} - \frac{Q'_0}{C'_{OX}} + 1{,}5\ \phi_F + \gamma\sqrt{1{,}5\ \phi_F}$$

Cette tension V_X est donc équivalente à la tension de seuil en régime de forte inversion.

La relation donnée au début de ce paragraphe permet de calculer la charge d'inversion :

$$Q'_I = -\sqrt{2eN_A\varepsilon_s}\,\frac{\phi_t \exp^{\frac{-0{,}5\,\phi_F}{\phi_t}}}{2\sqrt{1{,}5\ \phi_F}} \exp^{\frac{V_G - \phi_{MS} + \frac{Q_0}{C_{OX}} - 1{,}5\,\phi_F - \gamma\sqrt{1{,}5\,\phi_F}}{\left(1 + \frac{\gamma}{2\sqrt{1{,}5\,\phi_F}}\right)\phi_t}}$$

Il est maintenant possible d'exprimer cette charge en fonction de la tension V_X. On obtient alors :

$$Q'_I = Q'_{I0} \exp^{\frac{V_G - V_X}{n_0\phi_t}} \tag{3.25}$$

Dans cette formule, plus lisible, les paramètres sont définis par :

$$Q'_{I0} = -\sqrt{2e\varepsilon_s N_A}\,\frac{\phi_t \exp^{\frac{-0{,}5\,\phi_F}{\phi_t}}}{2\sqrt{1{,}5\ \phi_F}}$$

$$n_0 = 1 + \frac{\gamma}{2\sqrt{1{,}5\,\phi_F}}$$

Il est également intéressant d'exprimer le facteur n appelé *ideality* en fonction des capacités du dispositif. Il est défini à partir de l'équation obtenue précédemment :

$$V_G = \phi_{MS} - \frac{Q_0'}{C_{OX}'} + V_s + \gamma\sqrt{V_s}$$

$$n = \frac{dV_G}{dV_s} = 1 + \frac{\gamma}{2\sqrt{V_s}}$$

Comme le coefficient γ est donné par $\sqrt{2e\varepsilon_s N_A}/C_{OX}'$, on obtient :

$$\frac{\gamma}{2\sqrt{V_s}} = \frac{\sqrt{2e\,\varepsilon_s N_A}}{2C_{OX}'\sqrt{V_s}}$$

soit,

$$\frac{\gamma}{2\sqrt{V_s}} = \frac{C_{si}'}{C_{OX}'}$$

Dans cette formule, on reconnaît la capacité de la zone de charge d'espace et la capacité de la couche d'oxyde.

$$n = \frac{dV_G}{dV_s} = 1 + \frac{C_{si}'}{C_{OX}'} \tag{3.26}$$

Cette relation importante sera reprise dans l'étude du MOSFET. La valeur de n est idéalement 1, car dans ce cas la grille contrôle totalement le potentiel dans le silicium. En pratique, elle est comprise entre 1 et 1.6. On comprend l'intérêt d'augmenter la capacité de l'oxyde et de diminuer la capacité de la zone de charge d'espace dans le silicium. Le paramètre n_0 est la valeur particulière de n pour V_S égal à $1{,}5\,\phi$.

3.4 Annexe : calcul du potentiel dans le dispositif MOS

On part de l'équation de Poisson.

$$\frac{d^2V}{dy^2} + e\left(\frac{p(y) - n(y) - N_A}{\varepsilon_s}\right) = 0$$

avec,

$$\frac{n(y)}{n_b} = \exp^{\frac{V(y)}{\phi_t}}$$

$$\frac{p(y)}{p_b} = \exp^{\frac{V(y)}{\phi_t}}$$

De plus,

$$p_b - n_b = N_A$$

On obtient alors,

$$\frac{\mathrm{d}^2 V}{\mathrm{d}y^2} + \frac{e}{\varepsilon_s}\left[p_b\left(\exp^{-\frac{V(y)}{\phi_t}} - 1\right) - n_b\left(\exp^{\frac{V(y)}{\phi_t}} - 1\right)\right] = 0$$

Si on suppose N_A largement supérieur à n_b, on obtient :

$$\frac{\mathrm{d}^2 V}{\mathrm{d}y^2} + \frac{eN_A}{\varepsilon_s}\left[\exp^{-\frac{V(y)}{\phi_t}} - 1 - \exp^{-\frac{2\phi_F}{\phi_t}}\left(\exp^{\frac{V(y)}{\phi_t}} - 1\right)\right] = 0$$

Si on multiplie cette équation par $2\frac{\mathrm{d}V}{\mathrm{d}y}$, on obtient alors :

$$\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{\mathrm{d}V}{\mathrm{d}y}\right)^2 + \frac{2eN_A}{\varepsilon_s}\left[\frac{\mathrm{d}V}{\mathrm{d}y}\exp^{-\frac{V(y)}{\phi_t}} - \frac{\mathrm{d}V}{\mathrm{d}y} - \frac{\mathrm{d}V}{\mathrm{d}y}\exp^{-\frac{2\phi_F}{\phi_t}}\left(\exp^{\frac{V(y)}{\phi_t}} - 1\right)\right] = 0$$

Cette équation s'intègre entre un point y et un point en profondeur dans le silicium pour lequel le champ $\frac{\mathrm{d}V}{\mathrm{d}y}$ est nul ainsi que le potentiel $V(y)$.

$$\left(\frac{\mathrm{d}V}{\mathrm{d}y}\right)^2 = \frac{2eN_A}{\varepsilon_s}\left[\phi_t \exp^{-\frac{V(y)}{\phi_t}} + V(y) - \phi_t + \exp^{-\frac{2\phi_F}{\phi_t}}\left(\phi_t \exp^{\frac{V(y)}{\phi_t}} - V(y) - \phi_t\right)\right]$$

Chapitre 4

Le transistor MOSFET et son évolution

Le but de ce chapitre est d'expliquer le fonctionnement du dispositif de base de la micro-électronique actuelle : le transistor MOS. La très grande majorité des circuits intégrés logiques et analogiques sont fabriqués en combinant des transistors de type MOS (*Metal Oxyde Semiconductor*) et des composants passifs (principalement des résistances et des condensateurs). Le fonctionnement du transistor MOS est simple dans le principe mais assez complexe quand on rentre dans le détail et quand on s'intéresse à son optimisation. Les résultats de ce chapitre permettent d'établir des modèles électriques qui seront utilisés dans les chapitres suivants. La modélisation du transistor s'appuie sur les résultats extraits de la référence [7].

4.1 Principe de base et brève histoire

L'idée initiale fut de contrôler la conduction d'un fil par une grille de la même manière qu'une grille contrôle le courant émis par le filament d'une triode et qu'un robinet contrôle le débit de l'eau. Un fil isolant ne pouvait convenir car le courant traversant un isolant est nul. Un fil conducteur semblait peu utilisable car le champ ne pénètre pas dans un conducteur. Il était donc difficile d'imaginer dans ce cas un mode de contrôle. Le semi-conducteur apparut alors comme le bon matériau puisqu'il offrait les deux propriétés de base : possibilité de laisser passer un courant et pénétration du champ à l'intérieur. Le premier dispositif a été imaginé et breveté par Lilienfield en 1933.

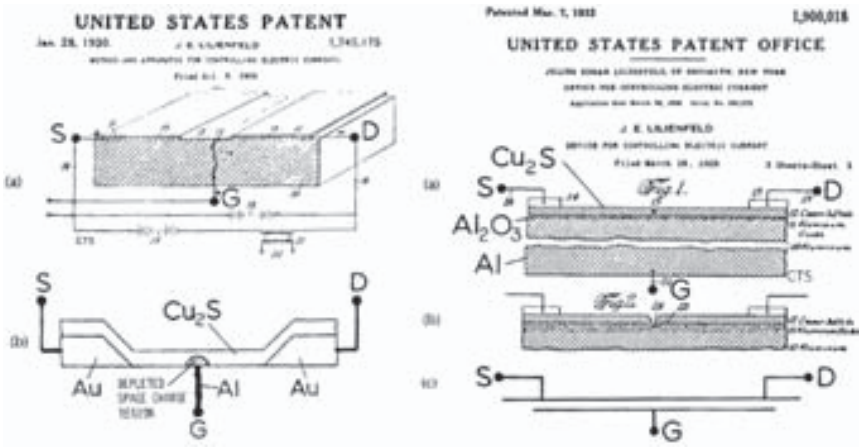


Figure 4.1 – Les brevets de Lilienfield.

Ces brevets illustrent le principe général du contrôle de la conduction d'un semi-conducteur. Le matériau semi-conducteur choisi était le sulfure de cuivre présentant un comportement de type *p*. Il n'a pas connu un grand avenir par la suite mais le principe était posé. Ces dispositifs étaient du type MOS à appauvrissement. Le principe du MOS à appauvrissement est de rejeter les porteurs hors du canal de conduction en appliquant une tension. Ce n'est qu'en 1948 que Bardeen eut l'idée de reprendre ce principe mais en créant une couche d'inversion, c'est-à-dire formée par les porteurs minoritaires et cette fois en augmentant la densité de porteurs par application d'une tension. Les deux principes de fonctionnement sont illustrés *figure 4.2*.

Étudions, dans un premier temps, le fonctionnement du MOSFET à appauvrissement. Le dispositif est un morceau de semi-conducteur appelé *bulk* dans lequel sont créées deux régions fortement dopées de type *n* qui jouent le rôle de réservoirs d'électrons. Ce sont la source et le drain. Dans le MOS à appauvrissement une zone supplémentaire de type *n* est créée entre source et drain. On suppose, pour simplifier, que la face arrière du semi-conducteur et la source sont reliées électriquement. Une tension positive V_{DS} entre drain et source a pour effet de faire passer un courant de conduction (courant de dérive) à la condition qu'il y ait des électrons dans la zone de type *n*. Quand on applique une tension nulle sur la grille, les électrons sont en grand nombre puisqu'ils proviennent de la zone dopée *n*. Un courant circule de la source vers le drain. Quand une tension négative est appliquée sur la grille, elle attire les trous du semi-conducteur qui se recombinent aux électrons du canal si bien que la densité d'électrons de conduction diminue. En conséquence, le courant diminue.

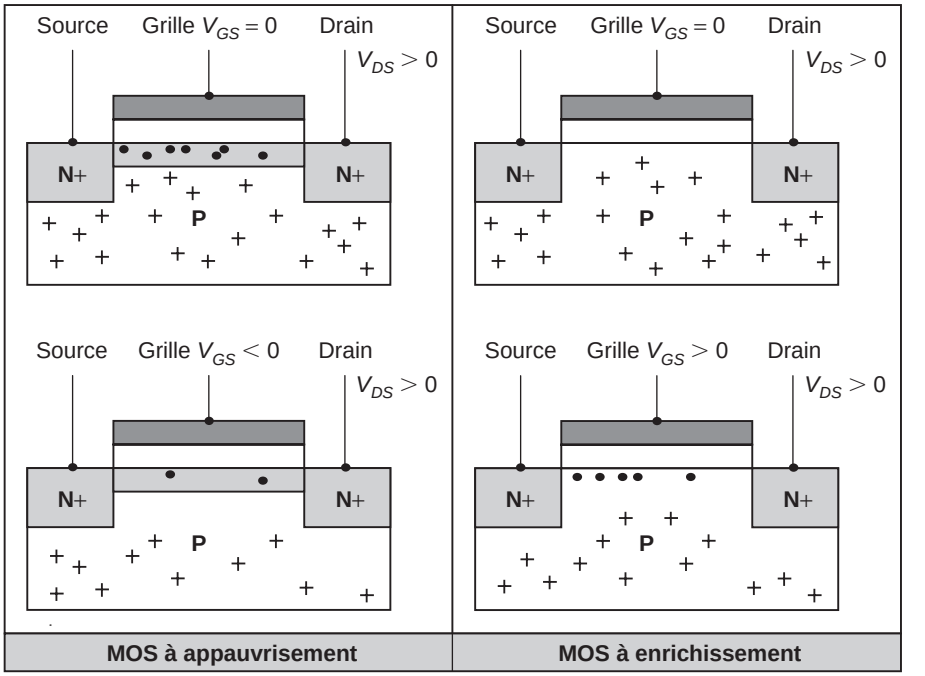


Figure 4.2 – Les deux types de MOS.

Remarquons que les jonctions du dispositif sont soit polarisées en inverse (jonction bulk-drain) soit non polarisées (jonction source-bulk). Les courants de jonction sont très faibles. Le courant de drain est uniquement dû à la conduction dans le canal. Il est contrôlé par la tension de grille.

Dans le MOS à enrichissement, il n'y a plus de zone dopée servant de canal de conduction. Les trous du matériau de base ne peuvent donner lieu à un courant puisque les deux jonctions source-bulk et bulk-drain sont respectivement non polarisée et polarisée en inverse. Seuls, les électrons peuvent créer un courant dans ce type de dispositif. Quand une tension nulle est appliquée sur la grille, les électrons ne sont pas injectés dans le semi-conducteur et aucun courant ne circule de la source vers le drain. Quand une tension positive est appliquée sur la grille, elle attire des électrons fournis par la source et le drain et un courant peut alors s'établir.

Les transistors MOS à appauvrissement ont été progressivement abandonnés et les transistors à enrichissement se sont imposés dans l'industrie micro-électronique. La fabrication est en effet plus simple. De plus, les MOS à enrichissement permettent de réaliser des circuits consommant très peu ce qui a donné lieu à la technologie CMOS avec laquelle on réalise aujourd'hui la majorité des circuits intégrés. Les dispositifs à canal *n* ne sont pas les seuls à être fabriqués. Des dispositifs équivalents peuvent être réalisés en jouant sur la conduction des trous. Le canal est alors de type *p*. Une tension négative de grille est dans ce cas appliquée pour enrichir le canal. Ces deux types de transistors à enrichissement, MOS canal *n* et MOS canal *p*, sont les deux briques de base de la technologie CMOS.

Étudions maintenant le fonctionnement du transistor à enrichissement. La *figure 4.3* reprend le fonctionnement du MOS à enrichissement et explique comment la conduction varie avec les tensions appliquées.

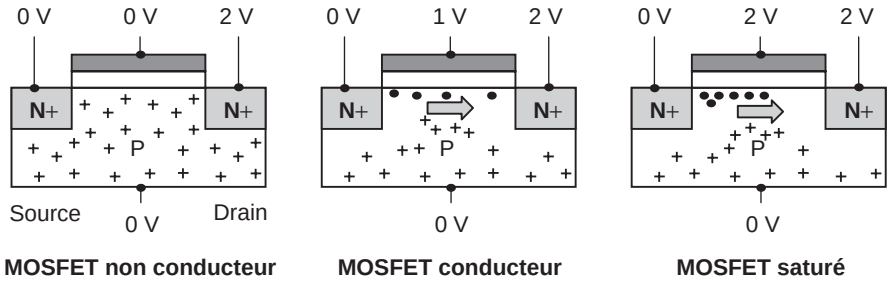


Figure 4.3 – Fonctionnement du FET canal n à enrichissement.

Dans une première étape, la grille est à tension nulle et le drain est polarisé positivement. Comme il a été vu dans le chapitre 3, aucun électron n'est confiné à l'interface semi-conducteur isolant, et de ce fait il n'y a pas de courant. Le MOSFET est non conducteur.

Quand une tension positive est appliquée sur la grille, une couche d'inversion se forme et on passe du régime de faible inversion au régime de forte inversion au fur et à mesure que la tension de grille augmente. Ce phénomène apparaît quand la tension de grille est supérieure à une tension dite de seuil de l'ordre de 0,4 V. Les électrons sont fournis par la source et un courant peut circuler de la source vers le drain sous l'effet du champ électrique présent dans le dispositif. Le MOSFET est conducteur.

Si maintenant, la tension de grille restant constante, on augmente la tension du drain, la différence de potentiel entre la grille et la zone du canal proche du drain peut devenir inférieure à la tension de seuil. La couche d'inversion est alors nulle en bout de canal. Ce dernier régime est appelé régime de saturation. On peut considérer le canal comme la mise en série d'une zone de conduction faiblement résistive et d'une jonction polarisée en inverse. Toute augmentation supplémentaire de la tension de drain se traduit par une augmentation de la tension aux bornes de la jonction *pn* en bout de canal et aucune augmentation de tension ne peut alors se manifester aux bornes du canal de conduction. Le courant de drain reste donc constant.

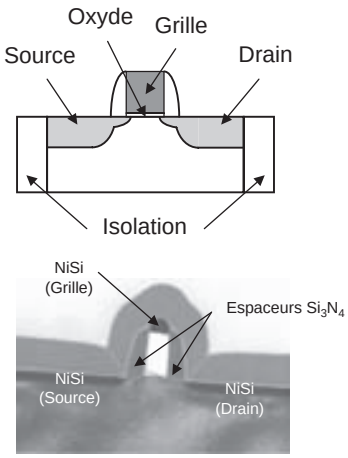


Figure 4.4 – Le MOSFET réel dans une technologie avancée (Crédit photo CEA Grenoble).

La géométrie décrite précédemment est en fait très simplifiée. La géométrie réelle est représentée *figure 4.4*. La longueur du canal est très faible (moins de 90 nm pour les technologies les plus avancées), l'épaisseur de l'oxyde est de quelques nanomètres pour que l'influence de la grille soit maximale. La largeur du transistor est définie par les concepteurs mais de l'ordre du micron pour les technologies numériques.

Remarquons la forme particulière des régions de drain et de source au niveau du canal. Cette forme adoucie du dopage permet de limiter le champ électrique dans ces régions.

4.2 Comment la structure MOS se modifie

Étudions plus en détail le fonctionnement du transistor à partir des schémas exposés dans le paragraphe 4.3. Le fonctionnement du transistor est compris quand on peut exprimer le courant sortant du drain en fonction des tensions appliquées. On distingue trois régimes : le régime statique, le régime quasi-statique et le régime dynamique.

- Dans le régime statique, les électrons sortant du drain sont exactement compensés par ceux qui entrent dans la source. Les charges sont en mouvement dans le canal mais tout se passe comme si elles étaient fixes pour le calcul des potentiels puisque la charge totale des électrons du canal est constante. Ce régime est également appelé régime continu. Le courant dans le canal est constant en fonction de la position le long du canal.
- Dans le régime quasi-statique, les tensions appliquées varient mais suffisamment lentement pour que les électrons sortants soient compensés par les électrons entrants. La relation donnant le courant en fonction des tensions appliquées est donc la même que dans le cas précédent mais les tensions sont des fonctions du temps.
- Dans le régime dynamique, les charges ne sont pas intégralement compensées et le courant ne peut plus être considéré comme constant dans le canal de conduction. Il faut alors écrire une relation locale de conservation du courant et résoudre le problème. Ce cas plus difficile sera traité dans le paragraphe 4.5.

Le cas statique peut sembler identique au système MOS décrit dans le chapitre 3. Il y a cependant une différence fondamentale dans l'origine des électrons qui constituent le canal de conduction. Dans la structure MOS, les électrons viennent du substrat et sont attirés par la grille. Dans le transistor MOS, les électrons sont fournis par la source. Cette différence se traduit également par une différence dans la valeur des potentiels chimiques ce qui conduit à des expressions différentes des concentrations de porteurs. La *figure 4.5* représente les deux structures : la structure MOS étudiée dans le chapitre 3 et le transistor MOS étudié dans ce chapitre.

Si on examine la concentration des électrons au voisinage de la source (ligne en pointillés sur la figure) on peut écrire dans le cas de la simple structure MOS :

$$n_s = n_i \exp^{\frac{E_F - E_{is}}{k_B T}}$$

Dans cette expression, le potentiel chimique E_F est supposé constant dans le dispositif car il n'y a pas transfert de charges entre régions. La concentration n_b des électrons en profondeur dans la région dopée *p* s'écrit :

$$n_b = n_i \exp^{\frac{E_F - E_{i0}}{k_B T}}$$

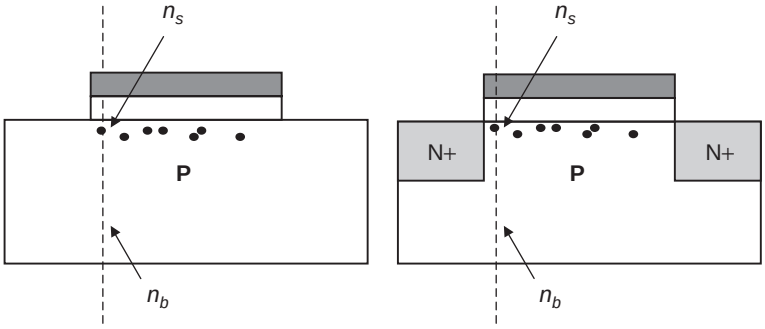


Figure 4.5 – Modification des concentrations.

Dans le dispositif MOSFET le potentiel chimique ne peut être considéré comme constant pour les électrons car il y a transfert d'électrons de la source au canal. On écrit alors en appelant E_{Fn} le potentiel chimique pour les électrons :

$$n_s = n_i \exp^{\frac{E_{Fn} - E_{is}}{k_B T}}$$

En fait, la grandeur E_{Fn} n'est plus un potentiel chimique pour l'ensemble des porteurs du dispositif, électrons et trous. On admet que chaque population est en équilibre avec son propre potentiel chimique. La grandeur E_{Fn} est un pseudo potentiel chimique défini pour les électrons. En profondeur, on suppose l'équilibre. Le potentiel chimique des électrons est égal à celui des trous. Il est noté E_F.

$$n_b = n_i \exp^{\frac{E_F - E_{ib}}{k_B T}}$$

Fort heureusement, il y a une relation simple entre le potentiel chimique dans le canal et le potentiel chimique en profondeur dans le dispositif comme il a été vu dans le chapitre 2.

$$E_{Fn} - E_F = -eV_{SB}$$

V_{SB} est la tension extérieure appliquée entre la source et le substrat.

Le potentiel chimique dans le canal au niveau de la source est en effet égal à celui qui règne dans la source car il y a échange d'électrons entre source et canal avec la formation d'un état d'équilibre local. On écrit alors :

$$\frac{n_s}{n_b} = \exp^{(E_{Fn} - E_F - E_{is} + E_{ib})/k_B T}$$

La différence d'énergie $E_{ib} - E_{is}$ est la différence de potentiel V_s entre la zone profonde et le canal, multipliée par l'inverse de la charge de l'électron. On obtient donc :

$$\frac{n_s}{n_b} = \exp^{\frac{-eV_{SB} + eV_s}{k_B T}}$$

Dans la structure MOS, la relation s'écrivait simplement :

$$\frac{n_s}{n_n} = \exp^{\frac{+eV_s}{k_BT}}$$

Les relations sont inchangées pour les trous car il n'y a pas d'échange avec la source. Ces considérations s'appliquent également pour exprimer la densité d'électrons non pas en surface mais en profondeur dans le silicium.

Pour calculer les densités de charge il suffit donc de remplacer $V(y)$ par $V(y) - V_{SB}$ quand on exprime $n(y)$ dans l'équation générale au niveau de la source. La relation générale devient :

$$\left(\frac{\mathrm{d}V}{\mathrm{d}y}\right)^2 = \frac{2eN_A}{\varepsilon_s}\left[\phi_t \exp^{-\frac{V(y)}{\phi_t}} + V - \phi_t + \exp^{-\frac{2\phi_F}{\phi_t}}\left(\phi_t \exp^{\frac{V(y)-V_{SB}}{\phi_t}} - V - \phi_t\right)\right] = 0$$

À cette différence près, toutes les relations établies dans le chapitre restent valables. On peut ainsi écrire la relation donnant le potentiel de surface au niveau de la source :

$$V_{GB} = \phi_{MS} - \frac{Q_0'}{C_{OX}'} + V_s(0) + \frac{1}{C_{OX}'}\sqrt{2eN_A\varepsilon_s}\sqrt{V_S(0) + \phi_t \exp^{\frac{V_s(0)-2\phi_F-V_{SB}}{\phi_t}}}$$

Si on fait le même raisonnement au niveau du drain, le potentiel chimique E_{Fn}' des électrons du canal au niveau du drain est alors :

$$E_{Fn}' - E_F = -eV_{DB}$$

Remarquons que le potentiel chimique des électrons au niveau du drain a une valeur différente de celui au niveau de la source. Il varie en fait tout le long du canal. Cela n'est pas étonnant si on reprend les équations générales donnant concentrations et courant d'une population d'électrons.

$$n(x) = n_i \exp^{\frac{E_{Fn}-E_i}{k_BT}}$$

$$J_n = e\mu_n n E_x + eD_n\frac{\mathrm{d}n}{\mathrm{d}x}$$

On écrit alors :

$$\frac{\mathrm{d}n}{\mathrm{d}x} = \frac{n}{k_BT}\left(\frac{\mathrm{d}E_{Fn}}{\mathrm{d}x} - \frac{\mathrm{d}E_i}{\mathrm{d}x}\right)$$

$$E_x = -\frac{1}{e}\frac{\mathrm{d}E_i}{\mathrm{d}x}$$

On en déduit :

$$J_n = \mu_n n\frac{\mathrm{d}E_{Fn}}{\mathrm{d}x}$$

On vérifie bien que le passage d'un courant s'accompagne d'une variation du pseudo-potentiel chimique de la population d'électrons. Une relation équivalente peut s'écrire pour une population de trous. La relation donnant le potentiel de surface s'écrit donc au niveau du drain :

$$V_{GB} = \phi_{MS} - \frac{Q_0'}{C_{OX}'} + V_s(L) + \frac{1}{C_{OX}'}\sqrt{2eN_A\varepsilon_s}\sqrt{V_S(L) + \phi_t \exp\frac{V_s(L) - 2\phi_F - V_{DB}}{\phi_t}}$$

Nous allons maintenant chercher à exprimer le courant fourni par le transistor en régime continu en fonction des tensions appliquées. Le point de départ est l'écriture de la densité de courant dans un semi-conducteur comme elle a été calculée dans le chapitre 2.

$$J_n = e\mu_n n E_x + e D_n \frac{\mathrm{d}n}{\mathrm{d}x}$$

On ne considère que la partie longitudinale du courant correspondant à l'axe des *x*. Dans les autres dimensions on admet que le courant est nul. Cette hypothèse n'est vraie que pour un canal long devant les dimensions transverses. On considère également que le courant transverse de diffusion compense exactement le courant transverse de dérive comme le montre la *figure 4.6*.

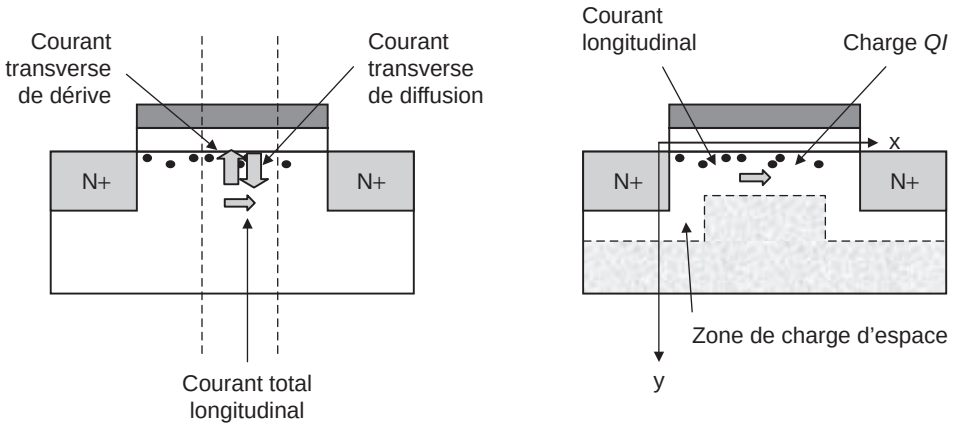


Figure 4.6 – Le courant dans un transistor.

Si on appelle Q_I' la charge des électrons du canal par unité de surface, *W* et *L* la largeur du dispositif et la longueur du canal comme le montre la *figure 4.7*, le courant traversant le dispositif s'écrit alors :

$$I_D = \mu_n W(-Q_I')\frac{\mathrm{d}V_s}{\mathrm{d}x} + \mu_n W \phi_t \frac{\mathrm{d}Q_I'}{\mathrm{d}x}$$

Le passage de J_D (densité de courant) à I_D (courant) n'est pas tout à fait évident. Il amène à remplacer *n* par WQ_I' comme le lecteur pourra le vérifier.

Comme le courant est supposé constant le long de l'axe des *x* (cas du régime continu), cette équation s'intègre très facilement de la source ($x = 0$) au drain ($x = L$). On obtient alors :

$$\int_0^L I_D \,\mathrm{d}x = \mu_n W \int_{V_s(0)}^{V_s(L)} (-Q_I')\frac{\mathrm{d}V_s}{\mathrm{d}x}\,\mathrm{d}x + \mu_n W \phi_t \int_{Q_I'(0)}^{Q_I'(L)} \frac{\mathrm{d}Q_I'}{\mathrm{d}x}\,\mathrm{d}x = I_D L \qquad (4.1)$$

Pour aller plus loin dans le calcul, il faut exprimer la charge par unité de surface et le potentiel de surface en fonction des tensions appliquées. Ce calcul sera effectué dans le paragraphe suivant dans deux cas : le régime de forte inversion et le régime de faible inversion.

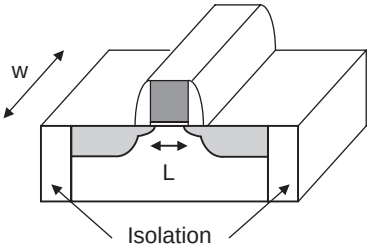


Figure 4.7 – Le transistor 3D et ses dimensions.

4.3 Le modèle canal long

4.3.1 *Le régime de forte inversion*

Reprenons l'équation 4.1 en explicitant les valeurs de la charge Q'_I et du potentiel de surface $V_s(x)$. les résultats sont tirés du paragraphe 4.2 et du chapitre 3.

$$Q'_G + Q'_0 + Q'_I + Q'_B = 0 \tag{3.19}$$

$$V_{GB} = V_{OX}(x) + V_S(x) + \phi_{MS} \tag{3.9}$$

$$Q'_G = C'_{OX} V_{OX}(x) \tag{3.14}$$

$$Q'_B = -\sqrt{2eN_A\varepsilon_s}\sqrt{V_S(x)}$$

On écrit facilement à partir de ces équations :

$$Q'_I = -C'_{OX}(V_{GB} - V_{FB} - V_S - \gamma\sqrt{V_s})$$

avec,

$$V_{FB} = \phi_{MS} - \frac{Q'_0}{C'_{OX}}$$

et,

$$\gamma = \frac{\sqrt{2eN_A\varepsilon_s}}{C'_{OX}}$$

L'équation 4.1 s'intègre alors sous la forme :

$$I_D L = \mu_n W \int_{V_s(0)}^{V_s(L)} C'_{OX}(V_{GB} - V_{FB} - V_S - \gamma\sqrt{V_s})\frac{dV_s}{dx}dx + \mu_n W \phi_t \int_{Q'_I(0)}^{Q'_I(L)} \frac{dQ'_I}{dx}dx$$

soit,

$$I_D L = \mu_n W \int_{V_s(0)}^{V_s(L)} C'_{OX}(V_{GB} - V_{FB} - V_S - \gamma\sqrt{V_s})\,dV_s + \mu_n W \phi_t \int_{Q'_I(0)}^{Q'_I(L)} dQ'_I$$

Le courant de drain s'exprime alors facilement en fonction du potentiel de surface.

$$I_D L = \mu_n W C'_{OX}\left[(V_{GB} - V_{FB})[V_s(L) - V_s(0)] - \frac{1}{2}(V_s^2(L) - V_s^2(0)) - \frac{2}{3}\gamma\left(V_s^{\frac{3}{2}}(L) - V_s^{\frac{3}{2}}(0)\right)\right]$$

$$+ \mu_n W C'_{OX}\phi_t\left[V_s(L) - V_s(0) + \gamma\left(V_s^{\frac{1}{2}}(L) - V_s^{\frac{1}{2}}(0)\right)\right]$$

Il faut alors faire les hypothèses de forte inversion :

- le deuxième terme (courant de diffusion) est négligé ;
- le potentiel de surface est supposé de la forme :

$$V_S(x) = \phi_B + V_{CB}(x)$$

V_{CB} est égale à V_{SB} au niveau de la source et égale à V_{DB} au niveau du drain. Le terme ϕ_B est classiquement choisi égal à $2\,\phi_F$ (voir chapitre 3). Cette relation est une conséquence du régime de forte inversion et généralise les résultats du chapitre 3. Le potentiel de surface tend vers une valeur constante quand la tension de grille augmente.

On peut alors calculer le courant de drain. Après quelques manipulations algébriques, on obtient :

$$I_D = \frac{W}{L}\mu_n C'_{OX}\left((V_{GS} - V_{FB} - 2\,\phi_F)V_{DS} - \frac{1}{2}V_{DS}^2 - \frac{2}{3}\gamma\left[(2\,\phi_F + V_{SB} + V_{DS})^{\frac{3}{2}} - (2\,\phi_F + V_{SB})^{\frac{3}{2}}\right]\right) \quad (4.2)$$

Il est maintenant possible de tracer le courant en fonction de V_{DS}, les autres paramètres étant fixés, en particulier la tension de grille.

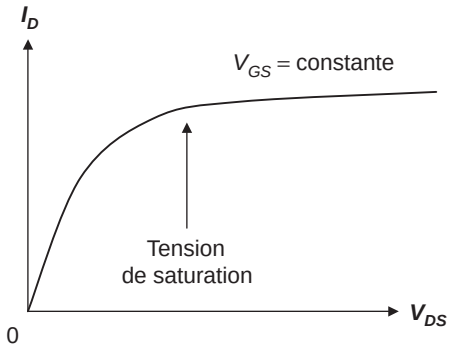


Figure 4.8 – Variation du courant de drain en forte inversion

On constate que la pente de cette courbe diminue quand la tension augmente. Elle s'annule pour une certaine valeur de V_{DS}. Cette valeur est la solution de l'équation :

$$\frac{dI_D}{dV_{DS}} = 0$$

On trouve alors :

$$V_{DSsat} = V_{GS} - 2\,\phi_F - V_{FB} + \frac{\gamma^2}{2} - \gamma\sqrt{V_{GS} - V_{FB} + V_{SB} + \frac{\gamma^2}{4}}$$

Rappelons que le paramètre γ a la dimension d'une racine carrée de tension et que sa valeur est d'environ $0{,}5\ \mathrm{V}^{0{,}5}$. Au-delà de cette valeur, le transistor n'est plus en régime de forte inversion au niveau du drain et la formule n'est plus valable. Si elle l'était, le courant de drain diminuerait ce qui est contraire au sens physique. Il reste à prouver que cette valeur de la tension de drain correspond à la valeur qui annule la charge d'inversion au niveau du drain.

Pour cela on utilise la relation générale 3.21 donnant la charge d'inversion au niveau du drain comme il a été expliqué dans le chapitre 3 mais en remplaçant V_s par $V_S - V_{DB}$:

$$Q'_I = -\sqrt{2e\varepsilon_s N_A}\left[\sqrt{V_s + \phi_t \exp\frac{V_S - V_{DB} - 2\phi_F}{\phi_t}} - \sqrt{V_s}\right]$$

Cette charge s'annule pour :

$$V_S = V_{DB} + 2\,\phi_F$$

En supposant nulle la charge d'inversion, les équations du dispositif au niveau du drain s'écrivent :

$$Q'_G + Q'_0 + Q'_B = 0$$
$$V_{GB} = V_{OX} + V_s + \phi_{MS}$$
$$Q'_G = C'_{OX} V_{OX}$$
$$Q'_B = -\sqrt{2eN_A\varepsilon_s}\sqrt{V_s}$$

On calcule alors V_S à partir de ces équations :

$$V_{GB} = V_{FB} + V_s + \gamma\sqrt{V_s}$$

Cette équation du second degré par rapport à $\sqrt{V_s}$ a comme solution :

$$V_s = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}}\right)^2$$

On en déduit V_{DB} puis V_{DS}. On retrouve bien la valeur de $V_{DS\,sat}$.

Ce calcul un peu long montre la cohérence de la représentation. **La valeur de saturation correspond à l'annulation de la charge d'inversion au niveau du drain.** En réalité, le calcul du courant de drain n'est pas valable en régime de saturation puisque le dispositif n'est plus en régime de forte inversion au niveau du drain. On supposera cependant que l'erreur introduite par cette approximation est négligeable ce qui est vérifié dans la pratique.

Quand la tension de drain continue d'augmenter, l'expérience montre que le courant de drain reste constant. En effet, la région au niveau du drain a une charge d'inversion nulle, si bien que la zone canal-drain peut être considérée comme une jonction *pn* polarisée en inverse et donc électriquement équivalente à une résistance de valeur élevée. Toute augmentation de tension appliquée à ce dispositif comprenant en série la partie conductrice du canal et la jonction polarisée en inverse au niveau du drain est donc intégralement appliquée à cette jonction. La tension appliquée aux bornes de la partie conductrice du canal n'augmentant pas, le courant de drain n'augmente pas également. Il ne faudrait

pas penser que la jonction *pn* polarisée en inverse au niveau du drain est un obstacle au passage d'un courant de la source vers le drain. Le champ appliqué a en effet la bonne orientation pour pousser les électrons du canal conducteur vers le drain. Cette zone en bout de canal dans laquelle la charge d'inversion est nulle est appelée région de pincement ou *pinch-off*.

La *figure 4.9* représente les variations du courant de drain en fonction de la tension appliquée entre drain et source pour différentes valeurs de la tension appliquée entre grille et source (1 V, 1,5 V et 2 V). Le transistor représenté a un canal de 0,18 micron de longueur et une largeur *W* de 10 microns.

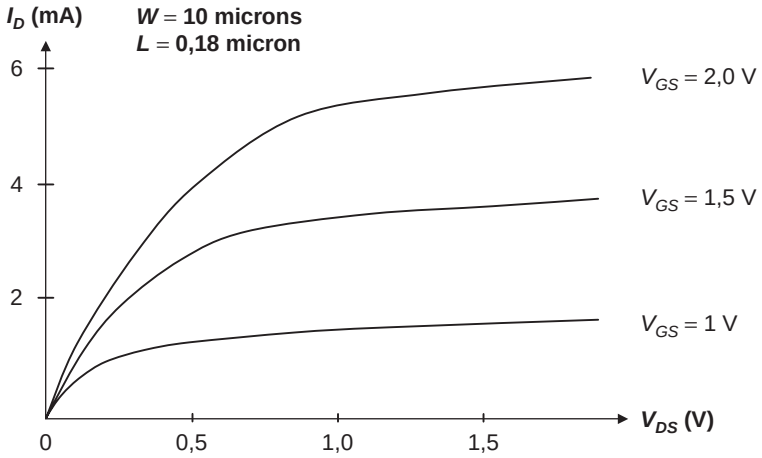


Figure 4.9 – Courbes courant de drain/tension drain source pour un transistor de 10 microns de large dans une technologie 0,18 micron.

Il est maintenant possible de simplifier l'expression 4.2 du courant de drain moyennant quelques approximations supplémentaires. Pour cela on reprend les équations de base rappelées en début de ce chapitre.

$$Q'_G + Q'_0 + Q'_I + Q'_B = 0 \tag{3.19}$$

$$V_{GS} = V_{OX} + V_S + \phi_{MS} \tag{3.9}$$

$$Q'_G = C'_{OX} V_{OX} \tag{3.14}$$

$$Q'_B = -\sqrt{2 e N_A \varepsilon_s}\sqrt{V_s}$$

Le courant s'écrit :

$$I_D L = \mu_n W \int_{V_s(0)}^{V_s(L)} C'_{OX} (V_{GB} - V_{FB} - V_S - \gamma\sqrt{V_s})\, \mathrm{d}V_s$$

Le dernier terme de l'intégrale sera approximé par une fonction linéaire :

$$\gamma\sqrt{V_S} \approx \gamma\sqrt{V_{SB} + 2\phi_F} + \delta \cdot [V_{CB}(x) - V_{SB}]$$

La valeur de δ sera discutée ultérieurement. Le calcul du courant conduit alors au résultat suivant :

$$I_D = \frac{W}{L}\mu_n C'_{OX}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}(1+\delta)V_{DS}^2\right] \tag{4.3}$$

Dans cette relation, le seuil V_T est défini par :

$$V_T = V_{T_0} + \gamma(\sqrt{2\phi_F + V_{SB}} - \sqrt{2\phi_F}) \tag{4.4}$$

avec :

$$V_{T_0} = V_{FB} + 2\phi_F + \sqrt{2\phi_F} \tag{4.5}$$

Cette relation est valable tant que V_{DS} est inférieure à V_{DSsat} c'est-à-dire avant le régime de saturation. La valeur V_{DSmax} est obtenue en annulant la dérivée du courant par rapport à la tension de drain, on obtient alors :

$$V_{DSsat} = \frac{V_{GS} - V_T}{1+\delta}$$

Le courant de saturation est alors :

$$I_{Dsat} = \frac{W}{L}\mu_n C'_{OX}\frac{(V_{GS} - V_T)^2}{2(1+\delta)} \tag{4.6}$$

Cette valeur est évidemment indépendante de V_{DS}.

Le calcul précédent a également mis en relief la dépendance de la tension de seuil V_T avec la polarisation de la source V_{SB}. Cet effet important est appelé effet *body* dans la littérature. Il est possible d'en donner une interprétation physique simple. Quand la source est polarisée positivement par rapport au substrat, elle attire les électrons de la couche d'inversion. Il faut donc compenser cet effet en augmentant le potentiel de grille ce qui est équivalent à une augmentation de la tension de seuil.

Pour terminer cet important paragraphe, il faut indiquer les valeurs de δ les plus utilisées. La valeur la plus immédiate est la dérivée de la fonction pour la valeur $V_{SB} + 2\phi_F$. On obtient alors :

$$\delta = \frac{\gamma}{2\sqrt{2\phi_F + V_{SB}}}$$

Comme cette valeur surestime le calcul, on utilise aussi les valeurs suivantes :

$$\delta = k\frac{\gamma}{2\sqrt{2\phi_F + V_{SB}}} \quad \text{ou} \quad \delta = \frac{\gamma}{2\sqrt{2\phi_F + V_{SB} + \phi_3}}$$

dans lesquelles k et ϕ_3 sont des paramètres d'ajustement.

Pour terminer, il faut également noter les relations approchées très souvent utilisées dans la littérature qui correspondent à δ égal à zéro :

$$I_D = \frac{W}{L}\mu_n C'_{OX}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right] \tag{4.7}$$

$$I_{Dsat} = \frac{W}{L}\mu_n C'_{OX}\frac{(V_{GS} - V_T)^2}{2} \tag{4.8}$$

Ces relations sont très utilisées mais introduisent des erreurs importantes.

4.3.2 *Le régime de faible inversion*

Nous allons maintenant reprendre le calcul du courant pour des tensions de drain faibles. Les deux hypothèses du régime de forte inversion (courant de diffusion négligeable et potentiel de surface ayant atteint une valeur de saturation) seront abandonnées et remplacées par les deux hypothèses suivantes :

- le courant de dérive est négligeable car la charge d'inversion a une valeur très faible dans tout le dispositif ;
- le potentiel de surface est de faible valeur.

À partir de la relation 4.1 donnant le courant et négligeant cette fois le courant de dérive on écrit :

$$I_D L = \mu_n W \phi_t \int_{Q_I'(0)}^{Q_I'(L)} \mathrm{d}Q_i' = \mu_n W \phi_t [Q_I'(L) - Q_I'(0)]$$

Il suffit alors de reprendre l'expression de la charge d'inversion calculée dans le chapitre 3 en régime de faible inversion en prenant bien garde de remplacer dans le terme exponentiel V_S par $V_S - V_{SB}$ au niveau de la source et V_S par $V_S - V_{DB}$ au niveau du drain.

$$Q_I'(0) = -\frac{\gamma C_{OX}'}{2} \frac{\phi_t}{\sqrt{V_s(0)}} \exp \frac{V_s(0) - V_{SB} - 2\phi_F}{\phi_t}$$

$$Q_I'(L) = -\frac{\gamma C_{OX}'}{2} \frac{\phi_t}{\sqrt{V_s(L)}} \exp \frac{V_s(L) - V_{SB} - 2\phi_F}{\phi_t}$$

Le calcul exposé dans le paragraphe précédent montre que quand la charge d'inversion est faible, le potentiel de surface dépend uniquement de la tension de grille et s'exprime par :

$$V_s(0) = V_s(L) = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}} \right)^2$$

Le potentiel de surface étant constant tout le long du canal, le champ de dérive est nul et le courant de dérive également ce qui est cohérent avec l'hypothèse initiale.

On calcule alors le courant de drain :

$$I_D L = \mu_n W \phi_t [Q_I'(L) - Q_I'(0)] = -\mu_n W \phi_t Q_I'(0) \left[1 - \frac{Q_I'(L)}{Q_I'(0)} \right]$$

Comme le potentiel de surface est constant le long du canal, on obtient :

$$I_D = -\frac{W}{L} \mu_n \phi_t Q_I'(0) \left[1 - \exp^{-\frac{V_{DS}}{\phi_t}} \right]$$

On peut alors reprendre la formule simplifiée 3.15 du chapitre 3 pour exprimer $Q_I'(0)$.

$$Q_I'(0) = -\sqrt{2e\varepsilon_s N_A} \frac{1}{2} \frac{\phi_t \mathrm{e}^{\frac{-0,5\phi_F}{\phi_t}}}{\sqrt{1,5\phi_F + V_{SB}}} \exp \frac{V_{GS} - \phi_{MS} + \frac{Q_0}{C_{OX}} - 1,5\phi_F - \gamma\sqrt{1,5\phi_F + V_{SB}}}{\left(1 + \frac{\gamma}{2\sqrt{1,5\phi_F + V_{SB}}}\right)\phi_t}$$

Cette relation un peu compliquée met en évidence la variation du courant avec le seuil défini par le terme V_X égal à la valeur suivante :

$$V_X = \phi_{MS} - \frac{Q'_0}{C'_{OX}} + 1{,}5\,\phi_F + \gamma\sqrt{1{,}5\,\phi_F + V_{SB}}$$

Il est possible d'écrire la relation donnant le courant sous une forme plus lisible.

$$I_D = I_S \exp^{\frac{V_{GS} - V_X}{n_0\,\phi_t}} \left(1 - \exp^{-\frac{V_{DS}}{\phi_t}}\right) \tag{4.9}$$

Dans laquelle,

$$I_S = \frac{W}{L}\mu_n\,\phi_t^2\sqrt{2e\,\varepsilon_s N_A}\,\frac{1}{2}\,\frac{\exp^{\frac{-0{,}5\,\phi_F}{\phi_t}}}{2\sqrt{1{,}5\,\phi_F + V_{SB}}}$$

Rappelons que d'après les résultats du chapitre 3, le coefficient *ideality n* s'exprime par :

$$n = \frac{\mathrm{d}V_G}{\mathrm{d}V_s} = 1 + \frac{\gamma}{2\sqrt{V_S}}$$

$$n = \frac{dV_G}{\mathrm{d}V_s} = 1 + \frac{C'_{si}}{C'_{OX}} \tag{3.26}$$

Le courant d'inversion s'exprime en fonction de la valeur de *n* pour une valeur du potentiel de surface égal à $1{,}5\,\phi_F$.

Il est intéressant de comparer les deux seuils que nous avons définis : le seuil V_T en régime de forte inversion et le seuil V_X en régime de faible inversion.

$$V_X = \phi_{MS} - \frac{Q'_0}{C'_{OX}} + 1{,}5\,\phi_F + \gamma\sqrt{1{,}5\,\phi_F + V_{SB}}$$

$$V_T = \phi_{MS} - \frac{Q'_0}{C'_{OX}} + 2\,\phi_F + \gamma(\sqrt{2\,\phi_F + V_{SB}} - \sqrt{2\,\phi_F})$$

Ces deux valeurs sont différentes ce qui n'est pas toujours pris en compte dans la littérature. On peut constater que ces valeurs sont très proches pour V_{SB} égal à zéro.

La formule 4.9 met en évidence un des phénomènes les plus importants du fonctionnement du MOSFET, l'augmentation de la consommation statique avec la diminution du seuil. Quand le MOSFET canal *n* est bloqué, la tension grille-source est alors nulle. Le courant de drain qui traverse le transistor est faible mais non nul. Il varie selon la formule 4.9 en $\exp(-V_X / n_0 \Phi t)$ ce qui montre bien son augmentation rapide quand la tension de seuil diminue. Au fur et à mesure que les technologies micro-électroniques progressent, les tensions diminuent car les dimensions diminuent. La tension de seuil suit également cette loi ce qui entraîne une augmentation du courant de non conduction du transistor. Cet effet est d'autant plus important que le facteur *n* est élevé. Un objectif de la technologie est donc de réduire ce coefficient lié aux capacités du système comme le montre la formule 3.26 rappelée dans ce paragraphe.

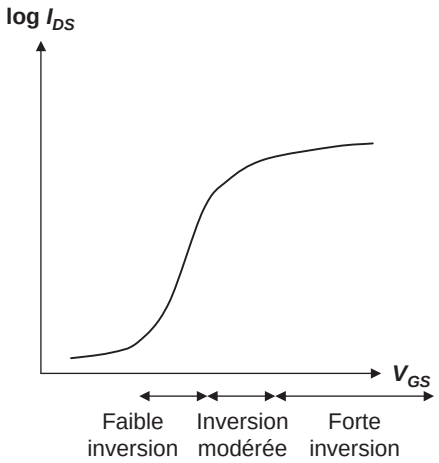


Figure 4.10 – Les trois régimes de fonctionnement du MOSFET.

Pour terminer cet important paragraphe, rappelons les différents régimes de fonctionnement du MOSFET sur la *figure 4.10*. L'échelle logarithmique met en relief le régime de faible inversion.

4.3.3 *La mobilité effective et ses effets*

La mobilité des porteurs dans le canal de conduction est différente de la mobilité dans le semi-conducteur en profondeur. Elle est plus faible car le champ électrique transverse attire les électrons ou les trous vers l'interface avec l'oxyde et favorise des interactions avec les impuretés présentes à cette interface. La prise en compte de ces effets dans le calcul du courant de drain est très complexe car la mobilité en surface est reliée au champ électrique transverse dans le dispositif par une relation de la forme :

$$\mu = \frac{\mu_0}{1 + \alpha E_y}$$

Dans cette relation, μ_0 est une valeur qui dépend de la température. Elle est égale à environ la moitié de la mobilité mesurée dans le volume d'un semi-conducteur. La constante α a une valeur d'environ 0,025 µm/V à la température ambiante. Il est assez difficile de calculer analytiquement le courant de drain avec cette hypothèse. Il est en général admis que les relations données dans le paragraphe précédent restent valables à la condition de remplacer la mobilité μ par une mobilité effective donnée en première approximation par :

$$\mu_{\text{eff}} = \frac{\mu}{1 + \theta(V_{GS} - V_T) + \theta_B V_{SB}}$$

Les paramètres θ et θ_B sont respectivement $0{,}002/d_{\text{OX}}$ et quelques centièmes de V^{-1} avec d_{OX} exprimé en micron.

4.3.4 *Le MOS canal p*

La littérature étudie généreusement le MOS canal *n* et ne fait que mentionner le MOS canal *p* en supposant qu'il est identique au premier à la condition d'inverser le signe des tensions. C'est vrai pour

l'essentiel mais le MOS canal *p* présente des particularités qu'il est bon de connaître et cela d'autant plus qu'il y a autant de MOS canal *p* que de MOS canal *n* dans les circuits intégrés. Reprenons donc le schéma de fonctionnement de base.

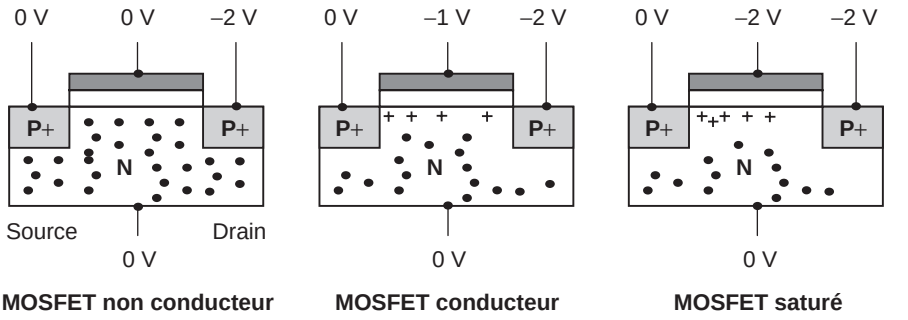


Figure 4.11 – Le fonctionnement du MOS canal p.

L'application sur la grille d'une tension négative attire à l'interface avec l'oxyde les trous provenant de la source fortement dopée. Un champ électrique longitudinal créé entre la source et le drain et convenablement orienté (tension de drain plus négative que celle de la source) permet de créer un courant de dérive dans le dispositif. Les relations 4.3 et 4.4 exprimant le courant et la tension de seuil deviennent :

$$I_D = -\frac{W}{L}\mu_p C'_{OX}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}(1+\delta)V_{DS}^2\right]$$

$$V_T = V_{T_0} - \gamma(\sqrt{-2\phi_F - V_{SB}} - \sqrt{-2\phi_F})$$

Il faut noter que la valeur de la mobilité à prendre en compte est celle des trous. Elle est environ trois fois plus faible que celle des électrons. Les tensions V_{GS}, V_{DS}, ϕ_F sont négatives dans ces formules. À dimensions égales et pour des tensions appliquées égales en valeur absolue, le courant de drain d'un MOS canal *p* est donc environ trois fois plus faible que celui d'un MOS canal *n*. La tension de seuil est également négative.

Donnons quelques précisions sur les signes car c'est souvent une source de confusion. En règle générale, on cherche à éviter les valeurs négatives dans les formules. On exprimera donc le courant de drain mesuré de la source vers le drain et non pas du drain vers la source comme dans le NMOS. Les tensions sont inversées par rapport à celles du NMOS. On exprime V_{SG} au lieu de V_{GS} et V_{SD} au lieu de V_{DS}. La grandeur ϕ_F est négative puisque le substrat est de type *n*.

Avec ces conventions, la formule du PMOS est la même que celle du NMOS.

$$I_D = \frac{W}{L}\mu_p C'_{OX}\left[(V_{GS} - |V_T|)V_{SD} - \frac{1}{2}(1+\delta)V_{SD}^2\right]$$

$$|V_T| = \left|V_{T_0} - \gamma(\sqrt{-2\phi_F - V_{SB}} - \sqrt{-2\phi_F})\right|$$

Il est maintenant possible de comparer NMOS et PMOS en représentant les courbes caractéristiques des deux composants réalisés dans une même technologie.

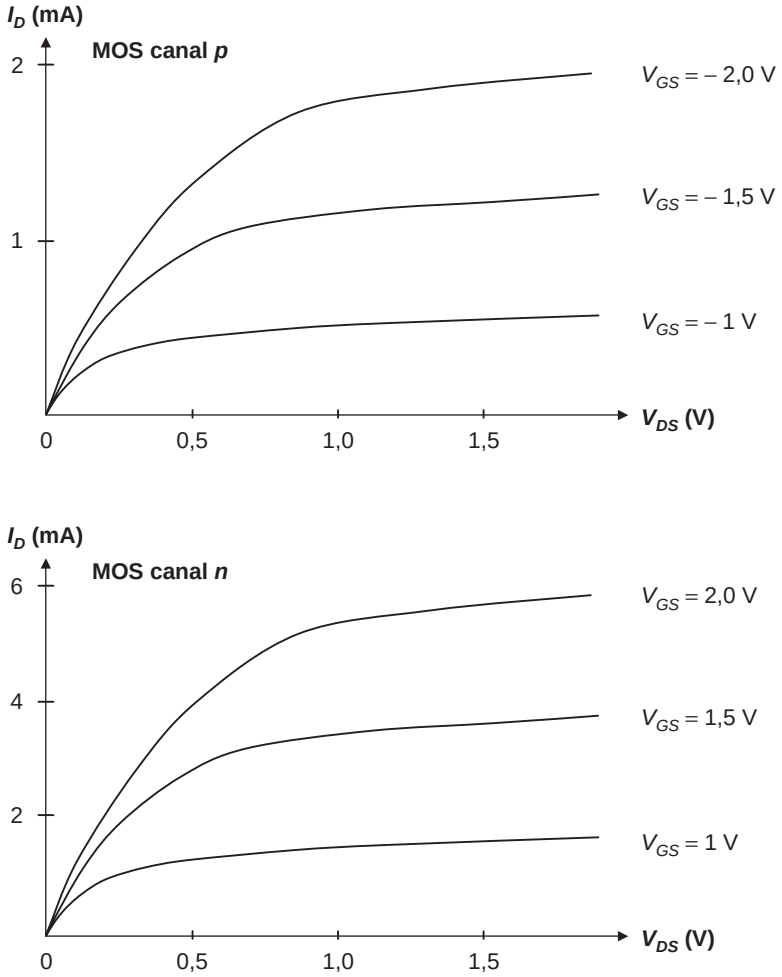


Figure 4.12 – Courant de drain d'un MOS canal p en 0,18 micron comparé au MOS canal n de même dimension dans la même technologie.

4.4 Le modèle canal court

En réalité et pour les technologies actuelles, le transistor MOS ne peut être considéré comme un dispositif de longueur élevée par rapport aux autres dimensions. Les effets de source et de drain ne peuvent être négligés et une véritable description à deux dimensions serait en théorie nécessaire. Pour simplifier, le modèle à une dimension est conservé avec quelques ajouts :

- prise en compte de la zone de charge d'espace au niveau du drain et diminution de la longueur effective du canal ;
- prise en compte de l'effet de saturation de la vitesse des électrons dans le canal ;
- introduction du seuil effectif et de l'effet d'abaissement de ce seuil avec la tension de drain.

4.4.1 *Effet de diminution de la longueur de canal*

L'examen des courbes représentant le courant de drain en fonction de la tension drain-source fait apparaître que le courant n'est pas strictement indépendant de la tension drain-source en régime de saturation mais augmente légèrement quand celle-ci augmente. Cet effet peut s'expliquer par la modulation de la longueur du canal avec la tension de drain.

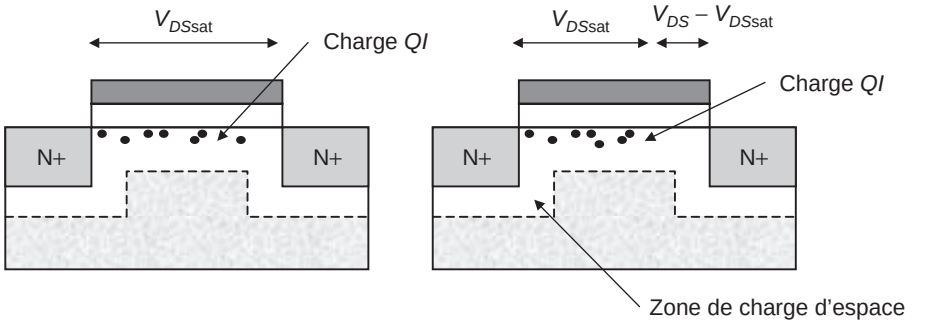


Figure 4.13 – Modulation de la longueur de canal.

Pour la valeur $V_{DS\,sat}$ de la tension de drain, le transistor entre en régime de saturation et la densité d'électrons au niveau du drain devient nulle. Toute augmentation de tension supplémentaire se traduit par l'apparition d'une zone de d'espace dans la jonction *pn* qui se forme au niveau du drain comme le montre la *figure 4.13*. La profondeur de la zone de charge d'espace augmente quand la tension de drain augmente et la longueur du canal de conduction (région dans laquelle on trouve des électrons) diminue de ΔL, ce qui a pour effet d'augmenter le courant de drain qui varie inversement proportionnellement à la longueur du canal.

La prise en compte de ces effets conduit à écrire de manière approximative la valeur du courant de drain en régime de saturation sous la forme :

$$I_D = I_{Dsat}\left(1 + \frac{V_{DS} - V_{DSsat}}{V_A}\right) \tag{4.11}$$

avec,

$$V_A = kL\sqrt{N_A}$$

Le paramètre k est une constante égale à environ $0{,}15\ \text{V}\,\mu\text{m}^{1/2}$, et N_A est le dopage du substrat.

4.4.2 *Effet de saturation de la vitesse*

Dans le calcul précédent, la vitesse des électrons était supposée proportionnelle au champ électrique longitudinal (parallèle à la direction du canal), la constante de proportionnalité étant la mobilité. En fait, à partir d'une certaine valeur du champ E_C cette loi n'est plus vérifiée et la vitesse tend vers une valeur limite indépendante du champ et notée v_{max}.

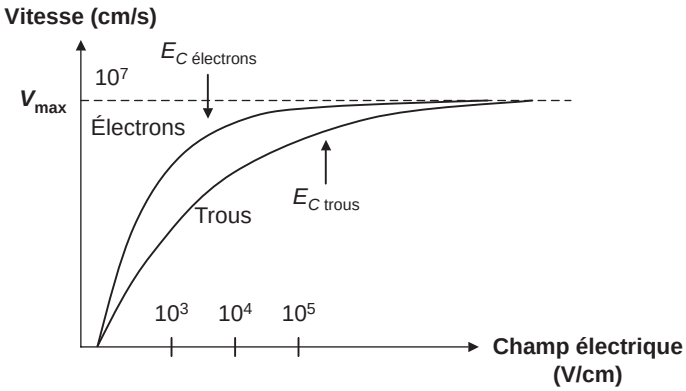


Figure 4.14 – Vitesse de saturation

La relation entre vitesse et champ peut alors s'écrire :

$$v(x) = v_{\max}\frac{\dfrac{1}{E_c}\dfrac{\mathrm{d}V_s}{\mathrm{d}x}}{1+\dfrac{1}{E_c}\dfrac{\mathrm{d}V_s}{\mathrm{d}x}}$$

La relation de base exprimant le courant était :

$$I_D = \mu_n W(-Q_I')\frac{\mathrm{d}V_s}{\mathrm{d}x} + \mu_n W \phi_t \frac{\mathrm{d}Q_I'}{\mathrm{d}x}$$

soit en négligeant le courant de diffusion :

$$I_D = \mu_n W(-Q_I')\frac{\mathrm{d}V_s}{\mathrm{d}x}$$

Intégrons dans cette expression la formule donnant la vitesse :

$$I_D = W(-Q_I')\, v_{\max}\frac{\dfrac{1}{E_c}\dfrac{\mathrm{d}V_s}{\mathrm{d}x}}{1+\dfrac{1}{E_c}\dfrac{\mathrm{d}V_s}{\mathrm{d}x}}$$

On obtient donc :

$$I_D\left(1+\frac{1}{E_c}\frac{\mathrm{d}V_s}{\mathrm{d}x}\right) = W(-Q_I')\, v_{\max}\frac{1}{E_c}\frac{\mathrm{d}V_s}{\mathrm{d}x}$$

En intégrant cette équation de $x = 0$ à $x = L$, on obtient alors :

$$I_D\left(L+\frac{V_{DB}-V_{SB}}{E_c}\right) = W\mu\int_{V_{SB}}^{V_{DS}}(-Q_I')\,\mathrm{d}V_s$$

$$I_D = \frac{W}{L}\frac{\mu}{1+\dfrac{V_{DS}}{LE_c}}\int_{V_{SB}}^{V_{DS}}(-Q'_I)\,\mathrm{d}V_s$$

Si on note I_{D0} le courant calculé en ne tenant pas compte de l'effet de saturation de la vitesse, on peut écrire :

$$I_D = \frac{1}{1+\dfrac{V_{DS}}{LE_c}}I_{D0}$$

Si on exprime dans cette formule le courant calculé sans effet de saturation par sa valeur 4.3, on obtient :

$$I_D = \frac{W}{L}\left(\frac{1}{1+\dfrac{V_{DS}}{LE_c}}\right)\mu_n C'_{OX}\left[(V_{GS}-V_T)V_{DS}-\frac{1}{2}(1+\delta)V_{DS}^2\right] \tag{4.12}$$

Cette relation est valable en régime non saturé, puis au-delà d'une tension $V_{DS\,sat}$ le courant reste constant quand la tension de drain augmente. Cette valeur $V_{DS\,sat}$ est obtenue quand la dérivée du courant par rapport à la tension V_{DS} est nulle. On obtient facilement :

$$V_{DSsat} = LE_c\left[\sqrt{1+\frac{2(V_{GS}-V_T)}{(1+\delta)LE_c}}-1\right] \tag{4.13}$$

Cette relation montre que le régime de saturation est atteint pour des valeurs plus faibles de la tension de drain. En effet, la valeur de la tension de saturation sans effet de saturation de la vitesse est :

$$V_{DSsat} = \frac{V_{GS}-V_T}{1+\delta}$$

Il ne faut pas confondre la saturation de la vitesse et le régime de saturation du transistor. Les deux valeurs de la tension de saturation sont égales pour un champ critique infini. En pratique, le champ critique est de l'ordre de 10^4 V/cm. Le produit LE_C est donc de 0,1 V pour une longueur de 100 nm et de 0,8 V pour une longueur de 0,8 micron. La tension de saturation est donc environ de 50 mV pour une longueur de 100 nm.

Que deviennent ces expressions quand le canal est très petit ce qui est le cas des technologies modernes ? Il est maintenant possible de calculer la valeur du courant de saturation à partir des relations 4.12 et 4.13 en faisant l'hypothèse que le terme V_{DS}/LE_C est très supérieur à 1.

$$I_{Dsat} = WE_c\mu_n C'_{OX}\left[(V_{GS}-V_T)-\frac{(1+\delta)}{2}V_{DSsat}\right] \tag{4.14}$$

La tension de saturation est donnée par la formule 4.13. Cette expression est parfois simplifiée en négligeant la tension de saturation.

$$I_{Dsat} = W\mu_n C'_{OX}[V_{GS}-V_T]E_c$$

La valeur du champ critique E_C est de 1 V/µm à 3 V/µm pour les électrons et de 2 V/µm à 10 V/µm pour les trous. On met ainsi en évidence la variation linéaire du courant avec la tension de grille. Rappelons que dans le cas d'un canal long, la dépendance était quadratique. Le courant est donc dans le cas du canal court indépendant de la longueur du canal.

Pour donner un sens physique à ce résultat qui peut sembler paradoxal, il suffit de considérer la dépendance de la charge et du temps de transit en fonction de la longueur du canal. Le temps de transit à vitesse constante est proportionnel à la longueur du canal. La charge stockée dans le canal est également proportionnelle à la longueur du canal. Le courant qui est le rapport des deux est donc indépendant de cette longueur. Notons également que les vitesses de saturation des électrons et des trous sont voisines (environ 100 microns par ns). Les performances en vitesse des NMOS et des PMOS sont donc voisines en régime de canal court. Le tableau suivant illustre les deux régimes de fonctionnement :

Courant de drain pour un canal long en régime de saturation	Courant de drain pour un canal court en régime de saturation
$I_{Dsat} = \frac{W}{L}\mu_n C'_{OX}\frac{(V_{GS}-V_T)^2}{2(1+\delta)}$	$I_{Dsat} = W\mu_n C'_{OX}\left[V_{GS}-V_T-\frac{1+\delta}{2}V_{DSsat}\right]E_c$

Il est possible d'illustrer ces conclusions sur deux courbes représentant les variations du courant de drain en fonction de la tension de grille, la première pour un canal de 20 microns et la seconde pour un canal de 45 nm. Pour un transistor à canal court le courant de drain varie proportionnellement à la tension de grille.

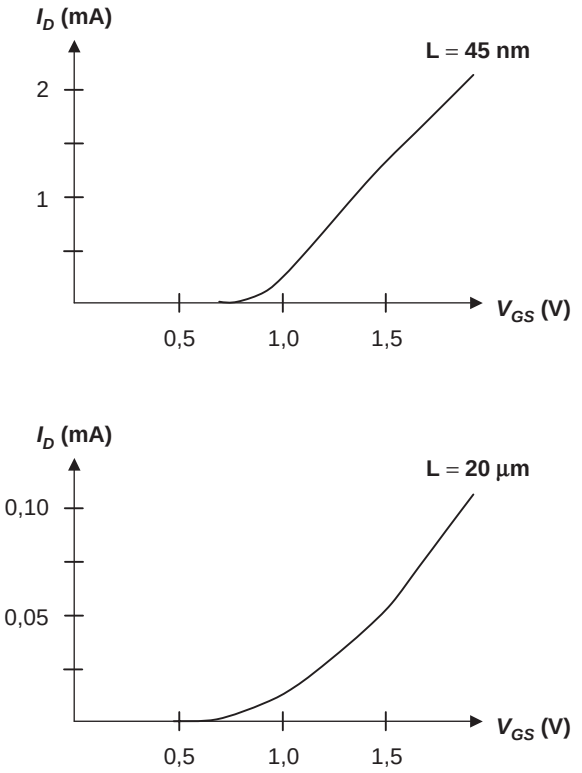


Figure 4.15 – Variation du courant de drain pour L = 45 nm et pour L = 20 microns.

4.4.3 *Effet de diminution du seuil effectif*

Les effets de canal court se manifestent également par une déformation des lignes de champ dans la zone de charge d'espace du dispositif comme le montre la *figure 4.16*.

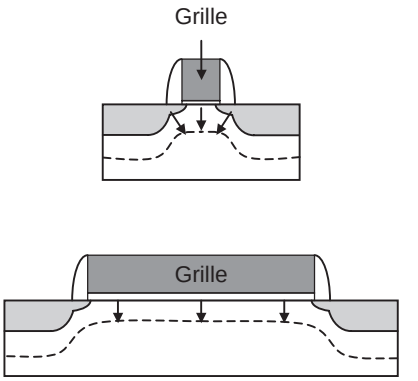


Figure 4.16 – Modification de la zone de charge d'espace.

Quand le dispositif devient plus court, pour une même tension de grille la zone de charge d'espace s'étend en profondeur (effets de bord dus aux zones de diffusion de la source et du drain). La charge Q_B de la zone de charge d'espace est alors plus importante que si le dispositif était idéal. Le potentiel de surface augmente donc et en conséquence le courant de drain. À tension de grille égale, le courant de drain est plus important dans un dispositif à canal court. Cet effet peut également se représenter par une diminution de la tension de seuil. Il est important puisqu'il peut atteindre plusieurs dixièmes de volts. Des calculs approchés et basés sur des considérations purement géométriques de la déformation de la zone de charge d'espace conduisent à l'expression suivante de la diminution de la tension de seuil.

$$\Delta V_T = 2\beta \frac{\varepsilon_s}{\varepsilon_{OX}} \frac{d_{OX}}{L} (2\phi_F + V_{SB})$$

Dans cette expression on notera l'effet des permittivités du silicium et de l'oxyde de grille (ε_s et ε_{OX}) ainsi que l'effet de l'épaisseur de l'oxyde (d_{OX}). Le coefficient β est peu différent de 1.

Dans les considérations précédentes, la tension entre drain et source n'a pas été prise en compte. Elle a pourtant un rôle important exprimé dans le paramètre DIBL (*Drain Induced Barrier Lowering*). Il exprime la diminution de la tension de seuil en fonction de la tension de drain. Quand la tension de drain augmente, les électrons sous l'oxyde de grille sont attirés par le drain ce qui augmente le courant dans le canal. On peut exprimer cette augmentation par une diminution de la tension de seuil. Cet effet est d'autant plus important que le canal est court car l'influence électrostatique du potentiel de drain dépend de la distance. Ce terme qui conduit à faire varier la tension de seuil doit être le plus faible possible car les variations de tension de seuil peuvent conduire à dégrader les performances des circuits logiques réalisés avec des MOSFET. La relation précédente devient donc :

$$\Delta V_T = 2\beta \frac{\varepsilon_s}{\varepsilon_{OX}} \frac{d_{OX}}{L} [(2\phi_F + V_{SB}) + \chi V_{DS}] \qquad (4.15)$$

Le coefficient χ est proportionnel à l'inverse de la longueur du canal. La *figure 4.17* illustre comment la tension de seuil varie avec la tension de drain dans diverses technologies.

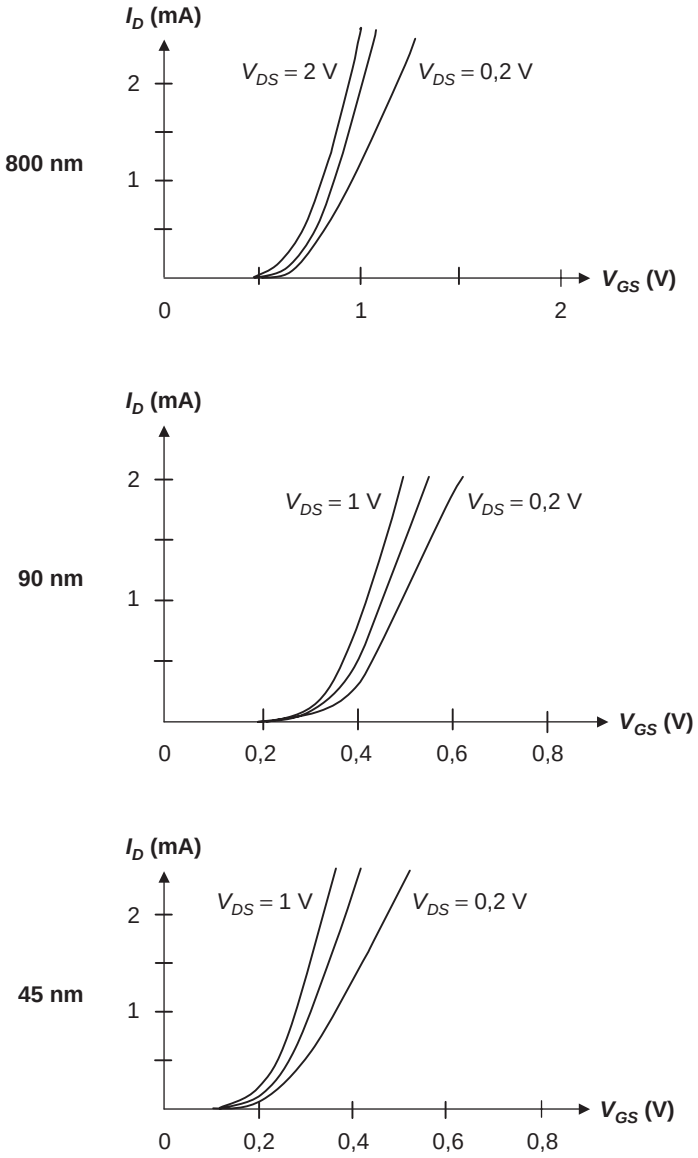


Figure 4.17 – Diminution de la tension de seuil avec la tension de drain.

4.5 Le fonctionnement dynamique du MOS

4.5.1 *Régime quasi-statique*

Comme il a été expliqué dans l'introduction, la première étape est de décrire le régime quasi-statique du transistor MOS. L'hypothèse principale est de supposer le courant de conduction constant dans le canal. Les expressions précédentes obtenues pour le calcul des charges seront donc utilisées, à la différence près qu'elles sont maintenant des fonctions du temps. Les raisonnements seront effectués en considérant la partie active du transistor, c'est-à-dire la zone de canal. Source, drain, grille et les paramètres électriques associés ne seront pas pris en compte dans cette analyse.

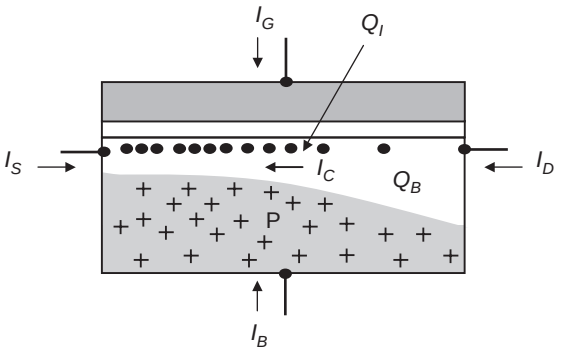


Figure 4.18 – Schéma de base pour l'analyse quasi-statique.

On écrit les équations de conservation du courant en faisant la différence entre le courant de conduction I_C (diffusion et dérive des charges) et les courants transitoires ou de déplacement I_{DV} et I_{SV} créés par les variations de charges dans le dispositif. La définition précise de ces charges n'est pas évidente et sera expliquée ultérieurement.

$$I_D(t) = I_C(t) + I_{DV}(t)$$
$$I_S(t) = -I_C(t) + I_{SV}(t)$$

Le régime quasi-statique suppose que la différence de courant est due à la variation de charge du canal. Les autres charges n'ont pas le temps de changer de manière significative. On écrit alors :

$$I_{DV}(t) + I_{SV}(t) = \frac{\mathrm{d}Q_I}{\mathrm{d}r}$$

Il nous faut maintenant calculer ces deux composantes du courant. Pour faire ce calcul de manière rigoureuse, oublions les hypothèses du régime quasi-statique et considérons le problème de manière générale en prenant une partie du canal de conduction de longueur dx et de largeur W, comme le montre la *figure 4.19*. On néglige les courants ayant des directions différentes de celle du canal comme cela a été expliqué précédemment.

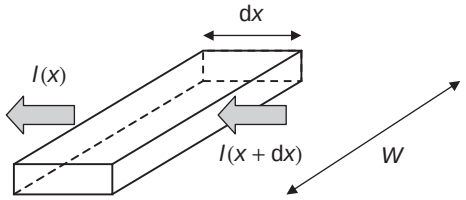


Figure 4.19 – Équation de continuité.

Si on applique le principe de conservation de la charge, on peut écrire en fonction de la charge par unité de surface dans le canal :

$$I(x + \mathrm{d}x) - I(x) = \frac{\partial}{\partial t}(Q'_I W \mathrm{d}x)$$

On en déduit donc,

$$\frac{\partial I(x,t)}{\partial x} = W \frac{\partial Q'_I(x,t)}{\partial t} \tag{4.16}$$

On peut également écrire dans le canal en négligeant le courant de diffusion.

$$I(x,t) = -\mu_n W Q'_I(x,t) \frac{\partial V_{CB}(x,t)}{\partial x}$$

Rappelons que le potentiel $V_{CB}(x,t)$ est la partie variable du potentiel de surface dans le canal.

$$V_s(x,t) = \phi_B + V_{CB}(x,t)$$

La résolution de l'équation (4.16) permet d'écrire :

$$I(x,t) - I(0,t) = W \int_0^x \frac{\partial Q'_I(x',t)}{\partial t} \mathrm{d}x'$$

Le courant $I(0,t)$ n'est autre que le courant $-I_S(t)$. La relation précédente peut donc s'écrire :

$$I_S(t) = W \int_0^x \frac{\partial Q'_I(x',t)}{\partial t} \mathrm{d}x' + \mu_n W Q'_I(x,t) \frac{\partial V_{CB}(x,t)}{\partial x}$$

On intègre alors cette équation de $x = 0$ à $x = L$ et on obtient :

$$I_S(t)L = W \int_0^L \int_0^x \frac{\partial Q'_I(x',t)}{\partial t} \mathrm{d}x' \mathrm{d}x + \mu_n W \int_0^L Q'_I(x,t) \frac{\partial V_{CB}(x,t)}{\partial x} \mathrm{d}x$$

Cette équation s'écrit aussi :

$$I_S(t)L = W \frac{\partial}{\partial t} \int_0^L \int_0^x Q'_I(x,t) \, \mathrm{d}x' \mathrm{d}x + \mu_n W \int_0^L Q'_I(x,t) \frac{\partial V_{CB}(x,t)}{\partial x} \mathrm{d}x$$

En intégrant par parties la première expression de la forme $\int_0^L F(x)\mathrm{d}x$, dans laquelle la fonction $F(x)$ est définie par $\int_0^x Q'_I \, \mathrm{d}x'$, on obtient :

$$I_S(t)L = W\frac{\partial}{\partial t}\int_0^L (L-x)Q_I'(x,t)\,\mathrm{d}x + \mu_n W\int_0^L Q_I'(x,t)\frac{\partial V_{CB}(x,t)}{\partial x}\,\mathrm{d}x$$

soit,

$$I_S(t) = W\frac{\partial}{\partial t}\int_0^L\left(1-\frac{x}{L}\right)Q_I'(x,t)\,\mathrm{d}x + \mu_n\frac{W}{L}\int_0^L Q_I'(x,t)\frac{\partial V_{CB}(x,t)}{\partial x}\,\mathrm{d}x \qquad (4.17)$$

Cette équation peut se comparer à l'équation établie en début de ce paragraphe pour le courant au niveau de la source. On obtient donc pour le courant au niveau de la source :

$$I_S(t) = -I_C(t) + I_{SV}(t)$$

avec,

$$I_{SV}(t) = W\frac{\partial}{\partial t}\int_0^L\left(1-\frac{x}{L}\right)Q_I'(x,t)\,\mathrm{d}x \qquad (4.18)$$

Ce courant est donc la partie additionnelle au courant de conduction vue au niveau de la source. En régime quasi-statique, la formule de la charge par unité de surface est supposée être la même qu'en régime purement statique. Elle est donc donnée par les relations établies dans les paragraphes précédents. En régime statique, ce terme additionnel est nul.

De la même manière, on peut calculer le courant au niveau du drain, il s'écrit :

$$I_D(t) = I_C(t) + I_{DV}(t)$$

avec,

$$I_{DV}(t) = W\frac{\partial}{\partial t}\int_0^L\frac{x}{L}Q_I'(x,t)\,\mathrm{d}x \qquad (4.19)$$

Il reste à exprimer la charge d'inversion en fonction des tensions appliquées pour pouvoir écrire les courants transitoires calculés précédemment. En régime de forte inversion et de canal long, on obtient donc :

$$Q_I' = \mu_n C_{OX}'(V_{GS} - V_{FB} - V_S - \gamma\sqrt{V_S})$$

Rappelons que V_{FB} est défini par :

$$V_{FB} = \phi_{MS} - \frac{Q_0'}{C_{OX}'}$$

On reprend l'approximation classique :

$$\gamma\sqrt{V_S} \approx \gamma\sqrt{V_{SB} + 2\phi_F} + \delta(V_{CB}(x) - V_{SB})$$

La valeur de la tension de saturation est alors :

$$V_{DSsat} = \frac{V_{GS} - V_T}{1+\delta}$$

Tous ces résultats ont été établis dans le paragraphe 4.3.1 et seuls les résultats utiles sont repris.

Si maintenant on introduit le paramètre α qui exprime la position de la tension de drain par rapport à la tension de saturation, on écrit :

$$\alpha = 1 - \frac{V_{DS}}{V_{DS\text{sat}}} \tag{4.20}$$

Il est égal à 0 au début du régime saturé et égal à 1 quand le point de fonctionnement est éloigné du régime de saturation.

Il est alors possible de calculer la charge d'inversion par unité de surface Q'_I puis la charge d'inversion totale Q_I en fonction du paramètre α. La charge d'inversion totale est donnée par la relation suivante :

$$Q'_I = \int_0^L Q'_I(x) W \mathrm{d}x$$

Comme la dépendance de la charge d'inversion en fonction de la position n'est pas connue, il est préférable d'exprimer les charges en fonction de la variation de potentiel de surface. Pour cela on utilise la relation donnant le courant de conduction en supposant qu'il est exclusivement créé par la dérive des électrons en régime de forte inversion.

$$I_D \mathrm{d}x = -\mu_n W Q'_I \mathrm{d}V_{CB}(x)$$

On écrit alors :

$$Q_I = \frac{\mu_n W^2}{I_D} \int_{V_{SB}}^{V_{DB}} Q'^2_I \mathrm{d}V_{CB}$$

Le calcul est un peu long et n'est pas détaillé dans cet ouvrage :

$$Q_I = -WLC'_{\text{OX}}(V_{GS} - V_T)\frac{2}{3}\frac{1 + \alpha + \alpha^2}{1 + \alpha} \tag{4.21}$$

On peut calculer de la même manière les autres charges du dispositif.

$$Q_B = -WLC'_{\text{OX}}\left[\gamma\sqrt{\phi_B + V_{SB}} + \frac{\delta}{1+\delta}(V_{GS} - V_T)\left(1 - \frac{2}{3}\frac{1 + \alpha + \alpha^2}{1 + \alpha}\right)\right] \tag{4.22}$$

On en déduit la charge Q_G en fonction de la charge Q_0 de l'interface.

On peut également donner les expressions des courants de transition I_{SV} et I_{DV}.

$$I_{SV}(t) = W\frac{\partial}{\partial t}\int_0^L \left(1 - \frac{x}{L}\right) Q'_I(x,t)\mathrm{d}x$$

On obtient après des calculs assez longs :

$$I_{SV}(t) = -\frac{\partial}{\partial t} WLC'_{\text{OX}}(V_{GS} - V_T)\frac{6 + 12\alpha + 8\alpha^2 + 4\alpha^3}{15(1+\alpha)^2}$$

$$I_{DV}(t) = -\frac{\partial}{\partial t} WLC'_{\text{OX}}(V_{GS} - V_T)\frac{4 + 8\alpha + 12\alpha^2 + 6\alpha^3}{15(1+\alpha)^2}$$

Résumons les résultats de ce paragraphe. Tout se passe comme si les courants de transition étaient dus à des charges « virtuelles », Q_S et Q_D variant dans le temps. Ces charges sont respectivement :

$$Q_S(t) = W\int_0^L \left(1 - \frac{x}{L}\right) Q'_I(x,t)\mathrm{d}x \tag{4.23}$$

$$Q_D(t) = W\int_0^L \frac{x}{L} Q'_I(x,t)\,\mathrm{d}x \tag{4.24}$$

Le lecteur est sans doute convaincu de la complexité des calculs analytiques pour obtenir les expressions des courants en fonction du temps. En pratique, on se base sur les résultats donnés par les simulateurs électriques. Il est cependant utile de pouvoir fixer quelques ordres de grandeur et surtout d'être conscient des hypothèses de calcul.

En régime de faible inversion, les calculs sont plus simples puisque le potentiel de surface est constant et indépendant de la position.

$$V_s(0) = V_s(L) = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}}\right)^2$$

Les résultats du paragraphe 4.3.2 permettent de calculer la charge d'inversion par unité de surface. On constate qu'elle varie quasi-linéairement entre la source et le drain. Les charges virtuelles de source et de drain se calculent facilement :

$$Q_S = WL\left(\frac{Q'_{I\text{source}}}{3} + \frac{Q'_{I\text{drain}}}{6}\right) \qquad Q_D = WL\left(\frac{Q'_{I\text{source}}}{6} + \frac{Q'_{I\text{drain}}}{3}\right)$$

En pratique, ces charges sont négligeables en régime de faible inversion devant les charges extérieures au canal et elles ne seront pas prises en compte dans la modélisation dynamique.

4.5.2 *Régime dynamique*

Quelles sont les conditions de validité du régime quasi-statique ? Il faut que les distributions de charges s'établissent plus vite que les variations de tension. En pratique, on donne souvent la condition suivante : le temps de montée de la tension d'entrée (en général, la tension de grille) doit être au moins vingt fois plus important que le temps de transit des charges dans le canal. Le temps de transit τ est défini comme suit :

$$\tau = \frac{|Q_I|}{I_D}$$

En régime de forte inversion et en régime saturé, on trouve alors :

$$Q_I = -WLC'_{\mathrm{OX}}(V_{GS} - V_T)\frac{2}{3}$$

$$I_{D\mathrm{sat}} = \frac{W}{L}\mu_n C'_{\mathrm{OX}}\frac{(V_{GS} - V_T)^2}{2(1+\delta)}$$

On en déduit :

$$\tau = (1+\delta)\frac{L^2}{\mu_n(V_{GS} - V_T)} \cdot \frac{4}{3} \tag{4.25}$$

Quelques ordres de grandeur peuvent être donnés.

Pour une technologie ancienne :

$$L = 0{,}8\,\mu\mathrm{m}, \quad V_{GS} - V_T = 3\,\mathrm{V}, \quad \mu_n = 64\,\mu\mathrm{m}^2/\mathrm{V}\cdot\mathrm{ns} \quad \text{alors} \quad \tau = 3\,\mathrm{ps}$$

Pour une technologie moderne :

$$L = 0{,}05\,\mu\text{m}\,,\quad V_{GS} - V_T = 1\,\text{V}\,,\quad \mu_n = 64\,\mu\text{m}^2/\text{V}\cdot\text{ns}\quad \text{alors}\quad \tau = 0{,}04\,\text{ps}$$

Ces ordres de grandeur montrent que le domaine d'application du régime quasi-statique est très large. Revenons cependant à la manière la plus générale de traiter le comportement dynamique d'un MOSFET en reprenant les équations de base :

$$\frac{\partial I(x,t)}{\partial x} = W\frac{\partial Q_I'(x,t)}{\partial t}$$

$$I(x,t) = -\mu_n W Q_I'(x,t)\frac{\partial V_{CB}(x,t)}{\partial x}$$

$$Q_I' = \mu_n C_{OX}'(V_{GB} - V_{FB} - V_S - \gamma\sqrt{V_s})$$

$$V_s = \phi_B + V_{CB}(x,t)$$

Ce système de quatre équations permet en théorie de trouver les quatre inconnues du problème : $I(x,t)$, $Q'(x,t)$, $V_{CB}(x,t)$ et $V_s(x,t)$. Il est nécessaire pour résoudre ces équations de disposer d'un jeu de conditions initiales. Prenons l'exemple d'un transistor MOS initialement bloqué auquel on applique brusquement une tension sur la grille.

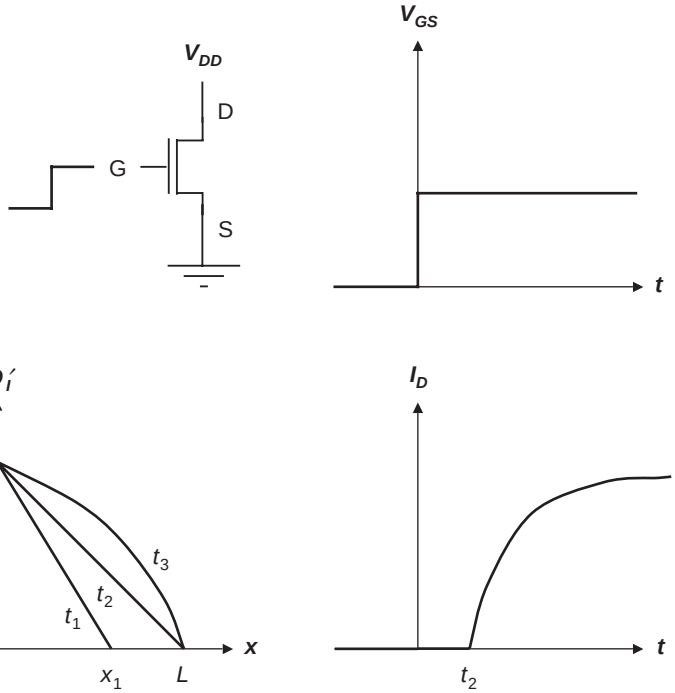


Figure 4.20 – Résolution des équations dynamique.

Dans un premier temps, le canal est vide. La tension est appliquée sur la grille et des électrons sont injectés de la source vers le canal. Au bout du temps t_1, les électrons ont atteint la distance x_1 puis au temps t_2 la distance L correspondant à la longueur du canal. Ensuite l'équilibre s'établit pour atteindre la répartition finale à partir du temps t_3. On comprend facilement que le temps t_2 est le retard dans l'apparition du courant et que le temps t_3 correspond au temps de montée de l'impulsion de courant.

Une étude plus détaillée des équations de base pourrait montrer que la grandeur $V_{CB}(x, t)$ est en fait la différence entre le pseudo potentiel chimique des électrons dans le canal et le pseudo potentiel chimique des trous en profondeur dans le matériau.

4.6 Les modèles du transistor MOSFET

Ce paragraphe est une sorte de résumé des paragraphes précédents. Il donne en plus les bases pour constituer un ensemble d'équations utilisables par un logiciel de simulation. Cet ensemble d'équations est un modèle. Les modèles sont définis en fonction de leur domaine d'application. On distingue classiquement les trois domaines suivants :

- le domaine statique : les courants sont exprimés en fonction des tensions appliquées ;
- le domaine dynamique dit petits signaux : ce ne sont plus les courants qui sont exprimés mais les variations de ces courants pour des variations faibles des tensions. Les courants considérés sont les courants qui entrent dans le dispositif : courant de source, courant de grille, courant de drain et courant de bulk ;
- le domaine dynamique dit petits signaux et haute fréquence. Les effets supplémentaires à prendre en compte sont d'une part les éléments parasites négligés dans les représentations précédentes (capacités et self-inductances) et d'autre part les effets liés à la représentation spatiale des phénomènes électriques. Les dispositifs sont alors représentés comme des ensembles comportant des éléments répartis dans l'espace et la dimension spatiale compte. Notons que la modélisation du comportement du canal de conduction s'inscrit dans ce mode de représentation.

4.6.1 *Rappels du modèle statique*

Le *tableau 4.1* illustre les différents cas possibles traités dans les paragraphes précédents.

Tableau 4.1

	Faible inversion	Forte inversion, non saturé	Forte inversion, saturé
Canal long	A	B	C
Canal court	D	E	F
Haute fréquence	G	H	I

Reprenons dans chaque cas l'expression du courant de drain en fonction des tensions appliquées.

Cas A : Le courant de drain s'écrit comme suit.

$$I_D = -\frac{W}{L}\mu_n \phi_T Q_I'(0)\left(1 - e^{-\frac{V_{DS}}{\phi_t}}\right)$$

$$I_D = I_S \exp^{\frac{V_{GS}-V_T}{n_0\phi_t}}\left(1 - \exp^{-\frac{V_{DS}}{\phi_t}}\right)$$

Cas B : Le courant de drain s'écrit donc.

$$I_D = \frac{W}{L}\mu_n C_{OX}'\left[(V_{GS}-V_T)V_{DS} - \frac{1}{2}(1+\delta)V_{DS}^2\right]$$

$$\delta = \frac{\gamma}{2\sqrt{2\,\phi_F + V_{SB}}}$$

Cas C : Le courant de drain s'écrit.

$$I_{D\text{sat}} = \frac{W}{L}\mu_n C_{OX}' \frac{(V_{GS}-V_T)^2}{2(1+\delta)}$$

Cas D : Le courant de drain est alors.

$$I_D = -\frac{W}{L}\mu_n \phi_t Q_I'(0)\left(1 - \exp^{-\frac{V_{DS}}{\phi_t}}\right)$$

$$I_D = I_S \exp^{\frac{V_{GS}-V_T}{n_0\phi_t}}\left(1 - \exp^{-\frac{V_{DS}}{\phi_t}}\right)$$

Cas E : Le courant de drain s'écrit comme suit.

$$I_D = \frac{W}{L}\left(\frac{1}{1+\frac{V_{DS}}{LE_C}}\right)\mu_n C_{OX}'\left[(V_{GS}-V_T)V_{DS} - \frac{1}{2}(1+\delta)V_{DS}^2\right]$$

Cas F : L'expression du courant de drain se simplifie.

$$I_{D\text{sat}} = W\mu_n C_{OX}'\left[V_{GS} - V_T - \frac{1+\delta}{2}V_{DS\text{sat}}\right]E_c$$

Les autres cas, cas de la haute fréquence, ne donnent pas de valeurs différentes du courant.

4.6.2 *Modèles petits signaux : généralités*

On s'intéresse aux variations des grandeurs électriques autour d'un point de polarisation donné. La méthode est adaptée à la description des circuits analogiques qui manipulent en général des signaux de faibles amplitudes. Elle est par extension également utilisée en numérique bien que les variations soient importantes. Le principe est simple : il suffit de dériver le courant par rapport aux

tensions de commande pour avoir des générateurs de courant équivalents. Prenons l'exemple du courant de drain dans le mode non saturé.

$$I_D = \frac{W}{L}\mu_n C'_{OX}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}(1+\delta)V_{DS}^2\right]$$

Si on superpose à la tension V_{GS} une variation de faible amplitude v_{gs}, on obtient :

$$I_D + i_d = \frac{W}{L}\mu_n C'_{OX}\left[(V_{GS} + v_{gs} - V_T)V_{DS} - \frac{1}{2}(1+\delta)V_{DS}^2\right]$$

La valeur de i_d s'obtient par dérivation :

$$i_d = \left(\frac{\partial I_D}{\partial V_{GS}}\right)_{\substack{V_{GS} = V_{GS0} \\ V_{DS} = V_{DS0}}} \cdot v_{gs}$$

Ces données ne suffisent pas pour obtenir un schéma électrique. Il faut également tenir compte des courants de transition que nous avons introduits dans le modèle quasi-statique du transistor. Enfin, il faut tenir compte des éléments parasites (résistances, condensateurs et inductances). L'établissement d'un modèle est donc une opération assez complexe et particulièrement si on impose de passer d'un régime de fonctionnement à un autre sans discontinuité.

Nous procéderons de manière itérative, tout d'abord en considérant la partie centrale du transistor à basse fréquence. Ensuite, nous étendrons cette description aux fréquences intermédiaires, ensuite nous introduirons les éléments parasites et finalement nous donnerons quelques éléments sur la modélisation à haute fréquence.

4.6.3 *Le MOS interne en basse fréquence*

Les équations conduisent naturellement au schéma suivant. Les courants de transition sont supposés nuls.

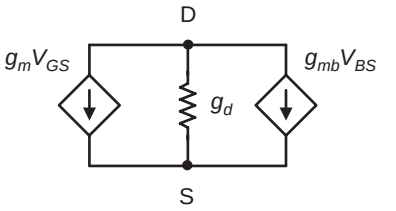


Figure 4.21 – Le modèle basse fréquence interne.

Les coefficients du modèle sont obtenus par simple dérivation :

$$g_m = \left(\frac{\partial I_D}{\partial V_{GS}}\right)_{V_{GS},\, V_{DS}}$$

La notation indique que les autres différences de potentiel (V_{DS} et V_{BS}) restent constantes. On obtient alors en fonction des régimes de fonctionnement :

Cas A et D : faible inversion et canal long ou court.

$$g_m = \frac{1}{n}\frac{I_D}{\phi_t} \tag{4.26}$$

On obtient de même pour le paramètre g_{mb}.

$$g_{mb} = \left(\frac{\partial I_{DS}}{\partial V_{BS}}\right)_{V_{GS},\, V_{DS}}$$

On obtient après quelques calculs :

$$g_{mb} \approx (n-1)g_m \approx \frac{\varepsilon_s}{\varepsilon_{OX}}\frac{d_{OX}}{y_B}g_m \tag{4.27}$$

Dans cette relation, y_B est la profondeur moyenne de la zone de charge d'espace.
On peut également calculer la conductance de sortie.

$$g_d = \left(\frac{\partial I_{DS}}{\partial V_{DS}}\right)_{V_{GS},\, V_{BS}}$$

$$g_d = \frac{\exp^{-\frac{V_{DS}}{\phi_t}}}{1-\exp^{-\frac{V_{DS}}{\phi_t}}}\frac{I_D}{\phi_t} \tag{4.28}$$

La conductance de sortie décroît rapidement quand la tension de drain augmente.

Cas B et C : forte inversion, canal long, régime non saturé et saturé.

Les mêmes calculs sont repris avec une expression différente pour le courant de drain.

$$g_m = \frac{W}{L}\mu_n C'_{OX} V_{DS} \quad \text{si } V_{DS} \le V_{DSsat}$$

$$g_m = \frac{W}{L}\mu_n C'_{OX} V_{DSsat} \quad \text{si } V_{DS} > V_{DSsat}$$

En régime saturé et uniquement dans ce cas, on peut écrire :

$$g_m = \frac{W}{L}\mu_n C'_{OX}\frac{V_{GS}-V_T}{1+\delta} \quad \text{si } V_{DS} > V_{DSsat}$$

$$g_m = \frac{2\,I_D}{V_{GS}-V_T} \quad \text{si } V_{DS} > V_{DSsat} \tag{4.29}$$

Cette dernière formule est d'un usage très courant car elle relie transconductance et courant consommé. On peut également écrire en exprimant la différence de tension $V_{GS}-V_T$ en fonction du courant de drain :

$$g_m^2 = \frac{W}{L}\mu_n\frac{C'_{OX}}{1+\delta}2\,I_D$$

On note,

$$\beta_n = \frac{W}{L}\mu_n C'_{OX}$$

alors,

$$g_m = \sqrt{2\beta_n \frac{I_D}{1+\delta}} \quad \text{si } V_{DS} > V_{DSsat} \tag{4.30}$$

Cette formule est très utilisée, souvent en négligeant δ. Rappelons qu'elle est valable uniquement en régime saturé et pour un MOS à canal long.

Le paramètre g_{mb} est plus difficile à calculer de manière rigoureuse. Il faut revenir à l'équation 4.2 de ce chapitre pour exprimer cette transconductance. Nous nous contenterons de donner le résultat.

$$\frac{g_{mb}}{g_m} = \gamma \frac{\sqrt{V_{DS} + V_{SB} + \phi_B} - \sqrt{V_{SB} + \phi_B}}{V_{DS}} \quad \text{si } V_{DS} \leq V_{DSsat}$$

$$\frac{g_{mb}}{g_m} = \gamma \frac{\sqrt{V_{DSsat} + V_{SB} + \phi_B} - \sqrt{V_{SB} + \phi_B}}{V_{DSsat}} \quad \text{si } V_{DS} > V_{DSsat}$$

Calculons maintenant la conductance g_d. C'est la dérivée du courant de sortie par rapport à la tension de drain, la tension de grille étant donnée.

La formule du courant en régime non saturé conduit à :

$$g_d = \frac{W}{L} \mu C'_{OX} [V_{GS} - V_T - (1+\delta) V_{DS}] \tag{4.31}$$

Quand le régime de saturation est atteint le courant reste constant et la transconductance est nulle. Il faut alors revenir aux résultats du paragraphe 4.4.1 pour expliquer l'origine de la conductance de sortie et pour en déduire sa valeur.

$$I_D = I_{Dsat} \left(1 + \frac{V_{DS} - V_{DSsat}}{V_A} \right)$$

On en déduit,

$$g_d = \frac{I_{Dsat}}{V_A} \tag{4.32}$$

Rappelons que le paramètre V_A est égal à $kL\sqrt{N_A}$, formule dans laquelle k vaut environ 0,15 Vμm^2.

Cas E et F : canal court, forte inversion, régime saturé et non saturé.

En régime saturé, on obtient simplement :

$$g_m = WC'_{OX} \mu_n E_C \tag{4.33}$$

La formule approchée est particulièrement simple. La transconductance ne dépend plus que de la largeur du MOS au premier ordre. On néglige dans ce calcul tiré de la relation 4.14 la variation de la tension de saturation avec la tension de grille.

La conductance de sortie est due à la variation de la longueur effective de canal et est donc donnée par la formule établie dans les cas B et C.

$$g_d = \frac{I_{Dsat}}{V_A}$$

On peut également montrer que dans le régime de canal court la relation suivante lie transconductance et conductance de sortie.

$$\frac{g_d}{g_m} \approx 0{,}5\frac{\varepsilon_s}{\varepsilon_{OX}}\frac{d_{OX}}{L} \tag{4.34}$$

Cette relation est d'un grand intérêt car elle illustre la compétition entre la grille et le drain pour commander le transistor. Les influences des permittivités et des dimensions sont évidentes. Pour réaliser un dispositif commandé par la grille, il faut minimiser ce ratio et donc réduire l'épaisseur de l'oxyde ou bien augmenter la permittivité de l'oxyde de grille. Cette remarque explique les efforts accomplis actuellement pour développer de nouveaux oxydes de permittivité plus élevée que celle de la silice.

4.6.4 *Le MOS interne à fréquence moyenne*

Un schéma électrique équivalent est une représentation des équations donnant les courants aux quatre bornes de sortie du transistor. Ces courants sont en fait les petites variations des courants globaux, ils sont notés avec des minuscules pour bien faire la différence. La méthode consiste à exprimer chaque variation par rapport aux tensions appliquées puis à sommer les contributions. Si nous reprenons maintenant les courants de source et de drain, rappelons les relations suivantes :

$$\begin{aligned} I_D(t) &= I_C(t) + I_{DV}(t) \\ I_S(t) &= -I_C(t) + I_{SV}(t) \end{aligned}$$

Rappelons que le courant I_c est le courant de conduction, somme du courant de diffusion et du courant de dérive. Les courants sont composés d'une partie continue et d'une partie variable autour de cette valeur continue et de faible amplitude appelée composante petit signal. On peut donc écrire :

$$\begin{aligned} I_D + i_d(t) &= I_C + i_c(t) + i_{dv}(t) \\ I_S + i_s(t) &= -I_C - i_c(t) + i_{sv}(t) \end{aligned}$$

Comme les courants I_{DV} et I_{SV} ne comportent pas de composantes continues puisqu'ils sont des courants liés à des variations de charge, ils seront notés i_{dv} et i_{sv} dans la suite de l'analyse.

L'écriture des variations demande quelques précisions. On superpose aux valeurs continues des tensions appliquées au dispositif des valeurs variables dans le temps et de faibles valeurs. Ces valeurs sont notées avec des minuscules comme par exemple v_g et on écrit :

$$V_G(x,t) = V_G + v_g(t)$$

La fonction $v_g(t)$ peut être une fonction sinusoïdale par exemple. On écrit de manière évidente :

$$\frac{\mathrm{d}V_G}{\mathrm{d}t} = \frac{\mathrm{d}v_g}{\mathrm{d}t}$$

Le dispositif est représenté *figure 4.22*. Il ne prend en compte que le canal et assimile source et drain à de simples contacts ponctuels.

On exprime alors les variations de courant de drain en séparant les parties conduction et transition. Pour compléter l'analyse du paragraphe 4.5.1 qui s'intéressait aux courants transitoires i_{dv} et i_{sv}, nous allons également calculer les courants transitoires de grille et de bulk.

Nous allons dans une première étape nous intéresser à un modèle complet. Un modèle complet est l'expression des variations des quatre courants de transition entrant dans le dispositif en fonction

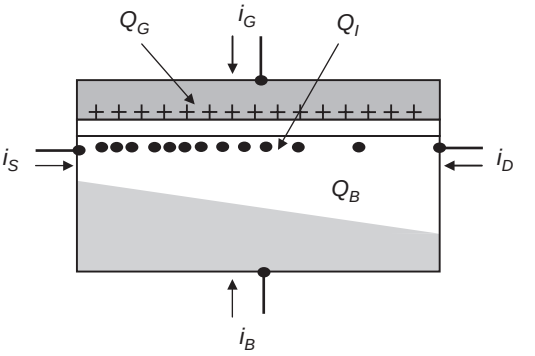


Figure 4.22 – La représentation du transistor.

des variations des quatre tensions de commande ce qui conduit à 16 capacités de couplage. Ce modèle global s'écrit :

$$i_{dv}(t) = C_{dd}\frac{dv_d}{dt} - C_{dg}\frac{dv_g}{dt} - C_{db}\frac{dv_b}{dt} - C_{ds}\frac{dv_s}{dt}$$

$$i_g(t) = -C_{gd}\frac{dv_d}{dt} + C_{gg}\frac{dv_g}{dt} - C_{gb}\frac{dv_b}{dt} - C_{gs}\frac{dv_s}{dt}$$

$$i_b(t) = -C_{bd}\frac{dv_d}{dt} - C_{bg}\frac{dv_g}{dt} + C_{bb}\frac{dv_b}{dt} - C_{bs}\frac{dv_s}{dt}$$

$$i_{sv}(t) = -C_{sd}\frac{dv_d}{dt} - C_{sg}\frac{dv_g}{dt} - C_{sb}\frac{dv_b}{dt} + C_{bs}\frac{dv_s}{dt}$$

En fait, neuf capacités sont indépendantes. Il y a en effet quatre relations du type suivant :

$$C_{gg} = C_{gs} + C_{gd} + C_{gb}$$

Cette relation se démontre à partir de l'expression générale du courant de grille en considérant le système particulier dans lequel les tensions varient mais sont toutes égales. Le courant transitoire de grille dans ce cas est nécessairement nul ce qui établit la relation. On obtient les trois autres équations en considérant les trois autres courants.

$$C_{dd} = C_{dg} + C_{db} + C_{ds}$$
$$C_{bb} = C_{bd} + C_{bg} + C_{bs}$$
$$C_{ss} = C_{sd} + C_{sg} + C_{sb}$$

On peut également obtenir quatre relations en écrivant que la somme des courants de transition entrant dans le dispositif est nul. En effet, la somme des courants globaux est nulle et la somme des courants de conduction est également nulle. Dans le cas particulier où toutes les tensions sont fixes sauf la tension de drain, on obtient à partir des équations générales :

$$\frac{dv_d}{dt}(C_{dd} - C_{gd} - C_{bd} - C_{sd}) = 0$$

Comme cette relation est vérifiée pour toute valeur de la tension de drain, on écrit :

$$C_{dd} - C_{gd} - C_{bd} - C_{sd} = 0$$

De même, on obtient :

$$\begin{aligned} C_{gg} - C_{dg} - C_{bg} - C_{sg} &= 0 \\ C_{bb} - C_{db} - C_{gb} - C_{sb} &= 0 \\ C_{ss} - C_{ds} - C_{gs} - C_{bs} &= 0 \end{aligned}$$

On peut aussi noter que la connaissance de trois courants suffit puisque le quatrième se calcule par la condition de nullité de la somme. De plus, les relations intéressantes font intervenir non pas les tensions absolues v_g, v_d, v_s, v_b mais les tensions mesurées relativement à la source, soit v_{ds}, v_{gs} et v_{bs}. Par exemple, la tension v_{ds} s'exprime par :

$$v_{ds} = v_d - v_s$$

On en arrive facilement au jeu d'équations suivantes :

$$\begin{aligned} i_{dv}(t) &= C_{dd}\frac{\mathrm{d}v_{ds}}{\mathrm{d}t} - C_{dg}\frac{\mathrm{d}v_{gs}}{\mathrm{d}t} - C_{db}\frac{\mathrm{d}v_{bs}}{\mathrm{d}t} \\ i_g(t) &= -C_{gd}\frac{\mathrm{d}v_{ds}}{\mathrm{d}t} + C_{gg}\frac{\mathrm{d}v_{gs}}{\mathrm{d}t} - C_{gb}\frac{\mathrm{d}v_{bs}}{\mathrm{d}t} \\ i_b(t) &= -C_{bd}\frac{\mathrm{d}v_{ds}}{\mathrm{d}t} - C_{bg}\frac{\mathrm{d}v_{gs}}{\mathrm{d}t} + C_{bb}\frac{\mathrm{d}v_{bs}}{\mathrm{d}t} \end{aligned} \tag{4.36}$$

Ce jeu d'équations suffit à résoudre notre problème puisque le courant $i_{sv}(t)$ s'exprime en fonction des trois autres. Il est maintenant possible de faire correspondre à ces équations un schéma équivalent. L'écriture des équations de Kirchhoff aux nœuds de ce schéma doit alors conduire au même jeu d'équations. C'est le critère de validité du modèle. Cette manière de procéder est rigoureuse mais conduit à un schéma complexe peu utilisé en pratique. On lui préfère un schéma plus simple.

En effet, on peut négliger un certain nombre de termes et aboutir au schéma classique de la *figure 4.23*. Revenons au schéma global du transistor MOS en oubliant provisoirement source et drain comme le montre la *figure 4.22*. Ce schéma a pour but d'expliquer les variations des charges stockées dans le dispositif, à savoir la charge de grille Q_G et la charge de la zone de charge d'espace Q_B. Ces charges varient quand les tensions appliquées sont modifiées pendant dt. Prenons l'exemple de la charge accumulée sur la grille. Quand elle varie sous l'effet d'une variation de l'une des tensions appliquées au système, un courant transitoire $i_g(t)$ circule dans le fil de grille. Pour s'en convaincre, on applique le principe de conservation de la charge à la grille.

$$i_g(t) = \frac{\mathrm{d}Q_G}{\mathrm{d}t}$$

On peut alors considérer les variations induites par les variations de V_S, V_B, V_G et V_D sur la charge de grille. Les tensions sont mesurées à partir d'une référence quelconque.

$$i_g(t) = \frac{\partial Q_G}{\partial V_S}\frac{\mathrm{d}v_S}{\mathrm{d}t} + \frac{\partial Q_G}{\partial V_B}\frac{\mathrm{d}v_B}{\mathrm{d}t} + \frac{\partial Q_G}{\partial V_D}\frac{\mathrm{d}v_D}{\mathrm{d}t} + \frac{\partial Q_G}{\partial V_G}\frac{\mathrm{d}v_G}{\mathrm{d}t}$$

On pose alors :

$$C_{gg} = \left(\frac{\partial Q_G}{\partial V_G}\right)_{V_S,\, V_D,\, V_B}$$

$$C_{gs} = -\left(\frac{\partial Q_G}{\partial V_S}\right)_{V_G,\, V_D,\, V_B}$$

$$C_{gd} = -\left(\frac{\partial Q_G}{\partial V_D}\right)_{V_S,\, V_G,\, V_B}$$

$$C_{gb} = -\left(\frac{\partial Q_G}{\partial V_B}\right)_{V_S,\, V_G,\, V_D}$$

On écrit ensuite :

$$i_g(t) = -C_{gs}\frac{\mathrm{d}v_S}{\mathrm{d}t} - C_{gb}\frac{\mathrm{d}v_B}{\mathrm{d}t} - C_{gd}\frac{\mathrm{d}v_D}{\mathrm{d}t} - C_{gg}\frac{\mathrm{d}v_G}{\mathrm{d}t}$$

Si nous ajoutons à cela la relation établie précédemment :

$$C_{gg} = C_{gs} + C_{gd} + C_{gb}$$

On vérifie facilement que la *figure 4.23* traduit correctement le jeu d'équations. Il suffit d'appliquer les lois de Kirchhoff au circuit représenté. Ce n'est pas tout à fait immédiat et il faut remplacer, dans la relation donnant le courant, le coefficient C_{gg} par la somme exprimée ci-dessus. Les courants symbolisés par des sources liées ($g_m v_{gs}$ *et* $g_{mb} v_{bs}$) sont les parties variables des courants de conduction.

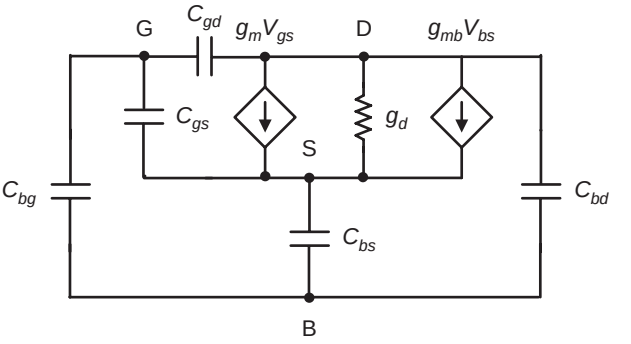


Figure 4.23 – Schéma équivalent petits signaux.

Le même calcul peut se faire en considérant la zone de bulk incluant la zone de charge d'espace de charge Q_B mais excluant la zone de canal. Le seul courant entrant dans cette zone est le courant transitoire de bulk. Rappelons une nouvelle fois que les courants de diffusion et de dérive sont répartis tout le long du canal et sont perpendiculaires à celui-ci. Ils se compensent et n'interviennent donc pas dans l'expression du courant total. On en déduit donc les capacités équivalentes :

$$C_{bb} = \left(\frac{\partial Q_B}{\partial V_B}\right)_{V_G, V_D, V_S}$$

$$C_{bs} = -\left(\frac{\partial Q_B}{\partial V_S}\right)_{V_G, V_D, V_B}$$

$$C_{bd} = -\left(\frac{\partial Q_B}{\partial V_D}\right)_{V_G, V_S, V_B}$$

$$C_{bg} = -\left(\frac{\partial Q_B}{\partial V_G}\right)_{V_D, V_S, V_B}$$

Cette dernière capacité n'est pas nécessairement égale à C_{gb}. Cette remarque importante illustre bien le fait que les capacités introduites sont des capacités équivalentes exprimant les influences électriques dans le dispositif et ne sont pas les capacités de condensateurs géométriquement identifiés.

Au niveau du drain et de la source, les relations entre courants transitoires et charges sont beaucoup moins évidentes et ont conduit à définir les charges équivalentes de source et de drain comme il a été expliqué dans le paragraphe 4.5.1.

Le calcul de ces capacités équivalentes peut se faire en fonction des expressions analytiques des charges dans les différents régimes de fonctionnement. Rappelons que l'hypothèse quasi-statique conduit à supposer que les expressions établies en régime continu sont encore valables. Pour obtenir la valeur des capacités équivalentes, il faut calculer les charges totales en fonction des charges par unité de surface. Prenons l'exemple de la charge de grille en inversion forte.

$$Q_G = W\int_0^L Q'_G \, \mathrm{d}x$$

Comme,

$$I_D = \mu_n W(-Q'_I)\frac{\mathrm{d}V_{CB}(x)}{\mathrm{d}x}$$

On exprime la variation spatiale en fonction de la variation de potentiel.

$$\mathrm{d}x = -\frac{\mu_n W}{I_D} Q'_I \, \mathrm{d}V_{CB}(x)$$

On écrit alors :

$$Q_G = -\frac{\mu_n W^2}{I_D}\int_{V_{SB}}^{V_{DB}} Q'_G \, Q'_I \, \mathrm{d}V_{CB}$$

On calcule de même Q_B et Q_G. Les formules approchées du régime de forte inversion conduisent aux valeurs suivantes :

$$Q_B = -WLC'_{OX}\left(\gamma\sqrt{\phi_B + V_{SB}} + \frac{\delta}{1+\delta}(V_{GS} - V_T)\right)\left(1 - \frac{2}{3}\frac{1+\alpha+\alpha^2}{1+\alpha}\right)$$

$$Q_I = -WLC'_{OX}(V_{GS} - V_T)\frac{2}{3}\frac{1+\alpha+\alpha^2}{1+\alpha}$$

$$Q_G = WLC'_{OX}\left[\gamma\sqrt{\phi_B + V_{SB}} + \frac{V_{GS} - V_T}{1+\delta}\left(\delta + \frac{2}{3}\frac{1+\alpha+\alpha^2}{1+\alpha}\right)\right] - Q_0$$

On en déduit donc les capacités équivalentes en négligeant les dérivées de δ par rapport aux tensions V_S et V_B. Les calculs sont assez fastidieux et ne sont pas détaillés.

$$C_{gs} = -\left(\frac{\partial Q_G}{\partial V_S}\right)_{V_G, V_D, V_B} = \frac{2}{3} C'_{OX} WL \frac{1 + 2\alpha}{(1 + \alpha)^2}$$

$$C_{bs} = -\left(\frac{\partial Q_B}{\partial V_S}\right)_{V_G, V_D, V_B} = \delta \frac{2}{3} C'_{OX} WL \frac{1 + 2\alpha}{(1 + \alpha)^2}$$

$$C_{gd} = -\left(\frac{\partial Q_G}{\partial V_D}\right)_{V_G, V_S, V_B} = \frac{2}{3} C'_{OX} WL \frac{\alpha^2 + 2\alpha}{(1 + \alpha)^2}$$

$$C_{bd} = -\left(\frac{\partial Q_B}{\partial V_D}\right)_{V_G, V_S, V_B} = \delta \frac{2}{3} C'_{OX} WL \frac{\alpha^2 + 2\alpha}{(1 + \alpha)^2}$$

$$C_{gb} = -\left(\frac{\partial Q_G}{\partial V_B}\right)_{V_G, V_S, V_B} = \frac{\delta}{3(1 + \delta)} C'_{OX} WL \left(\frac{1 - \alpha}{1 + \alpha}\right)^2$$

Rappelons qu'en régime saturé, le paramètre α est nul.

Il est maintenant possible de représenter les variations de ces capacités en fonction de la tension de drain et donc du régime de fonctionnement. La *figure 4.24* montre les variations des capacités équivalentes les plus importantes pour un transistor en régime de conduction. Les autres capacités ayant des valeurs plus faibles ne sont pas représentées.

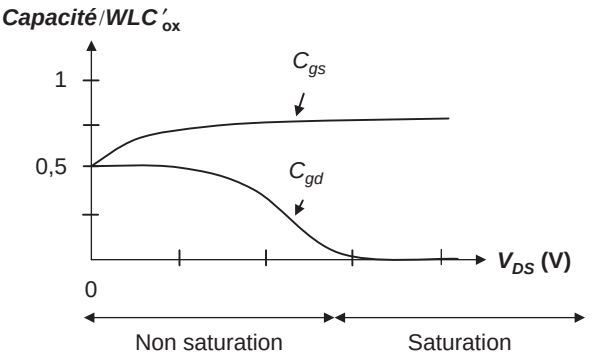


Figure 4.24 – Capacités équivalentes en fonction de la tension de drain.

Il est intéressant de noter que la capacité grille-source passe de $1/2\,WLC'_{OX}$ à $2/3\,WLC'_{OX}$ quand on passe du régime non saturé au régime saturé. Les capacités grille-drain et bulk-drain tendent vers 0 en régime saturé. La tension de drain n'a en effet plus d'action sur la valeur des charges du dispositif. La variation des capacités équivalentes peut également se représenter en fonction de la tension de grille pour une tension de drain donnée.

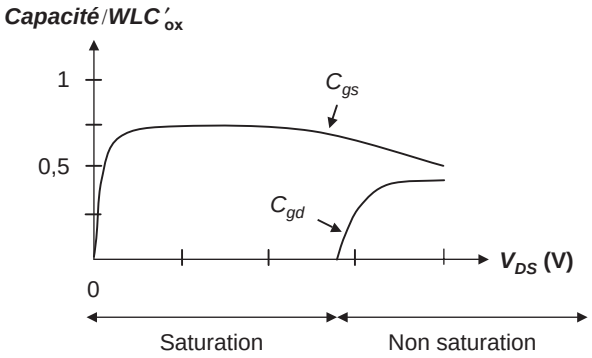


Figure 4.25 – Capacités équivalentes en fonction de la tension de grille.

4.6.5 *Le modèle du MOS complet aux fréquences moyennes*

Le schéma précédent n'est pas suffisant et doit être complété par les éléments suivants :

- prise en compte du recouvrement entre les extrémités et la grille ;
- prise en compte des capacités des jonctions substrat-source et drain-source ;
- prise en compte des résistances d'accès aux électrodes.

Le premier effet est dû aux capacités parasites de couplage entre la grille et les zones de source et de drain comme il est illustré *figure 4.26*. La technique d'auto-alignement de la source et du drain à la fabrication n'est pas parfaite et il subsiste une zone de recouvrement.

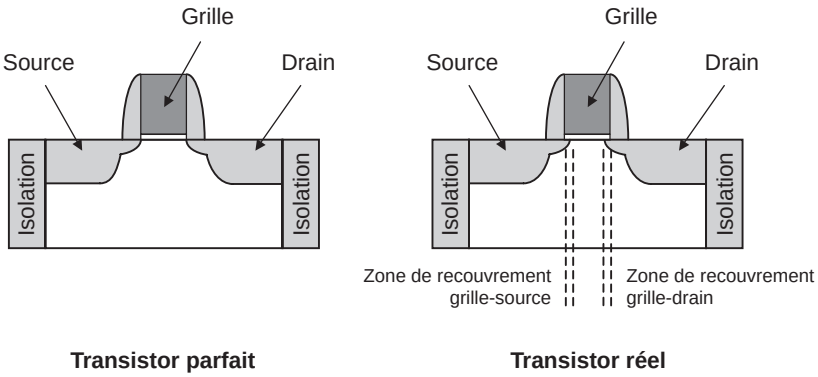


Figure 4.26 – Capacités parasites de recouvrement.

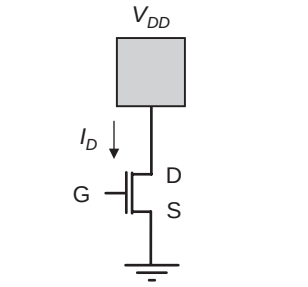
Si la longueur de la zone de recouvrement est *Lr*, les deux capacités parasites ont pour valeurs $WL_rC'_{OX}$. Elles sont notées respectivement C_{gsr} et C_{gdr}. Les effets sont très différents pour le fonctionnement du transistor car la capacité C_{gsr} est une simple augmentation de la capacité C_{gs} que nous avons calculée précédemment et qui est relativement faible. La capacité C_{gdr} a par contre un effet majeur sur le fonctionnement du MOS en saturation car la capacité électrique équivalente C_{gd} est nulle en régime de saturation si on ne tient pas compte du recouvrement.

Cela apparaît *figure 4.24*. Cette capacité de recouvrement injecte en entrée une partie de la tension de sortie ce qui introduit une contre-réaction dans le fonctionnement électrique du MOS. Cet effet conduit à une diminution importante de la bande passante du transistor comme il sera étudié dans le chapitre 7.

4.6.6 *Un résumé du modèle du MOS*

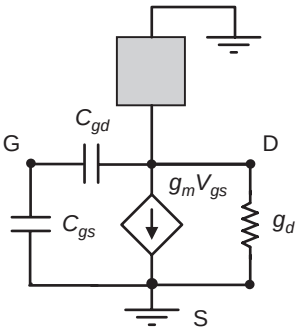
Nous allons maintenant donner un résumé des résultats précédents en simplifiant au maximum la représentation du transistor afin de pouvoir prévoir au premier ordre le comportement des circuits intégrés. Il est ensuite possible et souvent nécessaire d'utiliser les modèles élaborés pour obtenir des résultats précis.

Le premier tableau est relatif au transistor NMOS dans la configuration classique. La grille et le substrat sont reliés et mis à la masse du circuit. Les effets « body » sont alors annulés.



CANAL LONG

	g_m	g_d	C_{gd}	C_{gs}
Off	0	0	$WL_{gdr}C'_{OX}$	$WL_{gsr}C'_{OX}$
Faible inversion	$\frac{I_D}{n_0\phi_t}$	$\frac{\exp^{-\frac{V_{DS}}{\phi_t}}}{1-\exp^{-\frac{V_{DS}}{\phi_t}}}\frac{I_D}{\phi_t}$	$\frac{1}{2}WLC'_{OX}$	$\frac{1}{2}WLC'_{OX}$
Triode $\alpha = 0$	$\frac{W}{L}\mu_n C'_{OX} V_{DS}$	$\beta_n(V_{GS}-V_T-(1+\delta)V_{DS})$	$\frac{1}{2}WLC'_{OX}$	$\frac{1}{2}WLC'_{OX}$
Saturé $\alpha = 1$	$\sqrt{2\beta_n I_D}$	$\frac{I_{Dsat}}{V_A}$	$WL_{gdr}C'_{OX}$	$\frac{2}{3}WLC'_{OX}$



CANAL COURT

	g_m	g_d	C_{gd}	C_{gs}
Off	0	0	$WL_r C'_{OX}$	$WL_{gsr}C'_{OX}$
Faible inversion	$\frac{I_D}{n_0\phi_t}$	$\frac{\exp^{-\frac{V_{DS}}{\phi_t}}}{1-\exp^{-\frac{V_{DS}}{\phi_t}}}\frac{I_{DS}}{\phi_t}$	$\frac{1}{2}WLC'_{OX}$	$\frac{1}{2}WLC'_{OX}$
Triode $\alpha = 0$	$\frac{WC'_{OX}\mu_n\frac{V_{DS}}{L}}{1+(V_{DS}/LE_{Cn})}$	$\frac{\beta_n(V_{GS}-V_T-(1+\delta)V_{DS})}{1+(V_{DS}/LE_{Cn})}$	$\frac{1}{2}WLC'_{OX}$	$\frac{1}{2}WLC'_{OX}$
Saturé $\alpha = 1$	$W\mu_n C'_{OX} E_{Cn}$	$\frac{I_{Dsat}}{V_A}$	$WL_{gdr}C'_{OX}$	$\frac{2}{3}WLC'_{OX}$

Figure 4.27 – Modèle du NMOS.

Rappelons les termes de ces formules. Le courant de drain s'exprime de quatre manières en fonction du régime :

Faible inversion : $$I_D = I_S \exp^{\frac{V_{GS}-V_T}{n_0\phi_t}}\left(1-\exp^{-\frac{V_{DS}}{\phi_t}}\right)$$

Triode et forte inversion : $$I_D = \frac{W}{L}\mu_n C'_{OX}\left[(V_{GS}-V_T)V_{DS}-\frac{1}{2}(1+\delta)V_{DS}^2\right]$$

Saturation et forte inversion : $I_D = \frac{W}{L}\mu_n C'_{OX}\frac{(V_{GS}-V_T)^2}{2(1+\delta)}\left(1+\frac{V_{DS}-V_{DSsat}}{V_A}\right)$

Canal court et triode : $I_D = \frac{\frac{W}{L}\mu_n C'_{OX}}{1+\frac{V_{DS}}{LE_C}}\left[(V_{GS}-V_T)V_{DS}-\frac{1}{2}(1+\delta)V_{DS}^2\right]$

Canal court et saturation : $I_D = WC'_{OX}v_{nsat}\left[V_{GS}-V_T-\frac{1+\delta}{2}V_{DSsat}\right]\left(1+\frac{V_{DS}-V_{DSsat}}{V_A}\right)$

Dans ces relations les paramètres sont définis comme suit :

- W : largeur du transistor
- L : longueur du canal
- C'_{OX} : capacité de l'oxyde par unité de surface (égale à ε_{OX}/t_{OX} rapport de la constante diélectrique de l'oxyde sur l'épaisseur de l'oxyde)
- μ_n : mobilité effective des électrons à l'interface, différente de la mobilité dans le matériau.
- V_T : tension de seuil du NMOS, dépend de paramètres technologiques et de l'effet « body » caractérisé par le paramètre γ défini par $\gamma = \sqrt{2eN_A\varepsilon_s}/C'_{OX}$ mais aussi de la tension de drain de manière plus indirecte.
- δ : paramètre de la modélisation égal à $\delta = \gamma/2\sqrt{2\phi_F + V_{SB}}$
- N_A : dopage du silicium *p*
- ε_s : permittivité du silicium
- ϕ_F : potentiel de Fermi du silicium *p*
- ϕ_t : potentiel égal à k_BT/e
- n_0 : égal à $1+\gamma/2\sqrt{1{,}5\,\phi_F + V_{SB}}$
- $V_{DS\,sat}$: tension de saturation

  Les tensions de saturation sont respectivement :

$$V_{DSsat} = \frac{V_{GS}-V_T}{1+\delta} \quad \text{pour un canal long}$$

$$V_{DSsat} = LE_c\left[\sqrt{1+\frac{2(V_{GS}-V_T)}{(1+\delta)LE_c}}-1\right] \quad \text{pour un canal court}$$

- E_c : champ critique en relation avec v_{sat} par la relation : $v_{sat} = \mu_n E_c$
- V_A : paramètre pour la modulation de canal égal à $kL\sqrt{N_A}$ formule dans laquelle le coefficient k vaut environ $0{,}15\ \text{V}\,\mu\text{m}^{1/2}$
- β_n est défini par : $\beta_n = \mu_n C'_{OX} W/L$
- L_{gdr} et L_{gsr} sont les longueurs de recouvrement entre grille et drain et grille et source dues aux imperfections de la technologie.
- Le courant I_S exprimé dans la formule de la faible inversion est une formule complexe donnée dans le paragraphe 4.3.2.

Il est maintenant possible de donner quelques valeurs numériques dans deux cas : une technologie 0,8 micron et une technologie 45 nm : *tableau 4.2*. On considère deux transistors typiques pour des applications analogiques donc de dimensions non minimales.

Tableau 4.2

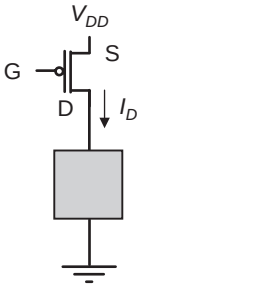
Paramètre	0,8 micron	45 nm
W (µm)	$10 \times 0{,}8$	$50 \times 0{,}045$
L (µm)	$2 \times 0{,}8$	$2 \times 0{,}045$
I_D (µA)	20	10
V_{GS} (V)	1	0,35
$V_{DS\ sat}$ (mV)	250	50
V_T (mV)	900	280
C'_{OX} (Ff/µm²)	1,8	25
$\mu_n C'_{OX}$ (µA/V²)	120	
g_m (µA/V)	150	150
g_d (µA/V)	0,2	6
C_{gd} (fF)	2	1,6
C_{gs} (fF)	23	4,2
C_{OX}	35	6,25
v_{sat} (m/s)		100×10^3

Il faut ajouter à ce tableau un certain nombre de remarques. Le paramètre $\mu_n C'_{OX}$ n'a pas de signification pour un MOS canal court. De la même manière, la vitesse de saturation est utilisée uniquement pour décrire un MOS canal court. Il faut noter une diminution de la résistance de sortie pour le MOS canal court. De plus, la tension de seuil diminue notablement pour le MOS de nouvelle génération ce qui entraîne un courant tunnel d'environ 5 A/cm^2.

Un modèle de type équivalent peut être donné pour le PMOS. On suppose que le transistor est de manière classique connecté entre l'alimentation positive V_{DD} et le reste du circuit comme l'indique la *figure 4.28*. La source est à la tension la plus positive et le caisson du transistor est relié lui aussi à ce potentiel. Tous les montages électriques ne conduisent pas à cette configuration mais elle est très souvent utilisée aussi bien en logique qu'en analogique.

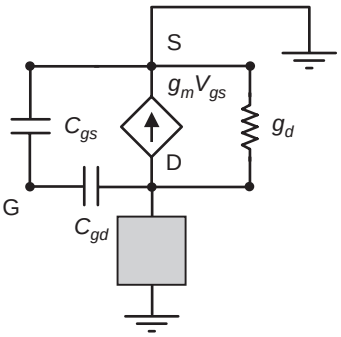
Le lecteur doit prendre garde à l'apparente inversion du sens de la source de courant. Ce n'est que l'effet de la représentation du schéma inversé par rapport à celui du NMOS puisque la source est maintenant en haut de la figure et non pas en bas.

On rappelle les équations donnant le courant de drain en fonction des tensions. Ce sont les mêmes que celle du NMOS à la condition de remplacer V_{GS} par V_{SG} et V_{DS} par V_{SD}. La tension de seuil du PMOS est, rappelons-le, négative quand elle est définie pour la tension entre la grille et la source. Elle est donc positive quand elle est relative à la tension entre source et grille. Ces conventions con-



CANAL LONG

	g_m	g_d	C_{gd}	C_{gs}
Off	0	0	$WL_{gdr}C_{OX}'$	$WL_{gsr}C_{OX}'$
Faible inversion	$\frac{I_D}{n_0\phi_t}$	$\frac{\exp\left(-\frac{V_{SD}}{\phi_t}\right)}{1-\exp\left(-\frac{V_{SD}}{\phi_t}\right)}\frac{I_D}{\phi_t}$	$\frac{1}{2}WLC_{OX}'$	$\frac{1}{2}WLC_{OX}'$
Triode $\alpha = 0$	$\frac{W}{L}\mu_n C_{OX}' V_{DS}$	$\beta_n(V_{SG}-V_T-(1+\delta)V_{SD})$	$\frac{1}{2}WLC_{OX}'$	$\frac{1}{2}WLC_{OX}'$
Saturé $\alpha = 1$	$\sqrt{2\beta_n I_D}$	$\frac{I_{Dsat}}{V_A}$	$WL_{gdr}C_{OX}'$	$\frac{2}{3}WLC_{OX}'$



CANAL COURT

	g_m	g_d	C_{gd}	C_{gs}
Off	0	0	$WL_t C_{OX}'$	$WL_{gsr}C_{OX}'$
Faible inversion	$\frac{I_D}{n_0\phi_t}$	$\frac{\exp\left(-\frac{V_{SD}}{\phi_t}\right)}{1-\exp\left(-\frac{V_{SD}}{\phi_t}\right)}\frac{I_D}{\phi_t}$	$\frac{1}{2}WLC_{OX}'$	$\frac{1}{2}WLC_{OX}'$
Triode $\alpha = 0$	$\frac{WC_{OX}'\mu_n\frac{V_{DS}}{L}}{1+(V_{SD}/LE_{Cp})}$	$\frac{\beta_p(V_{SG}-V_T-(1+\delta)V_{SD})}{1+(V_{SD}/LE_{Cp})}$	$\frac{1}{2}WLC_{OX}'$	$\frac{1}{2}WLC_{OX}'$
Saturé $\alpha = 1$	$W\mu_p C_{OX}' E_{Cp}$	$\frac{I_{Dsat}}{V_A}$	$WL_{gdr}C_{OX}'$	$\frac{2}{3}WLC_{OX}'$

Figure 4.28 – Modèle du PMOS.

duisent donc à des valeurs toutes positives pour le PMOS. Ce point est souvent source de confusion. On obtient alors :

Faible inversion : $$I_D = I_S \exp^{\frac{V_{SG}-V_T}{n_0\phi_t}}\left(1-\exp^{-\frac{V_{SD}}{\phi_t}}\right)$$

Triode et forte inversion : $$I_D = \frac{W}{L}\mu_p C_{OX}'\left[(V_{SG}-V_T)V_{SD}-\frac{1}{2}(1+\delta)V_{SD}^2\right]$$

Saturation et forte inversion : $$I_D = \frac{W}{L}\mu_p C_{OX}'\frac{(V_{SG}-V_T)^2}{2(1+\delta)}\left(1+\frac{V_{SD}-V_{SDsat}}{V_A}\right)$$

Canal court et triode : $$I_D = \frac{\frac{W}{L}\mu_p C_{OX}'}{1+\frac{V_{SD}}{LE_{Cp}}}\left[(V_{SG}-V_T)V_{SD}-\frac{1}{2}(1+\delta)V_{SD}^2\right]$$

Canal court et saturation : $$I_D = WC_{OX}'v_{psat}\left[V_{SG}-V_T-\frac{1+\delta}{2}V_{SDsat}\right]\left(1+\frac{V_{SD}-V_{SDsat}}{V_A}\right)$$

La différence fondamentale entre NMOS et PMOS est la mobilité des porteurs. La mobilité des trous est trois fois plus faible que celle des électrons en régime non saturé au sens de la saturation de la vitesse. On ne retrouve pas cette différence pour la vitesse de saturation qui est environ la même

pour les électrons et les trous. On observe cependant une différence équivalente pour les valeurs des champs critiques. Le champ critique du PMOS est environ trois fois plus élevé que celui du NMOS. En résumé, que ce soit en régime de canal court ou en régime de canal long, le PMOS doit être plus large que le NMOS pour délivrer le même courant. Cette règle fondamentale pour le design conduit à réaliser en général des PMOS trois fois plus larges que les NMOS dans les technologies de canal long et deux fois plus larges dans les technologies de canal court.

Pour terminer ce résumé, le *tableau 4.3* donne les valeurs typiques des paramètres de deux transistors PMOS dans les deux technologies que nous avons envisagées.

Tableau 4.3

Paramètre	0,8 micron	45 nm
W (µm)	$30 \times 0{,}8$	$100 \times 0{,}045$
L (µm)	$2 \times 0{,}8$	$2 \times 0{,}045$
I_D (µA)	20	10
V_{GS} (V)	1,2	0,35
$V_{DS\ sat}$ (mV)	250	50
V_T (mV)	900	280
C'_{OX} (Ff/µm²)	1,8	2
$\mu_p C'_{OX}$ (µA/V²)	40	
g_m (µA/V)	150	150
g_d (µA/V)	0,2	0,1
C_{gd} (fF)	6	3,8
C_{gs} (fF)	70	8,4
C_{OX}	110	13
v_{sat} (m/s)		90×10^3

Pour terminer ce paragraphe important, il est possible de commenter les circuits dans lesquels le « body » du transistor n'est pas relié à la source. Rappelons que le « body » du transistor est le substrat dopé *p* pour le NMOS et le puits dopé *n* pour le PMOS. La tension de seuil augmente quand V_{SB} passe de 0 à une valeur positive. Quand la tension V_{SB} augmente encore, la tension de seuil continue à augmenter mais plus lentement. Ce phénomène est illustré *figure 4.29*.

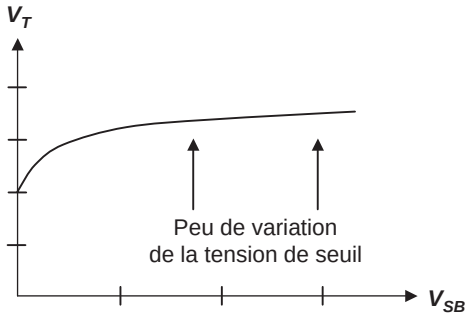


Figure 4.29 – Effet body en fonction de la tension de drain.

4.7 Les modèles électriques de la CAO

L'industrie micro-électronique utilise des modèles électriques dans des logiciels qui permettent de concevoir les circuits intégrés. Ces logiciels sont dérivés du logiciel SPICE mis au point à l'université de Berkeley. Ils résolvent les équations de Kirchhoff pour des circuits comportant des centaines et même des milliers de transistors.

Simuler un circuit c'est :

- trouver les points de fonctionnement statique (courant et tension) ;
- calculer les grandeurs dynamiques (gain, pôles et zéros des fonctions de transfert) autour de ces points de fonctionnement.

Pour cela, il faut disposer de modèles électriques statiques et dynamiques. Les paragraphes précédents ont permis de constituer des modèles statiques et dynamiques simples. Ces modèles analytiques permettent en général de fixer les ordres de grandeurs et d'expliquer le fonctionnement des circuits.

Les phénomènes physiques sont cependant plus complexes et de nombreux effets doivent être pris en compte : variation de la mobilité, recombinaison des porteurs, confinement quantique des électrons, effets tunnel, variations 3D du potentiel électrique… Les modèles se complexifient au fur et à mesure que les dimensions du transistor se réduisent et les modèles les plus complets peuvent comporter des centaines de paramètres. Ces paramètres sont, le plus souvent, des coefficients d'ajustement de fonctions et n'ont pas de signification physique.

Les trois modèles les plus utilisés sont les suivants :

- le modèle BSIM dérivé du modèle SPICE de Berkeley ;
- le modèle EKV ayant un sens physique plus avancé mais moins utilisé ;
- le modèle propriétaire de la société Philips.

Chapitre 5

Le transistor bipolaire

5.1 Principe de fonctionnement
5.2 Technologie de fabrication
5.3 Les applications

Ce chapitre court a pour but de décrire assez sommairement le principe de fonctionnement du transistor bipolaire. Historiquement, le transistor bipolaire a été le premier composant à être fabriqué, et les premières applications électroniques, analogiques ou logiques, ont largement utilisé des bipolaires. Au fil du temps, le transistor MOS a remplacé le transistor bipolaire tout d'abord dans les circuits logiques puis dans les circuits analogiques. La faible consommation du transistor MOS dans les circuits logiques en est la cause principale. Les transistors bipolaires sont parfois utilisés dans des applications logiques demandant des vitesses très élevées. Ils sont principalement utilisés dans les circuits radiofréquence pour des raisons qui seront expliquées dans le paragraphe 5.3. Il faut également noter que les transistors bipolaires sont présents comme composants « parasites » dans les circuits CMOS et servent à générer des sources de tension.

5.1 Principe de fonctionnement

Un transistor bipolaire est formé par deux jonctions *pn* tête-bêche comme le montre la *figure 5.1*. Les tensions appliquées par des générateurs extérieurs sont telles que le champ électrique est dirigé de la base vers l'émetteur dans la zone de charge d'espace de la jonction émetteur-base et de la base vers le collecteur dans la zone de charge d'espace de la jonction base-collecteur. Le champ est cependant faible pour la jonction polarisée en direct si bien qu'il s'oppose peu à la diffusion des trous de l'émetteur vers la base.

La particularité de ce dispositif est dans le fait que les trous ayant traversé la jonction émetteur-base, diffusent dans la zone non chargée de la base et sont ensuite attirés vers le collecteur par le champ de la jonction base-collecteur polarisée en inverse. Ce phénomène est possible car l'épaisseur physique de la base est supposée faible devant la longueur de diffusion des trous.

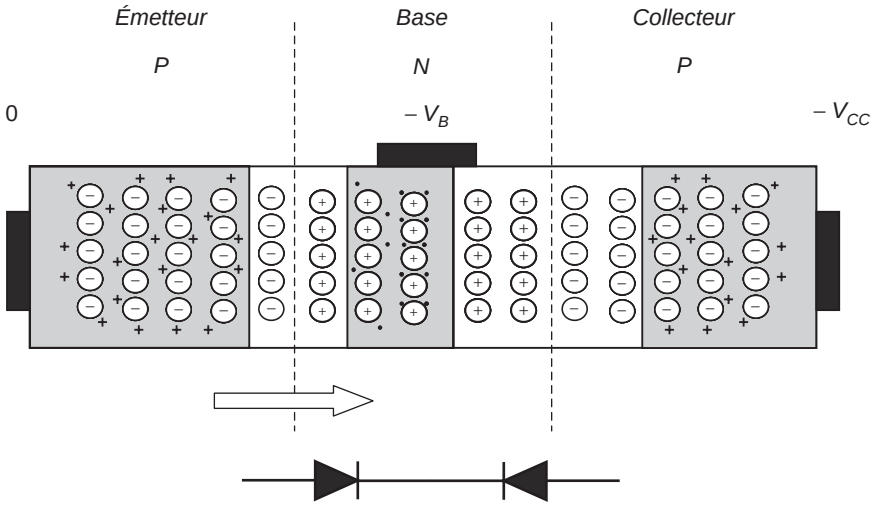


Figure 5.1 – Principe du transistor bipolaire.

En pratique, tous les trous ne traversent pas la base et un certain nombre d'électrons de la base se déplacent vers l'émetteur. On admet que ce flux est de faible valeur. On écrit donc classiquement la relation entre le courant I_E entrant dans l'émetteur et le courant I_C sortant du collecteur.

$$I_c = \alpha I_E \tag{5.1}$$

Le coefficient α est légèrement inférieur à 1 pour les raisons exprimées précédemment. La différence entre les courants d'émetteur et de collecteur est le courant de base supposé sortant du dispositif. On peut donc écrire :

$$I_E = I_C + I_B$$

On en déduit facilement :

$$I_c = \frac{\alpha}{1-\alpha} I_B \tag{5.2}$$

Le facteur de proportionnalité entre I_C et I_B est noté β. Sa valeur est élevée : quelques centaines.

$$\beta = \frac{\alpha}{1-\alpha} \tag{5.3}$$

Cette relation est parfois interprétée comme l'expression d'un gain entre le courant de collecteur et le courant de base. Cette interprétation n'a pas de sens physique car il n'y a pas de notion de contrôle du courant de collecteur par le courant de base. C'est la tension appliquée entre base et émetteur (V_{BE}) qui contrôle le courant I_E traversant la jonction et en conséquence le courant de collecteur. La relation entre le courant d'émetteur et la tension V_{BE} est donnée dans le chapitre 3.

$$I_E = I_S\left(\exp^{\frac{V_{EB}}{\Phi_t}} - 1\right) \tag{5.4}$$

$$I_S = eA\left(\frac{D_p N_{AE}}{L_p} + \frac{D_n N_{DB}}{L_n}\right)$$

$$L_n = \sqrt{D_n \tau_n}$$

$$L_p = \sqrt{D_p \tau_p}$$

Dans cette relation, les grandeurs N_{AE} et N_{DB} sont respectivement le dopage de l'émetteur et le dopage de la base. La grandeur A est la section du dispositif. Les coefficients D et τ sont respectivement le coefficient de diffusion et la durée de vie. Ils sont définis pour les électrons et pour les trous.

Il est également possible de relier le facteur α aux propriétés de la base. Ce calcul est fait dans un grand nombre d'ouvrages spécialisés. Un résultat approximatif est le suivant :

$$\alpha = \frac{1}{\mathrm{ch}\frac{L}{L_p}} \tag{5.5}$$

Dans cette relation, le dénominateur est un cosinus hyperbolique du rapport entre l'épaisseur L de la base et la longueur de diffusion des trous. L'épaisseur L est l'épaisseur de la zone de champ nul c'est-à-dire la zone grisée de la *figure 5.1* et non pas l'épaisseur physique de la base. Ces relations simples suffisent à comprendre l'essentiel des propriétés du transistor.

Quand on augmente la tension entre l'émetteur et la base (la base devenant plus négative), le courant d'émetteur augmente et en conséquence le courant de collecteur. On peut donc par analogie avec le MOS dire que le transistor bipolaire est un dispositif dans lequel le courant de collecteur est contrôlé par la tension émetteur-base. De la même manière, le courant de drain du MOS est commandé par la tension grille-source.

On peut alors tracer les courbes du courant de collecteur en fonction de la tension collecteur-émetteur pour différentes valeurs de la tension V_{EB}. Cette représentation n'est cependant pas très commode car la tension V_{EB} varie peu. L'examen des caractéristiques courant-tension d'une diode montre que la tension varie peu autour de 700 mV. Il est donc plus intéressant de prendre comme paramètre, non pas la tension d'entrée mais le courant d'entrée, c'est-à-dire le courant de base. On obtient donc un réseau de courbes comme le montre la *figure 5.2*.

On peut remarquer sur ces courbes une augmentation du courant de collecteur en fonction de la tension collecteur-émetteur pour un courant de base donné. En théorie, les courbes devraient être

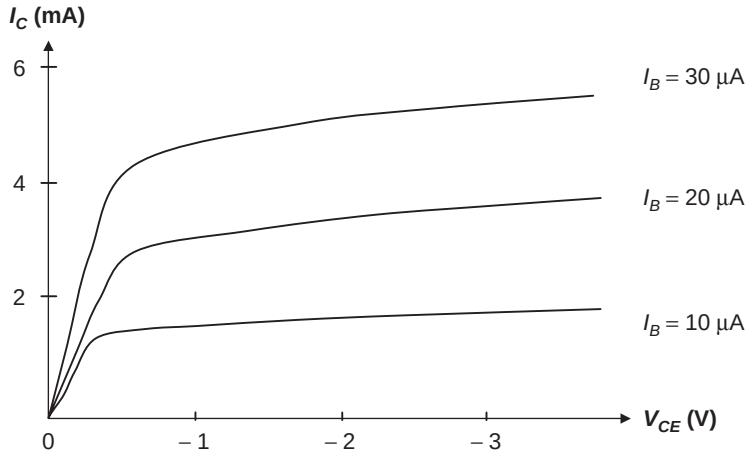


Figure 5.2 – Caractéristiques du transistor bipolaire.

horizontales. En effet, le courant de base étant donné, la relation (5.2) permet d'en déduire le courant de collecteur indépendamment de la tension collecteur-émetteur. Ce serait vrai si le coefficient α était indépendant de la tension V_{CE}. Mais ce n'est pas le cas car un changement de V_{CE} influe sur la jonction collecteur-base. L'étendue de la zone de charge d'espace est modifiée et donc l'épaisseur L de la base. Le coefficient α varie donc en accord avec la relation (5.5). Cet effet est appelé effet Early dans la littérature.

Les considérations précédentes peuvent également s'appliquer à un dispositif NPN. De la même manière la jonction *np* est polarisée en direct et la jonction *pn* en inverse comme le montre la *figure 5.3*.

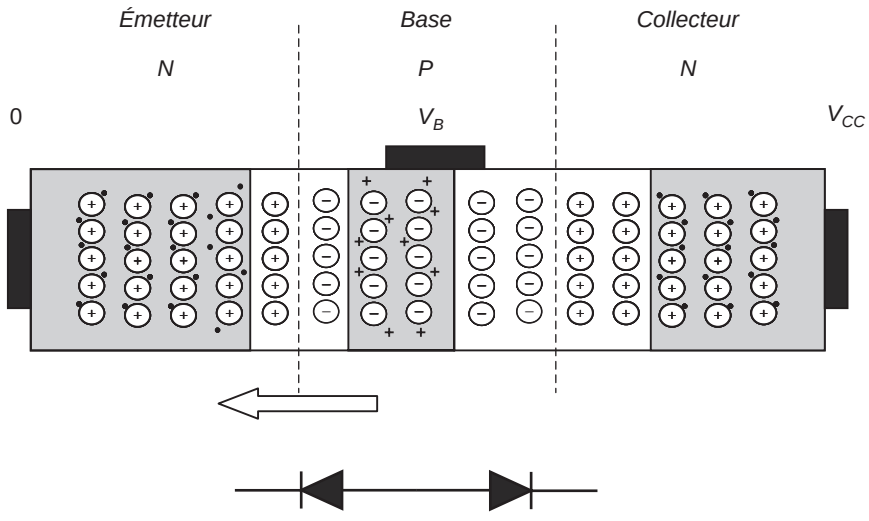


Figure 5.3 – Transistor NPN.

Le dispositif est symétrique. Les tensions appliquées sont positives par rapport à l'émetteur. La *figure 5.4* résume le fonctionnement du transistor PNP et du transistor NPN en indiquant les polarités des tensions de fonctionnement.

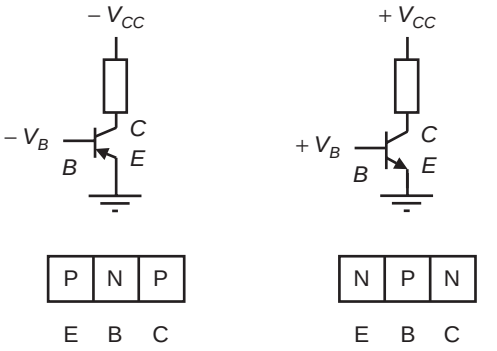


Figure 5.4 – Transistors NPN et PNP.

Notons que le sens de la flèche sur le schéma indique le sens du courant.

Pour compléter cette courte description, donnons quelques éléments sur le fonctionnement dynamique du transistor en petits signaux Dans une première étape, il est possible de calculer la transconductance du transistor PNP :

$$I_E = I_S\left(\exp^{\frac{V_{EB}}{\phi_t}} - 1\right)$$

$$g_m = \frac{\partial I_C}{\partial V_{BE}}$$

On en déduit donc pour V_{BE} très supérieur à ϕ_t :

$$g_m = -\frac{I_E}{\phi_t}$$

Si le courant petit signal i_c est orienté du collecteur vers l'émetteur, on écrit :

$$i_c = \frac{I_E}{\phi_t} v_{be}$$

On peut aussi calculer l'impédance de sortie :

$$r_c = \frac{\partial I_C}{\partial V_{CE}}$$

Le calcul est plus difficile car il est basé sur l'effet Early. Il faut calculer l'effet d'une variation de la tension collecteur-émetteur sur la longueur L de la zone neutre de la base puis en déduire la variation du coefficient α et finalement la variation du courant de collecteur.

Les effets temporels seront pris en compte de deux manières : premièrement en ajoutant au schéma équivalent les capacités des deux jonctions et deuxièmement en tenant compte du temps de transit des porteurs dans la base. Rappelons que la capacité base-émetteur est élevée puisque la jonction est polarisée en direct et que la capacité collecteur-base est de faible valeur puisque la jonction est polarisée en inverse. Ajoutons à cela la résistances $r_{bb'}$, résistance d'accès à la base, et nous obtenons le schéma élémentaire du transistor en petits signaux représenté *figure 5.5*.

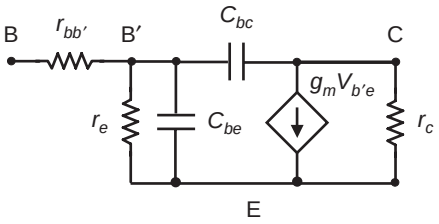


Figure 5.5 – Schéma petits signaux du transistor bipolaire.

Dans ce schéma, le seul paramètre que nous n'avons pas estimé est la résistance d'entrée r_e. Pour le calculer, il suffit d'exprimer :

$$r_e = \frac{\partial V_{EB}}{\partial I_B}$$

On obtient donc en fonction des équations 5.2 et 5.3 :

$$r_e = \frac{\partial V_{EB}}{\partial I_B} = \frac{\partial V_{EB}}{\partial I_E} \cdot \frac{1}{1-\alpha}$$

soit,

$$r_e = \frac{\phi_t}{I_E} \cdot (1+\beta) \qquad (5.6)$$

Ajoutons pour terminer ce paragraphe que le temps de transit dans la base se traduit par l'écriture d'une fréquence de coupure dans la transconductance g_m.

5.2 Technologie de fabrication

Le transistor bipolaire contrairement au transistor MOS est un dispositif de volume et non pas de surface. Cela est illustré *figure 5.6*.

Cette vue simplifiée des deux dispositifs montre une différence fondamentale :

- le courant d'un MOS est parallèle à la surface du circuit intégré et est proportionnel au rapport de dimension *W/L* ;
- le courant d'un transistor bipolaire est perpendiculaire à la surface de silicium et est proportionnel à la surface du dispositif.

Le fonctionnement du transistor bipolaire est plus complexe que celui exposé dans le modèle simpliste du paragraphe 5.1. Nous allons maintenant étudier quelques particularités.

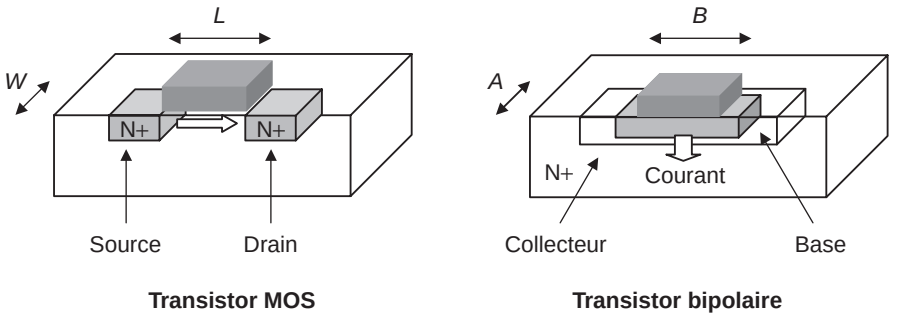


Figure 5.6 – MOS et bipolaire.

Quand les tensions appliquées augmentent, les zones de charge d'espace s'étendent et, à la limite, se rejoignent. Il y a alors perçage de la base et augmentation brutale du courant de collecteur. Les transistors ont des bases fortement dopées et des collecteurs peu dopés si bien que le perçage de la base survient généralement après le claquage par avalanche de la jonction base-collecteur. Ce claquage est dû au phénomène d'ionisation par impact. Les porteurs du transistor peuvent acquérir une énergie tellement élevée qu'ils sont capables d'ioniser les atomes de silicium et de créer des paires électron-trou.

Quand le courant débité par le transistor augmente, les chutes de tension dans les résistances d'accès augmentent. Le centre du transistor est moins polarisé que la périphérie et le courant circule principalement en périphérie. Ce phénomène est appelé défocalisation du courant émetteur.

Les effets thermiques sont également importants dans les transistors bipolaires. Dans les MOS une élévation de température conduit à une diminution du courant. Dans les bipolaires, le courant augmente avec la température et il est nécessaire d'évacuer la puissance pour éviter l'emballement.

Pour réaliser les transistors bipolaires, il est nécessaire comme pour les MOS de prévoir des dispositifs d'isolation évitant les courants de fuite. Les trois techniques d'isolation sont :

- la fabrication de jonctions d'isolement ;
- la fabrication de tranchées remplies de diélectrique ;
- la combinaison des deux.

Ces mêmes techniques sont mises en œuvre pour les circuits CMOS. Elles sont illustrées *figure 5.7*.

Les techniques d'isolement par jonctions sont les plus anciennes. Elles comportent cependant des transistors parasites dont il faut limiter la conduction. La solution purement diélectrique est en théorie la meilleure puisque le transistor est entouré par un isolant. Elle nécessite cependant un substrat spécifique de type SOI (*Silicon On Insulator*). La solution mixte intègre des tranchées mais conserve la jonction collecteur-substrat. Les tranchées sont profondes pour allonger les chemins de conduction. C'est la solution nominale dans la micro-électronique. La réalisation du collecteur doit résoudre le compromis suivant : un dopage élevé réduit la résistance d'accès mais augmente la capacité de jonction collecteur-base. La solution est d'empiler deux couches dopées *n*, une faiblement dopée en contact avec la base et une fortement dopée. Différentes techniques sont mises en œuvre pour ramener le contact de base en surface. Le but est de réduire la capacité collecteur-base qui contribue à l'effet Miller. Une couche de polysilicium et parfois deux couches sont utilisées pour réaliser les contacts de base et d'émetteur.

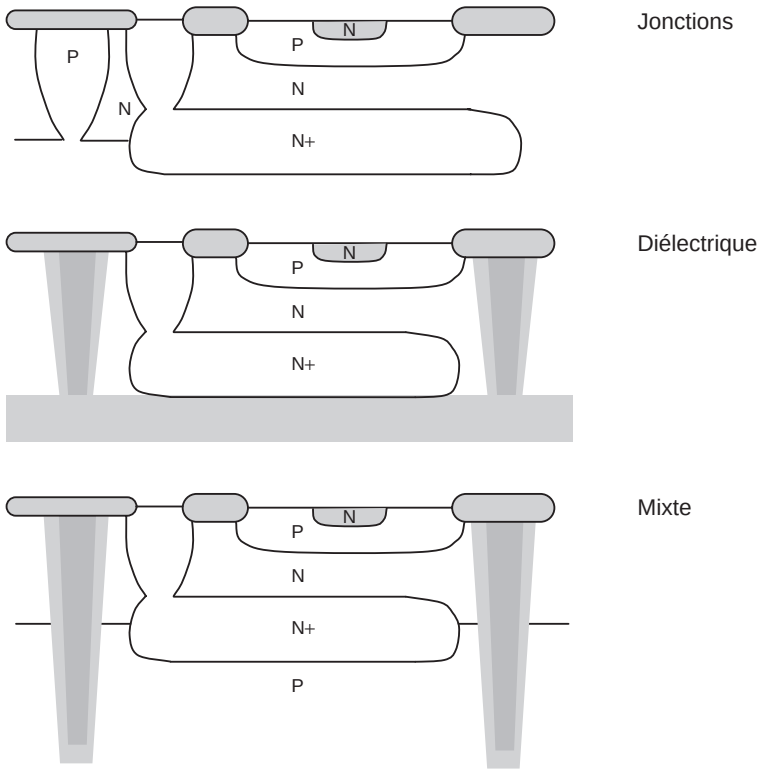


Figure 5.7 – Techniques d'isolement.

Ajoutons à cette brève description deux éléments technologiques décisifs pour les applications à fréquence élevée : la structure interdigitée et le matériau silicium-germanium.

La structure interdigitée consiste à réaliser un transistor en alternant différents émetteurs et bases. Cette technique permet alors de réduire l'effet de défocalisation en augmentant la part de conduction de surface et de réduire les résistances d'accès puisqu'elles sont en parallèle. Le matériau SiGe est un alliage à quelques pour-cent de Germanium. Il permet de doper plus fortement la base du transistor, de réduire la résistance d'accès, et donc d'améliorer les performances fréquentielles.

Notons pour terminer que contrairement au MOS, le transistor bipolaire est un composant de volume et est donc de ce fait beaucoup plus sensible aux défauts du cristal et à la présence d'impuretés, en particulier d'impuretés métalliques.

5.3 Les applications

La technologie bipolaire n'est jamais utilisée seule. Elle est en général associée à la technologie CMOS dans une technologie mixte appelée BICMOS. Cette technologie offre les avantages des deux types de transistors. Elle conduit cependant à augmenter le nombre de niveaux de masques et entraîne un surcoût par rapport à une technologie purement CMOS. Elle est donc réservée aux applications haute fréquence qui demandent des performances en terme de bande passante. Nous pouvons également noter une utilisation particulière des transistors bipolaires dans les circuits

MOS. C'est la fabrication des références de tensions. Cette technique utilise en fait le transistor bipolaire parasite.

5.3.1 *Références de tensions et utilisation du bipolaire parasite*

Il est nécessaire dans la conception des fonctions analogiques de concevoir des sources de tension qui ne varient pas avec la température. Ces sources serviront par exemple à polariser les amplificateurs opérationnels. La valeur absolue n'est pas nécessairement définie avec une très grande précision mais la stabilité avec la température est cruciale. Dans le cas contraire, la tension de sortie de l'amplificateur pourrait varier beaucoup avec la température. Il est difficile de réaliser des sources stables en utilisant uniquement des MOSFET et des résistances. Il est par contre possible de concevoir des systèmes dont la tension de sortie varie peu avec la température en associant des circuits à base de MOS et des transistors. Ces transistors ne sont pas fabriqués spécifiquement comme dans la technologie BICMOS mais sont simplement les transistors parasites de la technologie CMOS comme le montre la *figure 5.8*.

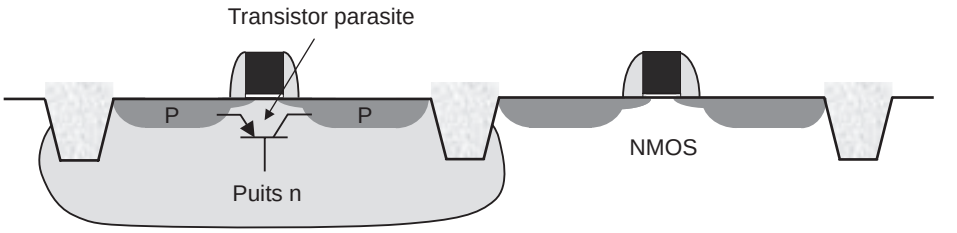


Figure 5.8 – Transistor parasite dans la CMOS.

Ce transistor PNP sera donc utilisé dans un schéma assez complexe servant à créer une référence de tension stable en température. Le principe général de ce schéma est de cascader deux dispositifs dont les sorties varient en sens opposés avec la température. Ce schéma est appelé « bandgap ».

5.3.2 *Utilisation en radiofréquence*

Dans les circuits radiofréquence, des fréquences supérieures à 100 MHz sont mises en œuvre allant jusqu'à des valeurs aussi élevées que quelques GHz et pouvant même atteindre 60 GHz pour les communications cellulaires. Le comportement haute fréquence des transistors est alors déterminant. Il dépend de la fréquence de coupure du transistor qui à deux origines :

- le temps de transit des porteurs dans le canal pour le MOSFET et dans la base pour le bipolaire ;
- les constantes de temps créées par les résistances d'accès et les capacités du composant.

Il faut ajouter à cela, l'effet des résistances de charge et les capacités des interconnexions. En fait, les performances fréquentielles des MOSFET sont élevées et assez peu différentes de celles des transistors bipolaires dans les technologies avancées. Des fréquences de coupure de plusieurs centaines de GHz sont données dans la littérature pour le nœud 90 nm. Ce n'est donc pas le comportement fréquentiel qui permet de choisir entre bipolaire et MOS pour les applications haute fréquence. Le choix se fait de manière plus rigoureuse à partir des considérations de bruit.

Les résultats du chapitre 7 montrent que les performances en bruit d'un étage amplificateur dépendent fondamentalement de la transconductance g_m du transistor utilisé, MOS ou bipolaire. Plus cette

valeur est élevée et plus le bruit ramené en entrée est faible. Le bon critère de choix est donc la comparaison de la transconductance pour un courant de polarisation donné. En effet, le compromis sera à faire entre performance en bruit et consommation.

Les transconductances du transistor MOS et du transistor bipolaire sont données par les relations suivantes :

Pour le MOSFET en saturation :

$$g_m = \frac{2 I_D}{V_{GS} - V_T} \quad \text{si} \quad V_{DS} > V_{DS\text{sat}}$$

Pour le transistor bipolaire :

$$g_m = \frac{I_E}{\phi_t}$$

Le ratio entre les deux transconductances est donc pour un même courant de polarisation ($I_E = I_D$).

$$\frac{g_{m\text{MOS}}}{g_{m\text{BIP}}} = \frac{2 \phi_t}{V_{GS} - V_T}$$

La fréquence de coupure du MOS devant être élevée dans les applications haute fréquence, une valeur minimale de la différence $V_{GS} - V_T$ est nécessaire, disons 500 mV pour fixer les idées. Le ratio est donc d'un facteur 10 environ. Les transistors bipolaires ont donc à courant de polarisation égal une transconductance 10 fois plus forte qu'un MOS.

De plus, les densités spectrales des bruits basse fréquence des transistors bipolaires sont notablement plus faibles que celles des transistors MOS comme il le sera étudié dans le chapitre 7. Cela peut avoir des conséquences importantes, en radio par exemple, si on pense au bruit de phase des oscillateurs commandés en tension. En résumé, les transistors bipolaires offrent un meilleur compromis bruit-consommation que les transistors MOS ce qui explique leur utilisation dans des applications exigeantes en termes de bruit et de consommation. Les circuits des téléphones portables en sont un exemple. D'autres cas peuvent être cités dans le domaine de la conversion analogique-numérique ultra rapide.

Chapitre 6

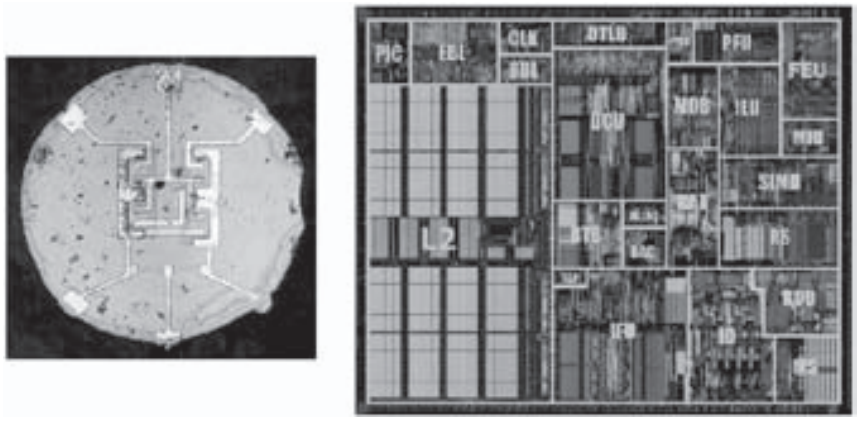
La fabrication collective des circuits intégrés

Le but de ce chapitre est d'expliquer de manière très simplifiée les procédés utilisés classiquement en micro-électronique et d'indiquer les évolutions possibles. L'industrie de la micro-électronique est arrivée à un degré de maturité élevé avec la fabrication de circuits intégrés comportant des millions de transistors de tailles submicrométriques tout en garantissant des rendements de fabrication supérieurs à 80 %.

La réduction de taille du transistor avec des longueurs de grille attendues très inférieures au micron conduit à une complexité croissante des équipements de fabrication et à des exigences draconiennes en terme de propreté et de contrôle des impuretés. Ces contraintes entraînent une croissance exponentielle du coût des équipements. De nouvelles techniques sont donc examinées pour limiter cette explosion des coûts, comme les techniques de nanolithographie et les techniques d'auto-assemblage.

6.1 Les principes généraux de fabrication des circuits intégrés

Les circuits intégrés sont des circuits électroniques présentant la particularité de rassembler sur un morceau de silicium la totalité des composants nécessaires. Ces composants sont des transistors, MOSFET et bipolaires mais aussi des composants passifs comme les résistances, les inductances et les condensateurs. N'oublions pas les fils d'interconnexions qui jouent un rôle majeur dans la technologie micro-électronique. En fait, il est souvent impossible d'intégrer tous les composants dans un seul morceau de silicium en particulier les condensateurs de fortes valeurs et les inductances de qualité. Certains composants seront donc placés à l'extérieur du circuit. Ce nombre doit être le plus faible possible pour réduire les coûts de connexion. La réalisation de résistances de valeurs élevées ou de résistances de précision est également difficile comme nous le verrons par la suite.



Le premier circuit intégré en 1961 **Le pentium 4**

Figure 6.1 – Évolution de la complexité des circuits intégrés.

La *figure 6.1* montre à partir de deux exemples la remarquable évolution des circuits intégrés en 40 ans. Le circuit intégré de 1961 ne comporte que quelques transistors. Le Pentium 4 comporte des millions de transistors. Ces deux circuits si différents ont cependant des points communs : ils sont en silicium et ils sont fabriqués collectivement. Remarquons que la taille du circuit intégré, appelé « puce » ou « chip », n'a pas augmenté en proportion de l'augmentation du nombre de transistors. Les circuits intégrés complexes actuels ont une surface de l'ordre du cm^2 et cette surface ne devrait pas augmenter de manière significative dans les années à venir. La raison à cela est la nécessité d'obtenir un excellent rendement de fabrication.

La fabrication collective des transistors signifie que plusieurs circuits identiques sont fabriqués simultanément. La fabrication des circuits intégrés utilise comme matériau de base un disque de silicium de 0,5 à 1 mm d'épaisseur environ appelé « wafer ». Le diamètre de ce disque est maintenant de 200 mm et parfois de 300 mm. Il est donc possible de fabriquer simultanément environ 1 000 circuits de 1 cm^2 de surface. Le diamètre des wafers n'a cessé de croître dans l'histoire de la micro-électronique et des diamètres de 450 mm sont attendus dans l'avenir.

La fabrication collective d'un grand nombre de circuits en même temps explique l'intérêt de ce type de technologie pour l'électronique et la réduction du coût de production des systèmes électroniques

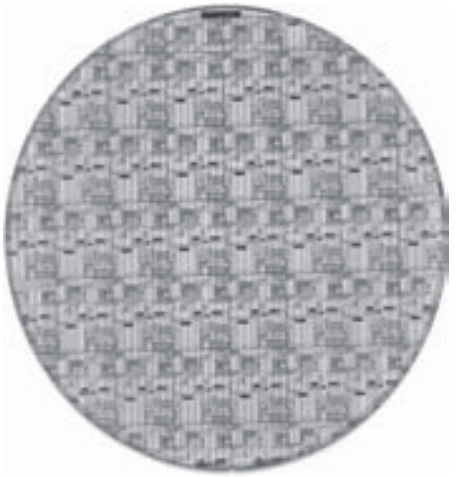


Figure 6.2 – Wafer 300 mm.

malgré la croissance de la complexité. Les paragraphes qui vont suivre expliquent comment cette fabrication est possible et quels sont les procédés utilisés.

6.2 Les procédés de base

6.2.1 *La lithographie*

Pour fabriquer des transistors à des endroits précis sur le wafer, il faut construire des régions présentant des propriétés différentes dans le silicium. Certaines sont fortement dopées, d'autres faiblement dopées et d'autres isolantes. De plus, il faudra réaliser des connexions conductrices entre zones. Ces régions seront réalisées sur le support silicium de base en ajoutant des dopants, en déposant du métal (aluminium ou cuivre) ou de l'oxyde. Dans tous les cas l'opération devra être strictement limitée aux régions choisies sur le wafer. La lithographie permet de faire ce choix.

Imaginons le cas simple de la fabrication d'une zone dopée *n* sur un wafer dopé *p*. La *figure 6.3* représente les étapes.

Reprenons les étapes du procédé. Dans une première phase, une résine est déposée sur le wafer. Dans une seconde étape, un rayonnement est envoyé sur la résine avec une modulation spatiale représentant le motif à réaliser. La façon de réaliser cette modulation sera vue dans la suite. Dans les procédés industriels, le rayonnement est généralement de la lumière et principalement des ultraviolets. Les régions exposées de la résine voient une modification de leurs propriétés chimiques. Elles seront par exemple plus résistantes à l'action d'un solvant. La quatrième étape est la dissolution des régions non exposées de la résine. L'inverse est également possible et dans ce cas, les régions insolées sont dissoutes plus facilement. Dans une cinquième étape, on fait diffuser les ions devant être inclus pour doper le silicium (du phosphore pour un dopage de type *n*) dans les régions non recouvertes de résine. Les techniques permettant cette opération seront étudiées ultérieurement. Enfin, et c'est la sixième étape, on enlève la résine pour obtenir le wafer structuré demandé. Les régions grisées sont dopées *n* et peuvent, par exemple, constituer les puits des transistors PMOS.

Il nous faut maintenant étudier comment moduler la lumière. Deux méthodes sont utilisées : les méthodes parallèles et les méthodes séquentielles. Le principe des méthodes séquentielles est de

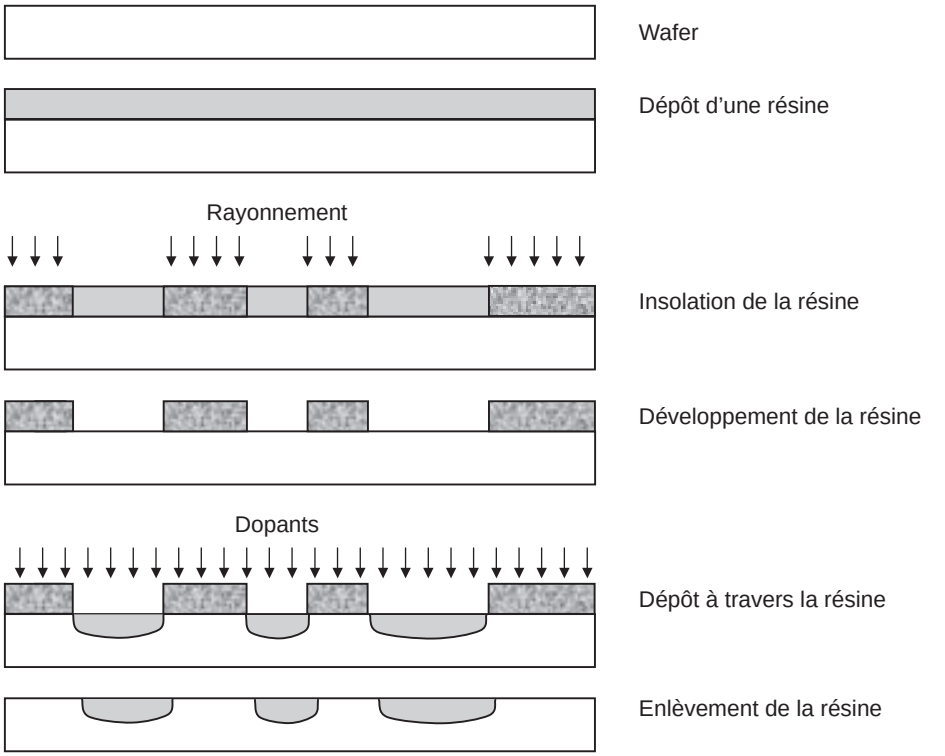


Figure 6.3. Principe de la fabrication par lithographie.

graver tous les motifs en même temps en se servant d'un masque. Le masque est un objet interposé entre la source de lumière et le wafer qui laisse passer la lumière de manière sélective. Le principe des méthodes séquentielles est d'inscrire point par point le motif à reproduire en guidant le faisceau lumineux. Les deux méthodes sont illustrées *figure 6.4*. Quand le faisceau est un faisceau d'électrons et non pas de lumière, la lithographie est dite à faisceau d'électrons et est appelée lithographie « e-beam ».

Dans la lithographie parallèle, un masque interposé entre la source de lumière et le wafer arrête les rayons lumineux aux endroits définis par le design. Il est constitué d'une plaque de quartz transparente aux photons recouverte d'une couche de chrome gravée en fonction des motifs à réaliser. Ce masque très précis est en général fabriqué par une méthode de gravure faisant usage de la lithographie à faisceau d'électrons permettant des précisions bien inférieures au micron. Cela explique cependant le coût élevé d'un jeu de masques.

Dans sa réalisation la plus simple, la lithographie par contact consiste à placer le masque en contact avec le wafer comme le montre la *figure 6.5*.

Cette technique très simple a été utilisée jusqu'au milieu des années 70. Elle est limitée par la diffraction de Fresnel qui prédit que le motif le plus fin ($a_{\min}$) réalisable est donné par :

$$a_{\min} = \frac{3}{2}\sqrt{\lambda\left(s + \frac{e}{2}\right)} \qquad (6.1)$$

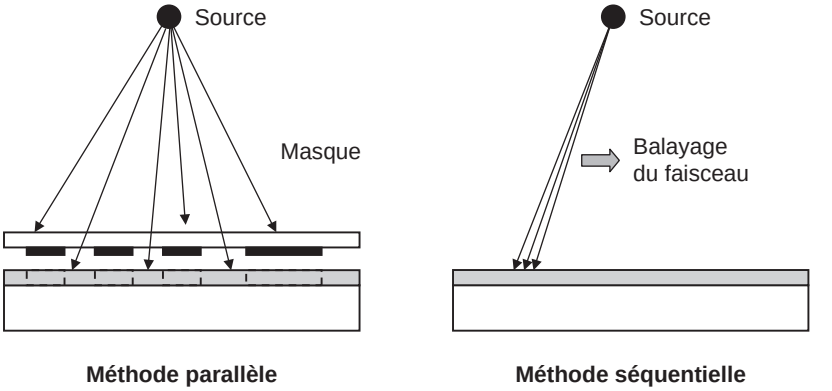


Figure 6.4 – Méthodes parallèle et séquentielle.

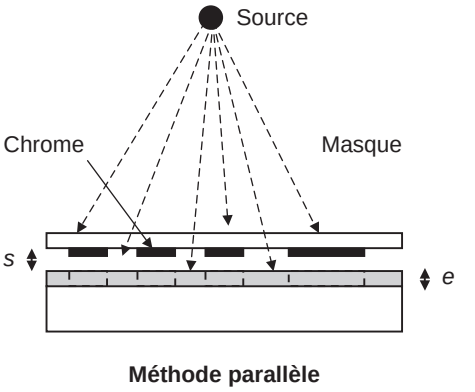


Figure 6.5 – Lithographie par contact.

Dans cette relation, λ est la longueur d'onde de la lumière utilisée pour insoler et *e* est l'épaisseur de la résine. La distance *s* entre le wafer et le masque est minimisée mais non nulle.

Avec une longueur d'onde de 400 nm et une distance de 10 microns entre wafer et masque, on atteint une résolution de 3 microns environ. Cette technique est limitée fondamentalement par la contrainte de planéité du wafer ce qui fixe la distance *s* minimale à 10 microns. La lithographie par contact n'est plus utilisée de nos jours de manière importante car sa résolution est insuffisante.

La lithographie par projection s'est imposée au fil du temps pour résoudre les problèmes de planéité évoqués précédemment. Elle consiste à interposer un objectif photographique entre le masque et le wafer comme le montre la *figure 6.6*. Elle est très contraignante pour l'optique qui doit être de grande qualité.

On considère alors le masque comme une source étendue. L'image de cette source est formée sur le wafer par l'optique avec un facteur de réduction dépendant de la distance focale de l'optique. La séparation minimale de deux objets est donnée par la relation classique.

$$a_{\min} = \frac{k\lambda}{n \sin i} \tag{6.2}$$

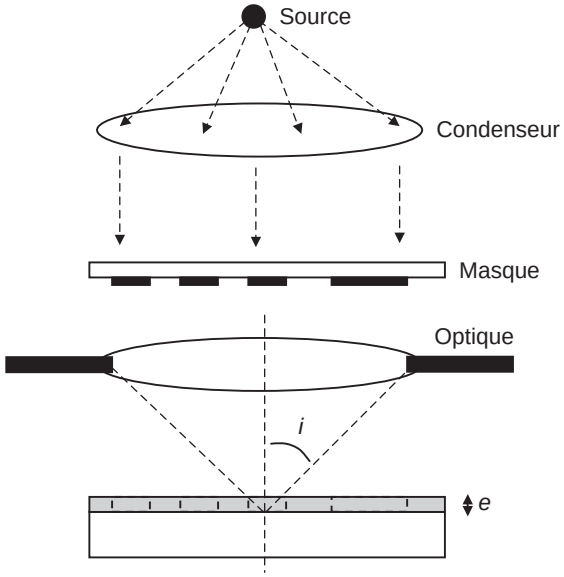


Figure 6.6 – Lithographie par projection.

Dans cette relation de base, λ est la longueur d'onde de la lumière. L'indice du milieu est n et i est l'angle maximum de collection de la lumière défini sur la *figure 6.6*. Le coefficient k a une valeur théorique de 0,61 mais est en pratique égal à 0,8 pour des installations classiques. Le produit $(n \cdot \sin i)$, appelé ouverture numérique, est passé de 0,3 à 0,9 avec des progrès constants dans la conception des optiques. Pour améliorer la précision de la lithographie, il est donc possible de diminuer la longueur d'onde. C'est le sens de la lithographie UV et à plus long terme de la lithographie à base de rayons X. Il est également possible avec une source UV d'utiliser des techniques plus sophistiquées comme les masques à contraste de phase ou la lithographie à immersion.

Il ne faudrait pas croire qu'un wafer entier puisse être insolé de cette manière. La zone utile appelée champ est bien plus petite que la surface du wafer. La solution généralement mise en œuvre est d'insoler une zone du wafer correspondant au champ puis de déplacer le wafer pour insoler une autre zone et ainsi de suite. Cette technique permet d'appliquer un facteur de grandissement G variant entre 5 et 20. Le motif sur le masque peut être G fois plus grand que le motif gravé sur le wafer. Le masque est donc plus facile à fabriquer que s'il était à la même échelle. Si le facteur de grandissement est élevé, il faut déplacer le wafer un grand nombre de fois pour une insolation totale du wafer. Le temps d'insolation sera donc plus important. Il y a un choix optimal à faire en fonction de la résolution nécessaire et du coût de l'opération.

En réalité, quand on fabrique un circuit on utilise non pas un masque mais un jeu de masques car les opérations à effectuer sont nombreuses. Le coût d'un jeu de masques dans une technologie avancée est un véritable problème pour l'industrie micro-électronique. Dans une technologie 50 nm des coûts de 4 millions d'euros sont annoncés pour réaliser un circuit intégré. Ce constat conduit à des évolutions importantes dans l'architecture des circuits comme il le sera étudié dans la suite de cet ouvrage.

6.2.2 *Les procédés de retrait de matériaux*

Ils permettent d'enlever de la matière dans des zones définies par la lithographie. Trois procédés sont possibles : la gravure humide, la gravure sèche et la gravure ionique réactive.

✦ La gravure humide

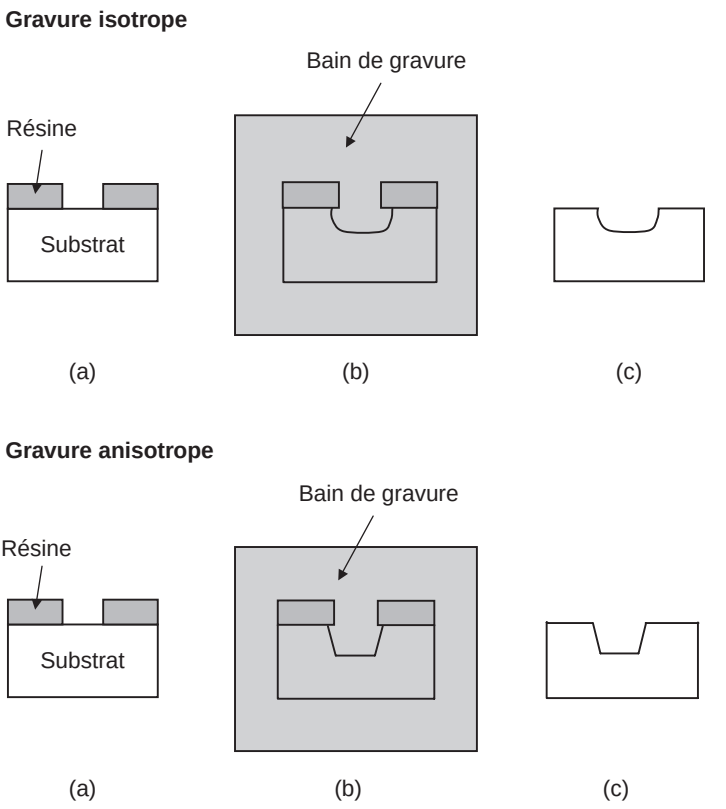


Figure 6.7 – Les procédés de gravure humide.

Prenons l'exemple de la gravure d'une tranchée dans le silicium. Les autres zones sont protégées par la résine. Le substrat est trempé dans un bain chimique attaquant le silicium mais pas la résine. Une zone gravée se forme d'autant plus profonde que le temps d'immersion est élevé. Des vitesses de gravure de plusieurs microns par minute sont possibles en fonction de la concentration de la solution d'attaque. La *figure 6.7* illustre ce procédé de base. La gravure isotrope est un procédé simple mais présente l'inconvénient de la sous-gravure c'est-à-dire la gravure sous la résine. Ce procédé est donc peu adapté à l'obtention de motifs fins et à l'obtention de flancs de gravure raides.

L'orientation du réseau cristallin change les vitesses de gravure. Les plans les plus denses du silicium sont par exemple gravés beaucoup plus lentement que les autres régions. Il est donc possible en orientant convenablement le wafer de réaliser des gravures non isotropes comme le montre la *figure 6.7*.

On peut graver le silicium mais aussi les oxydes et les métaux comme le montre le *tableau 6.1*.

Tableau 6.1

Matériau à graver	Agent de gravure
Silicium	Soude (KOH)
Dioxyde de silicium	Acide fluoridrique (HF)
résine	$H_2SO_4 + H_2O_2$

✦ La gravure sèche

Le principe n'est plus de contrôler une réaction chimique dans un bain mais de bombarder les zones non protégées de la surface avec des ions. L'opération se passe dans une enceinte sous vide comme le montre la *figure 6.8*. Trois techniques sont alors possibles : la gravure dite par *sputtering*, la gravure par plasma, la gravure ionique réactive.

Ces trois techniques font usage de deux principes physiques : le premier est l'arrachement d'un atome de la surface par collision élastique avec un ion incident (*sputtering*), le second est l'activité chimique en surface du wafer quand un plasma est créé dans l'enceinte à vide. Ce plasma est créé par un champ radiofréquence à 13,56 MHz.

La gravure par *sputtering* utilise le premier effet ; elle est purement mécanique. La gravure par plasma utilise le deuxième effet ; elle est purement chimique. La gravure ionique réactive combine les deux effets et permet ainsi de décaper la surface avec une grande efficacité. Elle est donc très largement mise en œuvre dans les procédés actuels de fabrication. Elle permet également d'atteindre d'excellentes résolutions de gravure (quelques dizaines de nm) et de réaliser des flancs quasi-verticaux. Les trois méthodes sont illustrées *figure 6.8*.

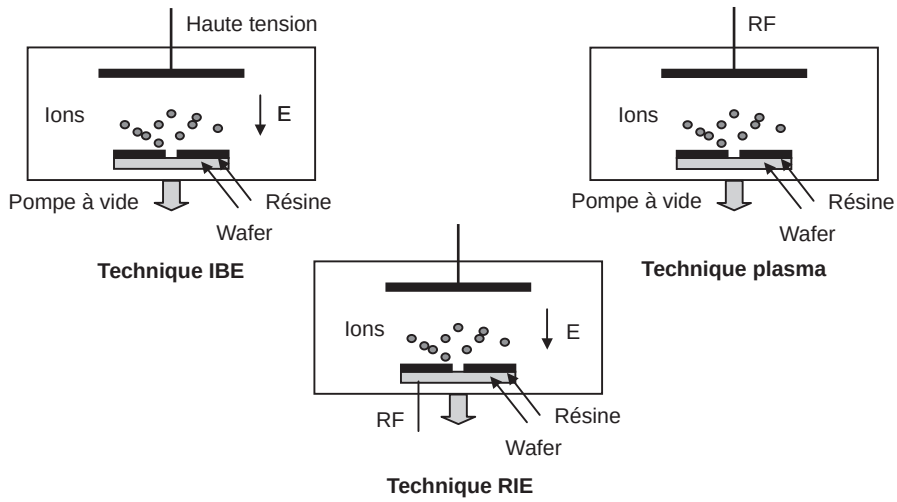


Figure 6.8 – Les trois méthodes de gravure sèche.

✦ Le polissage mécanico-chimique (CMP)

C'est un procédé inspiré du polissage des miroirs optiques qui permet d'obtenir une surface de silicium de planéité parfaite et présentant un état de surface proche de la surface polie d'un miroir. Il associe une attaque chimique et un polissage mécanique.

6.2.3 *Les procédés d'apport de matériaux*

Il s'agit non plus d'enlever un matériau existant mais d'amener dans des régions définies par la lithographie les matériaux nécessaires à la fabrication des composants. Les matériaux sont du silicium, des isolants (dioxyde de silicium ou nitrure de silicium) et des métaux (aluminium, cuivre, cobalt, titane et tungstène). Ce sont aussi des ions (Bore, arsenic et phosphore) qui sont ajoutés au silicium pour modifier le dopage.

Les procédés sont assez nombreux : implantation ionique, PVD, CVD, croissance thermique, croissance électrolytique. Ils sont classés dans la *figure 6.9* selon trois types d'apport. Les procédés de dépôt consistent à amener sur le wafer des matériaux existants. Les procédés de croissance consistent à déclencher la croissance du matériau à partir de ses constituants chimiques. Enfin, les procédés d'apport en profondeur consistent à introduire des atomes dans le wafer pour changer les propriétés électriques.

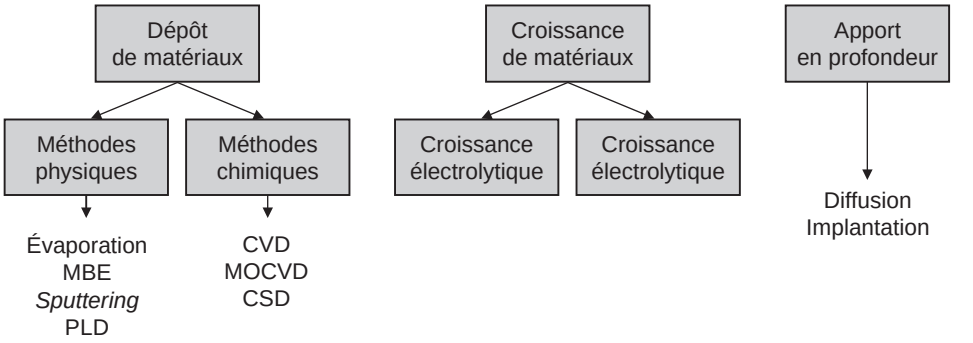


Figure 6.9. Les procédés d'apport de matériaux.

✦ Évaporation

C'est un procédé très simple qui consiste à chauffer un matériau dans un creuset, à le transformer en phase vapeur puis à provoquer la croissance du même matériau en phase solide sur les zones non protégées du wafer. Ce sont les métaux qui sont principalement déposés par cette technique car le point de fusion est relativement bas. Le matériau peut être chauffé par une résistance mais aussi par un faisceau d'électrons quand il est nécessaire d'atteindre des températures plus importantes. Les taux de déposition sont élevés, jusqu'à 5 000 nm par minute.

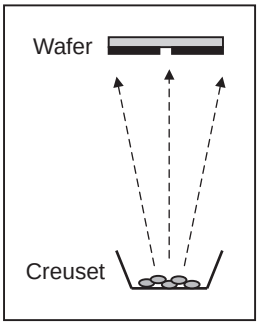


Figure 6.10 – Évaporation.

✦ Épitaxie par jets moléculaires (MBE)

C'est une amélioration de la technique d'évaporation qui bénéficie des techniques d'ultra vide. Différentes sources sont présentes dans l'enceinte sous vide. Des obturateurs rapides sont placés devant chaque source et des équipements de caractérisation sont en général ajoutés pour contrôler les dépôts. Un équipement type est symbolisé *figure 6.11*.

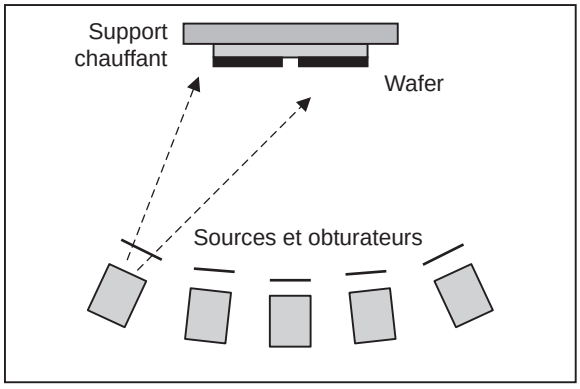


Figure 6.11 – Épitaxie par jets moléculaires.

Cette technique permet en particulier de réaliser des dépôts multicouches.

✦ Dépôt par *sputtering*

L'équipement est le même que celui de la gravure par *sputtering* à la différence près que l'anode est le wafer et que la cathode est une cible faite dans le matériau qu'il faut déposer. Un gaz inerte comme l'argon est ionisé et les ions sont accélérés vers la cible. Des atomes de la cible sont éjectés et vont se déposer sur le wafer.

Le *sputtering* RF peut également être utilisé pour déposer des isolants.

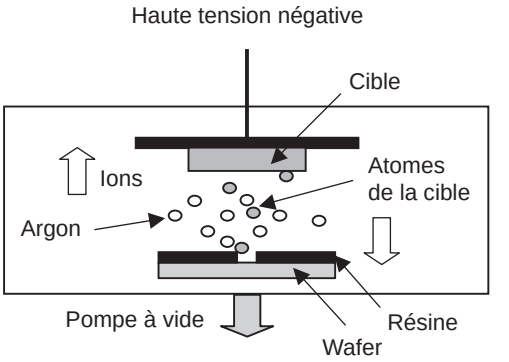


Figure 6.12 – Dépôt par sputtering.

✦ Dépôt par laser pulsé (PLD)

Cette technique plus récente est mise en œuvre pour déposer des isolants en multicouches. Un laser pulsé de puissance (un Joule par impulsion environ) forme un plasma au niveau de la cible. Ce plasma contient des atomes, des ions et des molécules de la cible. Ces composés se déposent sur le wafer. Le dispositif est représenté *figure 6.13*.

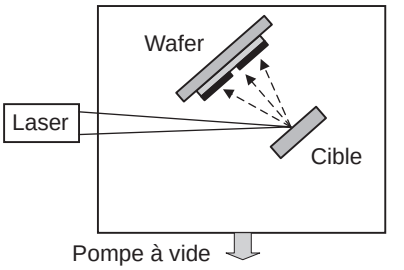


Figure 6.13 – Dépôt par ablation laser.

✦ Le dépôt en phase vapeur (CVD)

Le principe est de faire croître sur un substrat une couche relativement mince à partir de composants en phase vapeur appelés précurseurs. Le substrat est chauffé dans un dispositif comme celui de la *figure 6.14*.

Différentes techniques sont possibles :

- pression atmosphérique : APCVD ;
- basse pression : LPCVD ;
- haute pression : HPCVD.

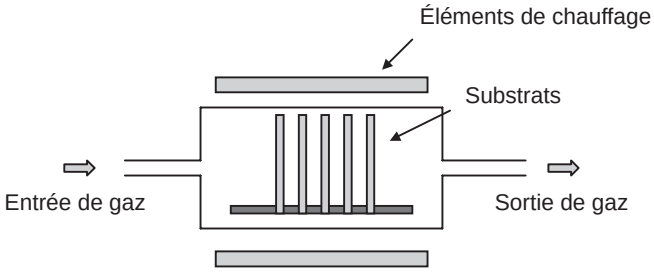


Figure 6.14 – Réacteur CVD.

Le dépôt à basse pression se fait à plus haute température. Le dépôt à haute pression peut se faire à plus basse température. Il est possible de déposer des semi-conducteurs et des isolants en exploitant les réactions chimiques suivantes :

- dépôt de silicium polycristallin :

  $SiH_4 \rightarrow Si + 2\,H_2$ pour une pression de 1 mbar

- dépôt de nitrure de silicium :

  $3\,SiH_4 + 4\,NH_3 \rightarrow Si_3N_4 + 12\,H_2$ 700-900 °C à la pression atmosphérique

- dépôt de silice :

  $SiH_4 + O_2 \rightarrow SiO_2 + 2\,H_2$ 450 °C

Ce sont les réactions les plus classiques. Elles utilisent un gaz appelé silane (SiH_4).

✦ Dépôt de type MOCVD

Le dépôt MOCVD (*Metal Organic Chimical Vapour Deposition*) est une évolution de la méthode CVD adaptée au dépôt de composés métalliques. Les précurseurs contiennent le métal à déposer et des composés organiques. Cette technique permet en particulier de déposer les matériaux à haute permittivité nécessaires à la fabrication des DRAM et autres composants de la micro-électronique. Prenons comme exemple le dépôt de PZT (composé à base de plomb, de titane et de zirconium). Les précurseurs sont les liquides suivants : $Pb(C_2H_5)_4$, $Ti(OC_3H_7)_4$, $Zr(OC_4H_9)_4$.

✦ Dépôt de type CSD (*Chemical Solution Deposition*)

Ce sont des méthodes chimiques qui consistent à partir de précurseurs généralement en phase liquide pour arriver à un film polycristallin ou cristallin en passant par une phase amorphe. Il faut également citer les méthodes de type Langmuir-Blodgett pour déposer des composés organiques. Des molécules organiques présentant une extrémité hydrophobe et une autre extrémité hydrophile peuvent être déposées sur un volume d'eau de la même manière qu'un film d'huile se forme au-dessus d'un volume liquide. Le film mince ainsi formé peut alors être transféré sur un substrat. Cette technique simple permet en particulier la fabrication des diodes électroluminescentes organiques et laisse espérer le développement d'une électronique grande surface.

✦ Croissance thermique

Ce sont les méthodes qui permettent en particulier de fabriquer le wafer lui-même. Leur principe a été donné dans le paragraphe 6.2.3 mais la méthode CVD est limitée à la fabrication d'une couche mince. Dans ce paragraphe, on étudie comment réaliser une tranche de 1 mm d'épaisseur. Les wafers sont découpées dans un lingot de silicium. Un lingot est un cylindre de diamètre élevé (300 mm dans les fabrications les plus avancées) qui présente la propriété importante d'être monocristallin. D'autres semi-conducteurs peuvent être fabriqués comme le germanium, l'arséniure de gallium ou le tellurure de cadmium. Le silicium a cependant pris une place largement majoritaire. L'obtention de cristaux monocristallins d'arséniure de gallium ou de tellurure de cadmium reste une opération difficile. Le procédé de croissance du lingot est le procédé Czochralski. Il est représenté de manière simplifiée *figure 6.15*.

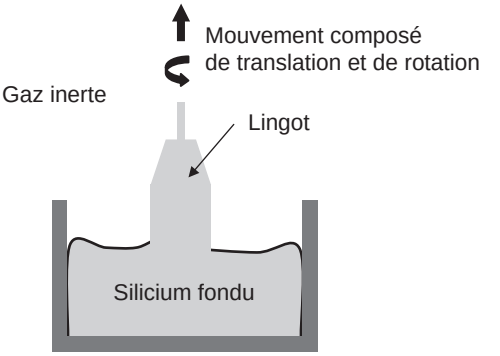


Figure 6.15 – Croissance thermique du silicium.

À partir d'un germe de silicium monocristallin, le silicium fondu apporté dans le creuset se solidifie autour du germe au fur et à mesure que le solide ainsi formé se déplace selon un mouvement combinant rotation autour de l'axe du cylindre et translation vers le haut. Le silicium dans le creuset est formé à partir de la décomposition de la silice, matériau très abondant dans la nature et d'un processus de purification permettant de contrôler les impuretés. Les déplacements du lingot (rotation et translation) sont très lents si bien que la fabrication de lingots de haute pureté avec un contrôle précis de la structure cristallographique reste une opération complexe et lente justifiant le prix relativement élevé des wafers. Le lingot de silicium est ensuite découpé en tranches de 1 mm d'épaisseur environ pour assurer un minimum de rigidité mécanique.

✦ La croissance électrolytique

Cette technique permet d'obtenir des couches relativement épaisses de métal sur un substrat. Prenons l'exemple du dépôt de nickel représenté *figure 6.16*.

La solution est dans ce cas du chlorure de nickel mélangé à du chlorure de potassium. À la cathode, il y a une réaction de réduction :

$$Ni^{2+} + 2\ e^- \rightarrow Ni$$

Au niveau de l'anode, il y a une réaction d'oxydation :

$$2\ Cl^- \rightarrow Cl^2 + 2\ e^-$$

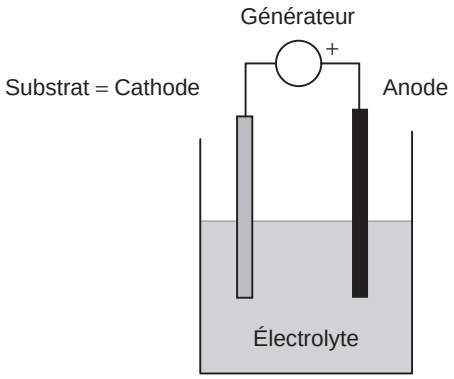


Figure 6.16 – Dépôt par électrolyse.

Il y a donc dépôt de nickel sur le wafer jouant le rôle de cathode. La réaction peut se faire de manière sélective au travers d'ouvertures gravées dans la couche de résine déposée sur le substrat. Si le substrat n'est pas un bon conducteur, il faut déposer par une autre méthode une mince couche de métal qui sert de base de déclenchement de la réaction électrochimique.

✦ Implantation ionique

Nous abordons maintenant les techniques qui permettent de modifier des matériaux en profondeur. La notion de profondeur est toute relative puisque les régions sont créées en général à moins d'un micron de la surface du wafer. La zone active du dispositif est donc située dans une région d'épaisseur négligeable devant l'épaisseur du wafer.

La première technique étudiée est l'implantation ionique. Ce n'est pas réellement une technique permettant de fabriquer en profondeur un matériau mais une technique qui permet de changer les propriétés électriques du matériau de base. Le principe est d'envoyer perpendiculairement à la surface du wafer un flux d'ions de haute énergie. L'énergie de ces ions est suffisamment élevée pour qu'ils pénètrent à l'intérieur de la matière avant d'être arrêtés sous l'effet des interactions avec les électrons des atomes de silicium du wafer. Le principe d'un implanteur ionique est illustré *figure 6.17*.

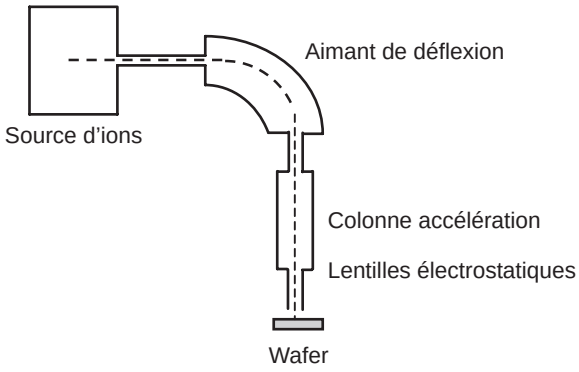


Figure 6.17 – Implantation ionique.

Des atomes sont ionisés dans la source puis extraits vers l'aimant de déflexion. Cet aimant est utilisé en spectromètre de masse. La valeur du champ magnétique créé permet de dévier les ions en fonction de leur masse. Les ions choisis seront les seuls à pouvoir être extraits de cette zone de champ magnétique pour être ensuite accélérés par la colonne d'accélération. L'énergie des ions est donc réglable ce qui permet d'ajuster la profondeur de pénétration dans la matière. Des dispositifs de focalisation et de balayage complètent le dispositif. En résumé, l'implanteur permet les réglages suivants :

- choix des dopants par le réglage du champ magnétique ;
- choix de la pénétration des ions par le réglage de la tension d'accélération ;
- choix de la dose implantée par le réglage de l'intensité du faisceau et du temps d'implantation.

Les énergies d'implantation sont comprises entre quelques keV et quelques MeV. Si on examine le wafer dans sa profondeur après implantation, on obtient une répartition des ions implantés du type de la courbe de la *figure 6.18*.

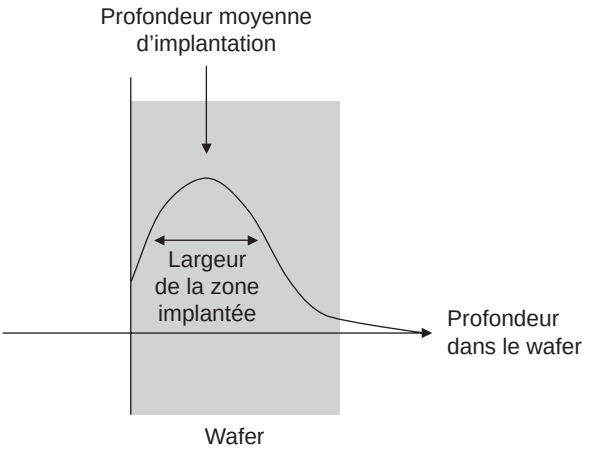


Figure 6.18 – Répartition des dopants après implantation.

La profondeur d'implantation ne peut être définie qu'en moyenne car tous les ions ne subissent pas le même nombre de collisions avec les électrons du wafer. La profondeur moyenne dépend de l'énergie des ions et est inférieure au micron. Pour les faibles énergies, les ions sont implantés au voisinage de la surface du wafer.

Il ne suffit pas d'implanter des ions pour créer en profondeur des zones de dopage donné, il faut que les dopants (Bore, Arsenic, Phosphore) occupent des positions interstitielles dans le réseau cristallin comme il a été expliqué dans le chapitre 2. C'est à cette condition que les dopants sont électriquement actifs. De plus, l'implantation crée de nombreux défauts dans le réseau cristallin en déplaçant les atomes. Il est donc nécessaire pour ces deux raisons de chauffer le substrat implanté pour guérir les défauts et pour rendre les dopants électriquement actifs et modifier les propriétés électriques. Cette opération s'appelle le recuit et se pratique vers 600 °C.

✦ **La diffusion thermique**

La diffusion thermique est une opération complémentaire de l'implantation ionique qui permet de distribuer des dopants dans un substrat. Si le matériau est chauffé à haute température (au-dessus de la température du recuit), les dopants peuvent avoir assez d'énergie pour se déplacer dans la matière en quittant leurs sites initiaux. Ces déplacements dépendent de la température et du temps pendant lequel la haute température est appliquée. La diffusion élargit la distribution initialement créée par l'implantation ionique. Il est également possible de faire diffuser dans le wafer des atomes déposés en surface.

6.2.4 *Les bondings*

En périphérie d'un circuit intégré sont disposés sur le silicium des plots de contact qui permettent de relier le circuit au monde extérieur. Ce sont en général des carrés conducteurs de 100 microns de côté environ. Dans le circuit intégré, ils sont reliés à la couche d'interconnexion la plus haute au sens de l'empilement des couches. À l'extérieur du circuit, ces plots de sortie sont reliés aux plots du boîtier contenant le circuit intégré par des fils de métal qui sont soudés de part et d'autre. La *figure 6.19* montre un circuit intégré et ses fils de raccordement au boîtier.

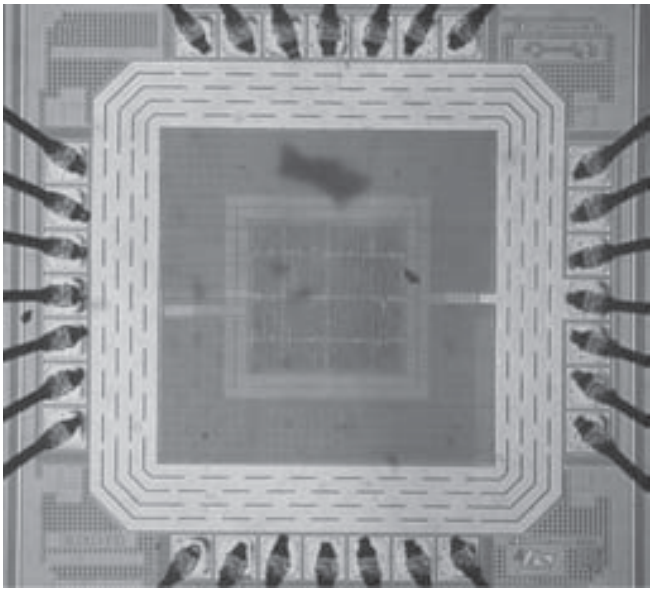


Figure 6.19 – Les bondings (Crédit photo CEA Grenoble).

6.3 Le flot simplifié pour la technologie CMOS

6.3.1 *Les grandes étapes de la fabrication*

Ce paragraphe est fondamental. Il montre comment sont réalisés les circuits CMOS. La description est simplifiée mais est conforme à un procédé industriel réel. Les étapes seront décrites par une série de dessins montrant comment il est possible de réaliser sur un wafer un PMOS, un NMOS et

les interconnexions. Dans la pratique, tous les transistors du circuit sont réalisés en même temps. Les dernières étapes permettent de réaliser les connexions entre transistors ce qui permet de définir les fonctions. Le NMOS sera implanté à droite du dessin et le PMOS à gauche. Dans un premier temps, montrons les grandes étapes de la fabrication sans trop se soucier des modes de réalisation comme cela est représenté sur la *figure 6.20*.

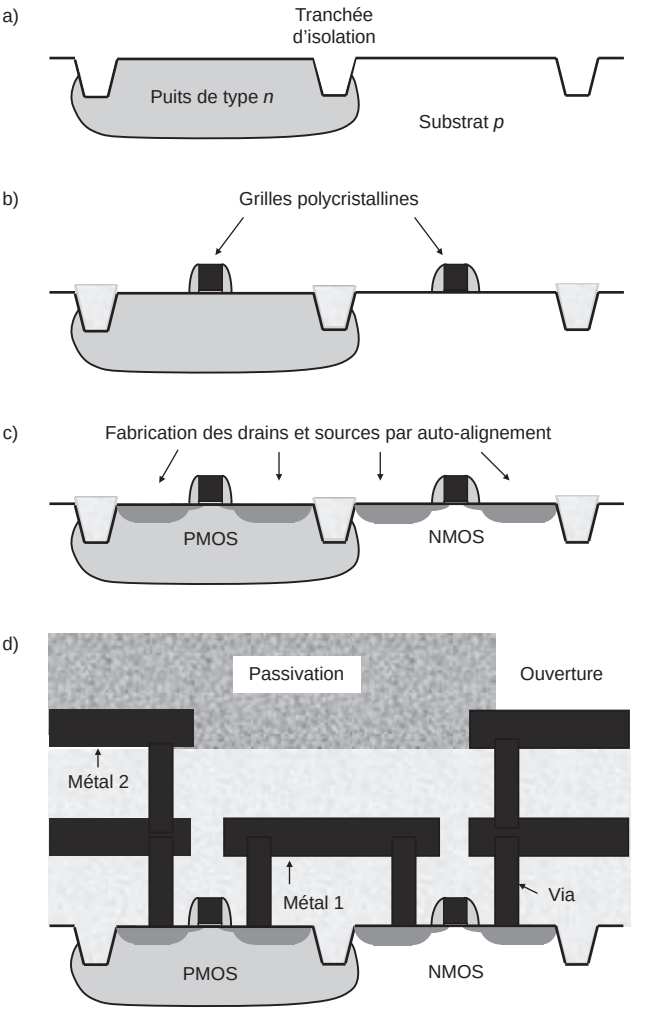


Figure 6.20 – a) Puits et tranchées. b) Isolants de grille et grilles. c) Fabrication des sources et des drains. d) Métallisations et passivation.

Dans une première étape représentée *figure 6.20a*, le substrat de type *p* est découpé puis poli et traité chimiquement pour être prêt à recevoir les traitements de la micro-électronique. Des tranchées d'isolation entre NMOS et PMOS sont gravées dans le matériau puis remplies d'un oxyde de haute qualité. Le but est d'éviter que des courants de fuite puissent circuler. Il y a en fait quatre technologies

possibles : substrat de type *p* et puits de type *n*, substrat de type *n* et puits de type *p*, substrat de type *p* et deux puits *n* et *p*, technologie triple puits. On décrira ici la technologie avec un puits de type *n*.

Dans une seconde étape, les oxydes de grille et les grilles en silicium polycristallin sont réalisés. Notons cependant que des procédés avancés introduisent des grilles métalliques afin de réduire les résistances d'accès.

Dans une troisième étape, les zones de grille et de drain sont réalisées par implantation ionique à travers le silicium polycristallin de grille jouant le rôle de masque. Le procédé est dit auto-aligné et son avantage est de réduire les capacités parasites drain-grille qui ont un effet négatif sur la vitesse de fonctionnement des transistors.

Dans une quatrième étape, on réalise les interconnexions entre transistors ainsi que les contacts traversant pour passer d'un niveau d'interconnexion à un autre. La *figure 6.20c* montre deux niveaux d'interconnexion mais dans les technologies avancées, on peut trouver jusqu'à 7 niveaux. Il ne faut pas oublier la fabrication de l'oxyde entre pistes et l'oxyde de passivation qui protège le circuit des effets chimiques venant de l'extérieur. Certaines zones de cet oxyde de passivation sont ouvertes pour permettre de relier par *bonding* une sortie du circuit à une broche du boîtier. Dans les technologies avancées, les derniers niveaux de métallisation sont utilisés pour les connexions à longue distance et pour réaliser des pseudo-plans de masse. Ils sont constitués d'alliages de cuivre ce qui permet de réduire les résistances des pistes. Cette évolution de la technologie CMOS s'est imposée difficilement à cause des effets chimiques introduits par l'électromigration du cuivre.

Rappelons que tous les transistors d'un circuit et que tous les circuits identiques sont réalisés en même temps sur le wafer ce qui permet d'envisager la fabrication simultanée de milliers de circuits intégrés comportant chacun des millions de transistors.

6.3.2 *Fabrication des tranchées d'isolation et des puits*

La première phase est la fabrication des caissons et des tranchées d'isolation. Avant de décrire en détail les opérations technologiques, il est nécessaire de justifier la réalisation de ces dispositifs d'isolation. Pour cela, imaginons deux transistors MOS en série, un de type *p* et un de type *n* reliés par leurs drains. C'est le schéma de base de la technologie CMOS comme il sera vu dans les chapitres 7 et 8.

Les deux transistors sont polarisés en reliant la source du NMOS au potentiel 0 et la source du PMOS au potentiel V_{DD}, la plus haute tension positive du schéma. Le substrat est relié au potentiel 0 et le puits dopé *n* au potentiel V_{DD} comme c'est souvent le cas. Les deux drains sont reliés. La *figure 6.21* représente le dispositif sans tranchée d'isolation. Les connexions aux potentiels 0 et V_{DD} sont représentées de manière symbolique pour ne pas alourdir le graphique.

La tension des deux drains reliés est intermédiaire entre 0 et V_{DD}. Dans le premier cas, le PMOS et le NMOS sont en fait reliés par deux diodes tête-bêche *np* et *pn* polarisées en inverse. Les surfaces de contact sont telles que les courants inverses, bien que faibles, sont très supérieurs aux courants permanents de conduction tolérés par la logique CMOS.

La réalisation de tranchées conduit à la *figure 6.21b*. Le PMOS et le NMOS sont alors reliés par deux MOS parasites respectivement de type *p* et de type *n* dont les oxydes de grille sont formés par l'oxyde déposé dans la tranchée. Les paramètres de ces MOS seront contrôlés de telle sorte qu'ils soient toujours bloqués, ce qui limite les courants de fuite du dispositif total.

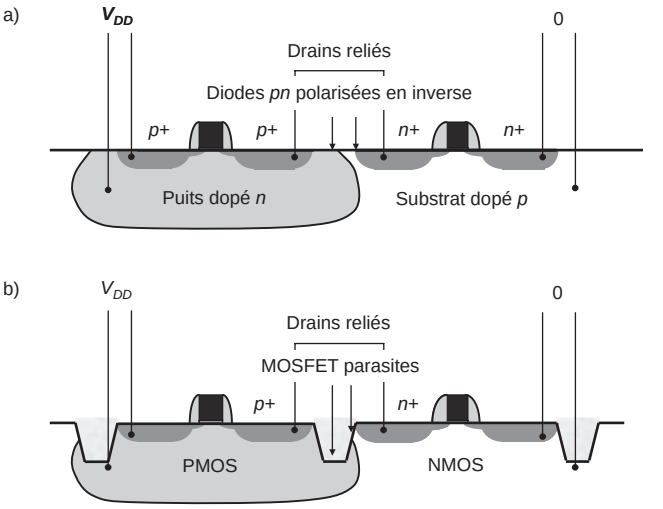


Figure 6.21 – Nécessité des tranchées d'isolation : a) Dispositif sans tranchée d'isolation. b) Dispositif avec tranchée d'isolation.

Voyons maintenant les étapes de réalisation de ces tranchées. Dans une première phase, un mince oxyde est créé à la surface du wafer, environ 100 Å de silice. Le procédé est la CVD.

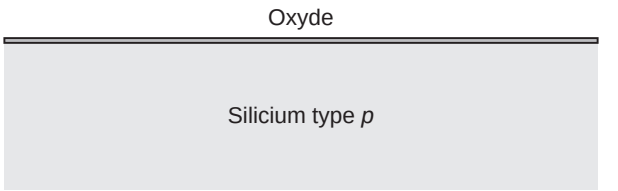


Figure 6.22 – Oxydation du silicium.

La deuxième étape est de faire croître par LPCVD une couche de nitrure de silicium qui sert à la fois de masque et de couche d'arrêt dans un procédé CPM ultérieur.

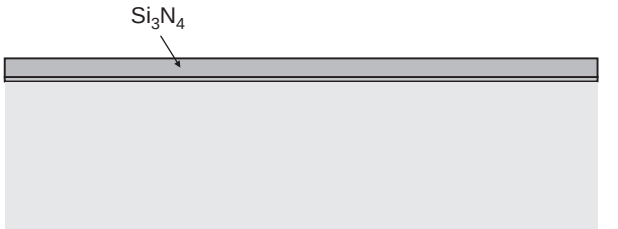


Figure 6.23 – Dépôt du nitrure de silicium.

La troisième étape consiste à graver les tranchées d'isolation entre NMOS et PMOS. Ces tranchées seront ensuite remplies d'oxyde et serviront à isoler les deux transistors. Pour les définir, on utilise un premier jeu de masques et on dépose de manière sélective une épaisseur de résine.

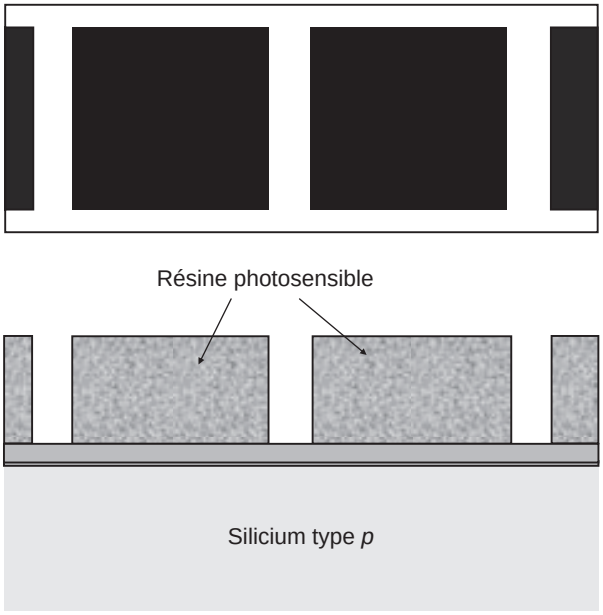


Figure 6.24 – Réalisation des tranchées d'isolation.

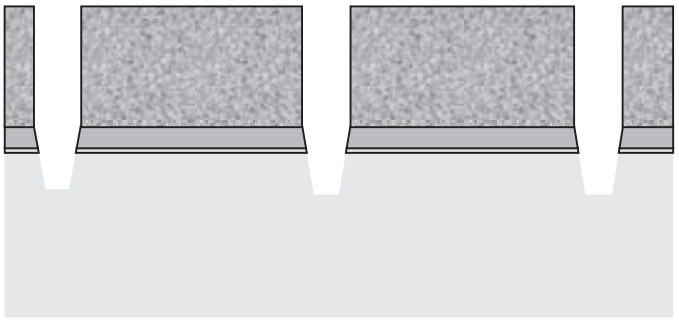


Figure 6.25 – Gravure des tranchées.

La phase suivante consiste à traiter les tranchées pour limiter les courants de fuite. Un oxyde thermique est déposé dans les tranchées après la gravure du silicium. Ensuite, des ions BF_2 sont implantés à travers un masque dans la partie droite de la tranchée et contribuent à limiter les courants de fuite en modifiant la tension de seuil du MOS parasite. L'oxyde de silicium ayant subi l'implantation est endommagé. Il est dissous et remplacé par un oxyde de qualité appelé « liner » déposé par crois-

sance thermique. De manière symétrique, la partie gauche de la tranchée est implantée avec des ions phosphore à travers un masque. Finalement, la tranchée est remplie par un matériau diélectrique en utilisant un dépôt de type HPCVD comme le montre la *figure 6.26*.

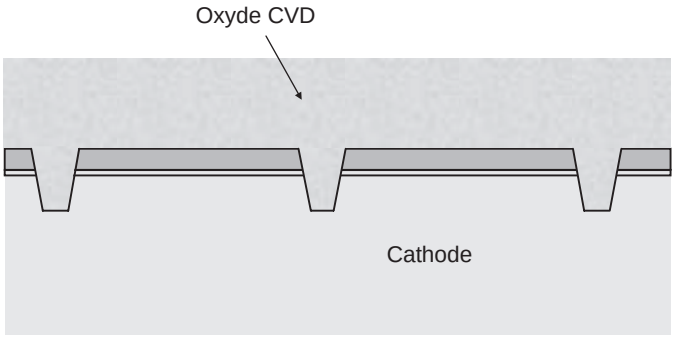


Figure 6.26 – Dépôt par CVD de l'oxyde de tranchée.

La dernière étape est une « planarisation » CMP de l'ensemble. La couche de nitrure sert de couche d'arrêt. Enfin, on enlève par attaque chimique sélective la couche de nitrure. On en arrive donc à la *figure 6.27*. Le résultat semble mince à l'issue de toutes ces étapes mais il faut comprendre que la réalisation de ces tranchées d'isolation est un point capital dans la fabrication des circuits CMOS.

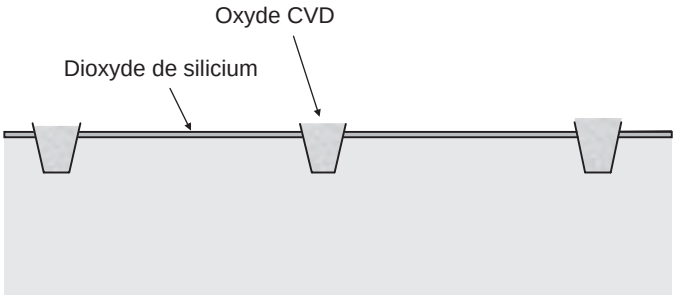


Figure 6.27 – Wafer après la formation des tranchées.

Il nous reste à étudier maintenant comment sont réalisés les puits dans le wafer. Dans certains procédés les deux types de puits *p* et *n* sont implantés. Nous ne considérons ici que les puits de type *n* permettant de réaliser les PMOS.

Il est possible d'utiliser le même masque que celui qui a servi dans les opérations d'implantation pour traiter les tranchées. Ce masque délimite une zone légèrement plus étendue que celle qui correspond à la tranchée comme le montre la *figure 6.28*.

Le profil de dopage voulu est obtenu par des implantations à différentes énergies. Les dimensions des puits dopés *p*, dans le procédé à double puits, sont différentes de celles des puits dopés *n*. Après attaque chimique de la résine, il reste la structure de la *figure 6.29*.

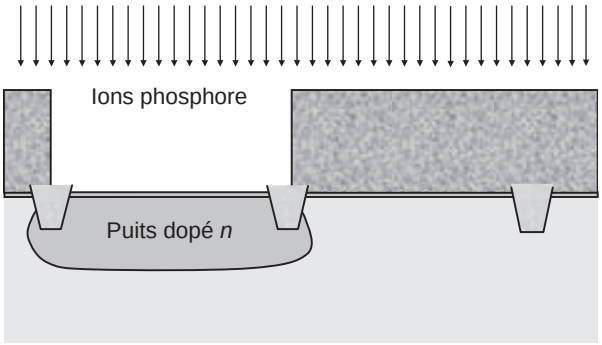


Figure 6.28 – Fabrication des puits.

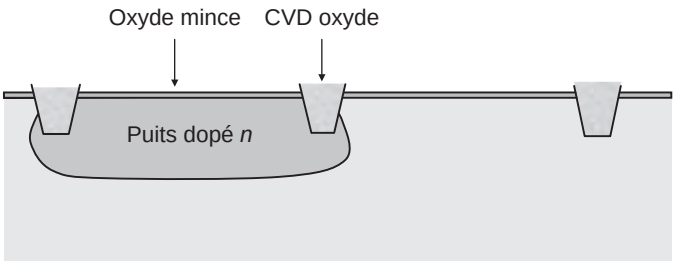


Figure 6.29 – Wafer après fabrication des tranchées et des puits.

Le wafer est donc prêt pour envisager la fabrication des transistors eux-mêmes.

6.3.3 *Fabrication des transistors NMOS et PMOS*

Cette phase appelée *front end* comprend les étapes 2 et 3 du flot décrit en début de chapitre. Le procédé d'auto-alignement conduit à réaliser dans une première étape la grille puis les dopages de source et de drain et non l'inverse. Ce procédé permet donc de minimiser les capacités de recouvrement entre grille et drain et entre grille et source.

Dans un premier temps, l'oxyde résiduel est enlevé sur toute la surface. Un autre dépôt est effectué pour servir de seuil aux implantations ioniques qui vont suivre. Ces implantations effectuées sur toute la surface permettent d'ajuster les tensions de seuil des NMOS et des PMOS en modifiant la résistivité du « body » comme il a été expliqué dans le chapitre 4. On utilise des ions BF_2 et des ions phosphore.

Dans un deuxième temps, on réalise l'oxyde de grille du transistor et le matériau conducteur de grille. Cette réalisation est cruciale pour le fonctionnement du circuit intégré car le fonctionnement du transistor est directement lié aux phénomènes de conduction sous l'oxyde de grille. Des études actuelles préconisent l'emploi d'oxydes ayant une permittivité plus élevée. La croissance de l'oxyde de grille est immédiatement suivie par la croissance du silicium polycristallin déposé par méthode LPCVD. La grille doit être fortement dopée, éventuellement par une implantation ionique supplémentaire.

Le silicium polycristallin déposé sur l'ensemble de la surface est ensuite gravé en dehors des régions protégées par la résine. Le masque utilisé dans cette étape est critique car il délimite la longueur du canal du transistor. Les *figures 6.30 et 6.31* illustrent cette opération.

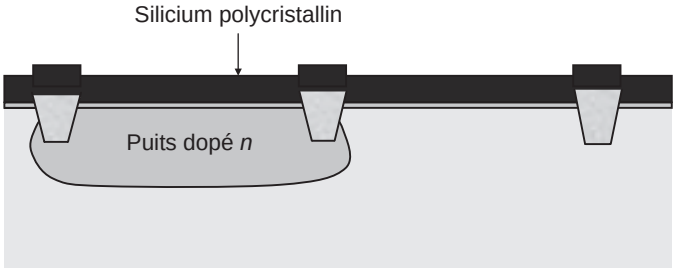


Figure 6.30 – Dépôt du matériau de grille.

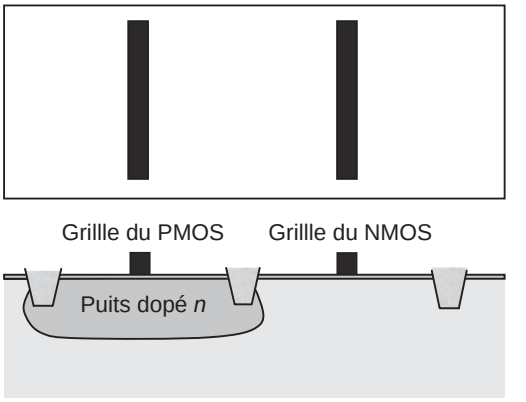


Figure 6.31 – Délimitation des canaux des transistors.

Ensuite le wafer est à nouveau oxydé sur toute sa surface. Il faut maintenant réaliser source et drain par implantation ionique.

Dans un premier temps, les régions faiblement dopées (zones dites LDD) sont réalisées par implantation d'ions à basse énergie (Bore pour les zones dopées *p* et Phosphore pour les zones dopées *n*). Ces zones à proximité du canal permettent de réduire le champ électrique aux jonctions source-canal et drain-canal. L'injection d'électrons de haute énergie dans la grille est alors minimisée. Notons encore une fois l'intérêt de l'auto-alignement. Le silicium polycristallin des grilles sert de masque pour les ions implantés. Deux jeux de masques sont nécessaires pour cette opération : l'un pour sélectionner les PMOS et l'autre pour sélectionner les NMOS.

L'étape suivante consiste à réaliser les « espaceurs » de part et d'autre des grilles. La technique LPCVD est utilisée pour faire croître un oxyde, en général du nitrure de silicium, à une température d'environ 800 °C. Une gravure anisotrope de cet oxyde conduit ensuite au dispositif de la *figure 6.32*. Notons que ces espaceurs serviront de masque pour les implantations à plus forte dose des régions de source et de drain.

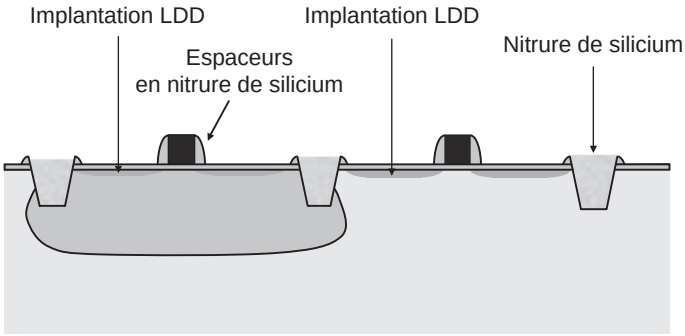


Figure 6.32 – Fabrication des espaceurs et des LDD.

La dernière étape est de finaliser la création des zones de source et de drain par implantation d'ions BF2 pour les zones *p* et d'Arsenic pour les régions *n*. On peut utiliser le même jeu de masques que pour les implantations LDD, les espaceurs font office de masques.

La dernière opération est un recuit à haute température pour rendre les zones implantées électriquement actives. On comprend alors la raison pour laquelle l'industrie micro-électronique a réalisé les grilles en silicium et pas en métal. Si les grilles étaient en métal, il y aurait un problème difficile à résoudre : la diffusion du métal à la température de recuit. À ce stade de la fabrication, les transistors PMOS et NMOS sont réalisés comme le montre la *figure 6.33*.

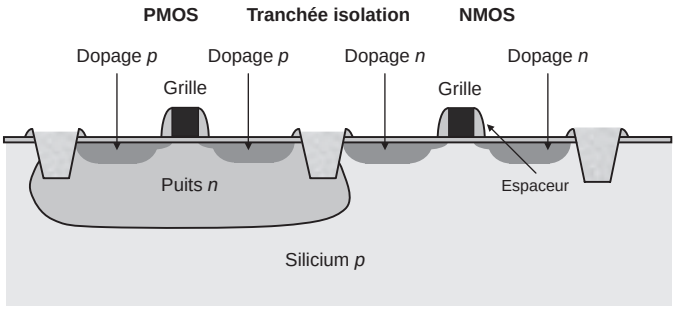


Figure 6.33 – Silicium à la fin du front end.

6.3.4 *Fabrication des interconnexions (back end)*

La dernière phase est de réaliser les interconnexions. Cette phase est appelée *back end*. Les couches métalliques sont des composés de titane, d'aluminium et de nickel. Les technologies avancées font également usage du cuivre bien connu pour ses bonnes propriétés de conduction. Le cuivre, considéré comme un dopant dangereux en micro-électronique, a été introduit récemment dans le flot de fabrication. On utilise également un procédé auto aligné pour réaliser les contacts. Le principe général est de déposer un métal sur l'ensemble du wafer. Les zones recouvertes d'oxyde ne formeront pas un matériau conducteur après les traitements thermiques. Par contre, les zones de silicium polycristallin ou de silicium dopé formeront avec le métal déposé des régions conductrices aptes à

réaliser des contacts électriques. Le procédé est dit auto-aligné car il n'y a pas besoin de masques supplémentaires.

Les opérations se déroulent alors de la manière suivante. Le mince oxyde résiduel à la surface du wafer est enlevé par une attaque chimique d'acide fluorhydrique. Le métal (titane ou cobalt) est déposé par *sputtering*. Un recuit basse température permet la formation de $TiSi_2$. Ce « siliciure » n'est pas encore conducteur et un deuxième recuit est nécessaire après avoir éliminé le surplus de métal déposé. Les espaceurs et les tranchées d'isolement sont protégés contre la formation de cette couche conductrice.

La base conductrice étant constituée, il est possible de réaliser la première couche de connexion permettant par exemple de relier drain du NMOS et drain du PMOS dans un inverseur. Les sources pourront être reliées à d'autres transistors par des connexions réalisées dans le même niveau mais seront assez souvent reliées à la masse ou à la tension d'alimentation. Dans ce cas, le deuxième niveau d'interconnexion sera mis à contribution. On considère dans cet exemple simple que deux niveaux suffisent. En pratique, d'autres niveaux sont nécessaires à cause de la faible taille des transistors car il faut donner aux connexions une largeur minimale pour conserver des liaisons faiblement résistives. Les technologies avancées utilisent parfois 7 niveaux d'interconnexion.

L'étape suivante consiste à déposer une fine couche de nitrure de silicium qui servira de couche d'arrêt dans une opération de gravure ultérieure. Ensuite, une couche épaisse d'oxyde de silicium (oxyde PMD) est déposée par une méthode plasma phase vapeur. La surface obtenue n'étant pas plane, une phase de polissage (CPM) est nécessaire. Le résultat final est représenté *figure 6.34*.

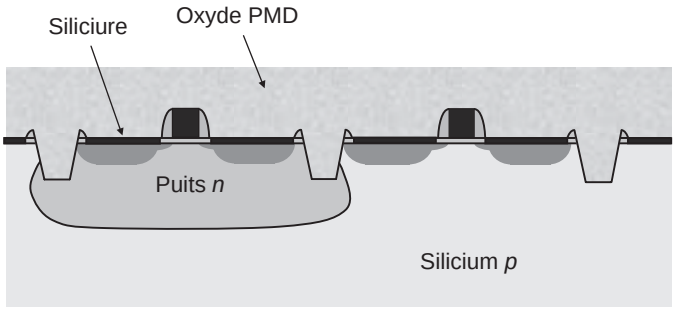


Figure 6.34 – Wafer oxydé.

L'opération suivante est la fabrication des contacts traversants entre les transistors et le premier niveau d'interconnexion. Ces contacts sont appelés « vias » dans la technologie micro-électronique. La lithographie permet de définir des ouvertures comme le montre la *figure 6.35*. La couche de nitrure sert de couche d'arrêt à la gravure de l'oxyde PMD.

La métallisation des ouvertures se fait alors en trois temps : dépôt d'une mince couche de titane par *sputtering* après une phase de nettoyage par des ions argon, dépôt d'une couche mince de *TiN* pour protéger le silicium dans la phase suivante, remplissage des ouvertures par du tungstène en utilisant un *procédé* CVD à base de WF_6. Le tungstène « déborde » des ouvertures et un procédé de polissage permet d'obtenir le dispositif représenté *figure 6.36*.

Dans l'étape suivante, une couche conductrice est formée au-dessus des vias. Elle sera gravée dans une étape ultérieure car les sources et les drains ne sont pas tous reliés. Fabriquée par *sputtering*,

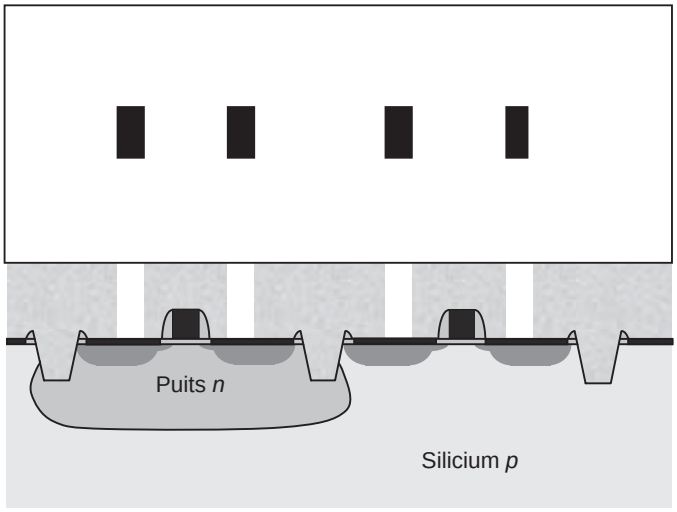


Figure 6.35 – Fabrication des vias par lithographie.

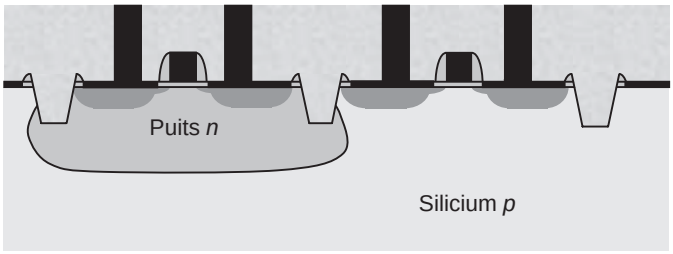


Figure 6.36 – Dispositif avec vias.

cette couche est un empilement *Ti/TiN/Al/TiN*. Cet empilement assure à la fois l'adhérence, la conductivité et limite l'électromigration. La couche de *TiN* est une couche d'arrêt pour les attaques chimiques qui suivent. Elle fait aussi fonction de couche optique. Cette couche métallique est ensuite gravée en utilisant un nouveau jeu de masques comme le montre la *figure 6.37*.

Ensuite, le diélectrique entre les niveaux 1 et 2 est déposé par CVD. Ce diélectrique est appelé IMD1. Le polissage CMP est également employé pour obtenir une surface plane. Un nouveau jeu de masques permet de définir les vias entre niveau 1 et niveau 2 comme le montre la *figure 6.38*. Notons qu'il n'est pas utile de faire passer les contacts de drain du niveau 1 au niveau 2, pour la réalisation de l'inverseur.

Le procédé se répète comme dans la fabrication du niveau 1 et on obtient finalement le dispositif de la *figure 6.39*.

D'autres niveaux peuvent être réalisés en fonction de la complexité des interconnexions. Enfin, la surface en contact avec l'extérieur est passivée. Les matériaux les plus utilisés sont des verres dopés et du nitrure de silicium. Des ouvertures peuvent être effectuées dans l'oxyde de passivation par

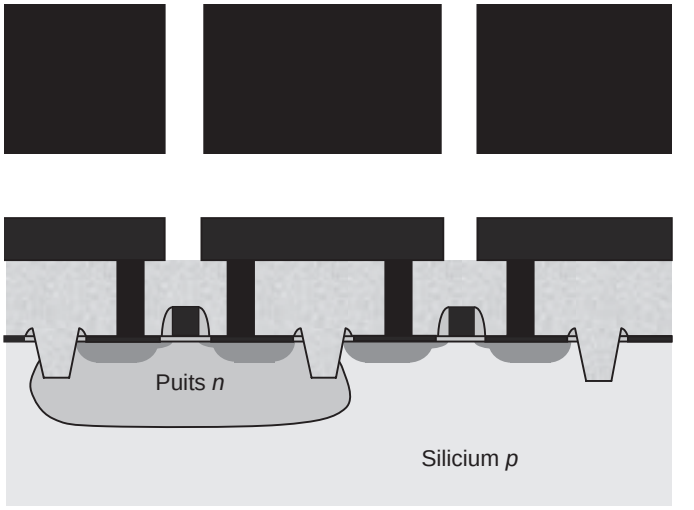


Figure 6.37 – Métallisation niveau 1.

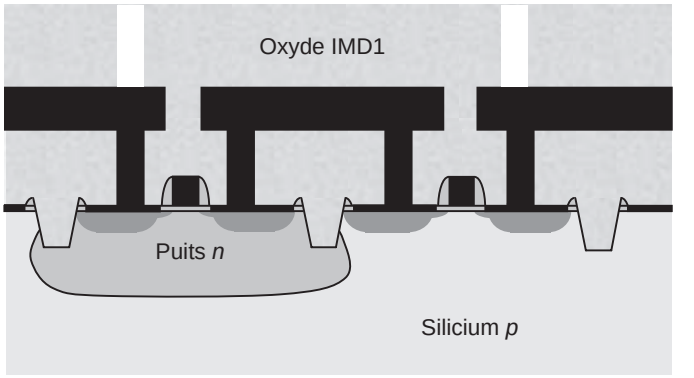


Figure 6.38. Vias entre métal 1 et métal 2.

gravure sèche en utilisant un masque adapté. Cette opération est pratiquée pour pouvoir faire les connexions avec les sorties du boîtier. Ces connexions sont appelées *bondings*. Le dispositif final est représenté *figure 6.40*.

Notons que la technique pour réaliser les interconnexions en cuivre n'est pas décrite dans ce chapitre. Le procédé est également du type *Damascene*, c'est-à-dire remplissage d'une tranchée puis érosion mécanique pour obtenir une surface plane. Ce procédé a été inventé dans la ville de Damascus aux USA.

Après fabrication collective et test, tous les circuits du wafer sont découpés avec une scie de précision puis conditionnés. Ils sont en général montés dans des boîtiers mais il est également possible de fournir des puces nues à condition de les conserver en atmosphère neutre.

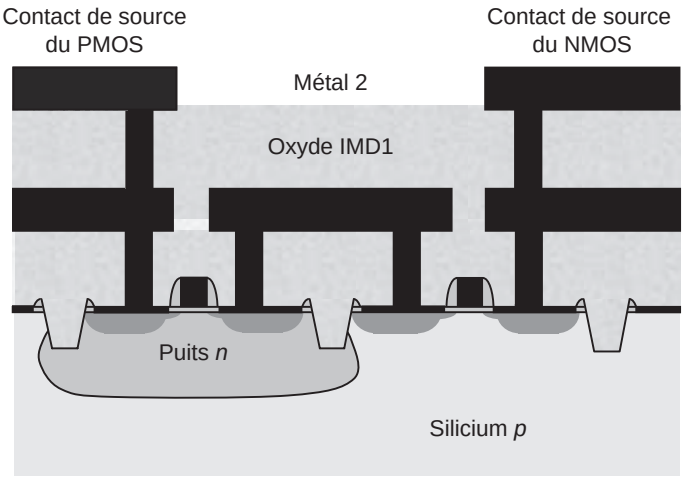


Figure 6.39 – Dispositif avec les deux niveaux d'interconnexion.

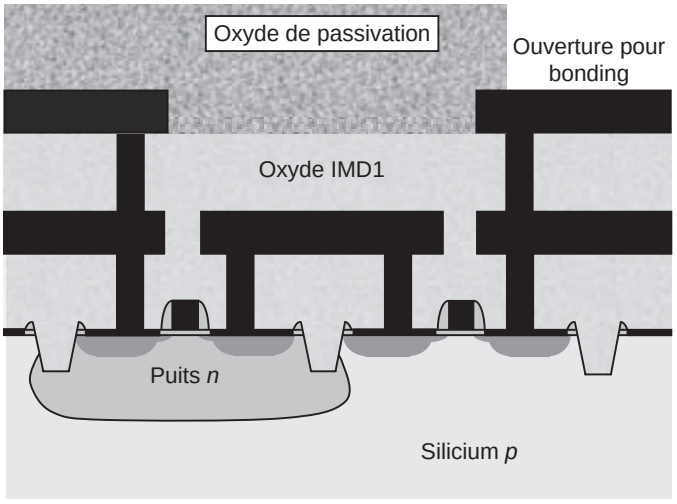


Figure 6.40 – Dispositif final avec deux niveaux de métallisation.

Pour terminer ce chapitre, les *figures 6.41 et 6.42* montrent la photographie au microscope électronique d'une véritable interconnexion réalisée sur un circuit intégré.

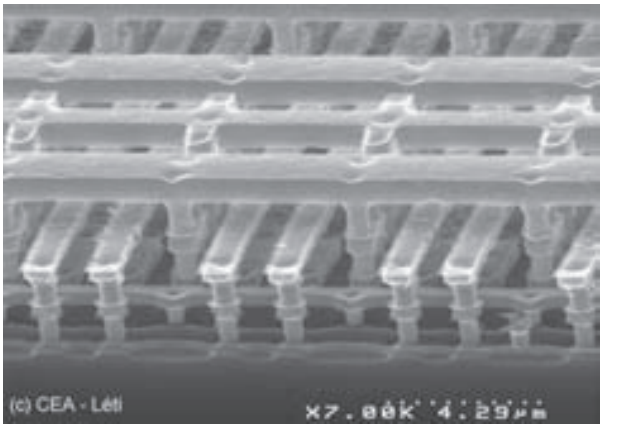


Figure 6.41 – Réseau d'interconnexions (Crédit photo CEA Grenoble).

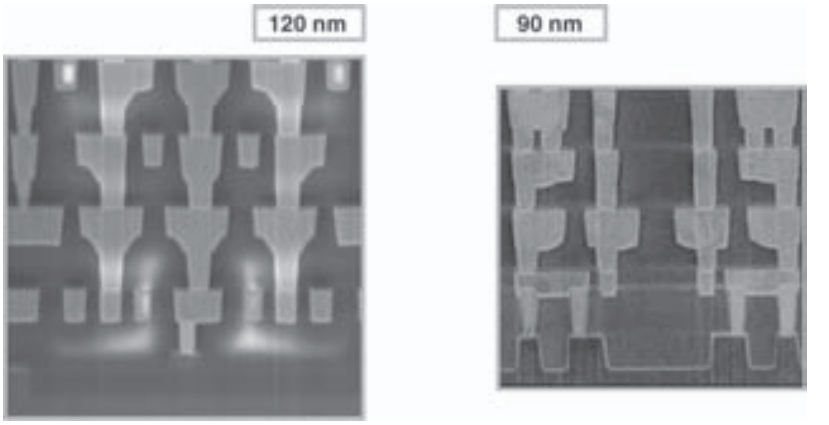


Figure 6.42 – Interconnexions dans différentes technologies (Crédit photo CEA Grenoble).

6.3.5 *Fabrication des condensateurs*

La technologie CMOS permet de fabriquer facilement des transistors performants. Il n'en est pas de même pour les autres composants de l'électronique : résistances, condensateurs, inductances. Tous les passifs sont difficiles à fabriquer et cela d'autant plus que leur valeur est élevée. La faible surface disponible en est la raison principale. Pour réaliser un condensateur, trois techniques sont utilisées :

- *Capacité MOS :* elle est formée au niveau de l'oxyde de grille, le silicium polycristallin formant une des armatures et une couche de silicium fortement dopée *n* l'autre armature. Cette couche dopée n'est pas prévue dans le flot classique et doit donc être réalisée moyennant un niveau de masque supplémentaire. L'intérêt de cette méthode est lié à la faible épaisseur de l'oxyde de grille ce qui permet de réaliser des capacités relativement importantes par unité de surface.

- *Capacité poly/poly :* elle est créée en ajoutant un second niveau de silicium polycristallin. Le condensateur se forme donc entre les deux surfaces de silicium polycristallin. Cette technique est la plus utilisée, bien qu'elle nécessite un niveau de masquage supplémentaire. Le *tableau 6.2* donne des valeurs typiques pour deux technologies, une technologie 0,8 micron et une technologie 45 nm.

Tableau 6.2

	Épaisseur oxyde (nm)	Capacité par unité de surface (fF/μm²)
0,8 micron	20	1,75
45 nm	1,5	25

- *Capacité poly/puits :* elle est créée entre une surface de silicium polycristallin et un puits dopé *n* fabriqué spécifiquement. L'oxyde de grille sert de couche diélectrique. Cette technique a comme avantage d'être disponible dans un flot CMOS standard ce qui n'est pas le cas des deux premières.

6.3.6 *Fabrication des résistances*

Le silicium dopé, le silicium polycristallin et les couches de siliciure sont des matériaux résistants pouvant être utilisés pour faire des résistances. Il est commode de définir la résistance d'une couche par la résistance carrée, c'est-à-dire la résistance d'un élément carré de côté *a* comme le montre la *figure 6.43*.

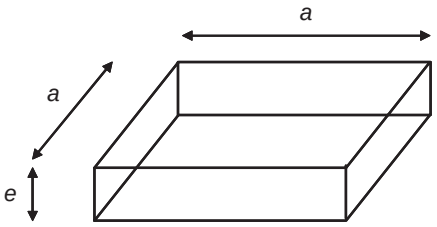


Figure 6.43 – Notion de résistance carrée.

La résistance entre deux faces du dispositif est alors :

$$R = \rho \frac{a}{a \cdot e} = \frac{\rho}{e} \tag{6.3}$$

Cette valeur exprimée en Ω/carré est donc indépendante de la grandeur du carré choisi mais dépend de la résistivité ρ et de l'épaisseur *e* de la couche. On comprend facilement que si *n* carrés sont mis bout à bout, la résistance du dispositif est *nR* Ω/carré. Le *tableau 6.3* donne quelques valeurs typiques de la résistance en Ω/carré ainsi que le coefficient de variation en fonction de la température exprimé en ppm par degré. Le ppm est une partie par million.

Il faut également noter que la valeur de la résistance réalisée à partir de l'un quelconque de ces matériaux dépend de la tension aux bornes. Cette dépendance est exprimée par le coefficient *VCR*

Tableau 6.3

Type de matériau	Résistance (Ω/carré)	Coefficient de température (ppm/degré)
Puits n	500	2 400
Polycristallin n+	200	20
Polycristallin p+	400	200
n+	100	1 500
p+	130	1 500
Polycristallin n+ siliciure	5	3 000
Polycristallin p+ siliciure	7	4 000
n+ siliciure	10	4 000
p+ siliciure	20	4 000

qui est en général de quelques milliers de ppm par volt. Cette variation est donc en général plus faible que celle induite par une variation de température mais dans quelques cas particuliers importants elle devient prépondérante. Par exemple, la valeur du rapport de tension créé par un pont diviseur résistif varie peu avec la température mais varie beaucoup avec la valeur absolue de la tension. Le lecteur pourra vérifier cette affirmation en écrivant :

$$R = R_0[1 + TCC(T - T_0)]$$
$$R = R_0[1 + VCR(V - V_0)]$$

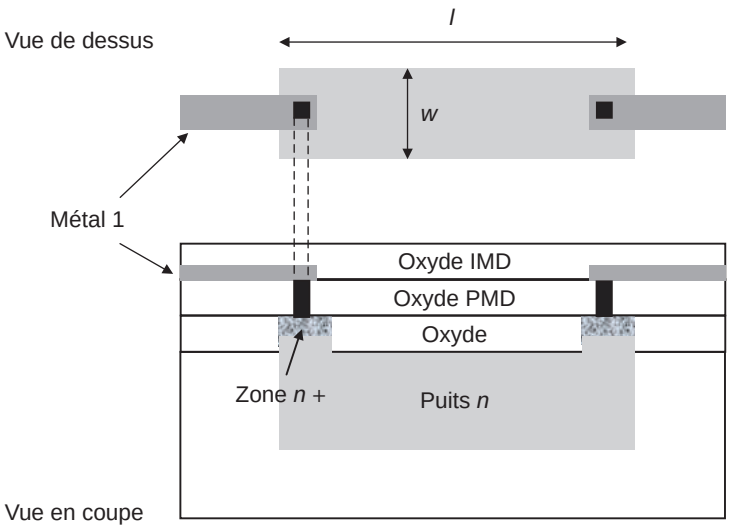


Figure 6.44 – Dessin d'une résistance à partir du silicium du puits.

La *figure 6.44* montre le dessin d'une résistance sur un circuit utilisant par exemple le silicium du puits dopé *n*. La résistance n'est pas toujours en ligne droite. Pour atteindre des valeurs élevées, il est possible de réaliser des motifs en forme de serpentins. Il est souvent nécessaire dans les circuits de réaliser des résistances appariées, par exemple égales. La conservation de cette égalité est souvent plus importante que la définition précise de la valeur absolue de la résistance. Une cause importante de dispersion est la non uniformité du dopage sur la surface du circuit. La *figure 6.45* montre trois manières de réaliser deux résistances égales à partir d'éléments résistifs implantés sur le wafer.

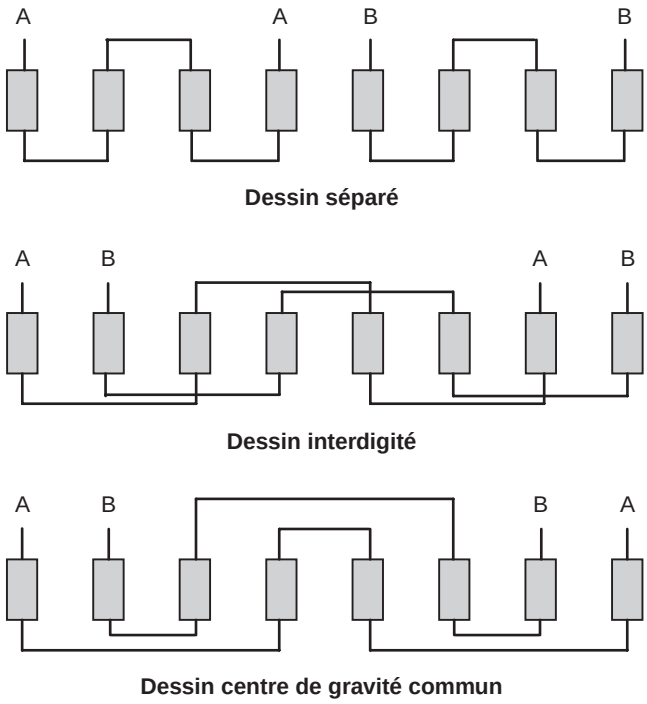


Figure 6.45 – Dessin de résistances.

Les deux implémentations (interdigitée et centre de gravité commun) permettent de s'affranchir plus facilement des variations de résistivité de la couche conductrice. Pour s'en convaincre, il suffit d'imaginer un gradient de résistivité par exemple de droite à gauche sur la figure. La première solution conduit à des valeurs différentes pour A et B mais les deux autres fournissent des valeurs proches. La dernière solution (centre de gravité) est en fait la plus performante. Ces concepts se transposent facilement dans des assemblages à deux dimensions. L'inconvénient est l'introduction d'éléments parasites de couplage.

6.3.7 *Architecture des circuits*

Dans un premier temps rappelons de manière simplifiée comment les transistors NMOS et PMOS sont réalisés sur le wafer.

Ce schéma simplifié montre la correspondance entre la structure en couches d'un circuit intégré et le dessin à deux dimensions des motifs permettant la création des dispositifs et leurs intercon-

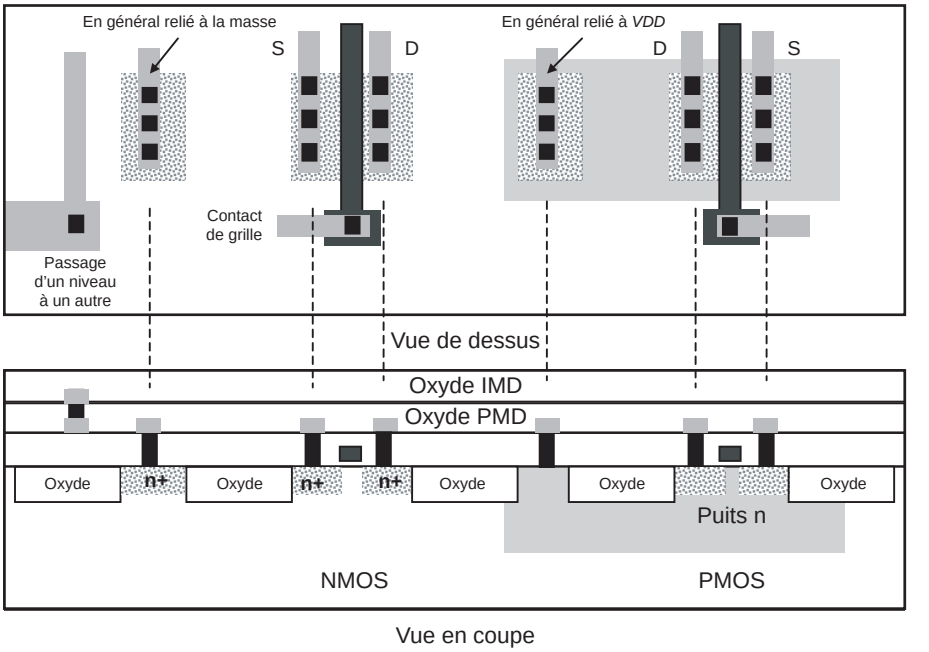


Figure 6.46 – Dessins des transistors et des interconnexions.

nexions. Tous les concepts technologiques ont été exposés précédemment. Notons simplement la symétrie totale du transistor, ce qui implique qu'il est possible de faire jouer au même élément physique le rôle de drain ou de source. Le placement-routage global choisira la solution la meilleure, celle qui minimise la longueur des interconnexions. Notons également qu'il est intéressant de relier le substrat du transistor NMOS à la masse pour éviter les effets de variation de tension de seuil avec la tension du « body ». De même, le puits dopé *n* du transistor PMOS est en général relié à la tension d'alimentation (V_{DD}) du circuit.

Il est également utile de donner l'ordre de grandeur des capacités par unité de surface entre niveaux de métallisation ou entre métal et silicium. Ces capacités jouent un rôle majeur dans les propriétés électriques des circuits car elles sont en général très supérieures aux capacités des transistors eux-mêmes. Les valeurs sont données dans deux cas : technologie 0,8 micron et technologie 45 nm.

Le *tableau 6.4* montre que les capacités parasites ont une valeur uniforme de 100 *aF* par micron carré pour les technologies avancées.

Un certain nombre de règles régissent le dessin des éléments et des interconnexions. Ce sont les règles de dessin de la technologie. La technologie est, rappelons-le, définie par une longueur λ appelée nœud de la technologie. Cette grandeur n'est pas la longueur minimale de la grille d'un transistor de la technologie mais la moitié de la distance minimale entre deux pistes de connexion. Cette valeur résulte directement des contraintes de la lithographie. Par exemple, un nœud 45 nm implique que deux connexions seront au moins distantes de 90 nm comme le montre la *figure 6.47*.

En fait, cette dimension correspond en général avec la largeur minimale d'une couche de silicium polycristallin et donc à la longueur dessinée de la grille minimale d'un transistor. La longueur électrique est cependant plus faible comme il a été expliqué dans le chapitre 4.

Tableau 6.4

	Technologie 0,8 micron (aF/μ^2)	Technologie 45 nm (aF/μ^2)
Polycristallin / substrat	60	90
Métal 1 / polycristallin	40	90
Métal 1 / substrat	20	80
Métal 1 / diffusion	40	80
Métal 2 / polycristallin	20	80
Métal 2 / diffusion	20	90
Métal 2 / métal 1	40	100
Métal 2 / diffusion	20	90

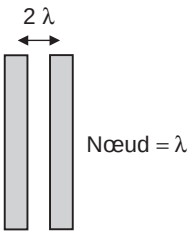


Figure 6.47 – Le nœud d'une technologie.

Les règles de dessin fixent les largeurs minimales des pistes et les distances entre éléments en fonction du niveau. Les niveaux « haut » correspondent à des pistes plus larges car elles transportent les signaux sur des distances élevées ou sont utilisées pour créer des plans de masse et des plans de tension d'alimentation. Les règles de dessin fixent également les distances minimales entre les pistes pour garantir une bonne isolation électrique. Enfin, elles fixent les zones minimales de recouvrement pour assurer les contacts entre couches différentes.

La *figure 6.48* donne une représentation graphique de règles de dessin pour une technologie donnée. Toutes les dimensions ont comme unité le nœud de la technologie. Une dimension 2 dans une technologie 90 nm est équivalente à 180 nm. Rappelons que la règle majeure de la micro-électronique est de chercher la surface minimale pour diminuer le coût.

Dans une deuxième étape, nous pouvons étudier comment les cellules de base sont réparties dans un circuit intégré. Insistons tout d'abord sur l'importance des contacts avec l'extérieur. Ils sont disposés tout autour du circuit intégré comme le montre la *figure 6.49*. Ce sont des carrés conducteurs d'environ 100 microns de côté.

Notons que la dimension du contact de sortie appelée « pad » n'évolue pas avec la réduction de taille du transistor. C'est le seul paramètre qui ne diminue pas de manière proportionnelle avec le nœud de la technologie. La raison en est simple : il faut souder un fil de connexion entre ce « pad » et le plot du boîtier. Les dimensions des fils de soudure ne peuvent diminuer en dessous d'une valeur

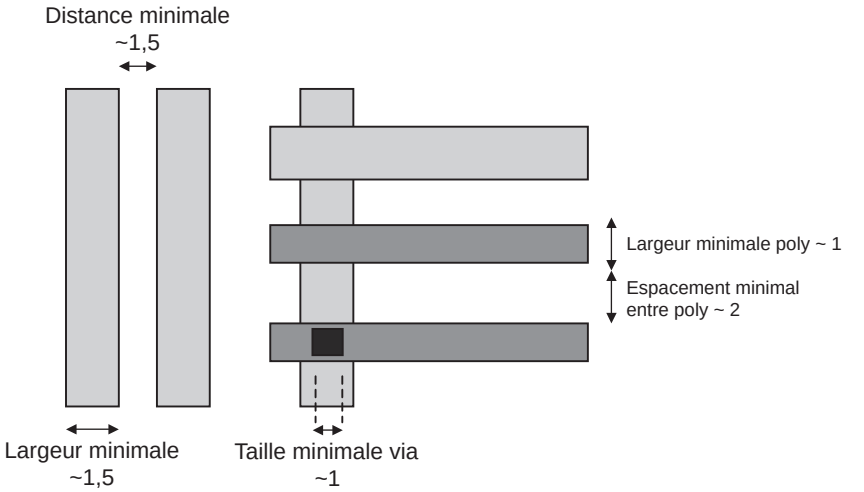


Figure 6.48 – Exemples de règles de dessin.

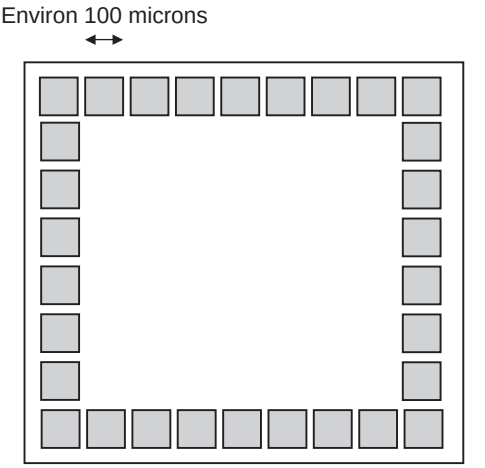


Figure 6.49 – Entrées et sorties d'un circuit intégré.

minimale pour ne pas trop augmenter la résistance et l'inductance du fil de contact. La fiabilité du contact impose également une surface minimale.

Pour terminer ce paragraphe, il est maintenant possible d'étudier comment sont réparties les cellules logiques de base dans un circuit intégré. Les cellules logiques seront détaillées dans le chapitre 8 mais on peut admettre qu'elles sont formées d'un nombre limité de PMOS et de NMOS en série. Le cas de l'inverseur est le plus simple puisque dans ce cas un NMOS est placé en série avec un PMOS comme le montre la *figure 6.50*.

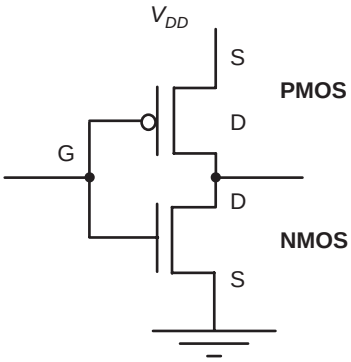


Figure 6.50 – Inverseur CMOS.

Des schémas plus complexes conduisent à placer plusieurs transistors en série pour réaliser des fonctions logiques plus élaborées. La *figure 6.51* illustre le cas d'un additionneur.

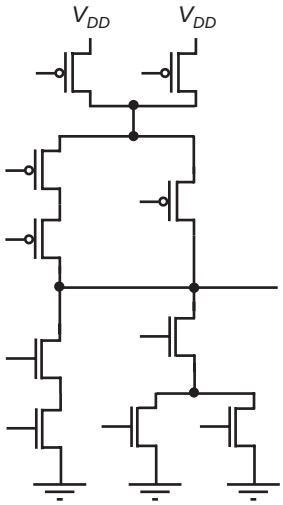


Figure 6.51 – Élément d'additionneur binaire.

On constate donc que toutes les cellules logiques sont formées de PMOS reliés à la tension positive d'alimentation et des NMOS reliés à la masse. On en arrive donc naturellement à une organisation des cellules logiques en lignes comme le montre la *figure 6.52*. Le cas des mémoires est tout à fait différent et sera traité dans le chapitre 9.

Les cellules de base ont donc toutes la même hauteur ce qui permet de les placer entre les lignes de masse et d'alimentation. Les largeurs varient en fonction de la complexité de la fonction. L'ensemble de ces cellules de base constitue une bibliothèque. Le nombre de cellules de base d'une bibliothèque varie d'une centaine à quelques centaines d'éléments.

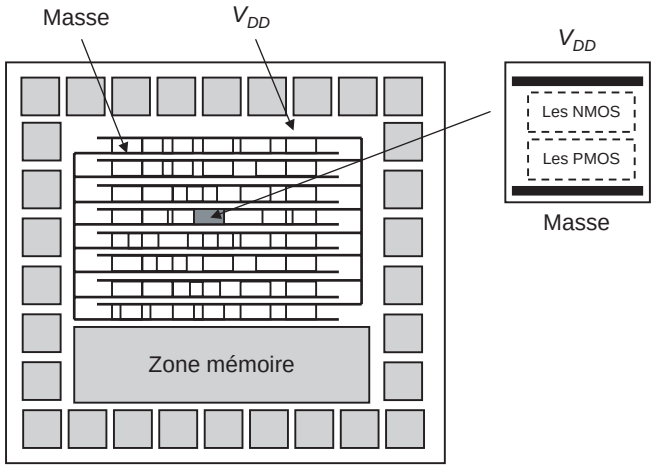


Figure 6.52 – Implantation de la logique sur un circuit intégré.

6.4 Les technologies micro-électroniques

Les fabricants de circuits intégrés offrent un large panel de technologies possibles. Le nœud de la technologie n'est pas le seul choix possible et de nombreuses options sont offertes.

Les technologies seront soit purement numériques soit mixtes, c'est-à-dire combinant transistors adaptés à la logique et transistors adaptés à l'analogique. Elles pourront aussi comporter des composants adaptés à l'implémentation de fonctions radiofréquence. Les technologies permettent de fabriquer soit uniquement des MOS soit à la fois des MOS et des transistors bipolaires. Il est également possible de prévoir la fabrication possible de composants pouvant délivrer des courants élevés ou commuter des tensions élevées. On parle alors de technologie de puissance. Les circuits intégrés pour l'automobile sont fabriqués en partie à l'aide de ces technologies de puissance. Enfin, certaines technologies permettent d'embarquer des mémoires particulières comme les mémoires non volatiles.

Pour illustrer cette offre, il est possible de parcourir les *roadmaps* d'un certain nombre de fondeurs. Les technologies de TSMC, fondeur asiatique, sont présentées *figures 6.53 et 6.54*.

L'offre de ce fondeur se répartit pour les technologies basse tension en trois familles : une mixte classique, une mixte avec des options RF et une mixte à base de MOS et de bipolaires Silicium-Germanium. Les technologies 90 nm sont mises sur le marché depuis 2004. Les tensions de fonctionnement de la partie CMOS sont autour de 1 V pour les technologies les plus avancées. Il faut également noter que plusieurs générations coexistent dans l'offre industrielle.

Les coûts de développement d'un circuit intégré dans une technologie avancée ne se justifient que si les performances ou le volume de production le demandent. La *figure 6.54* illustre l'offre du même fondeur mais cette fois dans le domaine des technologies haute tension.

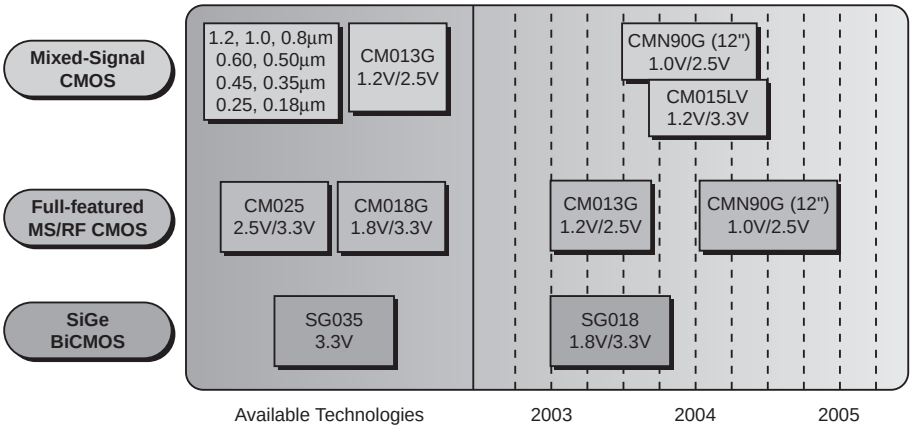


Figure 6.53 – Les technologies du fondeur TSMC en basse tension.

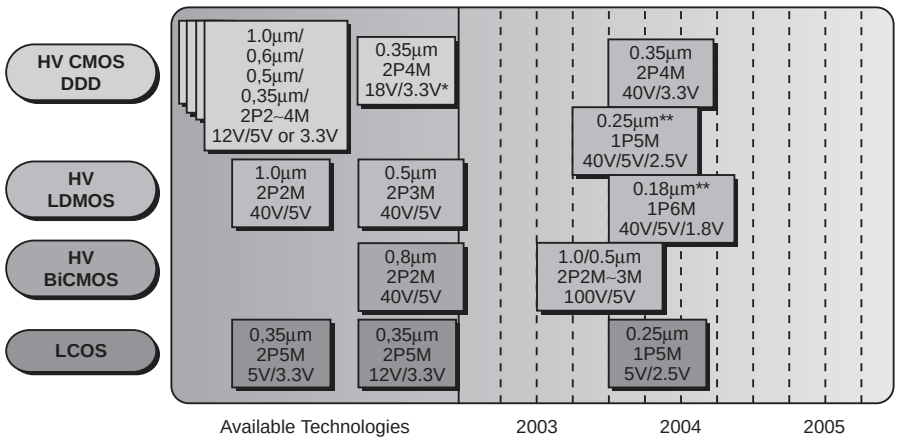


Figure 6.54 – Offre industrielle de circuits haute tension.

6.5 Les procédés alternatifs

La lithographie a fait la force de la micro-électronique en permettant la fabrication collective des circuits. Elle est cependant la principale source de difficultés pour l'avenir. En effet, les dispositifs de lithographie sont d'une complexité croissante au fur et à mesure que la résolution demandée s'affine. Le coût des équipements devient très élevé. Un dispositif d'insolation avec photorépétition est acheté autour de 10 millions d'euros et un jeu de masques pour réaliser un circuit complexe peut coûter quelques millions d'euros.

Les masques sont fabriqués à l'aide d'un appareil générant un faisceau d'électrons capable de balayer la surface du masque avec une précision extrême. Le procédé n'est plus limité par la longueur d'onde puisque la longueur d'onde associée à l'électron est très faible. La résolution possible de gravure

est de quelques nm mais le temps d'insolation est très long à cause du caractère séquentiel de l'opération ce qui explique en partie le coût du jeu de masques fabriqué par cette méthode.

On pourrait imaginer des procédés de fabrication permettant de se passer des masques. On pourrait par exemple utiliser la lithographie par faisceau d'électrons non pas pour fabriquer les masques mais pour produire directement les circuits intégrés. Cela est effectivement pratiqué pour réaliser des prototypes. La lithographie par faisceau d'électrons ne peut cependant être mise en œuvre dans un procédé industriel car le temps d'insolation est long par principe puisqu'il est séquentiel. Il est donc naturel que la micro-électronique investigue des techniques alternatives à la lithographie optique.

6.5.1 *La nanoimpression*

Ce procédé a été proposé par S.Y. Chou en 1995. Il est devenu une méthode de référence pour fabriquer des nanodispositifs car très simple de mise en œuvre. Son principe est hérité des technologies de l'impression. Il est illustré *figure 6.55*.

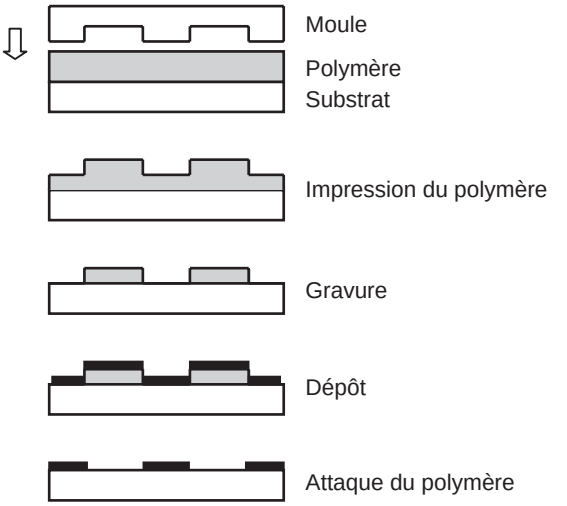


Figure 6.55 – Nanoimpression.

Le procédé s'explique simplement, Un moule réalisé par lithographie électronique sur un support de silicium écrase un polymère déposé sur le substrat. La pression est de quelques dizaines de bars. La gravure ionique réactive permet ensuite de graver le polymère en atteignant le substrat. Un film mince est ensuite déposé sur le polymère résiduel, un film métallique par exemple. L'étape suivante appelée *lift-off* consiste à enlever le polymère par dissolution chimique. On obtient alors le motif choisi imprimé dans le film déposé. Ce film peut par la suite servir de masque de gravure. Cette technique permet de réaliser des motifs avec une résolution de l'ordre de 10 nm. Le moule peut en effet être fabriqué par une technique mettant en œuvre la lithographie e-beam puisqu'il est fabriqué une seule fois. La nanoimpression est appliquée pour étudier des dispositifs de très faible dimension et pour réaliser des matériaux magnétiques ou optiques nanostructurés.

Des techniques dérivées sont également étudiées comme la nanoimpression sous irradiation, la nanocompression et la lithographie molle. La nanoimpression sous irradiation permet de travailler à

température ambiante. La nanocompression utilise le polymère sous forme de granulés. La lithographie molle est basée sur l'utilisation d'un élastomère pour réaliser une sorte de tampon encreur capable de déposer des molécules organiques. On peut également citer la lithographie en champ proche, technique de lithographie par contact mais qui utilise un masque souple en contact parfait avec le substrat. Toutes ces techniques sont encore au stade de la recherche et ne sont pas encore appliquées dans l'industrie.

6.5.2 *Techniques d'auto-assemblage*

Pour éviter les coûts de la lithographie, il est séduisant d'imaginer des structures organisées qui seraient réalisées par des procédés purement chimiques. Il est difficile d'imaginer que de telles méthodes soient capables de générer des systèmes non réguliers comme les unités de calcul mais il est envisageable de les appliquer à la fabrication de systèmes réguliers tels que les mémoires.

La fabrication atome par atome est possible en appliquant les principes de la microscopie en champ proche mais elle n'est pas applicable pour la réalisation de systèmes complexes. Il est donc nécessaire d'imaginer des méthodes collectives conduisant à la création de structures régulières de nano-objets. Ces méthodes sont très nombreuses mais peuvent se répartir en deux grandes familles. La première s'appuie sur les propriétés cristallographiques du support et la seconde intègre différentes techniques issues de la synthèse chimique.

Il est intéressant d'envisager les méthodes de croissance de nano-objets sur des surfaces préstructurées. Prenons l'exemple de la surface obtenue par section d'un cristal. La surface de coupe peut, si la coupe ne se fait pas dans la direction d'un plan cristallin, présenter une succession de marches à l'échelle nanométrique. Ce support structuré peut alors être la base pour faire croître des nanofils le long des marches.

Trois grandes techniques sont mises en œuvre pour former les surfaces préstructurées :

- profiter des propriétés intrinsèques des surfaces (réseaux de défauts, réseau de marches…) ;
- structurer la surface par des techniques de gravure particulières (géométrie de la surface ou création de réseaux de dislocations) ;
- coupler les deux méthodes précédentes.

Les méthodes de synthèse chimique s'inspirent des assemblages moléculaires qui ont conduit aux organismes vivants. Les assemblages peuvent alors se faire en volume ou en surface. Ces méthodes permettent en particulier de créer des réseaux de nanoparticules magnétiques mais aussi des monocouches organisées sur des surfaces.

Chapitre 7

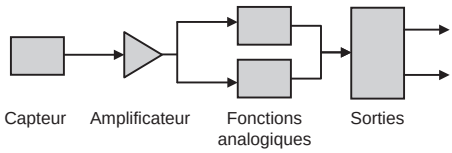
Les fonctions analogiques de base

Le but de ce chapitre est de montrer comment les fonctions électroniques analogiques de base (amplification, commutation, filtrage, conversion) sont réalisées dans un circuit intégré. L'optimisation du rapport signal sur bruit est un point très important dans la conception des fonctions analogiques. Ce chapitre montre également la relation entre les propriétés électriques des fonctions et les dimensions physiques des transistors, ce qui est un point fondamental de la technologie CMOS.

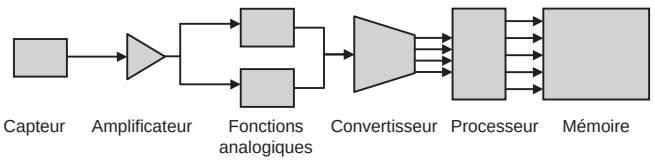
7.1 Les fonctions analogiques et les outils de conception

Les systèmes électroniques (cartes ou circuits intégrés) sont constitués de diverses fonctions dites élémentaires de type numérique ou analogique. Les systèmes numériques traitent des données binaires en effectuant des opérations en arithmétique binaire et en réalisant des fonctions logiques. Il est cependant nécessaire de faire l'interface avec le monde physique qui, lui, est de type analogique, c'est-à-dire qu'il fournit des signaux variant continûment.

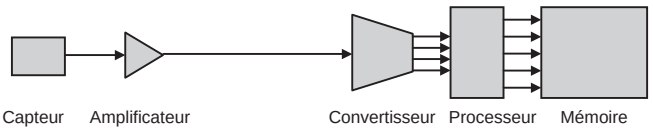
Le signal créé par un microphone ou une caméra, le signal détecté par une antenne sont autant d'exemples montrant le caractère analogique des interfaces avec le monde physique. Les signaux délivrés par le monde physique sont en général de faibles valeurs et souvent peu discernables dans un bruit important. C'est le cas du signal électrique aux bornes d'une antenne. Il est donc nécessaire d'amplifier ces signaux. L'amplification n'est cependant pas la seule fonction à réaliser, elle est le plus souvent associée à une fonction de filtrage permettant d'une part d'optimiser le rapport signal sur bruit et d'autre part d'éliminer les fréquences indésirables (signal parasite ou bandes de fréquences à rejeter). L'optimisation du rapport signal sur bruit est sans doute la technique la plus importante dans la conception des éléments analogiques d'entrée d'un système électronique.



a) Voie analogique



b) Voie mixte



c) Voie numérique

Figure 7.1 – Évolution de l'électronique analogique.

La généralisation des composants numériques (processeurs et mémoires) conduit à limiter les fonctions analogiques de filtrage et de contrôle au strict minimum. Pensons par exemple aux fonctions de réglage du son et aux compensations des défauts d'une image. Toutes ces opérations de traitement s'effectuant sur des valeurs numérisées, il est nécessaire de réaliser des convertisseurs analogique-numérique performants. Amplificateurs et convertisseurs constituent les fonctions analogiques les

plus importantes de l'électronique. Leur intégration dans des puces de silicium est donc l'aspect principal traité dans cet ouvrage. La *figure 7.1* illustre l'évolution de l'électronique intégrée et l'importance croissante des fonctions numériques.

La fonction amplification reste indispensable mais les opérations de filtrage sont le plus souvent effectuées par des calculs sur les échantillons numérisés du signal et non pas comme dans le passé sur les signaux eux-mêmes. Les techniques numériques offrent en effet plus de souplesse. Il suffit de changer le programme pour changer le filtre. Il en est de même pour toutes les opérations de correction ou de modification des données. La fonction de conversion analogique-numérique est la deuxième fonction de base fondamentale dans l'électronique moderne. Elle est plus complexe que la fonction amplification et sera évoquée à la fin de ce chapitre. À l'opposé, les données issues du traitement sont restituées au monde physique par un convertisseur numérique-analogique suivi quand cela est nécessaire par un amplificateur de puissance.

En résumé, on trouvera les fonctions de base suivantes :

- amplification ;
- conversion analogique-numérique ;
- conversion numérique-analogique ;
- amplification de puissance.

Pour réaliser ces fonctions à base de transistors, l'industrie micro-électronique a défini un flot de conception. Ce flot commence par une analyse de la fonction à réaliser soit de manière textuelle soit de manière plus formelle en utilisant des logiciels mathématiques de description et de simulation comme Matlab ou Mathematica.

Ensuite, il est nécessaire de faire le schéma électrique de la fonction à partir de transistors, de résistances, de condensateurs et d'inductances. Le composant le plus facile à réaliser dans un circuit intégré est le transistor. On limitera donc le plus possible la fabrication des résistances et des condensateurs qui consomment beaucoup de surface de silicium. Les inductances sont également difficiles à réaliser et leur coefficient de qualité est souvent médiocre. Ces considérations ont une grande importance dans la conception des fonctions analogiques. Les résistances seront dans la mesure du possible remplacées par des transistors et les schémas seront conçus pour utiliser des condensateurs de faibles valeurs, quelques picofarads au maximum. Les réalisations présentées dans ce chapitre font appel à la technologie CMOS qui s'impose au fil du temps aussi bien en analogique qu'en numérique. La technologie bipolaire est réservée à quelques applications exigeant une vitesse élevée et ne sera pas détaillée dans cet ouvrage.

À l'issue de cette phase de dessin, il est indispensable de simuler le comportement électrique du circuit pour vérifier que les choix effectués sont conformes aux performances à obtenir : vitesse, consommation, niveau de bruit… Des logiciels très performants permettent d'effectuer ces simulations électriques à partir des modèles électriques décrits dans le chapitre 4. Ils sont issus du logiciel SPICE inventé à l'université de Berkeley.

Enfin, il est nécessaire de dessiner les masques permettant de fabriquer les transistors et les interconnexions. Cette opération est le routage. Elle est d'une grande importance dans la conception des fonctions analogiques car les propriétés électriques des interconnexions ont un effet majeur sur les propriétés électriques globales.

7.2 Amplification et sources

7.2.1 *Remarques préliminaires*

Avant de détailler les amplificateurs les plus classiques de la micro-électronique, il est nécessaire de donner quelques explications globales.

Si on considère un amplificateur de la manière la plus générale, il peut se représenter sous forme d'une boîte noire alimentée par une tension continue V_{DD}. Il est donc évident que les composants doivent travailler dans la gamme de tension comprise entre 0 et V_{DD}. Cela est vrai pour le signal d'entrée et aussi pour le signal de sortie. On peut donc représenter graphiquement la valeur de la tension de sortie en fonction de la valeur de la tension d'entrée. On obtient donc nécessairement une courbe du type de la *figure 7.2*. Deux cas sont représentés.

- Le premier offre une zone de gain relativement étendue. Le gain y est exprimé comme le rapport entre une variation de la tension de sortie divisée par la variation correspondante de la tension d'entrée. Il est modéré.
- Le second offre une zone de gain plus restreinte mais le gain y est plus élevé. Il sera dans ce cas nécessaire de régler le point de fonctionnement continu avec précision.

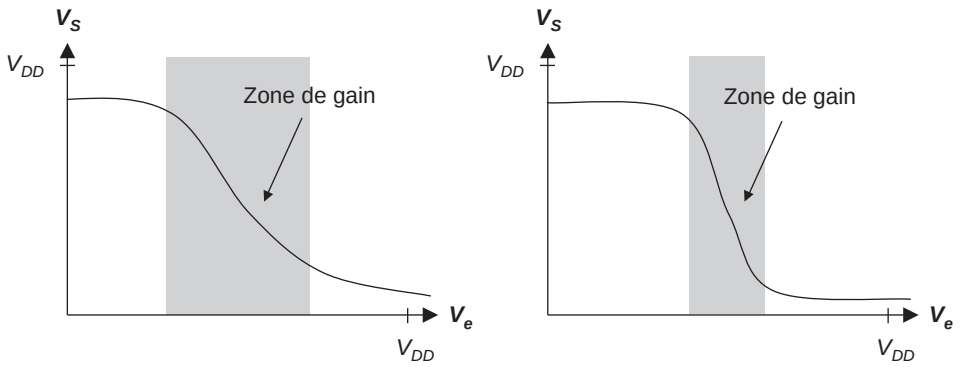


Figure 7.2 – Fonction de transfert d'un amplificateur.

Ce type de courbe est le seul à avoir un sens physique pour une fonction de type inverseur. On aurait pu considérer également une fonction de type non inverseur.

Si la tension d'entrée peut varier de 0 à V_{DD}, il n'en est généralement pas de même pour la tension de sortie qui varie dans une gamme plus étroite. Étendre au maximum la plage de fonctionnement de la tension de sortie est un objectif important dans le design des circuits. La zone de gain, c'est-à-dire la zone dans laquelle une variation de tension d'entrée induit une variation plus grande de la tension de sortie, est dans ce cas limitée à la zone centrale de la figure. La pente de la fonction est supérieure à l'unité en valeur absolue. Cette zone est nécessairement limitée. Graphiquement, on peut constater que plus le gain maximum est élevé, plus cette zone de gain est étroite.

Une autre remarque importante est liée aux conditions de polarisation des transistors. L'industrie électronique s'est ingéniée à développer des schémas faisant usage d'une seule polarité de tension, en général positive. Les premiers amplificateurs intégrés nécessitaient une tension d'alimentation positive et une tension d'alimentation négative. Cela compliquait la réalisation des sources d'ali-

mentation et ce type de schéma a progressivement disparu. Il est donc nécessaire de concevoir des schémas fonctionnant avec une source d'alimentation positive et uniquement positive. À l'origine de la technologie MOS, les tensions pouvaient atteindre des valeurs aussi élevées que 10 V. Au fur et à mesure du développement de cette technologie, les dimensions des transistors ont diminué et de ce fait il est devenu impossible de travailler avec des tensions élevées. Pour les technologies actuelles, des valeurs de 2 V sont habituelles, et dans l'avenir, des valeurs d'environ 1 V sont attendues. Cette réduction de la tension d'alimentation pose des difficultés de conception considérables car la plage de variation devient étroite.

Voyons maintenant comment polariser les transistors NMOS et PMOS avec une seule source de tension positive et comment assurer le régime de saturation. Le régime de saturation d'un transistor est intéressant car dans ce régime, le transistor a véritablement sa fonction idéale de source de courant commandée. La *figure 7.3* représente les deux types de transistor et les tensions à appliquer.

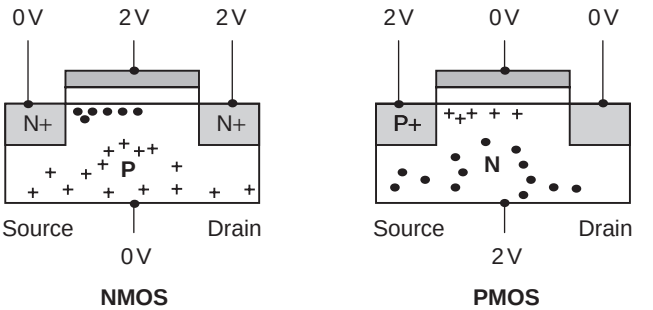


Figure 7.3 – Polarisation des transistors en mode saturé.

La polarisation du NMOS est classique avec des tensions positives appliquées sur les électrodes et le substrat polarisé à zéro. Si grille et drain sont au même potentiel, alors la tension de seuil n'est pas atteinte en bout de canal et le transistor est saturé. Quand on relie électriquement grille et drain, la condition de saturation est donc automatiquement assurée.

Dans le cas du PMOS, il faut que grille et drain soient polarisés négativement par rapport à la source et au substrat. Une manière facile de faire est de relier substrat et source à la tension la plus positive du circuit soit V_{DD}. Pour être en régime saturé, il est possible comme pour le NMOS de placer grille et drain au même potentiel, par exemple en les connectant électriquement.

7.2.2 *Comment amplifier avec un MOSFET ?*

Le transistor MOS étant un dispositif permettant de commander le courant de drain par la tension de grille, le premier schéma venant à l'esprit pour réaliser un amplificateur de tension est celui de la *figure 7.4*.

Ce schéma très simple conduit à écrire les équations suivantes :

$$V_s = V_{DD} - R_C I_{DS}$$

$$I_{DS} = k\frac{W}{2L}(V_e - V_T)^2(1 + \lambda V_{DS})$$

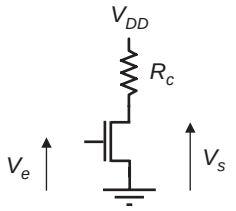


Figure 7.4 – Amplificateurs à MOSFET.

Dans cette dernière relation, tirée du chapitre 4, on a négligé la tension de seuil. Les paramètres λ et k sont donnés par les formules :

$$k = \mu_n C'_{OX}$$

$$\lambda = \frac{1}{V_A}$$

On exprime alors facilement le gain en tension :

$$A_V = \frac{dV_s}{dV_e} \approx -k\frac{W}{L}(V_e - V_T)R_c$$

On a négligé le terme en λ et la tension de seuil. Le transistor est supposé saturé. Dans une technologie 0,8 micron, la valeur de k est d'environ 110 µA par V^2. Si le transistor a un facteur de forme de 10 (valeur de W/L), on obtient pour une valeur typique de $(V_e - V_T)$ égale à 1 V.

$$A_V \approx 10^{-3} R_C$$

Un gain significatif peut être obtenu pour des résistances de charge supérieures à $10^4\ \Omega$. Comme il est difficile de réaliser des valeurs de résistance élevée, il faut trouver une autre solution.

7.2.3 *Amplificateur avec une source de courant comme charge*

L'idée est d'utiliser un transistor saturé comme charge. En effet, un transistor saturé est voisin d'une source de courant donc d'une résistance de valeur élevée. On en arrive naturellement à la *figure 7.5*.

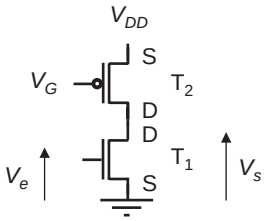


Figure 7.5 – Amplificateur à charge active.

Le transistor T2 sert de charge au transistor T1. C'est un PMOS afin de pouvoir fonctionner avec des tensions uniquement positives. La tension V_G fixe, appliquée sur la grille, est négative par rapport à la source reliée au V_{DD}. Le caisson est relié à la tension d'alimentation V_{DD} et le transistor est sup-

posé en régime de saturation. Il est équivalent à une source de courant soit une impédance de forte valeur.

La première étape dans l'étude d'un circuit est de fixer son point de fonctionnement, c'est-à-dire les valeurs continues des tensions et des courants. Dans les circuits à base de MOSFET il est impossible de calculer analytiquement ce point de fonctionnement car les équations sont trop complexes. Il est cependant possible de résoudre graphiquement le problème à partir des courbes caractéristiques. Dans cet exemple simple, on trace le réseau courant de drain fonction de la tension drain-source pour le transistor T1 et cela pour différentes valeurs de la tension continue d'entrée. Ensuite, on trace la courbe représentant le courant de drain en fonction de la tension drain-source pour le transistor T2. Dans ce cas, la tension de grille étant fixée, il y a une seule courbe possible. Le même courant traverse les transistors T1 et T2. Les tensions drain-source des transistors T1 et T2 sont reliées par la formule :

$$V_{DD} = V_{SD2} + V_{DS1}$$

On représentera donc sur le graphique non pas la tension V_{SD2} mais la valeur $V_{DD} - V_{SD2}$. Dans ce cas, une valeur en abscisse représente aussi bien V_{DS1} que $V_{DD} - V_{SD2}$. Les différents points de fonctionnement possibles sont obtenus par intersection du réseau de courbes correspondant au transistor 1 et de la courbe correspondant au transistor 2.

La valeur de V_{DD} est choisie à 2 V et la valeur appliquée sur la grille de T2 à 0,5 V pour assurer la saturation du transistor. Trois points de fonctionnement sont représentés pour trois valeurs de la tension sur la grille de T2, respectivement : 1 V, 1,5 V et 2 V. La *figure 7.6* représente ces points de fonctionnement.

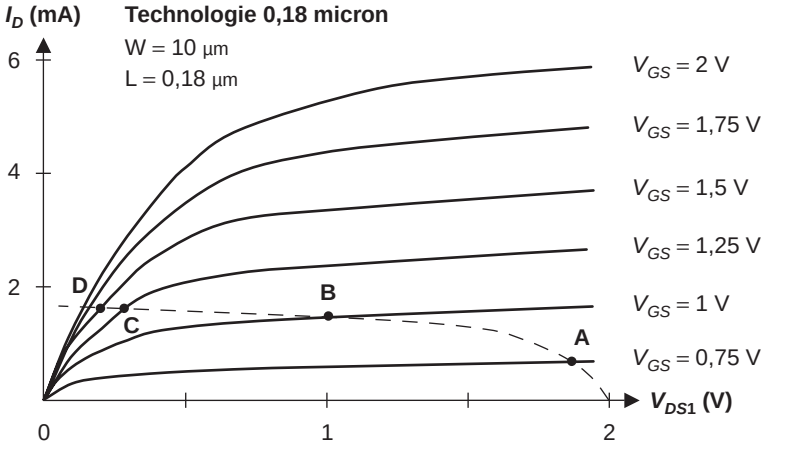


Figure 7.6 – Point de fonctionnement.

En fonction des valeurs obtenues sur ce graphique, il est possible de tracer la courbe donnant la tension continue de sortie (V_{DS1}) en fonction de la tension continue d'entrée (V_e).

L'examen de cette courbe montre que la plage de fonctionnement pour laquelle le gain en tension est élevé est relativement étroite. C'est la partie centrale de la courbe. Un réglage précis du point de fonctionnement est donc nécessaire.

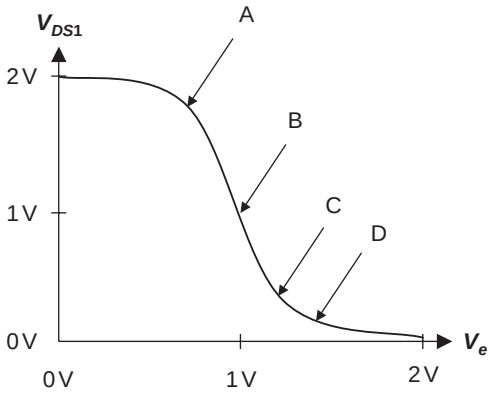


Figure 7.7 – Fonction de transfert.

En fonction des résultats établis dans le chapitre 5, on peut tracer le schéma équivalent en petits signaux correspondant à la *figure 7.8*. La tension de grille d'entrée du transistor T2 étant constante, la source de tension en petits signaux est nulle ainsi que la source de courant commandée correspondante. Le nœud correspondant à la tension d'alimentation V_{DD} est considéré à la masse dans la représentation petits signaux.

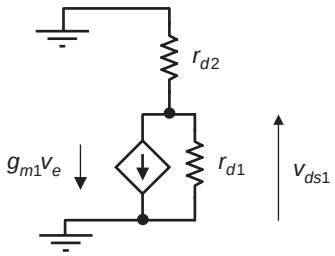


Figure 7.8 – Schéma équivalent en petits signaux.

Pour déterminer le gain de cet étage, il est nécessaire d'écrire la relation entre le courant traversant le transistor T2 et la tension drain-source de ce transistor. On utilisera le schéma équivalent petits signaux du transistor en régime saturé tel qu'il a été établi dans le chapitre 4.

Ce schéma est valable uniquement dans la région où les deux transistors sont saturés.

Un calcul élémentaire conduit à :

$$v_{ds1} = -g_{m1} v_e \frac{1}{\frac{1}{r_{d1}} + \frac{1}{r_{d2}}}$$

En exprimant les résistances en fonction du paramètre λ défini dans la formule 4.32 du chapitre 4, on obtient :

$$g_d = \frac{I_{D\text{sat}}}{V_A}$$

$$v_{ds1} = -g_{m1} v_e \frac{1}{g_{d1} + g_{d2}}$$

Le gain en tension est donc élevé et varie avec la transconductance du transistor T1. Elle est proportionnelle à la racine carrée du produit W_1/L_1 par le courant de drain en régime de canal long. En effet, le chapitre 4 nous apprend qu'en régime de canal long :

$$g_{m1} = \sqrt{2\beta_n \frac{I_{D1}}{1+\delta}} = \sqrt{2\mu_n C'_{OX} \frac{W_1}{L_1} \frac{I_{D1}}{1+\delta}} \quad (7.1)$$

Les termes g_{d1} et g_{d2} varient proportionnellement aux courants de saturation des transistors. Ces courants sont peu différents du courant de drain traversant les deux transistors en série. Le gain en tension varie donc au total de la manière suivante :

$$A_V \approx \sqrt{\frac{W_1}{L_1} \frac{1}{I_D}}$$

Ce résultat montre bien l'impact du choix des dimensions sur les performances.

7.2.4 *Amplificateur avec un transistor saturé comme charge*

Un autre schéma est fréquemment utilisé car il ne nécessite pas de source de tension supplémentaire. De plus, la plage dans laquelle le gain est élevé est plus large. Le principe est de charger le transistor T1 par un transistor T2 toujours en saturation. Deux cas sont possibles : T2 est un PMOS ou bien T2 est un NMOS. Dans un premier temps on envisage le cas du PMOS. Le schéma est celui de la *figure 7.9*. On notera que la saturation est assurée en reliant grille et drain. Dans ce cas, la tension grille-drain est nulle, donc toujours inférieure à la tension de seuil, ce qui implique que la charge d'inversion est nulle au niveau du drain de T2.

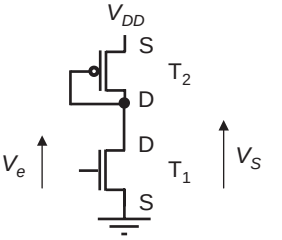


Figure 7.9 – Amplificateur avec transistor saturé comme charge.

Comme dans le montage précédent, on trace sur le même graphique les caractéristiques des deux transistors en tenant compte du fait que :

$$V_{DD} = V_{SD2} + V_{DS1}$$

La courbe exprimant le courant de drain en fonction de la tension drain-source pour le transistor T2, est extraite du réseau de courbes du transistor T2. Elle est obtenue quand la tension de grille et la tension de drain sont égales puisque grille et drain sont reliés électriquement. Cette courbe est tracée à partir des caractéristiques générales en se limitant aux points pour lesquels $V_{DS2} = V_{GS2}$.

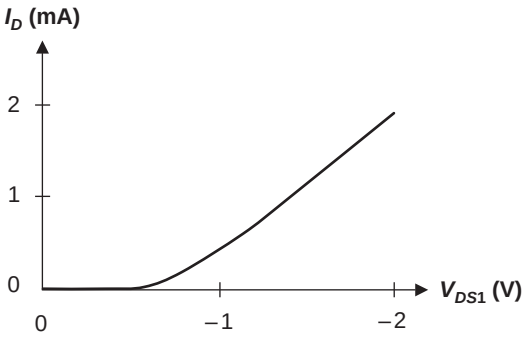


Figure 7.10 – Courant du PMOS saturé.

Il est alors possible de tracer sur le réseau de courbes de T1 la relation courant-tension de T2. Il faut simplement inverser la courbe et décaler de V_{DD} pour tenir compte de la relation suivante :

$$V_{DD} = V_{SD2} + V_{DS1}$$

On obtient alors la *figure 7.11*.

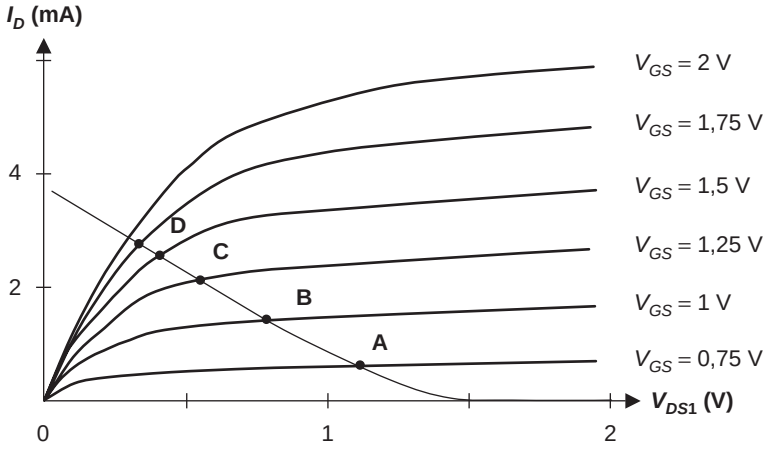


Figure 7.11 – Points de fonctionnement.

On en déduit alors la fonction de transfert de l'étage.

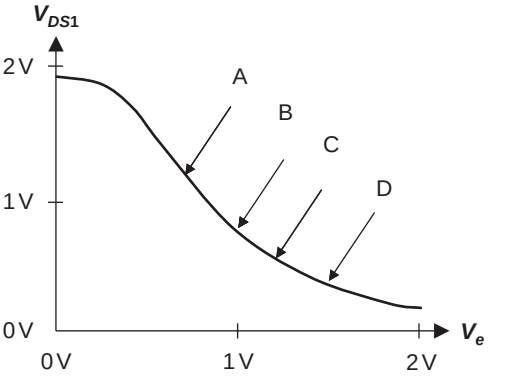


Figure 7.12 – Fonction de transfert.

La fonction de transfert est donc assez différente de celle du montage précédent. La plage dans laquelle le gain en tension est élevé est plus large. Le gain y est cependant plus faible.

Le calcul du gain dans la région centrale s'effectue à partir du schéma petits signaux.

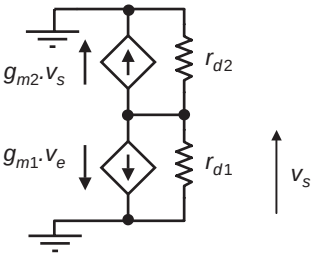


Figure 7.13 – Schéma équivalent petits signaux.

Il faut noter le sens de la source de courant pour le PMOS. Le chapitre 4 donne des précisions sur la manière d'obtenir ce schéma équivalent. Le fait que grille et drain de T2 soient reliés explique que la tension v_{GS2} est simplement *vs*. Le calcul du gain est alors immédiat en écrivant que la somme des courants est nulle au nœud central :

$$g_{m1}v_e + g_{d1}v_s + g_{m2}v_s + g_{d2}v_s = 0$$

soit,

$$\frac{v_s}{v_e} = -\frac{g_{m1}}{g_{m2} + g_{d1} + g_{d2}} \approx -\frac{g_{m1}}{g_{m2}} \qquad (7.2)$$

Les conductances de sortie sont en effet négligeables devant la transconductance. Le gain est donc fixé par le rapport de dimensions des transistors T1 et T2 et a de ce fait une valeur relativement modeste.

L'impédance de sortie de cet étage se calcule facilement.

$$Z_{out} = \frac{1}{g_{m2} + g_{d1} + g_{d2}} \approx \frac{1}{g_{m2}} \tag{7.3}$$

Elle est assez faible si le transistor T2 a un courant de polarisation suffisant ce qui confère à cet amplificateur des propriétés de rapidité convenables.

Pour calculer l'impédance de sortie de cet étage, on annule les sources de tension non liées et on calcule la tension qui apparaît en sortie quand on injecte un courant donné. Cette méthode classique est illustrée dans ce cas sur la *figure 7.14*.

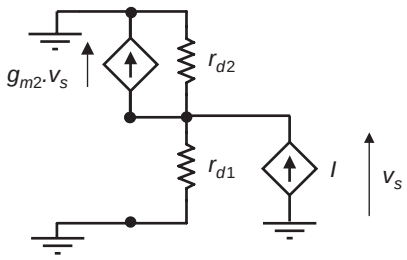


Figure 7.14 – Calcul de l'impédance de sortie.

On écrit facilement à partir du schéma 7.14 :

$$I = g_{m2}v_S + g_{d1}v_S + g_{d2}v_S$$

On obtient donc la relation 7.3 donnant l'impédance de sortie.

Le calcul qui suit, donne au premier ordre le comportement en fréquence de cet étage.

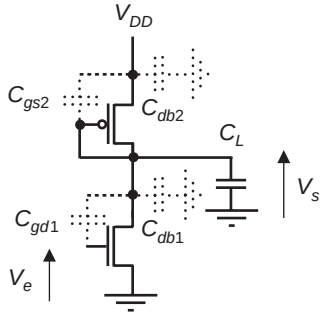


Figure 7.15 – Schéma haute fréquence de l'amplificateur.

Dans ce schéma, apparaissent les capacités parasites introduites dans le chapitre 4 ainsi que la capacité de charge C_L de l'étage, somme de la capacité d'entrée de l'étage suivant et de la capacité de la ligne de liaison. Ce schéma conduit à modifier le schéma équivalent petits signaux comme il est indiqué *figure 7.16*.

Ce schéma peut se représenter de manière plus lisible *figure 7.17*.

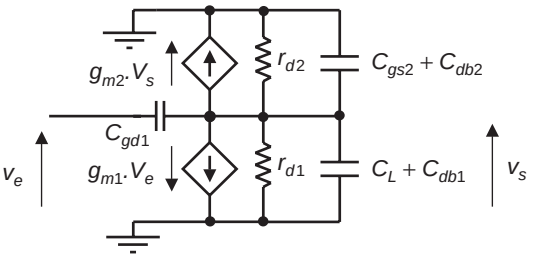


Figure 7.16 – Schéma petits signaux.

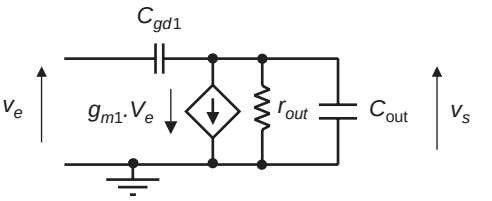


Figure 7.17 – Schéma équivalent de l'amplificateur.

Dans ce schéma les valeurs des paramètres électriques sont :

$$r_{out} = (g_{d1} + g_{d2} + g_{m2})^{-1}$$
$$C_{out} = C_{db1} + C_{db2} + C_{gs2} + C_L$$

La tension de sortie s'exprime alors en approximant r_{out} par $1/g_{m2}$. Les détails du calcul ne sont pas donnés dans ce paragraphe.

$$\frac{v_s}{v_e} = -g_{m1} \cdot r_{out} \frac{1 - \dfrac{s}{g_{m1}/C_{gd1}}}{1 + \dfrac{s}{g_{m2}/(C_{out} + C_{gd1})}}$$

La limitation en bande passante apparaît au dénominateur de la fonction de transfert dans le terme $g_{m2}/(C_{out} + C_{gd1})$ homogène à une pulsation. Les effets intégrateurs de la capacité drain-grille du transistor T1 et de la capacité de charge de l'étage (C_L) sont clairement mis en évidence. Le terme au numérateur est un zéro de la fonction de transfert et correspond à des fréquences beaucoup plus élevées.

Dans une technologie 0,18 micron, les ordres de grandeur sont les suivants :

$$g_{m2} = 150\,\mu s$$
$$C_L = 1\,pF$$

La fréquence de coupure est alors de 30 MHz. Pour une valeur de 50 fF de C_L, elle serait de 600 MHz. Il suffit d'augmenter la taille du transistor pour augmenter cette fréquence car la transconductance g_{m2} augmente avec la largeur du transistor que ce soit en régime de canal long ou en régime de canal court.

Pour terminer ce paragraphe, il est possible d'évoquer un autre schéma qui utilise cette fois un NMOS saturé comme charge.

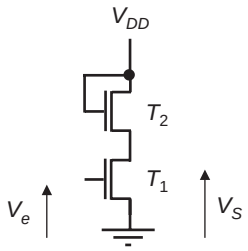


Figure 7.18 – Amplificateur à charge NMOS active.

Dans ce schéma, une importante différence de potentiel apparaît entre la source et le substrat du transistor T2. En effet, Il n'est pas question de relier la source au substrat. La différence de potentiel entre source et substrat se traduit alors par une variation de la tension de seuil, comme il est expliqué dans le chapitre 4. Cet effet « body » est assez gênant, si bien que ce montage est peu utilisé.

7.2.5 *Amplificateur push-pull*

Cet amplificateur n'est pas utilisé en analogique mais en logique pour la fonction inverseur. En effet, le régime dans lequel il y a un gain est très étroit et le circuit serait délicat à stabiliser dans des applications analogiques. Le schéma électrique est indiqué *figure 7.19*.

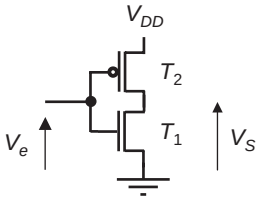


Figure 7.19 – Amplificateur push-pull.

Le signal d'entrée est appliqué simultanément sur les grilles des deux transistors, un PMOS et un NMOS. L'idée de ce montage est d'associer deux transistors qui ne peuvent conduire en même temps. La consommation statique de cet étage est donc nulle. Finalement, en régime continu aucun courant ne traverse les deux transistors en série. Ce schéma est le schéma de base de 99 % de la logique contemporaine. En réalité, si on examine le graphique de la fonction de transfert, il y a une étroite région dans laquelle les deux transistors conduisent simultanément. Le gain en tension est alors très élevé. Le graphique de la *figure 7.20* servira à déterminer la fonction de transfert. Les deux réseaux de courbes sont portés sur le même graphique en tenant compte des relations suivantes :

$$V_{DS1} + V_{SD2} = V_{DD}$$
$$V_{GS1} + V_{SG2} = V_{DD}$$

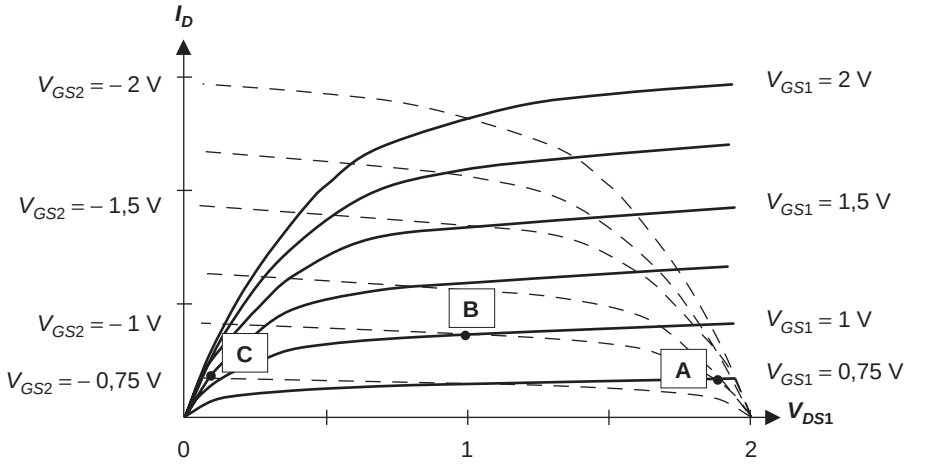


Figure 7.20 – Les caractéristiques du push-pull.

Les points A, B et C sont trois exemples de points de fonctionnement correspondant à une valeur donnée de la tension de grille. Quand la tension V_{GS1} est à 0,75 V, la tension V_{GS2} de T2 est à $-1{,}25$ V ce qui établit A comme point de fonctionnement. Quand la tension V_{GS1} est à 1 V, la tension V_{GS2} est égale à -1 V. Le point de fonctionnement est le point B. On peut construire point par point la fonction de transfert de l'étage. On obtient alors une fonction de transfert comme indiqué sur la *figure 7.21*.

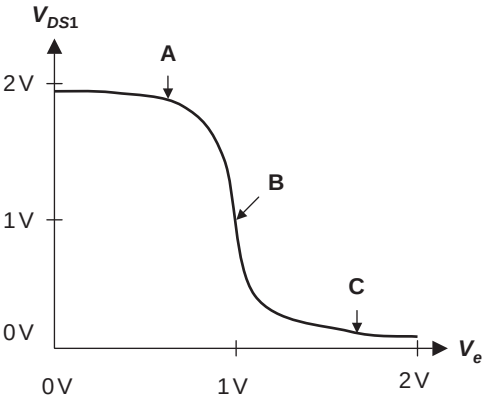


Figure 7.21 – Fonction de transfert de l'inverseur.

7.2.6 *Amplificateur différentiel*

Le principe général de ce schéma classique de l'électronique est d'amplifier non pas une tension mesurée par rapport à une référence (généralement la masse) mais une différence de deux tensions. Il faut évidemment concevoir les circuits en amont de telle sorte que le signal d'entrée puisse être fourni sous forme différentielle. L'intérêt de cette méthode d'amplification est de s'affranchir des

sources de perturbation qui affectent en général les deux entrées de la même manière. Le signal perturbateur est en effet appliqué sur les deux entrées et la différence qui apparaît est nulle. Il n'est donc pas amplifié. Le signal à prendre en compte apparaît quant à lui sous forme d'un signal de différence. Ce signal de différence est amplifié. Un mode classique de réalisation d'un amplificateur différentiel est représenté *figure 7.22*.

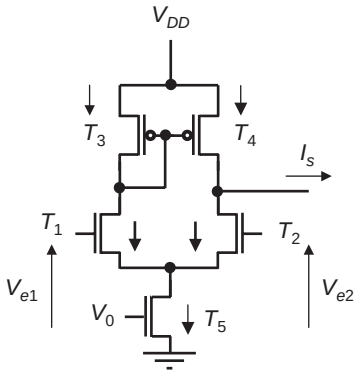


Figure 7.22 – Amplificateur différentiel.

Le fonctionnement de ce montage s'explique de la manière suivante. Le transistor T5 est une source de courant puisque sa tension de grille est fixe. Un montage à deux transistors de type cascode, comme il sera vu paragraphe 7.2.7, est également possible.

Les transistors T1 et T2 sont les transistors d'entrée du différentiel et sont chargés par les transistors T3 et T4 associés dans un montage particulier appelé miroir de courant. Ce montage a la propriété de maintenir égaux les courants I_3 et I_4 traversant les transistors T3 et T4. En effet, les courants I_3 et I_4 sont donnés par les relations :

$$\frac{I_4}{I_3} = \frac{W_4}{L_4}\frac{L_3}{W_3}\left(\frac{V_{GS4} - V_{T4}}{V_{GS3} - V_{T3}}\right)^2\left(\frac{1 + \lambda_4 V_{DS4}}{1 + \lambda V_{DS3}}\right)$$

Ces relations sont valables en régime de saturation ce qui est assuré par la connexion entre grille et drain du transistor T3. Les tensions V_{GS3} et V_{GS4} sont égales. Si les transistors ont mêmes dimensions, les courants I_4 et I_3 sont donc égaux. On suppose que les tensions drain-source des deux transistors sont voisines. On peut alors écrire les équations globales du montage :

$$I_4 = I_3$$

$$I_1 + I_2 = I_5$$

$$I_S = I_4 - I_2$$

$$I_3 = I_1$$

On en déduit :

$$I_S = I_1 - I_2$$

Exprimons maintenant les variations de ces grandeurs autour du point de fonctionnement en fonction de l'équation 4.8 du chapitre 4. Rappelons que les grandeurs en minuscules expriment des grandeurs de faibles amplitudes et que les grandeurs en majuscules sont les expressions continues correspondant au point de fonctionnement. Le coefficient β est donné par la relation :

$$\beta = \frac{W}{L}\mu C'_{OX}$$

On suppose les transistors identiques.

$$V_d = V_{e1} - V_{e2} = (V_{GS1} - V_T) - (V_{GS2} - V_T)$$

$$V_d = \left(\frac{2I_1}{\beta}\right)^{\frac{1}{2}} - \left(\frac{2I_2}{\beta}\right)^{\frac{1}{2}}$$

avec,

$$I_1 + I_2 = I_5$$

On calcule alors,

$$I_1 = \frac{I_5}{2} + \frac{I_5}{2}\left(\frac{\beta V_d^2}{I_5} - \frac{\beta^2 V_d^4}{4 I_5^2}\right)^{\frac{1}{2}}$$

$$I_2 = \frac{I_5}{2} - \frac{I_5}{2}\left(\frac{\beta V_d^2}{I_5} - \frac{\beta^2 V_d^4}{4 I_5^2}\right)^{\frac{1}{2}}$$

Ces relations montrent que les courants traversant T1 et T2 sont égaux quand les tensions appliquées sur les grilles sont égales. Quand ce n'est pas le cas, le montage est non équilibré et un des deux courants est supérieur à l'autre. Quand la différence est suffisamment importante, un des deux transistors est bloqué et tout le courant fourni par T5 passe dans le transistor conducteur. Le montage ne fonctionne plus alors en amplificateur différentiel. En résumé, ce montage n'effectue sa fonction que si la différence des tensions continues en entrée n'est pas trop importante.

La transconductance du montage est alors :

$$g_m = \frac{\partial I_1}{\partial V_d}(V_d = 0) = \left(\frac{1}{4}\beta I_5\right)^{\frac{1}{2}} \tag{7.4}$$

On peut également définir la transconductance différentielle de l'étage.

$$g_m = \frac{\partial I_S}{\partial V_d}(V_d = 0)$$

On calcule alors :

$$g_m = (\beta I_5)^{\frac{1}{2}}$$

Quelques questions peuvent se poser à propos de ce montage :

- comment sont reliés les substrats des NMOS T1 et T2 ?
- comment expliquer la dissymétrie entre T3 et T4 ?

Les transistors T1 et T2 sont en position intermédiaire dans le schéma. Il n'est pas possible de relier leurs sources au substrat dans ce type de montage. Si la technologie CMOS utilise des puits dopés *p*, les NMOS T1 et T2 ont un « body » qui peut être soit connecté aux sources des deux transistors et rester flottant soit être relié à la masse. Le choix entre les deux solutions dépend des applications. Si la technologie est à base de puits dopés *n*, le substrat est nécessairement relié à la masse et le choix n'est plus possible. Les deux sources de T1 et T2 ne peuvent être reliées au substrat et l'effet « body » est à considérer dans le fonctionnement.

La dissymétrie entre les transistors T3 et T4 est celle du miroir de courant. Cet étage a pour fonction de fournir deux courants égaux. Il ne participe pas directement à la création du gain différentiel. La symétrie de base du montage différentiel n'est donc pas brisée.

Il est possible de concevoir un amplificateur différentiel avec des PMOS en entrée comme le montre la *figure 7.23*. On pourrait penser, étant donné la différence de mobilité entre PMOS et NMOS, que cette architecture n'est pas d'un grand intérêt. À courant égal, un PMOS prend plus de place sur le circuit. En réalité, pour des raisons liées au bruit basse fréquence, il est plus intéressant quand on cherche un très faible niveau de bruit de choisir ce type de montage.

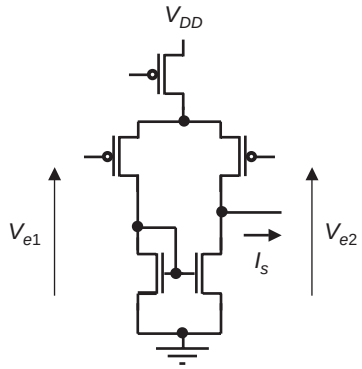


Figure 7.23 – Amplificateur différentiel avec des PMOS en entrée.

7.2.7 *Amplificateur cascode*

C'est le dernier schéma d'amplificateur que nous allons étudier. Son intérêt principal est d'obtenir une fréquence de coupure élevée comparée aux amplificateurs étudiés précédemment. Le principe de ce montage est de charger l'étage d'entrée T1 par un montage constitué par un transistor T2 monté en source de courant. Un troisième transistor T3 monté également en source de courant sert de charge au transistor T2.

Pour calculer le gain et le comportement en fréquence de cet étage, il faut déterminer son schéma équivalent en petits signaux. Pour simplifier le calcul, les sources de tension dues au potentiel de caisson du transistor T2 seront négligées. On obtient alors :

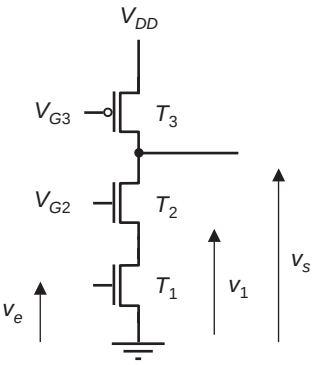


Figure 7.24 – Amplificateur cascode.

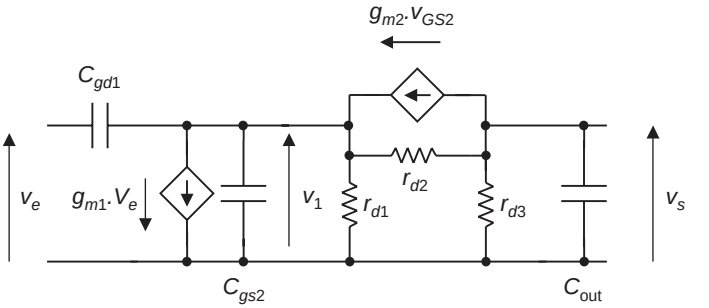


Figure 7.25 – Schéma petits signaux du cascode.

La tension v_{GS2} est égale à l'inverse de la tension v_1. En effet, en petits signaux, la tension de grille de T2 étant à un potentiel constant, elle est supposée à la masse. De même, la tension de grille du PMOS T3 étant à un potentiel constant ainsi que sa source, la source de courant liée du transistor T3 est nulle. Il est alors possible d'écrire les lois de Kirchhoff à partir de la *figure 7.24* aux deux nœuds du circuit.

$$(g_{m2} + g_{d1} + g_{d2} + sC_{gs2} + sC_{dg1})\, v_1 - g_{d2} v_s = -g_{m1} v_e + sC_{dg1} v_e$$
$$-g_{d2} v_1 + (g_{d2} + g_{d3} + sC_{\text{out}}) v_s = g_{m2} v_1$$

Ce système de deux équations à deux inconnues peut alors se résoudre en supposant les transconductances g_m très supérieures aux conductances g_d et en faisant l'hypothèse qu'un pôle est très supérieur à l'autre dans le polynôme du second degré obtenu au dénominateur. On obtient alors après un certain nombre de manipulations algébriques :

$$\frac{v_s}{v_e} = -\frac{g_{m1}}{g_{d3}} \frac{1 - \dfrac{s}{z_1}}{\left(1 + \dfrac{s}{p_1}\right)\left(1 + \dfrac{s}{p_2}\right)} \tag{7.5}$$

$$p_1 = \frac{g_{d3}}{C_{\text{out}}}$$

$$p_2 = \frac{g_{m2}}{C_{dg1} + C_{gs2}}$$

$$z_1 = \frac{g_{m1}}{C_{dg1}}$$

Le pôle le plus bas est p_1 car la conductance g_{d3} est faible et car la capacité C_{out} de sortie peut être élevée, en fonction de l'étage suivant. Le montage cascode permet de régler de manière quasi-indépendante le gain $gm1/g_{d3}$ et la bande passante g_{d3}/C_{out}.

Le véritable intérêt du montage cascode n'a cependant pas encore été réellement mis en évidence. Pour cela, il est nécessaire de supposer que la tension d'entrée n'est pas un simple générateur de tension mais une source présentant une impédance interne *Rs*. Le schéma petits signaux est alors légèrement modifié.

Avant de reprendre le calcul du montage cascode et à titre de comparaison, voyons l'effet de cette résistance *Rs* dans un montage plus élémentaire, par exemple l'amplificateur chargé par une simple source de courant décrit dans le paragraphe 2.3. Le schéma équivalent est présenté *figure 7.26*.

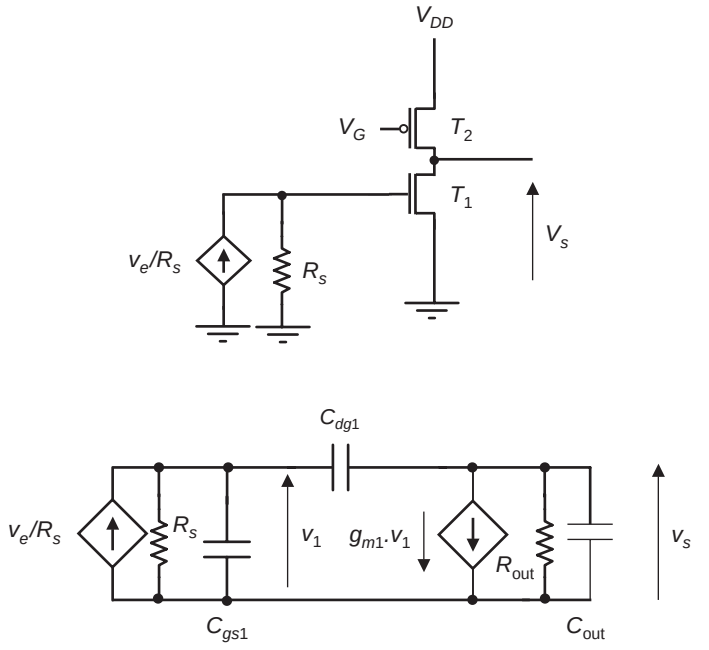


Figure 7.26 – Effet Miller.

Dans ce schéma, la résistance R_{out} est la résistance résultant de la mise en parallèle des résistances R_{d1} et d_{d2}.

Un calcul classique appliqué aux deux nœuds du circuit permet d'écrire :

$$(g_s + sC_{dg1} + sC_{gs1})v_1 - sC_{dg1}v_s = \frac{v_e}{R_s}$$

$$-sC_{dg1}v_1 + (g_{out} + sC_{out} + sC_{dg1})v_s = -g_{m1}v_1$$

La résolution de ce système conduit au résultat suivant :

$$\frac{v_s}{v_e} = -g_{m1}R_{out}\frac{1 - \dfrac{s}{z_1}}{\left(1 + \dfrac{s}{p_1}\right)\left(1 + \dfrac{s}{p_2}\right)}$$

La technique du pôle dominant a été utilisée pour faciliter la factorisation du dénominateur et les hypothèses habituelles sont faites sur les valeurs des conductances. On obtient alors :

$$p_1 = \frac{1}{g_{m1}R_sC_{dg1}}\frac{1}{R_{out}}$$

$$p_2 = \frac{g_{m1}C_{dg1}}{C_{gs1}C_{dg1} + C_{gs1}C_{out} + C_{gs1}C_{out}} \tag{7.6}$$

$$z_1 = -\frac{g_{m1}}{C_{dg1}}$$

On vérifie bien que la valeur de p_1 est très inférieure à p_2 et z_1. En effet, la capacité drain-grille du transistor T1 est multipliée par le gain en tension en basse fréquence de l'étage ($g_{m1}R_{out}$) dans l'expression de la constante de temps ce qui limite considérablement la bande passante de ce montage. **Rappelons que la capacité drain-grille est une capacité parasite due à l'imperfection de la technologie et qu'il est impossible de la réduire par le choix du point de fonctionnement.** Cet effet fondamental dans le design des circuits est appelé effet Miller.

Il serait possible de faire des calculs équivalents pour les autres types d'étages étudiés précédemment. Les résultats sont équivalents. Résumons le résultat obtenu : Quand un étage est commandé par une source de tension d'impédance de sortie R_S, la constante de temps d'intégration est le produit de cette résistance par la capacité vue en entrée. Cette capacité est le produit de la capacité parasite drain-grille par le gain en tension de l'étage.

Voyons maintenant en quoi le cascode permet d'augmenter la bande passante. Pour cela, reprenons le schéma petits signaux du cascode de la *figure 7.25* et calculons le gain en tension v_1/v_e intervenant dans l'effet Miller. Ce gain est faible, d'une valeur de $-2g_{m1}/g_{m2}$ environ. Ce calcul est proposé au lecteur à titre d'exercice. L'effet Miller est donc considérablement réduit. En reprenant le calcul du gain du montage cascode et en faisant un bon nombre d'approximations on obtient le résultat suivant. Si les deux transistors T1 et T2 sont de mêmes dimensions, le pôle dominant devient :

$$p_1 = \frac{1}{R_s(C_{gs1} + 3\,C_{dg1})} \tag{7.7}$$

La bande passante est donc considérablement augmentée. Il suffit de comparer la valeur de ce pôle à celle obtenue dans le montage précédent.

$$p_1 = \frac{1}{g_{m1} R_s C_{dg1}} \frac{1}{R_{\text{out}}} \tag{7.8}$$

Le montage cascode présente cependant l'inconvénient de placer trois transistors en série entre la tension d'alimentation et la masse, ce qui peut poser des problèmes de dynamique de fonctionnement quand la tension d'alimentation est trop faible. Ce cas est malheureusement inévitable avec les nouvelles technologies de la micro-électronique.

7.3 Le commutateur analogique

La fonction de commutation est également importante en électronique. Elle permet de diriger un signal d'un point à un autre du circuit, c'est la fonction de démultiplexage. La commutation est également nécessaire dans la synthèse des filtres à capacités commutées qui ont remplacé dans une large mesure les filtres passifs ou actifs traditionnels. La fonction de commutation du transistor est une simple application de la propriété de conduction commandée par la tension de seuil. La *figure 7.27* montre comment un MOS canal *n* peut servir de commutateur.

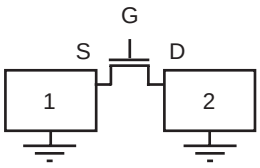


Figure 7.27 – Le MOS en commutateur.

Une tension positive sur la grille suffit à rendre le MOS conducteur. Les deux circuits 1 et 2 sont alors électriquement reliés. Il faut cependant s'assurer que le transistor est correctement polarisé. Pour que la résistance présentée par le MOS soit faible, il faut que le transistor ne soit pas en régime saturé. Les tensions présentes sur la source et le drain du transistor doivent donc être inférieures à la tension de grille d'une valeur au moins égale à la tension de seuil. La relation courant-tension s'écrit alors :

$$I_D = k \frac{W}{L} \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right]$$

L'inverse de la dérivée du courant par rapport à la tension drain-source fournit la valeur de la résistance du commutateur.

$$r_{\text{on}} = \frac{L}{W} \frac{1}{k} \cdot \frac{1}{V_{GS} - V_T - V_{DS}} \tag{7.9}$$

Cette formule simple montre l'influence de la taille du transistor et l'effet de la tension de grille. Quand la tension de grille est inférieure à la tension de seuil, le courant est idéalement nul. En fait, un faible courant de fuite dû au courant sous le seuil et dû également aux courants drain-bulk et source-bulk, traverse le dispositif. De plus, la résistance de conduction n'est jamais très faible comme on pourrait l'espérer d'un commutateur. Le MOS est donc un commutateur de qualité médiocre et ses imperfections devront être prises en compte.

Un autre problème sérieux dans l'utilisation du MOS en commutateur est lié aux charges injectées à chaque commutation de grille. Les voies de couplage sont les diverses capacités présentes dans le MOSFET et les capacités des connexions. Une méthode est parfois utilisée pour réduire cet effet : la méthode du transistor « dummy ». La *figure 7.28* illustre cette technique. Un transistor court-circuité est ajouté en série au transistor actif et des transitions en opposition de phase sont appliquées sur les grilles. Des charges opposées sont alors induites par les transitions en opposition de phase appliquées sur les deux grilles. Elles se compensent en partie.

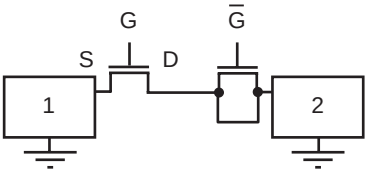


Figure 7.28 – Méthode du transistor « dummy ».

Un autre schéma est souvent proposé pour gagner cette fois en dynamique. Il associe deux transistors complémentaires commandés en opposition de phase : *figure 7.29*.

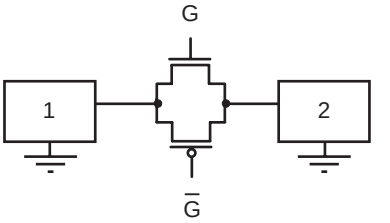


Figure 7.29 – Transistors complémentaires.

7.4 L'optimisation du rapport signal sur bruit

Les amplificateurs intégrés sont mis en œuvre pour amplifier des signaux faibles issus de capteurs ou pour amplifier le signal collecté par une antenne radio. La minimisation du bruit est donc un aspect fondamental. Le problème se traite de deux façons : la première est de réduire les sources de bruit, la seconde est de filtrer le mieux possible pour ne retenir que les composantes utiles du signal. En pratique, les deux approches sont liées et l'optimisation est globale.

L'origine du bruit s'explique par le fonctionnement des composants électroniques. Dans les ouvrages consacrés aux techniques de filtrage, il est souvent défini de manière assez théorique et caractérisé par sa densité spectrale de puissance qui est sa valeur dans une bande de fréquences données. Il est cependant utile de passer un peu de temps sur son origine, ce qui permet de mieux faire le lien entre théorie et origine physique. Dans cette approche, on considère simplement que le bruit est un signal. Sa particularité est d'être aléatoire et de faible niveau.

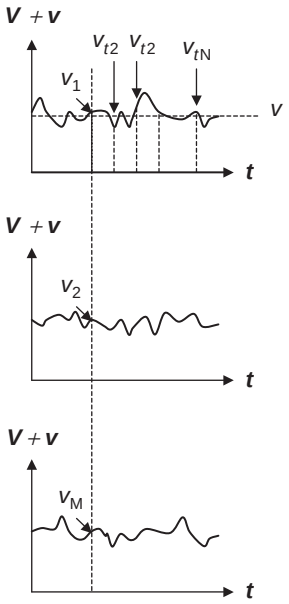
7.4.1 *Présentation phénoménologique du bruit électronique*

L'étude du bruit est un domaine présenté comme difficile en électronique. Son importance est cependant indiscutable. Ce sont les considérations relatives au rapport signal sur bruit qui sont à l'origine des architectures électroniques permettant de construire les fonctions analogiques et numériques de base. Si l'électronique était sans bruit, de nombreuses solutions seraient équivalentes.

L'expérience montre qu'en plaçant une sonde de mesure en un point quelconque d'un circuit, il se superpose au signal utile un signal de faible niveau *v* appelé bruit. On considère un signal de tension car c'est la mesure la plus commune mais des mesures de courant, de charge ou de temps présenteraient la même propriété. Dans le cas d'une mesure de tension, on observe donc des variations dans le temps du niveau mesuré comme le montre la *figure 7.30*.

Le système électronique au lieu de prendre en compte un niveau V donné à l'instant t prendra en compte la somme $V + v$. Pensons à la tension d'entrée d'une porte logique pour mesurer l'importance de ce phénomène si, par exemple, l'écart v est du même ordre de grandeur que la tension de seuil. On observe une variation aussi bien de l'amplitude que de la forme de ce signal qui est dit aléatoire. Le caractère aléatoire est dû aux multiples causes qui contribuent à la formation du bruit comme il sera expliqué dans les paragraphes suivants. On comprend donc qu'il ne sera pas possible, pour caractériser le signal de bruit, de définir autre chose que des grandeurs moyennes.

Les grandeurs moyennes peuvent se définir de manière statistique ou dans le temps comme le montre la *figure 7.30*. La moyenne statistique ou espérance mathématique est la moyenne de v sur plusieurs systèmes rigoureusement identiques, au même temps. On pourrait par exemple construire 100 amplificateurs identiques ayant le même signal en entrée et mesurer à un instant donné les 100 valeurs de la tension de sortie.



$$E(v) = \frac{v_1 + v_2 + \ldots + v_M}{M} = v_0$$

$$\bar{v} = \frac{v_{t_1} + v_{t_2} + \ldots + v_{t_N}}{N}$$

$$var(v) = \frac{(v_1 - v_0)^2 + (v_2 - v_0)^2 + \ldots + (v_M - v_0)^2}{M}$$

$$\bar{v}^2 = \frac{(v_{t_1} - \bar{v})^2 + (v_{t_2} - \bar{v})^2 + \ldots + (v_{t_N} - \bar{v})^2}{N}$$

Figure 7.30 – Le bruit et ses caractéristiques.

La moyenne statistique se définit alors par :

$$E(v) = \frac{1}{M}\sum_{i=1}^{i=M} v_i(t) \tag{7.10}$$

La moyenne est nulle si on étudie le signal de bruit créé par effet thermique car les signaux vont dans un sens ou dans l'autre. Elle est non nulle si on étudie le courant total résultant d'une multitude de contributions élémentaires. La moyenne est alors la valeur centrale du courant. Ces exemples seront détaillés dans les paragraphes suivants.

On peut également effectuer cette moyenne en prenant cette fois différentes valeurs obtenues sur le même système mais à des temps différents. On obtient alors la relation suivante :

$$\overline{v} = \frac{1}{N}\sum_{i=1}^{i=N} v_{t_i} \tag{7.11}$$

On considère généralement que ces deux manières de calculer conduisent au même résultat pour les sources de bruit envisagées en électronique. On dit alors que le bruit est ergodique. Il est possible de calculer facilement la moyenne temporelle pour un système électronique donné. Le calcul serait très complexe sans cette hypothèse. La moyenne n'est cependant pas la caractéristique intéressante. Elle est d'ailleurs souvent nulle pour beaucoup de sources de bruit. Ce qui compte réellement, ce sont les dispersions introduites par le bruit.

On définit donc la variance de la variable aléatoire v comme la moyenne des carrés des écarts à la moyenne.

$$\text{var } v = \frac{(v_1 - E(v))^2 + (v_2 - E(v))^2 + \ldots + (v_M - E(v))^2}{M} \tag{7.12}$$

De la même manière que pour la moyenne, il est plus simple de calculer cette grandeur dans le domaine temporel. On obtient alors :

$$\overline{v}^2 = \frac{(v_{t_1} - \overline{v})^2 + (v_{t_2} - \overline{v})^2 + \ldots + (v_{t_N} - \overline{v})^2}{N} \tag{7.13}$$

La variance du bruit est également une mesure de la puissance du signal de bruit.

Les exemples qui suivent montrent que ces grandeurs peuvent se calculer en tenant compte du contenu en fréquence des signaux et du filtrage. De manière qualitative, un signal de bruit tel que celui représenté *figure 7.30* peut être considéré comme une somme de sinusoïdes de fréquences différentes. C'est le principe de la décomposition d'un signal selon l'intégrale de Fourier. Il y aura d'autant plus de sinusoïdes de fréquences élevées que des transitions temporelles rapides seront présentes dans le signal de bruit. Si ce signal est filtré c'est-à-dire conduit dans un dispositif qui élimine un domaine de fréquences, certaines de ces sinusoïdes ne seront pas transmises et en conséquence la puissance de ce bruit mesurée par la variance sera réduite. L'influence de la bande passante sur le niveau de bruit se comprend donc assez facilement.

Pour terminer ce paragraphe d'introduction, insistons sur le fait que la grandeur importante n'est pas le bruit mais le rapport signal sur bruit. Rien ne sert de réduire le bruit si le signal à prendre en compte est réduit de la même manière. Pour mesurer le rapport signal sur bruit on calcule le rapport entre la puissance du signal et la variance du bruit. En sortie d'un amplificateur on mesure la puissance du signal par le carré de la valeur maximale de la tension de sortie et on mesure le bruit par

la variance de la partie aléatoire de la tension de sortie. La racine carrée de la variance est appelée valeur RMS du bruit.

7.4.2 *Bruit de grenaille d'un courant*

La première source de bruit est due aux fluctuations d'un courant franchissant une barrière de potentiel. C'est le cas du courant inverse d'un détecteur franchissant la barrière de potentiel de la jonction. C'est le cas du courant de base ou d'émetteur d'un transistor, franchissant dans un cas la barrière de potentiel inverse et dans l'autre cas la barrière de potentiel directe. C'est aussi le cas du courant créé par une photocathode. Il faut alors dans tous ces cas considérer le courant comme formé par la somme des courants créés par le passage d'une charge élémentaire, électron ou trou, à travers la barrière de potentiel. Cette somme n'est pas déterministe car les impulsions élémentaires sont créées irrégulièrement dans le temps. Cette dispersion temporelle est à l'origine de la fluctuation du signal créé. Pour calculer le bruit on se place donc en sortie de la voie considérée. On considère toutes les sources de bruit et chaque source est assimilée à une série d'impulsions de courant, chacune d'entre elles comportant une charge égale à la charge élémentaire *e*. On définit la fonction de transfert entre la source de courant et la tension en sortie.

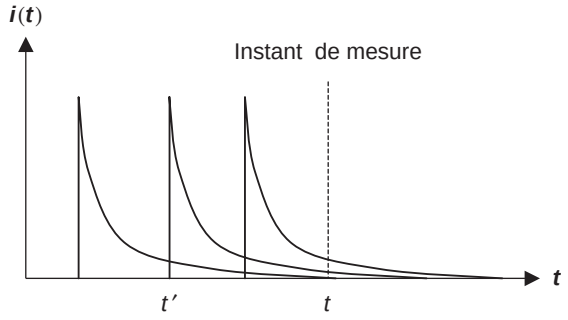


Figure 7.31 – Le bruit de grenaille.

Chaque courant élémentaire créé au temps t' s'écrit alors, comme le montre la *figure 7.31* :

$$i(t) = e \cdot \delta(t - t')$$

Dans cette relation $\delta(t - t')$ est la fonction de Dirac créée au temps t'. Le signal en sortie d'une voie électronique de fonction de transfert donnée s'écrit donc :

$$V(t) = \int_{-\infty}^{t} n \, \mathrm{d}t' \cdot e \, h(t - t')$$

Dans cette relation, le nombre d'impulsions de Dirac créées pendant $\mathrm{d}t'$ est $n \, \mathrm{d}t'$. C'est une variable aléatoire de valeur 1 ou 0, la valeur 0 étant la plus probable. La fonction h est la réponse de la voie électronique à une impulsion de Dirac, on l'appelle réponse percussionnelle. La variable $V(t)$ est aléatoire car $n \, \mathrm{d}t'$ est aléatoire. On peut calculer espérance mathématique et variance.

$$E[V(t)] = \int_{-\infty}^{t} \bar{n} \, \mathrm{d}t' \cdot e \, h(t - t')$$

L'espérance d'une somme est en effet la somme des espérances et l'espérance de $n\,\mathrm{d}t'$ est $\bar{n}\,\mathrm{d}t'$, $\bar{n}$ étant le nombre moyen d'impulsions créées par unité de temps. On calcule alors facilement :

$$E[V(t)] = \bar{n}\,e\int_{-\infty}^{t} h(t-t')\,\mathrm{d}t' = \bar{n}\,e\int_{0}^{\infty} h(u)\,\mathrm{d}u$$

La grandeur que nous venons de calculer est représentative du courant moyen formé. Dans ce cas, elle est non nulle mais on pourrait la soustraire du signal total pour obtenir un signal aléatoire de moyenne nul. On calcule de même la variance de la tension.

$$\mathrm{var}[V(t)] = \int_{-\infty}^{t} \mathrm{var}(n\,\mathrm{d}t')\cdot e^2\,h^2(t-t')$$

On utilise le fait que la variance d'une somme est la somme des variances quand les variables ne sont pas corrélées et on se sert de la propriété :

$$\mathrm{var}(aX) = a^2\,\mathrm{var}X$$

Il est facile de calculer la variance de $n\,\mathrm{d}t'$.

$$\mathrm{var}(n\,\mathrm{d}t') = \bar{n}\,\mathrm{d}t'\cdot(1-\bar{n}\,\mathrm{d}t')^2 + (1-\bar{n}\,\mathrm{d}t')\cdot(0-\bar{n}\,\mathrm{d}t')^2$$

soit,

$$\mathrm{var}(n\,\mathrm{d}t') = \bar{n}\,\mathrm{d}t'$$

En définitive,

$$\mathrm{var}[V(t)] = \bar{n}\,e^2\int_{-\infty}^{t} h^2(t-t')\mathrm{d}t'$$

soit,

$$\mathrm{var}[V(t)] = \bar{n}\,e^2\int_{0}^{\infty} h^2(u)\mathrm{d}u$$

On constate que l'espérance et la variance de la tension $V(t)$ sont indépendantes du temps, ce qui traduit la propriété de stationnarité du bruit. On peut également exprimer ces grandeurs en fonction de la fonction de transfert harmonique $H(f)$, transformée de Fourier de la réponse percussionnelle. On utilise la propriété que la norme d'une fonction est égale à celle de sa transformée de Fourier.

$$\mathrm{var}[V(t)] = \bar{n}\,e^2\int_{-\infty}^{+\infty} |H(f)|^2\mathrm{d}f$$

Cette relation s'exprime le plus souvent en tenant compte du fait que $|H|^2$ est une fonction paire :

$$\mathrm{var}[V(t)] = 2\,\bar{n}\,e^2\int_{0}^{\infty} |H(f)|^2\mathrm{d}f \qquad (7.14)$$

Le terme $2\,\bar{n}\,e^2$ est appelé densité spectrale du bruit. Il s'écrit aussi :

$$N_I(f) = 2\,e\cdot I \qquad (7.15)$$

La valeur du courant est en effet :

$$I = e\cdot\bar{n}$$

La valeur du courant à prendre en compte dans cette formule est la valeur continue du courant inverse du détecteur, la valeur du courant de base ou d'émetteur d'un transistor bipolaire, la valeur du courant d'émission d'une photodiode, le courant de fuite ou le courant tunnel dans un MOS. Il ne faut pas prendre en compte le courant traversant une résistance car il n'y a pas de barrière de potentiel transformant le courant en flux aléatoire. Dans un circuit électronique, les sources de bruit de ce type sont multiples.

7.4.3 *Bruit thermique d'une résistance*

Une deuxième source de bruit est le bruit aux bornes d'une résistance. Une observation sensible de la tension aux bornes d'une résistance traversée par un courant continu montre que cette tension n'est pas strictement constante mais affectée de légères fluctuations. Ces fluctuations sont indépendantes du courant continu mais augmentent avec la température. On peut facilement imaginer qu'elles sont dues aux déplacements aléatoires des électrons dans la résistance sous l'effet de la température. Comme ces déplacements sont quelconques, la valeur moyenne du signal de tension induit est nul. La variance par contre est non nulle et d'autant plus forte que la valeur de la résistance est élevée. Il est possible de montrer que ces fluctuations sont créées par une série aléatoire d'impulsions élémentaires de tension $\pm e \cdot R \cdot \delta(t - t')$, comme le représente la *figure 7.32*.

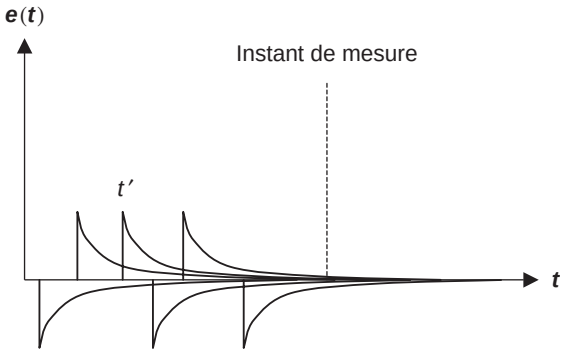


Figure 7.32 – Bruit thermique d'une résistance.

Ces impulsions sont en nombre moyen :

$$\bar{n} = \frac{2\,k_B T}{e^2 R}$$

Dans cette relation, k_B est la constante de Boltzmann égale à $1{,}38 \times 10^{-23}\ \mathrm{J \cdot K^{-1}}$. Il y a autant d'impulsions positives que d'impulsions négatives. La démonstration de cette formule sort du cadre de cet ouvrage. De la même manière que pour le bruit de grenaille, la variance du bruit en sortie d'une chaîne de mesure définie par sa réponse percussionnelle $h(t)$ s'écrit :

$$\mathrm{var}[V(t)] = \int_{-\infty}^{t} \mathrm{var}(n\,\mathrm{d}t')\,e^2 R^2 h^2(t - t')$$

La fonction h exprime la tension de sortie quand une tension $R\,e\,\delta(t - t')$ est appliquée en entrée au niveau de la résistance. Elle s'écrit alors $R\,e\,h(t - t')$. C'est une fonction de transfert tension-tension

et non plus courant-tension comme précédemment. Les contributions des impulsions positives et négatives sont égales. Compte tenu du nombre moyen d'impulsions par unité de temps, on obtient :

$$\text{var}[V(t)] = \int_0^\infty \frac{2\,k_B T}{e^2 R}\, e^2 R^2 h^2(t)\,\mathrm{d}t$$

soit,

$$\text{var}[V(t)] = 2\,k_B T R \int_0^\infty h^2(t)\,\mathrm{d}t$$

Dans le domaine fréquentiel,

$$\text{var}[V(t)] = 4\,k_B T R \int_0^\infty |H(f)|^2\,\mathrm{d}f \tag{7.16}$$

Le terme $4\,k_B TR$ est la densité spectrale de bruit de la résistance. Il faut également noter que l'espérance mathématique de la tension de bruit est nulle puisque les impulsions positives et négatives sont en nombre égal et créent des signaux identiques en valeur absolue.

De manière plus générale, on peut définir la densité spectrale d'une source de bruit quelconque pour obtenir une grandeur dépendant de la fréquence. La source de bruit génère un signal de forme générale $n(t)$. On suppose que ce bruit a deux propriétés : l'ergodicité et la stationnarité. Tout d'abord, il faut préciser que le signal de bruit est, en général, le signal observé après avoir soustrait le signal déterministe du signal total. La valeur moyenne est donc nulle.

L'ergodicité exprime que les propriétés statistiques peuvent être obtenues aussi bien en considérant différentes réalisations de ce signal à un temps donné qu'en considérant diverses valeurs obtenues à des temps différents.

La stationnarité exprime que les propriétés statistiques ne dépendent pas de la tranche temporelle d'examen choisie quand on raisonne sur une réalisation temporelle.

Nous allons raisonner par la suite sur un intervalle de temps allant de 0 à T. Cette valeur T doit être suffisamment élevée pour que les variations les plus lentes du bruit soient prises en compte. On fera donc tendre cette valeur vers l'infini. Le signal de bruit source étant noté $n(t)$ et le signal de bruit en sortie de la chaîne $V(t)$, on obtient pour la variance du bruit :

$$\text{var}(V) = \frac{1}{2T} \cdot \int_{-T}^{T} V(t)^2\,\mathrm{d}t$$

Cette formule donne une estimation de la variance, d'autant plus exacte que la valeur T est élevée. Introduisons maintenant le signal $n_T(t)$ égal au bruit de $-T$ à $+T$ et nul ailleurs. Ce signal a une transformée de Fourier $n_T(f)$. En introduisant la transformée de Fourier $V(f)$ du signal de sortie et en tenant compte de la conservation de la norme, on obtient :

$$\text{var}(V) = \frac{1}{2T} \cdot \int_{-\infty}^{+\infty} |V(f)|^2\,\mathrm{d}f$$

Le système étant linéaire de fonction de transfert $H(f)$, on obtient en assimilant le bruit et le signal $n_T(t)$, ce qui n'est vrai que pour les grandes valeurs de T.

$$\text{var}(V) = \int_{-\infty}^{+\infty} \frac{1}{2T} |n_T(f)|^2 \cdot |H(f)|^2\,\mathrm{d}f$$

La densité spectrale de puissance du bruit est alors définie par :

$$N(f) = \frac{1}{2T} \cdot |n_T(f)|^2 \tag{7.17}$$

La valeur de T tendant vers l'infini, la densité spectrale de bruit devient indépendante du choix de la période d'estimation T. Le bruit en sortie s'exprime alors par :

$$\text{var}(V) = \int_{-\infty}^{+\infty} N(f) \cdot |H(f)|^2 \mathrm{d}f$$

Les grandeurs étant paires, la variance s'exprime par :

$$\text{var}(V) = \int_{-\infty}^{+\infty} 2\,N(f) \cdot |H(f)|^2 \mathrm{d}f$$

C'est généralement ce terme $2\,N(f)$ qui est exprimé dans la littérature. Dans la suite de cet ouvrage, nous utiliserons cette définition de la densité spectrale de bruit.

7.4.4 *Bruit d'un transistor*

L'analyse est plus complexe et fait référence au fonctionnement du composant. Nous n'exposerons ici que les résultats. Les transistors sont classés en deux familles, les transistors bipolaires et les transistors à effet de champ. Les transistors à effet de champ sont eux mêmes classés en deux catégories, les MOSFET et les JFET. Il est d'usage d'exprimer les termes de bruit sous forme d'une tension de bruit et d'un courant de bruit comme le montre la *figure 7.33*.

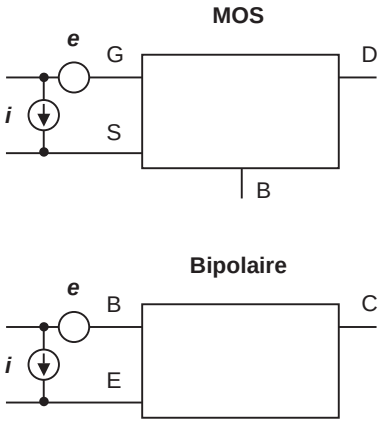


Figure 7.33 – Modèle de bruit du transistor.

Le principe général est de chercher une représentation sous forme de schéma équivalent qui conserve le schéma du transistor sans bruit et intègre les sources de bruit du transistor sous forme d'une source de tension unique et d'une source de courant unique en entrée. Pour en arriver à ce schéma, il faut identifier la source de bruit en un point donné du transistor puis ramener en entrée cette valeur divisée par la fonction de transfert entre l'entrée et ce point où est créé le bruit. Ce principe

général conduit à des calculs assez complexes. Nous décrirons uniquement le bruit thermique généré dans le canal d'un MOS.

Le principe du calcul est de considérer le bruit thermique $\Delta e\,(x_1)$ généré par un élément dR de résistance du canal de conduction placé à une abscisse x_1. La variation $\Delta i\,(x_1)$ du courant de drain causé par la tension de bruit peut se calculer en imaginant le système comme deux transistors en série.

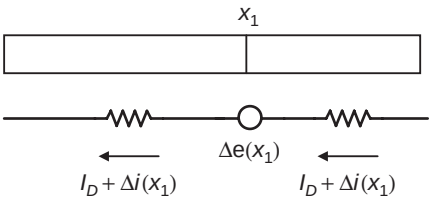


Figure 7.34 – Calcul de l'effet d'un élément de bruit thermique.

On écrit alors pour les parties gauche et droite de la *figure 7.34* en supposant le transistor en régime de forte inversion.

$$I_D + \Delta i = -\frac{W}{x_1}\mu_n \int_{V_{SB}}^{V_1} Q_I'[V_{CB}(x)]\,\mathrm{d}V_{CB}$$

$$I_D + \Delta i = -\frac{W}{L - x_1}\mu_n \int_{V_1 + \Delta e}^{V_{DB}} Q_I'[V_{CB}(x)]\,\mathrm{d}V_{CB}$$

En supposant que Δe est très petit, on peut établir la relation suivante en éliminant x_1 dans les deux équations précédentes.

$$I_D + \Delta i = -\frac{W}{L}\mu_n \int_{V_{SB}}^{V_{DB}} Q_I'[V_{CB}(x)]\,\mathrm{d}V_{CB} + \frac{W}{L}\mu_n Q_I'[V_{CB}(x_1)]\Delta e$$

On reconnaît dans cette relation l'expression du courant de drain, il reste donc :

$$\Delta i = \frac{W}{L}\mu_n Q_I'[V_{CB}(x_1)]\Delta e$$

Ce raisonnement établi pour une tension Δe continue reste vrai en régime quasi-statique dans un domaine de fréquence discuté dans le chapitre 4. Il est maintenant possible de calculer la densité spectrale de la variable aléatoire Δi. Il faut pour cela calculer la résistance de l'élément de canal de conduction considéré. L'expression générale du courant de drain s'écrit :

$$I_D = -W\mu_n Q_I'[V_{CB}(x_1)]\frac{\mathrm{d}V_{CB}(x)}{\mathrm{d}x_1}$$

L'élément de résistance est donc :

$$\mathrm{d}R = \frac{\mathrm{d}x_1}{-W\mu_n Q_I'[V_{CB}(x_1)]}$$

La densité spectrale de la variation de courant est donc :

$$N_{\Delta i}(f) = \left(\frac{W}{L}\mu_n Q_I'[V_{CB}(x_1)]\right)^2 4k_BT\frac{\mathrm{d}x_1}{-W\mu_n Q_I'[V_{CB}(x_1)]}$$

soit,

$$N_{\Delta i}(f) = -4k_BT\frac{W}{L^2}W\mu_n Q_I'[V_{CB}(x_1)]\mathrm{d}x_1$$

La variance du courant de drain est alors :

$$\mathrm{var}\Delta i = -4k_BT\frac{W}{L^2}W\mu_n Q_I'[V_{CB}(x_1)]\mathrm{d}x_1\Delta f$$

Remarquons que la bande de fréquence dans laquelle est calculée le bruit Δf n'est pas un infiniment petit comme $\mathrm{d}x_1$. Il reste maintenant à intégrer sur tous les éléments de résistance du canal en supposant que les bruits sont non corrélés ce qui n'est pas une hypothèse évidente puisque les mêmes charges traversent le canal.

$$\mathrm{var}\, I_D = -4k_BT\frac{1}{L^2}\mu_n\Delta f\int_0^L Q_I'(x_1)W\mathrm{d}x_1$$

On obtient alors une expression dépendant de la charge totale du canal.

$$\mathrm{var}\, I_D = -4k_BT\frac{1}{L^2}\mu_n Q_I\Delta f \tag{7.18}$$

Il est possible de reprendre les résultats du chapitre 4, en particulier la formule 4.21 pour obtenir finalement :

$$\mathrm{var}\, I_D = 4k_BT\frac{W}{L}\mu_n C_{OX}'(V_{GS}-V_T)\frac{2}{3}\cdot\frac{1+\alpha+\alpha^2}{1+\alpha}\Delta f$$

En régime de saturation, α est nul et on obtient simplement :

$$\mathrm{var}\, I_D = 4k_BT\frac{W}{L}\mu_n C_{OX}'(V_{GS}-V_T)\frac{2}{3}\Delta f$$

Il est maintenant possible de ramener ce bruit en entrée sous forme d'une tension de bruit si on applique la relation entre le courant et la tension grille-source en petits signaux :

$$i_d = g_m v_{GS}$$

La valeur de g_m en régime de forte inversion est indiquée dans le chapitre 4, formule 4.29.

$$g_m = \frac{W}{L}\mu_n C_{OX}'\frac{V_{GS}-V_T}{1+\delta}$$

La densité spectrale de la source de bruit représentant le bruit thermique dans le canal est donc :

$$N_e(f) = 4k_BT\frac{(1+\delta)^2}{\frac{W}{L}\mu_n C_{OX}'(V_{GS}-V_T)}\cdot\frac{2}{3}\frac{1+\alpha+\alpha^2}{(1-\alpha)(1-\alpha^2)} \tag{7.19}$$

Le cas ($\alpha = 1$) conduit à une densité spectrale infinie. Dans ce cas, la transconductance est nulle et l'opération consistant à ramener le bruit en entrée n'a plus de sens physique. On retient en général le cas du régime saturé et $\alpha = 0$. La densité spectrale de la tension de bruit s'écrit alors :

$$N_e(f) = 4 k_B T \frac{(1+\delta)^2}{\frac{W}{L}\mu_n C'_{OX}(V_{GS} - V_T)} \cdot \frac{2}{3} \tag{7.20}$$

On définit souvent dans la littérature la résistance de bruit. C'est une résistance fictive qui génère un bruit équivalent au bruit ramené en entrée. Dans le cas du régime saturé sa valeur est égale à :

$$R_e = \frac{(1+\delta)^2}{\frac{W}{L}\mu_n C'_{OX}(V_{GS} - V_T)} \cdot \frac{2}{3} \tag{7.21}$$

De manière simplifiée, δ est supposé nul et on écrit :

$$R_e = \frac{2}{3}\frac{1}{g_m}$$

Que deviennent ces valeurs de bruit en régime de faible inversion ?

Il faut revenir à la relation générale (7.18) exprimant le bruit du courant de drain et remplacer la valeur de la charge d'inversion par son expression en régime de faible inversion. On supposera que la relation 7.18 est valable en faible inversion, ce qui n'est pas évident. La charge d'inversion s'écrit :

$$Q_I = WL\left(\frac{Q'_{I\text{source}}}{2} + \frac{Q'_{I\text{drain}}}{2}\right)$$

À partir de cette relation, on peut calculer le courant de drain. Le calcul n'est pas détaillé dans cet ouvrage.

$$Q_I = \frac{L^2}{2\mu_n \phi_t} I_D \left(1 + \exp^{-\frac{V_{DS}}{\phi_t}}\right)$$

On en déduit l'expression du bruit :

$$\text{var}\, I_D = 2 e I_D \left(1 + \exp^{-\frac{V_{DS}}{\phi_t}}\right) \Delta f \tag{7.22}$$

Cette formule est remarquable car si on se limite aux valeurs de la tension de drain très supérieures à ϕ_t c'est-à-dire 26 mV, la densité spectrale du bruit thermique est la même que celle du bruit d'un courant franchissant une barrière de potentiel comme il a été vu dans le paragraphe 5.1.

Tout se passe comme si le courant de drain franchissait une barrière de potentiel et générait un bruit de grenaille. Ce résultat est assez remarquable dans la théorie du bruit des semi-conducteurs. On peut également ramener ce bruit sous forme d'une tension de bruit en entrée et la densité spectrale est dans ce cas :

$$N_e(f) = \frac{2 e n^2 \phi_t^2}{I_D} \frac{1 + \exp^{-\frac{V_{DS}}{\phi_t}}}{\left(1 + \exp^{-\frac{V_{DS}}{\phi_t}}\right)^2} \tag{7.23}$$

Pour terminer cette présentation du bruit des transistors, il est possible de résumer les résultats principaux dans un tableau simplifié. Rappelons que l'analyse du bruit du transistor bipolaire n'est pas faite dans cet ouvrage mais déroule une méthodologie équivalente à celle que nous avons mise en œuvre dans l'étude du MOS.

Deux exemples simples de schémas équivalents sont représentés, le premier correspond au transistor bipolaire et le second au transistor à effet de champ. Tension de bruit et courant de bruit sont définis par leur densité spectrale de puissance. Les valeurs des densités spectrales sont indiquées dans le *tableau 7.1*.

Le seul élément nouveau est l'introduction du bruit en 1/*f*. Les origines physiques de ce bruit sont multiples et complexes. Les auteurs s'accordent pour souligner l'importance des pièges à l'interface oxyde-semi-conducteur mais les fluctuations de mobilité jouent également un rôle. Le résultat final est l'apparition d'une densité spectrale très élevée pour les basses fréquences et une proportionnalité à l'inverse de la surface du dispositif. Il faut également noter que ce type de bruit est quasi-nul pour les transistors bipolaires ce qui leur confère un avantage indéniable de ce point de vue. Les effets de ce bruit sont, par principe, dans le domaine des basses fréquences. Il y a cependant des conséquences indirectes comme par exemple l'augmentation du bruit de phase des oscillateurs commandés en tension.

Tableau 7.1

	Transistor bipolaire	**Transistor à effet de champ en régime saturé**
Courant de bruit	$N_i(f) = 2eI_B$	$N_i(f) = 2eI_{GS}$
Tension de bruit	$N_e(f) = 4k_BT\left(r_{bb'} + \frac{0{,}5}{g_m}\right)$	$N_e(f) = 4k_BT \cdot \frac{0{,}7}{g_m} + \frac{1}{g_m^2}\frac{K_F I_D^k}{f \cdot C'_{OX} \cdot L^2}$

Dans le *tableau 7.1*, les variables sont définies de la manière suivante :

- I_B : courant de base du transistor bipolaire, dépendant de son point de polarisation, en pratique de quelques micro-ampères.
- $r_{bb'}$: résistance répartie de base du transistor bipolaire, de quelques dizaines d'ohms à quelques centaines d'ohms.
- g_m : transconductance du transistor, c'est-à-dire le rapport du courant de sortie à la tension d'entrée. Dans le cas du transistor bipolaire, sa valeur est pour un courant d'émetteur I_E :

$$g_m = \frac{eI_E}{k_BT} \tag{7.24}$$

- I_{GS} : courant grille-source du transistor à effet de champ, généralement de quelques nano-ampères.
- g_m : transconductance du transistor à effet de champ. Elle dépend du point de polarisation.
- K_F : constante du bruit dit en 1/*f*. L'origine de ce bruit est complexe mais liée aux effets de surface et d'interface. Dans une technologie 0,8 micron, K_F vaut 4×10^{-28} F.A pour le NMOS et $0,5 \times 10^{-28}$ F.A pour le PMOS.

La densité spectrale du bruit en $1/f$ s'exprime également en remplaçant la transconductance par sa valeur et en supposant que k a une valeur voisine de 1. La mobilité μ est celle des électrons pour le NMOS et celle des trous pour le PMOS. Les valeurs données en exemple et la différence de trois entre les mobilités font que le PMOS est, dans ce cas, quatre fois moins bruyant que le NMOS en basse fréquence.

$$N_i(f) = \frac{K_F I_D^k}{f \cdot C'_{OX} \cdot L^2} \tag{7.25}$$

En divisant par le carré de la transconductance, on obtient :

$$g_m^2 = 2\mu_n C'_{OX} \frac{W}{L} I_D$$

$$N_e(f) = \frac{K_F}{2f \cdot C'^2_{OX} \cdot WL\mu_n} \tag{7.26}$$

Cette formule très approchée permet cependant de constater que le bruit basse fréquence des transistors MOS est plus élevé que celui des bipolaires et que les transistors MOSFET offrent de ce point de vue les propriétés les moins bonnes. Il faut cependant tenir compte du fait que le filtrage atténue en général une grande partie du bruit basse fréquence.

Pour calculer le bruit en sortie d'un circuit quelconque, il faut sommer les contributions des résistances et des éléments actifs en utilisant les formules précédentes. La fonction de transfert à prendre en compte pour chaque terme est le ratio entre la tension de sortie et la source de bruit envisagée, représentée soit par une source de courant, soit par une source de tension. Il est facile de comprendre que seuls les étages d'entrée contribuent au bruit de manière significative, puisque le gain intervient dans la fonction de transfert. En pratique, les deux premiers étages d'une voie électronique contribuent au bruit.

7.4.5 *Mesure de charge et filtre adapté*

Les principales sources de bruit étant définies, il est possible de définir la chaîne électronique optimale. Avant toute chose, il faut définir précisément la quantité à optimiser. Dans de nombreux cas, c'est le rapport signal sur bruit en sortie dans la mesure d'une charge.

Le signal d'entrée de la voie électronique est représenté par son schéma équivalent sous forme d'une source de courant et d'une capacité équivalente. La *figure 7.35* représente une voie d'amplification de la manière la plus générale possible. Le lecteur peut s'interroger sur la pertinence de cette approche dans la mesure où les paragraphes précédents ont principalement traité l'amplification en tension.

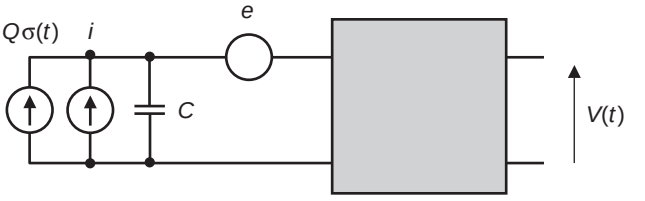


Figure 7.35 – Une voie d'amplification.

De nombreux phénomènes physiques conduisent à définir une charge plus qu'un courant ou une tension. L'information contenue dans un point mémoire l'est sous forme d'une charge. Un transistor qui change d'état commute la charge de son canal. Un détecteur transforme une énergie en une charge. La charge est souvent la grandeur physique la plus pertinente car elle est reliée à l'énergie.

Le signal de courant créé peut s'écrire :

$$i(t) = Q \cdot s(t)$$

Dans cette relation Q est la charge mise en jeu dans le processus. Le bruit est ramené en entrée sous forme d'une source de courant et sa densité spectrale est $N(f)$. Quelle est alors la fonction de transfert qui maximise le rapport signal sur bruit en sortie pour un temps de mesure donné ? Nous verrons qu'il est possible de transformer la source de tension de bruit en une source de courant de bruit et que finalement tout le bruit peut se représenter comme une source de courant.

Intuitivement, on imagine que la solution est celle qui laisse passer les fréquences constitutives du signal et élimine celles du bruit. Un calcul mathématique assez simple permet de préciser les choses. Le signal en sortie, mesuré au temps t_m, s'écrit, si $H(f)$ est la fonction de transfert et si $S(f)$ est la transformée de Fourier du signal d'entrée :

$$V(t_m) = Q\int_{-\infty}^{+\infty} S(f) \cdot H(f) \cdot \exp^{2\pi i \cdot f t_m} \mathrm{d}f$$

Cette formule est simplement la transformée de Fourier inverse. La variance du bruit en sortie s'écrit :

$$\mathrm{var}(V) = \int_0^{\infty} N(f) \cdot |H(f)|^2 \, \mathrm{d}f$$

Le rapport signal sur bruit est :

$$\left(\frac{S}{N}\right)^2 = \frac{V(t_m)^2}{\mathrm{var}(V)}$$

Cette quantité doit être maximisée. Elle s'exprime sous forme du quotient de deux intégrales. On utilise ensuite un théorème mathématique appelé inégalité de Schwartz. On trouve alors que l'optimum est atteint quand :

$$H(f) = k\frac{S(f)^*}{N(f)} \cdot \exp^{-2\pi i \cdot t_m}$$

Dans cette relation, k est une constante multiplicative quelconque et $S(f)^*$ est le complexe conjugué de la transformée de Fourier du signal. Transformons maintenant le schéma d'entrée de la manière suivante, comme le montre la *figure 7.36*.

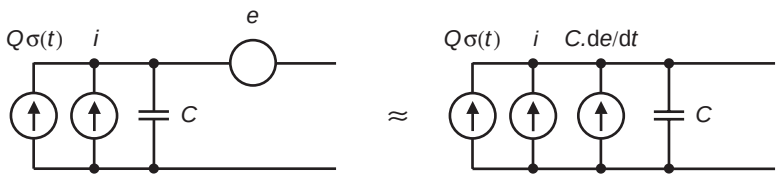


Figure 7.36 – Sources de bruit en entrée.

Pour obtenir la densité spectrale de bruit à partir des sources de courant et de tension, il suffit de sommer quadratiquement les deux composantes non corrélées. Pour faire intervenir la même fonction de transfert, on transforme par le théorème de Thévenin la source de tension e en une source de courant $C \cdot \mathrm{d}e/\mathrm{d}t$ comme le montre la *figure 7.36*. Il suffit pour cela de calculer le courant débitant dans une charge nulle. Tout se passe alors comme si le courant de bruit en entrée était la somme de deux termes :

$$i_b = i + C \cdot \frac{\mathrm{d}e}{\mathrm{d}t}$$

On montre alors que densité spectrale de bruit associée à cette source est donnée par la relation :

$$N(f) = N_i + 4\pi^2 C^2 f^2 \cdot N_e$$

La démonstration de cette formule est proposée au lecteur comme exercice. Elle utilise le fait que la transformée de Fourier d'une dérivée est le produit de la transformée de Fourier par $2\pi if$.

La synthèse de la fonction de transfert adaptée ou d'une fonction approchée est donc le principe de base de la conception d'une chaîne électronique de traitement. Il faut ajouter comme contrainte la nécessité d'amplifier le signal, ce qui dimensionne la constante k.

En résumé, un modèle simplifié permet d'obtenir les conclusions principales sur les chaînes d'amplification. Il suppose le signal en entrée très rapide, idéalement représenté par une impulsion de Dirac. On suppose également que la tension de bruit est équivalente au bruit d'une résistance de valeur R. Ce n'est qu'une manière de représenter les choses. On suppose également la présence en entrée d'un bruit de grenaille créé par un courant I. La densité spectrale du bruit ramené en entrée s'écrit alors :

$$N(f) = 2e \cdot I + 4\pi^2 C^2 f^2 \cdot 4k_B TR$$

La transformée de Fourier du signal d'entrée est Q. Le filtre optimum s'écrit donc :

$$H(f) = k' \frac{\exp^{-2\pi j \cdot t_m}}{eI + 4\pi^2 C^2 f^2 \cdot 2k_B TR}$$

La réponse percussionnelle du filtre est donc la transformée de Fourier inverse de la fonction de transfert soit :

$$V(t) = k' \cdot \exp^{\frac{|t - t_m|}{\tau}}$$

Le temps τ étant donné par :

$$\tau^2 = \frac{2k_B TRC^2}{eI} \tag{7.27}$$

Ce calcul demande quelques manipulations mathématiques qui ne sont pas détaillées dans cet ouvrage.

La *figure 7.37* montre la réponse idéale. Elle est acausale, ce qui veut dire que la réponse est non nulle avant même que l'entrée ne soit excitée. Le filtre idéal est donc non réalisable, du moins en temps réel. Il ne peut être réalisé que de manière approchée. La *figure 7.37* représente une approximation réalisable obtenue par troncature du filtre idéal. Ce filtre est d'autant plus proche du filtre idéal que le temps de mesure t_m est élevé.

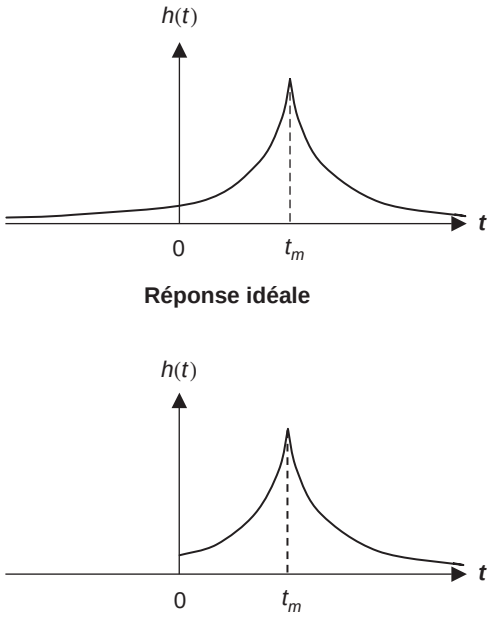


Figure 7.37 – Filtre adapté.

Les considérations précédentes permettent de définir une voie de mesure de charge ou d'énergie. Le filtre idéal n'est jamais réalisé puisqu'il est acausal, il faut donc se contenter de réalisations approximatives. En pratique, des solutions simples permettent de s'en approcher à 10 %.

Le calcul du rapport bruit sur signal peut se faire dans le cas optimum du filtre adapté. On obtient le résultat suivant :

$$\left(\frac{N}{S}\right)^2 = \frac{4\,k_B T R C^2}{Q^2}\,\frac{1}{\tau} \tag{7.28}$$

Ce résultat fondamental donne le meilleur rapport signal sur bruit que l'on puisse obtenir dans une chaîne d'amplification définie par sa résistance équivalente de bruit R et par la capacité totale C vue en entrée. Il peut sembler paradoxal que le courant I générant le bruit de grenaille n'intervienne pas dans cette relation. En fait, il intervient dans la valeur de la constante de temps optimale selon la formule suivante :

$$\tau^2 = \frac{2\,k_B T R C^2}{e\,I}$$

À l'optimum, les deux contributions sont égales, contribution venant de R et contribution venant de I. Le résultat final peut donc s'exprimer uniquement en fonction de l'une des sources de bruit.

7.4.6 *Exemples de calcul de bruit*

Le premier exemple est le calcul du bruit de l'amplificateur source commune, dont le schéma est rappelé *figure 7.38*. Les sources de bruit en sortie sont définies par les densités spectrales dans les formules données en 7.26.

Il est plus facile de raisonner à partir des densités spectrales de courant de bruit comme le montre la *figure 7.38*. Les courants de bruit en sortie des transistors sont notés i_1 et i_2. Les densités spectrales sont alors :

$$N_{i1}(f) = \frac{8}{3} k_B T \cdot g_{m1} + \frac{K_{Fn} I_D^k}{f \cdot C'_{OX} \cdot L_1^2}$$

$$N_{i2}(f) = \frac{8}{3} k_B T \cdot g_{m2} + \frac{K_{Fp} I_D^k}{f \cdot C'_{OX} \cdot L_2^2}$$

Ces deux courants créent en sortie une tension :

$$v_S = \frac{1}{g_{m2}} (i_1 + i_2)$$

On établit facilement ce résultat sur le schéma équivalent en supposant la transconductance largement supérieure aux conductances.

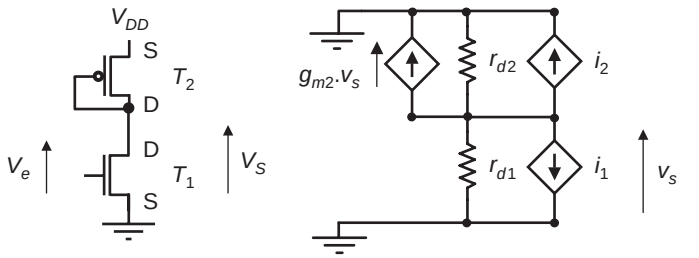


Figure 7.38 – Calcul du bruit de l'amplificateur source commune.

La tension de bruit ramenée en entrée est donc cette valeur divisée par le gain en tension du montage soit g_{m1}/g_{m2}. La densité spectrale de la tension de bruit ramenée en entrée est donc :

$$N_e(f) = \frac{1}{g_{m1}^2} \left(\frac{8}{3} k_B T \cdot g_{m1} + \frac{K_{Fn} I_D}{f \cdot C'_{OX} \cdot L_1^2} + \frac{8}{3} k_B T \cdot g_{m2} + \frac{K_{Fp} I_D}{f \cdot C'_{OX} \cdot L_2^2} \right)$$

Cette relation montre l'intérêt d'augmenter g_{m1} et de diminuer g_{m2}, c'est-à-dire d'augmenter le gain en tension de l'étage. On constate également qu'il est nécessaire d'augmenter les surfaces des deux transistors en choisissant pour les longueurs L des valeurs suffisantes.

Pour calculer le bruit total en sortie de la voie électronique, il faut ensuite calculer l'intégrale :

$$\text{var } V_s = \int_0^\infty N_e(f) |H(f)|^2 \, \mathrm{d}f$$

La fonction de transfert H est définie comme la transformée de Fourier du rapport entre la tension de sortie V_s et la tension d'entrée V_e. Le terme thermique de la densité spectrale de bruit étant indépendant de la fréquence, l'intégrale correspondante se calcule uniquement en tenant compte de $|H(f)|^2$. Pour les composantes en $1/f$, l'intégration prend en compte le ratio $|H(f)|^2/f$.

Reprenons ce calcul pour le montage le plus utilisé en entrée d'une voie électronique, l'amplificateur différentiel. Le schéma et le modèle électrique sont rappelés *figure 7.39*.

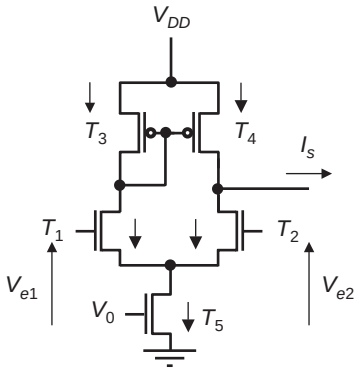


Figure 7.39 – Le bruit d'un amplificateur différentiel.

Un calcul du même type que celui effectué dans le cas du montage source commune conduit au résultat suivant.

$$N_e(f) = \frac{1}{g_m}[N_{i1}(f) + N_{i2}(f) + N_{i3}(f) + N_{i3}(f) + N_{i4}(f)]$$

Il faut noter deux choses :

- Le transistor source de courant T5 ne contribue pas au bruit. En effet, les impulsions de bruit sont considérées comme un signal de mode commun et la contribution à la sortie est nulle par effet différentiel.
- La transconductance g_m commune aux deux transistors d'entrée T1 et T2 joue un rôle majeur.

La relation précédente peut s'écrire :

$$N_e(f) = 2\,N_{e1}(f)\left[1 + \frac{g_{m3}^2}{g_{m1}^2}\frac{N_{e3}(f)}{N_{e1}(f)}\right]$$

On comprend donc l'intérêt d'augmenter la transconductance et la longueur du transistor d'entrée. Ces conclusions sont-elles valables pour les autres montages ? En toute rigueur, il faudrait effectuer un calcul équivalent dans les autres cas. Il est cependant assez intuitif de penser que plus les gains en tension sont élevés et plus les transconductances sont élevées, moins l'influence du bruit associé au courant de drain du MOS sera importante. Les conclusions relatives aux surfaces des transistors sont également généralisables. Pour limiter le bruit basse fréquence, il faut donner aux transistors des surfaces assez importantes. Cette condition conduit à optimiser non pas un rapport de dimensions mais à optimiser les dimensions absolues des transistors.

7.5 L'amplificateur opérationnel

Il n'est pas question dans ce paragraphe de détailler les multiples utilisations de l'amplificateur opérationnel car de nombreux ouvrages traitent de ce sujet (référence **[5]**). L'objectif est de relier les architectures des amplificateurs opérationnels et leurs performances aux possibilités de la technologie micro-électronique. De nombreux schémas sont possibles, et seuls les plus importants sont étudiés dans ce chapitre.

7.5.1 *Description et architecture*

L'amplificateur opérationnel est une sorte d'amplificateur idéal caractérisé par un gain en tension élevé, une grande impédance d'entrée, une faible impédance de sortie et une bande passante la plus élevée possible. Ajoutons à cela une entrée de type différentielle, c'est-à-dire deux entrées et une excellente immunité au mode commun et nous avons dressé le portrait d'une fonction idéale. En fait, obtenir un gain élevé et une grande bande passante sont deux objectifs contradictoires car le produit du gain par la bande passante est une grandeur à peu près constante pour une technologie donnée. La consommation électrique est un critère devenu essentiel au cours du temps et sa minimisation est souvent en contradiction avec l'obtention d'une bande passante élevée. Le schéma est classiquement représenté *figure 7.40*.

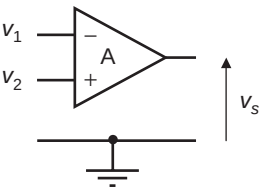


Figure 7.40 – Amplificateur opérationnel.

La relation de base de l'amplificateur opérationnel est alors :

$$v_s = A(v_2 - v_1)$$

Le facteur A est le gain dit en boucle ouverte. Cette relation est valable uniquement en petits signaux dans la zone de gain de l'amplificateur. La relation peut être plus complexe si on fait apparaître les pôles et les zéros de la fonction de transfert. On considère généralement que les courants d'entrée sont nuls car les entrées sont des grilles de MOSFET. Ce n'est plus le cas quand la technologie bipolaire est utilisée car les entrées sont des bases de transistors.

Le schéma général d'un amplificateur opérationnel est plus ou moins celui indiqué *figure 7.41*. Il comporte en entrée un étage différentiel puis un étage de gain élevé et enfin un étage de sortie capable de délivrer un courant important. Il faut y ajouter un circuit de compensation et les circuits nécessaires pour générer les tensions et les courants de polarisation.

Deux structures classiques seront détaillées : l'amplificateur à deux étages et le « folded » cascode.

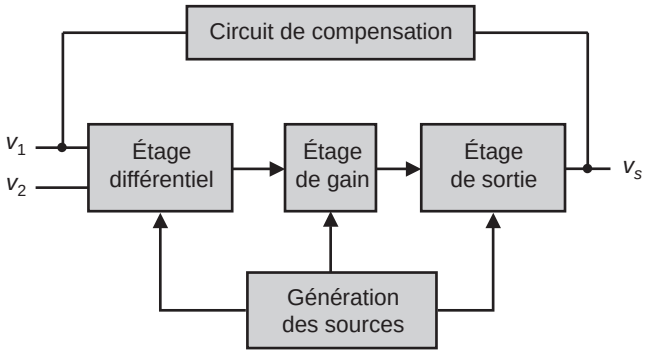


Figure 7.41 – Architecture d'un amplificateur opérationnel.

7.5.2 *Amplificateur à deux étages*

C'est le plus immédiat, il est formé de l'étage différentiel étudié dans le paragraphe 7.2 suivi d'un étage PMOS chargé par une source de courant. Le schéma est donné *figure 7.42*.

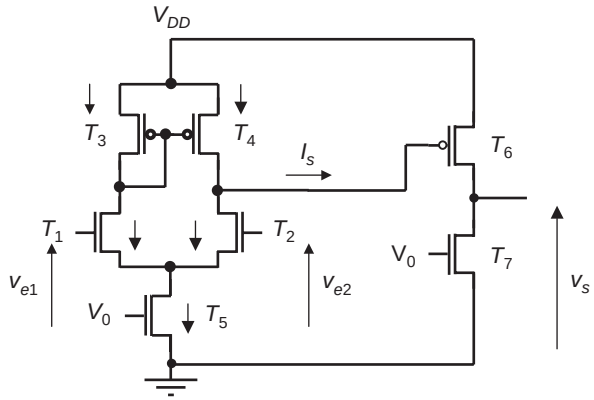


Figure 7.42 – Amplificateur à deux étages.

Les amplificateurs opérationnels sont en général utilisés en effectuant un rebouclage de la sortie sur l'entrée inverseuse. Cela permet par exemple de définir le gain en fonction des éléments de rebouclage. Le gain est alors indépendant des paramètres des transistors ce qui garantit une grande stabilité. La rétroaction peut cependant entraîner des instabilités si une valeur élevée du gain en boucle ouverte coïncide avec un déphasage de 180 degrés. Pour remédier à cela, une méthode simple est de limiter le déphasage en introduisant une fréquence de coupure plus faible, soit un pôle dans le gain en boucle ouverte.

Dans la théorie des systèmes bouclés, on montre qu'un signal appliqué de la sortie vers l'entrée avec un gain supérieur à un et déphasé de 180 degrés conduit à un état instable caractérisé par des oscillations. Cette méthode est décrite en détail dans les nombreux ouvrages traitant des amplificateurs

opérationnels (référence **[5]**). En pratique, pour l'amplificateur à deux étages, il suffit de placer une capacité entre le drain et la grille du transistor de gain T6. On en arrive alors à la *figure 7.43*.

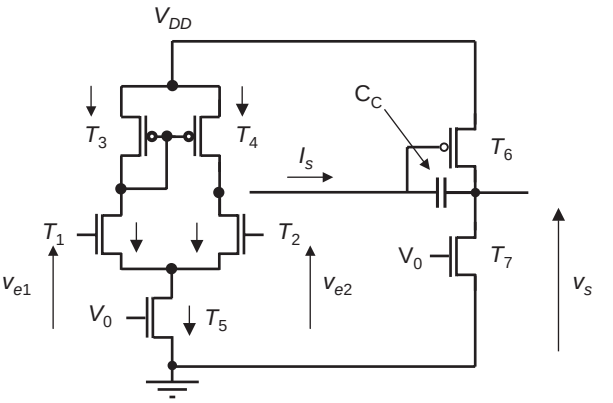


Figure 7.43 – Amplificateur compensé.

Il n'est pas utile dans cet ouvrage de donner tous les détails des techniques de compensation. Il est cependant possible de présenter l'aspect le plus élémentaire de cette méthode en considérant la représentation de Bode du gain de boucle, défini comme le produit du gain en boucle ouverte par la fonction de transfert du circuit de rétroaction. Fonctionner en boucle ouverte signifie que le réseau de contre-réaction est supprimé.

La *figure 7.44* représente le gain de boucle, produit du gain de l'amplificateur opérationnel par la fonction de transfert de l'étage de contre-réaction, en fonction de la fréquence. Les échelles sont logarithmiques. Le logarithme du module du produit est représenté en fonction du logarithme de la fréquence. Le nouveau pôle introduit évite de placer le système dans un état dans lequel un déphasage de 180 degrés est possible avec un gain supérieur à l'unité ce qui est une cause d'instabilité.

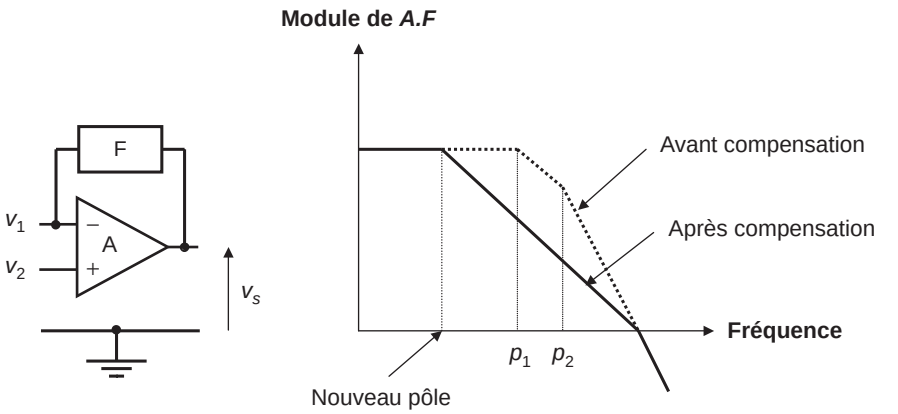


Figure 7.44 – Technique de compensation.

La conception d'un amplificateur de ce type réside dans le choix des paramètres du schéma. Ils sont au nombre de 9 :

- les facteurs de forme des 7 transistors (W/L) ;
- la tension V_0 de polarisation des sources de courant ;
- la capacité C_c de compensation.

Pour des raisons de symétrie évidentes, les facteurs de forme des transistors T1 et T2 du différentiel sont égaux ainsi que ceux de T3 et T4. Il y a donc 7 paramètres à choisir. La tension d'alimentation est supposée donnée pour une technologie donnée. Il y a donc la possibilité de satisfaire sept contraintes qui peuvent être des propriétés particulières de l'amplificateur. Ces propriétés peuvent être choisies dans la liste suivante :

- gain continu en boucle ouverte ;
- produit gain-bande passante ;
- le *slew rate* noté SR ;
- la capacité maximale de charge en sortie ;
- le temps d'établissement ;
- le ICMR (*Input Common Mode Range*) ;
- le CMRR (*Common Mode Rejection Ratio*) ;
- le PSSR (*Power Supply Rejection Ratio*) ;
- la dynamique de sortie ;
- l'impédance de sortie ;
- le niveau continu de sortie ;
- le niveau de bruit ;
- la taille sur la puce ;
- la consommation.

Quelques termes demandent explication. Le ICMR est la gamme dans laquelle il est possible d'appliquer une même tension sur les deux entrées tout en conservant les propriétés d'amplification différentielle. Le CMMR est le rapport entre le gain différentiel et le gain en mode commun, il est idéalement infini. Le PSSR mesure le rapport entre une variation de la tension d'alimentation et la variation du niveau continu de sortie associé. Il est idéalement infini. Le *slew rate* est défini pour une charge capacitive donnée, il mesure la pente maximale de la tension de sortie en fonction du temps.

Il n'est en général pas possible de respecter un jeu quelconque de valeurs de ces 14 grandeurs. Certaines sont le plus souvent des valeurs minimales à atteindre. Pour aider au choix, il est utile de donner un ensemble de relations simples entre quelques-unes de ces grandeurs et les paramètres de dimensionnement accessibles au concepteur.

L'amplificateur est divisé en deux étages *a* et *b*. Les éléments de schéma correspondant sont identifiés par les indices *a* et *b*. Le premier est l'étage différentiel et le second est l'étage de sortie.

Le calcul complet du gain de ce circuit est assez lourd et conduit au résultat suivant :

$$\frac{v_s}{v_e} = \frac{g_{ma}\, g_{mb} R_a R_b \left(1 - \dfrac{sC_c}{g_{mb}}\right)}{1 + s[R_a(C_a + C_b) + R_b(C_b + C_c) + g_{mb} R_a R_b C_c] + s^2 R_a R_b (C_a C_b + C_c C_a + C_c C_b)}$$

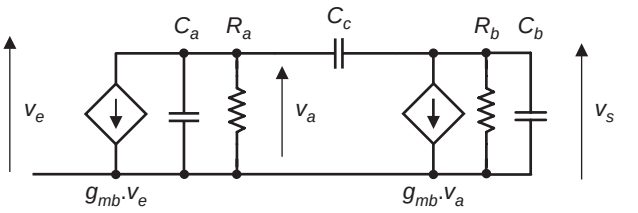


Figure 7.45 – Schéma de l'amplificateur opérationnel à deux étages.

La technique du pôle dominant (voir chapitre 2) consiste à supposer que les deux racines du dénominateur sont très différentes. Il est donc possible de simplifier la relation. Les pôles et le zéro du gain sont donc :

$$p_1 = \frac{1}{g_{mb}R_aR_bC_c}$$

$$p_2 = \frac{g_{mb}C_c}{C_aC_b + C_bC_c + C_aC_c}$$

$$z_1 = \frac{g_{mb}}{C_c}$$

En général, le pôle p_2 s'écrit simplement :

$$p_2 = \frac{g_{mb}}{C_b}$$

On peut maintenant exprimer les pôles et le zéro en fonction des paramètres des transistors du schéma. On obtient alors en supposant que la capacité C_b est la capacité de charge C_L de l'amplificateur (capacité de la piste de sortie et de l'étage suivant).

$$p_1 = \frac{1}{g_{mb}R_aR_bC_c} = \frac{(g_{d2} + g_{d4})(g_{d6} + g_{d7})}{g_{m6}C_c}$$

$$p_2 = \frac{g_{mb}}{C_b} = \frac{g_{m6}}{C_L}$$

On utilise le fait que la transconductance du deuxième étage est égale à la transconductance du transistor T6 puisque la tension de sortie de l'étage différentiel est appliquée entre grille et source de T6.

$$z_1 = -\frac{g_{mb}}{C_c} = -\frac{g_{m6}}{C_c}$$

On peut également exprimer le gain basse fréquence en tension.

$$A_0 = g_{ma}g_{mb}R_aR_b$$

soit,

$$A_0 = \frac{2\,g_{m1}}{I_5(\lambda_2+\lambda_4)}\,\frac{g_{m6}}{I_6(\lambda_6+\lambda_7)}$$

$$SR = \frac{I_5}{C_C}$$

Cette relation s'obtient en considérant que le courant maximum fourni par l'étage différentiel est le courant I_5. Ce courant est considéré comme le courant de charge de la capacité interne de contre-réaction.

À partir de la formule du gain continu et de la valeur du pôle p_1, on calcule le produit gain-bande passante :

$$GB = \frac{g_{m1}}{C_C}$$

$$p_2 = \frac{g_{m6}}{C_L}$$

$$z_1 = -\frac{g_{m6}}{C_c}$$

$$V_{e\max} = V_{DD} - \sqrt{\frac{I_5}{\beta_3}} - |V_{T3}| + V_{T1}$$

$$V_{e\min} = \sqrt{\frac{I_5}{\beta_1}} + V_{T1} + V_{DS5\text{sat}}$$

La détermination des points de fonctionnement peut alors se faire en fonction des relations précédentes. Il faut y ajouter les relations classiques des transistors saturés :

$$g_m = \sqrt{2\,k\,\frac{W}{L}I_D} = \sqrt{2\,\beta\,I_D}$$

$$g_{ds} = \lambda \cdot I_{D\text{sat}}$$

$$I_D = k\,\frac{W}{L}\,\frac{1}{2}\,(V_{GS}-V_T)^2(1+\lambda V_{DS})$$

ou sous une autre forme,

$$V_{GS} = V_T + \sqrt{\frac{2\,I_D}{\beta}}$$

Le terme $\sqrt{\frac{2\,I_D}{\beta}}$ est également la tension de saturation $V_{DS\text{sat}}$.

La première étape est le choix de la capacité de compensation C_c. Un raisonnement basé sur la stabilité du montage (choix de la marge de phase) conduit à choisir :

$$C_c > 0{,}22\,C_L$$

Cette relation empirique dépend de la capacité de charge de l'amplificateur qui est la somme des capacités vue du nœud de sortie. Ensuite la donnée de *GB* permet de déterminer g_{m1}. La valeur de SR conduit à la valeur de I_5. La relation donnant la valeur de $V_{e\max}$ en fonction de I_5/β_3 permet alors de calculer β_3 donc de déterminer le rapport de forme du transistor T3. Le rapport de forme de T1 est obtenu à partir de g_{m1} et de I_5 puisque le transistor T1 est traversé par $I_5/2$. La tension de saturation du transistor T5 soit $V_{DS5\text{sat}}$ est ensuite obtenue à partir de la relation donnant $V_{e\min}$. Ensuite, le transistor T5 peut être dimensionné à partir de la valeur de $V_{DS5\text{sat}}$ puisque son courant est déjà connu. Les dimensions des transistors T5, T1, T2, T3 et T4 sont donc fixées.

Les considérations relatives à la stabilité ont amené à placer le pôle de sortie à 2,2 fois *GB*. On peut donc écrire :

$$g_{m6} = 2{,}2\, g_{m2} \frac{C_L}{C_C}$$

Pour finir de dimensionner T6, il suffit d'écrire que le courant I_6 satisfait les contraintes de consommation statique de l'amplificateur. Le transistor T7 peut alors être dimensionné à partir de T5. Rappelons que T5 et T7 ont même tension de grille.

$$\left(\frac{W}{L}\right)_7 = \left(\frac{W}{L}\right)_5 \frac{I_6}{I_5}$$

Il faut alors vérifier que le gain en boucle ouverte a une valeur suffisante et que les taux de réjection vis-à-vis du mode commun et vis-à-vis de l'alimentation sont convenables. Si ce n'était pas le cas, il faudrait modifier les paramètres généralement en augmentant les facteurs de forme et en diminuant les courants.

Les considérations de bruit n'ont pas encore été prises en compte. Les résultats obtenus dans l'étude de l'amplificateur différentiel sont applicables. La seule chose à prendre en compte à ce stade est le lien qui existe entre la taille absolue des transistors (valeurs de *W* et de *L*) et le niveau de bruit basse fréquence. Des transistors de tailles trop faibles peuvent conduire à un bruit basse fréquence trop élevé. Notons que c'est la première fois dans notre analyse que la taille absolue des transistors intervient. Dans toutes les relations précédentes, seul comptait le rapport *W/L* soit le rapport relatif de taille des transistors. En pratique et pour réduire la taille de la puce, la valeur minimale de la longueur est choisie dans une technologie donnée et la largeur est ajustée. Les considérations relatives au bruit peuvent amener à choisir des transistors de tailles supérieures. Pour des technologies avancées comme le 45 nm, une longueur minimale de 100 nm est généralement choisie.

7.5.3 *L'amplificateur de type folded-cascode*

L'amplificateur à deux étages décrit précédemment est très utilisé, il a cependant les limitations suivantes :

- gain boucle ouverte insuffisant ;
- zone de stabilité réduite ;
- faible réjection vis-à-vis de la tension d'alimentation.

L'amplificateur de type *folded-cascode* permet d'améliorer ces trois points. Son schéma de principe est représenté *figure 7.46*.

Le schéma du miroir de courant cascodé est détaillé *figure 7.47*.

Il est alors possible de tracer le schéma équivalent à petits signaux de ce montage (*figure 7.48*).

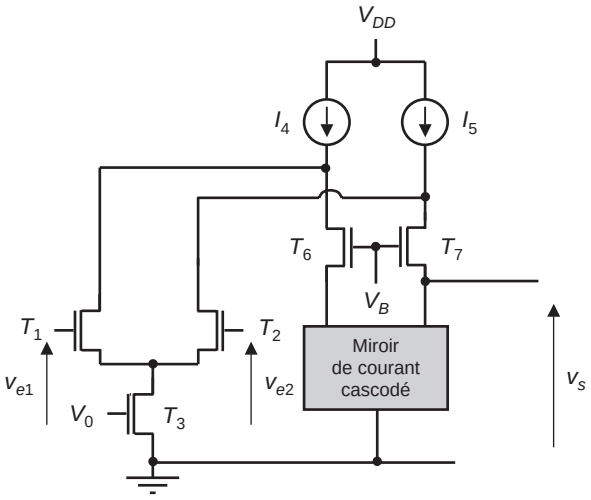


Figure 7.46 – Montage folded-cascode.

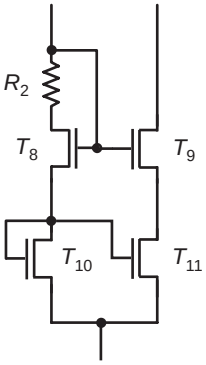


Figure 7.47 – Miroir de courant du folded-cascode.

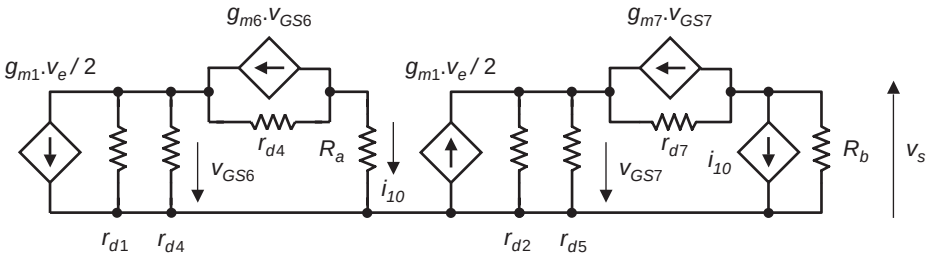


Figure 7.48 – Schéma petits signaux du folded-cascode.

Dans ce schéma, obtenu à partir des figures précédentes, les résistances *Ra* et *Rb* sont respectivement :

$$R_a = R_2 + \frac{1}{g_{m10}}$$

$$R_b = g_{m9} r_{d9} r_{d11}$$

Obtenir ce schéma et ces deux relations n'est pas immédiat mais découle des analyses précédentes. Le gain différentiel boucle ouverte se calcule alors facilement et seul le résultat sera donné :

$$\frac{v_s}{v_e} = \left(\frac{g_{m1}}{2} + \frac{g_{m2}}{2(1+k)} \right) R$$

avec,

$$R \approx g_{m9}\, r_{d9}\, r_{d11} \parallel g_{m7}\, r_{d7} \cdot \left(\frac{r_{d2}\, r_{d5}}{r_{d2} + r_{d5}} \right)$$

et,

$$k = \frac{R_b (g_{d2} + g_{d4})}{g_{m7}\, r_{d7}}$$

Pour étudier le comportement fréquentiel, on appliquera une règle empirique valable pour la CMOS : les pôles sont à chaque nœud donnés par le produit de la résistance par la capacité vue de ce nœud. Le pôle dominant est vu de la sortie. Si C_L est la capacité de charge de l'amplificateur, sa valeur est :

$$p = \frac{1}{RC_L}$$

Les autres pôles peuvent se calculer de la même façon. Il faut remarquer que, contrairement à l'amplificateur à deux étages, aucune capacité n'a été ajoutée pour limiter la bande passante. Les propriétés de réjection sont alors améliorées car la capacité de compensation de l'amplificateur à deux étages offre un chemin facile de couplage entre une variation de la tension d'alimentation et la sortie.

7.5.4 *Amplificateurs opérationnels rapides*

Si on considère un amplificateur opérationnel contre-réactionné, la bande passante est reliée au gain et au produit gain-bande passante *GB* par la relation :

$$B = \frac{GB}{G}$$

Diminuer *G* en augmentant la contre-réaction est donc le moyen le plus évident pour augmenter la bande passante. La *figure 7.49* détaille cette propriété fondamentale.

Si le gain différentiel boucle ouverte est $-A$, le gain du montage contre-réactionné est :

$$G = \frac{-A}{1 + AF}$$

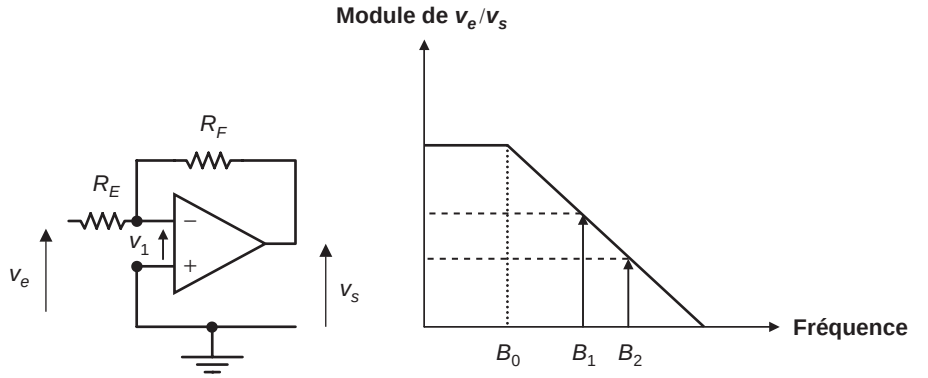


Figure 7.49 – Bande passante d'un amplificateur contre-réactionné.

On a simplement admis que les impédances d'entrée étaient infinies. Pour obtenir cette formule, on utilise les deux relations :

$$\frac{v_s}{v_1} = -A$$

et,

$$\frac{v_1 - v_e}{R_E} = \frac{v_s - v_1}{R_F}$$

$$F = \frac{R_E}{R_E + R_F}$$

Si on suppose que le gain boucle ouverte a un seul pôle p_0 alors,

$$A = \frac{A_0}{1 + \frac{s}{p_0}}$$

et donc,

$$G = \frac{-1}{F} \frac{1}{1 + \frac{s}{p_0 A_0 F}}$$

Le nouveau pôle est donc $p_0 A_0 F$ alors que le gain est $1/F$ ce qui signifie que le produit gain-bande passante est inchangé et égal à $p_0 A_0$. Le calcul précédent justifie donc la technique de contre-réaction pour augmenter la bande passante. La limite est donnée par le produit gain-bande passante *GB* égal à $p_0 A_0$.

Cette caractéristique fondamentale s'exprime de la manière suivante. Pour l'amplificateur à double étage, c'est g_{m1}/Cc tandis que pour le *folded-cascode* c'est g_{m1}/C_{out}. Rappelons que g_{m1} est la transconductance du transistor d'entrée et que C_c et C_{out} sont respectivement la capacité de compensation de l'amplificateur double étage et la capacité en sortie du *folded-cascode*.

Il est cependant indispensable de s'assurer que les autres pôles sont très au-delà du pôle principal. Comme la transconductance d'un MOSFET est de l'ordre du mA/V, le produit gain-bande passante et les bandes passantes maximales seront de l'ordre de 160 MHz par picofarad de charge. Notons au passage la supériorité des transistors bipolaires qui offrent des transconductances de l'ordre de 10 mA/V pour un même courant de polarisation.

Pour obtenir des bandes passantes élevées, la technologie des amplificateurs à contre-réaction de courant est intéressante. Le principe général est exposé *figure 7.50*.

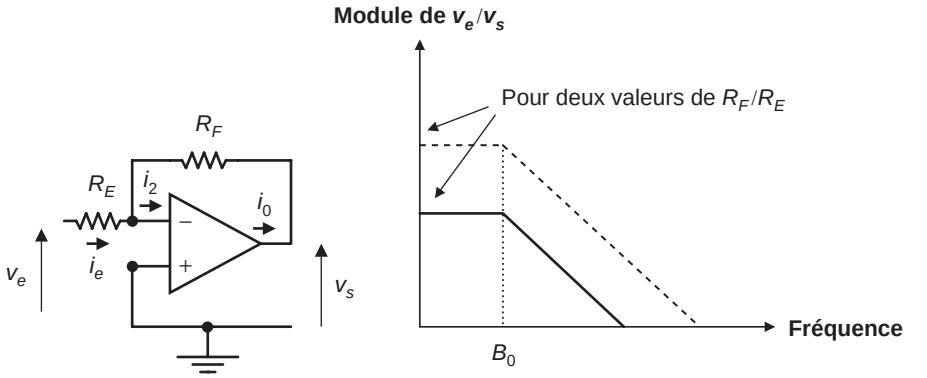


Figure 7.50 – Amplificateur à contre-réaction de courant.

La seule différence avec le montage précédent est dans la fonction de transfert de l'amplificateur qui s'écrit dans ce cas : $i_0 = -A \cdot i_2$.

On peut également écrire en supposant les deux entrées au même potentiel puisque le gain de l'amplificateur en boucle ouverte est élevé :

$$i_e = i_2 - i_0$$

$$i_e = \frac{v_e}{R_E}$$

$$i_0 = \frac{v_s}{R_F}$$

On calcule alors facilement :

$$\frac{v_s}{v_e} = \left(-\frac{R_F}{R_E}\right)\left(\frac{A}{A+1}\right)$$

Si le gain en boucle ouverte a un pôle p_0, alors le gain devient :

$$\frac{v_s}{v_e} = \left(-\frac{R_F}{R_E}\right)\left(\frac{A_0}{A_0+1}\right)\left(\frac{1}{1+\dfrac{s}{p_0(1+A_0)}}\right)$$

Il est donc possible en modifiant le rapport des résistances d'augmenter le gain tout en conservant la même bande passante, en l'occurrence le produit gain-bande passante de l'amplificateur de courant. Cette propriété est très intéressante car elle dissocie gain et bande passante contrairement au schéma précédent. La limite de l'exercice est la stabilité du montage.

7.6 Les filtres à capacités commutées

Les filtres analogiques sont difficiles à réaliser en micro-électronique étant donné la complexité ou l'impossibilité de réaliser des composants passifs ayant les valeurs et les précisions souhaitées. Il est en particulier difficile de réaliser des résistances de très fortes valeurs et il est impossible d'intégrer des condensateurs de capacités supérieures à quelques dizaines de pF. Ces contraintes sont particulièrement pénalisantes dans le domaine des basses fréquences. L'idée est donc apparue de développer d'autres techniques pour synthétiser des filtres.

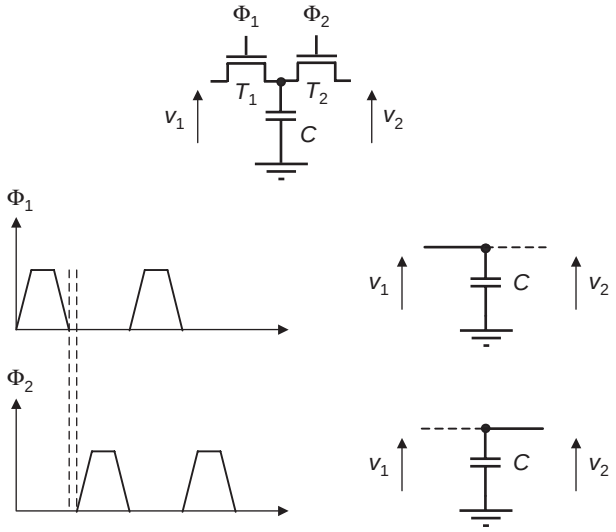


Figure 7.51 – Capacité commutée.

Le principe du filtre à capacité commutée est d'insérer une capacité *C* entre deux MOSFET fonctionnant en interrupteurs. Dans la première phase, le transistor T1 est conducteur et le transistor T2 est non passant. La capacité a donc une charge :

$$Q_1 = C\,v_1(t_1)$$

Dans la deuxième phase, après une période *T*, les tensions de commande sont inversées sur les grilles. T1 est bloqué alors que T2 est conducteur. On suppose que les signaux de commande ne se recouvrent pas comme il est indiqué sur le graphique. Le condensateur a alors une charge.

$$Q_2 = C\,v_2(t_2)$$

Le courant moyen qui a transité du nœud 1 vers le nœud 2 est donc :

$$I = \frac{C(v_1 - v_2)}{T}$$

Le système est donc équivalent à une résistance comme le montre la *figure 7.52*, avec la valeur suivante pour la résistance :

$$R = \frac{T}{C}$$

On comprend alors l'intérêt de ce système : il suffit de faire varier *T* pour choisir la résistance équivalente et donc la constante de temps d'un filtre réalisé avec cette résistance.

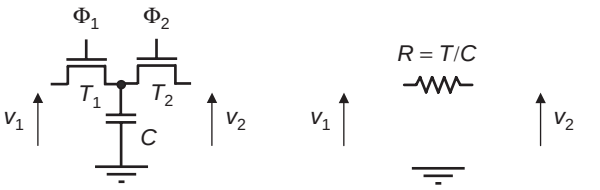


Figure 7.52 – Filtrage à capacité commutée.

Ce type de montage est cependant peu utilisé car la capacité *C* doit être largement supérieure aux capacités parasites présentes dans le circuit intégré. On lui préfère donc un autre montage qui est le montage de base des filtres à capacités commutées : l'intégrateur à capacités commutées. Le schéma de ce montage fondamental est donné *figure 7.53*.

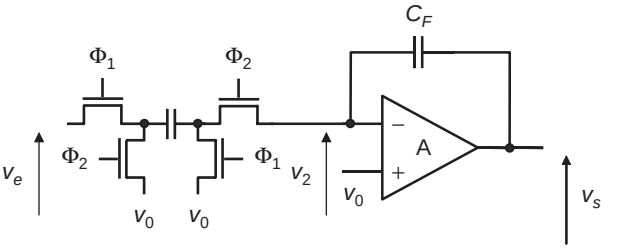


Figure 7.53 – L'intégrateur à capacités commutées.

De la même manière que pour le premier schéma, quand une tension positive est appliquée sur les grilles repérées par Φ_1, la tension $(v_e - v_0)$ est appliquée aux bornes du condensateur. Quand une tension positive est appliquée sur les grilles repérées par Φ_2, la tension $(v_2 - v_0)$ est appliquée. La différence de charge stockée correspond à la circulation d'un courant moyen égal à :

$$I = \frac{C(v_e - v_2)}{T}$$

Le système d'entrée est donc équivalent à une résistance R de valeur :

$$R = \frac{T}{C}$$

Le schéma équivalent de l'intégrateur est donc celui de la *figure 7.54*.

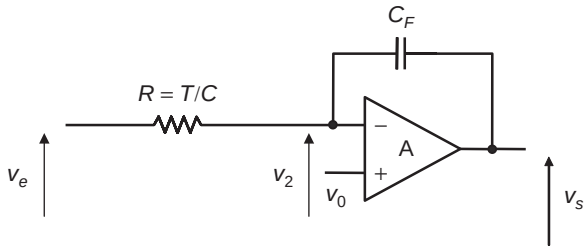


Figure 7.54 – Filtre passe-bas à capacités commutées.

La fonction de transfert s'écrit facilement en effectuant le même type de calcul que celui présenté dans l'étude de l'amplificateur contre-réactionné.

$$\frac{v_e - v_2}{R} = \frac{v_2 - v_s}{1/sC_F}$$

En supposant le gain de l'amplificateur élevé, la tension v_2 est donc égale à v_0. En petits signaux, elle est donc nulle puisque v_0 est constant.

$$\frac{v_s}{v_c} = -\frac{1}{sRC_F} = -\frac{C}{C_F} \cdot \frac{1}{sT}$$

La constante de temps d'intégration ne dépend donc que d'un rapport de capacités ce qui est relativement facile à réaliser en technologie micro-électronique.

On peut également comprendre que ce schéma fournit une valeur de la résistance équivalente indépendante des capacités parasites en étudiant la *figure 7.55* dans laquelle on a représenté deux capacités parasites par rapport à la masse.

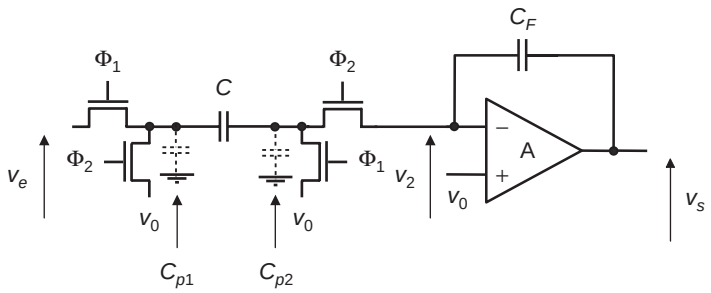


Figure 7.55 – Insensibilité de l'intégrateur actif aux capacités parasites.

Comme la capacité parasite C_{p2} est dans les deux phases de fonctionnement reliée à la tension de référence v_0, soit directement par le transistor monté en interrupteur soit à la borne négative de l'amplificateur opérationnel également à la tension de référence, la différence de charge de cette capacité est nulle entre les deux phases. Un raisonnement du même type montre que la capacité parasite C_{p1} n'intervient pas dans la formation du signal de sortie.

Pour terminer ce paragraphe, voyons comment un filtre simple est réalisé par la technique des capacités commutées. Le schéma de base et son implémentation dans une architecture à base de capacités commutées sont illustrés *figure 7.56*.

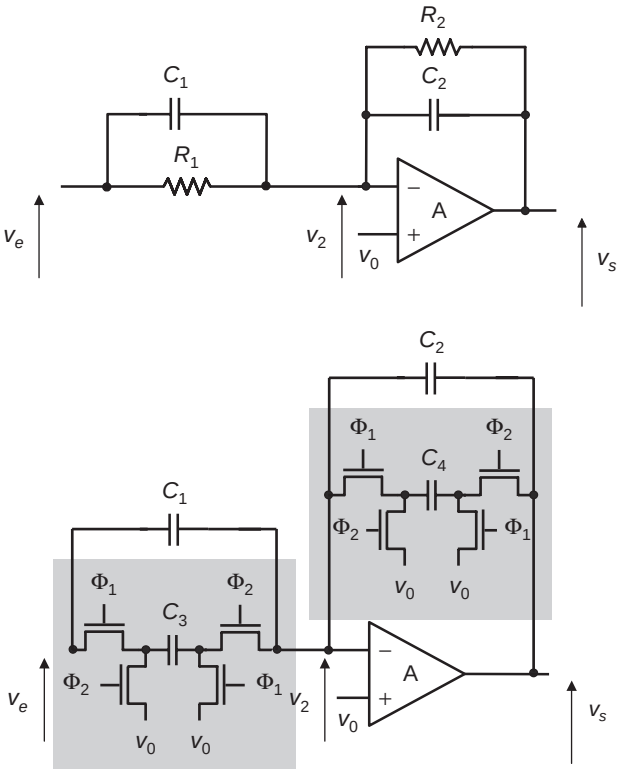


Figure 7.56 – Réalisation d'un filtre du premier ordre.

La fonction de transfert de ce filtre se calcule facilement.

$$F(s) = \frac{R_2(1 + sR_1C_1)}{R_1(1 + sR_2C_2)}$$

La technique des capacités commutées permet de le réaliser avec les capacités C_3 et C_4 commutées à la fréquence f_c, inverse de la période T.

$$R_1 = \frac{1}{C_3 f_e} \quad \text{et} \quad R_2 = \frac{1}{C_4 f_c}$$

On obtient alors :

$$F(s) = \frac{C_3}{C_4} \frac{\left(1 + s \dfrac{C_1}{C_3 f_e}\right)}{\left(1 + s \dfrac{C_2}{C_4 f_c}\right)}$$

Quelques valeurs typiques permettent de donner des ordres de grandeur. Si le pôle de la fonction de transfert doit être 1 kHz et le zéro 10 kHz, en choisissant une fréquence de commutation de 100 kHz, on obtient les rapports suivants de capacités.

$$10^3 = \frac{1}{2\pi \dfrac{C_2}{C_4} 10^{-5}}$$

$$10^4 = \frac{1}{2\pi \dfrac{C_1}{C_3} 10^{-5}}$$

On en déduit,

$$\frac{C_2}{C_4} = 16$$

$$\frac{C_1}{C_3} = 1{,}6$$

Si on impose le gain à fréquence nulle, 10 par exemple, on peut écrire la relation supplémentaire :

$$\frac{C_3}{C_4} = 10$$

La capacité la plus faible est C_4. Une valeur minimale de 100 fF est choisie. Des valeurs plus faibles pourraient être choisies en théorie. En pratique, des considérations liées au bruit, conduisent à choisir des valeurs minimales pour les capacités.

En effet, si on considère le bruit thermique créé par une résistance R quelconque mise en série avec une capacité C et si on calcule le bruit aux bornes de la capacité en supposant que la bande passante est limitée par le seul produit RC, on obtient facilement que la variance du bruit est donnée par la relation :

$$\text{var } v_R = k_B T / C$$

Ce calcul très simple est proposé comme exercice au lecteur. On obtient alors une valeur RMS du bruit (racine de la variance) égale à environ 64 µV pour un pF à température ambiante et 200 µV pour 100 fF. Ces considérations montrent qu'il est nécessaire de choisir une valeur minimale pour les capacités mises en œuvre dans les systèmes à capacités commutées.

En définitive, les valeurs correspondant à l'exemple sont les suivantes :

$C_1 = 1{,}6$ pF $\qquad C_3 = 1$ pF

$C_2 = 1{,}6$ pF $\qquad C_4 = 100$ fF

Ces valeurs sont des valeurs typiques pour des filtres à capacités commutées. Notons également que la fréquence de commutation doit être largement supérieure aux fréquences de coupure du filtre. Une valeur 10 est choisie dans l'exemple. Si la fréquence de commutation était du même ordre que les fréquences du filtre, des phénomènes de « repliement » de spectre seraient à prendre en compte et affecteraient les propriétés du filtre. Pour bien comprendre cela il faut appliquer la théorie des systèmes échantillonnés, ce qui est hors du cadre de cet ouvrage.

7.7 Comment passer de l'analogique au numérique

Dans ce dernier paragraphe, seront donnés quelques principes généraux à propos des architectures des convertisseurs intégrés dans les puces. Ces fonctions permettent de passer du mode analogique au monde numérique et ont donc une importance fondamentale.

7.7.1 *La conversion numérique-analogique*

La fonction est de convertir une série de « 0 » et de « 1 » représentant la valeur binaire d'une grandeur en une valeur électrique variant continûment, en général une tension. Quelques principes de conversion seront étudiés.

✦ Convertisseur à chaîne de résistances

C'est l'architecture la plus immédiate. Des interrupteurs à base de MOSFET commutent les points de mesure répartis le long d'une chaîne de résistances comme le montre la *figure 7.57*.

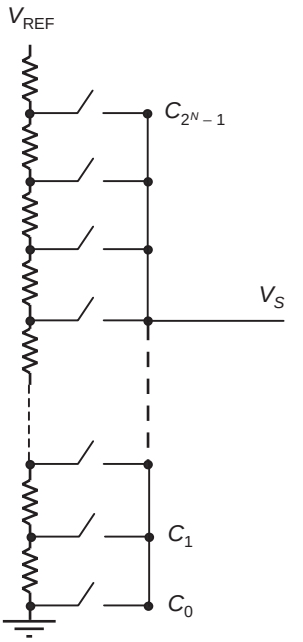


Figure 7.57 – Convertisseur numérique-analogique à chaîne de résistances.

Le code binaire de la valeur est transformé en une suite de signaux de commande D_i dont un seul est « 1 » et correspond à la valeur binaire. Le signal positif associé rend le transistor T_i conducteur et transfère le potentiel du pont de résistances vers la sortie. Les autres interrupteurs sont ouverts. Ce système simple demande cependant l'implémentation d'un grand nombre de résistances (2^N) si les données sont codées sur N bits. De plus, la forte capacité de sortie interdit des fréquences de conversion trop élevées. Il faut également inclure le circuit numérique de décodage qui transforme le code binaire en une série de signaux de commande pour les interrupteurs.

✦ Convertisseur R-2R

C'est un schéma très utilisé car il demande un nombre limité de composants. Son principe est illustré *figure 7.58*.

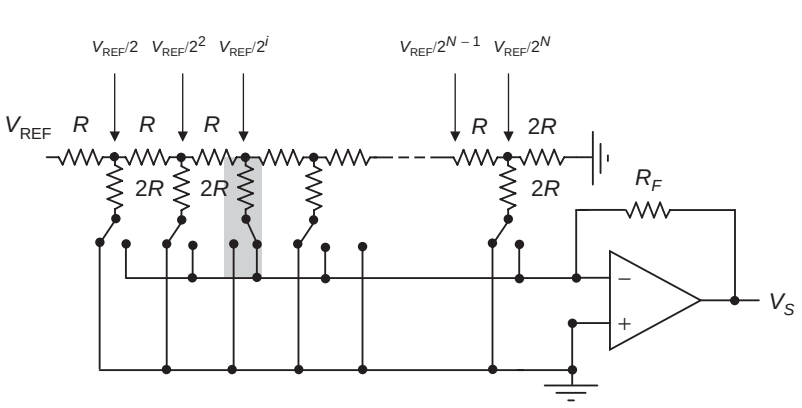


Figure 7.58 – Convertisseur R-2R.

La première chose à constater est que la tension de référence se répartit en fonction des puissances de 2. Pour le démontrer, il suffit de remarquer que la résistance à droite de chaque nœud est toujours $2R$ pour toutes les positions des interrupteurs. On commence par le nœud le plus proche de l'amplificateur et on raisonne de nœud en nœud en allant vers la gauche. Les valeurs de la tension en chaque nœud s'en déduisent alors facilement. Quand un interrupteur est positionné à droite, un courant $V_{REF}/2^i$ circule donc dans la résistance $2R$ puis dans la résistance R_F. Il crée donc une tension en sortie de l'amplificateur égale à :

$$V_{Si} = -R_F \frac{V_{REF}}{2R \cdot 2^i}$$

Chaque bit i à l'état « 1 » crée donc un signal proportionnel à 2^{-i} ou 2^{j-N} si on change l'indice. Le raisonnement s'applique pour toutes les cellules et finalement on obtient un signal de sortie proportionnel à la valeur correspondant au mot binaire servant de commande aux interrupteurs.

Ce schéma est intéressant car il n'y a pas besoin de circuit de décodage. Le nombre de composants est limité, il se prête donc bien à l'intégration.

✦ Commutation de courant

Le principe est de générer à partir de sources de courants de référence un courant commandé par des interrupteurs et correspondant à un mot binaire donné. Le schéma est donné *figure 7.59*.

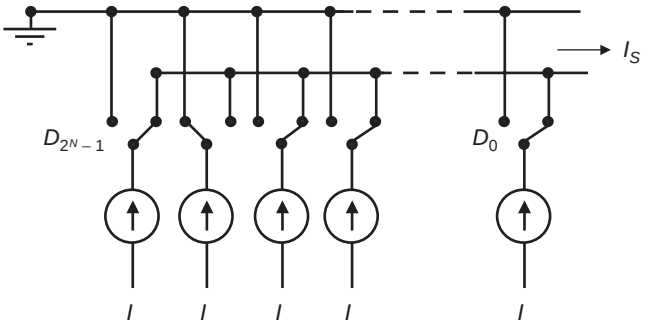


Figure 7.59 – Commutation de courant.

Dans ce cas, il faut un décodeur pour transformer le code binaire en une série de bits D_i indiquant le nombre de sources de courant à commuter. Ce code est appelé code thermométrique. Dans le cas d'un codage sur 8 bits par exemple, la valeur binaire 01000000 correspond à la valeur décimale 128. Il faut alors commuter 128 sources de courants. Pour réduire le nombre de sources et pour simplifier le décodage, les sources de courant peuvent être proportionnelles à 2^i. Dans ce cas, il suffit de commuter les sources en fonction de la valeur de i.

Ce convertisseur est intéressant à cause de sa rapidité. Les sources de courant peuvent être réalisées à partir de montages de type miroirs de courant. Sa consommation est cependant importante.

7.7.2 *La conversion analogique-numérique*

C'est la deuxième fonction de base, inverse de la précédente. Elle permet de passer du mode analogique au mode numérique. De nombreuses architectures sont possibles et nous ne citerons que les plus importantes.

✦ Le convertisseur parallèle

C'est l'architecture la plus naturelle. Elle est basée sur la comparaison du signal à coder avec une série de tensions continues échelonnées entre 0 et la tension maximale. La *figure 7.60* détaille l'architecture interne de ce type de convertisseur.

Le signal d'entrée est comparé simultanément aux $2^N - 1$ valeurs de tension réparties uniformément et fournies par le pont résistif. Les comparateurs sont des amplificateurs différentiels à gain élevé qui fournissent en sortie un signal qui ne dépend pas de l'amplitude du signal d'entrée mais uniquement de la polarité de la tension différentielle d'entrée. Le signal est inférieur ou supérieur au seuil. Un amplificateur différentiel fonctionnant en saturation permet de représenter comment travaille un comparateur.

Les signaux de sorties sont donc d'amplitudes définies, nuls ou d'une amplitude donnée V_0. Ils sont nuls quand l'amplitude du signal d'entrée est inférieure au seuil du comparateur et d'amplitude V_0 quand elle est supérieure au seuil. Ces signaux sont donc de type « logique » et sont pris en compte par le décodeur qui fournit le code binaire de sortie. Les $2^N - 1$ valeurs en sortie de la batterie de

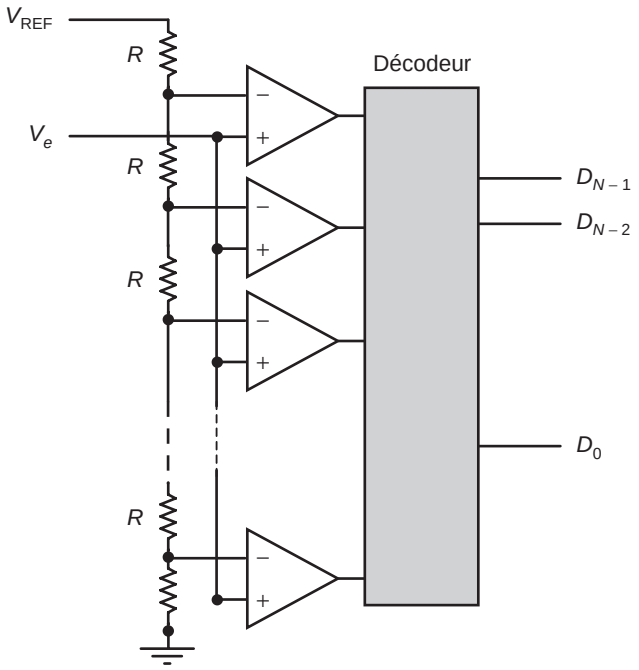


Figure 7.60 – Convertisseur parallèle ou « flash ».

comparateurs ne donnent pas directement la série de bits du code binaire mais le codage « thermométrique » du signal d'entrée. Par exemple, la valeur $V_{REF}/2$ fournit dans un convertisseur à 3 bits la séquence suivante : 0, 0, 0, 0, V_0, V_0, V_0, V_0. Cette séquence se symbolise par la suite « 00001111 ». On commence par le seuil le plus haut. La valeur binaire étant « 100 », il faut un circuit de décodage. Les convertisseurs parallèles ont pour avantage principal leur rapidité et des temps de codage de 1 ns sont maintenant possibles sur 10 bits. Leur consommation est élevée. Ils sont donc réservés aux applications rapides : codage vidéo, traitement des signaux radar…

Ils peuvent être utilisés en deux passes. Un premier convertisseur convertit par exemple le signal sur 8 bits. Cette valeur numérique grossière est convertie par un convertisseur numérique-analogique en une valeur continue. Un amplificateur de précision mesure la différence entre cette valeur et le signal d'entrée. La différence est amplifiée d'un facteur 256 puis elle est à nouveau codée sur 8 bits. Le système a, en théorie, codé le signal sur 16 bits. Ces opérations sont cependant très délicates à mettre en œuvre.

✦ Convertisseur pipe-line

Le principe précédent peut être généralisé à un nombre N de passes en utilisant à chaque étape un convertisseur un bit. C'est le principe du convertisseur pipe-line. Son fonctionnement peut se décrire par la séquence suivante.

Dans une première phase, l'entrée est comparée à $V_{REF}/2$. Si elle est supérieure, le bit de sortie est positionné à un, la valeur $V_{REF}/2$ est retranchée au signal d'entrée et la différence est transférée à l'étage suivant. Si la valeur de l'entrée est inférieure à $V_{REF}/2$, le bit de sortie est 0 et la valeur est transmise à l'étage suivant.

Dans une seconde phase, la valeur transmise est multipliée par deux et appliquée à un échantillonneur-bloqueur. Ce circuit mesure la valeur du signal à un temps donné, le temps d'échantillonnage puis maintient la valeur pendant la période du processus.

Dans une dernière phase, la sortie de l'échantillonneur bloqueur est appliquée à l'étage suivant et le même processus de comparaison se déroule mais en opérant sur la valeur amplifiée. La *figure 7.61* donne une représentation spatiale de ce processus.

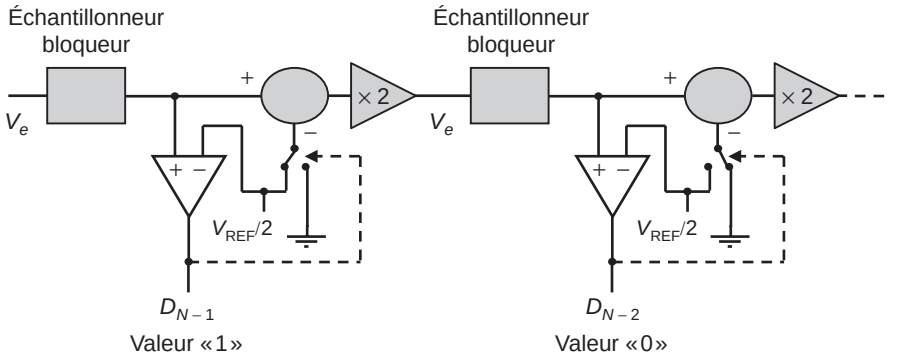


Figure 7.61 – Convertisseur pipe-line.

Ce schéma est assez séduisant car il utilise un simple comparateur un bit donc linéaire par principe. Un autre avantage est lié à la possibilité d'entamer une deuxième conversion avant même que la première ne soit terminée.

✦ Convertisseur à rampe

C'est également un schéma classique. Le signal d'entrée est comparé à un signal variant linéairement avec le temps. Une horloge rapide est déclenchée au début de la rampe. Quand les deux signaux sont égaux on arrête l'horloge et on compte le nombre de coups proportionnel au niveau du signal mesuré. Le principe est illustré *figure 7.62*.

On constate sur le schéma que deux valeurs de la tension d'entrée conduisent à deux nombres différents N_1 et N_2 d'impulsions d'horloge. Ce convertisseur est assez simple de réalisation. Pour obtenir une bonne précision, il faut soit une fréquence d'horloge élevée soit une rampe de pente faible. Dans ce cas, le temps de conversion est élevé.

✦ Convertisseur à approximations successives

Le principe de ce convertisseur est de rechercher par essais successifs la valeur la plus proche de la tension d'entrée. Le principe est illustré *figure 7.63* dans le cas d'une conversion sur 3 bits.

La tension est comparée à $V_{REF}/2$. Si elle est supérieure, une valeur $V_{REF}/4$ est ajoutée à $V_{REF}/2$ et la comparaison se fait à nouveau. Ensuite, on ajoute ou non $V_{REF}/4$ en fonction de la comparaison. Le processus se poursuit ainsi et la valeur binaire est de plus en plus proche de la valeur de la tension à mesurer.

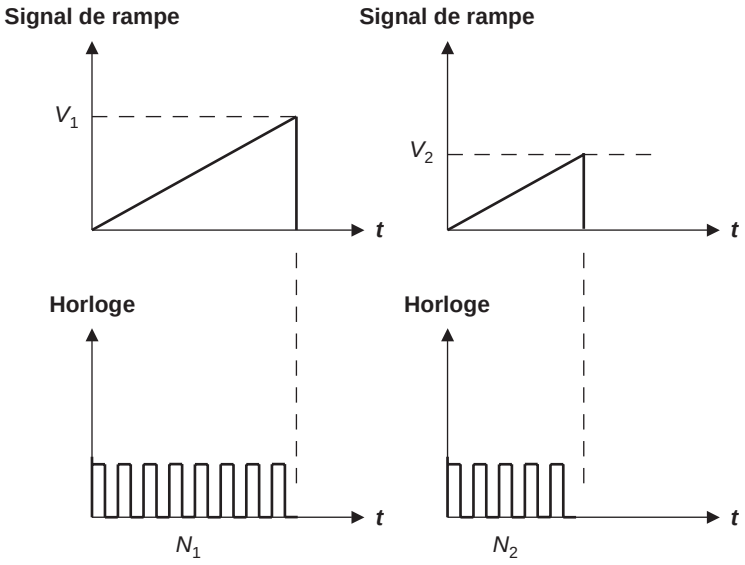


Figure 7.62 – Principe du convertisseur à rampe.

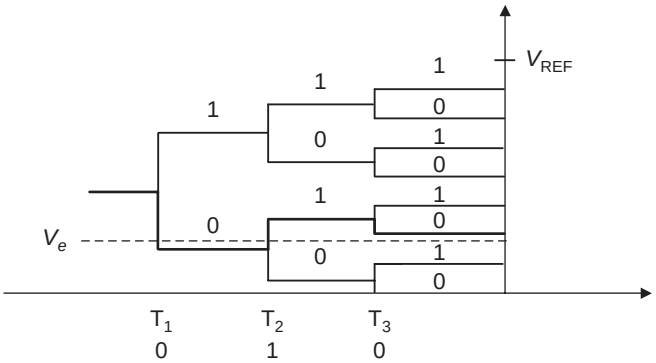


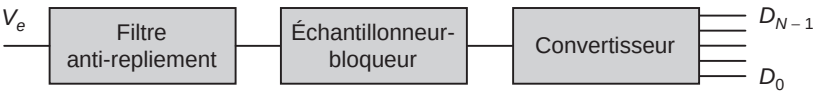
Figure 7.63 – Convertisseur à approximations successives.

✦ Conversion Sigma-Delta

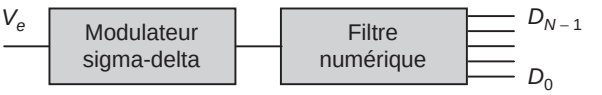
Apparue plus récemment, la conversion Sigma-Delta a connu un grand succès car c'est la seule technique qui permet de coder relativement facilement sur un grand nombre de bits, 14 bits et plus. Cette technique est particulièrement adaptée aux technologies micro-électroniques car elle tolère les dispersions. Elle ne demande pas les opérations d'ajustement des valeurs des composants, comme dans les architectures précédentes, pour obtenir de bonnes propriétés de linéarité. Elle utilise en effet un convertisseur un bit qui est, par principe, linéaire.

Cette technique est également appelée conversion suréchantillonnée. En effet, la fréquence d'échantillonnage du signal est bien supérieure à la fréquence de Nyquist, le double de la fréquence la plus

élevée contenue dans le spectre. Des facteurs de 100 ou plus sont utilisés. La *figure 7.64* montre la différence fondamentale entre une chaîne de conversion classique et une chaîne de conversion de type Sigma-Delta.



a) Chaîne classique



b) Chaîne sigma-delta

Figure 7.64 – Conversion Sigma-Delta.

La conversion classique demande un filtre anti-repliement en entrée à flancs raides car la fréquence d'échantillonnage n'est pas très élevée comparée aux fréquences du signal. De plus, le signal ne doit pas varier pendant la conversion, ce qui explique la présence de l'échantillonneur-bloqueur qui maintient pendant la durée de la conversion le signal en entrée du convertisseur à la valeur qu'il avait à l'instant de début de conversion. Ces deux fonctions sont assez délicates à réaliser en micro-électronique. La conversion Sigma-Delta par principe échantillonne le signal à fréquence très élevée par rapport aux fréquences du signal et du bruit. Un filtre simple du premier ordre suffit pour éviter les phénomènes de repliement de spectre.

Le principe de la conversion Sigma-Delta n'est pas évident. Il est basé sur un modulateur qui fournit en sortie à fréquence élevée une série d'impulsions carrées positives ou négatives dont la valeur moyenne est une bonne approximation de la valeur d'entrée. Il suffit donc de compter la différence entre les impulsions positives et négatives pour obtenir la valeur codée du signal d'entrée. Le schéma du modulateur est indiqué *figure 7.65*.

Le modulateur utilise des sommateurs, un convertisseur analogique-numérique un bit qui fournit « 1 » quand l'entrée est positive et « 0 » quand elle est négative, un convertisseur numérique-analogique qui fournit $+V_{REF}$ quand l'entrée est « 1 » et $-V_{REF}$ quand l'entrée est « 0 ». Pour simplifier l'écriture, on notera +1 et –1 les deux états logiques de la variable logique v_s au lieu de 1 et 0.

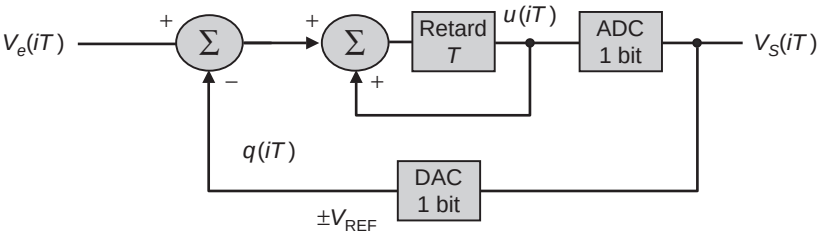


Figure 7.65 – Principe de la conversion Sigma-Delta.

Ce fonctionnement se résume par les équations aux différences :

$$u(iT) = v_e(iT - T) - q(iT - T) + u(iT - T)$$

On notera E l'erreur de quantification :

$$E(iT) = V_{REF}\, v_s(iT) - u(iT)$$

Rappelons que cette équation s'écrit à la condition d'affecter à v_s les deux valeurs + 1 et – 1.

Pour simplifier la suite, on supposera que V_{REF} est égal à 1 V. Étudions sur un exemple comment le modulateur converge vers la valeur précise convertie. On représente sur un tableau à chaque période, la valeur de la tension d'entrée (v_e), la sortie du modulateur (u), l'erreur de quantification E et la sortie numérique (v_s) du système. Les relations de passage d'une période à la suivante deviennent :

$$u(i) = v_e - v_s(i-1) + u(i-1)$$

$$E(i) = v_s(i) - u(i)$$

On calculera également la valeur moyenne $\bar{v}_s$ des impulsions de sortie depuis le début de la conversion. Pour cela, on effectue la différence entre les impulsions positives et négatives. On suppose que la tension d'entrée est 0,5 V et que la sortie du modulateur est initialement à 0,2 V.

Tableau 7.2

Temps	v_e	u	v_s	E	$\bar{v}_s$
1	0,5	0,2	1	0,8	1
2	0,5	–0,3	–1	–0,7	0
3	0,5	1,2	1	–0,2	1/3
4	0,5	0,7	1	0,3	2/4
5	0,5	0,2	1	0,8	3/5
6	0,5	–0,3	–1	–0,7	2/6
7	0,5	1,2	1	–0,2	3/7
8	0,5	0,7	1	0,3	4/8
9	0,5	0,2	1	0,8	5/9
10	0,5	–0,3	–1	–0,3	4/10
11	0,5	1,2	1	–0,2	5/11
12	0,5	0,7	1	0,3	6/12
13	0,5	0,2	1	0,8	7/13
14	0,5	–0,3	–1	–0,7	6/14
15	0,5	1,2	1	–0,2	7/15
16	0,5	0,7	1	0,3	8/16

On constate sur cet exemple que la valeur moyenne du signal formé par les impulsions de sortie du convertisseur un bit converge vers le signal à mesurer. Il suffit de compter la différence entre les impulsions positives et négatives pour obtenir la valeur codée du signal.

Pour comprendre l'intérêt de cette technique, on peut écrire le signal de sortie en fonction de l'erreur de quantification.

$$E(iT) = V_{REF}\, v_s(iT) - u(iT)$$
$$u(iT) = v_e(iT - T) - q(iT - T) + u(iT - T)$$

Comme,

$$q(iT) = V_{REF} v_s(iT)$$
$$V_{REF}\, v_s(iT) = v_e(iT - T) - v_{REF} v_s(iT - T) + u(iT - T) + E(iT)$$
$$V_{REF}\, v_s(iT) = v_e(iT - T) - E(iT - T) + E(iT)$$

Le signal quantifié de sortie est donc le signal d'entrée plus la différence des erreurs de quantification entre deux périodes consécutives.

Cette formule montre, de manière évidente, l'intérêt de cette technique. En effet, si on suppose que le signal varie peu pendant le temps total de codage, les erreurs de quantification s'annulent en grande partie. Il suffit de sommer cette relation sur un grand nombre de périodes. Les valeurs $v_e(iT)$ s'ajoutent alors que les grandeurs $E(i)$ s'annulent deux à deux. Pour s'en convaincre écrivons les relations pour 10 périodes.

$$V_{REF} v_s(10T) = v_e(9T) - E(9T) + E(10T)$$
$$V_{REF} v_s(9T) = v_e(8T) - E(8T) + E(9T)$$
$$V_{REF} v_s(8T) = v_e(7T) - E(7T) + E(8T)$$
$$\cdots$$
$$V_{REF} v_s(2T) = v_e(T) - E(T) + E(2T)$$
$$V_{REF} v_s(T) = v_e(0) - E(0) + E(T)$$

En sommant ces relations, on obtient :

$$\sum_{i=1}^{i=10} V_{REF} v_s(iT) = 10 \cdot v_e + E(10T) - E(0)$$

$$\frac{1}{10} \sum_{i=1}^{i=10} V_{REF} v_s(iT) = v_e + \frac{1}{10}\left[E(10T) - E(0)\right]$$

Le signal somme de sortie est donc une bonne approximation de l'entrée quand le nombre de périodes augmente. Le signal d'entrée est supposé constant pendant ce calcul. On comprend donc que si le signal d'entrée varie dans une bande de fréquence donnée, il est nécessaire d'augmenter la fréquence de suréchantillonnage pour avoir une bonne précision de codage.

Pour terminer ce paragraphe, donnons un exemple de schéma de modulateur Sigma-Delta en reprenant le schéma de l'intégrateur à capacités commutées. On peut remarquer la simplicité du schéma.

Des modulateurs plus complexes peuvent être réalisés pour diminuer encore le bruit de quantification. Ils sont du deuxième ou du troisième ordre.

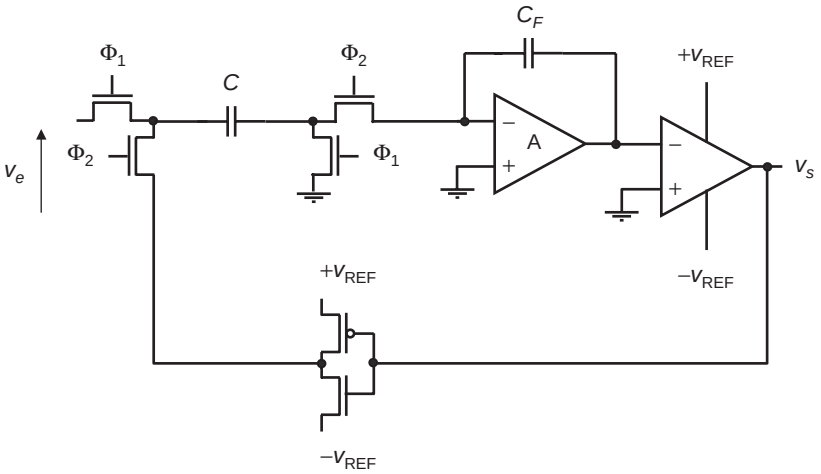


Figure 7.66 – Réalisation d'un modulateur Sigma-Delta.

Pour terminer ce paragraphe dédié à la conversion, il est possible de classer les différentes architectures sur un graphique faisant apparaître deux caractéristiques complémentaires : le nombre de bits et le temps de conversion.

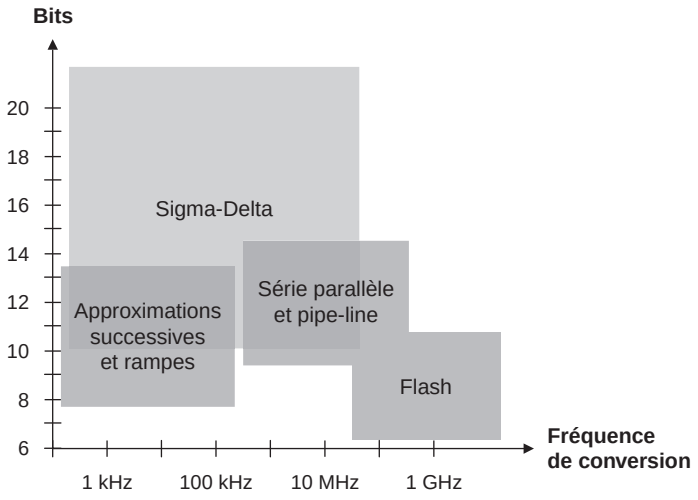


Figure 7.67 – Les technologies de conversion.

Chapitre 8

Les fonctions numériques de base

Le but de ce chapitre est d'introduire les fonctions logiques de base de l'électronique et de montrer comment elles sont réalisées dans les circuits intégrés. Le domaine est très vaste et ce chapitre ne traite que les principes de base illustrés par quelques exemples. Des ouvrages entiers sont consacrés à ce sujet de première importance en électronique étant donné l'importance du traitement numérique de l'information. La méthode d'analyse temporelle des circuits est extraite de la référence **[6]**.

8.1 Logique combinatoire et logique séquentielle

Tous les circuits électroniques de nos jours fonctionnent à partir de la logique binaire à deux états. Elle permet de réaliser des fonctions logiques quelconques qui peuvent être considérées comme des automates définissant un ou plusieurs états de sortie en fonction des entrées et éventuellement des états précédents. Elle permet également d'implémenter des opérations de calcul dans le système binaire bien adapté à cette logique à deux états. Elle permet aussi de coder des lettres ou des images et d'effectuer des traitements. Quelques travaux de laboratoire ont permis d'investiguer des logiques à plusieurs états, comme il sera vu dans le chapitre 11, mais les applications ne sont pas encore présentes dans les systèmes électroniques.

Une variable logique A ne peut prendre que deux valeurs possibles « vrai » ou « faux ». On notera aussi ces états par « 0 » et « 1 » par commodité d'écriture. Cela ne veut pas dire que les niveaux électriques correspondant sont 0 V et 1 V. On parle de logique positive quand l'état « vrai » ou état « 1 » correspond à la valeur de tension la plus élevée et de logique négative dans l'autre cas. Les deux valeurs de tension associées aux deux états caractérisent la logique.

Les progrès technologiques ont conduit à diminuer ces valeurs au fur et à mesure que les technologies ont évolué. Elles sont passées de 5 V dans les années 70 à 1 V en 2004. L'intérêt de réduire ces tensions se comprend facilement si on pense à la consommation électrique des circuits. Notons que la valeur la plus basse des deux valeurs possibles est en général 0 V et que les deux états sont définis par rapport à des valeurs de tensions et non pas de courants. Les logiques en courant sont possibles mais n'ont pas conduit à des applications industrielles.

Enfin et c'est un point fondamental de la micro-électronique, il faut constater que la logique de type CMOS s'est imposée à l'ensemble des circuits intégrés en éliminant progressivement les autres familles. La logique CMOS se caractérise par deux propriétés :

- elle utilise des transistors MOSFET de type n et p ;
- les circuits sont conçus de telle manière qu'en régime statique le courant consommé soit nul.

En réalité, et particulièrement pour les technologies avancées, cette deuxième propriété n'est pas tout à fait exacte et constitue un point d'évolution fondamental.

Revenons maintenant à la notion de variable logique pour expliquer la différence entre logique combinatoire et logique séquentielle. Une variable logique F dite de sortie dépend d'un nombre fini de variables logiques dites d'entrée $A, B, C\ldots$ On écrira donc :

$$F = F(A, B, C\ldots)$$

Rappelons que les variables n'ont que deux états possibles notés « 0 » et « 1 ». La fonction est définie si, pour toutes les combinaisons des variables d'entrée, on est capable de définir l'état de la variable F. Le tableau donnant ce résultat est appelé table de vérité. Il définit complètement la fonction logique. On définit ainsi la logique combinatoire.

Quand la donnée des variables d'entrée n'est pas suffisante pour définir l'état de sortie on parle de logique séquentielle. De manière générale, une fonction en logique séquentielle dépend des états des variables d'entrées au moment où on veut déterminer sa valeur mais aussi de la séquence des états précédents pour toutes les variables. Un circuit qui détermine si le nombre de changements d'états en entrée est pair ou impair est un exemple de logique séquentielle. On pourrait également dire que la logique séquentielle dépend de l'histoire des événements. Dans le chapitre 9 consacré aux architectures nous verrons que cette notion peut se généraliser en définissant une machine d'états, ce qui constitue une notion fondamentale des architectures électroniques.

Toutes les fonctions de la logique combinatoires peuvent s'exprimer en combinant des fonctions de base : le « et » logique, le « ou » logique et l'inverseur. Pour s'en convaincre, il suffit de prendre une table de vérité quelconque. Définissons le « et » et le « ou » par leurs tables de vérité : *tableaux 8.1 et 8.2*.

Tableau 8.1

A	B	A « et » B
0	0	0
0	1	0
1	0	0
1	1	1

Tableau 8.2

A	B	A « ou » B
0	0	0
0	1	1
1	0	1
1	1	1

On notera A et B par AB. Il y a une certaine analogie avec les règles du calcul algébrique.

On notera A ou B par $A + B$. L'analogie avec le calcul algébrique n'est pas totale puisqu'en logique $1 + 1$ donne 1.

On définit également l'inverseur, noté $\bar{A}$ et défini par la *table 8.3*.

Tableau 8.3

A	« non » A
0	1
1	0

Une fonction quelconque peut s'exprimer à l'aide de ces fonctions de base. Pour s'en convaincre, prenons une fonction logique F définie par sa table de vérité : *tableau 8.4*.

La fonction F a un état 1 pour les cas indiqués en gris sur le tableau. Elle s'exprime alors par des fonctions « et », « ou » et des inversions comme suit. Une valeur vraie ou « 1 » est obtenue par une simultanéité de conditions et cela dans les trois cas indiqués par le tableau.

$$F = \bar{A}BC + A\bar{B}C + AB\bar{C}$$

Dans tous les autres cas, la fonction est nulle au sens logique du terme.

Tableau 8.4

A	B	C	F
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0

Cet exemple simple montre comment on construit une fonction à partir des briques logiques de base. Le schéma électrique peut s'en déduire facilement comme il est indiqué *figure 8.1*.

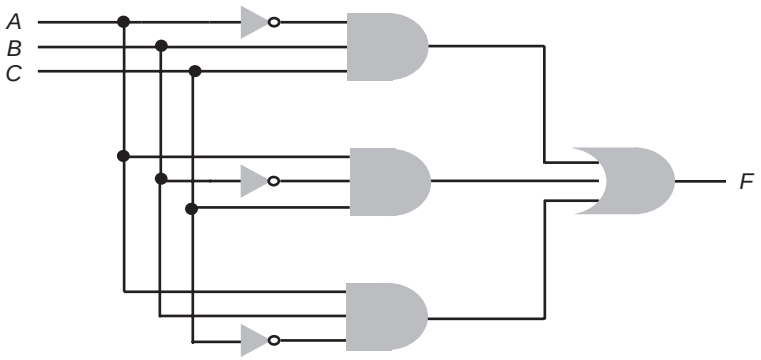


Figure 8.1 – Exemple de logique combinatoire.

En fait, les trois fonctions de base ne sont pas indispensables et il est possible de réaliser une fonction quelconque avec un inverseur et une des deux fonctions « et » ou « ou ». On utilise pour cela les relations logiques :

$$\overline{A + B} = \bar{A}\,\bar{B}$$

$$\overline{AB} = \bar{A} + \bar{B}$$

Ces relations se démontrent facilement en construisant les tables de vérité. Elles se généralisent pour trois et un nombre quelconque de variables logiques.

En pratique, la micro-électronique fait usage de deux autres fonctions : le « nand » qui est un « et » suivi d'un inverseur et le « nor » qui est un « ou » suivi d'un inverseur. Ces deux fonctions sont représentées sur la *figure 8.2* avec le petit cercle en sortie pour signifier l'inversion. Elles sont plus faciles à réaliser en technologie CMOS et sont les fonctions de base de la logique.

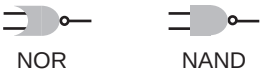


Figure 8.2 – Les fonctions « nor » et « nand ».

Toutes les fonctions logiques peuvent se réaliser avec des NOR car l'inverseur est un NOR particulier. Il suffit de positionner une entrée à l'état « 1 ». Il en est de même avec le NAND en positionnant une entrée à « 1 ».

La logique séquentielle ajoute une notion de base supplémentaire : la fonction mémoire. Pour en comprendre la signification, il suffit de considérer le schéma très simple de la *figure 8.3* formé par deux portes NOR bouclées.

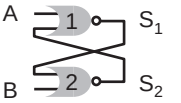


Figure 8.3 – Un exemple de logique séquentielle.

Supposons l'entrée *A* dans l'état 0 et l'entrée B dans l'état 0. Deux états sont alors possibles : si la deuxième entrée de la porte 1 est à l'état 1, la sortie S_1 est dans l'état 0 et la sortie S_2 est dans l'état 1. Si la deuxième entrée de la porte 1 est dans l'état 0, la sortie S_1 est dans l'état 1 et la sortie S_2 dans l'état 0. C'est donc la situation inverse. Les deux cas sont possibles et l'état dans lequel se trouve le système dépend non seulement des entrées mais des changements d'états qui se sont passés antérieurement. Bien évidemment, un nouveau changement d'état sur une des entrées a pour conséquence un changement des états de sortie. Dans tous les cas, les deux sorties sont dans des états différents : quand S_1 est à l'état 1, S_2 est à l'état 0 et réciproquement.

Un système équivalent peut être construit avec deux portes NAND. Le lecteur est invité à étudier la faisabilité d'un tel système avec deux portes ET ou deux portes OU. Ce système simple illustre deux fonctions fondamentales de l'électronique numérique :

- la fonction bistable ;
- la fonction mémoire.

La fonction bistable se caractérise par deux états possibles. La fonction mémoire se caractérise par la prise en compte des événements précédents.

8.2 Le modèle de transistor utilisé

Pour étudier les circuits logiques, il serait possible de calculer les circuits correspondant aux différentes portes définies précédemment ainsi que les assemblages de portes, comme il a été fait pour les fonctions analogiques. Cette méthode a rapidement des limites si on pense aux circuits complexes de la logique formés de milliers de portes. Il est donc nécessaire de définir les propriétés les plus importantes du transistor MOS utilisé en commutateur. Les simulateurs logiques feront de même mais avec le niveau d'automatisation nécessaire dans la conception de circuits complexes.

Le principe général de la logique MOS est de commuter la tension d'alimentation V_{DD} ou la tension nulle de référence en jouant sur la valeur de la tension de grille d'un MOSFET. En logique, on adopte

une représentation simple pour distinguer les MOS canal *n* et canal *p*. La *figure 8.4* rappelle les deux schémas.

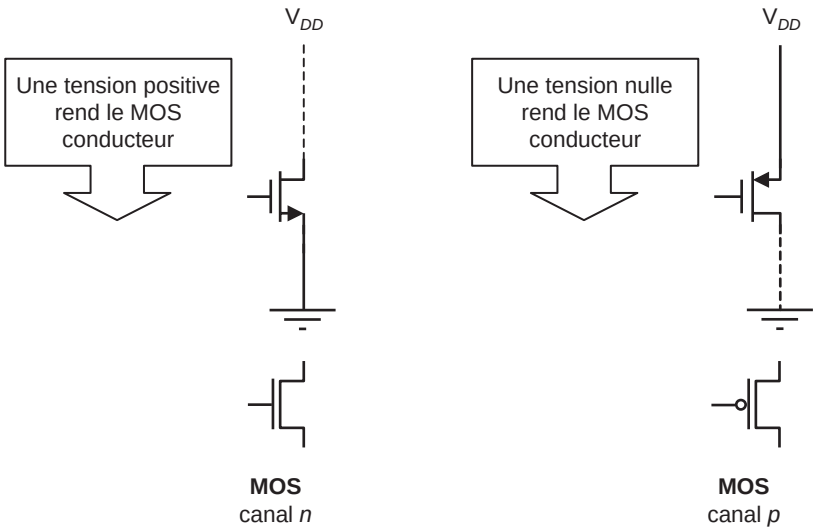


Figure 8.4 – Les schémas des MOS en logique.

Rappelons que le MOS canal *p* fonctionne généralement en polarisant la source et le puits à la tension positive d'alimentation V_{DD}. Les potentiels de grille et de drain sont donc négatifs par rapport au potentiel de la source.

La *figure 8.5* illustre le principe de fonctionnement d'une cellule logique de base. On ignore pour l'instant la charge du MOS et on considère seulement une capacité de charge due par exemple à la piste de connexion et à une autre cellule logique reliée à celle que nous étudions.

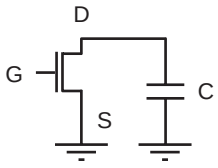


Figure 8.5 – Cellule logique de base.

L'examen des caractéristiques du transistor permet d'observer ce qui se passe quand la tension de grille passe de 0 V à V_{DD}.

La *figure 8.6* illustre la commutation d'un MOS dans une technologie avancée de 50 nm. Initialement, le transistor est non passant car une tension nulle est appliquée sur sa grille. Il est alors relié à la tension d'alimentation, 1 V dans le cas présent. Ensuite, sa grille est portée au potentiel 1 V et le MOS devient conducteur. Le point de fonctionnement passe de B à C. Pendant la transition, il y

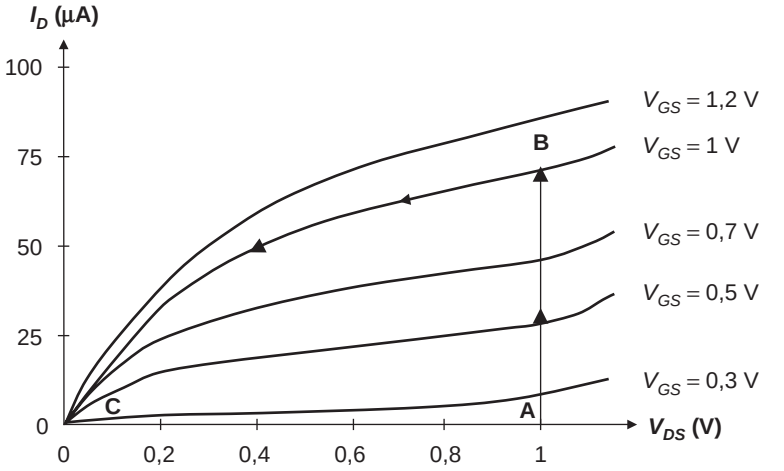


Figure 8.6 – Commutation d'un transistor.

a un courant d'environ 70 µA puis le régime s'établit et le courant revient à zéro en suivant la courbe V_{GS} = 1 V. Si on mesure la pente de la droite BC, on obtient une valeur moyenne de la résistance du MOS pendant la commutation. Dans le cas présent, elle est de 14 kΩ. Il faut noter que les courbes sont relatives à un NMOS de petite taille. Le ratio *W/L* est de 2. Cette méthode est grossière mais elle donne pourtant les ordres de grandeur des caractéristiques électriques. Rappelons les formules donnant le courant en régime de saturation à partir des résultats du chapitre 4. Dans le cas d'un canal court, c'est-à-dire pour les technologies numériques actuelles :

$$I_D = W\mu C'_{OX}[V_{GS} - V_T - V_{DSsat}]E_C$$

Pour les technologies plus anciennes, le modèle canal long s'applique et :

$$I_D = \frac{W}{L}\mu C'_{OX}\frac{(V_{GS} - V_T)^2}{2(1+\delta)}$$

Rappelons que V_T est la tension de seuil et que les paramètres physiques sont approximativement :

- C'_{OX} = 1,75 fF/µm² pour un NMOS dans une technologie 0,8 µm
- C'_{OX} = 25 fF/µm² pour un NMOS dans une technologie 45 nm
- µ est de 600 cm²/V · s pour les électrons et 200 cm²/V · s pour les trous
- E_C est de l'ordre de 1 V/µm

Si on exprime ces valeurs dans le calcul précédent on obtient la valeur de la résistance moyenne du MOS pendant la commutation en fonction de sa largeur, pour des valeurs typiques de tension d'alimentation de 3 V et 1 V.

Tableau 8.5

	Technologie 0,8 micron	Technologie 45 nm
NMOS	15 kΩ *L/W*	30 kΩ/*W*
PMOS	45 kΩ *L/W*	70 kΩ/*W*

Les dimensions *L* et *W* sont exprimées en unités relatives pour simplifier l'utilisation des formules, c'est-à-dire en prenant comme unité la grandeur caractéristique de la technologie exprimée en micron. Dans une technologie 45 nm, les valeurs *W* et *L* d'un transistor de largeur 450 nm et dont le canal mesure 45 nm de long seront simplement $W = 10$ et $L = 1$.

Le schéma du MOS en commutation prend donc la forme très simple indiquée *figure 8.7*. Ce modèle grossier donne cependant des résultats assez proches de ceux que l'on peut obtenir à l'aide d'un simulateur de type SPICE.

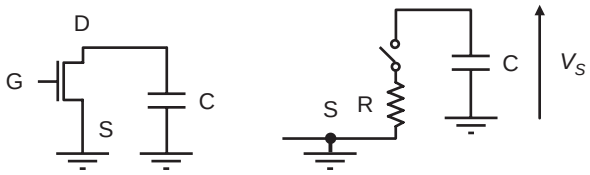


Figure 8.7 – Modèle grossier du MOS en commutation.

Il est naturel de s'intéresser à l'aspect temporel de la commutation car la vitesse de fonctionnement des circuits logiques est une propriété fondamentale. D'autres caractéristiques devront être également étudiées : la consommation électrique et la tolérance au bruit et aux dispersions.

Le comportement temporel d'un étage logique peut être caractérisé par deux paramètres temporels : le retard et le temps de montée. Ces deux valeurs sont indiquées *figure 8.8* en comparant la tension appliquée à l'entrée et la tension en sortie.

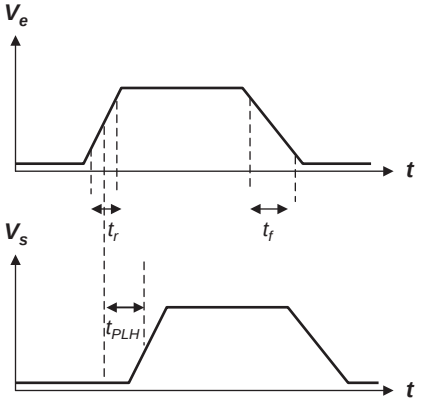


Figure 8.8 – Propriétés temporelles.

Les impulsions logiques ne sont pas parfaites. Elles sont caractérisées par un temps de montée t_r et un temps de descente t_f. Les valeurs sont mesurées en prenant le temps correspondant à 10 % et 90 % du signal. Si l'impulsion d'entrée est définie par son temps de montée et par son temps de descente, l'impulsion de sortie est caractérisée de la même manière par les temps de montée (t_{LH}) et de descente (t_{HL}). Ils peuvent éventuellement être plus courts que ceux d'entrée. Pensons à un étage

logique ayant un gain en tension élevé. Si la tension d'entrée sature l'étage, la pente de la tension de sortie peut être plus forte.

Une autre caractéristique apparaît, notée t_{PLH} : le retard de l'impulsion de sortie. Il est facile de constater que ce retard est lié à la valeur de la constante de temps *RC* de la *figure 8.7*. En effet, supposons une variation rapide de la tension d'entrée. La tension aux bornes de la capacité est donnée par la relation :

$$V_s(s) = \frac{\frac{1}{Cs}}{R + \frac{1}{Cs}} \frac{\Delta V_{DS}}{s}$$

On applique les lois de calcul des circuits et on suppose que la tension d'entrée varie brusquement, ce qui induit une variation de la tension V_{DS} égale à ΔV_{DS}.

On obtient donc :

$$V_s(s) = \left(\frac{1}{s} - \frac{1}{s + \frac{1}{RC}} \right) \Delta V_{DS}$$

Dans le domaine temps, on obtient :

$$V_s(t) = \left(1 - \exp^{-\frac{t}{RC}}\right) \Delta V_{DS}\, \gamma(t)$$

Les signaux sont représentés *figure 8.9*.

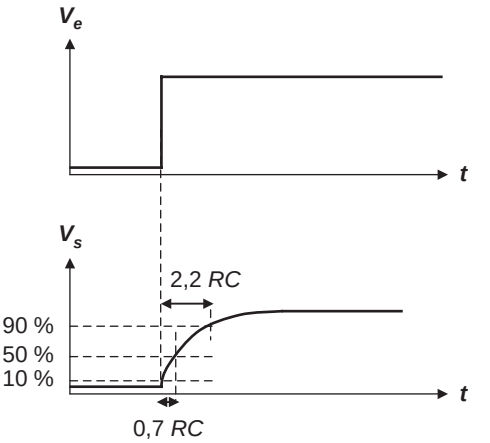


Figure 8.9 – Calcul du retard et du temps de montée.

Le retard apparaît alors comme le temps correspondant à une variation de 50 % de l'amplitude soit 0,7 *RC* et le temps de montée comme le temps correspondant à 90 % de l'amplitude soit 2,2 *RC*. On suppose que la porte située en aval commute quand son entrée est à un potentiel supérieur à $V_{DD}/2$.

Rappelons que seule la capacité de charge du MOS a été prise en compte dans ce modèle simpliste. Il faut pour affiner le modèle tenir compte des capacités du MOSFET lui-même.

Les résultats établis dans le chapitre 3 montrent qu'en régime triode les capacités entre grille et drain et entre grille et source s'expriment de la manière suivante.

$$C_t = \frac{1}{2} C'_{OX} WL$$

La valeur C'_{OX} est la capacité par unité de surface.

Le schéma de la *figure 8.7* se modifie alors comme il est indiqué *figure 8.10*. Ce résultat n'est pas tout à fait immédiat et demande quelques précisions. Les capacités sont valables en régime triode uniquement. La capacité drain-source est donc surestimée notablement.

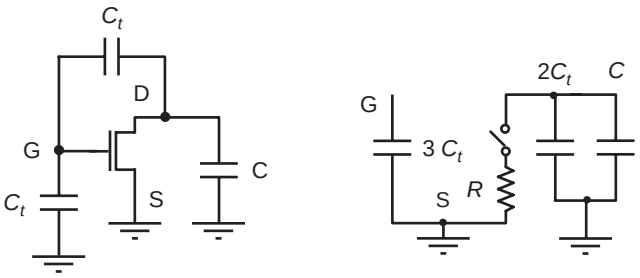


Figure 8.10 – Le MOS en commutation.

Pour passer du schéma de gauche à celui de droite, calculons par exemple le courant d'entrée dans la commutation. La grille passe de 0 à V_{DD} pendant Δt et le drain de V_{DD} à 0. Le courant qui circule dans la capacité drain-grille est donc $2\,C_t V_{DD}/\Delta t$ alors que le courant qui circule dans la capacité grille-source est $C_t V_{DD}/\Delta t$. Au total, un courant $3\,C_t V_{DD}/\Delta t$ circule dans l'entrée pendant la commutation. De même, on peut calculer le courant qui circule en sortie. C'est $2\,C_t V_{DD}/\Delta t$. Le schéma équivalent de droite est donc justifié.

L'interrupteur est fermé quand le MOS passe de l'état bloqué à l'état conducteur et cela pour une tension qui est appelée tension de seuil de l'étage. Elle ne doit pas être confondue avec la tension de seuil du transistor qui est un paramètre physique. Dans le paragraphe suivant nous verrons comment calculer la tension de seuil de l'étage.

On constate alors que les valeurs de retard et de temps de montée deviennent :

- Retard : $0{,}7\,R\,(C + 2\,C_t)$
- Temps de montée : $2{,}2\,R\,(C + 2\,C_t)$

La commutation d'un PMOS se calcule de la même manière en s'appuyant sur la *figure 8.11*. Dans ce cas, comme il sera vu par la suite, le transistor est relié par sa source à la tension d'alimentation V_{DD} si bien que toutes les tensions sont négatives par rapport à la source.

Les constantes de temps de commutation s'expriment de la même manière que pour la commutation du NMOS en fonction de la valeur $R(2C_t + C)$.

On peut alors exprimer la capacité typique du transistor C_t en fonction des dimensions relatives du transistor pour les deux technologies envisagées.

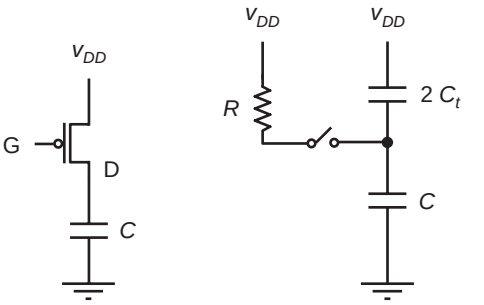


Figure 8.11 – Commutation du PMOS.

Tableau 8.6

Technologie	R (kΩ)	C_t
0,8 µm (canal long) NMOS	15 L/W	0,9 $W \cdot L$ (fF)
0,8 µm (canal long) PMOS	45 L/W	0,9 $W \cdot L$ (fF)
45 nm (canal court) NMOS	35/W	30 $W \cdot L$ (aF)
45 nm (canal court) PMOS	68/W	30 $W \cdot L$ (aF)

Rappelons que les dimensions W et L sont en unité relative. Un transistor de 450 nm de large dans la technologie 45 nm a par exemple une largeur relative de 10. Ce tableau ne fait que donner des ordres de grandeur et il est facile d'interpoler les valeurs pour d'autres technologies en se basant sur les résultats du chapitre 3.

La constante de temps qui caractérise le retard et le temps de montée doit intégrer la capacité du circuit de charge du transistor qui commute. Cette capacité, notée C, est en général très supérieure à la capacité C_t du transistor lui-même. Elle est due aux transistors interconnectés et surtout aux liaisons entre transistors. Une valeur typique de 50 *fF* permet de donner des ordres de grandeur.

Le transistor permet de commuter la tension V_{DD} ou la tension 0 définissant ainsi un état logique. Il peut également être utilisé pour laisser ou ne pas laisser passer une tension. Cette fonction a été expliquée dans le chapitre 7 dans le cas d'une tension continue mais les conclusions restent valables pour une tension de type numérique ne prenant que deux états possibles.

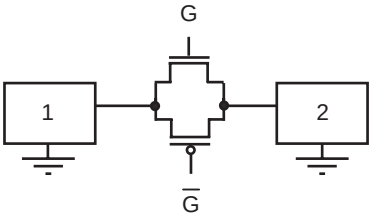


Figure 8.12 – Le MOS comme interrupteur.

Comme il a été expliqué dans le chapitre 7, il est avantageux de placer deux transistors complémentaires en parallèle pour réaliser cette fonction. Le retard apporté par cette fonction dans la transmission du signal est calculé comme précédemment. On obtient facilement :

$$t_D = 0{,}7\,RC$$

La résistance R est $R_nR_p/R_n + R_p$, chacune des deux résistances étant la résistance équivalente de conduction du transistor considéré. La capacité C est donnée par la relation :

$$C = C_L + 2\,C_t$$

Dans cette formule, C_L est la capacité de l'entrée de l'étage 2.

8.3 L'étage inverseur

C'est véritablement la brique de base de la logique CMOS et toutes les autres fonctions logiques sont dérivées de cette cellule élémentaire. La fonction logique est le complément : un état « 0 » devient un état « 1 » et réciproquement. Elle est en pratique réalisée par la mise en série d'un PMOS et d'un NMOS comme le montre la *figure 8.13*.

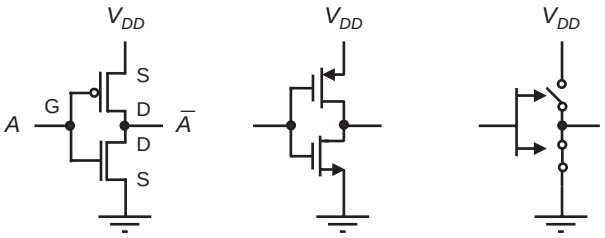


Figure 8.13 – L'inverseur CMOS.

La *figure 8.13* donne les trois représentations de l'inverseur : la représentation logique, électrique et fonctionnelle. Quand une tension positive, de l'ordre de V_{DD} est appliquée sur la grille, le transistor NMOS conduit et le transistor PMOS est bloqué. La connexion de sortie est reliée à la masse et isolée de la tension d'alimentation. Elle est nulle. Il y a donc bien changement d'état logique. Quand une tension nulle est appliquée sur l'entrée, c'est l'inverse. Le transistor NMOS est bloqué et le transistor PMOS conduit.

On remarque que, dans les deux états, un transistor est conducteur et l'autre est bloqué. Comme les deux transistors sont en série, le courant traversant l'ensemble est nul, en fait, très faible dans la réalité. La notion de transistor conducteur peut sembler inadaptée dans ce raisonnement mais il faut imaginer le fonctionnement en dynamique. Les capacités du schéma se chargent et se déchargent à travers les transistors MOS. Pour préciser le fonctionnement de l'inverseur, il est nécessaire de raisonner plus précisément sur les courbes caractéristiques. Cet exercice a déjà été fait dans le chapitre 7 mais il est utile de le reprendre dans le cas d'un système logique à deux états.

On trace les courbes du transistor NMOS (le transistor en bas de la figure) puis sur le même graphique on trace les courbes du transistor PMOS. On suppose que les caractéristiques sont identiques au signe près. De plus, la somme des deux tensions drain-source est égale à la tension d'alimentation.

$$V_{DSn} + V_{SDp} = V_{DD}$$

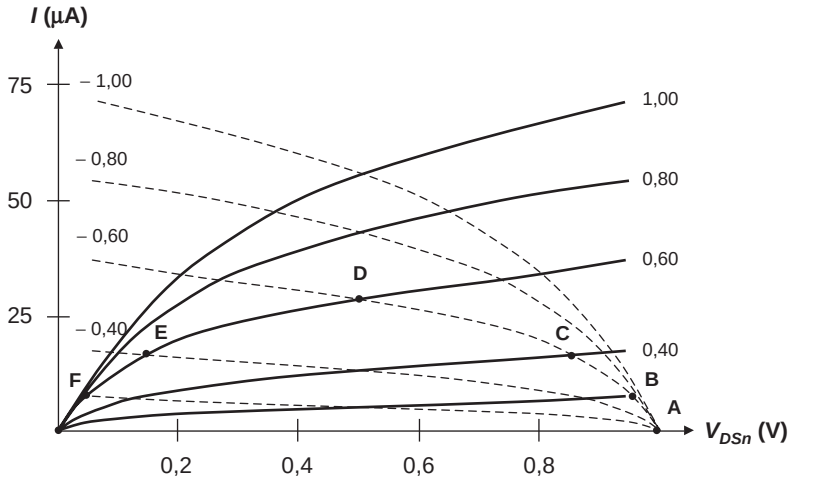


Figure 8.14 – Inverseur CMOS.

En abscisse, on mesure la tension V_{DSn} aux bornes du NMOS à partir de l'origine. La tension aux bornes du PMOS est mesurée à partir de V_{DD} égal à 1 V dans cet exemple. La somme de ces deux tensions est V_{DD}. Quand la tension grille-source est V_G pour le NMOS, elle est égale à $V_{DD} - V_G$ pour le PMOS. On peut donc tracer point par point l'ensemble des états possibles du système.

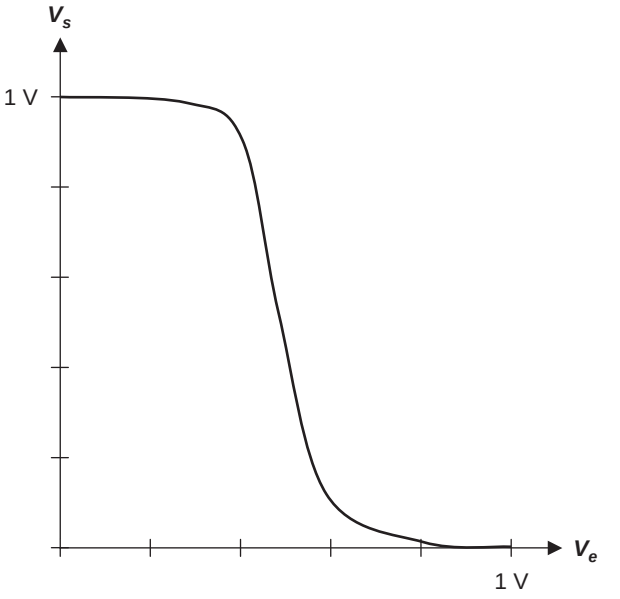


Figure 8.15 – Fonction de transfert.

Quand la tension d'entrée est nulle, le point de fonctionnement est le point A. Quand la tension d'entrée est 0,2 V, la tension de grille du PMOS est 0,8 V et le point de fonctionnement se déplace en B et ainsi de suite par incréments de 0,2 V. On peut également tracer la fonction de transfert, tension de sortie en fonction de la tension d'entrée. On obtient la courbe de la *figure 8.15*.

La *figure 8.14* montre que le point de fonctionnement se déplace pendant la commutation avec apparition d'un courant uniquement pendant le temps du changement d'état. En régime établi, c'est-à-dire en dehors de la phase de commutation, le courant est bien nul. Ce n'est plus vrai dans les technologies très avancées comme il sera vu dans le chapitre consacré à l'évolution du transistor. Dans l'exemple donné, le courant peut atteindre la valeur de 25 μA.

Il est intéressant de chercher la tension pour laquelle la tension d'entrée est égale à la tension de sortie. Cette tension est la tension de commutation de l'inverseur. Elle ne doit pas être confondue avec la tension de seuil du transistor. Les deux transistors sont alors en régime de saturation comme le montre la *figure 8.14*. Dans une technologie à canal long on écrit :

$$I_D = \left(\frac{W}{L}\right)_n \mu_n C'_{OX} \frac{(V - V_{Tn})^2}{2(1+\delta)}$$

soit,

$$I_D = \frac{1}{2}\beta_n (V - V_{Tn})^2$$

De même, on écrit pour le PMOS :

$$I_D = \frac{1}{2}\beta_p (V_{DD} - V - V_{Tp})^2$$

Dans cette formule, toutes les grandeurs sont positives. On obtient donc :

$$V = \frac{\sqrt{\frac{\beta_n}{\beta_p}}\, V_{Tn} + (V_{DD} - V_{Tp})}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}$$

Si on souhaite que la tension de commutation soit égale à la moitié de la tension d'alimentation, ce qui a pour effet de symétriser le système, on obtient pour des valeurs numériques typiques la condition suivante :

$$\frac{\beta_n}{\beta_p} \approx 1$$

Comme la mobilité des trous est environ trois fois plus faible que celle des électrons, on en déduit :

$$\left(\frac{W}{L}\right)_p \approx 3\left(\frac{W}{L}\right)_n$$

À longueur de grille égale, le PMOS sera donc trois fois plus large que le NMOS. On comprend que cette condition est avantageuse pour assurer une garantie maximum de fonctionnement vis-à-vis du bruit. En effet, si la tension de commutation de l'étage était proche de V_{DD} ou de 0 V, une variation

de la tension d'alimentation ou une variation du potentiel local de référence pourrait plus facilement être à l'origine d'un changement d'état non lié à un changement de l'état logique en entrée.

Dans une technologie canal court, la relation correspondant à cette même condition devient en négligeant la tension de saturation :

$$I_D = W_n \mu_n C'_{OX} [V - V_{Tn}] E_{Cn}$$
$$I_D = W_p \mu_p C'_{OX} [V_{DD} - V - V_{Tp}] E_{Cp}$$

On en déduit donc :

$$V = \frac{\frac{W_p \mu_p E_{Cp}}{W_n \mu_n E_{Cn}} (V_{DD} - V_{Tp}) + V_{Tn}}{\frac{W_p \mu_p E_{Cp}}{W_n \mu_n E_{Cn}} + 1}$$

De la même manière, un choix des dimensions tel que $W_p \mu_p E_{Cp} / W_n \mu_n E_{Cn}$ soit égal à 1 conduit à une valeur de V égale à $V_{DD}/2$ ce qui est le but recherché. On obtient donc une condition géométrique en régime de canal court assez proche de celle obtenue en régime de canal long. Il faut tenir compte de la différence de valeur de la mobilité et du champ critique. Finalement, on choisit un rapport de deux entre le PMOS et le NMOS.

Il est maintenant possible de calculer le temps de commutation de l'inverseur en reprenant la méthode et les schémas du paragraphe précédent. La *figure 8.16* représente le modèle électrique simplifié de l'inverseur incluant les capacités dans les deux états possibles.

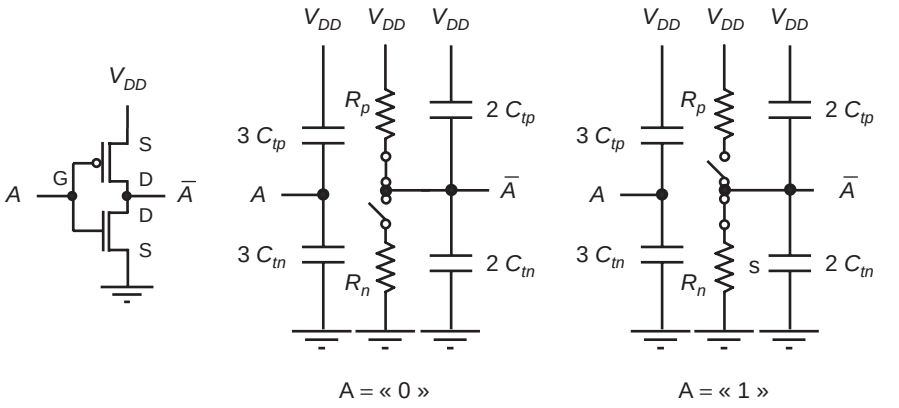


Figure 8.16 – Schéma électrique de l'inverseur.

Quand l'entrée passe de « 1 » à « 0 », on suppose que le NMOS devient instantanément non conducteur et que le PMOS devient instantanément conducteur. Le schéma correspondant à $A =$ « 0 » s'applique. La constante de temps caractéristique de la montée de la tension de sortie est alors $R_p(2\,C_{tp} + 2\,C_{tn})$. Le temps de propagation de l'inverseur est donc :

$$t_{PLH} = 0{,}7\,R_p(2\,C_{tp} + 2\,C_{tn})$$

De manière symétrique, quand l'entrée passe de « 1 » à « 0 », le retard s'écrit :

$$t_{PHL} = 0{,}7\, R_n(2\, C_{tp} + 2\, C_{tn})$$

Si l'inverseur commande un autre étage équivalent à une capacité C, il faut ajouter cette valeur aux capacités des formules précédentes. Quelques exemples numériques illustrent les possibilités des technologies du point de vue de la rapidité.

Un inverseur dans une technologie 45 nm réalisé avec un NMOS ayant un *W/L* de 10/1 et un PMOS ayant un *W/L* de 20/1 se caractérise donc par les temps suivants :

$$t_{PHL} = 0{,}7\, R_p(2\, C_{tp} + 2\, C_{tn}) = 0{,}7 \cdot 3{,}4\,\text{k} \cdot (0{,}6 + 1{,}2)\,\text{fF} = 4{,}5\,\text{ps}$$

Une simulation plus précise conduit cependant à des valeurs notablement supérieures. À cette échelle de temps, d'autres phénomènes doivent être pris en compte. Si la capacité de charge de l'étage est de 50 fF, les temps de propagation deviennent :

$$t_{PLH} = 0{,}7\, R_p C = 0{,}7 \cdot 3{,}4\,\text{k} \cdot 50\,\text{fF} = 120\,\text{ps}$$

$$t_{PHL} = 0{,}7\, R_n C = 0{,}7 \cdot 3{,}4\,\text{k} \cdot 50\,\text{fF} = 120\,\text{ps}$$

Le dernier calcul à effectuer est celui de la consommation. Quand les états sont établis, la consommation statique est nulle car le courant traversant l'inverseur est nul. En fait, il y a un courant résiduel appelé courant sous le seuil et un courant tunnel entre grille et drain. Ces valeurs ne sont plus négligeables dans les technologies avancées au-delà du 90 nm et ce point sera vu en détail dans le chapitre 10. Ce phénomène est d'autant plus important qu'il est permanent : un inverseur qui ne commute pas consomme en permanence de l'énergie statique.

Il est maintenant possible de calculer l'énergie dynamique c'est-à-dire l'énergie dissipée dans un changement d'état. Nous avons vu en effet que lors d'un changement d'état les deux transistors sont simultanément conducteurs pendant la commutation. Ce calcul fondamental peut se faire dans un cadre assez général et est illustré *figure 8.17*.

Le circuit de charge est simplement représenté par une capacité. La piste de connexion entre l'inverseur et le circuit de charge est modélisée par une résistance R et une capacité dont la valeur est intégrée à la capacité d'entrée C du circuit de charge. On suppose que la transition d'entrée fait conduire le PMOS et bloque le NMOS. Ce circuit peut se représenter sous une forme très simple illustrée *figure 8.18*. Le schéma peut se simplifier pour aboutir au schéma de droite.

Il est maintenant possible de calculer l'énergie dissipée lors de la commutation. L'énergie est dissipée dans les résistances. La source de tension est supposée être un échelon commutant de la valeur 0 V à la valeur V_{DD}. Le principe de ce calcul est de déterminer le courant qui circule dans ce dispositif puis de mesurer l'énergie dissipée par effet joule dans la résistance.

$$E = (R + R_p)\int_0^{\infty} i(t)^2\,\mathrm{d}t$$

Dans le formalisme de Laplace, on écrit étant donné la forme de la tension d'entrée :

$$\frac{V_{DD}}{s} = \left(R + R_p + \frac{1}{(C + C_t)s}\right) i(s)$$

On en déduit facilement :

$$i(s) = \frac{V_{DD}}{(R + R_p)} \frac{1}{s + \dfrac{1}{(R + R_p)(C + C_t)}}$$

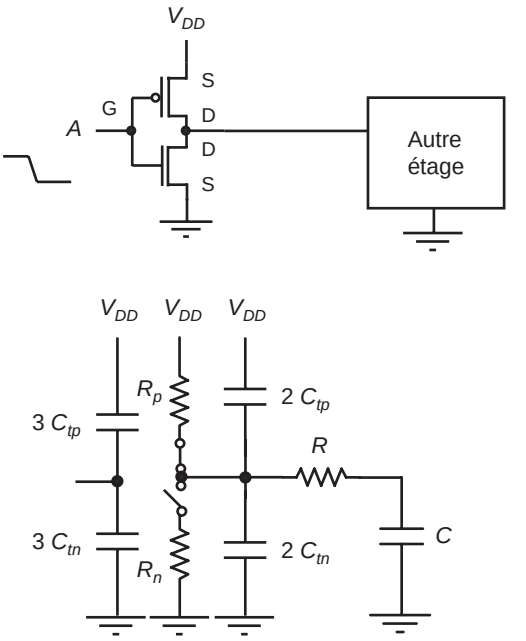


Figure 8.17 – Deux circuits logiques interconnectés.

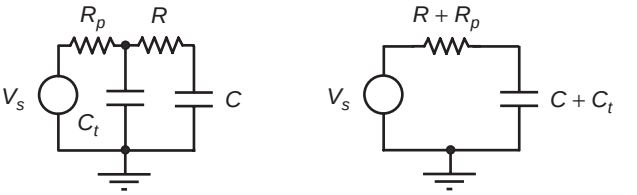


Figure 8.18 – Schéma simplifié de l'inverseur.

soit,

$$i(t) = \frac{V_{DD}}{(R + R_p)} \gamma(t) \cdot \exp^{-\frac{t}{(R + R_p)(C + C_t)}}$$

On en déduit donc :

$$E = (R + R_p) \int_0^{\infty} \frac{V_{DD}^2}{(R + R_p)^2} \cdot \exp^{-\frac{2t}{(R + R_p)(C + C_t)}} \, dt$$

Un calcul simple conduit à :

$$E = \frac{1}{2} (C + C_t) V_{DD}^2$$

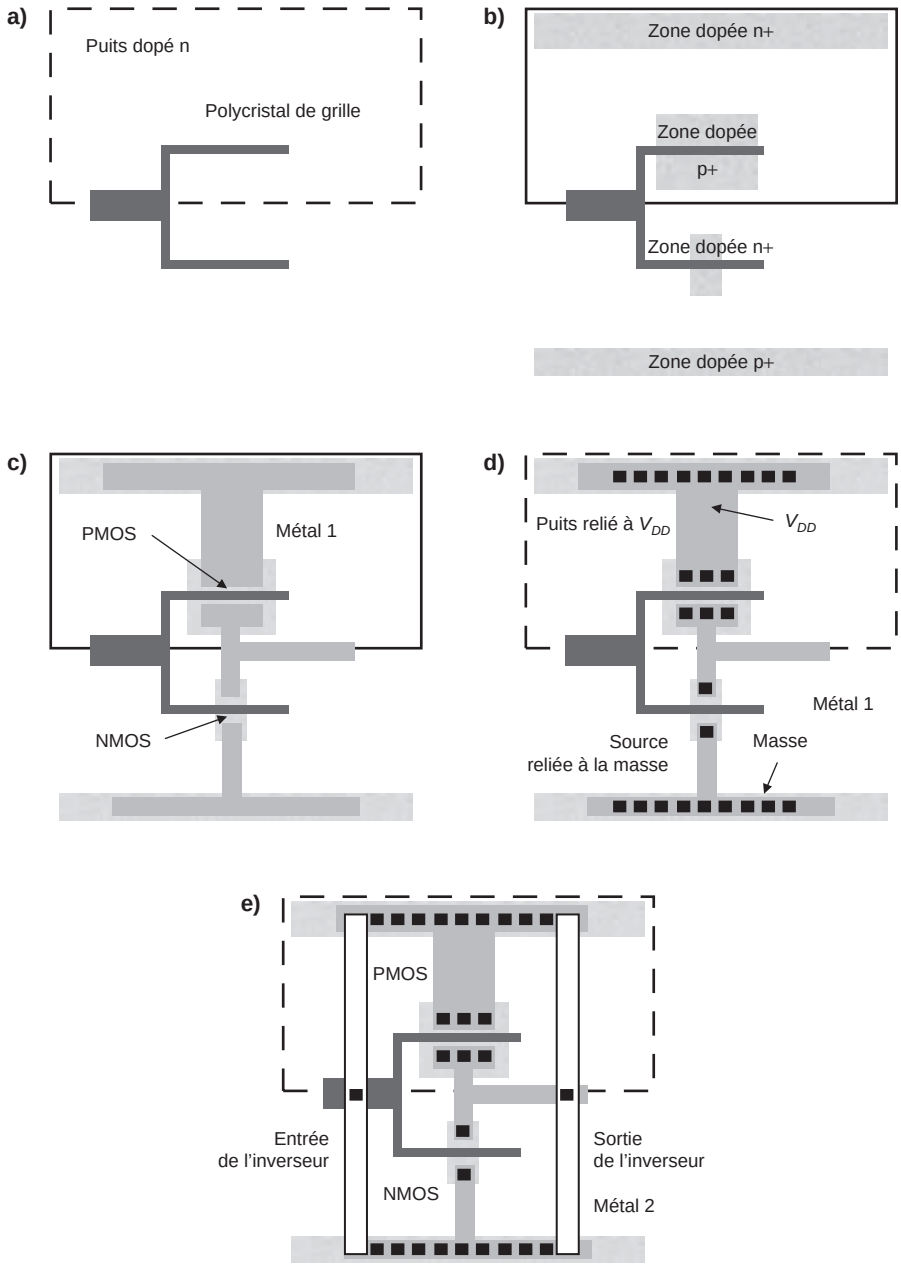


Figure 8.19 – Réalisation de l'inverseur CMOS.

Il est remarquable que la puissance dissipée dans les résistances ne dépende pas de la valeur de ces résistances. Il est possible de justifier physiquement ce résultat. C'est en fait l'énergie fournie par l'alimentation qui se dissipe pour la moitié dans les résistances et qui, pour l'autre moitié, est stockée dans la capacité $C + C_t$.

L'énergie dissipée en chaleur dans la commutation est proportionnelle à la capacité totale vue en sortie et au carré de la tension d'alimentation. Quand la tension de sortie revient à 0, il y a une dissipation de même valeur, et quand ces deux transitions se produisent à une fréquence *f*, l'énergie dissipée par unité de temps, c'est-à-dire la puissance, devient :

$$P = f(C + C_t)V_{DD}^2$$

On comprend immédiatement à la vue de cette formule l'intérêt de réduire la tension d'alimentation et l'impact de la fréquence de fonctionnement sur la consommation d'un circuit.

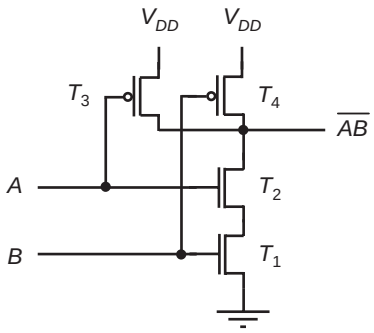
Pour terminer ce paragraphe important, il faut donner quelques éléments sur l'implantation physique de l'inverseur en technologie CMOS. La *figure 8.19* illustre étape par étape la fabrication de l'inverseur dans un procédé très simplifié. Les détails des opérations sont donnés chapitre 6.

- Dans une première étape (*figure 8.19a*), on réalise le caisson dopé *n* sur le silicium de base dopé *p*. Dans cette partie du silicium, on réalisera le PMOS. On dépose aussi le polycristal qui servira à réaliser les grilles des deux transistors de l'inverseur.
- Dans une seconde étape (*figure 8.19b*), on implante les zones dopées ou actives qui réalisent les régions dopées de source et de drain. Elles sont dopées *p*+ pour le PMOS et dopée *n*+ pour le NMOS. On réalise aussi les dopages de contact pour le puits du PMOS et pour la connexion de masse du NMOS.
- Dans une troisième étape (*figure 8.19c*), on réalise les pistes d'interconnexion en métal 1 qui servent à relier les deux transistors entre eux et à matérialiser les pistes d'alimentation et de masse.
- Dans une quatrième étape (*figure 8.19d*), on réalise les « vias » traversantes qui mettent en liaison électrique drain, source et piste de connexion du métal 1 de même que puits, substrat de base et pistes de masse et d'alimentation.
- Dans une dernière étape (*figure 8.19e*), les connexions d'entrée et de sortie de l'inverseur sont réalisées en métal 2 et connectées aux transistors.

8.4 Les autres fonctions logiques

Le lecteur peut facilement imaginer que des schémas inspirés de celui de l'inverseur permettent de réaliser d'autres fonctions logiques. Il n'est pas possible dans cet ouvrage général de décrire toutes les implémentations possibles. Nous nous limiterons à la description du NAND et du NOR à deux entrées, et de l'étage additionneur.

La *figure 8.20* montre le schéma d'un NAND à deux entrées ainsi que sa table de vérité. Si les deux entrées sont dans l'état « 1 », les deux NMOS conduisent et les deux PMOS sont bloqués. Dans ce cas, la sortie est à la masse soit dans l'état « 0 ». Dans tous les autres cas, la sortie est dans l'état « 1 ». Le dimensionnement des transistors est basé sur les règles de dimensionnement de l'inverseur. Il suffit de savoir que deux transistors en parallèle sont équivalents à un transistor de largeur égale à la somme des largeurs et que deux transistors en série sont équivalents à un transistor dont la longueur est la somme des longueurs. La tension de commutation de l'étage devient donc :



B \ A	0	1
0	1	1
1	1	0

Figure 8.20 – Réalisation d'un NAND.

$$V = \frac{\sqrt{\frac{\beta_n}{4\,\beta_p}}\,V_{T_N} + (V_{DD} - V_{T_p})}{1 + \sqrt{\frac{\beta_n}{4\,\beta_p}}}$$

En effet, les deux NMOS en série ont un *W/L* divisé par deux si les MOS sont identiques et les deux PMOS en parallèle ont un *W/L* multiplié par deux. Ce résultat se généralise à un système ayant *N* entrées. Le facteur de correction est alors N^2.

La tension de commutation de l'étage est donc différente de celle de l'inverseur et peut s'éloigner de la valeur optimale de $V_{DD}/2$ dans l'hypothèse où on a gardé le même rapport de taille entre le PMOS et le NMOS. Rien n'interdit cependant de changer cette règle qui n'était optimale que pour l'inverseur. Choisissons, par exemple, de dessiner les deux transistors de telle sorte que les dimensions soient égales. Comme le β_n du NMOS est plus élevé à dimension égale que le β_p du PMOS, le ratio $\beta_n/4\,\beta_p$ est proche de 1 et la tension de commutation de l'étage est proche de $V_{DD}/2$.

Le circuit NAND est donc intéressant puisqu'il permet de réaliser des fonctions performantes avec des transistors de tailles voisines. Il est donc volontiers utilisé dans la conception des circuits intégrés logiques. Un schéma d'implantation est représenté *figure 8.21* pour un NAND à trois entrées.

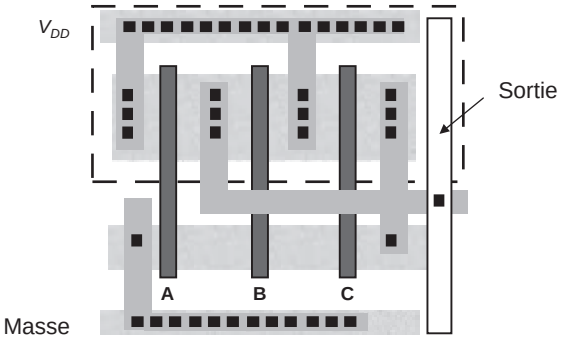
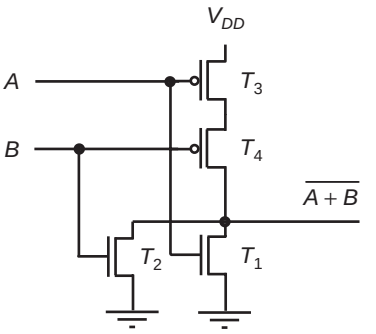


Figure 8.21 – NAND à trois entrées.

On peut de même réaliser une fonction NOR comme le montre la *figure 8.22*. Le fonctionnement en est très simple et les deux MOS de sortie sont en parallèle au lieu d'être en série comme dans la porte NAND.



B \ A	0	1
0	1	0
1	0	0

Figure 8.22 – Porte NOR à deux entrées.

Quand une des deux entrées est à l'état « 1 », un des deux NMOS est conducteur alors que l'un des PMOS est bloqué. La tension de sortie est alors à la masse. On peut également exprimer la tension de commutation de l'étage. On obtient :

$$V = \frac{\sqrt{\frac{4\beta_n}{\beta_p}} V_{T_N} + (V_{DD} - V_{T_p})}{1 + \sqrt{\frac{\beta_n}{4\beta_p}}}$$

La différence avec l'inverseur est cette fois défavorable puisque le ratio β_n/β_p est multiplié par 4 ce qui accentue encore la différence entre NMOS et PMOS. Pour un étage à N entrées la tension de commutation s'écrit :

$$V = \frac{\sqrt{\frac{N^2\beta_n}{\beta_p}} V_{T_N} + (V_{DD} - V_{T_p})}{1 + \sqrt{\frac{N^2\beta_n}{\beta_p}}}$$

Il est possible de donner quelques éléments pour estimer les retards et temps de montée de ce type de portes. On utilisera pour cela quelques résultats généraux.

Quand N PMOS identiques sont mis en parallèle, alors la constante de temps caractéristique est égale à $R_p/N(NC_t + C)$ en reprenant les notions du paragraphe 8.2. Le temps de propagation est alors :

$$t_{PLH} = 0{,}7 \frac{R_p}{N} (2NC_{tp} + C)$$

Quand *N* NMOS sont en série, le retard devient :

$$t_{PHL} = 0{,}35\, R_n\, C_{tn} \cdot 2\, N^2 + 0{,}7\, R_n \cdot NC$$

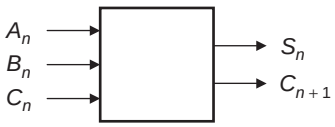
Rappelons que *C* est la capacité de charge extérieure. Dans cette relation le facteur 0,35 n'est pas évident. Il est le résultat de la résolution approximative du problème formé par une cascade de cellules *RC*. À l'aide de ces relations, il est facile d'estimer le retard des portes NAND et NOR étudiées dans ce paragraphe.

Les fonctions ET et OU sont en général réalisées avec des NAND et des NOR en appliquant les équations logiques suivantes :

$$\overline{A} + \overline{B} = \overline{AB}$$
$$\overline{A}\,\overline{B} = \overline{A + B}$$

Les fonctions logiques quelconques s'exprimant avec des ET, des OU et des inverseurs, comme il a été vu dans le paragraphe 8.1, on peut donc réaliser une fonction quelconque avec les briques de base que nous venons d'étudier.

Pour terminer ce chapitre, nous pouvons étudier comment il est possible à l'aide des portes de base de réaliser les opérateurs binaires permettant d'implémenter les fonctions de calcul dans un circuit intégré. Partant du fait que toutes les opérations peuvent se faire à partir de l'addition binaire, il suffit d'examiner comment il est possible de réaliser un additionneur. Dans une première étape examinons comment on effectue bit à bit une addition. Comme en décimal on fait la somme des deux bits en ajoutant le report si nécessaire. La table d'addition est la suivante :



A_n	B_n	C_n	S_n	C_{n+1}
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Figure 8.23 – Additionneur entier.

La *figure 8.24* montre une implémentation possible de cet additionneur. Un additionneur complet de *N* bits sera formé de *N* circuits de ce type.

Pour en arriver à ce schéma, il faut partir de la table de vérité et écrire :

$$S_n = \overline{A_n B_n} C_n + \overline{A_n} B_n \overline{C_n} + A_n \overline{B_n C_n} + A_n B_n C_n$$
$$\overline{C_{n+1}} = A_n B_n + C_n (A_n + B_n)$$

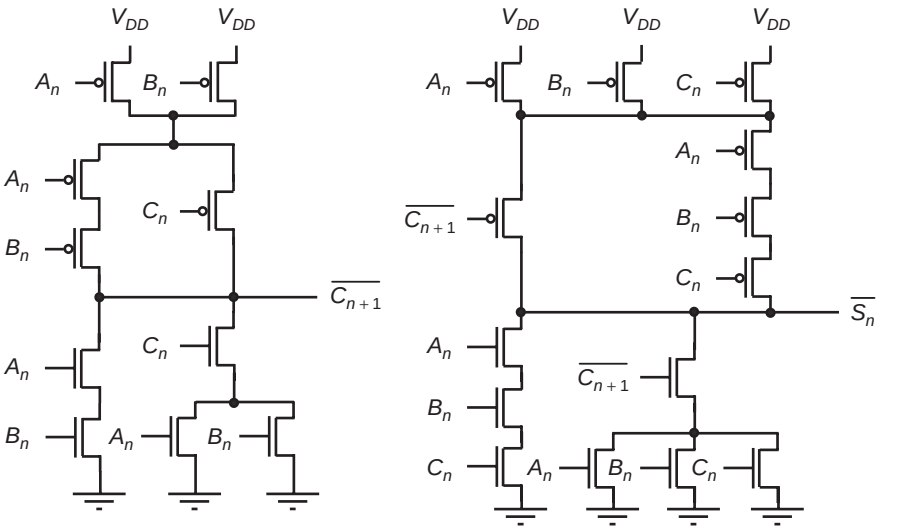


Figure 8.24 – Additionneur entier.

On peut montrer que :

$$\overline{C_{n+1}} = (\overline{A_n} + \overline{B_n})(\overline{C_n} + \overline{A_n B_n})$$

On peut aussi écrire S_n sous la forme suivante :

$$S_n = (A_n + B_n + C_n)\overline{C_{n+1}} + A_n B_n C_n$$

Ces deux dernières expressions expliquent les schémas proposés.

Il était possible de réaliser ces fonctions avec des portes de base directement à partir des équations logiques initiales. Cette manière de faire aurait conduit à un nombre de transistors beaucoup plus important.

8.5 *Flip-flop* et *latches*

Nous pouvons maintenant étudier les fonctions de base de la logique séquentielle évoquée dans le paragraphe 8.1. L'examen de ces fonctions de base permettra d'introduire les notions d'horloge et de logique synchrone.

La logique séquentielle est la logique dans laquelle il est nécessaire de prendre en compte les états logiques antérieurs dans la détermination des états d'un système. C'est finalement la forme la plus générale de la logique. La logique combinatoire est un cas particulier de la logique séquentielle.

La logique synchrone est un mode particulier de fonctionnement de la logique électronique (séquentielle ou combinatoire). Son principe est de définir les changements d'états en fonction de deux conditions distinctes : les états d'entrée changent et un autre signal est présent appelé horloge. En général, ce signal est une transition (du niveau bas vers le niveau haut ou l'inverse) et les changements d'états se produisent en synchronisme avec cette transition. La *figure 8.25* illustre cette notion fondamentale.

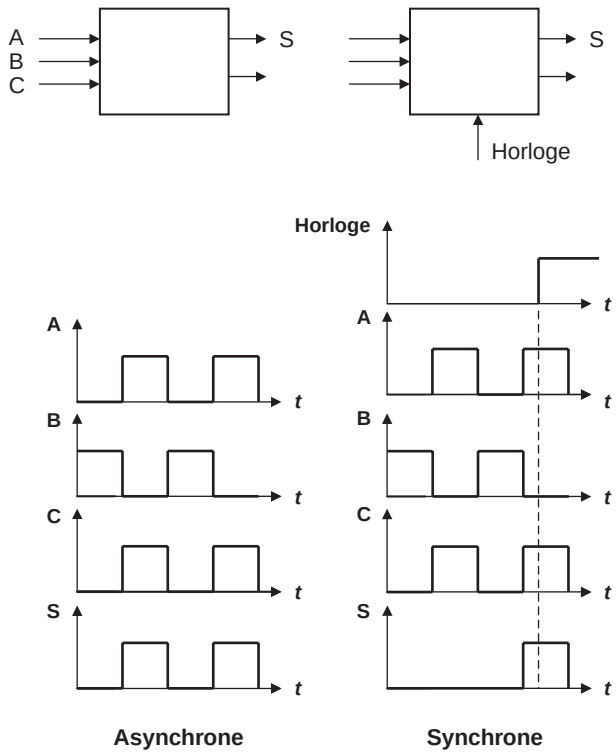


Figure 8.25 – Logique synchrone et asynchrone.

Dans le système asynchrone, un simple changement d'état d'une des entrées suffit à induire un changement d'état d'une sortie. Dans un système synchrone, il faut en plus que ce changement soit associé à une transition du signal d'horloge comme le montre le graphique de droite. Les systèmes électroniques actuels sont en grande majorité synchrones car ce type de logique est beaucoup plus robuste qu'une logique asynchrone. En effet, un signal perturbateur d'intensité suffisamment forte peut créer de manière transitoire un changement d'état. Dans un système asynchrone, ce changement d'état sera transmis. Dans un système synchrone, il ne sera pris en compte que si l'effet perturbateur survient au moment précis d'une transition de l'horloge ce qui est beaucoup plus improbable. Les systèmes asynchrones sont cependant envisagés dans les technologies avancées pour des raisons qui seront exprimées dans le chapitre 11.

La *figure 8.25* pourrait laisser penser que les changements d'états des entrées-sorties sont plus fréquents que les transitions de l'horloge. En réalité, il n'en est rien et les changements d'états représentés sur la *figure 8.25* ne sont aussi fréquents que pour illustrer la différence de fonctionnement entre circuits synchrones et circuits asynchrones. Dans les circuits logiques, le signal d'horloge est celui qui varie le plus souvent et de plus les transitions se font de manière très régulière comme il est représenté *figure 8.26*. On peut ainsi définir une période et une fréquence de fonctionnement. À chaque transition, les blocs logiques réalisent leur fonction logique à la condition que les signaux soient stables au moment de la transition.

La fonction logique est réalisée au moment de la transition (de l'état bas vers l'état haut dans l'exemple) à la condition que les états logiques soient bien établis. Le niveau *D* doit par exemple être établi

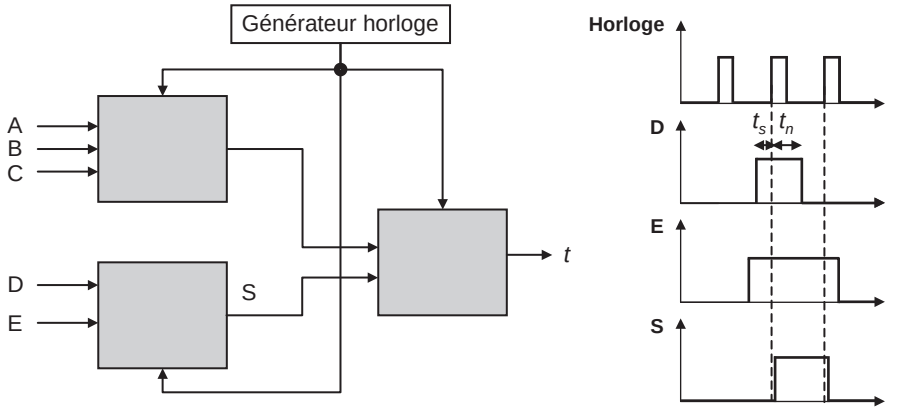


Figure 8.26 – Circuit synchrone.

avant la transition un temps t_s supérieur à une valeur minimale et doit être maintenu un temps t_h supérieur à une valeur minimale. Ces deux temps sont appelés temps de « set-up » et temps de « hold ». Quand ces conditions ne sont pas satisfaites, l'état de sortie n'est pas défini de manière certaine. Il est donc nécessaire dans la conception des circuits logiques de s'assurer que ces conditions soient respectées. Les temps de transit des signaux dans les connexions doivent être pris en compte ainsi que les temps de propagation dans les portes logiques.

La logique séquentielle conduit également à réaliser des banques de registres. Le chapitre 9 montre que toute logique séquentielle peut être réalisée avec des blocs combinatoires (des portes NOR, des portes NAND, des inverseurs) et des registres. Un registre est un ensemble d'éléments de mémorisation de un bit chacun. Un registre 8 bits par exemple comporte 8 éléments de mémorisation un bit. La fonction de ces registres est de conserver un état logique donné. Une sortie d'un bloc combinatoire est par exemple mise en mémoire au moment d'une transition de l'horloge. Pour réaliser ces blocs de mémorisation on peut utiliser des *flip-flops* et des *latches*.

- Le *latch* garde en mémoire la valeur de son entrée quand un signal d'activation appelé *enable* l'autorise. La mémoire est active sur un état.
- Le *flip-flop* mémorise également la valeur de l'entrée mais quand un signal d'activation (l'horloge) change d'état. La mémoire est active sur un flanc. La transition peut se faire de l'état bas vers l'état haut ou de l'état haut vers l'état bas selon le type de logique.

Les exemples de signaux donnés sur la *figure 8.27* montrent la différence de comportement entre les *latches* et les *flip-flops*.

Il est maintenant possible de donner quelques éléments sur la réalisation de ces deux fonctions de base. Il existe de nombreuses manières de réaliser ces fonctions et leur schéma exact est une partie importante du savoir faire des fondeurs de silicium. Nous donnerons quelques exemples de réalisation sans trop entrer dans les détails.

Le *latch* est très simple à réaliser puisque deux portes NAND ou NOR suffisent comme le montre la *figure 8.28*.

On en déduit immédiatement la réalisation dans une technologie CMOS en utilisant les résultats du paragraphe 8.4.

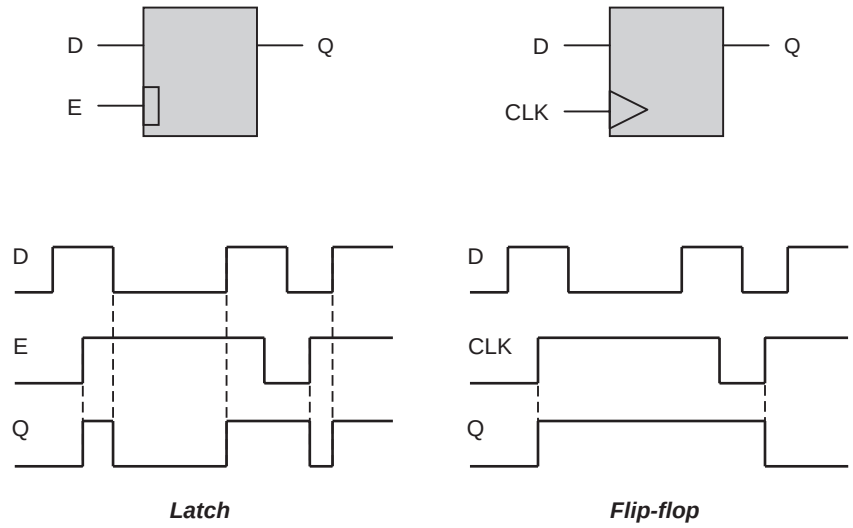


Figure 8.27 – Latches et flip-flops.

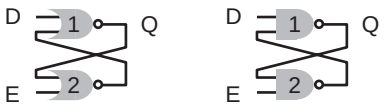


Figure 8.28 – Schémas des latches.

La réalisation des *flip-flops* est plus complexe. Nous ne donnerons ici qu'un exemple de *flip-flop* sensible à une transition d'horloge et réalisé à partir d'inverseurs couplés.

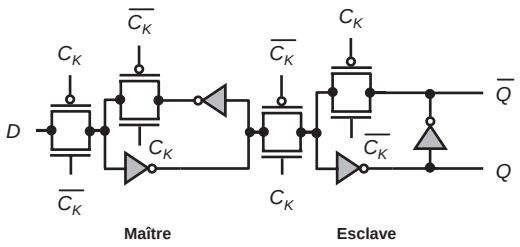


Figure 8.29 – Exemple de flip-flop.

Quand l'horloge est à l'état bas, le premier étage (le maître) capture la donnée d'entrée tandis que le second étage (l'esclave) conserve la valeur de son entrée. Quand l'horloge passe à l'état haut, la donnée passe deux inverseurs et est transmise sur la sortie *Q*. Quand l'horloge revient au niveau bas, la valeur de *Q* est maintenue. Il faut que l'horloge revienne au niveau haut pour qu'une autre valeur de l'entrée puisse être prise en compte. Pour bien comprendre le fonctionnement de ce circuit, on

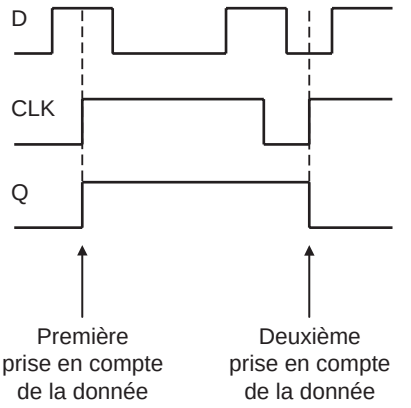


Figure 8.30 – Fonctionnement du flip-flop.

peut dans un premier temps considérer les MOS comme des interrupteurs ouverts ou fermés. Les propriétés plus subtiles relatives à la stabilité de ce montage nécessitent de détailler le circuit en transistors élémentaires et de calculer le comportement électrique comme on le ferait pour un circuit analogique. Nous nous contenterons dans ce paragraphe de représenter les signaux en fonction du temps comme le montre la *figure 8.30*.

Rappelons les conditions de bon fonctionnement d'un *flip-flop*. La donnée pour être prise en compte au moment de la transition doit être présente un temps minimum avant la transition et doit être maintenue un temps minimum après la transition. On comprend donc l'intérêt d'utiliser des signaux d'horloge ayant des flancs très raides. En pratique, ces temps minimaux sont de quelques dizaines de picosecondes pour des technologies avancées.

8.6 La logique dynamique

Pour terminer ce chapitre consacré à la logique, il est nécessaire d'introduire la logique dite dynamique mise en œuvre quand les contraintes de consommation et de vitesse sont fortes. La logique dite « domino » sera également introduite.

L'idée de base de la logique dynamique est d'utiliser la capacité d'entrée d'un MOSFET pour mettre un état en mémoire. La *figure 8.31* illustre le principe de base de la logique dynamique. Un état logique donné est conduit à travers un transistor monté en commutateur lors d'une transition d'horloge et charge la capacité d'entrée d'un MOSFET. Cet état est maintenu tant que la capacité ne peut pas se décharger.

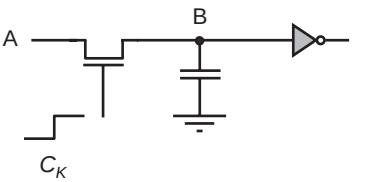


Figure 8.31 – Logique dynamique.

Si on appelle I_{off} le courant de fuite du transistor d'entrée, l'équation de décharge du condensateur d'entrée s'écrit :

$$\frac{dV}{dt} = \frac{I_{off}}{C}$$

Des valeurs typiques dans une technologie 50 nm conduisent à :

$$\frac{dV}{dt} = \frac{4\,nA}{50\,fF} = 80\,mV/\mu s$$

Il est alors possible d'imaginer des systèmes logiques à la condition d'ajuster les phases des horloges. La notion d'horloges non recouvrantes sera naturellement introduite. Le circuit représenté *figure 8.32* est un registre à décalage piloté par des horloges soigneusement synchronisées. Le signal logique d'entrée se propage d'étage en étage à chaque front d'horloge. Les inverseurs permettent de restaurer le niveau du signal quand il se propage.

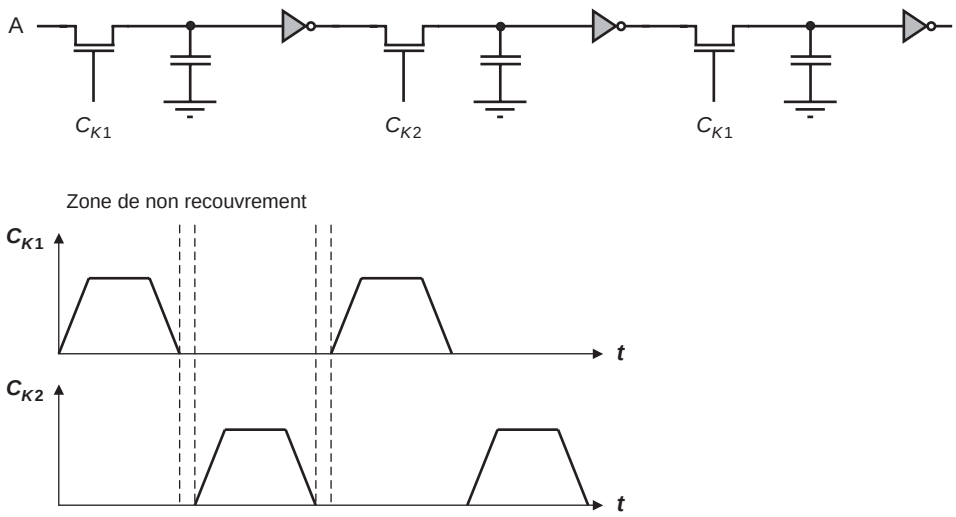


Figure 8.32 – Registre à décalage en logique dynamique.

Pour un fonctionnement correct du circuit, il est impératif que les signaux d'horloge ne se recouvrent pas car sinon un état et son contraire pourraient être présents sur la même capacité de stockage ce qui conduit à un état indéterminé.

La logique dynamique synchrone est basée sur le même principe de mémorisation d'un état sur une capacité d'entrée mais utilise en plus les notions de « précharge » et d'« évaluation ». Ce principe est illustré *figure 8.33* dans l'exemple d'une porte NAND à trois entrées.

Le fonctionnement est alors le suivant :

- L'horloge est initialement à l'état bas, le transistor de précharge est passant et la sortie est à l'état haut.

- L'horloge passe à l'état haut et la porte entre dans la phase d'évaluation. Si toutes les entrées sont dans l'état « 1 », le transistor d'évaluation étant passant, la sortie est à la masse. Cet état est maintenu sur la capacité de stockage tant que l'horloge reste à l'état haut. Les entrées peuvent alors changer d'état, la sortie reste à la masse.
- Quand l'horloge revient à l'état bas, la sortie est au potentiel V_{DD} car le transistor de précharge est passant.

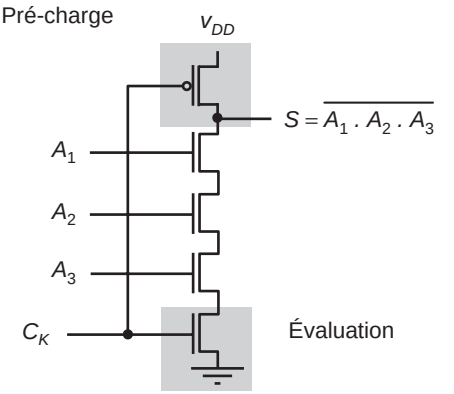


Figure 8.33 – NAND en logique dynamique.

La logique précharge-évaluation, que nous noterons PE dans la suite, est intéressante car elle utilise peu de transistors. On peut également prouver que la consommation est plus faible que celle de la logique CMOS classique. En effet, une entrée est reliée à un seul transistor et non à deux ce qui réduit les capacités.

Le dessin des transistors est plus souple que dans la logique CMOS classique car la définition d'une tension de commutation de l'étage n'a pas de sens dans ce type de logique. Les transistors seront donc dessinés pour satisfaire les contraintes de consommation et de vitesse. Les états logiques sont réalisés uniquement pendant une fraction du temps, le temps pendant lequel l'horloge est à l'état haut dans l'exemple du NAND. Toutes ces raisons expliquent que la logique PE est de plus en plus utilisée dans le design des circuits logiques avancés.

La logique dite domino utilise la logique PE pour définir une architecture générale comme il est indiqué *figure 8.34*, page suivante. Remarquons que les blocs logiques sont réalisés avec des NMOS si on généralise l'exemple du NAND. Ils pourraient l'être avec des PMOS.

Le fonctionnement global de cette architecture est cependant problématique dans cette première version et des états mal définis peuvent apparaître suite à des décalages temporels entre l'horloge et les signaux de sortie. Des étages inverseurs sont donc ajoutés entre les blocs logiques pour corriger ce défaut.

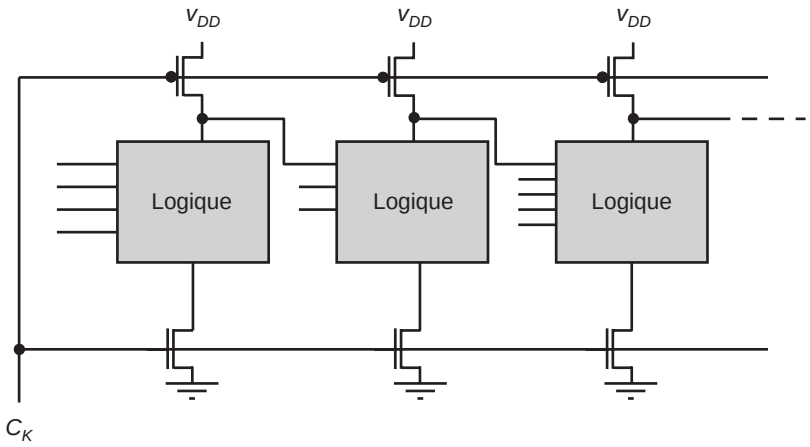


Figure 8.34 – Principe de la logique domino.

Chapitre 9

Les circuits intégrés complexes

Tous les circuits intégrés sont maintenant complexes. Pensons aux puces utilisées dans la majorité des objets de la vie quotidienne : téléphones, téléviseurs, ordinateurs personnels… Cependant, ces systèmes intégrés complexes sont construits à partir de deux fonctions de base : la fonction mémoire et la fonction calcul. Ce chapitre a pour objectif de donner les grands principes architecturaux associés à chaque fonction et de faire la liste des technologies permettant de les réaliser. L'évolution de la micro-électronique montre que de plus en plus souvent les deux fonctions de base (mémoire et calcul) coexistent sur une même puce, ce qui introduit le concept de système sur puce ou *system on chip* en anglais.

9.1 Les différents types de circuits intégrés

9.1.1 *Les fonctions*

Il y une grande diversité de circuits intégrés. Il est cependant possible de les classer en deux grandes familles : les circuits de type mémoire et les circuits de type calcul. Cette classification est effectuée selon la fonction réalisée par le circuit. La fonction calcul doit être comprise au sens large : capacité à transformer les grandeurs continues ce qui constitue le calcul analogique ou capacité à effectuer des calculs binaires ce qui constitue le calcul numérique. La fonction calcul comprend aussi la détermination des expressions booléennes en logique combinatoire ou en logique séquentielle. De nombreux circuits intégrés combinent aujourd'hui fonctions analogiques et fonctions numériques, ils sont appelés circuits mixtes. La *figure 9.1* illustre cette classification.

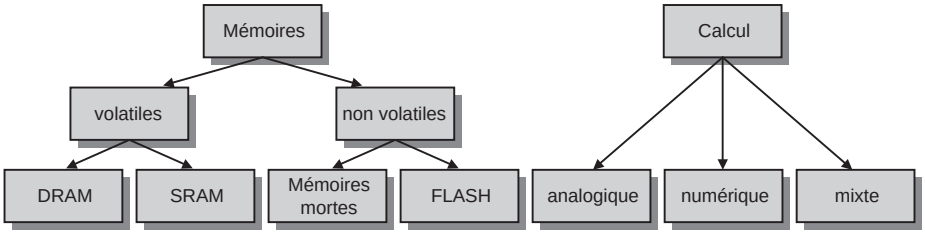


Figure 9.1 – Classement des circuits intégrés par fonction.

Les mémoires ont pour fonction principale de conserver des données : nombres, lettres, images numérisées, texte de programme, fichiers binaires… Elles fonctionnent en deux temps :

- Dans une première phase, on écrit la donnée (un ensemble de bits) dans la mémoire à un emplacement donné identifié par un nombre binaire appelé adresse. La donnée *X* est par exemple stockée à l'adresse « 01110001110010 ». La donnée est fondamentalement un bit mais peut être un ensemble de bits, en général 8, 16, 32, 64 bits. On parle alors de mots et on appelle byte un mot de 8 bits. On peut également employer le terme octet pour désigner 8 bits. La taille d'une mémoire est le nombre d'adresses différentes possibles car on ne peut stocker deux données différentes à la même adresse. Un espace d'adressage de 8 bits permet de stocker 256 données distinctes.
- Dans une seconde phase, on lit la donnée en appliquant en entrée de la mémoire une adresse. Le contenu de l'emplacement mémoire correspondant, un bit ou un mot, est alors placé sur un bus de sortie et peut être utilisé par le système électronique.

La structure la plus générale d'une mémoire est indiquée *tableau 9.1* en se limitant à 8 adresses.

Les tailles des mémoires, le nombre d'emplacements distincts et également le nombre d'adresses différentes, sont exprimées en bits quand la donnée est un bit et en octets quand la donnée est un octet. La micro-électronique produit en 2005 des mémoires de plusieurs gigabits.

Les mémoires sont classées en deux familles : les mémoires volatiles et les mémoires non volatiles. Les mémoires volatiles ne conservent les données que si la tension d'alimentation est maintenue. Dans le cas contraire, les données sont perdues. Les mémoires non volatiles conservent les données enregistrées même en cas de coupure de tension d'alimentation. Les clés USB et les cartes mémoires pour appareils photographiques numériques sont des exemples de mémoires non volatiles.

Tableau 9.1 – Structure d'une mémoire.

Adresse	Donnée
000	mot 1
001	mot 2
010	mot 3
011	mot 4
100	mot 5
101	mot 6
110	mot 7
111	mot 8

Les mémoires volatiles sont composées de deux types : les mémoires statiques ou SRAM et les mémoires dynamiques ou DRAM. La différence est dans la technologie de fabrication comme il sera expliqué dans le paragraphe 9.2.

Les mémoires non volatiles sont classées en deux familles : les mémoires mortes et les mémoires Flash. Les mémoires mortes sont destinées à conserver des données qui ne sont jamais modifiées. La configuration d'un système électronique ou l'identifiant d'un produit peuvent être mémorisés dans une mémoire morte. À l'inverse, on peut facilement et rapidement modifier le contenu d'une mémoire flash de manière électrique. Les données sont alors conservées en cas de coupure de l'alimentation électrique.

Donnons maintenant quelques commentaires sur les fonctions de calcul. Les circuits intégrés analogiques réalisent des fonctions sur des grandeurs électriques variant continûment. La fonction amplification, par exemple, consiste à multiplier une tension ou un courant par un nombre supérieur à l'unité. Les fonctions de filtrage sont également très utilisées comme la dérivation ou l'intégration. Elles transforment le signal en un signal qui est soit la dérivée soit l'intégrale du signal d'entrée au sens mathématique du terme. Ces opérations sont équivalentes dans le domaine fréquentiel à un filtrage passe-haut pour la dérivation et passe-bas pour l'intégration. Des filtres plus complexes peuvent être réalisés avec des transistors et des composants passifs. Enfin, il est souvent nécessaire de transformer une tension en un mot binaire représentatif de son amplitude. Cette opération est la conversion analogique numérique.

Donnons maintenant quelques éléments sur les fonctions numériques de calcul. Les circuits intégrés numériques réalisent des opérations booléennes combinatoires ou séquentielles. Les fonctions combinatoires sont réalisées avec des portes logiques comme il a été vu dans le chapitre précédent alors que les fonctions séquentielles demandent en plus des registres de mémorisation d'états. Ce point important sera illustré par un exemple dans le paragraphe 9.3.

Les circuits intégrés mixtes comportent à la fois des fonctions analogiques et numériques. Ils se sont considérablement développés à cause de la croissance des techniques de traitement numérique des signaux et des images dans le domaine des technologies de l'information : son numérique, DVD, photographie numérique, télévision numérique… La *figure 9.2* montre dans un circuit intégré mixte les différentes parties constitutives : fonctions analogiques, fonctions numériques de calcul et fonctions mémoires.

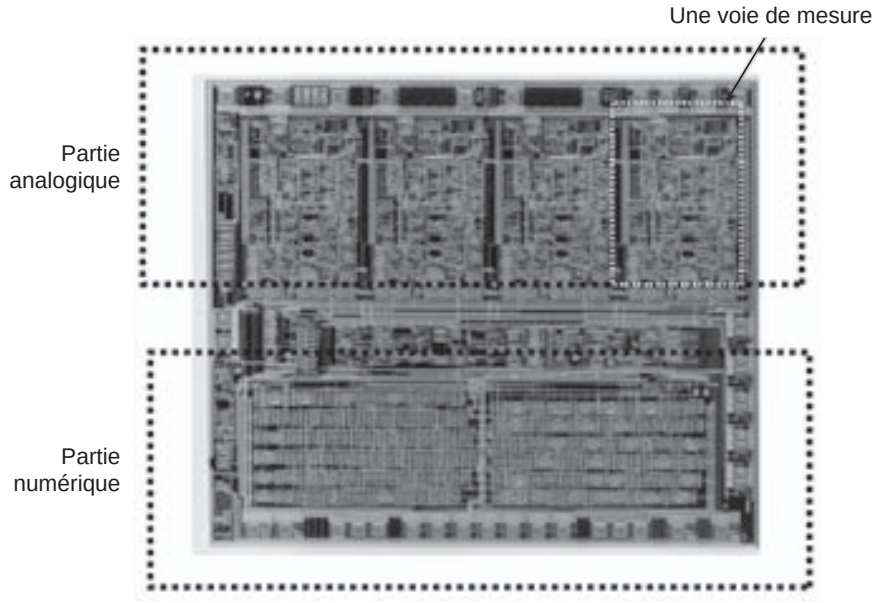


Figure 9.2 – Circuit intégré mixte (Crédit photo CEA Grenoble).

9.1.2 *Les technologies*

Les circuits intégrés peuvent également se classer en fonction des technologies de fabrication.

On distingue alors quatre familles : les circuits programmables de type FPGA (*Field Programmable Gate Array*), les circuits prédiffusés, les circuits précaractérisés et les circuits *full custom*.

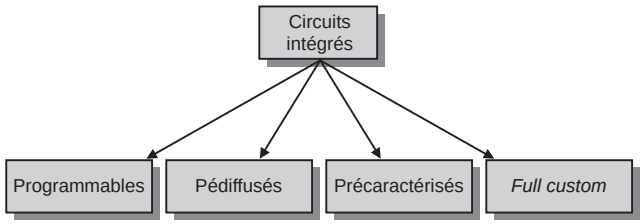


Figure 9.3 – Classification des circuits intégrés selon les technologies.

Cette classification correspond à différentes technologies de conception et de fabrication des circuits. Tous les circuits définis précédemment peuvent être fabriqués dans une technologie choisie parmi les quatre. En pratique, certaines technologies sont mieux adaptées à certaines fonctions.

Les circuits programmables sont un ensemble de fonctions logiques de base qu'il est possible d'interconnecter à la demande par programmation électrique. Ces fonctions de base ne sont pas des portes logiques élémentaires mais des fonctions plus complètes : opérateurs arithmétiques sur un bit, fonctions logiques de plusieurs variables, registres d'un bit. Ces blocs de base appelés CLB peuvent être

programmés individuellement puis interconnectés à la demande par des interrupteurs programmables.

En pratique, l'utilisateur décrit l'architecture qu'il souhaite réaliser. Cette description est écrite dans un langage adapté, le langage VHDL par exemple. Ensuite, le logiciel de synthèse fourni par le vendeur de FPGA détermine les interconnexions à réaliser. Les CLB et les interconnexions sont alors programmés électriquement de manière automatique. Si l'utilisateur désire modifier l'architecture, il suffit de décrire cette nouvelle architecture puis de passer par une nouvelle phase de synthèse.

On conçoit donc la grande souplesse d'utilisation de ce type de circuit, ce qui explique la croissance exceptionnelle de la technologie FPGA depuis les années 90. Les FPGA peuvent réaliser des circuits équivalents à des millions de portes logiques. Ajoutons à cela que certains modèles intègrent des fonctions complexes et des fonctions mémoire dans le silicium selon une architecture décrite *figure 9.4*.

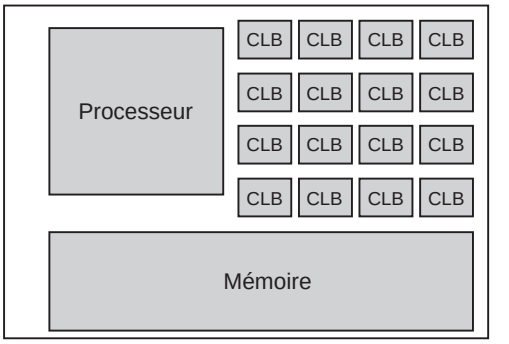


Figure 9.4 – Architecture d'un FPGA.

La *figure 9.4* représente de manière très simplifiée l'architecture d'un FPGA comprenant un circuit processeur, une partie mémoire et un ensemble de blocs programmables, les CLB. Le réseau d'interconnexions programmables n'est pas représenté sur la figure.

Ces circuits présentent cependant un certain nombre d'inconvénients. Ils consomment beaucoup plus de portes pour une fonction donnée qu'un circuit fabriqué spécifiquement à partir de portes élémentaires. La consommation est élevée à cause du nombre de portes et des longueurs d'interconnexions non optimisées. Le coût est élevé. La technologie à base de PFGA est donc réservée à des séries relativement modestes ou quand le critère de choix principal est la réduction du temps de développement. Des progrès constants dans les technologies de réalisation et de programmation font cependant que les FPGA sont de redoutables concurrents des circuits intégrés spécifiques (prédiffusés, précaractérisés ou circuits *full custom*) quand le nombre de pièces à produire est inférieur à quelques dizaines de milliers de pièces.

Les circuits prédiffusés comportent un ensemble de cellules logiques réalisées de manière régulière dans le silicium. On emploie également le terme de mer de portes pour désigner ce type de circuit. Ces cellules ne sont pas interconnectées et il est nécessaire de personnaliser le circuit à la fabrication en implémentant les interconnexions entre portes. Une reprise de conception implique donc une nouvelle fonderie. On comprend facilement que la surface de silicium n'est pas optimisée puisque certaines portes peuvent être inutilisées.

Les circuits précaractérisés, contrairement aux prédiffusés, n'utilisent pas de fonctions de base réalisées dans le silicium mais uniquement des schémas électriques et des dessins qui permettent de réaliser ces fonctions de base. Ces éléments sont appelés des bibliothèques. Ces fonctions sont par exemple des portes logiques, des opérateurs binaires, des registres, des compteurs, des éléments mémoire… La phase de conception consiste donc à choisir parmi ces éléments prédéfinis ceux qui sont nécessaires à la réalisation du circuit puis à prévoir leur placement et les interconnexions associées. À l'issue de cette phase de conception, les différents blocs et les interconnexions sont ensuite réalisés dans la phase de fonderie. On comprend que ce type de circuit permet une utilisation optimale de la surface du silicium à la condition que les bibliothèques utilisées soient bien adaptées aux fonctions à réaliser.

Quand il est nécessaire de concevoir et de dessiner les bibliothèques de fonctions, on parle alors de circuits *full custom* ou spécifiques. Les fonctions numériques et analogiques sont alors définies au niveau du transistor. Les transistors eux-mêmes sont dessinés de manière spécifique. En réalité, un design n'est jamais totalement *full custom* et certains blocs sont réutilisés. La capacité d'une entreprise à réutiliser les blocs et les fonctions participe de manière décisive à la productivité de la conception des circuits intégrés.

Le flot de conception des circuits intégrés est résumé dans la *figure 9.5*, qui indique de manière simplifiée les différentes étapes.

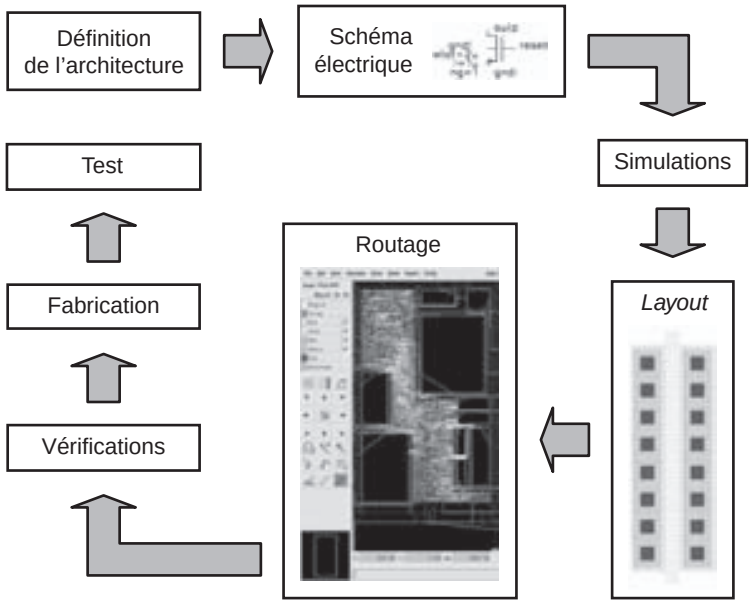


Figure 9.5 – Les étapes dans la conception d'un circuit intégré.

Le cycle décrit dans la *figure 9.5* s'applique aussi bien à une fonction de base (élément d'une bibliothèque) qu'à un système complexe composé de multiples fonctions de base. Les outils de description et de simulation seront différents en fonction du type de fonction à réaliser : fonction de base ou système complexe.

Pour terminer cette introduction, il est possible de croiser les deux classifications des circuits intégrés, classification selon la fonction et selon la technologie de fabrication. Cet exercice conduit à la *figure 9.6*. Les cases grisées de la matrice indiquent les technologies possibles pour une fonction donnée.

	DRAM	SRAM	Mortes	Flash	Analogique	Numérique	Mixte
Programmable							
Prédiffusé							
Précaractérisé							
Full custom							

Figure 9.6 – Fonctions et technologies.

9.1.3 *Les technologies nouvelles*

Les technologies actuelles de conception et de fabrication des circuits intégrés connaîtront-elles de fortes évolutions dans les années futures ? L'évolution la plus importante est la convergence des circuits intégrés en deux grandes familles assez différentes des familles vues dans le paragraphe précédent. Ces deux familles de circuits intégrés sont : les circuits complexes et les circuits simples.

Par circuit complexe, il faut comprendre un circuit intégrant différents processeurs, des mémoires réparties et une partie de logique programmable. Ces circuits à usage assez général pourraient remplacer les processeurs d'aujourd'hui et les circuits spécifiques à haute performance que l'on peut trouver dans des produits comme le téléphone portable ou la console de jeux. Les contraintes de capacité de traitement, de consommation et de flexibilité pourraient être satisfaites dans des architectures de ce type à la condition de pouvoir y porter sans trop de difficulté les logiciels applicatifs. Les technologies de type précaractérisé sont alors bien adaptées.

Par circuit simple, il faut entendre les circuits associés aux objets de la vie quotidienne et offrant une capacité de calcul et de communication plus réduite. Ce sont les circuits des étiquettes RFID, les circuits dans les objets domestiques et les capteurs de surveillance de l'environnement. Le faible coût et la très faible consommation sont alors les contraintes principales. Dans ce cas, les circuits spécifiques *full custom* permettent une optimisation complète.

Il est également possible de s'interroger sur l'avenir de la technologie CMOS. Rappelons que cette technologie silicium s'est progressivement généralisée dans tous les domaines d'application à l'exception peut-être du domaine radiofréquence, domaine dans lequel les transistors bipolaires gardent un avantage de principe lié à leur transconductance élevée. L'hypothèse la plus probable cependant est que la technologie CMOS poursuivra sa progression inéluctable en réduisant la taille du transistor et que les autres technologies ne viendront qu'en complément pour réaliser des fonctions particulières. Parmi ces fonctions, on peut citer la fonction mémoire non volatile et les fonctions d'interface avec l'environnement comme la détection de la lumière ou les interfaces avec le vivant. Pour réaliser ces fonctions, le matériau silicium n'est pas nécessairement le meilleur candidat.

La *figure 9.7* montre comment au fil du temps la technologie CMOS s'est enrichie en intégrant d'autres technologies ayant des avantages spécifiques.

D'autres semi-conducteurs ont été intégrés au silicium sous forme de puces rapportées pour la réalisation de photodiodes ou de diodes laser d'émission. Des matériaux magnétiques sont ajoutés

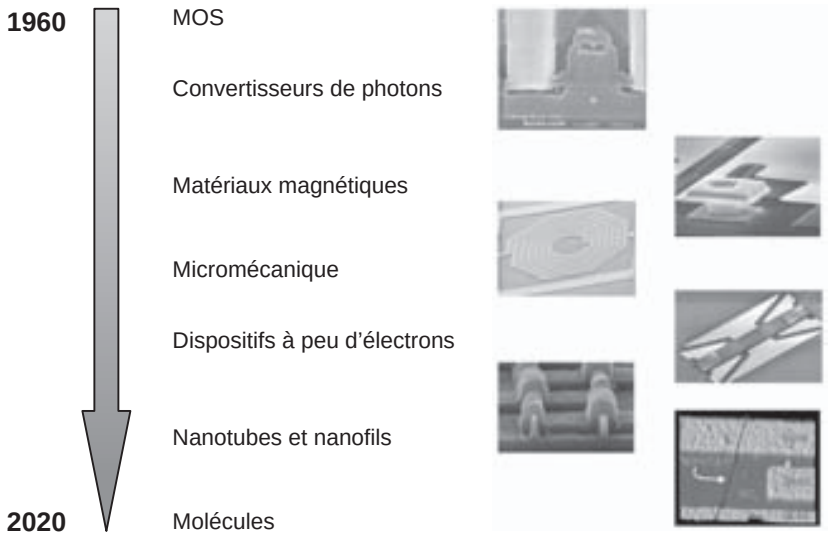


Figure 9.7 – Évolution de la technologie CMOS.

pour réaliser des mémoires non volatiles. Des dispositifs mécaniques à l'échelle de quelques microns permettent d'ajouter à un circuit CMOS des interrupteurs ou des oscillateurs de très bonne qualité que la technologie CMOS seule ne saurait faire. À plus long terme, on peut envisager d'intégrer des dispositifs à peu d'électrons, des nanofils ou des nanotubes et même des molécules comme il sera étudié dans le chapitre 12.

9.2 Les mémoires électroniques

Les mémoires électroniques ont pour fonction principale de conserver des données binaires. Les contraintes de coût, de vitesse et de consommation électrique conduisent à une optimisation de la cellule mémoire élémentaire. En dépit des progrès technologiques, le temps d'écriture et de lecture des mémoires reste notablement plus élevé que le temps de cycle d'un processeur. Ce problème appelé *memory gap* dans la littérature reste une des difficultés de la micro-électronique de demain. Il faut ajouter à cela les difficultés de réalisation des mémoires non volatiles pour comprendre les défis à relever dans la production industrielle des mémoires.

9.2.1 *Les mémoires dynamiques ou DRAM*

Les mémoires dynamiques permettent d'atteindre les densités les plus élevées car la cellule mémoire de base est réduite au minimum, un condensateur de stockage et un transistor MOS d'isolation permettant d'adresser la capacité. En contrepartie, il est nécessaire de rafraîchir la cellule mémoire de manière régulière car la charge stockée dans le condensateur s'écoule par les courants de fuite du transistor. Cette contrainte conduit à ajouter des fonctions électroniques capables d'effectuer cette opération de rafraîchissement et introduit des limitations temporelles dans les cycles de lecture et d'écriture des données. La cellule de base et l'architecture générale de la mémoire DRAM sont indiquées *figure 9.8*.

Les emplacements possibles sont au nombre de 2^{m+n}. La donnée est un bit prenant deux états possibles. Dans la phase d'écriture, une ligne et une colonne sont choisies parmi 2^m et 2^n valeurs pos-

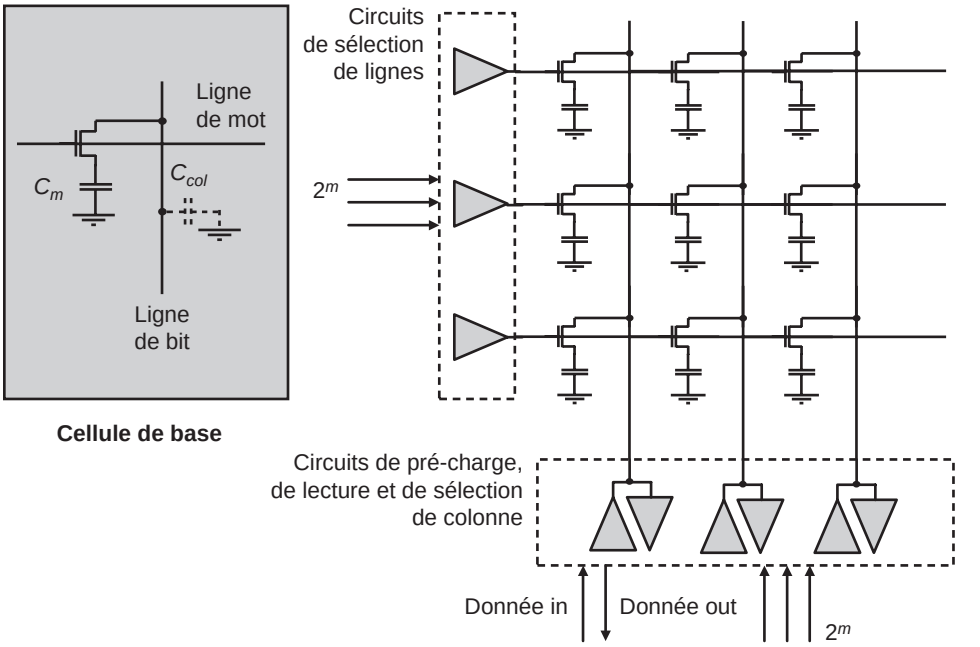


Figure 9.8 – Mémoire DRAM.

sibles. La charge correspondant à l'état de la donnée (« 0 » ou « 1 ») est stockée dans la capacité C_m, à l'intersection de la ligne et de la colonne. Dans la phase de lecture, une ligne et une colonne sont choisies puis la charge stockée est lue par l'amplificateur de colonne. Les *figures 9.9* et *9.10* illustrent les séquences temporelles.

Les bits d'adresse désignent à la fois lignes et colonnes. Ils sont supposés appliqués de manière synchrone. Au moment du front du signal *R/W* les bits d'adresses sont pris en compte. Un temps t_w après, correspondant au traitement des signaux par les circuits, la valeur de la donnée est appliquée sur la capacité de stockage à l'intersection de la ligne et de la colonne. De la même manière, un temps t_r après le front du signal de sélection C_S, la valeur de la tension aux bornes de la capacité de stockage correspondant aux adresses présentes est lue par le circuit en bout de colonne et conditionne la valeur de la donnée de sortie. Les mêmes opérations peuvent se faire pour toutes les adresses de la mémoire.

Pour estimer la consommation électrique et les vitesses d'écriture et de lecture, il faut détailler un peu plus les circuits de commande et de lecture. On raisonnera avec les données d'une technologie avancée soit avec une tension d'alimentation de 1 V.

Quand la donnée est « 0 », une tension nulle est appliquée aux bornes de la capacité. Quand la donnée est à l'état « 1 », une tension égale à V_{DD} est appliquée aux bornes de la capacité C_m. À la lecture, il ne faut pas croire que ces valeurs apparaissent sur les lignes de colonne. En effet, la charge stockée se répartit entre la capacité de stockage et la capacité de la colonne qui est loin d'être négligeable.

Imaginons par exemple une ligne de 100 nm de large et de 100 microns de long. La capacité est alors d'environ 1 fF. Si on suppose que 256 lignes sont dans le bloc mémoire envisagé et que chaque transistor présente une capacité d'entrée de 0,5 fF, la capacité de la colonne C_c est donc égale à

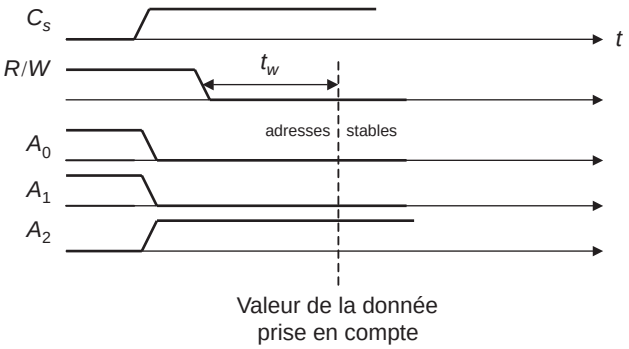


Figure 9.9 – Écriture dans la DRAM.

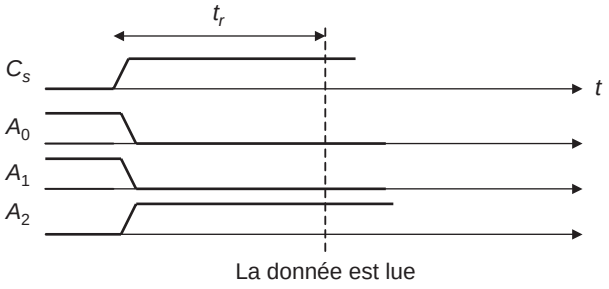


Figure 9.10 – Lecture de la DRAM.

$256 \times 0{,}5$ fF + 1 fF, soit environ 120 fF. On suppose également que la tension de la colonne est réglée à $V_{DD}/2$. Pour trouver la nouvelle valeur V_b, lue sur la ligne de bit, il faut écrire que la charge stockée s'est conservée.

$$C_m V_{DD} + C_c \frac{V_{DD}}{2} = (C_c + C_m) V_b$$

On en déduit :

$$V_b = \frac{C_m V_{DD} + C_c \dfrac{V_{DD}}{2}}{C_c + C_m}$$

Les valeurs de l'exemple conduisent à :

$$V_b = \frac{20 + 60}{120 + 20} = 0{,}58 \text{ V}$$

La tension de colonne n'a varié que de 80 mV par rapport à sa valeur initiale. De même, si la valeur stockée était 0 V, ce qui correspond à l'état « 0 », on obtiendrait :

$$V_b = \frac{60}{120 + 20} = 0{,}42 \text{ V}$$

La tension de colonne a diminué de 80 mV. L'amplificateur en bout de colonne doit donc être sensible à de faibles variations de la tension.

Ce calcul élémentaire montre qu'il est difficile de placer un nombre trop élevé de transistors en parallèle sur une ligne de bit. En effet, la capacité de la colonne augmente avec le nombre de transistors connectés et est très supérieure à la capacité de stockage si le nombre de transistors connectés est important. Dans ce cas, les deux valeurs de la tension lue correspondant aux deux états possibles de la charge stockée sont très voisines et le circuit de lecture est dans l'incapacité de mesurer l'état mémorisé. Ce calcul montre également l'intérêt d'augmenter la capacité C_m autant que possible.

Le schéma de la *figure 9.11* est en général utilisé pour la lecture. Il est simple car formé de deux transistors MOS.

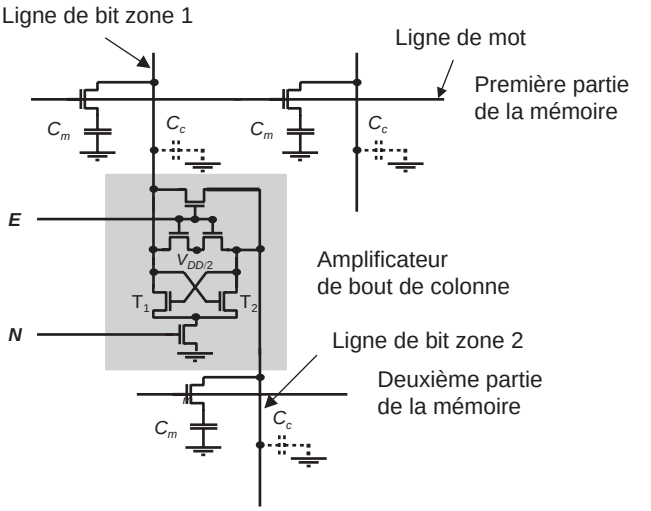


Figure 9.11 – Amplificateurs de colonne.

Pour des raisons liées à la prise en compte des perturbations électriques, la tension de lecture est mesurée non pas par rapport à la masse mais par rapport à la ligne voisine. La cellule lue est dans une partie du circuit appelée zone 1 alors que la ligne voisine de référence est dans une région appelée zone 2. L'amplificateur de lecture est un système simple de type différentiel. Quand les deux lignes sont au même potentiel, le système est équilibré. Quand une des deux lignes est à un potentiel légèrement supérieur à l'autre, la dissymétrie s'amplifie quand on fournit un courant à l'ensemble. Cela peut se faire en portant la ligne de commande *N* à une valeur positive. Finalement, un transistor (T1 ou T2) conduit et l'autre est bloqué.

Un examen des signaux dans le temps peut aider à la compréhension de ce mécanisme de d'écriture et de lecture. Supposons que l'état mémorisé soit « 0 » dans la cellule de stockage de la zone 1, en trait plus foncé sur la *figure 9.12*. Dans un premier temps, le système est rendu symétrique par le signal *E* qui place les deux lignes de bit au même potentiel. Ensuite, la ligne de mot sélectionne la cellule devant être lue. La charge se répartit alors comme il a été vu précédemment entre la capacité C_m et la capacité de la colonne C_c. La dissymétrie est amplifiée quand le signal *N* est appliqué. Le potentiel de la ligne de bit de la zone tend alors vers le potentiel 0. Les transistors à l'état passant sont identifiés par un trait plus foncé sur la figure.

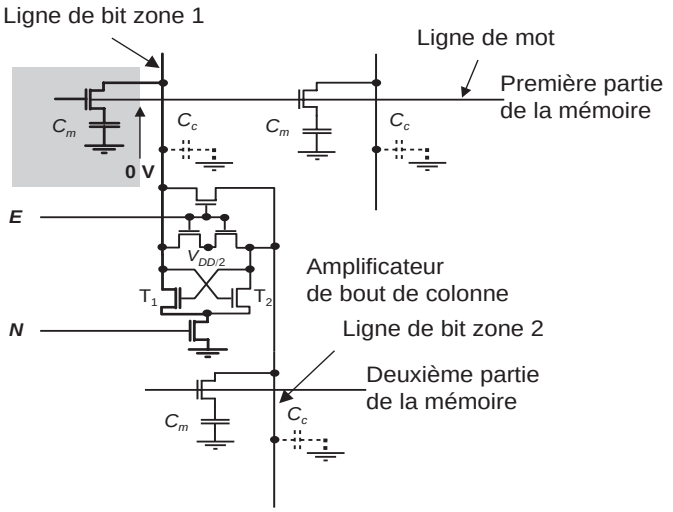


Figure 9.12 – Lecture d'un « 0 ».

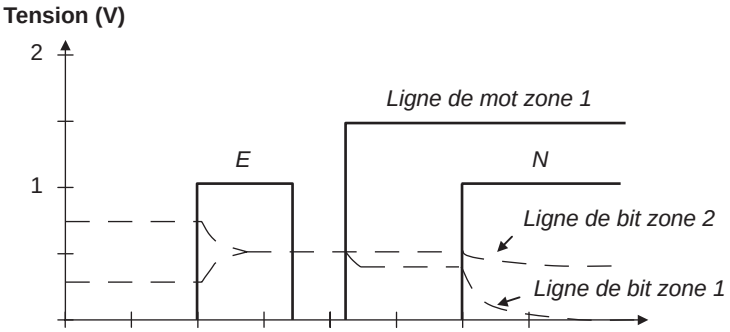


Figure 9.13 – Signaux dans la lecture d'un état « 0 ».

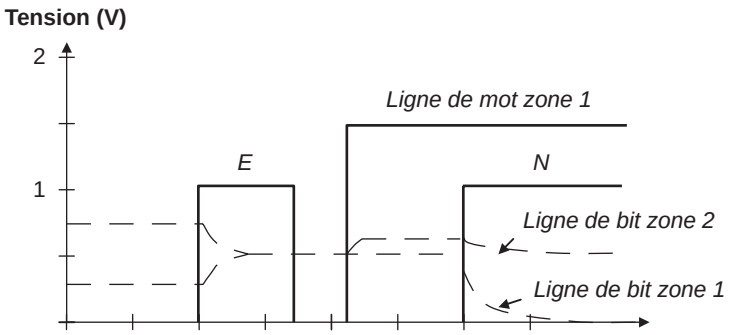


Figure 9.14 – Signaux de lecture d'un état « 1 ».

De la même manière, on peut représenter les signaux quand un signal « 1 » est mémorisé dans la même cellule.

Dans ce fonctionnement, il faut bien comprendre que la ligne de bit 2 n'a pas d'autre fonction que de servir de référence. Pour être tout à fait complet, il faut ajouter que les lignes de bit ne sont pas reliées directement à la tension V_{DD} mais reliées à un circuit en bout de lignes à base de PMOS qui permet de fournir les courants voulus.

L'architecture décrite ci-dessus a l'avantage de la densité mais présente l'inconvénient d'être sensible aux perturbations. En effet, les lignes de bit en entrée de l'amplificateur de mesure sont physiquement éloignées et peuvent donc collecter des signaux de bruit très différents venant du substrat. L'amplificateur peut donc prendre une mauvaise décision.

Il est possible d'imaginer une autre architecture moins dense mais plus résistante aux perturbations. La *figure 9.15* illustre les deux alternatives. La version (a) dite ouverte est la plus dense car une seule ligne de bit passe entre deux colonnes de transistors. La version (b) dite repliée a pour avantage de rapprocher la ligne de bit lue et la ligne de référence. Le bruit substrat induit donc le même signal sur les deux lignes. Cette architecture est moins dense car deux lignes de bit passent entre les colonnes de transistors.

Il faut maintenant étudier l'implantation physique des éléments car la densité d'intégration est une propriété fondamentale de la mémoire.

Les condensateurs de stockage sont réalisés en gravant des trous dans le silicium, en déposant un isolant particulier (une alternance oxyde de silicium nitrure de silicium et oxyde de silicium) puis en réalisant le contact par croissance de silicium polycristallin. Une valeur d'au moins 25 fF est recherchée pour les générations actuelles. L'épaisseur du diélectrique est environ 5 nm. Si l'oxyde était plus mince, l'effet tunnel serait trop important. La profondeur du condensateur peut atteindre 7 microns pour une largeur de 250 nm dans les technologies actuelles. Ce facteur d'aspect est tout à fait surprenant dans la micro-électronique qui est historiquement une technologie planaire.

La surface de la cellule mémoire est alors de 12 λ^2, λ étant la dimension minimale de la technologie. La surface d'une cellule de stockage est de 0,75 μm^2 pour une technologie 250 nm.

Remarquons que les lignes de bit sont en métal alors que les lignes de mot sont en silicium polycristallin car elles servent aussi à fabriquer les grilles des transistors. Le retard de propagation le long d'une ligne de mot peut être calculé comme suit si on considère *n* colonnes (ou *n* lignes de bit) reliées à cette ligne.

$$t_d \cong n\,(C'_{OX} WL + C_p) \cdot n\,R_G$$

On reconnaît dans cette formule approchée :

- C'_{OX} est la capacité d'entrée d'un transistor bloqué, environ 400 aF
- C_p est la capacité parasite associée à une cellule environ 100 aF
- R_G est la résistance de l'élément de ligne polycristallin associée à une cellule soit 4 Ω

Cette valeur élevée de la constante de temps est un facteur limitatif important. Si on considère un ensemble de 512 lignes de bit, on en arrive à environ 500 ps. Ce calcul élémentaire montre que les points mémoire doivent être groupés en blocs de tailles relativement limitées si on veut conserver des temps d'accès courts. Le groupement en blocs de 512 lignes de mots et 512 lignes de bit est typique, ce qui offre une capacité de 2^{18} bits ou 256 kbits.

D'autres techniques permettent de réaliser des condensateurs de stockage. Les condensateurs peuvent par exemple être réalisés au-dessus de la zone de transistors et non en dessous comme dans le cas de la *figure 9.16*. La technologie est plus simple mais les capacités parasites sont plus élevées.

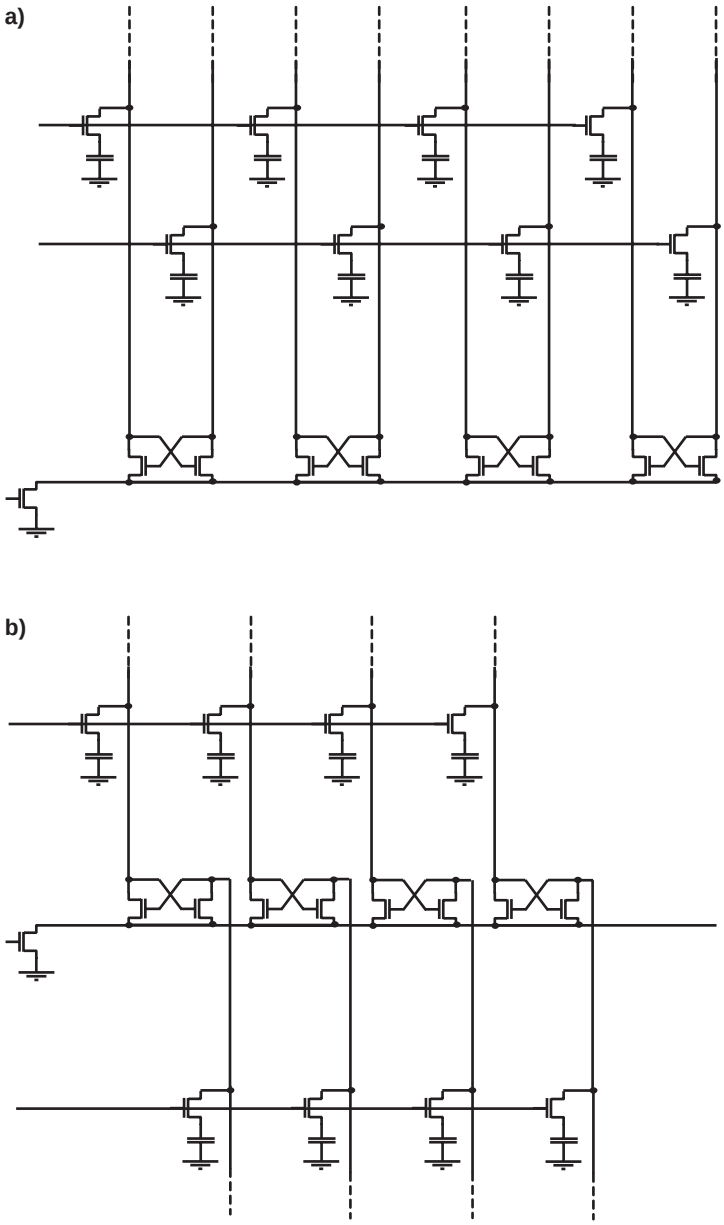


Figure 9.15 – a) Architecture ouverte. b) Architecture repliée.

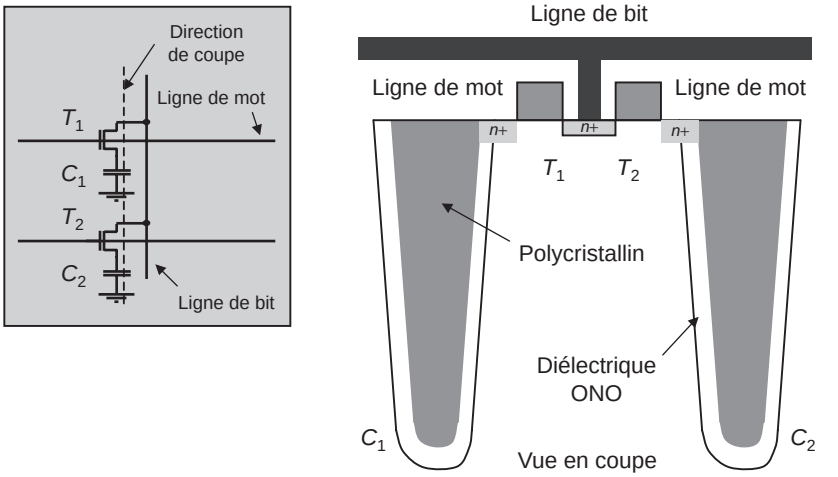


Figure 9.16 – Mémoire DRAM avec des capacités en tranchées.

Des progrès constants sont effectués permettant de réaliser en laboratoire des mémoires de plusieurs gigabits avec un objectif à moyen terme de 16 Gbits dans une technologie 22 nm.

Les considérations précédentes ont amené à envisager l'architecture d'une mémoire de grande capacité non pas comme une simple matrice de cellules à un bit mais comme un ensemble de matrices de stockage de tailles réduites. Les contraintes temporelles et les contraintes de consommation, également liées aux capacités totales des éléments mis en parallèle, expliquent cette évolution architecturale.

Il y a de nombreuses façons d'organiser la mémoire globale. Elle peut par exemple être organisée en mot de *l* bits. Pour une adresse donnée, *l* emplacements sont écrits ou lus simultanément. Des groupements en 4, 8,16 et 32 bits sont utilisés.

Enfin, pour terminer cette introduction aux mémoires DRAM notons la nécessité du rafraîchissement de ces mémoires. Les condensateurs se déchargent. Au bout d'un certain temps, la charge stockée est nulle. Pour éviter cela, il faut lire la mémoire périodiquement et écrire une nouvelle fois les données.

9.2.2 *Les mémoires statiques ou SRAM*

Le principe de stockage de l'information est différent. Une valeur binaire est mémorisée, non pas sous forme d'une charge dans un condensateur, mais sous la forme d'un des deux états possibles d'un bistable électronique. Contrairement à la DRAM, il n'est plus nécessaire de rafraîchir la mémoire et l'information est conservée tant que la tension d'alimentation est maintenue. Les SRAM sont également plus rapides que les DRAM. Par contre, il est nécessaire d'utiliser 6 transistors par point mémoire ce qui explique que la densité de ce type de mémoire est plus faible. Le coût est donc plus élevé. Les SRAM sont en général utilisées comme mémoires de données ou de programme quand les échanges avec le processeur sont fréquents et rapides. La cellule de base est représentée *figure 9.18*.

Dans la phase d'écriture, la cellule est sélectionnée en activant la ligne de mot. Pour écrire « 1 » dans la cellule, la ligne « Bit » est portée à 0 V. Le PMOS de droite conduit et la sortie de l'inverseur de droite est à l'état haut. L'état « 1 » est donc écrit dans la cellule. Si la ligne de bit est désactivée,

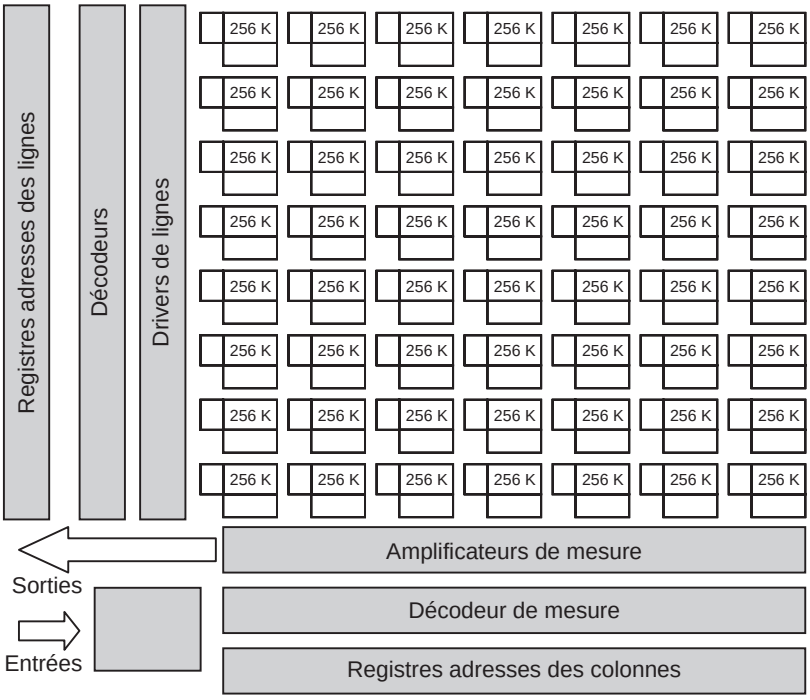


Figure 9.17 – Architecture mémoire globale.

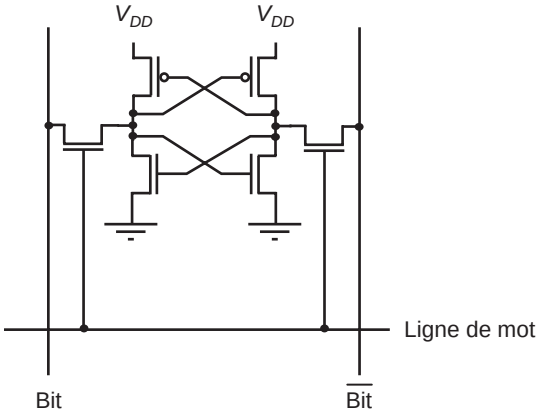


Figure 9.18 – Cellule de base de la SRAM.

ce niveau est maintenu. Pour écrire « 0 », il suffit d'appliquer une tension haute sur la ligne de bit. Les deux cas sont illustrés *figures 9.19* et *9.20*.

Dans la phase de lecture, la ligne de mot est activée. Si la ligne de mot est activée et si le niveau « 1 » est en mémoire, alors la ligne « Bit » est mise à zéro. Si le niveau « 0 » est en mémoire, c'est la ligne

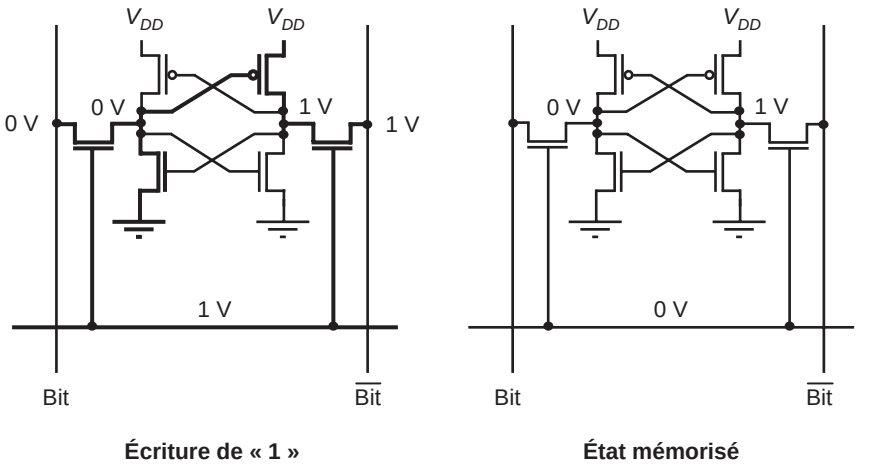


Figure 9.19 – Écriture de « 1 » dans une SRAM.

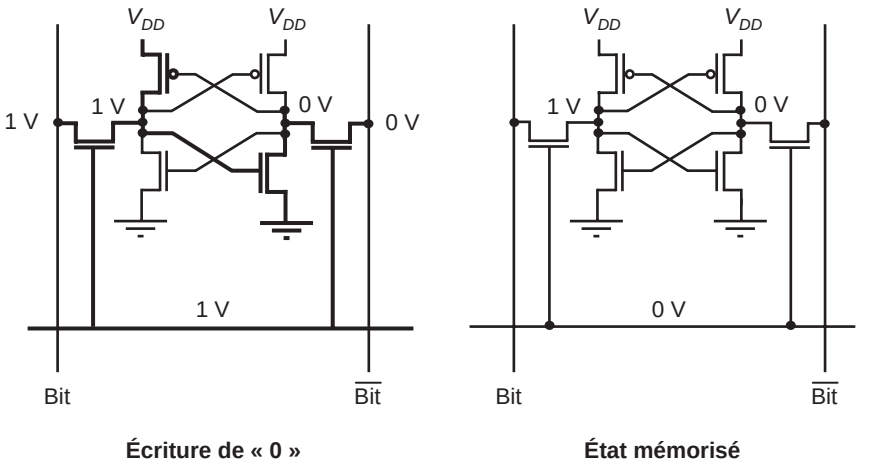


Figure 9.20 – Écriture de « 0 » dans une SRAM.

$\overline{\text{Bit}}$ qui est à zéro. La valeur de l'état mémorisé est faite par mesure de la différence de tension entre la ligne de bit et la ligne de bit complémentaire. Le choix de la taille des transistors est très important pour un fonctionnement correct de la cellule. Les détails peuvent se trouver dans les ouvrages spécialisés.

Contrairement à la DRAM, le signal de lecture a une valeur égale à l'intégralité de la tension d'alimentation. Cela explique que la mesure est plus facile et que finalement le cycle de lecture peut être plus rapide. De nombreuses études ont été faites pour réduire le nombre de transistors dans la cellule par exemple en remplaçant les PMOS par des résistances faites en silicium polycristallin. Les valeurs de résistances sont alors d'une dizaine de MΩ. La consommation statique augmente si bien que cette solution reste peu utilisée.

Les considérations formulées sur les aspects architecturaux pour les DRAM restent applicables aux SRAM. Les SRAM ont beaucoup progressé ces dernières années et les temps de cycle typiques sont de quelques nanosecondes. Rappelons que le temps de cycle est le temps minimum entre deux opérations.

9.2.3 *Les mémoires mortes*

Ce sont des mémoires qui sont programmées de manière définitive. Elles servent à stocker des constantes ou des instructions fixes dans les systèmes numériques. Le schéma de principe est représenté *figure 9.21*.

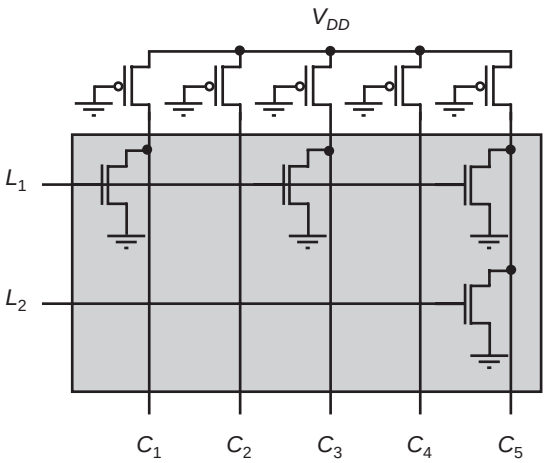


Figure 9.21 – Mémoire morte.

Dans l'exemple de la *figure 9.21*, le mot « 01010 » est mémorisé sur la ligne L_1 et le mot « 11110 » sur la ligne L_2. Le fonctionnement de cette mémoire est simple. Quand la ligne L_1 est activée, les transistors sont conducteurs et les sorties reliées à un transistor sont au potentiel 0 V. Les transistors PMOS reliés au V_{DD} sont des transistors ayant une longueur de canal élevée. Ils délivrent peu de courant. Quand un transistor de la matrice est conducteur, c'est lui qui impose la tension de la colonne. La mémoire est configurée soit en fabriquant les transistors correspondant aux valeurs choisies soit en fabriquant des transistors à chaque point mémoire. Ce transistor est ensuite relié ou non à la colonne.

Notons que, après un arrêt de tension, la mémoire morte restitue les données puisqu'elles sont définies par la position des transistors de la matrice et non pas par la polarisation de ces transistors. On dit alors que ces mémoires sont non volatiles.

9.2.4 *Les mémoires non volatiles*

Les mémoires électroniques ROM sont non volatiles comme il a été vu dans le paragraphe précédent. Cependant, il est d'usage de réserver le terme « non volatile » aux mémoires conservant les données après coupure de l'alimentation mais offrant la possibilité d'effacement et de réinscription des données. Le principe général de ces dispositifs est d'ajouter dans un MOS une grille isolée électriquement des autres parties du MOS. Cette grille est appelée grille flottante. Diverses technologies sont alors possibles :

- La technologie EPROM : les données peuvent être effacées par insolation aux ultraviolets.
- La technologie EEPROM : les données peuvent être effacées par application d'une tension électrique en général de forte valeur.
- La technologie « Flash ». C'est une version plus récente de la technologie EEPROM.
- La technologie MRAM. C'est la technologie la plus récente et la plus prometteuse. Elle est basée sur des principes physiques différents.

Dans un premier temps, décrivons un transistor à grille flottante tel que celui représenté *figure 9.22*.

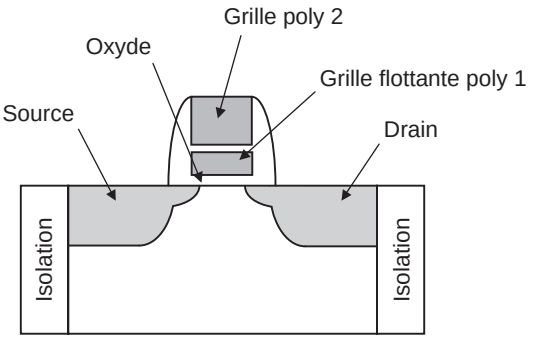


Figure 9.22 – Transistor à grille flottante.

Le fonctionnement de ce transistor s'explique alors sur la *figure 9.23* montant la caractéristique courant-tension de grille dans les deux cas : aucune charge sur la grille flottante et une charge donnée stockée sur la grille flottante.

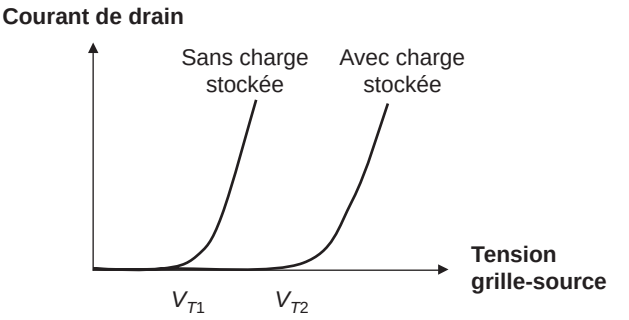


Figure 9.23 – Effet de la grille flottante.

Le chapitre 3 (relation 3.24) nous a appris que la tension de seuil d'un transistor s'exprime par :

$$V_T = \phi_{MS} - \frac{Q_0'}{C_{OX}'} + \phi_B + \frac{1}{C_{OX}'}\sqrt{2eN_A\varepsilon_s}\sqrt{\phi_s}$$

Si on considère un transistor de même taille mais pourvu d'une grille flottante supplémentaire située à une distance du silicium égale à la distance entre la grille flottante et la grille supérieure, tout se passe comme si la capacité par unité de surface était divisée par deux. La tension de seuil augmente donc par rapport à celle d'un MOS sans grille flottante. En pratique, elle augmente de 50 mV environ.

$$V_{T1} = \phi_{MS} - \frac{2\,Q_0'}{C_{OX}'} + \phi_B + \frac{2}{C_{OX}'}\sqrt{2\,e\,N_A\,\varepsilon_s}\sqrt{\phi_B}$$

Si maintenant, on ajoute des charges négatives sur la grille flottante, elles vont avoir un effet équivalent aux charges d'interface Q'_0 mais de signe opposé. Appelons Q' cette charge stockée, la nouvelle tension de seuil est alors :

$$V_{T2} = \phi_{MS} - \frac{2\,Q_0'}{C_{OX}'} + \phi_B + \frac{2}{C_{OX}'}\sqrt{2\,e\,N_A\,\varepsilon_s}\sqrt{\phi_s} + \frac{2|Q'|}{C_{OX}'}$$

Comment créer cette charge sur la grille flottante et comment l'évacuer pour effacer la mémoire ?

Dans le cas des EPROM, la charge sur la grille flottante est créée par injection d'électrons « chauds ». Les électrons chauds sont créés par une tension drain élevée (de l'ordre de 25 V), ce qui leur procure une énergie suffisamment élevée pour qu'ils puissent traverser la barrière de potentiel dans l'oxyde et s'accumuler sur la grille flottante. Il faut que le potentiel de grille (la grille non flottante) soit assez positif pour les attirer.

Notons que pour déclencher ce mécanisme de création de charge sur la grille flottante, il est nécessaire de porter simultanément la grille et le drain à des tensions élevées. Ce mécanisme se limite automatiquement. En effet, l'augmentation de tension de seuil qui en résulte entraîne une diminution du courant de drain et donc du nombre d'électrons supplémentaires piégés sur la grille flottante.

Le mécanisme d'effacement par UV consiste à créer une zone de conduction dans l'oxyde capable d'évacuer la charge stockée. Cette manière peu commode d'effacer la mémoire a été remplacée par un procédé électrique.

Dans le cas des mémoires EEPROM ou Flash, l'épaisseur de l'oxyde est réduit à une dizaine de nm environ. Les opérations de création et d'évacuation de la charge sur la grille flottante peuvent se faire par des moyens purement électriques. L'effet physique mis en œuvre pour créer et évacuer la charge est l'effet tunnel Fowler-Nordheim du nom des inventeurs. Le schéma complet de la cellule mémoire est représenté *figure 9.24* pour les deux opérations d'écriture et d'effacement.

Dans la phase d'écriture, l'application d'une tension de 20 V sur la grille permet l'injection d'électrons sur la grille flottante par effet tunnel. Le drain doit être à la masse pour que cette opération puisse se faire. S'il était à une tension élevée ce serait impossible. Remarquons que ce mécanisme se limite de lui-même car au fur et à mesure que la charge de la grille flottante augmente, le courant tunnel diminue.

Pour effacer la mémoire, la grille est cette fois à 0 V mais les potentiels des puits *n* et *p* sont à 20 V. Les électrons stockés sur la grille flottante peuvent alors être injectés dans le substrat par effet tunnel. Le champ est inversé par rapport au cas précédent. La source et le drain sont flottants. De la même manière que pour l'écriture, le mécanisme se limite de lui-même.

L'organisation des mémoires Flash est différente de celle des DRAM et des SRAM car le but recherché est le maximum de densité. Les transistors de mémorisation sont alors mis en série comme il est montré *figure 9.25,* page 316. Cette architecture est dite de type NAND. L'exemple montre une cellule de 4 bits mais, en pratique, les cellules sont de 8 ou 16 bits.

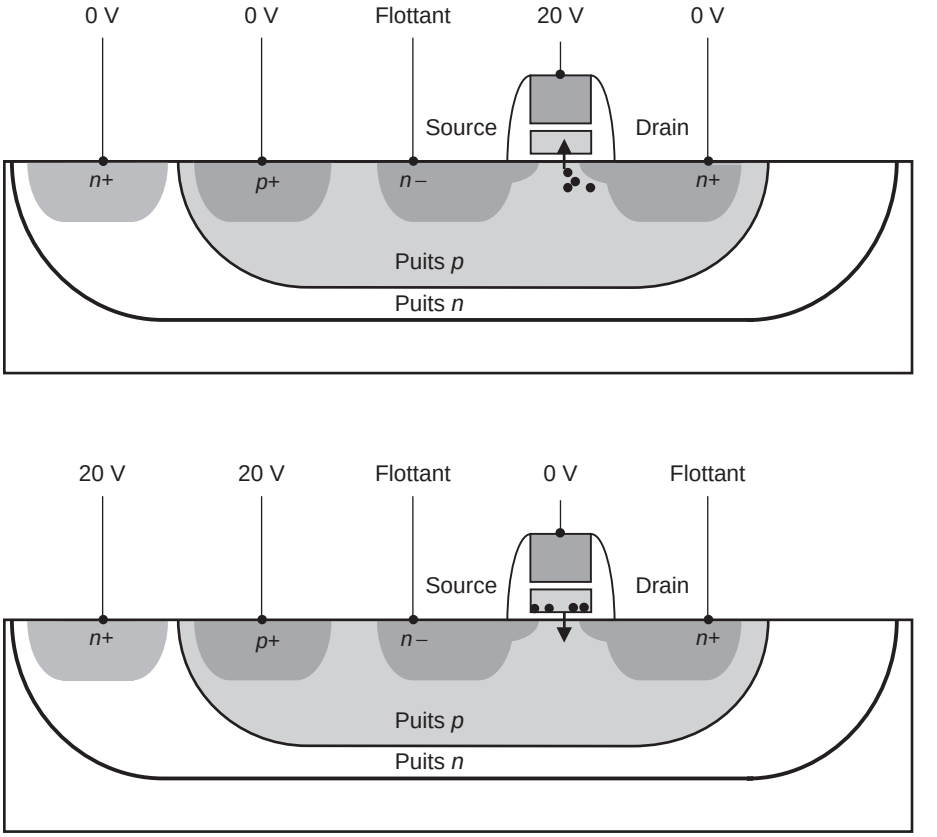


Figure 9.24 – Les deux opérations d'écriture et d'effacement.

La *figure 9.25* représente à la fois le schéma et le *lay-out* d'un ensemble de 4 bits.

Examinons la programmation du bit A_0. Une tension de 20 V est appliquée sur la ligne de sélection haute ainsi que sur la ligne A_0. Des tensions de 5 V sont appliquées en A_1, A_2 et A_3. Une tension de 0 V est appliquée sur la ligne de sélection basse. Les tensions de 5 V sont suffisantes pour rendre les transistors conducteurs mais insuffisantes pour déclencher un effet tunnel. La cellule A_0, et elle seule, est donc écrite.

Pour lire une cellule, il faut se souvenir du fait que la charge accumulée sur la grille flottante modifie la courbe courant-tension du transistor impliqué. Une tension de 0 V est appliquée sur la grille du transistor devant être lu alors que les grilles des autres transistors sont portées à 5 V.

La *figure 9.26* montre la différence de courant à tension de grille nulle selon que la grille flottante est chargée ou non. Il suffit donc de lire cette différence. En pratique, on injecte un courant I_{lec} sur la ligne de bit égal à la moyenne entre les deux valeurs possibles du courant de drain. Si la grille est chargée, le potentiel de la ligne monte à l'état haut. Si la grille n'est pas chargée, le potentiel de la ligne est au niveau bas.

La *figure 9.27* illustre le mécanisme de lecture. Pour bien comprendre ce mécanisme, il faut remarquer que les trois transistors correspondant aux cellules non lues sont passants quel que soit leur état

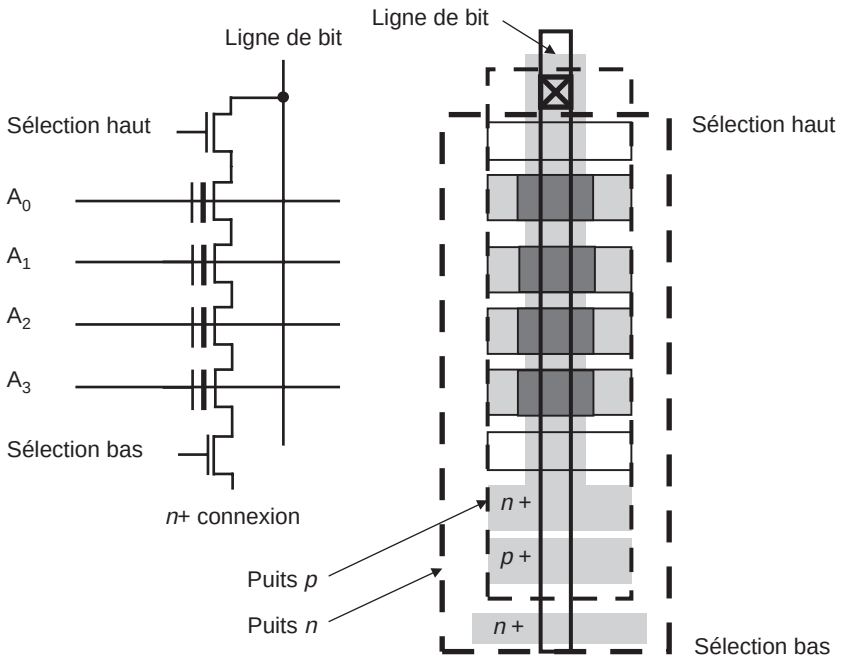


Figure 9.25 – Mémoire Flash NAND 4 bits.

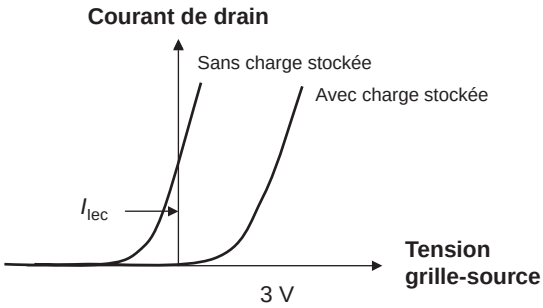


Figure 9.26 – Lecture d'une mémoire Flash.

de charge car une tension de 5 V est appliquée sur leur grille. Le courant de ligne imposé par un générateur extérieur détermine alors la tension de la ligne comme le montrent les caractéristiques de la *figure 9.27*.

Diminuer les tensions de programmation et d'effacement est un objectif permanent pour les fabricants de mémoires Flash. Le problème est cependant difficile car il faut réduire l'épaisseur de l'oxyde pour maintenir un effet tunnel. Il est alors très difficile de maintenir la charge stockée pendant 10 ans comme le veut la norme, quand l'oxyde devient trop mince. En pratique, les tensions de programmation et d'effacement peuvent atteindre des valeurs aussi basses que 8 V à condition d'utiliser les valeurs positives et négatives.

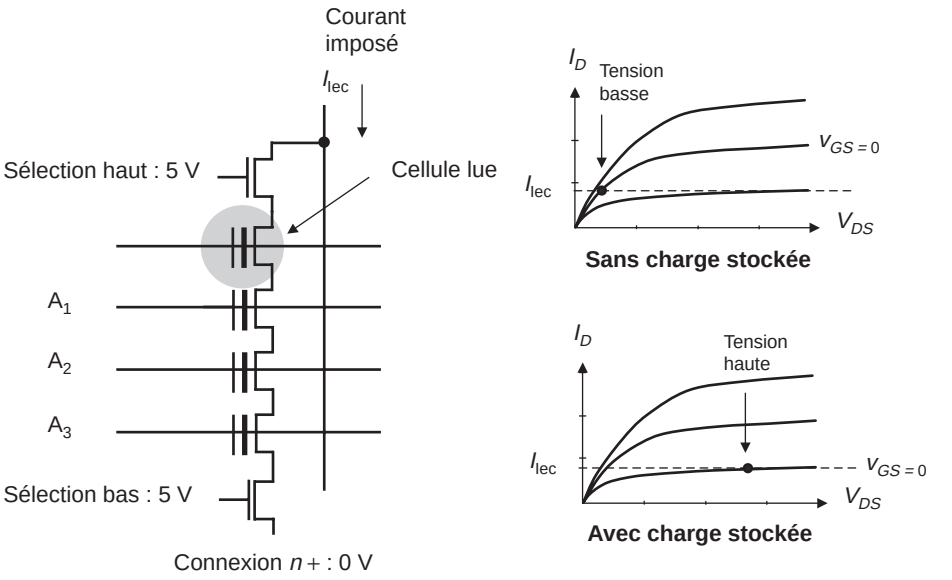


Figure 9.27 – Lecture d'une mémoire Flash.

La technologie décrite est du type NAND. Elle est très compacte et a remplacé dans de nombreux cas une autre technologie appelée NOR. La technologie NOR est historiquement la première à avoir vu le jour et est encore utilisée. Il est possible d'en dire quelques mots à partir de la *figure 9.28*.

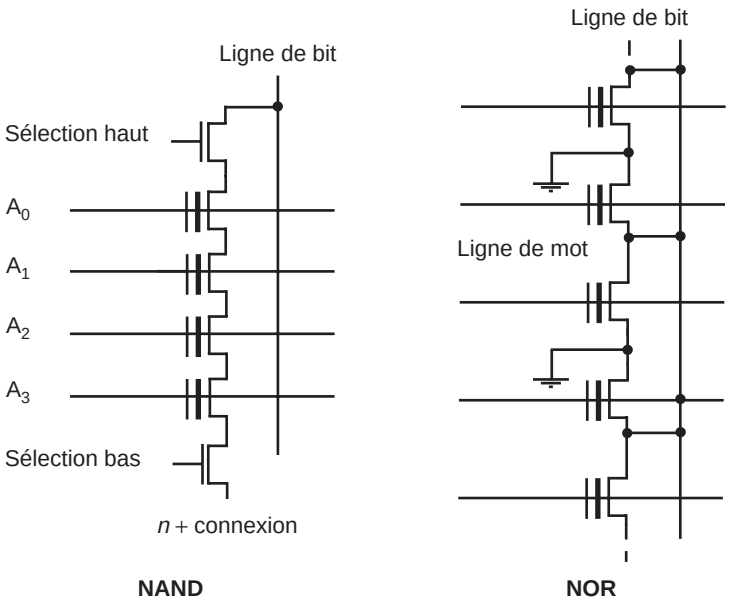


Figure 9.28 – Architectures NAND et NOR.

L'architecture est voisine de celle des DRAM et chaque cellule est indépendante. Il est cependant nécessaire de relier les sources des transistors à la masse pour chaque cellule ce qui conduit à insérer des contacts entre chaque groupe de deux transistors pour les relier à la ligne de bit. Dans la structure NAND, il n'y pas cette contrainte car un certain nombre de transistors (quatre dans l'exemple) sont reliés en série. Les mémoires de type NOR sont moins denses. Elles sont cependant plus rapides.

Réduire les tensions de programmation et d'effacement guide l'évolution de la technologie des mémoires Flash. Il faut en effet implémenter sur la puce des éléments capables de générer des tensions plus élevées que celles utilisées dans la logique. Ces éléments consomment à la fois de la surface de silicium et de la puissance, ce qui est défavorable en terme économique. Cette diminution des tensions conduit nécessairement à réduire les épaisseurs d'oxyde ce qui pose des difficultés en terme de temps de rétention. Les constructeurs se tournent donc vers d'autres principes physiques pour réaliser des mémoires non volatiles comme les MRAM qui utilisent des matériaux magnétiques ou les PCRAM dont le principe est le changement local d'état cristallin.

9.2.5 *Les mémoires adressables par le contenu*

Dans une mémoire classique, une adresse est appliquée en entrée et la donnée correspondante est lue. Dans une mémoire adressable par le contenu, appelée CAM, c'est le contenu qui est placé en entrée de la mémoire. La mémoire doit être capable d'indiquer très rapidement si ce mot est ou non présent. On comprend facilement l'intérêt de ce type de mémoire quand il s'agit de faire une recherche par mot clé dans une base de données. La mémoire est alors dite adressable par le contenu. Elle doit également fournir l'adresse à laquelle le mot recherché est mémorisé. Le schéma de principe de la mémoire CAM est donné *figure 9.29*.

Dans cet exemple simplifié, le mot recherché « 101 » est appliqué en entrée de la mémoire. Ce mot étant stocké dans la deuxième ligne, la ligne de « match » est à l'état haut et indique la présence du

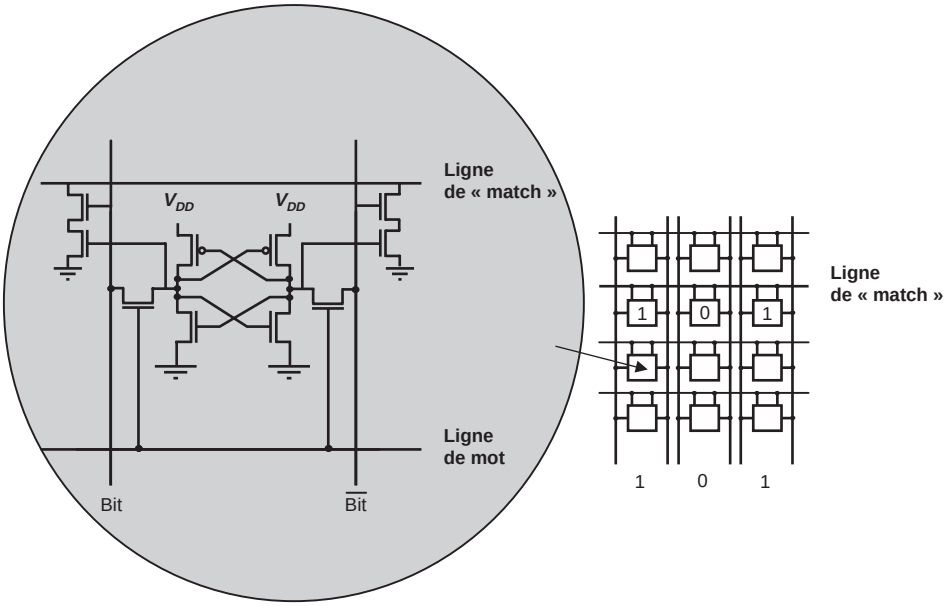


Figure 9.29 – Mémoire adressable par le contenu.

mot recherché ainsi que sa localisation codée par le numéro de la ligne de « match ». Cette ligne est à l'état haut si toutes les paires de transistors reliées à cette ligne sont non conductrices vers la masse. La condition de *matching* dans la cellule doit donc conduire à cet état électrique. Le circuit fonctionne bien de cette manière. Le lecteur pourra le vérifier facilement en considérant les différents cas possibles.

Les mémoires adressables par le contenu sont utiles quand il s'agit de réaliser des mémoires « cache » comme il sera vu dans le paragraphe 9.3.4.

9.3 Les opérateurs de calcul

Les circuits intégrés ne sont pas uniquement des mémoires même si cette fonction représente une part croissante de la surface de silicium produit. Deux autres fonctions sont nécessaires pour intégrer un système complet : prendre des décisions par rapport à des états logiques donnés et faire des calculs binaires. On définit alors les fonctions de calcul et de contrôle.

9.3.1 *Fonctions de contrôle et automates*

Le chapitre précédent a montré que les fonctions logiques pouvaient se réaliser par une combinaison de portes logiques. Quand le problème à résoudre n'est pas posé en ces termes, la première étape est d'en donner une représentation logique.

Prenons l'exemple d'un ascenseur dans un immeuble à trois étages et imaginons la machine logique capable de commander cet ascenseur en fonction de la demande. On note r_1, r_2, r_3 les étages demandés. On note u_1 et u_2 les deux actions de sortie possibles selon que l'ascenseur monte d'un ou deux étages. On note d_1 et d_2 les deux actions de sortie selon que l'ascenseur descende d'un ou deux étages. On note n quand l'ascenseur ne se déplace pas.

Il est alors possible de représenter l'ensemble des cas par un graphe comme celui de la *figure 9.30*. Les états sont les positions de l'ascenseur : premier, deuxième et troisième étage. Les flèches indiquent comment on passe d'un état à un autre et sous quelles conditions.

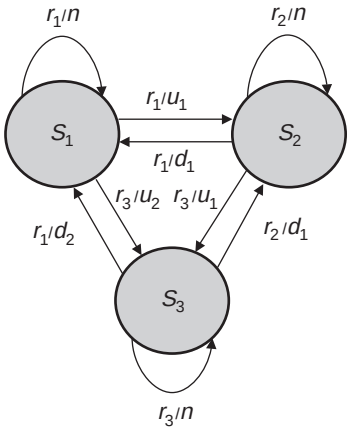


Figure 9.30 – Machine d'états finis.

Le problème est représenté par une machine d'états. Par exemple, si l'ascenseur est au deuxième étage, et si le premier est demandé, alors l'action de sortie est de descendre d'un étage. Comme la donnée de sortie est définie à la fois par l'état dans lequel est le système et par la valeur de la donnée d'entrée (l'étage demandé dans ce cas), on dit que la machine d'états est du type machine de Mealy. Si la donnée de sortie ne dépendait que de l'état, la machine serait appelée machine de Moore.

Un autre exemple permet d'aller jusqu'à la synthèse de la logique. C'est celui des feux de circulation. Une route principale croise un chemin. Le but du système de contrôle est de donner l'avantage à la route principale mais sans toutefois interdire le trafic de la route secondaire. Le feu est en général vert pour la route principale et par voie de conséquence le feu du chemin est rouge. Un détecteur permet d'indiquer la présence éventuelle d'une voiture sur le chemin. Dans ce cas, une temporisation courte se met en route et le feu du chemin passe à l'orange puis au vert pendant un temps plus long. Ensuite, le feu est à nouveau rouge sur le chemin. Le système de feux a quatre états possibles comme il est indiqué *figure 9.31*. Tout autre état est interdit par les règles élémentaires de sécurité.

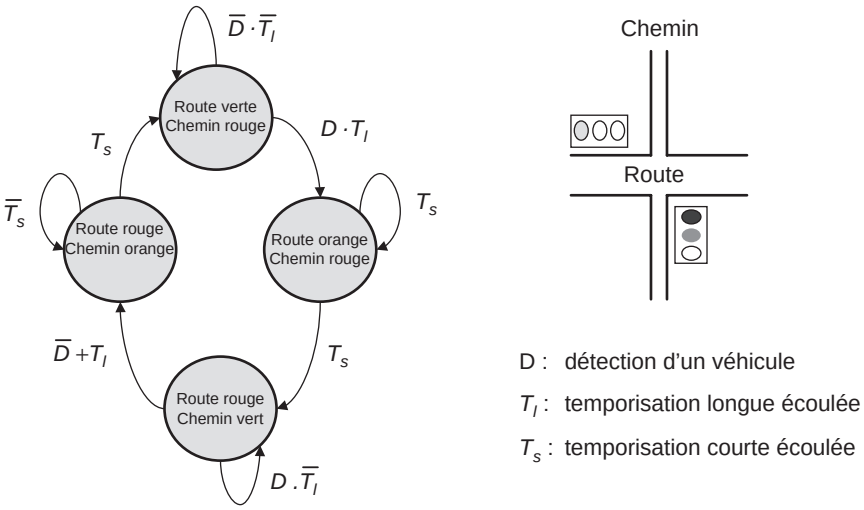


Figure 9.31 – Le problème des feux de croisement.

Le graphique explique de lui-même les différents cas possibles. Par exemple, si le feu de la route principale est rouge et celui du chemin vert, si aucune voiture n'est détectée sur le chemin, le feu du chemin passe à l'orange. Pour favoriser le trafic de la route principale, après un temps T_l, le feu du chemin passe à l'orange. La variable logique T_l est égale à 0 tant que le temps n'est pas écoulé. Les autres cas s'expliquent de la même manière. Le temps T_s est le temps pendant lequel l'orange est maintenu.

Il y a en tout quatre états possibles identifiables avec deux bits. On conçoit donc que les états puissent être mis en mémoire comme les sorties de deux registres.

Le passage de la *figure 9.31* à la *figure 9.32* est immédiat. Il faut noter que les états sont repérés par le couple de variables logiques Q_1, Q_2. Les équations logiques donnent les conditions de passage d'un état à l'autre. On peut ensuite former le tableau permettant de passer d'un état à un autre. Ce tableau est indiqué *figure 9.33*. La notation *X* signifie que l'état est indifférent.

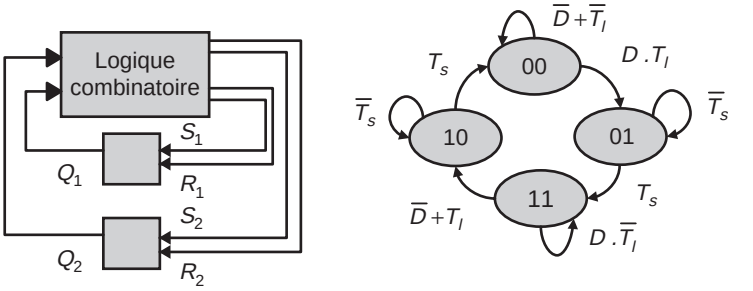


Figure 9.32 – Représentation logique du problème.

Tableau 9.2 – Le tableau des états.

Valeurs					Valeurs suivantes					
Q_1	Q_2	D	T_s	T_l	Q_1	Q_2	S_1	R_1	S_2	R_2
0	0	0	X	0	0	0	0	0	0	0
0	0	0	X	1	0	0	0	0	0	0
0	0	1	X	0	0	0	0	0	0	0
0	0	1	X	1	0	1	0	0	1	0
0	1	X	0	X	0	1	0	0	0	0
0	1	X	1	X	1	1	1	0	0	0
1	1	0	X	0	1	0	0	0	0	1
1	1	0	X	1	1	0	0	0	0	1
1	1	1	X	0	1	1	0	0	0	0
1	1	1	X	1	1	0	0	0	0	1
1	0	X	0	X	1	0	0	0	0	0
1	0	X	1	X	0	0	0	1	0	0

Pour comprendre comment est construit ce tableau, prenons l'état (0,0). Si la variable logique $\overline{D} + \overline{T}_l$ est égale à 1, le système reste dans le même état. Il y a trois cas possibles correspondant aux trois premières lignes du tableau. Quand elle est égale à 0, le système passe dans l'état (0,1). Les états initiaux sont les deux premières colonnes du tableau et les états suivants sont les colonnes six et sept du tableau. Notons que la valeur de la variable T_s peut être quelconque dans ce changement d'état particulier.

On peut immédiatement transformer les changements d'états en valeurs logiques pour les entrées *R* et *S* des bistables. Les registres fonctionnent de la manière suivante. L'entrée *S* à l'état haut positionne la sortie *Q* à l'état haut et l'entrée *R* à l'état haut positionne la sortie à l'état bas. Dans la transition examinée, l'état passe par exemple de (0,0) à (0,1) quand la variable logique $\overline{D} + \overline{T}_l$ est

égale à 0 ce qui implique que S_2 prenne la valeur 1. Les autres lignes se construisent de la même manière par examen des transitions entre états.

L'étape suivante est l'écriture des équations logiques des variables S_1, R_1, S_2, R_2. Elle est immédiate à partir du tableau qui est la table de vérité de ces variables. On écrit :

$$S_1 = \overline{Q_1} Q_2 T_s \qquad R_1 = Q_1 \overline{Q_2} T_s$$

$$S_2 = \overline{Q_1}\,\overline{Q_2} D T_l \qquad R_2 = Q_1 Q_2 (\overline{D}\,\overline{T_l} + \overline{D} T_l + D T)$$

Ces fonctions logiques se synthétisent facilement avec des portes élémentaires comme le montre la *figure 9.33*.

Les blocs de temporisation créent des signaux logiques un certain temps après que l'entrée soit à l'état haut. Ce sont des retards ajustables. Le bloc de temporisation courte correspond au temps pendant lequel le feu reste à l'orange. Le bloc de temporisation longue correspond au temps pendant lequel le feu reste au vert sur le chemin.

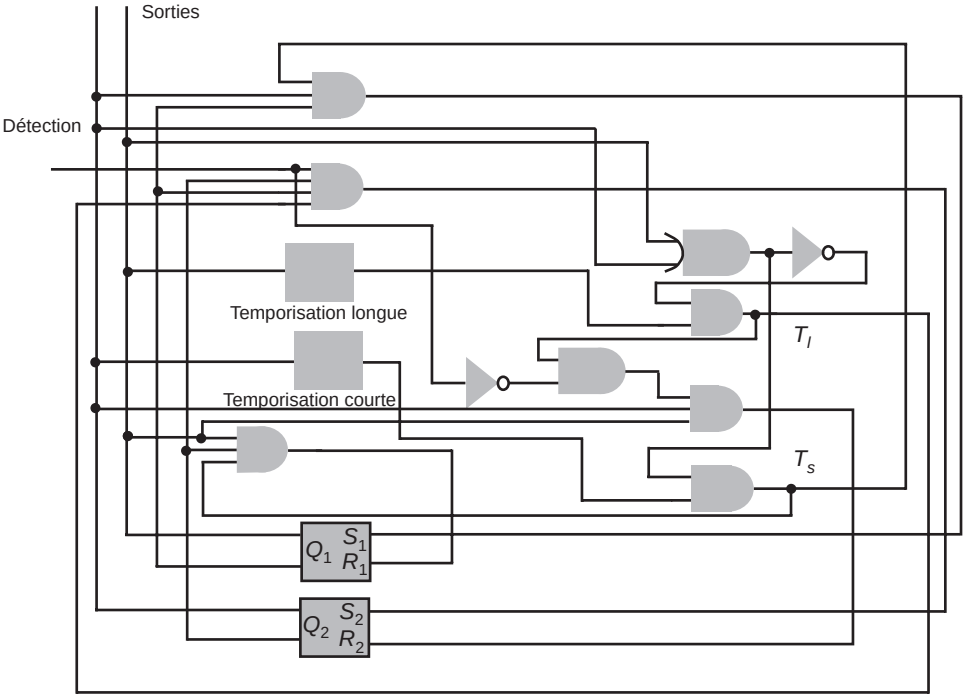


Figure 9.33 – Schéma du système de contrôle des feux.

Cet exemple simple a permis de comprendre les étapes dans la synthèse des automates à états finis. Quelques étapes ne sont pas détaillées comme par exemple la simplification des équations logiques. Dans un cas plus général, on comprend qu'il est possible de réaliser un automate de contrôle à 2^n états à l'aide de n registres et de fonctions combinatoires de base.

On distinguera ensuite les automates asynchrones dans lesquels les états se succèdent uniquement suite aux variations des variables logiques de contrôle (les signaux *S* et *R* de l'exemple précédent) et

les automates synchrones dans lesquels les changements d'états ne peuvent se faire que sur les fronts d'horloge comme il a été vu dans le chapitre 7.

9.3.2 *Les fonctions de calcul et le chemin de données*

C'est la deuxième fonction des architectures électroniques : effectuer des calculs en séquence. Prenons l'exemple du calcul de la fonction $\sqrt{x^2 + y^2}$. Le principe du calcul est donné *figure 9.34*.

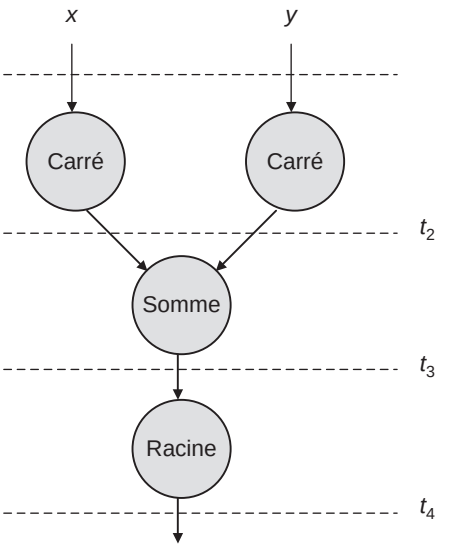


Figure 9.34 – Graphe de chemin de données.

Chaque flèche du graphe est associée à une valeur numérique. L'ensemble des valeurs est un vecteur v noté (v_1, v_2, v_3, v_4, v_5, v_6). Les valeurs se remplissent au fur et à mesure de la progression du calcul. Si par exemple les valeurs 3 et 4 sont entrées, on obtient la séquence suivante :

(3,4,X,X,X,X)
(3,4,9,16,X,X)
(3,4,9,16,25,X)
(3,4,9,16,25,5)

Dans cet exemple simple, nous avons supposé que les valeurs obtenues restaient mémorisées. En pratique, ce n'est pas le cas et les différentes opérations s'enchaînent dans le temps selon un mécanisme appelé « pipe-line ». Chaque étage du pipe-line traite les données au fur et à mesure qu'elles accèdent à l'entrée et délivre le résultat à l'étage suivant. Il est alors nécessaire d'avoir une représentation du système dans l'espace et dans le temps.

On représente le chemin de données comme un ensemble de blocs fonctionnels : registres, mémoires, unités de calcul arithmétique, blocs de multiplexage ou de démultiplexage, bus de données. La *figure 9.35* est un exemple de chemin de données.

La description en chemin de données est également appelée description RTL (*Register Transfer Level*) du système. Elle est intéressante pour la conception car elle est synthétisable c'est-à-dire transfor-

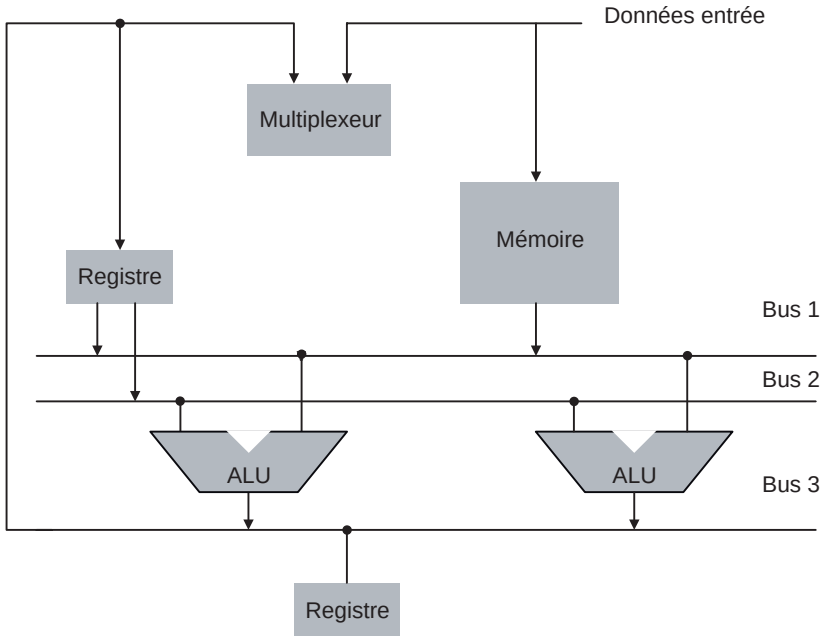


Figure 9.35 – Exemple de chemin de données.

mable en portes logiques. Des logiciels de synthèse permettent d'effectuer cette opération de manière plus ou moins automatique. Les systèmes électroniques comportent en général une partie contrôle et une partie chemin de données si bien que leur architecture est du type de celle indiquée *figure 9.36*.

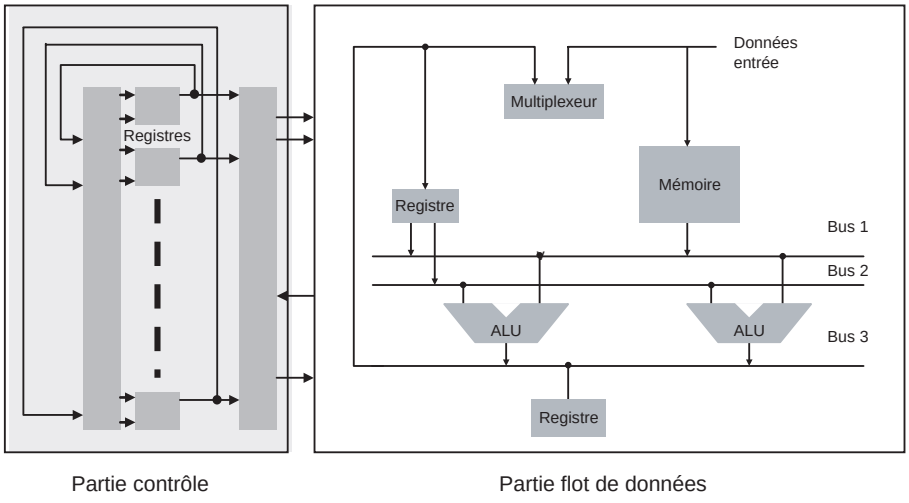


Figure 9.36 – Architecture d'un système électronique.

9.3.3 *Le processeur le plus simple*

Tous les circuits intégrés modernes comportent au moins un processeur. Cette affirmation n'est pas tout à fait juste mais tend à être vérifiée de plus en plus.

L'architecture la plus simple dite Von Neumann est formée de deux blocs interconnectés : un processeur et une mémoire. À l'origine, la mémoire est extérieure au processeur et deux circuits intégrés différents sont associés à ces deux fonctions. Aujourd'hui, cette distinction n'a plus de sens. Les processeurs comportent également une ou plusieurs mémoires, les mémoires cache. Nous verrons plus loin les raisons de cette évolution. Dans l'avenir, on peut penser que les circuits intégrés de type processeur seront en fait principalement des mémoires rapides de plusieurs dizaines de mégaoctets et même de plusieurs centaines de mégaoctets, comportant en plus des unités de calcul.

Revenons au cas le plus simple et examinons un cycle de base. Le principe de fonctionnement du processeur est d'exécuter un programme enregistré dans une partie de la mémoire. Pour introduire les notions utiles nous prendrons l'exemple simple de l'addition binaire de deux nombres. L'architecture du système est détaillée *figure 9.37*.

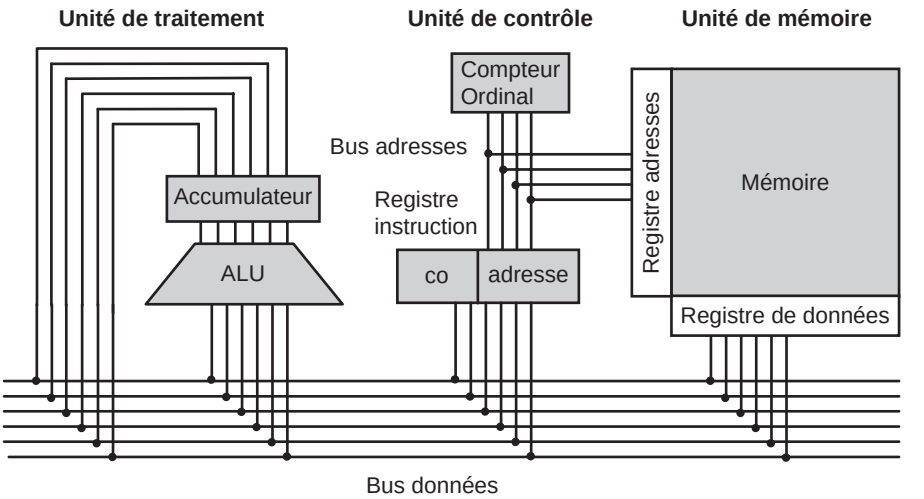


Figure 9.37 – Architecture Von Neumann simplifiée.

Cette architecture simple comporte un bus de communication de 6 bits et un bus d'adresses de 4 bits. La mémoire est organisée en mots de 6 bits adressés par 4 bits. En pratique, les mots sont de 64 bits dans les processeurs actuels. Le bus PCI transporte alternativement adresses et données sur ses 64 lignes électriques.

L'unité arithmétique et logique (ALU) effectue des opérations booléennes logiques ou arithmétiques. Elle est associée à un registre de sortie appelé accumulateur. Elle pourra par exemple faire l'addition binaire entre le contenu de l'accumulateur et la valeur sur le bus d'entrée. Le résultat sera stocké dans l'accumulateur qui prendra donc une nouvelle valeur.

Elle peut également faire le « et » logique entre deux mots binaires, placés respectivement en entrée et dans l'accumulateur. Elle peut faire des décalages et bien d'autres opérations. Toutes ces opérations correspondent à des fonctions logiques électroniques implémentées physiquement dans l'ALU sous

forme de portes. Il faut orienter les signaux d'entrée vers les blocs capables de réaliser les fonctions et synchroniser les opérations. C'est le rôle des signaux de contrôle, non représentés sur le schéma.

Les diverses opérations possibles sont identifiées par un certain nombre de bits, 2 dans l'exemple donné. De plus, on associe à une opération un code plus facile à retenir qu'un nombre binaire, AD pour l'addition par exemple. Une instruction comprendra donc une partie identifiant l'opération (2 bits dans l'exemple) et une partie correspondant à l'adresse de l'opérande (4 bits dans l'exemple).

Notons que c'est l'adresse qui est indiquée dans l'instruction et non pas la donnée. En effet, le principe de la programmation est d'effectuer la même opération pour des valeurs différentes. Il suffit donc d'écrire des valeurs différentes dans la mémoire au même emplacement c'est-à-dire à la même adresse. Cet emplacement est associé au nom de la variable dans le programme. L'opération « $A + B$ » sera par exemple effectuée en associant à la variable A un emplacement mémoire bien défini et à la variable B un autre emplacement. En écrivant des valeurs différentes dans ces deux emplacements, on obtiendra donc les valeurs différentes de la somme des deux nombres.

Prenons l'exemple de l'addition pour expliquer comment interviennent les différents éléments de l'architecture. L'instruction se représente de la manière suivante :

AD	adresse

Elle correspond à un mot binaire du type :

01	1010

Le code opération « 01 » identifie l'addition et « 1010 » identifie l'adresse du nombre de 6 bits dans la mémoire parmi 16 emplacements possibles.

L'opération se déroule en quatre phases :

✦ Phase 1 : recherche de l'instruction

L'adresse de l'instruction est supposée dans le compteur ordinal par exemple « 0100 ». Le compteur ordinal est une fonction que nous n'avons pas encore décrite. C'est un registre qui contient les adresses du programme à effectuer. Le compteur ordinal présentera ces adresses en sortie dans l'ordre prévu par la programmation. L'adresse de l'instruction à effectuer est placée sur le bus d'adresses et est appliquée en entrée adresses de la mémoire. Après un cycle de lecture de la mémoire, le contenu de l'adresse c'est-à-dire l'instruction du programme, est placé sur le bus de données du système. Ce contenu contient le code de l'opération et l'adresse de la donnée à additionner appelée opérande.

✦ Phase 2 : recherche de l'opérande

Le registre d'instruction décode la valeur présente sur le bus de données. Il en extrait le code opération qui servira à générer les signaux dont l'ALU a besoin pour effectuer l'addition. Il extrait également l'adresse de l'opérande et la transfère au registre d'entrée adresses de la mémoire. Le registre de sortie de la mémoire est effacé puis le contenu de l'adresse de l'opérande est lu et placé dans le registre de données de la mémoire.

✦ Phase 3 : réalisation de l'addition

Le contenu du registre de données de la mémoire est placé sur le bus de données. Les signaux de commande correspondant à l'addition sont ensuite envoyés dans l'ALU pour effectuer l'opération.

Le contenu de l'accumulateur est ajouté à la valeur présente sur le bus, qui est la valeur de l'opérande. Le résultat est stocké dans l'accumulateur.

✦ Phase 4 : préparation de l'instruction suivante

Le compteur ordinal est incrémenté de 1. La nouvelle valeur est donc « 0101 » ce qui donne l'adresse de l'instruction suivante dans le programme qui est la suite des instructions rangées dans la mémoire. Cette opération est, par exemple, le transfert du résultat précédent à un emplacement mémoire donné.

La *figure 9.38* montre le déroulement temporel des opérations au rythme des signaux de synchronisation fournis par une horloge.

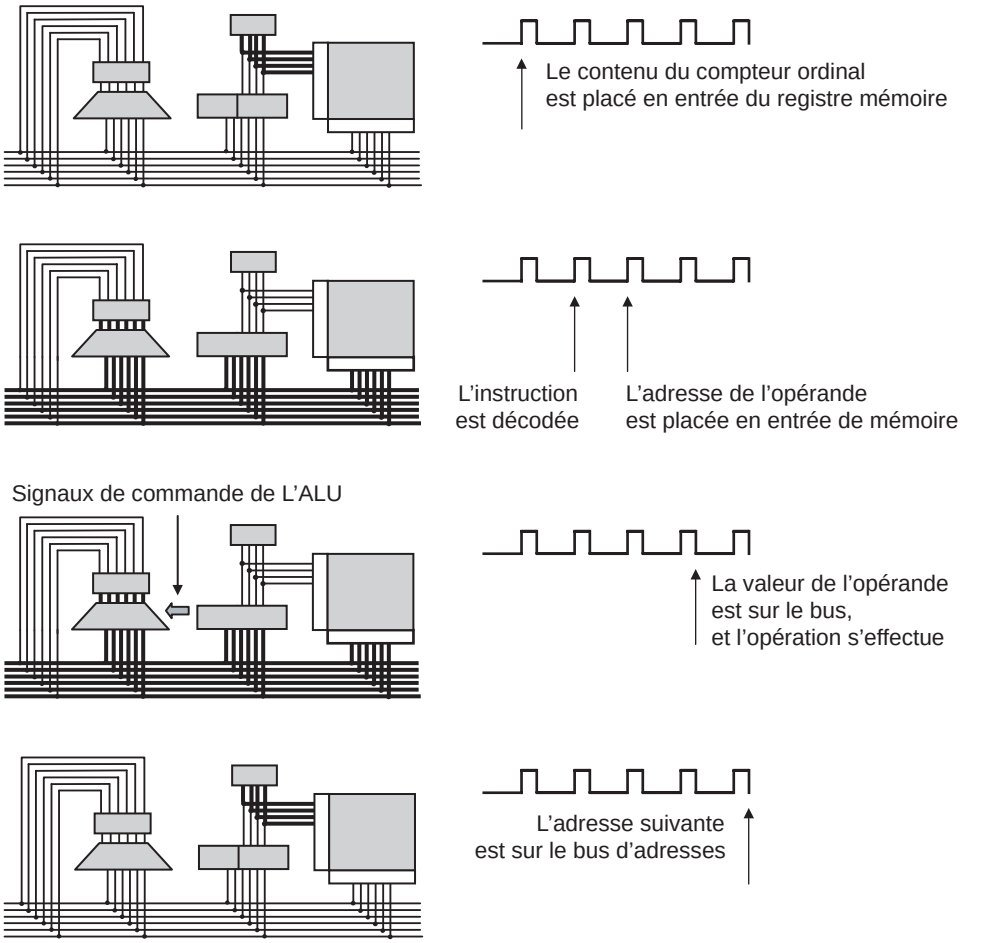


Figure 9.38 – Fonctionnement du processeur Von Neumann.

Une autre instruction très souvent utilisée dans un programme est l'instruction de rangement. Elle se présente sous la forme :

R	adresse

Elle consiste à copier le contenu de l'accumulateur dans la mémoire à l'adresse indiquée dans l'instruction. Les deux premières phases sont identiques à celles de l'addition. La troisième phase consiste à placer le contenu de l'accumulateur sur le bus de données et à lancer une opération d'écriture dans la mémoire. La quatrième phase est la préparation de l'instruction suivante.

Enfin, on peut noter l'instruction de branchement qui consiste à rompre la séquence du programme pour passer directement à une adresse donnée du programme. Elle s'écrit :

B	adresse

L'adresse indiquée est alors l'adresse de l'instruction que l'on désire effectuer dans le programme.

L'exécution d'une instruction conduit à générer différents signaux de commande pour écrire dans les registres, orienter les données dans l'ALU, déclencher les opérations de lecture ou d'écriture des mémoires. Ces signaux doivent être parfaitement définis dans le temps pour que les opérations s'effectuent convenablement. Les données doivent par exemple être établies de manière stable avant les opérations de lecture ou d'écriture. La génération de ces signaux peut s'effectuer de deux manières :

- par une logique séquentielle câblée ;
- par un séquenceur microprogrammé.

Le séquenceur câblé est un automate générant des impulsions et des états à partir d'un graphe du même type que celui que nous avons tracé dans l'exemple des feux de circulation. Le concept de séquenceur microprogrammé a été introduit par Wilkes en 1951. Il est symbolisé dans la *figure 9.39*.

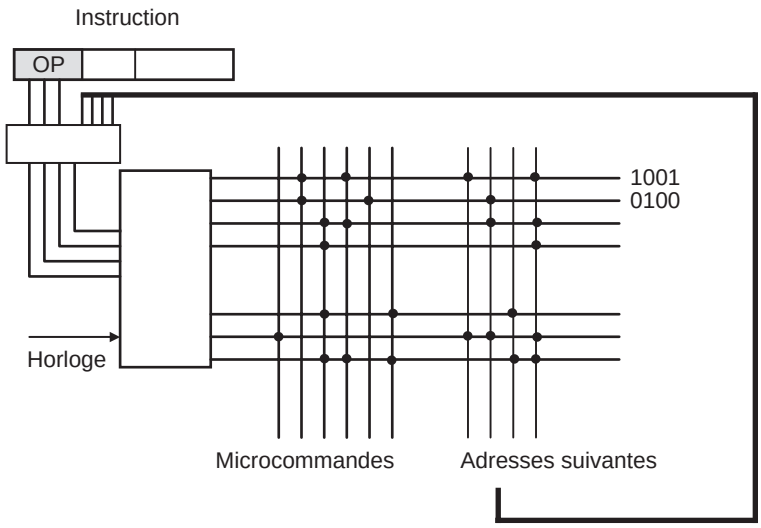


Figure 9.39 – Modèle de Wilkes d'un séquenceur.

Le code opération est considéré comme l'adresse de la première instruction d'un microprogramme associé à cette opération. Ce microprogramme est représenté par une série de lignes générant à la fois l'ensemble des signaux nécessaires mais aussi l'adresse de la micro-instruction suivante. On comprend facilement qu'une mémoire ROM peut être utilisée pour réaliser cette fonction.

9.3.4 *Le processeur le plus complexe*

L'architecture décrite précédemment s'est compliquée avec le temps en intégrant les évolutions suivantes :

✦ **Introduction de registres supplémentaires**

L'architecture devient celle de la *figure 9.40*.

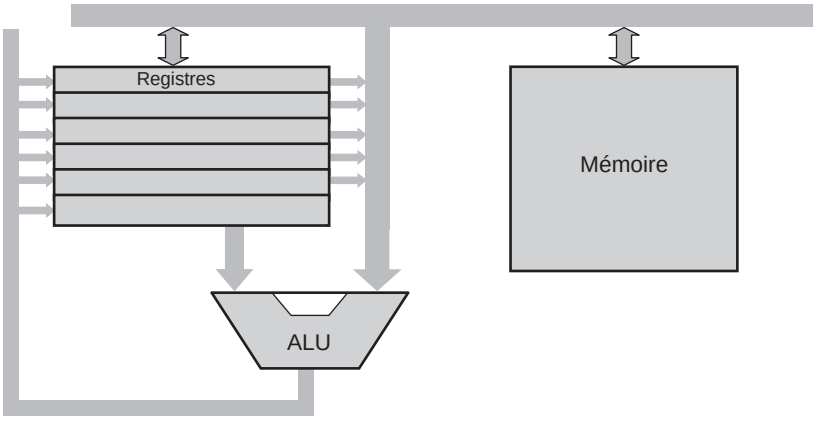


Figure 9.40 – Inclusion de registres.

Les registres jouent donc le rôle de mémoires locales. Il est évidemment très intéressant de maximiser les échanges entre registres dans l'exécution d'un programme et de limiter les échanges avec la mémoire. Ce n'est pas toujours possible. L'organisation des systèmes conduit à la définition d'une hiérarchie de mémoires plus ou moins faciles d'accès dans lesquelles les données et les instructions seront rangées de manière optimale. Le nombre de registres a beaucoup augmenté et les processeurs actuels comprennent au moins 32 registres.

✦ **Introduction des techniques de pipe-line**

Le temps le plus pénalisant dans l'exécution des programmes est le temps relatif à la recherche des instructions dans la mémoire principale. Il est donc intéressant d'anticiper au maximum et de disposer des données au plus vite. En décomposant le cycle de fonctionnement en tâches, on en arrive à la classification suivante :

- Étage 1 : il va chercher les instructions en mémoire et les stocke dans un registre
- Étage 2 : il décode l'instruction
- Étage 3 : il va chercher les opérandes soit en mémoire soit dans un registre
- Étage 4 : il réalise l'opération en conduisant les opérandes à travers le chemin de données
- Étage 5 : il transfère le résultat de l'opération soit dans un registre soit en mémoire principale

On peut alors représenter en fonction du temps le travail effectué par chaque étage, l'instruction est repérée par un indice et le temps est divisé en 9 périodes d'horloge.

Les instructions 1, 2, 3, 4, 5, 6, 7, 8, 9 sont exécutées dans cet ordre.

Tableau 9.3 – Le pipe-line.

E1	1	2	3	4	5	6	7	8	9
E2		1	2	3	4	5	6	7	8
E3			1	2	3	4	5	6	7
E4				1	2	3	4	5	6
E5					1	2	3	4	5

Le processus se poursuit ainsi au fur et à mesure que les instructions sont traitées. Un système sans pipe-line devrait attendre la fin d'un cycle (cinq périodes d'horloge dans l'exemple) avant de démarrer un cycle suivant.

✦ Généralisation de l'architecture RISC

De nombreux débats ont alimenté l'histoire de l'informatique autour de ce sujet. L'évolution des processeurs à partir des années 70 conduisit à des architectures capables de traiter des instructions très complexes qui le plus souvent étaient interprétées par le processeur en utilisant par exemple les techniques de microprogrammation. Chaque instruction complexe était transformée en une série d'opérations plus simples. Ces machines furent appelées CISC.

En 1980, les concepteurs de machine, en opposition aux processeurs CISC, introduisirent des processeurs exécutant des instructions simples, en nombre limité mais très rapidement. La difficulté était alors reportée sur le compilateur qui devait fournir un ensemble optimisé d'instructions machine à partir d'un programme écrit dans un langage de haut niveau. Ces processeurs furent appelés processeurs RISC.

Cette évolution bénéficiait également des continuels progrès effectués sur la vitesse des mémoires vives alors que la vitesse des ROM ne progressait pas aussi vite. Des processeurs hybrides apparurent avec des instructions de type RISC et d'autres de type CISC. Finalement, et sans doute suite aux progrès obtenus sur la vitesse des mémoires vives, la technologie RISC s'est imposée au fil du temps.

✦ Optimisation des techniques de cache

Le principe d'une mémoire cache est simple. Les mots les plus utilisés sont conservés dans le cache et non pas dans la mémoire centrale. Le cache est une mémoire rapide donc de faible taille. Quand le processeur a besoin d'un mot donné, il commence par inspecter le cache. Si la mémoire cache ne contient pas ce mot et uniquement dans ce cas, il va le chercher dans la mémoire centrale.

Notons à ce sujet la nécessité de disposer de mémoires ayant la propriété d'indiquer très rapidement si un contenu donné est ou non présent. Ces mémoires dites adressables par le contenu ont été étudiées paragraphe 9.2.5. L'efficacité de cette technique est basée sur le principe de localité. Les programmes ne génèrent pas des accès aléatoires à la mémoire. Les instructions du programme lui-même sont par exemple situées à des emplacements voisins dans la mémoire. À l'exception des branchements et des appels à des sous-programmes, les instructions proches physiquement de l'instruction en cours seront exécutées avec une probabilité plus forte à court terme. De même, un

calcul effectué sur une matrice de données utilisera de manière plus probable à court terme les données stockées à proximité de la donnée en cours de traitement.

L'évolution actuelle est d'utiliser non pas un cache mais des caches de différentes tailles. Les processeurs actuels comportent trois mémoires cache : une mémoire très rapide de quelques kilo-octets, une mémoire plus importante de quelques centaines de kilo-octets et une mémoire de grande capacité de quelques mégaoctets.

✦ Une évolution à moyen terme vers le parallélisme

Dans un contexte général où les applications sont de plus en plus complexes et traitent des flux de données de plus ne plus importants, le parallélisme apparaît comme une évolution assez naturelle des architectures Von Neumann. Ce n'est cependant qu'à partir de 2005 que les premiers circuits commerciaux à usage général et multiprocesseurs apparaissent sur le marché.

Citons le processeur « cell » développé par IBM, SONY et TOSHIBA construit à partir de 8 processeurs élémentaires. Les avantages sont immédiats : possibilité d'effectuer des tâches en parallèle et donc gain en vitesse, distances réduites entre processeur et mémoires et donc gain en vitesse et en consommation. Cette évolution n'est pourtant pas sans soulever des difficultés de taille si on pense au portage des logiciels existants.

Le principe de compatibilité ascendante s'est imposé dans l'informatique. Il stipule que toute nouvelle version de logiciel ou de matériel doit être compatible avec la version précédente. En pratique, un nouveau processeur doit être capable de supporter des applications développées pour un processeur de génération antérieure. Les architectures Von Neumann sont bien adaptées à cette contrainte puisque tous les temps caractéristiques de l'exécution d'un programme sont divisés par le même facteur au fur et à mesure que la technologie progresse : temps d'exécution d'une opération, temps de lecture et d'écriture des mémoires. Pour les architectures parallèles cette condition est plus difficile à remplir. De plus, porter une application développée pour un processeur Von Neumann sur une architecture parallèle demande un effort important pour réellement bénéficier des avantages du parallélisme. La solution de ce problème majeur est vraisemblablement dans l'évolution des méthodologies de conception des logiciels et particulièrement des logiciels embarqués sur les puces.

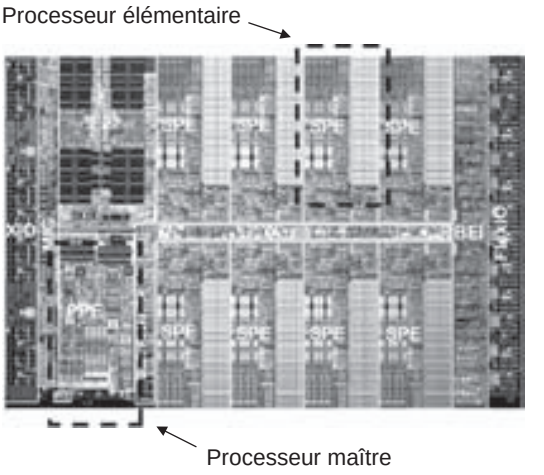


Figure 9.41 – Le processeur « cell ».

9.4 Évolution des circuits intégrés complexes

On appelle système sur puce le dispositif résultant de la miniaturisation d'un ensemble électronique précédemment réalisé sous forme d'une carte. En première analyse, ce n'est qu'une question de taille et de technologie d'intégration. En pratique, la conception et le test des systèmes sur puce posent des problèmes spécifiques assez difficiles. Le premier système que nous allons étudier est une mémoire. Cet exemple peut sembler surprenant car la mémoire est une structure régulière ne présentant pas en première analyse un fort potentiel d'évolution dans son architecture. Cette idée doit être maintenant remise en cause dans un certain nombre de cas. La troisième partie de ce paragraphe sera consacrée à une description plus générale de l'évolution des architectures des circuits intégrés.

9.4.1 *Évolution des technologies mémoires actuelles*

Avant d'étudier de nouvelles architectures de mémoires, rappelons les propriétés principales des grandes familles de mémoires dans la *figure 9.42*.

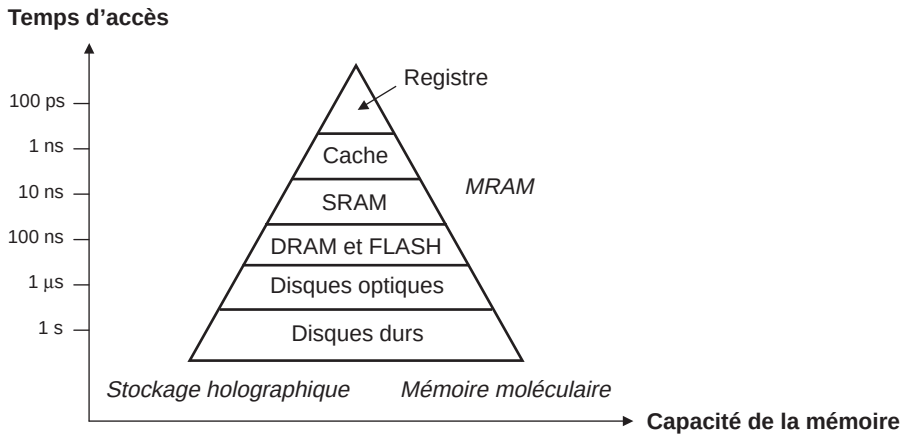


Figure 9.42 – Les mémoires.

Les tailles maximales sont de 100 Go pour les disques durs. Des capacités de 1 To sont attendues dans l'avenir, Les disques optiques de type DVD proposent des capacités de stockage de 10 Go. Les DRAM offrent une capacité de stockage de 4 Gbits et visent 100 Gbits dans des technologies très avancées. Les SRAM offrent une capacité de 256 Mbits et visent à long terme 10 Gbits. Les mémoires Flash atteignent en 2005 des capacités de 4 Gbits et, à terme, visent des capacités de 100 Gbits. Les technologies optiques et magnétiques ne sont pas en reste en espérant atteindre des capacités de 100 Go. Les diverses technologies progressent de manière parallèle et les écarts actuels en terme de capacité et de temps d'accès seront conservés de manière relative.

Ce qui n'est pas conservé, par contre, c'est la différence de rapidité entre les mémoires et les processeurs. Cet écart se creuse au cours du temps comme le montre la *figure 9.43* car la vitesse de fonctionnement des processeurs augmente bien plus rapidement que la vitesse des mémoires. Ce phénomène est appelé « memory gap ».

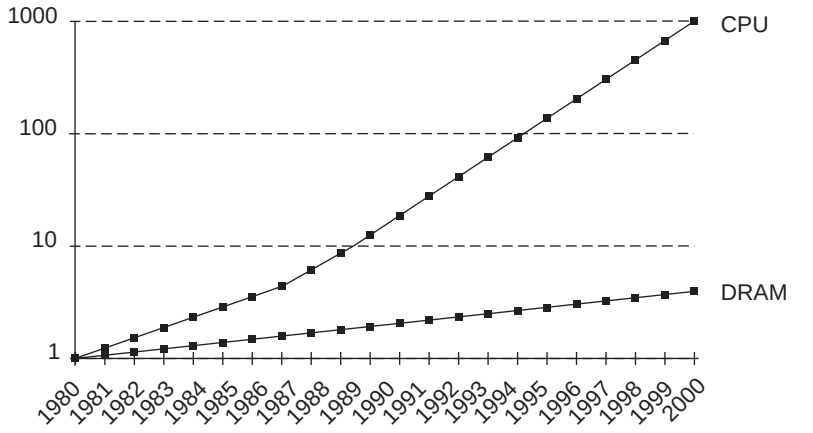


Figure 9.43 – Évolution des processeurs et des mémoires.

Les raisons à cela ne sont pas toujours simples. Elles sont de deux ordres :

Le temps consommé pour adresser la mémoire (décoder l'adresse, mettre la donnée sur le bus et la propager) est en fait lié à la dimension de l'espace d'adressage et à la distance physique entre mémoire et processeur. Il n'est donc pas fondamentalement réduit quand le transistor est de plus faible dimension.

Les critères d'optimisation des processeurs et des mémoires sont différents. La technologie la mieux adaptée aux processeurs réduit la taille des transistors et multiplie les niveaux d'interconnexion pour augmenter la vitesse. La technologie la mieux adaptée aux DRAM utilise au maximum les liaisons en silicium polycristallin et cherche avant tout à réduire la surface de la cellule de stockage.

Des améliorations dans l'architecture des DRAM permettent de gagner en vitesse. Il est par exemple possible de maintenir constante l'adresse de la ligne et de faire varier uniquement les adresses des colonnes. On réalise alors un adressage par page de la mémoire. Il est également possible d'associer à la mémoire un contrôleur qui gère le trafic et les rafraîchissements de manière optimale. On atteint alors des débits de lecture et d'écriture de l'ordre de l'octet par nanoseconde.

Des modifications plus fondamentales conduisent au concept de mémoire intelligente introduite par l'université de Berkeley en 1997. Dans ce cas, les unités de calcul et de mémoires sont démultipliées en éléments de tailles réduites et associés en minimisant les interconnexions. Ce concept est en fait proche du concept de processeur parallèle. La *figure 9.44* illustre la classification actuelle des mémoires.

Les principales performances des mémoires de la *figure 9.44* et de quelques technologies nouvelles sont représentées au *tableau 9.4*.

Dans le *tableau 9.5* figurent de nouvelles technologies qui seront évoquées à la fin de ce paragraphe et dans le chapitre 12. Avant d'évoquer leurs potentialités, il est nécessaire de donner quelques éléments sur les prévisions de performances attendues par les fondeurs pour les technologies classiques : DRAM, SRAM et Flash.

Le *tableau 9.5* illustre la loi de Moore pour les mémoires. Il confirme que les technologies avancent au même rythme, consolide l'avantage des Flash de type NAND pour le stockage à haute densité et

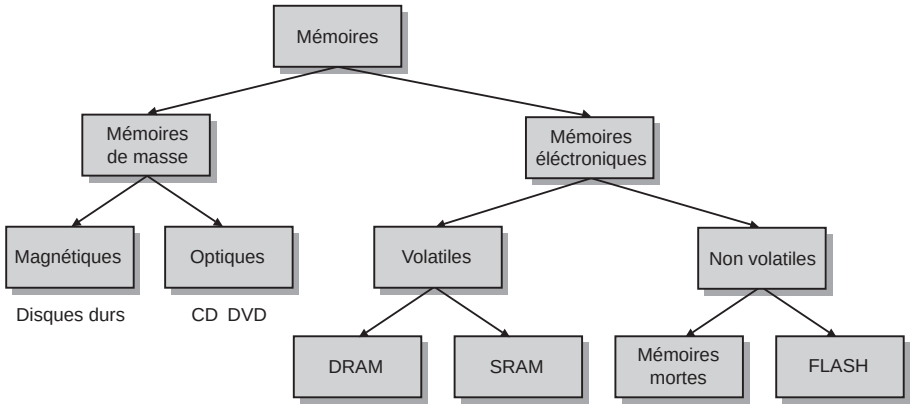


Figure 9.44 – Les technologies mémoire en 2006.

Tableau 9.4 – Performances des technologies mémoire.

	Capacité	Densité (Gbits/cm²)	Temps d'accès	Temps de rétention
Disque dur	800 Gbits	3,5	7 ms	10 ans
CD	5,6 Gbits	0,05	80 ms	10 ans
DVD	37,6 Gbits	0,4	80 ms	10 ans
Micropointes holographiques	1 Tbit	60-200	?	?
DRAM	4 Gbits	1	10 ns	70 ms
SRAM	20 Mbits	0,1	quelques ns	0
Flash	4 Gbits	0,5	quelques µs	10 ans
FRAM	64 Mbits	0,2	30 ns	10 ans
MRAM	4 Mbits	0,5	20 ns	10 ans
PCRAM		0,5	10 ns	10 ans
Moléculaire	1 Tbit	1 000	20 ns	?

annonce les évolutions technologiques majeures que sont l'utilisation de nouveaux oxydes et le remplacement du transistor MOS classique par le FINFET ou un dispositif équivalent (voir chapitre 10).

Pour terminer ce paragraphe, il est possible de passer en revue les technologies candidates pour fabriquer les mémoires du futur. De nombreux matériaux sont en développement et de multiples principes sont envisagés. Quelques-uns seulement seront retenus.

Tableau 9.5 – Évolution des mémoires classiques.

	Année	2004	2006	2008	2010	2012	2014
	Nœud (nm)	**80**	**65**	**50**	**40**	**30**	**25**
DRAM	Densité (Gbits)	1	2	4	8	16	16
	Diélectrique	MIS	MIM	MIM	MIM	MIM	NIM
	Architecture	MOS	MOS	FinFET	FinFET	FinFET	FinFET
SRAM	Densité (Mbits)	256	512	1 024	2 048	4 096	4 096
	Diélectrique						
	Architecture	MOS	MOS	FinFET	FinFET	FinFET	FinFET
Flash NAND	Densité (Gbits)	2	4	8	16	32	32
	Diélectrique	ONO	ONO	ONO	High k	High k	High k
	Architecture	Grille flottante	Grille flottante	Grille flottante	FinFET	FinFET	FinFET
Flash NOR	Densité (Mbits)	256	512	512	1 024	4 096	4 096
	Diélectrique	ONO	ONO	High k	High k	High k	High k
	Architecture	Grille flottante	Grille flottante	Grille flottante	FinFET	FinFET	FinFET

9.4.2 *Les technologies mémoires alternatives*

Les évolutions se font dans deux directions :

- augmenter la densité des mémoires de masse ;
- améliorer les mémoires non volatiles.

Pour améliorer de manière significative la densité des mémoires de masse, deux techniques nouvelles sont étudiées : le stockage holographique et l'utilisation de nanotêtes d'écriture.

Le principe du stockage holographique est de créer dans un matériau des changements locaux de propriétés optiques en faisant interférer un laser de référence et un laser dont le faisceau est modulé par les données à mémoriser. Les figures d'interférence sont créées dans un bloc de matière. De manière symétrique, à la lecture, le laser de référence sera capable en utilisant le bloc gravé de créer avec un laser de lecture un signal modulé représentant l'information stockée. L'avantage de cette méthode est qu'elle permet un véritable stockage en volume de l'information. Des densités aussi élevées que 10^{15} bits par cm^3 sont attendues.

Le principe du stockage par nanotêtes d'écriture est basé sur le principe du microscope à force atomique. Une matrice de têtes d'écriture est amenée à proximité d'un matériau de stockage. Le courant tunnel qui circule quand la distance est très proche chauffe localement le matériau et change sa structure cristallographique. Il en résulte une différence de résistance détectable à la lecture. La densité de stockage espérée est également importante.

Ce même principe est repris dans les mémoires de type PCRAM qui utilisent un matériau sensible à un échauffement local. Dans une PCRAM, mémoire à accès aléatoire, cet échauffement est obtenu électriquement par le courant fourni par un transistor.

Notons également les MRAM qui sont des mémoires non volatiles basées sur un matériau magnétique dont on change l'aimantation dans une cellule. Ces dispositifs peuvent fonctionner avec des tensions plus faibles que les mémoires Flash mais n'ont pas encore atteint une maturité industrielle.

9.4.3 *Les systèmes sur puce*

Les systèmes sur puce sont présentés comme une évolution majeure des circuits intégrés. Ils sont en fait l'intégration de diverses fonctions électriques réalisées précédemment sur des puces différentes. On intègre ainsi processeurs, processeur de signal, mémoires et fonctions analogiques sur la même puce. Les difficultés de l'opération sont de deux ordres.

- La technologie choisie n'est pas nécessairement optimale pour toutes les fonctions. La technologie bipolaire est par exemple optimale pour la RF et la technologie CMOS est optimale pour réaliser des processeurs.
- Il convient de vérifier très soigneusement le fonctionnement de l'ensemble en particulier en simulant les interférences entre blocs logiques et analogiques.

Cette solution est parfois abandonnée et remplacée par une technologie jugée plus sûre. C'est celle du SIP (*System On Pakage*). Dans ce cas, différentes puces, en général réalisées dans des technologies différentes, sont assemblées sur un support commun, par exemple en céramique.

Le coût de fabrication est cependant plus élevé que celui de la solution SOC équivalente. Il est également possible de concevoir des puces capables de supporter des éléments rapportés et connectés soit par des billes conductrices soit par des *bondings*.

Chapitre 10

Limites à la réduction de taille du transistor et nouveaux composants

La réduction de la taille du transistor a pour but d'intégrer de plus en plus de fonctions sur une surface donnée et donc de réduire le coût. Cette réduction de taille s'accompagne naturellement d'une augmentation de la vitesse de fonctionnement, ce qui est favorable. Cependant, d'autres phénomènes moins favorables apparaissent : l'augmentation du courant sous le seuil et l'augmentation de la dispersion de la tension de seuil et du temps de propagation. La question se pose donc : quelle est la plus petite taille possible du transistor dans les années futures ? La deuxième question qui en découle est la suivante : quel est le dispositif qui pourra un jour remplacer le transistor MOS ? La réponse à la première question est l'objet de ce chapitre. La réponse à la deuxième question sera traitée dans les chapitres 11 et 12 de cet ouvrage.

10.1 Les règles de réduction de taille

La règle de conception qui a été appliquée dans les premiers temps de la micro-électronique est de maintenir le champ électrique constant dans le canal du transistor au fur et à mesure que la taille diminue.

Les principales grandeurs physiques intervenant dans le fonctionnement du transistor sont représentées *figure 10.1*. Ce sont : la longueur physique du canal (L), l'épaisseur de l'oxyde de grille (d_{OX}), le dopage du caisson (N_A), la profondeur de la zone de charge d'espace (y_b) des jonctions source-caisson et drain-caisson, la largeur (l) des contacts de source et de drain. Il est important de noter que la largeur (W) du transistor est considérée comme une variable indépendante et peut être choisie uniquement en fonction des impératifs du design, principalement en fonction de la valeur du courant de commutation.

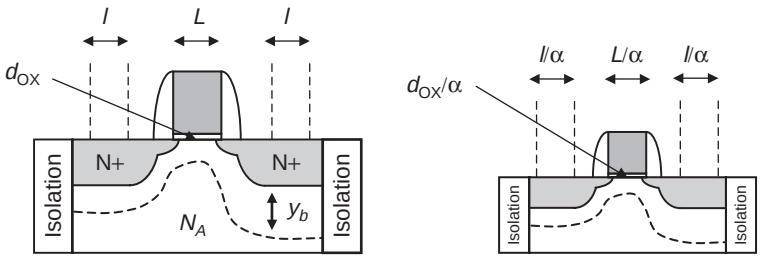


Figure 10.1 – La réduction de taille du transistor.

Il est assez naturel de penser que si la dimension du canal diminue d'un facteur α, la tension d'alimentation doit diminuer du même facteur pour garder le champ constant dans le canal. Le facteur α de réduction est égal à 1,37 quand on passe d'une génération à une autre, soit tous les 18 mois. Cette règle est appelée la loi de Moore. Elle n'a rien d'une loi car elle ne repose sur aucun principe physique. C'est une sorte de compromis entre la résolution des problèmes techniques liés à la réduction de taille et la nécessité économique de proposer des produits plus performants ou moins chers.

Si on admet que les contraintes de design imposent de réduire la largeur W du transistor de la même valeur pour conserver le rapport W/L constant, on en arrive donc à un dispositif dont la surface a été réduite d'un facteur α^2. Remarquons que les dimensions des contacts de source et de drain et donc l'encombrement de l'interconnexion doivent être réduits du même facteur. On peut donc réduire la surface d'un circuit d'un facteur α^2 en changeant de génération ce qui correspond à un facteur 2 environ.

Il faut maintenant expliquer pourquoi il est nécessaire de réduire l'épaisseur de l'oxyde (d_{OX}), ce qui pose par ailleurs des difficultés technologiques considérables. Cette dimension influe directement sur la capacité par unité de surface C'_{OX} et sur la valeur du courant traversant le transistor en régime de saturation.

$$I_{D\,\mathrm{sat}} = W \mu_n C'_{OX} [V_{GS} - V_T] E_C \tag{10.1}$$

Cette relation tirée du chapitre 4 exprime le courant d'un transistor saturé en régime de canal court. La tension de saturation est négligée dans cette formule.

Les tensions diminuant d'un facteur α ainsi que la dimension W, le courant diminuerait en α^2 si l'épaisseur de l'oxyde était constante. Si on admet que les contraintes du design imposent que la diminution du courant de conduction soit au maximum d'un facteur α, il est donc nécessaire d'augmenter C'_{OX} d'un facteur α. Un argument possible pour justifier cette règle est de considérer la charge d'une capacité par le courant d'un transistor. La capacité diminuant d'un facteur α à cause de la réduction de taille, on peut tolérer que le courant de charge diminue du même facteur. Dans ce cas, la vitesse de commutation n'est pas réduite. En résumé, il est nécessaire d'augmenter la capacité C'_{OX} d'un facteur α, ce qui conduit à diminuer l'épaisseur de l'oxyde du même facteur.

Expliquons l'évolution du dopage N_A du caisson du transistor. La profondeur de la zone de charge d'espace est donnée par la relation suivante tirée du chapitre 3.

$$y_b = \sqrt{\frac{2\varepsilon_s}{eN_A}\Delta V} \tag{10.2}$$

La variation de potentiel est la différence entre le potentiel en surface et la partie profonde du silicium. Pour réduire cette dimension d'un facteur α et maintenir les mêmes conditions de fonctionnement du point de vue de l'électrostatique, il faut donc augmenter le dopage d'un facteur α, puisque la variation de tension diminue d'un facteur α.

En pratique, la règle consistant à maintenir le champ constant dans le canal n'est plus strictement appliquée car elle conduirait à réduire les tensions en dessous d'un seuil acceptable pour le bon fonctionnement des circuits. Une autre raison pour modifier cette règle est l'augmentation du courant sous le seuil quand la tension de seuil diminue trop. Cet effet que nous avons déjà noté apparaît dans la relation suivante tirée du chapitre 4.

$$I_D = \frac{W}{L} k \exp^{-\frac{V_{GS}-V_T}{n\phi_t}} \left(1 - \exp^{-\frac{V_{GS}}{\phi_T}}\right) \tag{10.3}$$

La règle pratique appliquée est donc de diminuer la tension non pas en divisant par α mais en divisant par α/ε avec ε supérieur à un. La conséquence est une augmentation du champ électrique d'un facteur ε.

Les conséquences de la réduction de taille du transistor sont résumées dans le *tableau 10.1* qui exprime la variation des principaux paramètres physiques et électriques en fonction des coefficients α et ε. Rappelons que les deux coefficients sont supérieurs à 1.

Les paramètres électriques comme le retard ou la puissance dissipée sont estimés au premier ordre de la manière suivante. Le retard intrinsèque est le temps de transit dans le canal c'est-à-dire le rapport entre la longueur du canal et la vitesse de saturation. La puissance consommée est la puissance dynamique, produit de la capacité du dispositif par le carré de la tension et par la fréquence de fonctionnement. La capacité varie en $1/\alpha$, mais la fréquence de fonctionnement varie en α.

Il est maintenant possible de donner les valeurs absolues de ces paramètres en se basant sur les prévisions de l'ITRS, organisation internationale regroupant tous les fabricants de semi-conducteurs. Ce tableau donne à la fois des informations techniques et économiques relatives à l'évolution des transistors et des circuits intégrés.

On notera que la longueur du canal est plus faible que le nœud de la technologie, défini comme le demi-pas minimum entre deux pistes conductrices. On notera l'évolution de la tension d'alimentation et de la fréquence de l'horloge locale ainsi que l'augmentation du nombre de niveaux d'interconnexions. Enfin, il est important de constater l'augmentation exponentielle du prix d'un jeu de masques. Quand deux chiffres figurent dans une case du tableau, cela veut simplement dire qu'il y a deux familles de composants.

Tableau 10.1

	Règle standard	Règle modifiée
Longueur du canal	$1/\alpha$	$1/\alpha$
Épaisseur de l'oxyde de grille	$1/\alpha$	$1/\alpha$
Largeur du transistor (*W*)	$1/\alpha$	$1/\alpha$
Champ électrique	1	e
Tension d'alimentation	$1/\alpha$	ε/α
Courant de conduction	$1/\alpha$	ε/α
Dopage du silicium	α	$\varepsilon\alpha$
Surface du transistor	$1/\alpha^2$	$1/\alpha^2$
Capacités du dispositif	$1/\alpha$	$1/\alpha$
Retard intrinsèque	$1/\alpha$	$1/\alpha$
Dissipation de puissance	$1/\alpha^2$	ε^2/α^2
Dissipation de puissance par unité de surface	1	ε^2

Tableau 10.2

	2004	2007	2010	2013
Nœud (nm)	90	65	45	32
Longueur canal (nm)	37 53	25 32	18 22	16 22
Densité en millions de transistors par cm²	77 390	150 830	300 1700	600 3400
Taille de la puce (mm²)	140	140	140	140
Fréquence maximale de l'horloge locale (MHz)	4 100	9 200	15 000	23 000
Nombre de couches de connexion	10 à 14	11 à 15	12 à 16	12 à 16
Tension alimentation (volts)	1,2 0,9	1,1 0,8	1,0 0,7	0,9 0,6
Coût du transistor (µcent)	34	12	4	1,5
Coût du jeu de masques en millions de $	1	1,5	3	7

10.2 Dégradation des performances électriques

Une évolution importante de la micro-électronique est la remise en cause de l'architecture classique du transistor MOS. Pour comprendre cette évolution, il faut revenir de manière plus précise sur le comportement du transistor en régime de canal court. Une théorie analytique simplifiée du transistor à deux dimensions permet de comprendre l'essentiel du problème.

On considère donc le schéma simplifié du transistor à canal court comme le montre la *figure 10.2*.

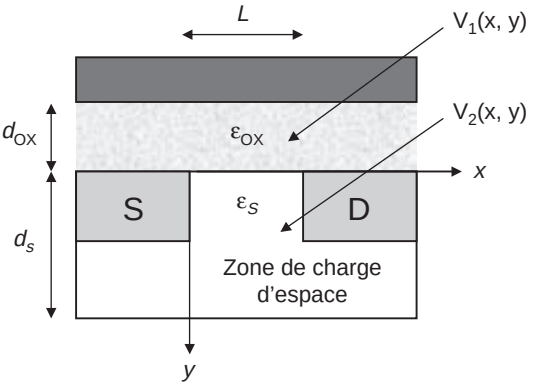


Figure 10.2 – Fonctionnement à deux dimensions.

On peut calculer le potentiel électrique dans les deux régions indiquées sur la *figure 10.2*. Un raisonnement électrostatique non détaillé dans ce paragraphe et tiré de la référence **[8]** conduit à écrire les deux potentiels correspondant aux deux régions sous la forme suivante :

$$V_1(x,y) = V_1(x) + u_{g1}(x,y) + u_{d1}(x,y)$$
$$V_2(x,y) = V_2(x) + u_{g2}(x,y) + u_{d2}(x,y)$$

Le potentiel $V_1(x)$ est le potentiel obtenu dans le chapitre 4 quand on néglige les effets à deux dimensions. Les fonctions $u_{g1}(x,y)$ et $u_{d1}(x,y)$ sont les solutions gauche et droite de l'équation de Laplace quand on applique les potentiels sur les électrodes. Elles s'écrivent sous forme de séries infinies, mais on peut se limiter aux premiers termes de la série.

$$u_{g1} = b_{11}\frac{\sinh(k_1(L-x))}{\sinh(k_1 L)}\sin(k_1(y+d_{OX}))$$

$$u_{d1} = c_{11}\frac{\sinh(k_1 x)}{\sinh(k_1 L)}\sin(k_1(y+d_{OX}))$$

$$u_{g2} = b_{21}\frac{\sinh(k_1(L-x))}{\sinh(k_1 L)}\sin(k_1(y-d_s)+\pi)$$

$$u_{d2} = c_{21}\frac{\sinh(k_1 x)}{\sinh(k_1 L)}\sin(k_1(y-d_s)+\pi)$$

Les coefficients *b* et *c* figurant dans ces formules dépendent des tensions appliquées mais pas de la position. On vérifie que ces fonctions de correction sont nulles aux limites du dispositif. Ce sont les cas : $y = -d_{OX}$ et $y = d_S$ ainsi que $x = 0$ et $x = L$.

En appliquant maintenant la condition aux limites en $y = 0$ c'est-à-dire en exprimant l'égalité des deux fonctions $V_1(x, 0)$ et $V_2(x, 0)$, on obtient.

$$b_{11} \sin(k_1 d_s) = b_{21} \sin(\pi - k_1 d_s)$$

La conservation du produit εE quand on passe d'une région à l'autre permet d'écrire :

$$\varepsilon_{OX} b_{11} \cos(k_1 d_s) = \varepsilon_s b_{21} \cos(\pi - k_1 d_s)$$

En divisant membre à membre les deux équations, on obtient :

$$0 = \varepsilon_s \tan(k_1 d_{OX}) + \varepsilon_{OX} \tan(k_1 d_s) \tag{10.4}$$

Cette formule simple fixe la valeur du paramètre k_1 exprimant les variations spatiales du potentiel en deux dimensions dues aux effets de canal court. Exprimons par exemple le potentiel de surface au centre du canal. Il peut s'écrire :

$$V_2(L/2,0) = V_2(x) + (b_{21} + c_{21}) \cdot \frac{\sinh\left(k_1 \frac{L}{2}\right)}{\sinh(k_1 L)} \sin(\pi - k_1 d_s)$$

On constate donc la dépendance du potentiel en $\exp^{-k_1 \frac{L}{2}}$ et en conséquence un certain nombre d'effets néfastes pour le fonctionnement du transistor. Ces effets sont :

- augmentation de l'effet d'abaissement de la barrière induite par le drain (DIBL) ;
- dérive de la tension de seuil ;
- augmentation de la conductance de sortie.

L'abaissement de la barrière (DIBL) traduit l'influence de la tension de drain sur les électrons du canal. Il est équivalent à une diminution de la tension de seuil. Il est défini en considérant deux tensions de drain, 0,05 V et 1 V pour une technologie avancée. C'est alors la différence entre les deux tensions de seuil.

La dérive de la tension de seuil est due à la modification de la zone de charge d'espace en comparaison du cas dans lequel le canal est long.

Les conductances de sortie et la transconductance sont mesurées pour une tension moyenne et exprimées en Siemens par cm car elles dépendent de la largeur du transistor.

Les courbes sont exprimées en fonction du paramètre Λ défini par :

$$\Lambda = \pi / k_1$$

Les différentes grandeurs électriques significatives sont représentées *figure 10.3*.

La valeur 0,4 du paramètre L/Λ est un minimum absolu car dans ce cas, la valeur de la transconductance est égale à la conductance de sortie. Le gain en tension d'un étage est alors au maximum égal à un. Une valeur minimale de L/Λ est estimée à 1,5 ce qui conduit à une valeur inférieure à 150 mV pour le DIBL et à un rapport de 10 pour g_m/g_{out}.

La relation entre les effets de canal court sur la répartition du potentiel et la valeur des paramètres électriques précédents n'est pas toujours évidente. Les résultats sont obtenus à partir d'un modèle

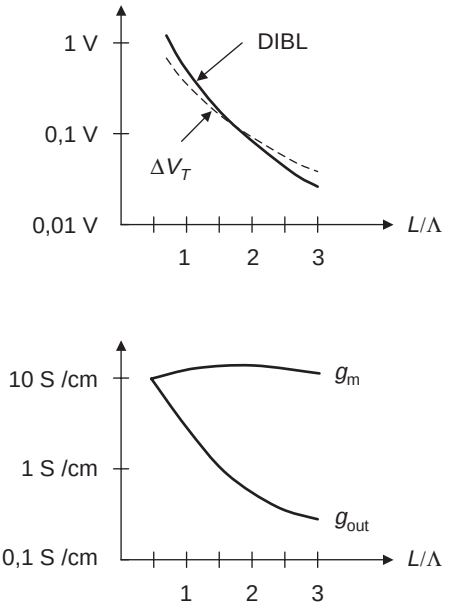


Figure 10.3 – Effet électrique de canal court.

du transistor. Il est cependant facile de comprendre que plus le canal est court, moins la conduction dans le canal est contrôlée par la grille et plus elle dépend de la tension de drain. Le transistor perd alors sa propriété d'être une source de courant commandée en tension.

Il est plus facile d'interpréter les effets de canal court si on exprime de manière plus simple le paramètre k_1 dans l'équation 10.4. Un développement limité de cette équation permet d'écrire quand l'épaisseur d'oxyde est mince :

$$\frac{\pi}{k_1} = d_S + \frac{\varepsilon_S}{\varepsilon_{OX}} d_{OX} - \frac{\pi^2}{3} \frac{\varepsilon_S}{\varepsilon_{OX}} \left(\frac{\varepsilon_S^2}{\varepsilon_{OX}^2} - 1 \right) \frac{d_{OX}^2}{d_S^2} d_{OX}$$

Cette équation se simplifie pour les faibles valeurs de l'épaisseur d'oxyde :

$$\frac{\pi}{k_1} = d_S + \frac{\varepsilon_s}{\varepsilon_{OX}} d_{OX} \tag{10.5}$$

Le rapport à prendre en compte dans l'analyse des effets de réduction des dimensions est donc :

$$k_1 L / \pi = \frac{L}{d_S + \frac{\varepsilon_S}{\varepsilon_{OX}} d_{OX}}$$

Pour augmenter ce rapport, il faut soit diminuer la profondeur de la zone de charge d'espace, soit diminuer l'épaisseur de l'oxyde soit augmenter la permittivité de l'oxyde soit passer à des architectures de transistors plus complexes avec les transistors double grille ou triple grille. Nous allons maintenant examiner ces quatre techniques.

La diminution de la profondeur de la zone de charge d'espace peut se faire en augmentant le dopage comme l'indique la formule 10.2. Cette technique est cependant limitée car elle conduit à augmenter le courant sous le seuil. En effet, dans l'expression 10.3 donnant la valeur du courant en régime de faible inversion, le paramètre n peut s'exprimer au premier ordre par :

$$n = 1 + \frac{\varepsilon_S d_{OX}}{\varepsilon_{OX} d_S} \tag{10.6}$$

Pour que le courant sous le seuil soit faible, il faut minimiser la valeur de n ce qui conduit à augmenter d_S Diminuer d_S conduit donc à une augmentation du courant sous le seuil. L'augmentation du dopage a également des effets néfastes sur la valeur de la mobilité des électrons dans le canal. La solution retenue est donc de maintenir un dopage assez faible en surface et d'augmenter la valeur du dopage en profondeur afin de limiter les effets de canal court. L'implantation ionique permet de réaliser un tel profil de dopage.

La deuxième technique est de diminuer l'épaisseur de l'oxyde. Cette technique est effectivement mise en œuvre si bien que les épaisseurs d'oxyde atteignent des valeurs aussi faibles que 1,5 nm. Il n'est pas possible d'aller beaucoup plus loin car outre les difficultés de réalisation et de contrôle de la surface, des difficultés considérables apparaissent consécutives à l'effet tunnel à travers l'oxyde de grille. Cet effet de nature quantique a été présenté dans le chapitre 2. Il dépend de deux paramètres principaux : la valeur de la tension de grille et l'épaisseur de l'oxyde. Cet effet sera étudié plus en détail dans le paragraphe suivant. Retenons simplement que pour une épaisseur d'oxyde de 2 nm, le courant tunnel est de 0,1 A/cm^2 avec une tension de grille de 1,2 V.

La troisième technique est l'augmentation de la permittivité de l'oxyde. Beaucoup d'études sont menées dans la micro-électronique pour introduire des oxydes ayant des permittivités plus élevées. On les appelle les « high k ». Le plus prometteur est l'oxyde d'Hafnium (HfO_2). La permittivité passe alors de 3,8 à 25. Ce changement d'oxyde serait cependant une évolution majeure de la technologie micro-électronique car il serait nécessaire de revoir également le matériau de grille et le contact de grille avec cet oxyde.

La quatrième technique est de changer l'architecture du dispositif. L'idée générale est d'augmenter le pouvoir de contrôle par la grille et de limiter celui du drain. On retrouve ainsi un comportement idéal du transistor, celui d'une source de courant commandée par la tension de grille. Le transistor double grille représenté *figure 10.4* est un exemple de réalisation.

Le calcul du potentiel dans ce dispositif peut se faire comme dans le transistor classique. On définit de la même manière trois formes de potentiel sous forme de série dans les trois régions du dispositif. On obtient une relation entre le paramètre k_1 et les caractéristiques du transistor de la forme.

$$1 = \frac{\varepsilon_S}{\varepsilon_{OX}} \tan(d_{OX} k_1) \tan\left(d_S \frac{k_1}{2}\right) \tag{10.7}$$

Il est possible de prouver que cette architecture est plus favorable pour maintenir de bonnes performances électriques quand les dimensions et les constantes physiques sont données. Le facteur n défini en régime de faible inversion est par exemple plus proche de l'unité. La *figure 10.5* montre un dispositif réel à double grille.

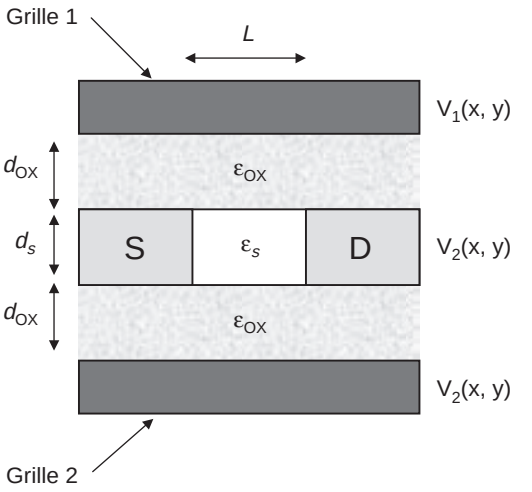


Figure 10.4 – Transistor double grille.

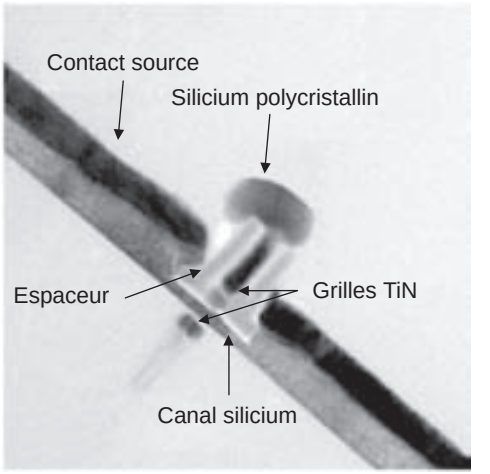


Figure 10.5 – Transistor double grille (Crédit photo CEA Grenoble).

10.3 Les limitations physiques

10.3.1 *Limitations dues à l'effet tunnel*

Le courant tunnel à travers l'oxyde de grille est une des plus sérieuses limites à la réduction de taille du transistor. Le graphique de la *figure 10.6* montre la valeur de ce courant par unité de surface en fonction de l'épaisseur de l'oxyde et de la tension appliquée.

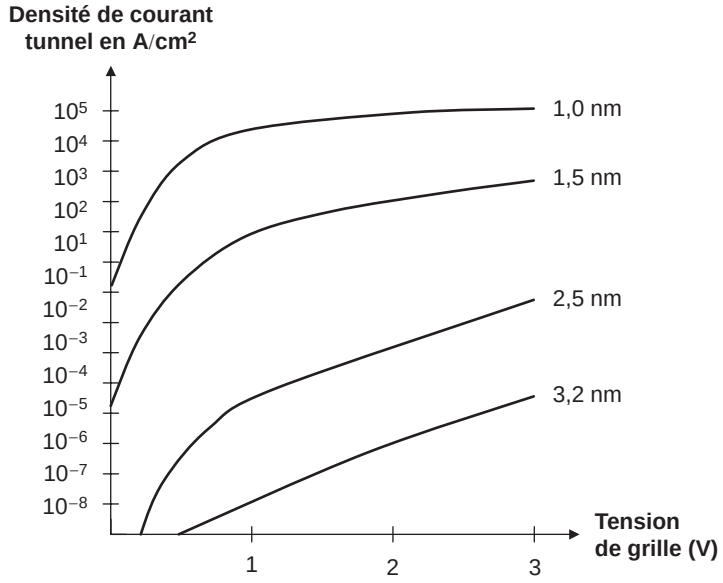


Figure 10.6 – Courant tunnel à travers l'oxyde de grille.

Selon cette figure, une épaisseur d'oxyde de 2 nm conduit à un courant tunnel de 0,1 A/cm^2 pour une tension de grille de 1,2 V. Cette valeur amène à une consommation statique de quelques mW pour l'ensemble du circuit. Pour des épaisseurs de l'ordre du nm, des densités de 10^3 à 10^4 A/cm^2 sont prévues pour une tension de grille de l'ordre du volt. Ces valeurs beaucoup trop importantes conduisent donc à renoncer à utiliser des épaisseurs trop faibles d'oxyde.

Le courant tunnel à travers l'oxyde de grille n'est pas le seul à envisager. Il est également possible qu'un courant puisse circuler du drain vers le caisson du transistor à travers la jonction drain-caisson du transistor. La surface correspondant à cette zone de champ élevé est cependant plus faible que la surface totale de la grille, un tiers environ si bien que des densités de courant plus élevées sont tolérables. Des valeurs maximales de 1 000 A/cm^2 sont données pour des applications à haute performance et de 1 A/cm^2 pour des applications basse consommation. Ces valeurs conduisent alors à définir une valeur minimale de 15 nm pour la zone de charge d'espace soit un dopage au plus égal à 10^{19} dopants par cm^3.

Enfin, il faut également prendre en compte le courant tunnel de la source vers le drain pour des dimensions très faibles du transistor. Cet effet a été mis en évidence sur des dispositifs ayant un canal de 8 nm de long. Cette limite apparaît cependant plus lointaine.

10.3.2 *Tension minimale*

Les difficultés évoquées précédemment pour limiter la consommation statique des circuits amènent à s'interroger sur la valeur minimale de la tension d'alimentation utilisable. L'analyse sera menée en deux temps. Dans une première étape, on étudie l'abaissement de la tension de seuil. Dans une seconde étape, on étudie l'abaissement de la tension d'alimentation.

Diminuer la tension de seuil a deux effets principaux : le premier est la diminution du temps de commutation, le deuxième est l'augmentation du courant sous le seuil qui est le courant traversant un transistor MOS quand il est bloqué. Il faut donc gérer le compromis entre la consommation et

la vitesse comme c'est souvent le cas en électronique. La *figure 10.7* montre l'évolution de la puissance consommée par les processeurs et les processeurs de signaux au fur et à mesure des progrès de la micro-électronique.

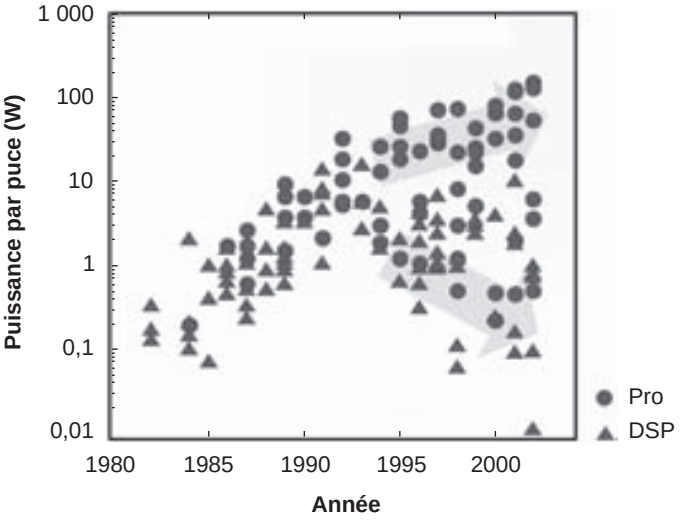


Figure 10.7 – Évolution de la consommation des processeurs.

Cette courbe montre que le marché se segmente en deux familles : les processeurs à haute performance et consommant beaucoup et les processeurs faible consommation mais moins rapides. Pour comprendre le compromis entre consommation et vitesse, il faut rappeler les expressions de la consommation et du retard d'une porte logique.

$$P = a f C V_{DD}^2 + I_0 V_{DD} \exp^{-\frac{V_T}{S}} \tag{10.8}$$

$$t_r = k \frac{C V_{DD}}{(V_{DD} - V_T)^{1,3}} \tag{10.9}$$

Dans ces relations de base, les variables sont définies de la manière suivante.

- a est le facteur d'activité de la porte
- f est la fréquence de l'horloge. La porte commute $a f$ fois par seconde
- C est la capacité totale vue en sortie de porte
- V_{DD} est la tension d'alimentation
- I_0 est une constante définie dans la conduction du transistor en faible inversion
- S est la pente sous le seuil définie par :

$$S = n \frac{k_B T}{e} \tag{10.10}$$

$$n = 1 + \frac{\varepsilon_S d_{OX}}{\varepsilon_{OX} d_S}$$

La relation donnant le retard est semi-empirique. Elle s'explique facilement si on considère la charge du condensateur en sortie de porte par le courant de commutation. Les deux relations 10.8 et 10.9 montrent de manière claire les effets de la tension d'alimentation et de la tension de seuil sur la consommation totale et la vitesse de commutation.

Fixons quelques ordres de grandeur. Avec un courant I_0 de 0,1 A/cm et une pente sous le seuil de 90 mV par décade, on obtient des valeurs tolérables de la puissance pour des seuils compris entre 0,3 V et 0,5 V dans le cas des systèmes faible consommation. Le seuil doit avoir une valeur minimale de 0,1 V pour les systèmes à haute performance qui admettent une consommation plus élevée. Ces considérations conduisent donc les fondeurs à proposer trois types de transistors dans une technologie donnée : un transistor à haute performance, un transistor de performance moyenne, et un transistor faible consommation. On peut également noter l'intérêt des transistors double grille offrant une pente sous le seuil de 70 mV par décade.

La deuxième partie de l'analyse est l'étude de la valeur minimale de la tension d'alimentation.

Cette étude peut se faire en considérant uniquement la caractéristique non linéaire d'une porte en fonction de la configuration des entrées. En effet, la forme de la fonction de transfert d'une porte varie en fonction de la configuration des entrées. Selon les phases de fonctionnement, une seule entrée ou plusieurs peuvent varier. La fonction de transfert entrée-sortie se modifie alors de manière importante. Pour des tensions d'alimentation élevées, la variation relative de la caractéristique est faible et donc la variation du seuil de basculement de la porte. Pour des tensions d'alimentation faibles, la variation relative de la caractéristique de transfert est importante et donc la variation du seuil de basculement de la porte.

Le *tableau 10.3* indique la valeur minimale de la tension d'alimentation pour une porte NAND en fonction du nombre d'entrées. La tension est exprimée en fonction du paramètre significatif nk_BT/e. Le paramètre n est défini dans la relation 10.10.

Tableau 10.3

NAND	Deux entrées	Quatre entrées	Huit entrées	Unité
V_{DD} minimale	2,27	3,01	3,72	nk_BT/e

De manière plus générale, on peut montrer que la tension minimale à appliquer sur une porte comportant F entrées est environ :

$$V_{\min} = n \frac{k_B T}{e} \ln(F) \qquad (10.11)$$

Rappelons que cette expression est obtenue uniquement pour tenir compte des variations de forme de la fonction de transfert en fonction de la configuration des entrées appliquées. Ce raisonnement n'intègre pas les perturbations électriques, le bruit thermique et les variations technologiques. La valeur de la tension minimale obtenue peut donc être considérée comme une limite purement théorique. On obtient ainsi pour une porte à quatre entrées une valeur de 75 mV.

10.3.3 *Limites dues aux résistances d'accès*

Les résistances série apparaissent au niveau des contacts de drain et de source et au niveau de la grille réalisée en silicium polycristallin. Les technologies de fabrication de contacts auto-alignés à base de siliciure réduisent ces résistances à des valeurs aussi faibles que 100 Ω µm. Cette valeur est faible comparée à la résistivité du dispositif total qui est environ de 1 000 Ω µm. Cette dernière valeur est obtenue simplement par :

$$\rho = \frac{V_{DD} - V_T}{I_{on}/W}$$

Si on envisage une réduction de la taille du transistor, le libre parcours moyen des électrons ou des trous peut devenir supérieur à la longueur du canal. Dans ce cas, les porteurs ne subissent plus de collisions avec le réseau. Ce régime dit balistique conduit à une valeur limite de la résistivité équivalente de l'ordre de 500 Ω µm. En résumé, les résistances d'accès n'introduisent pas une contrainte majeure dans la miniaturisation du transistor.

Si on considère la grille polycristalline, la résistance d'accès présentée par cette électrode ne constitue pas un obstacle majeur susceptible de limiter les performances en vitesse du transistor. Elle est par contre responsable d'un effet beaucoup plus gênant, la désertion de grille. Quand une tension est appliquée sur la grille, il se forme dans le silicium polycristallin une zone de charge d'espace que nous avons négligée dans les analyses précédentes. Tout se passe alors comme si l'épaisseur de l'oxyde augmentait ce qui, comme il a été vu précédemment, a un effet négatif sur le pouvoir de contrôle du courant par la grille. Cet effet est faible en valeur absolue étant donné la faible résistivité du silicium de la grille. L'épaisseur de l'oxyde est très faible également, de l'ordre du nm pour les technologies avancées, si bien que l'effet relatif de désertion de la grille peut être non négligeable. Cet effet justifie l'intérêt des recherches menées dans la micro-électronique ayant pour objectif le remplacement du silicium polycristallin de grille par un métal. Cette modification fondamentale de la technologie de fabrication du transistor doit cependant être validée avant de devenir un standard industriel.

10.3.4 *Quelques dispositifs avancés*

L'objectif de ce paragraphe est de décrire quelques dispositifs avancés illustrant les principes généraux exposés précédemment. Ces dispositifs ne sont donnés qu'à titre d'exemples et ne prétendent pas préfigurer les dispositifs du futur.

Le premier dispositif est un transistor à grille enterrée. Il est représenté *figure 10.8*. Il est le plus petit transistor réalisé à ce jour puisque la longueur de la grille est de 8 nm.

La grille supérieure crée des zones d'inversion dans les régions de drain et de source ce qui minimise l'influence du drain sur le courant par rapport à un dispositif classique. La zone de charge d'espace a une profondeur d'environ 25 nm et l'épaisseur de l'oxyde est de 5 nm. La longueur équivalente de grille n'est cependant que de 20 nm à cause de l'effet de la grille supérieure. Ce transistor expérimental a permis d'étudier les effets tunnel de la source vers le drain.

La *figure 10.9* montre diverses manières de réaliser des transistors à plusieurs grilles. Rappelons que l'intérêt de ces structures est d'améliorer le pouvoir de contrôle du canal de conduction par le potentiel de grille. Une grille entourant totalement le canal de conduction peut être considérée comme une architecture idéale. Ces dispositifs sont en général réalisés sur un substrat de type SOI, c'est-à-dire composé d'un empilement silicium/oxyde/silicium. La couche de silicium servant à réaliser les dispositifs est très mince.

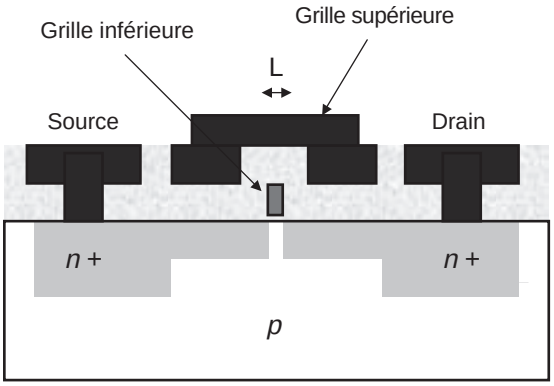


Figure 10.8 – Transistor de très faible taille à grille enterrée.

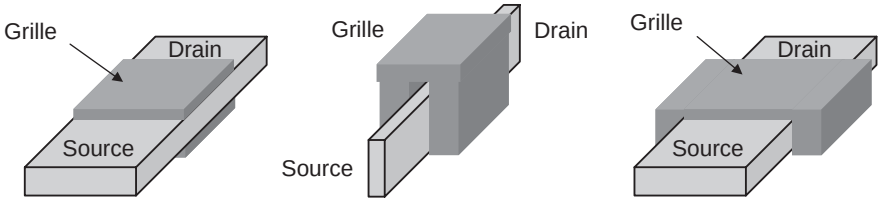


Figure 10.9 – Transistors multigrille.

10.4 Les limitations dues aux dispersions

Les limites théoriques sont une chose, les contraintes de fabrication en sont une autre. Les limites de fabrication seront étudiées pour le transistor classique et pour le transistor double grille. Enfin, les fluctuations de dopage et les problèmes de dissipation de la puissance seront pris en compte.

10.4.1 *Fabrication du transistor classique*

Pour donner un caractère plus concret à cette étude, il est possible d'imaginer un transistor avancé, par exemple un transistor prévu en 2008 correspondant au nœud 45 nm. Ce transistor présente une longueur de canal d'environ 25 nm et une épaisseur d'oxyde de 1,5 nm, grandeur supérieure à la valeur de 1 nm préconisée pour éviter les effets de canal court. La tension d'alimentation est de 1 V et la tension de seuil est maintenue à 200 mV pour maintenir le courant sous le seuil à une valeur tolérable.

Il est nécessaire de contrôler de manière précise le dopage dans le substrat pour limiter les effets de canal court. Ce dopage particulier, non uniforme en profondeur et dans la dimension latérale, est appelé « superhalo ». Il permet de maintenir une tension de seuil plus ou moins indépendante des variations de longueur du canal. Cette propriété est fondamentale car la lithographie ne permet

pas de garantir une longueur de canal constante pour des dimensions aussi faibles que 25 nm. Les variations de tension de seuil poseraient alors des problèmes sérieux pour le dessin de fonctions numériques. Le profil de dopage recherché est obtenu par les techniques classiques d'implantation ionique mais la mise au point est assez délicate. Moyennant ces précautions, une relative stabilité de la tension de seuil est attendue. Un retard intrinsèque aussi faible que 4 ps est prévu pour ce type de dispositif.

10.4.2 *Fabrication du transistor double grille*

Cette géométrie présente des avantages indéniables du point de vue électrostatique comme il a été vu précédemment. Il est cependant nécessaire d'étudier plus en détail les effets de la non uniformité de l'épaisseur de silicium. Une variation d'épaisseur du canal de silicium a deux effets principaux : une modification de la longueur électrique équivalente du canal et une variation de la tension de seuil. Ces deux effets peuvent être estimés par les relations suivantes :

$$\frac{\Delta L}{L} = \frac{2\,\varepsilon_S\, d_{OX} + \varepsilon_{OX}\, d_S}{4\,\varepsilon_S\, d_{OX} + \varepsilon_{OX}\, d_S} \cdot \frac{\Delta d_S}{d_S} \qquad (10.12)$$

$$\Delta V_T = -\frac{h^2}{4\, m^*\, e\, d_S^2} \frac{\Delta d_S}{d_S}$$

La démonstration de ces relations sort du cadre de cet ouvrage. Elles conduisent donc à une valeur minimale de 5 nm pour l'épaisseur de la couche de silicium utilisée. Il faut alors assurer une précision de 0,5 nm sur l'épaisseur de silicium.

La deuxième cause de variation de la tension de seuil est la variation de la valeur du dopage. La tension de seuil peut être ajustée en jouant sur le dopage. Pour un substrat silicium de type *p*, on peut écrire :

$$\Delta V_T = \frac{e\, N_A\, d_S\, d_{OX}}{2\,\varepsilon_{OX}}$$

Le contrôle du dopage et des épaisseurs est donc nécessaire pour garantir une valeur de seuil donnée.

La fabrication des transistors double grille est en résumé très délicate puisqu'elle doit satisfaire les contraintes suivantes : contrôle précis de l'épaisseur de la zone de silicium, alignement précis des grilles supérieures et inférieures pour éviter les capacités parasites, contrôle précis du dopage dans le cas où il est utilisé pour régler la tension de seuil.

10.4.3 *Les fluctuations du dopage et leurs effets*

Le nombre d'atomes dopants dans la zone de charge d'espace d'un transistor MOS décroît avec la réduction de taille du transistor comme le montre la *figure 10.10*.

Le nombre de dopants implantés *N* fluctue en fonction de la loi de Poisson. En dessous de 1 000 dopants, les fluctuations statistiques, mesurées par leur écart type en racine de *N*, ne sont plus négligeables en valeur relative. Cette fluctuation du dopage a pour conséquence une fluctuation des tensions de seuil. Notons à ce propos que les profils non uniformes de dopage atténuent cet effet. Dans une implantation profonde non uniforme, le maximum de dopage est en profondeur si bien que l'effet électrostatique des ions est écranté par les porteurs libres du canal. Des simulations numériques conduisent pour un transistor de 25 nm de longueur de grille à un écart type de 10 mV pour une largeur de un micron et pour un dopage uniforme.

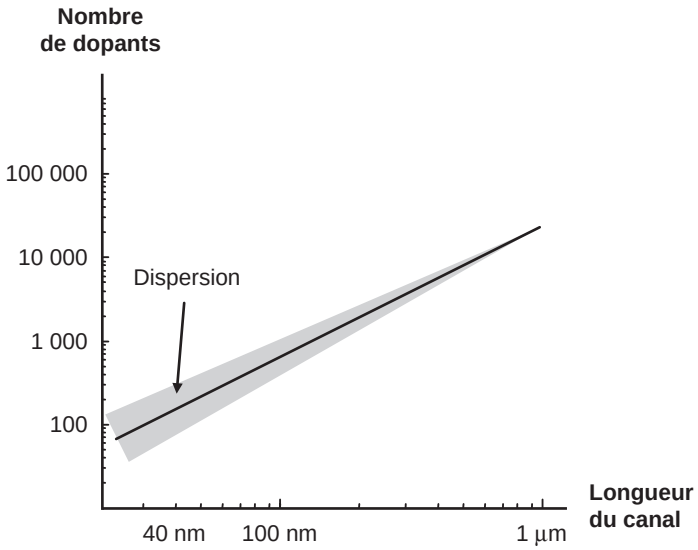


Figure 10.10 – Variation du nombre de dopants.

Cette valeur est divisée par deux pour un dopage non uniforme. Les transistors des circuits logiques, de tailles relativement importantes, sont peu affectés par cet effet mais les transistors de petite taille des mémoires SRAM sont affectés de variations de seuil de l'ordre de 40 mV.

10.4.4 *La dissipation de puissance*

La puissance dissipable est intrinsèquement élevée, de l'ordre du kW par cm^2 avec un liquide de refroidissement. Par contre, la puissance consommée est limitée par les spécifications des produits. Cette contrainte est particulièrement forte pour les systèmes portables alimentés par des batteries.

10.5 Limites et applications

Les paragraphes précédents ont mis en évidence que la miniaturisation amenait à un certain nombre de compromis dans le dimensionnement du transistor en particulier pour limiter l'augmentation de puissance consommée et pour maintenir les dispersions de la valeur de la tension de seuil et du retard intrinsèque dans une gamme convenable. Il est maintenant possible de reprendre cette analyse pour chaque classe d'application.

10.5.1 *Miniaturisation du transistor et DRAM*

Dans la fabrication des DRAM, la réduction de taille du transistor est l'objectif premier car il s'agit de baisser le coût du bit stocké. Aujourd'hui, les technologies DRAM utilisent des transistors plus longs et des tensions d'alimentation plus élevées que les technologies microprocesseurs.

En effet, comme le montre la *figure 10.11*, le seuil du transistor doit être fixé à une valeur élevée (500 mV environ) pour éviter la décharge du condensateur de stockage quand la ligne de bit est à la tension basse. La tension appliquée sur le condensateur doit être plus faible que celle appliquée sur la grille, avec au mois 1,5 V de différence. L'épaisseur de l'oxyde du condensateur de stockage est plus faible que celle de l'oxyde de grille du transistor. Des épaisseurs de 3 nm sont courantes et

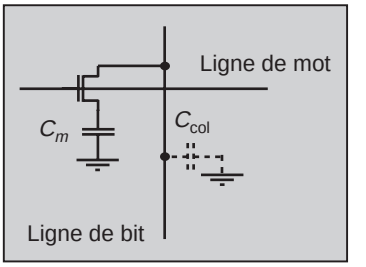


Figure 10.11 – Cellule mémoire d'une DRAM.

permettent d'atteindre des valeurs de capacités suffisantes pour stocker une charge importante. Le diélectrique ONO permet de réduire l'épaisseur à 2,5 nm.

La réduction de taille des DRAM est donc maintenant une opération complexe. Les tailles actuelles des cellules DRAM sont de $8\lambda^2$, formule dans laquelle λ est le nœud de la technologie.

Les DRAM peuvent évoluer dans le futur en améliorant l'architecture en profondeur du condensateur de stockage comme il a été expliqué dans le chapitre 9 mais aussi en les intégrant aux circuits logiques ce qui introduit le concept de DRAM embarquée.

10.5.2 *Miniaturisation du transistor et SRAM*

Les SRAM jouent un rôle de plus en plus important dans les circuits intégrés car elles sont associées aux processeurs sous forme de mémoires cache de différents niveaux comme le montre la *figure 10.12*.

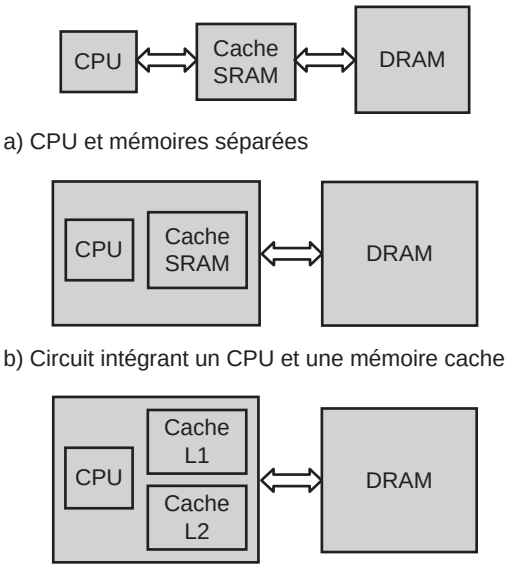


Figure 10.12 – Évolution des SRAM dans les circuits intégrés.

Initialement unité de calcul et mémoires étaient séparées si bien que les transistors de la mémoire SRAM pouvaient être optimisés spécifiquement. Ensuite, une mémoire cache de petite taille a été ajoutée à l'unité de calcul dans une optique d'optimisation de la vitesse de lecture et d'écriture.

Les circuits intégrés actuels comportent en plus de l'unité de calcul au moins deux mémoires SRAM notées L1 et L2 et de caractéristiques différentes. La première est de taille limitée mais très rapide et conserve les données les plus souvent utilisées. Les transistors mis en œuvre ont les mêmes contraintes de dimensionnement que les transistors de l'unité de calcul et l'optimisation de la vitesse est le critère principal. La mémoire L2 est également une SRAM et sa taille peut être de quelques méga-octets.

La cellule SRAM est de manière classique réalisée à partir de 6 transistors comme le montre la *figure 10.13*. L'encombrement est de 50 λ^2 environ. Les transistors de la mémoire L1 sont dimensionnés principalement en fonction du critère de vitesse (canal court et seuil faible) alors que les transistors de L2 sont dimensionnés en fonction de la consommation statique (canal plus long et seuil plus élevé).

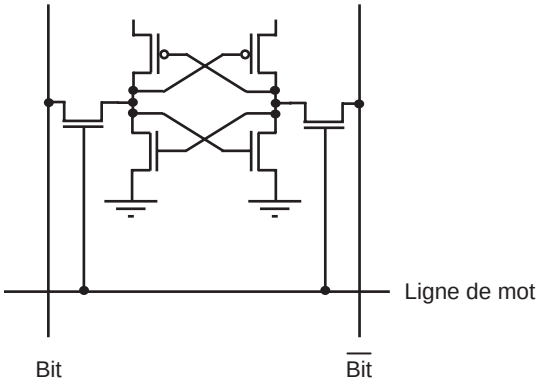


Figure 10.13 – Cellule SRAM.

Une évolution possible est d'aller vers plus de souplesse dans les conditions de fonctionnement des transistors des mémoires SRAM embarquées en ajustant les tensions d'alimentation et de seuil en fonction des demandes du logiciel. Une évolution à plus long terme est de remplacer les SRAM par des mémoires non volatiles ce qui permettrait de gagner encore en terme de consommation et de facilité d'utilisation. Les mémoires de type MRAM ou PCRAM semblent les alternatives les plus prometteuses.

10.5.3 *Miniaturisation du transistor et fonctions de calcul*

L'objectif est clairement d'augmenter la vitesse si bien que les transistors sont dimensionnés avec des tensions de seuil les plus faibles possibles. Le problème majeur est alors de contrôler les dispersions de ces tensions de seuil, dispersion d'un wafer à un autre et dispersions d'un circuit à un autre sur un même wafer. L'effet de la dispersion des seuils est mis en évidence en écrivant le courant de repos d'un ensemble de N transistors.

$$I_{\text{OFF}} = N \cdot I_{\text{OFF moyen}} \cdot \exp^{\frac{\sigma^2 V_T}{2(\ln(10)S)^2}} \qquad (10.13)$$

Dans cette relation, on exprime le courant sous le seuil moyen $I_{OFF\ moyen}$, la dispersion de la tension de seuil σ^2_{VT} mesurée sur un même circuit et la pente sous le seuil S exprimée en mV par décade. Cette relation s'obtient par simple somme de tous les courants individuels. Les transistors ayant des tensions de seuil faibles contribuent de manière déterminante au courant total, et plus la dispersion est grande, plus ces transistors sont nombreux. Ces considérations conduisent à choisir des tensions de seuil autour de 25 % de la tension d'alimentation.

On peut alors classer les circuits de calcul en trois familles : haute performance, performance moyenne et basse consommation. La première famille comprend les processeurs pour les ordinateurs individuels et les consoles de jeux, la deuxième famille comprend les circuits pour les ordinateurs portables ou les mobiles téléphoniques et la troisième famille comprend les circuits des objets communicants devant fonctionner sur batterie de manière autonome.

Les performances des transistors optimisés pour ces différentes classes apparaissent dans le *tableau 10.4*.

Tableau 10.4 – Transistor optimal et applications.

Type	Applications	L (nm)	d_{OX} (nm)	L (nm)	V_T (mV)	S (mV/déc)	V_{DD} (V)	I_{OFF} (nA/µm)	P (W/cm²)
MOS	Haute performance	15	1,3	9	140 230	100	0,9	1 000 100	1 000 30
DG-MOS		13	1,3	9	155 255	85	0,8	1 000 75	1 000 30
MOS	Performance moyenne	20	1,4	13	300 390	100	0,8	25 2	30 5
DG-MOS		15	1,4	10	300 390	85	0,8	20 2	5 0,5
MOS	Très basse consommation	30	2,3	21	550	87	0,8	0,008	0,001
DG-MOS		19	2,3	13	530	75	0.7	0,005	0,001

Il est donc possible de définir une sorte de transistor idéal en fonction de l'application recherchée. Les architectures simple et double grille sont les seules à figurer dans ce tableau par souci de simplicité. Les solutions triple grille ou FINFET sont également envisageables mais offrent des résultats voisins de ceux obtenus avec le double grille. L'émergence des diélectriques à haute permittivité peut également modifier les choses en limitant les difficultés liées au courant tunnel sans résoudre pour autant l'augmentation du courant sous le seuil.

Cette vision est assez nouvelle dans la micro-électronique puisqu'elle conduit à une sorte d'asymptote dans la réduction de taille du transistor avec une longueur de grille optimale de 10 nm pour les applications à hautes performances. Des facteurs de progrès importants restent possibles dans les technologies mémoire en particulier en optimisant le couplage entre les conditions de fonctionnement du circuit et le logiciel.

10.5.4 *Résumé des évolutions technologiques liées à la miniaturisation du transistor*

Il est possible en conclusion de reprendre les différents items étudiés dans ce chapitre pour dresser une possible évolution des techniques de fabrication du transistor. Ces prévisions sont largement inspirées des études de l'ITRS et ne sont données qu'à titre indicatif. Les différents effets physiques à traiter sont portés en ordonnée et les solutions technologiques sont indiquées.

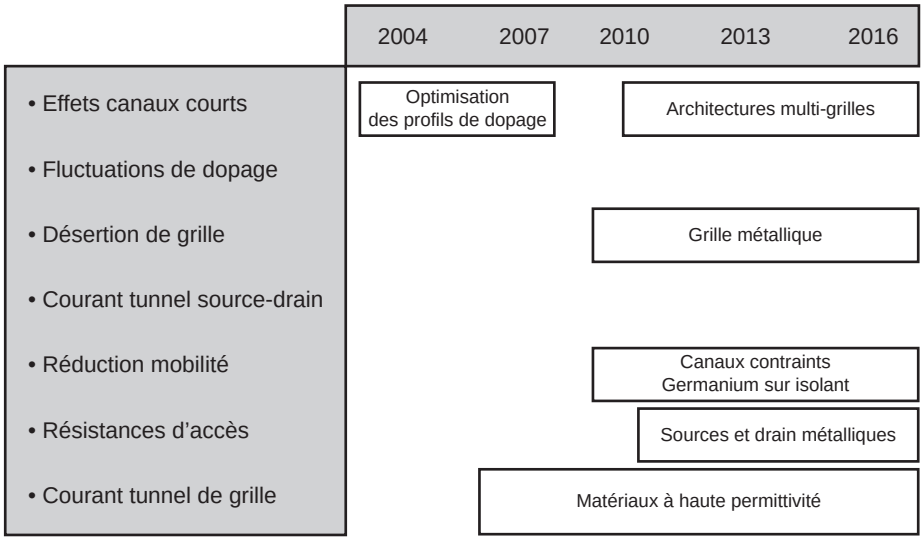


Figure 10.14 – Panorama des évolutions technologiques.

Chapitre 11

Traitement de l'information et nanotechnologies

Ce chapitre étudie l'évolution des architectures. Les nouveaux composants, éventuels remplaçants du transistor MOS, et les transistors MOS eux-mêmes seront-ils assemblés dans des systèmes équivalents à ceux que nous connaissons aujourd'hui ? Les techniques de fabrication envisagées et les contraintes du monde nanométrique laissent penser qu'il sera de plus en plus difficile de réaliser à grande échelle des composants présentant un comportement déterministe. Les défauts seront nombreux et les dispersions ne cesseront de croître sous l'effet de la miniaturisation. La position et l'interconnexion des composants pourront même être en partie inconnues si les techniques d'auto-assemblage et d'auto-organisation sont utilisées. De nouvelles architectures sont alors nécessaires pour tirer parti des propriétés des nanocomposants. Cette condition est indispensable pour que les nanotechnologies soient appliquées à grande échelle dans les systèmes de traitement de l'information.

11.1 Les limites physiques en micro-électronique

Ce paragraphe a pour but de déterminer l'impact des lois physiques de base sur les limites en performances des systèmes électroniques. Les aspects énergétiques et les considérations liées aux interconnexions sont les plus importants.

11.1.1 *Énergie minimale de commutation*

La question est la suivante : quelle est l'énergie minimale pour qu'un élément logique de base change d'état ? Elle peut se traiter de diverses manières qui conduiront toutes à calculer cette énergie comme un multiple de la grandeur de base $k_B T$ dont la valeur est 26 meV.

La première manière de traiter le problème est d'appliquer le théorème de Shannon, théorème de base en théorie de l'information. Ce théorème relie le débit maximum d'un canal de transmission C à la bande passante B de ce canal et au rapport entre l'énergie du signal E et celle du bruit N. Les énergies du signal et du bruit seront définies comme les énergies dissipées pendant la transmission du signal.

$$C = B \log_2 \left(\frac{E + N}{N} \right) \tag{11.1}$$

Appliquée à la transmission d'un bit d'information dans une cellule de traitement (un inverseur logique par exemple), cette formule devient :

$$\frac{C}{B} = \frac{\ln\left(1 + \frac{E}{N}\right)}{\ln 2}$$

Comme $\ln(1 + x)$ est inférieur à x, on en déduit :

$$\frac{E}{C} > \frac{N}{B} \ln 2$$

Si le bruit est thermique, la valeur de N/B est $k_B T$. On obtient donc :

$$\frac{E}{C} > k_B T \ln 2 \tag{11.2}$$

Cette relation donne une valeur minimale à l'énergie par bit transmis dans une unité logique.

La même relation peut s'obtenir par un raisonnement inspiré de la thermodynamique statistique. L'entropie d'un système étant défini par la relation de Boltzman :

$$S = k_B \ln \Omega$$

La grandeur Ω exprime le nombre total d'états possibles du système. Appliquée à une unité logique ayant 2 entrées et une sortie, la variation d'entropie s'écrit :

$$\Delta S = k_B \ln 4 - k_B \ln 2 = k_B \ln 2$$

Cette formule se généralise facilement à un système ayant n entrées et $n - 1$ sorties. La thermodynamique nous apprend que cette variation d'entropie est associée à une dissipation de chaleur selon la formule :

$$E = T \Delta S$$

soit,

$$E = k_B T \ln 2$$

Certains auteurs ont tenté d'aller au-delà de cette limite en introduisant le calcul réversible. Cette technique suppose qu'il n'y a pas perte d'information quand elle se propage dans un élément logique. En pratique, il y a autant de sorties que d'entrées. Il faut cependant avoir à l'esprit que les systèmes électroniques actuels ont besoin, pour transmettre un bit d'information, d'une énergie bien supérieure à cette limite théorique.

Une dernière manière de traiter ce problème est de considérer le transfert d'un bit entre deux portes et de calculer le taux d'erreur de la transmission. Cette approche pragmatique conduit à des valeurs de l'énergie minimale de commutation bien supérieure à la limite théorique de $k_B T \ln 2$ puisqu'il faut environ 165 $k_B T$ pour transmettre un bit avec un taux d'erreur inférieur à 10^{-19}. Un schéma simple représenté *figure 11.1* permet d'établir ce résultat.

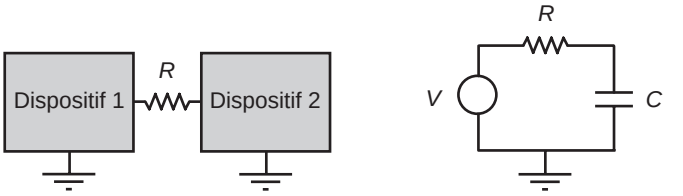


Figure 11.1 – Énergie de commutation et taux d'erreur.

On considère deux dispositifs quelconques reliés par un élément de connexion présentant une résistance de valeur R et une capacité C par rapport à la masse. On suppose que les dispositifs sont parfaits. Le dispositif 1 est une source de tension parfaite qui génère par bit transmis une variation de tension égale à V. Le dispositif 2 est un récepteur parfait d'impédance d'entrée infinie et non capacitif. Un calcul analogue à celui effectué dans le chapitre 8 montre que l'énergie E dissipée dans la résistance pendant la transition est :

$$E = \frac{1}{2} C V^2$$

La variance σ_N^2 de la tension de bruit en entrée du dispositif 2 est alors donnée par la relation suivante, en supposant la bande passante limitée uniquement par le produit RC :

$$\sigma_N^2 = 4 k_B T R \int_0^\infty \frac{1}{1 + 4\pi^2 R^2 C^2} \mathrm{d}f = \frac{k_B T}{C}$$

Le rapport V^2/σ_N^2 permet de mesurer la capacité du dispositif 2 à décider quelle est la valeur d'un bit transmis. Ce rapport s'écrit :

$$\eta = \frac{V^2}{k_B T / C} = \frac{2E}{k_B T}$$

Ce rapport est directement relié au taux d'erreur dans la transmission de l'information comme le montre la *figure 11.2*. On peut alors calculer la probabilité d'erreur en fonction de ce rapport. On trouve ainsi une valeur minimale de 165 $k_B T$ pour un taux d'erreur inférieur à 10^{-19}.

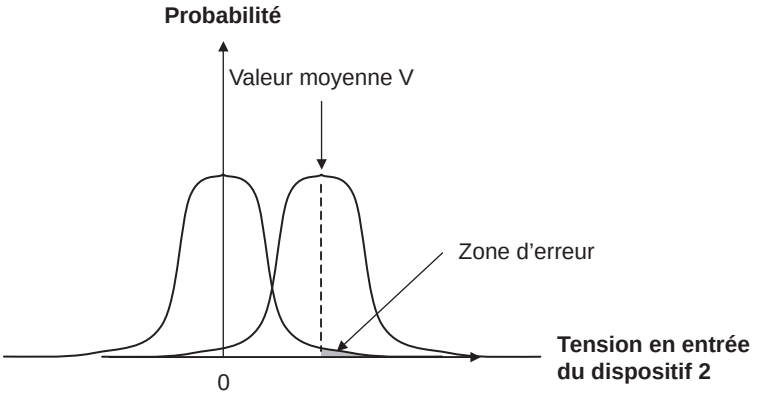


Figure 11.2 – Probabilité d'erreur.

11.1.2 *Volume minimum d'un dispositif*

Combien d'atomes faut-il pour faire un dispositif logique ? Cette question a déjà été posée dans le chapitre 10. Dans le cas des transistors, c'est l'effet de la fluctuation statistique du nombre de dopants sur les dispersions de la tension de seuil qui est l'élément principal d'analyse.

Une autre grandeur à prendre en compte est la tension de claquage dans le dispositif. Un cube de silicium de 1 µm de côté comporte environ 10^4 atomes de dopants pour des dopages classiques de la micro-électronique. Pour une tension appliquée donnée, la valeur du dopage détermine la valeur du champ maximum dans le dispositif et finalement la tension de claquage. Ce n'est pas le nombre total de dopants qui compte pour déterminer la tension de claquage du dispositif mais le dopage local. On considère généralement une unité de volume formée par un cube de côté a égal à la profondeur de la zone de charge d'espace. Le nombre de dopants à considérer est donc Na^3 dans laquelle N est la densité de dopants. Ce nombre est affecté d'une fluctuation de Poisson égale à $\sqrt{Na^2}$. La dispersion relative du nombre de dopants doit alors rester acceptable (inférieure à 10 %) pour éviter des claquages locaux.

Une deuxième raison pour définir un volume minimum du dispositif est la résistance aux radiations naturelles. Une particule de haute énergie (proton ou rayon cosmique) est capable de créer dans un semi-conducteur un grand nombre de paires électron-trou. Dans le silicium, il suffit de diviser l'énergie de la particule par 3,6 eV. La charge créée doit être comparée à celle mise en œuvre dans un changement d'état logique. Le cas des mémoires est particulièrement critique.

11.1.3 *Les limites apportées par les interconnexions*

Le chapitre 6 a montré que les interconnexions occupaient la plus grande partie de la surface d'un circuit intégré. La *figure 11.3* représente comment la longueur totale des interconnexions d'un circuit intégré a évolué depuis les années 70.

Cette courbe montre de manière évidente l'importance de l'interconnexion dans l'évolution des circuits intégrés. Robert N. Noyce, un des inventeurs du circuit intégré, a ainsi déclaré :

"*The integrated circuit is the component industry's solution to the interconnection problem.*"

L'interconnexion est aussi la source de problèmes difficiles et constitue la véritable limite à l'évolution de la complexité des systèmes électroniques intégrés. Quand un signal se propage sur une ligne con-

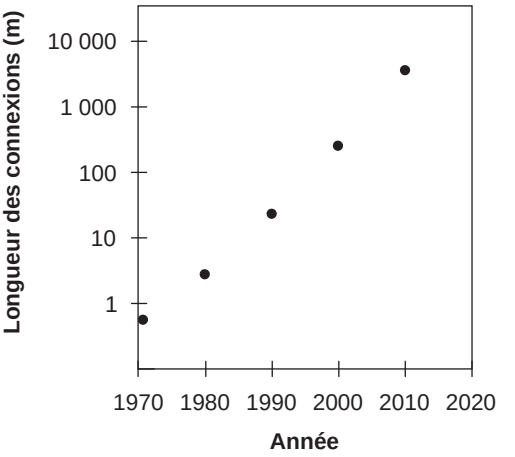


Figure 11.3 – Évolution de l'interconnexion d'un circuit intégré.

ductrice de résistance nulle, la vitesse limite de propagation de l'information est la vitesse de propagation d'une onde électromagnétique dans un milieu ayant une constante diélectrique ε_r soit :

$$v = \frac{c}{\sqrt{\varepsilon_r}} \tag{11.3}$$

Dans cette formule, c est la vitesse de la lumière dans le vide soit 300 000 km/s. Cette limite n'est pas atteinte en pratique car la ligne présente une résistance non nulle et la constante de temps RC présentée par la ligne est le principal facteur de ralentissement.

De manière plus rigoureuse, il faut comparer le temps de montée du signal au temps de propagation v/L de la connexion, L étant la longueur de la connexion. Quand il est très supérieur, ce qui est souvent le cas, la connexion peut être considérée comme un simple circuit RC intégrant le signal et introduisant un retard en conséquence. Quand le temps de montée est inférieur au temps de propagation v/L, il faut considérer la connexion comme une ligne. Le signal se propage alors à la vitesse v.

La *figure 11.4* montre comment les paramètres électriques de base d'une liaison par piste conductrice évoluent avec la miniaturisation.

Une piste conductrice de longueur L et de largeur W est placée à une distance h de la couche inférieure considérée comme un plan de masse. L'épaisseur de métal de la piste est e. Il est facile à partir de ce modèle simple d'estimer la résistance R et la capacité C de cet élément.

$$R = \rho \frac{L}{We}$$

$$C = \varepsilon \frac{LW}{h}$$

La constante de temps RC qui exprime le retard apporté par cette connexion s'écrit donc :

$$\tau = RC = \rho\varepsilon \frac{L^2}{e\,h} \tag{11.3}$$

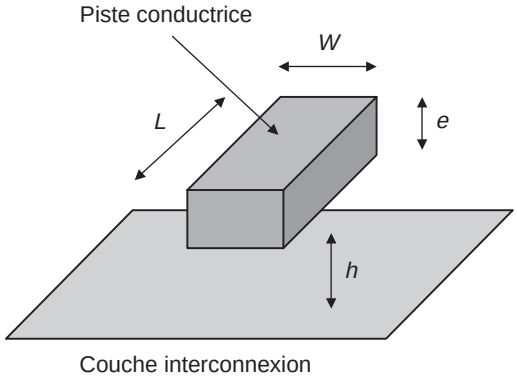


Figure 11.4 – Paramètres électrique d'une piste conductrice dans un circuit intégré.

Quand la technologie évolue avec le paramètre α de réduction de taille, les paramètres technologiques *e* et *h* diminuent proportionnellement à α. La longueur *L* ne diminue pas du même facteur car la dimension totale du circuit intégré ne diminue pas. Il est nécessaire de faire des liens sur des longueurs qui restent du même ordre de grandeur. Pensons par exemple à la liaison entre un processeur et une mémoire. En définitive, la constante de temps τ augmente du facteur α^2.

La *figure 11.5* illustre ce problème en faisant figurer le retard introduit par l'interconnexion comparé au retard apporté par une porte logique et cela pour différentes technologies. Cette figure montre en particulier pourquoi la technologie cuivre est devenue indispensable pour conserver les propriétés de rapidité nécessaires. Elle compare la technologie cuivre à la technologie aluminium pour des longueurs de connexion typiques de 40 µm environ entre composants. Au-delà du nœud 250 nm, la technologie cuivre est une nécessité.

On peut également noter l'intérêt de réaliser des isolants à faible permittivité diélectrique. Ces nouveaux matériaux, le plus souvent réalisés à partir de matières poreuses, font également l'objet d'intenses recherches en micro-électronique.

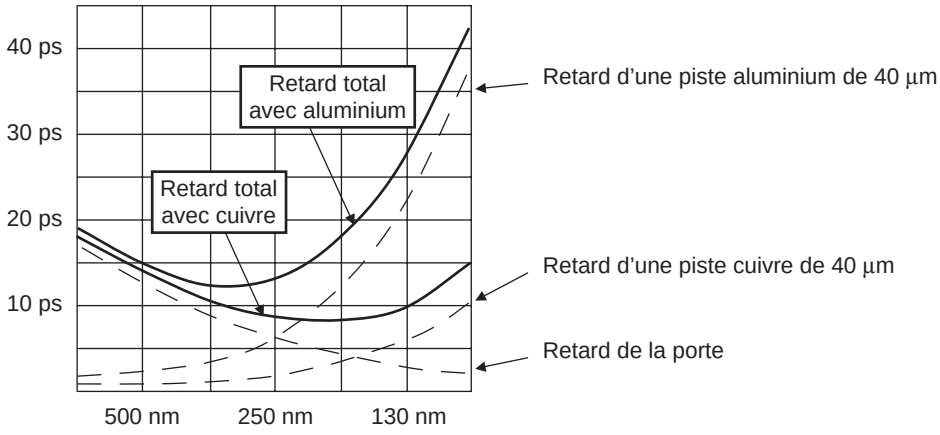


Figure 11.5 – Retard apporté par les interconnexions.

Voyons maintenant comment l'interconnexion influe sur la densité de transistors dans un circuit intégré. Il est possible d'estimer la part de la surface d'un chip occupée par l'interconnexion. On suppose qu'un composant élémentaire, typiquement un transistor, est associé à une surface *a*. La distance moyenne entre composants est donc $\sqrt{a}$. Si on note $m\sqrt{a}$ la longueur moyenne d'une connexion par composant et si *W* est la distance minimale entre deux connexions, on peut dire que la surface d'une connexion par composant est $mW\sqrt{a}$. Si le circuit n'avait qu'une couche de connexions, on obtiendrait, en supposant les transistors entièrement recouverts par l'interconnexion :

$$a = mW\sqrt{a}$$

Le circuit a en fait *K* couches de connexions. La relation devient :

$$Ka = mW\sqrt{a}$$

La surface minimale attribuée à chaque composant est donc :

$$a = \left(\frac{mW}{K}\right)^2 \qquad (11.4)$$

En fait, les paramètres *m* et *W* sont différents de couche à couche. Cette formule simple montre que la densité de composants est largement déterminée par l'interconnexion.

11.2 Les logiques

11.2.1 *La logique multivaluée*

Tous les systèmes électroniques actuels font usage de la logique binaire à deux niveaux. Certains dispositifs nanométriques introduisent cependant la possibilité de traiter plusieurs états. C'est le cas des dispositifs de stockage à nombre limité d'électrons. Il est donc naturel de s'interroger sur la faisabilité d'une logique à plusieurs états. La *figure 11.6* décrit deux fonctions logiques standard relatives à une logique à quatre états.

x	***NON*** **x**
0	1
1/3	2/3
2/3	1/3
1	0

x	y	x ***ET*** y
0	0	0
0	1/3	0
0	2/3	0
0	1	0
1/3	0	0
1/3	1/3	1/3
1/3	2/3	1/3
1/3	1	1/3
2/3	0	0
2/3	1/3	1/3
2/3	2/3	2/3
2/3	1	2/3
1	0	0
1	1/3	1/3
1	2/3	2/3
1	1	1

Figure 11.6 – Logique à quatre états.

Les fonctions *NON*, *ET* et *OU* sont définies comme suit :

$$NON\, x = 1 - x$$
$$x\, ET\, y = \min(x, y)$$
$$x\, OU\, y = \max(x, y)$$

On peut également représenter la fonction *OU* comme il est indiqué dans la *figure 11.7*.

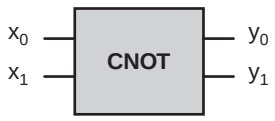
x	y	x OU y
0	0	0
0	1/3	1/3
0	2/3	2/3
0	1	1
1/3	0	1/3
1/3	1/3	1/3
1/3	2/3	2/3
1/3	1	1
2/3	0	2/3
2/3	1/3	2/3
2/3	2/3	2/3
2/3	1	1
1	0	1
1	1/3	1
1	2/3	1
1	1	1

Figure 11.7 – Fonction OU dans une logique à quatre états.

Une technologie adaptée à la logique multivaluée est celle des mémoires non volatiles à base de nano-dots capables de stocker quatre états de charge différents.

11.2.2 *La logique réversible*

Les considérations théoriques relatives à la dissipation de puissance ont amené l'idée suivante : réaliser des systèmes logiques avec des portes de base ayant autant de sorties que d'entrées. Plus précisément, ces portes permettent de retrouver les valeurs des entrées à partir des valeurs des sorties. Cela n'est évidemment pas possible avec une porte OU ou une porte ET classique.



	00	01	10	11
00	1	0	0	0
01	0	1	0	0
10	0	0	0	1
11	0	0	1	0

Figure 11.8 – Porte CNOT en logique réversible.

La *figure 11.8* représente la porte CNOT introduite par Benett. La table de vérité est représentée par la matrice de transition entre les états d'entrée et les états de sortie. Le coefficient de la matrice est égal à 1 quand un couple donné en entrée donne le couple indiqué en sortie.

Le NOT est identique à l'inverseur. Le C-NOT est représenté également *figure 11.9*.

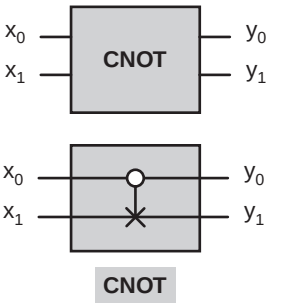


Figure 11.9 – Représentation du C-NOT.

La croix indique l'opération NOT mais cette opération est contrôlée par le signal indiqué par un cercle. Si l'entrée correspondant au cercle est « 1 », alors l'entrée correspondant à la croix est inversée. Dans le cas contraire, elle est transmise sans changement. La sortie correspondant au cercle est identique à l'entrée.

On peut montrer qu'il est impossible de synthétiser une logique complète avec uniquement des NOT et des CNOT. Les concepteurs de la logique réversible ont donc introduit une porte supplémentaire, le CC-NOT. Le fonctionnement du CC-NOT est décrit par les *figures 11.10* et *11.11*.

Les deux lignes de contrôle doivent être à l'état « 1 » pour que l'état correspondant à la croix change. Dans le cas contraire, l'état correspondant à la croix est transmis sans changement. Les états de contrôle sont transmis sans modification. Il est également intéressant de constater que la matrice de transition est formée d'une matrice identité associée à une matrice d'inversion et à deux matrices nulles.

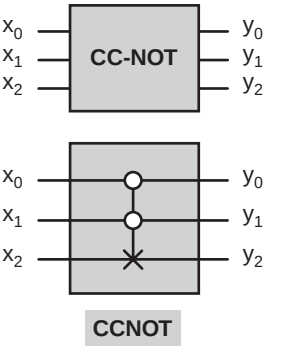
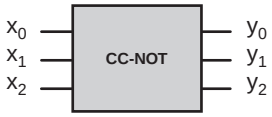


Figure 11.10 – Représentation du CC-NOT.



$x_0 x_1 x_2$ \ $y_0 y_1 y_2$	000	001	010	011	100	101	110	111
000	1	0	0	0	0	0	0	0
001	0	1	0	0	0	0	0	0
010	0	0	1	0	0	0	0	0
011	0	0	0	1	0	0	0	0
100	0	0	0	0	1	0	0	0
101	0	0	0	0	0	1	0	0
110	0	0	0	0	0	0	0	1
111	0	0	0	0	0	0	1	0

Figure 11.11 – Fonctionnement du CC-NOT.

La porte CC-NOT est très puissante. On peut facilement, en choisissant une entrée, la transformer en un NOT ou en un C-NOT. De plus, la sortie du C-NOT correspondant à la croix peut s'interpréter facilement comme un OU exclusif des deux entrées. En résumé, il est possible de réaliser n'importe quelle fonction logique avec des CC-NOT.

Un autre acteur de la logique réversible, Fredkin, a introduit une porte réversible appelée porte de Fredkin. La porte de Fredkin fonctionne comme le CC-NOT mais introduit une permutation entre les entrées et les sorties comme l'indique la *figure 11.12*. Cette porte a la propriété suivante : le nombre de « 1 » et de « 0 » ne change pas. Cette propriété est utilisée en informatique quantique. Il est également possible, comme pour le CC-NOT, de réaliser n'importe quelle fonction logique avec des portes de Fredkin.

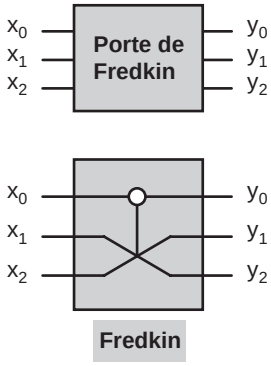


Figure 11.12 – Porte de Fredkin.

11.2.3 *La logique neuromorphique*

Inspirée par le fonctionnement des cellules nerveuses des organismes vivants, la logique neuromorphique propose une alternative à la logique booléenne. Ces concepts introduits en 1943 ont conduit à des applications relativement limitées jusqu'à ce jour. Ils suscitent cependant un intérêt croissant pour la mise en œuvre des nanotechnologies à cause d'une propriété remarquable des réseaux de neurones. Une logique neuromorphique tolère un taux de défauts relativement élevé. La fabrication des nanocomposants telle qu'elle est imaginée aujourd'hui conduit à prévoir des taux de défauts importants. L'intérêt suscité par les réseaux de neurones est donc assez naturel.

Le fonctionnement des neurones biologiques est à l'origine du concept de neurone formel. Il est intéressant de donner quelques éléments explicatifs sur le sujet.

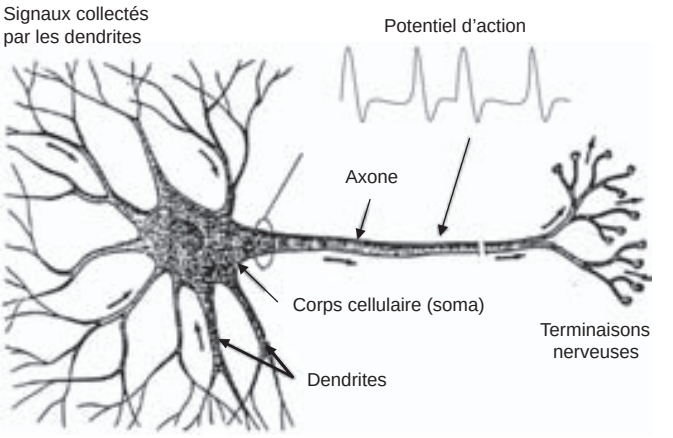


Figure 11.13 – Fonctionnement d'un neurone biologique.

Les neurones sont composés d'un corps cellulaire et d'une grande quantité d'excroissances appelées dendrites. L'une d'elle, l'axone, est longue et permet la communication à longue distance. Certains axones peuvent mesurer un mètre de long. Au repos, il y a une différence de potentiel d'environ 80 mV entre l'extérieur et l'intérieur du neurone. Les neurones sont fortement interconnectés entre eux. Un neurone peut être directement en contact avec des milliers d'autres.

Les échanges d'information se font par variation de ce potentiel. Quand les dendrites d'un neurone reçoivent des signaux en provenance d'autres neurones ou de centres sensitifs (rétine par exemple), ils peuvent réagir de deux manières : soit transmettre le signal sans amplification soit, à partir d'un certain seuil d'excitation, déclencher un « potentiel d'action ». La différence de potentiel entre l'intérieur et l'extérieur devient alors nulle suite à des échanges importants d'ions à travers la paroi du neurone. Le signal de sortie se présente alors comme un train d'impulsions qui se propagent dans l'axone vers les autres neurones du réseau. Les amplitudes de ces impulsions sont plus ou moins constantes (autour de 100 mV) mais leur nombre varie ainsi que les intervalles temporels, en fonction des entrées. Les vitesses de propagation sont faibles, une dizaine de mètres par seconde.

Ce principe de fonctionnement est plus ou moins repris par le neurone formel, comme le montre la *figure 11.14*.

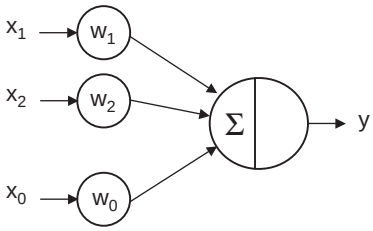


Figure 11.14 – Neurone formel.

La dynamique du réseau est définie de la manière suivante. La sortie $y(t + \Delta t)$ est une fonction non linéaire de la somme pondérée des entrées, par exemple la fonction tangente hyperbolique.

$$y(t + \Delta t) = \tanh(\beta\, h(t))$$

avec,

$$h(t) = \sum_i w_i\, x_i(t)$$

Quand le paramètre β est largement supérieur à un, la fonction $\tanh(\beta h)$ tend vers la fonction échelon $\gamma(h)$, nulle pour h négatif et égale à 1 pour h positif. Dans ce cas, la sortie y passe à 1 à l'instant $t + \Delta t$ si la somme pondérée des entrées est positive. Ce critère peut être remplacé par la condition suivante : la somme pondérée des entrées est supérieure à un seuil donné. La sortie d'un neurone est ensuite considérée comme une des entrées d'un autre neurone.

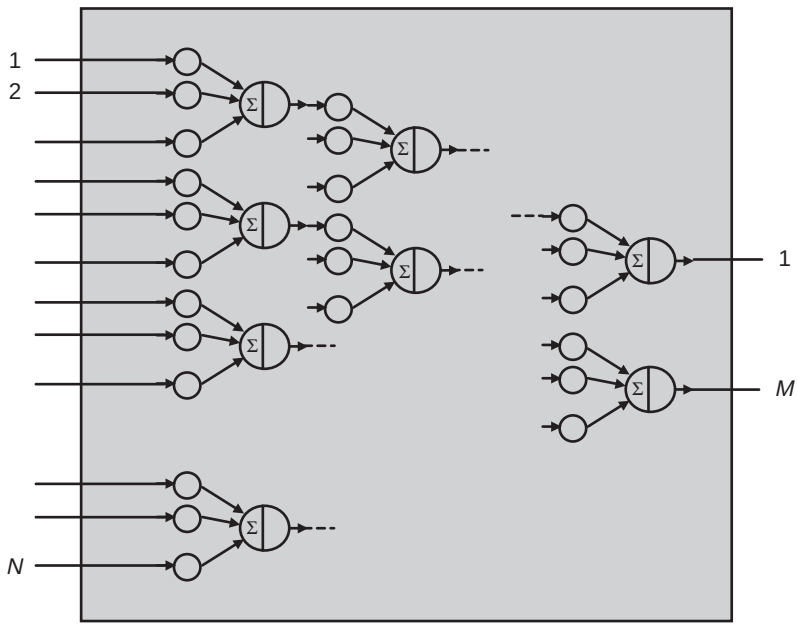


Figure 11.15 – Réseau de neurones formels.

Un réseau de neurones peut servir dans une opération de classification. Imaginons un réseau ayant *N* entrées et *M* sorties. Le nombre de neurones du système ainsi que leur connectivité est laissé au choix du concepteur. Un réseau est dit complet quand toutes les sorties sont reliées à toutes les entrées.

Les entrées sont par exemple les valeurs binaires des pixels d'une image et les sorties sont les caractères d'un alphabet. Le réseau de neurones peut alors « reconnaître » l'image d'un caractère. Encore faut-il avoir déterminé tous les coefficients de pondération du réseau et éventuellement les seuils. Cette opération est appelée l'apprentissage du réseau. En pratique, il faut présenter un certain nombre d'entrées au réseau puis optimiser le choix des coefficients pour que la réponse de sortie soit la plus proche possible de la réponse souhaitée.

Une propriété remarquable des réseaux de neurones est leur tolérance aux défauts. Les coefficients de pondération étant trouvés, le réseau fonctionnera de manière satisfaisante même si un certain nombre de ses neurones ne sont pas opérationnels. La proportion de neurones défectueux peut atteindre 10 % dans certains cas. Il n'est donc pas étonnant que ce type d'architecture inspire les concepteurs de dispositifs à base de nanocomposants. Les réalisations sont encore au stade de la recherche.

11.3 Conditions pour faire un système logique

Un certain nombre de conditions sont nécessaires pour réaliser un système logique à partir de composants élémentaires. Le principe d'assemblage est de les cascader pour synthétiser une fonction donnée.

11.3.1 *Non linéarité de la fonction de transfert*

Si on examine la fonction de transfert d'une porte logique, par exemple un inverseur, on constate que la caractéristique est fortement non linéaire. Que se passerait-il si la fonction de transfert était linéaire ?

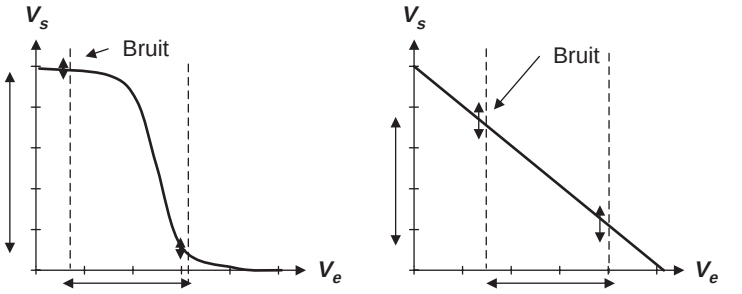


Figure 11.16 – Contrainte de non linéarité.

Il faut comprendre que cascader des dispositifs fonctionnant avec la même tension d'alimentation impose des contraintes fortes. Si le dispositif était linéaire, la pente serait nécessairement égale à un pour que l'excursion logique soit conservée. Imaginons alors une série de *N* dispositifs équivalents. Chaque étage ajoute du bruit alors que le signal est maintenu à sa valeur. Le rapport signal sur bruit se dégrade donc après la traversée d'un nombre important de portes.

Dans le cas d'un système non linéaire, la situation est différente car le dispositif a un gain en tension dans la partie centrale. Le bruit est en quelque sorte écrêté et le rapport signal sur bruit ne se dégrade pas. Remarquons que cette condition est également respectée dans les réseaux de neurones biologiques puisque le déclenchement du potentiel d'action est fortement non linéaire.

11.3.2 *Gain en puissance*

La non linéarité ne suffit pas et il est nécessaire que le dispositif présente un gain en puissance. En effet, un dispositif logique commande en général un ou plusieurs autres dispositifs équivalents. Ces dispositifs ainsi que les connexions associées sont équivalents à une capacité C de valeur plus ou moins élevée. Le dispositif doit être capable de fournir le courant de charge de cette capacité.

$$i(t) = C\frac{\mathrm{d}V}{\mathrm{d}t}$$

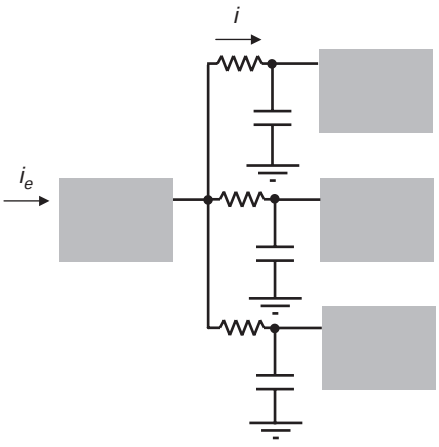


Figure 11.17 – Nécessité du gain en puissance.

Un dispositif doit être capable d'en commander N autres. Cette propriété est le « fan out » du dispositif. Si on considère N dispositifs identiques, le courant de charge est divisé par N. Ce courant est le courant fourni par la porte, produit du courant d'entrée par le gain. Si on veut conserver la vitesse de commutation il faut donc un gain en courant au moins égal à N.

Associée à la condition précédente, cette condition conduit à imposer que le dispositif ait un gain en puissance. Dans les neurones biologiques le potentiel d'action est régénéré au fur et à mesure qu'il se propage et le gain en puissance est également nécessaire.

11.3.3 *Concaténabilité*

Cette propriété veut simplement dire que les signaux de sortie d'un dispositif sont « compris » par l'entrée d'un autre dispositif cascadé avec le premier. Cette propriété semble évidente mais on pourrait imaginer des dispositifs logiques ayant des signaux électriques en entrée et optiques en sortie. Dans les réseaux de neurones, le potentiel d'action a pour effet de faire diffuser des molécules appelées neurotransmetteurs qui influencent les synapses d'entrée du neurone récepteur.

11.3.4 *Protection contre la rétro-propagation de l'information*

Dans un système de traitement de l'information, le signal porteur de l'information ne doit pas se propager avec la même probabilité vers la sortie ou vers l'entrée. Dans la logique CMOS, le signal se propage principalement des entrées vers les sorties et non pas des sorties vers les entrées.

Cette propriété est liée au caractère non symétrique du transistor MOS : l'entrée influe sur la sortie par l'action de la transconductance mais la sortie a peu d'effet sur l'entrée. Le signal de sortie appliqué en entrée n'est cependant pas nul. C'est la fraction de la tension transmise par la capacité parasite drain-grille comme le montre la *figure 11.18*.

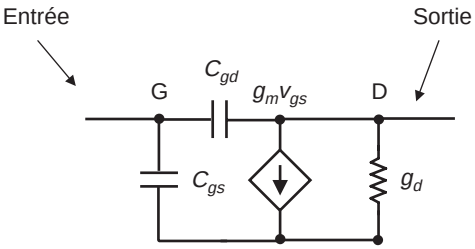


Figure 11.18 – La rétroaction dans un transistor.

Quand la fréquence augmente, la protection du transistor contre la rétroaction diminue puisque le condensateur drain-grille C_{gd} se présente de plus en plus comme un court-circuit.

Dans les neurones biologiques, le potentiel d'action ne peut se propager que dans une direction, du corps du neurone vers l'extrémité.

11.3.5 *Fournir un jeu complet d'opérateurs*

Dans une logique Booléenne, il suffit d'avoir à sa disposition des inverseurs et des portes ET pour pouvoir réaliser n'importe quelle fonction logique. Un inverseur et une porte OU sont également suffisants. Il suffit de disposer de NAND ou de NOR. Il est en effet facile de faire des inverseurs à partir de ces deux portes en fixant l'état de l'une des deux entrées. En revanche, avoir à sa disposition un OU et un ET ne suffit pas. Le CC-NOT suffit à lui seul pour créer une logique de même que la porte de Fredkin. Ces notions de base sont indispensables pour évaluer le potentiel des nouveaux dispositifs.

La logique neuronale utilise des opérateurs analogiques : multiplication par un coefficient, somme et comparaison à un seuil. Ces trois opérations suffisent pour créer un réseau complet.

11.4 Évolution des systèmes électroniques

L'objectif de ce paragraphe est d'identifier les évolutions majeures des systèmes électroniques en fonction des principes généraux étudiés dans les paragraphes précédents.

11.4.1 *Les architectures GAL*

Les architectures GAL sont une réponse aux difficultés posées par les interconnexions électriques. Comme il a été vu dans le paragraphe 11.1, le temps de propagation dans une connexion augmente au fur et à mesure que la technologie progresse. La *figure 11.19* illustre ce phénomène en montrant

sur un circuit intégré la zone accessible pendant une période d'horloge et cela pour différentes technologies. Deux effets sont pris en compte : l'augmentation de la fréquence de l'horloge et l'augmentation du produit *RC* de la ligne de transmission.

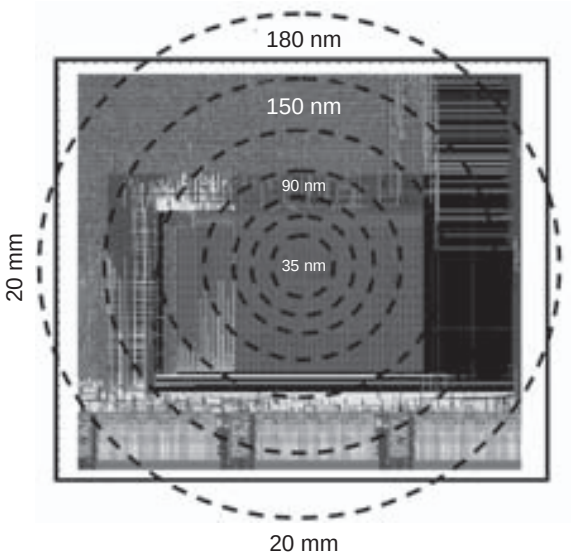


Figure 11.19 – Régions accessibles pendant une période d'horloge.

Ce schéma montre clairement qu'à partir du nœud 90 nm, il devient difficile d'atteindre pendant une période d'horloge toutes les régions d'un circuit intégré. Il est alors impossible de synchroniser en une période d'horloge tous les éléments logiques d'un circuit comme le veut la logique synchrone. Une solution envisagée est de découper le circuit en blocs fonctionnant chacun de manière synchrone avec leur propre horloge et de faire communiquer ces différents blocs par des protocoles asynchrones comme dans les réseaux de communication. Ces architectures sont appelées GAL : localement synchrones et globalement asynchrones.

11.4.2 *Réseau sur puce*

Ce deuxième concept est fortement lié aux considérations du paragraphe précédent. Les architectures classiques de circuit intégré sont bâties autour d'un ou de plusieurs bus transportant les données et les adresses des unités de calcul aux unités de mémoire. Cette architecture est devenue plus complexe au fil du temps avec l'apparition de mémoires cache de tailles différentes et de processeurs de calcul spécialisés. Le bus apparaît alors comme un goulot d'étranglement. De plus, il est avantageux de remplacer une mémoire de grande taille par des mémoires de plus petites tailles pour des raisons liées à la vitesse et à la consommation.

Les architectures de circuit évoluent donc vers des systèmes plus complexes dans lesquels blocs de calcul et blocs mémoires communiquent de manière plus souple selon un schéma indiqué *figure 11.20*.

Il ne faudrait pas penser que le réseau de communication est une simple ligne. Il est formé d'un ensemble de connexions en parallèle transportant par exemple les 32 ou 64 bits de données. Les

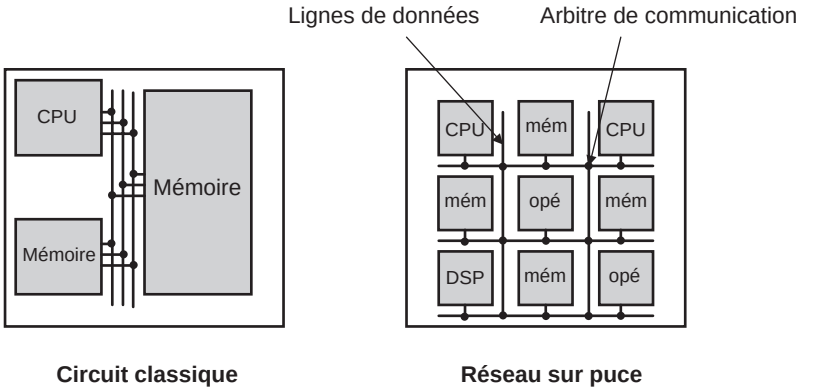


Figure 11.20 – Réseaux sur puce.

arbitres de bus sont des circuits capables de guider convenablement les données d'un point à un autre en évitant la congestion du réseau. Ce mode de communication est dans son principe assez semblable à l'acheminement de données dans un réseau comme Internet.

En pratique, les circuits de type réseau sur puce peuvent comporter une centaine de blocs. Aller au-delà n'est pas d'un grand intérêt car la surface occupée par le réseau et les arbitres de communication ne doit pas représenter plus de 10 % de la surface totale de la puce.

11.4.3 *Architectures reconfigurables*

Les architectures reconfigurables ont été imaginées pour résoudre un problème crucial de la microélectronique à savoir le coût des masques. Des coûts aussi élevés que 4 millions d'euros sont prévus pour les technologies futures. À ce prix, il n'est pas rentable de produire des quantités de circuits inférieures à plusieurs millions de pièces par an. Il faut donc concevoir des circuits intégrés adressant une très large classe d'applications. C'est le cas des microprocesseurs à usage général mais ce n'est pas le cas des circuits destinés à des applications particulières comme la téléphonie, le multimédia ou l'automobile. Le concept de plate-forme est donc apparu.

C'est un circuit dédié à une application mais présentant une flexibilité suffisante pour satisfaire à plusieurs standards ou pour anticiper des variations de normes. Une grande partie de la flexibilité vient de la programmation des processeurs et des processeurs de signaux inclus dans la puce. Cette technique n'est cependant pas suffisante quand il faut optimiser la surface de silicium, en particulier pour réduire la consommation électrique. Elle est également insuffisante quand les temps de traitement doivent être réduits et sont incompatibles avec le temps de traitement d'un processeur à usage général. Cela explique l'intérêt d'interconnecter à la demande un certain nombre de portes ou d'opérateurs pour réaliser une fonction spécifique dans le circuit intégré. C'est la même idée qui a conduit à développer la technologie des FPGA. Les circuits reconfigurables sont cependant différents.

Les FPGA sont des réseaux de portes logiques élémentaires qu'il est possible d'interconnecter par programmation électrique. Il est difficile d'intégrer des FPGA de grande taille dans un circuit intégré car l'extrême souplesse de cette technologie (toutes les fonctions logiques peuvent être synthétisées) se paye par une augmentation importante de la surface de silicium (un facteur 10 par rapport à une solution dédiée) et par une augmentation de la consommation électrique. Les technologies reconfigurables cherchent donc des solutions intermédiaires entre le FPGA et le circuit dédié. Une des techniques consiste à interconnecter à la demande des blocs de calcul élémentaires.

11.4.4 *Architectures tolérantes aux fautes*

Si on pense à l'intégration des nouveaux composants nanométriques (nanofils, nanotubes, molécules), l'utilisation d'architectures tolérantes aux fautes est une nécessité étant donné le taux de défaillance attendu pour ces composants et les difficultés prévues pour les interconnecter. Cette règle pourrait également s'appliquer aux systèmes à plusieurs milliards de transistors envisagés avec des MOSFET de très faibles tailles. Les principales techniques envisagées sont : la redondance, le multiplexage et la reconfiguration.

La technique de redondance consiste à dupliquer *R* fois la fonction puis à prendre une décision majoritaire. Le cas *R* égal à trois est le plus largement pratiqué. On peut grouper de manière redondante des modules redondants et augmenter encore la fiabilité comme le montre la *figure 11.21*.

Le vote majoritaire effectué sur *R* éléments consiste à comparer les résultats des *R* éléments et à choisir celui obtenu le plus souvent. Quand le résultat de l'opération est donné sur *B* bits, il faut

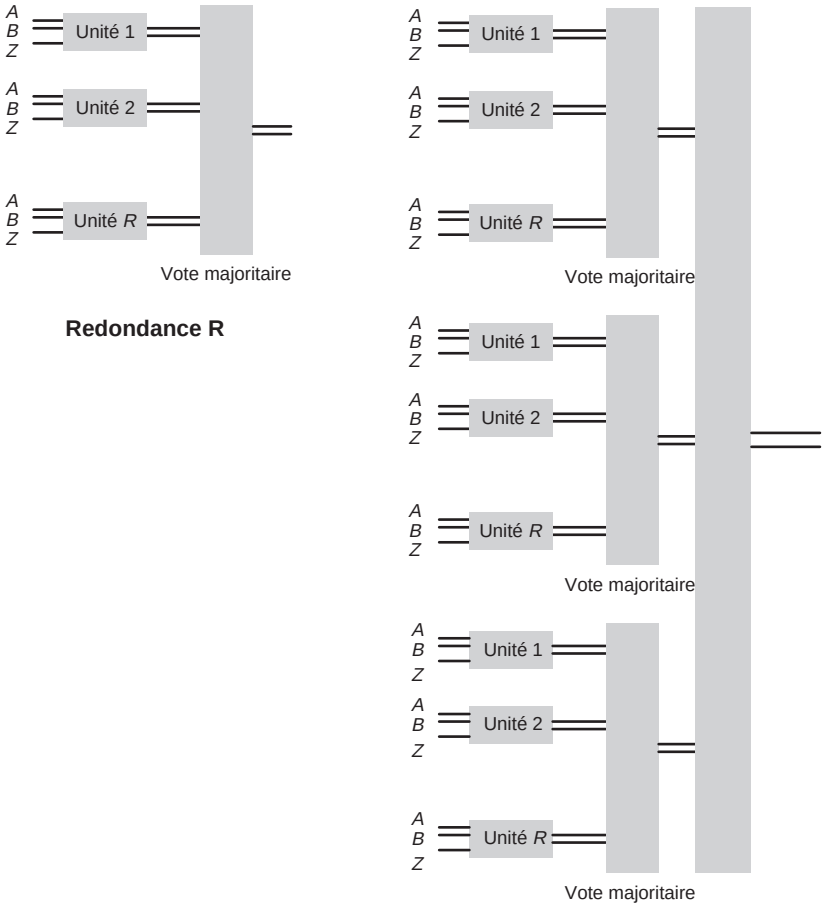


Figure 11.21 – Systèmes redondants.

effectuer la comparaison sur toutes les configurations possibles de ces B bits. On comprend facilement que la surface de silicium augmente considérablement par l'usage de cette technique.

Si le système global comporte N éléments (des transistors par exemple), il est divisé en R blocs logiques comprenant chacun N_c éléments. On appelle p la probabilité de défaillance d'un élément, supposée très inférieure à l'unité. La probabilité qu'un bloc fonctionne est :

$$P_{\text{bloc}} = (1-p)^{N_C}$$

soit si p est très inférieur à 1,

$$P_{\text{bloc}} \approx \exp^{-N_C p}$$

La probabilité qu'un module soit défaillant est alors :

$$P_{\text{def bloc}} \approx 1 - \exp^{-N_C p} \approx N_C p$$

Les dispositifs de vote sont supposés comporter mB composants, formule dans laquelle B est le nombre de sorties par bloc. Un groupe logique comporte R blocs identiques de N_C éléments et un bloc de vote de mB éléments. Il fonctionne si au moins $(R+1)/2$ blocs fonctionnent et si le module de vote fonctionne. Le nombre de groupes est $N/(RN_C + mB)$. La probabilité de défaillance de la puce P_{puce} est la somme des probabilités des défaillances des blocs.

$$P_{\text{puce}} = \frac{N}{RN_C + mB}\left[C(R)(N_C p)^{\frac{(R+1)}{2}} + mBp\right] \tag{11.5}$$

Dans cette formule, C est le nombre de cas possibles pour lesquels $(R+1)/2$ blocs sont défaillants parmi les R possibles.

$$C(R) = \frac{R!}{\left(\frac{R-1}{2}\right)!\left(\frac{R+1}{2}\right)!}$$

Il est possible de déterminer le nombre N_C optimal en dérivant la probabilité P_{puce}.

Les résultats sont groupés dans le *tableau 11.1*.

Tableau 11.1 – Optimisation de la redondance.

R	N_C optimal
1	N
3	$\left(\frac{mB}{3}\right)^{1/2} p^{1/2}$
5	$\left(\frac{mB}{20}\right)^{1/3} p^{2/3}$
7	$\left(\frac{mB}{105}\right)^{1/4} p^{3/4}$
infini	$\frac{1}{p}$

Une autre technique a été introduite il y a cinquante ans par le mathématicien Von Neumann, c'est la technique du multiplexage. Elle est proche des techniques de redondance. Le circuit de vote majoritaire est remplacé par un circuit générant les différents résultats possibles.

Une troisième technique fait usage de la reconfiguration. La logique est groupée en blocs qu'il est possible d'interconnecter à la demande. Les blocs défectueux sont éliminés après tests.

Ces trois techniques sont comparées pour une puce hypothétique comportant 10^{12} éléments selon les résultats donnés en référence **[12]**. Le graphique indique le niveau de redondance nécessaire, c'est-à-dire le facteur d'augmentation de la surface, en fonction du taux de défaillance d'un élément et cela pour obtenir un taux de fiabilité globale de 90 %.

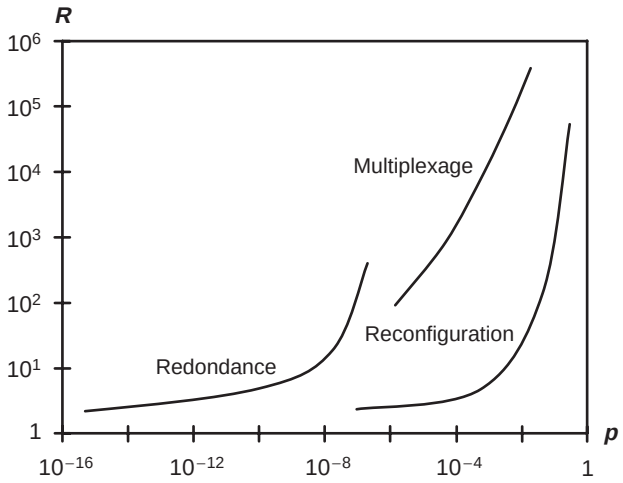


Figure 11.22 – Comparaison des solutions.

La reconfiguration offre les perspectives les plus intéressantes en conservant une augmentation de surface modérée. Elle a cependant l'inconvénient d'imposer au système des temps de test et de reconfiguration relativement importants. Cela peut être très pénalisant en particulier en présence de défaillances dynamiques c'est-à-dire survenant à des instants aléatoires pendant le fonctionnement de la puce.

11.5 L'informatique quantique

Inspiré par les considérations relatives à la dissipation de puissance dans un calcul, l'ordinateur quantique a pour but d'effectuer des calculs beaucoup plus rapidement qu'un ordinateur classique en tirant parti de propriétés strictement quantiques comme la superposition des états et l'enchevêtrement.

La représentation des portes logiques, donnée dans le chapitre 2, montre que l'on passe d'un état à un autre par une matrice. Si on associe l'état d'un bit à l'état d'un système physique prenant deux valeurs possibles, par exemple un atome ayant deux états, la matrice représentant la table de vérité de la fonction correspond à la matrice faisant passer l'atome d'un état à un autre.

Un calculateur classique prend un état d'entrée représenté par exemple par une série (0,1,0,1) et le transforme en un état de sortie, par exemple la somme binaire des deux mots de deux bits contenus dans l'entrée. Un calculateur quantique prend un état d'entrée et le fait évoluer vers un état de sortie. Le passage entre les deux états est déterminé par l'équation de Schrödinger.

L'informatique traditionnelle manipule des variables logiques prenant deux valeurs possibles « 0 » ou « 1 » alors que l'informatique quantique manipule des « qubits » représentés comme une combinaison linéaire de deux états possibles.

$$|x\rangle = a|0\rangle + b|1\rangle$$

Les deux états $|0\rangle$ et $|1\rangle$ pourront par exemple être deux états possibles d'un atome.

Un registre quantique est un ensemble de n qubits. Un état de ce registre est noté :

$$|y\rangle = \sum_{x=0}^{x=2^n-1} c_x|x\rangle$$

Par exemple, un registre de 2 qubits s'écrit :

$$|y\rangle = c_0|00\rangle + c_1|01\rangle + c_2|10\rangle + c_4|11\rangle$$

Un algorithme quantique est une succession d'opérations appliquées à partir d'un état initial. Si on pense maintenant aux états physiques, ils évoluent selon l'équation de Schrödinger et donc en fonction de l'opérateur hamiltonien. Celui-ci sera choisi de telle manière qu'il génère le calcul souhaité. À la fin des opérations, il est nécessaire de choisir parmi toutes les valeurs possibles le résultat du calcul.

La puissance du calcul quantique vient du fait que le nombre inscrit dans un registre peut prendre plusieurs valeurs à la fois ce qui introduit un parallélisme important dans le calcul.

Cette propriété est liée à la décomposition de l'état en une somme linéaire. Pour que cette propriété se conserve, il faut que le système interagisse peu avec l'extérieur ce qui est difficile à obtenir en pratique.

Quelques algorithmes seulement se prêtent à un traitement quantique. Ce sont la décomposition d'un nombre en produit de nombres premiers, le calcul des transformées de Fourier, la recherche dans des bases de données désordonnées et la simulation de systèmes physiques.

La généralisation de cette méthode pour traiter l'ensemble des problèmes soumis aux calculateurs électroniques n'est donc pas évidente d'autant plus que la réalisation matérielle des systèmes capables de supporter ce type de calcul est également difficile.

Chapitre 12

Les nouveaux composants nanométriques

Il s'agit dans ce chapitre de présenter de nouveaux composants, différents du transistor MOS ou du transistor bipolaire. Ces composants sont pour la plupart des objets de laboratoire mais pourraient à terme suppléer ou même remplacer les composants actuels. Les trois objectifs qui guident les nouveaux développements sont les suivants :

- s'affranchir d'une lithographie extrême beaucoup trop coûteuse en imaginant de nouveaux procédés de fabrication collective ;
- réduire au maximum la taille du composant en allant encore en dessous de la taille des transistors les plus miniaturisés ;
- réduire la consommation d'énergie à chaque changement d'état logique.

Ces objectifs sont cependant très liés aux considérations architecturales comme il sera vu dans le dernier paragraphe.

Les physiciens et les technologues se sont donc intéressés à un certain nombre de dispositifs qu'il est possible de classer comme suit :

- les nanotubes de carbone ;
- les nanofils semiconducteurs ;
- les dispositifs à peu d'électrons ;
- les molécules fonctionnalisables.

12.1 Les nanotubes

12.1.1 *La structure des nanotubes*

Ce sont actuellement les composants qui font l'objet de la recherche la plus intense. Une des raisons est sans doute la relative facilité avec laquelle il est possible de les fabriquer. Ils ont été découverts en 1991 par Sumio Iijma dans une expérience dont le but était d'étudier les filaments de carbone. Ce sont des plans de graphène qui s'enroulent sous forme de cylindre comportant une ou plusieurs feuilles.

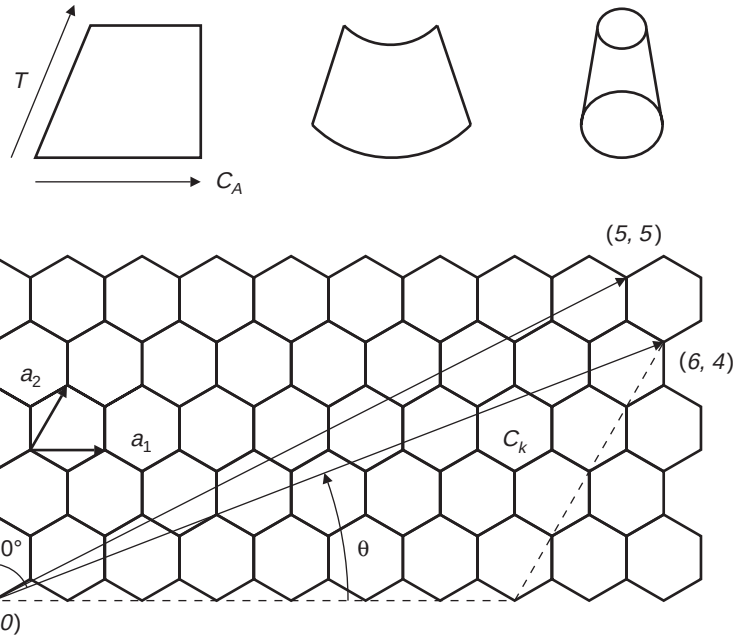


Figure 12.1 – Formation des nanotubes.

Ces objets ont la propriété remarquable de présenter un rapport entre la longueur et le diamètre exceptionnellement élevé puisque le diamètre est compris entre 1 et 10 nm et puisque la longueur peut atteindre des dizaines de microns. La *figure 12.2* représente deux types possibles de nanotubes monoparoi.

La conduction dans ces dispositifs est totalement différente de celle habituellement observée dans les systèmes de plus grande taille. Il est donc nécessaire de donner quelques principes de base. Le chapitre 2 a permis de comprendre comment les densités d'états énergétiques pouvaient se former dans des structures à deux, à une et à zéro dimension. Il faut maintenant partir de ces résultats pour en déduire le courant électrique traversant de tels dispositifs.

Le résultat sera très différent de la loi d'ohm, uniquement valable quand les électrons subissent de nombreuses collisions dans le dispositif. Dans les dispositifs de taille nanométrique (au moins dans une dimension), la distance entre deux collisions peut être grande devant la dimension du dispositif. De manière plus précise, les systèmes de petite taille peuvent être caractérisés par trois longueurs :

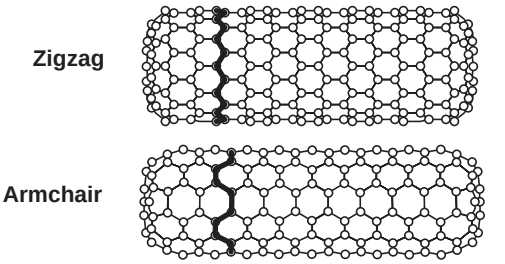


Figure 12.2 – Les nanotubes de carbone.

- La longueur d'onde de Fermi est la longueur d'onde correspondant à l'énergie de Fermi des électrons soit :

$$\lambda_F = \frac{2\pi}{k_F} \quad \text{et} \quad E_F = \frac{\hbar^2 k_F^2}{2 m^*}$$

- Elle vaut 10 nm pour les semi-conducteurs et environ 1 nm pour les métaux.
- Le libre parcours moyen est la distance entre deux collisions. Il est environ de 1 µm pour des semi-conducteurs de bonne qualité.
- La longueur de cohérence de la phase est la longueur que parcourt l'électron avant que sa phase initiale ne soit modifiée. Elle est plus grande que le libre parcours moyen.

Quand les dispositifs ont des dimensions plus faibles que le libre parcours moyen, le régime est dit balistique et la loi d'ohm ne s'applique plus puisqu'elle est due aux collisions. Quand les dimensions du dispositif sont faibles devant la longueur de cohérence, des interférences entre les différentes trajectoires possibles peuvent avoir lieu et produire des effets de nature purement quantique. Les collisions inélastiques de l'électron avec le milieu détruisent la cohérence de phase et le comportement classique de la conduction est à nouveau vérifié. Pour expliquer la conduction dans les systèmes nanométriques, il faut faire appel à la théorie de la conduction due à Landauer. Cette théorie est assez difficile et nous ne donnerons ici qu'une interprétation très simplifiée.

12.1.2 *Une nouvelle théorie de la conduction*

On considère deux réservoirs d'électrons de grande taille, caractérisés par leurs potentiels chimiques, en liaison par un système unidimensionnel comme le montre la *figure 12.3*. Dans ce système unidimensionnel, les fonctions d'onde des électrons se propagent et on définit un canal de propagation pour un mode de propagation donnée.

Ce modèle peut être comparé à un mode de propagation dans un guide d'onde. Le fil unidimensionnel est traité comme un milieu de propagation idéal présentant un centre de diffusion pour la fonction d'onde de l'électron. On suppose de plus que la fonction d'onde se transmet en partie et se réfléchit en partie avec des paramètres T et R qui expriment respectivement le coefficient de transmission et le coefficient de réflexion.

Dans un premier temps, on considère un seul mode de propagation. Le vecteur d'onde de la fonction d'onde est quantifié quand on écrit les conditions périodiques dans la dimension transverse. Si A est la section du fil, il y a environ A/λ_F^2 valeurs possibles du vecteur d'onde.

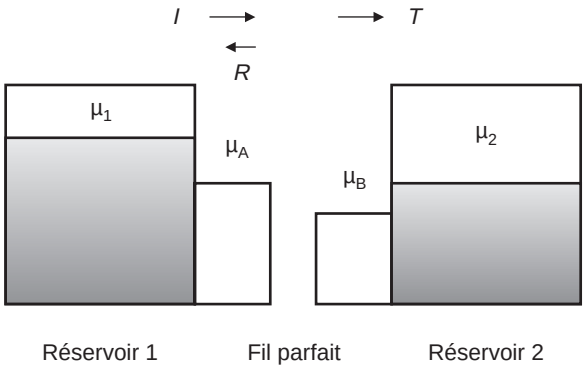


Figure 12.3 – Modèle de Landauer de la conduction.

Le paramètre λ_F est la longueur d'onde d'un électron dont l'énergie est celle de Fermi. Un mode de propagation, ou un canal, correspond à une valeur particulière du vecteur d'onde.

Pour les énergies en dessous de μ_2, il y a autant d'électrons qui passent de gauche à droite que l'inverse. Le courant est donc nul. Le courant est donc dû aux n électrons d'énergies entre μ_2 et μ_1. La valeur de ce courant, si v_F est la vitesse au niveau de l'énergie de Fermi, est alors :

$$I = e\, v_F\, n T$$

Le nombre d'électrons impliqués n est :

$$n = 2(\mu_1 - \mu_2)\frac{\mathrm{d}n}{\mathrm{d}E}$$

Pour un système à une dimension, le chapitre 2 a permis d'établir que :

$$\frac{\mathrm{d}n}{\mathrm{d}E} = \frac{2}{h v_F}$$

On obtient donc :

$$I = \frac{2e}{h} T(\mu_1 - \mu_2)$$

Si la tension entre les deux réservoirs est V, la relation classique établie dans le chapitre 2 s'applique :

$$eV = \mu_1 - \mu_2$$

et donc,

$$I = \frac{2\,e^2}{h} TV \qquad (12.1)$$

Il faut noter que même pour un fil parfait, transmission totale et T égal à un, la résistance n'est pas nulle. La valeur la plus faible de la résistance ou quantum de résistance est donc :

$$R = \frac{h}{2\,e^2}$$

La valeur de ce quantum de résistance est 12,9 kΩ. C'est la valeur minimale de la résistance d'un tel dispositif. Cette propriété quantique est bien différente de ce que notre intuition pourrait prévoir.

Si maintenant nous mesurons la différence de potentiel V' au niveau des fils on peut écrire :

$$V' = e\,(\mu_A - \mu_B)$$

Les potentiels μ_A et μ_B sont les potentiels chimiques des électrons dans le fil de part et d'autre de la barrière. En raisonnant sur les échanges d'électrons entre réservoirs et fils, on obtient :

$$I = \frac{2e^2}{h}\frac{T}{R}V' \qquad (12.2)$$

Cette fois, quand le dispositif est parfait ($T = 1$ et $R = 0$), la résistance est bien nulle.

La formule de Landauer peut se généraliser à plusieurs canaux. On définit alors une matrice de transition entre les canaux du fil de gauche et ceux du fil de droite. Les coefficients T_{ij} expriment les amplitudes de probabilité pour qu'un électron dans un mode i à gauche soit transmis dans un mode j à droite. On montre alors que la conductance s'écrit :

$$G = \frac{I}{V} = \sum_i \frac{2e^2}{h} T_i$$

avec,

$$T_i = \sum_j T_{ij}$$

La mesure de la conductance fait donc apparaître une variation marquée par des sauts au fur et à mesure que les canaux s'ouvrent comme le montre la *figure 12.4*. Cette figure est un résultat expérimental illustrant les variations de la conductance entre deux gaz d'électrons situés à 250 nm, en fonction de la tension appliquée.

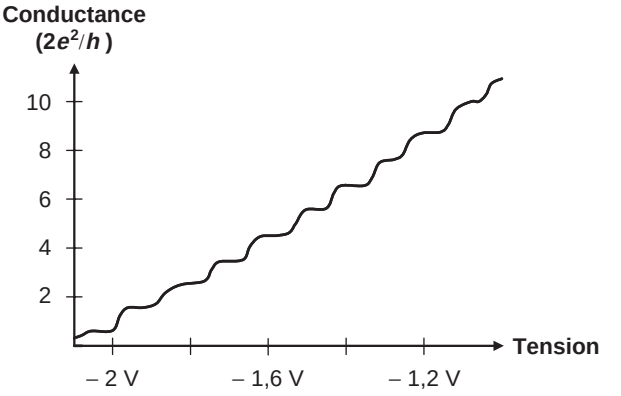


Figure 12.4 – Mise en évidence de la quantification de la conductance.

12.1.3 *Propriétés et fabrication des nanotubes*

Les propriétés des nanotubes découlent directement de leur filiation avec le graphite. De plus, des propriétés spécifiques sont apportées par l'enroulement et la faible dimension obtenue pour leur diamètre. En fonction de la direction de l'enroulement, on peut obtenir un comportement métallique ou semi-conducteur. Quand aucune précaution n'est prise, la résistance d'un nanotube en contact avec deux conducteurs métalliques est de l'ordre du MΩ, mais avec une évaporation appropriée on peut approcher la valeur théorique de 6,5 kΩ expliquée précédemment. Il faut aussi tenir compte du fait qu'il y a deux sous-bandes au niveau de Fermi.

De plus, les nanotubes sont capables de transporter des densités de courant considérables de l'ordre de 10^{10} A/cm^2 soit 100 fois plus que les métaux. Cette propriété justifie l'intérêt présenté par ces dispositifs pour réaliser des fils de connexion dans les circuits intégrés ou pour réaliser le canal de conduction d'un transistor. Dans un autre registre, les nanotubes présentent également des propriétés mécaniques exceptionnelles. Ils peuvent se courber et se tordre très facilement.

Il est possible de dire quelques mots sur les techniques de fabrication. Deux procédés sont utilisés : la synthèse à haute température et la synthèse à moyenne température.

À haute température, l'opération se fait en deux étapes : évaporation du graphite puis condensation dans une enceinte dans laquelle il y a de fortes variations de températures. Il faut dépasser 3 200 °C et pour cela on peut faire usage d'un arc électrique ou d'un laser de puissance. On obtient en général des nanotubes multiparois. Pour obtenir des nanotubes monoparoi, qui présentent des propriétés plus séduisantes, il faut ajouter un catalyseur métallique à la poudre de graphite.

Dans le procédé à moyenne température, un gaz carboné est décomposé entre 500 et 1 000 °C en contact avec des particules de catalyseur métallique. Le carbone libéré par la décomposition du gaz donne naissance au niveau des particules métalliques à des nanotubes. Cette voie semble plus compatible avec un procédé micro-électronique. La *figure 12.5* montre des nanotubes obtenus par ces techniques.

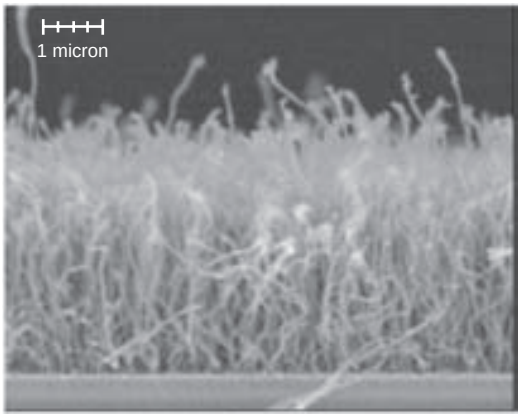


Figure 12.5 – Nanotubes obtenus par synthèse.

12.1.4 *Applications des nanotubes*

La première application envisagée est d'utiliser les nanotubes comme conducteurs dans les futurs circuits intégrés. Ils permettent en effet de transporter des courants avec des densités remarquables de 10^{10} A/cm^2 ce qui est largement supérieur aux meilleurs conducteurs actuels, le cuivre par exemple. Il est cependant nécessaire de savoir réaliser des contacts de faible résistance avec d'autres matériaux utilisés en micro-électronique. Quand aucune précaution n'est prise, des contacts de plusieurs MΩ sont obtenus. Il a été démontré cependant qu'il était possible de s'approcher de la résistance théorique de 6,5 kΩ.

En fonction des conditions de fabrication, des propriétés métalliques ou semi-conductrices peuvent être obtenues ce qui ouvre le champ à la fabrication de dispositifs actifs comme des diodes ou même des transistors. Les propriétés de conduction dépendent fondamentalement de la manière dont la feuille de graphène s'enroule sur elle-même pour former un tube et le contrôle de cette propriété à la fabrication reste un défi pour les années futures. Aujourd'hui la seule possibilité est de les trier après fabrication.

La forme tout à fait particulière des nanotubes laisse envisager des applications intéressantes en émission. Les nanotubes sont des émetteurs d'électrons à faible tension. En effet, le champ électrique intense à l'extrémité d'un nanotube peut favoriser l'émission d'électrons par effet de pointe.

Enfin, les nanotubes peuvent être utilisés comme éléments de conduction dans un transistor dont le canal de conduction serait alors remplacé par ces éléments cylindriques. La *figure 12.6* représente un tel dispositif.

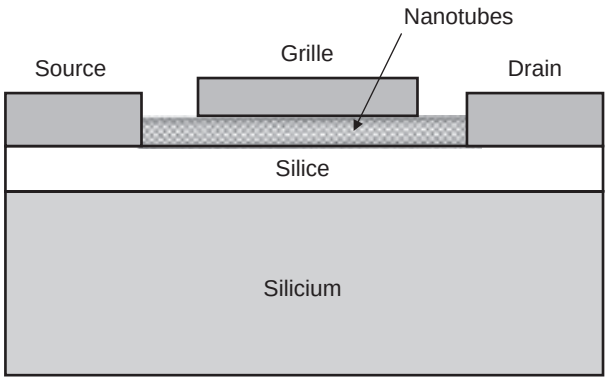


Figure 12.6 – Transistor à nanotubes.

Ce dispositif est très semblable à un transistor MOS classique. La seule différence est dans la constitution de la zone de canal. Différentes géométries ont été expérimentées en plaçant la grille soit dessus soit dessous. Le fonctionnement de ces dispositifs n'est pas encore parfaitement compris mais il est admis que l'effet transistor est dû à la modulation de la barrière de potentiel au niveau des contacts et non pas à la modulation de la conductance dans les nanotubes eux-mêmes. Les contacts entre les nanotubes et les électrodes de drain et de source doivent être considérés comme des contacts Schottky entre un métal et un semi-conducteur. Les performances électriques obtenues pour la transconductance sont remarquables, ce qui pousse l'industrie électronique à s'intéresser à des dispositifs avancés.

Le *tableau 12.1* compare ces nouveaux transistors à nanotubes à leurs homologues réalisés de manière plus classique.

Tableau 12.1

	CNT-FET 1 µm	CNT-FET 1,4 µm	CNT-FET 3 µm	MOSFET 100 nm	MOSFET 14 nm
Longueur de grille (nm)	1 030	1 400	2 000	130	14
Courant « on » (mA/µm)	0,700	3	3,5	1,4	0,4
Courant « off » (nA/µm)	7	7	1	3	100
Transconductance (mS/µm)	200	6 700	6 000	1 000	360
Pente sous le seuil (mV/décade)	700	80	70	90	70
Résistance « on » (Ω/µm)	1 400	360	340	2 600	4 200

Les propriétés des transistors à nanotubes sont donc très intéressantes mais leur fabrication doit être rendue compatible avec les techniques de fabrication collective de la micro-électronique.

12.2 Les nanofils

12.2.1 *Structure et fabrication des nanofils*

Ce sont des dispositifs nanométriques assez proches des nanotubes de carbone. Ils sont également candidats pour réaliser de nouveaux dispositifs dans le futur. On peut les définir comme des objets dont le rapport longueur sur largeur est supérieur à 10 avec une largeur de quelques dizaines de nanomètres. Leur facteur de forme est sans commune mesure avec celui des nanotubes de carbone. Ils sont en général fabriqués dans un matériau semi-conducteur ce qui laisse envisager une possible intégration à une technologie micro-électronique. Leur intérêt est dans leur faible dimension. Les méthodes de fabrication appartiennent à deux familles : les méthodes dites *top-down* empruntées à la micro-électronique et les méthodes dites *bottom-up* inspirées des procédés chimiques.

L'approche *top-down* fait usage de la lithographie extrême et doit être considérée comme une première approche pour fabriquer des nanofils. Ce n'est pas une méthode envisageable pour une production industrielle. Un objectif majeur est en effet de réaliser des dispositifs à base de fils en se passant le plus possible de la lithographie de très grande précision. Des techniques sophistiquées comme la microscopie en champ proche permettent également en manipulant les atomes de réaliser des motifs nanométriques. Il faut également citer la nano-impression qui, à l'image des procédés de l'imprimerie, permet de reproduire un motif fabriqué à l'aide d'une lithographie de haute précision. Ces procédés sont détaillés dans le chapitre 6 de cet ouvrage.

L'approche *bottom-up* a pour ambition d'être applicable dans le futur à l'échelle industrielle. L'objectif est alors de réaliser sur une surface un grand nombre de dispositifs identiques en utilisant des procédés chimiques. On parle d'auto-assemblage quand la surface présente une structure périodique régulière et peut servir de support à la croissance d'un matériau donné. La technique VLS (*Vapor Liquid Solid*) est souvent mise en œuvre. Le principe est de créer une goutte en fusion du semi-conducteur choisi dans un réacteur à haute température (1 000 degrés environ) puis de faire croître les fils en s'appuyant sur les irrégularités précédemment décrites.

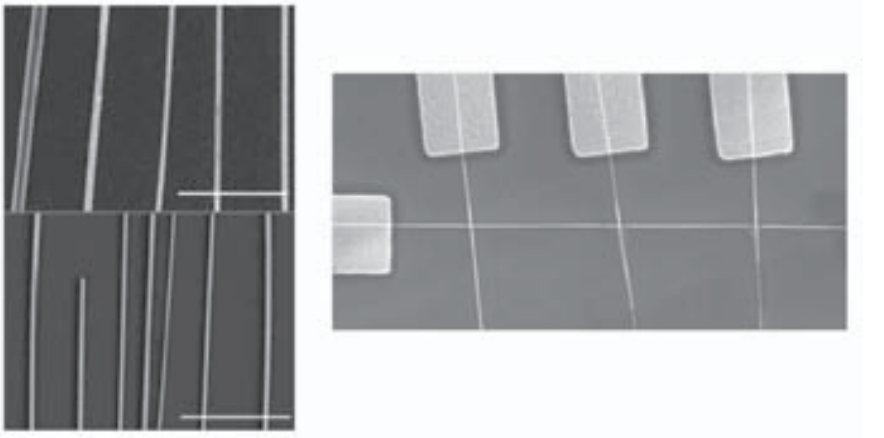


Figure 12.7 – Réseau de nanofils.

Les semi-conducteurs habituels (silicium, arséniure de gallium, phosphure d'indium…) sont expérimentés dans les laboratoires. Le dopage peut être contrôlé par la composition de la phase vapeur d'entrée. Des fils semi-conducteurs de quelques microns de long pour une dizaine de nm de marge sont ainsi obtenus. La *figure 12.7* montre des réalisations de laboratoire.

12.2.2 *Fonctions obtenues à base de nanofils*

Les résultats présentés sont en grande partie issus des travaux de Charles Lieber et de ses collaborateurs. Un nanofil seul n'a pas de fonctionnalité particulière mais quand on réalise un croisement de nanofils, de nombreuses possibilités sont offertes.

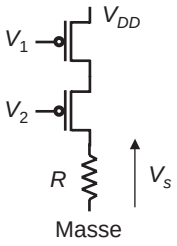
La première est de réaliser une simple connexion. Dans une version plus sophistiquée, il est possible de rendre cette connexion programmable par exemple en faisant jouer les forces électrostatiques capables de coller les deux fils quand les valeurs des tensions appliquées sont convenablement choisies. Si les deux fils sont dopés différemment alors le croisement des deux fils matérialise une jonction *pn*. Il est donc facile d'imaginer une logique à diodes utilisant de tels dispositifs.

Faisons maintenant croître un oxyde à la surface de l'un des fils de type semi-conducteur et appliquons un autre fil supposé conducteur et placé perpendiculairement. Le dispositif réalisé est de type transistor. Si le fil semi-conducteur est de type *p*, une tension positive appliquée sur le fil conducteur a pour effet de repousser les trous du fil de part et d'autre de la zone de croisement. La conduction est alors impossible dans le fil considéré et son comportement est analogue à celui d'un transistor bloqué.

Ces différents principes ont permis de réaliser les principales fonctions logiques de la micro-électronique. L'exemple du NOR est donné *figure 12.8* et illustre les travaux de Charles Lieber.

Les caractéristiques de cette nouvelle logique semblent séduisantes. Il faut cependant être conscient des difficultés de la fabrication collective de tels dispositifs et être attentifs aux difficultés de positionnement des fils entre eux.

Une autre application intéressante est le dopage des fils dans le sens de la longueur. Le paragraphe consacré aux nouvelles architectures montrera tout l'intérêt de réaliser des fils semi-conducteurs avec une variation du profil de dopage dans le sens de la longueur.



v_1 (V)	v_2 (V)	v_s (V)
0	0	4,85
5	0	0,02
0	5	0,06
5	5	0,01

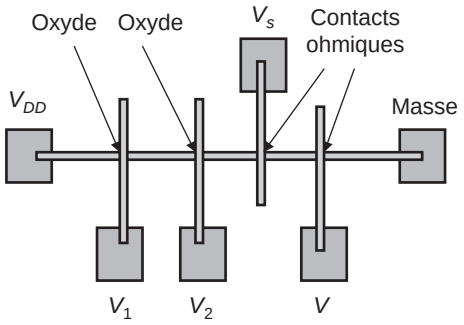


Figure 12.8 – Exemple de fonction logique à base de nanofils.

12.3 Les dispositifs à peu d'électrons

La recherche d'une énergie de commutation minimale amène à s'intéresser à des dispositifs manipulant de faibles nombres d'électrons et à la limite manipulant un électron unique. Il faut cependant garder à l'esprit qu'il reste nécessaire de faire communiquer ces composants entre eux et si possible à température ambiante.

12.3.1 *Fonctionnement de la boîte à un électron*

Dans une première étape, il est possible d'étudier le comportement d'une boîte à électron unique. Le système étudié est représenté *figure 12.9*. Il se compose d'une électrode de référence, d'une électrode de grille et d'un îlot appelé boîte quantique isolé électriquement des électrodes. Cet îlot est considéré comme un condensateur de très faible taille.

Les deux dispositifs sont modélisés sous forme de capacités. La capacité entre boîte et grille est classique mais celle entre boîte et électrode de référence est particulière car les dimensions sont si faibles qu'un effet tunnel peut avoir lieu. La charge de l'îlot est $Q_B - Q_G$ soit $-ne$. L'énergie électrostatique stockée dans l'îlot est :

$$W_C = \frac{1}{2}\frac{n^2e^2}{C_B}$$

La charge de la boîte varie de manière quantique si la différence d'énergie correspondant à une variation d'un électron stocké est largement supérieure à k_BT, énergie thermique moyenne d'un électron. Pour cela, il faut que la température soit basse ou que la capacité de la boîte soit très fai-

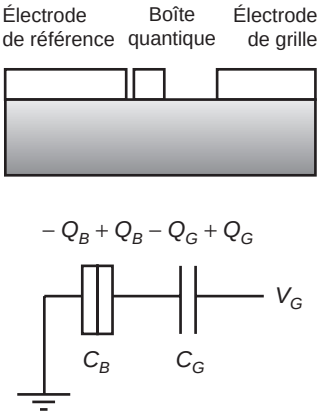


Figure 12.9 – La boîte quantique.

ble. En appliquant la formule du condensateur plan de surface A et d'épaisseur d, on en déduit la condition pour observer des variations discontinues de charge :

$$C_B = \varepsilon \frac{A}{d}$$

soit,

$$A \ll \frac{e^2 d}{2 \varepsilon k_B T}$$

Une application numérique à température ambiante pour un dispositif sur de la silice et de 3 nm d'épaisseur conduit à un carré de moins de 16 nm de côté.

Les deux équations du dispositif sont :

$$Q_B - Q_G = -ne \quad \text{et} \quad \frac{Q_G}{C_G} + \frac{Q_B}{C_B} = V_G$$

Il reste maintenant à écrire la condition pour que l'îlot conserve une charge de n électrons pour une valeur de la tension de grille donnée.

Pour cela, il faut revenir à des notions assez générales de thermodynamique et faire intervenir l'énergie libre. L'énergie libre est définie comme la différence entre l'énergie totale et le travail effectué par la source de potentiel pour placer n électrons dans l'îlot. La thermodynamique montre qu'un système se place dans l'état correspondant à son énergie libre minimale.

De plus, la théorie dite « orthodoxe » relie la probabilité de transition d'un état à n électrons à un état à $n + 1$ électrons à la variation d'énergie libre. L'énergie du système d'électrons sera supposée uniquement due à l'énergie électrostatique ce qui veut dire que l'énergie des électrons en $\hbar^2 k^2 / 2 m^\star$ est négligée, ce qui est légitime à basse température.

L'énergie libre s'écrit alors quand n électrons sont dans l'îlot.

$$F(n) = \frac{Q_G^2}{2 C_G} + \frac{Q_B^2}{2 C_B} - Q_G V_G$$

Les deux équations du dispositif permettent d'exprimer les charges en fonction du potentiel de grille et du nombre d'électrons.

$$F(n) = \frac{e^2 n^2}{2(C_B + C_G)} + \frac{C_B C_G V_G^2}{2(C_G + C_B)} - e n \frac{C_G V_G}{(C_G + C_B)} - \frac{C_G C_B V_G^2}{(C_G + C_B)}$$

L'état à n électrons est alors stable si :

$$F(n) < F(n+1)$$
$$F(n) < F(n-1)$$

On en déduit alors :

$$\left(n - \frac{1}{2}\right)\frac{e}{C_G} < V_G < \left(n + \frac{1}{2}\right)\frac{e}{C_G} \tag{12.2}$$

Cette relation donne la gamme de tension appliquée sur la grille pour maintenir n électrons dans la boîte quantique. On peut également calculer la variation d'énergie libre.

$$F(n+1) - F(n) = \frac{e}{C_B}(Q_B - Q_C) \tag{12.3}$$

La charge critique Q_C est définie par :

$$Q_C = \frac{e}{2\left(1 + \frac{C_G}{C_B}\right)} \tag{12.4}$$

12.3.2 *Le transistor à un électron*

Le dispositif précédent permettait de contrôler une charge, électron par électron. Il est maintenant possible de construire un commutateur comme il est représenté *figure 12.10*.

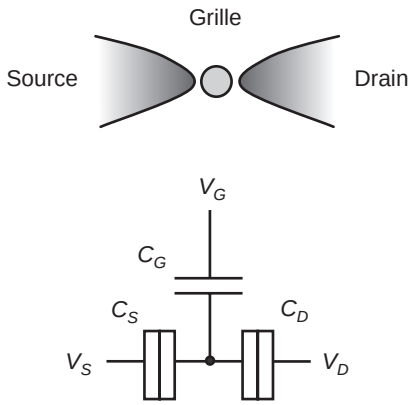


Figure 12.10 – Le transistor à un électron.

Les capacités sont celles formées entre l'îlot et les électrodes. On suppose qu'il y a des effets tunnel possible entre source et îlot et entre îlot et drain. La grille ne fait que contrôler le potentiel et il n'y

a pas d'effet tunnel possible entre grille et îlot. Le principe du calcul qui va suivre est de se ramener au cas plus simple de la jonction étudié dans le paragraphe précédent. On utilise le classique théorème de Thévenin de la théorie des circuits. Si on considère le circuit de la source à la grille, on obtient le schéma de la *figure 12.11*.

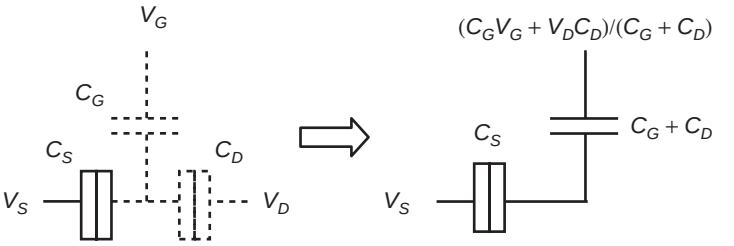


Figure 12.11 – Transformation du schéma par Thévenin.

On en déduit donc la condition pour que n électrons soient dans l'îlot.

$$\left(n-\frac{1}{2}\right)\frac{e}{C_G+C_D} < \frac{C_G V_G + C_D V_D}{C_G+C_D} < \left(n+\frac{1}{2}\right)\frac{e}{C_G+C_D}$$

soit,

$$\frac{1}{C_D}\left(n\,e-\frac{e}{2}-C_G V_G\right) < V_D < \frac{1}{C_D}\left(n\,e+\frac{e}{2}-C_G V_G\right) \qquad (12.5)$$

Il est possible de faire un calcul équivalent en considérant cette fois la jonction drain-grille. On obtient alors :

$$\frac{e}{C_G+C_S}\left(n-\frac{1}{2}\right) < \frac{C_G V_G}{C_G+C_S} - V_D < \frac{e}{C_G+C_S}\left(n+\frac{1}{2}\right)$$

soit,

$$\frac{1}{C_G+C_S}\left(-n\,e+\frac{e}{n}+C_G V_G\right) > V_D > \frac{1}{C_G+C_S}\left(-n\,e-\frac{e}{n}+C_G V_G\right) \qquad (12.6)$$

Il est maintenant possible de tracer les domaines de validité des équations 12.5 et 12.6 pour les petits nombres d'électrons.

Les nombres indiqués dans les zones grisées donnent les nombres possibles d'électrons en fonction des inégalités précédentes. Le symbole « 2, 1, 0 » signifie qu'il peut y avoir 2, 1 ou 0 électrons dans l'îlot.

Quand la tension de drain est faible, on constate que l'augmentation de la tension de grille fait passer le nombre d'électrons stockés dans l'îlot de 0 à 1 puis de 1 à 2 puis de 2 à 3 et ainsi de suite. Ces changements d'états sont accompagnés par des sauts quantifiés de courant comme il est indiqué *figure 12.13*.

Cette variation périodique du courant est appelée oscillation de Coulomb. La théorie dite orthodoxe relie le taux d'effets tunnels dans une jonction tunnel et la variation d'énergie libre du système. Les deux jonctions source-îlot et îlot-drain doivent être prises en compte.

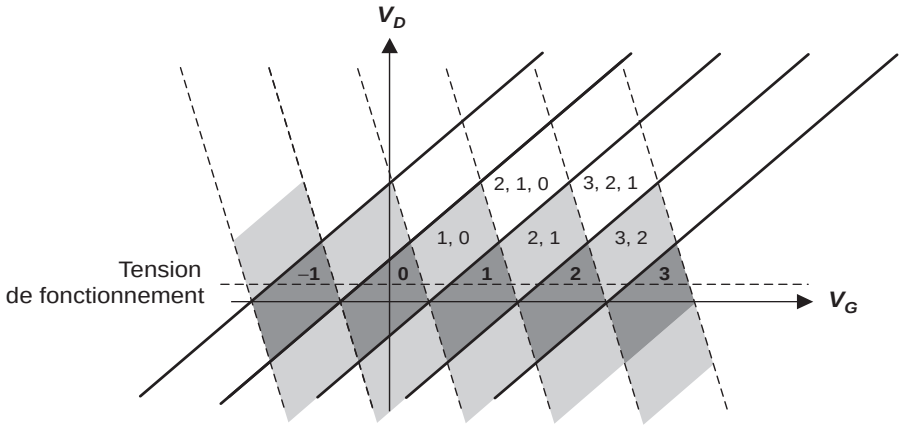


Figure 12.12 – Zones de stabilité du transistor à un électron.

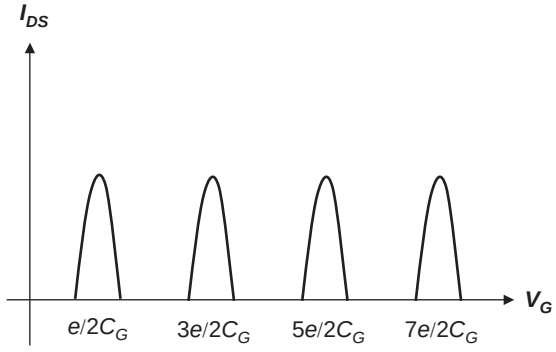


Figure 12.13 – Courant dans un transistor à un électron.

$$\Gamma(n, n+1) = \frac{1}{e^2 R_t} \frac{\Delta F(n, n+1)}{1 - \exp - \frac{\Delta F(n, n+1)}{k_B T}} \tag{12.7}$$

La résistance de la jonction tunnel au sens quantique du terme est notée R_t. On définit alors les taux d'effets tunnels pour les deux jonctions tunnel du système. Le calcul du courant à partir de ces considérations reste cependant complexe.

En pratique, il est difficile d'imaginer des composants fonctionnant à tension de drain très faible comme nous l'avons supposé précédemment. Il est possible de calculer le courant en fonction de la tension de grille pour une tension V_{DS} quelconque. Rappelons les deux conditions pour observer ce type de conduction :

- L'énergie mise en œuvre pour ajouter un électron dans l'îlot doit être supérieure à l'énergie d'agitation thermique ce qui signifie que le dispositif est très petit (dimension caractéristique de 10 nm) ou que le dispositif est à très basse température.

- La résistance des jonctions tunnel est très supérieure au quantum de résistance de 25,8 kΩ, cela pour rendre négligeables les fluctuations quantiques du nombre d'électrons dans l'îlot. Cette condition n'est pas très simple à comprendre et sa justification rigoureuse est donnée dans les ouvrages spécialisés. Le résultat à retenir est que les dispositifs de ce type sont nécessairement à haute impédance.

On obtient alors une relation donnant le courant d'un transistor à un électron en fonction des tensions de drain et de grille, utilisable dans le calcul des circuits.

$$I_{DS} = \frac{e}{2RC} \frac{(v_{GSn}^2 - v_{DS}^2)\sinh\left(\frac{v_{DS}}{\hat{T}}\right)}{v_{GSn}\sinh\left(\frac{v_{DSn}}{\hat{T}}\right) - v_{DS}\sinh\left(\frac{v_{DS}}{\hat{T}}\right)} \tag{12.8}$$

avec,

$$v_{GSn} = \frac{2C_G V_{GS}}{e} - \frac{(C_G + C_S - C_D)V_{DS}}{e} - 1 - 2n$$

$$v_{DS} = \frac{CV_{DS}}{e}$$

$$\hat{T} = \frac{2k_B TC}{e^2}$$

$$C = C_G + C_D + C_S$$

$$R = R_{tS} + R_{tD}$$

Il faut maintenant donner des ordres de grandeur pour apporter à cette formule un sens plus physique. Les capacités réalisables sont de l'ordre de quelques aF et les résistances tunnel sont de l'ordre de 10 MΩ. La température réduite est d'environ 0,2 à basse température. On mesure alors un courant de l'ordre de 100 pA, quand la tension de grille varie de quelques dizaines de mV. Le courant traversant le dispositif est finalement représenté *figure 12.14*.

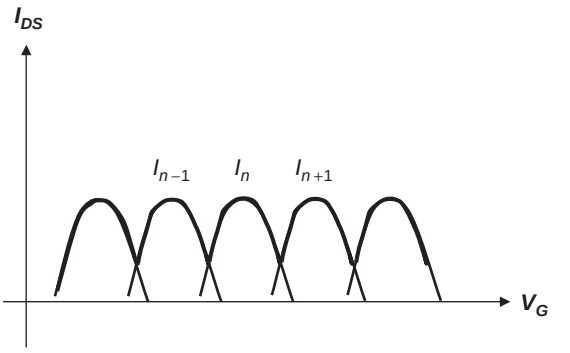


Figure 12.14 – Le courant dans un dispositif réel.

Un modèle complet permet d'obtenir la forme du courant pour deux températures.

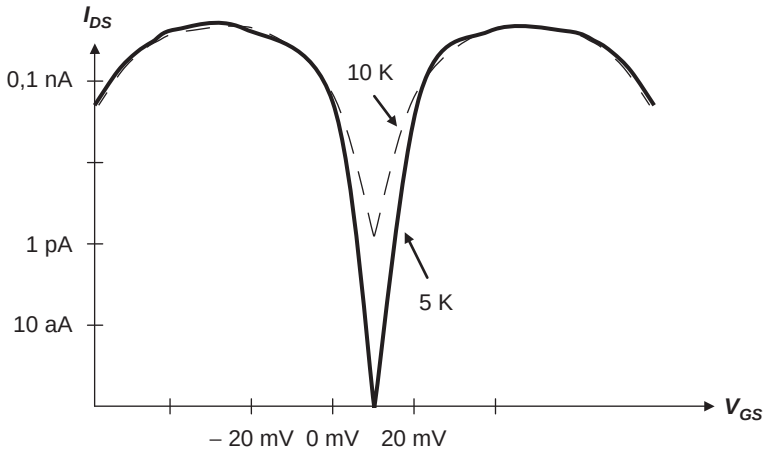


Figure 12.15 – Résultats de simulation.

Il est maintenant possible d'imaginer des fonctions logiques. La *figure 12.16* représente un inverseur obtenu avec deux dispositifs de ce type. On s'appuie sur un fonctionnement du dispositif de base obtenu avec des valeurs typiques. Le schéma de l'inverseur est proposé *figure 12.16*. Il est formé de deux dispositifs en série et polarisés positivement et négativement.

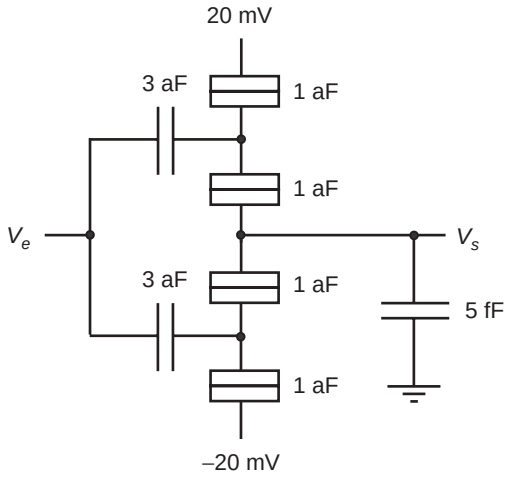


Figure 12.16 – L'inverseur à base de transistors à un électron.

Il suffit alors de faire varier la tension d'entrée et de reporter les valeurs de la courbe de la *figure 12.15* pour trouver les différents modes de fonctionnement.

Quand par exemple l'entrée est à – 20 mV, la tension grille-source du SET supérieur est de – 40 mV et le courant est de l'ordre de 0,1 nA. La tension grille-source du SET inférieur est à 0 V et son courant est alors de l'ordre de 10 pA. On a supposé que la valeur de la tension de drain avait peu d'influence sur la valeur du courant ce qui est une approximation grossière. L'examen de la tension de sortie montre alors que cette répartition des courants conduit à une tension de sortie d'environ 5 mV.

Si la tension d'entrée passe à – 10 mV, le SET supérieur est encore « passant » mais le courant du SET inférieur est encore plus faible. La tension de sortie est alors proche de 20 mV. Quand la tension de sortie passe à 10 mV, la situation est inverse et le SET inférieur conduit plus que le SET supérieur. La tension de sortie est alors voisine de – 20 mV. Il en est de même quand la tension d'entrée est de 20 mV. L'ensemble de ces résultats est représenté *figure 12.17*.

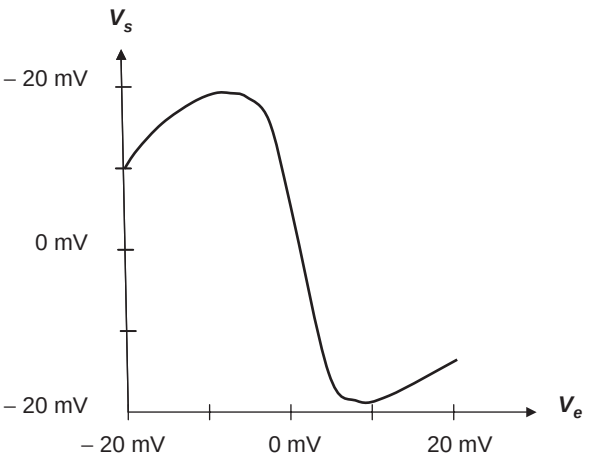


Figure 12.17 – Fonction de transfert de l'inverseur à base de SET.

Notons que la gamme de fonctionnement est de 40 mV, ce qui est largement plus faible que la gamme de fonctionnement de la logique à base de MOS. Cette gamme est définie à une température de 5 K. Elle serait encore plus faible à plus haute température.

Pour terminer cette introduction, il est possible de donner quelques critères relatifs à la faisabilité de la logique à base de SET. Pour cela, on examine deux cas. Dans le premier cas le SET charge une capacité, par exemple celle d'une interconnexion et des dispositifs reliés. Dans le second cas, le SET est un chemin de décharge d'une capacité présente dans la logique.

L'analyse de ces cas conduit à la condition suivante : la tension d'alimentation doit être inférieure à l'excursion en tension de la tension de grille.

On en arrive alors à proposer une logique dont l'architecture est indiquée *figure 12.18*.

Le fonctionnement de cette logique s'effectue en deux temps :

- dans une première phase, la capacité de charge est mise au potentiel de l'alimentation ;
- dans une seconde phase, la logique est mise à la masse, la tension d'alimentation étant coupée et les entrées sont appliquées.

Ce principe est comparable à celui de la logique dynamique ou logique domino étudiée dans le chapitre 8 de cet ouvrage. Une amplification par de la CMOS classique reste indispensable.

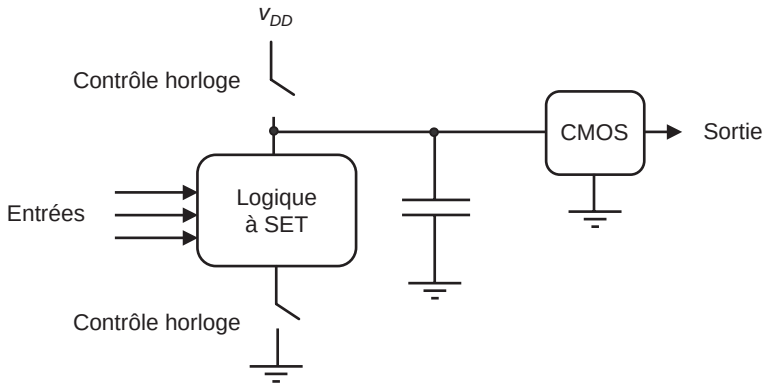


Figure 12.18 – Architecture d'une logique à SET.

Cependant, on peut espérer une consommation globale plus faible que dans une logique traditionnelle et un fonctionnement à température ambiante si les dispositifs SET sont de tailles très faibles.

12.4 Les molécules fonctionnalisables

12.4.1 *La conduction de molécules uniques*

La micro-électronique réduit la taille du transistor pour lui donner des dimensions de l'ordre d'une dizaine de nanomètres. L'idée peut être poussée plus loin et certains imaginent de créer des fonctions électroniques, numériques ou analogiques, au niveau de la molécule elle-même. Cette approche séduisante pose cependant un grand nombre de difficultés pour réaliser les molécules envisagées et encore plus pour les assembler et réaliser des fonctions globales. Les avantages attendus par rapport aux technologies plus classiques sont d'une part une augmentation de la densité d'intégration et d'autre part la possibilité de réaliser des systèmes en mettant en œuvre des procédés d'auto-assemblage supposés moins coûteux que les procédés lithographiques.

Ces molécules peuvent être des molécules de synthèse, des molécules présentes dans la nature, des nanoparticules métalliques ou semi-conductrices. Les premières molécules organiques ont été proposées dans les années 70 et présentaient des caractéristiques courant-tension asymétriques. Les termes « redresseur moléculaire » puis « transistor moléculaire » ont été successivement introduits. Une étape importante a été franchie quand une molécule unique a été connectée électriquement. La microscopie à effet tunnel a permis cet exploit.

Une expérience fondamentale illustrée en *figure 12.19* a été réalisée en 1997 pour mesurer, à l'aide d'un microscope à effet tunnel, le courant émis par une molécule de C_{60} déposée sur une couche d'or recouverte d'un dépôt isolant. Rappelons le principe de la microscopie tunnel. Une pointe très fine est approchée d'une surface et le courant tunnel entre la pointe et la surface est mesuré. Cette mesure fait apparaître à 4,2 K des marches de courant en fonction de la tension appliquée entre la pointe et la surface d'or. Ces résultats peuvent s'interpréter en supposant deux effets tunnel, un premier entre l'or et la molécule à travers l'isolant et un second de la molécule à la pointe du microscope.

La situation est différente si la molécule est déposée directement sur le métal. Les caractéristiques sont alors linéaires et la résistance de la jonction molécule-métal est d'environ 55 MΩ. Ces expériences montrent que le couplage entre molécules et substrat est difficile à maîtriser.

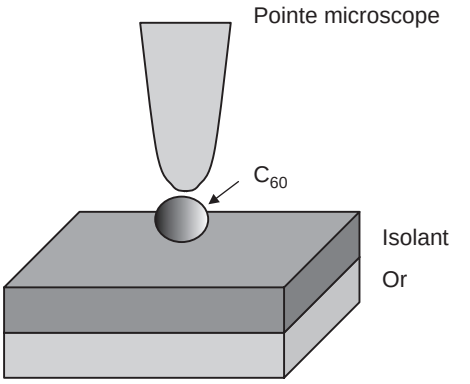


Figure 12.19 – Conduction d'une molécule.

Nous pouvons maintenant décrire sommairement le principe de fonctionnement d'un dispositif électrode-molécule-électrode représenté *figure 12.20*. Le couplage entre molécule et substrat peut se décrire par deux mécanismes comme il est expliqué dans la référence **[11]**.

- Le couplage fort. Il faut considérer le dispositif comme un guide d'onde pour la fonction d'onde de l'électron.
- Le couplage faible. Il faut alors interpréter la conduction comme le passage d'un électron d'une électrode à une autre en faisant intervenir un état intermédiaire dans lequel la molécule est chargée.

12.4.2 *Description de la conduction en régime de couplage fort*

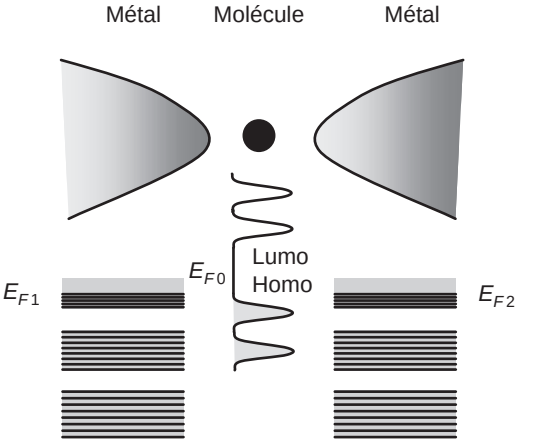


Figure 12.20 – Le modèle de couplage.

Sur la *figure 12.20*, on reconnaît les niveaux d'énergie dans les deux électrodes et dans la molécule. Les énergies sont quantifiées dans les électrodes et les états sont occupés en dessous des valeurs

limites appelées potentiel chimique et notées E_F. Les états énergétiques de la molécule sont répartis plus continûment. Le niveau Homo est le niveau le plus haut occupé tandis que le niveau LUMO est le niveau le plus bas non occupé. On peut définir pour la molécule un niveau de Fermi et non pas un potentiel chimique car le potentiel chimique a un sens uniquement pour une population de particules. Le courant se calcule alors en considérant les équilibres statistiques entre les électrodes et la molécule. Le détail du calcul est donné dans la référence **[11]**.

La *figure 12.21* montre le type de résultat que l'on peut obtenir.

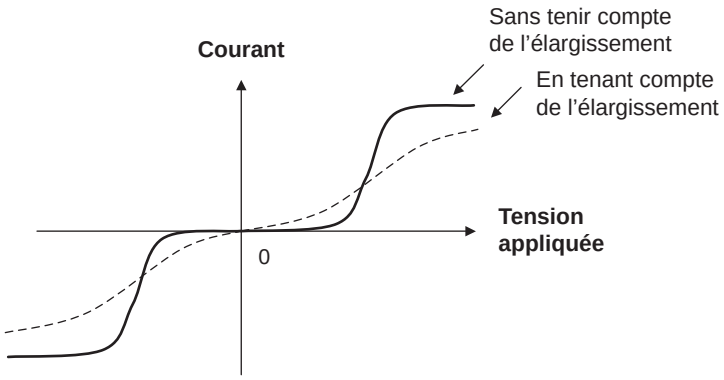


Figure 12.21 – Courbe courant tension d'un dispositif moléculaire.

12.4.3 *Description de la conduction en régime de couplage faible*

Dans ce cas, les transferts des électrons de l'électrode à la molécule et de la molécule à l'électrode ont lieu par effet tunnel.

La *figure 12.22* représente les niveaux énergétiques dans les électrodes et les niveaux énergétiques dans la molécule correspondant soit à l'émission d'un électron (potentiel d'ionisation) soit à l'absorption d'un électron (affinité électronique). La molécule est supposée couplée capacitivement avec les électrodes.

On obtient alors une forme du courant analogue à celle de la *figure 12.21* mais faisant intervenir les coefficients de couplage par effet tunnel.

12.4.4 *Dispositifs à base de molécules*

La première fonction à créer est celle de fil conducteur. Les oligomères π-conjugués sont considérés comme des prototypes. La conductance d'un système métal-molécule-métal de ce type peut s'exprimer sous la forme :

$$G = G_0 \exp^{-\beta L}$$

Dans cette relation, G_0 est la conductance de contact, β est le coefficient d'atténuation tunnel et L la longueur de la molécule. Les caractéristiques courant-tension sont fortement non linéaires et les courants sont compris entre 10 et 10^4 pA par molécule.

Des dispositifs redresseurs ont été également réalisés. En plus de la non linéarité, ils montrent une forte dissymétrie en fonction de la polarité de la tension appliquée.

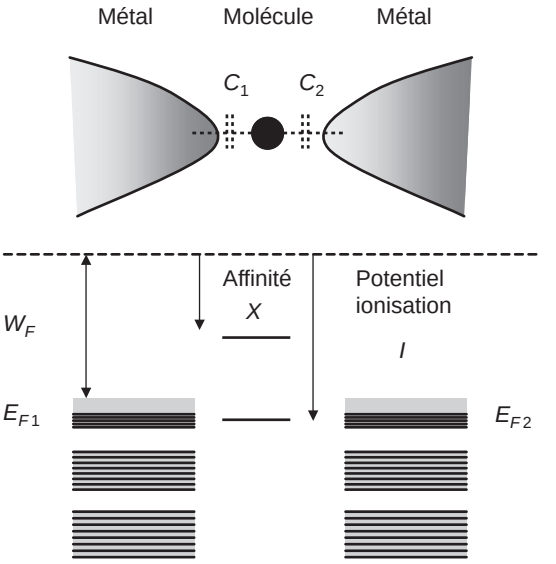


Figure 12.22 – Modèle pour la liaison faible.

Des éléments de mémoire ont été étudiés avec des molécules de type caténanes ou rotaxanes. Ces molécules sont constituées de deux parties entrelacées et mobiles et peuvent adopter des positions différentes après une excitation chimique ou optique. Des travaux ont été menés dans ce sens par les laboratoires de HP pour arriver à un premier prototype de mémoire moléculaire. Une impulsion de tension permet de fixer la molécule dans un état donné et la lecture se fait par mesure de la conductance. Ces travaux donnent lieu à débat pour expliquer précisément les mécanismes mis en jeu.

Quelques travaux ont également permis de mettre au point des transistors moléculaires. Drain et source métallique sont en contact avec une molécule tandis qu'une troisième électrode modifie la conduction de la molécule.

Dans tous les cas, il reste difficile d'obtenir des dispositifs stables et reproductibles et les réalisations restent encore dans le domaine de la recherche.

12.5 Les architectures associées

12.5.1 *Contraintes générales imposées à ces architectures*

Réaliser des architectures électroniques à partir des composants nanométriques n'est pas un mince problème. La difficulté principale est d'imaginer des architectures compatibles avec des procédés d'interconnexion réalisables à grande échelle. Pour juger de l'intérêt de ces nouvelles approches, le plus simple est de partir des difficultés de la micro-électronique de demain. Ces difficultés sont principalement au nombre de quatre :

- surmonter le coût de la lithographie extrême ;
- résoudre le problème de l'augmentation des temps de propagation ;
- maîtriser l'augmentation de la consommation ;
- maîtriser l'accroissement des dispersions.

Les solutions sont, dans chaque cas, ou technologiques ou architecturales et le plus souvent une combinaison des deux.

Face à l'augmentation exponentielle des coûts de la lithographie, deux attitudes sont possibles. La première est d'imaginer des technologies d'auto-assemblage permettant alors de réaliser des composants nanométriques avec une densité d'intégration très élevée sans faire usage de la lithographie. Il faut non seulement réaliser ces composants mais aussi les interconnecter.

La seconde attitude est de conserver une lithographie avancée dans la fabrication des systèmes nanométriques mais d'imaginer des architectures telles que le même circuit puisse être utilisé dans un grand nombre d'applications différentes. C'est l'approche reconfigurable.

Pour maîtriser l'augmentation des temps de propagation et résoudre le problème de la distribution d'une horloge à fréquence élevée sur une puce, des solutions technologiques sont possibles, par exemple en utilisant des composants optoélectroniques et en envisageant une distribution optique du signal d'horloge. Des solutions sont également possibles au niveau de l'architecture en imaginant des circuits globalement asynchrones et localement synchrones.

Si on pense maintenant à la maîtrise de la consommation, on peut espérer que le traitement de l'information par des dispositifs à un électron ou par les molécules peut conduire à minimiser l'énergie mise en œuvre. Les considérations du chapitre 11 ont cependant montré que l'interconnexion amenait des contraintes fortes pour la valeur minimale de l'énergie mise en œuvre dans un tel mécanisme. Les solutions architecturales sont basées sur une meilleure coopération entre le logiciel et le matériel. Le bloc impliqué dans une tâche choisit ses paramètres physiques de fonctionnement (tension, fréquence horloge, tension de seuil) pour satisfaire les contraintes de la tâche logicielle tout en consommant le moins possible.

La prise en compte des dispersions et des défauts, inéluctables avec une réduction de la taille du composant élémentaire en dessous de 100 nm, conduit à envisager les architectures tolérantes aux fautes. Ces aspects ont été étudiés en détail dans le chapitre 11.

En résumé, les solutions ne sont pas purement technologiques ou architecturales mais le plus souvent une habile combinaison des deux approches. En considérant les aspects purement architecturaux, deux approches sont possibles pour utiliser les dispositifs nanométriques :

- remettre en cause radicalement les architectures actuelles, par exemple en envisageant des systèmes auto-organisés s'inspirant des réseaux d'automates couplés ou des systèmes neuronaux. Ces concepts sont en accord avec les technologies nanométriques mais il faut pouvoir porter sur de telles architectures les problèmes traités aujourd'hui sur des ordinateurs électroniques. Ce n'est pas une mince affaire et le monde de l'informatique et des services associés est concerné ;
- associer de manière habile les dispositifs à base de nanotechnologies aux architectures CMOS conventionnelles. Il est par exemple possible d'ajouter à un circuit CMOS réalisé à l'aide d'une lithographie « raisonnable », des éléments mémoire obtenus par une méthode *bottom-up* à base de nanofils ou de nanotubes. Cette approche appelée *nano inside* semble aujourd'hui la plus réaliste.

12.5.2 *Architectures à base de nanofils ou de nanotubes*

Les fonctions élémentaires à base de nanofils ou de nanotubes ont été présentées dans le paragraphe 12.2. Nous allons compléter cette présentation par l'étude de quelques fonctions logiques de base dont la fonction NOR représentée *figure 12.23*.

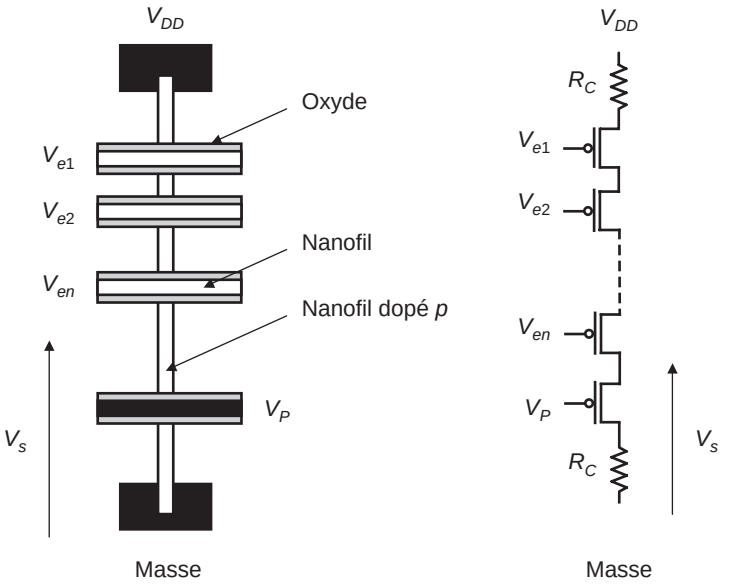


Figure 12.23 – NOR à base de nanofils.

On reconnaît dans cette figure un nanofil ou un nanotube dopé *p* et des nanofils de commande utilisés en simples conducteurs et reliés aux entrées. Ils ne sont pas en contact électriques avec le nanofil dopé mais séparés par une mince couche d'oxyde. Cette couche pourrait se fabriquer en faisant circuler un courant de forte intensité dans le nanofil. On suppose que les techniques de fabrication collective des nanofils ou des nanotubes permettent de construire à grande échelle des structures de ce type, ce qui reste à démontrer.

Le fonctionnement de cette porte est le suivant. Les croisements de nanofils conducteurs avec le nanofil dopé sont équivalents à des PMOS. En effet, une tension positive appliquée sur un nanofil conducteur horizontal repousse les trous du fil dopé *p* en dehors de la zone d'influence du fil qui joue donc le rôle d'une grille. Le fil dopé devient donc non conducteur dans cette région. Le PMOS relié à la masse joue le rôle d'une charge. La valeur de sa résistance (R_p) est réglée par la tension V_P.

Appelons R_C la valeur de la résistance de contact entre les pistes conductrices et les nanofils. La valeur est au minimum de 6,5 kΩ comme il a été expliqué dans le paragraphe 12.1. Les résistances des PMOS commandés par les entrées sont notées $R(V_{ei})$. La tension de sortie de ce dispositif s'écrit donc :

$$V_S = V_{DD} \frac{R_C + R_p}{2R_C + R_p + \sum_{i=1}^{i=n} R(V_{ei})}$$

Quand toutes les entrées sont à 0 V, le fil est conducteur et la tension de sortie est :

$$V_S = V_{DD} \frac{R_C + R_p}{2R_C + R_p}$$

Cette tension est peu différente de V_{DD} si la résistance *Rp* est largement supérieure à R_C.

Quand une des entrées est à l'état haut, le PMOS associé est non passant et si on suppose que sa résistance à l'état non passant est suffisamment élevée, la tension de sortie est à la masse.

Le circuit remplit bien la fonction NOR. Les autres fonctions de la logique peuvent se réaliser par des dispositifs équivalents.

Il faut maintenant étudier comment des dispositifs peuvent être reliés au monde extérieur. Comme nous l'avons noté en introduction, il est peu réaliste d'imaginer un circuit entier réalisé à base de nanofils ou de nanotubes. Les dimensions de ces dispositifs sont d'ailleurs incompatibles avec des circuits ayant des côtés de l'ordre du cm. Il est donc plus réaliste de supposer que le circuit total est formé d'un nombre élevé de nanoblocs intégrant des cellules à base de fils comme la cellule NOR que nous avons prise en exemple. Ces différents blocs sont reliés à un circuit CMOS classique qui gère les échanges entre nanoblocs.

Ce type d'architecture est représenté *figure 12.24*.

Les nanoblocs communiquent entre eux deux à deux et les fils de sortie de l'un sont les fils d'entrée de l'autre. Les conducteurs de la partie CMOS peuvent servir à mettre en contact un fil d'un nanobloc avec un fil d'un autre nanobloc. Ils peuvent également permettre de mettre un fil sous tension ou de l'isoler. On imagine donc facilement que de tels dispositifs sont capables d'implémenter de la logique programmable à haute densité.

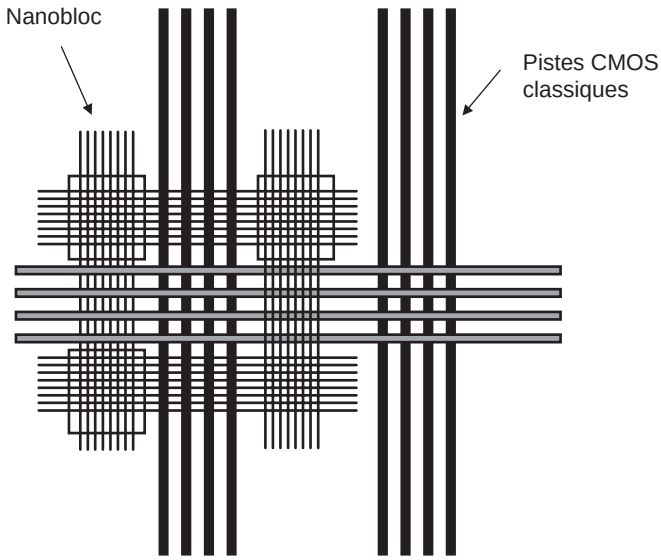


Figure 12.24 – Architecture de type nano inside.

Un problème important a cependant été passé sous silence. C'est la manière de réaliser les contacts entre les pistes CMOS typiquement de 100 nm de large avec les nanofils ou les nanotubes typiquement de 10 nm de diamètre. Il ne faudrait pas bien sûr que cette technique fasse usage de procédés lithographiques de haute précision (10 nm) car le but des nanotechnologies est précisément de se passer de la lithographie extrême. Nous allons donc passer en revue quelques techniques qui per-

mettent de faire ce lien électrique entre la nano-électronique et la micro-électronique. Les deux techniques principales sont illustrées *figure 12.25*.

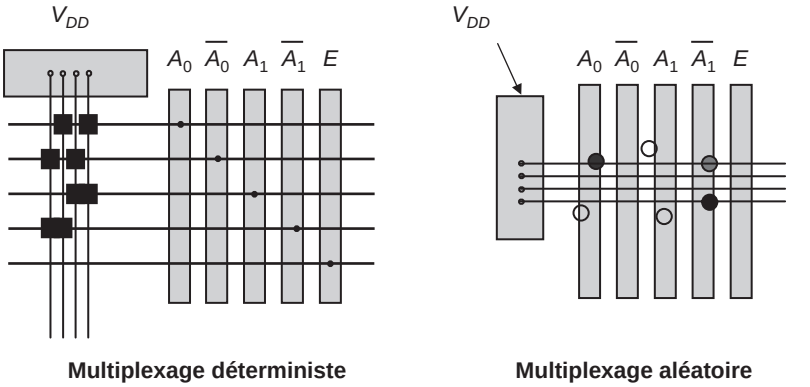


Figure 12.25 – Techniques de multiplexage.

Dans le multiplexage déterministe, les nanofils verticaux sont ou ne sont pas reliés aux nanofils horizontaux. La présence d'un carré noir sur la figure montre qu'un isolement électrique est réalisé. Les isolants représentés par les carrés réalisent en fait un adressage physique des nanofils. L'application de tensions sur les lignes d'adressage permet de relier un nanofil et un seul à l'alimentation. Le mécanisme sera étudié en détail par la suite. La réalisation des isolants est un problème difficile car il faut se passer de la lithographie. Des solutions de type nano-impression peuvent être envisagées.

Dans le multiplexage aléatoire, certains nanofils sont reliés et d'autres non, aux pistes d'adressage. Par exemple, des molécules d'or sont déposées avec une densité choisie de telle manière que la probabilité qu'il y ait contact entre la piste et le nanofil soit environ 1/2. Cette technique semble plus simple à mettre en œuvre. Elle nécessite cependant plus de lignes d'adressage que la technique déterministe. Rappelons que le but de ce dispositif est d'adresser un fil et un seul pour une adresse donnée. Dans ce cas, cet objectif ne pourra être totalement atteint et une même adresse pourra sélectionner plusieurs fils dans quelques cas rares.

Revenons sur le mécanisme d'adressage déterministe en supposant que les nanofils sont dopés *p* et que chaque bit d'adresse se traduit par une tension positive sur une piste et une tension nulle sur la piste complémentaire. Il est en effet indispensable dans ce type de multiplexage d'avoir les deux signaux, le signal et son complément. La *figure 12.26* montre l'application de deux codes différents et indique le fil adressé c'est-à-dire conducteur.

Le lecteur vérifiera facilement le fonctionnement de ce circuit en tenant compte du fait que le nanofil est coupé au sens électrique quand un des nanofils horizontaux, jouant le rôle de grille, est polarisé et quand l'influence électrique n'est pas annihilée par un isolant (carré noir sur la figure).

Pour s'affranchir des difficultés de fabrication des nano-isolants, une autre technique d'adressage physique du nanofil a été imaginée. Il est possible, à la fabrication, de faire varier le dopage du fil en fonction de la position. Imaginons donc le nanofil comme une succession de zones fortement dopées et de zones faiblement dopées. L'effet des nano-isolants est reproduit de manière plus simple. En effet, une zone fortement dopée est insensible à la tension électrique appliquée et la conduction électrique reste importante dans cette zone. Ce principe est illustré *figure 12.27*.

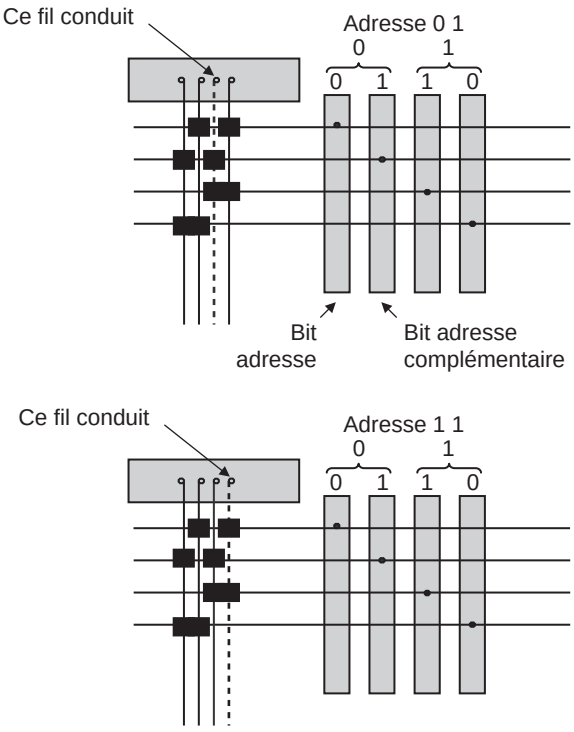


Figure 12.26 – Multiplexage déterministe.

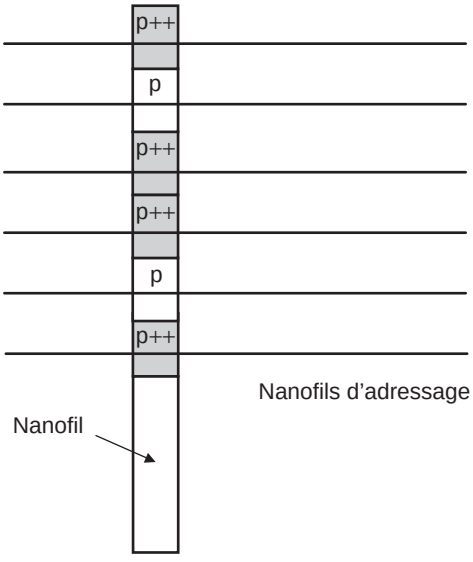


Figure 12.27 – Adressage physique par dopage.

On comprend alors comment il est possible de concevoir des mémoires à base de nanofils et de nanotubes. Le point mémoire est obtenu par croisement entre deux nanofils. On admet que le passage d'un courant élevé dans les deux fils crée un phénomène particulier à l'intersection, par exemple une modification de la résistance électrique au niveau du croisement. Les techniques de multiplexage que nous avons étudiées précédemment permettent de choisir ces fils. Pour fonctionnaliser le point mémoire, les effets électrostatiques peuvent être utilisés mais aussi les propriétés de certaines molécules placées à l'intersection. Il est alors possible, toujours en utilisant les techniques de multiplexage de choisir deux fils qui se croisent et de mesurer la résistance du contact. Il faut également prévoir un mécanisme d'effacement de l'effet mémoire obtenu.

Toutes ces techniques sont pour l'instant du domaine de la recherche car il n'y a pas encore de procédé permettant de réaliser de telles architectures à grande échelle.

12.5.3 *Les architectures à base de molécules*

Utiliser directement les molécules pour réaliser les dispositifs laisse envisager des techniques de fabrication chimique simples sans avoir à mettre en œuvre des procédés lithographiques complexes. Une approche intéressante est proposée par James Tour et son équipe (référence **[14]**). Le principe est de fonctionnaliser des dispositifs appelés *nanocells* représentés *figure 12.28*.

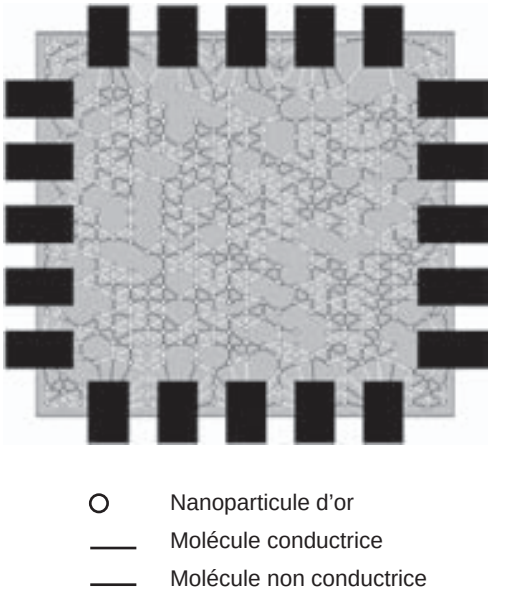


Figure 12.28 – Nanocell fonctionnalisable.
J.M. Tour et al., Nanocell logic gates for molecular computer, *© IEEE 2002.*

La *nanocell* est composée d'un ensemble de molécules qui peuvent être en liaison chimique entre elles par l'intermédiaire de nanoparticules d'or. Ces molécules ont deux états possibles, l'un équivalent à un état de conduction et l'autre à un état de non conduction. Le système total est donc équivalent à un réseau de résistances interconnectées de manière aléatoire. Cet ensemble de molécules est relié à un certain nombre de plots électriques disposés en périphérie et servant d'interface avec une élec-

tronique intégrée plus classique par exemple dans une technologie 90 nm. Les molécules ont la propriété de voir leur état changer en fonction de la tension appliquée à leurs bornes.

Le système est, après fabrication, dans un état non défini et le réseau de résistances est aléatoire. Il est alors possible de le programmer. En effet, en appliquant des combinaisons de tensions différentes sur les électrodes périphériques, on modifie la résistance de certaines molécules en fonction du principe énoncé précédemment. Si cette opération est répétée un grand nombre de fois en suivant un algorithme déterminé, il est possible d'établir une relation logique entre les tensions appliquées sur certains plots et les courants mesurés sur d'autres. La *nanocell* a été fonctionnalisée. Des fonctions logiques de base ont été ainsi créées comme le montre la *figure 12.29*.

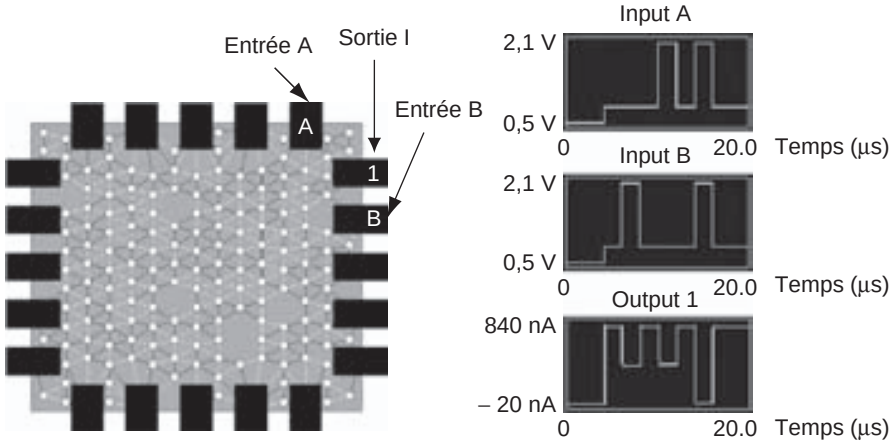


Figure 12.29 – Porte NAND à base de molécules.
J.M. Tour et al., Nanocell logic gates for molecular computer*, © IEEE 2002.*

Ajoutons qu'il est nécessaire que la courbe courant-tension de la molécule présente au moins dans une région une résistance négative. En effet, un système logique complet ne peut se faire sans une fonction de type inverseur. Le principe de fonctionnement de la *nanocell* étant uniquement basé sur une division du potentiel appliqué en fonction des résistances, il est indispensable de disposer de résistances négatives pour réaliser une fonction de type inverseur.

Bibliographie

Physique générale

[1] Berkeley (Université de), *Cours de physique de Berkeley – Tome 2 : Électricité et Magnétisme*, Dunod, 1998

[2] Albert Messiah, *Mécanique quantique, 2e édition* (2 tomes), Dunod, 1995

[3] Charles Kittel, *Physique de l'état solide, 7e édition*, Dunod, 2005

[4] Christian et Hélène Ngô, *Les Semi-conducteurs – De l'électron aux dispositifs*, Dunod, 2003

Électronique et circuits intégrés

[5] Hervé Fanet, François De Dieuleveult, *Principes et applications de l'électronique* (2 tomes), Dunod, 1997

[6] R. Jacob Baker, *CMOS Circuit design, Layout and Simulation, 2nd edition*, John Wiley and Sons, 2005

[7] Yannis P. Tsividis, *Operation and Modelling of the MOS transistor, 2nd edition*, McGraw-Hill, 1999

Articles de synthèse

[8] D.J. Franck *et al.*, *Device scaling limits of Si MOSFETs*, Proceedings of the IEEE, Vol. 89, march 2001

[9] R.W. Keyes, *Fundamental limits of silicon technology*, Proceedings of the IEEE, Vol. 89, march 2001

[10] Richard P. Feynman, Robin W. Allen and Tony Hey, *Feynman lectures on computation*, Perseus Books, 2000

[11] Marcel Lahmani, Claire Dupas, Philippe Houdy, *Les nanosciences*, Belin, 2004

[12] K. Nikolic, A. Sadeck and M. Forshaw, *Fault-tolerant techniques for nanocomputers*, Nanotechnology 13 2002

[13] André DeHon, Patrick Lincoln, John E. Savage, *Stochastic assembly of sublithographic nanoscale interfaces*, IEEE transactions on Nanotechnology, sept. 2003

[14] J.M. Tour *et al.*, *Nanocell logic gates for molecular computing*, IEEE transactions on Nanotechnology, june 2002

[15] Dmitri B. Strukov, Konstantin K. Likharev, *Prospects for terabit-scale nanoelectronic memories*, Nanotechnology 16, 2005

Index

049141 - (I) - (1,2) - OSB 80° - ALL-MMC
Imprimerie Chirat - 42540 Saint-Just-la-Pendue
Dépôt légal Mars 2006 - N° 8888

Imprimé en France

Collection EEA
Série Technologie électronique

PRINCIPES ÉLECTRONIQUES
TECHNOLOGIE ÉLECTRONIQUE
ÉLECTROTECHNIQUE
AUTOMATISMES

Hervé Fanet

MICRO ET NANO-ÉLECTRONIQUE

Bases • Composants • Circuits

Cet ouvrage présente de façon exhaustive l'art de la **miniaturisation** en électronique intégrée, en étudiant l'ensemble des **aspects physiques**, **technologiques et architecturaux** de la micro-électronique et de la nano-électronique. Les **bases** de ces disciplines ainsi que **les progrès les plus récents** sont exposés :

- principes de fonctionnement des composants électroniques (physique des semi-conducteurs, électromagnétisme, mécanique quantique) ;
- transistor MOS (le plus utilisé) et transistor bipolaire ;
- procédés de fabrication de l'industrie micro-électronique ;
- fonctions analogiques et numériques de base ;
- circuits intégrés complexes, processeurs et mémoires ;
- composants nanométriques et architectures associées (nanofils, nanotubes…) ;
- perspectives d'utilisation, circuits électroniques de demain.

Ce livre s'adresse principalement aux ingénieurs et chercheurs en électronique, ainsi qu'aux enseignants et étudiants du domaine.

HERVÉ FANET

Ingénieur de recherche au CEA LETI. Il est en charge de coopérations scientifiques et d'actions de formation au sein du pôle MINATEC (micro et nanotechnologies) à Grenoble.

ISBN 2 10 049141 5

www.dunod.com

DUNOD