République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Djillali Liabes Sidi Belabbes

Faculté De Génie Electrique

THESE

Présenté

En vue de l'obtention du diplôme de

DOCTEUR EN SCIENCE

Spécialité: Electronique
Option : matériaux et composants

Présenté Par

MOSTEFAI ABDELKRIM

Contribution à l'étude et la modélisation du transistor à effet de champ à grille isolée de haute permittivité diélectrique

Soutenue le 18/ 12 /2017 devant le jury composé de :

SOUDINI Belabbes	Professeur	Université de Sidi belabbés	Président
ABID Hamza	Professeur	Université de Sidi Belabbés	Encadreur
BERRAH Smail	Professeur	Université de Bejaïa	Co-encadreur
ARBOUCHE Omar	MCA	Université de Saida	Examinateur
SAHNOUN Mohamed	Professeur	Université de Mascara	Examinateur
BENSAAD Zouaoui	Professeur	Université de Sidi Belabbés	Examinateur


بسم الله الرحمن الرحيم

Remerciements

En premier je remercie Allah le tout puissant de m'avoir donné la volonté et le courage de mener à bien ce travail.

Je tiens à remercier :

Mon encadreur Prof ABID Hamza, qui m'a guidé et apporter son savoir tout le long de ce travail.

Mon Co-encadreur Prof BERRAH Smaïl, qui m'a guidé et apporter son savoir tout le long de ce travail.

Je tiens à exprimer toute ma reconnaissance à Monsieur SOUDINI Belabbes pour avoir accepté de présider le jury de ma thèse.

Que tous les membres du jury qui ont bien voulu examiner mon travail, je remercie :

Monsieur SAHNOUN Mohammed
Monsieur ARBOUCHE Omar
Monsieur BENSAAD Zouaoui

A mon père.
A ma mère.
A mes frères et ses femmes.
A ma sœur et ça famille.
A grande et petite famille.
Je remercie également tous mes amis.

Table des matières

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES SYMBOLES

1- Listes des principaux paramètres du transistors MOSFETs

Paramètres	Unité	Commentaire
μ_n	$m^2/V.s$	Mobilité de surface des électrons
ε_0	F/m	Constant diélectrique de l'espace libre (permittivité du vide) ($8.85e^{-12}$ F/m)
ε_{ox}	-	Constant diélectrique de l'oxyde
ε_{high-k}	-	Constant diélectrique de matériaux high-k
t_{ox}	m	Epaisseur d'oxyde
t_{high-k}	m	Epaisseur d'oxyde high-k
L	m	La longueur du canal
W	m	La largeur du canal
C_{ox}	$F.m^{-2}$	Capacité de l'oxyde de grille par unité de surface
V_{DS}	V	Tension Drain-Source
V_{GS}	V	Tension Grille-Source
V_T	V	Tension de seuil du transistor
I_D	A	Courant de drain
m_{eff}	Kg	Masse effective des porteurs
q	C	Valeur absolue de la charge de l'électron ($1.602*10^{-19}$ C)
ħ	eV.s (ou J.s)	Constante de planck
ϕ_B	eV (ou J)	Hauteur de barrière du « high-k »
V_{high-k}	V	La perte de tension à travers le diélectrique
i	-	Courant de drain normalisé
I_S	A	Courant spécifique
I_F	A	Courant direct normalisé
I_R	A	Courant inverse normalisé
J_g	A/m^2	Densité de courant
V_P	V	Tension de pincement
U_T	V	Tension thermique
K	$J.K^{-1}$	Constant de Boltzmann ($k=1.38*10^{23}$ $j.K^{-1}$)
T	K	Température absolue
q_F	-	Densité de charge mobile normalisée
n	-	Facteur de pente (slope factor)
Q'_{inv}	$C.m^{-2}$	charge de la zone d'inversion

2- Listes des abréviations

MOSFET	Metal Oxide Semiconductor Field Effect Transistor
PDA	Personal Digital Assistant
AG	Algorithme Génétique
ANN	Artificial Neural Network
RN	Réseau de Reurone
EOT	Equivalent Oxide Thickness
DFT	Density Functional Theory
DRAM	Dynamic Random Access Memory
MPU	MicroProcessor Unit
CMOS	Complémentary Metal Oxide Semiconductor
ITRS	International Technology Roadmap for Semiconductors
LOP	Low Operating Power
VEGA	Vector Evaluated Genetic Algorithm

Introduction Générale

Introduction Générale

Depuis l'invention du circuit intégré en 1958 [1-2], le principal vecteur de croissance de l'industrie microélectronique Silicium est la miniaturisation des transistors MOSFET. Ce composant est la brique élémentaire de tout circuit logique et concentre à lui seul une grande part des éfforts de recherche et développement menés ces dernières décennies pour accroître les performances des circuits intégrés.

Le premier avantage de la réduction d'échelle (ou scaling) se comprend intuitivement : une plus grande densité d'intégration permet de réaliser des fonctions électroniques plus complexes sur une même surface, et réduit ainsi le coût de fabrication unitaire de chaque transistor. En 1965, Gordon E. Moore édicta sa fameuse loi éponyme, prédisant une augmentation de la densité d'intégration des circuits d'un facteur 2 tous les 2 ans [1-3]. Cinquante années plus tard, force est de constater que cette conjecture s'est révélée exacte : un circuit Silicium contient aujourd'hui près d'un milliard de transistors (contre moins de 100 dans les années 1960) et l'électronique fait aujourd'hui partie intégrante de notre quotidien.

L'intégration continue du transistor MOSFET conventionnel requiert de nouvelles innovations pour contrecarrer ces limites physiques obligeant les chercheurs à trouver des solutions pour pouvoir réaliser des transistors toujours performants. Afin de réduire les effets indésirables dus à la miniaturisation des transistors MOSFETs, plusieurs solutions ont été utilisées dans les technologies les plus avancées pour améliorer les performances du transistor. Parmi ces innovations, on peut citer l'utilisation de nouveaux oxydes de grille à haute permittivité (high-k) pour réduire le courant de grille.

Ces composants, qui atteignent maintenant une taille nanométrique, doivent respecter les règles établies par l'ITRS (International Technology Roadmap for Semiconductors).

Suivant les spécifications de l'International Technology for Roadmap Semiconductor (ITRS), le matériau retenu, en plus d'avoir une permittivité plus élevée que celle du SiO_2, doit présenter une largeur de bande interdite suffisante et un bon alignement de bande avec le silicium.

De nombreux candidats sont potentiellement aptes à remplir le cahier des charges imposé par l'ITRS. L'un d'entre eux, l'oxyde d'hafnium (HfO_2) s'est distingué. Ce matériau est largement étudié depuis ces 10 dernières années du fait de ses propriétés physiques et électriques encourageantes et de sa compatibilité avec la technologie du silicium.

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.

Le domaine de la modélisation et la simulation des dispositifs fortement submicroniques peut être considéré comme un champ important d'applications des réseaux de neurones et des algorithmes génétiques.

Les réseaux de neurones sont apparus dans les années cinquante mais n'ont reçu cependant un intérêt considérable qu'à partir des années 80 avec l'apparition de l'algorithme de rétropropagation (Rumelhart et McClelland, 1986). Leur capacité d'apprentissage et d'approximation de fonctions leur procure un intérêt considérable de la part des chercheurs. Il suffit de voir les nombreuses applications industrielles qui en découlent à partir des années 90 et de consulter l'abondante littérature sur le sujet pour s'en convaincre.

Les algorithmes génétiques sont des méthodes stochastiques basées sur une analogie avec des systèmes biologiques. Ils reposent sur un codage de variables organisées sous forme de structures chromosomiques et prennent modèle sur les principes de l'évolution naturelle de Darwin pour déterminer une solution optimale au problème considéré. Ils ont été introduits par Holland (Holland, 1975) pour des problèmes d'optimisation complexe. Contrairement aux méthodes d'optimisation classique, ces algorithmes sont caractérisés par une grande robustesse et possèdent la capacité d'éviter les minimums locaux pour effectuer une recherche globale. De plus, ces algorithmes n'obéissent pas aux hypothèses de dérivabilité qui contraignent pas mal de méthodes classiques destinées à traiter des problèmes réels.

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 et l'approche génétique pour le développement d'un modèle du transistor MOSFET fortement submicronique.

Ce mémoire est organisé en trois chapitres comme suit:

Le premier chapitre expose les principales notions de fonctionnement du transistor MOS**,** la problématique liée à la miniaturisation dans le domaine de la microélectronique et l'étude sur les matériaux à haute permittivité « high-k ». Ainsi qu'une justification du choix des matériaux étudiés.

Le deuxième chapitre est consacré aux Intelligence artificielle (les réseaux de neurone et les algorithmes génétiques). Il est constitué de deux parties principales. La première traite les réseaux de neurones et leur modalité d'utilisation pour les problèmes de modélisation, la deuxième présente une description détaillée des algorithmes génétiques dans laquelle nous rappelons les définitions relatives à leur fonctionnement.

Le troisième chapitre présente la simulation prédictive du transistor MOSFET (oxyde HfO_2 monoclinique, oxyde HfO_2 tétragonal) fortement submicronique par les réseaux de neurones et les algorithmes génétiques pour identifier les différentes paramètres du transistor MOSFET.

Chapitre I

Le transistor MOSFET éléments de théorie et l'état de l'art

I-1 Introduction :

Depuis la réalisation du premier transistor par William Shockley, Walter Brattain et John Bardeen en 1948 [1], la microélectronique a parcouru un chemin incroyable. D'une taille des dispositifs de l'ordre du centimètre dans les années 50, nous sommes rentrés, en ce début du XXIème siècle, dans l'ère du nanomètre. Les premiers sont réalisés dans un même monocristal semiconducteur, divisé en trois régions dopées différemment (pnp ou npn) et formant deux jonctions montées en opposition avec une zone commune [1-2-3]. Ici, deux types de porteurs participent à la conduction : les électrons (n) et les trous (p).

De nos jours, le composant le plus utilisé dans les circuits intégrés est le transistor MOS (*Métal Oxyde Semi-conducteur*). Apparu dans les années 60 et généralisé dans les années 80 en association avec la logique CMOS (Complémentaire *Métal Oxyde Semi-conducteur*), l'amélioration de ses performances, conformément à la célèbre loi de Moore, est passée et passe encore par la miniaturisation de ses dimensions. Les raisons de cette miniaturisation sont simples. Elles sont :

- Financières :
  - Réduction du coût par fonction.
  - Augmentation de la productivité des usines microélectroniques.

- Technologiques :
  - Augmentation de la densité de composants par « wafer ».
  - Réduction de la consommation électrique.
  - Vitesse des circuits.

Dans le cadre de notre travail, nous allons nous intéresser uniquement aux transistors à effet de champ et plus précisément les transistors MOS. L'objectif de ce premier chapitre est de présenter un état de l'art général sur la problématique de miniaturisation des transistors MOS (Metal Oxide Semiconductor). Pour ce faire, nous allons procéder comme suit :

- Dans un premier temps, nous présenterons les transistors MOS et les limitations des technologies actuelles.

- Dans un deuxième temps, nous considérerons les problèmes posés par la substitution à la silice d'un oxyde à forte permittivité diélectrique. Entre autres, seront présentés les critères auxquels doivent répondre les oxydes high-*k* pour être considérés comme candidats potentiels au remplacement de la silice. Les propriétés spécifiques des oxydes high-*k* sur lesquelles s'est portée notre étude seront discutées.

I-2 Transistor MOS : historique et définition :

Les transistors à effet de champ correspondent à des dispositifs constitués de quatre électrodes : la source, le drain, la grille et le substrat [3-4]. Contrairement aux transistors bipolaires, un seul type de porteurs de charge (électrons ou trous) participe à la conduction [3-4-5]. De plus, le mode de conduction est surfacique, à l'opposé des transistors bipolaires où l'on assiste plutôt à une conduction en volume [5].

Les premiers travaux sur le concept du transistor Metal Oxide Semiconductor Field Effect Transistor (MOSFET ou MOS) datent de 1926-1930 avec le brevet déposé par Lilienfeld et Heil [4-5]. En 1955, I. Ross mit en place la structure actuelle du transistor MOS ; mais, à la place de l'oxyde de grille, on a alors plutôt un cristal ferroélectrique. C'est en 1960 que Kahng et Atalla présentent le premier transistor MOS en silicium complètement opérationnel, avec, comme oxyde de grille, l'oxyde de silicium [4-5]. C'est le point de départ de l'essor de l'industrie du circuit intégré [4-5]. Dans la vie de tous les jours, les transistors se retrouvent dans les cartes à puces (environ 1000 transistors), les agendas électroniques PDA (environ 5 millions de transistors), les téléphones portables (environ 10 millions), les clés USB (environ 1 milliard de transistors pour une clé de 128 Mo), la principale application étant bien sûr celle des microprocesseurs (environ 820 milliards de transistors pour le « Core 2™ Quad » d'Intel).

I-3 Le transistor MOS à effet de champ :

I-3.1 Présentation de la structure MOS :

Le transistor MOS est un commutateur qui permet de commander sous l'action d'un champ électrique, le passage d'un courant à travers un canal séparant deux réservoirs de porteurs de charge distincts : la source et le drain (S et D) [2-3-4-6]. Sa structure est la suivante : sur un substrat semi-conducteur de type donné (p ou n), on vient créer par diffusion d'impuretés (dopantes), deux contacts ou électrodes (la source et le drain) dont le type de conduction diffère de celui du substrat. L'ensemble est ensuite isolé d'une électrode métallique, la grille, par une fine couche de diélectrique (oxyde de grille) [2-3-4] (Figure I-1). Habituellement, c'est l'oxyde de silicium (SiO_2) qui est utilisé.

Les dimensions d'un transistor MOS sont définies par la longueur L de la grille, la largeur transversale W du transistor et l'épaisseur t_{ox} de l'oxyde diélectrique.

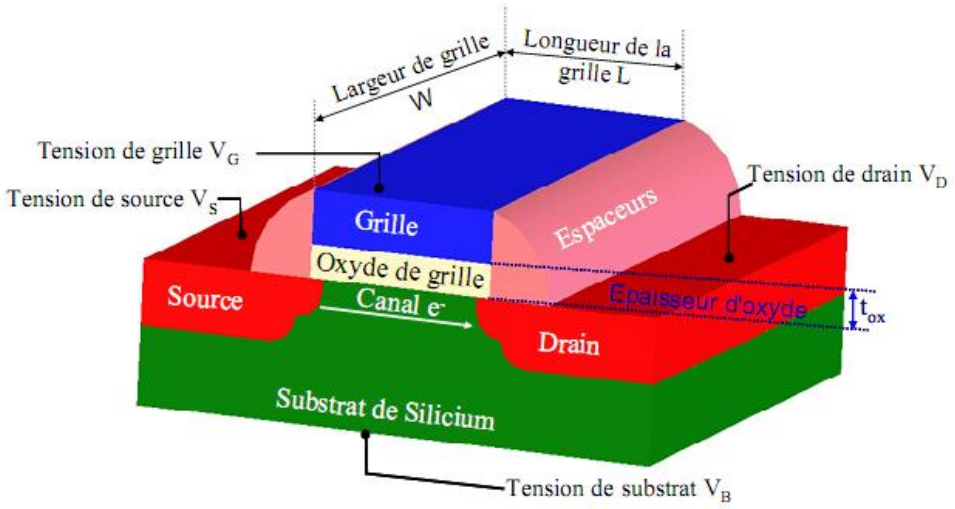


Figure I-1 : Structure d'un transistor MOS

Selon la nature des porteurs assurant la conduction électrique, on distingue deux types de transistors MOS : Le transistor MOS à canal N appelé N-MOS, où le passage du courant est gouverné par le transport des électrons, et le transistor MOS à canal P ou P-MOS, pour lequel ce sont les trous qui assurent le passage du courant électrique [2-6]. Notons toutefois que dans le cas du transistor N-MOS, le substrat semiconducteur est de type P et qu'il est de type N pour un P-MOS [2-6].

I-3.2 Principe et régimes de fonctionnement du MOSFET :

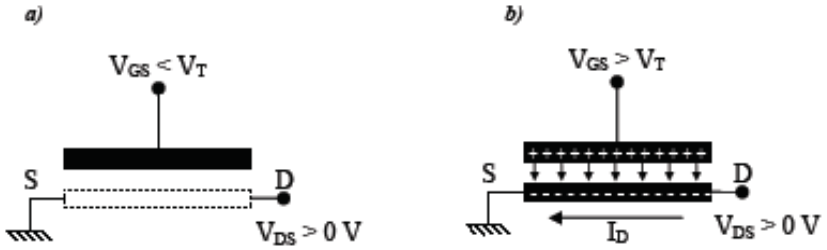


Figure I-2 : Principe de fonctionnement d'un transistor MOS [7]. a) Etat bloqué. b) Etat passant.

Le principe de fonctionnement d'un tel dispositif est schématisé en Figure I-2. La tension de grille crée un champ vertical qui, par l'intermédiaire de la capacité MOS, module la densité de porteurs libres à l'interface Substrat/Oxyde permettant ainsi de commander sa conductivité. Ce phénomène est appelé « effet de champ ».

L'un des paramètres essentiels de fonctionnement d'un transistor MOS est la tension seuil(V_T). Elle représente la tension à appliquer entre la grille et le substrat afin que le transistor commence à conduire. Elle constitue la délimitation entre les deux régimes de fonctionnement du transistor : régime de faible inversion ou sous seuil et régime de forte inversion ou passant.

Lorsqu'on applique sur la grille une tension (V_{GS}) positive (transistor N-MOS) et supérieure à la tension seuil (V_T), on dit que le transistor se trouve en régime de forte inversion. Ici, une accumulation de charges (électrons) est créée par effet de champ, à la surface du substrat, juste en dessous du diélectrique de grille. On forme ainsi entre la source et le drain, à la surface du semiconducteur, un canal dit d'inversion qui est riche en électrons [2-4].

L'application d'une différence de potentiel V_{DS} entre le drain et la source va induire le passage d'un courant plus ou moins important (I_{DS}), du drain vers la source.

Si $V_{DS} < V_{GS} - V_T$, l'effet de champ est uniforme. Le transistor fonctionne donc en régime linéaire et l'expression du courant est donnée par la relation :

$$I_D = \mu_n C_{ox} \left(\frac{W}{L}\right) \left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] \qquad \mathbf{I-1}$$

Ou

$$I_D = k_n[2(V_{GS} - V_T)V_{DS} - V_{DS}^2] \qquad \mathbf{I-2}$$

Avec

$$k_n = \frac{\mu_n \varepsilon_0 \varepsilon_{ox}}{2t_{ox}} \frac{W}{L} = \frac{\mu_n C_{ox}}{2} \left(\frac{W}{L}\right) \qquad \mathbf{I-3}$$

Avec

μ_n mobilité de surface des électrons
ε_0 constant diélectrique de l'espace libre (permittivité du vide) ($8.85e^{-14}$ F/cm)
ε_{ox} constant diélectrique de SiO_2
t_{ox} épaisseur de l'oxyde
L la longueur du canal
W la largeur du canal

Par contre, si $V_{DS} > V_{GS}-V_T$, le canal d'inversion est pincé du côté du drain et le courant I_{DS} est à son maximum et n'augmente plus avec V_{DS}. On dit que le transistor est en régime saturé. L'expression du courant de saturation ($I_{on}=I_D$) est alors donnée par la relation :

$$I_D = k_n (V_{GS} - V_T)^2 \qquad \mathbf{I-4}$$

Quand

$$V_{DS} = V_{GS} - V_T \qquad \mathbf{I-5}$$

À partir de l'équation **(I-2)** et **(I-4)**

$$I_D = k_n {V_{DS}}^2$$

Toutefois, si la tension V_{GS} est inférieure à V_T, le transistor se trouve en régime de faible inversion, et on parle de déplétion. La concentration des porteurs dans le canal de conduction est pratiquement nulle et un courant I_{DS} circule entre la source et le drain indépendamment de la valeur de la tension V_{DS}. C'est le courant de fuite, couramment noté I_{off}. On dit que le transistor est en régime bloqué.

Dans le cas du transistor P-MOS, on a la situation inverse. En effet, ici, la grille est polarisée négativement ($V_{GS} < 0$). On a alors création d'un canal de conduction de type trous à la surface du substrat semiconducteur qui est de type N. Ainsi, lorsqu'on applique une tension V_{DS} négative entre les contacts source et drain, on a une circulation des trous de la source vers le drain [2-5].

A l'heure actuelle, la technologie la plus répandue est celle associant les deux types de transistors MOS. On parle alors de CMOS (Complementary MOS ou MOS complémentaires).

Son principal intérêt est de réduire de façon considérable la puissance consommée, par l'association appropriée des transistors MOS de différents types [2-4].

Du fait de sa technologie de fabrication et d'intégration assez simple, couplée à un faible encombrement sur substrat, on assiste depuis plusieurs années à une réduction continuelle et spectaculaire des dimensions des transistors CMOS. Le but premier de cette miniaturisation est de réduire les coûts de fabrication (plus de puces par plaque

de silicium fabriquée) tout en augmentant les performances (vitesse accrue, plus de fonctions et grande capacité mémoire).

En effet, le coût de fabrication est inversement proportionnel à la densité d'intégration de transistors par plaque de silicium – on atteint aujourd'hui environ 12 milliards de transistors sur des wafers de 300 mm de diamètre - ; Il est donc proportionnel aux dimensions du transistor. L'une des voies de miniaturisation passe ainsi par la réduction de la longueur de la grille. Ceci implique une réduction de l'épaisseur de l'oxyde de grille afin de maintenir constant ou d'augmenter le couplage capacitif grille/substrat :

$$C_{ox} = \frac{\varepsilon_{ox}\varepsilon_0}{t_{ox}} \qquad \mathbf{I-6}$$

Avec

C_{ox} Capacité de l'oxyde de grille par unité de surface
ε_0 Constant diélectrique de l'espace libre (permittivité du vide)
ε_{ox} Constant diélectrique de SiO_2
t_{ox} Epaisseur de l'oxyde

I-3.3 Structure CMOS et Miniaturisation avec la silice comme oxyde de grille :

Il est important de souligner que pour l'on peut intégrer les deux types de transistors dans un même substrat (structure complémentaire appelée CMOS). A la figure (I-3), on montre ce type de structure [8].

Les transistors à canal p sont intégrés directement dans le substrat de type n, alors que la source et le drain du transistor à canal n sont intégrés dans un caisson d'isolation de type p.

Afin de satisfaire la polarisation inverse des jonctions BS et BD, le substrat de type n est connecté au potentiel le plus positif et le caisson de type p au potentiel le plus négatif du circuit [8].

Les diffusions sont généralement réalisées en utilisant la technique de l'implantation ionique.

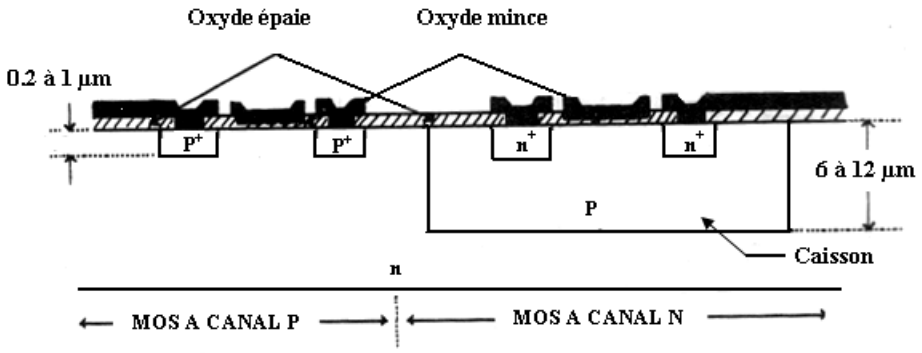


Figure I-3 : Représentation schématique d'une structure CMOS [8].

Durant plusieurs années la silice a été utilisée comme isolant de grille dans les structures CMOS; c'est elle qui a rendu possible le développement accéléré et poussé de l'industrie de la microélectronique.

En effet, elle présente une bonne stabilité thermique et, malgré sa faible permittivité (K_{SiO2}= 3.9) [9-10-11], elle a d'excellentes propriétés diélectriques, notamment une largeur de bande interdite de 8.9 eV [9-10-11] conduisant à un offset approprié avec les bandes du silicium. Enfin, elle possède une faible densité de défauts (en volume $<10^{16}$ cm^{-3} et $< 10^{11}$ cm^{-2} à l'interface Si/SiO_2), une haute résistivité électrique ($\geq 10^{15}$ Ω.cm) et un champ de claquage diélectrique élevé ($> 10^7$ V/cm) [12-13]. Elle a donc tout pour être l'oxyde de grille idéal du silicium.

Mais, lorsqu'on diminue son épaisseur en dessous d'une certaine valeur (~ 12 Å) [11-14], de nombreux problèmes se posent: effets de canaux courts, diffusion des dopants (en particulier le bore) dans la couche d'oxyde [9-12], dispersion des dopants dans le canal [9-12], courants de fuite de grille atteignant des valeurs excessives.

Jusqu'à présent, la silice a été utilisée comme oxyde de grille.

I-4 Le « scaling » - Loi de Moore :

Le monde de la microélectronique est en constante recherche d'amélioration de la rapidité et de la densité d'intégration de ses composants. L'évolution de la technologie CMOS vers la miniaturisation (ou downscaling), qui doit suivre des règles strictes, a entraîné de très nombreux défis technologiques et en particulier des solutions pour diminuer la longueur de grille des transistors. Nous présenterons dans cette partie l'évolution des dimensions du transistor depuis que Gordon Moore a prédit sa célèbre « Loi de Moore ». Nous présenterons également la « Roadmap » des semi-conducteurs éditée régulièrement par « l'International Technology Roadmap for Semiconductors : ITRS » dont le rôle est d'évaluer les technologies à mettre en place pour les années futures, afin que l'évolution suive la loi de Moore.

I-4.1 Loi d'échelle :

Plus les technologies avancent, plus le nombre de composants dans les équipements électroniques augmente. Le but est de miniaturiser ces équipements électroniques (exemple des équipements portatifs) et par conséquent, les composants doivent être de plus en plus petits et leur consommation doit être réduite. Ces composants doivent cependant conserver le même mode de fonctionnement et surtout les mêmes propriétés électriques de base.

En 1965 [Moore 65], Gordon Moore a formulé des hypothèses sur les progrès technologiques concernant la fabrication des circuits intégrés. Il publia sa célèbre « loi de Moore » qui permit par la suite de mesurer l'évolution extrêmement rapide du nombre de transistors inclus dans une puce. Gordon Moore affirma que la densité des transistors doublerait tous les ans pour une puce de même taille et par conséquent, serait accompagnée d'un abaissement des coûts de fabrication par transistor. En 1975, la loi fut réévaluée et adaptée au microprocesseur pour lequel Moore annonça que le rythme de croissance des transistors doublerait tous les 18 mois.

L'amélioration des techniques de fabrication, et en particulier les outils de photolithographie, fut un point clef pour la réduction d'échelle. Autour des années 2000, cette tendance s'approche plutôt d'un doublement tous les 2 ans (Figure I-4).

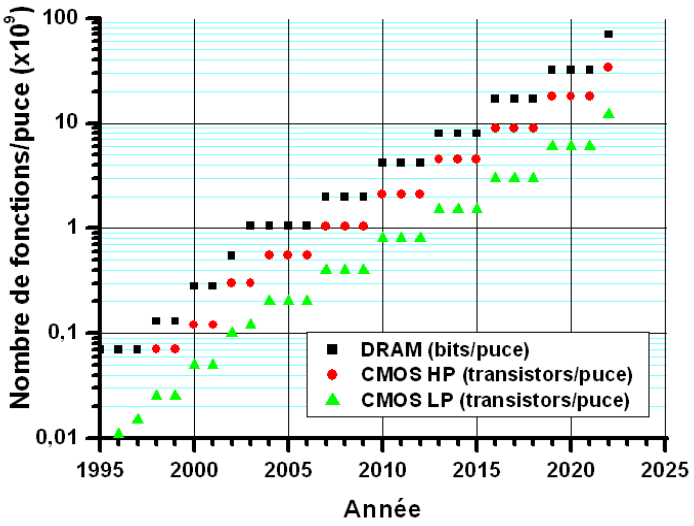


Figure I-4 : Le nombre de transistors par unité de surface double en moyenne tous les 2 ans [d'après ITRS2007].

Intéressons-nous plus en détail au transistor MOS. Le défi consiste à trouver des solutions pour continuer à diminuer la longueur de la grille des transistors. La réduction de cette longueur suit des règles strictes (tableau I-1) qu'on appelle les lois d'échelle, et doit s'effectuer en réduisant, en même temps que la taille, d'autres paramètres géométriques, électriques et physiques afin de préserver les bonnes caractéristiques électriques des transistors.

	Transistor MOS et circuit	**Facteur multiplicatif K>1**
Dimensionnement du transistor	Dimensions du transistor : T_{ox}, L_G,W Tension Concentrations de dopants	1/K 1/K K
Effets sur les transistors	Courant [A] Capacité [F]	1/K 1/K
Effets sur les circuits	Temps de retard (CV/I) Puissance dissipée par circuit (VI) Densité Puissance (P/Surface)	1/K $1/K^2$ 1

Tableau I-1 : Lois d'échelles des transistors (facteur à champ constant) [17].

I-5 Problématique du SiO_2 et solutions à la miniaturisation :

I-5.1 L'effet de diminution de l'épaisseur d'oxyde de grille SiO_2 :

La miniaturisation des transistors se traduit par la diminution de la longueur de grille. En parallèle, le couplage capacitif entre la grille et le substrat doit augmenter, ou rester constant. En conséquence, pour conserver ces propriétés, l'épaisseur d'oxyde de grille t_{ox} est réduite.

Pour augmenter la capacité de l'oxyde de grille, la première solution serait de diminuer l'épaisseur physique de SiO_2. C'est ce qui a été effectivement appliqué jusqu'au nœud technologique de 90 nm. Dans ce cas, l'épaisseur du SiO_2 est supérieure ou égale à 2 nm, et le courant de fuite n'est pas un facteur limitant. Pour les nœuds technologiques de 65 nm et 45 nm, les limites du SiO_2 sont atteintes, et la recherche de nouveaux matériaux est nécessaire. La figure I-5 représente les performances du SiO_2 pour les applications « basse consommation» (LOP), et met en évidence, l'écart entre le courant de fuite du SiO_2 et les spécifications en fonction des générations technologiques.

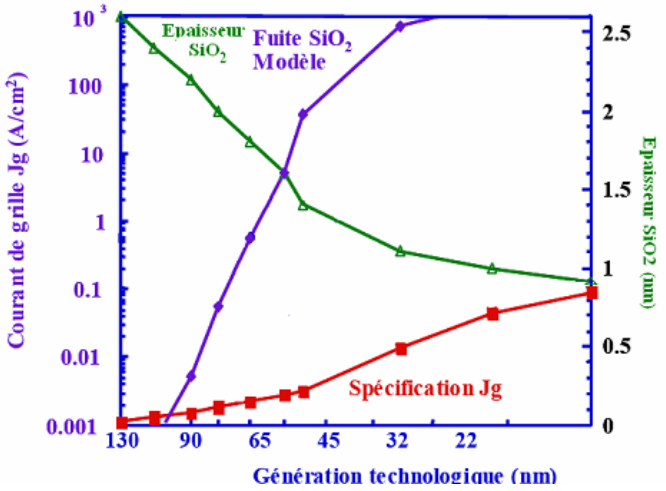


Figure I-5 : Données ITRS pour les transistors pour les applications « basse consommation » (LOP). Le courant de grille croît exponentiellement lorsque l'épaisseur d'oxyde SiO_2 diminue [18].

Lorsque l'épaisseur de l'oxyde de grille est inférieure à 2 nm, les différentes limitations sont :

1- Les courants de fuite, liés au courant tunnel (Annexe I) des électrons à travers l'oxyde SiO_2, excèdent les 10 A/cm² à 1V (10 A/cm^2 pour une épaisseur de silice de 15 Å à une tension d'environ 1 V), et deviennent inacceptables [10-15-16-19-20]. Afin d'illustrer ces propos, les densités de courant de la grille, en fonction de la tension de grille, sont tracées sur la figure I-6. Pour des transistors fabriqués avec un diélectrique de grille inférieur à 3,5 nm, le courant de fuite augmente exponentiellement [21] ce qui dégrade la fiabilité du composant.

2- La mobilité des porteurs dans le canal diminue [22-23]. La diminution de la mobilité est liée à la difficulté de créer des interfaces de bonnes qualités [24-25]. Cette dégradation est associée à la présence de charges dans le diélectrique et aux interfaces.

3- Le phénomène d'électrons chauds qui dégrade les propriétés du diélectrique. Ces électrons, d'une énergie variant de 1 à 3 eV, traversent le canal de la source au drain et irradient l'interface Si/SiO_2. Cela enclenche une succession de phénomènes physiques (effet tunnel, création de pièges, ionisation par impact, génération d'états d'interfaces) ou chimiques (réactions avec l'hydrogène). Au final, la tension de claquage du diélectrique diminue et la fiabilité du système est réduite [21].

4- La fiabilité des films SiO_2 face au claquage électrique diminue avec l'épaisseur.

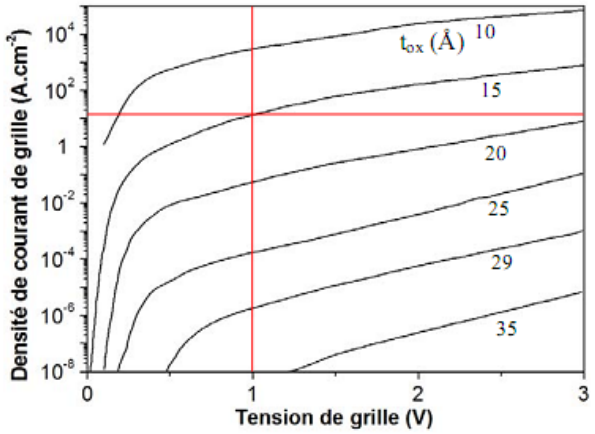


Figure I-6 : Densité de courant de la grille en fonction de la tension de grille pour une capacité CMOS avec différentes épaisseurs d'oxyde de grille SiO_2.

Les problèmes rencontrés lors de la diminution de l'épaisseur de l'oxyde de grille SiO_2 vont donc affecter les performances électriques (réduction de la tension de seuil, résistance au claquage, courant de fuite). Les propriétés des composants sont dégradées ainsi que le contrôle de fabrication et la reproductibilité. L'ensemble de ces arguments démontre donc l'intérêt de remplacer l'oxyde de grille SiO_2 par des matériaux à plus forte permittivité ε.

I-5.2 L'introduction des matériaux « high-k » :

Pour augmenter la capacité de l'oxyde de grille, est l'utilisation de matériaux avec un constant diélectrique k, plus élevée que celle du SiO_2. Il est ainsi possible de conserver un dispositif aminci fonctionnel et, de plus, d'avoir une meilleure flexibilité sur l'épaisseur de l'oxyde déposé. L'utilisation des matériaux « high-k » permet de déposer une épaisseur t_{ox} plus élevée comparée au SiO_2 tout en conservant les mêmes propriétés capacitives.

Pour une capacité donnée, si l'épaisseur du diélectrique, t_{HK}, est plus élevée, le courant de fuite, Jg, sera réduit (Figure 1.7) car le courant tunnel est inversement proportionnel à l'exponentielle de l'épaisseur physique [27-28-29] (équation [I-7]). Il est donc possible, en remplaçant l'oxyde de grille par les matériaux « high-k », de résoudre une partie des limitations électriques du SiO_2.

$$J_g = \frac{A}{t_{high-k}} exp\left\{-2t_{high-k}\sqrt{\frac{2m_{eff}q}{\hbar^2}\left\{\phi_B - \frac{V_{high-k}}{2}\right\}}\right\} \qquad \mathbf{I-7}$$

Avec :

A : constante
$t_{high\text{-}k}$: épaisseur physique du diélectrique
m_{eff} : masse effective des porteurs
q : charge électronique
ħ : constante de planck

ϕ_B : hauteur de barrière du « high-k »
V_{high-k} : la perte de tension à travers le diélectrique

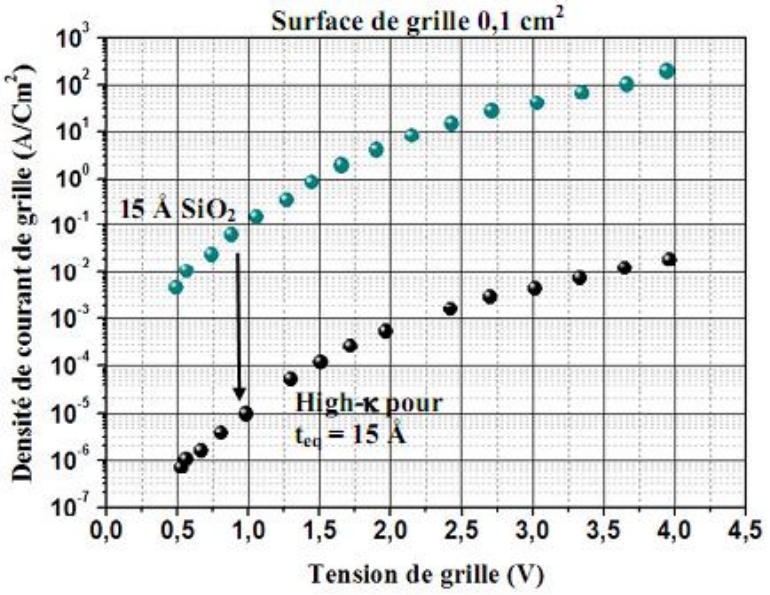


Figure 1-7 : Comparaison de la densité de courant de fuite entre une puce d'épaisseur 15Å d'oxyde de grille SiO_2 et une puce de diélectrique ayant la même EOT (Surface totale de grille = 0.1 cm^2) [9].

I-5.3 Critères de sélection des oxydes high-*k* :

La réduction des dimensions des transistors MOS s'accompagne de la réduction de l'épaisseur de l'oxyde de grille, de manière à augmenter le couplage capacitif entre la grille et le canal et garder ainsi le contrôle de la couche d'inversion. Or, l'épaisseur des oxydes de grille actuels correspond à quelques couches atomiques seulement, ce qui se traduit par l'apparition d'un courant de fuite important, mettant en péril les futurs générations 0,1µm et au-delà.

Pour pallier ce problème, on envisage de remplacer le diélectrique de grille (actuellement du SiO_2) par un diélectrique de permittivité supérieure.

Les critères de sélection des oxydes high-*k* candidats au remplacement de la silice comme oxyde de grille dans les structures CMOS sont définis par un consortium, fixant chaque année, une feuille de route appelée ITRS (International Technology Roadmap for Semiconductors) [30]. Avec la fabrication en 1999 de CMOS dont la longueur de grille était de 65 nm, l'ITRS a mis en exergue la nécessité d'introduire de nouveaux matériaux comme oxyde de grille. Un certain nombre d'oxydes high-*k* candidats ont ainsi été répertoriés (Figure 1-8).

Année deproduction		2006	2007	2010	2013	2017	2020
Nœud téchnologique	DRAM	70	65	45	32	20	14
	MPU	78	68	45	32	20	14
Application haute performance	L_g(nm)	28	25	18	13	8	5
	EOT(nm)	1.1	1.1	0.65	0.5	0.5	0.5
	V_{dd}(V)	1.1	1.1	1	0.9	0.7	0.7
	J_g(A/cm^2)	536	800	1560	2230	1380	2200
Application basse consomation en fonctionnement	L_g(nm)	37	32	22	16	10	7
	EOT(nm)	1.3	1.2	0.9	0.8	0.7	0.7
	V_{dd}(V)	0.8	0.9	0.7	0.6	0.5	0.5
	J_g(A/cm^2)	41	78	110	310	1000	1400
Oxyde de grille	Oxyde nitrurés		HfO_X ; Si, N			Oxyde de terres rares	

Tableau 1-2 : Résumé des critères physiques (EOT, longueurs Lg de grille et courants de fuite de l'oxyde de grille Jg) et des matériaux requis par l'ITRS 2007 pour une miniaturisation toujours plus poussée des architectures CMOS. Cas des applications logiques Hautes Performances et Basse Consommation en fonctionnement. Les données en rouge représentent les nœuds technologiques pour lesquels les solutions ne sont pas connues. (DRAM : Dynamic Random Access Memory ; MPU : MicroProcessor Unit).

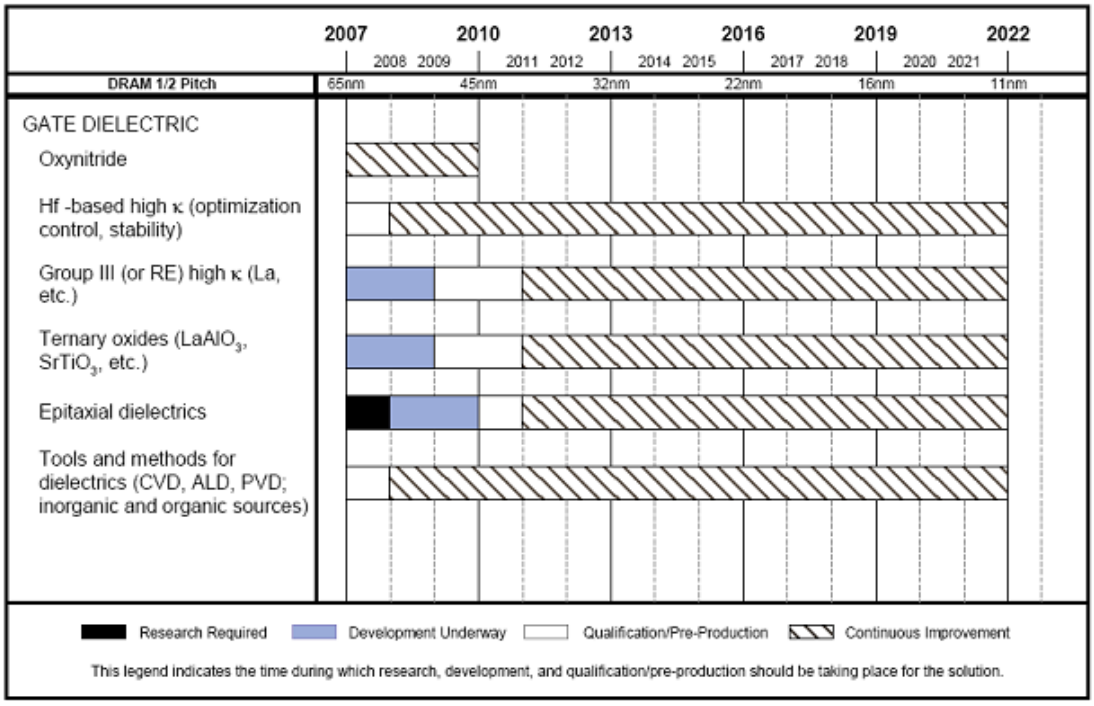


Figure 1-8 : Evolution des matériaux requis par l'ITRS 2007 pour les nouvelles architectures CMOS.

Les diélectriques de grille doivent remplir un certain nombre de critères, notamment par rapport au substrat de silicium. Les propriétés à considérer pour la sélection de l'oxyde high-*k* sont les suivantes :

- La permittivité diélectrique (*k*)
- Compatibilité avec la technologie silicium

- La discontinuité et l'alignement des bandes par rapport à celle du silicium (Offset = 1eV),

I-5.3.1 La permittivité diélectrique :

C'est évidemment le premier facteur à prendre en compte lorsqu'on parle d'oxyde à haute permittivité. En effet, pour être intéressants, les oxydes doivent posséder une permittivité relative supérieure à 10. Or, d'après la littérature [10], la permittivité des oxydes candidats est globalement inversement proportionnelle à leur bande interdite.

Ainsi, certains oxydes ayant une permittivité très élevée se trouvent être parfois mis à l'écart, qui possède une permittivité très élevée (2200), mais qui ne peut cependant pas, être considéré comme potentiel candidat car sa bande interdite est relativement faible (2.2 eV) [9] au regard de ce seul critère. C'est le cas de $BaTiO_3$. Aussi, une permittivité trop élevée serait néfaste pour le bon fonctionnement des structures CMOS car elle conduirait à la défocalisation des lignes de champ (« fringing fields effect») au niveau des électrodes (source et drain) du transistor [10]. La conséquence directe de ce phénomène est une diminution considérable de la mobilité des porteurs dans le canal séparant les deux électrodes [10].

Toutefois, un pas important a été franchi en 2007 avec l'introduction de HfO_2 comme oxyde de grille dans les nouvelles générations de transistors [30] (Figure 1-8).

I-5.3.2 Compatibilité avec la technologie silicium :

En plus du respect des performances électriques, le matériau sélectionné doit être compatible avec la technologie silicium. Il doit donc respecter certains paramètres pour être compatible avec les critères requis par l'ITRS [20-21-26].

- Il doit être compatible avec l'ensemble de la technologie d'intégration afin d'éviter la mise en place de nouvelles étapes.

- L'oxyde doit former une interface de très bonne qualité avec le canal en silicium, afin d'éviter la dégradation de la mobilité des porteurs liée à de la rétrodiffusion de ces porteurs par les défauts ou la rugosité d'interface.

- Concernant les propriétés intrinsèques, les « high k » doivent avoir :

  - Une forte permittivité (> 15),
  - Une hauteur de barrière avec le silicium supérieur à 1.1 eV. Il s'agit du minimum d'énergie requis entre les bandes de conductions du silicium et de l'oxyde (figure I-9). Cette valeur correspond à l'écart énergétique minimum, pour éviter le transport d'un électron par émission thermique ou par effet tunnel.

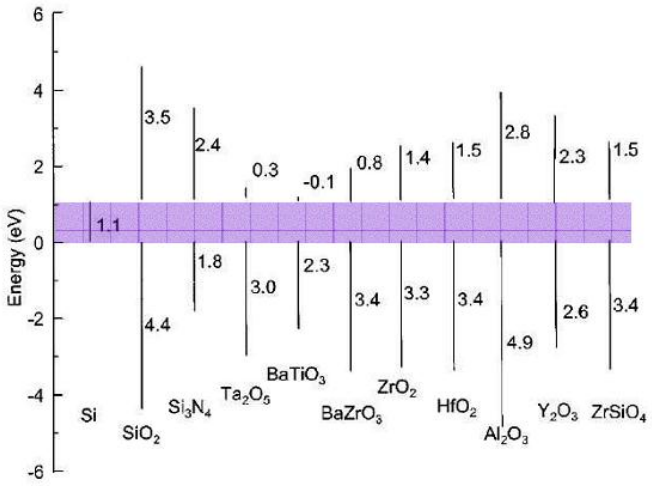


Figure I-9 : Ecarts de bandes d'énergie pour certains matériaux diélectriques « high-k » [20].

I-5.3.3 Discontinuité et alignement des bandes par rapport au silicium :

Pour éviter tout transfert de charge à l'interface, les bandes d'énergie du diélectrique doivent être suffisamment décalées par rapport à celles du silicium. La valeur minimale requise de cet "offset" est généralement estimée à 1 eV pour les bandes de conduction et de valence (Figure I-10).

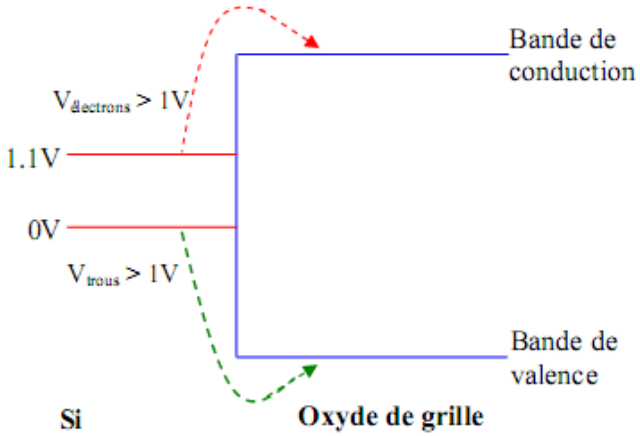


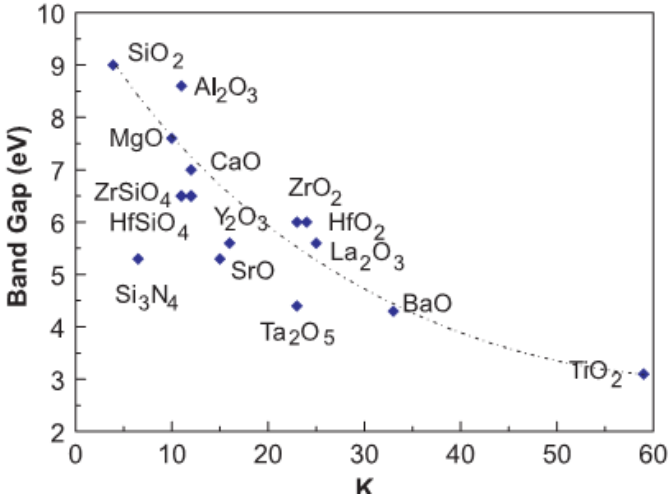
Figure I-10 : Schéma des discontinuités (offset) de bandes inhibant le passage des porteurs dans les bandes de l'oxyde.

Pour des oxydes dans lesquels la bande de valence dérive des orbitales 2p de l'oxygène, situées à environ -8 eV par rapport au vide, la condition d'un décalage de bandes supérieur à 1 eV est toujours largement satisfaite pour les bandes de valence [31].

En revanche, pour assurer un "offset" suffisant du côté bande de conduction, l'oxyde doit avoir un gap supérieur à 4 eV. Pour réaliser ces conditions, il doit contenir des éléments fortement électropositifs [31].

I-5.4 Les diélectriques de grille à haute permittivité (High-k) :

De nombreux matériaux sont actuellement étudiés pour remplacer le SiO_2 pour les longueurs de grille inférieures à 0,1µm. Les principaux sont : Al_2O_3 [32], La_2O_3 [33], Ta_2O_5, TiO_2 [34], HfO_2 [35-36], ZrO_2 [37], et Y_2O_3 [38]. Ils ont chacun un ε_{ox} différent. Néanmoins, ils ont un gap généralement moins important que le SiO_2, ce qui se traduit par des hauteurs de barrière plus faible pour les trous et les électrons, les performances des High-k par rapport au SiO_2 (figure I-11).



a)

Matériel	Diélectrique constant	Bande Gap E_0 (eV)	ΔE_0 (eV) to Si	Cristal structure
SiO_2	3.9	8.9	3.2	Amorphous
Si_3N_4	7	5.1	2	Amorphous
AL_2O_3	9	8.7	2.8 a	Amorphous
Y_2O_3	15	5.6	2.3 a	Cubic
La_2O_5	30	4.3	2.3 a	Hexagonal, cubic
TiO_2	80	3.5	1.2	Tetragonal, rutile, anatase
HfO_2	25	5.7	1.5 a	Monoclinic, Tetragonal, cubic
ZrO_2	25	7.8	1.4 a	Monoclinic, Tetragonal, cubic

- a Calculated by Robertson.

b)

Figure I-11 : L'énergie de bande interdite est représentée en fonction de la constante diélectrique pour les matériaux envisageables en tant qu'oxydes de grille. Figure (a) issue de Peacock et Robertson [42]. Figure (b) : issue de Wilk, Wallace et Anthony [9].

On peut alors augmenter l'épaisseur et par conséquent réduire le courant de grille, tout en gardant un C_{ox} élevé. On introduit alors la notion épaisseur équivalente d'oxyde (EOT).

I-5.5 Epaisseur d'Oxyde Equivalent (EOT – Equivalent Oxide Thickness):

I-5.5.1 Cas d'un isolant de grille mono-couches et notion d'EOT[39] :

Le principe est donc de remplacer le SiO_2 par un matériau de plus grande permittivité. Cela permet du point de vue électrostatique, d'obtenir une épaisseur (EOT) équivalente (ou plus faible) à l'épaisseur de l'oxyde SiO_2 alors que l'épaisseur physique « T_{High-k} » du matériau high-k est plus grande. A même capacité et avec un T_{High-k} plus grand, le courant de fuite par effet tunnel sera donc fortement réduit. L'épaisseur physique d'oxyde équivalente est donnée par :

$$C_{SiO2} = C_{High-k} \Leftrightarrow \frac{\varepsilon_{SiO2}}{t_{SiO2}} = \frac{\varepsilon_{High-k}}{t_{High-k}} \Leftrightarrow t_{SiO2} = EOT = \frac{\varepsilon_{ox}}{\varepsilon_{High-k}} t_{High-k} \qquad \mathbf{I-8}$$

Avec C_{SiO2} et C_{High-k} respectivement les capacités d'oxyde de grille avec du SiO_2 et du high-k, ε_{ox} la permittivité relative du diélectrique de référence, ε_{High-k} la permittivité du matériau high-k et t_{High-k}, l'épaisseur physique du diélectrique high-k.

La figure I-12 illustre l'intégration des diélectriques high-k.

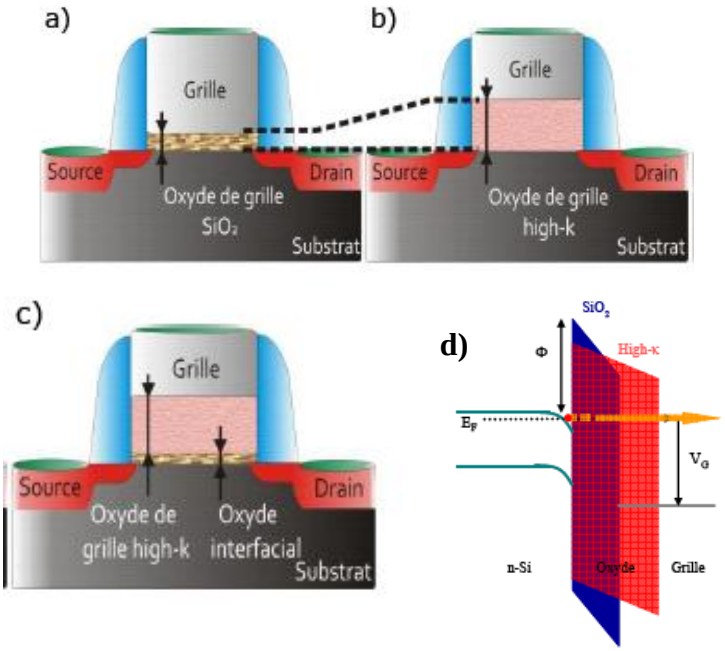


Figure I-12 : Illustration de l'intégration d'un diélectrique high-k dans une structure MOSFET permettant de comparer les différents EOT à capacités équivalentes : a) Oxyde de référence SiO_2 ; b) Intégration d'un oxyde high-k ; c) Structure réelle d**)** Cette différence d'épaisseur permet de limiter les courants de fuite par effet tunnel à travers la grille. [39,40].

Prenons un exemple pour illustrer l'EOT. Un matériau « high-k », avec un constant diélectrique de 25 est choisi. Il est déposé sur le silicium pour une épaisseur de 5 nm. Cela correspond, grâce à l'utilisation de l'équation [I-8], à une EOT de 0,78 nm. La figure I-13 résume l'intérêt de l'emploi de l'EOT.

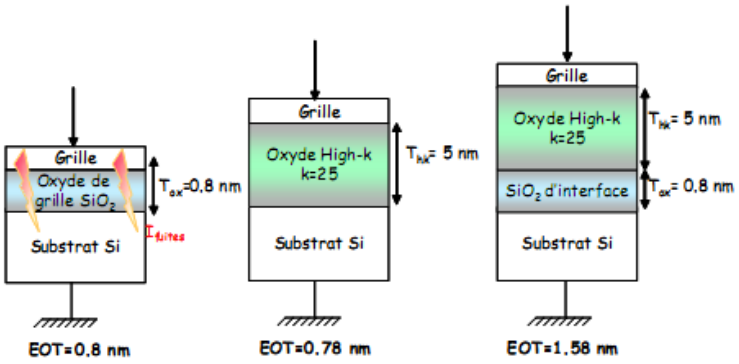


Figure I-13 : Illustration de l'EOT, valeur permettant de comparer l'épaisseur électrique de l'oxyde diélectrique avec l'oxyde de référence SiO_2.

I-5.5.2 Cas d'un isolant de grille multicouches et notion d'EOT :

Lorsque nous avons évoqué la structure MOS idéale, nous avons considéré que l'isolant de grille n'était constitué que d'une seule couche d'oxyde. Cependant, les raisons technologiques ont obligé l'industrie microélectronique à remplacer cette couche unique de SiO_2 par un empilement à deux couches, constitué d'un oxyde inter-facial (SiO_x ou SiON) et d'une couche d'oxyde de forte permittivité dit high-k (HfO_2 ou ZrO_2….)[41].

En figure I-12.c et figure I-13 le dernier cas, la structure MOSFET à diélectrique high-k tient compte d'une couche inter-faciale (SiOx). Cette couche est souvent volontairement déposée avant le dépôt du diélectrique high-k car dans le cas contraire, elle peut être créée lors du dépôt par la diffusion d'oxygène à travers le matériau ou lors d'une réaction thermodynamique. Dans ce cas, l'épaisseur ainsi que la qualité d'interface Si/SiOx ne pourront pas être contrôlées. En considérant une couche inter-faciale, la relation [I-8] se réécrit [39]:

$$EOT = \frac{\varepsilon_{ox}}{\varepsilon_{High-k}} t_{High-k} + \frac{\varepsilon_{ox}}{\varepsilon_{inter}} t_{inter} \qquad \mathbf{I-9}$$

Considérons maintenant un isolant de grille constitué de deux couches superposées d'oxyde. Cet empilement peut être vu comme l'association en série de deux capacités, l'une liée à l'oxyde inter-facial C_{inter} et l'autre liée au matériau de forte permittivité C_{high-k} Avec les mêmes conventions que dans l'équation (I-8), la capacité surfacique totale s'écrit [41]:

$$\frac{1}{C_{tot}} = \frac{1}{C_{inter}} + \frac{1}{C_{High-k}} = \frac{t_{inter}}{\varepsilon_{inter}\,\varepsilon_0} + \frac{t_{High-k}}{\varepsilon_{High-k}\,\varepsilon_0} \qquad \mathbf{I-10}$$

Nous pouvons alors exprimer la capacité totale :

$$C_{tot} = \frac{\varepsilon_{ox}\,\varepsilon_0}{EOT} \qquad \mathbf{I-11}$$

Lorsque l'isolant de grille est constitué d'un oxyde unique, cette couche présente une capacité par unité de surface valant : avec ε_{ox} la permittivité diélectrique relative de l'oxyde (3.9 pour le SiO_2) et t_{ox} l'épaisseur physique d'oxyde.

I-5.6 Quelques exemples de matériaux « high-k » :

Actuellement, la recherche se porte plutôt vers des matériaux comme Al_2O_3, ZrO_2 et HfO_2 qui présentent l'avantage d'être thermodynamiquement stables sur le silicium. On cherche à obtenir avec ces matériaux une qualité d'interface équivalente à celle du SiO_2 La recherche sur les High-k se base essentiellement sur la caractérisation physique et électrique de ces films minces, dans le but de mieux comprendre leurs propriétés physiques (stabilité thermique, réactivité avec le substrat...) et leur microstructure, afin de les corréler avec leurs propriétés électriques.

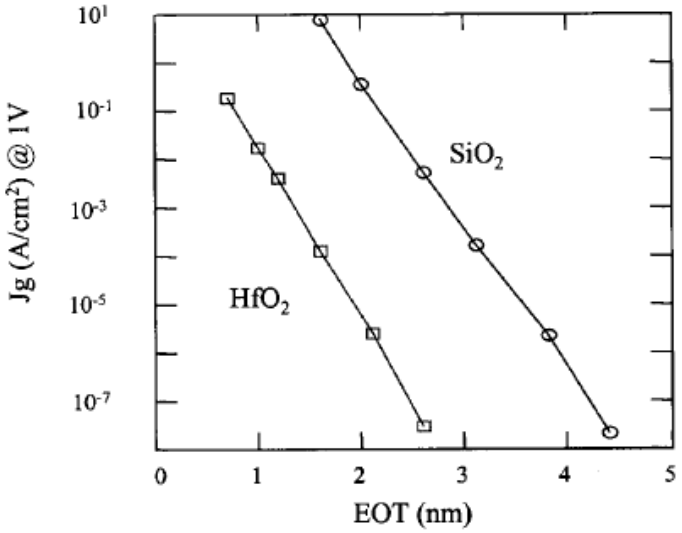


Figure I-14: Comparaison de la Densité de courant de fuite en fonction de l'épaisseur équivalente d'oxyde entre le HfO_2 et le SiO_2. D'après Lee et Al [36]

I-5.6.1 Dioxyde de zirconium ZrO_2 :

La zircone est un matériau bien connu de la communauté scientifique travaillant sur les matériaux oxydes, du fait de ses nombreuses propriétés remarquables (haute constante diélectrique, grande conductivité ionique, fort indice de réfraction, grandes propriétés réfractaires, de grandes propriétés mécaniques…) qui lui offrent de larges applications dans des domaines très divers.

Leur constant diélectrique théorique est évalué entre 20 et 25 [20-43], avec une énergie de bande interdite entre 5,1 eV et 7,8 eV [44]. Les phases cristallines du ZrO_2 (phase monoclinique, quadratique et cubique), dépendent des conditions de température et de pression.

Naturellement, la largeur de la bande interdite évolue aussi en fonction de la cristallographie. Le tableau de la figure I-3 rapporte les valeurs calculées par la théorie de la fonctionnelle de la densité (DFT) [45]. Une forte anisotropie de la permittivité a été montrée: la valeur donnée est une moyenne.

phase	ε_r	E_g(eV)
Cubic	37	2.63
Tétragonal	38	3.31
Monoclinique	20	2.98

Tableau I-3: permittivité relative et largeur de bande interdite calculées pour les trois phases ZrO2 [45].

I-5.6.1.1 Les différentes phases du ZrO_2 :

La zircone existe à pression atmosphérique sous trois variétés polymorphiques, la phase monoclinique, tétragonal et cubique. La forme naturelle de la zircone (baddeleyte), stable à faible température est cristallisée dans le système monoclinique. A 1170°C est se transforme en phase tétragonal et devient cubique à partir de 2370°C. Cette dernière forme est stable jusqu'à sa température de fusion qui est de 2680°C. La phase cubique est de type fluorine (groupe d'espace Fm3m) et présente un ion Zr^{4+} au centre d'un cube parfait de huit anions oxyde, que l'on peut considérer comme la somme de deux tétraèdres réguliers identiques. La structure de la variété quadratique (groupe d'espace $P4_2/nmc$) dérive de celle de la phase cubique et est obtenue par une distorsion du réseau de la fluorine, l'atome de zirconium est toujours en coordinence 8 mais le cube anionique est cette fois déformé. Enfin, la structure de la variété monoclinique correspond également à une déformation de la structure fluorine et admet le groupe d'espace $P2_1/c$. Cette fois, l'atome de zirconium prend une coordinence de 7. Une représentation schématique de ces différentes structures est donnée en figure I-15.

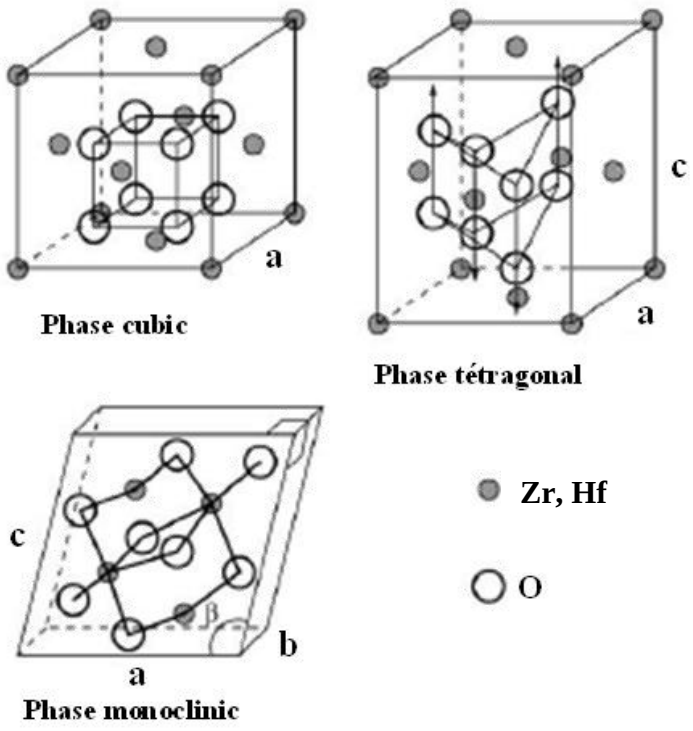


Figure I-15 : Les trois phases cristallines de ZrO_2 et de HfO_2 [46].

I-5.6.2 Dioxyde d'hafnium HfO_2 :

Le HfO_2 est un matériau largement étudié depuis ces dernières années dans le milieu de la microélectronique, en vue de son intégration dans les générations technologiques futures. Ainsi, un grand nombre de publications ont été et sont encore éditées. L'intérêt du HfO_2 provient de sa compatibilité avec la technologie du silicium, comme sa stabilité thermique, mais aussi de sa haute permittivité (constant diélectrique théorique), variant entre 17 et 25 d'après la littérature [20].

La densité théorique du HfO_2 est de 9,68 g/cm^3 mais lors de son application en couche mince une densité plus faible peut être obtenue. Par exemple, les études de Ferrari indiquent une densité de 9,17 g/cm^3 [47].

Le HfO_2 présente une énergie de bande interdite théorique Eg de 5,62 eV. Cependant, dans la littérature, différentes valeurs de Eg sont évaluées et semblent différer selon la structure dans laquelle se trouve le matériau. Ainsi des valeurs variant de 4,19 eV pour les films amorphes, à 5,65 eV pour les films cristallins ont été estimées [48-49-50]. Pour les films cristallisés, la taille des grains, influencent aussi la valeur de Eg obtenue [51].

Ceci est confirmé par les valeurs calculées par DFT [45-52]. L'anisotropie constatée sur ZrO_2 se retrouve ici, les valeurs données étant des moyennes (Tableau I-4).

phase	ε_r	E_g(eV)
Cubic	29	3.15
Tétragonal	70	3.84
Monoclinique	16 à 18	3.45

Tableau I-4 : permittivité relative et largeur de bande interdite calculées pour les trois phases HfO_2 [45].

De plus, la valeur moyenne de la largeur de sa bande interdite est relativement grande, de l'ordre de 3,48 eV, et enfin, HfO_2 présente une meilleure stabilité thermodynamique sur le silicium que ZrO_2. Par ailleurs, il cristallise à des températures plus élevées que ce dernier.

HfO_2 semble être le meilleur candidat pour remplacer l'oxyde de silicium.

I-5.6.2.1 Les différentes phases du HfO_2 :

Le dioxyde d'hafnium, aussi connu sous le nom de hafnone, ressemble à la zircone (ZrO_2) de part ses propriétés physiques et chimiques. La similitude structurale entre ces deux oxydes s'explique par la similitude entre les rayons ioniques de Hf et de Zr (i.e. les rayons ioniques Zr^{+4} et Hf^{+4} de 0.78 et 0.79 Å respectivement) [53-54].

Le seul oxyde d'hafnium stable, HfO_2, existe à pression atmosphérique sous trois polymorphes dont les structures sont de type fluorite plus ou moins déformée (figure I-15). La phase la plus stable est la phase monoclinique, présente à basse température (groupe spatial $P2_1/c$). Au-dessus de 2000 °C la phase quadratique (tétragonal) ($P4_2/nmc$) apparaît et à partir de 2700 °C se forme la phase fluorite cubique ($Fm3m$).

Monoclinique (*$P2_1/c$*) ⇄ **Quadratique (*$P4_1/nmc$*)** ↔ **Cubique (*$Fm3m$*)** ↔ **Liquide**

1510-2000°C (chauffage) 2700°C

L'ajout d'un oxyde sous forme d'alliage dans HfO_2 est une méthode commune pour stabiliser une phase (monoclinique, tétragonal ou cubique) du HfO_2. Autrement dit, si la phase initiale est la phase tétragonal, il est possible de la stabiliser par du dopage [50].

Le Tableau I-5 regroupe les groupes d'espace et les paramètres de maille des trois phases de HfO_2 d'après [53-55].

	HfO_2 **cubique**	***HfO_2*** **Quadratique**	***HfO_2*** **monoclinique**
Groupe spatial	Fm3m	$P4_2/nmc$	$P2_1/c$
Paramètres de maille (Å)	a=b=c=5.3	a=b=3.659 c=5.325	a=5.117 b=5.175 c=5.291 β=99.22°

Tableau I-5 : Données cristallographiques de l'oxyde d'hafnium à pression atmosphérique

I-5.7 Modèle Continue "modèle EKV" :

Dans certains cas, tels que pour la simulation de circuit analogique, il est nécessaire de disposer d'un modèle qui est continue dans tout régime de fonctionnement (faible ou forte inversion). Pour résoudre le problème, Le modèle EKV [56] est un modèle mathématique de transistors à effet de champ métal-oxyde-semiconducteur (MOSFET) est destinée à la simulation et la conception des circuits analogiques. La figure suivante représente les caractéristiques *log* I_D $-V_{GS}$ d'un MOSFET standard.

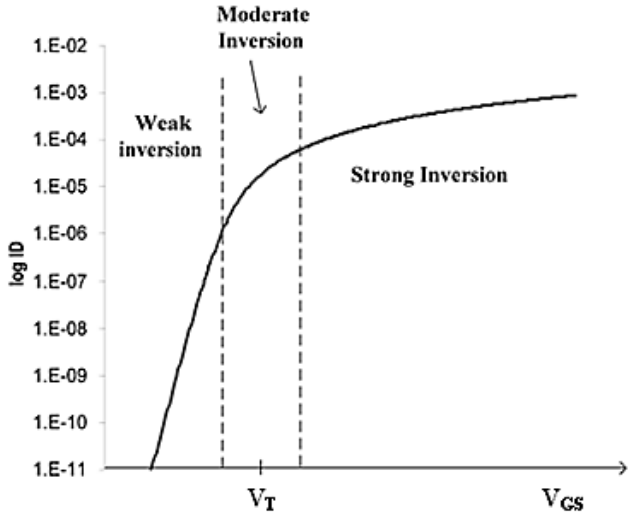


Figure I-16 : Discontinuité des caractéristiques à $V_{GS} \cong V_T$.

EKV définit le courant de drain en tant que combinaison d'un courant direct commandé par la source, et un courant de retour commandé par le drain. Suite les travaux de Enz, Krummenacher et Vittoz [57-58] (la soi-disant «modèle EKV"), on peut réécrire le courant de drain, comme suit:

$$i = q_F^2 + q_F \Rightarrow i = \frac{I_D}{I_S} \Rightarrow I_D = i.I_S \qquad \mathbf{I-12}$$

$$I_S = 2nU_T^2\mu_n C_{ox}\frac{W}{L} \qquad \mathbf{I-13}$$

Avec $U_T = (KT/q)$

La tension de pincement V_P est un nombre positif défini comme la valeur du potentiel de canal pour lequel la charge d'inversion est nulle dans un état de non-équilibre. V_P dépend de la tension V_G de grille et représente la tension appliquée au canal pour équilibrer l'effet du V_G. V_P est liée à V_G et V_T par:

$$\frac{V_P - V}{U_T} = 2(q_F - 1) + log(q_F) \qquad \mathbf{I-14}$$

$$\Rightarrow q_F = invq_F\left(\frac{V_P - V}{U_T}\right) \Rightarrow V_P = \frac{(V_{GS} - V_T)}{n} \qquad \mathbf{I-15}$$

$$V_P = \frac{(V_{GS} - V_T)}{n} \Rightarrow V_{GS} = n.V_P + V_T \qquad \mathbf{I-16}$$

La figure suivante représente la densité de charge d'inversion par rapport la tension de canal.

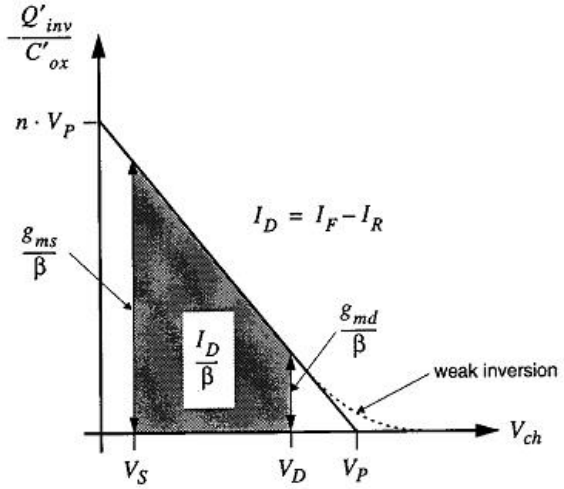


Figure I-17 : la densité de charge d'inversion vs. la tension de canal.

$$I_D = \overbrace{\beta.\int_{V_S}^{\infty}\left[-\frac{Q'_{inv}(V_{ch})}{C_{ox}}\right].dV_{ch}}^{=I_F\ forward\ current} - \underbrace{\beta.\int_{V_D}^{\infty}\left[-\frac{Q'_{inv}(V_{ch})}{C_{ox}}\right].dV_{ch}}_{=I_F\ reverse\ current} \qquad \mathbf{I-17}$$

$$I_D = I_F - I_R \qquad \mathbf{I-18}$$

$$I_D = 2n\mu_n C_{ox}\frac{W}{L}\left(\frac{KT}{q}\right)^2\left[\left\{ln\left[1+exp\left(\frac{V_P - V_S}{\frac{2KT}{q}}\right)\right]\right\}^2 - \left\{ln\left[1+exp\left(\frac{V_P - V_{DS}}{\frac{2KT}{q}}\right)\right]\right\}^2\right] \qquad \mathbf{I-19}$$

D'autre part, $V_S = 0$, $V_{DS} < V_P$ et $V_{GS} > V_T$ (par exemple le transistor fonctionne en régime de non-saturé). Dans ce cas, les termes exponentiels sont beaucoup plus grands que l'unité, et l'on peut écrire:

$$I_D = 2n\mu_n C_{ox}\frac{W}{L}\left(\frac{KT}{q}\right)^2\left[\left(\frac{V_P}{\frac{2KT}{q}}\right)^2 - \left(\frac{V_P - V_{DS}}{\frac{2KT}{q}}\right)^2\right] \qquad \mathbf{I-20}$$

$$I_D = \frac{1}{2}n\mu_n C_{ox}\frac{W}{L}[2V_{DS}V_P - V_{DS}^2]$$

$$I_D = \frac{1}{2}n\mu_n C_{ox}\frac{W}{L}\left[2\frac{(V_{GS}-V_T)V_{DS}}{n} - V_{DS}^2\right]$$

$$I_D = \mu_n C_{ox}\frac{W}{L}\left[(V_{GS}-V_T)V_{DS} - \frac{1}{2}nV_{DS}^2\right] \qquad \mathbf{I-21}$$

Avec
i: courant de drain normalisé
I_S: courant spécifique
I_F: courant direct normalisé
I_R: courant inverse normalisé
V_P: tension de pincement
U_T: la tension thermique

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 [59].

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 l'épaisseur de l'oxyde et 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 qui détériorent les caractéristiques courant-tension.

Dans notre cas, nous avons intéressé aux nouveaux diélectriques à constante élevée, appelés 'high-k diélectrics'. Ces isolants permettent de réduire significativement les courants de fuite et de progresser dans la miniaturisation.

Après avoir quelques exemples de matériaux « high-k », nous prenons le cas de dioxyde d'hafnium HfO_2 pour remplacer le SiO_2 comme oxyde de grille.

Le dioxyde d'hafnium HfO_2 a été beaucoup étudié en vue du remplacement de la silice dans les technologies CMOS. En effet, sa grande constante diélectrique et son gap relativement grand en font un oxyde high-k prometteur dans la diminution des dimensions des MOS silicium.

Chapitre II

Les Réseaux de Neurones et Les Algorithmes Génétiques

II-1 Introduction :

L'utilisation des réseaux de neurones (RN) et les algorithmes génétiques (AG) pour la modélisation de systèmes complexes a connu un essor important au cours de ces dernières années. Dans ce chapitre, nous allons décrire ces outils ainsi que les modalités de leur utilisation. Le chapitre est organisé en deux parties. La première partie fait l'objet d'une étude détaillée sur l'emploi des réseaux de neurones. Après une brève présentation de quelques notions générales sur les réseaux de neurones, nous présentons les grandes familles de structures neuronales les plus utilisées. Nous abordons également le problème d'apprentissage des paramètres de ces structures, nous nous intéressons particulièrement à l'algorithme de rétro propagation.

La deuxième partie est consacrée aux algorithmes génétiques. La première sous-section décrit le fonctionnement général des algorithmes génétiques. La seconde est consacrée aux différents aspects d'un problème qui doivent être formulés pour appliquer les algorithmes génétiques, les opérateurs génétiques sont présentés dans la troisième sous-section. La dernière sous-section décrit la manière dont les algorithmes génétiques gèrent les solutions dans les problèmes multi-objectifs.

II-2 Réseaux de Neurones :

II-2.1 Introduction :

La reconnaissance du fait que le cerveau fonctionne de manière entièrement différente de celle d'un ordinateur conventionnel a joué un rôle très important dans le développement des réseaux de neurones artificiels. Les travaux effectués pour essayer de comprendre le comportement du cerveau humain ont menés à représenter celui-ci par un ensemble de composants structurels appelés neurones, massivement interconnectés entre eux. Le cerveau humain en contiendrait plusieurs centaines de milliards, et chacun de ceux-ci serait, en moyenne, connecté à dix mille autres. Le cerveau est capable d'organiser ces neurones, selon un assemblage complexe, non-linéaire et extrêmement parallèle, de manière à pouvoir accomplir des tâches très élaborées. Par exemple, n'importe qui est capable de reconnaître des visages, alors que c'est là une tâche quasiment impossible pour un ordinateur classique. C'est la tentative de donner à l'ordinateur les qualités de perception du cerveau humain qui a conduit à une modélisation électrique de celui-ci. C'est cette modélisation que tentent de réaliser les réseaux de neurones artificiels [1].

Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau.

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 semi-conducteur sont limitées [2]. 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-2.2 Historique :

En 1943, deux biophysiciens de l'université de Chicago, Mac Culloch et Pitts proposent le premier modèle de neurone biologique [3]. Ce neurone formel, aussi appelé neurone à seuil, est inspiré des récentes découvertes en biologie. Ce sont des neurones logiques (0 ou 1).

En 1949, le psychologue Donald Hebb introduit le terme connexionnisme pour parler de modèles massivement parallèles et connectés [4]. Il propose de nombreuses règles de mise à jour des poids dont la célèbre " règle de Hebb ".

En 1958, le psychologue Frank Rosenblatt, combinant les idées de ses prédécesseurs, propose le premier perceptron [5]. Ce réseau, capable d'apprendre à différencier des formes simples et à calculer certaines fonctions logiques, est inspiré du système visuel.

Au début des années 60, les travaux de Rosenblatt suscitent un vif enthousiasme dans le milieu scientifique. Mais en 1969, deux scientifiques américains de renom, Minsky et Papert, publient un livre [6] qui démontre les limites du perceptron proposé par Rosenblatt. En particulier, son incapacité à résoudre les problèmes non linéairement séparables, dont la fonction logique XOR est un célèbre exemple.

Les travaux ralentissent considérablement jusqu'aux années 80. En 1982, Hopfield démontre l'intérêt des réseaux entièrement connectés [7]. Parallèlement, Werbos conçoit un mécanisme d'apprentissage pour les réseaux multicouches de type perceptron: la rétropropagation (Back-Propagation). Cet algorithme, qui permet de propager l'erreur vers les couches cachées sera popularisé en 1986 dans un livre " Parallel Distributed Processing " par Rumelhart *et* al [8].

Depuis ces travaux, les applications des réseaux de neurones n'ont cessé de croître. Il a d'ailleurs été démontré qu'un réseau MLP (Multi Layer perceptron) avec seulement deux couches peut approximer n'importe quelle fonction de R^n dans R^m avec une précision arbitraire.

II-2.3 Le modèle neurophysiologique (neurone biologique) :

Le neurone biologique est composé de quatre parties distinctes (figure II-1) :

- ***le corps cellulaire***, qui contient le noyau de la cellule nerveuse; c'est en cet endroit que prend naissance l'influx nerveux, qui représente l'état d'activité du neurone;

- ***les dendrites***, ramifications tubulaires courtes formant une espèce d'arborescence autour du corps cellulaire; ce sont les entrées principales du neurone, qui captent l'information venant d'autres neurones;

- ***l'axone***, longue fibre nerveuse qui se ramifie à son extrémité; c'est la sortie du neurone et le support de l'information vers les autres neurones;

- ***la synapse***, qui communique l'information, en la pondérant par un poids synaptique, à un autre neurone; elle est essentielle dans le fonctionnement du système nerveux.

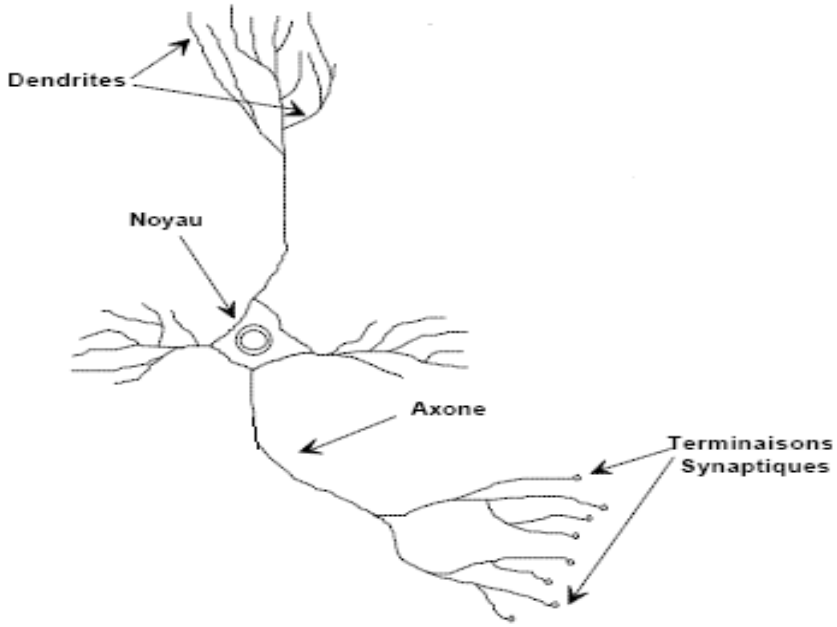


Figure II-1 : Un neurone avec son arborisation dendritique [1].

II-2.4 Les modèles mathématiques (neurone artificiel) :

La figure II-2 montre la structure d'un neurone artificiel. Chaque neurone artificiel est un processeur élémentaire. Il reçoit un nombre variable d'entrées en provenance de neurones amont. A chacune de ces entrées est associé un poids *w* abréviation de weight (poids en anglais) représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie ensuite pour alimenter un nombre variable de neurones avals. A chaque connexion est associé un poids.

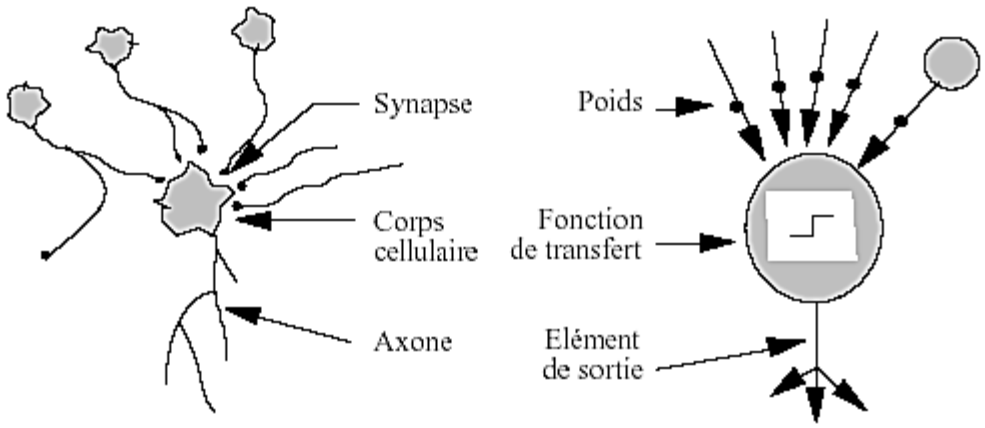


Figure II-2 : Mise en correspondance neurone biologique / neurone artificiel

II-2.4.1 Modèle de base de neurone artificiel: le neurone formel :

Un neurone est considéré comme un dispositif qui reçoit, à partir d'autres neurones ou de l'extérieur, des stimulations par des entrées (inputs), au nombre de n, et les pondère grâce à des valeurs réelles appelées coefficients synaptiques ou poids synaptiques. Ces coefficients peuvent être positifs, et l'on parle alors de synapses excitatrices, ou négatifs pour des synapses inhibitrices. Un neurone j calcule ainsi, un potentiel P_j, égal à la somme de ses entrées (inputs) ($x_1, x_2, \ldots, x_n$) pondérées, par les coefficients synaptiques respectifs ($w_1, w_2, \ldots, w_n$), à laquelle on ajoute un terme constant : le biais b_j. La valeur du potentiel P_j est donnée par l'équation suivante:

$$P_j = \sum_{i=1}^{n} w_{ij}\, x_i + b_j = W'X + b_j \qquad \mathbf{II-1}$$

A ce potentiel, le neurone applique une fonction d'activation *f*, de manière à ce que la sortie y_j, calculée par le neurone, soit égale à $f(P_j)$, tel que :

$$y_j = f(P_j) \qquad \mathbf{II-2}$$

La valeur de sortie y_j (output) est émise par le neurone vers d'autres neurones ou vers l'extérieur. Ainsi, un neurone est caractérisé par trois concepts : son état interne qui est son potentiel, ses connexions avec d'autres neurones et sa fonction de transfert.

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

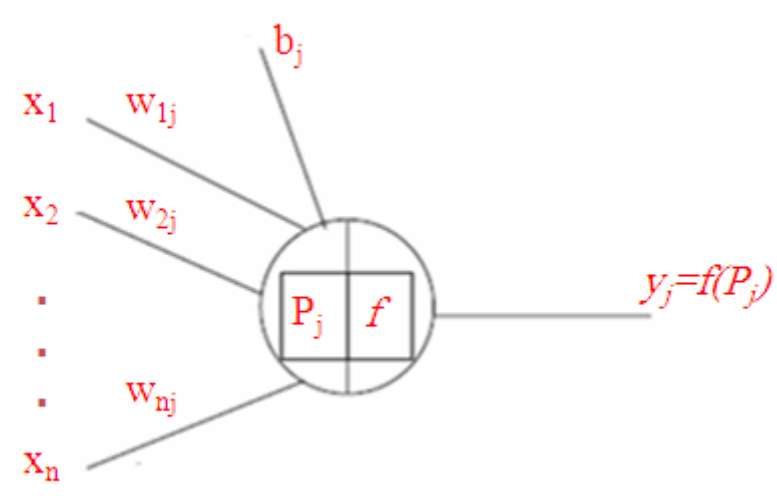


Figure II-3 : Modèle général d'un neurone

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

- Pondération : multiplication de chaque entrée par un paramètre appelé poids de connexion (poids synaptique).
- Sommation : une sommation des entrées pondérées est effectuée
- Activation : passage de cette somme dans une fonction, appelée fonction d'activation.

La valeur calculée est la sortie du neurone qui est transmise aux neurones suivants.

Plusieurs types de fonctions d'activation peuvent être utilisés, les plus courants sont donnés sur la figure II-4

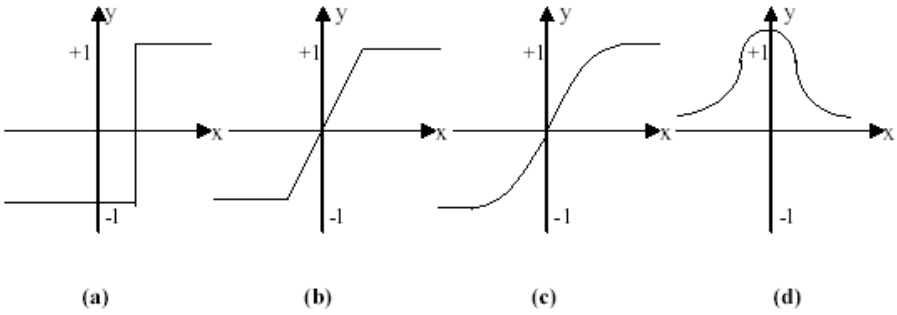


Figure II-4 : Différents types de fonctions de transfert pour le neurone artificiel. a) fonction à seuil du neurone de Mc Culloch et W. Pitts (1949), b) linéaire par morceaux du modèle Adaline de Widrow et Hoff (1960), c) sigmoïde d'un réseau perceptron Multi Couches de Rosenblatt (1962), d) gaussienne du réseau RFR de Moody et Darken (1989) [9].

II-2.4.1.1 Définition [10] :

Un réseau de neurones peut être considéré comme un modèle mathématique de traitement réparti, composé de plusieurs éléments de calcul non linéaire (neurones), opérant en parallèle et connectés entre eux par des poids.

Les neurones artificiels sont souvent utilisés sous forme de réseaux qui diffèrent selon le type de connections entre les neurones, Il existe de nombreux types de réseaux neuronaux, par exemples nous citons : les réseaux de Hopfield, le perceptron de Rosemblat.

Ces derniers sont les plus utilisés, ils sont constitués d'une succession de couches dont chacune prend ses entrées sur les sorties de la précédente. Les neurones de deux couches adjacentes sont interconnectés par des poids.

L'information dans le réseau se propage d'une couche à l'autre, on dit qu'ils sont de type « Retropropagation ». Nous distinguons trois types de couches : Couche d'entrée : les neurones de cette couche reçoivent les valeurs d'entrée du réseau et les transmettent aux neurones cachés. Chaque neurone reçoit une valeur, il ne fait pas donc de sommation.

Couches cachées : chaque neurone de cette couche reçoit l'information de plusieurs couches précédentes, effectue la sommation pondérée par les poids, puis la transforme selon sa fonction d'activation qui est en général une fonction sigmoïde. Par la suite, il envoie cette réponse aux neurones de la couche suivante.

Couche de sortie : elle joue le même rôle que les couches cachées, la seule différence entre ces deux types de couches est que la sortie des neurones de la couche de sortie n'est liée à aucun autre neurone.

II-2.5 Architecture des réseaux de neurones :

On distingue deux structures de réseau, en fonction du graphe de leurs connexions, c'est-à-dire du graphe dont les nœuds sont les neurones et les arêtes les «connexions» entre ceux-ci :

- Les réseaux de neurones statiques (ou acycliques, ou non bouclés).
- Les réseaux de neurones dynamiques (ou récurrents, ou bouclés).

II-2.5.1 Les réseaux de neurones non bouclés (Réseaux proactifs) :

Les études biologiques montrent que la structure des réseaux de neurones réels est très complexe. Il a été montré que le cortex est divisé en différents couches, connectées entre elles, et que les neurones d'une même couches sont également connectés entre eux.

L'architecture des réseaux de neurones mathématiques est inspirée de la réalité tout en étant simplificatrice, avec une structure régulière facilitant l'utilisation. Elle peut aller d'une connectivité totale (réseaux entièrement interconnectés) à une connectivité locale (réseaux à couches, les neurones ne sont connectés qu'avec leurs voisins). Les réseaux à couches, dont la structure simple est suffisante à la résolution des problèmes posés.

La structure des réseaux à couches est schématisée sur la Figure II-5. Les neurones d'une même couche ne sont pas connectés ; chaque neurone d'une couche envoie une information à tous les neurones de la couche suivante. Les deux couches extrêmes correspondent à la réception des données extérieures d'une part, et à la production du résultat du traitement d'autre part. Les couches intermédiaires sont appelées les couches cachées, leur nombre est variable.

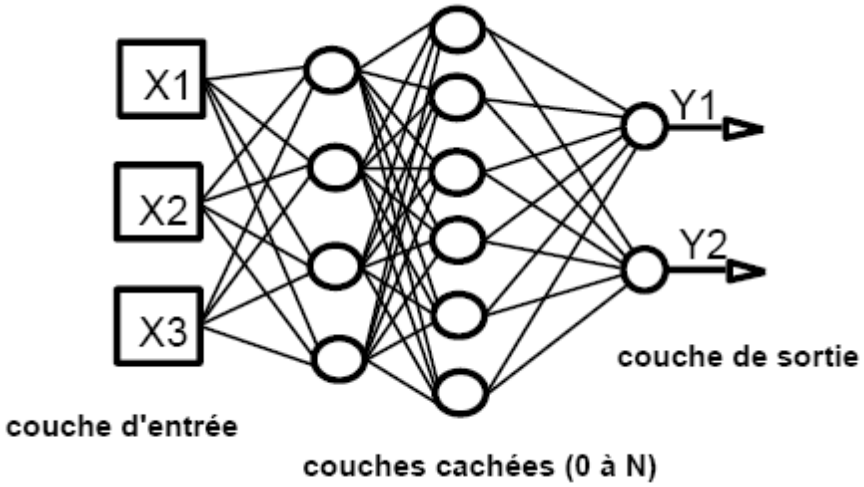


Figure II-5 : Structure d'un réseau.

II-2.5.1.1 Perceptron :

Présenté originellement par Rosenblatt, en 1958, le perceptron est la forme la plus simple de réseau de neurones, Il consiste en un seul neurone qui possède un seuil ainsi qu'un vecteur de poids synaptiques ajustables, tout comme le modèle de neurone de McCulloch & Pitts.

Le perceptron est limité dans la résolution des problèmes et certaines situations comme les problèmes de classifications non linéaires, ne peuvent être résolus. Une solution consiste à fournir au réseau la possibilité de se reformuler le problème avec une représentation interne propre, à partir d'une structure avec couches cachées (perceptron multicouche (MLP)).

II-2.5.1.1.1 Perceptron multicouche (MLP) :

La mise en cascade de perceptron conduit à ce qu'on appelle le perceptron multicouches "multilayer feedforward networks" (figure II-6). Les perceptrons employés ici diffèrent cependant de celui de Rosenblatt, par le fait que la non-linéarité utilisée est à présent une fonction continue, d'allure sigmoïdale par exemple, et non

plus la fonction de signe. Lorsque le vecteur de caractéristiques d'un objet est présenté à l'entrée du réseau, il est communiqué à tous les neurones de la première couche. Les sorties des neurones de cette couche sont alors communiquées aux neurones de la couche suivante, et ainsi de suite. La dernière couche du réseau est appelée *couche de sortie*, les autres étant désignées sous le terme de *couches cachées* car les valeurs de sortie de leurs neurones ne sont pas accessibles de l'extérieur [1].

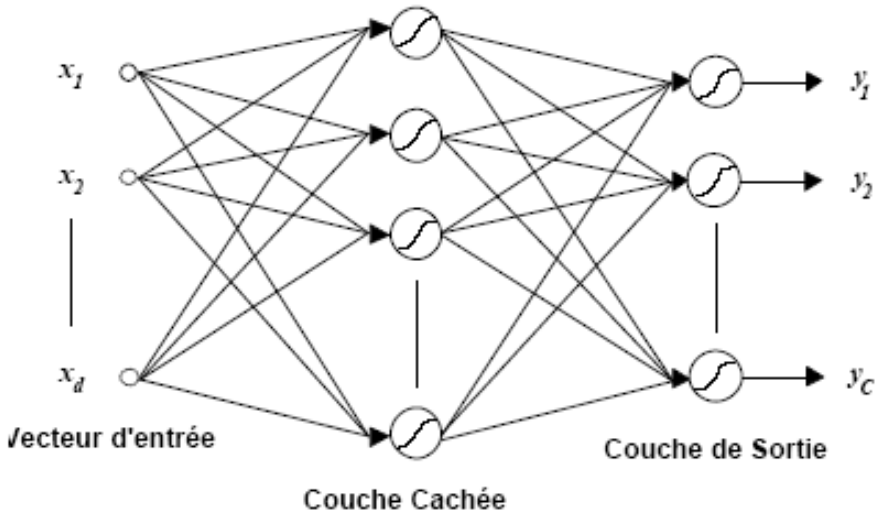


Figure II-6 : Perceptron Multicouches à une couche cachée [1].

II-2.5.2 Les réseaux de neurones bouclés (Réseaux récurrents) :

L'architecture la plus générale pour un réseau de neurones est le « réseau bouclé », dont le graphe des connexions est cyclique : lorsqu'on se déplace dans le réseau en suivant le sens des connexions, il est possible de trouver au moins un chemin qui revient à son point de départ (un tel chemin est désigné sous le terme de « cycle »). La sortie d'un neurone du réseau peut donc être fonction d'elle-même; cela n'est évidemment concevable que si la notion de temps est explicitement prise en considération.

Ainsi, à chaque connexion d'un réseau de neurones bouclé (ou à chaque arête de son graphe) est attaché, outre un poids comme pour les réseaux non bouclés, un retard, multiple entier (éventuellement nul) de l'unité de temps choisie. Une grandeur, à un instant donné, ne pouvant pas être fonction de sa propre valeur au même instant, tout cycle du graphe du réseau doit avoir un retard non nul.

Les connexions récurrentes ramènent l'information en arrière par rapport au sens de propagation défini dans un réseau multicouche. Ces connexions sont le plus souvent locales.

Pour éliminer le problème de la détermination de l'état du réseau par bouclage, on introduit sur chaque connexion « en retour » un retard qui permet de conserver le mode de fonctionnement séquentiel du réseau (figure II-7).

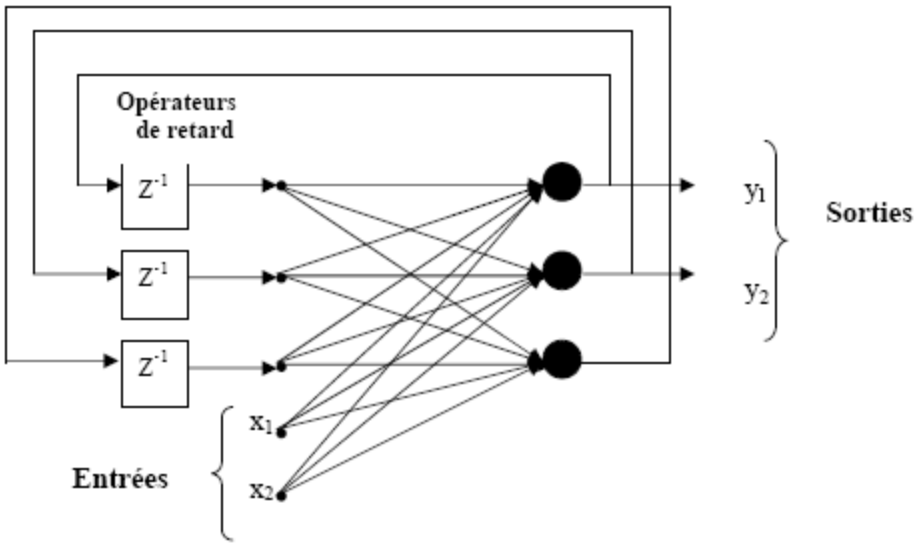


Figure II-7 : Réseau de neurone bouclé.

II-2.6 Apprentissage des réseaux de neurones :

Le point crucial du développement d'un réseau de neurones est son apprentissage. Il s'agit d'une procédure adaptative par laquelle les connexions des neurones sont ajustées face à une source d'information [4-8].

Dans le cas des réseaux de neurones artificiels, on ajoute souvent à la description du modèle l'algorithme d'apprentissage. Le modèle sans apprentissage présente en effet peu d'intérêt.

Dans la majorité des algorithmes actuels, les variables modifiées pendant l'apprentissage sont les poids des connexions. L'apprentissage est la modification des poids du réseau dans l'optique d'accorder la réponse du réseau aux exemples et à l'expérience. Les poids sont initialisés avec des valeurs aléatoires. Puis des exemples expérimentaux représentatifs du fonctionnement du procédé dans un domaine donné, sont présentés au réseau de neurones. Ces exemples sont constitués de couples expérimentaux de vecteurs d'entrée et de sortie. Une méthode d'optimisation modifie les poids au fur et à mesure des itérations pendant lesquelles on présente la totalité des exemples, afin de minimiser l'écart entre les sorties calculées et les sorties expérimentales.

II-2.6.1 Type d'apprentissage :

Il existe de nombreux types de règles d'apprentissage qui peuvent être regroupées en trois catégories: les règles d'apprentissage supervisé, non supervisé, et renforcé. Mais l'objectif fondamental de l'apprentissage reste le même : soit la classification, l'approximation de fonction ou encore la prévision.

II-2.6.1.1 Apprentissage supervisé :

Un apprentissage est dit supervisé lorsque l'on force le réseau à converger vers un état final précis, en même temps qu'on lui présente un motif. Ce genre d'apprentissage est réalisé à l'aide d'une base d'apprentissage, constituée de plusieurs exemples de type entrées-sorties (les entrées du réseau et les sorties désirées ou encore les solutions souhaitées pour l'ensemble des sorties du réseau).

La procédure usuelle dans le cadre de la prévision est l'apprentissage supervisé (ou à partir d'exemples) qui consiste à associer une réponse spécifique désirée à chaque signal d'entrée. La modification des poids s'effectue progressivement jusqu'à ce que l'erreur (ou l'écart) entre les sorties du réseau (ou résultats calculés) et les résultats désirés soient minimisés.

Cet apprentissage n'est possible que si un large jeu de données est disponible et si les solutions sont connues pour les exemples de la base d'apprentissage.

II-2.6.1.2 Apprentissage renforcé :

L'apprentissage renforcé est une technique similaire à l'apprentissage supervisé à la différence qu'au lieu de fournir des résultats désirés au réseau, on lui accorde plutôt un grade (ou score) qui est une mesure du degré de performance du réseau après quelques itérations. Les algorithmes utilisant la procédure d'apprentissage renforcé sont surtout utilisés dans le domaine des systèmes de contrôle [11].

II-2.6.1.3 Apprentissage non supervisé :

L'apprentissage non supervisé consiste à ajuster les poids à partir d'un seul ensemble d'apprentissage formé uniquement de données. Aucun résultat désiré n'est fourni au réseau.

Qu'est-ce que le réseau apprend exactement dans ce cas ? L'apprentissage consiste à détecter les similarités et les différences dans l'ensemble d'apprentissage. Les poids et les sorties du réseau convergent, en théorie, vers les représentations qui capturent les régularités statistiques des données. Ce type d'apprentissage est également dit compétitif et (ou) coopératif. L'avantage de ce type d'apprentissage réside dans sa grande capacité d'adaptation reconnue comme une auto organisation, « self-organizing» [12]. L'apprentissage non supervisé est surtout utilisé pour le traitement du signal et l'analyse factorielle.

II-2.6.2 Règles d'apprentissage :

Un réseau de neurones artificiel, comme le cerveau animal, apprendre à réagir correctement à un stimulus provenant de l'extérieur. Le principe de l'apprentissage consiste à soumettre le réseau à un stimulus dont on connaît la réponse souhaitée, autant de fois qu'il lui est nécessaire à la modification des poids des connexions, jusqu'à obtention de la bonne réponse. Il existe plusieurs règles de modification des poids, les principales sont :

- La règle de Hebb
- La règle de Widrow-Hoff
- L'algorithme de rétropropagation du gradient de l'erreur

II-2.6.2.1 Algorithme de rétropropagation du gradient :

II-2.6.2.1.1 Introduction :

La rétropropagation est actuellement l'outil le plus utilisé dans le domaine de réseaux de neurones. C'est une technique de calcul des dérivées qui peut être appliquée à n'importe qu'elle structure de fonctions dérivables.

Mathématiquement, cette méthode est basée sur l'algorithme de descente du gradient et utilise les règles de dérivation des fonctions dérivables. Dans cette méthode, l'erreur commise en sortie du réseau sera rétropropagée vers les couches cachées d'où le nom de rétropropagation.

II-2.6.2.1.2 Equation du réseau :

Avant de définir la règle d'apprentissage, on doit définir la relation entre les sorties du réseau d'une part, et les entrées et les poids d'autre part.

Dans un réseau à (L) couches ayant (n) entrées et (m) sorties les états des différents neurones.

$$u_i^l(k) = f^l\left(P_i^l(k)\right) \qquad \mathbf{II-3}$$

$$P_i^l(k) = \sum_{j=1}^{N_{l-1}} W_{ij}^L \, u_j^{l-1}(k) \qquad \mathbf{II-4}$$

Ou i 1,2,..... N_L
et j 1,2,..... $N_{L\text{-}1}$
et q 1,2,..... N_{L+1}

N nombre de neurones dans la couche L.
$N_{L\text{-}1}$ nombre de neurones dans la couche L -1.
N_{L+1} nombre de neurones dans la couche L +1.
L nombre de couches.

$u_i^l(k)$ Sortie du neurone i de la couche L.
$W_{ij}^L(k)$ Coefficient synaptique (poids) de la j^{eme} entrée du neurone (i) de la couche (l).

$$u_i^0(k) = x_i(k) \qquad \text{i } 1,2,\ldots\ldots \text{n} \qquad \mathbf{II-5}$$

$$u_i^l(k) = y_i(k) \qquad \text{i } 1,2,\ldots\ldots \text{m} \qquad \mathbf{II-6}$$

Ou : $x_i\ (k)$ et $y_i\ (k)$ sont respectivement les entrées et les sorties du réseau.

L'objectif de la méthode de la rétropropagation est d'adapter les paramètres W_{ij}^{L} de façon a minimisé une fonction de coût donné par :

$$E(W) = \sum_{p=1}^{T} E_P(W) \qquad \mathbf{II-7}$$

$$E_P(W) = \frac{1}{2}\sum_{i=1}^{m}\left[y_i^d(k) - y_i(k)\right]^2 \qquad \mathbf{II-8}$$

Ou $y^d(k)$ est le vecteur de sortie désiré, $y(k)$ le vecteur de sortie de réseau et T le nombre d'exemples ou longueur de l'ensemble d'entraînement.

II-2.6.2.1.3 Principe de la rétropropagation :

L'approche la plus utilisée pour la minimisation de la fonction E est basée sur la méthode du gradient. On commence l'entraînement par un choix aléatoire des vecteurs initiaux du poids.

On présente le premier vecteur d'entrée, une fois on a la sortie du réseau, l'erreur correspondante et le gradient de l'erreur par rapport à tous les poids sont calculés. Les poids sont alors ajustés. On refait la même procédure pour tous les exemples d'apprentissage. Ce processus est répété jusqu'à ce que les sorties du réseau soient suffisamment proches des sorties désirées.

II-2.6.2.1.4 Adaptation des poids :

L'adaptation des poids se fait par la méthode du gradient basée sur la formule itérative suivante :

$$W_{ij}^{l}(k+1) = W_{ij}^{l}(k) - \Delta W_{ij}^{l} \qquad \mathbf{II-9}$$

$$\text{Avec} \quad \Delta W_{ij}^{l} = \eta \frac{\partial E(W)}{\partial W_{ij}^{l}(k)} \qquad \mathbf{II-10}$$

Où ***k*** : représente le numéro d'itération.
η est une constante appelée facteur ou pas d'apprentissage.

La vitesse de convergence dépend de la constante μ. Sa valeur est généralement choisie expérimentalement.

La dérivée de la fonction du coût par rapport au poids W_{ij}^{l} est données par :

$$\frac{\partial E(W)}{\partial W_{ij}^{l}(k)} = \sum_{P=1}^{T} \frac{\partial E_P(W)}{\partial W_{ij}^{l}(k)}$$
$$\frac{\partial E_P(W)}{\partial W_{ij}^{l}(k)} = \frac{\partial E_P(W)}{\partial \mu_i^{l}(k)} \cdot \frac{\partial \mu_i^{l}(k)}{\partial W_{ij}^{l}(k)} \qquad \mathbf{II-11}$$

Pour la couche de sortie :

$$\frac{\partial E_P(k)}{\partial \mu_i^l(k)} = -(y_i^d(k) - y_i(k)) \qquad \mathbf{II-12}$$

Pour les couches cachées :

$$\frac{\partial E_P(W)}{\partial \mu_i^l(k)} = \sum_{q=1}^{N_{l+1}} \frac{\partial E_P(W)}{\partial \mu_q^{l+1}(k)} . \frac{\partial \mu_q^{l+1}(k)}{\partial \mu_i^l(k)} \qquad \mathbf{II-13}$$

$$\frac{\partial \mu_i^l(k)}{\partial W_{ij}^l} = f^{l^*}\left(P_i^l(k)\right) \mu_j^{l-1}(k) \qquad \mathbf{II-14}$$

Donc l'expression (3-11) s'écrit sous la forme :

$$\frac{\partial E_P(W)}{\partial W_{ij}^l} = \frac{\partial E_P(W)}{\partial \mu_j^l(k)} f^{l^*}\left(P_i^l(k)\right) \mu_j^{l-1}(k) \qquad \mathbf{II-15}$$

Pour minimiser l'erreur totale sur l'ensemble d'entraînement, les poids du réseau doivent être ajustés après présentation de tous les exemples.

II-2.6.2.1.5 Algorithme de la rétropropagation :

Etape 1: Initialiser les poids W_{ij}^L et les seuils internes des neurones à des petites valeurs aléatoires.

Etape 2 : Calculer le vecteur d'entrée et de sortie désirée, correspondant.

Etape 3 : Calculer la sortie du réseau en utilisant les expressions (II-3) et (II-4)

Etape 4 : Calculer l'erreur de sortie en utilisant l'expression (II-12)

Etape 5 : Calculer l'erreur dans les couches en utilisant l'expression (II-13)

Etape 6 : Calculer le gradient de l'erreur par rapport aux poids en utilisant l'expression (II-10)

Etape 7 : Ajuster les poids selon l'expression (II-9).

Etape 8 : Si la condition sur l'erreur ou sur le nombre d'itérations est atteinte, aller à **l'étape 9,** sinon aller **à l'étape 2.**

Etape 9: Fin.

Les exemples sont présentés d'une manière récursive, lorsque tous les exemples sont présentés, le test s'effectue sur l'erreur de sortie et les poids sont ajustés au fur et à mesure, jusqu'à ce que l'erreur de sortie se stabilise à une valeur acceptable.

Le gradient local de l'erreur, pour chaque neurone, dépend de la dérivée de la fonction d'activation. Ainsi, cette fonction est supposée dérivable, donc, continue. Les fonctions d'activation non linéaires et différentiables, couramment utilisées dans les perceptrons multicouches, sont la fonction sigmoïde et la fonction tangente hyperbolique.

II-2.6.2.1.6 Accélération de la rétropropagation :

Bien que l'algorithme de rétropropagation soit le plus utilisé pour l'apprentissage supervisé des MLP, son implantation se heurte à plusieurs difficultés techniques. Il n'existe pas de méthodes permettant de :

- Trouver une architecture appropriée (nombres de couches, nombre de neurones).

- Choisir une taille et une qualité adéquate d'exemples d'entraînement.

- Choisir des valeurs initiales satisfaisantes pour les poids, et des valeurs convenables pour les paramètres d'apprentissage permettant d'accélérer la vitesse de convergence.

- Problème de la convergence vers un minimum local, qui empêche la convergence et cause l'oscillation de l'erreur.

Plusieurs approches ont été proposées pour remédier à ces problèmes.

Une des techniques d'accélération est celle de la création dynamique des neurones, un neurone est ajouté chaque fois que l'erreur se stabilise à un niveau inacceptable.

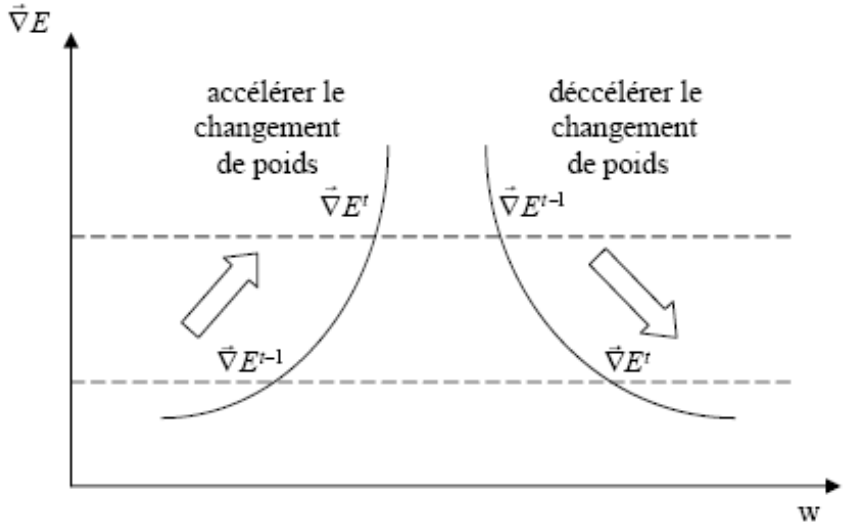


Figure II-8 : Mode de changement des poids dans la rétropropagation des poids dans la rétropropagation rapide.

II-2.7 Conception d'un réseau de neurones :

Les réseaux de neurones réalisent des fonctions non linéaires paramétrées. Leurs mises en œuvre nécessitent :

- La détermination des entrées et des sorties pertinentes, c'est à dire les grandeurs qui ont une influence significative sur le phénomène que l'on cherche à modéliser.

- La collecte des données nécessaires à l'apprentissage et à l'évaluation des performances du réseau de neurones.

- La détermination du nombre de neurones cachés nécessaires pour obtenir une approximation satisfaisante.

- La réalisation de l'apprentissage

- L'évaluation des performances du réseau de neurones à l'issue de l'apprentissage.

II-2.7.1 Détermination des entrées/sorties du réseau de neurones :

Pour toute conception de modèle, la sélection des entrées doit prendre en compte deux points essentiels :

- Premièrement, la dimension intrinsèque du vecteur des entrées doit être aussi petite que possible, en d'autre terme, la représentation des entrées doit être la plus compacte possible, tout en conservant pour l'essentiel la même quantité d'information, et en gardant à l'esprit que les différentes entrées doivent être indépendantes.

- En second lieu, toutes les informations présentées dans les entrées doivent être pertinentes pour la grandeur que l'on cherche à modéliser : elles doivent donc avoir une influence réelle sur la valeur de la sortie.

II-2.7.2 Choix et préparation des échantillons :

Le processus d'élaboration d'un réseau de neurones commence toujours par le choix et la préparation des échantillons de données. La façon dont se présente l'échantillon conditionne le type de réseau, le nombre de cellules d'entrée, le nombre de cellules de sortie et la façon dont il faudra mener l'apprentissage, les tests et la validation [13]. Il faut donc déterminer les grandeurs qui ont une influence significative sur le phénomène que l'on cherche à modéliser.

Lorsque la grandeur que l'on veut modéliser dépend de nombreux facteurs, c'est-à-dire lorsque le modèle possède de nombreuses entrées, il n'est pas possible de réaliser un « pavage » régulier dans tout le domaine de variation des entrées : il faut donc trouver une méthode permettant de réaliser uniquement des expériences qui apportent une information significative pour l'apprentissage du modèle. Cet objectif

peut être obtenu en mettant en œuvre un plan d'expériences. Pour les modèles linéaires, l'élaboration de plans d'expériences est bien maîtrisée, par ailleurs, ce n'est pas le cas pour les modèles non linéaires.

Afin de développer une application à base de réseaux de neurones, il est nécessaire de disposer de deux bases de données, une pour effectuer l'apprentissage et l'autre pour tester le réseau obtenu et déterminer ses performances.

Notons qu'il n'y a pas de règle pour déterminer ce partage d'une manière quantitative, néanmoins chaque base doit satisfaire aux contraintes de représentativité de chaque classe de données et doit généralement refléter la distribution réelle, c'est à dire la probabilité d'occurrence des diverses classes.

II-2.7.3 Elaboration de la structure du réseau :

La structure du réseau dépend étroitement du type des échantillons. Il faut d'abord choisir le type de réseau : un perceptron standard, un réseau de Hopfield, un réseau à décalage temporel (TDNN), un réseau de Kohonen, un ARTMAP etc...

Par exemple, dans le cas du perceptron multicouches, il faudra aussi bien choisir le nombre de couches cachées que le nombre de neurones dans cette couche.

- ***Nombre de couches cachées :***

Mis à part les couches d'entrée et de sortie, il faut décider du nombre de couches intermédiaires ou cachées. Sans couche cachée, le réseau n'offre que de faibles possibilités d'adaptation. Néanmoins, il a été démontré qu'un Perceptron Multicouches avec une seule couche cachée pourvue d'un nombre suffisant de neurones, peut approximer n'importe quelle fonction avec la précision souhaitée [14].

- ***Nombre de neurones cachés :***

Chaque neurone peut prendre en compte des profils spécifiques de neurones d'entrée.

Un nombre plus important permet donc de mieux ″coller″ aux données présentées mais diminue la capacité de généralisation du réseau. Il faut alors trouver le nombre adéquat de neurones cachés nécessaire pour obtenir une approximation satisfaisante.

Il n'existe pas, à ce jour, de résultat théorique permettant de prévoir le nombre de neurones cachés nécessaires pour obtenir une performance spécifique du modèle, compte tenu des modèles disponibles. Il faut donc nécessairement mettre en œuvre une procédure numérique de conception de modèle.

II-2.7.4 Apprentissage :

L'apprentissage est un problème numérique d'optimisation. Il consiste à calculer les pondérations optimales des différentes liaisons, en utilisant un échantillon. La méthode la plus utilisée est la rétropropagation, qui est généralement plus économe que les autres en termes de nombres d'opérations arithmétiques à effectuer pour évaluer le gradient.

II-2.7.5 Validation et Tests :

Alors que les tests concernent la vérification des performances d'un réseau de neurones hors échantillon et sa capacité de généralisation, la validation est parfois utilisée lors de l'apprentissage. Une fois le réseau de neurones développé, des tests s'imposent afin de vérifier la qualité des prévisions du modèle neuronal.

Cette dernière étape doit permettre d'estimer la qualité du réseau obtenu en lui présentant des exemples qui ne font pas partie de l'ensemble d'apprentissage. Une validation rigoureuse du modèle développé se traduit par une proportion importante de prédictions exactes sur l'ensemble de la validation.

Si les performances du réseau ne sont pas satisfaisantes, il faudra, soit modifier l'architecture du réseau, soit modifier la base d'apprentissage.

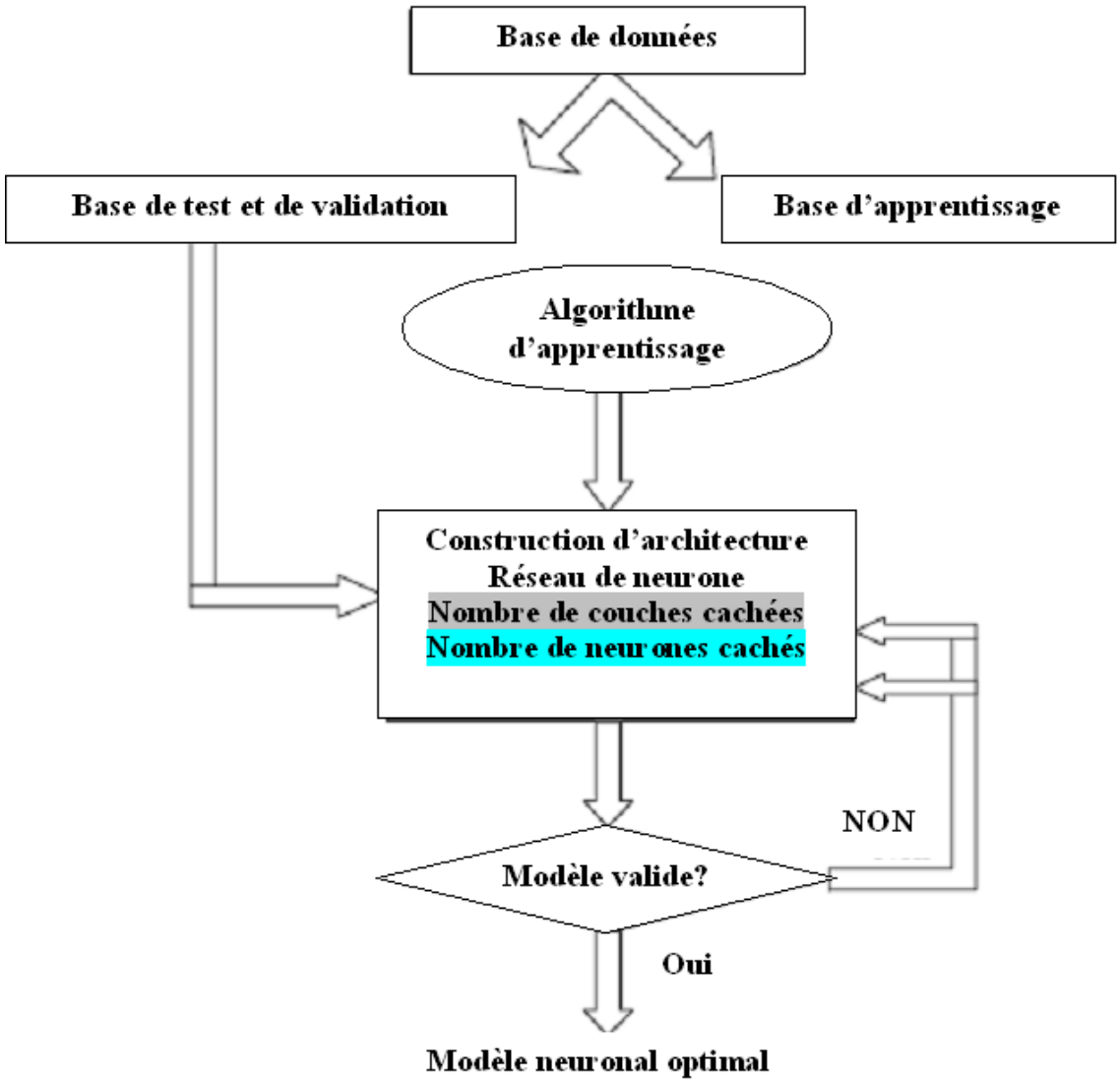


Figure II-9 : Organigramme de conception d'un réseau de neurones.

II-3 Algorithme Génétique :

II-3.1 Introduction :

Les algorithmes génétiques sont dans la famille des algorithmes métaheuristiques dont le but est d'obtenir une solution convenable dans un temps acceptable, et de concevoir des systèmes artificiels possédant des propriétés similaires aux systèmes naturels [15-16]. Ces algorithmes s'inspirent de l'évolution génétique des espèces, schématiquement, ils copient de façon extrêmement simplifiée certains comportements des populations naturelles. Ainsi, ces techniques reposent toutes sur l'évolution d'une population de solutions qui sous l'action de règles précises optimisent un comportement donné, exprimé sous forme d'une fonction, dite *fonction sélective* (*fitness function*) ou adaptation à l'environnement.

II-3.2 Généralité :

C'est en 1860 que Charles Darwin publie son livre intitulé '*L'origine des espèces au moyen de la sélection naturelle ou la lutte pour l'existence dans la nature'* [17]. Dans ce livre, Darwin rejeté l'existence «de systèmes naturels figés», déjà adaptés pour toujours à toutes les conditions extérieures, et expose sa théorie de l'évolution des espèces : sous l'influence des contraintes extérieurs, les êtres vivants se sont graduellement adaptés à leur milieu naturel au travers de processus de reproductions.

Darwin proposa une théorie qui clarifie l'évolution des espèces en mettant en avant quatre lois:

- La loi de croissance et de reproduction.
- La loi d'hérédité qu'implique quasiment la loi de reproduction
- La loi de variabilité, résultant des conditions d'existence.
- La loi de multiplication des espèces qui amène la lutte pour l'existence et qui a pour conséquence la sélection naturelle.

C'est alors à partir du 20ème siècle que la mutation génétique a été mise en évidence. Les problèmes de traitement de l'information sont résolus de manières figés : lors de sa phase de conception, le système reçoit toutes les caractéristiques nécessaires pour les conditions d'exploitations connues au moment de sa conception, ce qui empêche une adaptation à des conditions d'environnement inconnues, variables ou évolutives. Les chercheurs en informatique étudient donc des méthodes pour permettent aux systèmes d'évoluer spontanément en fonction de nouvelles conditions : c'est l'émergence de la programmation évolutionnaire *(Figure II-10).*

Dans les années 1960, John Holland étudie les systèmes évolutifs et, en 1975, il introduit le premier modèle formel des algorithmes génétiques (*the canonical genetic algorithm AGC*) dans son livre '*Adaptation in Natural and Artificial Systems'* [18]. Il expliqua comment ajouter de l'intelligence dans un programme informatique avec les croisements (échangeant le matériel génétique) et la mutation (source de la diversité génétique). Ce modèle servira de base aux recherches ultérieures et sera plus particulièrement repris par Goldberg qui publiera en 1989, un ouvrage de vulgarisation des algorithmes génétiques, et ajouta à la théorie des algorithmes génétiques les idées suivantes:

- un individu est lié à un environnement par son code d'ADN.
- une solution est liée à un problème par son indice de qualité.

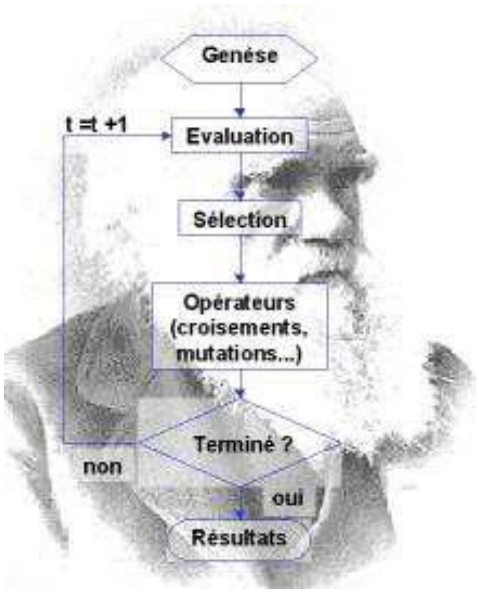


Figure II-10 : Organigramme d'un algorithme évolutionnaire.

La figure Ci-dessus est présenté l'organigramme d'un algorithme évolutionnaire. Il s'agit de simuler l'évolution d'une population d'individus divers (généralement tirée aléatoirement au départ) à laquelle on applique différents opérateurs (recombinaisons, mutations…) et que l'on soumet à une sélection, à chaque génération. Si la sélection s'opère à partir de la fonction d'adaptation, alors la population tend à s'améliorer [19]. Un tel algorithme ne nécessite aucune connaissance du problème : on peut représenter celui-ci par une boîte noire comportant des entrées (les variables) et des sorties (les fonctions objectif). L'algorithme ne fait que manipuler les entrées, lire les sorties, manipuler à nouveau les entrées de façon à améliorer les sorties, etc… [20]

Les algorithmes évolutionnaires constituent une approche originale : il ne s'agit pas de trouver une solution analytique exacte, ou une bonne approximation numérique, mais de trouver des solutions satisfaisant au mieux à différents critères, souvent contradictoires. S'ils ne permettent pas de trouver à coup sûr la solution optimale de l'espace de recherche, du moins peut-on constater que les solutions fournies sont généralement meilleures que celles obtenues par des méthodes plus classiques, pour un même temps de calcul.

Ils font parties du champ de la vie artificielle. La vie artificielle est l'étude des systèmes conçus par l'homme, qui présentent des comportements similaires aux systèmes vivants naturels. Elle complète l'approche traditionnelle de la biologie, définie étymologiquement par *étude des êtres vivants,* en essayant de synthétiser leurs comportements sur support artificiel. La modélisation, s'ajoutant à l'observation, à la théorie et à l'expérience, est un nouvel outil scientifique qui s'est fait valoir depuis l'avènement de l'informatique. Celle-ci peut contribuer à la biologie théorique en la plaçant dans un contexte plus vaste.

L'objectif est double: d'une part, la modélisation de ces phénomènes permet de mieux les comprendre, et ainsi mettre en évidence les mécanismes qui sont à l'origine de la vie ; d'autre part, on peut exploiter ces phénomènes de façon libre et peuvent donc être diverses.

Le domaine de l'évolution artificielle n'a connu une réelle expansion qu'à partir de ces 15 dernières années. Pourtant, l'idée de simuler sur ordinateurs des phénomènes évolutionnaires remonte aux années 50. Des concepts tels que la représentation des chromosomes par des chaînes binaires étaient déjà présents.

L'essor de l'évolution artificielle, depuis les années 80, peut s'expliquer par deux phénomènes concurrents. Premièrement, cet essor est principalement dû à l'accroissement exponentiel des moyens de calculs mis à la disposition des chercheurs, ce qui leur permet d'afficher des résultants expérimentaux pertinents et prometteurs. Le deuxième point est l'abandon du biologiquement Plausible.

Trois types d'algorithmes évolutionnaires ont été développé isolément et à peu près simultanément, par différents scientifiques : la programmation évolutionniste [21], les Stratégies d'évolution [22] et les Algorithmes Génétiques [18].

Dans les années 90, ces trois champs ont commencé à sortir de leur isolement et ont été regroupés sous le terme anglo-saxon d'Evolutionnary Computation.

II-3.3 Algorithmes génétiques :

Nous traiterons seulement ici les algorithmes génétiques fondés sur le Néo-Darwinisme, c'est-à-dire l'union de la théorie de l'évolution et de la génétique moderne. Ils s'appuient sur différentes techniques dérivées de cette dernière : croisements, mutation, sélection...

Un algorithme génétique recherche le ou les extrema d'une fonction définie sur un espace de données. Pour l'utiliser, on doit disposer des cinq éléments suivants :

1) Un principe de codage de l'élément de population. Cette étape associe à chacun des points de l'espace d'état une structure de données. Elle se place généralement après une phase de modélisation mathématique du problème traité. La qualité du codage des données conditionne le succès des algorithmes génétiques. Les codages binaires ont été très utilisés à l'origine. Les codages réels sont désormais largement utilisés, notamment dans les domaines applicatifs pour l'optimisation de problèmes à variables réelles.

2) Un mécanisme de génération de la population initiale. Ce mécanisme doit être capable de produire une population d'individus non homogène qui servira de base pour les générations futures. Le choix de la population initiale est important car il peut rendre plus ou moins rapide la convergence vers l'optimum global. Dans le cas où l'on ne connaît rien du problème à résoudre, il est essentiel que la population initiale soit répartie sur tout le domaine de recherche.

3) Une fonction à optimiser. Celle-ci retourne une valeur appelée fitness ou fonction d'évaluation de l'individu.

4) Des opérateurs permettant de diversifier la population au cours des générations et d'explorer l'espace d'état. L'opérateur de croisement recompose les gènes

d'individus existant dans la population, l'opérateur de mutation a pour but de garantir l'exploration de l'espace d'états.

5) Des paramètres de dimensionnement : taille de la population, nombre total de générations ou critère d'arrêt, probabilités d'application des opérateurs de croisement et de mutation.

II-3.3.1 Fonctionnement général des Algorithmes génétiques :

L'algorithme génétique commence avec une population initiale dont les individus sont évalués en fonction des objectifs et des contraintes du problème à résoudre. Cette population est ensuite diversifiée en utilisant des opérateurs génétiques (sélection, croisement, mutation) pour reproduire les individus. La reproduction est faite à plusieurs reprises jusqu'à ce que le critère d'arrêt (nombre d'itérations en générale) de l'algorithme soit atteint.

Le principe général de fonctionnement d'un algorithme génétique se présente suivant les étapes décrites ci-dessous et illustrée par la figure n°11.

Etape1 : Génération de la **population initiale** Pour passer d'une génération **k** à la génération **k+1**, les étapes suivantes sont répétées pour tous les individus de la génération **k**.

Etape2 : Evaluation de la population d'adaptation (fonction objective) ou « force » de chaque individu de la population.

Etape3 : Sélection des meilleurs individus en fonction de leurs forces.

Etape4 : Application stochastique des opérateurs génétiques :

*** Croisement :** Des couples de parents P1 et P2 sont formes de façon aléatoire et l'opérateur de leur croisement est appliqué avec une probabilité **Pcros** pour générer des enfants E1 et E2.

*** Mutation :** L'opérateur de mutation est appliqué avec une probabilité **Pmut** sur chaque individu.

Etape5 : Si le(s) **critère(s) d'arrêt** sont satisfaits, la génération **k** est considérée comme étant la solution (Population **finale)** et la procédure est arrêtée ; sinon, on retourne à l'étape2.

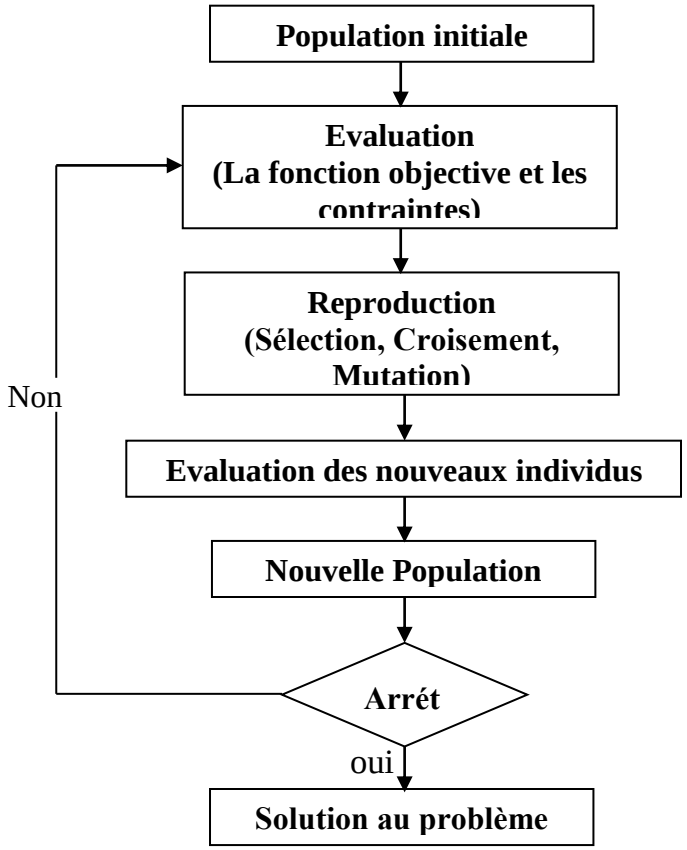


Figure II-11 : Fonctionnement des algorithmes génétiques

De manière plus formelle, Voice un algorithme génétiques de base:

Début

1: Générer une population aléatoire de n chromosomes.

2: Evaluer la fitness des chromosomes avec la fonction f (x)

3 : **Répéter**

4: Calculer la fonction fitness f (x), pour tout chromosome x

5: Appliquer l'opération de sélection

6: Appliquer l'opération de croisement avec une probabilité PC

7: Appliquer l'opération de mutation avec une probabilité PM

8: Ajouter les nouveaux chromosomes à la nouvelle population

9: Calculer la fonction fitness f (x), pour tout chromosome x

10: Appliquer l'opération de remplacement

11 : **Jusqu'à** la satisfaction des conditions de terminaison

Fin

II-3.3.2 Formulation du problème d'optimisation :

Les algorithmes génétiques travaillent sur la formulation d'un problème et non sur le problème lui-même. La formulation consiste en deux étapes essentielles : la représentation des individus et la formalisation des fonctions d'évaluations.

II-3.3.2.1 Codage des individus :

Une population est composée de chromosomes ou individus et chaque individu se caractérise par un ensemble de gènes. Chaque individu représente une solution possible au problème. Représenter un individu revient alors à définir ce qui caractérise une solution possible au problème. N'importe quel type de représentation peut être défini à condition qu'on s'assure de la définition des opérateurs génétiques pouvant traiter la représentation.

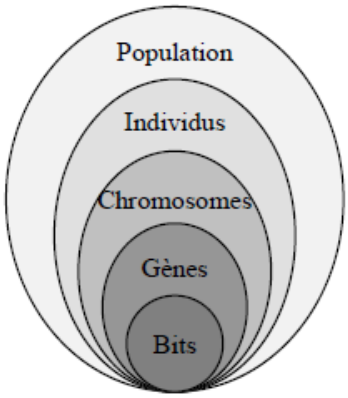
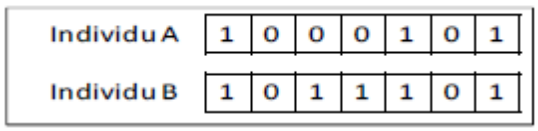


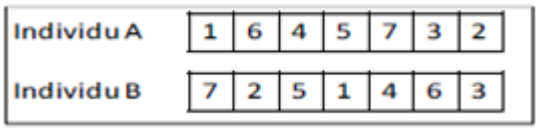
Figure II-12 : Les cinq niveaux d'organisation d'un algorithme génétique

Les représentations les plus courantes sont :

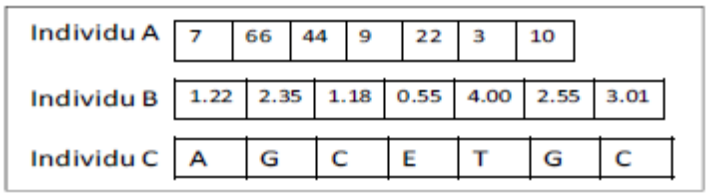
La représentation binaire : Il s'agit d'une suite de 0 et 1. Un exemple d'utilisation est dans le cas où chaque gène représente une caractéristique qui peut être présente (1) ou non (0) dans l'individu. La figure II.13.a illustre un exemple de deux individus codés en binaire.



a. Codage binaire



b. Codage par permutation



c. Codage par valeur

Figure II-13 : Codage des solutions

La permutation : Il s'agit de la permutation de valeurs entières égales au nombre de gènes dans les chromosomes. Ce codage est beaucoup utilisé pour les problèmes d'ordonnancement de ressources. Des exemples du codage par permutation sont présentés dans la figure II-13.b.

La représentation par valeurs : Dans ce codage on associe à chaque gène une valeur prise dans un ensemble fini ou infini. Ces valeurs dépendent du problème étudié. La figure II-13.c montre des exemples de codage en entier, réel, et caractère.

II-3.3.2.2 Formalisation des fonctions d'évaluation :

La formulation consiste en l'expression sous forme mathématique des objectifs d'optimisation de la manière la plus fidèle possible. Ces fonctions d'évaluation sont utilisées pour déterminer le degré de pertinence de chaque solution. L'évaluation de chaque solution est indépendante du reste de la population. Elle permet de s'assurer qu'on garde les individus les plus pertinents en éliminant les moins pertinents progressivement de la population.

Dans certains problèmes, les solutions doivent satisfaire des contraintes pour faire parties des solutions finales. Ces solutions sont représentées sous forme de fonctions de contraintes qui permettent de garantir la cohérence de chaque solution.

Le problème est alors formulé comme suit:

- $S = \{X_1, X_2, X_3, \ldots, X_n\}$ un ensemble d'individus représentant une population. X étant un ensemble de composants $X = \{x_1, x_2, x_3, \ldots, x_k\}$.

- L'objectif est de minimiser F (X)= {f_1 (X), f_2 (X),…, f_m(X)} pout tout X de S.
- Telles que les contraintes G (X)= {g_1 (X), g_2 (X),.. , g_m(X)} soient satisfaites.

II-3.3.3 Opérateurs génétiques :

Les algorithmes génétiques utilisent quatre opérateurs pour générer de nouvelles solutions :

(i) L'opérateur de sélection qui permet de choisir des solutions parentes sur lesquelles la reproduction va être faite pour générer des nouvelles solutions.

(ii) L'opérateur de croisement qui permet de croiser les deux solutions parentes et créer de nouvelles solutions.

(iii) L'opérateur de mutation qui permet de diversifier les nouvelles solutions afin qu'elles ne ressemblent pas trop aux solutions parentes. L'utilisation des coefficients de probabilistes permet de faire des bonnes diversifications en introduisant l'effet du hasard.

(iv) L'opérateur de remplacement, Cet opérateur est le plus simple, son travail consiste à réintroduire les descendants obtenus par application successive des opérateurs de sélection, de croisement et de mutation (la population P') dans la population de leurs parents (la population P).

II-3.3.3.1 L'opérateur de sélection :

La sélection consiste à choisir des individus qui permettront de générer de nouveaux individus. Plusieurs méthodes existent pour sélectionner des individus destinés à la reproduction.

On citera les deux méthodes classiques les plus utilisées.

La sélection par tournoi : Cette méthode consiste à choisir aléatoirement une paire d`individus dans la population. Le meilleur des deux individus sera sélectionné pour la reproduction avec une probabilité p supérieure à 0.5. Cette méthode est en général satisfaisante. La figure II-14 illustre ce type de sélection. Les paires d'individus (1,5) et (2,4) sont choisis, puis on sélectionne les individus 5 et 2 pour la reproduction.

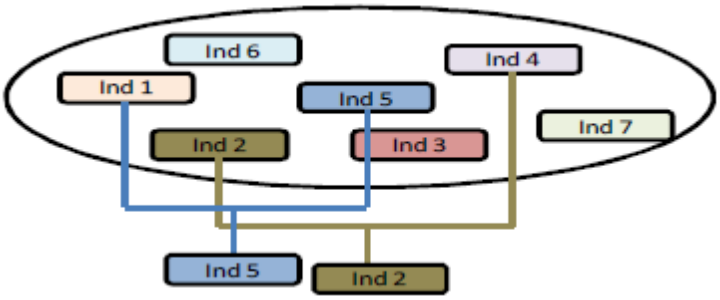


Figure II-14 : Sélection par tournoi.

La sélection par la roulette : Les individus de la population se voient allouer des longueurs proportionnelles à leurs performances. Ils sont représentés dans une roulette dont la surface totale représente les performances des individus et chaque individu est représenté par sa longueur sous forme de portion de roulette. Un nombre tiré aléatoirement permet de déterminer l'individu sélectionné selon la portion de roulette à laquelle il correspond. Cette méthode favorise les meilleurs individus ce qui n'est pas nécessairement souhaité puisque la reproduction avec de mauvais individus peut rapprocher de la solution optimale. La figure II-15 illustre une population de 7 individus dont les performances sont représentées en roulette.

Individu	Fonction Objective	Longueur
1	20	0.20
2	5	0.05
3	30	0.3
4	10	0.1
5	15	0.15
6	7	0.07
7	13	0.13

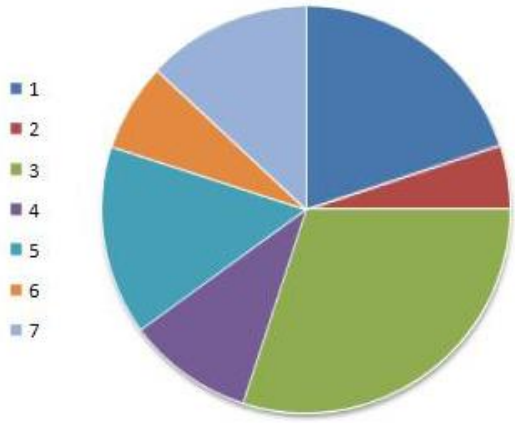


Figure II-15 : Sélection par roulette

II-3.3.3.2 L'opérateur de Croisement :

Le croisement consiste, à partir de deux parents P1 et P2 sélectionnés, à définir deux nouveaux individus appelés enfants E1 et E2. Ce qui permet de diversifier l'espace de solutions. Il existe plusieurs méthodes de croisement.

Le croisement à un point (slicing crossover) : Il a été initialement défini pour le codage binaire. Le principe consiste à tirer aléatoire une position pour chaque parent et à échanger les sous chaines des parents à partir des positions tirées. Ce qui donne naissance à deux nouveaux individus E1 et E2 Figure II-16.

Figure II-16 : Croisement à un point

Le croisement à deux-points (2-points crossover) : elle reprend le mécanisme de la méthode de croisement à un point en généralisant l'échange à 3 ou 4 sous chaines (Figure II-17).

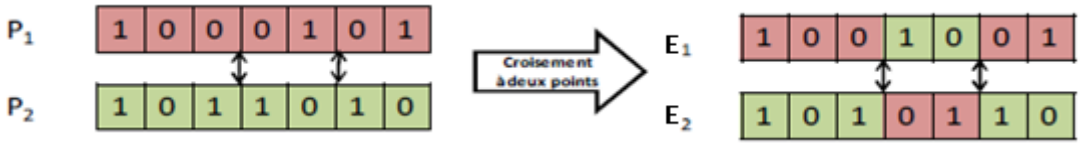


Figure II-17 : Croisement à deux-points

Le croisement Uniforme : On définit un « Masque » de manière aléatoire, de même longueur que les chromosomes parents. Pour un locus, si le locus du masque est 0 il hérite du parent 1, si 1 il hérite du parent 2, et de manière symétrique pour le deuxième fils.

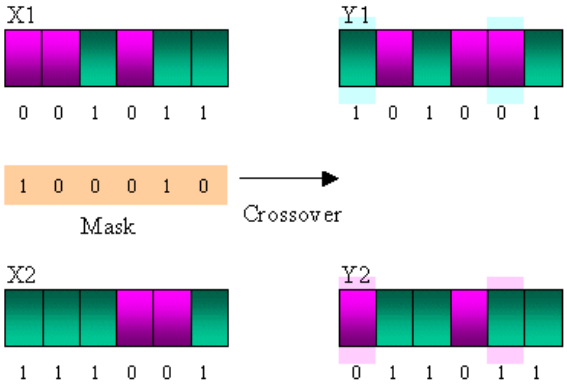


Figure II-18 : Croisement Uniforme

II-3.3.3.3 La mutation :

La mutation empêche l'algorithme de converger vers un extrema local en introduisant des modifications aléatoires à une petite proportion des individus. Cela augmente la diversité des solutions mais peut aussi détruire de bons individus. Si le taux de probabilité de mutation est trop élevé, l'algorithme dégénère en une recherche aléatoire. Cette caractéristique ne présente aucun intérêt dans notre cas.

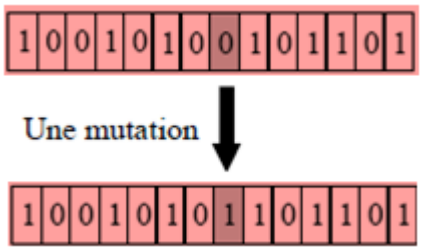


Figure II-19 : Principe de la mutation.

Après la mutation, l'algorithme reprend le cycle à partir de l'évaluation avec la nouvelle population produite.

Cet opérateur dispose de 4 grands avantages :

• Il garantit la diversité de la population, ce qui est primordial pour les algorithmes génétiques.

• Il permet d'éviter un phénomène connu sous le nom de ***dérive génétique***. On parle de dérive génétique quand certains gènes favorisés par le hasard se répandent au détriment des autres et sont ainsi présents au même endroit sur tous les chromosomes. Le fait que l'opérateur de mutation puisse entraîner de manière aléatoire des changements au niveau de n'importe quel locus permet d'éviter l'installation de cette situation défavorable.

• Il permet de limiter les risques d'une convergence prématurée causée par exemple par une méthode de sélection élitiste imposant à la population une pression sélective trop forte.

En effet, dans le cas d'une convergence prématurée on se retrouve avec une population dont tous les individus sont identiques mais ne sont que des optimums locaux. Tous les individus étant identiques, le croisement ne changera rien à la situation. En effet, l'échange d'informations par croisement entre des individus strictement identiques est bien sûr totalement sans conséquences; on aura beau choisir la méthode de croisement qu'on veut on se retrouvera toujours à échanger des portions de chromosomes identiques et la population n'évoluera pas. L'évolution se retrouvant bloquée on n'attendra jamais l'optimum global.

La mutation entrainant des inversions de bits de manière aléatoire permet de réintroduire des différences entre les individus et donc de nous extirper de cette situation. Il est quand même utile de garder à l'esprit que ceci n'est pas une solution "miracle" et qu'il est bien entendu plus intelligent de ne pas utiliser de méthodes de sélection connues pour entrainer ce type de problème.

• La mutation permet d'atteindre la propriété d'***ergodicité***.

L'ergodicité est une propriété garantissant que chaque point de l'espace de recherche puisse être atteint.

En effet, une mutation pouvant intervenir de manière aléatoire au niveau de n'importe quel locus, on a la certitude mathématique que n'importe quel permutation de notre chaîne de bits peut apparaître au sein de la population et donc que tout point de l'espace de recherche peut être atteint.

Grâce à cette propriété on est donc sûr de pouvoir atteindre l'optimum global.

On notera que la mutation règle donc le problème exposé à la fin du Section sur le croisement.

II-3.3.3.4 L'élitisme :

À chaque itération, des individus de la population peuvent être remplacés par de nouveaux individus. Lors de ce remplacement, il y' a de grandes chances que les meilleurs individus soient perdus. Pour éviter cela, l'élitisme permet à chaque itération de copier un ou plusieurs individus dans la nouvelle population avant de faire la reproduction.

II-3.3.3.5 Operateur de remplacement :

Cet opérateur est basé, comme l'operateur de sélection, sur la fitness des individus. Son travail consiste à déterminer les chromosomes parmi la population courante qui constituant la population de la génération suivante. Cette opération est appliquée après l'application successive des opérateurs de sélection, de croisement et de mutation. On trouve essentiellement 2 méthodes de remplacement différentes :

- **Le remplacement stationnaire :** dans ce cas, les enfants remplacent automatiquement les parents sans tenir compte de leurs performances respectives. Le nombre d'individus de la population est constant tout au long du cycle d'évolution, ce qui implique donc d'initialiser la population initiale avec un nombre suffisant d'individus.

- **Le remplacement élitiste :** dans ce cas, on garde au moins les individus possédant les meilleures performances d'une génération à la suivante. En général, on peut partir du principe qu'un nouvel individu (enfant) prend place au sein de la population que s'il remplit le critère d'être plus performant que le moins performant des individus de la population précédente. Donc les enfants d'une génération ne remplaceront pas nécessairement leurs parents comme dans le remplacement stationnaire et par conséquent la taille de la population n'est pas figée au cours du temps. Ce type de stratégie améliore les performances des algorithmes évolutionnaires dans certains cas. Mais présente aussi un désavantage en augmentant le taux de convergence prématuré.

II-3.3.4 Critères d'arrêt :

Le critère d'arrêt indique que la solution est suffisamment approchée de l'optimum. Plusieurs critères d'arrêt de l'algorithme génétique sont possibles. On peut arrêter l'algorithme après un nombre de générations suffisant pour que l'espace de recherche soit convenablement exploré. Ce critère peut s'avérer coûteux en temps de calcul si le nombre d'individus à traiter dans chaque population est important. L'algorithme peut aussi être arrêté lorsque la population n'évolue plus suffisamment rapidement.

On peut aussi envisager d'arrêter l'algorithme lorsque la fonction d'adaptation d'un individu dépasse un seuil fixé au départ. Nous pouvons également faire des combinaisons des critères d'arrêt précédents.

II-3.3.5 Les avantages et les limites des algorithmes génétiques :

Un des grands avantages des algorithmes génétiques est qu'ils autorisent la prise en compte de plusieurs critères simultanément, et qu'ils parviennent à trouver de bonnes solutions sur des problèmes très complexes. Le principal avantage des AG par rapport aux autres techniques d'optimisation combinatoire consiste en une combinaison de :

- l'exploration de l'espace de recherche, basée sur des paramètres aléatoires, grâce à une recherche parallèle,
- l'exploitation des meilleures solutions disponibles à un moment donné.

Ils doivent simplement déterminer entre deux solutions quelle est la meilleure, afin d'opérer leurs sélections.

L'inconvénient majeur des algorithmes génétiques est le coût d'exécution important par rapport à d'autres métaheuristiques. Les AG nécessitent de nombreux calculs, en particulier au niveau de la fonction d'évaluation. Mais avec les capacités calculatoire des ordinateurs récents, ce problème n'est pas grand. D'un autre côté, l'ajustement d'un algorithme génétique est délicat : des paramètres comme la taille de la population ou le taux de mutation sont parfois difficiles à déterminer. Or le succès de l'évolution en dépend et plusieurs essais sont donc nécessaires, ce qui limite encore l'efficacité de l'algorithme. Un autre problème important est celui des optima locaux. En effet, lorsqu'une population évolue, il se peut que certains individus qui à un instant occupent une place importante au sein de cette population deviennent majoritaires. À ce moment, il se peut que la population converge vers cet individu et s'écarte ainsi d'individus plus intéressants mais trop éloignés de l'individu vers lequel on converge. Il faut mentionner également le caractère indéterministe des AGs. Comme les opérateurs génétiques utilisent des facteurs aléatoires, un AG peut se comporter différemment pour des paramètres et population identiques. Afin d'évaluer correctement l'algorithme, il faut l'exécuter plusieurs fois et analyser statistiquement les résultats.

II-3.3.6 Gestion des solutions dans les Algorithmes Génétiques multiobjectifs :

Dans les problèmes multicritères, les objectifs sont en général contradictoires. Les solutions peuvent optimiser certains objectifs sans pour autant être bonnes sur d'autres objectifs. Selon les situations, il est possible de faire des compromis et de choisir des solutions avantageuses sur certains objectifs même si elles ne sont pas bonnes sur d'autres. Il est alors important de pouvoir fournir, comme solutions des problèmes multi-objectifs, un ensemble de choix possibles.

Les algorithmes génétiques sont très adaptés pour régler les problèmes multi-objectifs. Ils sont basés sur une approche évolutionnaire manipulant une population de solutions et faisant des explorations sur différentes régions de cette population. Il existe deux approches pour gérer les problèmes multi-objectifs : les approches Pareto et les approches non Pareto.

II-3.3.6.1 Les approches non Pareto :

Schaffer [23] a défini le premier algorithme génétique multi-objectif : VEGA (Vector Evaluated Genetic Algorithm). Il utilise un processus de recherche qui traite séparément les objectifs. Cette approche se fait en deux étapes. Dans la première, les objectifs sont traits séparément. Si la taille de la population est de *n* et qu'on a *k* objectifs, une sélection de *n/k* individus est effectuée pour chaque objectif. On obtiendra alors K sous-populations chacune contenant *n/k* meilleurs individus pour chaque objectif. Ces sous-populations sont regroupées en une nouvelle population sur laquelle vont se faire les opérations de croisement et de mutation.

Les approches non Pareto sont faciles à implémenter, mais donnent parfois des solutions extrêmes qui ne permettent pas d'avoir de bon compromis [24].

II-3.3.6.2 Les approches Pareto :

L'approche Pareto consiste à regrouper les solutions en un ensemble de solutions non dominées qui constituera le front de Pareto.

II-4 Conclusion :

Dans ce chapitre, nous avons introduit les définitions essentielles relatives aux réseaux de neurones et les algorithmes génétiques.

Nous avons mis l'accent sur l'utilisation des réseaux de neurones comme outils de modélisation par apprentissage. Ces derniers permettent d'ajuster des fonctions non linéaires très générales à des ensembles de points. Comme toute méthode qui s'appuie sur des techniques statistiques, l'utilisation de réseaux de neurones nécessite que l'on dispose de données suffisamment nombreuses et représentatives.

Nous avons aussi présenté les concepts fondamentaux de la modélisation à l'aide de réseaux de neurones et une méthodologie complète de conception et de mise en œuvre de modèles neuronaux.

Ensuite, nous avons présenté les concepts des algorithmes génétiques, leur fonctionnement, la formulation mathématique qui leur est associée, les opérateurs de sélection, de croisement et de mutation qui les caractérisent et enfin les différentes approches pour la gestion de solutions.

Chapitre III

Modélisation du transistor MOSFET à permittivité élevé en utilisant Les RNs et Les AGs

III-1 Introduction:

Les outils intelligents sont de plus en plus utilisés dans la conception, la modélisation et la commande de systèmes complexes. On entend par outils intelligents les techniques suivant : les réseaux de neurones et les algorithmes génétiques.

Dans ce contexte, notre travail consiste à étudier les transistors MOSFET à permittivité élevée (HfO_2) en utilisant les Réseaux de neurones et les Algorithmes génétiques.

III-2 Les Réseaux De Neurones :

Les neurones artificiels sont souvent utilisés sous forme de réseaux qui diffèrent selon le type de connections entre les neurones, une cinquantaine de types peut être dénombrée. En guise d'exemples nous citons : le perceptron de Rosemblatt [1], les réseaux de Hopfield [2] etc.…..

Ces derniers sont les plus utilisés et recommandé en science des matériaux à semiconducteurs (modélisation des composants électroniques). Ils sont constitués d'un nombre fini de neurones qui sont arrangés sous forme de couches. Les neurones de deux couches adjacentes sont interconnectés par des poids. L'information dans le réseau se propage d'une couche à l'autre, on dit qu'ils sont de type « feed-forward ».

Le but de notre travail est de mettre en place:

Des programmes de réseaux de neurones sous Matlab [3], qui utilise la rétropropagation du gradient pour le test du modèle ANN-MOSFET High-k et identifier différentes courbes.

III-2.1 Test et validation ANN-MOSFET High-k sous MATLAB:

III-2.1.1 Optimisation du prédicteur neuronal:

Après avoir choisie le type du réseau il faudra trouver une architecture optimale du réseau de neurones (MLP), nombre de couches cachées, nombre de neurones dans chaque couche, ainsi que la fonction d'activation pour chaque couche.

L'apprentissage et l'optimisation du réseau précédent sont accomplis par un programme structuré en MATLAB, pour améliorer l'apprentissage et pour obtenir une courbe plus proche du modèle, on joue alors sur différents paramètres tel le nombre de neurones de la couche cachée. Le réseau optimisé possède 4 entrées pour un isolant de grille monocouches et 5 entrées pour un isolant de grille multicouches, et un neurone dans la couche de sortie qui est déterminé par le nombre de sorties du système à modéliser, (le notre possède une seule sortie I_D), quant aux couches cachées et après optimisation, le réseau à deux couches de 10 neurones pour la première et 8 neurones pour la deuxième couche représente la structure choisie pour notre modélisation.

La fonction d'activation de la sortie est la fonction linéaire, pour les couches cachées , nous avons choisi la fonction sigmoïde (logsig).

Pour améliorer l'apprentissage et pour obtenir une courbe plus proche du modèle, on joue alors sur différents paramètres tels que le nombre de neurones de la couche cachée.

La figure III-1 représente l'évolution de l'erreur moyenne d'apprentissage de notre prédicteur (EQM de test= 0.0001et Nombre d'itérations= 4000).

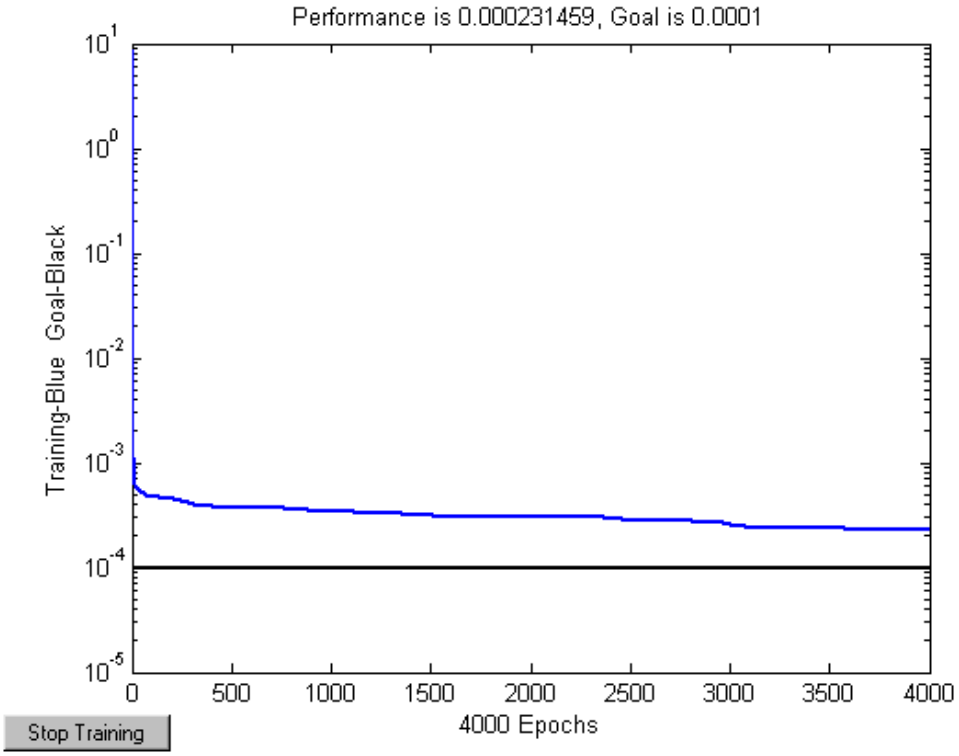


Figure III-1 : L'évolution de l'erreur moyenne d'apprentissage de notre prédicteur

La procédure utilisée dans cette étude pour l'optimisation du réseau de neurones est détaillée par l'organigramme présenté sur la figure III-2. 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.

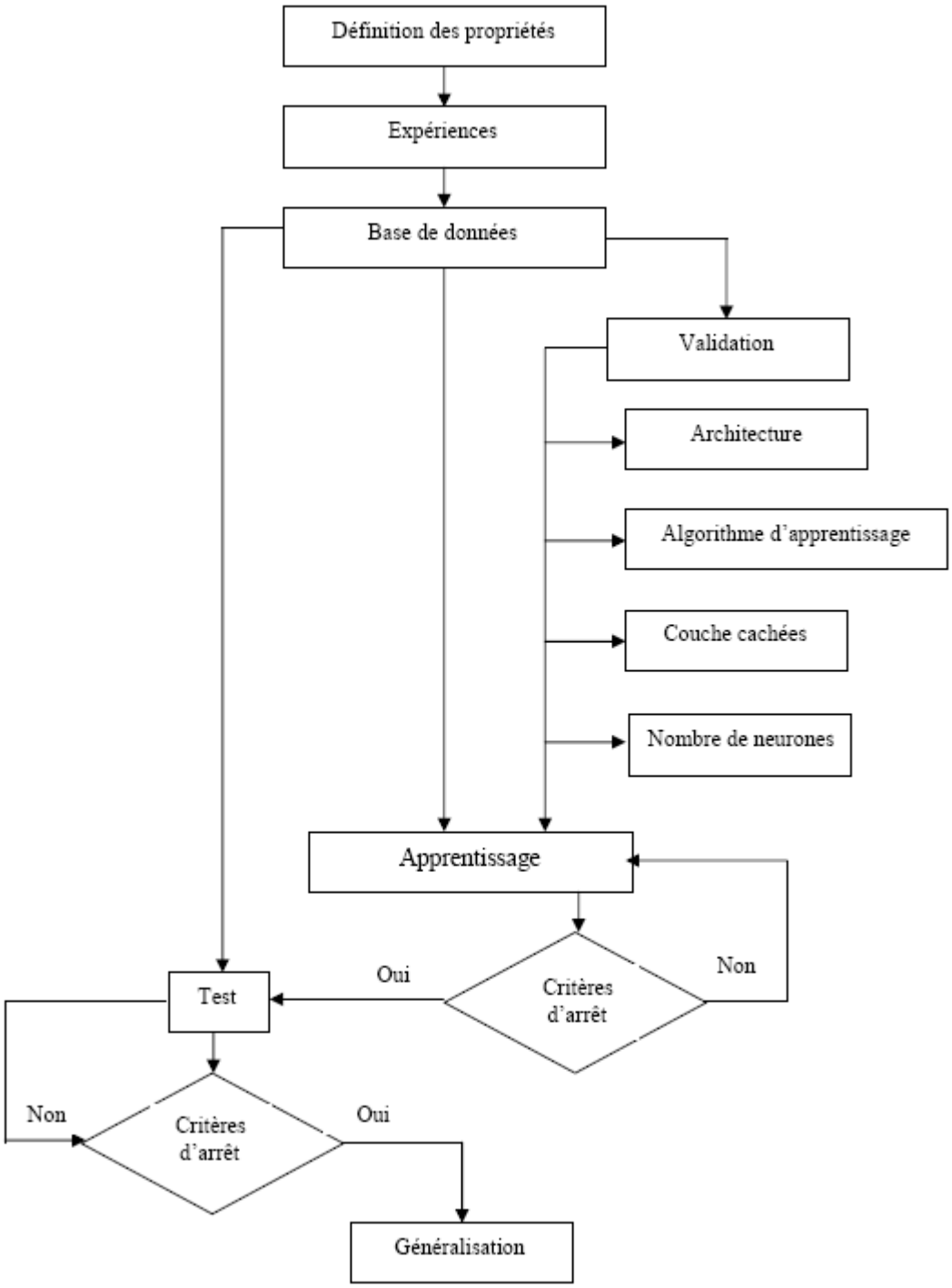
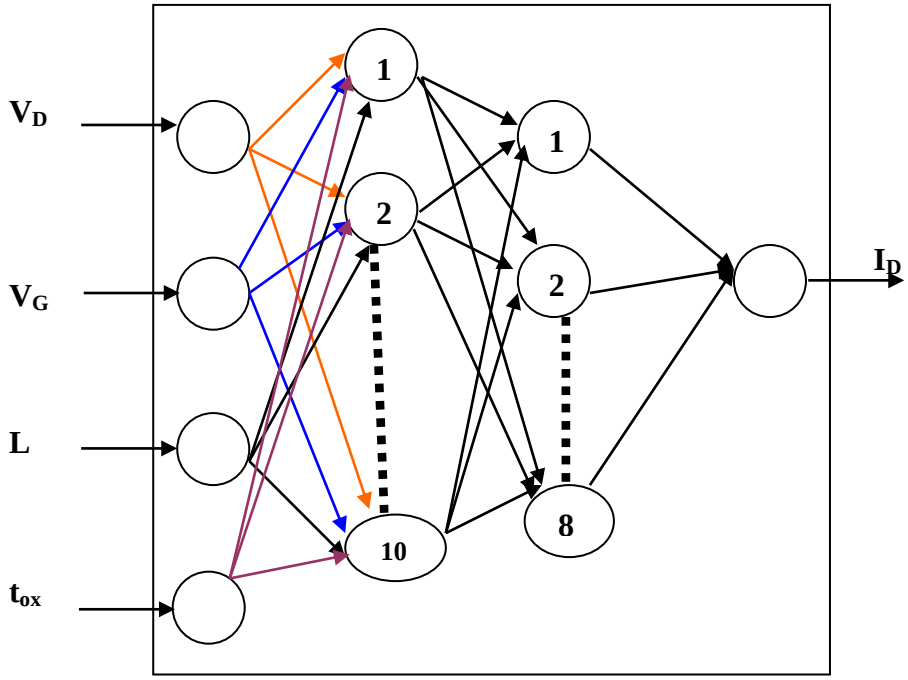


Figure III-2 : Organigramme de l'optimisation du prédicteur neuronal.

III-2.1.2 Simulation et test du modèle sous MATLAB :

III-2.1.2.1 Modèle neuronal (Cas d'un isolant de grille monocouches HfO_2 monoclinique et tétragonal):

Le réseau de neurone développé est conçu afin de relier le vecteur d'entrée (V_D, V_G, t_{ox} et L) au vecteur de sortie I_D. Chacun de ces paramètres est indexé par un neurone (Figure III-3).



Nombre du neurone de la première couche cachée : 10

Nombre du neurone de la deuxième couche cachée : 8

Figure III-3 : Le modèle neuronal du transistor MOSFET

III-2.1.2.1.1 Simulation et test du modèle sous MATLAB :

Les figures III-4, III-5, III-6 et III-7 montrent une comparaison entre les résultats prédits par le modèle neuronal (ANN) des différentes caractéristiques I-V (I_D-V_D et I_G-V_G) avec ceux calculés par le model analytique pour un transistor MOSFET faiblement dopé avec L=0.1 µm, t_{ox}=14 nm (HfO_2 monoclinique EOT=3 nm) et t_{ox}=22 nm (HfO_2 tétragonal EOT=3 nm).

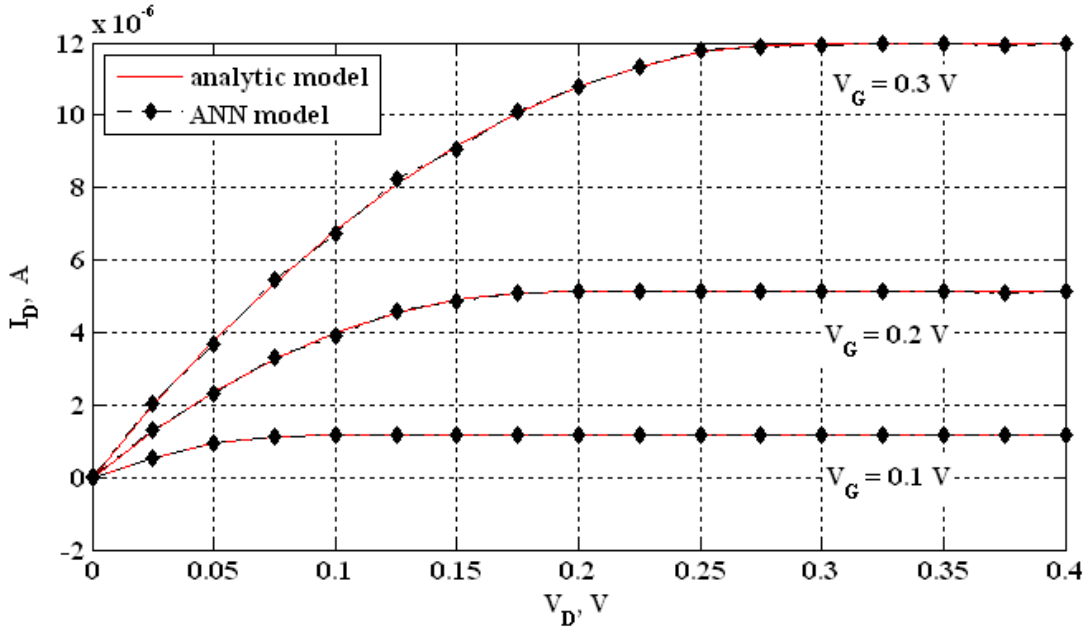


Figure III-4 : Caractéristique $I_D(V_D)$ du MOSFET high-k (HfO_2 monoclinique).

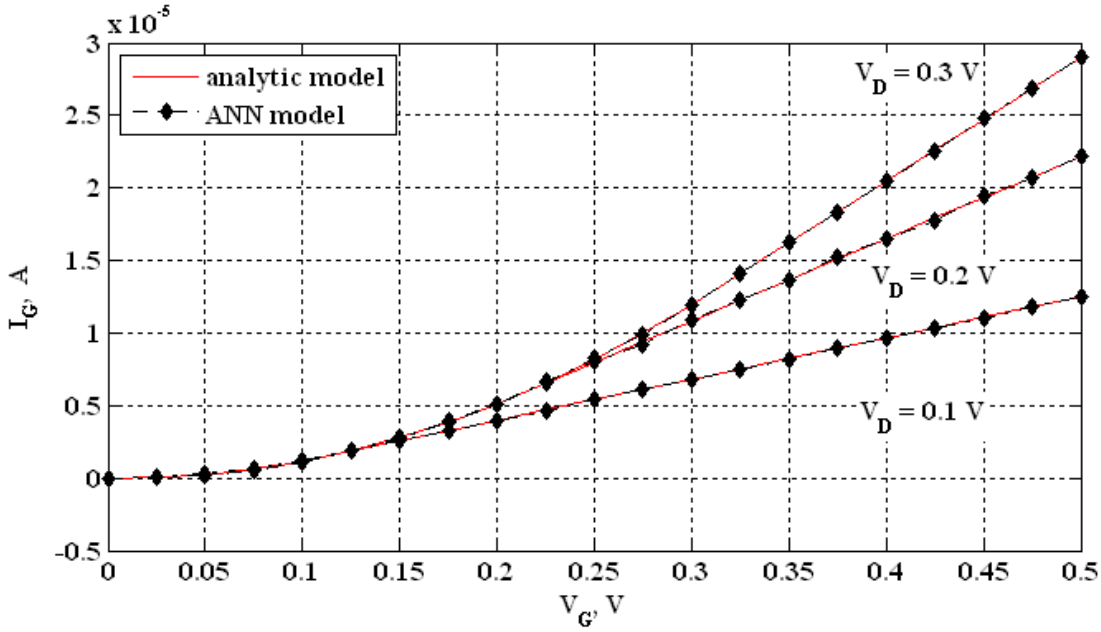


Figure III-5 : Caractéristique $I_G(V_G)$ du MOSFET high-k (HfO_2 monoclinique).

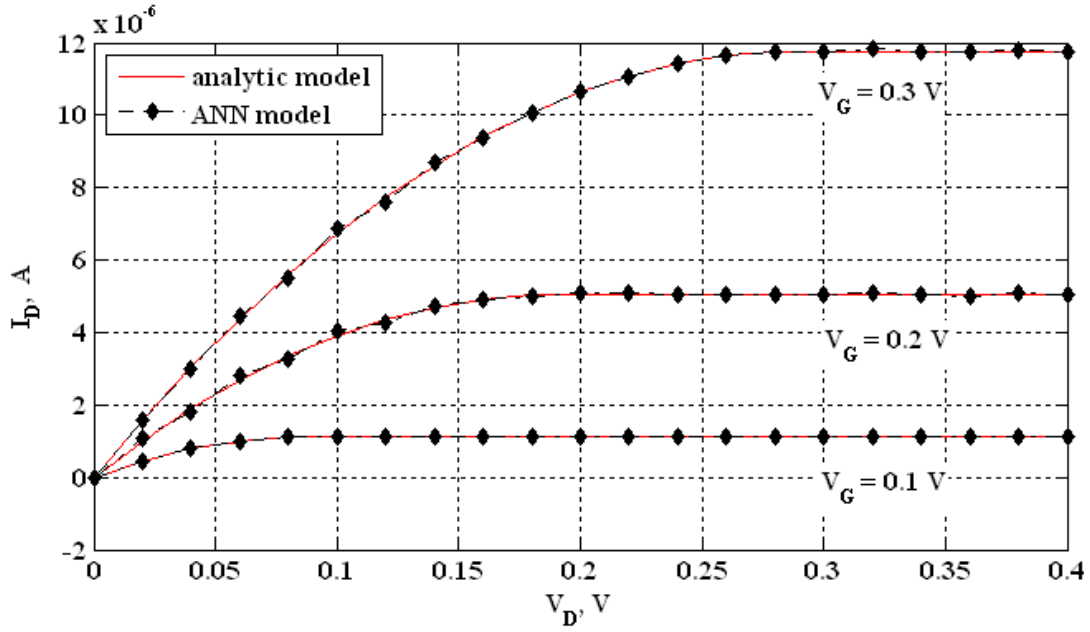


Figure III-6 : Caractéristique $I_D(V_D)$ du MOSFET high-k (HfO_2tétragonal).

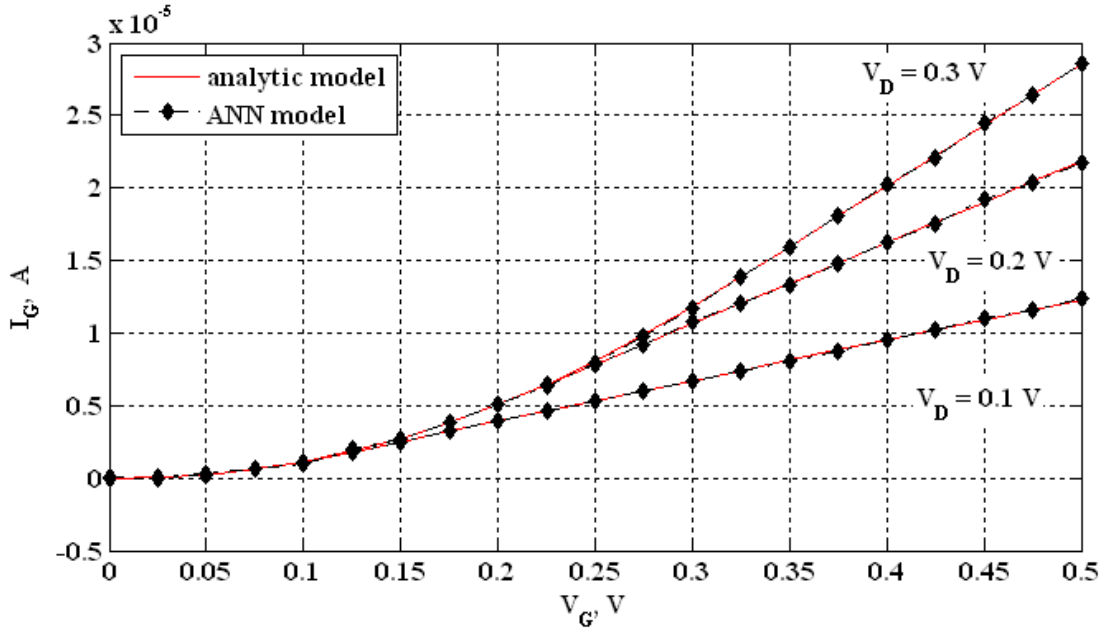
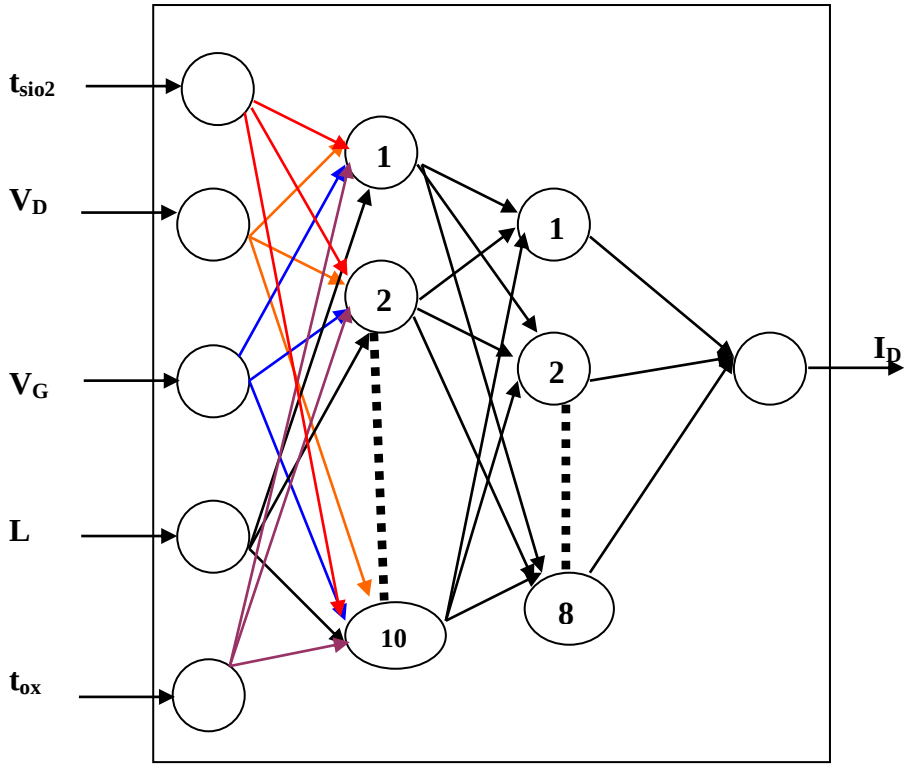


Figure III-7 : Caractéristique $I_G(V_G)$ du MOSFET high-k (HfO_2 tétragonal).

III-2.1.2.2 Modèle neuronal (Cas d'un isolant de grille multicouches SiO_2/HfO_2 monoclinique et tétragonal):

Le réseau de neurone développé est conçu afin de relier le vecteur d'entrée (V_D, V_G, t_{sio2}, t_{ox} et L) au vecteur de sortie I_D. Chacun de ces paramètres est indexé par un neurone (Figure III-8).



Nombre du neurone de la première couche cachée : 10

Nombre du neurone de la deuxième couche cachée : 8

Figure III-8 : Le modèle neuronal du transistor MOSFET

III-2.1.2.2.1 Simulation et test du modèle sous MATLAB :

Les figures III-9, III-10, III-11 et III-12, montrent une comparaison entre les résultats prédits par le modèle neuronal (ANN) des différentes caractéristiques I-V (I_D-V_D et I_G-V_G) avec ceux calculés par le model analytique pour un transistor MOSFET faiblement dopé avec L=0.1 µm, t_{sio2}=1 nm, tox=9 nm (HfO_2 multicouches SiO_2/HfO_2 monoclinique EOT=3 nm) et tox=15 nm (HfO_2 multicouches SiO_2/HfO_2 tétragonal EOT=3 nm) .

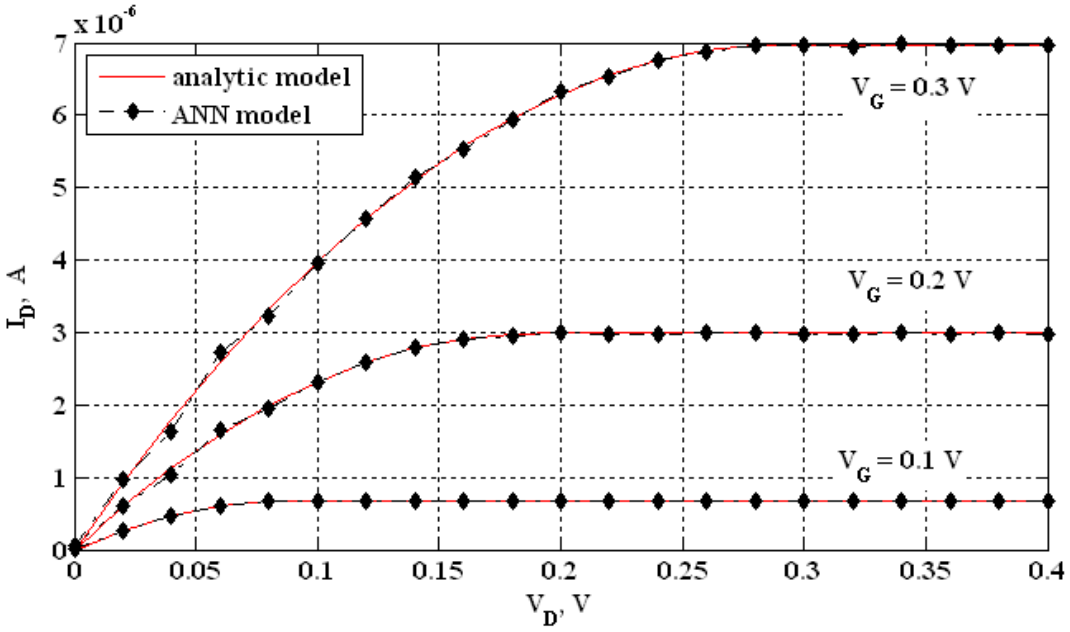


Figure III-9 : Caractéristique $I_D(V_D)$ du MOSFET high-k (SiO_2/HfO_2 multi couche monoclinique).

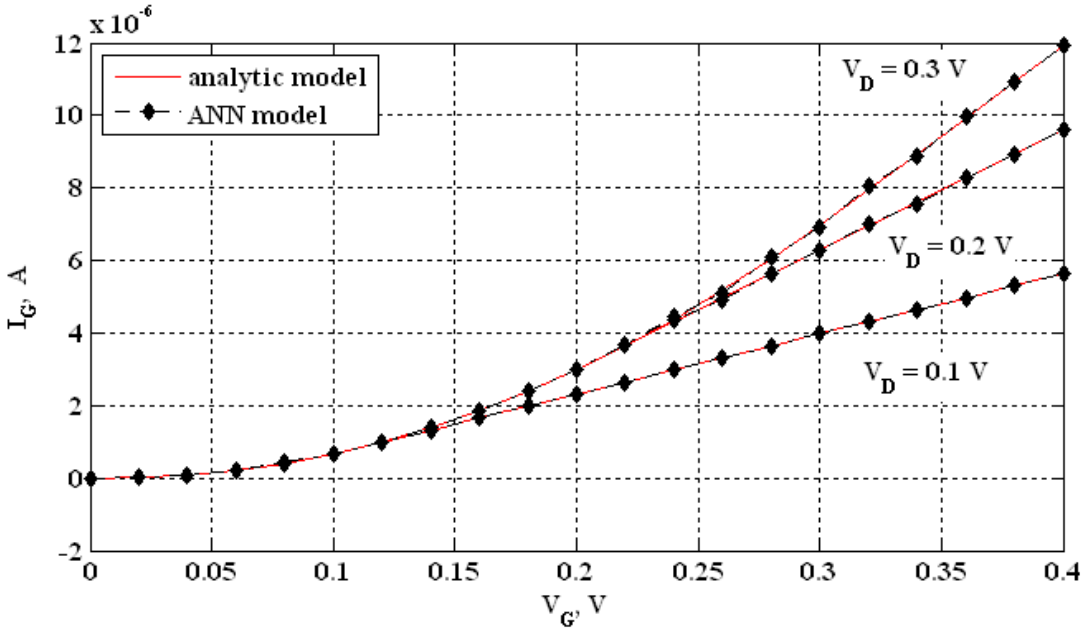


Figure III-10 : Caractéristique $I_G(V_G)$ du MOSFET high-k (SiO_2/HfO_2 multi couche monoclinique).

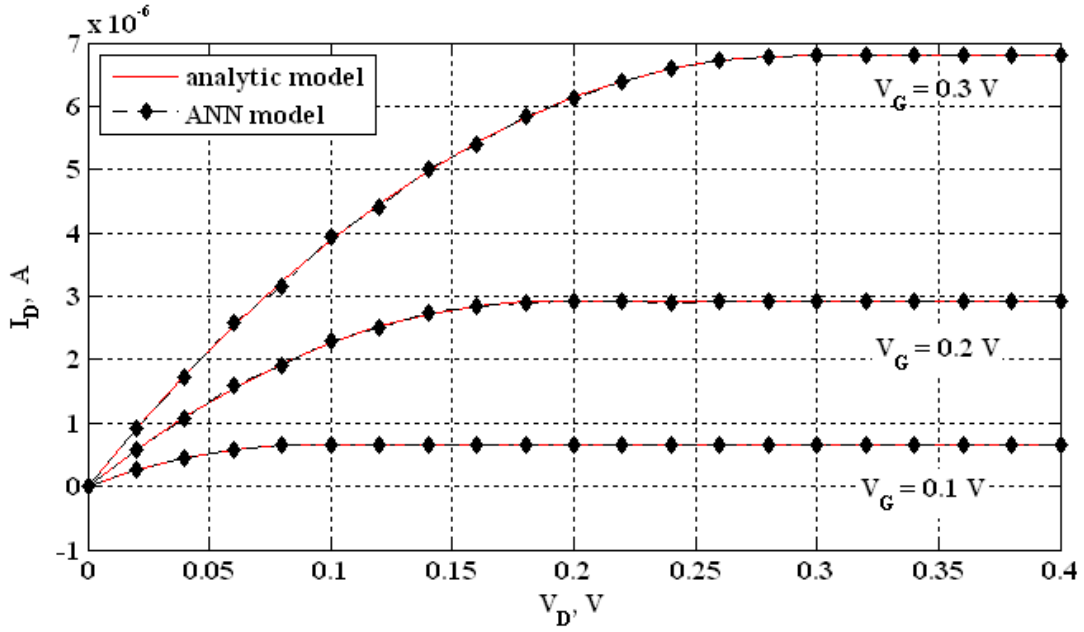


Figure III-11 : Caractéristique $I_D(V_D)$ du MOSFET high-k (SiO_2/HfO_2 multi couche tétragonal).

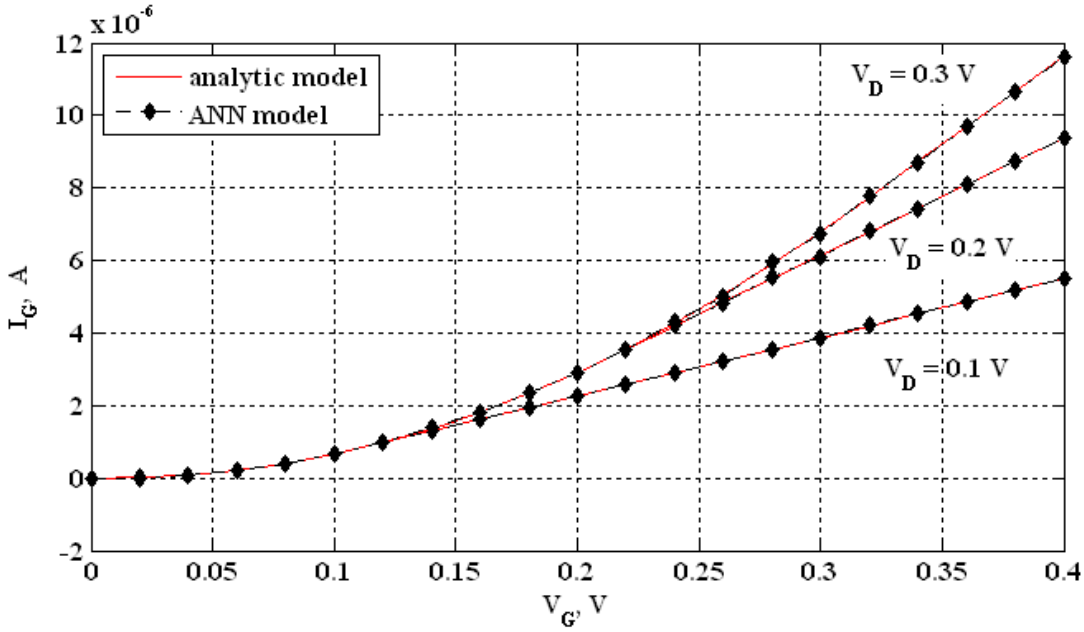


Figure III-12 : Caractéristique $I_G(V_G)$ du MOSFET high-k (SiO_2/HfO_2 multi couche tétragonal).

III-2.2 Commentaire des résultats :

Caractéristique Courant-Tension du MOSFET a permittivité élevée (HfO_2 mono couche monoclinique, figure III-4, figure III-5) et caractéristique Courant-tension du MOSFET a permittivité élevée (HfO_2 monocouche tétragonal, figure III-6, figure III-7) ont les même résultats mais avec différente épaisseur.

Même observation pour MOSFET a permittivité élevée (HfO_2multicouche, figure III-9, figure III-10, figure III-11, figure III-12).

Les résultats obtenus avec les paramètres d'apprentissage est tout à fait satisfaisant, le Tableau III-1 résume les caractéristiques du réseau optimisé pour la modélisation du MOSFET.

Propriété	**Caractéristique**
Architecture	10-8-1 MLP
Fonctions d'activation	Logsig-Logsig-linéaire
Règle d'apprentissage	Rétro propagation des erreurs rapide
EQM de test	0.0001
Nombre d'itérations	4000

Tableau III-1 : Caractéristiques du réseau optimisé.

Le Tableau III-2 résume le pourcentage d'erreur entre le modèle analytique et le modèle neuronal.

Différents cas d'oxyde	**Pourcentage d'erreur %**
monolayer HfO_2, monoclinic-HfO_2	1.37 %
monolayer HfO_2, tetragonal-HfO_2	1.2 %
multilayer SiO_2/HfO_2, monoclinic-HfO_2	2.11 %
multilayer SiO_2/HfO_2, tetragonal-HfO_2	2.24 %

Tableau III-2 : pourcentage d'erreur entre le modèle analytique et le modèle neuronal.

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 [4-5-6].

III-3 Les Algorithmes génétiques :

Algorithme génétique (GA) est une stratégie de recherche évolutionnaire basée sur des sélections naturelles et des processus génétiques, et il a été jugé hautement comme un algorithme d'optimisation robuste pour trouver la solution optimale globale. Contrairement à certaines stratégies de recherche traditionnelle. Il couvre une très large gamme des applications du monde réel [7-8-9-10-11], etc.

Dans cette étude, les paramètres du transistor MOS ont été extraits et optimisés avec l'algorithme génétique. Afin de caractériser les propriétés de MOSFET avec précision, différents modèles compact sont été proposés pour submicronique et la simulation de dispositif MOSFET nanométrique. Chaque modèle comprend des équations et des paramètres d'administration divers. Le modèle de dispositif pour la simulation de circuit doit être précis et robuste. Pour cette raison les paramètres qui sont inclus dans le modèle doivent être modélises, donnant les résultats optimal.

Pour réaliser l'extraction précise des paramètres, nous n'avons proposé une approche simple et rapide, afin de mettre en œuvre des modèles d'algorithmes génétiques pour le transistor MOS (oxyde HfO_2) dans MATLAB, Le procédé d'optimisation arrête le calcul évolutionnaire lorsque le critère d'arrêt est satisfait, les solutions sont trouvées.

III-3.1 Fonction de fitness :

La fonction fitness f utilisée pour l'évaluation des chromosomes et l'ajustement des paramètres est définie par:

$$f = \frac{1}{M}\sum_{V_{GS}}\sum_{V_{DS}}\left[\frac{I_{DS,NUM} - I_{DS,GA}}{I_{DS,NUM}}\right]^2 + \frac{1}{M}\sum_{V_{GS}}\sum_{V_{DS}}\left[\frac{\log(I_{DS,NUM}) - \log(I_{DS,GA})}{\log(I_{DS,NUM})}\right]^2$$

Où M représente la taille de la base de données de la trajectoire (modèle numérique), *'NUM'* et *'GA'* indiquent les données numériques et les données calculées par la technique des AGs respectivement.

L'expression de l'erreur normalisée est donnée comme suit:

$$Erreur\ norm = \varepsilon = \frac{1}{M}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|\frac{I_{DS,NUM} - I_{DS,GA}}{I_{DS,NUM}}\right|$$

Notre problème d'optimisation de paramètres comprend m points de la courbe I-V, chacun représente le courant du drain pour un V_{DS} appliqué (V_{GS} fixe), et n points de la courbe I-V, chacun représente le courant du drain pour un V_{GS} appliqué (V_{DS} fixe). Par conséquent, il possède totalement m* n nœuds calculés à adapter et à optimiser.

Notre objectif est de minimiser la fonction fitness (minimisation de RMS) afin d'obtenir la meilleure solution (meilleur chromosome) dans la population pour assurer l'exactitude et la précision de notre modèle analytique du courant de drain.

III-3.2 Simulation du modèle sous MATLAB :

Les figures III-13, III-14, III-15 et III-16 montrent le comportement de convergence vers la solution optimale (l'évolution de la fonction fitness avec le taux de reproduction utilisé dans cet essai égal à 0.8). Différentes combinaisons des paramètres de GA ont été employées pour trouver la meilleure solution. Les résultats des essais ont été employés pour étudier l'exécution des paramètres.

Les figures III-13, III-14, III-15 et III-16 montrent l'exécution de l'extraction de GA avec des différents taux de mutation et des différentes tailles de population.

Notre processus d'optimisation est assuré par une taille de population de (5, 20, 50) individus, taux de mutation de (0.01, 0.05, 0.1) pour chaque génération et 5000 générations. Les paramètres utilisés dans cette étude sont résumés dans le tableau III-3.

Les paramètres d'AG	**Values**
La taille de la population	5-20-50
Le nombre Maximum de générations	5000
Type de la fonction fitness	Proportionnel
La sélection	Tournoi
Le croisement	Dispersé
La mutation	Uniforme
Taux de mutation	0.01-0.05-0.1
Taux de reproduction	0.8

Tableau III-3 : Paramètres de GA utilisés dans cette application

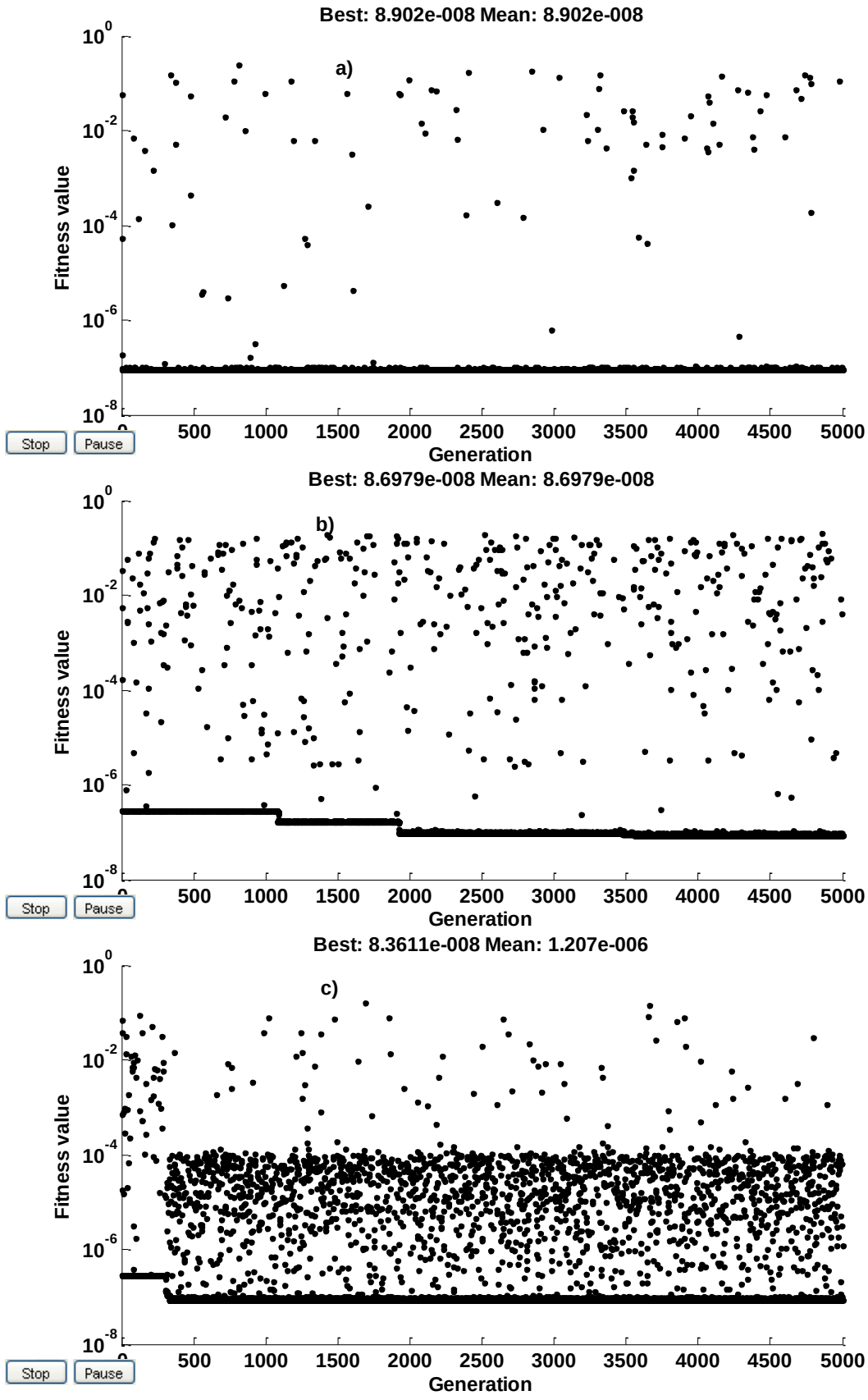


Figure III-13 : Evolution de la fonction fitness en fonction de nombre de générations pour différents taux de mutation (taille de population égal à 20), a) 0.01, b) 0.05, c) 0.1 (MOSFET HfO_2 monoclinique).

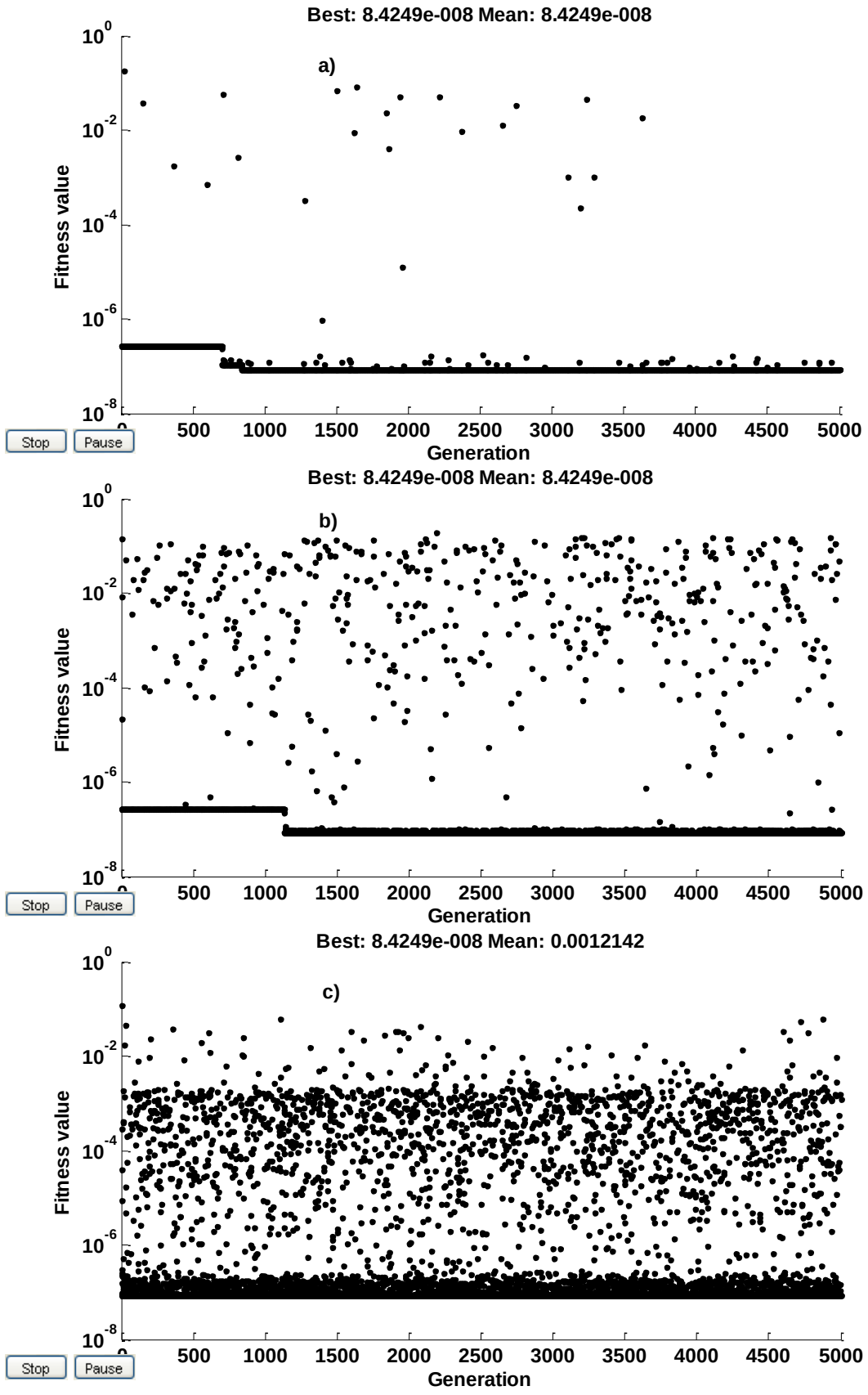


Figure III-14 : Evolution de la fonction fitness en fonction de nombre de générations pour différents taille de population (taux de mutation égal à 0.05), a) 5, b) 20, c) 50 (MOSFET HfO_2 tétragonal).

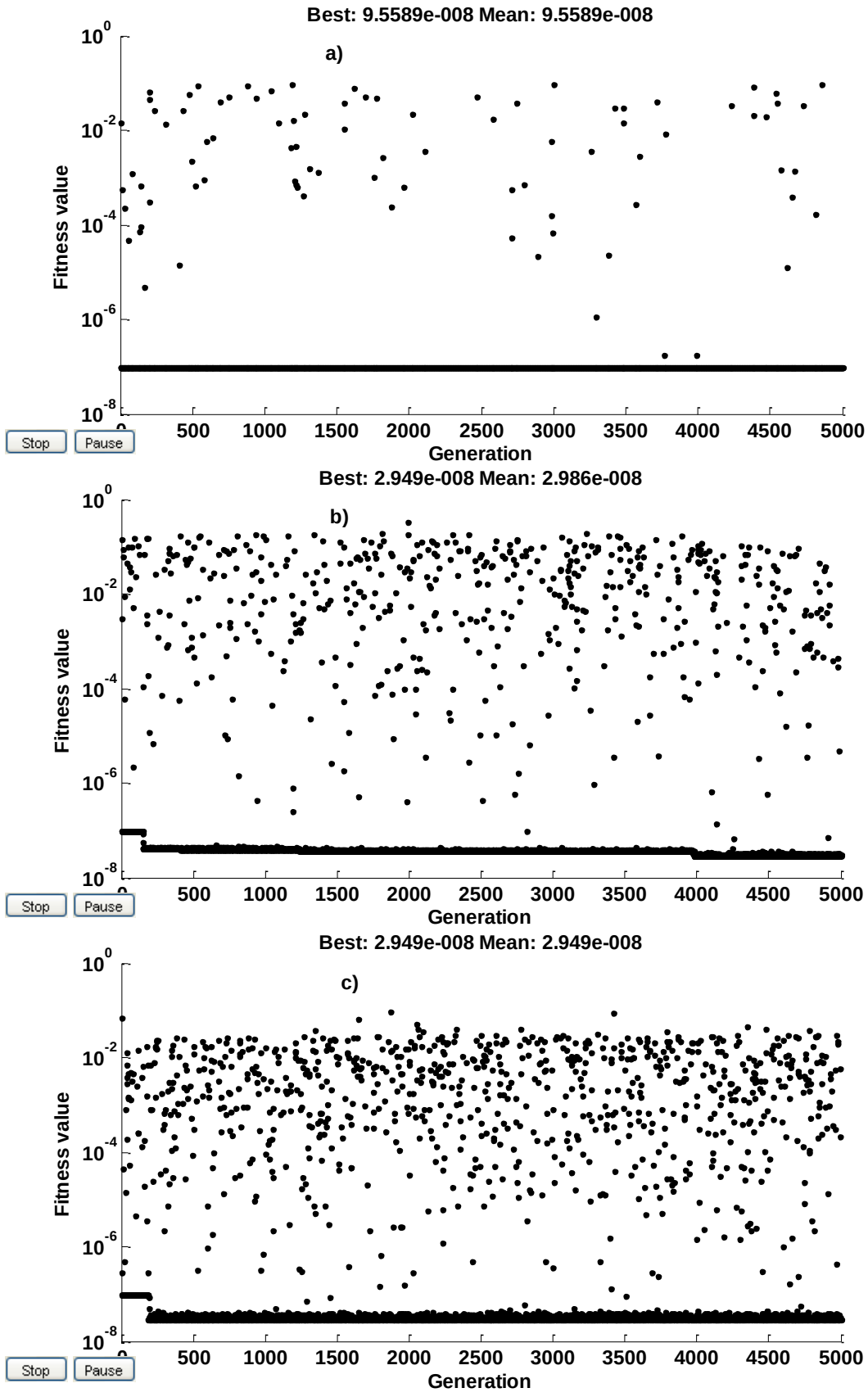


Figure III-15 : Evolution de la fonction fitness en fonction de nombre de générations pour différents taux de mutation (taille de population égal à 20), a) 0.01, b) 0.05, c) 0.1 (MOSFET SiO_2/HfO_2 monoclinique).

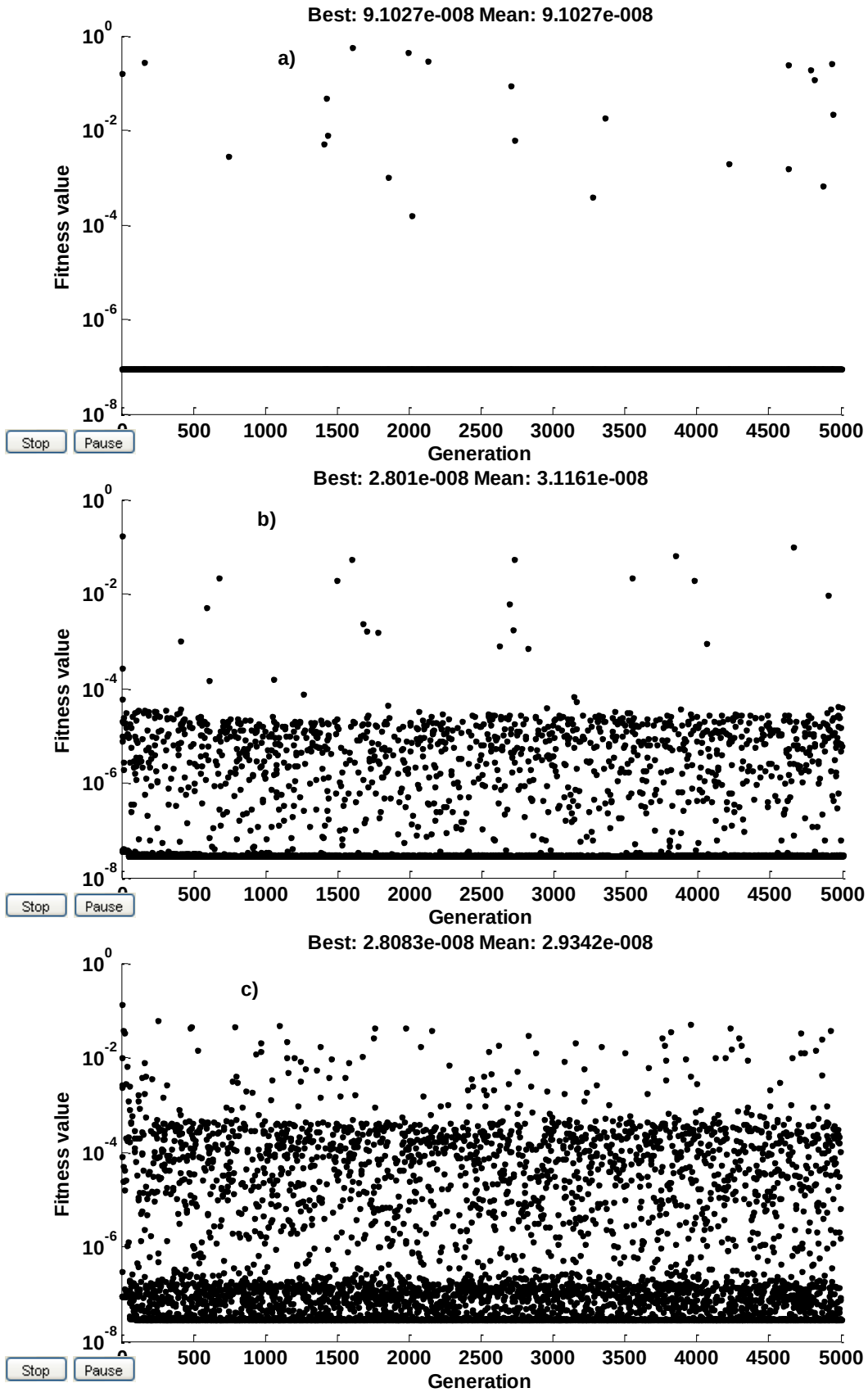


Figure III-16 : Evolution de la fonction fitness en fonction de nombre de générations pour différents taille de population (taux de mutation égal à 0.05), a) 5, b) 20, c) 50 (MOSFET SiO_2/HfO_2 tétragonal).

III-3.2.1 Commentaire des résultats :

Suivant les indications des figures III-13, III-14, III-15 et III-16:

Le Taux de Mutation : L'opérateur de mutation est appliqué avec une probabilité P_m. GA a été amélioré par augmentation de taux de mutation de 0.01 à 0.1.

- Si ce taux est trop grand, le processus de recherche devient purement aléatoire, et donc, la performance sera dégradée (figures III-13(c), III-15(c)).

- S'il est faible la population est moins diversifiée et en plus il y a risque de stagnation (figures III-13(a), III-15(a)).

La Taille de la population : On n'a également constaté que nous avons une population trop petite ou trop grande dégraderait l'exécution de GA. GA a été amélioré par augmentation de la taille de population de 5 à 50.

- Si elle est trop grande**,** la diversité augmente et la convergence vers un optimum local diminue, le système passerait la majeure partie de temps dans l'évaluation de fonction autre que la recherche (figures III-14(c), III-16(c)).

- Si elle est trop petite, la probabilité de s'attarder sur des minima locaux est grande,(figures III-14(a), III-16(a)).

Il convient de noter que l'exécution de GA dépend principalement des processus du choix et de la recombinaison.

Après l'analyse des figures, nous constatons que l'évolution de la fonction fitness donne un bon résultat avec taux de mutation égal à 0.05 et la taille de la population égal à 20 pour le processus d'optimisation (figures III-13(b), III-14(b), III-15(b) et III-16(b)).

III-3.3 Validation du modèle pour un dispositif à canal court (Cas d'un isolant de grille monocouches HfO_2 monoclinique et tétragonal) :

Les figures III-17, III-18, III-19 et III-20 montrent une comparaison entre les résultats prédits par le modèle compact des différentes caractéristiques I-V (I_D-V_D et I_G-V_G) avec ceux calculés par le model des paramètres optimisés (GA) pour un transistor MOSFET faiblement dopé avec L=0.1 µm, tox=14 nm (HfO_2 monoclinique EOT=3 nm) et tox=22 nm (HfO_2 tétragonal EOT=3 nm).

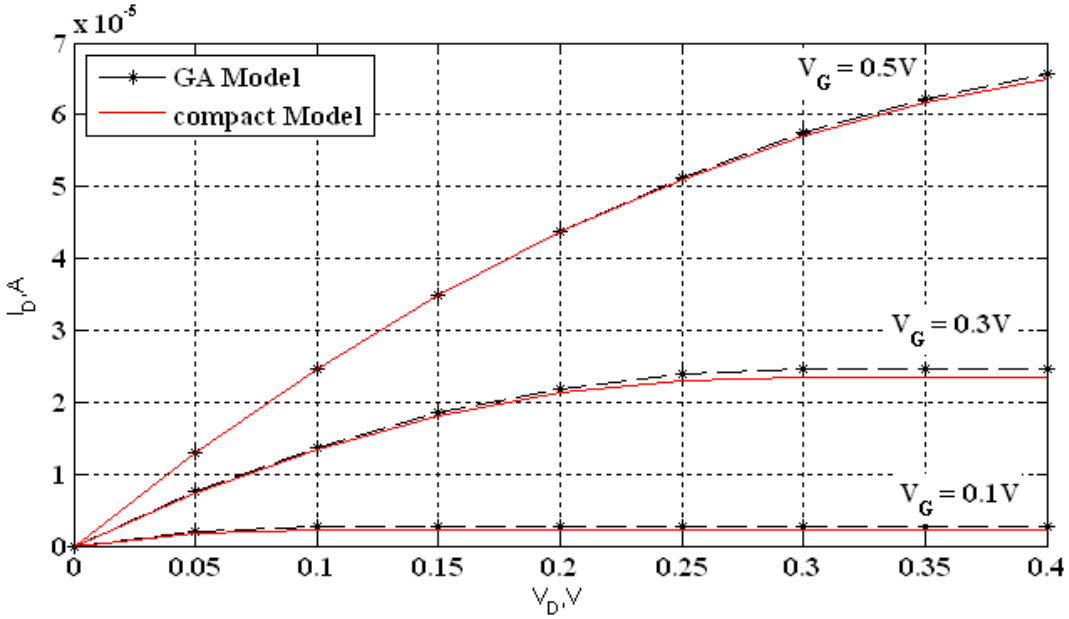


Figure III-17 : Caractéristique $I_D(V_D)$ du MOSFET high-k (HfO_2 tétragonal).

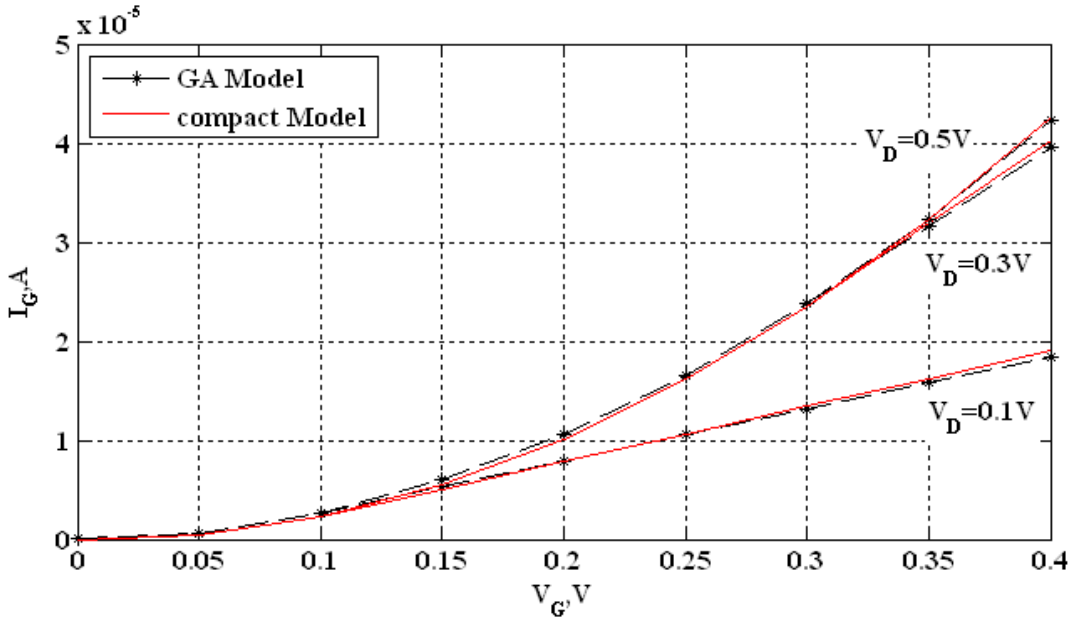


Figure III-18 : Caractéristique $I_G(V_G)$ du MOSFET high-k (HfO_2 tétragonal).

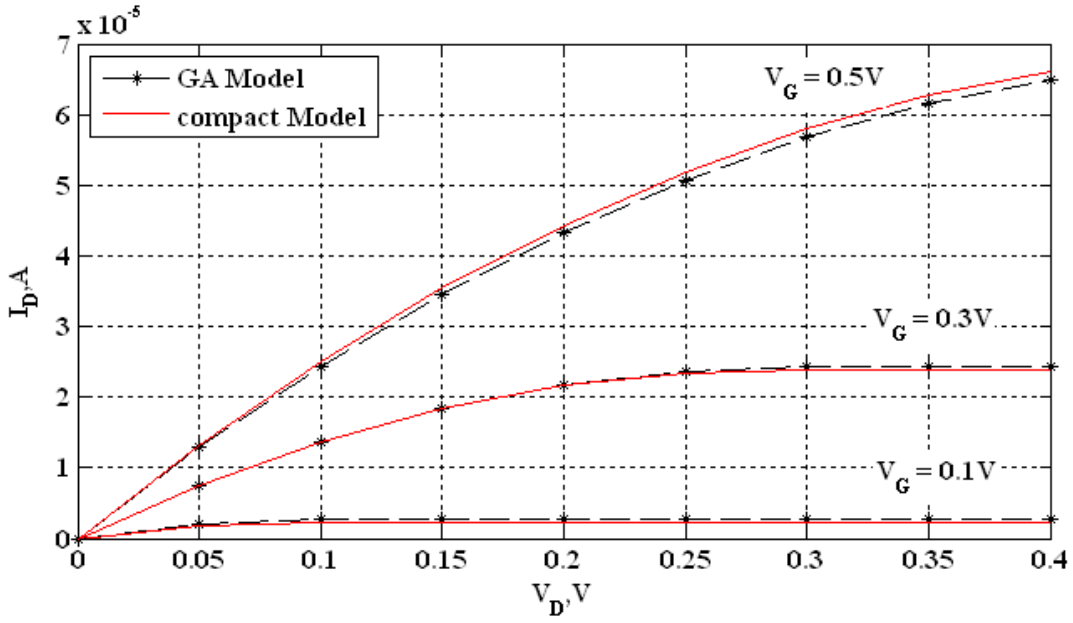


Figure III-19 : Caractéristique $I_D(V_D)$ du MOSFET high-k (HfO_2 monoclinique).

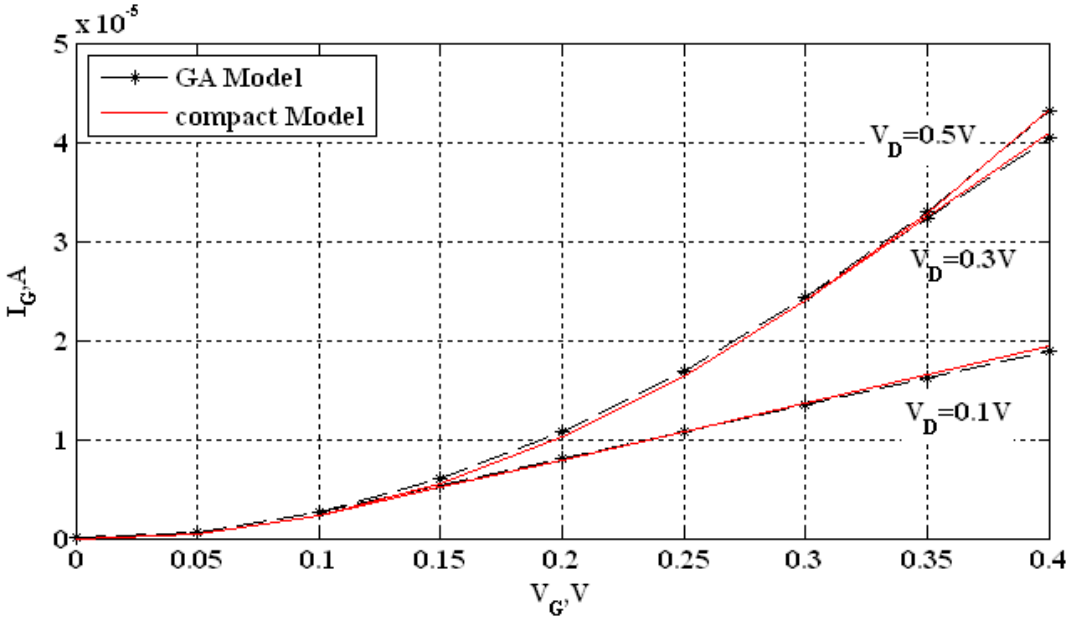


Figure III-20 : Caractéristique $I_G(V_G)$ du MOSFET high-k (HfO_2 monoclinique).

Les paramètres optimisés de notre modèle compact du courant à canal court sont récapitulés dans les tableaux III-4 et III-5

	Paramètres	**Compact**	**Optimisé GA**
MOSFET L=0.1µm, monoclinique	V_T	0.01	0.01+6e-15
	k_n	0.2844e-3	0.2844e-3
MOSFET L=0.1µm, tétragonal	V_T	0.01	0.01+2e-15
	k_n	0.2799e-3	0.2799e-3

Tableau III-4 : La configuration finale des paramètres obtenus pour un MOSFET (HfO_2 monocouche) ($I_D(V_D)$)

	Paramètres	**Compact**	**Optimisé GA**
MOSFET L=0.1µm, monoclinique	V_T	0.01	0.01+2e-15
	k_n	0.2844e-3	0.2844e-3
MOSFET L=0.1µm, tétragonal	V_T	0.01	0.01-2e-15
	k_n	0.2799e-3	0.2799e-3

Tableau III-5 : La configuration finale des paramètres obtenus pour un MOSFET (HfO_2 monocouche) ($I_G(V_G)$)

III-3.4 Validation du modèle pour un dispositif à canal court (Cas d'un isolant de grille multicouches SiO_2/HfO_2 monoclinique et tétragonal) :

Les figures III-21, III-22, III-23 et III-24 montrent une comparaison entre les résultats prédits par le modèle compact des différentes caractéristiques I-V (I_D-V_D et I_G-V_G) avec ceux calculés par le model des paramètres optimisés (GA) pour un transistor MOSFET faiblement dopé avec L=0.1 µm, t_{sio2}=1 nm, tox=9 nm (HfO_2 multicouches SiO_2/HfO_2 monoclinique EOT=3 nm) et tox=15 nm (HfO_2 multicouches SiO_2/HfO_2 tétragonal EOT=3 nm) .

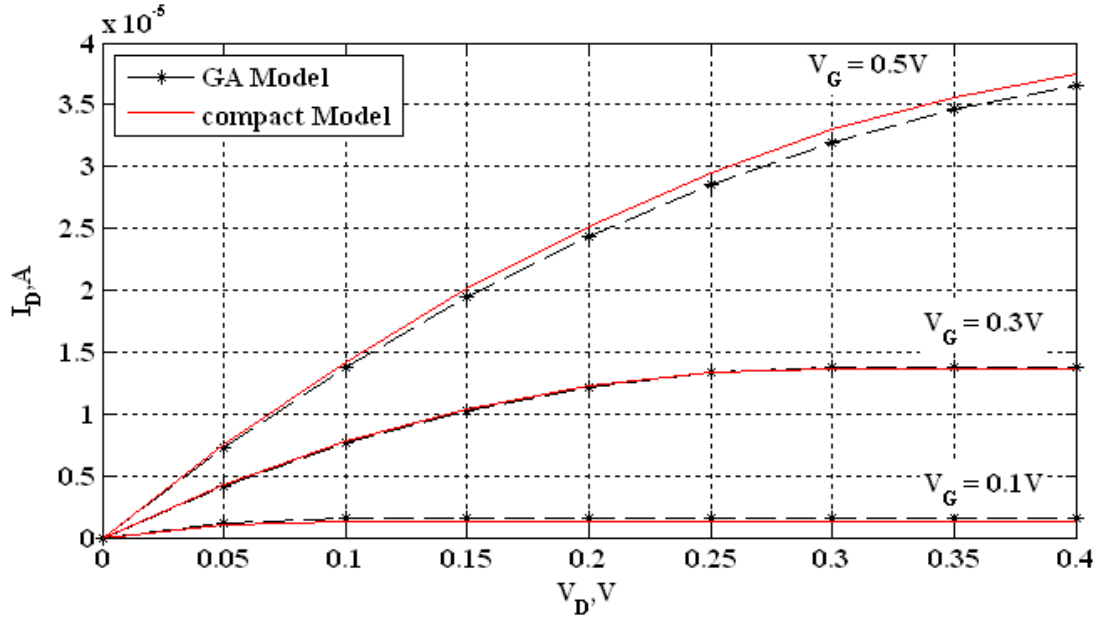


Figure III-21 : Caractéristique $I_D(V_D)$ du MOSFET high-k (SiO_2/HfO_2 tétragonal).

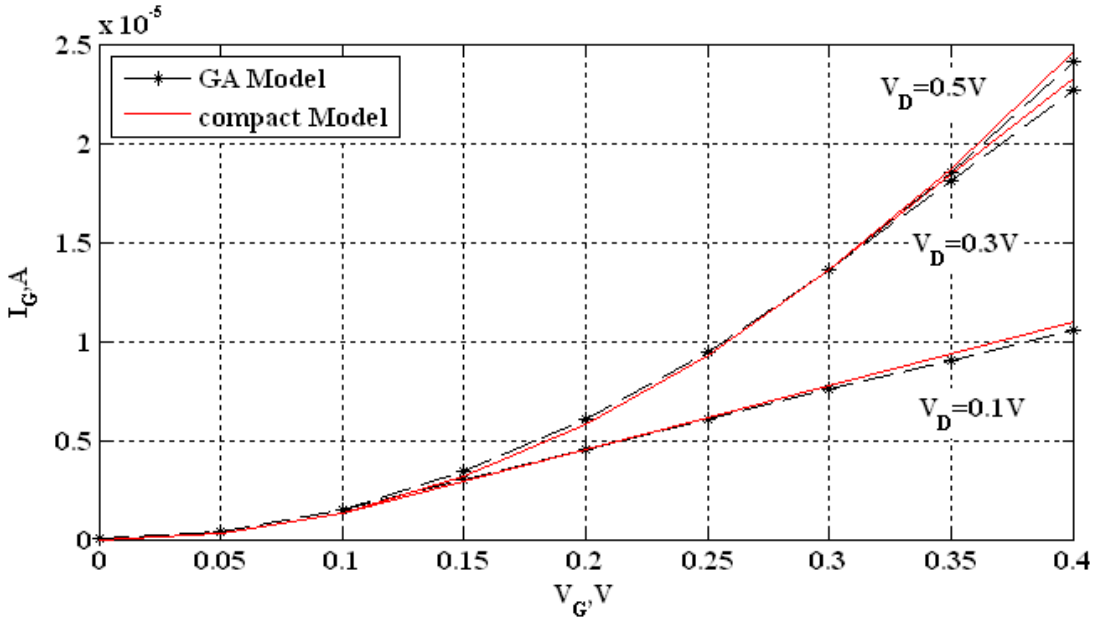


Figure III-22 : Caractéristique $I_G(V_G)$ du MOSFET high-k (SiO_2/HfO_2 tétragonal).

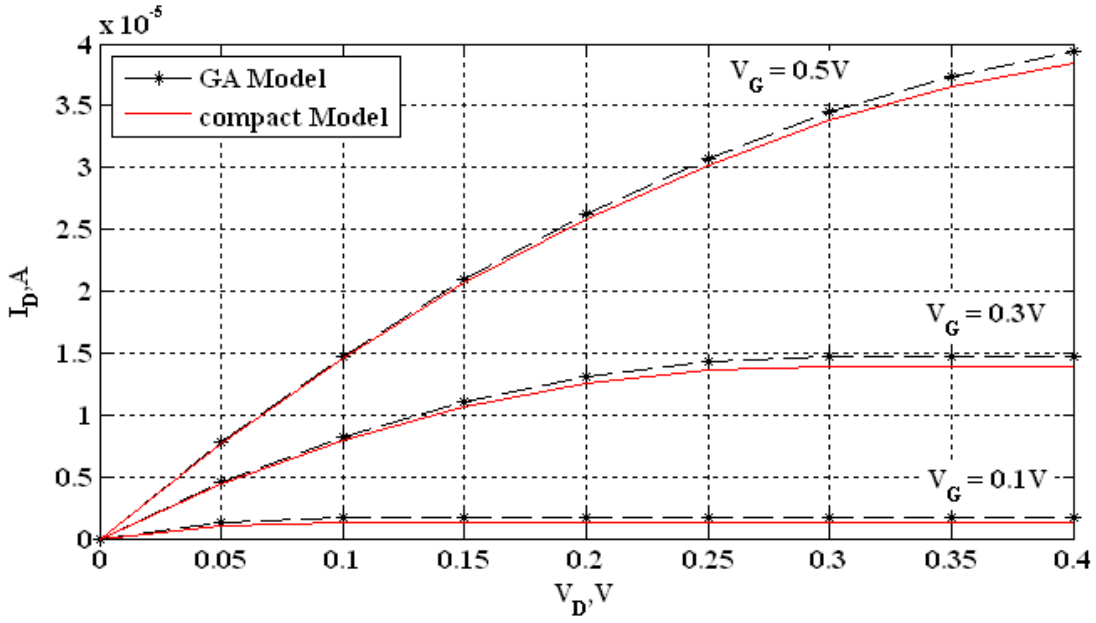


Figure III-23 : Caractéristique $I_D(V_D)$ du MOSFET high-k (SiO_2/HfO_2 monoclinique).

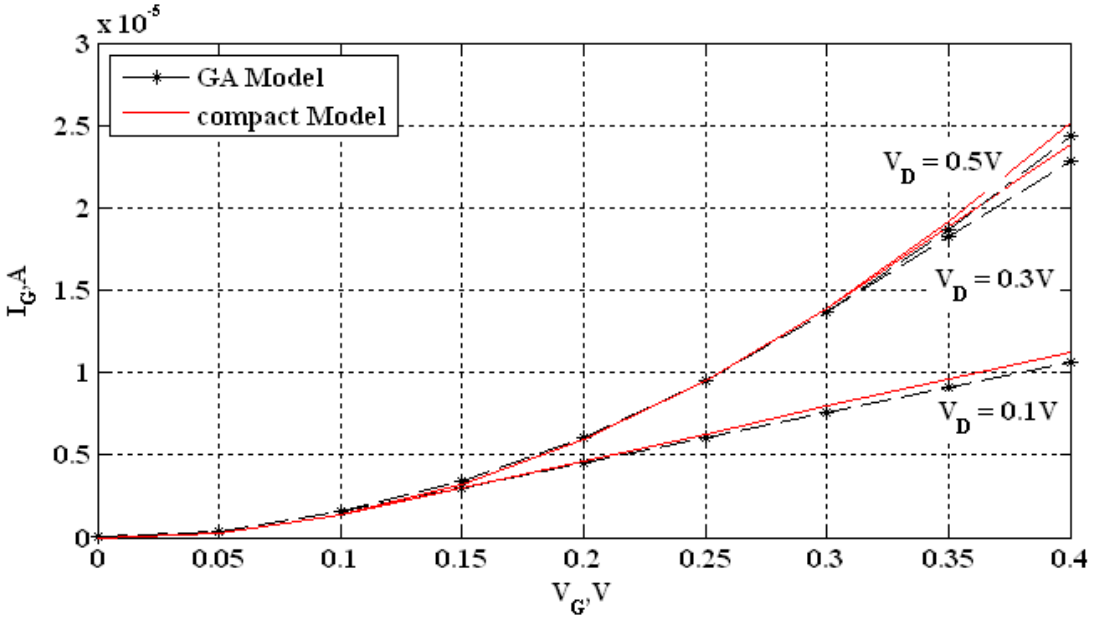


Figure III-24 : Caractéristique $I_G(V_G)$ du MOSFET high-k (SiO_2/HfO_2 monoclinique).

Les paramètres optimisés de notre modèle compact du courant à canal court sont récapitulés dans les tableaux III-6 et III-7

	Paramètres	**Compact**	**Optimisé GA**
MOSFET L=0.1µm, monoclinique	V_T	0.01	0.01+3e-15
	k_n	0.1656e-3	0.1656e-3
MOSFET L=0.1µm, tétragonal	V_T	0.01	0.01-3e-15
	k_n	0.1616e-3	0.1616e-3

Tableau III-6 : La configuration finale des paramètres obtenus pour un MOSFET (HfO_2 multicouche) ($I_D(V_D)$)

	Paramètres	**Compact**	**Optimisé GA**
MOSFET L=0.1µm, monoclinique	V_T	0.01	0.01+2e-15
	k_n	0.1656e-3	0.1656e-3
MOSFET L=0.1µm, tétragonal	V_T	0.01	0.01+3e-15
	k_n	0.1616e-3	0.1616e-3

Tableau III-7 : La configuration finale des paramètres obtenus pour un MOSFET (HfO_2 multicouche) ($I_G(V_G)$)

III-3.5 Commentaire des résultats :

Caractéristique I-V du MOSFET high-k (HfO_2 monocouche monoclinique, figure III-19, figure III-20) et caractéristique I-V du MOSFET high-k (HfO_2 mono couche tétragonal, figure III-17, figure III-18) ont les même résultats mais avec différente épaisseur.

Même remarque pour MOSFET high-k (HfO_2 multicouche, figure III-21, figure III-22, figure III-23, figure III-24).

Dans cette recherche, basée sur un algorithme d'optimisation, l'algorithme génétique a été utilisé pour trouver les valeurs des paramètres modèles pour le transistor MOSFET. Le problème a été abordé, comme le problème d'optimisation par les réseaux de neurone. GA est considéré pour être efficace pour résoudre les problèmes de la miniaturisation des dispositifs nanométriques [12],[13],[14],[15],[16], [17],[18] .

III-4 Conclusion :

La résolution d'un problème d'optimisation de transistor MOSFET à permittivité élevée est un problème complexe car de nombreux facteurs interviennent et interagissent entre eux. Néanmoins, l'optimisation appliquée au domaine de l'électronique permet de résoudre des problèmes qui étaient insolubles auparavant et aboutit souvent à des solutions originales.

Dans ce chapitre, nous avons démontré l'applicabilité de l'approche neuronale et de l'optimisation par l'algorithme génétique à l'étude des circuits CMOS nanométriques.

Une approche analytique basée sur un prédicteur neuronal a été développée dans le cas du transistor MOSFET à permittivité élevé (HfO_2). Cette dernière nous a permis de prévoir l'évolution de courant en fonction des différents paramètres (tension du drain, tension de grille, longueur du canal et épaisseur d'oxyde).

Les résultats obtenus par l'approche neuronal, elle était vérifié par la technique GA, est les deux méthodes présentent une meilleure stratégie conventionnelle d'extraction des paramètres, en terme de convergence elles fournissent des solutions optimales globales.

À partir des résultats obtenus HfO_2 est le bon candidat pour remplacer SiO_2.

Conclusion et Perspectives

Conclusion et Perspectives

La réduction des dimensions des composants impose aujourd'hui des changements radicaux dans la manière d'appréhender l'élaboration des dispositifs micro électroniques du futur, aussi bien d'un point de vue technologique que théorique. C'est le cas, par exemple, de l'oxyde de silicium dont les limites physiques intrinsèques sont actuellement atteintes. En effet, la prochaine génération de transistors MOS nécessiterait d'avoir des épaisseurs d'oxyde de grille inférieures au nanomètre, ce qui conduirait à des niveaux de courants de fuite et tunnel fortement nuisibles au fonctionnement du composant.

C'est dans ce contexte qu'ont été menés les travaux présentés dans ce manuscrit. Ils ont tout particulièrement été concentrés sur le remplacement de la couche diélectrique SiO_2 par un matériau de permittivité plus élevée (matériaux high-k). Cette solution alternative, déjà adoptée par Intel. Le remplacement de la silice par un oxyde à permittivité plus forte ou oxyde high-k devient inéluctable, permettant ainsi d'obtenir une épaisseur physique plus importante. Parmi les matériaux candidats HfO_2.

Dans le premier chapitre, on présentera le transistor MOSFET conventionnel et son évolution vers de nouvelles architectures innovantes. Dans un premier temps, le fonctionnement du transistor MOSFET, ses paramètres électriques importants, les effets de miniaturisation des dispositifs et leurs limites ainsi que les diverses solutions technologiques qui ont été utilisées pour améliorer les performances du transistor à canal. Nous avons discuté l'alternative du remplacement de SiO_2, en tant qu'oxyde de grille dans les applications CMOS, par des oxydes à permittivités plus élèves, ainsi que les principaux critères de choix pour les futurs diélectriques. TiO_2, HfO_2 et ZrO_2 ont démontré une compatibilité thermodynamique remarquable avec le substrat Si et sont considérés aujourd'hui parmi les plus intéressants candidats pour remplacer SiO_2.

Le deuxième chapitre est constitué de deux parties :

La première partie est consacrée 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.

La deuxième partie destinée à la présentation de principe des algorithmes génétiques, expose les bases nécessaires à la compréhension des méthodes d'optimisation par les algorithmes génétiques (AGs) et donne leurs applications dans les différents domaines d'optimisation.

Dans le troisième chapitre, nous avons appliqués les méthodes décrites au chapitre deux pour mettre en place :

1. Des programmes de réseaux de neurones sous Matlab, qui utilise la rétropropagation du gradient pour identifier différentes courbes (caractéristique $I_D(V_D)$ et $I_G(V_G)$ du MOSFET, l'évolution de l'erreur moyenne d'apprentissage de notre prédicteur), les neurones peuvent être organisés de différentes manières, c'est ce qui définit l'architecture et le modèle du réseau. L'architecture la plus courante est celle dite du perceptron multicouche. Avant de pouvoir utiliser les capacités de classification d'un réseau de neurones, il faut le construire. Ceci se déroule en quatre temps:

    - La construction de la structure du réseau.

    - La constitution d'une base de données de vecteurs représentant au mieux le domaine à modéliser. Celle-ci est scindée en deux parties : une partie servant à l'apprentissage du réseau (on parle de base d'apprentissage) et une autre partie aux tests de cet apprentissage (on parle de base de test).

    - Le paramétrage du réseau par apprentissage. Au cours de l'apprentissage, les vecteurs de données de la base d'apprentissage sont présentés séquentiellement et plusieurs fois au réseau. Un algorithme d'apprentissage ajuste le poids du réseau afin que les vecteurs soient correctement appris. L'apprentissage se termine lorsque l'algorithme atteint un état stable.

    - La phase de reconnaissance qui consiste à présenter au réseau chacun des vecteurs de la base de test. La sortie correspondante est calculée en propageant les vecteurs à travers le réseau. La réponse du réseau est lue directement sur les unités de sortie et comparée à la réponse attendue. Une fois que le réseau présente des performances acceptables, il peut être utilisé pour répondre au besoin qui a été à l'origine de sa construction.

2. Implémentation des paramètres de la technique (AG): Pour l'implémentation de la technique AG, des règles de la boîte d'outil GATOOL sous MATLAB sont utilisées (le tournoi, le croisement disperse, la mutation uniforme), pour validé les résultats obtenus par les réseaux de neurones.

En conclure, l'approche basée sur les réseaux de neurones et l'algorithme génétique pour l'étude de l'évolution de courant du drain et de courant du grille en fonction des différents paramètres (tension du drain, tension de grille, longueur du canal et épaisseur d'oxyde) est efficace pour résoudre les problèmes de la miniaturisation des dispositifs nanométriques.

A partir des résultats obtenus, plusieurs perspectives de ce travail peuvent être envisagées :

- La miniaturisation des transistors MOS a augmenté la densité d'intégration et la vitesse de fonctionnement des circuits. Cette miniaturisation a conduit à des phénomènes parasites qui dégradent les caractéristiques courant-tension. Le MOSFET EKV modèle résout ce problème, ce composant permet de progresser dans la miniaturisation.

- La miniaturisation continue des composants microélectronique nécessitera de disposer, à court terme, d'isolants alternatifs au SiO_2. Le remplacement de la silice par un matériau high-k disposant d'un constant diélectrique plus élevée permet de repousser les limites fondamentales de l'intégration et en particulier de limiter les courants de fuites traversant les dispositifs. La recherche actuelle se tourne de plus en plus vers le dopage ou l'élaboration d'alliage avec comme matériau de base le HfO_2. Ainsi des thèses et des publications sur l'incorporation d'oxyde d'yttrium, d'oxyde de lanthane ou encore d'oxyde de zirconium voient le jour. L'objectif de ces composés est de stabiliser la phase cubique ou tétragonale du HfO_2 afin d'obtenir des constantes diélectriques plus élevées.

Références bibliographiques

Références bibliographiques (Introduction)

[1] J. Coignus, Etude de la conduction électrique dans les diélectriques à forte permittivité utilisés en microélectronique, thèse doctorat, L'Université de Grenoble France, 2010.

[2] J. Kilby, Miniaturized Electronic Circuits, American Patent #3,138, 743, 1959.

[3] G. E. Moore, Cramming more Components onto Integrated Circuits, Electronics, Vol. 38, N°. 8, pp. 114 – 117, 1965.

Références bibliographiques (Chapitre I)

[1] D. A. Neamen, Semiconductor physics and devices: Basic principles, Third edition, McGraw-Hill, 2003.

[2] C. Ngô, H. Ngô, Les semiconducteurs De l'électron aux dispositifs, Dunod, 2003.

[3] H. Mathieu, Physique des semiconducteurs et des composants électroniques, Dunod, 2004.

[4] J-P. Colinge, F. Van De Wiele, Physique des dispositifs semiconducteurs, De Boeck, 1996.

[5] T. Skotnicki, Transistor MOS et sa technologie de fabrication, Techniques de l'Ingénieur, E2430, 2000.

[6] S. M.Sze, Physics of Semiconductor Devices, 2nd edition, New York: Wiley, 1981.

[7] J. Gautier, Physique des dispositifs pour circuits intégrés silicium, Paris, France Lavoisier, 367p, 2003.

[8] T. Bouttchacha, Les composants actifs à semi-conducteurs, Office des Publications Universitaire, 1995.

[9] G. D. Wilk, R. M. Wallace, J. M. Anthony, J. Appl. Phys. 89, 5243, 2001.

[10] J. Robertson, Eur. Phys. J. Appl. Phys. 28**,** 265, 2004.

[11] M. L. Green, T. W. Sorsch, G. L. Timp, D. A. Muller, B. E. Weir, P. J. Silverman, S. V.Moccio, Y. O. Kim, Microelectronic Engineering. 48, 25, 1999.

[12] M. L. Green, E. P. Gusev, R. Degraeve, E. L. Garfunkel, J. Appl. Phys. 90, 2057, 2001.

[13] K. Dabertrand, Croissance de diélectrique à forte permittivité par la technique MOCVD en phase liquide pulsée : Elaboration et caractérisation de films de HfO_2, thèse doctorat, Université Joseph Fourier, 2006.

[14] D. A. Muller, T. Sorsch, S. Moccio, F. H. Baumann, K. Evans-Lutterodt, G. Timp, Nature 399, 758, 1999.

[15] J. H. Stathis, IBM. J. Res. Dev. 46**,** 265, 2002.

[16] R. M. Wallace, G. Wilk, MRS Bulletin, 192, 2002.

[17] Y. Taur, Tak H. Ning, Fundamentals of modern VLSI devices, Cambridge University Press, 1998.

[18] Xavier Garros, Caractérisation et modélisation de l'oxyde d'hafnium comme alternative à la silice pour les futures technologies CMOS submicroniques, thèse doctorat, Université de Provence, Aix Marseille I, 2004.

[19] R.M. Wallace and G.D. Wilk, High-k dielectric materials for microelectronics, Critical Review in Solid State and Material Sciences, 28, 231, 2003.

[20] R.M. Wallace and G.D. Wilk, High-k gate dielectric materials, MRS Bulletin 27, 2002.

[21] R.M.C. de Almeida, I.J.R. Baumvol, Reaction-diffusion in high-k dielectrics on Si, Surface Science Reports 49, 1-114, 2003.

[22] D. Buchanan, E. Gusev, E. Cartier, H. Okon-Schmidt, K. Rim, M. Gribeliuk, A. Mocuta, 80 nm poly-silicon gated n-FETs with ultra-thin Al_2O_3 gate dielectric for ULSI applications, IEDM Technology Digest, p.223, 2000.

[23] S.H. Lo, D.A. Buchanan, Y. Taur and W. Wang, Quantum-mechanical modelling of electron tunnelling current from the inversion layer of ultra-thin oxide nMOSFET's, IEEE Electron Device Letters, 18 (5), 209, 1997.

[24] K. Rim, E.P. Gusev, C.D. Emic, T. Kanarsky, H. chen, J. Chu, J. Ott, K. Chan, D. boyd, V. Mazzeo, B.H. Lee, A. Mocuta, J. Welser, S.L. Cohen, M. Ieong and H.S. Wong, Mobility enhancement in strained Si NMOSFETs with HfO_2 gate dielectrics, VLSI Technology digest of technical paper, p.12, 2002.

[25] O. Weber, M. Cassé, L. Thevenod, F. Ducroquet, T. Ernst, B. Guillaumot and S. Deleonibus, Experimental determination of mobility scattering mechanism in metal gate MOSFETs, ESSDERC Technology Digest, p.379, 2005.

[26] S.J. Lee, C.H. Lee, Y.H. Kim, High-k gate dielectrics for sub-100 nm CMOS technology, Solid- State and Integrated-Circuit Technology, Proceedings. 6th International Conference, Vol. 1, 22-25, 303 – 308, IEEE, 2001.

[27] D.A. Buchanan, Scaling the gate dielectric: material, integration and reliability, IBM Journal of research development, 43 (3), 245, 1999.

[28] Schenk and G. Heiser, Modeling and simulation of tunneling through ultra-thin gate dielectrics, Journal of Applied Physics, 81, 7900, 1997.

[29] R.M. Wallace, G.D. Wilk, Exploring the limit of gate dielectric scaling, Semiconductor international, 153, 2001.

[30] http://www.itrs.net/Links/2007ITRS/ExecSum2007.pdf

[31] J. -P. Doumerc, Rapport interne ICMCB-ST, Programme Nano2008, 2005.

[32] Y. Shang et al, Al_2O_3 thin films deposited on silicon by atomic layer chemical vapor deposition, J. Non Cryst. Sol, 2000.

[33] Y.H. Wu, M.Y. Yang, Albert Chin, W.J. Chen and M Kwei. Electrical characteristics of high quality La_2O_3 gate dielectric with equivalent oxyde thikness of 5 A, IEEE Electron Device Letters, Vol. 21, N°. 7, Juillet 2000.

[34] C. Chaneliere, J.L. Autran, J.P. Reynard, J. Michailos, K. Barla, A. Hiroe, K. Shimomura, A. Kakimoto, Deposition and characterization of ultra-thin Ta_2O_5 layers deposited on silicon from a $Ta(OC_2H_5)_5$ precursor, Mat. Res. Soc. Symp. Proc., p 592, 2000.

[35] G.D. Wilk, R.M. Wallace, J.M. Anthony, Hafnium and Zirconium silicates for advanced gate dielectrics, J Appl. Phys., Vol. 87, p.484, 2000.

[36] B.H. Lee, L. Kang, W.J. Qi, R. Nieh, Y. Jeon, K. Onishi, J.C. Lee, Ultrathin hafnium oxide with low leakage and excellent reliability for alternative gate dielectric application, IEDM Tech. Dig, p.556, 1999.

[37] W.J. Qi, R. Nieh, B.H. Lee, L. Kang, Y. Jeon, K. Onishi, T. Ngai, S. Banerjee, et J.C. Lee, MOSCAP and MOSFET characteristics using ZrO_2 gate dielectric deposited directly on Si, IEDM Tech. Dig, p.605, 1999.

[38] M.E. Hunter, M.J. Reed, N.A. El-Masry, J.C. Roberts, S.M. Bedair, Epitaxial Y2O3 films grown on Si(111) by pulsed-laser ablation, Appl. Phys. Lett., 76, Vol. 1935, 2000.

[39] Théodore NGUYEN, Caractérisation, modélisation et fiabilité des diélectriques de grille à base de HfO_2 pour les futures technologies CMOS, thèse doctorat, l'Institut des Nanotechnologies de Lyon - INSA de Lyon, N° d'ordre 2009-ISAL-0067, 2009.

[40] Loïc BECERRA, Hétérostructures et Dispositifs Microélectroniques à Base d'Oxydes High-k Préparés sur Silicium par EJM, thèse doctorat, Ecole Centrale de Lyon, N° d'ordre : 2008-36, 2008.

[41] Laurent THEVENOD, Etude de la mobilité dans des transistors intégrant un oxyde de grille de forte permittivité et une grille métallique, thèse doctorat, Institut Polytechnique De Grenoble, 2009.

[42] P.W. Peacock and J. Robertson, Band offsets and schottky barrier heights of high dielectric constant oxides, J. Appl. Phys, 92, 8, 2002.

[43] M. Copel, M. Gribelyuk and E. Gusev, Structure and stability of ultrathin zirconium oxide layers on Si(001), Appl. Phys. Lett. 76, 4, 2000.

[44] M. Houssa, A. Stesmans and M. Naili, Trap-assisted tunneling in high gate dielectric stacks, J. Appl. Phys, 87 (12), 8615, 2000.

[45] X. Zhao and D. Vanderbilt, Proceeding of the 2002 MRS Fall Meeting, Vol. 745, p. N7.2.1, 2002.

[46] X. Zhao and D. Vanderbilt, Phys. Rev. B 65, p. 75105, 2002.

[47] S. Ferrari, M. Modreanu, G. Scarel, M. Fanciulli, X-Ray reflectivity and spectroscopic ellipsometry as metrology for the characterization of interfacial layers in high-k materials, Thin Solid Films, 450, 124, 2004.

[48] J.M. Leger, J. Haines, B. Blanzat, Materials potentially harder than diamond: Quenchable high pressure phases of transition metal dioxides, Journal of Material Science Letters, 13, 1688, 1994.

[49] Y.J. Cho, N.V. NGuyen, C.A. Richter, J.R. Ehrstein, B.H. Lee, J.C. Lee, Spectroscopic ellipsometry characterization of high-k dielectric HfO_2 thin films of the high-temperature annealing effects on their optical properties, Applied Physics Letters, 80, 1249, 2002.

[50] J. Wang, H.P. Li, R. Stevens, Hafnia and Hafnia-toughened ceramics, Journal of Materials Science – 27, 5397-5430, 1992.

[51] J. Aarik, H. Mandar, M. Kirm, L. Pung, Optical characterization of HfO_2 thin films grown by atomic layer deposition, Thin Solid Films, 466, 41-47, 2004.

[52] X. Zhao and D. Vanderbilt, Phys. Rev. B 65, 233106, 2002.

[53] Maricela VILLANUEVA-IBANEZ, HfO_2 et $SrHfO_3$ dopés terres rares réalisés par procédé sol gel : analyses structurales, propriétés optiques et potentialités en scintillation, thèse doctorat, Université Claude Bernard – LYON 1, France 2005.

[54] H. Ibégazène, S. Alpérine, C. Diot, Journal of Materials Science, 30, 938, 1994.

[55] G.B. Senft and V.S. Stuibican, Materials Research Bulletin, 18, 1163-1170, 1983.

[56] J.-P. Colinge, C.A. Colinge, Physics of Semiconductor Devices, Springer: Science & Business Media, 31 mai 2002.

[57] C.C. Enz, The EKV model: a MOST model dedicated to low-current and low voltage analogue circuit design and simulation", in Low-power HF microelectronics: a unified approach. Edited by G.A.S. Machado, IEE Circuits and Systems Series 8, the Institution of Electrical Engineers, p.247, 1996.

[58] C. Enz, F. Krummenacher, E.A. Vittoz, An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low current applications. Analog Integrated Circuit and Signal Processing, Vol. 8, N°.1, p.83, 1995.

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

Références bibliographiques (Chapitre II)

[1] Bernard GOSSELIN**,** Application de réseaux de neurones artificiels à la reconnaissance automatique de caractères manuscrits, thèse doctorat, Faculté Polytechnique de Mons, France 1996.

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

[3] W.S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity. Bull Math. Biophysics, 5, 113-115, 1943.

[4] D.O. Hebb, The organisation of behavior, Wiley, New-york, 1949.

[5] F. Rosenblatt, The Perceptron: a Probabilistic Model for Information Storage and Organisation in the Brain, Psychological Review, p. 386- 408, 1958.

[6] M. Minsky and S. Papert, Perceptron, The MIT Press, Cambridge, 1969.

[7] J.J. Hopfield, Neural Networks and Physical Systems with Emergent Collective Computational Abilities, Proceedings of the National Academy of Sciences, p. 460-464, 1982.

[8] D.E. Rumellhart and J.L. Mc Clelland, Parallel Distributed Processing, The MIT Press, Vol. 1 et 2, Cambridge, 1986.

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

[10] Mohamed Yessin AMMAR, Mise en œuvre de réseaux de neurones pour la modélisation de cinétiques réactionnelles en vue de la transposition batch/continu, thèse doctorat, Institut National Polytechnique De Toulouse France 2007.

[11] H. White, artificiel neural networks, Blackwell, New York, 1992.

[12] T. Kohonen**,** Self organized formation of topologically correct feature maps, Biol Cybernetics, Vol. 43, pp. 59-69, 1982.

[13] C. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995.

[14] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks, Vol. 4, pp. 251-257, 1991.

[15] J.-S. Lacroix, S. Terrade, Algorithmes Génétiques, MATH 6414, 17 novembre 2004.

[16] J.-B. Mouret, Concepts fondamentaux des algorithms évolutionnistes, 15 novembre 2005

[17] C. Darwin**,** On the Origin of Species by Means of Natural Selection, John Murray, London, U.K, 1958

[18] J.H. Holland, Adaptation in Natural and Artificial Systems, Ann Arbor: The University of Michigan Press, USA, 1975

[19] T. Back, Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms (Hardcover), Oxford University Press, USA, 1996

[20] D. Whitley, Foundations of Genetic Algorithms 2, US Edition, California, USA, 1993

[21] L. Fogel, A.J. Owens, M.J. Walsh, Artificial Intelligence through Simulated Evolution, Wiley, Chichester, UK, 1966

[22] I. Rechenberg, Cybernetic Solution Path of an Experimental Problem, Royal Aircraft Establishment Library Translation, Farnborough, U.K, 1965

[23] D. Schaffer, Multiple Objective Optimizations with Vector Evaluated Genetic Algorithm, [Conference] // Proceedings of the First International Conference on Genetic Algorithm, Vols. p.93-100, 1985.

[24] A. Konak, D.W. Coit and A.E. Smithc, Multi-objective optimization using genetic algorithms: A tutorial [Journal]. - [s.l.]: Reliability Engineering & System Safety, 2006 йил. - Vol. 91 (9), 992-1007, 2006.

Références bibliographiques (Chapitre III)

[1] F. Rosenblatt, The Perceptron: a Probabilistic Model for Information Storage and Organisation in the Brain, Psychological Review, p. 386- 408, 1958.

[2] J.J. Hopfield. Neural Networks and Physical Systems with Emergent Collective Computational Abilities, Proceedings of the National Academy of Sciences, p. 460-464, 1982.

[3] www.mathworks.com: site officiel de Matlab.

[4] Ph. Lindorfer, C. Bulucea, Modeling of VLSI MOSFET Characteristics Using Neural Networks, simulation of semiconductor devices and processes Vol. 5 Edited by S. Selberherr, H. Stippel, E. Strasser, September 1993.

[5] S. Hatami, M.Y. Azizi, H.R. Bahrami, D. Motavalizadeh, A. Afzali-Kusha, Modeling of drain current characteristics of SOI MOSFET using Neural Networks. Proc. Int. Conf. Microelectron, pp.114–117, 2002.

[6] P. Suresh, L. Sheela, Double Gate Nanoscale MOSFET Modeling by a Neural Network Approach. Proceedings of the 9th WSEAS International Conference on microelectronics, nanoelectronics, optoelectronics, ISBN: 978-954-92600-3-8.

[7] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, New York: Addison Wesley, 1989.

[8] K. A. De Jong, An analysis of the behavior of a class of genetic adaptive systems, Ph.D. thesis, University of Michigan, 1975.

[9] M. Dorigo and U. Schnepf, Genetic-based machine learning and behavior-based robotics: A new synthesis, IEEE Transactions on System, Man, and Cybernetics, Vol. SMC-23, N°. 1, pp. 141-154, 1993.

[10] K.A. De Jong, Learning with genetic algorithms: An overview, Machine Learning, Vol. 3, N°. 2/3, pp. 121-138, 1988.

[11] J.J Grefenstette and. al, Genetic Algorithms for the Traveling Salesman Problem, Proc. Int'l Conference on Genetic Algorithms and their Applications, 1985.

[12] Shraddha Ajit Badnikar, Anuj Vengurlekar, Prashant Kasambe, Electrical Performance Optimization of MOSFETs Using Multi-objective Genetic Algorithms, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 2 Issue 2, March 2013.

[13] M.Emin BAŞAK, Ayten KUNTMAN, Hakan KUNTMAN, MOS parameter extraction and optimization with genetic algorithm. Istanbul university – journal of electrical & electronics engineering, Vol. 9, N°. 2, 1101-1107, 2009.

[14] X. Cai, H. wang, X. Gu, G. Gildenblat, P. Bendix, Application of the genetic algorithm to compact MOSFET Model development and parameter extraction. Nanotech, Vol. 2, 2003.

[15] M. Taherzadeh-Sani, A. Abbasian, B. Amelifard, A. AfzaliKusha, A. MOS Compact I-V Modeling with Variable Accuracy Based on Genetic Algorithm and Simulated Annealing, In Proceedings of the 16th International Conference on Microelectronics, Tunis, Tunisia, December 6-8, pp. 364-367, 2004.

[16] J. Watts, C. Bittner, D. Heaberlin, J. Hoffmann, Extraction of Compact Model Parameters for ULSI MOSFETs Using a Genetic Algorithm, Proceedings of the Second International Conference on Modeling and Simulation of Microsystems, Computational Publications, Cambridge, MA, 176-179, 1999.

[17] M. Keser, K. Joardar, Genetic Algorithm Based MOSFET Model Parameter Extraction, nanotech2000, the Nanotechnology Conference & Trade Show, San Diego, California, U.S.A, March, 2000.

[18] F. Djeffal, T. Bendib, Multi-objective genetic algorithms based approach to optimize the electrical performances of the gate stack Double Gate (GSDG) MOSFET. J. Microelectronics, Vol. 42, pp. 661-666, 2011.

Annexe

ANNEXE I : Mécanismes de conduction dans les oxydes de grilles

Dans une structure métal isolant semi-conducteur idéale, la conductance de l'oxyde est supposée nulle. Dans les cas réels, la conduction des porteurs existe dans les isolants, quand la température ou le champ électrique sont suffisamment élevés. Le champ électrique dans un diélectrique, lorsqu'il est polarisé, est décrit par la relation :

$$\mathbf{E_i = E_S\,(\varepsilon_s / \varepsilon_i)}$$

Ou E_i et E_s sont les champs électriques dans l'isolant et le semiconducteur ; ε_i et ε_s les permittivités correspondantes.

1. Paramètres des équations :

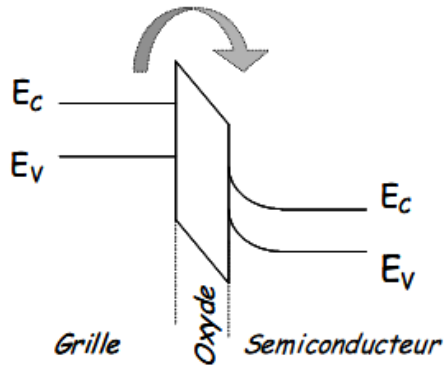
E : le champ électrique appliqué
C_1 : constante liée à la masse effective
E_0 : constante liée à la hauteur de barrière
ϕ_B : hauteur de barrière
m* : masse effective

V = Ed
d = épaisseur de l'isolant
A* : Constante effective de Richardson
ε_i : permittivité de l'isolant
$a = \sqrt{q/4\pi\varepsilon id}$

2. Quelques exemples de mécanismes de conductions :

2.1 Emission Schottky : conduction thermoïonique :

Le mécanisme dépend principalement du passage des électrons, d'énergie cinétique élevée, entre la bande de valence et la bande de conduction de l'oxyde de grille (voir schéma).



La densité des courant des électrons J est exprimée par :

$$J = A^* T^2 exp\left[\frac{-q\left(\phi_B - \sqrt{qE/4\pi\varepsilon_i}\right)}{KT}\right] \sim T^2 \exp\left(+\frac{a\sqrt{V}}{T} - q\phi_B/KT\right)$$

2.2. Conduction tunnel direct :

Le courant dépend des électrons qui passent directement à travers la barrière de l'oxyde si celui est suffisamment mince. L'émission tunnel est causée par l'ionisation du champ des électrons piégés dans la bande de conduction ou par les électrons qui traversent l'énergie de Fermi du métal pour aller dans la bande de conduction de

l'isolant, par effet tunnel. Ce mécanisme a la plus forte dépendance vis-à-vis de la tension appliquée et est essentiellement indépendante de la température.

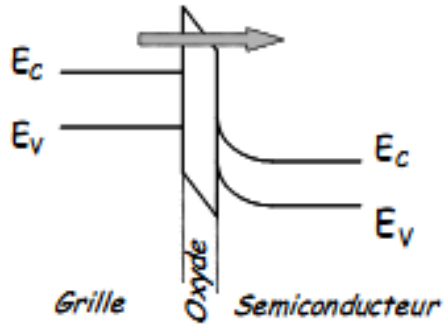
La densité de courant J prend alors la forme:

$$J = E^2 exp\left[-\frac{4\sqrt{2m^*}(q\phi_B)^{3/2}}{3q\hbar E}\right] \sim V^2 \exp(-b/V)$$

2.3. Conduction tunnel de type Fowler-Nordheim :

Le courant tunnel est assisté par le champ électrique. En présence de forts champs, les électrons ont une probabilité non nulle de passer à travers une barrière triangulaire d'épaisseur inférieure à celle de l'oxyde. Pour un champ électrique élevé : le transport du courant dépend fortement du champ électrique, et la densité de courant est décrite par la relation :

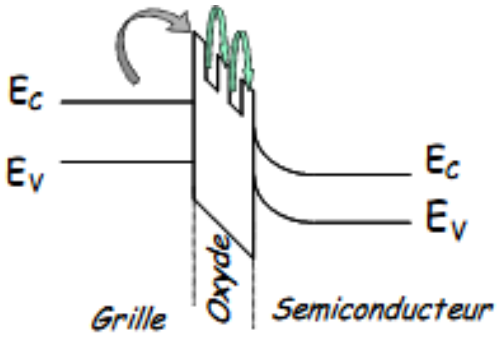
$$J = \frac{q^2}{16\pi\hbar\phi_B} E^2 exp\left[-\frac{4\sqrt{m^*}(q\phi_B)^{3/2}}{3q\hbar E}\right]$$



2.4. Courant de conduction type Pool-Frenkle :

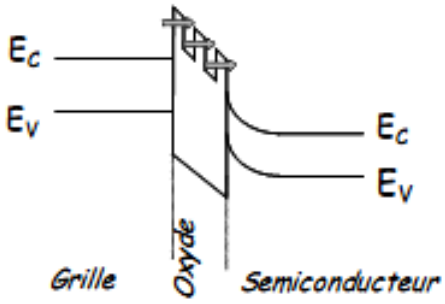
Les électrons ont une énergie cinétique suffisante pour passer par-dessus la barrière d'énergie de l'oxyde. Ils migrent ensuite dans la bande de conduction de l'oxyde jusqu'à ce qu'ils soient capturés par un autre défaut. Ce mécanisme de conduction est dû à l'augmentation du champ électrique par l'excitation thermique des électrons piégés dans la bande de conduction. Dans ce cas la hauteur de barrière est la profondeur du puit de potentiel et l'expression de J est de la forme suivante :

$$J = Eexp\left[\frac{-q\left(\phi_B - \sqrt{qE/4\pi}\varepsilon_i\right)}{KT}\right] \sim V \exp(+\frac{2a\sqrt{V}}{T} - q\phi_B/KT)$$



2.5. Conduction type Hopping :

Le mécanisme de type Hopping regroupe de nombreux mécanismes de conduction, comme le Poole-Frenkle. Dans ce cas, l'énergie des électrons est inférieure au maximum d'énergie entre deux défauts voisins. Les électrons traversent par effet tunnel la barrière d'énergie entre deux pièges de l'oxyde.



ANNEXE II : Jonction métal-oxyde-semiconducteur à l'équilibre

Cette jonction permet de comprendre les transistors MOSFET qui sont les composants les plus utilisés en électronique.

1- Jonction Métal-vide-semiconducteur.

1.1 Structure à l'équilibre avec un semiconducteur dopé N.

On considère un barreau de métal et un barreau de semi-conducteur dopé N que l'on tient à distance assez grande l'un de l'autre. On va étudier ce qui se passe au niveau des deux faces les plus proches quand on relie par un fil d'impédance nulle les deux extrémités les plus éloignées des barreaux.

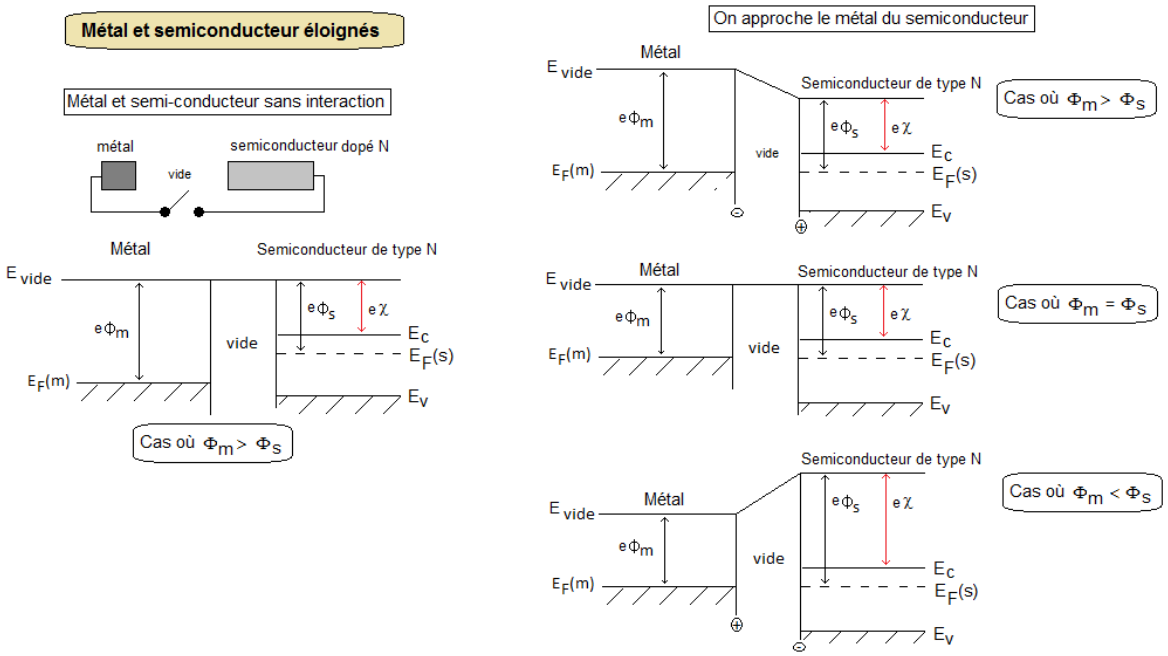
● Sous-Systèmes éloignés et non reliés électriquement:

Tant que les deux systèmes sont indépendants (éloignés et non reliés électriquement), on va supposer que seul le niveau d'un électron sans énergie cinétique est commun (vide), mais les autres niveaux sont indépendants.

● Sous-Systèmes éloignés et reliés électriquement:

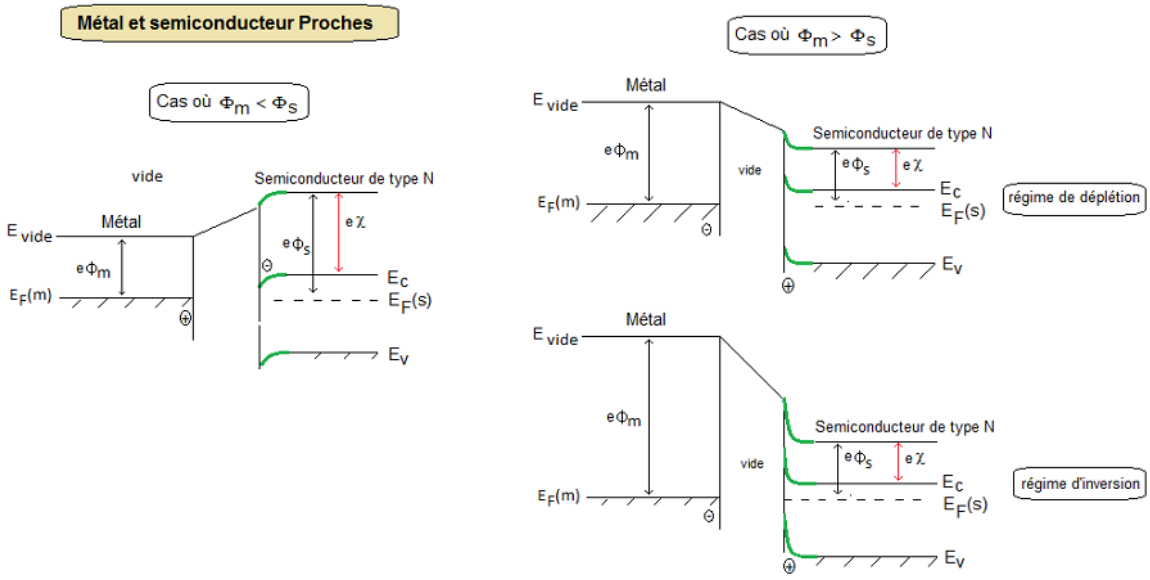
Dès que l'on relie électriquement les deux barreaux, ils ne constituent plus qu'un seul système et à l'équilibre, les niveaux de Fermi vont s'égaliser. Il en résulte l'apparition d'une différence de potentiel et donc d'un champ électrique dans le vide séparant les deux barreaux qui se font face. Le sous-système dont le niveau de Fermi a été augmenté par rapport à l'autre va recevoir des électrons supplémentaires, alors que l'autre sous système va se retrouver avec un déficit d'électrons. Les deux sous-systèmes se retrouvant séparés par un milieu isolant, les charges vont s'accumuler sur les deux surfaces en regard, plus ou moins profondément, suivant le matériau considéré.

Si on adopte un modèle de capacité plane, sachant que la différence de potentiel V_d entre les deux faces reste constante quelle que soit la distance (elle ne dépend que de la différence entre le travail de sortie du métal et le travail de sortie du semiconducteur), et que la capacité augmente quand on rapproche les barreaux, la charge qui apparaît sur les faces en regard des deux barreaux sera d'autant plus faible que les barreaux sont éloignés. Tant que les barreaux sont assez éloignés et on pourra donc négliger les zones de charges d'espace dans les barreaux.



● Sous-systèmes proches et reliés électriquement:

Ce ne sera plus le cas quand les barreaux seront proches. Dans ce cas, les zones de charges d'espace vont influer sur les niveaux d'énergie.



- Quand Φm = Φs, aucune charge n'apparaît et le semiconducteur est en régime de bandes plates.

- Quand Φm<Φs, le niveau d'énergie des électrons du métal est inférieur à celui des électrons de la bande de conduction du semiconducteur. Le potentiel des électrons du métal est supérieur à celui des électrons de la bande de conduction du semiconducteur. On aura alors des charges positives dans le métal à la surface et négatives dans le semiconducteur sur une épaisseur plus importante (cf ordres de grandeur des densités d'états considérées). Les charges dans le métal proviennent du départ des électrons, alors que dans le semiconducteur, il s'agit d'une accumulation d'électrons. Les bandes se courbent vers le bas compte tenu du signe de la tension de diffusion.

- Quand Φm>Φs, le niveau d'énergie des électrons du métal est supérieur à celui des électrons de la bande de conduction du semiconducteur. Le potentiel des électrons du métal est inférieur à celui des électrons de la bande de conduction du semiconducteur. Des charges positives apparaissent à la surface du semiconducteur et des charges négatives à la surface du métal. Dans le métal, il s'agit d'une accumulation d'électron. Dans le semiconducteur, il s'agit d'un départ d'électrons.

Si la différence de potentiel reste assez faible entre les deux faces (Φm <Φs, mais pas trop) pour que la densité de trous reste inférieure à celle des électrons, le semiconducteur reste de type N au voisinage de la jonction et l'essentiel des charges positives sont liées aux ions donneurs. L'accumulation de charges positives correspond alors à un régime de déplétion. Les charges sont alors distribuées sur une distance importante (jusqu'à qq 100 nm).

Si la différence de potentiel devient suffisamment importante pour que la densité des trous dépasse celle des électrons, le semiconducteur devient de type P au voisinage de la surface du semiconducteur qui est dit en régime d'inversion.

En augmentant encore davantage la différence de potentiel (Φm <<Φs), le nombre de trous libres devient supérieur au nombre d'ions donneurs et on est alors en régime de forte inversion. Les charges sont alors concentrées à la surface du semiconducteur (qq nm).

1.2 Structure à l'équilibre avec un semiconducteur dopé P.

On suppose directement que les barreaux sont assez rapprochés.

Quand Φm = Φs, aucune charge n'apparaît et le semiconducteur est en régime de bandes plates.

Quand Φm>Φs, on a accumulation de charges positives dans le semiconducteur et d'électrons dans le métal.

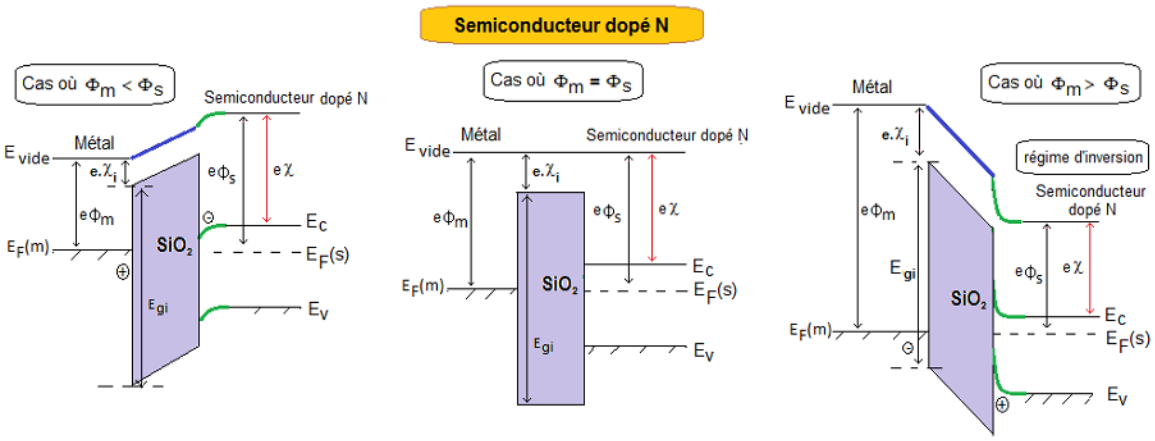
Quand Φm <Φs, si la différence reste assez faible, on a une charge négative dans le semiconducteur qui résulte d'une déplétion (ions accepteurs), Si la différence devient plus marquée, on finit par obtenir un régime d'inversion (le semiconducteur devient de type N au voisinage de la surface).

2- Jonction Métal-oxyde-semiconducteur.

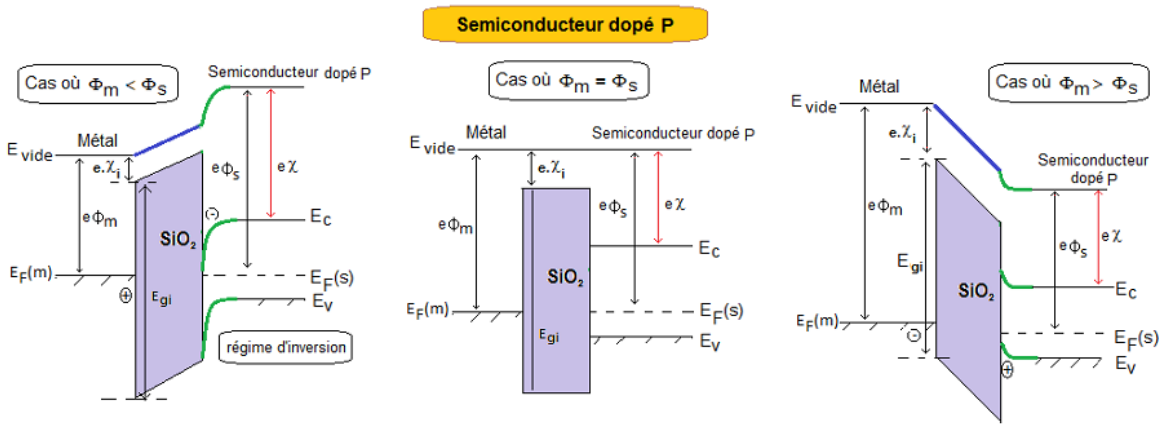
Dans les jonctions présentant un isolant, ce dernier est en général de l'oxyde de silicium (SiO_2).Cet isolant présente un gap E_{gi} et une affinité électronique χ_i.

2.1. Jonction à l'équilibre.

Les figures sont alors légèrement modifiées par rapport au cas précédent. Si on relie les deux faces les plus éloignées des semiconducteurs par un fil conducteur avec de l'oxyde de silicium comme isolant entre les deux autres faces, on se retrouve dans la configuration suivante pour un semiconducteur dopé N:



Pour un semiconducteur dopé P, on a de la même façon les configurations suivantes:



2.2. Jonction polarisée.

● En présence de polarisation, contrairement à la jonction métal-semiconducteur, la zone isolante ne permet pas de comportement ohmique. Excepté au voisinage de l'interface isolant-semiconducteur, le niveau des bandes va rester constant dans le semiconducteur.

● Si on applique une différence de potentiel V_G au barreau métallique appelé grille, par rapport au semiconducteur, la polarisation va alors contribuer à atténuer ou amplifier les effets vus précédemment en l'absence de polarisation. En jouant sur la

valeur de V_G, on pourra passer d'un régime d'accumulation à un régime d'inversion, ou inversement. Le régime de bande plates est atteint quand $V_G = \Phi m - \Phi s$.

● Pour une tension de grille V_G positive:

On va décaler le niveau de Fermi du métal vers le bas par rapport à celui du semiconducteur.

Pour le semiconducteur de type N, si $\Phi m < \Phi s$, on va augmenter l'effet d'accumulation. Si $\Phi m > \Phi s$, on va diminuer le niveau d'inversion ou de déplétion puis passer en accumulation.

Pour le semiconducteur de type P, si $\Phi m < \Phi s$, on va augmenter le niveau d'inversion.

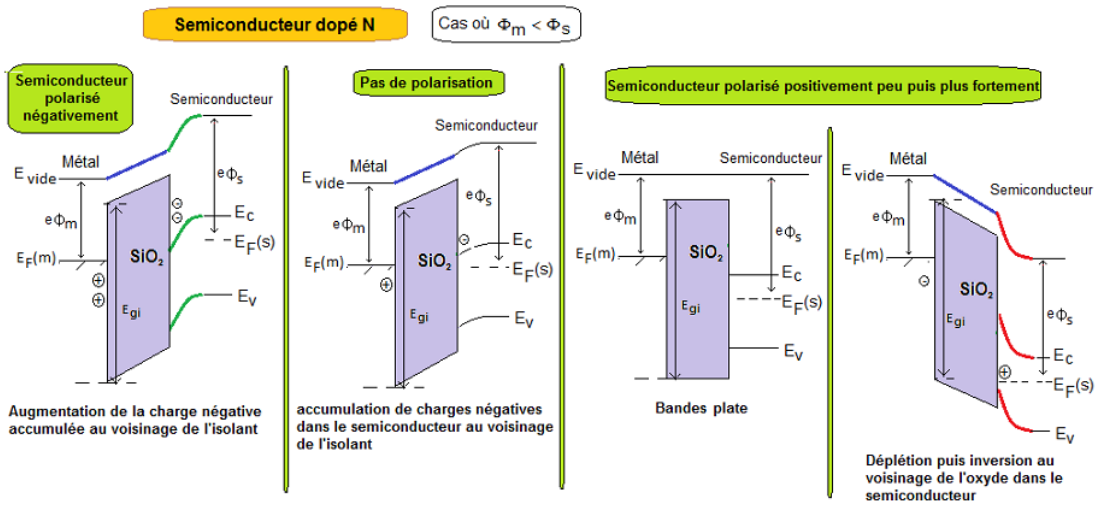
Si $\Phi m > \Phi s$, on va diminuer l'effet d'accumulation puis passer en déplétion puis en inversion pour le semiconducteur de type P.

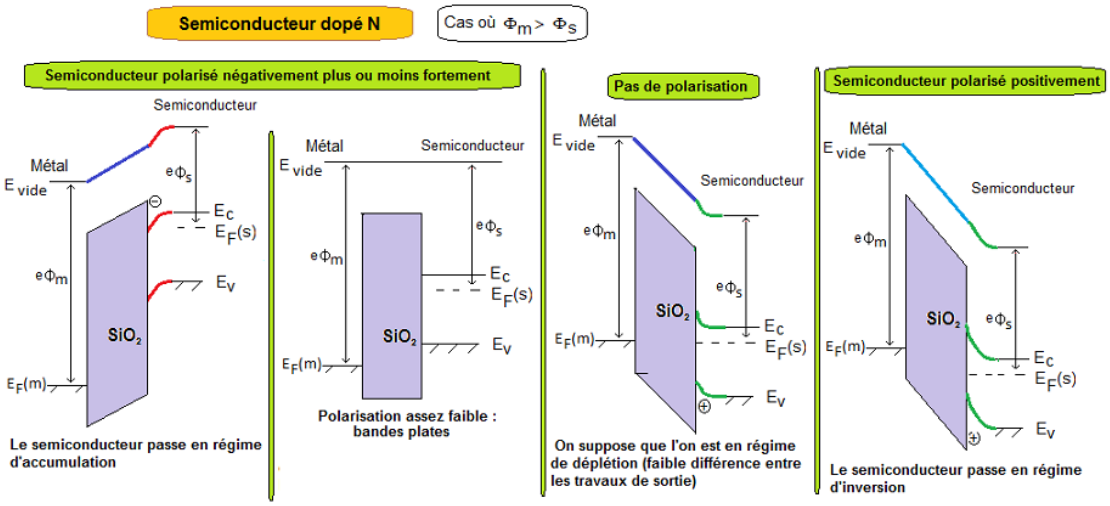
● Pour une tension de grille V_G négative:

On va décaler le niveau de Fermi du semiconducteur vers le bas par rapport à celui du métal.

Pour le semiconducteur de type N, si $\Phi m < \Phi s$, on va diminuer l'effet d'accumulation puis passer en déplétion puis en inversion. Si $\Phi m > \Phi s$, on va augmenter le niveau d'inversion.

Pour le semiconducteur de type P, si $\Phi m < \Phi s$, on va diminuer le niveau d'inversion pour passer en déplétion puis en inversion. Si $\Phi m > \Phi s$, on va augmenter l'effet d'accumulation pour le semiconducteur de type P.





2.3. Bilan.

Pour une même structure, en passant d'un régime d'inversion à un régime d'accumulation, ou inversement, on peut créer, dans le semiconducteur, au voisinage de la couche d'oxyde un gaz d'électrons ou de trous assez dense. Ainsi, en agissant sur la tension de grille, on pourra contrôler la densité de porteurs dans le semiconducteur au voisinage de l'isolant et envisager de contrôler le courant qui circule parallèlement à l'isolant, à son voisinage dans le semiconducteur... C'est ce qui sera fait dans les transistors MOS.

Résumé:

La miniaturisation des transistors Métal-Oxyde-Semi-conducteur à effet de champ (MOSFET) ne suffit plus à satisfaire les spécifications de performances de l'International Technology Roadmap for Semiconductors (ITRS). Dans ce cadre, les oxydes de grille MOS atteignent des épaisseurs limites qui les rendent perméables aux courants de fuite. Une solution est de remplacer le SiO_2 par un matériau de permittivité plus élevée. Une approche analytique basée sur un prédicteur neuronal a été développée dans le cas du transistor MOSFET à permittivité élevé (HfO_2). Cette dernière nous a permis de prévoir l'évolution de courant du drain en fonction des différents paramètres (tension du drain, tension de grille, longueur du canal et épaisseur d'oxyde). Les résultats obtenu par l'approche neuronal, elle était vérifié par la technique GA, est les deux méthodes présentent une meilleure stratégie conventionnelle d'extraction des paramètres, en terme de convergence elles fournissent des solutions optimales globales.

MOTS CLES: Microélectronique, MOSFET, technologie CMOS, SiO_2, HfO_2, réseau de neurone, algorithme génétique.

Abstract:

The miniaturization of Metal-Oxide-Semiconductor field effect transistors (MOSFET) is insufficient to satisfy the performance specifications of International Technology Roadmap for Semiconductors (ITRS). In this context, the gate oxide thicknesses reach limitations that make them permeable to leakage currents. The solution is to replace SiO_2 by high-k dielectrics (for example HfO_2 oxide). An analytical approach based on a neural predictor has been developed in the case of MOSFET has high permittivity (HfO_2). The latter allowed us to predict the evolution of the drain current as a function of various parameters (drain voltage, gate voltage, and channel length and oxide thickness). The results obtained by the neural approach, it was verified by the GA technique, is the two methods have a better conventional extraction strategy parameter, in terms of convergence they provide global optimal solutions.

KEY WORDS: Microelectronic, MOSFET**,** CMOS technology**,** SiO_2, HfO_2, neural network, Genetic Algorithms.

الملخص:

التصغير من الترانزستورات (المعدنية - أكسيد- أشباه الموصلات - تأثير الحقل) (MOSFET) ليست كافية لتلبية مواصفات الأداء بالنسبة لي ITRS (l'International Technology Roadmap for Semiconductors). وفي هذا السياق، أكاسيد البوابة تصل إلى سمك القيود التي تجعلها قابلة للاختراق لتيارات التسرب. من بين الحلول هو استبدال SiO_2 بأكسيد له سماحية أعلى. وقد تم تطوير نهج تحليلي على أساس الشبكة العصبية الاصطناعية في حالة ترانزستور ذو سماحية عالية (ثاني اكسيد الهافنيوم HfO_2), يسمح هذا الأخير لنا بالتعرف علي تغيرات التيار وفقا لمعايير مختلفة (الجهد الكهربي، جهد البوابة، وطول القناة و سمك الأكسيد). النتائج التي حصل عليها النهج العصبي، تم التحقق منها بواسطة تقنية الخوارزميات الجنية. وكلتا الطريقتين هي أفضل استراتيجية لدراسة خصائص التيار كما أنها توفر الحلول المثلى بشكل عام.

الكلمات المفتاحية : ميكروالكترونيك, الترانزستور, اكسيد سليسيوم, الهافنيوم, الشبكة العصبية الاصطناعية, الخوارزميات الجنية.