République Algérienne Démocratique et populaire
Ministre de l'enseignement Supérieur et de la recherche scientifique

Université de Batna
Faculté des Sciences de l'Ingénieur
Département d'Electronique

Laboratoire de l'Electronique Avancée
LEA Batna
Laboratoire des Etudes Physico-Chimique des Matériaux
LEPCM Batna

THESE DE DOCTORAT EN SCIENCES
EN ELECTRONIQUE
Option : **Contrôle**

Présentée par

Fayçal DJEFFAL
Ingénieur d'état en Electronique, Magister en Electronique

THEME

Modélisation et simulation prédictive du transistor MOSFET fortement submicronique. Application à la conception des dispositifs intégrés

Thèse soutenue le 19 / 04 / 2006
Devant le jury composé de :

N.E. Bouguechal	Professeur	Université de Batna	Président
M. Chahdi	Professeur	Université de Batna	Rapporteur
A. Benhaya	M.C	Université de Batna	Co-Rapporteur
M.S. Aida	Professeur	Université de Constantine	Examinateur
N.E.Sengouga	Professeur	Université de Biskra	Examinateur
M. Bouchemat	Professeur	Université de Constantine	Examinateur

2005/2006

Table des matières

Liste des symboles

A_S et A_D	fonctions spectrales associées à la source et au drain
$ETrn$	erreur moyenne d'apprentissage
$ETst$	erreur moyenne de test
$ETrnMax$	erreur maximale d'apprentissage
$ETstMax$	erreur maximale de test
$E_i(x)$ $et \psi_i(x,z)$	énergie et la fonction d'onde pour le mode i à la tranche x
E_l	energie longitinale
$f\left[I(z_k)\right]$	fonction d'activation
$f'\left[I(z_k)\right]$	première dérivée de la fonction d'activation
f_t	fréquence de coupure
f_{td}	fréquence de coupure dégradée
$J_t = \frac{1}{2}(\vec{r}_0 - \vec{r}_t)^2$	erreur quadratique au temps t entre $\vec{r}_t$ (vecteur de sortie prédit à l'instant t) et $\vec{r}_0$ (le vecteur de sortie du cas soumis au réseau).
H	constante Planck
h_i	hamiltonien
I	nombre des entrées du réseau de neurones
I_{ch}	courant du canal
I_D	courant du drain
I_{DE}	courant du drain dégradé
$I(x_i)$	somme des flux provenant des autres neurones connectés en amont
Gmax	transconductance maximale
g_{md}	transconductance dégradée
K_B	constante Boltzmann
Lg	longueur du canal
m_y^*	masse effective d'électron
N_1	nombre des neurones de la première couche cachée
N_2	nombre des neurones de la deuxième couche cachée
N_{Ptr}	taille de la base de données d'apprentissage
N_{Ptst}	taille de la base de données de test
$N_{D/S}$	dopage de source/ drain (n+ type)
N_{CH}	dopage du canal
$o(z_k)$	représente un élément de la réponse expérimentale
O	nombre des sorties du réseau de neurones
$O(I_i, W)$	vecteur de sortie prédit correspondant le i^{ème} échantillon dans la base de données en fonction de l'entrée I_i et les poids de pondération W
q	charge électron
$\vec{r}_t\ (o(z_k))_{k=1,N_Z}$	vecteur de sortie prédit comprenant N_z neurones du réseau de neurones au temps t, fonction des sorties des neurones z_k de la couche z

$\vec{r}_0\,(r_k)_{k=1,N_Z}$	vecteur de sortie réel (expérimental) comprenant N_z réponses, fonction des réponses individuelles r_k
Tstress	temps du stress
t_{ox}	épaisseur d'oxyde
t_{si}	épaisseur du canal
T	température absolue
$T_{SD}(E_l)$	coefficient de transmission de la source au drain
$-t_{x,I}$	énergie de couplage
Vd	tension du drain
Vg	tension de grille
V_{DS}	tension drain-source
V_{GS}	tension grille-source
$V(x,z)$	potentiel électrostatique
x_{min}, x_{max}	limites physiques du paramètre x
x_y_z_t	structure d'un réseau de neurones comprenant x entrées, t sorties, y neurones dans la première couche cachée et z neurones dans deuxième couche cachée
z_i	sortie correcte qui correspond le $i^{ème}$ échantillon dans la base de données
W_{kj}	poids des connexions reliant le neurone j (source) au neurone k
$w(x_i, y_j)$	poids entre le neurones x_i de la couche x et le neurone y_j de la couche y
θ_k	seuil du neurone
$\nabla J_z(y_j, z_k)$	énergie à un nœud donné du réseau
α	vitesse d'apprentissage
β	$=KT/q$ potentiel thermique
μ_D	niveau de fermi pour le drain
μ_S	niveau de fermi pour la source
Σ_S et Σ_D	la self-énergie de la source et du drain

INTRODUCTION GENERALE

La croissance de l'industrie des semi-conducteurs dépend pour l'instant de sa capacité à miniaturiser les transistors. L'objectif de la démarche est de délivrer de meilleures performances à moindre coût. Des circuits plus petits réduisent la surface globale de la puce électronique et permettent donc de produire plus de transistors sur un même wafer et réduire le prix de fabrication. L'augmentation de la densité d'intégration et l'amélioration des performances sont rendues possibles par la diminution de la taille des transistors. La grandeur caractéristique d'un transistor qui distingue une génération de transistors de la suivante est la longueur de grille L_G. Actuellement, la plus petite longueur de grille produite est de 37 nm et la réduction des dimensions demeure essentielle pour réaliser de nouvelles innovations technologiques. L'ITRS (International Technology Roadmap for Semiconductors) dont les prévisions sont établies par les principales industries des semi-conducteurs annonce une longueur de grille de 9 nm en 2016 [1-3]. Cependant, de telles dimensions génèrent des problèmes technologiques bien plus complexes et surtout différents de ceux rencontrés dans les transistors actuels. En plus de la nécessité d'optimiser les procédés de fabrication, il est indispensable de connaître les limites physiques de la technologie CMOS (pour l'anglais "Complementary Metal-Oxide-Semiconducteur), au delà desquelles un transistor technologiquement parfait n'aurait plus des propriétés électriques tolérables. Bien que la réponse ne soit pas encore évidente, il apparaît clairement que de nouvelles architectures devront remplacer à plus ou moins longue échéance, les transistors conventionnels. Il a déjà été décidé par exemple de ne plus réaliser un type de transistor valable pour toutes les applications, mais une architecture particulière pour les applications "hautes performances" (microprocesseurs) ou pour les applications "basses consommations" (téléphones portables). Le développement de transistors nanométriques ajoute aux problèmes expérimentaux, des défis théoriques qu'il est crucial de surmonter pour que perdure l'élan de miniaturisation. La diminution

soutenue des dimensions accélère la rencontre de la microélectronique avec la mécanique quantique et d'autres lois régissent désormais le transport des électrons. La simulation des transistors a donc besoin de nouvelles théories et techniques de modélisation (l'intelligence artificielle) améliorant la compréhension physique des dispositifs de taille nanométrique.

Les réseaux de neurones artificiels, connus généralement sous l'acronyme ANN (Artificial Neural Networks), constituent une approche fondamentalement nouvelle dans l'études des dispositifs microélectroniques. Ce sont des systèmes parallèles, adaptatifs et distribués dont le fonctionnement imite celui de neurone biologique, reproduisant ses caractéristiques de base. Les ANNs offrent des solutions compactes et rapides pour une large gamme de problèmes, dont certaines sont difficiles à traiter par les approches classiques (analytiques, numériques,…), en particulier les problèmes avec des contraintes temps réel, ou ceux dont la résolution met en jeu des règles inconnues ou difficiles à expliciter ou à formaliser.

Le domaine de la modélisation et la simulation des dispositifs fortement submicroniques peut être considéré comme un champ important d'applications des ANNs. Par conséquent, l'étude de la possibilité d'utilisation des réseaux de neurones artificiels dans le domaine de la microélectronique, notamment sous forme des prédicteurs et des simulateurs des dispositifs nanométriques (Double-Gate MOSFET, Tri-gate MOSFET, gate-all-around MOSFET,….)[4], s'avère nécessaire. Dans ce contexte, les principaux objectifs de cette thèse sont : 1) d'étudier le transistor MOSFET fortement submicronique et l'effet de la miniaturisation, 2) d'utiliser l'approche neuronale pour le développement d'un modèle de dégradation temporel du transistor MOSFET fortement submicronique, 3) d'examiner et de comprendre le transport de charges dans les DGMOSFETs afin de développer un modèle neuronal de ce composant, 3) le développement d'un prédicteur neuronal permettant d'étudier les possibilités et les limites de la graduation dimensionnelle vers l'échelle atomique du transistor DGMOSFET en fonction des différents paramètres (La longueur de grille, l'épaisseur du canal, le dopage,….).

Cette thèse s'articulera autour de cinq grands axes:

Le premier chapitre explique le principe de fonctionnement du transistor MOSFET, rappelle la problématique actuelle liée à la miniaturisation des transistors, présente les réalisations les plus significatives des MOSFETs conventionnels (bulk, SOI,….) et détaille le cas prometteur d'une structure nanométrique à double grille: le transistor DGMOSFET.

Le deuxième chapitre est consacré aux réseaux de neurones: il en donne, les principes, expose les différents types d'implantations et domaines d'applications existants et décrit l'état de l'art sur leurs propriétés de modéliser les systèmes complexes.

Dans le troisième chapitre, nous proposons un modèle de prédiction à base des réseaux de neurones capable de prédire les variations de dégradation de transistors MOSFETs fortement submicroniques en fonction des quatre paramètres, à savoir: la longueur de la grille, la tension de drain, la tension de grille et le temps du stress. Ce chapitre peut être divisé en deux parties. Dans la première partie, on propose un dispositif expérimental assisté par ordinateur permettant d'étudier les aspects expérimentaux des phénomènes du vieillissement des transistors MOSFETs fortement submicroniques. La deuxième partie est consacrée au développement d'une approche analytique à base des réseaux de neurones artificiels; elle permet de prédire les variations de la dégradation des transistors MOSFETs fortement submicroniques.

Le quatrième chapitre présente l'applicabilité des réseaux de neurones artificiels pour la simulation des circuits électroniques nanométriques. Cette étude est basée sur la modélisation numérique bidimensionnelle des caractéristiques courant-tension d'un transistor DGMOSFET symétrique utilisant les fonctions de Green (Non-Equilibrium Green's Function).

Le dernier chapitre est consacré au développement d'un abaque basé sur un prédicteur neuronal permettant d'étudier les possibilités de graduation dimensionnelle du transistor DGMOSFET en fonction de la longueur de grille L_G, l'épaisseur du canal t_{si}, l'épaisseur de l'oxyde t_{ox}, le type de la structure (symétrique ou asymétrique) et la dopage du canal N_A.

Une conclusion synthétise les résultas obtenus et donne un aperçu de perspectives qui peuvent être développées pour mieux comprendre ce sujet.

I.1 Evolution de la technologie CMOS

Le développement de la microélectronique depuis ces 30 dernières années est véritablement spectaculaire. Ce succès résulte en grande partie d'un savoir-faire et d'une maîtrise technologique de plus en plus poussés de l'élément fondamental de la microélectronique : le silicium. Le transistor MOS (Metal Oxide Semicondictor) est à la fois le principal acteur et le vecteur de cette évolution technologique. Il est à la base de la conception des circuits intégrés à très large et ultra large échelle (VLSI- ULSI), et a mené la technologie CMOS (Complementary MOS) au rang incontesté de technologie dominante de l'industrie du semiconducteur. Au fil des années, la complexité des circuits intégrés a augmenté de façon contenue, grâce aux performances accrues des nouvelles générations de transistors MOS (TMOS). La réduction constante des dimensions de ces composants est le moteur de cette course à la performance; en fait, c'est cette volonté de toujours réduire la taille des transistors MOS qui a entraîné toute l'industrie des semiconducteurs à se surpasser et à se projeter en permanence dans le futur.

En 1973, G. Moor, l'un des co-fondateurs d'Intel avait observé que le nombre de transistors intégrés sur une même puce doublait tous les 18 mois. Cette observation l'avait alors conduit à prédire que le nombre de transistors intégrés sur une puce continuerait à doubler tous les 18 mois, jusqu'à ce que les limites physiques soient atteintes. La véracité de sa prédiction durant ces 30 dernières années a été telle que l'on s'y réfère maintenant en tant que 'Loi de Moor'. La figure I.1 illustre la validité de cette prévision. Aujourd'hui, des circuits intégrés (IC) comprenant plus de 40 millions de transistors sont produits de façon industrielle (microprocesseur). La longueur de grille des TMOS utilisés pour ces dernières générations de microprocesseurs est égales à 0.1µm, tandis que la surface de la puce varie de 80 à 150mm^2.En fait ; la diminution de longueur de grille des dispositifs a deux avantages décisifs pour les fabricants : d'une part, à puissance égale, elle permet de réduire la surface de silicium de la puce, ce qui en termes de coût, est bénéfique, et d'autre part, elle permet d'augmenter la fréquence des circuits, cette dernière étant inversement proportionnelle à la longueur de grille.

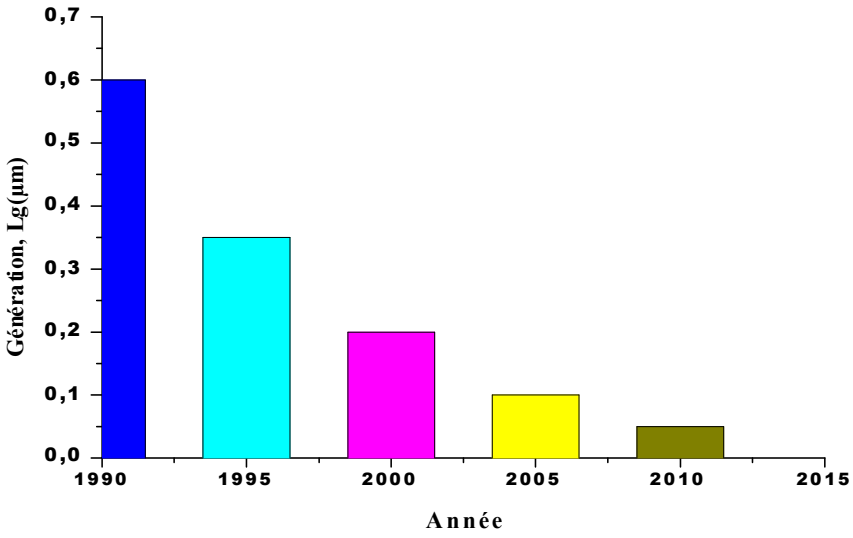


Figure I.1: Réduction d'échelle de la technologie CMOS, en accord avec la loi de Moore [5].

I.2 Contraintes pour les générations futures

A chaque nouvelle génération de transistors, la réalisation du défi lancé par la loi de Moore apparaît comme un casse-tête de plus en plus difficile à réaliser. Un compromis complexe entre la physique, la technologie et la rentabilité concentre ainsi toute l'attention des ingénieurs et des chercheurs. Des paramètres et contraintes souvent contradictoires, telles que la performance, la consommation et la fiabilité sont à prendre en compte [6,7]. Pour résumer, disons que le jeu consiste à augmenter les performances en diminuant les dimensions, sans trop augmenter la puissance dissipée à l'état du transistor.

Parier sur une croissance au rythme de la loi de Moore pour la décennie à venir relève d'un défi ambitieux. De plus, les architectures devenant très complexes, la conception, la fabrication et la vérification voient leurs coûts croître exponentiellement. Il est actuellement admis que la loi de Moore sera encore valide pour les 10-12 ans à venir (pour 3 à 4 générations de microprocesseurs). En effet, les projections industrielles pour le développement de la technologie CMOS suggèrent que cette dernière est proche des limites fondamentales physiques. L'association de l'industrie du semiconducteur : SIA (Semiconductor Industry Association), publie depuis 1998<< The International Technology Roadmap for semiconductors, ITRS >> qui est un guide de référence pour l'industrie mondiale du semiconducteur [8] (voir Tableau I.1).

Année	1999	2002	2005	2008	2011	2014
Lg (nm)	180	130	100	70	50	35
Vdd (V)	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6	0.3-0.6
Vth (V)	0.5	0.4	0.35	0.3	0.3	0.2
Tox (nm)	1.9-2.5	1.5-1.9	1.0-1.5	0.8-1.2	0.6-0.8	0.5-0.6
Na (cm^{-3})	$<10^{18}$	$<10^{18}$	10^{18}	10^{18}	10^{18}	10^{18}
X_j (nm)	45-70	30-50	25-40	20-28	13-20	10-14
E (MV/cm)	<5	5	>5	>5	>5	>5
Ion (µA/µm)	750/350	750/350	750/350	750/350	750/350	750/350
Ioff (µA/µm)	2	10	20	40	80	160

Tableau I.1: Prévision SIA de l'évolution de la technologie CMOS [8].

Selon l'édition 1999, malgré l'utilisation de nouveau matériel, il sera difficile de maintenir l'augmentation des performances électriques des composants au rythme de la loi de Moore. Il convient, cependant, de rappeler que les données du tableau I.1 sont basées sur de simples projections des progrès passés. Ceci ne garantit pas forcément qu'un dispositif plus court pourra être fabriqué, ni qu'il présentera les mêmes performances.

La figure I.2 illustre graphiquement l'évolution espérée des principales caractéristiques des TMOS, à savoir, la longueur de grille (Lg), la tension d'alimentation (Vdd), l'épaisseur d'oxyde de grille (Tox) et les profondeurs de jonctions des extensions de source et drain (X_j). Une première analyse de ces valeurs permet d'annoncer quelques possibles limitations et freins technologiques à la réduction d'échelle énoncée selon les critère de la SIA:

- La diminution de la longueur de grille en dessous de 50 nm semble difficile, compte tenu du contrôle nécessaire du courant de fuite à l'état bloqué du transistor.
- En raison de la réduction de la résistance du canal à l'état passant, il faut veiller à ce que les résistances source/drain, placées en série avec celle du canal, soient suffisamment faibles pour ne pas dégrader sérieusement les performances du composant. Cette contrainte impose donc de ne pas choisir des profondeurs de jonctions X_j trop faibles, et conduit à adopter un très fort dopage de source et de drain. Cela est cependant défavorable du point de vue des effets canaux courts car

la réduction des profondeurs de jonctions source/drain permet en fait un meilleur contrôle de la charge du canal à l'état bloqué [9].

- La tension d'alimentation ne peut que difficilement être réduite en dessous de 60V, en raison de la nécessité du maintien de la tension de seuil (Vth) à un niveau suffisant pour garantir des marges de bruit acceptables dans les circuits logiques [10].
- La réduction de l'épaisseur d'oxyde en dessous de 2nm résulte en un important courant tunnel, or vu les épaisseurs annoncées (Figure I.2) de sérieux problèmes risquent de se poser au niveau de la consommation statique. Il est admis que pour une tension d'alimentation de 1V, la limité maximale admise pour le courant de fuite de grille est de l'ordre de 1A/cm^2, ce qui situe l'épaisseur minimale d'oxyde aux environs de 1.8nm [11]. Cependant, on sait que ces courants de fuite ne perturberont pas le fonctionnement élémentaire des transistors MOS de longueur de canal inférieur à 1µm, mais en revanche, augmenteront la puissance dissipée à l'état bloqué [12]. Par ailleurs, il est clair également que la réduction des dimensions ne peut se faire sans réduire l'oxyde de grille, sous peine de ne plus parvenir à contrôler les effets canaux court [13].

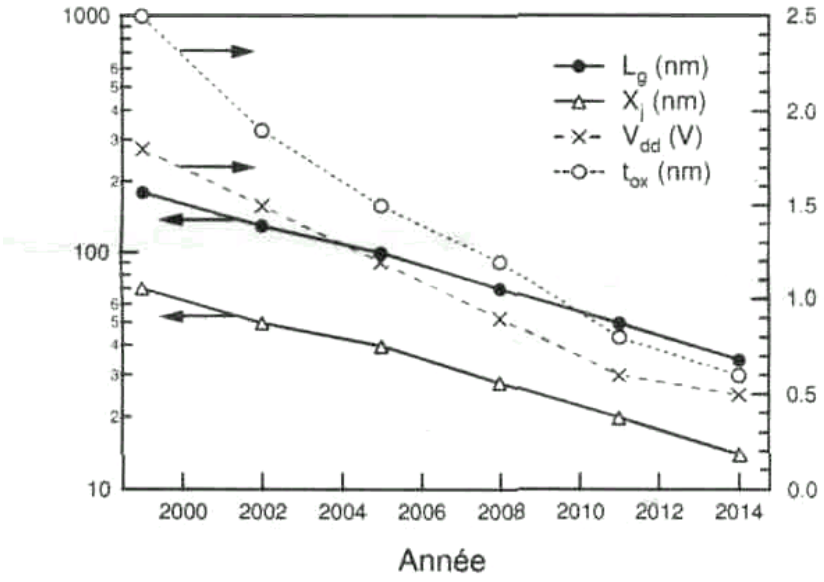


Figure I.2 : Evolution des grandeurs caractéristiques de la technologie CMOS, selon les prévision de la SIA [8].

I.3. Transistor MOSFET

Le transistor MOSFET est organisé autour d'une structure MOS suivant la disposition type représentée sur la figure. I.3. Ici, l'isolant est un oxyde (silice SiO2); deux îlots, de type opposé à celui du substrat, la source et le drain, délimitent la région active du dispositif qui se situe précisément sous l'électrode de grille. La source est le plus souvent reliée au substrat.

Le principe du MOSFET est très simple. Pour un substrat de type P par exemple, l'application sur la grille d'une tension positive par rapport au substrat fait apparaître une charge d'espace négative en surface du semiconducteur. Dés que la tension grille dépasse un certain seuil V_T, une couche d'inversion est créée et le canal N ainsi formé, de même type donc que la source et le drain et d'autant plus conducteur que Vg est grand, autorise le passage d'un courant entre les deux îlots de source et de drain. Hors de la présence de ce canal, c'est-à-dire pour $V_{GS} < V_T$, les îlots de source et de drain forment avec le substrat des jonction PN dont une obligatoirement se trouve être non passante quelle que soit la polarité de la tension de drain ; la conduction entre drain et source se trouve donc interrompue [14]. Pour le moment, nous supposons que la source et le drain sont à la masse ($Vsb = V_{db} = 0$) ; dans ce cas, trois situations peuvent être distinguées (dans la région du canal) : accumulation, déplétion et inversion.

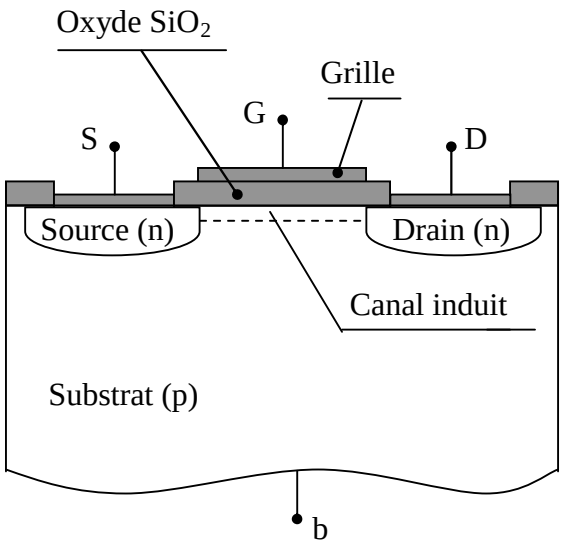


Figure I.3: Structure de base d'un transistor MOS de type n

Pour des tensions de grille négatives, les trous sont attirés à la surface et une très fine couche de charges positives (la couche d'accumulation) est alors formée avec

l'augmentation de V_{gb}, la courbure des bandes devient plus faible, jusqu'à une certaine valeur où il n'y a plus de courbure des bandes. Cette valeur particulière de tension de grille est appelée la tension de bandes plates V_{fb}. Au delà de ce point, la courbure des bandes est opposée à celle en accumulation, une charge négative est en train de se former. En fait, la charge positive à la grille repousse les trous de la surface de Silicium et fait apparaître une charge négative (due aux ions accepteurs immobiles), appelée charge de déplétion. Quand la tension de grille augmente encore plus, la courbure des bandes vers la bas devient plus prononcée. Cette courbure peut résulter en un croisement du niveau de Fermi intrinsèque Ei avec le niveau de Fermi E_{fb}. Dans cette situation, la surface du semiconducteur se comporte comme un matériau de type n, d'où le nom de région d'inversion. Une couche conductrice composée de charges négatives mobiles (électrons) est alors formée : c'est la charge d'inversion. Cette charge écartant la couche de déplétion, cette dernière n'est alors plus que faiblement dépendante de la polarisation de la grille. En conséquence, le couplage entre l'extension de la courbure des bandes dans le Silicium et l'augmentation de la tension de grille est alors fortement réduit.

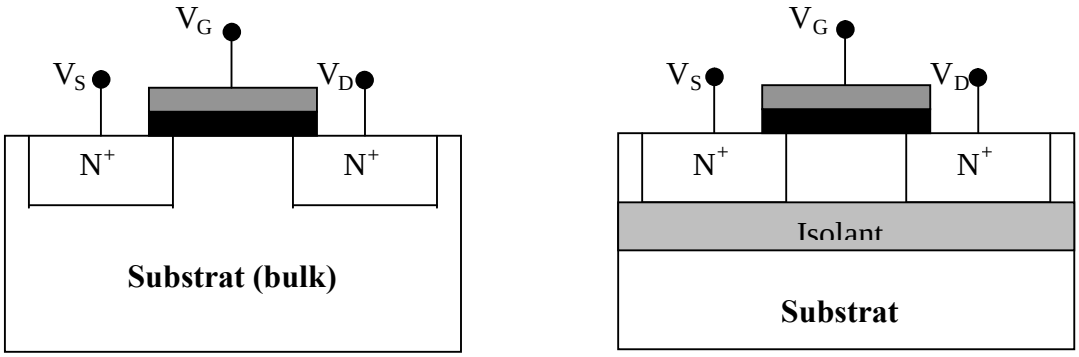
On parle d'inversion forte lorsque la densité de charge mobile dans la couche d'inversion est supérieure à la densité de charge fixe dans la couche de déplétion.
La charge d'inversion peut alors être mise en contact via les régions de source et de drain, et ainsi, un courant peut circuler dans le canal lorsqu'une différence de potentiel V_{ds} est appliquée entre le drain et la source. Puisque la charge d'inversion dépend fortement du potentiel appliqué à la grille, cette dernière peut alors être utilisée pour moduler le niveau du courant circulant dans le canal [15].

I.3.1. Technologie de fabrication du MOSFET

L'industrie de la microélectronique utilise divers types de matériaux semi-conducteurs sous forme monocristalline tel que le silicium (le plus utilisé), mais aussi des composés d'éléments des familles III/V et II/VI du tableau périodique.

Dans la grande majorité des applications, seule la partie superficielle des plaquettes du matériau semi-conducteur est utile à l'implantation de la partie active du circuit. La plus grande partie de l'épaisseur sert en effet essentiellement de support mécanique.

Deux technologies utilisant le silicium peuvent être distinguées suivant le type de supports sur lesquels sont implantées: la technologie bulk (Fig.I.4.a) et la technologie silicium sur isolant (Fig.I.4.b).



(a) (b)

Figure I.4: Vue en coupe de transistors NMOS en technologie Si-bulk (a) et SOI (b)[16]

I.3.1.a. Les méthodes de fabrication SOI

Différentes techniques ont été développées pour obtenir une couche de silicium actif isolée du substrat. Certaines méthodes d'obtention sont basées sur la croissance épitaxiale du silicium sur un isolant cristallin (héteroepitaxie). Le SOI peut être réalisé à partir d'une tranche de silicium – bulk en isolant une fine couche de Si du substrat par la formation et l'oxydation de silicium poreux (FIPOS) ou bien par implantation ionique (SIMOX). Il existe d'autres méthodes de fabrication SOI et chacune d'entre elles a ses avantages et inconvénients. Ci-dessous, nous allons décrire les trois principales méthodes.

- **Silicium sur corindon (dit aussi silicium sur saphir (SOS))**

Pour obtenir la couche de silicium active, on fait croître celle-ci par hétéroépitaxie (procédé qui consiste à faire croître une substance en couche mince sur un cristal mono cristallin de nature différente) sur un substrat monocristallin d'alumine (α-Al_2O_3). La différence de paramètre de maille cristalline (de l'ordre de 10%) conduit à une couche de silicium de faible qualité cristalline au voisinage de l'interface. Toutefois, lorsque l'épaisseur de la couche de silicium augmente, la densité de défauts cristallins décroît.

Pour cette méthode, la densité de défauts est élevée. L'inconvénient principal est l'auto-dopage du silicium par les atomes d'aluminium du substrat d'Al_2O_3.

- **SIMOX (Separation by Implanted Oxygen)**

Pour le procédé SIMOX, on crée sous la surface d'une plaquette de silicium une couche d'oxyde de silicium (SiO_2) appelée oxyde enterré par une implantation ionique

d'oxygène (Fig. I.5) à très forte dose ($1.8.10^{18}$ cm^{-2}) suivie de recuits à très haute température (>1300°C). Des couches de silicium de bonne qualité sont obtenues par cette méthode [17].

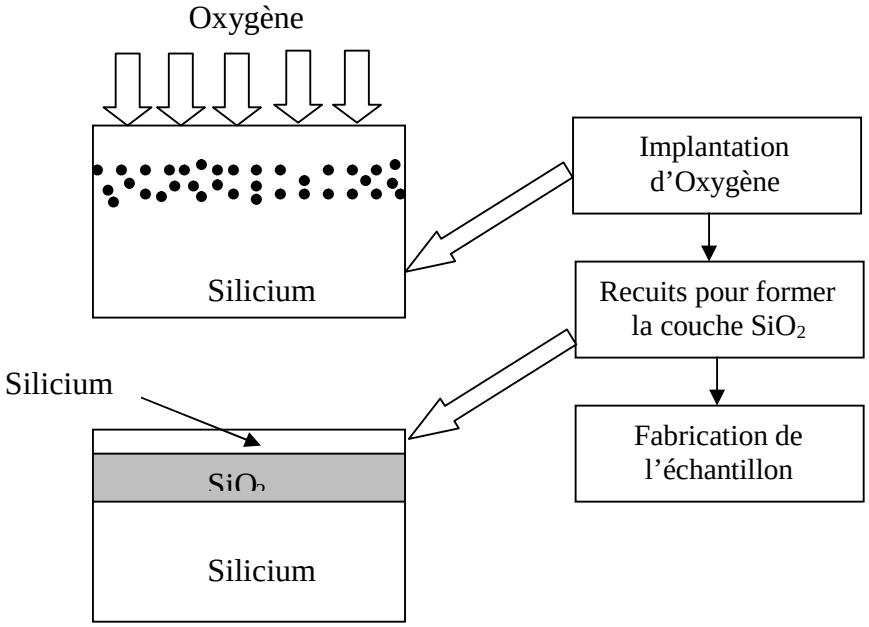


Figure I.5: Implantation ionique d'oxygène [17]

I.3.1.b Comparaison entre la technologie Bulk et la technologie SOI

Dans le cas d'un transistor MOS conventionnel en technologie Si-bulk ou massif (Fig.I.4.a), on peut constater que les zones dopées (zones actives) sont directement implantées dans une masse (bulk) de silicium épaisse dénommée « substrat ».

La présence d'un substrat épais en continuité électrique avec les couches superficielles induit des phénomènes parasites dans celles-ci et les rend sensibles notamment à des perturbations électriques (par ex. des courants de fuite vers le substrat).

Les différents inconvénients de la technologie bulk sont principalement ceux que la technologie SOI a permis de minimiser, voire supprimer grâce à l'isolation des parties actives par rapport au substrat (Fig.I.4.b).

Les courants de fuite apparaissent essentiellement au niveau des jonctions drain(source)-substrat, drain(source)-caisson et caisson-substrat. Ils provoquent une augmentation de la puissance consommée par le circuit.

Toutes les jonctions P-N étant polarisées en inverse, le courant de fuite provient de porteurs minoritaires.

Comme conséquence aux courants de fuite, nous citerons un phénomène fréquemment rencontré en technologie bulk/CMOS appelé «latchup». Celui-ci consiste au déclenchement indésirable d'un thyristor PNPN présent dans une structure constituée de deux transistors MOS complémentaires juxtaposés.

La présence de capacités parasites est liée aux différentes jonctions, C_{jSB} et C_{jDB} (Fig. I.6). En effet, lorsque deux semi-conducteurs dopés N et P sont mis en contact, il se crée de part et d'autre de l'interface une zone de charge d'espace qui aboutit à la formation d'une capacité «dynamique» lorsque la tension varie à ses bornes. Les capacités de grille (C_{GC}), de recouvrement (C_{ov}) sont inévitablement identiques pour des transistors MOS de même génération en technologies bulk et SOI. L'existence de la capacité C_{CB} en SOI dépend de l'épaisseur de la zone active du silicium.

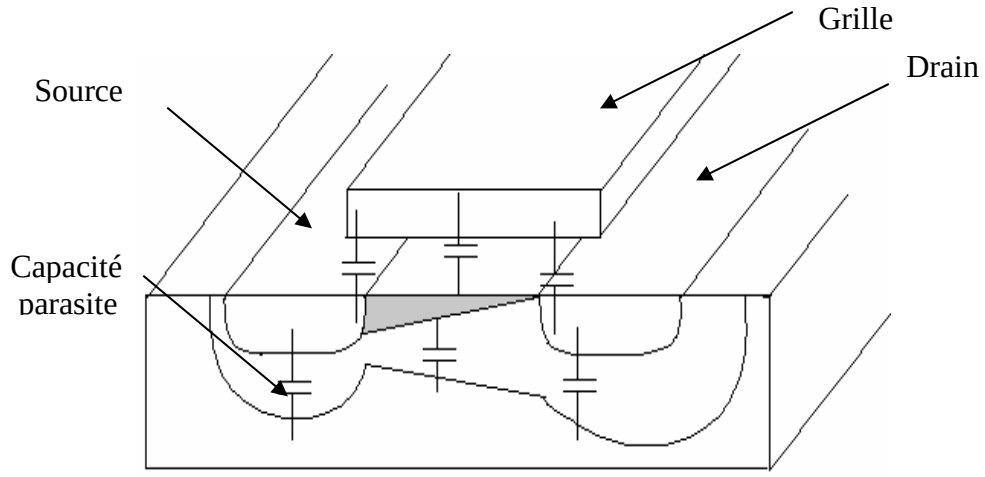


Figure I.6: Capacités parasites en technologie bulk.

En technologie SOI, la couche active superficielle est séparée électriquement de la partie massive du substrat. Cela permet non seulement de supprimer ou minimiser les inconvénients précités pour la technologie « bulk », c'est-à-dire élimination des problèmes de « latchup » et diminution de l'influence des capacités parasites C_{jSB} et C_{jDB} (Fig. I.7), mais apporte aussi des avantages comme une plus haute densité d'intégration, donc un gain de surface non négligeable.

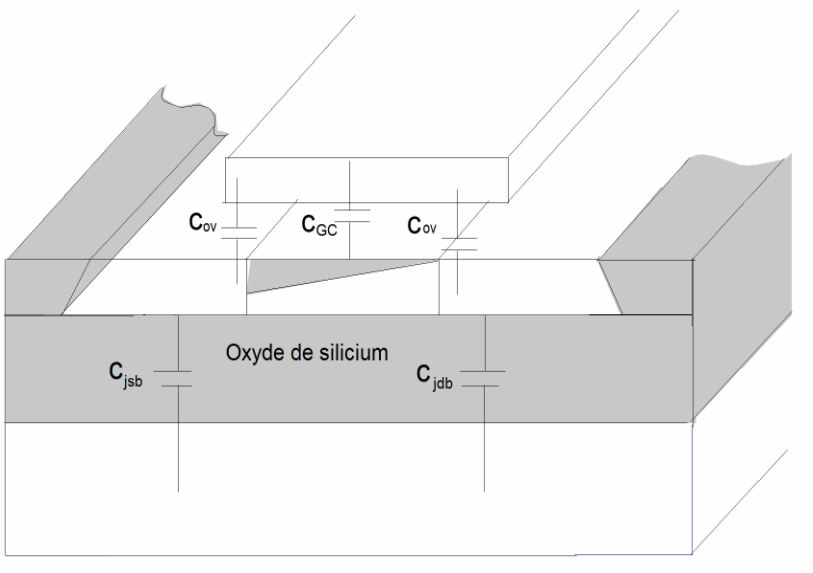


Figure I.7: Capacités parasites en technologie SOI.

I.3.1.c Technologie nanométrique

Dans la technologie nanométrique, une étape de lithographie à l'échelle nanométrique est requise pour fabriquer des nanostructures avec un contrôle de leur taille et de leur positionnement inférieurs au nanomètre [18].

Ces conditions très strictes de dimensions, sont-elles possibles à atteindre en utilisant les techniques de lithographie actuellement utilisées en microélectronique ?

L'étape de lithographie est cruciale en micro fabrication car c'est elle qui définit la géométrie et les cotes des structures. Les techniques de lithographie se scindent en deux familles: les procédés parallèles inspirés des techniques photographiques, basés sur l'utilisation d'une onde plane qui vient impressionner simultanément une surface sensible, et les procédés 'série', faisant appel au balayage d'un spot ou d'une sonde.

La lithographie UV et celle par rayons X appartiennent à la première famille, alors que les lithographies par faisceau d'électrons (EBL) ou par faisceau d'ions focalisés (FIBL) appartiennent à la deuxième.

Chaque technique de lithographie possède sa propre limite de résolution ultime et pratique. La limite pratique correspond à la résolution que les spécialistes espèrent atteindre en production, alors que la limite ultime correspond à ce qu'il serait possible d'atteindre théoriquement. Le Tableau I.2 [19], résume cet aspect:

Techniques	Type	Limite pratique	Limite ultime
Ultraviolet	Projection	150 nm	10 nm
Rayon X	Proximité	70 nm	10 nm
Faisceau d'électrons	Ecriture directe	30 nm	1 nm
Faisceau d'ions	Ecriture directe	30 nm	10 nm

Tableau I.2: Limites pratiques et théoriques des techniques de lithographies conventionnelles [19]

L'origine de la limite de résolution pratique pour l'utilisation de ces techniques à grande échelle (fabrication de plusieurs millions de transistors à la fois), est différente selon les techniques [20]:

UV: La résolution *lm* de ce procédé est directement liée à la longueur d'onde λ ainsi qu'à la distance *z* entre le masque et le substrat recouvert de résine, par la relation $lm=\sqrt{\lambda z}$. Cependant au plus la longueur d'onde utilisée est courte, au plus il devient difficile de trouver des résines photosensibles.

Rayons X : Ce procédé nécessite l'utilisation d'un masque à l'échelle 1 à cause des difficultés liées à la fabrication d'optiques performantes pour ces longueurs d'onde.
La fabrication de masque à l'échelle 1 représente un coût très élevé. De plus l'alignement du masque par rapport au substrat, devient de plus en plus critique lorsque les tailles des motifs du masque diminuent.

EBL et FIBL : Ces procédés série présentent des difficultés pour être convertis en procédés parallèles, seule configuration possible afin d'obtenir la fabrication de millions de motifs en un temps raisonnable. De plus en ce qui concerne le FIBL, seul un couplage avec un SEM (Scanning Electron Microscope) permet la visualisation et le contrôle in-situ des structures réalisées. Ce couplage rend encore plus délicat la mise au point d'un procédé série. D'ailleurs jusqu'à présent, ces deux techniques sont seulement utilisées dans les étapes de fabrication ou de réparation de prototypes.

A partir de ce tableau et des remarques précédentes, on se rend compte qu'aucune technique de lithographique ne pourra de manière certaine, atteindre la résolution nécessaire pour fabriquer soit les transistors MOSFET à leur taille minimum, soit les dispositifs élémentaires pré-sentis pour remplacer les transistors dans l'ère de la nano-électronique.
Il est donc impératif de mettre au point de nouvelles techniques permettant de relever ces

défis technologiques. Une des solutions envisagées est l'utilisation des microscopies en champ proche.

Il existe 3 principales microscopies en champ proche, la microscopie à force atomique (AFM), la microscopie par effet tunnel (STM) et la microscopie en champ proche optique (SNOM). Leur point commun repose sur l'utilisation du phénomène très local, d'interactions entre une pointe de faible rayon de courbure (de 5 à 100 nm) et une surface, interactions qui apparaissent lorsque la distance pointe-échantillon est très faible (quelques nanomètres).

La modification locale d'une surface à l'échelle nanométrique est possible sous la pointe d'un microscope à champ proche (AFM, STM et SNOM). Cette technique a même été utilisée pour positionner un unique atome sur une surface [19]. Il est alors possible d'imaginer utiliser les sondes en champ proche pour la fabrication de motifs à l'échelle nanométrique et pouvant même atteindre le contrôle atome par atome [19].

Les sondes locales ont été utilisées pour graver des résines, pour induire une oxydation locale sélective sur du silicium ou sur des couches ultra minces de métaux.

I.4 Effets de la miniaturisation

Etant donné que les dimensions des dispositifs ont considérablement diminué, beaucoup plus de transistors peuvent être intégrés sur une même puce avec la même surface. Il faut également noter que lorsque la dimension des transistors subit une réduction, ceci a une influence considérable sur les performances du circuit et notamment sur l'amélioration de la vitesse de fonctionnement.

L'amélioration de la vitesse est due essentiellement à la diminution de la longueur de canal. Il ne faut toutefois pas perdre de vue que cette diminution engendre quelques effets indésirables.

I.4.1 Injection d'électrons chauds

En régime de saturation, il existe à la jonction canal-drain polarisée en inverse, un champ électrique d'autant plus important que la longueur du canal diminue (E=V/L).
Les électrons qui pénètrent dans la zone de déplétion sont accélérés par ce champ et certains d'entre eux acquièrent suffisamment d'énergie cinétique pour créer l'ionisation par impact.

Il est à noter que l'ionisation par impact ou par choc apparaît dans un matériau pour des champs électriques intenses. Lorsqu'un électron dérive dans un solide sous l'effet d'un

champ électrique, celui-ci gagne de l'énergie cinétique et la transmet au cristal par le biais des nombreux chocs qu'il effectue avec les atomes du réseau. Ce processus assure la dissipation thermique de l'énergie potentielle perdue par les électrons (L'énergie cinétique et l'impulsion sont conservées au cours de collisions entre électrons).

Cependant, si le champ électrique est suffisamment intense, certains électrons de la bande de conduction peuvent acquérir une énergie telle que leur impact sur un atome du réseau cristallin aboutisse à la rupture d'une liaison de valence. On obtient donc deux électrons dans la bande de conduction et un trou dans la bande de valence. Ce processus peut devenir cumulatif et conduire au phénomène d'avalanche.

Les trous générés par l'ionisation peuvent emprunter différents chemins (Fig. I.8):

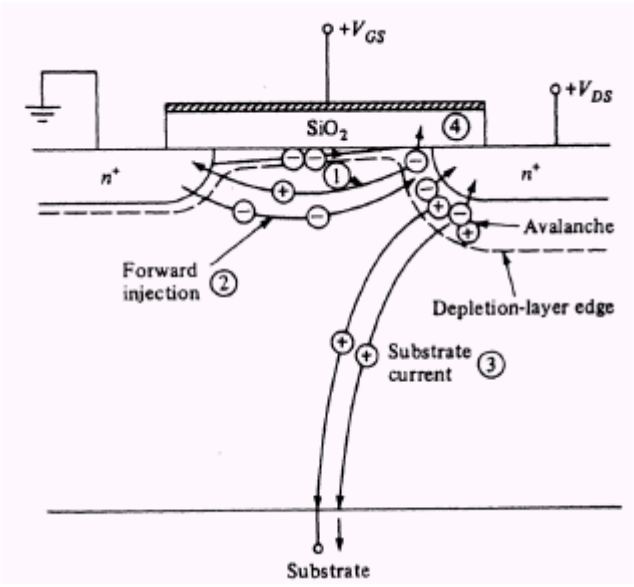


Figure I.8: Porteurs générés par ionisation par impact à la jonction canal-drain et les différentes composantes de courant parasites.

- Ils peuvent être attirés par l'électrode de substrat et donner suite à un important courant de substrat
- Certains d'entre eux peuvent également migrer vers la source et créer un abaissement de la barrière à la jonction source-canal. Il se produit alors une injection d'électrons supplémentaires de la source vers le canal.
- Cet ensemble source - canal - drain travaille comme un transistor n-p-n dont la base (canal) est flottante et le collecteur (drain) se trouve dans des conditions d'avalanche. L'accroissement du nombre d'électrons injectés de la source vers

le drain provoque une augmentation du nombre de pairs électrons-trous dans la zone de déplétion à proximité du drain. Ceci implique un abaissement de la barrière à la jonction source-canal encore plus prononcé et donc encore plus d'électrons injectés de la source. Ce phénomène est d'autant plus important que la longueur de canal (=base) diminue (effet transistor).

I.4.2 DIBL (Drain Induced Barrier Lowering)

Le phénomène de DIBL (« surface DIBL ») est pris en compte lorsque le transistor travaille en régime sous seuil (ou faible inversion) et concerne le potentiel de surface. En faible inversion, le potentiel de surface dans le canal pour des dispositifs à canal long est à peu près constant et le courant est dû à la diffusion des porteurs minoritaires (Cette diffusion est due au gradient de concentration longitudinal aux jonctions). Le courant de drain dépend exponentiellement de la tension de grille. Il existe également, en régime de faible inversion, une barrière de potentiel à la jonction entre la source et le canal qui résulte de l'équilibre entre le courant de diffusion et de dérive (cas similaire à une jonction PN à l'équilibre). Si la tension au drain augmente, la couche de déplétion s'étend de plus en plus dans le canal vers la source et il se produit un abaissement de la barrière source-canal. L'abaissement de la barrière à la source permet l'injection d'électrons au travers du canal (en surface) et ceci indépendamment de la tension de grille. Comme conséquence, en régime sous seuil, la grille perd le contrôle du courant de drain. Un courant sous seuil important peut être observé quand la longueur du canal est inférieure à 1.5µm. Cet effet est d'autant plus marqué que la tension de drain augmente et que la longueur de canal diminue Figure. I.9 [17].

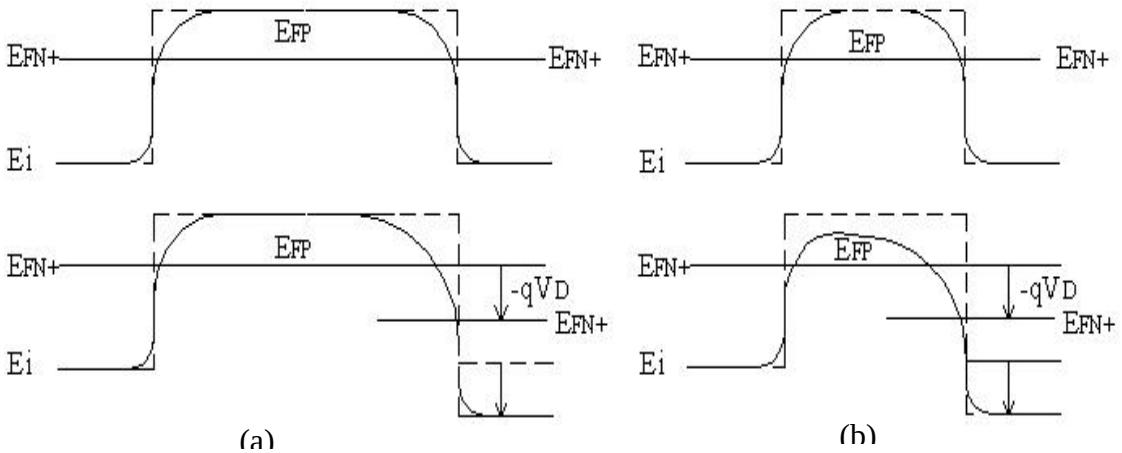


Figure I.9: Profile du potentiel de surface pour des transistors n-MOS à canal (a) long et (b) court.

L'effet DIBL est habituellement mesuré par le décalage de la courbe de transfert en régime sous seuil ΔV_{th} divisé par le ΔV_D entre deux courbes résultant de deux tensions de drain différentes :

$$DIBL = \frac{\Delta V_{th}}{\Delta V_D} \quad (mV/V) \tag{I.1}$$

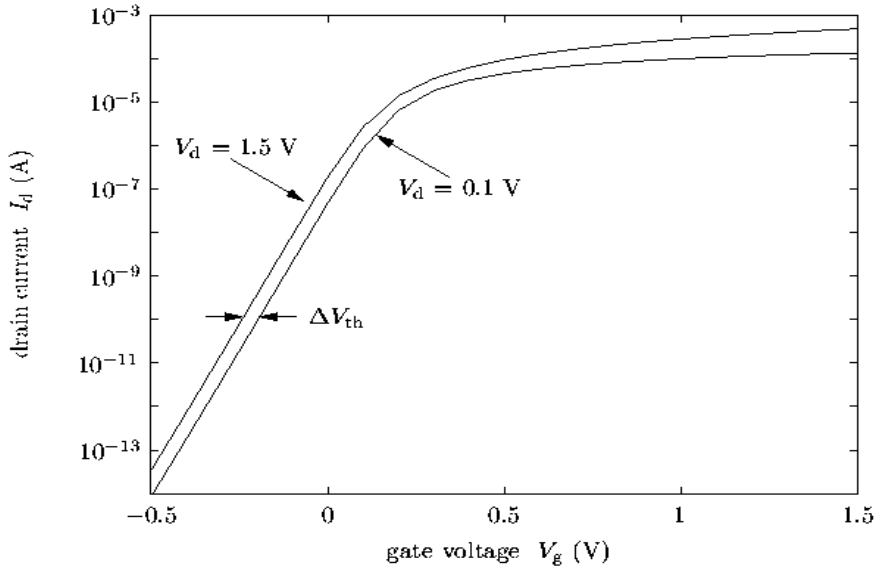


Figure I.10: Courbes de transfert pour des tension de drain de 0.1V (régime linéaire) et 1.5V (régime de saturation).

Il faut remarquer que le phénomène de DIBL se produit avant que la zone de déplétion du côté du drain (plus en profondeur dans le substrat) ne rencontre celle du côté de la source sous l'effet d'une augmentation de la tension de drain.

En vue d'atténuer les effets de canal court, il est courant d'augmenter le dopage du canal. Ceci permet de réduire l'extension de la zone de charge d'espace et donc réduit l'interaction entre le drain et la source.

I.4.3 Courant sous seuil

Le courant sous seuil est un courant qui circule entre le drain et la source du transistor, alors que la tension de grille V_{GS} est inférieure à la tension de seuil V_t. Le comportement de ce courant a évolué au fur et à mesure de l'apparition des technologies submicroniques; c'est pourquoi, il convient de distinguer deux cas selon la longueur du canal du transistor.

Dans une approximation au premier ordre, pour les transistors à canaux longs, le courant sous seuil est donné par l'équation suivante:

$$I_{DS} = \mu\, C_{ox}\, \frac{W}{L}\, \Phi_t^2 \exp^{\left(\frac{VGS - Vt}{\Phi_t}\right)} \left(1 - \exp^{\left(\frac{VDS}{\Phi_t}\right)}\right) \qquad \text{(I.2)}$$

Où $\Phi_t = \frac{kT}{q}$

$n = 1+C_D/C_{ox}$

avec :

C_D : la capacité de la couche déplétée.

C_{ox} : la capacité de l'oxyde de grille.

Comme la tension de polarisation inverse, V_{DS} , est grande devant Φ_t (=25mV à 300K), l'expression I.2 peut se mettre sous la forme mieux connue de (I.3):

$$I_{DSth} = \mu\, C_{ox}\, \frac{W}{L} \left(\frac{kT}{q}\right)^2 \exp^{\left(\frac{q\,(VGS - Vt)}{n\,kT}\right)} \qquad \text{(I.3)}$$

La caractéristique du courant sous seuil est généralement représentée par une échelle semi-logarithmique ($\log_{10}$ (I_{DSth})) en fonction de V_{GS}. Cette caractéristique est par conséquent une droite. L'inverse de la pente sous seuil (output swing) est donnée par:

$$S = \frac{n\,kT}{q} \ln(10) \qquad \text{(I.4)}$$

Plus le facteur S est petit, plus le courant sous seuil sera négligeable pour une tension de grille donnée. Il faut donc minimiser le coefficient d'effet de substrat n ($=1+C_D/C_{ox}$), c'est-à-dire n proche de 1. On peut minimiser l'effet de substrat en réduisant C_b, c'est-à-dire en diminuant le dopage du substrat (Formule I.5 et I.6).

$$C_D = \frac{\varepsilon_{SI}}{W_D} \qquad \text{(I.5)}$$

$$W_D = \sqrt{\frac{4\,\varepsilon_{si}\,\Phi}{q\,N_D}} \qquad \text{(I.6)}$$

Où

W_D : est l'épaisseur de la couche de déplétion dans le substrat .

Φ : est le potentiel de Fermi dans le substrat .

N_D : la concentration en impureté du substrat.

Habituellement, on mesure la valeur de la pente de la partie linéaire de la courbe correspondant à la conduction sous la tension de seuil. Plus la pente est élevée, plus le courant de fuite sera faible.

Pour les transistors à canal court (géométries fortement submicroniques), l'effet d'abaissement de la barrière de potentiel induite par le drain intervient, le courant sous seuil devient :

$$I_{DS} = \mu C_{ox} \frac{W}{L} \Phi_t^2 \exp\left(\frac{V_{GS} - V_t + \eta V_{DS}}{n \Phi_t}\right) . \left(1 - \exp^{-\left(\frac{V_{DS}}{\Phi_t}\right)}\right) \qquad \text{(I.7)}$$

On obtient également (pour $V_{DS} >> \Phi_t$), l'expression simplifiée:

$$I_{DS} = \mu C_{ox} \frac{W}{L} \left(\frac{kT}{q}\right)^2 \exp\left(\frac{V_{GS} - V_t + \eta V_{DS}}{\frac{n k T}{q}}\right) \qquad \text{(I.8)}$$

Le facteur ηV_{DS} diminue la valeur effective de $(-V_t)$ et par conséquent augmente la valeur du courant de fuite. Ceci équivaut à une réduction de la tension de seuil et donc à décaler la courbe $\log(I_{DSth})$ en fonction de V_{GS} (Figure I.11) vers la gauche.

La figure I.11 montre que la réduction de la longueur du canal augmente le courant sous seuil et par conséquent une dégradation au niveau de fonctionnement du transistor MOSFET sera observée.

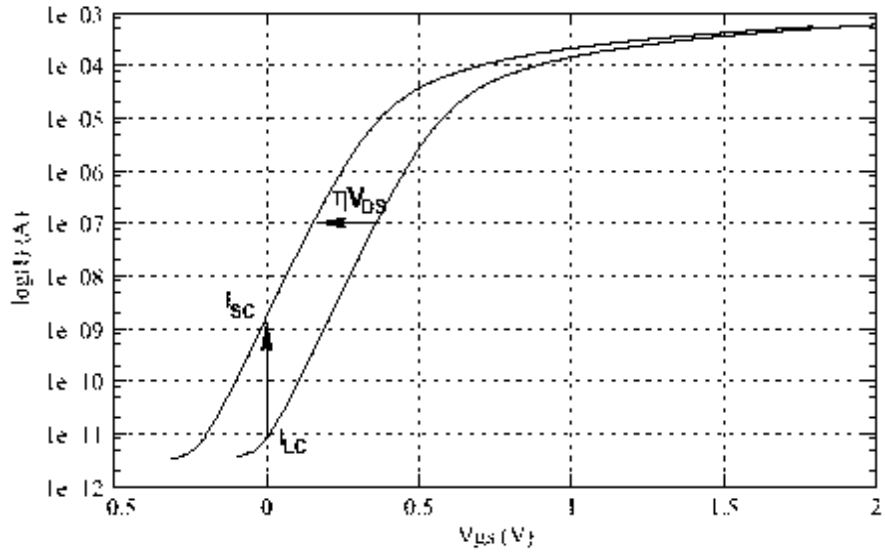


Figure I.11: Effet canal court sur le courant sous-seuil.

I.4.4 Résistances parasites à la source et au drain

Pour les transistors fortement submicroniques, les résistances intrinsèques de la source et du drain deviennent de plus en plus importantes. Etant donné que le courant augmente parallèlement à la miniaturisation, la chute de potentiel aux bornes des résistances intrinsèques devient non négligeable. De plus, ces résistances ne sont pas réduites parallèlement à la miniaturisation et peuvent même s'aggraver. Les composantes de la résistance associées à une structure MOSFET sont mises en évidence à la figure I.12. Celles-ci incluent la résistance de contact (R_{co}), la résistance de diffusion (R_{sh}), la résistance de dispersion (R_{sp}) et la résistance de la couche d'accumulation (R_{ac}).

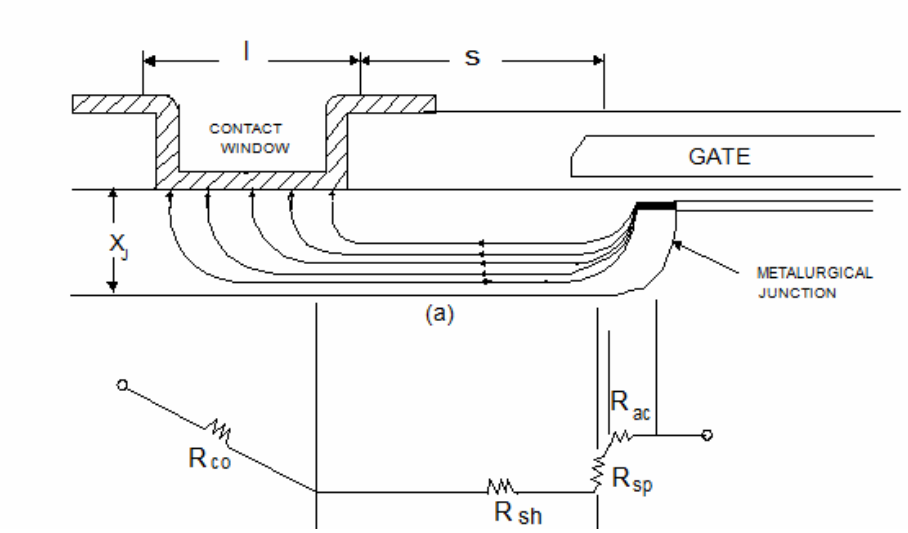


Figure I.12: Diagramme schématique pour (a) le modèle des lignes de courant dans la région source/drain et (b) les composantes de la résistance associées [21].

L'importance de la résistance de contact R_{co} dépend fortement de la répartition des lignes de courant le long de la fenêtre de contact. En effet, les porteurs cherchant à emprunter le chemin le moins résistif, la densité des lignes de courant sera beaucoup plus importante à l'extrémité droite du contact (crowded current). Le modèle utilisé pour la résistance R_{co} est par conséquent non linéaire et est établi suivant la formule:

$$R_{co} = \frac{\sqrt{\rho_c \rho_0}}{W} \coth\left(l \sqrt{\frac{\rho_0}{\rho_c}} \right) \qquad (1.9)$$

Où : l et W sont respectivement la longueur du contact et sa largeur

ρ_c est la résistivité spécifique à l'interface entre le Si et le contact

ρ_0 est la résistivité initiale à l'extrémité droite de la fenêtre de contact.

Lorsque l'on réduit la taille des dispositifs, si on veut garder une résistance de contact faible, la longueur de la fenêtre de contact ne peut diminuer dans les mêmes proportions. La dimension des contacts est un << obstacle>> à la miniaturisation.
La résistance R_{sh} est, quant à elle, simplement donnée par:

$$R_{sh} = \frac{\rho_{sh} S}{W X_j} \qquad (1.10)$$

Où : S est la distance entre le coin droit du contact et le canal.

X_j est la hauteur de la diffusion de drain (source).

En ce qui concerne l'expression de la résistance R_{sp}, celle-ci est obtenue par intégration [21] dans la zone où les lignes de courant commencent à se disperser. L'expression de cette résistance fait intervenir la somme de deux termes qui tiennent compte: d'une part (expression 1.11) de la variation de la résistivité avec le gradient de concentration (Fig.I.13) et d'autre part (expression 1.12) du profile de dispersion des lignes de courant.

La première composante de la résistance de dispersion R_{sp}, est donnée par :

$$\frac{\rho_0}{W} I_{NX} \text{ avec } I_{NX} \text{ un facteur proportionnel à } \exp(-Kx) \qquad (1.11)$$

Où : ρ_0 est la résistivité au point où les lignes de courant commencent à se disperser

K est une constante technologique pour un gradient de dopage déterminé.

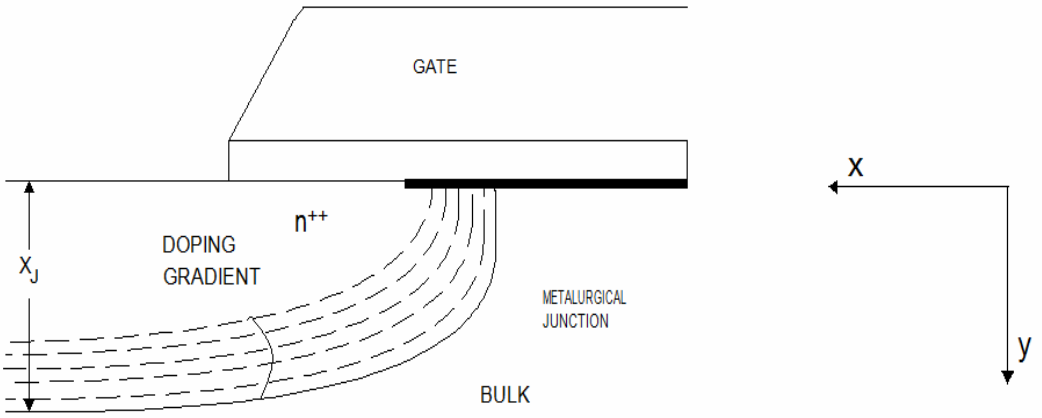


Figure I.13 : Diagramme schématique des lignes du gradient de dopage à proximité de la jonction métallurgique. Ces lignes sont aussi des lignes (équi-résistantes) [21].

La variation du dopage depuis la jonction métallurgique vers le contact est donnée par :

$$N_d(x)=N_{d0}\exp(Kx) \qquad (1.12)$$

Où *Nd0* est la concentration au point où lignes de courant commencent à se disperser. Pour une jonction abrupte, on passe d'une concentration d'atomes donneurs *Nd0* à la concentration maximum *Ndmax* sur un intervalle x très court. Pour ce type de jonction, il faut donc que le paramètre K soit négatif et à la limite égal à l'infini.

La résistivité étant inversement proportionnelle à la concentration *Nd*, on se rend compte sans difficulté (1.11) que, pour une jonction abrupte, cette composante de la résistance devient négligeable (lorsque K tend vers l'infini).

La deuxième composante de la résistance de dispersion Rsp, est donnée par :

$$\frac{\rho_b}{W}\left(\frac{X_j}{X_c}\right)I_{NX} \quad \text{avec } I_{NX} \text{ un facteur proportionnel à } \exp(-Kx) \qquad (1.13)$$

Où : ρ_b est la résistivité où la concentration en atomes donneurs est maximum.

X_c est l'épaisseur de la couche d'accumulation sous la zone de recouvrement de grille sur le drain.

Un moyen de réduire cette résistance est de diminuer l'épaisseur de la jonction X_j.

I.4.5 Effet de la géométrie du transistor sur la tension de seuil

Comme observé expérimentalement, la tension de seuil ne reste pas la même si les dimensions W et L sont réduites. Ce genre de phénomène peut être modélisé en utilisant un logiciel de simulation par éléments finis à deux dimensions comme ATLAS en vue de résoudre des équations telles que les équations de Poisson et de transport.

Toutefois, un modèle plus simple, développé par Yau [22], a permis de déterminer « graphiquement » les relations qui lient la répartition de la charge de déplétion et la tension de seuil. Ce modèle porte le nom de répartition de charge (Charge-sharing model).

La vue en coupe à la figure I.14 montre la répartition de la charge de déplétion d'un transistor MOS à canal court.

La relation habituellement utilisée pour la tension de seuil des transistors à canal long est la suivante :

$$V_t = V_{FB} + 2\Phi_F \pm \frac{Q_D}{C_{ox}} \qquad (1.14)$$

Avec :

+ pour un nMOSFET
- pour un pMOSFET

V_{FB} : est la tension de bande plate.

$2\Phi_f$: est le potentiel de surface maximum en forte inversion ($V_{GS}>V_t$).

La charge de déplétion est donnée comme:

$$Q_D = -qX_{dm}N_A \quad (Cb/m) \tag{1.15}$$

Où :

X_{dm} : est la profondeur de la charge de déplétion.

La ZCE est créée par un champ qui possède une composante longitudinale et transversale. Pour les transistors à canal long, la composante transversale du champ, c'est-à-dire le champ crée par la grille, contrôle pratiquement toute la charge de déplétion.

La composante longitudinale du champ, c'est-à-dire le champ crée par le drain, n'a que peu d'effet sur la charge de déplétion figure I.14.

Si la longueur du canal diminue, la charge de déplétion contrôlée par le drain prend de plus en plus d'importance par rapport à celle contrôlée par la grille. Cette diminution de la charge de déplétion va entraîner une diminution de la tension de seuil.

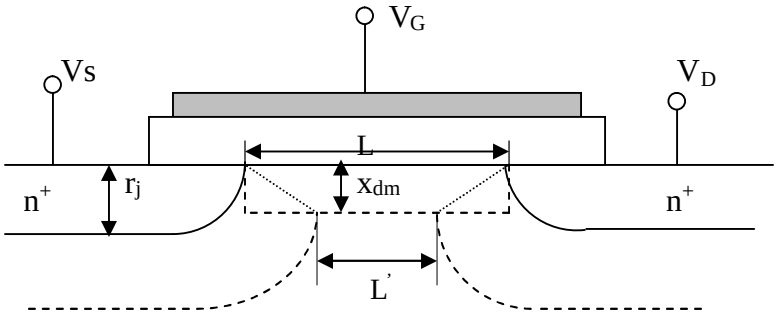


Figure I .14 : Partage de la charge de déplétion entre grille, source et drain.

En effet, la charge de déplétion effective Q'_D qui est représentée par la surface de trapèze ($S' = X_{dm}\frac{(L+L')}{2}$) est inférieure à la surface de la charge de déplétion qui est utilisée en première approximation pour les canaux longs ($S=X_{dm}.L$).

La charge Q'_D contrôlée par la grille est donnée par :

$$Q_D L = qN_A X_{dm}\frac{(L+L')}{2} \tag{1.16}$$

Par des considérations d'ordre géométrique, on peut prouver que:

$$\frac{(L+L')}{2L} = 1-\left(\sqrt{1+\frac{2x_{dm}}{r_j}}-1\right)\frac{r_j}{L} \tag{1.17}$$

La tension de seuil est donc une fonction du dopage, de L et de la profondeur de la jonction (r_j).

En remplaçant Q_D par Q'_D (1.16) dans l'expression de la tension de seuil (1.14), on obtient:

$$V_t = V_{FB} + 2\Phi_F \pm \frac{Q_D}{C_{ox}}\left[1 - \left(\sqrt{1 + \frac{2x_{dm}}{r_j}} - 1\right)\frac{r_j}{L}\right] \qquad (1.18)$$

Le model de Yau prévoit avec assez de justesse la chute de tension de seuil expérimentale (Figure 1.15)

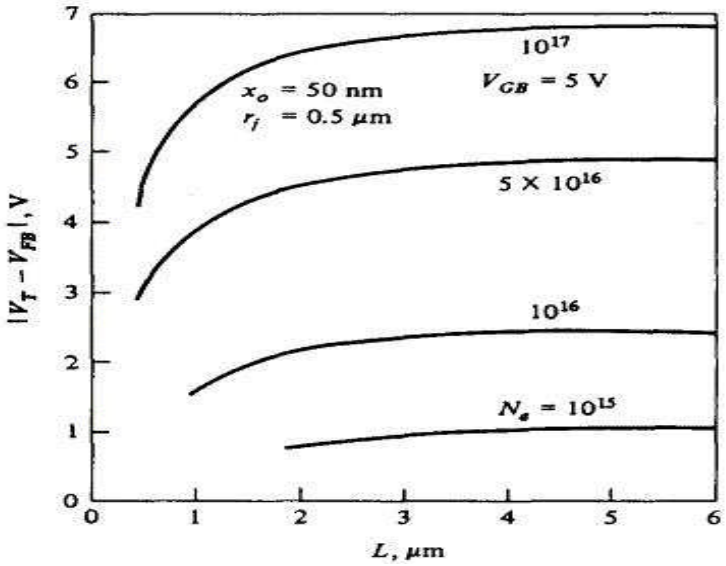


Figure I .15 : Evolution de la tension de seuil théorique en fonction de la longueur de canal selon le modèle de Yau [22].

I.5 Solutions apportées à certains effets indésirables de la miniaturisation

Nous exposerons ici les principales améliorations apportées aux dispositifs fortement submicroniques par la technologie SOI et quelques nouvelles architectures (Double-Gate MOSFET) utilisées en vue de minimiser des effets tels que la diminution de la tension de seuil, l'amélioration de l'inverse de la pente sous seuil (S), le contrôle de la grille, l'augmentation du courant d'entraînement et diminution de l'effet des électrons chauds.

I.5.1. Amélioration du contrôle de la grille sur la charge de déplétion grâce à la technologie SOI

Comme mentionné au paragraphe 1.4.5, le modèle de séparation de charge fournit la proportion de la charge de déplétion Qd contrôlée par la grille par rapport à celle

contrôlée par la source et le drain. Si on compare des transistors MOS en technologie bulk et SOI possédant les mêmes dimensions (même longueur de canal, même épaisseur de jonction…), on constate, que le rapport entre la charge de déplétion contrôlée par la grille et le drain (la source), est beaucoup plus important en technologie SOI. Ce phénomène est d'autant plus marqué que la longueur de canal diminue (Fig I.16).

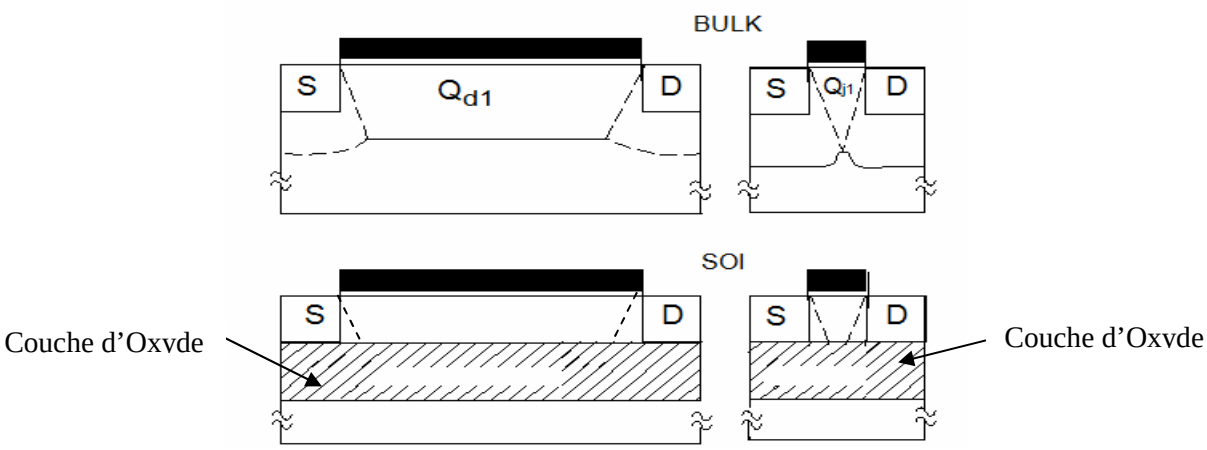


Figure I.16: Distribution de la charge de déplétion pour des transistors MOS en technologie bulk et SOI [5].

I.5.2. Diminution de l'effet des porteurs chauds

En vue de garder une tension de drain relativement élevée tout en diminuant la taille des transistors, il a été nécessaire d'introduire à proximité des source et drain une zone « tampon » dopée N mais avec une concentration en atome donneur plus faible que celle de la source et du drain.

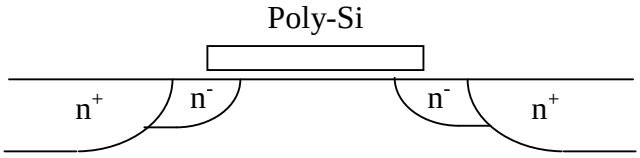


Figure I.17: Structure LDD MOS

La présence de ces régions peu dopées près du canal autour des diffusions de drain et de source (ces profiles sont désignés sous le nom de LDD ou Lightly Doped Drain) permet une meilleure répartition des zones de déplétion et donc du champ dans la structure. Les porteurs ne seront plus suffisamment accélérés pour engendrer le phénomène d'ionisation

par impact. En effet, la ZCE pourra alors s'étendre principalement dans les régions de contact et plus exclusivement dans le canal (Fig. I.17) [23].

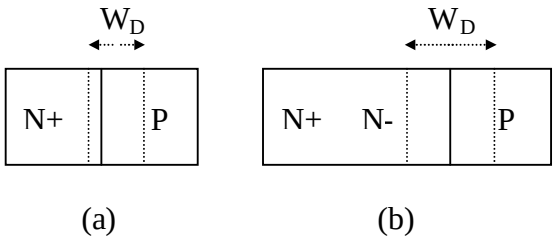


Figure I.18: Zone de charge d'espace dans (a) MOS (b) LDD MOS

I.5.3. Amélioration de la pente sous seuil

L'inverse de la pente sous seuil est donnée en toute généralité par :

$$S=\frac{\partial V_{GS}}{\partial \log_D} \qquad (1.18)$$

C'est à dire par la variation de la tension de grille par rapport au courant de canal sous seuil.

En technologie bulk, on obtenait pour S :

$S=\frac{nKT}{q}\ln(10)$ où $n=1+\frac{C_D}{C_{OX}}$ (coefficient d'effet de substrat)

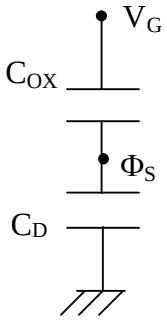


Figure I.19: Circuit capacitif équivalent d'un transistor MOS/bulk.

C_D était non négligeable et par suite $n>1$. On pouvait toutefois améliorer n en diminuant le dopage du substrat.

En technologie SOI sur film mince déplété (l'entière du film de silicium est complètement déplété avant que la tension n'atteigne la tension de seuil dans ce cas C_{SI} est une constante), S est donné par:

$S=\frac{nKT}{q}\ln(10)$ où $n=\left(1+\frac{C_{Si}}{C_{OX1}}\right)-\frac{\frac{C_{Si}}{C_{OX2}}-\frac{C_{Si}}{C_{OX1}}}{1+\frac{C_{Si}}{C_{OX2}}}$ (coefficient d'effet de substrat).

Avec C_{Si}, la capacité du film de silicium complètement déplété

$C_{ox}1$, la capacité de grille avant

$C_{ox}2$, la capacité de grille arrière

Habituellement, pour les films minces, $Cox2 << Cox1$ et $Cox2 << Csi$. On obtient donc n très proche de 1 et la pente sous seuil optimale.

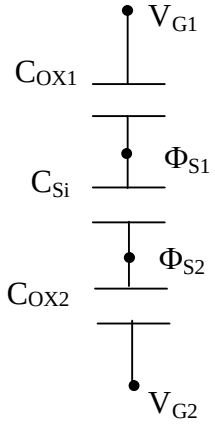


Figure I.20: Circuit capacitif équivalent d'un transistor MOS/bulk

La pente sous seuil à température ambiante est pratiquement maximale (=60mV/dec pour n=1) en SOI (Fig. I.21).

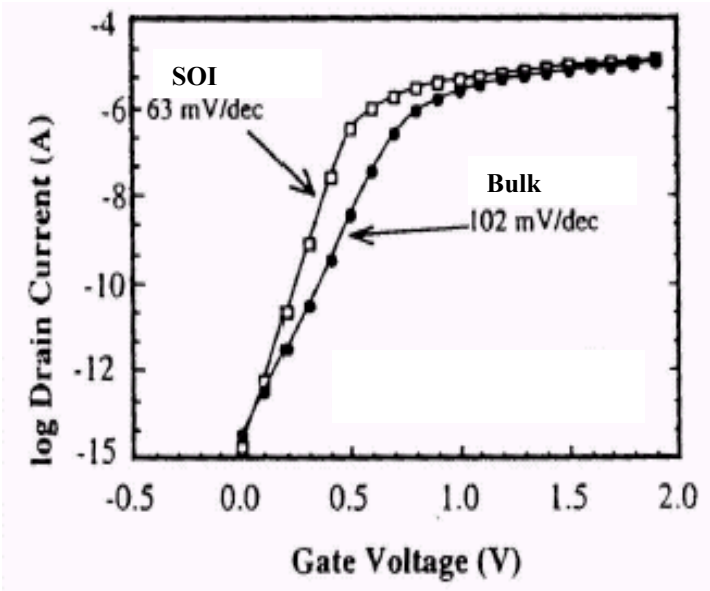
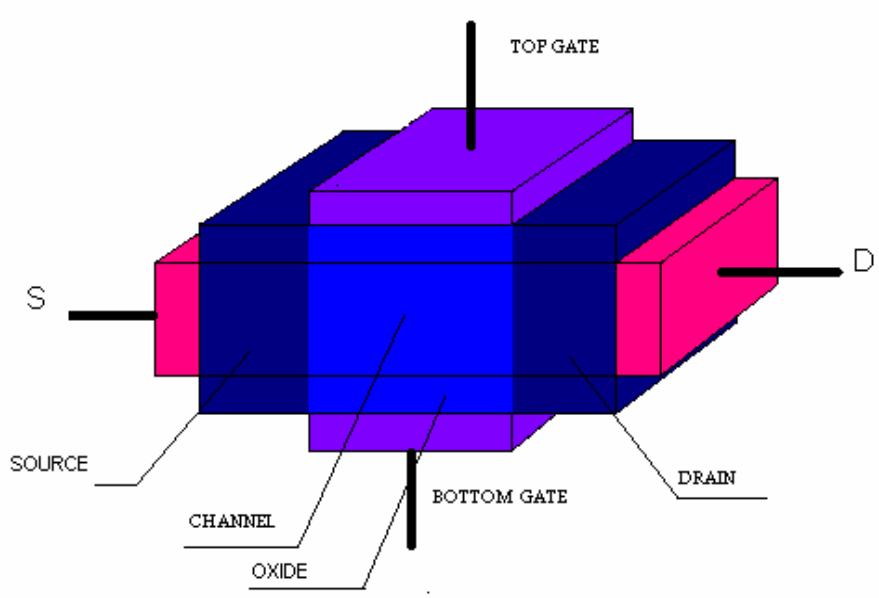


Figure I.21: Comparaison de la pente sous seuil en techno SOI et bulk [23].

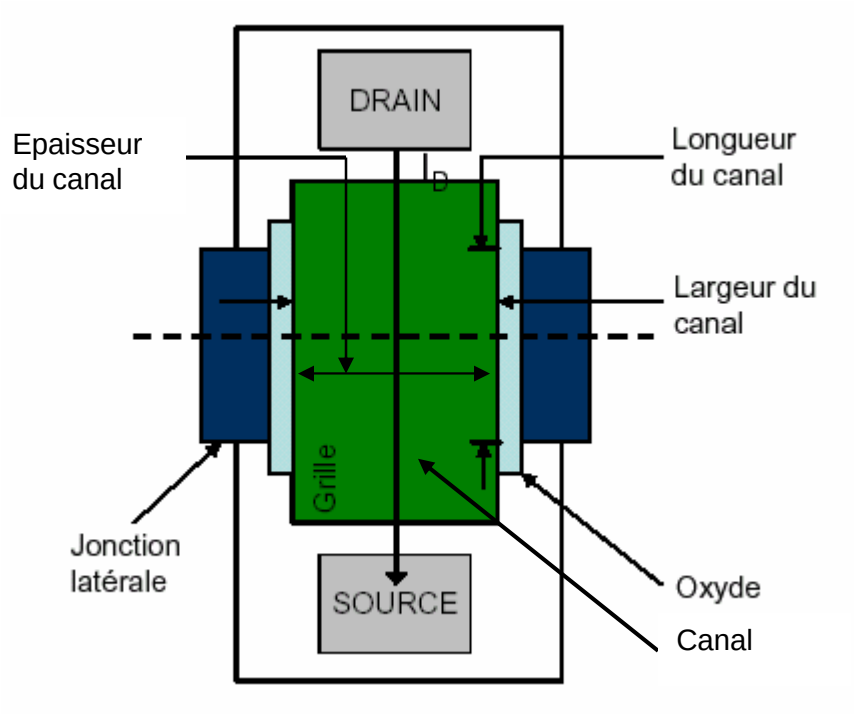
I.5.4 Transistor MOSFET à double grilles (Double-Gate MOSFET)

Cependant, les MOSFET SOI simple grille présentent un oxyde enterré très épais ($10<T_{BOX}<100$nm) dans lequel pénètrent les lignes équipotentielles de champ électrique

de la source et du drain. Il s'en suit une perte de potentiel dans l'oxyde enterré et un mauvais contrôle électrostatique de la grille pour les petites dimensions: les effets canaux courts apparaissent et dégradent à nouveau la pente sous le seuil [24]. Depuis une dizaine d'années, la plupart des études semblent indiquer que les transistors SOI MOSFET à deux, trois, voir quatre grilles sont plus adaptés à la réduction ultime des dimensions. La multiplication des grilles contrôle mieux le potentiel de canal et immunise le transistor contre les effets canaux courts observés dans une géométrie SOI simple grille [25, 26]. Pour ces géométries, un fort dopage de canal n'est plus nécessaire. L'utilisation d'une partie active en silicium ultra-fine et faiblement dopée stabilise les variations de la tension de seuil V_{TH} [27], améliore la mobilité du canal (diminution de la diffusion avec les impuretés ionisées) et fait tendre la pente sous le seuil S vers la valeur idéale de 60 mV/décade. Dans ce contexte, et afin de surmonter les contraintes imposées par la miniaturisation du transistor MOSFET en deçà de 30nm. Le transistor MOSFET à double grille (DG MOSFET) illustré dans la figure I.22 a été identifié par ITRS (International Technology Roadmap for semiconductors) en tant que la structure la plus prometteuse qui permet davantage de graduation dimensionnelle de CMOS au-delà de 65nm pour son courant d'entraînement plus élevé, la pente sous seuil améliorée, la conductivité pour les canaux courts et la flexibilité remarquable de conception des circuits intégrés à l'échelle nanométrique [28-29].



(a)



(b)

Figure I.22: (a) Transistor DGMOSFET (b) Vue de dessus du transistor DGMOSFET

I.6 Intelligence artificielle

Partant du principe que le transistor MOSFET fortement submicronique est un dispositif électronique complexe, de phénomènes électriques et physiques (la dégradation, l'effet de la miniaturisation, l'effet quantique,….) non linéaire et présentant des difficultés de son étude et de sa modélisation, il peut être étudié en employant des méthodes statistiques relevant du domaine de l'intelligence artificielle [30].

En effet, cette modélisation intelligente s'affranchit potentiellement des difficultés des modèles standard qui consistant à:

- admettre des hypothèses simplificatrices pour établir les équations différentielles représentant les relations entre les paramètres et les réponses;
- résoudre ces équations différentielles et leur donner un sens physique ou être confronté aux problèmes de convergence des méthodes numériques;
- adopter une approximation faible (dépendance linéaire) autour de certaines plages de paramètres, comme c'est le cas des plans d'expériences;
- choisir une meilleure approximation (non linéaire) en étant confronté au problème de la stabilité des solutions et leur forte dépendance à la nature des paramètres d'entrée.

I.7 Conclusion

La miniaturisation des transistors MOS et plus particulièrement la diminution de la longueur de canal a permis d'augmenter la densité d'intégration et la vitesse de fonctionnement des circuits. Cette réduction des dimensions a engendré des phénomènes parasites (DIBL, modification de la tension de seuil, augmentation du phénomène de porteurs chauds,…) qui détériorent les caractéristiques courant-tension. Toutefois, les technologues ont imaginé des procédés de fabrication particuliers en vue de conserver ces caractéristiques (technologie SOI, LDDMOSFET, DGMOSFET).

II.1 Introduction

Les réseaux de neurones artificiels ont été introduits en 1943 par Mcculloch et Pitts [31]. C'est un concept mathématique, dit neuromimétique, qui s'inspire du mode d'analyse et de transmission de données dans les cellules neurobiologiques pour permettre la résolution de problèmes complexes [32].

Du point de vue structural, un réseau de neurones est composé d'un certain nombre d'unités de traitement simples appelées neurones formels ou artificiels. Ces derniers sont connectés entre eux de façon à produire la réponse correspondant aux entrées reçues par le réseau. Plusieurs modèles de neurones artificiels ont été développés, s'inspirant du principe de fonctionnement de neurone biologique qui assure essentiellement les fonctions suivantes:

- réception des signaux provenant des neurones voisins;
- intégration de ces signaux;
- génération d'une réponse;
- transmission de celle-ci à d'autres neurones.

L'approche neuronale parfois appelée 'connexionniste', s'oppose à l'approche symbolique basée sur l'hypothèse sur laquelle le raisonnement modélisant la pensée est une combinaison de symboles à des règles logiques. Elle privilégie les avantages suivants:

- l'activité parallèle et en temps réel pour de nombreux composants;
- la représentation distribuée des connaissances;
- l'apprentissage par modification des connections.

Les applications des réseaux de neurones artificiels dans le domaine d'étude des dispositifs à semiconducteur sont limitées [33]. Ces applications concernent des problèmes de nature non linéaires (mobilité, l'effet du champ appliqué sur vitesse des porteurs de charge,…) avec un nombre important de paramètres à prendre en compte.

II.2 Eléments de base des réseaux de neurones

II.2.1 Le modèle neurophysiologique

L'élément de base du système nerveux central est le neurone. Le cerveau se compose d'environ mille milliards de neurones, avec 1000 à 10000 synapses (connexions) par neurone. Le neurone est une cellule composée d'un corps cellulaire et d'un noyau (Fig.2.1). Le corps cellulaire se ramifie pour former ce que l'on nomme les dendrites. Celles-ci sont parfois si nombreuses que l'on parle alors de chevelure dendritique ou

d'arborisation dendritique. C'est par les dendrites que l'information est acheminée de l'extérieur vers le soma (corps du neurone). L'information est traitée alors par le corps cellulaire. Si le potentiel d'action dépasse un certain seuil, le corps cellulaire répond par un stimulus. Le signal transmis par le neurone chemine ensuite le long de l'axone (unique) pour être transmis aux autres neurones. La transmission entre deux neurones n'est pas directe. En fait, il existe un espace intercellulaire de quelques dizaines d'Angströms entre l'axone du neurone afférent et les dendrites du neurone efférent. La jonction entre deux neurones est appelée synapse.

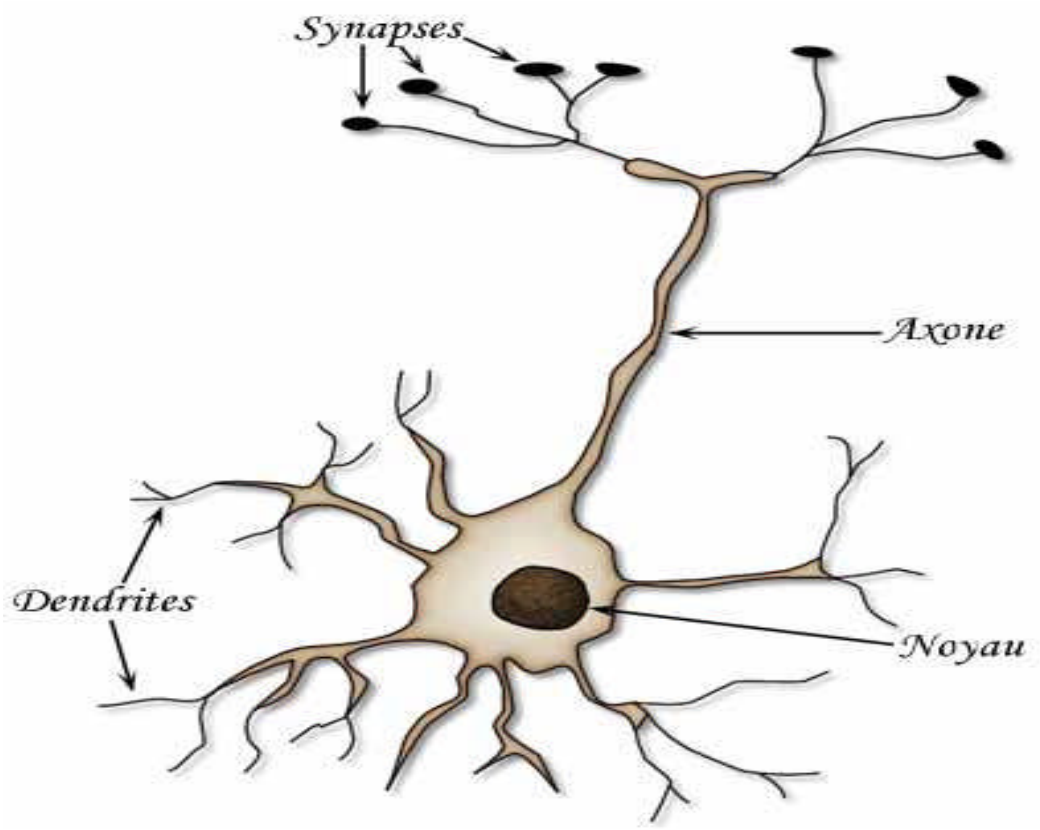


Figure II.1: Un neurone biologique [34]

II.2.2 Modèle de base de neurone artificiel: le neurone formel

Le modèle couramment utilisé du neurone formel est le suivant: un neurone est une unité de calcul élémentaire recevant ses entrées du milieu extérieur et/ou d'autres neurones et transmettant sa sortie à d'autre neurones et/ou au milieu extérieur. Ces entrées sont pondérées par un poids synaptique qui modélise l'intensité des connexions (synapses) entre les neurones. Chaque neurone additionne ses entrées, préalablement multipliées par les

poids associés, et génère la sortie correspondante à travers une fonction non-linéaire appelée fonction d'activation.

Un neurone artificiel, qu'il soit émulé d'une manière logicielle ou implanté matériellement, comprend donc les éléments suivants:

1. L'ensemble des connexions (synapses) permettant la réception des signaux d'entrée et la transmission du signal de sortie. Chacune des connexions est caractérisée par un poids synaptique, de façon que le signal transmis par un neurone source soit multiplié par le poids associé à la connexion avant d'être reçu par le neurone destination. Les poids synaptiques peuvent être entiers ou réels, positifs ou négatifs selon que la synapse correspondante soit excitatrice ou inhibitrice respectivement.
2. L'additionneur qui réalise la somme des signaux d'entrée pondérés par les poids synaptiques.
3. La fonction d'activation qui est une fonction non-linéaire de saturation servant à limiter l'amplitude du signal de sortie du neurone. Plusieurs types de fonctions d'activation peuvent être utilisés, les plus courants sont donnés sur la figure II.2.

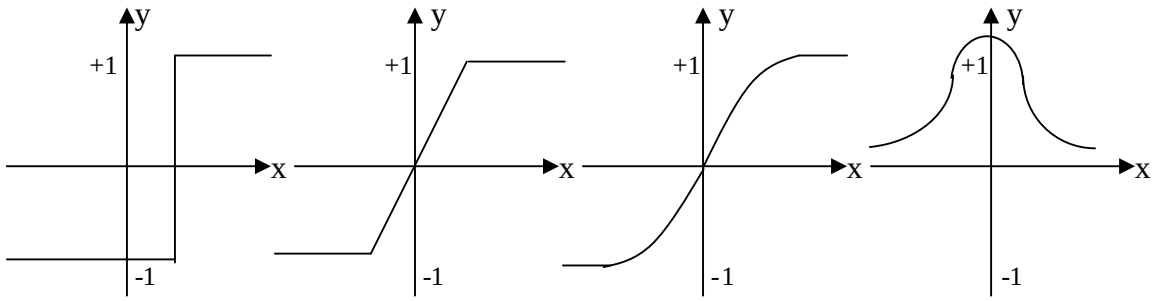


Figure II.2: Différents types de fonctions de transfert pour le neurone artificiel. a) fonction à seuil du neurone de Mc Culloch et W. Pitts (1949), b) linéaire par morceaux du modèle Adaline de Widrow et Hoff (1960), c) sigmoïde d'un réseau Perceptron Multi Couches de Rosenblatt (1962), d) gaussienne du réseau RFR de Moody et Darken (1989) [38].

Du point de vue mathématique, un neurone k est décrit par les deux équations suivantes:

$$U_k = \sum_{j=1}^{n} W_{kj} . I_j \qquad (2.1)$$

$$O_k = \Theta(U_k - \theta_k) \qquad (2.2)$$

Où:

- les I_j et O_j représentent respectivement les signaux d'entrée et la sortie du neurone;
- les W_{kj} sont les poids des connexions reliant le neurone j (source) au neurone k (destination);
- U_k est la somme des signaux d'entrée I_j pondérés par les poids W_{kj} ;
- θ_k est le seuil du neurone, il peut être considéré comme une entrée externe ayant pour effet de modifier la valeur d'entrée de la fonction d'activation;
- $(U_k - \theta_k)$ est l'entrée totale de la fonction d'activation.

La figure II.3 décrit le fonctionnement global d'un neurone artificiel.

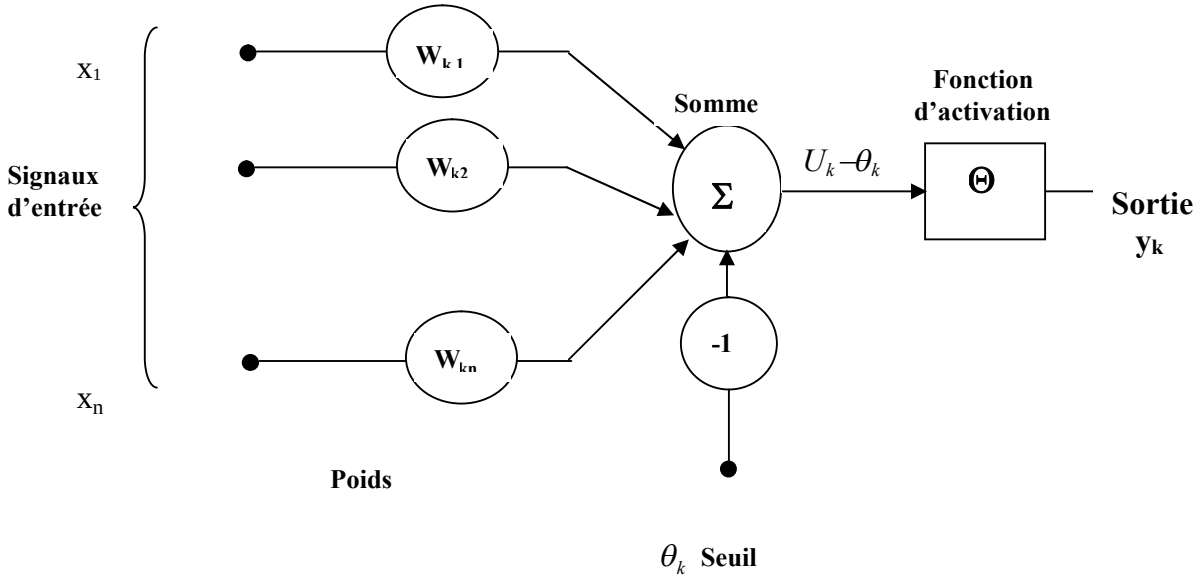


Figure II.3: Modèle général d'un neurone [35]

II.3 Architecture d'un réseau de neurone

Nous entendons par architecture ou topologie d'un réseau de neurones artificiels la manière selon laquelle les neurones sont organisés. Les structures qui peuvent être utilisées sont très variées mais beaucoup moins complexes que celles des réseaux de neurones biologiques.

D'une façon générale, l'architecture des réseaux des neurones artificiels peut aller d'une connectivité totale (chacun des neurones du réseau est relié à tous les autres) à une connectivité locale où les neurones ne sont liés qu'à leurs plus proches voisins. Il est

courant d'utiliser des réseaux à structure régulière pour faciliter leur utilisation (Fig.2.4, 2.5 et 2.6).

Deux classes différentes d'architectures de réseaux de neurones peuvent être distinguées :

1. les réseaux proactifs (feed-forward).
2. les réseaux récurrents.

II.3.1 Réseaux proactifs

Cette classe se distingue par l'absence de toute boucle de rétroaction de la sortie vers l'entrée, d'où l'appellation 'feed-forward'. En d'autres termes, la propagation des signaux s'y fait uniquement dans le sens de l'entrée ver la sortie.

Ce type de réseaux comprend deux groupes d'architectures : les réseaux mono-couches et les réseaux multicouches. Ils différent par l'existence ou non de neurones intermédiaires appelées neurones cachés entre les unités d'entrées et les unités de sorties appelées nœuds sources ou nœuds d'entrée et nœuds de sortie respectivement.

II.3.1.1 Réseaux proactifs monocouches

Ce type de réseau possède une couche d'entrée recevant les stimuli à traiter par l'intermédiaire des nœuds sources. Cette couche se projette en une couche de sortie composée de neurones (nœuds de calcul) transmettant les résultats du traitement au milieu extérieur. La figure II.4 montre, par exemple, un réseau proactif monocouche à 4 nœuds d'entrée et 3 nœuds de sortie. La désignation monocouche est attribuée à la couche de sortie (nœuds de calcul). La couche d'entrée n'est pas comptée dans ce sens vu qu'il n'y a pas de calcul fait par les nœuds sources, ils servent uniquement à recevoir les signaux d'entrée et à les transmettre à la couche suivante. Un exemple classique de réseau monocouche est le perceptron qui est un réseau proactif monocouche à une seule sortie. Ce type de réseau ne peut réaliser que la discrimination de classes d'entrées linéairement séparables.

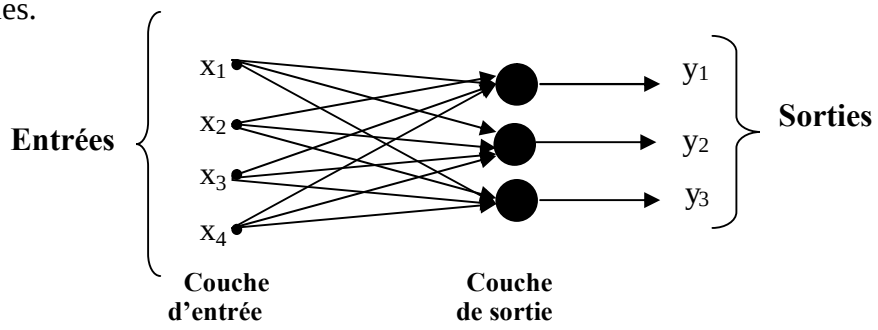


Figure II.4 : Réseau proactif monocouche (perceptron)

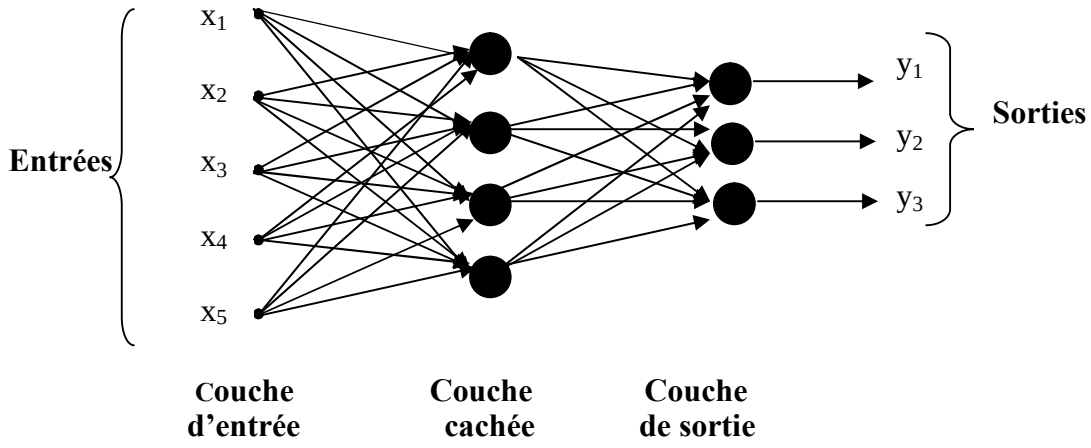
II.3.1.2 Réseaux proactifs multicouches

Ce type de réseaux proactifs se caractérise par la présence d'une ou plusieurs couches cachées, dont les nœuds de calcul correspondants s'appellent neurones cachés ou unités cachées. Les couches cachées s'interposent entre l'entrée du réseau et sa sortie. Leur rôle est d'effectuer un prétraitement des signaux d'entrées, reçus par la couche d'entrée en provenance du milieu extérieur, et de transmettre les résultats correspondants à la couche de sortie où sera déterminée la réponse finale du réseau avant qu'elle soit transmise au milieu extérieur.

Ce rôle de prétraitement fait que, en ajoutant une ou plusieurs couches cachées, le réseau est capable d'extraire plus de propriétés statistiques que celles extraites d'un réseau similaire ayant moins de couches cachées. Ceci est utile pour réaliser des fonctions plus complexes que de simples séparations linéaires.

Dans ce type de réseaux, les entrées des neurones d'une couche particulière proviennent uniquement des sorties de la couche adjacente précédente. Les réseaux les plus fréquemment utilisés de cette catégorie sont les perceptrons multicouches (Multilayered Perceptrons-MLP) [35].

Dans la figure II.5 est montré l'exemple d'un réseau à une seule couche cachée comportant 5 unités d'entrée, 4 unités cachées et 3 unités de sortie (réseau 5-4-3).Ce réseau est dit complètement connecté dans le sens que chaque nœud d'une couche est connecté à tous les nœuds de la couche adjacente suivante. Si éventuellement, des connections manquaient entre des neurones de deux couches voisines, le réseau serait dit partiellement connecté.



FigureII.5 : Réseau proactif complètement connecté avec une seule couche cachée

II.3.2 Réseaux récurrents

Les réseaux récurrents se distinguent des réseaux proactifs par le fait qu'ils contiennent au moins une boucle de contre-réaction des nœuds de sortie vers les nœuds d'entrée (ou, au moins d'une couche vers une couche précédente, adjacente ou non).
Dans la figure II.6 est illustré l'exemple d'un réseau de neurones récurrent ayant 2 unités d'entrée, 3 unités cachées et 2 unités de sortie. Dans ce réseau, les connexions de rétroaction proviennent aussi bien des unités cachées que les unités de sorties.

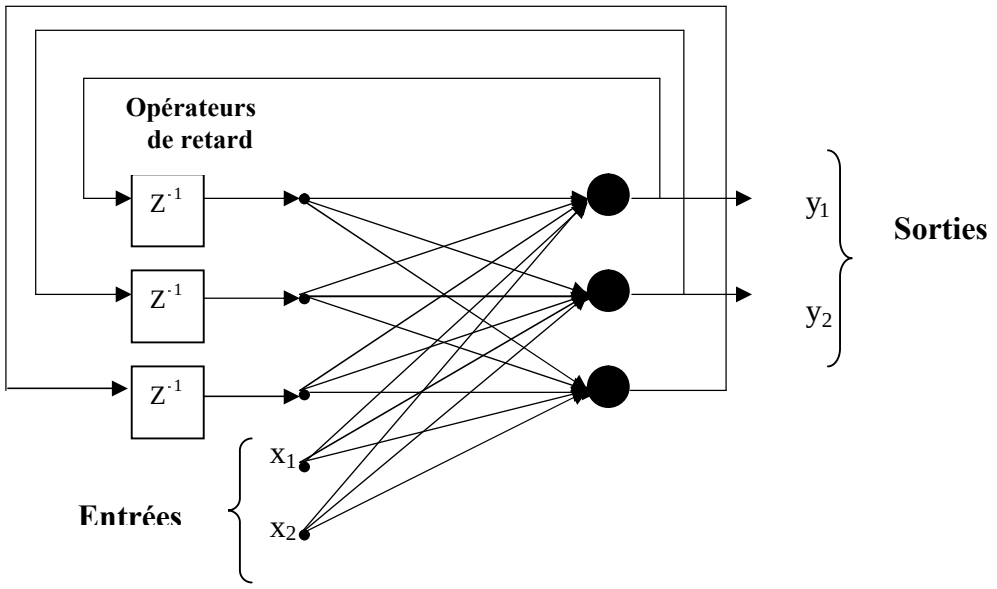


Figure II.6 : Réseau récurrent avec neurones cachés [36]

Contrairement aux réseaux proactifs, les propriétés dynamiques du réseau récurrent sont importantes. En effet, la présence des boucles de contre-réaction implique l'insertion, dans le graphe correspondant à la topologie du réseau, de branches particulières composées d'opérateurs de retard. Ces derniers résultent en un comportement dynamique non linéaire du réseau dû à la nature non linéaire des neurones. Dans certains cas, les états des neurones subissent un processus de relaxation faisant évoluer le réseau vers un état stable dans lequel ces états ne peuvent plus changer. Dans l'autre cas, le changement des états des neurones de sortie est important de façon à ce que le comportement dynamique constitue la sortie du réseau. Des exemples de ce type de réseaux sont les réseaux de Hopfield [37] et la machine de Boltzmann [38].

II.3.3 Mode de fonctionnement du réseau de neurones

Un neurone donné est relié aux autres neurones par des connexions (poids) qui sont des nombres représentant l'intensité des liens dans la structure. Si un poids donné s'approche de 0, le lien dans le réseau est inutile.

Les neurones du vecteur d'entrée représentent une exception vu qu'ils sont caractérisés par une seule entité, l'entrée et la sortie du neurone se confondent dans ce cas. La valeur que prend un neurone de la couche d'entrée est donc directement proportionnelle à la valeur du paramètre à prendre en compte dans la structure.

Dans ces conditions, la notion du flux dans un réseau de neurones est reliée directement à la relation (2.3) :

$$I(x_i)=w(x_i,y_j)O(y_j)=O(x_i,y) \quad j=1,N_y \qquad (2.3)$$

où $w(x_i,y_j)$ représente le poids entre le neurone i et le neurone j.

La fonction d'activation la plus courante permettant de traiter les paramètres réels sans débordement de leur valeur est la fonction sigmoïdale (Fig.2.2-c-). Compte tenu de cette transformation, la sortie d'un neurone peut s'écrire :

$$O(x_i)=f\left[I(x_i)\right] \qquad (2.4)$$

Le mode de fonctionnement des neurones impose l'ajustement de paramètres pour optimiser la structure [39]. Ces paramètres peuvent se grouper en plusieurs catégories [39-40], le mode de connexion, la définition des couches et l'énergie du système :

- **mode de connexion :** cette catégorie spécifie les connexions entre les neurones. Le mode le plus connu et recommandé en science des matériaux à semiconducteurs (modélisation des composants électroniques) est le mode proactif (alimentation en avant) [39].
- **Définition des couches :** cette définition comprend plusieurs paramètres :
  - *nombre de couches* : c'est un paramètre critique qui n'est pas connu à l'avance [41]. En général, une couche cachée ne suffit pas de répondre complètement à un problème non linéaire [32]. Cependant, des études montrent que plus la taille du réseau augmente, plus l'optimisation du réseau devient difficile [42];
  - *nombre de neurones* : c'est un paramètre également critique qui peut être optimisé pendant ou auprès la phase d'apprentissage ;

- *algorithme d'apprentissage* : il constitue un domaine de recherche à lui seul. C'est l'algorithme qui permet d'ajuster les valeurs des poids dans la structure. Cet ajustement intervient lors de l'alimentation du réseau par des exemples entrée/sortie issus d'une base de données reliant les paramètres du problème aux réponses attendues. Le plus connu des paradigmes d'apprentissages s'appelle la rétropropagation [32] dont une variante, la rétropropagation rapide, a été utilisée dans cette étude [42-36].
- *Prétraitement des données* : il s'agit de normaliser les exemples entrée/sortie pour homogénéiser les valeurs propagées dans le réseau de neurones. Ces données sont formatées en tenant compte de la dimension du système, c'est à dire en considérant les valeurs minimales et maximales des paramètres et des réponses [42]. L'expression (2.5) est utilisée dans tous les calculs relatifs à cette étude:

$$y=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \qquad (2.5)$$

où y est l'expression formatée du paramètre permettant d'exprimer les valeurs de ce dernier entre 0 et 1. $x_{\min}$ et $x_{\max}$ représentent les limites physiques du paramètre x;

- énergie du système : l'ajustement des poids des neurones se fait à un cycle donné dans le sens de la minimisation de l'erreur entre les cas soumis au réseau de neurones et la réponse donnée par celui-ci. Cette erreur exprime l'énergie du système qui peut s'écrire sous la forme:

$$J_t=\frac{1}{2}(\vec{r}_0-\vec{r}_t)^2 \quad , \quad \vec{r}_0\,(r_k)_{k=1,N_Z} \quad , \quad \vec{r}_t\,(o(z_k))_{k=1,N_Z} \qquad (2.6a)$$

$$\Delta w_{kl}=-\alpha\frac{\partial J_t}{\partial w_{kl}} \qquad (2.6b)$$

où l'expression (2.6a) représente l'erreur quadratique au temps t entre $\vec{r}_t$ (vecteur de sortie prédit à l'instant t) et $\vec{r}_0$ (le vecteur de sortie du cas soumis au réseau). L'équation (2.6b) exprime la variation du poids w_{kl} (lien entre le neurone k et le neurone l) en fonction du gradient de l'énergie selon le modèle de rétropropagation. r_k représente un élément de la réponse prédite. $o(z_k)$ représente un élément de la réponse expérimentale. k est l'indice de la réponse dans le vecteur de sortie. Le

paramètre α représente la vitesse d'apprentissage, c'est à dire l'amplitude de changement dans la valeur du poids.

II.4 Avantages et inconvénients

Les réseaux de neurones présentent les caractéristiques suivantes :

- la capacité à représenter n'importe quelle corrélation entre les paramètres d'un problème et ses réponses ;
- aucune hypothèse de dépendance n'est requise pour le traitement des données ;
- diverses représentations de la réponse sont possibles comme le classement parmi plusieurs catégories, la réponse par oui ou non, la réponse par valeur réelle, etc.. ;
- la prise en compte de la variabilité des paramètres et des réponses est effective ;
- la possibilité de représenter des évolutions intermédiaires constituant l'aspect prédictif est intégrée dans l'optimisation ;
- l'apprentissage est continuel, garantissant un système ouvert.

Par contre, les réseaux de neurones présentent quelques inconvénients, parmi lesquels :

- la découverte des corrélations sans analyse des phénomènes physiques qui sont à la charge de l'expert ;
- la nécessité d'une connaissance préalable du problème à traiter, pour choisir avec pertinence les paramètres, dimensionner la base de données et pour définir les limites du systèmes.

II.5 Algorithme d'apprentissage

L'algorithme d'apprentissage par rétropropagation rapide est utilisé dans cette étude pour optimiser la valeur des poids dans les réseaux de neurones [42]. Le formalisme développé ci-dessous s'applique à un réseau de neurones artificiel présentant deux couches cachées, mais il peut se généraliser pour un nombre plus important de couches.

Soient x, y et z, les indices respectifs relatifs à la première couche cachée, à la deuxième couche cachée et au vecteur de sortie, respectivement. L'énergie à un nœud donnée du réseau peut s'exprimer par :

$$\nabla J_z(y_j, z_k) = \frac{\partial J_z}{\partial w(y_j, z_k)} \qquad (2.7)$$

où z_k est un neurone du vecteur de sortie, y_j un neurone de la deuxième couche cachée, J est l'énergie du système définie par l'équation (2.6a) et $w(y_j,z_j)$ le poids liant le neurone y_j et z_k.

En utilisant la propriété des dérivées partielles, l'expression (2.7) peut s'écrire alors :

$$\nabla J_z(y_j,z_k)=\frac{\partial J_z}{\partial I(z_k)}\frac{\partial I(z_k)}{\partial w(y_j,z_k)} \qquad (2.8)$$

En remplaçant dans (2.8) les expressions (2.3), (2.4) et (2.6a), il s'en suit :

$$\nabla J_z(y_j,z_k)=f'\left[I(z_k)\right](r_k-O(z_k)).\frac{\partial\left[w(y_j,z_k)O(y_j)\right]}{\partial w(y_j,z_k)} \qquad (2.9a)$$

$$\nabla J_z(y_j,z_k)=f'\left[I(z_k)\right](r_k-O(z_k)).O(y_j) \qquad (2.9b)$$

où f' exprime la première dérivée da la fonction d'activation et $O(y_j)$ la valeur de sortie de la couche y vers le neurone j.

L'expression (2.9b) exprime la variation de l'énergie du système à la sortie du réseau. Cette variation se calcule facilement puisque tous les termes sont connus. Toutefois, la procédure devient compliquée quand il s'agit de rétropropager l'énergie dans le réseau jusqu'à l'entrée de ce dernier. Le protocole suivant présente le mode de calcul retenu permettant de préciser l'énergie à chaque couche du réseau de neurones.

La variation de l'énergie entre le neurone x_i de la première couche cachée et le neurone y_j de deuxième couche cachée peut s'exprimer par une équation similaire à l'expression (2.8) suivant :

$$\nabla J_z(x_i,y_j)=\frac{\partial J_y}{\partial O(y_j)}\frac{\partial O(y_j)}{\partial I(y_j)}\frac{I(y_j)}{\partial w(x_i,y_j)} \qquad (2.10)$$

En considérant des expressions similaires à (2.3) et (2.4) pour la deuxième couche cachée y, il s'en suit :

$$\nabla J_z(x_i,y_j)=\frac{\partial J_y}{\partial O(y_j)}f'\left[I(y_j)\right]O(x_i) \qquad (2.11)$$

La dérivée partielle du second membre peut alors se réécrire selon l'expression:

$$\frac{\partial J_y}{\partial O(y_j)}=\frac{\partial J_y}{\partial I(z_k)}\frac{\partial I(z_k)}{\partial O(y_j)} \quad k=1.N_z \qquad (2.12)$$

En introduisant la population des poids entre la couche y et la couche z, (2.12) devient :

$$\frac{\partial J_y}{\partial O(y_j)}=\frac{\partial J_y}{\partial I(z_k)}\frac{\partial\left[w(y_j,z_k)O(y_j)\right]}{\partial O(y_j)}=\frac{\partial J_z}{\partial I(z_k)}\,w(y_j,z_k) \quad k=1,N_z \qquad (2.13)$$

En remplaçant les dérivées partielles, (2.13) devient :

$$\frac{\partial J_y}{\partial O(y_j)} = f'\left[I(z_k)\right](r_k - O(z_k))\ w(y_j, z_k) \quad k=1,N_z \tag{2.14}$$

D'où :

$$\nabla J_y(x_i, y_j) = f'\left[I(z_k)\right](r_k - O(z_k))\ w(y_j, z_k)\ f'\left[I(y_j)\right] O(x_i) \quad k=1,N_z \tag{2.15}$$

Cette dernière équation a une signification globale, en cela qu'elle exprime la variation d'énergie à un nœud donné (entre les neurones x_i et y_j dans le cas présent) en fonction des paramètres donnés en sortie du réseau, d'où l'appellation 'rétropropagation'.

Compte tenu de l'équation (2.15), la correction des poids dans le modèle 'rétropropagation rapide' peut s'écrire :

$$\Delta w(x_i, y_j)^{tn} = \frac{\nabla J_y(x_i, y_j)^{tn}}{\nabla J_y(x_i, y_j)^{tn-1} - \nabla J_y(x_i, y_j)^{tn}}\ \Delta w(x_i, y_j)^{tn-1} \tag{2.16}$$

où tn représente le nombre de cycles.

Dans cette expression, qui est une amélioration de (2.6b), la variation d'un poids est fonction de l'amplitude de la variation dans le cycle précédent et du sens de la variation de l'énergie. Cette expression permet donc de modérer la vitesse de changement des poids en fonction de la position sur la courbe de l'énergie (FigureII.7).

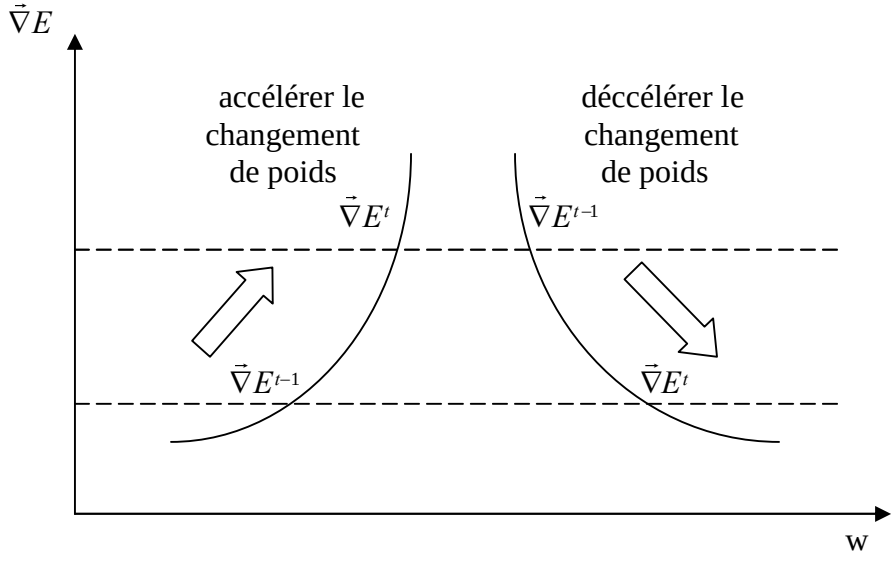


Figure II.7: Mode de changement des poids dans la rétropropagation des poids dans la rétropropagation rapide.

III.1 Introduction

La technologie VLSI est basée sur la réduction des dimensions du transistor MOS. Bien que cette miniaturisation ait de grand intérêt, elle est à l'origine de l'apparition de nouveaux problèmes liés à la fiabilité du composant. Cette réduction accroît les champs électriques à l'intérieur de TMOS, et les porteurs du canal acquièrent alors une énergie suffisante pour surmonter la barrière du potentiel de l'interface Si-SiO_2 [43]. Un pourcentage de ces porteurs est alors piégé dans l'oxyde et/ou à l'interface. Il en résulte une dégradation globale des paramètres déterminant les performances du TMOS. Cette dégradation, au cours du fonctionnement, est appelée vieillissement [44].

Pour atténuer le phénomène de vieillissement. Dans le présent chapitre, on propose un modèle de prédiction à base des réseaux de neurones capable de prédire les variations de la dégradation de transistors MOSFETs fortement submicroniques en fonction des quatre paramètres à savoir : la longueur de la grille, la tension de drain, la tension de grille et le temps du stress. Ce chapitre peut être divisé en deux parties; dans la première partie, on propose un dispositif expérimental assisté par ordinateur permettant d'étudier les aspects expérimentaux des phénomènes du vieillissement des transistors MOSFETs fortement submicroniques, la deuxième partie est consacrée au développement d'une approche analytique à base des réseaux de neurones artificiels qui permet de prédire les variations de la dégradation des transistors MOSFETs fortement submicroniques.

III.2 Conception et réalisation du dispositif expérimental

III.2.1 Partie Hardware

La mesure de la dégradation des transistors MOSFETs submicroniques nécessite un appareillage adéquat pour obtenir des résultats précis et reproductibles. Pour atteindre cet objectif, nous avons essayé, lors de la conception du dispositif expérimental, de tirer profit des possibilités offertes par l'informatique industrielle. Cette technique nous a permis d'améliorer la précision des résultats par rapport à ceux obtenus par les techniques classiques.

Notons aussi que lorsqu'on procède à la mesure par les techniques classiques, on passe par une longue suite d'opération pour déterminer les différents paramètres. C'est pour ces raisons que l'utilisation de l'informatique industrielle devient plus qu'indispensable.

Dans le présent chapitre, on propose un dispositif expérimental assisté par ordinateur permettant d'étudier les aspects expérimentaux des phénomènes du vieillissement des transistors MOSFETs submicroniques.

Le schéma bloc du dispositif expérimental conçu et réalisé est illustré par la Fig.III.1.

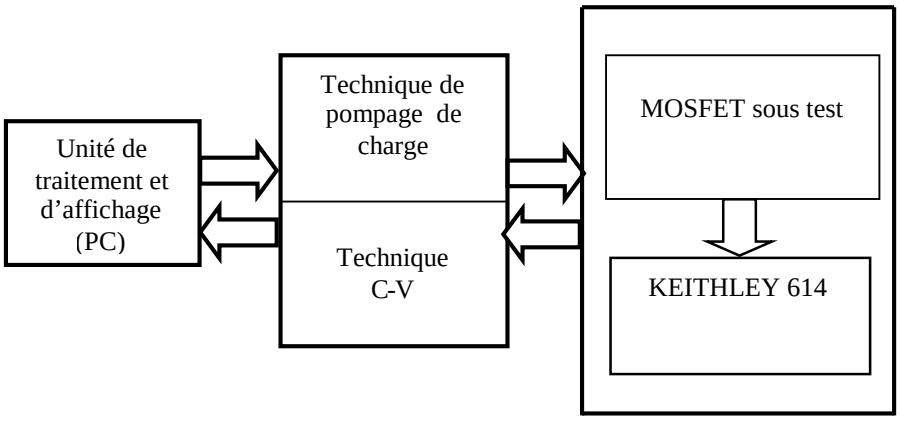


Figure III.1: Schéma bloc du dispositif expérimental

III.2.1.1 Technique C-V

Le schéma synoptique suivant illustre le système de contrôle et d'acquisition conçu et réalisé pour remplir la tache prévue par cette technique (figure III.2).

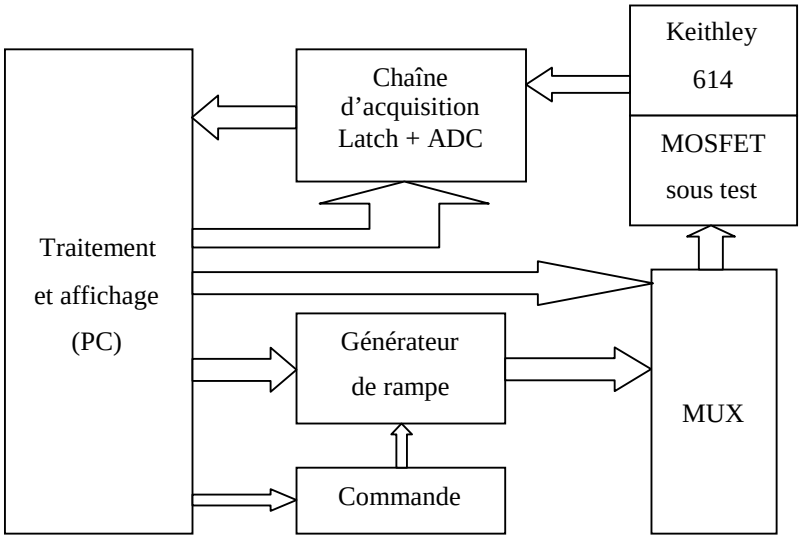


Figure III.2 : Schéma synoptique du système de pilotage et d'acquisition (Technique C-V)

III.2.1.2 Principe de fonctionnement du système

Le système réalisé est destiné à générer :

- Une rampe avec une possibilité de faire varier la vitesse de balayage et l'amplitude (basse fréquence).
- Une rampe modulée par un signal HF avec la possibilité de faire varier la vitesse de balayage et l'amplitude pour la caractérisation haute fréquence.
- Les tensions de stress pour chaque mode de mesure.
- Le temps du stress fixé par l'utilisateur.

Le montage décrit dans ce travail utilise un générateur de rampe entouré par de nombreux périphériques intégrés (amplification, multiplexage, latch, convertisseur analogique-numérique,…), voir figure III. 3.

L'information issue (la variation de la capacité de la structure MOS en fonction de V_{gs} mesurée par le Keithley 614) se présentera alors au circuit de conditionnement où elle sera amplifiée et convertie par un ADC0804.

Le déclenchement de la conversion parviendra du PC via la broche D4 du PC connectée à la ligne $\overline{WR}$ du CAN ($\overline{RD}$,$\overline{CS}$ toujours à la masse). L'octet transféré sera stocké dans le latch 74HC241, afin d'être transmis sous forme de quartet vers le PC.

L'élément principal du dispositif est le générateur de rampe que nous avons conçu et réalisé au laboratoire. Les caractéristiques principales de ce générateur sont les suivantes :

- Une tension minimale Vmin ajustable dans le domaine [–5 , 0V].
- Une tension maximale Vmax ajustable dans le domaine [0 ,+5V].
- Une vitesse de balayage ajustable dans le domaine [50, 550mV/s].
- La remise à Vmin.

III.2.1.3 Conception du générateur de rampe

Le générateur que nous avons conçu et réalisé peut être illustré par la figure suivante:

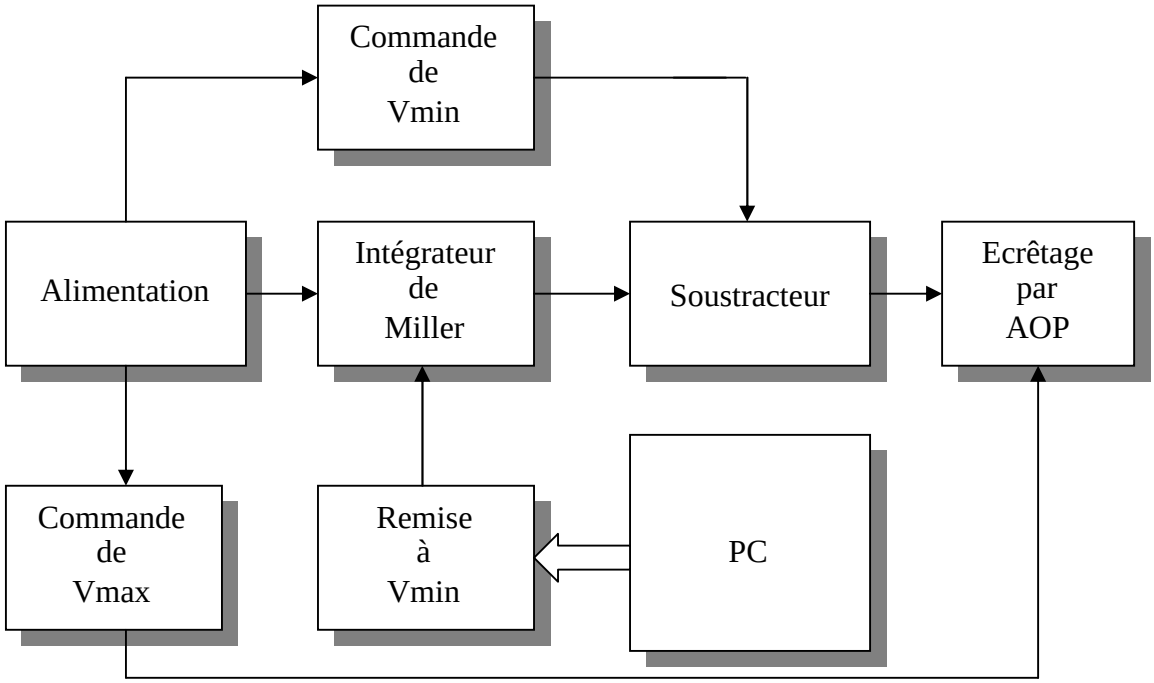


Figure III.3 : Schéma synoptique du générateur de rampe

III.2.1.4 Principe de fonctionnement du générateur de rampe

Le générateur que nous avons conçu est destiné à générer une rampe possédant les caractéristiques mentionnées dans le paragraphe II.2.1.2. Le générateur ainsi conçu est basé sur trois modules :

- Le module d'alimentation qui assure l'alimentation et la polarisation des différents blocs et composants.
- Le module principal (analogique), qui est le module le plus important dans notre système ; est responsable de la génération de la rampe. Cette dernière passe par les étapes suivantes :

  Après avoir attaqué l'intégrateur de MILLER (Figure III.6. (B1)) avec une tension constante, on obtient une rampe à la sortie qui a une pente ajustable par l'intermédiaire d'une résistance variable.

  Cette rampe va attaquer une entrée d'un soustracteur après avoir été inversée. La deuxième entrée du soustracteur est une tension constante qu'on peut ajuster grâce au système potentiomètre-suiveur (Figure III.6. (B1)). Pour cela, on peut

fixer le niveau du départ de la rampe (commande de Vmin). La commande de la tension maximale Vmax est assurée par le système potentiomètre-suiveur. Ce dernier délivre une tension ajustable qu'on exploite dans la polarisation positive d'un autre suiveur de tension qui a pour entrée le signal provenant du soustracteur. Alors, on obtient à la sortie de ce module une rampe ayant les caractéristiques déjà mentionnées auparavant. Voir la figure.III.6.B1.

- Le module de commande de la rampe (la remise à Vmin) parviendra du PC à travers la broche D2 du PC connectée au transistor.

III.2.1.5 Multiplexage

Si l'on veut réaliser la polarisation de la structure étudiée par plusieurs signaux analogiques différents, il est possible d'utiliser des circuits appelés 'Multiplexeurs Analogiques' (Figure III.6.B2) permettant d'aiguiller une entrée parmi toutes les entrées vers la structure à caractériser [45].

III.2.1.6 Conversion analogique numérique

Effectuer une conversion analogique-numérique, c'est rechercher une expression numérique dans un code déterminé, pour représenter une information analogique. Un convertisseur A/N est un dispositif qui reçoit un signal analogique et le transforme en un signal numérique [45] (Figure III.6.B3).

III.2.1.7 Technique de pompage de charges

Cette technique est fondée sur l'existence d'un courant pompé apparaissant au niveau du substrat d'un transistor MOS lorsque la grille est soumise à une polarisation faisant passer alternativement la structure d'un état bloqué à un état saturé (Figure III.4). Le mécanisme physique qui est à l'origine de ce courant est un processus de génération-recombinaison par l'intermédiaire des états électroniques localisés à l'interface silicium-isolant de grille. Pour détecter ces niveaux on balaye énergiquement la bande interdite, et en variant la polarisation de la grille de l'inversion à l'accumulation, afin de remplir puis vider ces pièges [44].

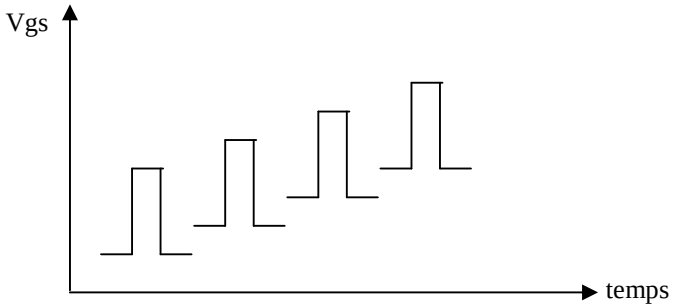


Figure III.4: Signal de polarisation de la grille

Le schéma synoptique suivant illustre le système de pilotage et d'acquisition conçu et réalisé pour remplir la tache prévue par cette technique.

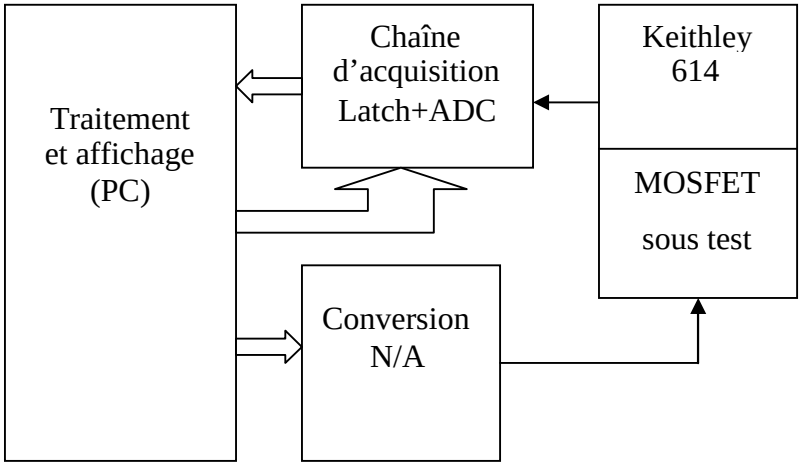


Figure III.5: Schéma synoptique du système de pilotage et d'acquisition (Technique de pompage de charges)

III.2.1.8 Principe de fonctionnement du système

Notre dispositif utilise une technique de génération de signaux (tension de la grille) appelée synthèse numérique directe ou DDS. Le principe de cette technique peut être donné comme suit:

Les données contenant les valeurs d'amplitudes d'un cycle complet du signal à générer sont téléchargées à partir de l'ordinateur. Au fur et à mesure que les données changent, un convertisseur numérique-analogique (CNA) 8 bits convertit les données en signal de polarisation de la grille (Figure III.7.B1). Afin d'avoir une résolution en tension du signal, les données d'amplitudes créées par le programme utilisent la plage complète

des codes du CNA (allant de 0 à 256). On affecte donc la valeur 0 au niveau bas, et la valeur 256 au niveau haut. Le courant substrat est mesuré à partir d'une unité de source et de mesure de courant-tension Keithley 614, qui permet de mesurer des courants de l'ordre de 10^{-14}A. Le déclenchement de la conversion parviendra du PC via la broche PIN16 du PC connectée à la ligne $\overline{WR}$ du CAN ($\overline{RD}$,$\overline{CS}$ toujours à la masse). L'octet transféré sera stocké dans le latch 74HC241, afin d'être transmis sous forme de quartet vers le PC (Figure III.7.B2).

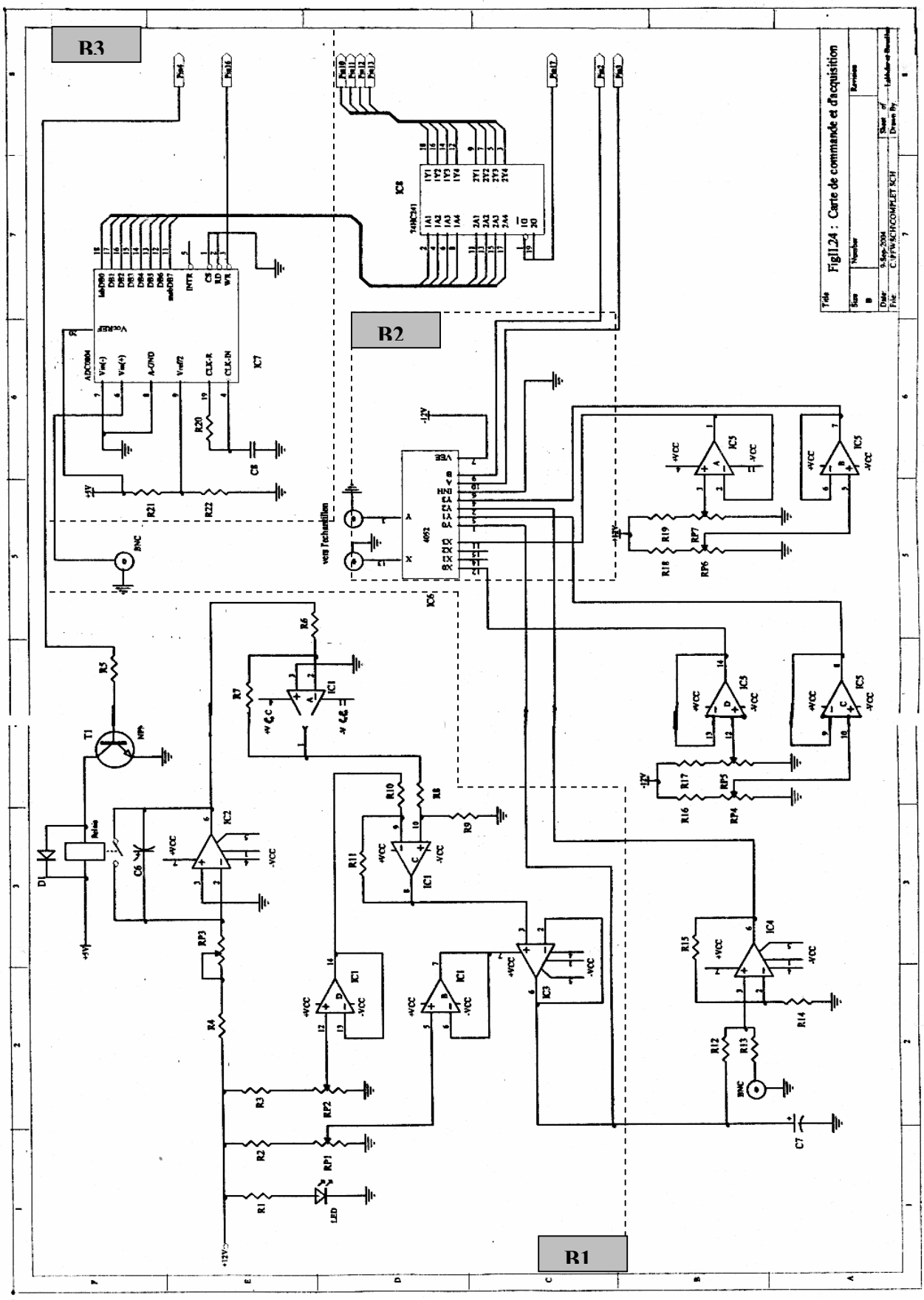


Figure III.6: Circuit électronique du système de pilotage et d'acquisition (Technique C-V) [44]

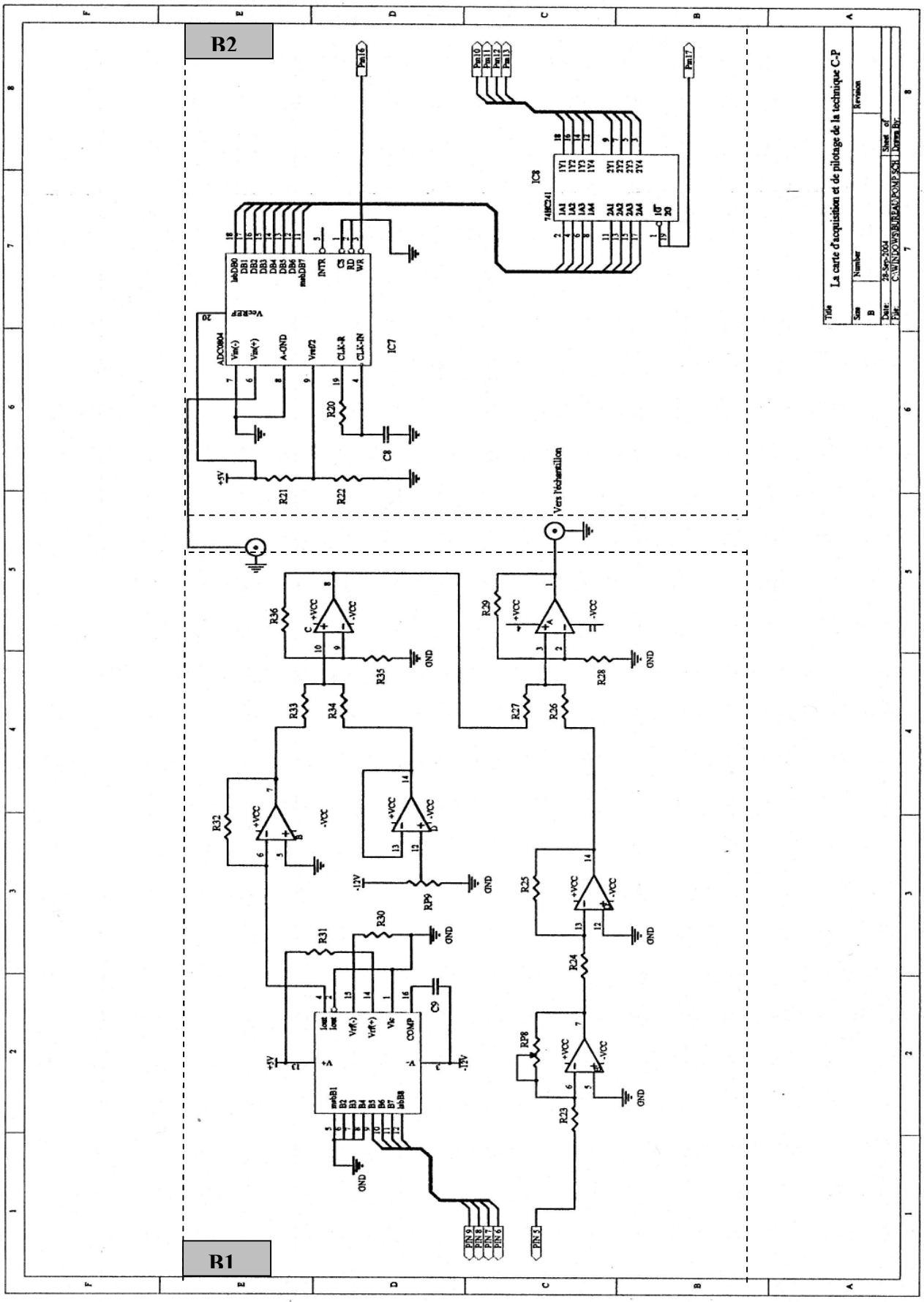


Figure III.7: Circuit électronique du système de pilotage et d'acquisition (Pompage de charges) [44]

III.2.2 Partie software

Pour bien remplir les tâches prévues, un développement software accompagne le développement hardware. Nous allons donc présenter la partie programme de notre système et expliquer les différentes parties qui le constituent. Notre programme est structuré sous forme de sous programmes auxquels des tâches bien précises ont été assignées.

III.2.2.1 organisation du programme principal

Nous avons développé une série de programme en langage C++ Builder, liés entre eux par la méthode d'appel à des sous programmes d'interruptions, à l'intérieur desquels ont été introduites les différentes commandes de contrôle des blocs du dispositif expérimental. Chacun de ces programmes a une tache bien spécifique à accomplir: technique C-V, pompage de charge et mesure de la dégradation de la transconductance en fonction du temps de stress. Le programme principal comprend le sous-programme de pilotage (génération des tensions de polarisation Vg, Vd,…) de la structure TMOS et le sous programme d'acquisition et de calcul (capacité de la structure MOS, le courant pompé et la variation du courant de drain I_D en fonction de V_{gs}).

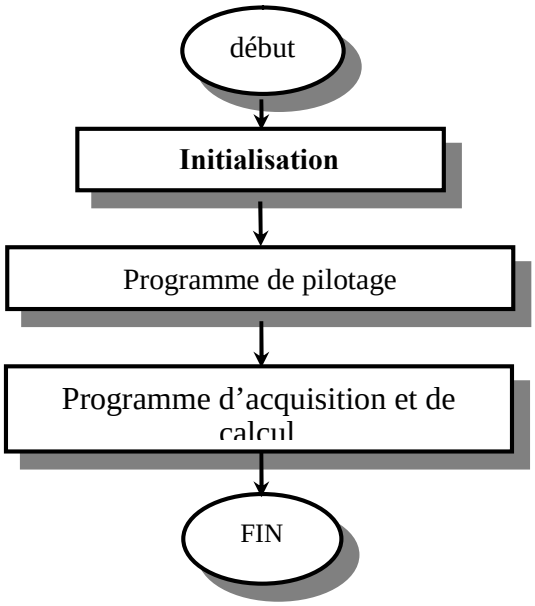


Figure III.8: Organigramme du programme principal

III.2.3 Présentation et interprétation des résultats

Pour une bonne mise en œuvre de notre dispositif et dans le but d'accomplir les tâches assignées, un programme en C++ Builder a été développé pour assurer le bon fonctionnement et la synchronisation de notre dispositif avec le micro-ordinateur (P4 1.2 GHZ).

III.2.3.1 Protocole expérimental

III.2.3.1.a Le dispositif de mesure

Le dispositif de mesure utilisé se compose des éléments suivants:

- Le système de pilotage et d'acquisition que nous avons conçu et réalisé au laboratoire.
- Un échantillon de type MOSFET BF960.
- Un micro ordinateur.

III.2.3.1.b L'échantillon

Les transistors nMOSFETs utilisés dans notre travail sont de technologie SIMOX de faible dose et partiellement dépeuplés (l'épaisseur du film du silicium égale à 100nm), avec des longueurs de grille comprises entre 0.1µm et 0.5µm. Les conditions de stress sont choisies dans le courant maximum de substrat (Vg=Vd/2), pour la mesure de la dégradation de la transconductance, le drain doit être relié au système de mesure de courant du drain I_D Keithley 614 comme il est illustré par la figure III.9.

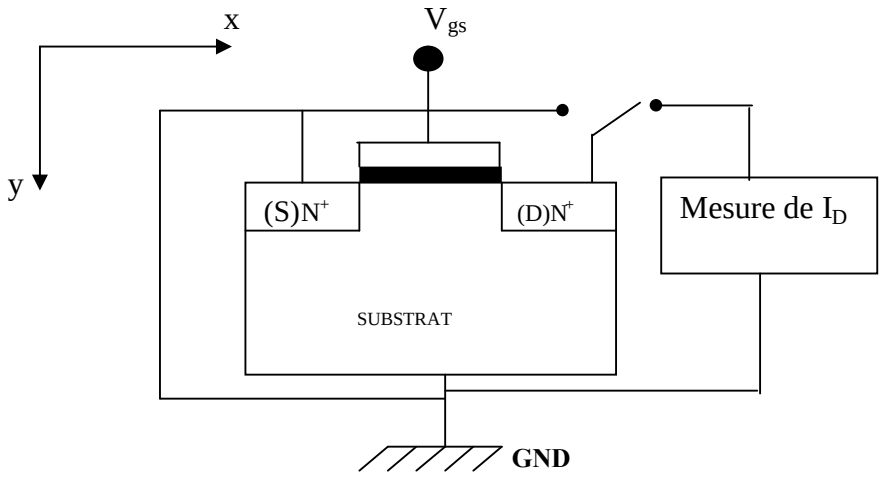


Figure III.9: Structure et brochage de l'échantillon utilisé

III.2.3.2 Principe de la méthode

a. Technique C-V

La caractérisation de la structure MOS par la méthode C(V) a été développée par TERMAN[48]. Cette technique permet la détermination de la densité des états d'interface D_{it} par comparaison de la caractéristique réelle et la caractéristique C(V) idéale (quantique) **(paragraphe I.)** Fig.III.10.

Le principal inconvénient de cette technique est dû au fait qu'elle ne permet pas d'explorer la totalité de la bande interdite [49]. Cependant, elle offre une réponse immédiate quant à la détermination du type du semiconducteur, l'épaisseur de l'oxyde, le dopage,…, à partir des valeurs limites de capacité.

b. Technique de pompage de charges

Comme nous avons déjà vu, la technique C-V ne permet pas d'explorer la totalité de la bande interdite. Cependant, la technique de pompage de charges peut encore être employée pour extraire la densité des états d'interface, et l'effet de courant tunnel de la grille peut être compensé [50]. Donc, l'utilisation de la technique de pompage de charges devient indispensable pour la détermination de D_{it} (Fig.III.10).

c. Mesure de dégradation de la transconductance

La transconductance d'un transistor MOSFET peut être définie comme la loi de variation du courant du canal (courant du drain) en fonction de la tension de contrôle Vgs, la détermination de ce paramètre est basée sur la mesure de la variation du courant du drain en fonction de tension de contrôle (I-V). La connaissance de la transconductance permet la détermination de la création du canal en fonction de la tension de contrôle.

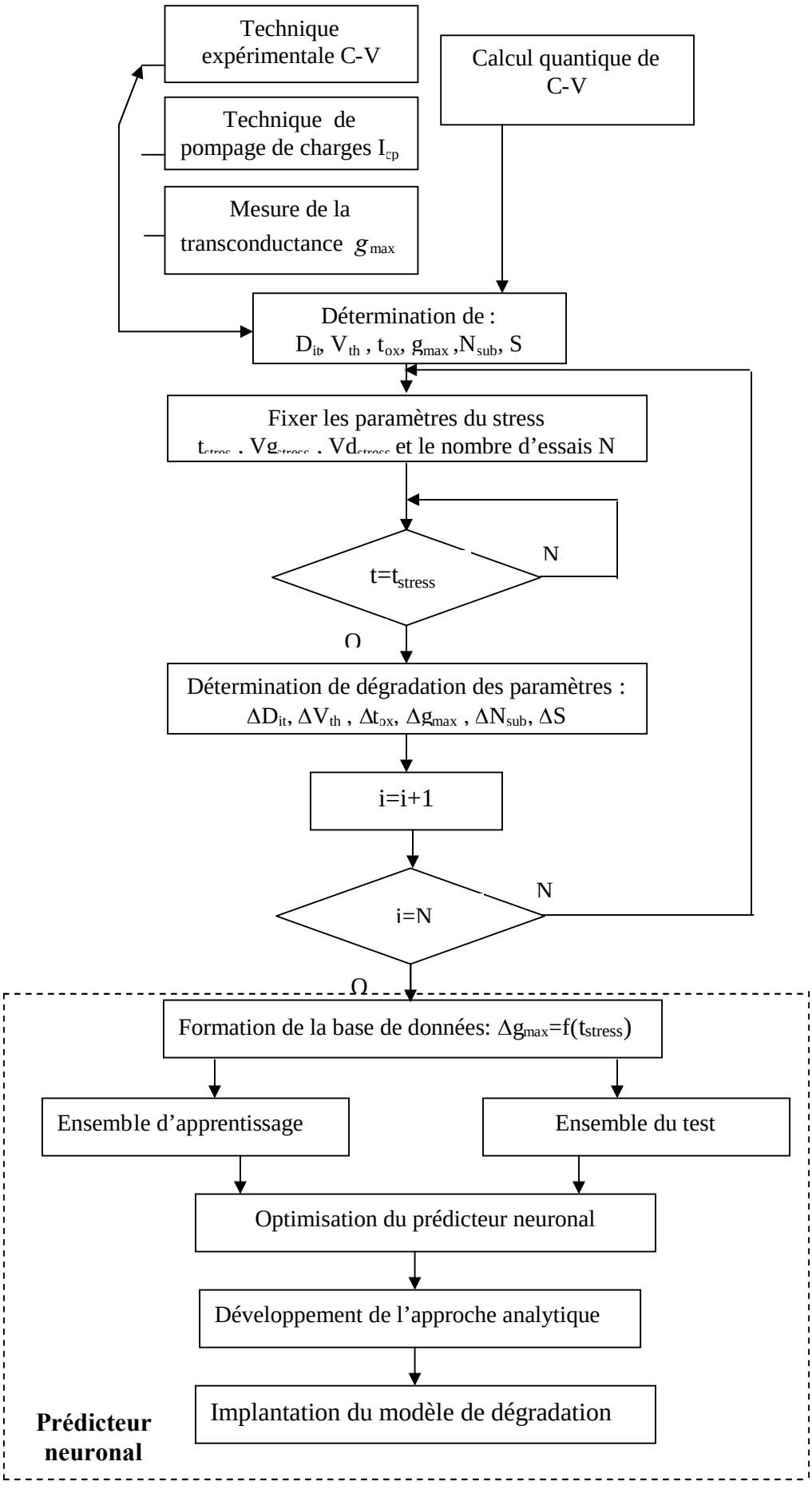


Figure III.10: Organigramme du principe de la méthode

III.2.3.2.1 Calcul quantique de la caractéristique C-V

La réduction des dimensions des dispositifs MOS se traduit par un amincissement de l'oxyde de grille et par une augmentation du dopage dans le canal. En effet, ces deux conditions permettent de maintenir un certain contrôle des effets canaux courts tout en préservant la valeur de la tension de seuil. Ainsi, la largeur caractéristique de la couche d'inversion ou d'accumulation (courbure des bandes d'énergie) devient comparable à la longueur d'onde associée aux porteurs de charge de la couche, ce qui conduit à l'émergence des effets quantiques. Le code de calcul développé sous MATLAB a pour nom MOSBAT et a pour objectif la prise en compte des effets quantiques.

La statistique classique de Fermi-Dirac pour un gaz tridimensionnel d'électrons n'est plus valable pour modéliser ce phénomène [51]. Il est nécessaire de faire appel à des équations issues de la mécanique quantique. L'approche la plus rigoureuse consiste à résoudre simultanément les équations de Poisson et de Schrödinger (auto-cohérent), rappelées ci-dessous. Celles-ci décrivent respectivement la courbure des bandes et la répartition des niveaux d'énergie pour chaque type de porteurs.

Le système d'équations non linéaire s'écrit, dans l'hypothèse de la masse effective, et en négligeant le potentiel image, comme suit

$$\frac{d^2\psi_i}{dy^2}+\frac{2m_y}{h^2}\left[E_i - V(y)\right]\psi_i(y) = 0 \qquad (3.1)$$

$$\frac{d^2V(y)}{dy^2} = -\frac{q}{\varepsilon_{si}}\rho(y) \qquad (3.2)$$

Où V(y) désigne l'énergie potentielle extérieure, m_y la masse effective dans la direction y et E correspond à l'énergie totale du système dans la direction y, $\psi_i(y)$ la fonction d'onde, q représente la charge des particules considérées (électrons ou trous), ε_{si} la constante diélectrique du silicium et ρ la concentration de charges du milieu où évolue la particule considérée.

$$\rho(y) = -q.[n(y)+N_a(y)-p(y)-N_d(y)] \qquad (3.3)$$

Dans le cas d'un transistor n-MOS, le semiconducteur est du type p et (4.5) se réduit alors à :

$$\rho(y) = -q.[n(y)+N_a(y)-p(y)] \qquad (3.4)$$

Intéressons nous maintenant plus en détail aux équations des densités d'électrons n(y) et de trous p(y). Nous donnerons ici les résultats pour les électrons (sachant qu'il est facile de l'étendre ensuite au cas des trous). Puisqu'en mécanique quantique, le module au carré de la fonction d'onde $\psi_i(y)$ représente la densité de probabilité de présence, nous obtenons simplement [52]:

$$n(y)=\sum_i N_i \left|\psi_i(y)\right|^2 \tag{3.5}$$

N_i représente le nombre total d'électrons occupant un niveau i donné, par unité de surface (ici la surface de la capacité MOS) avec:

$$N_i = \frac{m_e K_B T}{\pi h^2} Ln\left(1+\exp\left(\frac{E_F - E_i}{K_B T}\right)\right) \tag{3.6}$$

Finalement, les densités d'électrons n(y) et trous p(y) sont données par:

$$n(y)=\sum_i \frac{m_e K_B T}{\pi h^2} Ln\left(1+\exp\left(\frac{E_F - E_i}{K_B T}\right)\right)\left|\psi_i(y)\right|^2$$

$$n(y)=\sum_i \frac{m_h K_B T}{\pi h^2} Ln\left(1+\exp\left(\frac{E_F - E_i}{K_B T}\right)\right)\left|\psi_i(y)\right|^2 \tag{3.7}$$

Comme les fonctions d'onde sont normées, nous avons par ailleurs :

$$\int_0^{+\infty} n(y)dy = \sum_i N_i \tag{3.8}$$

La grandeur Q, donnée par :

$$Q = -q.\sum_i N_i \tag{3.9}$$

représente alors la charge liée à la présence des électrons attirés (ou repoussés)à la surface de la structure MOS.

Connaissant les densités de porteurs n et p, il est alors possible de calculer la densité de charge $\rho(y)$ donnée par (3.4) et de déduire l'énergie potentielle V en intégrant l'équation différentielle (3.2). L'énergie potentielle étant maintenant connue, il est alors possible de résoudre l'équation de Schrödinger (3.1). Ceci illustre bien le caractère auto-cohérent d'une résolution Schrödinger-Poisson couplée. En effet, il apparaît que pour résoudre l'équation de Schrödinger, il faut déjà en connaître le résultat, puisque les fonctions d'onde et les niveaux d'énergie servent à établir l'équation dont ils sont issus, par le biais de (3.6). En conséquence, l'obtention de résultats par une simulation Schrödinger-Poisson couplée nécessite forcément une résolution numérique.

Calcul numérique

La résolution numérique du système formé par les équations de Schrödinger et Poisson pour une couche d'inversion de l'hétérojonction Si(p)/SiO_2 est traité en général par un processus itératif (self-consistent), c'est à dire un potentiel d'essai est choisi (un potentiel initial obtenu par exemple à partir d'un modèle analytique simple (dans notre cas le modèle choisi est triangulaire)) et par utilisation de ce potentiel d'essai, on obtient les niveaux d'énergie et les fonction d'ondes électroniques, à partir desquelles on modifie le potentiel d'essai, et on recommence le calcul. La solution numérique du problème est obtenue quand le potentiel calculé est suffisamment proche du potentiel d'essai, c'est à dire la différence entre le dernier potentiel d'essai et le potentiel calculé à l'issue de la n^{éme} itération devient négligeable.

Pour un potentiel d'essai V_0 (y), l'équation de Schrödinger est discrétisée et peut s'écrire sous la forme:

$$(h^2/2m_e).(\psi_{j+1}+\psi_{j-1}-2\psi_j)/(\Delta y)^2+(E_i-V_{0j}).\psi_j=0 \qquad (3.10)$$

soit encore sous la forme matricielle: $[A][\psi]=-(\Delta y^2.2m_e/h^2)E_i.[\psi]$

où [A] est une matrice tridiagonale dont les éléments de la diagonale principale α_i ont pour valeur $\alpha_j=-2-(\Delta y^2.2m_e/h^2)V_{0j}$, et dont les éléments des deux diagonales secondaires valent un.

Le problème réduit donc à un problème de la forme Y= [A].X=λ.X $\Rightarrow$ ([A]- λ.I)X=0 c'est à dire à un problème classique de recherche de valeurs propres de la matrices [A].
Lorsque chacune des valeurs propres est connue, les vecteurs propres associés sont alors calculés, c'est à dire les niveaux d'énergie, et les fonctions d'onde associées, sont calculés, et tous les éléments sont réunis pour résoudre l'équation de Poisson.

A partir d'une solution initiale V_0 (y) donnée, les niveau d'énergie et les fonctions d'onde sont calculés, ainsi que la valeur du niveau de Fermi déduite de l'équation (3.8).
L'intégration de l'équation de Poisson fournit alors une nouvelle valeur du potentiel $U_0^n(y)$.
Tant que la valeur du niveau de fermi donnée par (3.8) est différent de 0 une nouvelle itération (n+1) est effectuée en utilisant comme nouveau potentiel $V_0^{n+1}(y)$, donné par la relation suivante :

$$V_0^{n+1}(y)=V_0^n(y)+f^n.[U_0^n(y)-V_0^n(y)] \qquad (3.11)$$

avec $f^n=\dfrac{f^{n-1}}{1-(w^n/w^{n-1})}$ et $w^n=Max(|U_0^n(y)-V_0^n(y)|)$ (3.12)

Le facteur de convergence f^n est indépendant de y et compris entre 0 et 1.

Cette expression de f^n a l'avantage d'accélérer la convergence par rapport au cas où f^n est choisi constant et indépendant de n. Cette méthode s'est révélée convergente dans tous les cas, même lorsque la solution initiale choisie arbitrairement est très éloignée de la solution finale, au prix d'un grand nombre d'itérations.

L'organigramme général de la procédure numérique est donné par la Figure III.11:

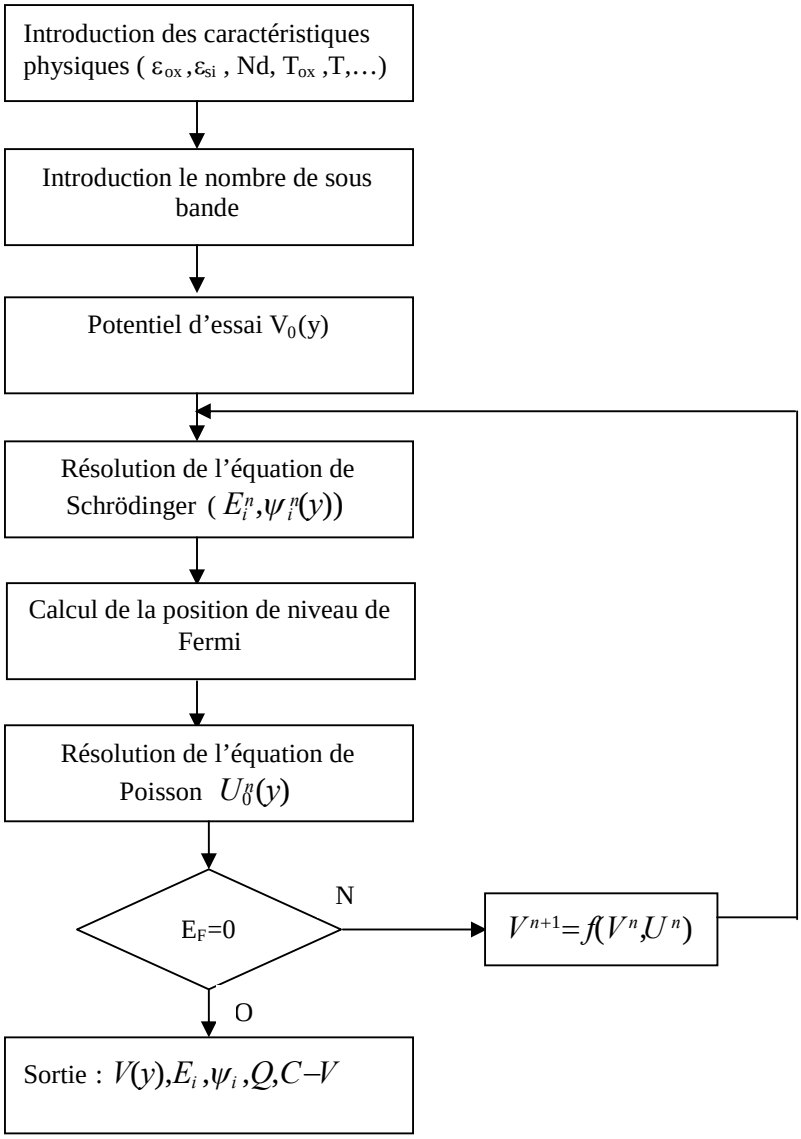


Figure III.11: Organigramme général de la procédure numérique de la résolution du système Schrödinger Poisson [53].

III.3 Développement du prédicteur neuronal

L'industrie des composants VLSI exige des investissements financiers de plus en plus lourds pour mesurer la sophistication grandissante des produits fabriqués ainsi que pour les équipements nécessaires à leur élaboration. De ce fait, la modélisation électrique des composants électriques constitue actuellement un axe de recherche très convoité à travers le monde. Pour suivre cette évolution, les modèles existants doivent être améliorés et de nouveaux modèles doivent être développés. C'est ainsi que nous assistons régulièrement à des améliorations des logiciels de simulation. Dans cette partie, on présente l'applicabilité des réseaux de neurones (ANNs) pour le développement d'une approche analytique permettant la prédiction de la durée de vie des transistors MOSFETs fortement submicronique.

Comme les réseaux de neurones peuvent modéliser les système fortement non linéaire, ils se sont avérés très utiles dans le domaine de la microélectronique, où ces conditions sont très communes. En particulier, la variation de la transconductance maximale (Gm) en fonction des différents paramètres (longueur du Canal Lg, tension du drain Vd, tension de grille Vg et temps du stress Tstress) est un exemple typique d'une réponse non-linéaire multivariables.

Les transistors nMOSFETs utilisés dans notre travail sont de technologie SIMOX de faible dose et partiellement dépeuplés (l'épaisseur du film du silicium égale à 100nm), avec des longueurs de grille comprises entre 0.1µm et 0.5µm. Les conditions de stress sont choisies dans le courant maximum de substrat (Vg=Vd/2) [54].

III.3.1 Calcul neuronal

Le réseau de neurone développé est conçu afin de relier le vecteur d'entrée (Vd, Vg, $\log(T_{stress})$ et Lg) au vecteur de sortie Gm. Chacun de ces paramètres est indexé par un neurone (Figure III.12) et présenté dans la structure neuronale comme une valeur formatée donnée par l'expression (2.5).

Les valeurs d'entrée et de sortie pour chaque neurone de la structure sont de la forme suivante:

$$\begin{cases} I_{ki} = O_{ki} & ; \ k = 1; i = 1,4 & /* \text{ Input layer} \\ I_{ki} = w_{ijk} I_{kj} + w_{i0k} & ; \ k = 2; i = 1,..., N_1; j = 1,4 & /* 1^{st} \text{ hidden layer} \\ O_{ki} = \dfrac{1}{1+e^{-I_{ki}}} & & \\ I_{ki} = w_{ijk} O_{kj} + w_{i0k} & ; \ k = 3; i = 1,..., N_2; j = 1,..., N_1 & /* 2^{nd} \text{ hidden layer} \\ O_{ki} = \dfrac{1}{1+e^{-I_{ki}}} & & \\ I_{ki} = w_{ijk} O_{kj} + w_{i0k} & ; \ k = 4; i = 1; j = 1,..., N_2 & /* \text{ Output layer} \\ O_{ki} = \dfrac{1}{1+e^{-I_{ki}}} & & \\ O_{ki} = y_i & & \end{cases} \quad (3.13)$$

Où N_1 =5 et N_2 =3 sont les nombres de neurone dans les premières et deuxièmes couches cachées (Figure III.12), k est la $k^{ème}$ couche, y_i est la valeur formatée de la sortie correspondante au paramètre Gm, w_{ijk} est la valeur de poids correspondant à la force du raccordement entre le neurone j de la couche k-1 et le neurone i de la couche k.

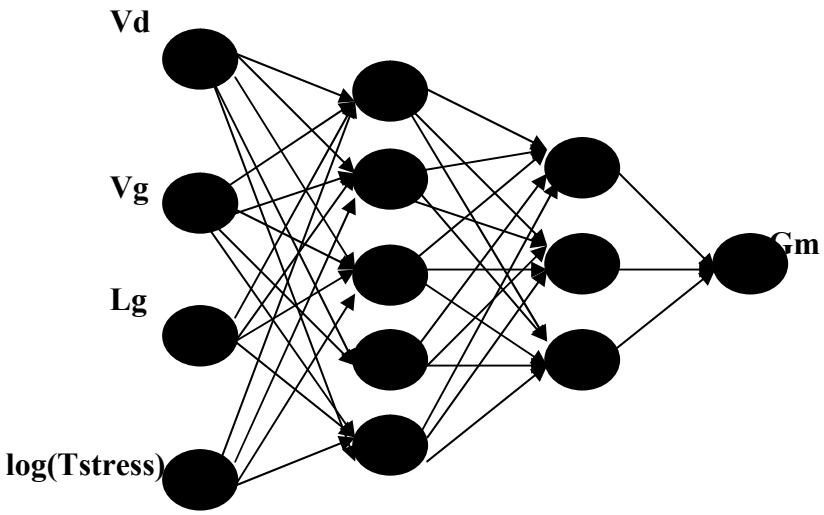


Figure III.12: Prédicteur neuronal optimisé

III.3.1.1 Optimisation du prédicteur neuronal

La procédure utilisée dans cette étude pour l'optimisation du réseau de neurones est détaillée par l'organigramme présenté sur la figure III.13. Le processus d'optimisation comprend plusieurs étapes : la constitution de la base de données, la validation de la structure du réseau de neurones, la correction de ses poids et son apprentissage.

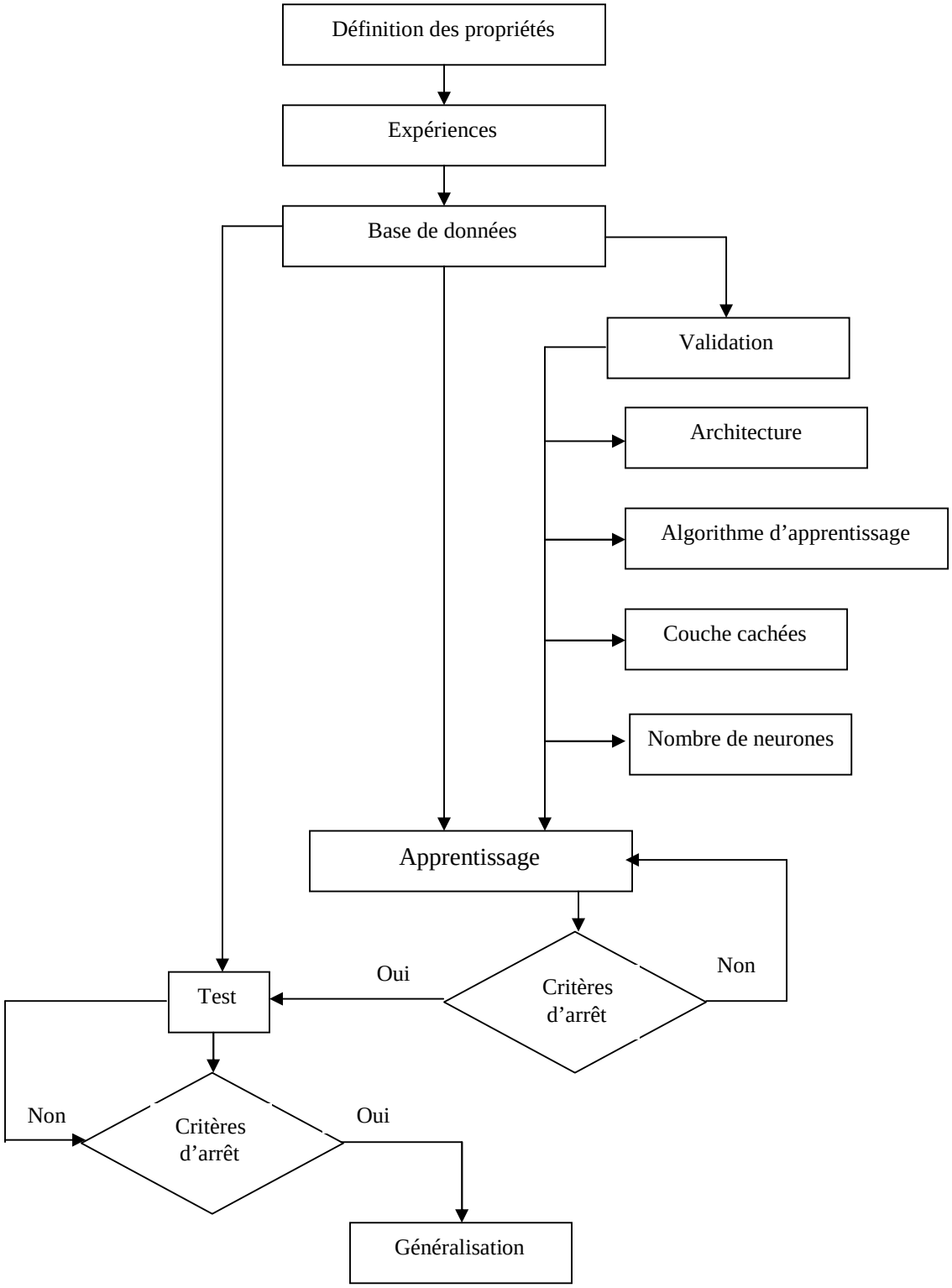


Figure III.13 : Organigramme de l'optimisation du prédicteur neuronal

L'intégration des données expérimentales est nécessaire à l'optimisation des réseaux de neurones puisque ces données encodent les corrélations paramètres-propriétés qui doivent être mémorisées par les poids de la structure [55]. Le nombre d'expériences requis pour assurer la propriété de généralisation des résultats du réseau de neurones est proportionnel au nombre de connexions dans le réseau. Par contre, il n'existe pas de règle précise d'estimation du nombre de neurones, car cette estimation dépend directement des paramètres des réseaux choisis et de la nature du problème considéré. Toutefois, une estimation connue, dérivant de la théorie PAC (Probably Approximatly Correct) d'apprentissage introduite par Valiant [57], s'exprime comme suivant:

$$N_p = 10W \qquad (3.14)$$

où N_p est le nombre de cas à considérer pour l'apprentissage et W est le nombre de connexions dans le réseau.

Brightwell et coll [58] proposent quant à eux une limite supérieure et inférieure du nombre de neurones requis selon l'expression:

$$N_H \in \left[\frac{N_p N_s}{N_I + N_S}, \frac{2N_S N_P}{N_I + N_S} \right] \qquad (3.15)$$

où N_H est le nombre de neurones dans les couches cachées, N_P est le nombre de cas soumis au réseau de neurones et N_I et N_S sont les nombres de neurones dans les vecteurs d'entrées et de sortie.

Dans cette étude, un compromis a été obtenu après plusieurs essais en tenant compte du nombre limite des expériences qui pouvaient être envisagées. En effet, quand le nombre de connexions est supérieur à la taille de la base de données, les calculs divergent. Avec une dimension comprise entre une fois et deux fois le nombre de connexions, la base de données permet un apprentissage avec un taux élevé de classification des cas [58]. L'estimation la plus optimiste du nombre de cas requis pour un apprentissage efficace peut alors s'exprimer par [59]:

$$W < N_P < 5*W \qquad (3.16)$$

Pour assurer un nombre suffisant de cas, un élargissement des cas a donc été considéré. Cet élargissement associe les déviations standard ou les tolérances mesurées ou calculées de chacune des entrées/sorties. Les nouveaux cas se calculent en adoptant la relation [59]:

$$(O_j, I_j)_{nouveau} = (O_k, I_k)_{original} \pm ALEA()*(\sigma(o_k), \sigma(i_k))$$

où O et I sont relatifs aux éléments des vecteurs d'entrée et de sortie, j et k sont, respectivement, les indices des cas dans la base de données originale et nouvelle, ALEA()

est un générateur de nombre aléatoires produisant une distribution Gaussienne auteur de la moyenne et σ est la déviation standard ou la tolérance associée à l'élément des vecteurs entrée ou sortie.

Cependant, l'élargissement de la base de données ne contribue pas de façon significative dans la découverte des relations entrée-sortie puisqu'il ne rajoute pas de cas indépendants. Il représente plutôt la flexibilité du système autour d'un point donné dans l'espace des paramètres. Cette possibilité d'introduire le bruit dans le réseau de neurones est appelée « jitter » [59]. Cette approche est donc un moyen simple d'éviter l'excès d'apprentissage, de vérifier la reproductibilité des essais et d'améliorer les propriétés de généralisation des résultats prédits [60].

La base de données élargie est divisée alors en trois catégories [61]:

- La catégorie de validation permettant de fixer les paramètres du réseau de neurones, soit le type d'architecture, le nombre de neurones, les couches cachées, etc;
- La catégorie d'apprentissage dédiée à l'ajustement des poids suivant l'algorithme d'apprentissage ;
- La catégorie de test permettant de vérifier les prédictions du réseau de neurones.

L'optimisation du réseau de neurones commence par la validation de la structure. Cette étape n'est pas requise pour toutes les optimisations conduites dans cette étude puisque la nature du problème n'a pas changé, seules les variables changent. Les paramètres de validation ont été déjà décrits en détails[54]. Ces paramètres concernent notamment le type d'architecture, la définition des couches (fonction d'activation, type d'erreur, etc.) et leur nombre et le type d'algorithme d'apprentissage.

Après la validation de la structure, l'étape suivante considère la correction des poids à l'aide de l'algorithme d'apprentissage jusqu'à l'obtention de la généralisation des corrélations prédites par le réseau de neurones (Figure III.18).

Après initialisation des poids, la soumission des cas pour l'apprentissage peut se faire suivant plusieurs organisations possibles, dont quelques schémas importants sont:

- L'apprentissage croisé ou 'cross-validation', où la base de données est divisée en k sous-groupes. Pour k cycles d'apprentissage avec (k-1) groupes choisis au hasard, la correction des poids est effectuée par rapport au groupe omis [62]. Cette

approche s'applique aux cas des bases de données importantes. Elle est abandonnée dans cette étude puisqu'elle s'accompagne d'un apprentissage très lent [65];

- L'apprentissage par série ou 'batch training', dans lequel la correction des poids se fait après soumission au réseau d'un nombre donné de cas. Cela a pour effet de diminuer l'effet d'un cas de la base de données si celui-ci présente une aberration ;
- l'apprentissage cas par cas ou 'per sample updating', avec lequel la mise à jour est effectuée après passage de chaque cas de la base de données. Ce dernier cas a été souvent utilisé dans cette étude pour éviter l'accumulation du bruit dans les poids.

Pour garantir une meilleure optimisation, la procédure d'apprentissage et la procédure de test ont été couplées. Pour arrêter la procédure d'apprentissage en même temps que la procédure de test, deux critères cumulatifs ont été choisis pour permettre une meilleure optimisation : le nombre de cycles et l'erreur d'apprentissage [64].

Un choix correct du nombre de cycles est important pour garantir un apprentissage efficace. Un nombre de cycles élevé provoque une approximation de la fonction et de son bruit (overfitting). A l'opposé, un nombre faible de cycles ne donne pas assez de temps aux poids de reconnaître les corrélations que présente le problème (underfitting). Dans cette étude, le nombre choisi de cycles a varié en fonction de la construction des réseaux et des paramètres opératoires (Vg, Vd,…) pour permettre la prédiction de comportements moyens [64].

L'erreur d'apprentissage est décrite par six attributs spécifiques [64]:

- l'erreur moyenne d'apprentissage (Etrn), qui est une forme dérivée de l'expression de l'énergie du système donnée par la relation (2.6a). Elle peut prendre plusieurs formes [64]. Une des expressions utilisées dans cette étude est donnée par [65]:

$$ETrn=\frac{1}{N_{Ptr}}(z_i - O(I_i, W))^2 \qquad i=1,N_{Ptr} \qquad (3.17)$$

Où ETrn est l'erreur moyenne d'apprentissage, N_{Ptr} est le nombre de cas de la base de données d'apprentissage soumis à l'apprentissage, I_i et z_i sont les vecteurs d'entrée et de sortie correspondant au cas i, $O(I_i, W)$ est le vecteur de sortie prédit par le réseau de neurones, W symbolise la population des poids, la figure (III.14) montre l'évolution de l'erreur moyenne d'apprentissage de notre prédicteur;

- l'erreur maximale d'apprentissage (EtrnMax), qui correspond au cas le plus défavorable de la liste de cas soumise à l'apprentissage. EtrnMax est donnée par la relation

$$ETrnMax = MAX\left|(z_i - O(I_i, W))\right| \qquad i=1,N_{Ptr} \qquad (3.18)$$

- Le taux de classification des cas dans la catégorie des cas d'apprentissage (CTrn), qui donne le pourcentage de cas réalisant u écart par rapport au résultat prédit inférieur à une tolérance donnée (5% en générale) ;
- L'erreur moyenne de test (Etst), qui est l'équivalente à Etrn pour la catégorie des cas de test;
- L'erreur maximale de test (EtstMax), qui est l'équivalente mais concernant les cas choisis pour le test ;
- Le taux de classification des cas dans la catégorie des cas de test (CTst), qui est l'équivalent de CTrn.

Après consommation du nombre total de cycles, une mise à jour de la structure (modification du nombre de neurones) est effectuée si le seuil n'est pas atteint. Le seuil est calculé sur la base d'un changement inférieur à 2% de l'erreur moyenne entre deux cycles successifs [65]. Dans le cas où le seuil est atteint, la structure est retenue comme structure optimale. Dans certains cas, plusieurs configurations équivalentes ont pu être obtenues [64]. La figure (III.15) illustre le choix de la structure optimale de notre prédicteur.

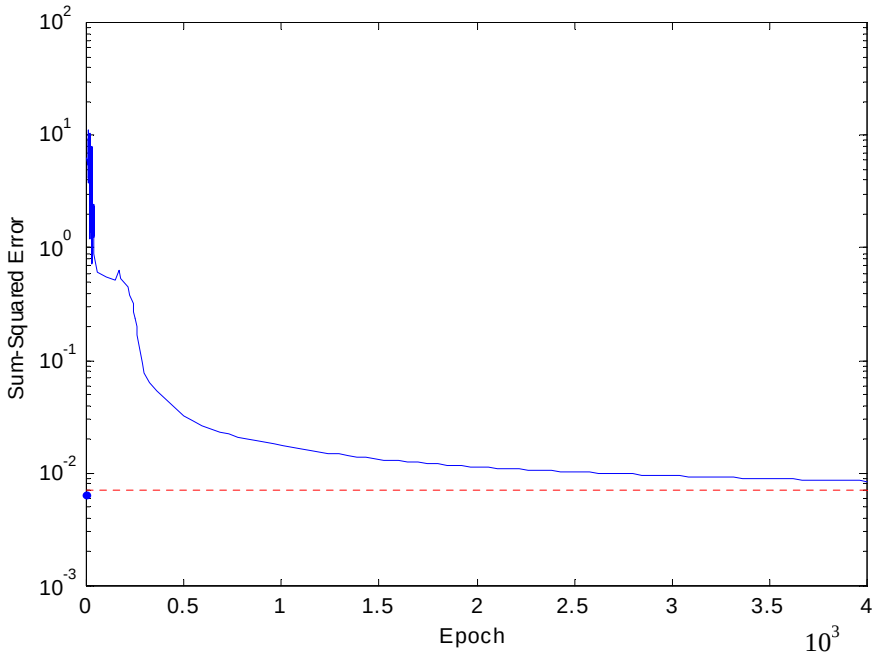


Figure III.14: L'évolution de l'erreur moyenne d'apprentissage de notre prédicteur

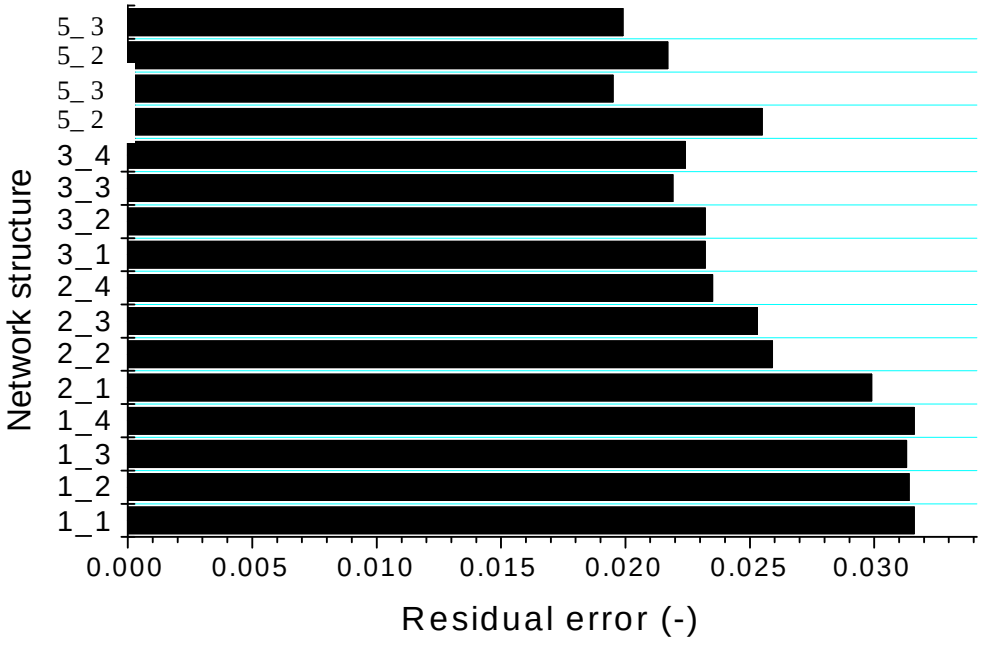


Figure III. 15: Histogramme de l'erreur moyenne après la phase d'optimisation montrant la structure optimale de notre prédicteur

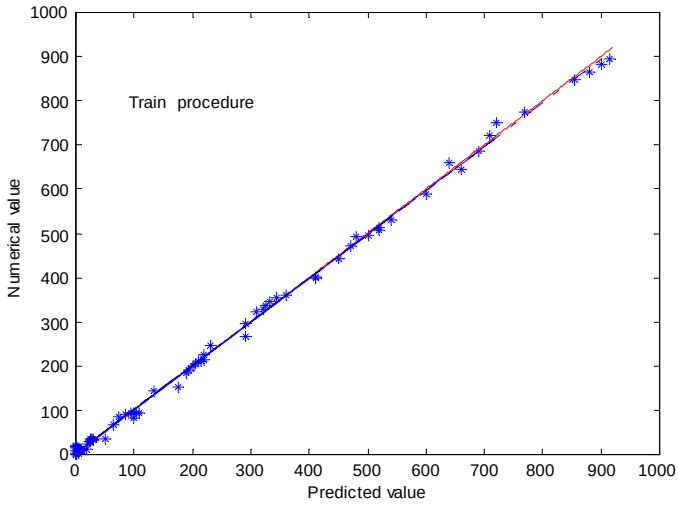
Le tableau (III.1) donne les paramètres de notre prédicteur après la procédure de l'optimisation.

<table>
<tr><td colspan="2">Paramètres</td><td colspan="6">Valeurs optimisées</td></tr>
<tr><td colspan="2">Architecture</td><td colspan="6">Normal feed-forward MLP</td></tr>
<tr><td rowspan="9">Layer definition</td><td>Couches cachées</td><td colspan="6">2</td></tr>
<tr><td>apprentissage</td><td colspan="6">Rétropropagation rapide (Quick propagation)</td></tr>
<tr><td rowspan="5">Nombre de neurones</td><td rowspan="3">Entrées (I=4)</td><td></td><td>Vd()</td><td>Vg ()</td><td>Log t ()</td><td>lg ()</td></tr>
<tr><td>Min</td><td>2.5</td><td>1.25</td><td>0</td><td>0.1</td></tr>
<tr><td>Max</td><td>3</td><td>1.95</td><td>7</td><td>0.3</td></tr>
<tr><td colspan="5">Première couche (N_1): variée</td></tr>
<tr><td colspan="5">Deuxième couche (N_2): variée</td></tr>
<tr><td rowspan="3">Sorties (O=1)</td><td></td><td colspan="4">G(%)</td></tr>
<tr><td>Min</td><td colspan="4">0</td></tr>
<tr><td>Max</td><td colspan="4">100</td></tr>
<tr><td>Population des poids</td><td colspan="6">$I*N_1+N_1*N_2+N_2*O$</td></tr>
<tr><td>Fonction d’entrée</td><td colspan="6">Produit scalaire entre les sorties de neurone
$w_{ij}o_j$</td></tr>
<tr><td>Traitement à l’entrée</td><td colspan="6">Variables entre 0 et 1
$\frac{x-x_{\min}}{x_{\max}-x_{\min}}$</td></tr>
<tr><td>Fonction d’activation</td><td colspan="6">Sigmoïd
$\frac{1}{1+\exp(-x)}$</td></tr>
<tr><td colspan="2">Type d’erreur du réseau (Mean square error)</td><td colspan="6">Training process
$ETrn=\frac{1}{N_{Ptr}}(z_i-O(I_i,W))^2 \qquad i=1,N_{Ptr}$
Test process
$ETst=\frac{1}{N_{Ptst}}(z_i-O(I_i,W))^2 \qquad i=1,N_{Ptst}$</td></tr>
<tr><td rowspan="4">Apprentissage et test</td><td>Batch size</td><td colspan="6">1</td></tr>
<tr><td>Itération maximale</td><td colspan="6">4000</td></tr>
<tr><td>Tolérance</td><td colspan="6">0.001 (not reached)</td></tr>
<tr><td>séquence</td><td colspan="6">apprentissage + test passe après la mise à jour des poids</td></tr>
<tr><td>Taille de la base de données</td><td colspan="6">54 échantillons pour l’apprentissage
25 échantillons pour le test</td></tr>
</table>

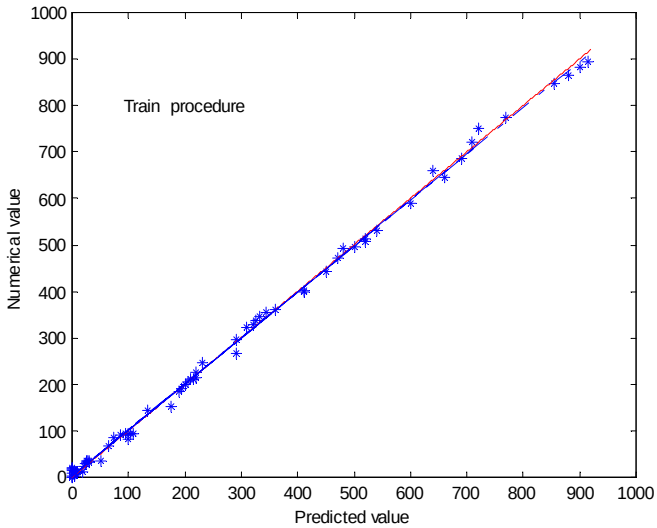
Tableau III.1: paramètres du prédicteur neuronal après la procédure d’optimisation

III.3.2 Résultats et discussion

Pour notre structure, l'erreur résiduelle était de 0,01 et presque 100% des cas soumis ont été saisis correctement (les erreurs de sortie étaient moins de 5%). Afin de valider la propriété prédictive de la structure optimisée de réseau, les soties de test et d'apprentissage ont été comparées à la réponse de réseau. La figure III.16 montre que dans les deux cas, un bon accord entre les résultats expérimentaux et les résultats prédits a été obtenu. Par conséquent, la structure optimisée peut être utilisée pour prédire d'autres combinaisons des variables d'entrée.



(a)



(b)

Figure III. 16: validation du prédicteur neuronal pour les deux ensembles (a) apprentissage et (b) test

Les transistors nMOSFETs fortement submicronique avec les différentes longueurs de grille ont été électriquement soumis à un stress où le courant de substrat est maximum (Vg= Vd/2). Les mesures de la durée de vie sont déterminées à partir de la courbe D=f(t, Lg). Dans notre cas, la durée vie correspondante est à 10 % de la dégradation de transconductance maximale (Gmax(%)=10, voir la figure III.17). La Figure III.17 montre l'évolution prédite et expérimentale de la dégradation de la transconductance maximale en fonction du temps de stress pour un transistor nMOSFET avec Lg=0.1µm et tension de stress Vg=Vd/2=1.5V.

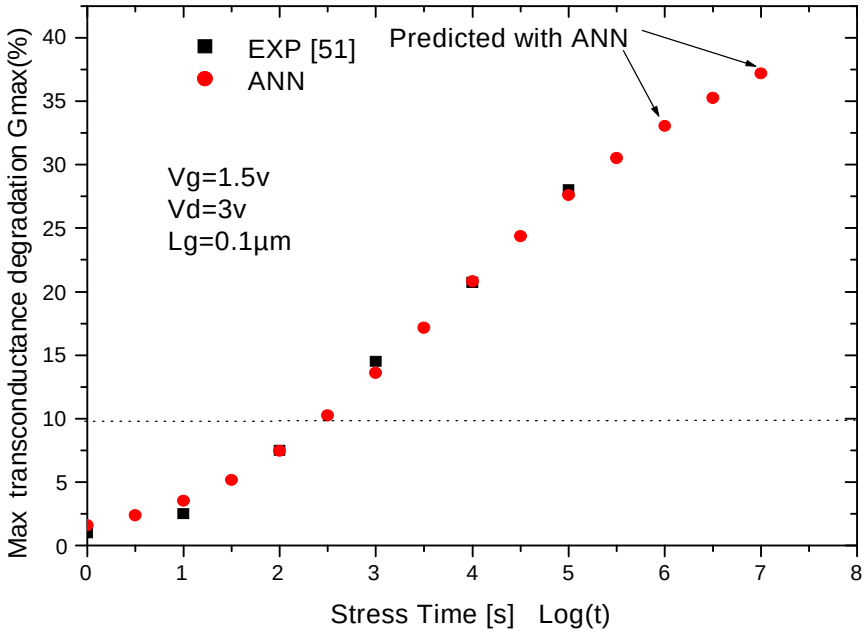


Figure III. 17: L'évolution de la transconductance maximal prédite et expérimentale en fonction du temps de stress

En se basant sur les valeurs générées par le prédicteur neuronal (ANN), la variation de la dégradation D(t) peut être représentée par une fonction logarithmique polynomiale:

$$D(t)=B_3 \log^3(t)+B_2 \log^2(t)+B_1 \log(t)+D_0 \qquad (3.19)$$

Où, B1,B2 et B3 sont les paramètres de la fonction polynomiale ainsi définie et t représente le temps de stress.

Pour différentes longueurs de grille, l'expression (3.19) peut être donnée comme:

$$D(t,Lg)=B_3(Lg)\log^3(t)+B_2(Lg)\log^2(t)+B_1(Lg)\log(t)+D_0(Lg) \qquad (3.20)$$

Dans ce cas $D_0(Lg)$, $B_1(Lg)$, $B_2(Lg)$ et $B_3(Lg)$ sont des paramètres qui varient en fonction de la longueur de grille. La connaissance de la dégradation en fonction du temps de stress pour différentes longueurs de grille nous permet de déterminer ces paramètres. Il est important de noter qu'une meilleure interpolation par une fonction polynôme est obtenue pour une approximation polynomiale du cinquième ordre (l'erreur de l'interpolation pour ce cas est minimale). Ces dernières fonctions peuvent être représentées comme:

$$D_0(Lg)=\sum_{i=0}^{5} d_i\, Lg^i \qquad (3.21)$$

$$B_1(Lg)=\sum_{i=0}^{5} b_{1i} Lg^i \qquad (3.22)$$

$$B_2(Lg)=\sum_{i=0}^{5} b_{2i} Lg^i \qquad (3.23)$$

$$B_2(Lg)=\sum_{i=0}^{5} b_{2i} Lg^i \qquad (3.24)$$

Les paramètres d_i, b_{1i}, b_{2i} *et* b_{3i} sont des coefficients donnés pour chaque mode de stress et de chaque type du transistor MOSFET. Ces paramètres sont récapitulés, dans le cas du transistor MOSFET de N-canal, dans le tableau (III.2).

Les expressions établies ci-dessus, donnent la dégradation de la transconductance maximale en fonction du temps de stress et de la longueur de grille pour différentes tensions de stress (Vg = Vd/2) (Figure III.17). Afin de valider notre approche analytique, la figure (III.17) compare les résultats expérimentaux et nos résultats analytiques (résultats prédits) basés sur le prédicteur neuronal pour une certaine condition de stress, un bon accord est remarquable pour la gamme complète. Cette dernière observation montre l'applicabilité des réseaux de neurones pour l'étude de la dégradation des dispositifs MOSFETs fortement submicroniques sous les conditions de stress.

d_i $i=0:5$	-1.7040	78.8030	-983.2682	5424.5000	13949	13672
b_{1i} $i=0:5$	18.9576	-670.6683	8081.9	-44222	113440	-111080
b_{2i} $i=0:5$	-7.2592	293.8180	-3749.9	21183	-55461	55094
b_{3i} $i=0:5$	0.4196	-18.938	271.2878	-1665.4	4627.3	-4801.4

Tableau III.2: valeurs des coefficients de la fonction de dégradation pour le mode de stress Vg=Vd/2=1.25V (Lg en µm) [54].

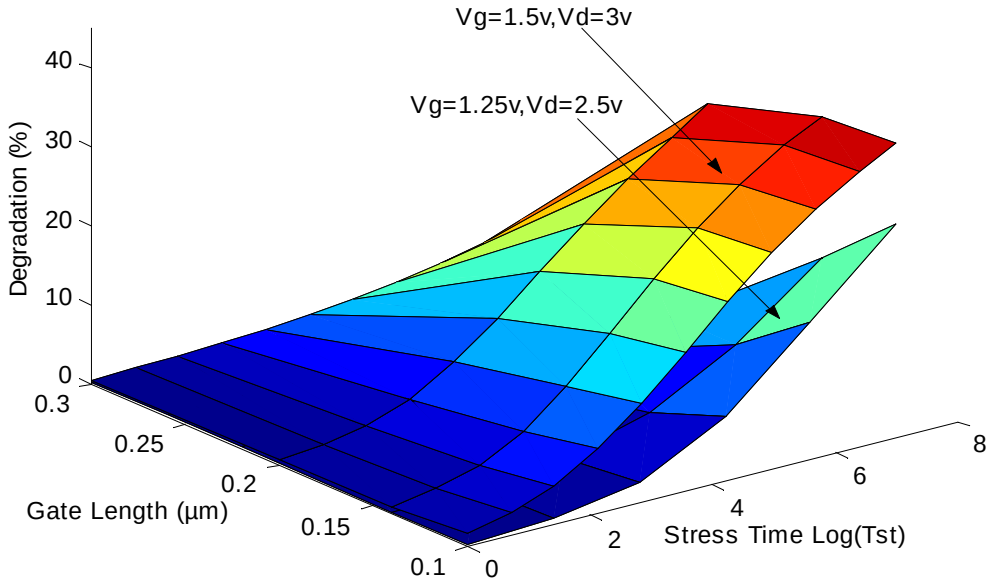


Figure III. 18: Graphe 3D de l'évolution de la dégradation (D(%)) en fonction du temps de stress et longueur de grille (Vd=3V,Vg=1.5V et Vd=2.5V,Vg=1.25V).

III.3.2.1 Estimation de la durée de vie

Le temps de dégradation peut être déterminé en résolvant l'équation (3.20) pour une dégradation fixe. Le calcul de la durée de vie, correspondant à une dégradation de la transconductance de 10% (Figure (III.21)), est obtenu donc pour $D(t,L_g)=10$ (figure (III.18)).

Finalement, l'équation gouvernant la durée de vie des dispositifs MOSFETs fortement submicroniques est donnée par [54]:

$$B_3(Lg)\log^3(t_{lifetime})+B_2(Lg)\log^2(t_{lifetime})+B_1(Lg)\log(t_{lifetime})+D_0(Lg)-10=0 \quad (3.25)$$

Comme application, nous avons appliqué notre approche analytique pour estimer la durée de vie de plusieurs N-Transistors MOSFETs. La figure (III.19) donne la durée de vie en fonction de la longueur de grille pour une tension de stress de Vg=Vd/2=1.5V. Le bon accord entre les résultats expérimentaux et les nôtres montre que le phénomène de dégradation peut être étudié en utilisant notre approche analytique.

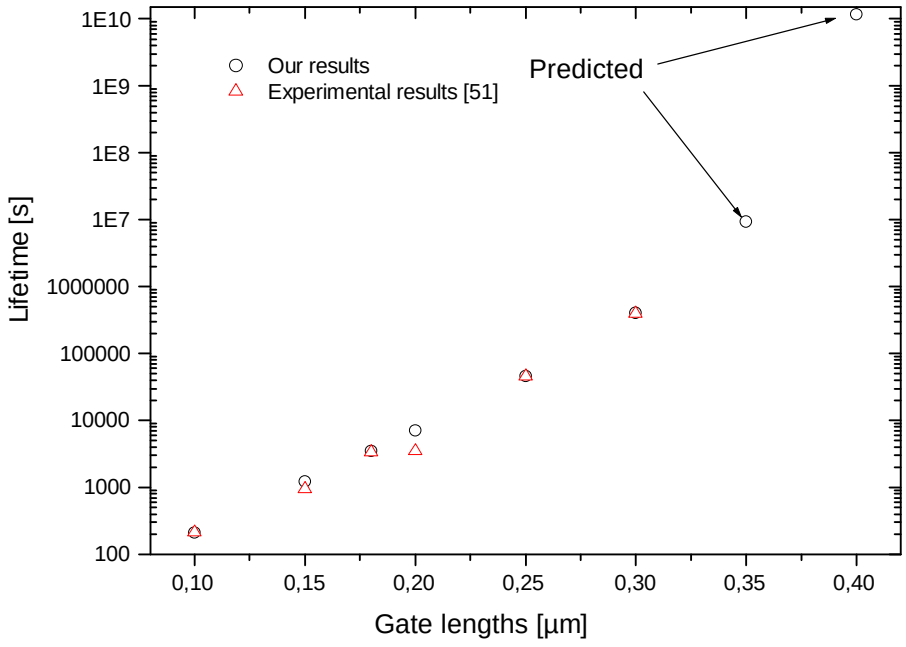


Figure III. 19: La durée de vie prédite et expérimentale comme fonction de la longueur de grille pour une tension de stress Vg=Vd/2=1.5V.

III.3.2.2 Implantation du modèle de dégradation

Les modèles existants du transistor MOSFET dans les simulateurs électroniques (SPICE3F4, PSPICE, CADENCE...) ne tiennent pas compte du phénomène de dégradation temporelle. Le dernier est purement expérimental et très efficace dans le traitement du phénomène de vieillissement.

Vu l'importance de ce phénomène, nous avons développé un modèle qui est lié à l'effet de dégradation. Le schéma électrique équivalent du transistor MOSFET fortement

submicronique avant et après l'implantation de l'effet de la dégradation temporelle est représenté sur la figure III.20, avec:

Dans la figure III.20.a, le courant du canal donné par le modèle EKV [66] comme:

- pour $Vg < Vth$:

$$I_{ch}=I_D=0 \qquad (3.26.a)$$

- pour $Vg > Vth$, en saturation:

$$I_{ch}=I_D=g_m \frac{nU_t\sqrt{IC}}{1-e^{-\sqrt{IC}}} \qquad (3.26.b)$$

où n est le facteur d'idéalité, , U_t est la tension thermique, g_m est la transconductance et IC est le coefficient d'inversion défini comme:

$$IC=\frac{I_D}{2n\beta U_t^2}, \text{ avec } \beta=\mu_{eff}C_{ox}\frac{W}{L} \quad (3.27)$$

μ_{eff} est la mobilité effective, C_{OX} est la capacité de l'isolant par unité de surface, et à IC égale à 1 correspond la transition faible/forte d'inversion (IC est inférieur à 1 pour le régime de faible inversion et plus grand dans le régime de forte inversion).

En régime de faible inversion :

$$I_{ch}=I_D=g_m nU_t \qquad (3.28)$$

En régime de forte inversion:

$$I_{ch}=I_D=g_m^2\frac{n}{2\beta} \qquad (3.29)$$

Dans la Figure III.20.b, le courant du canal est donné par:

- pour $Vg < Vth$: $I_{ch}=I_D=0$
- For $Vg > Vth$, en saturation.

Dans ce cas, on peut définir une nouvelle transconductance appelée la transconductance dégradée g_{md} donnée comme le suivant:

$$g_{md} = g_m (1 - 0.01D(t, Lg)) \qquad (3.30)$$

Remplaçant (3.30) dans (3.29) et (3.28), on aura un nouveau model du courant:

En régime de faible inversion:

$$I_{ch} = g_m(1 - 0.01D(t,Lg))nU_t = I_D - I_{DE} \qquad (3.31a)$$

Où:

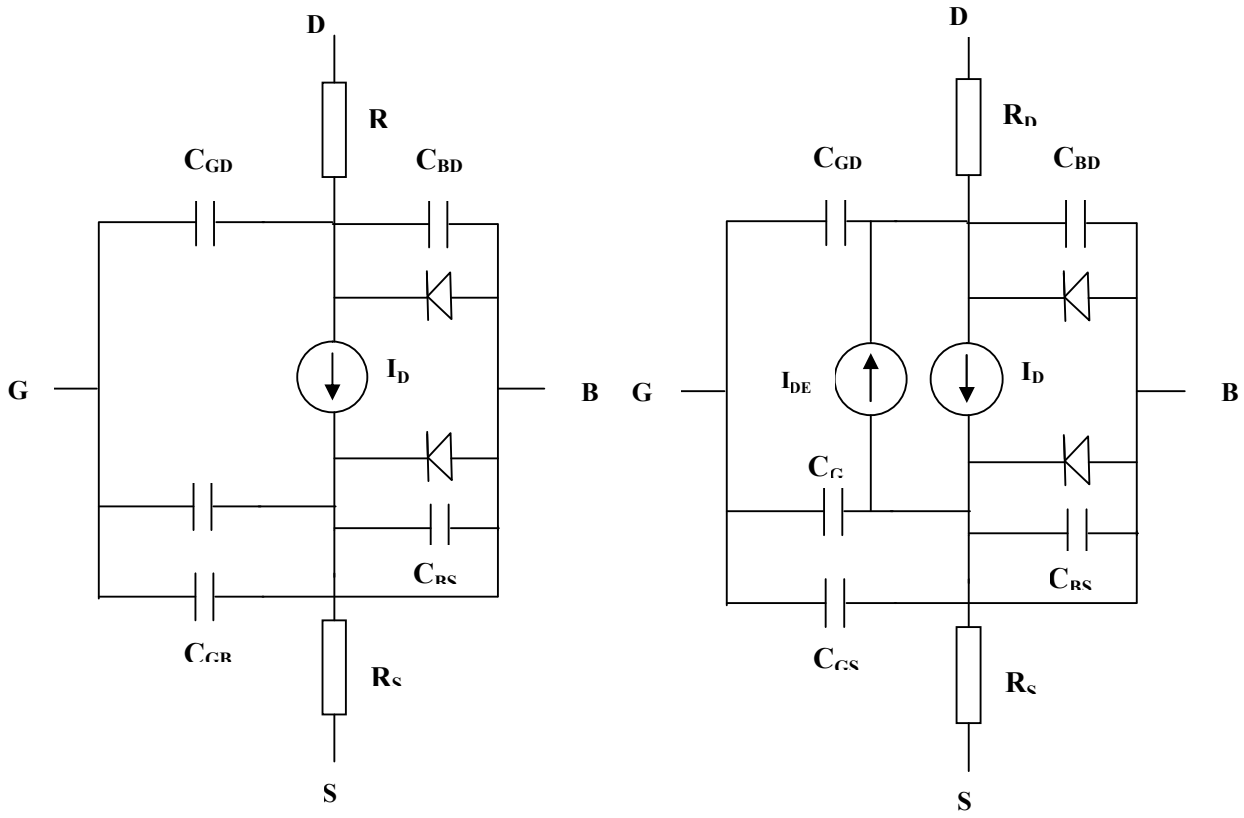
$$I_{DE} = 0.01 g_m D(t,Lg) nU_t \qquad (3.31b)$$

En régime de forte inversion:

$$I_{ch} = g_m^2(1 - 0.01D(t,Lg))^2 \frac{n}{2\beta} = I_D - I_{DE} \qquad (3.32a)$$

Où:

$$I_{DE} = g_m^2 \frac{n}{2\beta} (0.02D(t,Lg) - 10^{-4} D^2(t,Lg)) \qquad (3.32b)$$



(a) Avant **(b) Après**

Figure III. 20: Schéma électrique du transistor MOSFET fortement submicronique (a) avant l'implantation de notre modèle (b) après l'implantation de l'effet de dégradation temporelle [54].

III.3.2.3 Impact du modèle de dégradation sur la conception des circuits intégrés

La transconductance du canal g_m est fortement liée à la conception des circuits intégrés.

Comme première application, considérons l'exemple d'une porte inverseuse donnée par la figure (III.21):

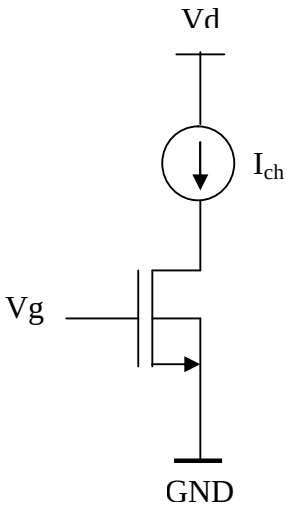


Figure III.21: Etage inverseur à base du transistor MOSFET fortement submicronique

La fréquence de coupure f_t peut être approximativement donnée par [66] comme:

$$f_t = g_m \Big/ 2\pi C_g \qquad (3.33)$$

C_g est la capacité de la grille.

Après l'implantation de notre modèle de dégradation, on peut définir une nouvelle fréquence de coupure appelée la fréquence de coupure dégradée donnée par:

$$f_{td} = \frac{g_{md}}{2\pi C_g} = \frac{g_m(1-0.01D(t,Lg))}{2\pi C_g} \qquad (3.34)$$

La figure III.26 illustre l'évolution de la fréquence de coupure normalisée (f_{td}/f_t) comme fonction du temps. On remarque que la fréquence de coupure diminue de façon non linéaire avec l'augmentation du temps du stress et la diminution de la longueur de grille (Lg).

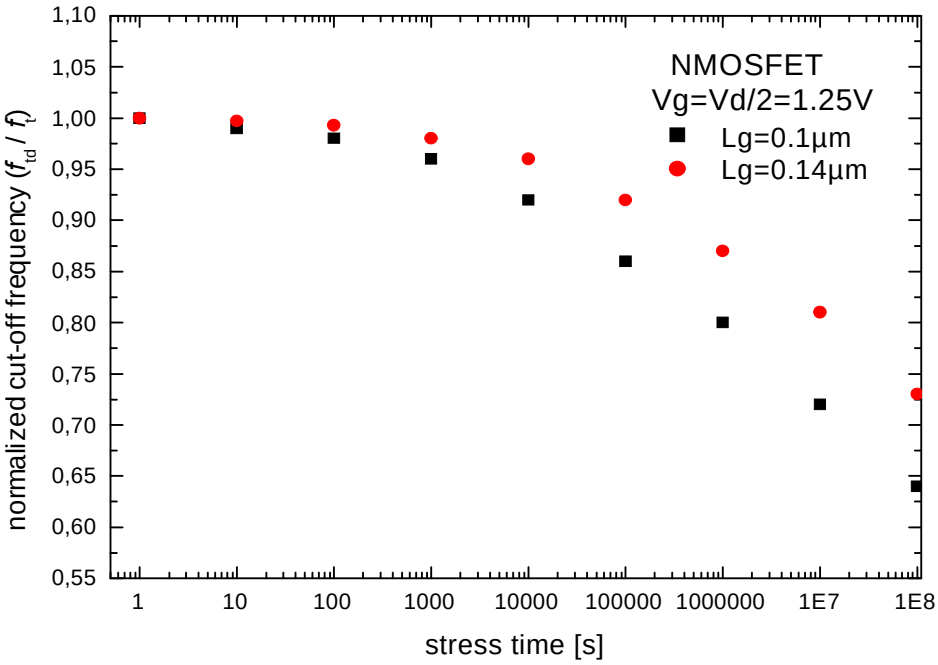


Figure III. 22: Evolution de la fréquence de coupure normalisée (f_{td}/f_t) en fonction du temps de stress [54]

La deuxième application est consacrée à l'étude de l'effet de la dégradation temporelle sur les performances du fonctionnement d'un amplificateur opérationnel basse fréquence fortement submicronique (W/L=0.8/0.18), on a opté pour une structure utilisant trois étages (source du courant, amplificateur différentiel et un amplificateur Push-Pull) [66] montés en cascade (Fig. III.23).

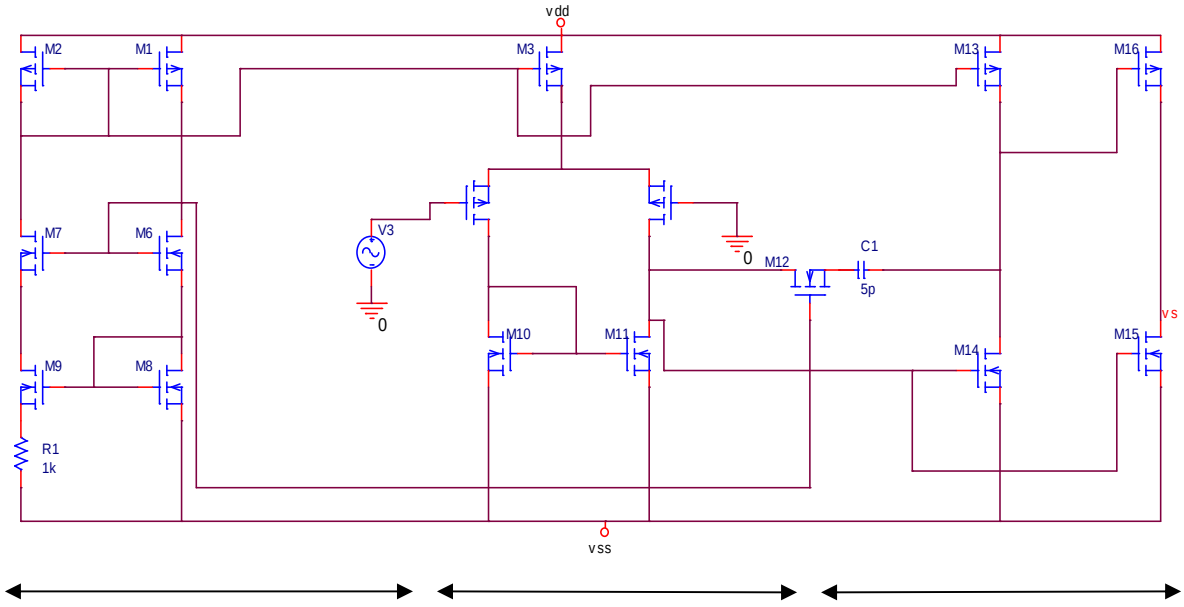


Figure III. 23: Schéma électrique de notre amplificateur opérationnel

Le gain en boucle ouverte de notre amplificateur opérationnel peut être donné par l'expression suivante [66]:

$$A_v = A_1 . A_2 . A_3 \qquad (3.35)$$

où A_1 est le gain de l'étage de l'amplificateur différentiel, A_2 est le gain de l'étage de l'amplificateur et A_3 est le gain de l'étage Push-Pull avec:

$$A_1 = \frac{g_m}{g_{ds2} + g_{ds4}}, \quad A_2 = \frac{-g_m}{g_{ds13} + g_{ds14}} \text{ et } A_3 = \frac{g_m}{g_m + g_{ds15} + g_{ds16}} \qquad (3.36)$$

avec: g_{ds15} *et* $g_{ds16} \prec\prec g_m$

donc:

$$A_v = \frac{g_m}{g_{ds2} + g_{ds4}} . \frac{-g_m}{g_{ds13} + g_{ds14}} \qquad (3.37)$$

En négligeant la dégradation de g_{ds} devant g_m [66] et en remplaçant (3.30) dans (3.37), la dégradation du gain devient:

$$A_{vd} = A_v(1 - 0.01D(t))^2 \qquad (3.38)$$

La bande passante de notre amplificateur opérationnel f_c peut être donnée comme [66]:

$$f_t = g_m / 2\pi C_{eq} \qquad (3.39)$$

En remplaçant (3.30) dans (3.39), la dégradation de la bande passante est donnée par:

$$f_{cd} = f_c\,(1 - 0.01D(t)) \qquad (3.40)$$

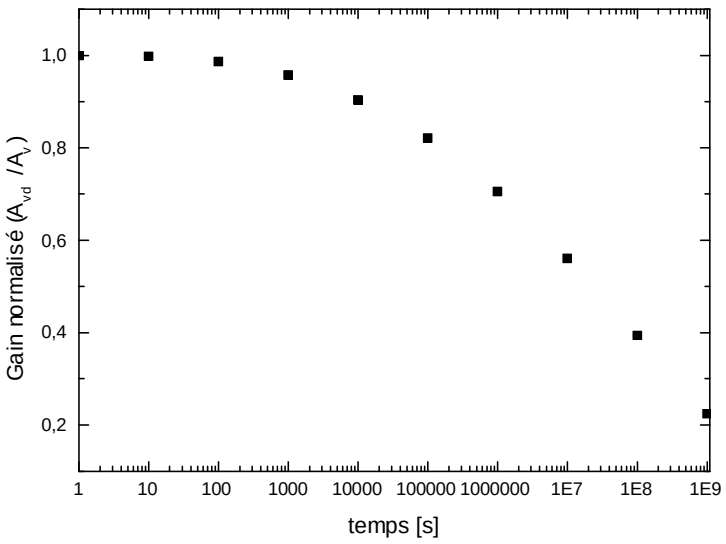
L'impédance de sortie de notre amplificateur opérationnel Z_{out} peut être donnée comme [66]:

$$Z_{out} \approx \frac{1}{2g_m} \qquad (3.41)$$

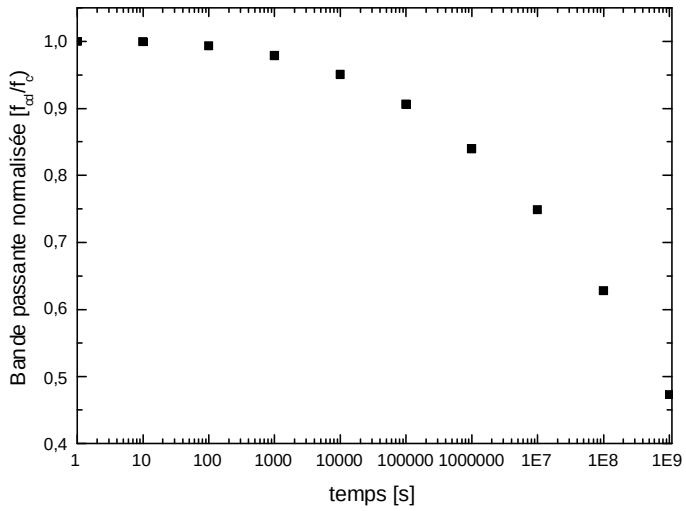
Donc l'impédance de sortie dégradée peut être donnée par l'expression suivante:

$$Z_{outd} = \frac{Z_{out}}{1 - 0.01D(t)} \qquad (3.42)$$

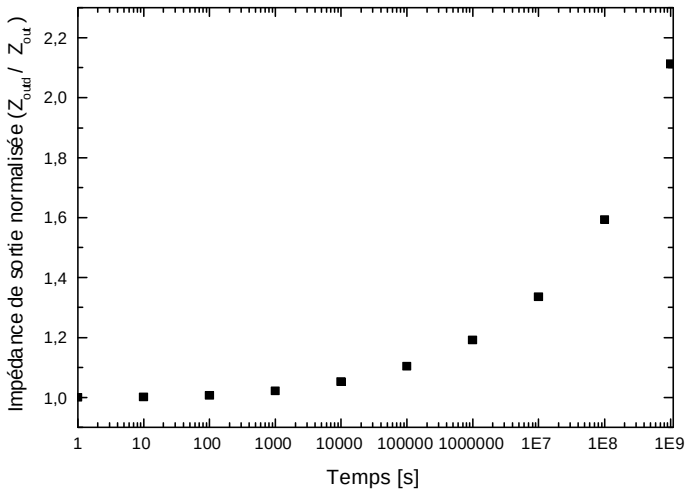
La figure ci-dessous montre l'évolution des différents paramètres normalisés de performance de notre amplificateur opérationnel comme fonction du temps.



(a)



(b)



(c)

Figure III. 24: Evolution des paramètres de performance normalisés en fonction du temps de stress (a) dégradation du gain A (b) dégradation de la bande passante (c) dégradation de l'impédance de sortie

III.4 Conclusion

Dans ce chapitre, nous avons démontré l'applicabilité de l'approche neuronale au problème de prédiction de la durée de vie des dispositifs intégrés fortement submicroniques. Une approche analytique basée sur un prédicteur neuronal a été développée dans le cas du transistor MOSFET fortement submicronique. Cette dernière nous a permis de prévoir l'évolution de la dégradation de la transconductance en fonction des différents paramètres (tension du drain, tension de grille, longueur du canal et temps de stress). L'approche développée peut également être implémentée dans les simulateurs des circuits électronique (SPICE, CADENCE,…) afin d'étudier la dégradation des circuits intégrés fortement submicroniques sans impact sur le temps de calcul et l'espace de stockage.

IV.1 Introduction

Les progrès de la technologie métal-oxyde-semiconducteur (MOS) conduisent à des transistors de taille nanométrique. A ce niveau de miniaturisation, les effets quantiques ne sont plus négligeables et modifient sensiblement les propriétés de transport des matériaux. Dans ce contexte, le formalisme des fonctions de Green hors-équilibre (NEGF) constitue une méthode pertinente pour décrire le comportement quantique des nano-transistors.

Parmi les différentes architectures émergentes, le transistor MOSFET double-grille apparaît comme un des meilleurs candidats pour relever le défi de réduction des transistors (paragraphe I.5.4). L'une des particularités de ce composant est sa très bonne immunité face aux effets canaux courts. La double-grille permet de mieux contrôler le potentiel du canal et de résister au courant tunnel source-drain. L'objectif de ce chapitre est de modéliser un MOSFET double-grille aux dimensions ultimes en utilisant le formalisme des fonctions de Green hors-équilibre afin de développer une approche basée sur les réseaux de neurones permettant la simulation des circuits électroniques à l'échelle nanométrique (balistique). Le transistor DG MOSFET symétrique illustré par la figure (IV.1) a été identifié par ITRS (International Technology Roadmap for semiconductors) en tant que la structure la plus prometteuse qui permet davantage de graduation dimensionnelle de CMOS en deçà de 65nm pour son courant d'entraînement plus élevé, la pente sous seuil améliorée, la conductivité pour les canaux courts et la flexibilité remarquable de conception des circuits intégrés à l'échelle nanométrique [28-29].

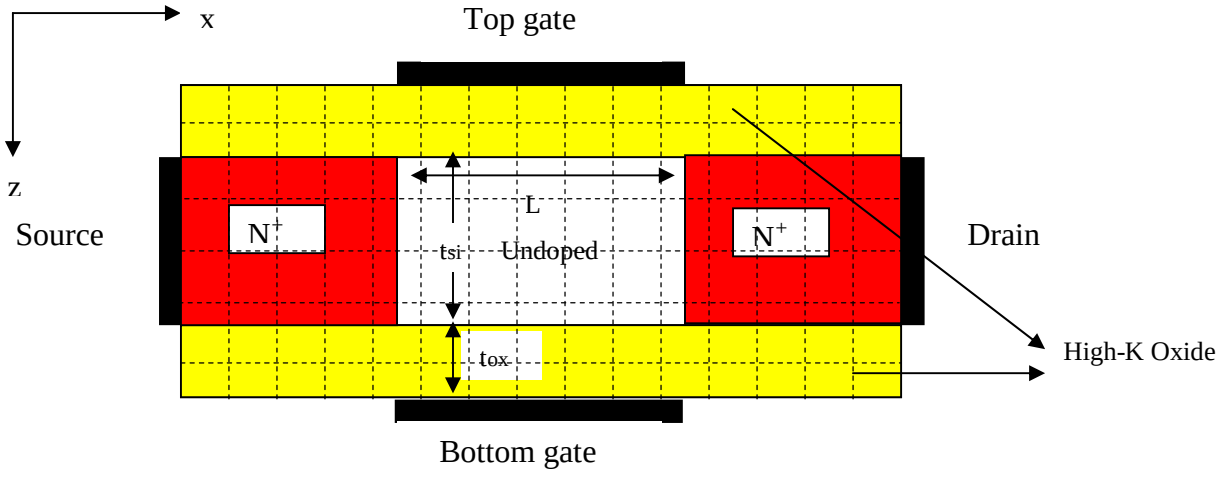


Figure IV.1: Transistor Double-Gate MOSFET ($N_{D/S}$ =10^{20} cm^{-3}, $L_{D/S}$=10nm et N_{CH}=10^{16} cm^{-3})

Le fonctionnement du transistor DG MOSFET est basé sur deux aspects fondamentaux: 1) la charge du canal induite par la grille sur la surface du substrat (les caractéristiques capacité- tension (C-V)) et 2) le transport des porteurs de charges à travers le canal (les caractéristiques courant- tension (I-V)). Une information exacte de la distribution des charges dans le canal exige la résolution de système d'équations Schrödinger-Poisson en se basant sur le formalisme des fonctions de Green hors-équilibre qui permet de calculer tout les effets quantiques (QM). Mais, pour la simulation des circuits nanoélectroniques, la résolution numérique du système d'équations Schrödinger-Poisson est indésirable à cause de sa complexité et de son temps de calcul élevé. Pour la modélisation analytique, en général, il est difficile ou presque impossible d'obtenir des modèles analytiques pour les nanocomposants. Cette modélisation demande plusieurs hypothèses simplificatrices, généralement nécessaires pour aboutir à des expressions analytiques afin d'étudier les différentes caractéristiques du transistor nanométrique [67]. La précision et la simplicité des modèles sont importantes pour la conception et la simulation des systèmes complexes. La modélisation basée sur les techniques l'intelligence artificielle qui constitue l'essentiel de ce chapitre, n'utilise en principe aucune hypothèse simplificatrice. Donc cette modélisation permet de fournir des solutions pratiques (précision et temps du calcul moins élevé) [68]. Dans ce chapitre, nous présentons l'applicabilité des réseaux de neurones artificiels (ANN) pour la simulation des circuits CMOS nanométriques. La base de données utilisée pour l'optimisation de notre réseau de neurone est établie sur la base d'un modèle numérique des caractéristiques courant-tension d'un transistor DG MOSFET développé en utilisant le formalisme des fonctions de Green hors-équilibre (NEGF) [69]. Ce modèle neuronal peut être utilisé comme une interface entre la modélisation numérique du composant et le simulateur électronique comme Berkeley SPICE, Philips'Pastar, Cadence Spectre, Anacad's Eldo afin d'avoir un simple simulateur des circuits électroniques nanométriques. Le schéma de notre approche peut être donné par la figure (IV.2).

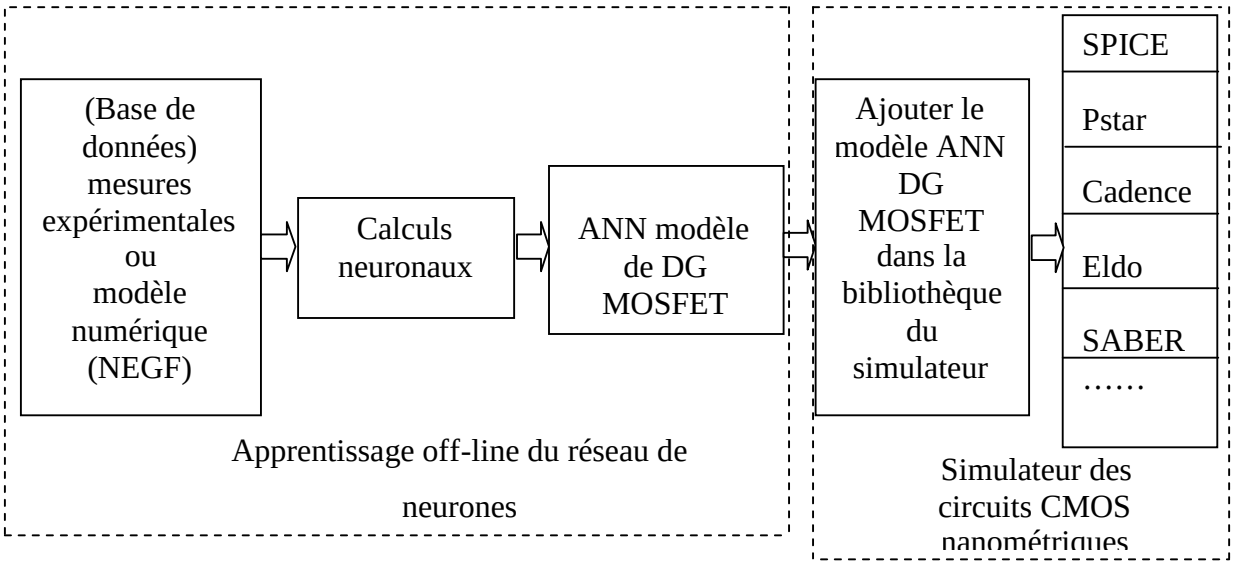


Figure IV.2: Schéma représentant notre approche pour la simulation des circuits CMOS nanométriques

IV.2 Méthodologie de modélisation

IV.2.1 Formalisme des fonctions de Green hors-équilibre (NEGF)

La structure de base du transistor DG MOSFET utilisé dans notre étude est montrée dans figure IV.1.

La réduction constante des dimensions des composants microélectroniques intégrés sur silicium conduit aujourd'hui à des structures de taille nanométrique faisant intervenir des dimensions de l'ordre de grandeur des distances inter atomiques. A ce stade de la miniaturisation, les méthodes semi-classiques couramment utilisées deviennent par conséquent inadaptées à la description des propriétés de transport de la matière, notamment sur le plan électrique. Les nanotransistors doivent donc être modélisés par une théorie qui puisse tenir compte des rapides variations spatiales et temporelles du champ électrique et qui inclue les effets quantiques. Plusieurs modèles ont été proposés. *Baccarani et Reggiani* ont par exemple présenté une approche qui couple le modèle de dérive-diffusion à celui hydrodynamique [70]. Le modèle de dérive-diffusion est utilisé pour décrire les effets quantiques tandis que l'approche hydrodynamique tient compte de la survitesse. En 1997 *Chang et Fossum* [71] ont également développé un modèle compact basé sur le second moment de l'équation de transport de Boltzmann afin de traiter l'effet de survitesse dans les MOSFETs ultimes. Néanmoins, la combinaison du modèle de dérive-diffusion et de l'hydrodynamique est inadaptée à la limite du transport purement balistique [72]. *Natori*

[73] fut le premier à présenter un modèle qui s'applique à ce type de régime en supprimant le concept de mobilité. Entre ces deux régimes extrêmes (diffusif et balistique), il existe un régime de transition dit "quasi-balistique". *Lundstrom et al.* [73, 74, 75] ont donc proposé un modèle plus général basé sur la théorie de la diffusion et capable de modéliser une large gamme de régimes. Dans cette approche, une partie du flux incident d'électrons provenant de la source traverse le canal et est récoltée par le drain. Le flux d'électrons restant est rétro-diffusé depuis le canal et retourne vers la source. La rétro-diffusion, définie par la fraction du flux d'électrons qui retourne vers la source, dépend de la diffusion des porteurs par le réseau et de la forme du potentiel électrostatique du canal. Les modèles *de Natori* et *Lundstrom* et *al* ne tiennent cependant pas compte de l'effet tunnel à travers la barrière de potentiel du canal, entre la source et le drain. Ce dernier, qui s'intensifie pour des longueurs de canal inférieures à 15 nm, a été inclus par la suite par plusieurs groupes. Citons tout d'abord le modèle quasi-2D de *Pirovano et al.* [76] qui couple le système auto-cohérent Schrödinger-Poisson aux équations semi-classiques de dérive-diffusion. *Picus et Likharev* [77] ont également proposé un modèle analytique pour décrire des MOSFETs de 10 nm. Enfin, *Svizhenko et al*. [78, 79] ont présenté une simulation 2D d'un MOSFET double-grille basée sur le formalisme des fonctions de Green hors-équilibre. Cette approche, numériquement améliorée par *Venugopal et al.* [80, 81, 82], a été successivement appliquée pour traiter l'influence de la forme des réservoirs (source et drain) sur le courant.

La fonction de Green est résolue pour obtenir la densité d'électron dans le dispositif et le courant sur les bornes dans la limite balistique. Dans des conditions balistiques, le formalisme des fonctions de Green hors-équilibre est mathématiquement équivalent à la résolution de l'équation de Schrödinger avec des conditions aux limites ouvertes [81, 82, 83]. La résolution de l'équation de Schrödinger nous permet d'obtenir les niveaux d'énergies et les fonctions d'ondes dans la direction du confinement quantique. Donc de l'équation de Schrödinger à résoudre est donnée comme:

$$-\frac{h^2}{2m_x^*}\frac{\partial^2}{\partial^2 x}\psi_i(x,z)-\frac{h^2}{2m_z^*}\frac{\partial^2}{\partial^2 z}\psi_i(x,z)-qV(x,z)\psi_i(x,z)=E_i(x)\psi_i(x,z) \qquad (4.1)$$

Où m_z^* est la masse effective de l'électron dans la direction verticale (z), $V(x,z)$ est le potentiel électrostatique, et $E_i(x)$ $et\,\psi_i(x,z)$, l'énergie et la fonction d'onde pour le mode i à la tranche x, respectivement (voir figure IV.1). Les fonctions d'ondes sont égales à zéro à l'interface oxyde /Si dans le cas où, la pénétration des électrons dans l'oxyde est

négligeable (autrement, la frontière zéro est prolongée aux interfaces de contact / oxyde). Pour résoudre le système d'équations donné par le formalisme de Green, une représentation en mode d'espace est utilisée dans la direction de confinement (z). Cette approche, réduit considérablement la taille du problème et fournit des très bons résultats [81,82]. La procédure de modélisation de notre structure (DG MOSFET) peut être donnée sous la forme suivante:

1. En commençant par le choix d'une base appropriée (ou représentation) dans laquelle tous les opérateurs sont discrétisés où le pas de discrétisation a=3A° (Fig IV.1).
2. nous multiplions les deux membres de l'équation de Schrödinger par l'opérateur du mode d'espace [$\delta^*(x-x')\varphi_i^*(x,z)$], l'expression (4.1) devient:

$$\int[\delta^*(x-x')\varphi_i^*(x,z)].[-\frac{h^2}{2m_x^*}\frac{\partial^2}{\partial^2 x}\psi_i(x,z)]dxdz + \int[\delta^*(x-x')\varphi_i^*(x,z)].[-\frac{h^2}{2m_z^*}\frac{\partial^2}{\partial^2 z}-qV(x,z)]\psi_i(x,z)dxdz$$

$$=E\int[\delta^*(x-x')\varphi_i^*(x,z)].\psi_i(x,z)dxdz \qquad (4.2)$$

où * présente le conjugué de la fonction. Nous savons que [$\delta^*(x-x')\varphi_i^*(x,z)$] est une fonction réelle, donc son conjugué reste la même fonction. D'après la définition de la fonction de Dirac (δ), le deuxième membre de l'équation (4.2) devient:

$$E\int \varphi_i^*(x',z)].\psi_i(x',z)dz=E\,\tilde{\psi}_i(x'). \qquad (4.3a)$$

Où $\tilde{\psi}_i(x')$ est le coefficient d'expansion de la fonction $\psi_i(x',z)$ par le vecteur de la représentation en mode d'espace $\varphi_i(x',z)$ (mode-space representation) donné comme:

$$\psi(x',z)=\sum_{i=1}^{\infty}\tilde{\psi}_i(x')\varphi_i(x',z)\text{, et }\int\varphi_i^*(x',z)\varphi_j(x',z)dz=\delta_{ij}\ , \qquad (4.3b)$$

avec δ_{ij} est la fonction de Dirac. En conservant les propriétés de la fonction de Dirac et les équations (4.1) et (4.3b), la deuxième partie du premier membre de l'équation (4.2) peut être donnée comme:

$$\int\varphi_i^*(x',z)].[-\frac{h^2}{2m_z^*}\frac{\partial^2}{\partial z^2}-qV(x',z)]\psi_i(x',z)dz=E_i(x')\tilde{\psi}_i(x'). \qquad (4.3c)$$

Finalement la première partie du premier membre peut être donnée comme [81]:

$$\int[\delta^*(x-x')\varphi_i^*(x,z)].[-\frac{h^2}{2m_x^*}\frac{\partial^2}{\partial x^2}\psi_i(x,z)]dxdz\ =\ -\frac{h^2}{2m_x^*}\frac{\partial^2}{\partial x'^2}\tilde{\psi}_i(x') \qquad (4.3d)$$

Donc l'équation (4.2) devient:

$$-\frac{h^2}{2m_x^*}\frac{\partial^2}{\partial x'^2}\tilde{\psi}_i(x')+E_i(x')\tilde{\psi}_i(x')=E\tilde{\psi}_i(x') \quad (4.4)$$

L'équation (4.4) est la transformation en mode d'espace de l'équation (4.1) où on peut voir une réduction du problème 2D à un problème 1D. L'équation (4.4) peut être résolue facilement par la méthode des différences finies (Fig.IV.1).

3. La troisième étape consiste à développer la matrice de l'hamiltonien (H) de l'équation de Schrödinger (4.4). Dans cette étude, le développement de l'opérateur (H) est basé sur l'approximation de la masse effective. Cette approximation est souvent appliquée aux semiconducteurs pour décrire le voisinage du minimum de bande de conduction. Dans cette approximation, la bande de conduction est parabolique ($E(k)=\frac{h^2k^2}{2m_e^{*2}}$), l'hamiltonien de l'équation (4.4) pour le mode i après la discrétisation par méthode des différences finies (Fig.IV.1) peut être donnée comme une matrice tridiagonale donnée comme le suivant:

$$h_i = \begin{bmatrix} -t_{x,i} & 2t_{x,I}+E_i(1) & -t_{x,i} & 0 & 0 & 0 \\ 0 & . & . & . & 0 & 0 \\ 0 & 0 & . & . & . & 0 \\ . & . & -t_{x,i} & 2t_{x,I}+E_i(N_X) & -t_{x,i} & \end{bmatrix} \quad (4.5)$$

où N_X est le nombre des points de la grille de discrétisation (Fig.IV.1) , $-t_{x,I}$ est l'énergie de couplage entre les points adjacents dans le canal donnée comme [83]:

$$t_{x,i}=\frac{h^2}{2m_{x,i}^* a^2} \quad (4.6)$$

où a est le pas de discrétisation et $m_{x,i}^*$, la masse effective longitudinale de l'électron pour le mode i.

4. Cette étape consacré au calcul des fonctions de Green et Self-énergie; le formalisme des fonctions de Green est particulièrement pertinent pour décrire des systèmes ouverts dont font partie les nano-structures connectées à des contacts semi-infinis. En effet, les fonctions de Green permettent de se concentrer sur la partie active du composant et de remplacer l'influence des contacts externes par des self-énergies. Le concept de self-énergie peut également servir à décrire les interactions électron-électron et électron-phonon [81].

Pour la sous bande i, la matrice G de la fonction de Green retardée de l'axe source-drain à une énergie E est donnée comme [81]:

$$G_i(E)=[EI-h_i-\Sigma]^{-1} \qquad (4.7)$$

où I représente la matrice unitaire, h_i est la matrice de l'hamiltonien du système donné par l'expression (4.6) et Σ_S et Σ_D sont la self-énergie de la source et du drain qui représentent l'effet du couplage du nano-composant (dans notre structure est le canal) à un réservoir (drain ou source). Cet effet peut être simplement décrit en ajoutant une self-énergie Σ à l'hamiltonien h_i. La self-énergie renormalise l'hamiltonien du nono-composant et remplace son spectre discret par un continuum (figure IV.3). Il s'agit d'un concept très général permettant d'éliminer les grands réservoirs et de travailler exclusivement dans le sous-espace de la zone active dont les dimensions sont nettement plus petites.

La matrice de la self-énergie dans notre cas peut être donnée par [84]:

$$\Sigma = \begin{bmatrix} \Sigma_S & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 0 \ldots & & \ldots. & & 0 \\ 0 & 0 & \ldots & 0 & & \Sigma_D \end{bmatrix} \qquad (4.7a)$$

où Σ_S et Σ_D sont définies dans le cas du transistor DG MOSFET par [84]:

$$\begin{aligned} &\Sigma_S(E)=-te^{jka} \text{ , où } E=E_i(1)+2t(1-\cos ka), \\ &\Sigma_S(E)=-te^{jka} \text{ , où } E=E_i(N_X)+2t(1-\cos ka), \end{aligned} \qquad (3.7b)$$

$E_i(1)$ est l'énergie au contact source/canal et $E_i(N_X)$ est l'énergie au contact canal/drain

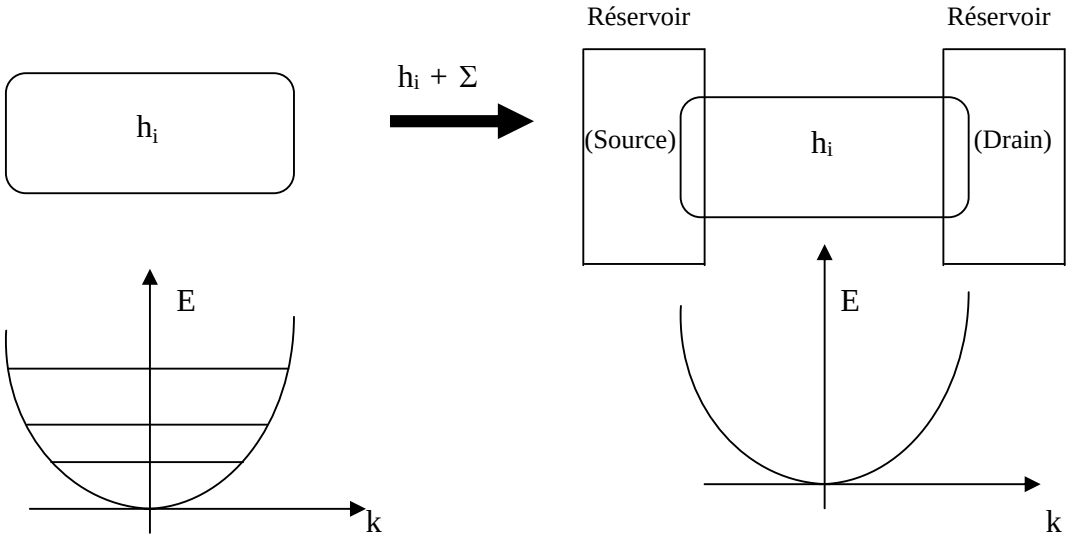


Figure IV.3: L'interaction du nano-système avec des contacts entraîne une renormalisation de son hamiltonien par une self-énergie Σ et la création d'un continuum d'énergie.

Une fois la fonction de Green retardée calculée, la densité d'électrons et le courant de drain sont déterminés. Définissons pour cela deux nouvelles quantités, fonctions des self-énergies:

$$\Gamma_S = i(\Sigma_S - \Sigma_S^T) \qquad \Gamma_D = i(\Sigma_D - \Sigma_D^T) \qquad (3.8)$$

où i^2=-1, et Σ_S^T est la matrice transposée de Σ_S

Physiquement, les fonctions Γ fixent la vitesse à laquelle un électron initialement dans un état particulier du canal se propage vers le réservoir de droite ou celui de gauche [85]. Les fonctions spectrales associées à la source et au drain sont alors exprimées par :

$$A_S = G\Gamma_S G^T \quad \text{et} \quad A_D = G\Gamma_D G^T \qquad (3.9)$$

La fonction spectrale de la source étant remplie selon la distribution de Fermi-Dirac de la source f_S (E), et la fonction spectrale du drain étant remplie selon la distribution de Fermi-Dirac du drain f_D (E), la matrice de la densité électronique s'écrit [81]:

$$n(E_l) = \frac{1}{ha}\sqrt{\frac{m^* K_B T}{2\pi^3}}[F_{-1/2}(\mu_S - E_l)A_S - F_{-1/2}(\mu_D - E_l)A_D] \qquad (3.10)$$

où $F_{-1/2}$ est l'intégrale de Fermi-Dirac (pour l'approximation analytique de $F_{-1/2}$, voir [86]). L'équation (3.10) rétro-agit alors sur l'équation de Poisson à deux dimensions afin d'obtenir une solution auto-cohérente. Le courant du canal est ensuite calculé. Ce dernier est donné par l'expression [81]:

$$I(E_l)=\frac{q}{h^2}\sqrt{\frac{m_y^* K_B T}{2\pi^3}}[F_{-1/2}(\mu_S-E_l)-F_{-1/2}(\mu_D-E_l)]T_{SD}(E_l) \qquad (3.11)$$

où $T_{SD}(E_l)$ est le coefficient de transmission de la source au drain [15,16,19] donné par:

$$T_{SD}(E_l)=Trace[\Gamma_S G \Gamma_D G^T] \qquad (3.11a)$$

Aspects numériques

Résolution de l'équation de Poisson à deux dimensions (2D)

L'équation de Poisson s'écrit:

$$\Delta V(x,z)=-\frac{\rho(x,z)}{\varepsilon_{SI}} \qquad (3.12)$$

avec ρ la densité de charges. La solution d'un tel système peut s'obtenir en maillant le domaine étudié en $N_X \times N_Z$ nœuds, où N_X et N_z représentent le nombre de nœuds suivant les directions x et z respectivement. La solution 2D de l'équation de Poisson est ainsi composée de $N_X \times N_Z$ valeurs de potentiels, initialement inconnus, correspondant à chaque nœud du réseau. Le canal du double-grille, intrinsèque, est connecté à deux réservoirs dopés n+. Néanmoins, la résolution de l'équation de Poisson se limite à la région du canal et de ses oxydes, le potentiel étant considéré constant au delà de ces régions. On conserve ainsi, à l'équilibre (V_G = V_{DS} = 0V), une barrière de potentiel entre la source et le drain due aux jonctions n+-int, tout en négligeant la pénétration du potentiel du canal dans les contacts (hypothèse des réservoirs parfaitement conducteurs présentant des réflexions aux interfaces). Afin d'obtenir les équations susceptibles de résoudre le système d'inconnues, nous devons appliquer l'équation (3.12) (dans laquelle les termes N_A, N_D et p sont maintenant nuls) aux nœuds internes et utiliser des conditions particulières aux limites pour les nœuds frontaliers. Etudions tout d'abord le cas d'un nœud interne quelconque [m, n] (ligne m et colonne n) de la figure (IV.4). L'approximation des différences finies aux dérivées spatiales exprime l'équation (3.12) sous la forme:

$$\frac{1}{a^2}V_{m-1,n}+\frac{1}{a^2}V_{m,n-1}-2\left(\frac{1}{a^2}+\frac{1}{a^2}\right)V_{m,n}+\frac{1}{a^2}V_{m,n+1}+\frac{1}{a^2}V_{m+1,n}=-\frac{1}{\varepsilon_{si}}q(N_D-N_A-n+p)_{m,n} \qquad (3.13)$$

où a est le pas du réseau dans les directions x et y respectivement. Suivant que le nœud [m, n] se situe dans les oxydes ou le silicium, la constante diélectrique ε est ε_{SiO_2} ou ε_{Si} . Dans

le cas où le nœud est positionné sur une interface Si/SiO_2, la continuité de la composante perpendiculaire $\varepsilon\vec{E}$ s'écrit :

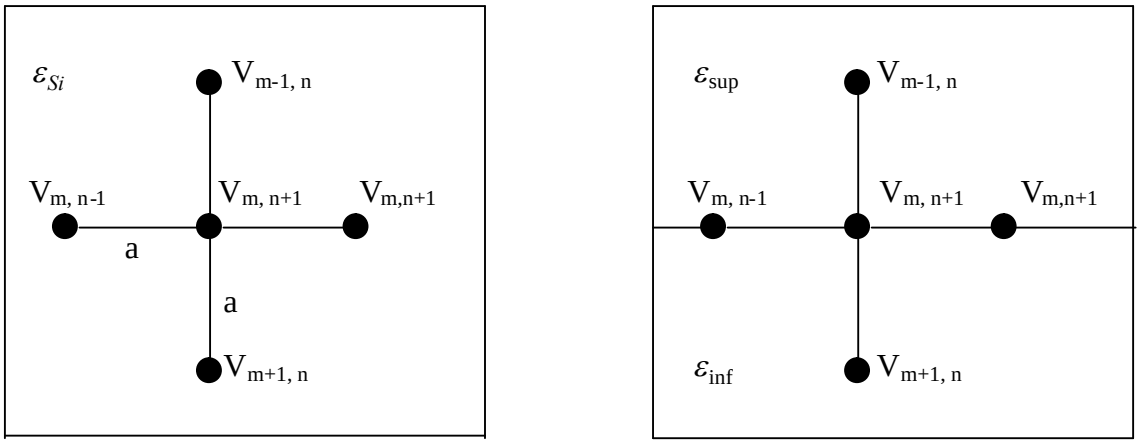
$$\varepsilon_{\text{sup}} E_{\text{sup}} = \varepsilon_{\text{inf}} E_{\text{inf}} \tag{3.14}$$

et également sous la forme:

$$\varepsilon_{\text{sup}} \left(\frac{\partial V}{\partial z}\right)_{\text{sup}} = \varepsilon_{\text{inf}} \left(\frac{\partial V}{\partial z}\right)_{\text{inf}} \tag{3.15}$$

où ε_{sup} et ε_{inf} sont les constantes diélectriques du matériau respectivement au-dessus et au-dessous de l'interface. En utilisant les notations de la figure (IV.4b), nous obtenons:

$$(4V_{m-1,n} - 3V_{m,n} - V_{m-2,n})\varepsilon_{\text{sup}} = (4V_{m+1,n} - 3V_{m,n} - V_{m+2,n})\varepsilon_{\text{inf}} \tag{3.16}$$



a) b)

Figure IV.4: Discrétisation de l'équation de Poisson en différences finies. a) Dans un matériau homogène de constante diélectrique ε b) A l'interface entre deux matériaux aux constantes diélectriques distinctes ε_{sup} et ε_{inf}.

Les conditions aux limites de Dirichlet sont imposées sur les nœuds appartenant au contour de la structure. L'équation à laquelle doit satisfaire le potentiel de grille est donc :

$$V_{m,n} = V_G - V_{FB} \tag{3.17}$$

où V_{FB} est la tension de bande plate qui traduit la différence des travaux de sortie du métal de grille et du silicium.

Les potentiels des extrémités latérales du canal (qui traduisent le début des réservoirs), respectent les égalités suivantes :

$$V_{m,n}=V_S=0V \tag{3.18}$$

$$V_{m,n}=V_D=V_{DS} \tag{3.19}$$

Connaissant la charge électronique n, les équations (3.12), (3.13), (3.16), (3.17) et (3.18) constituent un système linéaire dont la résolution peut être directement effectuée.

L'auto-cohérence

La méthode précédemment décrite est une approche mono-électronique auto-cohérente dans laquelle chaque électron, traité séparément, est soumis à un potentiel électrostatique résultant de son interaction coulombienne avec la densité électronique du système. La boucle d'auto-cohérence consiste à résoudre l'équation de Schrödinger, exprimée dans le formalisme des fonctions de Green, et à la coupler avec l'équation de Poisson. A partir d'un potentiel électrostatique d'essai, l'équation de Schrödinger fournit une densité de charges (équation (3.10)) qui est injectée dans l'équation de Poisson. Par une double intégration, nous obtenons un nouveau potentiel qui servira d'entrée à l'équation de Schrödinger. L'opération est ainsi réitérée jusqu'à convergence du potentiel électrostatique et des charges (figure IV.5).

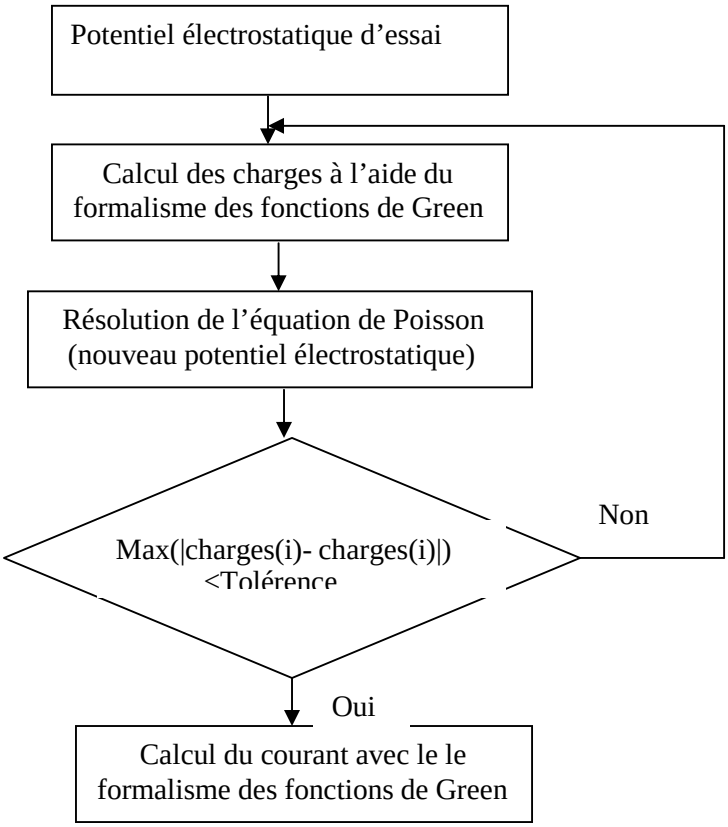


Figure IV.5: Etapes numériques de la simulation. L'indice i représente le nombre d'itérations de la boucle Schrödinger-Poisson.

La figure ci-dessous présente la caractéristique I_D=f (V_{DS}, V_{GS}) du transistor étudié (DG MOSFET) utilisant le formalisme des fonctions de Green.

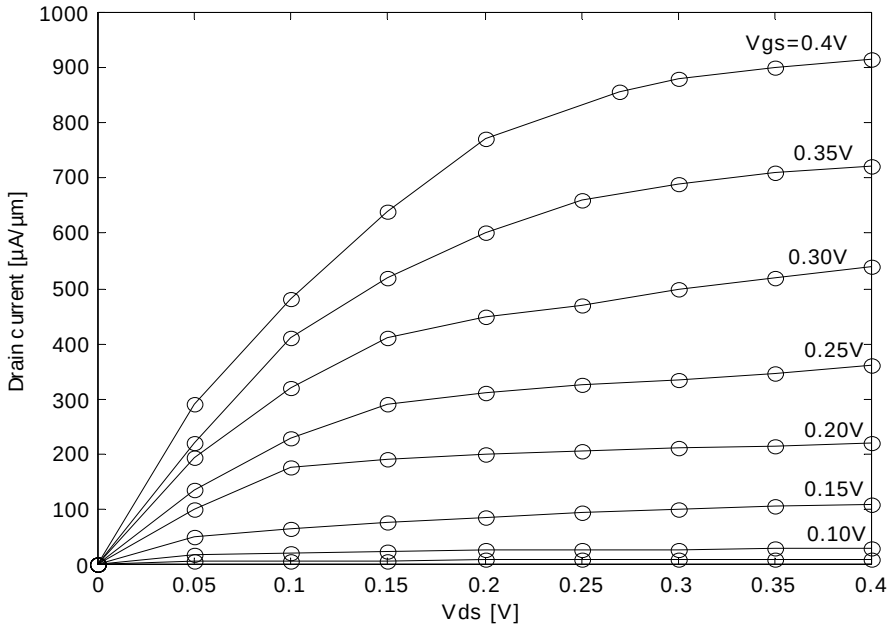


Figure IV.6: La caractéristique I-V du transistor DG MOSFET modélisé par le formalisme des fonctions de Green.

Dans ce travail, nous avons utilisé le formalisme des fonctions de Green afin de former une base de données qui va être utilisée pour l'optimisation de notre modèle neuronal du transistor DG MOSFET.

IV.2.2 Calcul neuronal

Le réseau de neurone développé est conçu afin de relier le vecteur d'entrée (L, Vds, Vgs, t_{ox} et t_{si}) au vecteur de sortie I_D. Chacun de ces paramètres est indexé par un neurone (figure IV.7) et présenté dans la structure neuronale comme une valeur formatée donnée par l'expression (2.5). La fonction d'activation utilisée dans notre structure est la fonction sigmoïdale (Tableau IV.1)

La base de données utilisée pour l'optimisation de la structure neuronal se compose de 460 échantillons (ce nombre d'échantillons est suffisant pour notre modélisation). Cette base de données a été divisée en deux ensembles: ensemble d'apprentissage (345 échantillons) et ensemble d'essais (115 échantillons). Le premier est utilisé pour ajuster les valeurs des poids et le second est utilisé pour examiner la configuration du réseau. Les échantillons d'essai et d'apprentissage doivent être différents et sont choisis aléatoirement à partir de la base de données originale (NEGF).

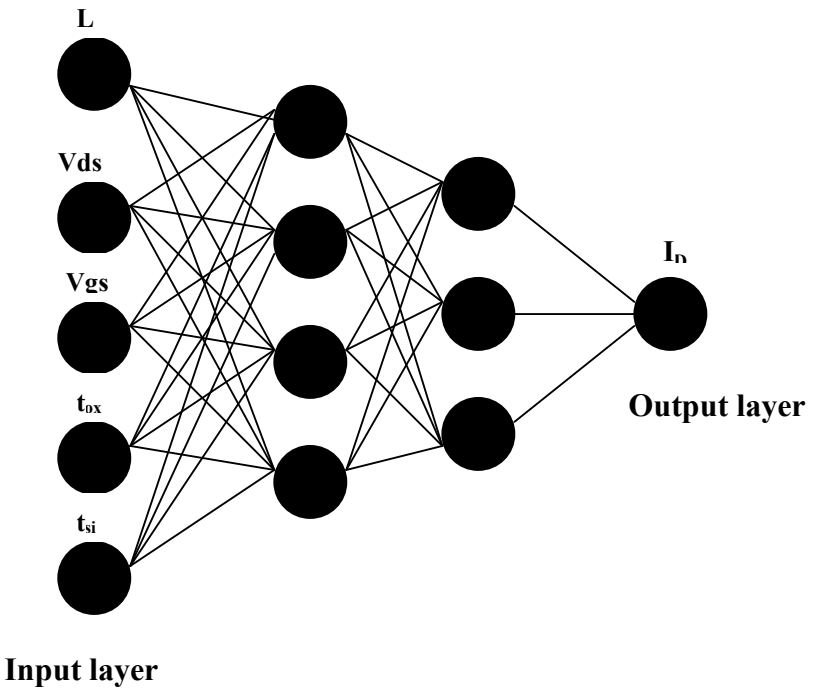


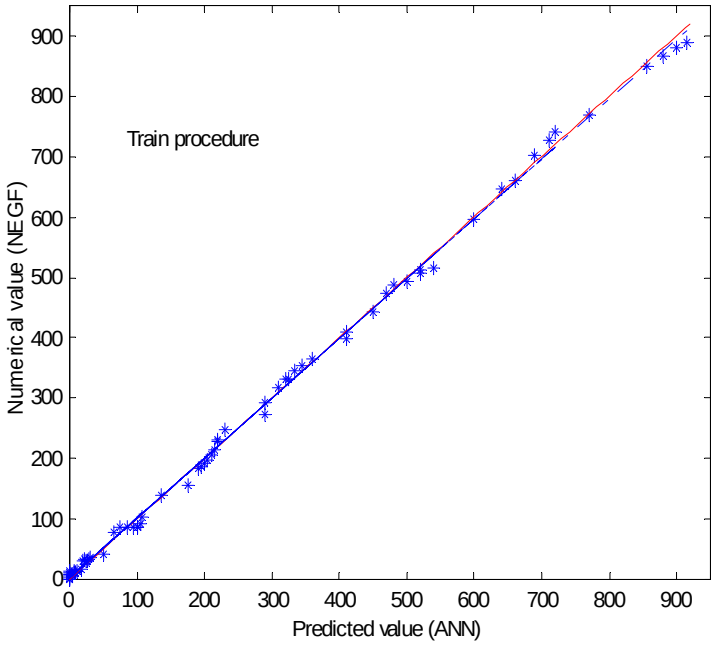
Figure. IV.7: Le modèle ANN optimisé du Transistor DGMOSFET

<table>
<tr><td colspan="2">Paramètres</td><td colspan="7">Valeurs optimisées</td></tr>
<tr><td colspan="2">Architecture</td><td colspan="7">Normal feed-forward MLP</td></tr>
<tr><td rowspan="8">Définition de la couche</td><td>Couches cachées</td><td colspan="7">2</td></tr>
<tr><td>apprentissage</td><td colspan="7">Rétro propagation rapide (Quick propagation)</td></tr>
<tr><td rowspan="5">Nombre de neurones</td><td rowspan="3">Entrées (I=4)</td><td></td><td>Vds()</td><td>Vgs ()</td><td>tsi</td><td>tox</td><td>L</td></tr>
<tr><td>Min</td><td>0</td><td>0</td><td>1</td><td>0.5</td><td>5</td></tr>
<tr><td>Max</td><td>0.5</td><td>0.6</td><td>10</td><td>2</td><td>50</td></tr>
<tr><td colspan="7">Première couche (N_1): variée</td></tr>
<tr><td colspan="7">Deuxième couche (N_2): variée</td></tr>
<tr><td rowspan="3">Sorties (O=1)</td><td></td><td colspan="5">I_D(µA/µm)</td></tr>
<tr><td>Min</td><td colspan="5">0</td></tr>
<tr><td>Max</td><td colspan="5">950</td></tr>
<tr><td>Population des poids</td><td colspan="7">$I*N_1+N_1*N_2+N_2*O$</td></tr>
<tr><td>Fonction d’entrée</td><td colspan="7">Produit scalaire entre les sorties de neurone
$w_{ij}o_j$</td></tr>
<tr><td>Traitement à l’entrée</td><td colspan="7">Variables entre 0 et 1
$\frac{x-x_{\min}}{x_{\max}-x_{\min}}$</td></tr>
<tr><td>Fonction d’activation</td><td colspan="7">Sigmoïdale
$\frac{1}{1+\exp(-x)}$</td></tr>
<tr><td colspan="2">Type d’erreur du réseau (Mean square error)</td><td colspan="7">Training process
$ETrn=\frac{1}{N_{Ptr}}(z_i-O(I_i,W))^2$ i=1, N_{Ptr}
Test process
$ETst=\frac{1}{N_{Ptst}}(z_i-O(I_i,W))^2$ i=1, N_{Ptst}</td></tr>
<tr><td rowspan="4">Apprentiss-age et test</td><td>Batch size</td><td colspan="7">1</td></tr>
<tr><td>Itération maximale</td><td colspan="7">10000</td></tr>
<tr><td>Tolérance</td><td colspan="7">0.001 (not reached)</td></tr>
<tr><td>séquence</td><td colspan="7">apprentissage + test passe après la mise à jour des poids</td></tr>
<tr><td>Taille de la base de données</td><td colspan="7">345 échantillons pour l’apprentissage
115 échantillons pour le test</td></tr>
</table>

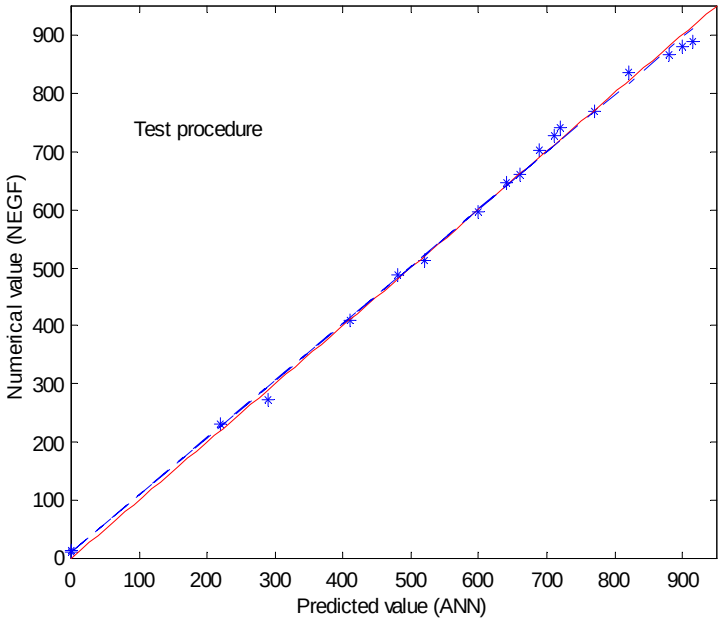
Tableau IV.1: paramètres de notre modèle neuronal après la procédure d’optimisation

IV.3 Résultas et discussion

Le processus d'optimisation (apprentissage et essai) a été aboutit pour 10000 cycles pour lesquels la stabilisation de l'erreur a été obtenue. Des nombres de neurone dans les premières et deuxièmes couches ont été changés et l'erreur associée d'optimisation a été enregistrée. Ce qui nous a permis d'obtenir une configuration optimale (figure IV.7). Pour cette structure, l'erreur résiduelle était 0,004 et presque 100% des cas soumis ont été appris correctement. Afin de valider la propriété prédictive de la structure optimisée de réseau, les ensembles d'essai et d'apprentissage ont été comparés à la réponse du modèle neuronal. Figure IV.8 montre que dans les deux cas, un très bon accord entre les résultats numériques (NEGF) et les résultats prévus (ANN) a été obtenu. Par conséquent, la structure optimisée peut être utilisée pour prédire d'autres combinaisons des variables d'entrée.



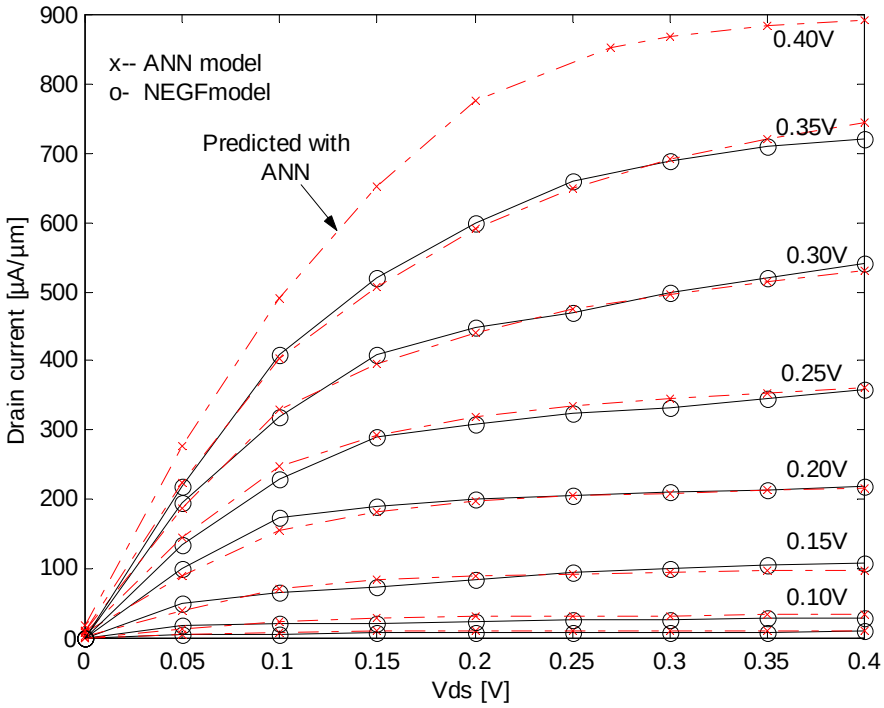
(a)



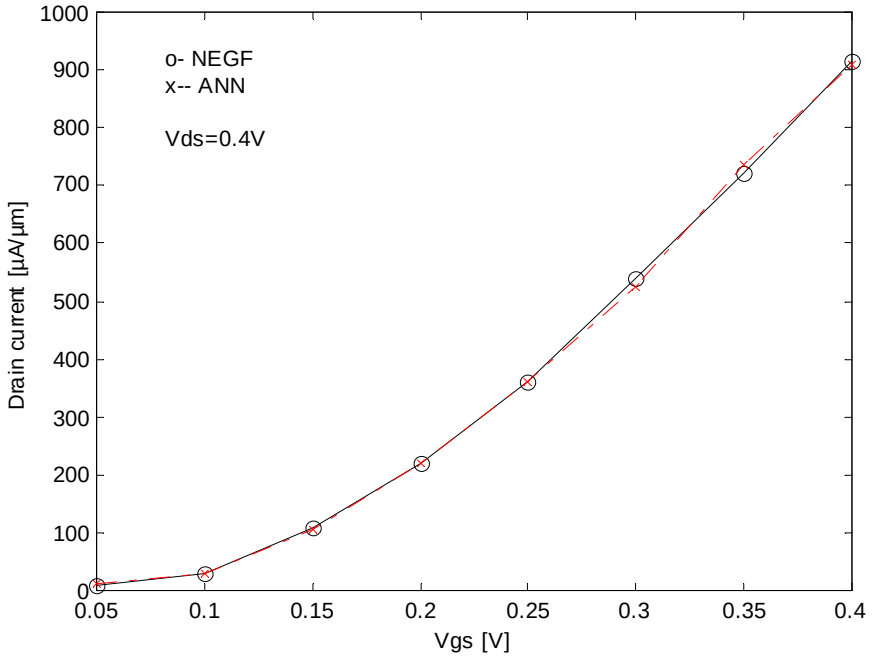
(b)

Figure IV. 8: validation de notre modèle neuronal pour les ensembles (a) apprentissage et (b) test

La figure IV.9 montre une comparaison entre les résultats prédits par le modèle neuronal (ANN) des différentes caractéristiques I-V (I_D-Vds et I_D-Vgs) avec ceux calculés par le formalisme des fonctions de Green (NEGF) pour un transistor DG MOSFET symétrique faiblement dopé avec L=10nm, t_{si}=3nm et t_{ox}=1nm. Comme il est montré, un très bon accord entre eux peut être observé pour toute la gamme de simulation. Cette dernière observation montre l'applicabilité des réseaux de neurones artificiels à l'étude des circuits CMOS nanométriques.



(a)

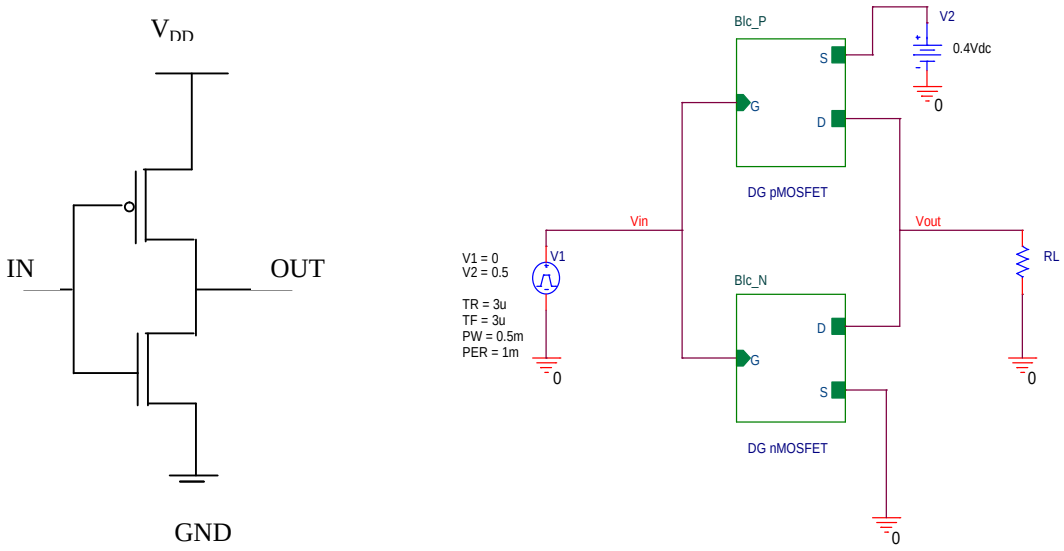


(b)

Figure IV.9: Caractéristiques Courant-Tension (I-V). (a) I_D-Vds (b) I_D-Vgs

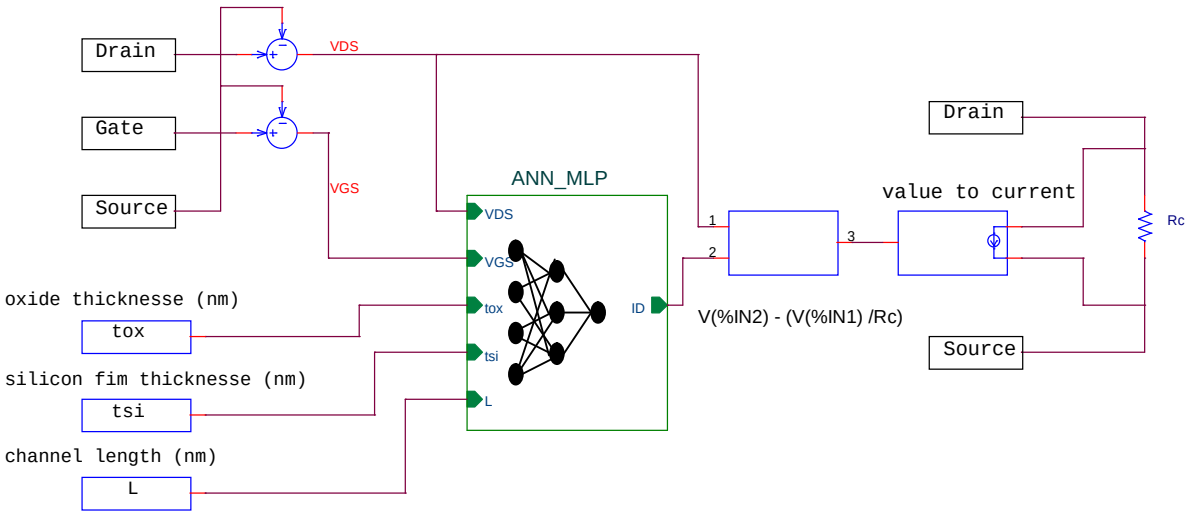
IV.3.1 Implémentation du modèle ANN

Afin de valider notre modèle neuronal, nous proposons l'étude et la simulation d'un circuit inverseur nanométrique. Ce dernier est considéré comme le plus important bloc pour la conception des circuits numériques VLSI [87]. Donc la modélisation et la simulation de l'inverseur nanométrique joue un rôle très important dans la conception des circuits électroniques VLSI. Comme l'architecture de l'inverseur est donnée par l'assemblage des deux transistors de type différent (nMOS et pMOS) (Figure IV.10a), donc la simulation de ce circuit est basée sur la modélisation de chaque type de transistor. Comme nous l'avons mentionné, le transistor MOSFET conventionnel peut être modélisé par des approches semiclassiques basées sur les équations de Boltzmann-Poisson [81]. Cependant, comme les progrès de la technologie MOS conduisent à des transistors de taille nanométrique, les effets quantiques ne sont plus négligeables et modifient ostensiblement les propriétés de transport des matériaux. Dans ce travail, en utilisant le modèle neuronal développé du transistor DG MOSFET, nous avons simulé un circuit inverseur. L'objectif de cette simulation est d'étudier l'évolution du gain de l'inverseur en fonction de la longueur du canal afin de tirer des conclusions pratiques concernant les limitations de la conception des circuits inverseurs en fonction de la longueur du canal. Chaque inverseur se compose de deux transistors DG MOSFETs. Les caractéristiques (I-V) de chaque transistor MOSFET ont été prévues en utilisant le modèle neuronal développé (ANN DG MOSFET) (Figure IV.10b). Dans les figures (IV.11a et IV.11b), nous montrons respectivement le modèle ABM (Analog Behavioral Modelling) de notre transistor DG MOSFET implémenté dans le logiciel de simulation PSPICE et le programme PSPICE qui correspond à notre modèle neuronal.



(a) **(b)**

Figure IV.10 : a) l'architecture du circuit inverseur b) modèle neuronal du circuit inverseur



(a)

*The ANN_MLP implementation :

```
      * source ANN_MLP
.EXTERNAL INPUT VDS
.EXTERNAL INPUT VGS
.EXTERNAL INPUT L
.EXTERNAL INPUT tox
.EXTERNAL INPUT tsi
.EXTERNAL OUTPUT ID
        *  first hidden layer
E_ANN_MLP_ABM1        ANN_MLP_N18934 0 VALUE {
+  1/(1+exp(-(B11+W111*V(VDS)+W112*V(VGS)+W113*V(TOX)+W114*V(TSI)+W115*V(L))))
+  }
E_ANN_MLP_ABM2        ANN_MLP_N18954 0 VALUE {
+  1/(1+exp(-(B12+W121*V(VDS)+W122*V(VGS)+W123*V(TOX)+W124*V(TSI)+W125*V(L))))
+  }
E_ANN_MLP_ABM3        ANN_MLP_N18978 0 VALUE {
+  1/(1+exp(-(B13+W131*V(VDS)+W132*V(VGS)+W133*V(TOX)+W134*V(TSI)+W135*V(L))))
+  }
E_ANN_MLP_ABM4        ANN_MLP_N19014 0 VALUE {
+  1/(1+exp(-(B14+W141*V(VDS)+W142*V(VGS)+W143*V(TOX)+W144*V(TSI)+W145*V(L))))
+  }
      * second  hidden layer
E_ANN_MLP_ABM5        ANN_MLP_N19042 0 VALUE {
+  1/(1+exp(-(B21+W211*V(ANN_MLP_N18934)+W212* V(ANN_MLP_N18954)+W213*
+  V(ANN_MLP_N18978)+W214* V(ANN_MLP_N19014))))    }
E_ANN_MLP_ABM6        ANN_MLP_N19054 0 VALUE {
+  1/(1+exp(-(B22+W221*V(ANN_MLP_N18934)+W222* V(ANN_MLP_N18954)+W223*
+  V(ANN_MLP_N18978)+W224* V(ANN_MLP_N19014))))    }
E_ANN_MLP_ABM7        ANN_MLP_N19058 0 VALUE {
+  1/(1+exp(-(B23+W231*V(ANN_MLP_N18934)+W232* V(ANN_MLP_N18954)+W233*
+  V(ANN_MLP_N18978)+W234* V(ANN_MLP_N19014))))    }
        * Output layer
E_ANN_MLP_ABM8        ANN_MLP_N19050 0 VALUE {
+  1/(1+exp(-(B31+W311*V(ANN_MLP_N19042)+W312*V(ANN_MLP_N19054)+W313*V(ANN_MLP_N19058))))
+    }
        * Drain current

E_ANN_MLP_ABM9        ID 0 VALUE { (V(ANN_MLP_N19050)*0.01)*0.001    }
```

(b)

Figure IV.11 : a) Modèle ABM de ANN DG MOSFET (b) Programme PSPICE de notre modèle neuronal [88]

La caractéristique de transfert (Vout - Vin) de l'inverseur peut être générée à partir des caractéristiques (I-V) de chaque transistor nMOSFET et pMOSFET.

Le gain d'un inverseur peut être donné comme [89]:

$$A_v = -\frac{\partial Vout}{\partial Vin}\Big|_{Vin=V_{DD}/2} \qquad (3.19)$$

De l'expression (3.19), le gain peut être calculé à partir de la fonction de transfert de l'inverseur (Vout-Vin).

Si on définit :

$$g_m = \frac{\partial I_{DS}}{\partial V_{gs}}\Big|_{Vgs=V_{ds}=V_{DD}/2} \qquad (3.20)$$

et

$$g_d = \left.\frac{\partial I_{DS}}{\partial V_{ds}}\right|_{V_{gs}=V_{ds}=V_{DD}/2} \qquad (3.21)$$

De (3.19), (3.20) et (3.21), le gain A_v peut être donné comme [89]:

$$A_V = \frac{g_m^n}{g_d^n} \qquad (3.22)$$

g_m^n et g_d^n représentent la transconductance et l'admittance de sortie du transistor DG nMOSFET (Figure IV.10a). Les signaux d'entrée/sortie de notre inverseur neuronal implémenté dans le logiciel PSPICE peuvent être représentés dans la figure (IV.12) où on peut remarquer que l'opération de l'inversion a été bien effectuée.

Pour voir l'effet de la longueur du canal sur le gain A_v, nous présentons dans la figure (IV.13a) les fonctions de transfert de deux inverseurs neuronaux pour L=10nm et L=5nm. Les gains de ces inverseurs peuvent être tracés en fonction de la longueur de canal (Figure IV.13b). Cette loi de variation peut être exprimée par une fonction linéaire comme le suivant:

$$A_v = \alpha . L + A_{v0} \qquad (3.23)$$

où α et A_{v0} sont des fonctions qui dépendent de la longueur de canal (L) et de l'épaisseur de l'oxyde (tsi). La connaissance de la variation du gain en fonction de la longueur du canal pour différentes épaisseurs d'oxyde et épaisseurs du canal nous permet la détermination de ces fonctions.

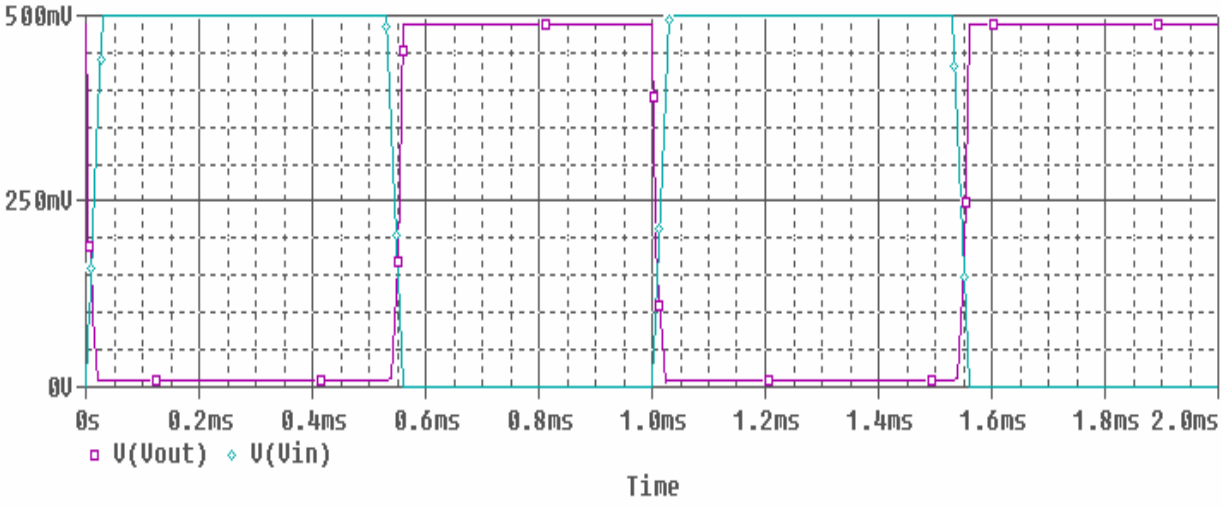
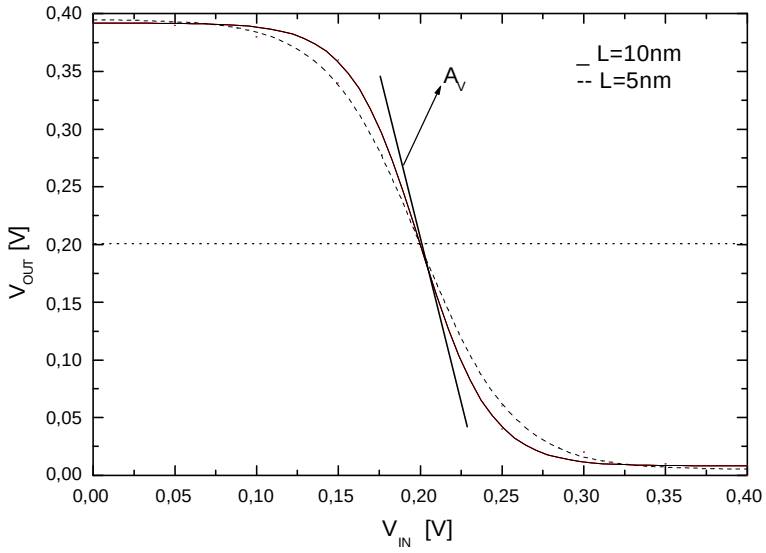
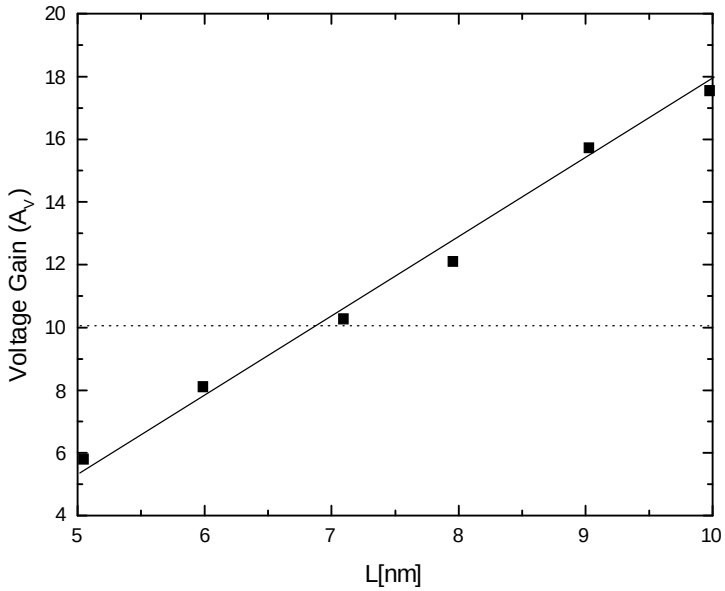


Figure IV.12: Les signaux PSPICE d'entrée/sortie de notre inverseur neuronal



(a)



(b)

Figure IV.13 : a) les fonctions de transfert des inverseurs neuronaux (L=5nm et L=10nm) b) la variation du gain en fonction de la longueur de canal pour (t_{si}=3nm and t_{ox}=1nm) [88].

L'expression (3.23) donne la loi de variation de gain de tension de notre inverseur neuronal en fonction de la longueur du canal, l'épaisseur du canal et l'épaisseur de l'oxyde, les résultats prédits par cette expression montrent bien la décroissance du gain de tension avec la diminution de la longueur du canal. Cette décroissance du gain peu être expliquée par l'apparition de l'effet du phénomène (DIBL) (Paragraphe I.4.2) [90]. Ce phénomène devient plus important quand la longueur du canal est diminuée. Comme conséquence, une dégradation des performances de l'inverseur nanométrique peut être observée (Figure IV.13b). En pratique, un gain de tension important peut assurer une vitesse de transition plus grande et des performances mieux de fonctionnement pour les opérations numériques (porte inverseuse, circuit buffer,..). Si on exige un gain A_V>10, la longueur de canal doit être supérieure à 6.8nm afin d'assurer le bon fonctionnement du circuit inverseur (Figure IV.13b).

IV.4 Conclusion

Dans ce chapitre, nous avons montré l'applicabilité de l'approche neuronale pour la conception des circuits électroniques nanométriques. Un modèle numérique des caractéristiques courant-tension (I-V) du transistor DG MOSFET a été développé en utilisant le formalisme des fonctions de Green hors-équilibre (NEGF). L'utilisation de ce modèle numérique nous a permis de former une base de données qui sera utilisée pour l'optimisation de notre structure neuronale (ANN). L'algorithme d'apprentissage retenu (quick propagation) a permis d'améliorer la convergence des résultas et de limiter le nombre de cycles d'apprentissage. Après l'optimisation, l'ensemble des poids généré peut être implémenté dans les logiciels de simulation (PSPICE, CADENCE,….) afin d'avoir notre modèle neuronal du transistor DG MOSFET. Les résultats obtenus nous ont encouragé d'adopter notre approche pour l'étude et la simulation des circuits électroniques nanométriques.

V.1 Introduction

La réduction constante des dimensions du transistor DG MOSFET conduit aujourd'hui à des structures de taille nanométrique faisant intervenir des dimensions de l'ordre de grandeur des distances inter atomiques. A ce stade de la miniaturisation, les méthodes classiques couramment utilisées deviennent par conséquent inadaptées à la prévision des limites technologiques (les contraintes imposées par les paramètres géométriques et physiques de la structure DG MOSFET) de fabrication des circuits intégrées nanométriques. Donc, en utilisant les techniques de l'intelligence artificielle (Réseaux de Neurones Artificiels), ce chapitre est consacré au développement des nouvelles abaques qui permettent de décrire la loi de la réduction dimensionnelle de la structure DG MOSFET en fonction des différents paramètres (longueur du canal L_i, du dopage Na et de l'épaisseur t_{si} du canal).

La modélisation du transistor DG MOSFET dans le régime sous seuil se fait actuellement de façon analytique [28]. Cette modélisation demande plusieurs hypothèses simplificatrices, généralement nécessaires pour aboutir à des expressions analytiques afin d'étudier les différentes caractéristiques du transistor nanométrique. Cependant, la modélisation semi-analytique à deux dimensions qui constitue l'essentiel de ce chapitre, n'utilise en principe aucune hypothèse simplificatrice. Dans ce cas, on résout le système d'équations non-linéaire Poisson-Boltzmann bidimensionnelle (2D) dans la région du canal, en utilisant la méthode des éléments finis pour développer un modèle semi-analytique qui permet de décrire la loi de transport des électrons dans le régime sous seuil (l'inverse de la pente sous seuil (S)). Donc, cette modélisation permet de fournir des solutions pratiques (précision et temps du calcul moins élevé) [91]. Dans ce chapitre, nous présentons l'applicabilité des réseaux de neurones artificiels (ANN) pour le développement des abaques qui permettent de décrire les limites technologiques imposées à la conception des transistors DG MOSFET. La base de données utilisée pour l'optimisation de notre structure neuronale est établie sur la base d'un modèle semi-analytique du transport des électrons dans le régime sous seuil (S) du transistor DG MOSFET développé en utilisant la méthode des éléments finis [91].

V.2 Méthodologie de modélisation

V.2.1 Formulation des éléments finis

Pour le canal totalement dépeuplé (FigureV.1), le modèle semianalytique de l'inverse de la pente sous seuil (S) pour différents modes de fonctionnement (symétrique ou asymétrique) peut être développé en se basant sur l'analyse du potentiel électrostatique dans le canal par la résolution du système d'équations non-linéaire bidimensionnelle (2D) Poisson-Boltzmann de la forme suivante:

$$\Delta\Psi=\frac{q}{\varepsilon_{si}}(N_A+n) \qquad (5.1a)$$

Où :

Ψ : le potentiel électrostatique .

n : la concentration d'électrons libres suit la distribution classique de Boltzmann:

$$n = n_i e^{\beta(\Psi-\phi_F)} \qquad (5.1b)$$

Φ_F : la différence entre le niveau de Fermi et le niveau quasi-Fermi d'électrons.

q : la charge électronique.

ε_{si} : la constante diélectrique de silicium.

N_A : le dopage du canal.

β : la tension thermique.

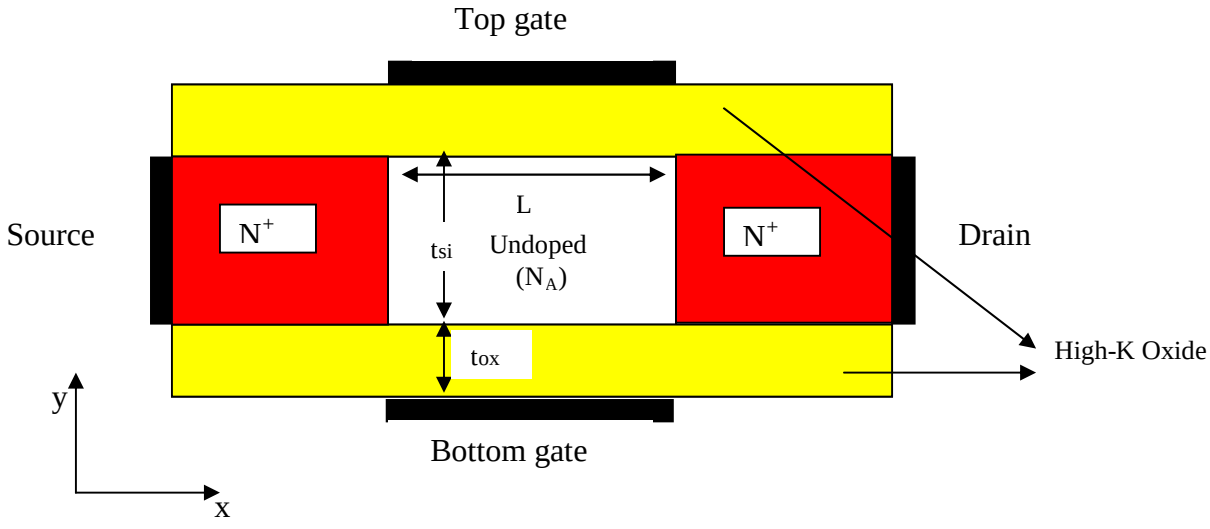


Figure V.1: Transistor Double-Gate MOSFET

Les conditions aux limites pour Ψ devient satisfaire la continuité du potentiel et la composante normale de vecteur du déplacement électrique aux interfaces Si/SiO2, et la continuité du potentiel [92].

$$\varepsilon_{ox}\frac{V_{Feff}-\Psi(x,0)}{t_{ox}}=\varepsilon_{si}\frac{\partial\Psi(x,y)}{\partial y}\Bigg|_{y=0} \qquad (5.2a)$$

$$\varepsilon_{ox}\frac{V_{Beff}-\Psi(x,t_{si})}{t_{ox}}=\varepsilon_{si}\frac{\partial\Psi(x,y)}{\partial y}\Bigg|_{y=t_{si}} \qquad (5.2b)$$

$$\Psi(0,y)=V_{bii} \qquad (5.2c)$$

$$\Psi(L,y)=V_{bii}+V_{D/S} \qquad (5.2d)$$

Où :

V_{bii} : la tension de jonction entre la source / drain et le silicium intrinsèque (canal).

$$V_{bii}=\frac{K_B T}{q}\ln\left(\frac{N_{D/S}}{n_i}\right)$$

$N_{D/S}$: la concentration du dopage de la source et le drain.

V_{DS} : la tension drain-source.

Les tentions efficaces aux grilles supérieure (Front) et inférieure (Bottom), V_{Feff} et V_{Beff} sont présentées, pour simplifier les notations, tel que :

$$V_{Feff}=V_{GSF}-(\Phi_{MF}-\Phi_i) \qquad (5.3a)$$

$$V_{Beff}=V_{GSB}-(\Phi_{MB}-\Phi_i) \qquad (5.3b)$$

Où :

Φ_i : le travail de sortie du silicium intrinsèque.

Quand $V_{Feff}=V_{Beff}$ le champ électrique dans la direction verticale (y) est symétrique par rapport au centre du canal ($y=t_{si}/2$) qui correspond au transistor DG MOSFET symétrique. Autrement, le transistor DG MOSFET est dans une configuration électriquement asymétrique ; dans ce cas, il existe deux possibilités pour avoir cette configuration du transistor DG MOSFET. La première consiste à polariser les deux grilles par des tensions différentes ($V_{Feff}\neq V_{Beff}$), la deuxième est basée sur le changement de l'épaisseur de la couche isolante (t_{ox1} ou t_{ox2}) [93]. Dans le cas du transistor DG MOSFET asymétrique ($V_{Feff}\neq V_{Beff}$), le champ électrique dans la direction verticale (y) est asymétrique par rapport au centre du canal.

Il est à noter que l'étude de transistor DG MOSFET symétrique est similaire à celle de la configuration asymétrique sauf que les conditions de Neumann dans ce cas sont données comme[92]:

$$\varepsilon_{ox}\frac{V_{Feff}-\Psi(x,0)}{t_{ox1}}=\varepsilon_{si}\frac{\partial\Psi(x,y)}{\partial y}\Bigg|_{y=0} \qquad (5.4a)$$

$$\varepsilon_{ox}\frac{V_{Beff}-\Psi(x,t_{si})}{t_{ox2}}=\varepsilon_{si}\frac{\partial\Psi(x,y)}{\partial y}\Bigg|_{y=t_s} \qquad (5.4b)$$

Nous avons donc un problème bidimensionnel du second ordre défini à l'intérieur du canal exprimé par le système d'équations (5.1) et par les conditions aux limites (les expressions (5.2a), (5.2b) pour le mode symétrique et (5.4a), (5.4b) pour le mode asymétrique) aux interfaces Si/SiO2 (condition de Neumann) et (les expressions (5.2c) et (5.3d)) aux cotés source/drain (condition de Dirichlet).

La forme intégrale pour le formalisme éléments finis est:

$$R(\Psi)=-\iint\left[\frac{\partial w}{\partial x}\frac{\partial\Psi}{\partial x}+\frac{\partial w}{\partial y}\frac{\partial\Psi}{\partial y}-\frac{q(N_A+n)}{\varepsilon_{si}}\right]dA+\oint_{\Gamma}w\frac{\partial\Psi}{\partial n}ds=0 \qquad (5.5)$$

L'assemblage de cette équation sur le domaine de résolution aboutit au système matriciel:

$$R(\Psi)=[C].[\Psi]-[B]-[F(\Psi)]=0 \qquad (5.6)$$

[C]: la matrice de raideur ;

[Ψ]: le vecteur des potentiels inconnus ;

[B]: le vecteur résultant de la condition $\partial\Psi / \partial n$ sur la frontière ;

[F(Ψ)]: le vecteur des sources du champ .

Les termes élémentaires du système (5.5) sont calculés par:

$$C_{ij}=\iint_{\Omega}\nabla w_i\nabla w_j\,dxdy \qquad (5.6a)$$

$$F_i=\int w_i\left(\frac{q(N_A+n)}{\varepsilon_{si}}\right)dxdy \qquad (5.6b)$$

$$B_j=\int_{\Gamma}w_i\frac{\partial\Psi}{\partial n}\partial\Gamma \qquad (5.6c)$$

Les éléments du vecteur [B] sont non nuls sur les frontières Si/SiO2 (interface isolant / semi-conducteur).

Le système non-linéaire (5.6) est résolu par la méthode de Newton-Raphson [94](voir Annexe A), où la matrice Jacobienne [*J*] pour notre problème est donnée comme:

$$J_{ij}=\frac{\partial R_i}{\partial\Psi_j}=K_{ij}+\sum_{k=1}^{r}\frac{\partial K_{ik}}{\partial\Psi_j}\Psi_k-\frac{\partial F_i}{\partial\Psi_j} \qquad (5.7)$$

Cette expression peut être donnée sous forme matricielle comme:

$$[J]=[K]+[\Delta F] \tag{5.8}$$

où:

$$\Delta F_{ij}=\frac{\partial F_i}{\partial \Psi_j} \tag{5.8a}$$

Il est à noter que l'élément de maillage utilisé dans notre étude est triangulaire à trois nœuds (figure V.2) [95].

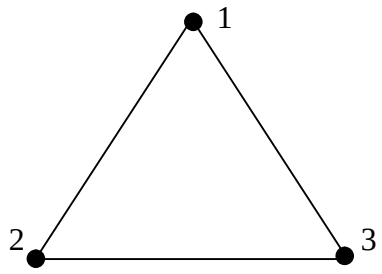


Figure V.2: Elément triangulaire à trois nœuds de maillage utilisé dans notre cas

V.2.2 Modélisation de l'inverse de la pente sous seuil (S)

L'inverse de la pente sous seuil est donné de manière générale par [28]:

$$S=\frac{\partial V_{GS}}{\partial \log_D} \tag{5.9}$$

C'est à dire par la variation de la tension de grille par rapport au courant de canal sous seuil.

En supposant que le courant de drain (I_D) est proportionnel aux porteurs libres montant à la cathode virtuelle (c.-à-d., où le potentiel électrostatique du canal atteint son minimum Ψ_m (y)), l'expression (5.9) peut être transformée [28] en :

$$S=\frac{K_B T}{q}\ln 10\left[\frac{\int_0^{t_{si}}\exp(\beta\Psi_m)\left(\frac{\partial\Psi_m}{\partial V_{GS}}\right)dy}{\int_0^{t_{si}}\exp(\beta\Psi_m)dy}\right]^{-1} \tag{5.10}$$

Par conséquent, le développement d'un modèle de S est basé sur la détermination du potentiel minimum de canal $\Psi_m(y)$ et sa dépendance de la tension de la grille (V_{GS}). Donc, le calcul du potentiel du canal par la méthode des Eléments Finis nous permet de déterminer la variation du potentiel minimum de canal $\Psi_m(y)$ en fonction de la tension de grille (V_{GS}) tel que:

Le potentiel minimum du canal $\Psi_m(y)$ peut être déterminé par la résolution de l'équation: $\partial\Psi(x,y)/\partial y=0$.

Pour différentes valeurs de (V_{GS}) et de dopage. La forme générale de $\Psi(x,y)$ (figure V.3) s'écrit comme:

$$\Psi(x,y)=a(y)x^2+b(y)x+c(y) \qquad (5.11)$$

avec:

$$a(y)=\sum_{i=0}^{2} a_i\, y^i \qquad (5.11a)$$

$$b(y)=\sum_{i=0}^{2} b_i\, y^i \qquad (5.11b)$$

$$c(y)=\sum_{i=0}^{2} a_i\, y^i \qquad (5.11c)$$

tel que:

a(y) , b(y) et c(y) sont des fonctions polynomiales qui varient en fonction de y (la position verticale du canal).

La dérivation de $\Psi(x,y)$ par rapport à la position verticale du canal (y) nous donne l'expression de $\partial\Psi(x,y)/\partial y$:

$$\partial\Psi(x,y)/\partial y=(2a_1 y+a_2)x^2+(2b_1 y+b_2)x=0 \qquad (5.12)$$

La résolution de cette équation nous donne :

$$x_1=0 \quad \text{et} \quad x_2=-\frac{2b_1 y+b_2}{2a_1 y+a_2} \qquad (5.13)$$

On n'accepte que la valeur non nulle de ces solutions (c'est à dire x_2).

De (5.11) et (5.13) ; le potentiel minimum de canal $\Psi_m(y)$ peut être exprimé par:

$$\Psi_m(y)=a(y)\left(-\frac{2b_1 y+b_2}{2a_1 y+a_2}\right)^2-b(y)\left(-\frac{2b_1 y+b_2}{2a_1 y+a_2}\right)+c(y) \qquad (5.14)$$

a. DG MOSFET symétrique

Dans cette configuration, les coefficients a_i , b_i et c_i sont donnés par le tableau suivant:

	i = 0	i = 1	i = 2
$a_i(y)$	$-82.13 \cdot 10^{-5}$	$-8.9 \cdot 10^{-7}$	$4 \cdot 10^{-7}$
$b_i(y)$	$213.0667 \cdot 10^{-4}$	$2.6667 \cdot 10^{-4}$	$13.33 \cdot 10^{-6}$
$c_i(y)$	0.5859	0	0

Tableau V.1: Les valeurs des coefficients de $\Psi_m(y)$ pour $N_A=5.10^{14}cm^{-3}$ et $V_{GS}=-0.1V$ (DG MOSFET symétrique).

Donc, la solution de l'équation (5.12) est donnée par :

$$x=-\frac{5.3.10^{-4}\ y+13.33.10^{-6}}{-17.8\ y+4.10^{-7}} \tag{5.15}$$

En remplaçant (5.15) dans (5.11), on aura :

$$\Psi_m(y)=a(y)\left(-\frac{5.3.10^{-4}\ y+13.33.10^{-6}}{-17.8\ y+4.10^{-7}}\right)^2-b(y)\left(-\frac{5.3.10^{-4}\ y+13.33.10^{-6}}{-17.8\ y+4.10^{-7}}\right)+c(y) \tag{5.16}$$

b . DG MOSFET asymétrique

Dans ce cas, les paramètres a_i, b_i et c_i sont donnés par le tableau suivant:

	i = 0	i = 1	i = 2
a_i	$-71.2 . 10^{-5}$	$-9.3 . 10^{-6}$	$2.3 . 10^{-7}$
b_i	$196.067 . 10^{-4}$	$1.92 . 10^{-4}$	$12.93 . 10^{-6}$
c_i	0.6135	0.001	0.003

Tableau V.2: Les valeurs des coefficients de $\Psi_m(y)$ pour $N_A=5.10^{14}cm^{-3}$ et V_{GS}=-0.1V (DG MOSFET asymétrique).

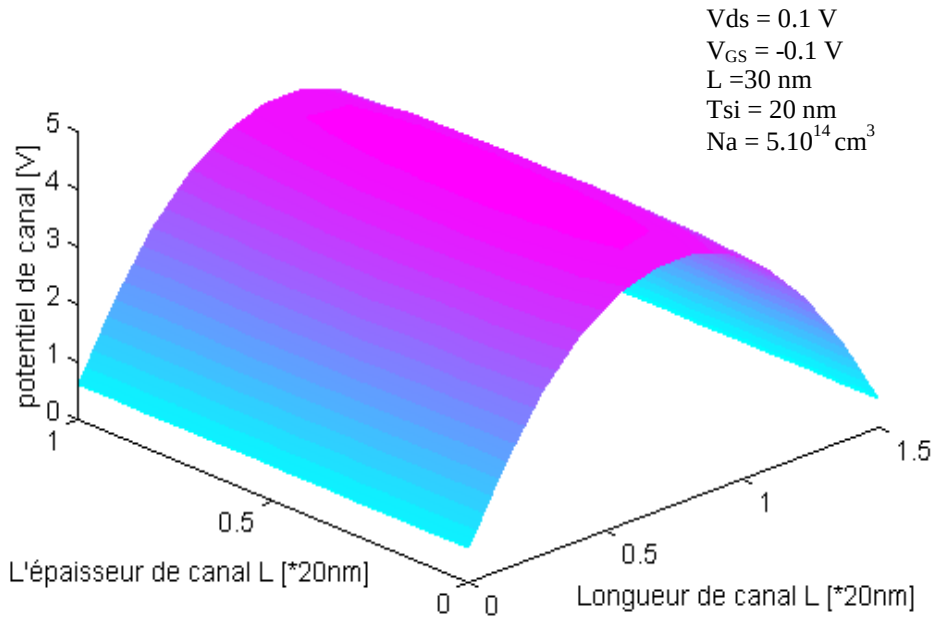
La solution de l'équation (5.12) peut être donnée comme:

$$x=-\frac{6.1.10^{-4}\ y+19.26.10^{-6}}{-16.9\ y+3.6.10^{-7}} \tag{5.17}$$

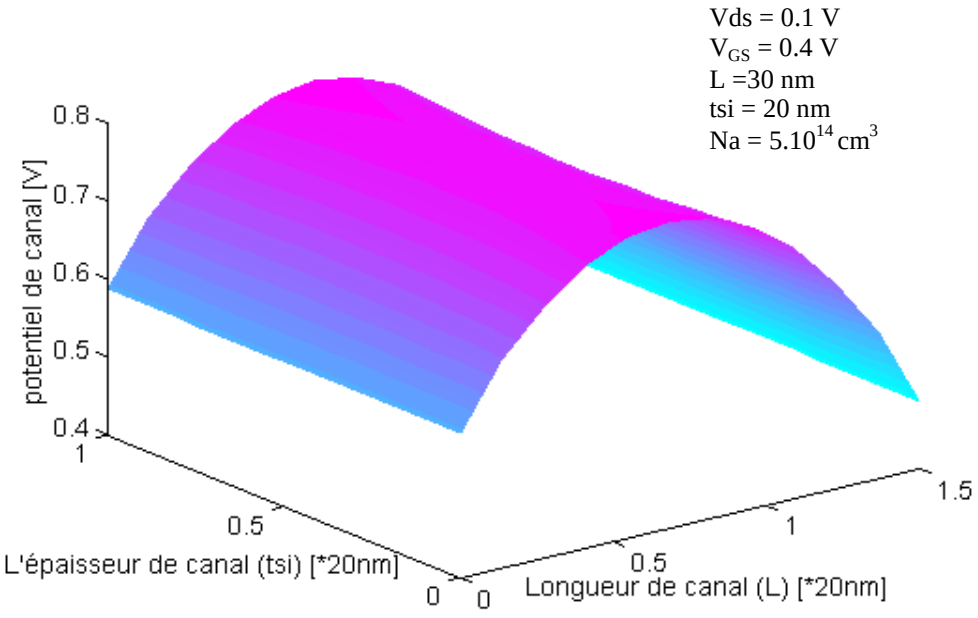
On remplace (5.17) dans (5.11), on aura:

$$\Psi_m(y)=a(y)\left(\frac{0.23\ y-1.93}{0.09-0.072\ y}\right)^2-b(y)\left(\frac{0.325\ y-3.152}{0.095-0.003\ y}\right)+c(y) \tag{5.18}$$

Les figures V.3 et V.4 représentent, respectivement, les variations du potentiel de canal pour différentes tensions de polarisation de grille (V_{GS}) pour le transistor DG MOSFET symétrique et asymétrique.

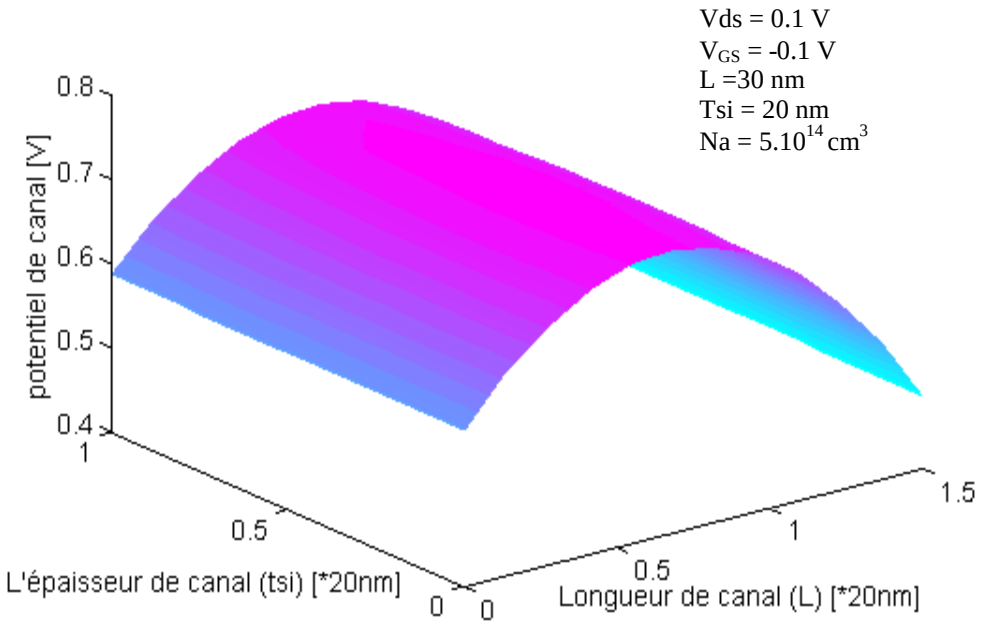


(a)

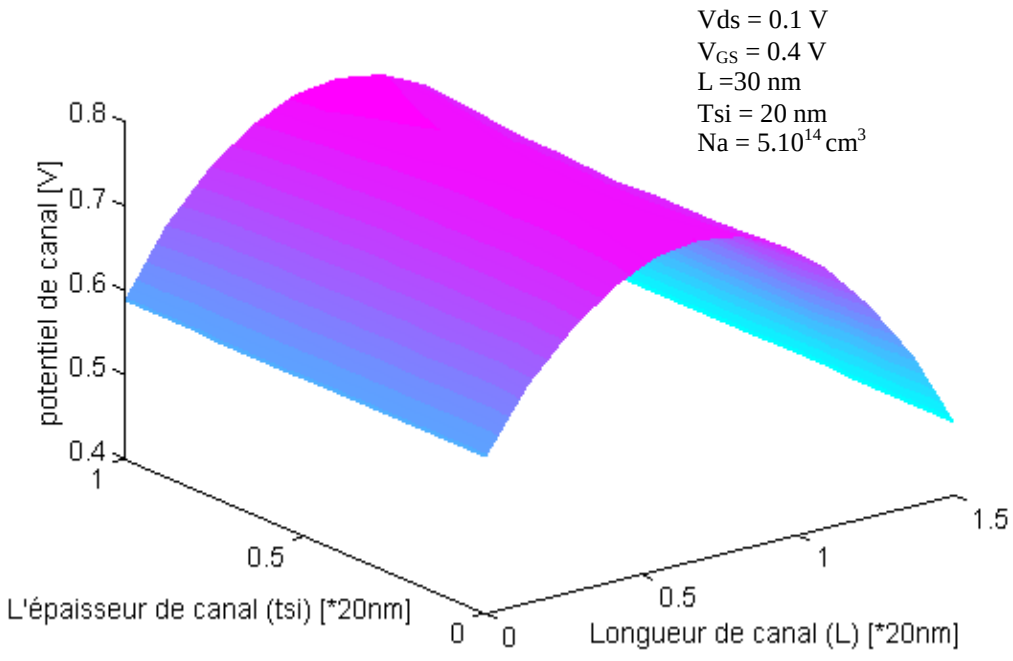


(b)

Figure V. 3 : Variation du potentiel du canal du transistor DG MOSFET symétrique.



(a)



(b)

Figure V. 4 : Variation de potentiel du canal de transistor DG MOSFET asymétrique

La variation du potentiel minimum de canal $\Psi_m(y)$ en fonction de la tension de commande (V_{GS}) est donnée pour les deux cas (symétrique et asymétrique) par les figures (V.5) et (V.6). Cette variation du potentiel minimum $\Psi_m(y)$ en fonction de la tension de commande V_{GS} peut être exprimée comme:

$$\Psi_m(y, V_{GS}, N_A) = d(V_{GS}, N_A)y^2 + e(V_{GS}, N_A)y + f(V_{GS}, N_A) \qquad (5.19)$$

où:

$d(V_{GS}, N_A)$ $e(V_{GS}, N_A)$ *et* $f(V_{GS}, N_A)$ sont des fonctions polynomiales qui varient en fonction de la tension de commande V_{GS} et le dopage N_A.

Les fonctions polynomiales $d(V_{GS})$, $e(V_{GS})$ *et* $f(V_{GS})$ peuvent être représentées pour différents dopages par les tableaux suivants :

	5.10^{14}	10^{15}	10^{16}	10^{17}	5.10^{17}	10^{18}	5.10^{18}
$d(V_{GS})$	d_1=56.43 10^{-5} d_0=5.54 10^{-5}	d_1=61.14 10^{-5} d_0=14.81 10^{-5}	d_1=0.0006 d_0=-0.0014	d_1=0.0018 d_0=-0.0192	d_1=0.0004 d_0=-0.0817	d_1=0.0005 d_0=-0.1635	d_1=0.0007 d_0=-0.7386
$e(V_{GS})$	e_1=-0.0113 e_0 =0.0011	e_1=-0.0122 e_0 =0.0030	e_1=-0.0118 e_0 =0.0289	e_1=-0.0366 e_0 =0.3850	e_1=-0.0084 e_0 =1.6344	e_1=-0.0094 e_0 =3.2702	e_1=-0.0132 e_0 =14.7729
$f(V_{GS})$	c_1=0.1390 c_0=0.7346	c_1=0.1390 c_0=0.9230	c_1=0.1390 c_0=4.3131	c_1=0.2018 c_0=38.0109	c_1=0.1340 c_0=188.9474	c_1=0.1390 c_0=377.3569	c_1=0.2 c_0=1884.5

Tableau V.3: Les fonctions $d(V_{GS})$, $e(V_{GS})$ *et* $f(V_{GS})$ pour différents dopages N_A (DG MOSFET symétrique)

Dopage m^{-3}	5.10^{14}	10^{15}	10^{16}	10^{17}	5.10^{17}	10^{18}	5.10^{18}
$d(V_{GS})$	d_1=18.53 10^{-5} d_0=-3.74 10^{-5}	d_1=21.83 10^{-5} d_0=-4.4 10^{-5}	d_1=71.69 10^{-5} d_0=- 6 . 10^{-4}	d_1=0.0005 d_0=– 0.0058	d_1=0.0006 d_0 =–0.0288	d_1=0.0005 d_0=-0.0576	d_1=-0.001 d_0=–0.2867
$e(V_{GS})$	e_1=-0.0035 e_0 =0.0007	e_1=-0.0042 e_0 =0.0008	e_1=-0.0134 e_0 =0.0068	e_1=-0.009 e_0 =0.0625	e_1=-0.0114 e_0 =0.3071	e_1=-0.0091 e_0 =0.6133	e_1=-0.0288 e_0 =3.0416
$f(V_{GS})$	c_1=0.1313 c_0=0.7361	c_1=0.1349 c_0=0.9259	c_1=0.1613 c_0=4.3986	c_1=0.1485 c_0=39.1643	c_1=0.1549 c_0=193.6484	c_1=0.1485 c_0=386.759	c_0=1931.6

Tableau V .4: Les fonctions $d(V_{GS})$, $e(V_{GS})$ *et* $f(V_{GS})$ pour différents dopages N_A (DG MOSFET asymétrique)

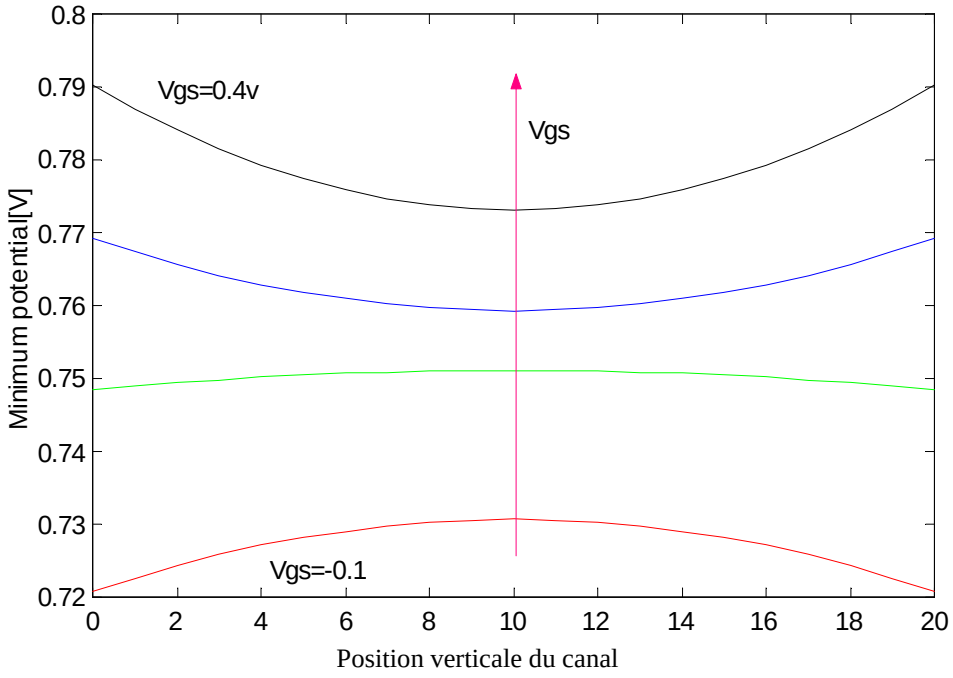


Figure V .5: Variation du potentiel minimum $\Psi_m(y)$ pour le transistor DG MOSFET symétrique

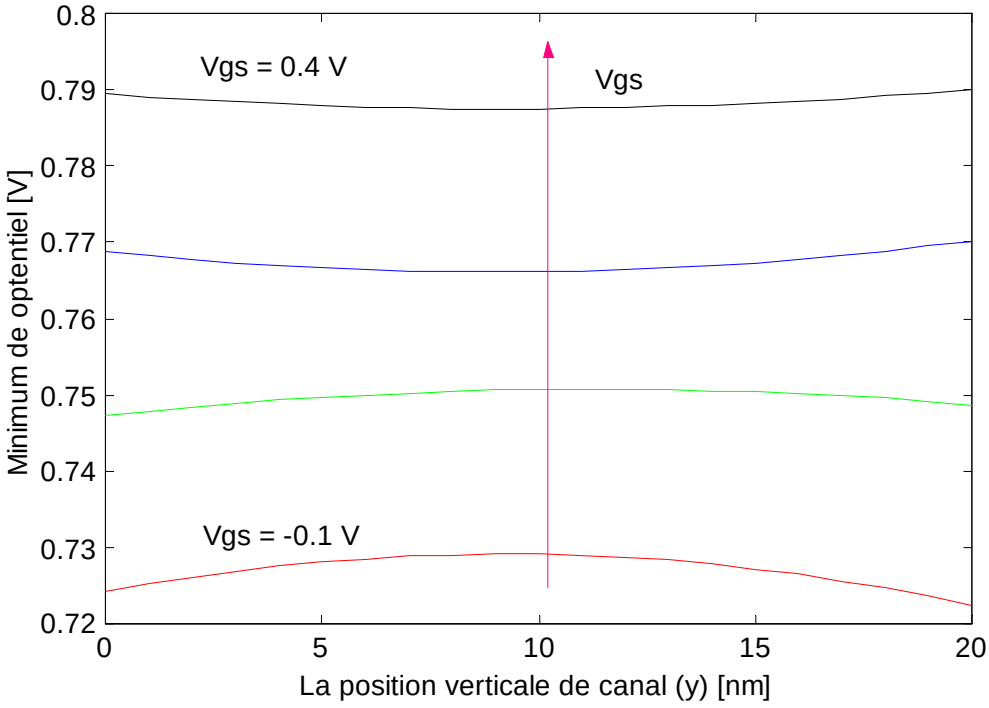


Figure V. 6: Variation du potentiel minimum $\Psi_m(y)$ pour le transistor DG MOSFET asymétrique

De (5.19) et (5.10), le modèle semianalytique de l'inverse de la pente sous seuil (S) peut être donné comme le suivant [96]:

$$S=\frac{KT}{q}\ln 10\left[\frac{\int_0^{t_{si}}\exp(\beta\,(d(V_{GS},N_A)y^2+e(V_{GS},N_A)y+f(V_{GS},N_A)))\,(d_1(N_A)y^2+e_1(N_A)y+f_1(N_A))dy}{\int_0^{t_{si}}\exp(\beta\,(d(V_{GS},N_A)y^2+e(V_{GS},N_A)y+f(V_{GS},N_A)))dy}\right]^{-1} \quad (5.20)$$

C'est cette dernière expression qui permet de déterminer l'évolution de la loi de l'inverse de la pente sous seuil (S) en fonction des différents paramètres (la longueur du canal L_i, le dopage N_A, l'épaisseur t_{si} du canal, la tension du contrôle V_{GS},…..). Contrairement à l'expression analytique de Qiang Chen et al, établie sur la base de nombreuses hypothèses simplificatrices [28], la notre a été déduite à partir d'un modèle numérique (Eléments Finis) plus élaboré, elle est censée d'être plus réaliste donc plus précise.

Le tableau ci-dessous illustre la variation de l'inverse de la pente sous seuil (S) en fonction de dopage pour les deux types du transistor DG MOSFET.

dopage cm^{-3}	$5 . 10^{14}$	10^{15}	10^{16}	10^{17}	$5 . 10^{17}$	10^{18}	$5 .10^{18}$
$S_{symétrique}$	100	102.1758	101.0321	99.4625	100.0977	100.2826	83.39
$S_{asymétrique}$	91.2872	91.0410	90.2612	88.8650	91.9107	89.6602	75.8190

Tableau V .5 : L'inverse de la pente sous seuil (S) en fonction de dopage (N_A) pour les deux types du transistor DG MOSFET .

Les figures V.7 et V.8 représentent les variations de l'inverse de la pente sous seuil S en fonction du dopage N_A pour les deux architectures symétrique et asymétrique.

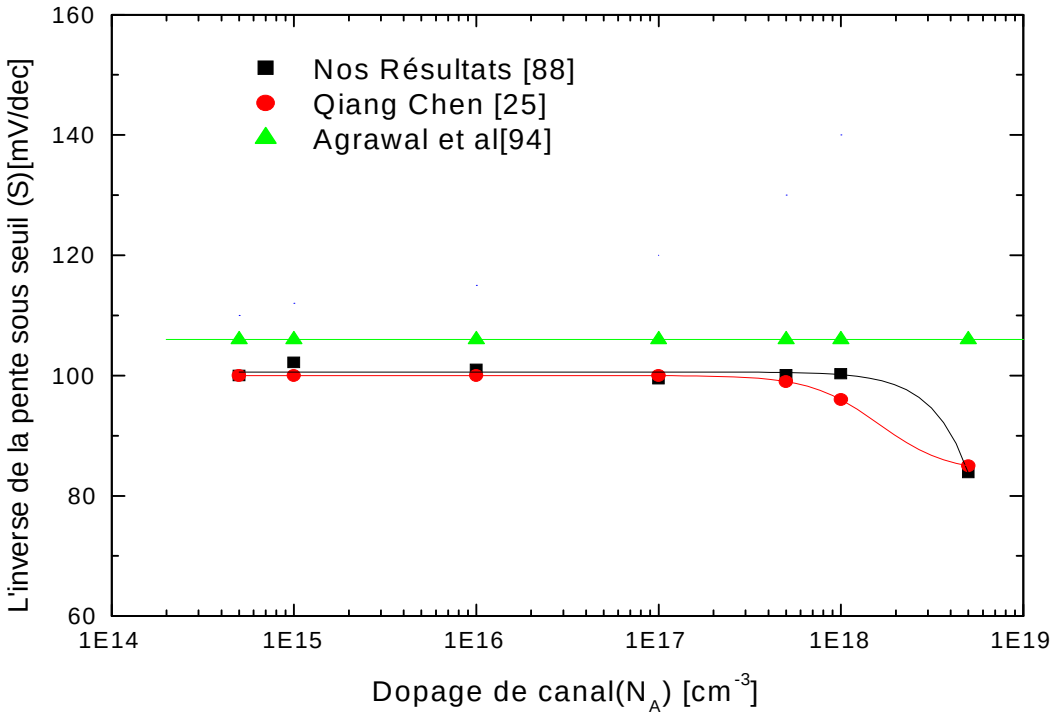


Figure V. 7: Variation de l'inverse de la pente sous seuil (S) en fonction du dopage N_A (DG MOSFET symétrique)

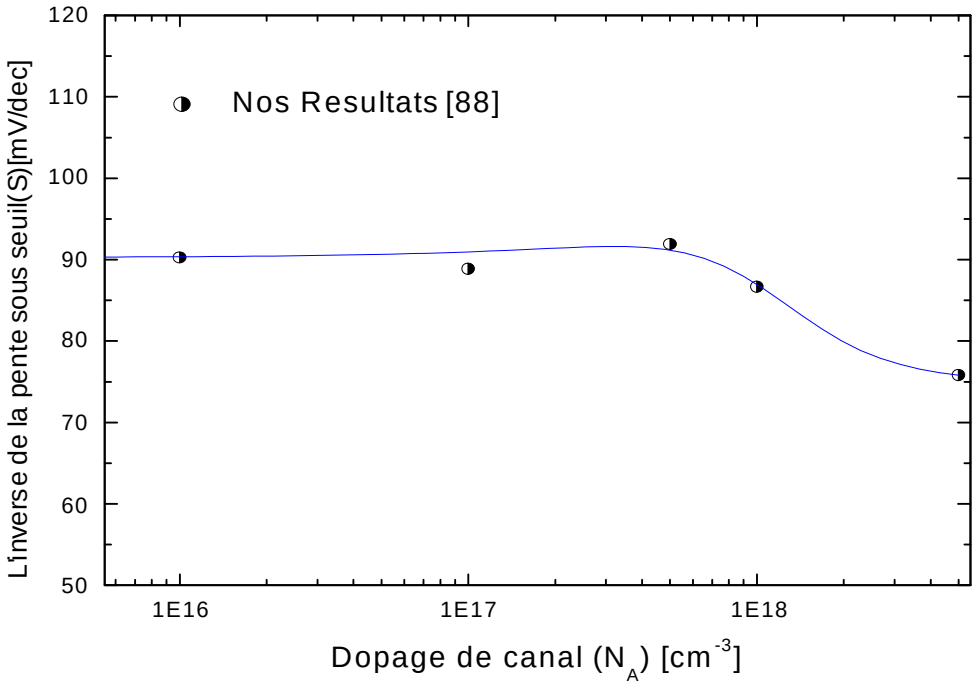


Figure V.8: Variation de la pente sous seuil (S) en fonction de dopage N_A (DG MOSFET asymétrique)

Pour un dopage faible du canal ($N_A = 10^{16}$ cm^{-3}), la variation de l'inverse de la pente sous seuil S en fonction de la longueur du canal (L) et de l'épaisseur (t_{Si}), est donnée par le tableau V.6 pour la configuration asymétrique

	L=30 nm	L=40 nm	L=50 nm	L=60 nm	L=110 nm
t_{SI}=10 nm	71.3626	65.0453	63.0893	61.2322	59.8713
t_{SI}=20 nm	100	77.2610	70.4023	68.3550	60.7863
T_{SI}=30 nm	Trés grande	103.7771	82.2923	69.5524	60.5661

Tableau V.6: Les valeurs de l'inverse de la pente sous seuil (S) en fonction de l'épaisseur (t_{SI}) et la longueur de canal (L) du DG MOSFET asymétrique

Le modèle semianalytique de l'inverse de la pente sous seuil (S) que nous avons développé pour un transistor DG MOSFET est comparé aux modèles analytiques de Qiang Chen [28] et Agrawal [97]. L'indépendance de (S) du dopage sur une large gamme (Figure.V.7), montre l'avantage de cette architecture par comparaison à celles obtenues avec la technologie classique (bulk technology) [48]. Pour des valeurs élevées de dopage (N_A =5.10^{18}), le potentiel de surface $\Psi_m(y=0 \; et \; y=t_{si})$ est beaucoup plus grand que le potentiel central $\Psi(y=t_{si}/2)$ et la conduction globale est fortement confinée dans les surfaces. Comme conséquence du confinement des linges de courant par rapport aux grilles, ces dernières assurent un contrôle efficace du canal ayant pour résultat un (S) amélioré. Avec la diminution du dopage N_A, la forme du profil de potentiel $\Psi_m(y)$ devient plus plate, dans ce cas, le contrôle du canal par les grilles devient plus faible et (S) plus grand. Finalement, pour des valeurs faibles de dopage ($N_A \leq 10^{16}$cm^{-3}), le profil de potentiel est pratiquement déterminé par la résolution de l'équation de Laplace $\Delta\Psi = 0$. En conclusion, le chemin de conduction effectif ne dépend plus de N_A; ce qui conduit à une valeur constante de (S).

Contrairement à l'expression analytique de Qiang Chen et al, établie sur la base de nombreuses hypothèses simplificatrices [28], la notre a été déduite à partir d'un modèle semianalytique plus élaboré, elle est censée d'être plus réaliste donc plus précise.

Le transistor DG MOSFET asymétrique montre un (S) plus amélioré par comparaison avec le transistor DG MOSFET symétrique puisque le chemin de conduction efficace dans le transistor asymétrique tend à se former de près d'une des surfaces Si/SiO2.

Comme conséquence, les grilles assurent un contrôle du canal mieux que l'architecture symétrique (S asymétrique < S symétrique) (figures V.8 et V.9).

La variation de (S) en fonction de la longueur de la grille joue un rôle particulier. Dans ce cas, on distingue deux domaines de variation de (S) en fonction de la longueur de la grille (figures V.10, V.11 et V.12):

a les canaux courts

Pour des longueurs du canal supérieures à une certaine valeur critique (aux environs de 100nm pour une épaisseur du canal tsi=30nm), l'inverse de la pente sous seuil affiche une très faible décroissance avec l'augmentation de la longueur du canal. Comparé aux résultats obtenus avec le paramètre du dopage, l'effet de ce paramètre est donc secondaire. En conclusion, le chemin de conduction effectif ne dépend plus de la longueur du canal; ce qui conduit à une valeur constante de (S).

b les canaux fortement courts

Pour des longueurs du canal inférieure à 100nm, les résultats de (S) montrent une évolution exponentielle avec une valeur minimale de (S) égale à 61 mV/dec (figure V.10). Cette augmentation de (S) peut être attribuée à un effet de l'apparition du courant tunnel source/drain qui affaiblit le contrôle du canal.

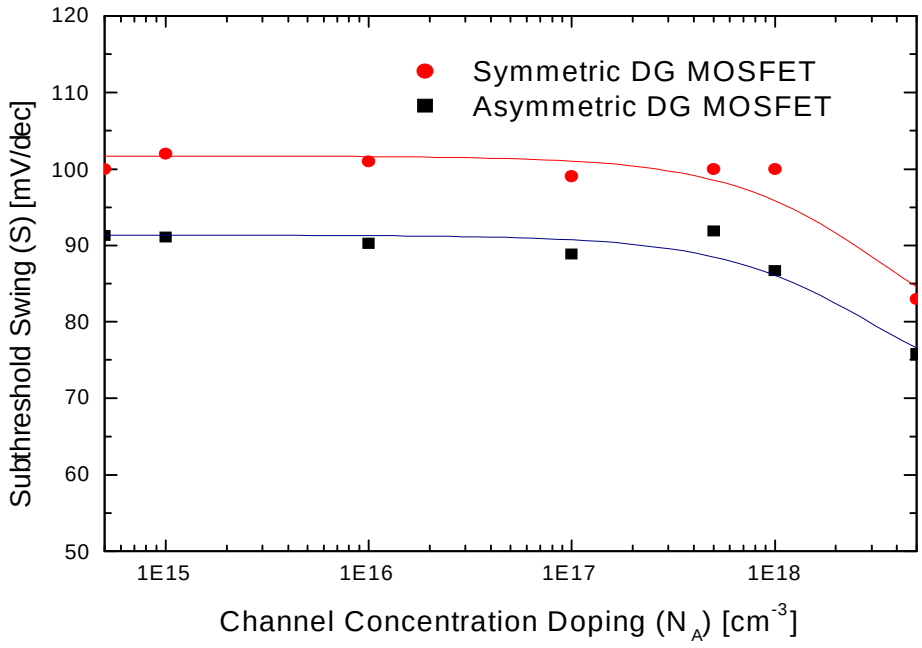


Figure V . 9 : La variation de L'inverse de la pente sous seuil (S) du transistor DG MOSFET symétrique et asymétrique

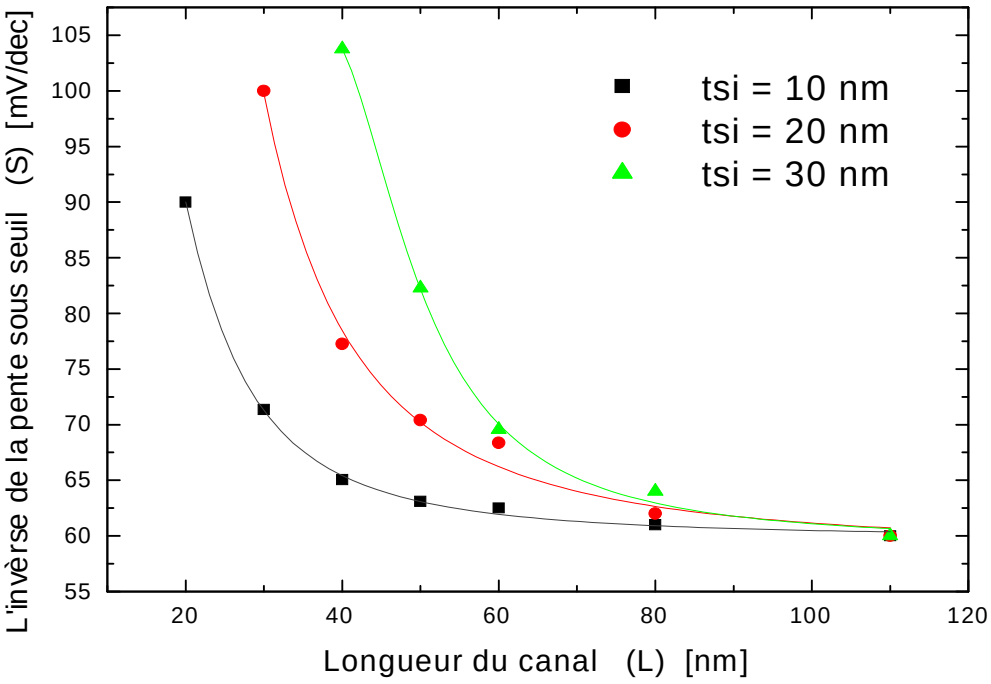


Figure V. 10: Variation de l'inverse de la pente sous seuil (S) en fonction de la Longueur de canal (L) pour le transistor DG MOSFET symétrique.

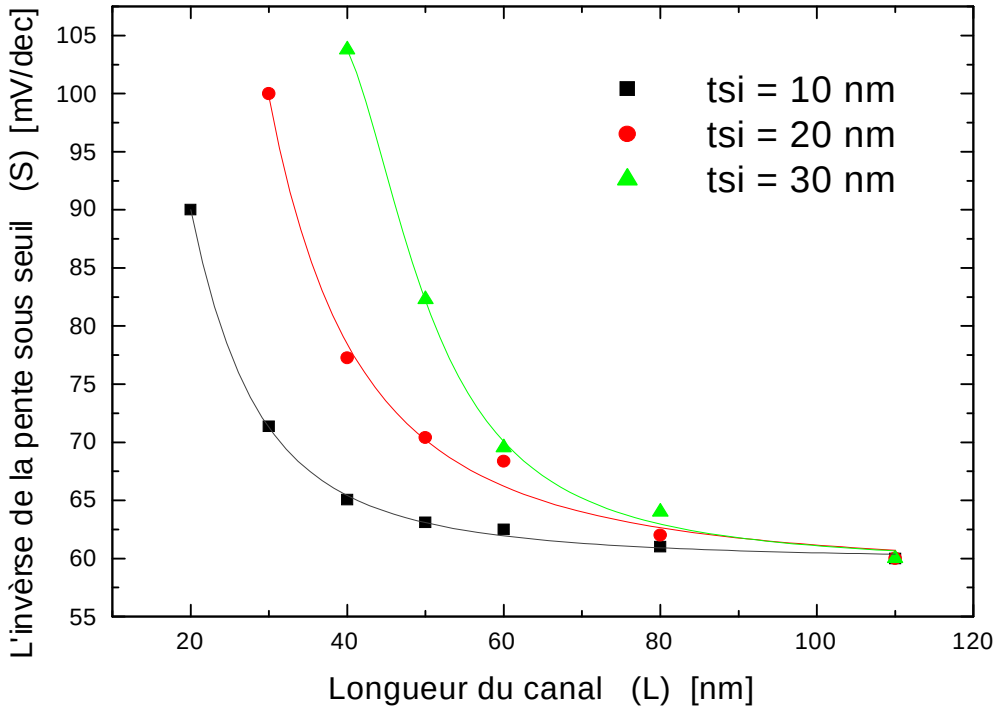


Figure V. 11: Variation de l'inverse de la pente sous seuil (S) en fonction de la Longueur de canal (L) pour le transistor DG MOSFET symétrique

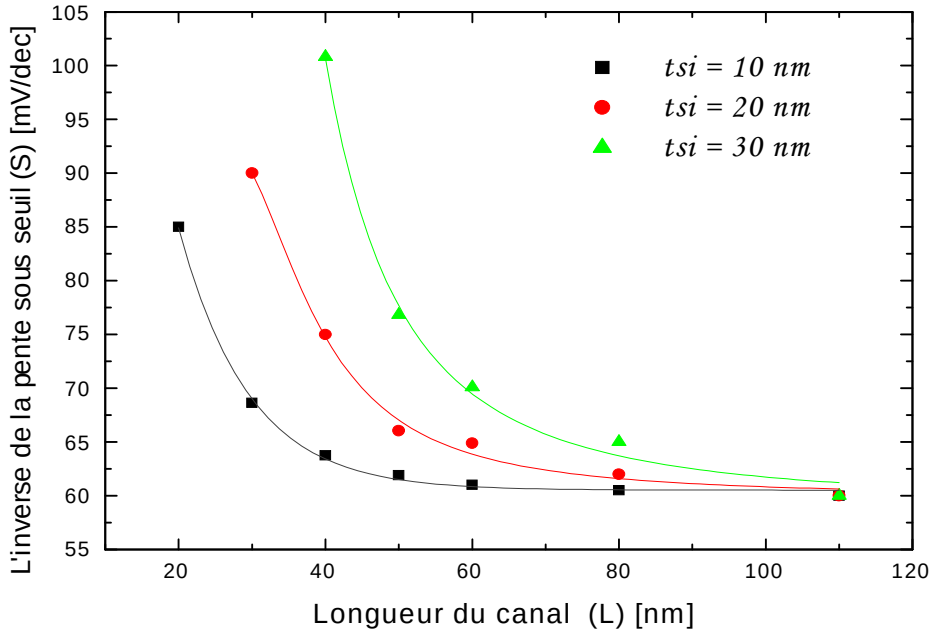


Figure V.12: Variation de l'inverse de la pente sous seuil (S) en fonction de la longueur du canal (L) (DG MOSFET asymétrique)

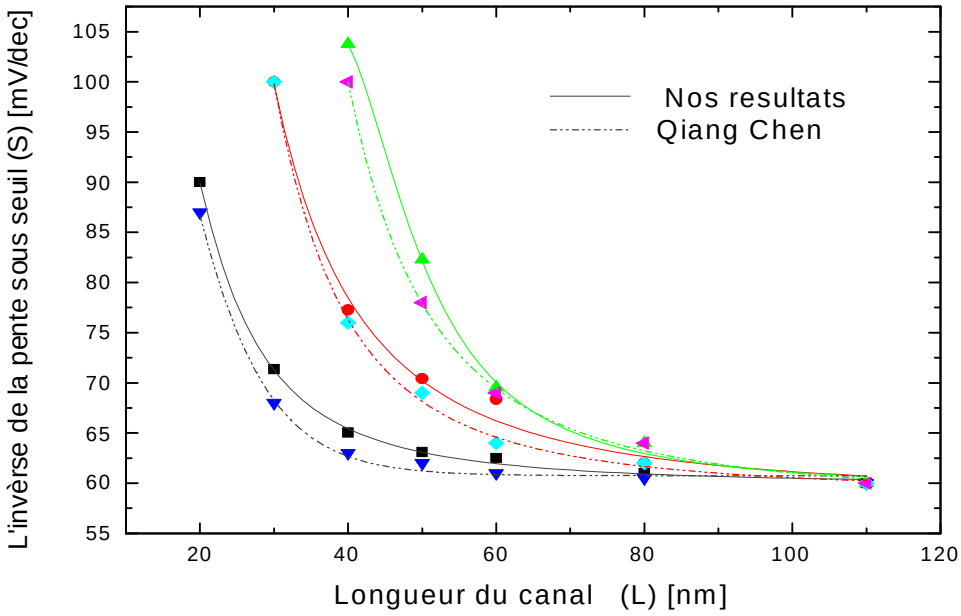


Figure V .13 : Comparaison entre nos résultats et les résultats de Qiang Chen (la variation de la pente sous seuil (S) en fonction de La longueur de canal (L))

V.2.3 Calcul neuronal

Le réseau de neurone développé est conçu afin de relier le vecteur d'entrée au vecteur de sortie. Dans cette étude, les variables d'entrées sont t_{ox1} (l'épaisseur de la couche isolante supérieure), t_{ox2} (l'épaisseur de la couche isolante inférieure), t_{si} (l'épaisseur du canal), L (la longueur du canal) et Sym (l'entrée de sélection de mode de fonctionnement, pour le mode symétrique Sym=1, pour le cas asymétrique Sym=0). La variable de la sortie est l'inverse de la pente sous seuil (S). Chacun de ces paramètres est indexé par un neurone et présenté dans la structure neuronale comme une valeur formatée donnée par l'expression (2.5).

La figure ci-dessous présente la structure optimisée (Paragraphe III.3.11) de notre prédicteur neuronal de l'inverse de la pente sous seuil (S). La base de données utilisée pour l'optimisation de notre structure neuronale est établie sur la base du modèle semi-analytique du transport des électrons dans le régime sous seuil (S) du transistor DG MOSFET développé (expression 5.20).

Le tableau V.7 résume les différentes caractéristiques de notre structure optimisée.

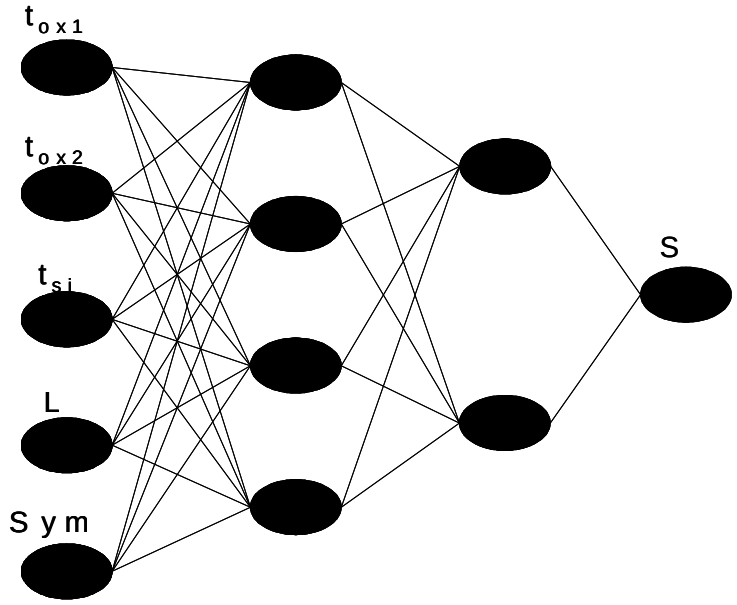


Figure V .14: Prédicteur neuronal de l'inverse de la pente sous seuil (S).

<table>
<tr><td colspan="2">Paramètres</td><td colspan="7">Valeurs optimisées</td></tr>
<tr><td colspan="2">Architecture</td><td colspan="7">Normal feed-forward MLP</td></tr>
<tr><td rowspan="8">Définition des couches</td><td>Couches cachées</td><td colspan="7">2</td></tr>
<tr><td>apprentissage</td><td colspan="7">Rétro propagation rapide (Quick propagation)</td></tr>
<tr><td rowspan="5">Nombre de neurones</td><td rowspan="3">Entrées (I=4)</td><td></td><td>Tox1(nm)</td><td>Tox2 (nm)</td><td>Tsi (nm)</td><td>L(nm)</td><td>Sym</td></tr>
<tr><td>Min</td><td>0.5</td><td>0.5</td><td>5</td><td>5</td><td>0</td></tr>
<tr><td>Max</td><td>3</td><td>3</td><td>50</td><td>100</td><td>1</td></tr>
<tr><td colspan="7">Première couche (N_1): variée</td></tr>
<tr><td colspan="7">Deuxième couche (N_2): variée</td></tr>
<tr><td rowspan="3">Sorties (O=1)</td><td></td><td colspan="5">S(mV/dec)</td></tr>
<tr><td>Min</td><td colspan="5">60</td></tr>
<tr><td>Max</td><td colspan="5">-</td></tr>
<tr><td>Population des poids</td><td colspan="7">$I*N_1+N_1*N_2+N_2*O$</td></tr>
<tr><td>Fonction d’entrée</td><td colspan="7">Produit scalaire entre les sorties de neurone
$w_{ij}o_j$</td></tr>
<tr><td>Traitement à l’entrée</td><td colspan="7">Variables entre 0 et 1
$\frac{x-x_{\min}}{x_{\max}-x_{\min}}$</td></tr>
<tr><td>Fonction d’activation</td><td colspan="7">Sigmoïdale
$\frac{1}{1+\exp(-x)}$</td></tr>
<tr><td colspan="2">Type d’erreur du réseau (Mean square error)</td><td colspan="7">Training process
$ETrn=\frac{1}{N_{Ptr}}(z_i-O(I_i,W))^2 \quad i=1,N_{Ptr}$
Test process
$ETst=\frac{1}{N_{Ptst}}(z_i-O(I_i,W))^2 \quad i=1,N_{Ptst}$</td></tr>
<tr><td rowspan="4"></td><td>Itération maximale</td><td colspan="7">5000</td></tr>
<tr><td>Tolérance</td><td colspan="7">0.001 (not reached)</td></tr>
<tr><td>séquence</td><td colspan="7">apprentissage + test passe après la mise à jour des poids</td></tr>
<tr><td>Taille de la base de données</td><td colspan="7">1000 échantillons pour l’apprentissage
250 échantillons pour le test</td></tr>
</table>

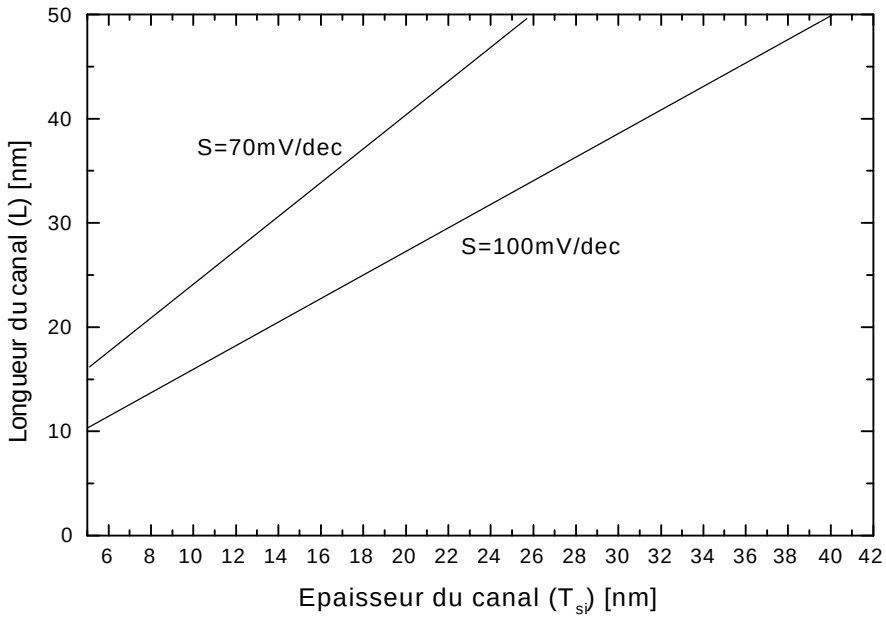
Tableau V.7: paramètres de notre prédicteur neuronal après la procédure d’optimisation

V.2.4 Abaque de la réduction dimensionnelle de transistor DG MOSFET

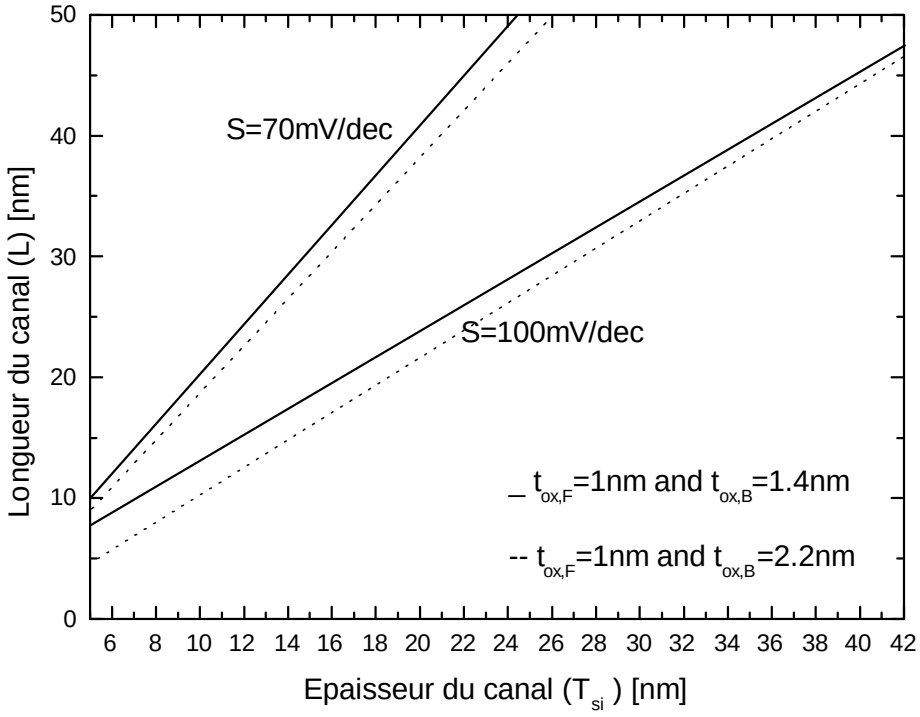
La base de données de l'inverse de la pente sous seuil (S) générée par notre prédicteur neuronal nous permet de construire un abaque graphique. Cet abaque permet de fixer les différents paramètres du transistor DG MOSFET afin d'optimiser la conception de ce dernier. La figure V.15a illustre notre abaque de la réduction dimensionnelle du transistor DG MOSFET symétrique faiblement dopé, où le minimum de la longueur du canal (L) en fonction de l'épaisseur du canal (t_{si}) est projeté pour S=100mV/dec et S= 70mV/dec avec une épaisseur d'oxyde t_{ox} =0.8nm, dans ce cas, une conception optimale du transistor DG MOSFET symétrique est obtenue pour une longueur minimale du canal (L=10nm) comme le suivant:

t_{si} = 5nm, t_{ox} =0.8nm, L=10nm, $N_A < 10^{17}$ cm^{-3} avec S=100mV/dec.

La figure V.15b illustre l'évolution de la loi de la réduction dimensionnelle dans le cas d'un transistor DG MOSFET asymétrique. Cette évolution montre l'effet de l'épaisseur de la couche isolante inférieure ($t_{ox,B}$) sur la loi de la réduction dimensionnelle du transistor DG MOSFET asymétrique faiblement dopé. Ce dernier nous permet d'avoir une loi d'évolution de la réduction dimensionnelle plus améliorée par comparaison avec le transistor DG MOSFET symétrique faiblement dopé, il est facile de remarquer qu'une technologie de 10nm (L=10nm) peut être atteinte avec S=70mV/dec (Figure V.15b).



(a)



(b)

Figure V .15: Abaque de la réduction dimensionnelle du (a) transistor DG MOSFET symétrique (b) transistor DG MOSFET asymétrique [91].

V.3 Conclusion

Dans ce chapitre, nous avons présenté un abaque graphique basé sur les réseaux de neurones artificiels pour l'étude et l'optimisation de la conception des transistors nanométriques symétriques et asymétriques. Cette étude est basée sur les effets induits par la longueur du canal, l'épaisseur du canal, le dopage du canal et l'épaisseur de la couche isolante sur l'inverse de pente sous seuil. Une approche semianalytique de l'inverse de pente sous seuil basée sur la résolution, dans la région du canal, du système d'équations bidimensionnel non linéaire (Poisson-Boltzmann) a été développée en utilisant la méthode des Eléments Finis et l'interpolation polynomiale. La base de données générée par notre approche semianalytique est utilisée pour l'apprentissage et l'optimisation de notre structure neuronale afin de développer nos abaques de la réduction dimensionnelle des transistors nanométriques.

CONCLUSION GENERALE

La miniaturisation du transistor MOSFET nécessite une modélisation qui permet de prendre en compte les effets quantiques, le transport non stationnaire et balistique dans le but d'évaluer de nouvelles architectures prometteuses. Dans le cadre de projet de cette thèse, nous avons présenté une contribution à l'étude et la modélisation du transistor MOSFET fortement submicronique. On s'intéresse principalement à l'élaboration de nouvelles approches basées sur les techniques de l'intelligence artificielle (les Réseaux de Neurones Artificiels) permettant:

- l'étude de l'impact de la miniaturisation sur les caractéristiques électriques des dispositifs MOSFETs émergents et plus particulièrement sur les architectures multi-grilles (DG MOSFET);
- l'estimation de la durée de vie des transistors MOSFETs fortement submicroniques et l'étude de l'effet du temps du stress sur le fonctionnement des dispositifs intégrés CMOS;
- le développement d'un nouveau modèle neuronal du DG MOSFET qui permet la modélisation et qui décret adéquatement le comportement des circuits intégrés nanométriques où les effets quantiques deviennent plus importants.
- l'étude de variation de l'inverse de la pente sous seuil en fonction de la longueur du canal, du dopage et de l'épaisseur du canal de la structure DG MOSFET;
- le développement des nouveaux abaques qui permettent de décrire la loi de réduction dimensionnelle de la structure DG MOSFET en fonction des différents paramètres (longueur du canal, du dopage et de l'épaisseur du canal). L'utilisation de ces abaques permet de prédire les limites technologiques (les contraintes imposées par les paramètres géométriques et physiques de la structure DG MOSFET) de fabrication des circuits intégrés à l'échelle nanométrique.

Le premier chapitre de ce manuscrit constitue un travail de synthèse des principaux effets parasites engendrés par la miniaturisation des transistors MOS et plus particulièrement la diminution de la longueur de canal. Cette réduction des dimensions a engendré des phénomènes parasites (DIBL, modification de la tension de seuil, augmentation du phénomène de porteurs chauds,…) qui détériorent les caractéristiques courant-tension. Toutefois, les technologues ont imaginé des procédés de fabrication particuliers en vue de conserver les caractéristiques électriques (technologie SOI, LDD MOSFET, DG MOSFET). Un accent particulier a été mis sur le cas prometteur du transistor à Double-Grille. Nous espérons qu'une telle synthèse, bien qu'elle soit connue pour certains, elle servira comme point de départ à d'autres travaux dans le domaine de la physique des composants.

L'étude et la description des Réseaux de Neurones Artificiels font l'objet du second chapitre. Les Réseaux de Neurones Artificiels font parties des approches implicites qui considèrent comme relations statistiques les corrélations entre les paramètres opératoires (les paramètres géométriques et électriques du transistor MOSFET) et les propriétés de sortie (la transconductance, les courants de sortie, l'inverse de la pente sous seuil,….). Les phénomènes qui contrôlent ces corrélations sont encodés dans un ensemble de neurones (entité de décision) qui permet de représenter les effets sans les expliquer. Un recours à une base de données représentant ces corrélations est impératif pour l'apprentissage de ces structures.

Le troisième chapitre a été consacré au développement d'une approche analytique permettant l'estimation de la durée de vie des transistors MOSFETs fortement submicroniques et l'étude de l'effet du temps du stress sur le fonctionnement des dispositifs intégrés. Ce chapitre a été divisé en deux parties; dans la première partie, nous avons proposé un dispositif expérimental assisté par ordinateur permettant d'étudier les aspects expérimentaux des phénomènes du vieillissement des transistors MOSFETs fortement submicroniques, les résultats expérimentaux générés par ce dispositif ont été utilisés comme une base de données d'optimisation de notre structure neuronale, la deuxième partie est consacrée au développement d'une approche analytique à base des Réseaux de Neurones Artificiels qui permet de prédire les variations de la dégradation des transistors MOSFETs fortement submicroniques. L'avantage d'une telle approche est d'être très facilement implémentable dans n'importe quel type de modèle compact.

Dans le quatrième chapitre, nous avons appliqué le formalisme des fonctions de Green hors-équilibre au MOSFET Double-Grille ultime dans lequel l'effet des réservoirs (source et drain) sur le composant intrinsèque (canal) peut être représenté par la matrice de la self-énergie. Ce modèle montre qu'un confinement vertical permet de conserver les propriétés électriques des MOSFETs conventionnels. Le modèle numérique des caractéristiques courant-tension (I-V) du transistor DG MOSFET développé en utilisant le formalisme des fonctions de Green hors-équilibre nous a permis de former une base de données afin d'optimiser notre structure neuronale. L'algorithme d'apprentissage retenu (quick propagation) a permis d'améliorer la convergence des résultas et de limiter le nombre de cycles d'apprentissage. Après l'optimisation, l'ensemble des poids générés ont été implémentés dans le logiciel de simulation (PSPICE) afin d'obtenir notre modèle neuronal du transistor DG MOSFET. Les résultats obtenus (la précision et le temps de calcul est moins élevé par comparaison avec les modèles numériques) sont encourageants à plus qu'un titre et ont permis d'adopter une nouvelle approche, la notre, pour l'étude et la simulation des circuits électroniques nanométriques plus complexes (nanoprocesseur, ….).

Le développement d'un abaque graphique basé sur les Réseaux de Neurones Artificiels pour l'étude et l'optimisation de la conception des transistors nanométriques symétriques et asymétriques fait l'objet de dernier chapitre. Cette étude est basée sur l'effet induit par la longueur du canal, l'épaisseur du canal, le dopage du canal et l'épaisseur de la couche isolante sur l'inverse de pente sous seuil. Une approche semianalytique de l'inverse de pente sous seuil basée sur la résolution d'un système d'équations bidimensionnel non linéaire (Poisson-Boltzmann) dans la région du canal a été développée en utilisant la méthode des Eléments Finis et l'interpolation polynomiale. La base de données générée par notre approche semianalytique a été utilisée pour l'apprentissage et l'optimisation de notre structure neuronale et a permis le développement de l'abaque de la réduction dimensionnelle des transistors nanométriques. L'utilisation de cet abaque a montré que l'évolution de la loi de la réduction dimensionnelle pour la configuration asymétrique est plus améliorée par comparaison avec la configuration symétrique.

En perspective à court et à moyen termes, certaines améliorations et certaines innovatrices idées peuvent contribuer à l'étude et à une meilleure connaissance des phénomènes abordés dans le cadre de ce travail:

- au niveau expérimental, il serait souhaitable d'intégrer des modules supplémentaires pour faire l'extension des mesures à des températures différentes. Cependant, cette étude nécessite d'acquérir un appareillage de mesure complémentaire (cryostat, thermostat,…);
- sur le plan théorique, les modèles développés dans ce manuscrit peuvent être améliorés en intégrant l'influence des différents défauts existent et qui sont de nature différente sur les paramètres électriques de structures MOSFETs émergentes. A l'échelle atomique, les propriétés électroniques des semiconducteurs sont particulièrement sensibles à la présence de ces défauts . Il est bien établi que les défauts perturbent la périodicité du réseau cristallin et introduisent des états localisés dans le gap du matériau. Donc, la modélisation de ces effets nécessite une description quantique basée sur le formalisme des fonctions de Green. Cependant, cette étude nécessite l'utilisation d'une machine plus puissante (macro-ordinateur) afin de pouvoir augmenter le nombre de nœuds de la grille (augmenter la taille de la matrice du système globale) pour bien étudier l'influence de défauts ponctuels sur les paramètres électriques de structures MOSFETs émergentes (à l'échelle atomique).

Bibliographie

[1] D. J. Frank, R. H. Dennard, E. Nowak, Device scaling limits of Si MOSFETs and their application dependencies, *Proc. IEEE*, vol. 89, no. 3, pp. 259–288, 2001.

[2] Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S.-H. Lo, G. Sai-Halasz, R. Viswanathan, H.-J. C. Wann, S. Wind and H.-S. Wong, CMOS scaling into the nanometer regime, Proc. IEEE, 85, pp. 468-504, 1997.

[3] H.-S. Wong, D. Frank and P. Solomon, Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SOI MOSFET's at the 25 nm Channel Length Generation, IEDM Tech. Digest, pp. 407-410, 1998.

[4] M.Bescond, Modélisation et simulation de transport quantique dans les transistors MOS nanométriques, Thèse de doctorat, université de Provence (Aix-MarseilleI), France, 2004.

[5] P.K.Bondyopadhyay, Moore's Law Governs the Silicon Revolution, *Proc.IEEE*, vol. 86, no. 1, pp. 78-81, 1998.

[6] F. Stern and Howard, Properties of semiconductor surface inversion layers in the electric quantum limit, Phys. Rev, vol. 163, pp. 816-835, 1967.

[7] F. Stern, Quntum properties of surface space-charge layer, CRC Crti Rev. Solid State Sci., vol.4, pp.499-514, 1974.

[8] R.Rios, N. Arora, et al., Aphysical compact MOSFET model, including quantum mechanical effects, for statistical circuit design applications, in Proc. IEEE Int. Electron Devices Meeting (IEDM), pp.937-940.

[9] T.Janik and B.Majkusiak, Analysis of MOSFET based on the self-consistent solution to the Schrodinger and poisson equations and the local mobility model, IEEE trans Electron Devices, vol. 45, no.6, pp.1263-1271, Juin 1998.

[10] G. Gildenblat, T.L. chen, and P.Bendix, Closed-form approximation for the perturbation of MOSFET surface potentiel by quantum mechanical, Electronics Letters, vol.36, no.12, pp.1072-1073, Juin 2000.

[11] Y.Ma, L.Liu, et al, A new charge model including quantum mechanical effects in MOS structure inversion layer, Solid Stat Electron, vol. 44, pp.1697-1702, 2001.

[12] R. Clerc, P.O'Sullivan, et al, A physical compact model for direct tunneling from NMOS inversion layers, Solid Stat Electron, vol. 45, pp.1705-1716, 2001.

[13] M. J. van Dort, P.H. Woerlee, et al., Influence of high subtrate doping levels on the threshold voltage and mobility of deep-submicrometer MOSFET's, IEEE trans Electron Devices, vol. 39, no.4, pp.932-938, Apr. 1992.

[14] M. J. van Dort, P.H. Woerlee, et al, A simple model for quantisation effects in heavily-doped silicon MOSFETs at inversion conditions, Solid Stat Electron, vol. 37, pp.411-414, 1994.

[15] J. Suné, P.Olivo, and B.Ricco, Quantum mechanical modeling of accumulation layers in MOS structure, IEEE trans Electron Devices, vol. 39, no.7, pp. 1732-1738, July. 1992.

[16] A**.** Neve , D. Flandre and J-J.Quisquater, Feasibility of Smart Cards in Silicon-On-Insulator (SOI) Technology, Proceeding of USENIX Workshop on Smartcard Technology, 1-7, 1999.

[17] M. Stockinger, Optimization of Ultra-Low-Power CMOS Transistors, Institut für ikroelektronik, PHD theses, 2000.

[18] A.Strass, Nano-MOSFETs for future ULSI applications, Solid State Technology, 65-74, 1996.

[19] A.N. Broers, Fabrication limits of electron beam lithography and UV, X-ray and ion beam lithographies, Phil. Trans. R.Soc.Lond. A, 291-311, 1995.

[20] M. Henini, Molecular Beam Epitaxy From Research to Mass-Production-Part1, III-V Review, Vol 9

[21] K. Kwok and T. William, Analysis of Gate-Voltage-Dependent Series Resistance of MOSDET's, IEEE Transactions on Electron Devices, vol. 7, n°7, pp. 965-972, 1986.

[22] L.D. Yau, A Simple Theory to Predict the threshold Voltage of Short-Channel IGFET's, IEEE J. of Solid State Electron., vol. 9, n°3, pp. 256-263, 1974.

[23] O. Seiki , P.J.Tsang et al., Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor, IEEE Transactions on Electron Devices, vol. 27, n°8, pp. 1359-1367, 1980.

[24] S. Wind, D. Frank, and H. Wong, Scaling silicon MOS device to their limits, Microelectronics Engg. 32, pp.271,1996.

[25] R.H. Yan, A. Ourmazd, and K.F. Lee, Scaling the Si MOSFET : From bulk to SOI to bulk, IEEE Trans. Electron Dev. 39,pp. 1704 ,1992.

[26] K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, Scaling theory for doublegate SOI MOSFETs, IEEE Trans. Electron Dev. 40,pp. 2326,1993.

[27] T. Tanaka, K. Suzuki, H. Horie, and T. Sugii, Ultrafast operation of VTH-adjusted p+-n+ double-gate SOI MOSFET, IEEE Electron Dev. Lett. 15,pp. 386, 1994.

[28] Qiang Chen, Bhavna Agrawal, and James D. Meindl, A Comprehensive Analytical Subthreshold Swing (S) Model for Double-Gate MOSFETs, IEEE Transaction Electron Devices, vol. 49, no. 6, June 2002.

[29] Z. Ren, R. Venugopal, S. Datta, M. Lundstrom, D. Jovanovic, and J. Fossum, The ballistic nanotransistor: A simulation study, IEDM Tech. Dig., pp. 715–718, 2000.

[30] M. Negnevitsky**,** Artificial Intelligence, Addison Wesley, 2 Edition , 2004.

[31] W.C. Mc Culloh, W.H.Pitts, A logical calculus ideas imanent in nervous activity, Bulletin of Mathematical Biophysics, 5, p.115, 1943.

[32] M.M.Nelson, W.T. Illingwotth, A practical guide to neural nets, Addision-Wesley Pub.,3rd edition, New York, NY, USA, 1991.

[33] P.B.L. Meijer, Neural Network Applications in Device and Subcircuit Modelling for Circuit Simulation, Proefschrift Technische Universiteit Eindhoven, Nederlands, 2003.

[34] Solomon Snyder, Les drogues et le cerveau, L'univers des Sciences, Pour la Science, pp 11-16,1987.

[35] Mohamed Ryad Zemouri, Contribution à la surveillance des systèmes de production à l'aide des réseaux de neurones dynamiques: Application à la e-maintenance, Thèse de Doctorat, Université de Franch Comte, France, N°d'ordre :986, 2003.

[36] Patterson D, Artificial Neural Networks, Prentice Hall, Singapore, 1996.

[37] J.J.Hopfield, Neural networks and physical systems with emergent collective computational abilities, In *Proc.* Nat'l Academy of Sciences, pages 2554-2558, 1982.

[38] J. A. Anderson and E. Rosenfeld, Neuro Computing Foundations of Research, MIT PRESS, Cambridge, 1988.

[39] S.Guessasma, Optimisation et contrôle de procédés de projection thermique par coopération de méthodes d'intelligence artificielle, Thèse de Doctorat, Université de Belfort, France, N°d'ordre :006, 2003.

[40] John Hertz, Anders Krogh, and Richard G. Palmer. Introduction to the theory of neural computation. Addison-Wesley, 1991.

[41] G. Brightwell, C.Kenyon, H. Paugam-Moisy, Multilayer neural network: One or tow hidden layers, NeuroCILOT Technical Report Series, NC-TR-97-001, ESPRIT Working Group in Neural and Computational learning, Department of Computer Science, University of London, 1997.

[42] S.E. Fahlman, Fast-learning variations on back-propagation: an ampirical study, Proc.of the 1988 Connectionist Models Summer School, San Mateo, CA, USA, p.38, 1988.

[43] C. Hu, F.-C. Tam, P.-K. Ko, T.-Y. Chan and K. W. Terrill, Hot - electron - induced MOSFET degradation - model, monitor, improvement, *IEEE Trans. Electron Devices,* vol. 32, pp. 375 - 385, 1985.

[44] N. Guenifi, F.Djahli, M.Hemissi., étude du vieillissement des TMOS ultra-courts, Sciences et technologie vol17, juin 2002.

[45] M.Boulmden, L'électronique analogique, Presse de L'université de Batna ,1997.

[46] BF960 : N-Channel dual gate MOSFET [TeleFunken Electronic], http://www.alldatasheet.co.kr/datasheet-pdf/pdf_kor/TFUNK/BF960.html.

[47] F.Djeffal, A. Benhaya, M.Chahdi, Un dispositif expérimental assisté par ordinateur permettant l'étude de la dégradation des transistors MOSFETs submicroniques, Congrès International sur la Génie Electrique, Sétif, Algérie, pp. 346-350, 10-12 Octobre 2004,.

[48] S.M.Sze, Physics of semiconductor devices, 2nd Edn, Wiley, New York, 1981.

[49] P.Masson, et al., On the tunneling component of charge pumping current in ultrathin gate oxide MOSFETs, IEEE Elect.Dev.LettVvol.20,No.2, pp. 92-94, 1999.

[50] G.Groeseneken, et al , a reliable approach to charge-Pumping measurements in MOS transitor, IEEE Elect.Dev., vol.ED-31, pp. 42-53, 1984.

[51] Farzin A, Zhibin R, Dragica V, Supriyo D, Mark SL. On the performance limits for Si MOSFETs: A theoretical study. IEEE Trans. Electron Dev, 47(1), pp. 232-240, 2000.

[52] N. D. Arora, R. Rios, and D. A. Antoniadis, Capacitance modeling for deep submicron thin gate oxide MOSFETs, European Solid State Devices Research Conference (ESSDERC), pp. 569–572, 1995.

[53] F.Stern, Self consistent results for n-type Si inversion layers, Physical review, 5, p.4891-4899, 1972.

[54] F.Djeffal, S.Guessasma, A.Benhaya and M.Chahdi, An analytical approach based on neural computation to estimate the lifetime of deep submicron MOSFETs, Semicond. Sci. Technol, (20), p. 158-164, 2005.

[55] F.Djeffal, A.Benhaya and M.Chahdi, Study of the scaling capability of DGMOSFET using the neural networks, article accépté pour publication au journal Material Sci and Eng, 2005 .

[56] L.Valiant, Atheory of the learnable, Communications of the ACM(CACM), 27(12), p.1134, 1984.

[57] G. Brightwell, C. Konyo, Multilayer neural network: one or two hidden layers, NeurCOLT Technical Repport Series, NC-TR-97-001, ESPRIT Working Group in Neural and Computational learning, Department of Computer Science, University of London, 1997.

[58] S.Guessama, G. Montavon, C.Coddet, Correlation between processing parameters, coatings properties and in flight particle characteristics, Thermal Spray 2003, ASM International Materials Park, USA, p.1139, 2003.

[59] L. Holmstrom, P. Koistinen, Using additive noise in back-propagation training, IEEE Transaction on Neural Networks, 3, p. 24, 1992.

[60] K-C Jim, C.L. Giles, B.G. Horne, An analysis of noise in recurrent neural networks: convergence and generalization, IEEE Transaction on Neural Networks, 7, p.1424, 1996.

[61] M.M. Nelson, W.T.Illingworth, Apractical guide to neural nets, Addisson-Wesley Pub., 3rd edition, New York, NY, USA, 1991.

[62] M.Stone, Comments on neural selection creteria of Akaike and Schwarz, Journal of the Royal Statistical Society, Series B, 41, p.276, 1979.

[63] C.Goutte, Note on free lunches and cross-validation, Neural Computation, 9, p. 1211, 1997.

[64] F.Djeffal, A.Benhaya and M.Chahdi, Un prédicteur à base des réseaux de neurones permettant l'estimation de la durée de vie des transistors Mosfets fortement submicroniques, Communication envoyée pour le Congrès International en Informatique Appliquée à Bordj Bou Arréridj (Algérie) du 22 au 24 novembre 2005.

[65] S.Guessama, G. Montavon, C.Coddet, Plasma spray process modeling using artificial neural networks: Application to Al_2O_3-TiO_2 (13% by weight) ceramic coating structure, 2nd International Conference on thermal Process Modelling and Computer Simulation, Nancy, France; Mars 31-Avril 1-2,2003.

[66] B. Razavi, Design of analog CMOS integrated circuits, Mc Graw Hill, Boston, 2001.

[67] Y. Taur, X. Liang, W. Wang, H Lu, A Continuous, Analytic Drain-Current Model for DG MOSFETs. IEEE Electron Device Lett, 25(2), p.107-109, 2004.

[68] F. Djeffal, M. Chahdi, A. Benhaya, M. Djouimaa, Modelling and simulation of nanoscale CMOS circuits using artificial neural network, article soumis à la conférence internationale des systèmes électroniques CISE 2005, Batna, Algérie, du 13 au 14 Décembre 2005.

[69] S. Datta, Nanoscale Device Modelling: the Green's Function Method, Superlattices and Microstructures, 28(4), p. 253-278, 2004.

[70] G. Baccarani, and S. Reggiani, A compact double-gate MOSFET model comprising quantum-mechanical and nonstatic effects, IEEE Trans. Electron Dev. 46, p. 232, 1999.

[71] D. Chang, and J.G. Fossum, Simplified energy-balance model for pragmatic multidimensional device simulation, Solid-State Electron. 41, p.1795, 1997.

[72] K. Banoo, and M.S. Lundstrom, Electron transport in a model silicon transistor, Solid-State Electron. 44,p.1689, 2000.

[73] K. Natori, Ballistic metal-oxide-semiconductor field-e.ect-transistor, J. Appl. Phys. 76, p.4879, 1994.

[74] M. Lundstrom, Elementary scattering theory of the Si MOSFET, IEEE Electron Dev. Lett. 18,p. 361, 1997.

[75] M.S. Lundstrom, and Z. Ren, Essential physics of carrier transport in nanoscale MOSFETs, IEEE Trans. Electron Dev. 49,p. 133 , 2002.

[76] A. Pirovano, A.L. Lacaita, and A.S. Spinnelli, Two-dimensional quantum e.ects in nanoscale MOSFETs, IEEE Trans. Electron Dev. 49,p. 25, 2002.

[77] F.G. Picus, and K.K. Likharev, Nanoscale field-effect transistors: an ultimate size analysis, Appl. Phys. Lett.p.71, 3661, 1997.

[78] A. Svizhenko, M.P. Anantram, T.R. Govindan, B. Biegel, and R. Venugopal, Nanotransistor modeling : two-dimensional Green's function method, J. Appl. Phys. 91, p.2343, 2002.

[79] A. Svizhenko, and M.P. Anantram, Role of scattering in nanotransistors, IEEE Trans. Electron dev. 50,p.1459, 2003.

[80] R. Venugopal, Z. Ren, S. Datta, M.S. Lundstrom and M.S. Jovanovic, Simulating quantum transport in nanoscale transistors : real versus mode-space approaches, J. Appl. Phys. 97,p.3730, 2002.

[81] R. Venugopal, M. Paulsson, S. Goasguen, S. Datta, and M. Lundstrom, A simple quantum mechanical treatment of scattering in nanoscale transistors, J. Appl. Phys. 93, p. 5613, 2003.

[82] R. Venugopal, S. Goasguen, S. Datta, and M.S. Lundstrom, Quantum mechanical analysis of channel access geometry and series resistance in nanoscale transistors, J. Appl. Phys. 95,p.292, 2004.

[83] P.Damle, Nanoscale Device Modelling: From MOSFETs to Molecules, Ph.D thesis, Purdue University, West Lafayette, IN, 2003.

[84] Z. Ren, Nanoscale MOSFETS: physics, simulation and design, PhD thesis, Purdue University, West Lafayette, IN, 2001.

[85] S. Datta, Nanoscale Device Modelling: the Green's Function Method, Superlattices and Microstructures, 28(4), p. 253-278, 2000.

[86] P. Van Halen and D.L. Pulfrey, Accurate short series approximations to Fermi-Dirac integrals of order -1/2, 1/2, 1, 3/2, 2, 5/2, 3, and 7/2, J. Appl. Phys, 59(6), p.5271-5274, 1985.

[87] Y. Yaur, T. H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, Cambridge, UK,1998

[88] F. Djeffal, M. Chahdi, A. Benhaya and L. Hafiane, An approach based on neural computation to design the nanoscale CMOS circuits, article soumis pour publication à Integration, the VLSI Journal, juillet 2005.

[89] J.G.Taylor, Neural Networks and their Applications, John Wiley& Sons Ltd, West Sussex, UK, 1996.

[90] G. Venkateshwar Reddy, M. Jagadesh Kumar, Investigation of the novel attributes of a single-halo double gate SOI MOSFET: 2D simulation study, Microelectronics Journal, 35,p.761–765, 2004.

[91] F.Djeffal, S.Guessasma, A.Benhaya et M.Chahdi, Study the scaling capability of the DG MOSFET using Artificial Neural Network, EMRS Spring meeting, Strasbourg, France; May31-Juin 1-3, 2005.

[92] F.Djeffal, M.Chahdi et A.Benhaya, Modélisation numérique du transistor DGMOSFET fortement submicronique, Edward A.Bouchet International Conference on Physics and high Technology EBIC'05, Hammamet 11-15 August 2003.

[93] C. H. Wann, R. Tu, Bin Yu, Chenming Hu, K. Noda, T. Tanaka, M. Yoshida, and K. Hui, A comparative study of advanced MOSFET structures, Symp. on VLSI Technology Digest of Technical Papers, pp. 32-33, 1996.

[94] C.F. Gerald, Applied Numerical Analysis, Addisson-Wesley Publishing Co., Inc., 1978.

[95] N. Brioua, Elaboration d'un code de calcul de champ 2D dans les structures électromagnétiques utilisant la méthode des Eléments Finis, Thèse de Magister, Université de Batna, Algérie, 2003.

[96] F.Djeffal, A. Benhaya et M. Chahdi, Modélisation semianalytique du transistor DG MOSFET symétrique fortement submicronique, Conférence Maghrébine en Génie Electrique' CMGE2004, Constantine 12-13 Avril, 2004.

[97] B. Agrawal, Comparative scaling opportunities of MOSFET structures for gigascale integration (GSI), Ph.D. dissertation, Rensselaer Polytech. Inst., Troy, NY, 1994.

ANNEXE A

Lorsque le phénomène physique à analyser par la méthode des éléments finis n'est pas linéaire, alors le système d'équations engendré n'est pas linéaire.

Supposons qu'on veut résoudre un système d'équations algébriques non linéaires:

$$\begin{cases} f_1(x_1, x_2, \ldots\ldots\ldots, x_n)=0 \\ f_2(x_1, x_2, \ldots\ldots\ldots, x_n)=0 \\ \quad\quad . \\ \quad\quad . \\ f_n(x_1, x_2, \ldots\ldots\ldots, x_n)=0 \end{cases} \tag{A.1}$$

sous forme vectorielle, le système (A.1) s'écrit:

$$F(X)=0 \tag{A.2}$$

Où

$$X=\begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix} \quad \text{et} \quad F(X)=\begin{pmatrix} f_1(X) \\ f_2(X) \\ . \\ . \\ f_n(X) \end{pmatrix}$$

La méthode de Newton-Raphson consiste à créer une suite de vecteurs $X^{(1)}$, $X^{(2)}$, …, $X^{(k)}$, .. tels que

$$X^{(k)}=X^{(k-1)}-J^{-1}(X^{(k-1)})F(X^{(k-1)}) \tag{A.3}$$

où $J^{-1}(X^{(k-1)})$ est la matrice inverse de la matrice jacobienne:

$$J^{-1}(X^{(k-1)})=\begin{pmatrix} \frac{\partial f_1(X^{(k-1)})}{\partial x_1} & \frac{\partial f_1(X^{(k-1)})}{\partial x_2} & \ldots\ldots & \frac{\partial f_1(X^{(k-1)})}{\partial x_n} \\ \frac{\partial f_2(X^{(k-1)})}{\partial x_1} & \frac{\partial f_2(X^{(k-1)})}{\partial x_2} & \ldots & \frac{\partial f_2(X^{(k-1)})}{\partial x_n} \\ . & . & & . \\ . & . & & . \\ \frac{\partial f_n(X^{(k-1)})}{\partial x_1} & \frac{\partial f_n(X^{(k-1)})}{\partial x_2} & \ldots\ldots & \frac{\partial f_n(X^{(k-1)})}{\partial x_n} \end{pmatrix} \tag{A.4}$$

Cette suite, si elle converge, tend vers un vecteur solution du système (A.1).

Les coefficients $\frac{\partial f_i}{\partial x_j}(X^{(k-1)})$ de la matrice J peuvent être calculés numériquement par l'introduction d'une petite perturbation δ (choisis) de la variable suivant la dérivée partielle considérée. La dérivée partielle de la fonction $f_i\,(X^{(k-1)})$ suivant la variable x_j sera donc:

$$DF_{ij} = \frac{\partial f_i}{\partial x_j}(X^{(k-1)}) \approx \frac{f_i\,(x_1^{(k-1)}, x_2^{(k-1)}, \ldots\ldots x_j^{(k-1)} + \delta, \ldots, x_n^{(k-1)}) - f_i\,(x_1^{(k-1)}, x_2^{(k-1)}, \ldots\ldots x_j^{(k-1)}, \ldots, x_n^{(k-1)})}{\delta}$$

pour δ petit. La matrice jacobienne $J\,(X^{(k-1)})$ est ainsi remplacée par une matrice $DF(X^{(k-1)})$ dont les coefficients sont les DF_{ij} .

Posons $DX^{(k)} = X^{(k)} - X^{(K-1)}$. Alors l'équation (A.3) devient

$$J(X^{(k-1)})(X^{(k)} - X^{(k-1)}) = -F(X^{(k-1)}) \qquad \text{(A.5)}$$

ou

$$J(X^{(k-1)})DX^{(k-1)} = -F(X^{(k-1)}) \cong DF(X^{(k-1)})DX^{(k-1)}$$

le système d'équations algébriques (A.5) est linéaire se résout par les méthodes classiques.

Abstract

The ULSI-component industry requires more financial investment than ever in order to measure the growing sophistication of the manufactured products and for the equipment necessary to their development. So, the modelling of electronic components constitutes a research field that is currently very important and very attracting throughout the world. To continue this trend, the existing models must be improved and new models have to be developed. Hence, we regularly see improvements of simulation software. In this work, we present the applicability of artificial neural networks for the development of an analytical approach allowing the assessment of the time degradation at deep submicron level of MOSFETs devices, the development of a neural model of DG MOSFET for the study of the nanoscale CMOS circuits and also the possibility of producing a graphical abacus for the study of scaling capability of the undoped DG MOSFET. The different developed neural models can be implemented in electronic simulators (SPICE, PSPICE, CADENCE, . . .). Our results are compared with those obtained experimentally and by numerical methods. To draw some useful and decisive information about ULSI technology, these results were throughly analysed and discussed.

Keywords: *Artificial neural network; Deep submicron; Degradation; DG MOSFET; Green's function; Nanoscale CMOS; Subthreshold swing; Scaling capability.*

Résumé:

L'industrie des composants ULSI exige des investissements financiers de plus en plus lourds pour mesurer la sophistication grandissante des produits fabriqués ainsi que pour les équipements nécessaires à leur élaboration. De ce fait, la modélisation électrique des composants électriques constitue actuellement un axe de recherche très convoité à travers le monde. Pour suivre cette évolution, les modèles existants doivent être améliorés et de nouveaux modèles doivent être développés. C'est ainsi que nous assistons régulièrement à des améliorations des logiciels de simulation. Dans ce travail, on présente l'applicabilité des réseaux de neurones artificiels pour le développement d'une approche analytique permettant l'évaluation de dégradation des transistors MOSFETs fortement submicroniques, le développement d'un modèle neuronal de DG MOSFET qui permet d'étudier les circuits CMOS nanométriques et ainsi la possibilité de produire des abaques graphiques pour l'étude et l'optimisation de la réduction dimensionnelle des transistors DG MOSFETs. Les différents modèles neuronaux développés peuvent être implémentés dans les simulateurs électroniques (SPICE, PSPICE, CADENCE, . .). Nos résultats sont comparés à des résultats expérimentaux et numériques, analysés et commentés de manière à pouvoir en tirer des conclusions pratiques qui sont de nature à intéresser tous ceux qui sont appelés à réaliser des circuits de technologie ULSI.

Mots clés : *Réseau de neurones artificiel; Fortement submicronique; Dégradation; DG MOSFET; Fonction de Green; CMOS nanométrique; Pente sous seuil; Réduction dimensionnelles.*