



Développement de modèles pour l'évaluation des performances circuit des technologies CMOS avancées sub-20nm

Joris Lacord

► To cite this version:

Joris Lacord. Développement de modèles pour l'évaluation des performances circuit des technologies CMOS avancées sub-20nm. Autre. Université de Grenoble, 2012. Français. NNT : 2012GRENT073 . tel-00820068

HAL Id: tel-00820068

<https://theses.hal.science/tel-00820068>

Submitted on 3 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Nano Electronique et Nano Technologies**

Arrêté ministériel : 7 août 2006

Présentée par

Joris LACORD

Thèse dirigée par **Gérard GHIBAUDO** et
co-encadrée par **Frédéric BOEUF**

préparée au sein de l'**Institut de Microélectronique,
Electromagnétisme et Photonique (IMEP-LAHC)** et
STMicroelectronics, Crolles.

dans l'**École Doctorale Electronique, Electrotechnique et
Automatique et Traitement du Signal (EEATS).**

**Développement de modèles pour
l'évaluation des performances circuit
des technologies CMOS avancées sub-
20nm.**

Thèse soutenue publiquement le **18 décembre 2012**,
devant le jury composé de :

Mr Francis CALMON

Pr, INSA de Lyon,

Président

Mr Jean Luc AUTRAN

Pr, Université de Aix-Marseille,

Rapporteur

Mr Jean Michel SALLESE

Pr, EPFL Lausanne (Suisse),

Rapporteur

Mr Frédéric BOEUF

Ing, STMicroelectronics, Crolles,

Co-encadrant

Mr Gérard GHIBAUDO

Dr, CNRS Alpes – IMEP-LAHC/INPG,

Directeur de thèse

Mr Thierry POIROUX

Ing, CEA-LETI, Grenoble,

Invité



Remerciements

Ces travaux de thèse ont été effectués en convention CIFRE entre STMicroelectronics et le laboratoire IMEP-LAHC, je tiens donc à remercier leurs dirigeants ainsi que toutes les personnes ayant rendu ces travaux possibles. Je tiens ensuite à remercier les membres du jury de ma soutenance, Francis Calmon en tant que président, Jean Luc Autran et Jean Michel Sallese en tant que rapporteurs et enfin Thierry Poiroux en tant qu'invité (de marque) pour avoir lu ce long manuscrit, d'y avoir posé un regard critique et bienveillant, mais également pour leurs questions auxquelles j'aurais sans doute pu répondre de manière moins concise..!

Un grand merci à Gérard Ghibaudo, mon directeur de thèse, mais aussi celui de nombreux autres thésards et également directeur de l'IMEP LAHC. Difficile d'être original pour remercier Gérard, tant les avis de ses thésards, anciens et nouveaux, sont unanimes. Comme tout le monde donc, je vais saluer son excellent encadrement, son impressionnante disponibilité malgré ses nombreuses activités et sa réactivité. Il est en effet rare qu'un mail envoyé à Gérard n'ait pas de réponse dans l'heure qui suit et qu'un papier soit lu et corrigé en plus d'une journée (deux pour un chapitre de manuscrit). Notons que l'échelle de temps à ST est tout autre. Je vais également saluer le sens physique aigu de Gérard dont une simple phrase peut nécessiter plusieurs semaines de réflexion pour la comprendre... L'exemple du fameux « C/2 » illustre parfaitement ces propos. Enfin merci à Gérard pour les nombreux coups de main lors de mes travaux de modélisation (électrostatique, quantique, courant et capacités parasites) mais également de simulation et de m'avoir fait découvrir les joies de la « TCAD du pauvre ».

L'IMEP-LAHC ne se limitant pas à Gérard, je tiens à remercier Raphaël Clerc et Quentin Rafhay pour nos discussions et nos « monthly » absolument pas mensuels concernant les III-V pour l'essentiel. Merci à eux deux également pour m'avoir impliqué dans l'organisation de la conférence ULIS 2012 à Grenoble (hormis la préparation des sacs). Enfin, merci à Annaick Moreau pour l'organisation de mes missions.

Basculons coté ST, où j'étais intégré dans l'équipe *advanced devices*, dans le service *Process Integration*. Je vais commencer par remercier Thomas Skotnicki, manager de l'équipe *advanced devices* lors des premiers mois de ma thèse. Merci Thomas d'avoir inventé MASTAR et la VDT sans quoi ma thèse aurait été très différente, si elle avait existé, ainsi que pour nos quelques discussions sur le courant et la tension de seuil de transistors divers. En réponse à la question piège, à laquelle je n'ai souvent que partiellement répondu, MASTAR a d'abord signifié **Model for Analog and digital Simulation of mos TrAnsistoRs** (cf papier IEDM de 1994), puis **Model for Assessments of cmoS Technologies And Roadmap** à peu près lorsqu'il a été adopté par l'ITRS pour l'établissement des roadmaps, début 2005. Merci également à Olivier Noblanc, manager du service *Process Integration* pour sa sympathie et sa bienveillance. Un grand merci à Estelle Di Rago, assistante du service, qui a forcément la solution à nos soucis administratifs et sans qui je n'aurais pas pu relier mon manuscrit. Je remercie également le centre de doc de ST Crolles (*Crolles Doccenter*) pour sa réactivité et pour m'avoir fourni de très anciens livres de mathématiques appliquées, qui m'ont, avec le recul, été indispensables.

Bien que des travaux de thèse se doivent d'être réalisés en grande partie seul, de surcroît sur un sujet de modélisation, ils se doivent également d'être le fruit d'interactions. Sachant que tout oubli fortuit sera sans doute volontaire, et dans un ordre certainement chaotique, en *Process Integration*, je remercie Manu Josse de m'avoir permis de mettre un premier pied à ST en m'accueillant en tant que stagiaire, de m'avoir encadré avec rigueur et pédagogie, de m'avoir familiarisé avec l'ambiance crolloise (notamment son vocabulaire « franglisé » difficilement accessible sans mode d'emploi) et d'avoir assisté à ma soutenance (chose promise trois ans plus tôt lors de son départ pour *Fishkill*). Un peu dans la même optique, merci à Michel Haond qui fut le manager de l'équipe où j'ai

effectué ce même stage et pour les discussions pendant ma thèse à propos du benchmark FDSOI vs Trigate où nous avons beaucoup abordé la fameuse question « normalisé par finpitch ou par Welec ? » (les puristes comprendront). Toujours en *Process Integration*, je remercie Olivier Weber et Nicolas Planes pour avoir répondu à mes petites questions sur l'architecture et la technologie FDSOI et Perrine Batude du LETI pour les discussions sur l'intégration 3D, le support pour l'étude et pour la rédaction du papier qui en a découlé (ce qui m'a permis de découvrir les US). Coté T2D je remercie David Hoguet pour les discussions et l'aide apportée pour les simulations numériques 3D ; Marie Anne Jaud (du LETI) pour la TCAD FDSOI et le support pour le modèle ; Clément Tavernier et François Wacquant pour le support TCAD et Denis Rideau et Davide Garetto pour les simulations quantiques. Toujours en T2D, mais cette fois côté modeling, je remercie Sophie Puget pour avoir rendu un peu moins longue (voire pénible) la formation ELDO et pour le temps qu'elle m'a consacré pour résoudre mes problèmes existentiels avec ELDO, UNIX et les netlists ; Thierry Poiroux pour les discussions sur le Trigate, la modélisation compacte, la modélisation du FDSOI, la relecture express (une semaine) de l'intégralité de ce manuscrit et pour la détection des dernières, du moins je l'espère, coquilles ; Patrick Scheer pour les discussions sur le Trigate, notons que Patrick a eu le plaisir et le privilège d'être le premier utilisateur « extérieur » de MASTAR VA (et le seul à ce jour); Olivier Rozeau (LETI) pour m'avoir aiguillé lors du début de l'implémentation de MASTAR VA et de m'avoir permis d'éviter quelques pièges; Clément Charbuillet pour le support express avec UNIX et SBENCH; Hervé Jaouen pour son regard critique sur mes travaux et « ses fameuses questions qui tuent » dont il a le secret; André Juge pour nos discussions techniques et pointues ainsi que pour l'aide apportée pour inclure la variabilité dans mes simulations de circuit; puis Vincent Quenette, Xavier Monagner, Salim El Gouhli, Michel Minondo, Alexandre Dartigues et Jean Rémi pour des coups de mains tant divers que variés. Merci également à Lidwine Chaize et Jean Claude Marin pour les discussions sur le PEX 22-Trigate, Anne Laure Mareau pour les netlists SRAM et Andrea Carmine Valente pour le support avec UNIX. Pour en finir avec T2D, un grand merci à Steve Colquhoun pour le temps qu'il m'a consacré pendant la partie la plus ingrate de ces trois ans, c'est-à-dire le débbugage de MASTAR VA. Steve m'a en effet appris à comprendre ce qu'essayait de me dire ELDO (tout comme sa réciproque) et pour les scripts, macro, simulations, extractions et netlists qui m'ont énormément simplifié la vie. Coté caractérisations électriques, merci à Antoine Cros, Cheikh Diouf et Imed Ben Akkez pour m'avoir fourni quelques mesures sur des vrais dispos qui m'ont permis de valider certains modèles (notamment I(V) et effet en température).

Je vais maintenant clore mes remerciements par l'équipe dans laquelle j'étais intégré: les modules avancés. Equipe qui est passée du statut de moribond village de quelques irréductibles gaulois où régnait un optimisme général et débordant au début de ma thèse à aujourd'hui une équipe plus nombreuse (plus de jeunes thésards, mais aussi des vieux), très pluridisciplinaire et menant des projets tant amont que de développement industriel. Je commencerai alors par remercier le chef de cette « grande » équipe, Frédéric Bœuf, qui fut surtout mon encadrant pendant ces trois ans (+2 mois d'intérim). Réflexion faite, je vais plutôt commencer par ne pas le remercier pour avoir été absent pour mes deux premiers jours: premier jour en tant que thésard (qui a en fait duré plusieurs semaines) et premier jour en tant qu'intérimaire. Je ne le remercie également pas pour m'avoir mis en retard à cause d'une sordide histoire de choix de restaurant pour ma première présentation en conférence, qui plus est dans un pays où le retard est difficilement envisageable (-> le Japon). A part ça, je remercie Fred pour son très bon encadrement, son humour (gras ?), sa bonne humeur, sa crainte du sport (des sport extrêmes à la marche à pied) et de la neige, son gros côté geek qui me dépasse souvent, sa connaissance approfondie de toutes les séries TV existantes, ayant existé ou allant exister (petit conseil: ne jamais lui parler d'une série ou d'un film que vous comptez voir prochainement) et le fait qu'il ait vu les deux trilogies Star Wars au cinéma dès leur sortie (oui, même l'épisode IV) ont contribué à alimenter nos nombreuses conversations

absolument pas professionnelles. Je salue également son « nez » pour remettre instantanément mes tentatives de présentation et de papier « dans l'ordre », sa grande maîtrise de Powerpoint, son aptitude à rendre les choses visuelles et sa capacité à modéliser divers phénomènes physiques par des droites (ou par des fonctions de Fermi dès que la continuité fait partie du problème). Le second élément historique de l'équipe est Stephane Monfray, que je remercie pour sa bonne humeur, les conversations surréalistes au café mais également pour nos discussions sur les dispos fumeux, pour m'avoir attendu chaque vendredi matin pour nos fameux weekly mastar et pour ses yeux à rayons X indispensables pour lire et analyser les photos TEM avec ou sans transparence (même si ce n'est pas coupé bien droit). Si je garde comme logique, l'ordre d'arrivée dans l'équipe, je devrais alors parler de Jean Luc. Je remercie donc Nathalie Vulliet, alias Nath, alias mamie (affectueusement) pour m'avoir payé un nombre conséquent de cafés et d'avoir donc, par ce biais, contribué au financement de ma voiture. Merci également à Nath d'avoir supporté mon humour douteux concernant, notamment, son grand âge, d'avoir toléré mon adresse toute relative avec la poubelle et d'avoir organisé la soirée barbecue à la frontale / gentiane-génépi qui reste mémorable. Dans le même bureau, merci à Aurélie Souhaite pour sa bonne humeur et son humour. Merci également à notre minorité visible, Charles Baudot (et sa petite famille), pour sa bonne humeur en pause-café et les soirées bière-pizza-série mais surtout ragots et sa faculté à changer de nationalité quand ça l'arrange. Merci également à Pierre Perreau, qui a quitté l'équipe à peu près au milieu de la thèse pour rejoindre l'obscur monde de l'intégration, pour sa sympathie, son humour et sa discrétion. Pour développer un peu le côté discret et introverti de Pierre, il faut bien comprendre que travailler à quelques bureaux de lui, comme cela a été mon cas, c'est un peu comme travailler à côté d'un marteau piqueur: du bruit en permanence et on a l'impression qu'il manque quelque chose quand il n'est plus là. Claire Fenouillet-Béranger, pour sa gentillesse, ses ragots et les discussions sur le FDSOI. Kinaou Hervé, qui a été mon premier voisin de bureau qui s'est évertué à toujours remettre mes stylos bien perpendiculaire au bord du bureau et qui dû supporter mon désordre ambiant. Je remercie Kinaou (c'est Kinaou son prénom) pour le temps qu'il m'a accordé pour m'expliquer le code MASTAR et pour avoir fait preuve d'une grande patience lors de nos séances de debugages. Mes travaux de thèse ont également fait l'objet de l'encadrement d'un stage de master 2. Malgré ses nombreuses tentatives de négociations et le fait qu'il ait essayé à de nombreuses reprises de m'enfumer comme un marchand de tapis, je remercie ma victime Madjid Akbal pour sa sympathie, son travail (je lui dois en partie le modèle de l'électrostatique du FDSOI du chapitre II de ce manuscrit), sa persévérance et pour avoir testé ma patience ainsi que mon endurance (surtout lors des nombreuses corrections du rapport de stage) jusqu'à leurs limites respectives. J'espère l'avoir suffisamment armé pour une bonne poursuite en thèse. Je vais maintenant saluer les thésards qui ont rejoint l'équipe durant ces trois ans Rémi, Onoriu, Boris (qui restera Pécore pour moi), Léopold, Arthur, Thomas et Gaspard. Je les remercie, modérément, pour leurs interminables présentations au weekly et leur souhaite bon courage pour terminer leurs thèses. J'ai gardé ces dernières lignes pour remercier Jean Luc Huguenin, qui a commencé puis terminé sa thèse un an avant moi. Après des débuts difficiles dans les locaux reculés du rez-de-chaussée du bâtiment B1-bis, nous nous sommes, après un déménagement et un changement de bâtiment, retrouvés dans l'adversité en partageant le même bureau. Bureau qu'on peut qualifier, sans exagération, d'hostile. Ce fut malgré tout nos débuts, favorisés par l'élaboration conjointe du modèle de seuil double grille. Comment résumer par des mots ce qui s'est passé depuis, difficile certes, mais nous avons réussi à poursuivre nos travaux respectifs de manière parfois conjointe dans la bonne humeur (et c'est un euphémisme). Je remercie également Jean Luc pour les à cotés, comme les nombreux trajets en bus mais également pour les éprouvantes soirées au Denfer, parfois riches en découvertes (surtout quand nous avons eu le malheur de traverser la rue). En bref, merci à Jean Luc pour les excellents moments que nous avons passé ensemble, à ST et

ailleurs, son soutien, pour avoir écouté mes nombreuses ralleries (via communicator et *IRL*), pour la patience dont il a fait preuve pour m'expliquer les « trucs de jeunes » et pour son côté gentil agneau que je crains lui avoir fait perdre (j'espère ne pas en être l'unique cause).

Enfin, je remercie mes parents, grands-parents et sœurs pour leur soutien et pour ne pas m'avoir posé trop de questions. Et pour terminer pour de bon ces remerciements, je remercie Rachel, ma très chère moitié, d'une part d'exister et d'autre part de m'avoir supporté et de continuer à le faire dans la vie de tous les jours.

Sommaire

REMERCIEMENTS	1
SOMMAIRE.....	7
INTRODUCTION GENERALE.....	11
CHAPITRE I : LE TRANSISTOR MOSFET: FONCTIONNEMENT, MINIATURISATION ET ARCHITECTURES.....	15
I.A. LE TRANSISTOR MOSFET IDEAL.....	16
I.A.1. Principe de fonctionnement.....	16
I.A.2. L'architecture conventionnelle sur substrat massif (BULK).....	16
I.A.3. La capacité MOS.....	18
I.A.4. La tension de seuil	21
I.B. LE TRANSISTOR MOSFET REEL.....	22
I.B.1. Régime sous le seuil : I_{off} et Pente sous le seuil	23
I.B.2. Régime au-delà du seuil : I_{on}	25
I.B.3. La mobilité effective μ_{eff} à fort champ de grille	28
I.B.4. Délai, Inverseur et Courant Effectif I_{eff}	29
I.C. REDUCTION DES DIMENSIONS	31
I.C.1. Règles de miniaturisation	31
I.C.2. Réduction de la longueur de grille	32
I.C.3. Réduction de l'épaisseur d'oxyde de grille	36
I.C.4. Réduction de la distance grille-contact.....	38
I.D. SOLUTIONS TECHNOLOGIQUES POUR POURSUIVRE LA COURSE A LA MINIATURISATION	40
I.D.1. Ingénierie de jonction	40
I.D.2. Empilement diélectrique haute permittivité-grille métallique.....	42
I.D.3. Résistance d'accès.....	44
I.D.4. Contact à prise ou Contact ruban ?.....	45
I.D.5. Epitaxie facettée et espaceur faible permittivité	45
I.D.6. Amélioration du transport.....	47
I.E. VERS DE NOUVELLES ARCHITECTURES	50
I.E.1. FDSOI	51
I.E.2. Double grille planaire.....	53
I.E.3. FinFET / Trigate	53
I.E.4. Nanofils.....	55
I.E.5. Matériau III-V et germanium	56
I.F. CONCLUSION DU CHAPITRE.....	57
CHAPITRE II: MODELISATION ANALYTIQUE DES CARACTERISTIQUES STATIQUES DES DIFFERENTES ARCHITECTURES CMOS.....	59
II.A. TENSION DE SEUIL ET ELECTROSTATIQUE DE L'ARCHITECTURE CONVENTIONNELLE SUR SUBSTRAT MASSIF ..	60

II.A.1.	Critère d'inversion	60
II.A.2.	Tension de seuil canal long	61
II.A.3.	Effets canaux courts (SCE) et DIBL	62
II.A.4.	Pente sous le seuil.....	70
II.A.5.	Comparaison aux mesures silicium :.....	73
II.A.6.	Validation du comportement en température :.....	74
II.B.	TENSION DE SEUIL ET ELECTROSTATIQUE DE L'ARCHITECTURE FDSOI	74
II.B.1.	Critère d'inversion	74
II.B.2.	Tension de seuil canal long	76
II.B.3.	Effets canaux courts (SCE) et DIBL	84
II.B.4.	Pente sous le seuil.....	90
II.B.5.	Validation du comportement en température :.....	92
II.C.	TENSION DE SEUIL ET ELECTROSTATIQUE DE L'ARCHITECTURE DOUBLE GRILLE FAIBLEMENT DOPEE.	93
II.C.1.	Critère d'inversion	93
II.C.2.	Tension de seuil canal long	93
II.C.3.	Effets canaux courts (SCE) et DIBL	96
II.C.4.	Pente sous le seuil.....	98
II.C.5.	Validation du comportement en température :.....	98
II.D.	MODELISATION DU COURANT DE DRAIN	99
II.D.1.	Méthodologie	99
II.D.2.	Transport.....	100
II.D.3.	Raccord faible-forte inversion.....	101
II.D.4.	Raccord régime linéaire-saturé.....	104
II.D.5.	Comparaison au silicium	106
II.E.	CONCLUSION DU CHAPITRE.....	110

CHAPITRE III: EVALUATION ANALYTIQUE DES CAPACITES PARASITES DANS LES STRUCTURES CMOS..... 113

III.A.	METHODOLOGIE	115
III.A.1.	Capacité formée par deux électrodes parallèles.....	116
III.A.2.	Capacité formée par deux électrodes perpendiculaires	116
III.B.	LA TRANSFORMATION CONFORME	116
III.B.1.	Définition du problème.....	116
III.B.2.	La fonction de transformation	118
III.B.3.	Expression de la capacité dans le nouveau repère	120
III.C.	CAPACITES PARASITES SUR UN DISPOSITIF PLANAIRE	122
III.C.1.	Composantes parasites sur structures planaires	122
III.C.2.	Capacité de recouvrement (C_{ov}).....	124
III.C.3.	Capacité de bords externes (C_{of}).....	125
III.C.4.	Capacité de bords internes (C_{if})	126
III.C.5.	Capacité grille-contact (C_{pcca})	127
III.C.6.	Capacité grille-source-drain epitaxiés (C_{gepi}).....	132
III.C.7.	Capacité de coin (C_{corner}).....	134
III.C.8.	Capacité de jonction ou d'oxyde enterré (C_j ou C_{box}).....	136
III.D.	CAPACITES PARASITES SUR UN DISPOSITIF NON-PLANAIRE	137

III.D.1.	Composantes parasites sur structures 3D	137
III.D.2.	Capacité de recouvrement (C_{ov})	139
III.D.3.	Capacité grille-fin (C_{gfin})	139
III.D.4.	Capacité de bords internes (C_{if})	140
III.D.5.	Capacité grille-contact (C_{pcca})	141
III.D.6.	Capacité grille-epitaxie (C_{gepi})	141
III.D.7.	Validation	143
III.D.8.	Capacité corner (C_{corner})	143
III.D.9.	Capacité de jonction ou d'oxyde enterré (C_j ou C_{box})	144
III.E.	RAFFINEMENT DU MODELE	144
III.E.1.	Source-drain épitaxiés avec facettes	144
III.E.2.	Espaceur multi-couche	145
III.F.	APPLICATIONS	146
III.F.1.	Estimation des parasites suivant la feuille de route ITRS	146
III.F.2.	Comparaison FinFET / Double grille planaire	148
III.G.	CONCLUSION DU CHAPITRE	149
CHAPITRE IV: APPLICATION DES MODELES DE CAPACITES PARASITES: ESTIMATION DU DELAI D'UN CIRCUIT SIMPLE		151
IV.A.	METHODE DE RESOLUTION	152
IV.A.1.	Détermination du délai pour $C=1fF$	152
IV.A.2.	Calcul de la capacité équivalente	153
IV.B.	EVALUATION DE PERFORMANCE DISPOSITIFS III-V/Ge: IFQW	155
IV.B.1.	Contexte	155
IV.B.2.	Adaptation du modèle CMOS silicium pour reproduire les simulations Monte-Carlo IFQW	156
IV.B.3.	Construction des références silicium	159
IV.B.4.	Evaluation des capacités parasites	160
IV.B.5.	Estimation du délai, et variation avec I_{off} et V_{dd}	161
IV.B.6.	Estimation de la fuite totale	162
IV.C.	EVALUATION DE PERFORMANCE D'UN INVERSEUR 3D MONOLITHIQUE.	164
IV.C.1.	Contexte de l'étude, réalisation technologique.	164
IV.C.2.	Présentation et définition des inverseurs 2D et 3D	165
IV.C.3.	Performance statique	167
IV.C.4.	Capacités des dispositifs dans les deux types d'inverseur	167
IV.C.5.	Capacités additionnelles dues à l'intégration 3D	168
IV.C.6.	Estimation du délai	171
IV.C.7.	Impact du couplage dynamique entre transistor du haut et transistor du bas dans l'inverseur 3D.	173
IV.D.	CONCLUSION DU CHAPITRE	174
CHAPITRE V: EVALUTATION DES PERFORMANCES AVEC DES OUTILS DE CAO CONVENTIONNELS		177
V.A.	METHODOLOGIE ET IMPLEMENTATION DU MODELE MASTAR VA	178
V.A.1.	Définir un dispositif dans un simulateur CAO	178
V.A.2.	Flot de simulation	182

V.A.3.	Circuits disponibles.....	182
V.A.4.	Prise en compte des parasites dans le back-end.....	184
V.A.5.	Définition de source de variation aléatoire des paramètres technologiques pour prise en compte de la variabilité.....	185
V.A.6.	Validation de la méthode par comparaison avec un modèle SPICE.....	186
V.B.	COMPARAISON DES PERFORMANCES AU NŒUD 20NM	187
V.B.1.	Contexte.....	187
V.B.2.	Définition des dispositifs	187
V.B.3.	Comparaison des performances statiques.....	196
V.B.4.	Robustesse à la variabilité	196
V.B.5.	Evaluation des capacités parasites	198
V.B.6.	Conditions de simulations :	199
V.B.7.	Ring d'inverseur FO1 avec charge additionnelle de sortie variable.....	200
V.B.8.	Ring d'inverseur FO3 avec charge additionnelle de sortie variable.....	204
V.B.9.	Chemin critique DDR3.....	211
V.C.	CONCLUSION DU CHAPITRE.....	211
CHAPITRE VI: EVALUATION DES PERFORMANCES AU NŒUD 14-16 NM		215
VI.A.	CONTEXTE	216
VI.B.	ESTIMATION DES PERFORMANCES LOGIQUES AU NŒUD 14NM	216
VI.B.1.	Définition des dispositifs	216
VI.B.2.	Performance statique.....	217
VI.B.3.	Evaluation des capacités parasites	218
VI.B.4.	Anneaux résonnants d'inverseur FO3 avec charge additionnelle de sortie variable.....	220
VI.C.	PERFORMANCE SRAM	224
VI.C.1.	Définition des dispositifs	224
VI.C.2.	Définition des dessins de cellule SRAM.....	226
VI.C.3.	Définition des sources de variabilité.....	228
VI.C.4.	Simulations de cellules SRAM, incluant la variabilité	229
VI.C.5.	Extraction de la tension minimale d'alimentation (V_{MIN})	230
VI.D.	CONCLUSION DU CHAPITRE.....	233
CONCLUSION GENERALE.....		235
BIBLIOGRAPHIE.....		241
PUBLICATIONS DE L'AUTEUR.....		255
ANNEXE.....		257

INTRODUCTION GENERALE

Le transistor à effet de champ MOSFET (Metal-Oxide-Semiconductor Field Effect Transistor) est le produit manufacturé le plus vendu sur le marché international. Il constitue l'élément central des circuits intégrés conçus à partir des technologies CMOS (Complementary Metal-Oxide-Semiconductor) et on dénombre aujourd'hui plus d'un milliard de transistors sur une même puce. Pour parvenir à cette prouesse technologique, l'industrie de la microélectronique vit au rythme effréné dicté par la loi empirique énoncée par Gordon Moore, co-fondateur d'Intel, en 1965. Celui-ci indiquait que la complexité des semiconducteurs devait doubler tous les ans à coût constant. Après la commercialisation du premier microprocesseur en 1971 par Intel (Intel 4004), qui comptait 2300 transistors sur une surface de 10mm^2 et une puissance de calcul équivalente au 30 tonnes et 167m^2 de l'ENIAC (Electronic Numerical Integrator Analyser and Computer, premier ordinateur entièrement électronique conçu en 1946), Moore réévalua sa prédiction. En 1975, il énonça que le nombre de transistors contenu sur une puce doit doubler tous les deux ans. Depuis maintenant presque quarante ans, l'industrie de la microélectronique s'est fixée comme leitmotiv de suivre cette loi et la densité d'intégration est passée de 230 transistors par mm^2 en 1971 à pratiquement 9 millions de transistors par mm^2 en 2012, soit pratiquement un facteur 40000 en 40 ans. La miniaturisation des transistors permet d'abord de réduire leur coût de production mais permet aussi un gain en performance [Dennard 74]. Associée à une réduction de tension d'alimentation la consommation est également abaissée. Ceci est resté vrai jusqu'au nœud 90nm, soit jusqu'au début des années 2000. En effet, après le passage de la barrière des 100nm de la longueur de grille minimale d'une technologie, les effets parasites liés à la miniaturisation sont devenus non négligeables et sont désormais un frein à la performance. Il s'agit principalement de l'effet canal court, des fuites de grille, de la dégradation du transport, des capacités parasites et des effets quantiques. Des solutions technologiques, telles que, par exemple, l'intégration d'empilement grille métallique/diélectrique haute permittivité et l'application de contraintes mécaniques, sont alors apportées afin de réduire l'impact de ces effets parasites sur le comportement électrique du transistor. Celles-ci vont cependant se révéler insuffisantes à partir des nœuds 28nm et 20nm. Ces derniers voient l'introduction de nouvelles architectures qui vont permettre de poursuivre la miniaturisation du transistor MOSFET, la croissance de sa densité d'intégration et finalement la loi de Moore.

Prenons maintenant un peu de hauteur et considérons le système, c'est-à-dire le circuit intégré, dans sa globalité. Au début des années 2000 est apparue une distinction entre deux types d'application:

- Les applications haute performance, recherchant une vitesse maximale sans (trop) tenir compte de la consommation (car celles-ci fonctionnent en étant branchées au réseau électrique et sont équipées de système d'évacuation de la chaleur). L'exemple représentatif de ce type d'application est le microprocesseur d'un ordinateur.
- Les applications basse consommation, recherchant une bonne autonomie, au détriment de la performance pure car fonctionnant sur batterie. Citons pour exemple le téléphone portable.

Ces considérations ne sont aujourd'hui plus valables. Concernant les applications haute performance, on observe à compter de 2002 une stagnation des fréquences d'horloge. En effet, le niveau de puissance dissipée, tant dynamique que statique, atteint est tel que les boîtiers ne peuvent plus évacuer suffisamment de chaleur et sont alors devenus limitants pour la performance. La consommation doit donc être considérée en tant que telle et devient une contrainte, même pour ce type d'application. Néanmoins, l'introduction des architectures multi-cœur en 2005 a permis de s'affranchir de cette limitation et de proposer des microprocesseurs toujours plus performants. Concentrons-nous maintenant sur les applications basse consommation et regardons le développement du marché du téléphone portable sur la dernière décennie mais également l'évolution de son

usage par l'utilisateur final. Aujourd'hui, on téléphone avec son téléphone mais on navigue aussi sur internet, on écoute de la musique et on regarde des vidéos. La vitesse d'exécution des applications multimédia est donc devenue un élément essentiel mais la consommation reste tout de même une contrainte forte, la question de l'autonomie existant toujours. Les niveaux de performance atteints aujourd'hui ont mené à la même limitation que celle rencontrée par les applications hautes performances quelques années plus tôt: la puissance thermique que le boîtier peut dissiper. Les applications basse consommation convergent donc vers le même modèle que les applications haute performance (l'architecture multi-cœur), et ce, pour la même raison : la recherche de la performance. La nuance entre ces deux types d'application est donc aujourd'hui très légère et les technologies ne sont plus, ou presque plus exclusivement développées pour l'une ou l'autre.

Le compromis vitesse/consommation est donc le paramètre clé d'un circuit intégré. Au niveau du transistor, on le traduit souvent par le compromis niveau de courant débité/niveau de fuite (I_{on}/I_{off}). En lui associant la capacité grille C on peut calculer le temps de transit τ des porteurs dans le transistor par la simple relation $\tau = CV_{dd}/I_{on}$, ce qui est représentatif de la vitesse d'une technologie. Dans l'objectif d'établir une feuille de route sur le long terme, STMicroelectronics a développé en interne un outil de modélisation nommé MASTAR [MASTAR] (Model for Assessment of cmoS Technologies And Roadmaps). Celui-ci propose d'évaluer de manière simple mais physique ces paramètres caractéristiques pour les principales architectures CMOS planaires à partir de leurs paramètres technologiques. Afin d'être plus représentatif de la réalité, MASTAR permet de tenir compte au premier ordre des capacités parasites (dont par exemple la capacité entre la grille et le contact) qui viennent augmenter la capacité totale du dispositif et donc augmenter le temps de transit τ . Les performances des différentes technologies CMOS planaires peuvent donc être évaluées, puis comparées de manière fiable grâce à MASTAR. C'est la raison pour laquelle l'ITRS [ITRS] (International Technology Roadmap for Semiconductor) a choisi MASTAR afin d'établir ses feuilles de route.

MASTAR ne permet cependant pas l'évaluation de performances des technologies CMOS non-planaires (donc 3D) qui émergent à compter du nœud technologie 20nm. De plus, il ne tient pas précisément compte de l'impact des capacités du dispositif tout comme l'effet des interconnexions métalliques entre transistors sur la vitesse et la consommation dynamique du circuit. De plus, MASTAR est intrinsèquement limité à des circuits très, voire trop, simples (des inverseurs) qui ne sont pas suffisamment représentatifs d'un véritable circuit.

On peut alors envisager d'utiliser des modèles SPICE industriels comme [BSIM] qui permettent de simuler et de concevoir les circuits intégrés. Cependant, ces modèles sont calibrés à partir de caractérisations électriques par l'intermédiaire de centaines de paramètres d'ajustement (pas toujours physiquement justifiable) pour reproduire parfaitement le comportement électrique d'une architecture donnée, d'un nœud technologique donné et d'une application donnée. On comprend alors aisément que ce type de modèle est peu flexible et peu prédictif, donc difficilement utilisable pour évaluer et comparer les performances de différentes architectures de transistor MOSFET sur le long terme.

Aujourd'hui, plusieurs architectures de transistors, planaires ou non, sont en concurrence. L'évaluation puis la comparaison de leurs performances au niveau du circuit est indispensable mais est seulement partielle à partir des métriques conventionnelles du transistor MOSFET et même à partir de MASTAR.

Les travaux de cette thèse s'inscrivent dans ce contexte et proposent la description puis l'utilisation d'un modèle compact, MASTAR_VA, prédictif, universel (c'est-à-dire utilisable pour chaque architecture de transistor MOSFET), flexible (c'est-à-dire adaptable à des dispositifs plus « exotiques ») et compatible avec les outils de CAO conventionnels, comme [ELDO].

Le premier chapitre est consacré à la description du fonctionnement du transistor MOSFET conventionnel, sa miniaturisation et les effets parasites qu'elle entraîne. Les solutions technologiques apportées à l'architecture conventionnelle sur substrat massif du transistor MOSFET sont ensuite explicitées. Enfin, cette architecture ne pouvant plus apporter le niveau de performance requis pour les nœuds technologiques les plus avancés, nous décrirons les architectures envisagées pour poursuivre la miniaturisation du transistor MOSFET.

Le second chapitre sera consacré à l'élaboration d'un modèle compact de tension de seuil et de courant drain-source utilisable pour chaque architecture CMOS, qui assure la continuité du courant et de ses dérivées entre

chaque régime. Nous commencerons par décrire les régimes sous le seuil des architectures conventionnelles sur substrat massif, FDSOI et double grille faiblement dopé par une modélisation prédictive et précise des paramètres électrostatiques: tension de seuil, effet canaux courts (SCE et DIBL) et pente sous le seuil. Celle-ci tiendra compte des variations de chaque paramètre technologique, de chaque type de dispositif. Le régime au-delà du seuil sera quant à lui évalué par la théorie classique du courant de dérive à partir des lois de mobilités universelles, corrigées pour tenir compte par exemple des contraintes mécaniques apportées par le procédé de fabrication. Enfin, la continuité entre chaque régime de fonctionnement sera garantie par l'utilisation de tension effective. Ces modèles seront ensuite validés par simulation numérique 2D et par caractérisations électriques lorsque celles-ci sont disponibles.

Le troisième chapitre sera dédié à l'évaluation des capacités parasites pour les principales architectures CMOS : planaires (architecture conventionnelle sur substrat massif, FDSOI et double grille planaire) et non-planaires (FinFET et Trigate). Après le développement d'une méthode générique d'estimation des capacités parasites, nous fournirons les expressions analytiques de chaque composante capacitive en tenant compte des toutes dernières briques technologiques.

Le quatrième chapitre proposera une méthode itérative et numérique d'estimation du délai de propagation d'une chaîne d'inverseur à partir des modèles analytiques développés dans les deuxième et troisième chapitres. Cette méthode sera ensuite utilisée pour évaluer la performance circuit de la co-intégration III-V/Ge de dispositif IFQW (Implant Free Quantum Well) puis de la comparer à une référence silicium (architecture conventionnelle sur substrat massif et FDSOI) au nœud 20nm. Cette même méthode sera à nouveau utilisée pour évaluer la performance circuit de l'intégration 3D monolithique et de la comparer à une référence FDSOI dans le cadre du nœud 28nm.

Le cinquième chapitre proposera dans un premier temps d'adapter les modèles des chapitres II et III aux contraintes d'un outil de CAO conventionnel. Ceci définira un modèle compact, que nous nommerons MASTAR_VA. A partir de l'état de l'art au nœud technologique 20nm, nous proposerons ensuite d'utiliser MASTAR_VA afin d'évaluer les performances logiques des trois architectures concurrentes pour ce nœud : l'architecture conventionnelle sur substrat massif, le FDSOI et le Trigate.

Enfin, le sixième et dernier chapitre de ce manuscrit sera dédié à la prédiction de performance logique et SRAM, en tenant compte de la variabilité due au procédé de fabrication, au nœud technologique 16nm des trois mêmes architectures. Celles-ci seront définies par extrapolation, selon la loi de Moore, à partir des dispositifs définis au chapitre V.

Ce travail de thèse a été réalisé au sein de l'équipe R&D avancée de STMicroelectronics à Crolles, en collaboration avec l'Institut de Microélectronique, Electromagnétisme et Photoniques (IMEP-LAHC) de Grenoble.

- CHAPITRE I -

LE TRANSISTOR MOSFET: FONCTIONNEMENT, MINIATURISATION ET ARCHITECTURES.

Ce chapitre a pour objectif de définir les différents concepts théoriques et les paramètres relatifs au transistor MOSFET (Métal-Oxyde-Semiconducteur). Nous commencerons par décrire l'architecture conventionnelle sur substrat massif classique (BULK) idéal, c'est-à-dire sans aucun effet parasite. Ensuite, nous expliciterons les effets de la course à la miniaturisation sur le fonctionnement du transistor MOSFET idéal. Nous détaillerons par la suite les solutions technologiques apportées à l'architecture conventionnelle sur substrat massif classique pour lutter contre ces effets, dits « parasites ». Enfin, la réduction des dimensions de nœuds en nœuds de l'architecture conventionnelle sur substrat massif ne procurant plus suffisamment de gain en performance, nous décrirons les nouvelles architectures de transistor MOSFET envisagées.

I.A. Le Transistor MOSFET idéal

I.A.1. Principe de fonctionnement

Le transistor à effet de champ Métal-Oxyde-Semiconducteur (MOSFET pour Metal-Oxide-Semiconductor Field Effect Transistor en anglais) est la brique de base des circuits intégrés produits en technologie CMOS (Complementary Metal Oxide Semiconductor). Ce transistor, très schématiquement représenté sur la Figure I-1-a, est composé de quatre électrodes : la grille (G), le drain (D), la source (S) et le substrat (B). Il a simplement pour fonction de, selon la polarisation de commande (appliquée sur la grille) laisser passer ou bloquer un courant entre la source et le drain. On résume alors souvent son fonctionnement à celui d'un interrupteur commandé en tension. En effet, le champ électrique vertical (dit champ de grille) dû à la polarisation de l'électrode de commande vient moduler la densité de porteurs dans le semiconducteur, ce qui mène aux deux états de fonctionnement fondamentaux du transistor MOSFET idéal :

- L'état bloqué : aucun courant ne circule entre source et drain (Figure I-1-b).
- L'état passant : un canal de conduction est formé par l'action du champ électrique vertical et, sous l'action du champ longitudinal (dû à la polarisation du drain et de la source) les porteurs du canal de conduction sont mis en mouvement. Un courant de porteurs circule alors entre source et drain (I_{ds}), il s'agit d'électrons pour un NMOS et de trous pour un PMOS (Figure I-1-c).

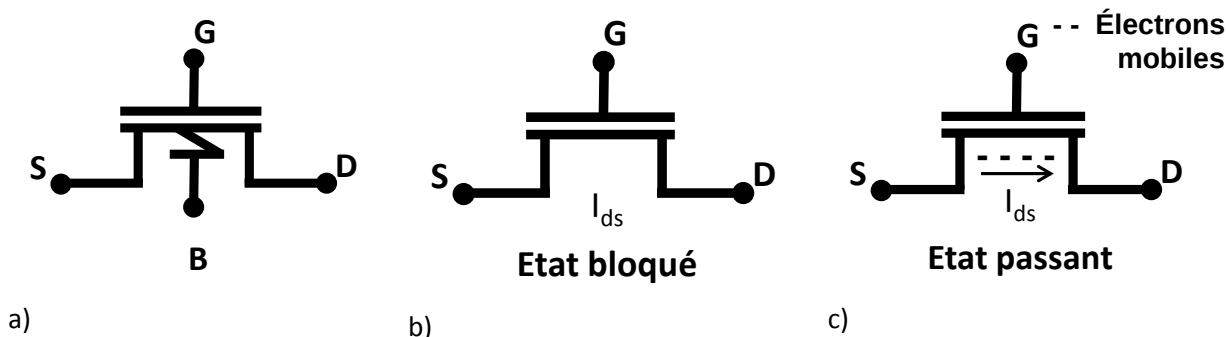


Figure I-1: a) Schéma très simplifié d'un transistor MOSFET. b) Illustration de l'état passant (dans le cas d'un NMOS) c) illustration de l'état bloqué.

I.A.2. L'architecture conventionnelle sur substrat massif (BULK)

Technologiquement, le transistor MOSFET (Figure I-2) est composé d'un empilement métal-oxyde-semiconducteur qui constitue une capacité MOS dans laquelle :

- Le métal est la grille, donc l'électrode de commande du transistor. Historiquement, elle était constituée de polysilicium dopé à dégénérescence, mais dans les technologies CMOS avancées (depuis les nœuds 45-

32nm), elle est en métal. Sa longueur, paramètre clé d'une technologie CMOS est notée L_g et sa polarisation est notée V_g .

- L'oxyde de grille constitue l'isolant de la capacité MOS. Historiquement, ce dernier est en oxyde de silicium (SiO_2). Son épaisseur, déterminante pour le niveau de performance du transistor, est notée t_{ox} . On notera également sa permittivité ϵ_{SiO_2} .
- Le semiconducteur (du silicium) constitue la seconde électrode de la capacité MOS et forme la zone active du transistor, dans laquelle le canal de conduction se forme. Dans le cas d'un NMOS, la zone active est dopée P alors que pour un PMOS elle est dopée N. dans les deux cas, le niveau de dopage N_{ch} est usuellement compris entre 10^{18} - 10^{19} cm^{-3} . La largeur de la zone active est notée W et la polarisation appliquée sur sa face arrière est notée V_b .

Pour qu'un courant circule, des réservoirs de porteurs sont nécessaires: c'est le rôle de la source et du drain qui sont deux zones fortement dopées (usuellement $5 \cdot 10^{19}$ - 10^{20} cm^{-3}) de type opposé à la zone active : dopage de type N pour un NMOS, P pour un PMOS. Leurs polarisations sont respectivement notées V_s et V_d . Afin de connecter le transistor au reste du circuit, des contacts métalliques sont indispensables et sont placés sur les quatre électrodes du dispositif :

- Le contact de grille est généralement placé sur le prolongement de la grille sur le STI (Shallow Trench Isolation : tranchée remplie d'oxyde utilisée pour isoler électriquement le transistor de ses voisins) qui est élargi pour pouvoir recevoir le contact (Figure I-2-b).
- Les contacts de source et de drain sont placés sur les zones implantées.
- Le contact de substrat, on parle souvent de prise caisson, est déporté et n'est pas représenté sur la Figure I-2.

Dans la course à la miniaturisation des technologies CMOS, un des enjeux est de réduire l'encombrement du transistor, donc sa surface. Cette dernière est égale au produit de la largeur du transistor, par son encombrement dans l'autre direction. Celle-ci est usuellement représentée par le CPP (« Contacted Poly Pitch »: pas de répétition d'une grille et d'un contact), représenté sur la Figure I-2. Un second enjeu est de fournir une technologie toujours plus performante (i.e. une vitesse des circuits accrue) d'un nœud technologique à l'autre tout en réduisant sa consommation. Il est usuel de réduire la tension d'alimentation d'un nœud à l'autre, et celle-ci est notée V_{dd} .

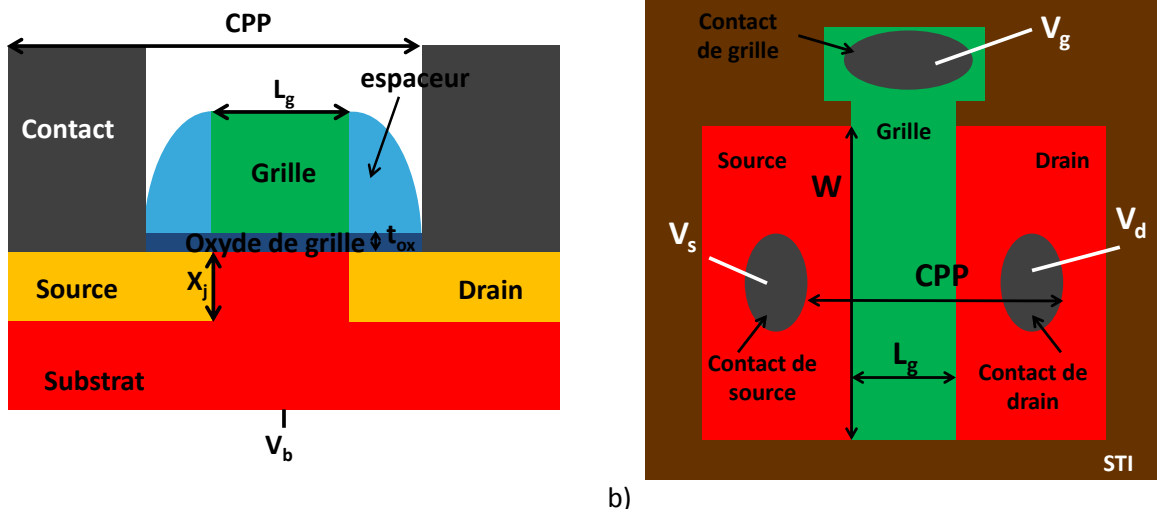


Figure I-2 : Schéma de l'architecture conventionnelle sur substrat massif. a) vue en coupe b) vue de dessus

Cela décrit l'architecture conventionnelle sur substrat massif, utilisée pendant les premières décennies de l'industrie de la microélectronique, et jusqu'au nœud technologique 20nm [Shang 12].

I.A.3. La capacité MOS

La capacité MOS étant l'élément central d'un transistor MOSFET, il faut donc commencer par décrire son fonctionnement avant d'entrer dans la physique du transistor MOSFET complet. Celui-ci se décompose en trois régimes distincts, dépendant de la valeur du potentiel φ_s en surface du semiconducteur, à l'interface avec l'isolant (i.e. l'oxyde de grille). Dans le cas d'un NMOS, donc pour du silicium dopé P (les porteurs majoritaires sont les trous, les minoritaires les électrons) :

- Régime d'accumulation: $\varphi_s < 0$. Le champ vertical dû à la grille est orienté vers le haut, les électrons sont repoussés de l'interface oxyde-silicium, alors que les trous sont attirés vers l'oxyde et viennent s'accumuler à cette interface. La variation de charge d'accumulation Q_{acc} est exponentielle et augmente avec $|\varphi_s|$.
- Régime de déplétion : $0 < \varphi_s < \varphi_f$ (φ_f est le potentiel de Fermi). Le champ de grille a changé de sens et est orienté vers le bas : les trous sont alors repoussés de l'interface oxyde-silicium et une zone sans charge mobile s'étend à partir de l'oxyde de grille. Cette zone, dite de déplétion (sa profondeur est notée T_{dep}) est chargée négativement, car elle est uniquement composée de dopants ionisés (des accepteurs ici car le silicium est de type P). La variation de charge de déplétion Q_{dep} est proportionnelle à $\sqrt{\varphi_s}$.
- Régime d'inversion $\varphi_s > \varphi_f$. La zone de déplétion ne s'étend presque plus et le champ vertical est suffisamment fort pour attirer des électrons (des porteurs minoritaires) à l'interface oxyde-silicium. La variation de charge du silicium de type P est alors dû aux porteurs minoritaires, c'est la raison pour laquelle on parle de régime d'inversion. Ce régime est divisé en deux parties :
 - $\varphi_s < 2\varphi_f$: régime d'inversion faible. La croissance de charge d'inversion Q_{inv} est d'abord faible, et reste proportionnelle à $\sqrt{\varphi_s}$.
 - $\varphi_s > 2\varphi_f$: régime d'inversion forte. La croissance de charge d'inversion Q_{inv} est à nouveau exponentielle et augmente avec φ_s .

Ces trois régimes de fonctionnement sont illustrés schématiquement sur la Figure I-3.

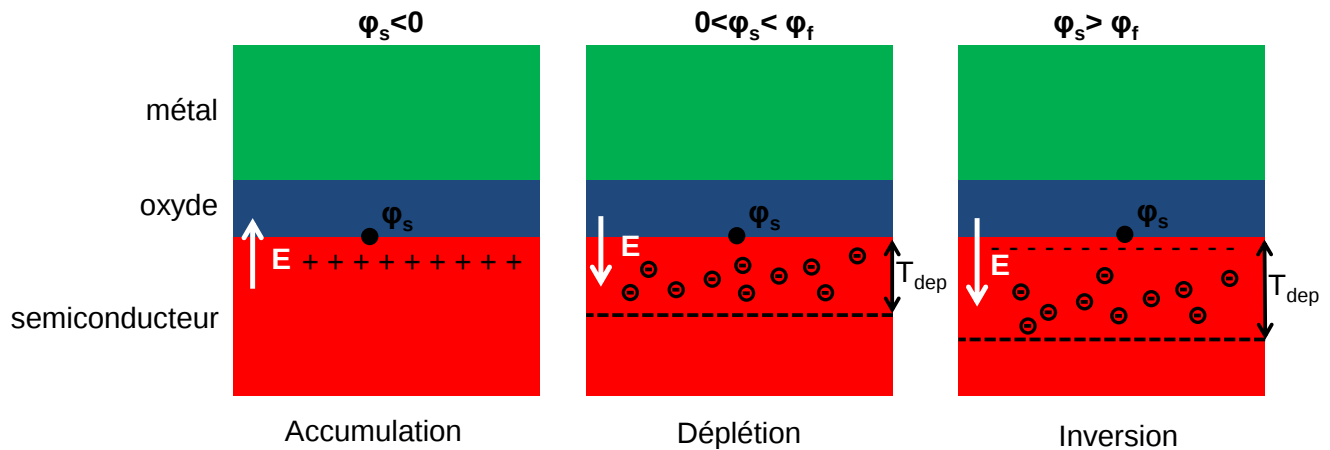


Figure I-3 : Illustration schématique des trois régimes de fonctionnement de la capacité MOS pour un semiconducteur de type P.

Pour un silicium de type N, les régimes de fonctionnement sont les mêmes, mais comme les charges changent de signe (i.e les dopants ionisés sont des donneurs et les porteurs minoritaires sont des trous donc des charges positives), les conditions sur le potentiel de surface sont modifiées. Par un raisonnement analogue, on a :

- Régime d'accumulation : $\varphi_s > 0$
- Régime de déplétion : $-\varphi_f < \varphi_s < 0$
- Régime d'inversion : $\varphi_s < -2.\varphi_f$

La Figure I-4 représente la variation de la valeur absolue de la charge dans le semiconducteur en échelle logarithmique en fonction de la valeur du potentiel de surface et met en évidence les trois régimes de fonctionnement pour du silicium de type P. Pour tracer cette courbe, les expressions analytiques de la charge dans le semiconducteur pour chaque régime de fonctionnement sont nécessaires et sont obtenues par la résolution de l'équation de Poisson couplée aux distributions de porteurs. Les détails de calcul sont développés dans [Mathieu 04].

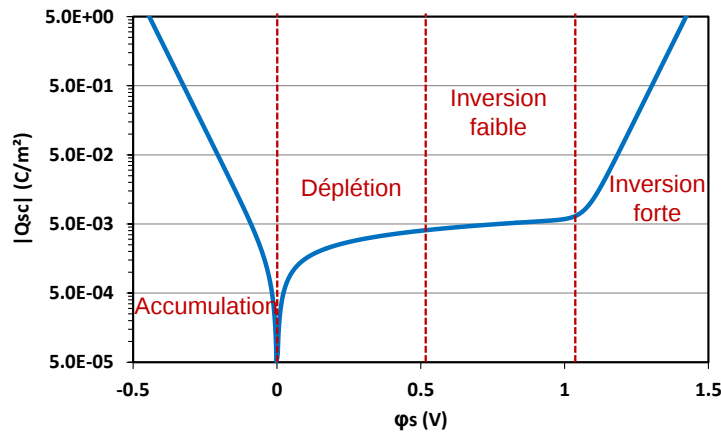


Figure I-4 : Variation de la valeur absolue de la charge dans le semiconducteur (de type P) Q_{sc} en fonction de la valeur du potentiel de surface φ_s , mettant en évidence les trois régimes de fonctionnement.

Pour tracer la variation de charge en fonction de la tension effectivement appliquée sur la grille, il faut tenir compte du travail de sortie du métal, mais également de la chute de potentiel dans l'oxyde de grille. Pour tenir compte du travail de sortie du métal, il faut évaluer la tension de bandes plates, qui correspond à la tension de grille à appliquer pour amener la structure MOS en régime de bandes plates. Le régime dit de « bandes plates » est atteint lorsque les bandes d'énergie de la structure MOS sont plates. A partir du diagramme de bandes d'énergie à l'équilibre thermodynamique, représenté sur la Figure I-5 dans le cas d'un NMOS, on obtient l'expression de la tension de bandes plates V_{FB} :

$$V_{FB} = \varphi_m - (\chi_s + \frac{E_g}{2} + \varphi_f) \quad \text{Eq. I-1}$$

De manière duale, on obtient la tension de bandes plates pour un PMOS :

$$V_{FB} = \varphi_m - (\chi_s + \frac{E_g}{2} - \varphi_f) \quad \text{Eq. I-2}$$

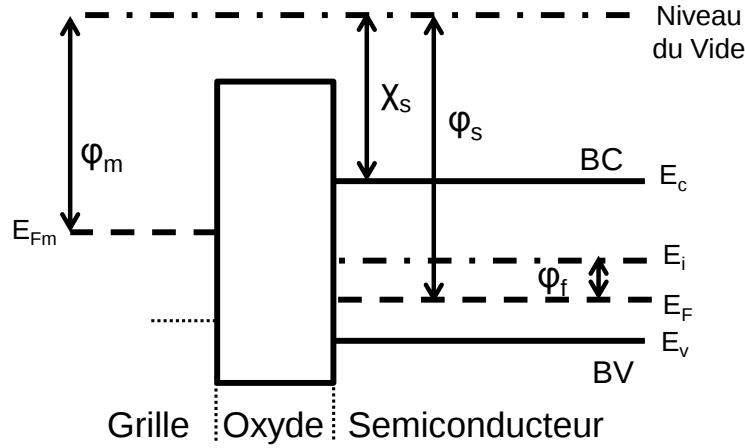


Figure I-5 : Diagramme de bandes d'énergie à l'équilibre thermodynamique obtenu pour un NMOS (silicium de type P).

Avec :

- ϕ_m travail de sortie du métal de la grille égal à la différence entre le niveau du vide et du potentiel de Fermi du métal.
- X_s affinité électronique du semiconducteur (pour le silicium: $X_s=4.05V$)
- E_g gap du semiconducteur qui est égal à la différence entre le niveau de la bande de conduction (BC) et celui de la bande de valence (BV), donc à E_c-E_v . Sa valeur dépend de la température T (en Kelvin) et une bonne approximation de sa valeur est donnée par l'expression [Sze 81] :

$$E_g = 1.17 - 4.73 \cdot 10^{-4} \frac{T^2}{T+636} \quad \text{Eq. I-3}$$

Pour le silicium, à une température de 300K on a $E_g=1.12V$.

- ϕ_f potentiel de Fermi du semiconducteur, qui est égal à la différence entre le niveau de Fermi E_f et le niveau de Fermi intrinsèque E_i (E_i correspond au milieu du gap, on a donc $(E_i-E_v = E_c-E_i = E_g/2)$ donné par l'équation :

$$\phi_f = \frac{kT}{q} \ln\left(\frac{N_{ch}}{n_i}\right) \quad \text{Eq. I-4}$$

Où N_{ch} est le dopage du semiconducteur, k la constante de Boltzman ($k=1.38062 \cdot 10^{-23} J \cdot K^{-1}$), q la valeur absolue de la charge de l'électron ($q= 1.602 \cdot 10^{-19} C$) et n_i la concentration de porteurs intrinsèques qui dépend aussi de la température et est donné, pour le silicium par l'équation :

$$n_i = \sqrt{N_c N_v} e^{-\frac{qE_g}{kT}} \quad \text{Eq. I-5}$$

Avec N_c et N_v les densités équivalentes dans les bandes de conduction et de valence, données par les équations :

$$N_c = 2 \left(\frac{2\pi m_c kT}{h^2} \right)^{3/2} \quad \text{Eq. I-6}$$

$$N_v = 2 \left(\frac{2\pi m_v kT}{h^2} \right)^{3/2} \quad \text{Eq. I-7}$$

Avec :

- $h=6.62.10^{-34} \text{ J.s}^{-1}$ est la constante de Planck.
- $m_e=9.66.10^{-31} \text{ kg}$ est la masse effective de l'électron.
- $m_h=5.37.10^{-31} \text{ kg}$ est la masse effective d'un trou.

La prise en compte de la dépendance de tous ces paramètres avec la température est fastidieuse, mais nécessaire pour prévoir la variation de tension de seuil.

I.A.4. La tension de seuil

La tension de seuil d'un transistor MOSFET est la tension de grille pour laquelle le transistor passe de l'état bloqué à l'état passant. Il s'agit donc de la tension pour laquelle le canal de conduction entre source et drain se forme. Dans un transistor MOSFET, le courant est dû aux porteurs minoritaires, par conséquent, l'état passant du transistor MOSFET correspond au régime d'inversion de la capacité MOS. Cette remarque ne suffit pas à connaître l'emplacement exact du seuil. Pour que le transistor soit bel et bien à l'état passant, la quantité de charge d'inversion doit être suffisante. De manière quantitative, cela signifie que la quantité de porteurs minoritaires, de densité n_s , doit être supérieure à la quantité de dopants ionisés, (des accepteurs pour un NMOS), de densité égale au niveau de dopage N_{ch} . Le seuil est donc donné par l'égalité entre ces deux quantités. Formellement, en utilisant la statistique de Fermi pour les distributions de porteurs [Mathieu 04], on obtient :

$$n_s = n_i e^{\frac{q(\phi_s - \phi_f)}{kT}} = N_{ch} = n_i e^{\frac{q\phi_f}{kT}} \quad \text{Eq. I-8}$$

En résolvant cette équation, on obtient la valeur du potentiel de surface au seuil :

$$\phi_s = 2 \times \phi_f \quad \text{Eq. I-9}$$

Le passage de l'état bloqué à l'état passant du transistor MOSFET correspond donc au passage en régime d'inversion forte de la capacité MOS. Afin d'obtenir l'expression de la tension de seuil, nous devons lier la polarisation de grille V_g , au potentiel de surface ϕ_s dont nous connaissons la valeur au seuil grâce à l'équation I-9. Ecrivons alors la loi de Gauss dans le semiconducteur de la capacité MOS (représentée sur Figure I-3):

$$C_{ox}(V_g - V_{FB} - \phi_s) = Q_{acc} + Q_{dep} + Q_{inv} \quad \text{Eq. I-10}$$

Avec C_{ox} capacité de l'oxyde grille, donné par la relation $C_{ox} = \epsilon_{SiO_2} / t_{ox}$. Or, au seuil, on se trouve à la limite entre les régimes d'inversion faible et d'inversion forte, la charge d'accumulation Q_{acc} est donc négligeable. De plus, la variation de charge d'inversion Q_{inv} est encore faible devant celle de déplétion Q_{dep} , et peut donc être négligée. L'équation I-10 peut donc être réduite comme:

$$C_{ox}(V_g - V_{FB} - \phi_s) = Q_{dep} \quad \text{Eq. I-11}$$

La tension de seuil V_{th} est la polarisation de grille V_g pour laquelle on a l'égalité $\phi_s = 2\phi_f$. En isolant V_g dans l'équation I-11 puis en se plaçant au seuil, on a :

$$V_{th} = V_{FB} + 2\phi_f + \frac{Q_{dep}}{C_{ox}} \quad \text{Eq. I-12}$$

La dernière inconnue est la valeur de la charge de déplétion au seuil. Celle-ci est évaluée avec son expression classique, donnée dans la littérature [Sze 81] :

$$Q_{\text{dep}} = qN_{\text{ch}}T_{\text{dep}} \quad \text{Eq. I-13}$$

Avec
$$T_{\text{dep}} = \sqrt{\frac{2\varepsilon_{\text{si}}2\varphi_f}{qN_{\text{ch}}}} \quad \text{Eq. I-14}$$

Finalement, la tension de seuil est donnée par l'expression:

$$V_{\text{th}} = V_{\text{FB}} + 2\varphi_f + \frac{\sqrt{2qN_{\text{ch}}\varepsilon_{\text{si}}2\varphi_f}}{C_{\text{ox}}} \quad \text{Eq. I-15}$$

La tension de seuil de la capacité MOS est donc celle du transistor MOSFET si on néglige l'existence de la source et du drain. Cela correspond au cas où la longueur de grille est suffisamment importante pour rendre l'impact de la source et du drain sur le canal de conduction, et donc sur la tension de seuil, négligeable. L'équation I-15 donne donc la tension de seuil V_{th} pour un transistor NMOS à canal long. De manière plus adaptée au transistor MOSFET, la tension de seuil est définie par la tension à appliquer sur la grille pour permettre aux porteurs minoritaires (provenant des source-drains) de franchir la barrière de potentiel existant entre la source et le drain.

A ce stade, en considérant que le transistor MOSFET se comporte comme un interrupteur commandé en tension parfait, avec un passage de l'état bloqué à l'état passant ayant lieu pour $V_g = V_{\text{th}}$, on obtient la caractéristique courant-tension idéale de la Figure I-6. Dans le cas idéal, le courant drain-source I_{ds} est nul lorsque $V_g < V_{\text{th}}$ puis passe brutalement à 1 (au sens logique) à $V_g = V_{\text{th}}$ et reste constant à 1 pour $V_g > V_{\text{th}}$.

Pour un transistor PMOS, la tension de seuil est négative du fait du changement de signe des charges dans le silicium et l'équation I-15 donne la valeur absolue de la tension de seuil du transistor.

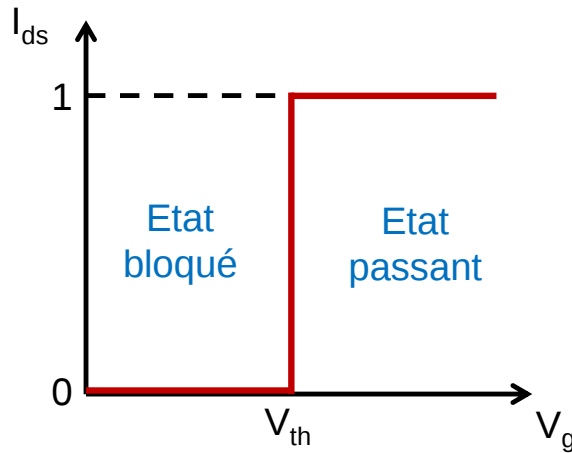


Figure I-6 : Caractéristique courant-tension d'un transistor MOSFET idéal.

I.B. Le transistor MOSFET réel

Le fonctionnement du transistor n'est cependant pas idéal :

- A l'état bloqué ($V_g < V_{\text{th}}$): le courant drain-source n'est pas nul. Son évaluation classique est développée au cours du paragraphe I.B.1.
- A l'état passant ($V_g > V_{\text{th}}$): le courant drain-source ne sature pas parfaitement. La description du comportement du courant I_{ds} dans cet état est donnée par le paragraphe I.B.2.

La non-idéalité du transistor MOSFET mène à la notion de délai de commutation (non nul, car le basculement entre l'état passant et l'état bloqué n'est pas instantané). Ceci fait l'objet du paragraphe I.B.4.

I.B.1. Régime sous le seuil : I_{off} et Pente sous le seuil

Sous le seuil, donc pour $V_g < V_{th}$, le courant drain-source n'est pas nul. Du fait du gradient de porteurs minoritaires entre source et drain, un courant de diffusion se met en place. En considérant que les porteurs minoritaires ne diffusent que selon une dimension (l'axe x), la densité de courant surfacique entre source et drain associée pour un NMOS est donnée par l'équation classique :

$$j_{\text{diff}} = -qD_n \frac{dn(x)}{dx} \quad \text{Eq. I-16}$$

Où $n(x)$ est la densité d'électron et D_n est la constante de diffusion des électrons donnée par la relation d'Einstein :

$$D_n = \frac{kT}{q} \mu_n \quad \text{Eq. I-17}$$

Où μ_n est la mobilité des électrons. En intégrant l'équation I-16 sur la surface du flux de courant, notée A , et en considérant que la variation de la densité d'électron est linéaire entre source et drain, on peut écrire l'expression du courant drain-source I_{ds} sous le seuil :

$$I_{ds} = qAD_n \frac{n(0) - n(L)}{L} \quad \text{Eq. I-18}$$

Avec $n(0)$ et $n(L)$ étant les densités d'électrons respectivement à la source et au drain données par [Sze 81] :

$$n(0) = n_0 e^{\frac{q\phi_s}{kT}} \quad \text{Eq. I-19}$$

$$n(L) = n_0 e^{\frac{q(\phi_s - V_d)}{kT}} \quad \text{Eq. I-20}$$

Pour obtenir l'expression du courant drain-source sous le seuil comme une fonction de V_g et V_d , on commence par substituer les équations I-19 et I-20 dans I-18. Ensuite, il faut évaluer la surface du flux de courant de diffusion A : sa largeur est égale à la largeur du transistor, soit W et son épaisseur est donnée par [Sze 81] et vaut $kT/(qE_s)$ (E_s est le champ en surface du semiconducteur, donc à l'interface oxyde-semi-conducteur, et vaut $\epsilon_{si}Q_{\text{dep}}$). Enfin, potentiel de surface ϕ_s et polarisation de grille V_g sont liés à partir de l'équation de Gauss (Eq. I-11). Après plusieurs étapes de calculs, détaillées notamment dans [Sze 81] et [Taur 98], on obtient l'expression analytique du courant drain-source sous le seuil :

$$I_{ds} = \mu_n C_{ox} \left(\frac{W}{L}\right) (m - 1) \left(\frac{kT}{q}\right)^2 e^{\frac{q(V_g - V_{th})}{mkT}} (1 - e^{-\frac{qV_d}{kT}}) \quad \text{Eq. I-21}$$

Avec $m = \Delta V_g / \Delta \phi_s$

Cette expression peut être également obtenue en intégrant l'expression analytique de la charge d'inversion en régime de faible inversion. Cette méthode est développée dans [Taur 98]. Sachant que V_{ds} est généralement supérieur à quelques kT/q , on peut considérer que l'expression donnée par l'Eq. I-21 est indépendante de la tension de drain V_d . En régime sous le seuil, le courant suit donc une variation exponentielle avec la polarisation de grille. Autrement dit, si le courant est représenté en échelle logarithmique, sa variation est linéaire (i.e. sous le seuil, $\log(I_{ds})$ est proportionnelle à V_g). Nous notons alors cette pente S^{-1} , ce qui définit l'inverse de la pente sous le seuil S :

$$S = \frac{dV_g}{d\log(I_{ds})} \quad \text{Eq. I-22}$$

En remplaçant dans l'équation I-22 I_{ds} par son expression donnée par l'équation I-21, on obtient son expression analytique :

$$S = \frac{kT}{q} \ln(10) \left(1 + \frac{C_{dep}}{C_{ox}} \right) \quad \text{Eq. I-23}$$

Où C_{dep} est la capacité de déplétion de la capacité MOS.

Une autre (et plus simple) manière de calculer le paramètre S est donnée dans [Pouydebasque 07] et consiste à réécrire la définition de la pente sous le seuil comme suit :

$$S = \frac{dV_g}{d \log(I_{ds})} = \frac{kT}{q} \ln(10) \frac{dV_g}{d\phi_s} \quad \text{Eq. I-24}$$

La dérivée de la tension de grille V_g par rapport au potentiel de surface ϕ_s est alors effectuée par un simple calcul de diviseur capacitif. Ce dernier est représenté sur la Figure I-7 et l'expression de l'inverse de la pente sous le seuil ainsi obtenue est l'équation I-23.

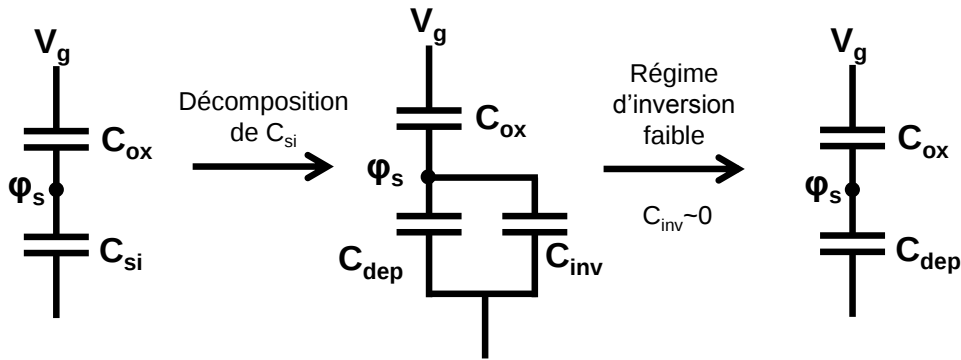


Figure I-7 : Diviseur capacitif utilisé pour le calcul de la pente sous le seuil.

La pente sous le seuil donne donc l'augmentation de polarisation de grille nécessaire pour que le courant drain-source en régime sous le seuil gagne une décade et s'exprime en mV/dec. Sa valeur idéale correspond au cas où la capacité de déplétion est négligeable (i.e. dopage faible) et vaut à $T=300K$ $(kT/q)\ln(10)=60\text{mV/dec}$. De manière très pragmatique, le courant drain-source en régime sous le seuil, noté I_{diff} peut être évalué simplement par l'expression :

$$I_{diff} = I_{th} e^{\frac{V_g - V_{th}}{S \ln(10)}} \quad \text{Eq. I-25}$$

Avec I_{th} valeur du courant au seuil (à $V_g=V_{th}$), qui vaut usuellement [Fleury 09] :

$$I_{th} = \frac{W}{L} 10^{-7} \quad \text{Eq. I-26}$$

A partir des équations I-25 et I-26, on obtient facilement l'expression du courant à polarisation de grille nulle, ce qui définit le courant de fuite I_{off} :

$$I_{off} = \frac{W}{L} 10^{-7} e^{\frac{-V_{th}}{S \ln(10)}} \quad \text{Eq. I-27}$$

Le courant de fuite I_{off} est le courant drain-source mesuré pour $V_g=0V$ et $V_d=V_{dd}$. La pente sous le seuil S est donc un paramètre clé pour minimiser le courant de fuite I_{off} . D'après l'équation I-27, on voit bien que, si deux transistors ont la même tension de seuil V_{th} mais des pentes sous le seuil S différentes, celui qui a la pente la plus forte aura le courant de fuite I_{off} le plus élevé (Figure I-8-a). De manière duale, si la fuite I_{off} est fixée, le transistor

avec la pente sous le seuil la plus élevée aura la tension de seuil la plus élevée (Figure I-8-b), ce qui peut être problématique pour les transistors MOS utilisés à très basse tension.

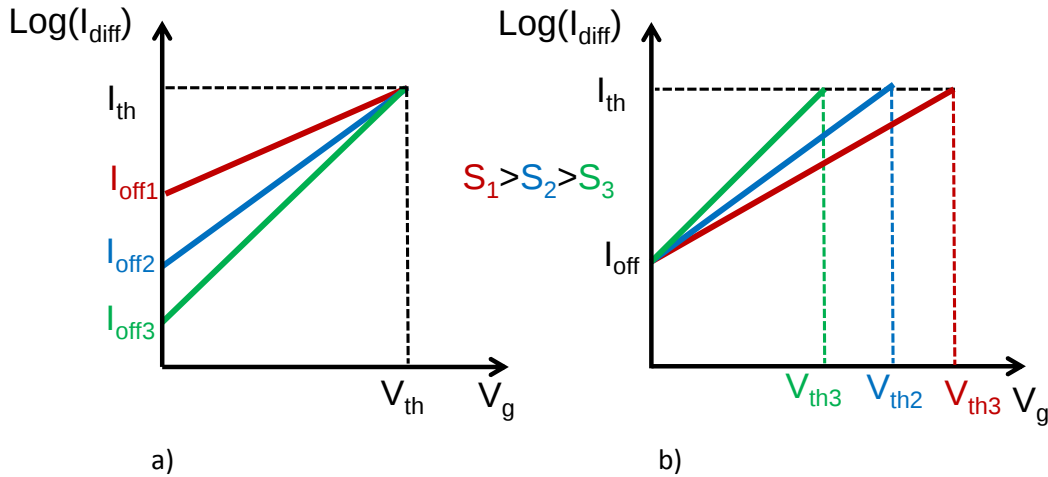


Figure I-8 : Illustration de l'impact de la pente sous le seuil sur le comportement électrique du transistor MOSFET. a) effet de la pente sous le seuil à tension de seuil V_{th} constante b) effet de la pente sous le seuil à courant de fuite I_{off} constant.

I.B.2. Régime au-delà du seuil : I_{on}

Au-delà du seuil, pour $V_g > V_{th}$, la capacité MOS passe en inversion forte. La densité de porteurs minoritaires et donc la charge d'inversion augmentent de manière exponentielle avec la polarisation de grille. Le courant drain-source est désormais dû aux porteurs minoritaires mis en mouvement par le champ électrique longitudinal E entre drain et source, et on parle alors de courant de dérive (drift en anglais). Sa densité surfacique est classiquement donnée pour un NMOS par :

$$j_{\text{drift}} = -qn\mu_n E \quad \text{Eq. I-28}$$

Où n est la densité de porteurs minoritaires, μ_n la mobilité de électrons et E le champ électrique.

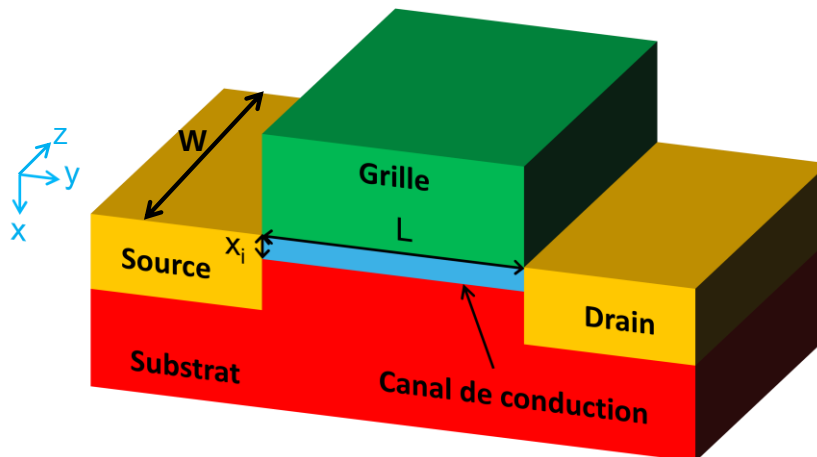


Figure I-9 : structure, dimensions et axes considérés pour le calcul du courant de dérive.

Pour obtenir l'expression du courant de dérive en un point du canal y , intégrons la densité surfacique de courant de dérive donnée par l'équation I-28 sur la surface du canal de conduction A . On considérera que le champ

électrique E est unidimensionnel, et qu'il est donné par la relation $E = -dV(y)/dy$ où $V(y)$ est le potentiel en un point y du canal. On considérera également que la densité de porteur ne dépend que de x et y , et sera noté $n(x,y)$. Cette surface A est donnée par le produit de la largeur du canal de conduction, donc à la largeur du transistor W (selon l'axe z) par l'épaisseur du canal de conduction que nous noterons x_i (selon l'axe x). Les dimensions et axes sont indiqués sur la Figure I-9. L'intégrale double à évaluer est alors la suivante:

$$I_{\text{drift}}(y) = \int_0^W \int_0^{x_i} q\mu_n n(x,y) \frac{dV(y)}{dy} dx dz \quad \text{Eq. I-29}$$

Tous les paramètres étant indépendant de z , l'intégration suivant cet axe est très simple. On considère ensuite que la mobilité des électrons μ_n est indépendante de x , et qu'elle peut être remplacée par la mobilité effective μ_{eff} [Taur 98]. La mobilité effective μ_{eff} fait l'objet du paragraphe suivant. On a alors:

$$I_{\text{drift}}(y) = -qW\mu_{\text{eff}} \frac{dV(y)}{dy} \int_0^{x_i} n(x,y) dx \quad \text{Eq. I-30}$$

Or, par définition, la charge d'inversion par unité de surface de grille en y est liée à la densité de porteurs minoritaires par:

$$Q_{\text{inv}}(y) = -q \int_0^{x_i} n(x,y) dx \quad \text{Eq. I-31}$$

Le courant de dérive en y s'exprime alors :

$$I_{\text{drift}}(y) = -qW\mu_{\text{eff}} \frac{dV(y)}{dy} Q_{\text{inv}}(y) \quad \text{Eq. I-32}$$

La variable y est interchangeable avec V tant que V est une fonction de y seulement. On a alors l'égalité :

$$I_{\text{drift}}(y) = -W\mu_{\text{eff}} \frac{dV(y)}{dy} Q_{\text{inv}}(y) = -W\mu_{\text{eff}} \frac{dV(y)}{dy} Q_{\text{inv}}(V) \quad \text{Eq. I-33}$$

En multipliant chaque terme de l'équation I-33 par dy , puis en intégrant sur la longueur du canal donc de $y=0$ à $y=L$ et de $V=0$ à $V=V_d$, on a alors l'intégrale suivante :

$$\int_0^L I_{\text{drift}}(y) dy = \int_0^{V_d} -W\mu_{\text{eff}} \frac{dV}{dy} Q_{\text{inv}}(V) dy \quad \text{Eq. I-34}$$

Finalement, le courant de dérive entre source et drain est donné par l'expression :

$$I_{\text{drift}} = \mu_{\text{eff}} \frac{W}{L} \int_0^{V_d} -Q_{\text{inv}}(V) dV \quad \text{Eq. I-35}$$

La procédure exposée ci-dessus, menant à l'équation I-35 est la double intégration de Pao et Sah [Pao 66]. Enfin, pour aboutir à l'expression du courant de dérive, nous écrivons la charge d'inversion avec la formule simple tirée de la théorie classique de la capacité MOS :

$$Q_{\text{inv}}(V) = C_{\text{ox}}(V_g - V_{\text{th}} - V) \quad \text{Eq. I-36}$$

Où V est le potentiel le long du canal, avec $V(0)=0$ (polarisation de la source) et $V(L)=V_d$ (polarisation du drain). Finalement, en reportant l'équation I-36 dans I-35, on a par une simple intégration :

$$I_{\text{drift}} = \mu_{\text{eff}} C_{\text{ox}} \frac{W}{L} \left(V_g - V_{\text{th}} - \frac{V_d}{2} \right) V_d \quad \text{Eq. I-37}$$

L'équation I-37 donne une variation de courant de dérivation est donc linéaire avec V_g , et parabolique avec V_d avec un maximum atteint pour $V_d = V_g - V_{th}$.

Tant que $V_d < V_g - V_{th}$, on parle de régime linéaire, ou de régime ohmique.

Mais, lorsque $V_d = V_g - V_{th}$, l'équation I-36 donne une charge d'inversion en $y=L$ (donc au drain) nulle. L'équation I-37 de courant de dérivation n'est alors plus utilisable pour $V_d = V_g - V_{th}$ (puis pour $V_d > V_g - V_{th}$) et on peut alors définir la tension de saturation de drain $V_{dsat} = V_g - V_{th}$. Le courant dit de saturation (I_{dsat}) en $V_d = V_{dsat}$ vaut alors :

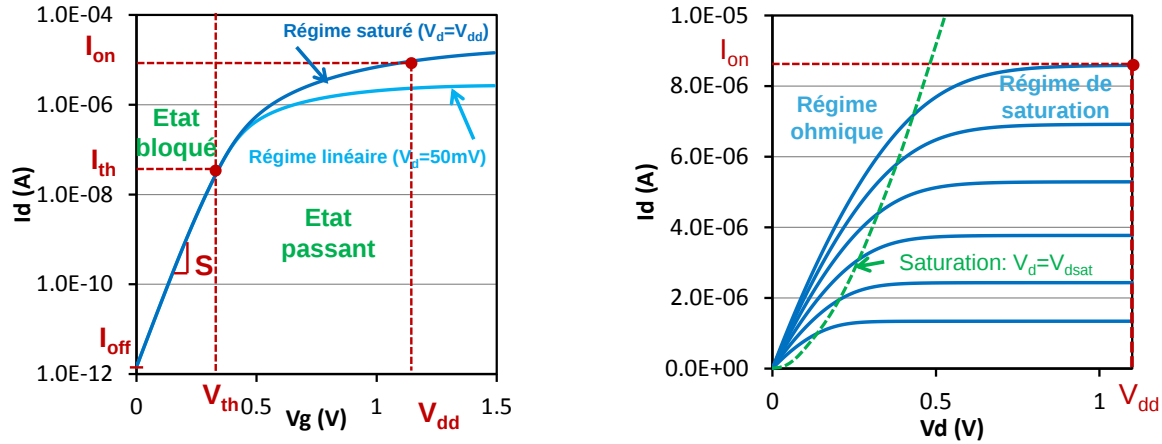
$$I_{dsat} = \mu_{eff} C_{ox} \frac{W}{L} V_{dsat}^2 \quad \text{Eq. I-38}$$

Nous pouvons alors définir le courant de saturation I_{on} , paramètre clé d'une technologie CMOS, qui sera égal au courant drain-source évalué avec l'équation I-38 pour $V_g = V_d = V_{dd}$. A tension d'alimentation fixée, on constate que le courant I_{on} est d'autant plus fort que la tension de seuil est faible ($V_{dsat} = V_g - V_{th}$). Nous avons vu dans le paragraphe précédent que la même règle s'applique au courant I_{off} . On estime qu'un transistor MOSFET est performant lorsqu'il présente un courant de fuite I_{off} faible, et un courant de saturation I_{on} élevé. On comprend alors aisément que le « choix » de la tension de seuil sera effectué pour satisfaire au mieux le compromis I_{on}/I_{off} , en fonction de l'application visée:

- Haute performance (type microprocesseur « CPU ») où on cherche à avoir le maximum de performance, donc de courant I_{on} . La fuite I_{off} n'est pas une contrainte très importante car ce type de système est connecté directement au réseau électrique. La minimisation de consommation n'est donc pas prioritaire et les tensions de seuil de ce type de dispositifs sont donc très basses.
- Basse consommation (type système sur puce « SOC ») où on cherche à minimiser la fuite, donc le courant I_{off} , car ce type d'application fonctionne sur batteries, et la question de l'autonomie du système est primordiale. Les tensions de seuil de ce type de dispositif sont donc élevées.

Il faut tout de même noter ici que les deux types d'application mentionnés ci-dessus tendent à converger vers une seule et même technologie du fait du développement des systèmes multimédias sans fils (notamment téléphones portables et tablettes numériques) qui doivent aujourd'hui fournir un niveau de performance très élevé tout en garantissant une autonomie importante.

La Figure I-10 représente les caractéristiques courant-tension d'un transistor MOSFET à canal long ($L=10\mu m$), obtenues par caractérisation électrique. Les équations de courant drain-source développées dans ce paragraphe pour le régime au-delà du seuil ($V_g > V_{th}$) et dans le paragraphe précédent pour le régime sous le seuil ($V_g < V_{th}$) peuvent reproduire chaque régime de fonctionnement, mais ne peuvent assurer les raccords continus entre chaque régime. Le développement des raccords de continuité fera l'objet des paragraphes II.D.3 et II.D.4. Nous avons également indiqué sur la Figure I-10 tous les paramètres électriques du transistor MOSFET d'ores et déjà définis.



a) b)
Figure I-10 : tracé du courant drain-source en fonction de la tension de grille V_g à tension de drain fixée à $V_d = V_{dd}$ (a) puis en fonction de la tension de drain V_d à tension de grille V_g fixée (à V_{dd} puis par pas de réduction de 100mV) (b)

I.B.3. La mobilité effective μ_{eff} à fort champ de grille

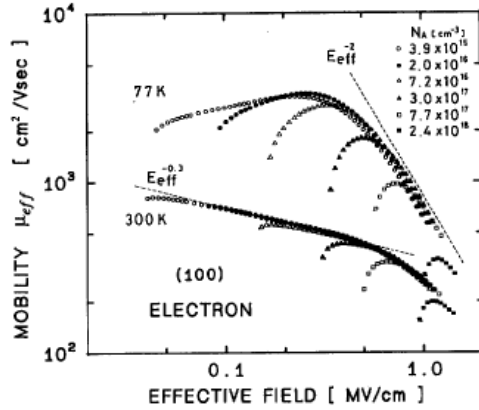
La mobilité effective μ_{eff} traduit la capacité des porteurs à être mis en mouvement par un champ électrique. Elle est donc représentative de du transport dans un transistor MOSFET. La vitesse de dérive v des porteurs est alors donnée par la relation :

$$v = \mu_{eff} E \quad \text{Eq. I-39}$$

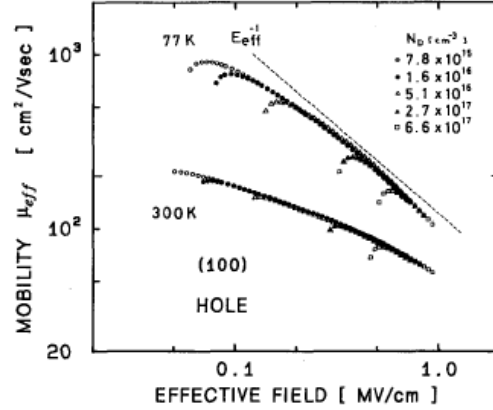
Où E est le champ électrique entre la source et le drain.
La mobilité effective est donnée par l'expression classique :

$$\mu_{eff} = \frac{q\tau_r}{m^*} \quad \text{Eq. I-40}$$

Où τ_r est le temps de relaxation (c'est-à-dire le temps moyen entre deux interactions) et m^* est la masse effective. Ce paramètre est crucial pour la performance d'un transistor car nous avons vu dans le paragraphe précédent que le courant de saturation I_{on} est proportionnel à la mobilité effective μ_{eff} . Une modélisation de ce paramètre se révèle très complexe [Fischetti 01] et difficile à mettre en place simplement. Il a cependant été démontré par [Takagi 94] que la mobilité effective μ_{eff} suit une loi universelle et que sa valeur peut donc être connue à partir de cette loi. La Figure I-11 montrent les lois universelles de variation de la mobilité effective pour les électrons (a) et pour les trous (b) dans le silicium extraites de [Takagi 94]. On remarque que la mobilité effective diminue quand le dopage augmente, ce qui était attendu car la mobilité est inversement proportionnelle à la probabilité d'interaction des porteurs, et que cette probabilité augmente avec la valeur du dopage. Formellement, si le dopage augmente le terme τ_r de l'équation I-40 diminue, donc la mobilité effective également. Les équations des lois universelles de mobilité seront données dans le chapitre II.



a) d'après [Takagi 94]

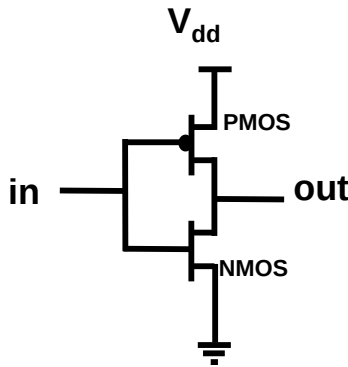


b) d'après [Takagi 94]

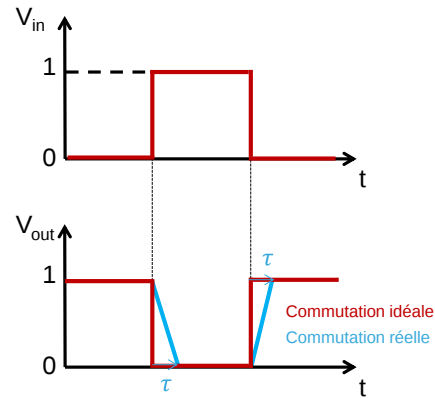
Figure I-11: Lois universelles de variation de la mobilité effective pour les électrons (a) et pour les trous (b) dans le silicium extraites de [Takagi 94].

I.B.4. Délai, Inverseur et Courant Effectif I_{eff}

Afin d'évaluer la vitesse d'une technologie CMOS, il est commun d'utiliser le délai τ d'un inverseur (schéma de principe donné sur la Figure I-12-a).



a)



b)

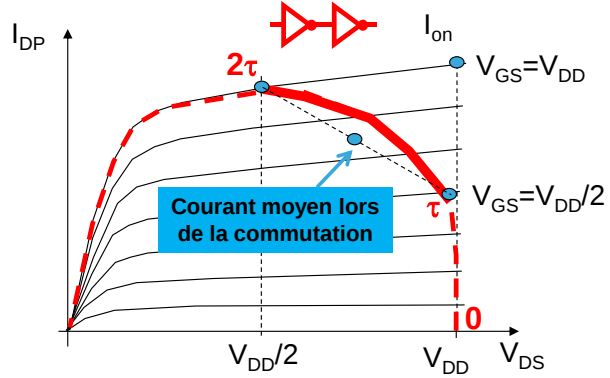
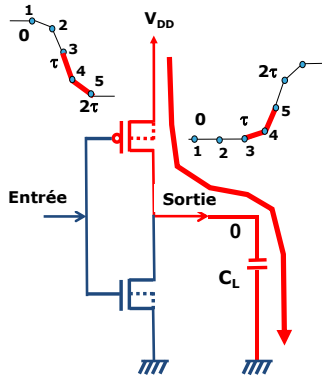
Figure I-12 : a) schéma d'un inverseur CMOS. b) illustration de la commutation d'un inverseur dans le cas idéal et dans le cas réel.

La commutation de l'inverseur est représentée sur la Figure I-12-b dans le cas idéal et dans le cas réel. Par un raisonnement très simple [Skotnicki 00], on a souvent évalué le délai de commutation d'un transistor par la relation :

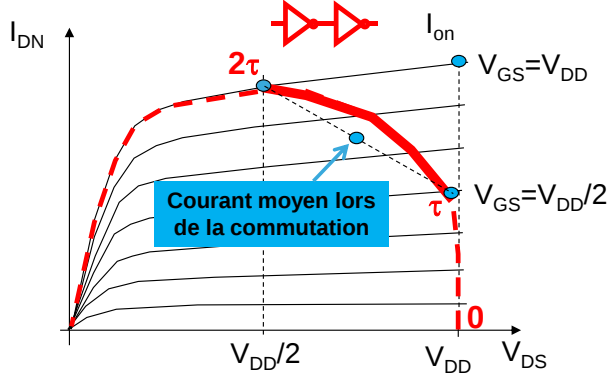
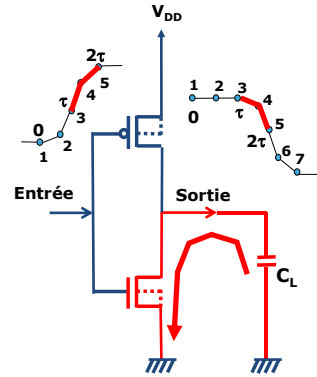
$$\tau = \frac{C_{ox}V_{dd}}{I_{on}} \quad \text{Eq. I-41}$$

En étudiant la trajectoire du courant drain-source lors de la commutation, il a été montré dans [Na 02] que dans le cas particulier d'une chaîne d'inverseur simple le courant n'atteint jamais la valeur I_{on} et que, par conséquent, l'équation I-39 n'est pas un bon indicateur du délai. En considérant que la commutation est irréversible lorsque la tension de drain a dépassé la valeur $V_{dd}/2$ [Skotnicki 09], on peut tracer les trajectoires des commutations de

l'inverseur vers l'état bas puis vers l'état haut. La Figure I-13-a montre le flux de courant sur le schéma de l'inverseur lors de la commutation à l'état bas et la trajectoire du courant de drain du PMOS, superposée à ses caractéristiques I_d-V_d . De la même façon, la Figure I-13-b montre le flux de courant sur le schéma de l'inverseur lors de la commutation à l'état haut et la trajectoire du courant de drain du NMOS, superposée à ses caractéristiques I_d-V_d . Sur ces deux graphes, les pointillés rouges indiquent la trajectoire totale du courant de drain et le trait plein rouge la trajectoire uniquement durant la commutation étudiée.



a) Passage à l'état haut.



b) Passage à l'état bas.

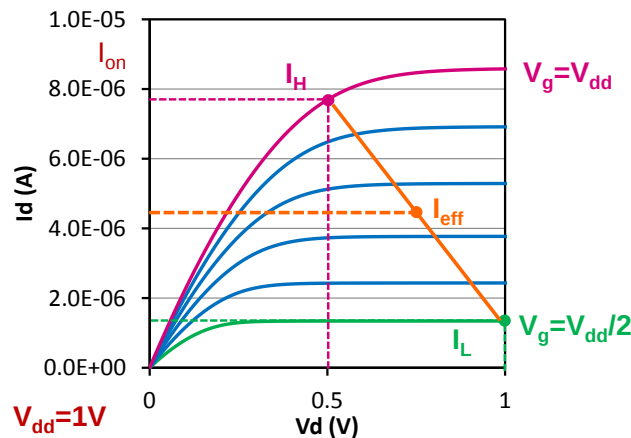


Figure I-13 : a) Passage à l'état haut d'un inverseur b) passage à l'état bas c) Illustration du courant effectif sur un réseau de courbes I_d-V_d .

On remarque bien que le courant de drain n'atteint jamais la valeur I_{on} et que le courant moyen durant la commutation est bien inférieur à I_{on} . [Na 02] propose alors de remplacer dans l'équation I-39 le courant I_{on} par ce courant moyen, qui sera donc plus représentatif du délai, noté I_{eff} pour courant effectif, donné par la formule:

$$\tau = \frac{C_{ox}V_{dd}}{I_{eff}} \quad \text{Eq. I-42}$$

$$I_{eff} = \frac{(I_H + I_L)}{2} \quad \text{Eq. I-43}$$

Avec $I_H = I_{ds}(V_g = V_{dd}, V_d = \frac{V_{dd}}{2})$ Eq. I-44

Et $I_L = I_{ds}(V_g = \frac{V_{dd}}{2}, V_d = V_{dd})$ Eq. I-45

Le courant effectif I_{eff} est représenté sur un réseau de courbes I_d - V_d sur la Figure I-13.

I.C. Réduction des dimensions

Le paragraphe précédent (I.B) donne le fonctionnement réaliste d'un transistor MOSFET à canal long où chaque dimension est relâchée (i.e. t_{ox} , X_j , CPP). Le dispositif ne souffre donc d'aucun effet parasite dû à la réduction des dimensions. L'objet de ce paragraphe est de décrire l'effet de la miniaturisation du transistor MOSFET sur ses paramètres électriques, puis sur ses caractéristiques courant-tension.

I.C.1. Règles de miniaturisation

Selon la loi de Moore, la densité d'intégration des transistors MOSFET doit doubler tous les deux ans, donc entre chaque nœud technologique. On peut traduire cela en disant que la surface d'un transistor doit être divisée par deux d'un nœud à l'autre. En considérant que la surface d'un transistor S est approximativement donnée par le produit de sa largeur W par CPP, on peut écrire que, pour le nœud suivant (S' , W' , CPP'), on a :

$$S' = \frac{S}{2} = \frac{WCPP}{2} = \frac{\sqrt{2}}{2} W \frac{\sqrt{2}}{2} CPP = W'CPP' \quad \text{Eq. I-46}$$

Par conséquent, on peut considérer la largeur W et le CPP du transistor sont réduits d'un facteur ~ 0.7 d'un nœud à l'autre. Si la dimension CPP est réduite d'un facteur 0.7 par nœud technologique, la longueur de grille doit suivre cette même loi pour ne pas trop complexifier le procédé de contact. Ceci se vérifie historiquement sur les anciens nœuds technologiques dont la dénomination correspondait à la longueur de grille minimale de la technologie : 130nm, 90nm, 65nm, 45nm [Bœuf 04] [Arnaud 04]. Cependant, pour les nœuds technologiques suivants (32nm, 22nm... [Arnaud09], [Cho 11]) les effets parasites dus à la réduction de longueur de grille (développés dans le paragraphe suivant I.C.2) sont trop difficiles à contrôler. La longueur de grille est donc relâchée et ne correspond plus à la dénomination du nœud. Pour réduire l'impact de la longueur de grille sur les paramètres électriques du transistor, l'épaisseur d'oxyde de grille est réduite de nœud en nœud afin d'améliorer le contrôle capacitif de la grille sur le canal (paragraphe I.C.3). Enfin, contrairement à la longueur de grille, le paramètre CPP poursuit sa réduction d'un facteur 0.7 de nœud en nœud. Par conséquent, pour les technologiques les plus avancés, la distance entre la grille et le contact de source ou de drain est réduite d'un facteur supérieur à 0.7, ce qui a pour effet d'augmenter le poids des capacités parasites sur la capacité totale du dispositif (paragraphe I.C.4).

I.C.2. Réduction de la longueur de grille

Lorsque la longueur de grille est réduite, l'impact des zones source et drain sur le canal de conduction ne peut plus être négligé. En effet, les interfaces entre source et substrat puis substrat et drain forment deux jonctions PN et une zone de déplétion se forme à chaque interface. La théorie classique des jonctions PN nous indique que cette zone de déplétion (ou zone de charge d'espace) s'étend du côté de la jonction qui est le moins dopé, donc dans le substrat. Les parties proches de l'oxyde de grille (donc l'emplacement du canal de conduction) des deux zones de charge d'espace ainsi formées sont alors moins bien contrôlées par la grille, car déplétées sous l'action de la source ou du drain. Pour les transistors à canal long, l'extension des zones de charges d'espace est négligeable devant la longueur de grille et ce phénomène n'a pas d'effet sur le comportement électrique du transistor (Figure I-14-a). Lorsque la longueur de grille est réduite, l'extension des zones de charge d'espace n'est plus négligeable devant la longueur de grille. La grille commence alors à perdre une partie de son contrôle électrostatique (Figure I-14-b). Enfin, si de plus une polarisation est appliquée sur le drain (la source est généralement à la masse), la zone de charge d'espace coté drain s'étend encore plus dans le substrat et son extension est d'autant plus grande que la polarisation de drain est forte (Figure I-14-c), jusqu'à, éventuellement, superposition entre les zones de charge d'espace. Le contrôle électrostatique est donc encore réduit par la polarisation de drain dans le cas d'un transistor à canal court.

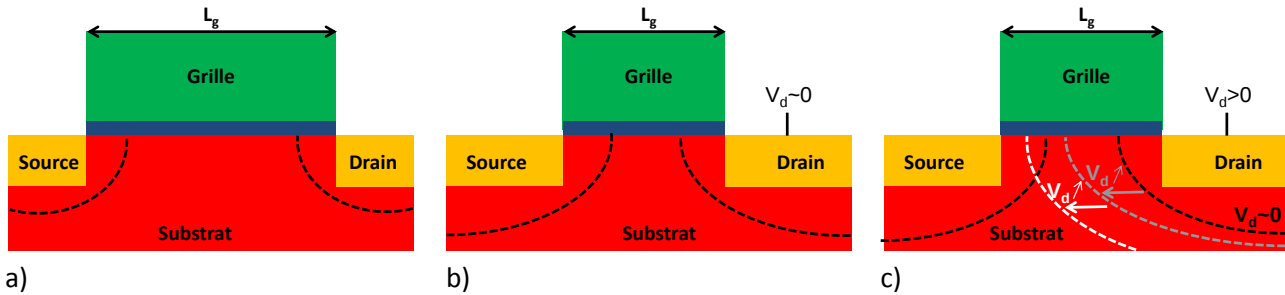


Figure I-14 : Illustration de l'extension de la zone de charge d'espace coté source et drain pour (a) un transistor à canal long (b) un transistor à canal court avec une polarisation de drain $V_d \sim 0$ (c) un transistor à canal court avec une polarisation de drain V_d forte. Les pointillés indiquent l'extension de la zone de charge d'espace.

La perte de contrôle électrostatique de la grille dans le cas d'un transistor à canal court se traduit par un abaissement de la tension de seuil. En effet, le rapprochement des zones de charge d'espace a pour effet d'abaisser la barrière de potentiel entre la source et le drain. La réduction de tension de seuil associée à ce premier effet définit le paramètre SCE (« Short Channel Effect », effet canal court en français). Si de plus une polarisation est appliquée sur le drain, la barrière de potentiel entre source et drain est encore réduite. La diminution de tension de seuil correspondante définit le paramètre DIBL (« Drain Induced Barrier Lowering » abaissement de barrière due à la polarisation de drain en français). La tension de seuil d'un transistor à canal court s'écrit alors :

$$V_{th} = V_{thlong} - SCE - DIBL \quad \text{Eq. I-47}$$

Où V_{thlong} est la tension de seuil du transistor à canal long, donnée par l'équation I-15.

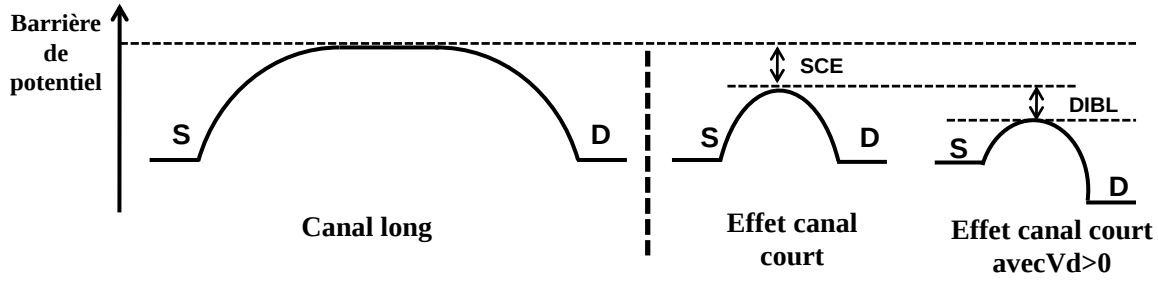


Figure I-15 : Illustration de l'abaissement de barrière de potentiel entre source et drain, donc de tension de seuil due à la réduction de longueur de grille (SCE) puis à la polarisation du drain (DIBL).

La tension de seuil dépend donc d'une part de la longueur de grille L_g , et d'autre part de la polarisation de drain. La transformation tension-dopage (VDT pour Voltage Doping Transformation [Skotnicki 88-a]) traduit l'abaissement de barrière de potentiel par une diminution du dopage canal. De cette manière, la résolution de l'équation de Poisson à deux dimensions est réduite à une dimension (les détails de cette méthode font l'objet du paragraphe II.A.3.a) et permet d'obtenir des expressions très simples des paramètres SCE et DIBL :

$$SCE = \frac{\epsilon_{si}}{\epsilon_{SiO_2}} \frac{t_{ox}}{L_g} \frac{T_{dep}}{L_g} \phi_d \quad \text{Eq. I-48}$$

$$DIBL = \frac{\epsilon_{si}}{\epsilon_{SiO_2}} \frac{t_{ox}}{L_g} \frac{T_{dep}}{L_g} V_d \quad \text{Eq. I-49}$$

Où ϕ_d est la tension de diode (formée par la jonction PN source-canal ou canal-drain), donnée par l'équation :

$$\phi_d = \frac{kT}{q} \ln\left(\frac{N_{ch} N_{sd}}{n_i^2}\right) \quad \text{Eq. I-50}$$

Avec N_{sd} niveau de dopage de la source et du drain.

La Figure I-1-a illustre la variation de tension de seuil avec la longueur de grille pour une polarisation de drain faible ($V_d=50\text{mV}$) ce qui définit la tension de seuil en régime linéaire V_{thlin} (la dépendance avec la longueur de grille ne vient que du paramètre SCE) et pour une polarisation de drain forte ($V_d=V_{dd}=1\text{V}$) ce qui définit la tension de seuil en régime de saturation V_{thsat} (la dépendance avec la longueur de grille provient des paramètres SCE et DIBL). Enfin, les équations I-45 et 46 montrent que les paramètres SCE et DIBL sont proportionnels à la profondeur de déplétion T_{dep} (Eq I-14) qui est d'autant plus faible que le dopage N_{ch} est élevé. SCE et DIBL sont plus faibles si le dopage est fort (Figure I-16-b), et, par conséquent, la grille a un meilleur contrôle électrostatique sur le canal si le dopage est élevé.

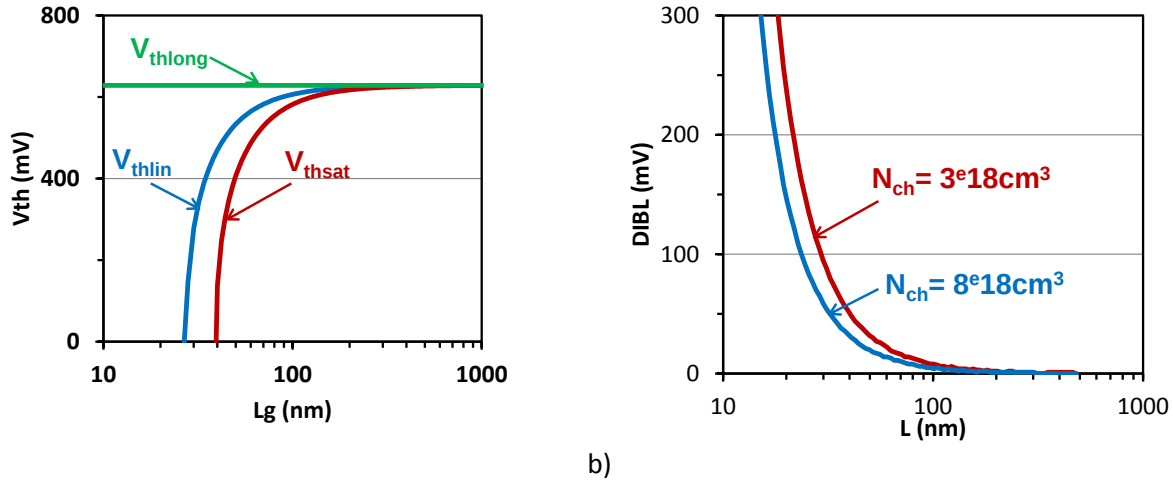


Figure I-16: a) illustration de la variation de la tension de seuil avec la longueur de grille et avec la polarisation de drain.

Examinons maintenant l'effet des paramètres SCE et DIBL sur les caractéristiques courant-tension du transistor MOSFET. La Figure I-17-a présente les courbes $\log(I_d)-V_g$ en régime linéaire ($V_d=50\text{mV}$) et saturé ($V_d=V_{dd}$) pour un transistor long et pour un transistor court. Comme attendu, pour le transistor long la polarisation de drain V_d n'a pas d'effet sur le régime sous le seuil ($V_g < V_{th}$): la tension de seuil et le courant de fuite I_{off} sont constants. Puis, lorsque la longueur de grille est réduite la tension de seuil du transistor diminue à cause de l'effet SCE (courbe rouge), puis baisse encore lorsqu'une polarisation de drain est appliquée à cause de l'effet DIBL (courbe verte). Ces deux abaissments successifs de tension de seuil provoquent l'augmentation du courant de fuite I_{off} , mais également du courant de saturation I_{on} . Il faut tout de même noter qu'une part de l'augmentation de courant drain-source entre transistor long et court est dû à la dépendance en $1/L$ de ce courant.

La Figure I-17-b représente deux réseaux d' I_d-V_d tracés pour deux transistors MOS présentant le même courant de saturation I_{on} . Pour le premier, le DIBL est nul (courbes bleues) alors que pour le second $\text{DIBL} \sim 100\text{mV}$ (courbes rouges). On remarque que si le DIBL est non nul, le courant drain-source en régime de saturation n'est plus constant, mais croissant avec V_d . Cette croissance de courant de saturation est justifiée par la réduction de tension de seuil dû au DIBL qui est de plus en plus importante quand la polarisation de drain augmente. Enfin, si on extrait le courant effectif I_{eff} de ces deux transistors, on constate que le courant effectif est supérieur pour le transistor MOSFET sans DIBL, bien que ces deux transistors présentent le même courant de saturation I_{on} . Le courant effectif étant représentatif du délai de l'inverseur (I.B.4), on peut donc en déduire que le DIBL a un effet sur le délai de commutation. Pour illustrer cette remarque, considérons deux technologies CMOS, l'une définie par (I_{on1} , I_{eff1} et DIBL_1) et la seconde par (I_{on2} , I_{eff2} et DIBL_2).

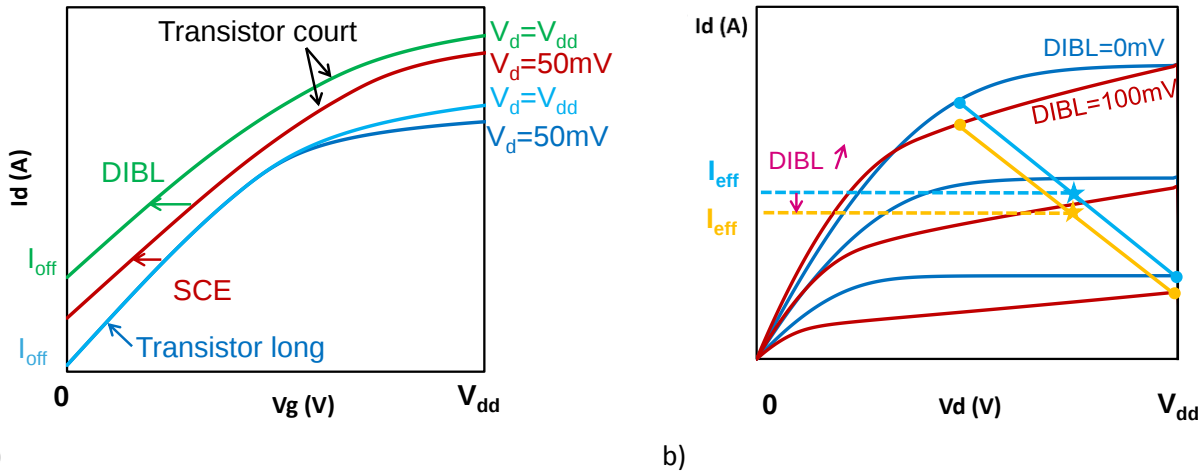
On a alors :

$$\text{Si} \quad \text{DIBL}_1 > \text{DIBL}_2 \quad \text{Eq. I-51}$$

$$I_{on1} > I_{on2} \quad \nRightarrow \quad I_{eff1} > I_{eff2} \quad \text{Eq. I-52}$$

$$\text{Donc} \quad I_{on1} > I_{on2} \quad \nRightarrow \quad \tau_1 < \tau_2 \quad \text{Eq. I-53}$$

Une technologie présentant le courant de saturation le plus élevé n'est donc pas forcément la plus rapides [Wei 09]. Ces simples constations démontrent la pertinence du courant effectif et mettent en exergue l'importance de comparer les technologies par le compromis courant effectif/courant de fuite (I_{eff}/I_{off}), en complément et non à la place du compromis courant de saturation/courant de fuite (I_{on}/I_{off}).



a) Figure I-17: a) Illustration de l'effet de la réduction de la longueur de grille sur une des courbes I_d - V_g en échelle logarithmique. b) Illustration de l'effet du DIBL sur un réseau d' I_d - V_d et sur le courant effectif I_{eff} .

On peut également constater sur la Figure I-17-a que la pente sous le seuil est dégradée par la réduction de longueur de grille. Ceci peut être également modélisé par la VDT [Skotnicki 88-a] et deux approches ont d'ores et déjà été proposées [Gwoziecki 02] et [Pouydebasque 07]. De plus amples détails théoriques seront donnés dans le paragraphe II.A.4 et la Figure I-18 donnent les variations de pente sous le seuil avec la longueur de grille données par [Pouydebasque 07]. Ces courbes démontrent également que la pente sous le seuil est moins dégradée par la longueur de grille si d'une part l'oxyde de grille est fin et d'autre part par si le dopage canal est fort.

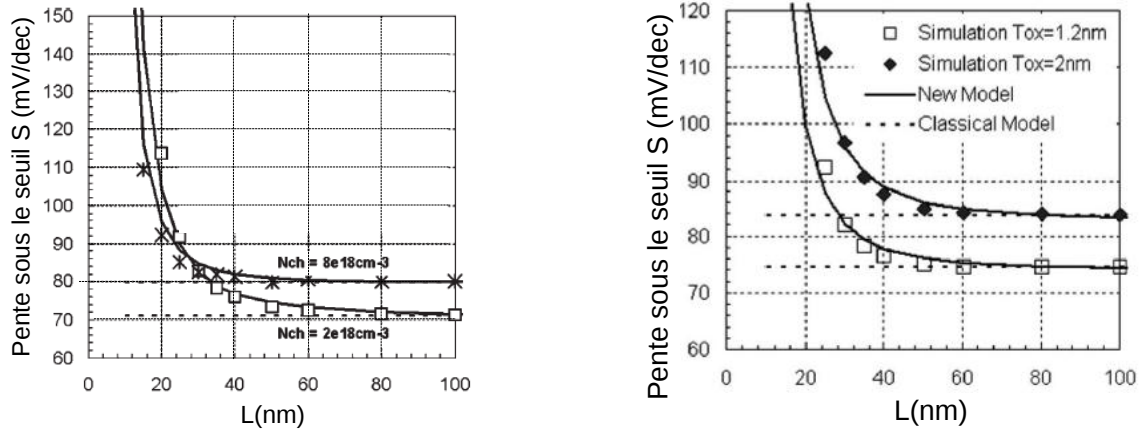


Figure I-18: Variation de pente sous le seuil avec la longueur de grille extraite de [Pouydebasque 07]

Enfin, il est démontré par exemple par [Antoniadis 01] que la réduction de longueur de grille a pour effet de réduire la mobilité effective. Cette dégradation de la mobilité a été modélisée par [Bidal 09].

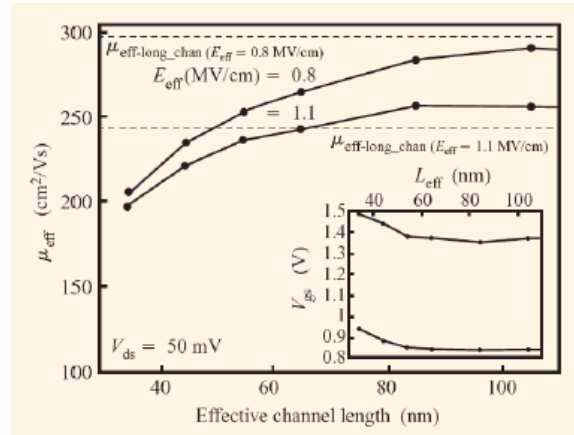


Figure I-19 : Variation de la mobilité effective en fonction de la longueur de grille donnée par [Antoniadis 01]

I.C.3. Réduction de l'épaisseur d'oxyde de grille

L'épaisseur de l'oxyde de grille t_{ox} est également réduite de nœud technologique en nœud technologique. Au vue des équations développées dans ce chapitre, ceci s'explique par trois raisons :

- Le courant de saturation (Eq. I-38) est proportionnel à $1/t_{ox}$, donc une réduction de t_{ox} induit un gain sur le courant I_{on} .
- Les paramètres SCE (Eq. I-45) et DIBL (Eq. I-46) sont proportionnels à t_{ox} , donc une réduction de t_{ox} induit une réduction des paramètres SCE et DIBL et le contrôle électrostatique de la grille est améliorée.
- La pente sous le seuil S est également améliorée, comme l'indique l'équation I-23.

Ces remarques sont vraies pour des épaisseurs de SiO_2 importantes (i.e. $t_{ox} > 2nm$, donc jusqu'au nœud 45nm où $t_{ox} = 1.7nm$ [Josse 06]). En dessous, il faut tenir compte de deux phénomènes parasites : la polydéplétion de grille pour les grilles en polysilicium, et la prise en compte des effets quantiques sur la distribution des porteurs minoritaires, plus communément appelé « darkspace ».

I.C.3.a) La polydéplétion de grille

Pour les grilles en polysilicium dopé à dégénérescence, une charge de déplétion se forme dans la grille, à l'interface grille-oxyde de grille en réaction à la charge présente dans le silicium. On notera l'épaisseur de cette zone déplétion $t_{polydep}$. Ce phénomène mis en évidence par exemple par [Josse 99] se traduit du point de vue électrique par un épaissement de l'oxyde de grille. Afin de garder le formalisme précédemment défini, il est courant de raisonner en épaisseur d'oxyde de grille équivalente EOT (Equivalent Oxide Thickness). L'EOT est calculée en considérant qu'il s'agit de l'épaisseur de la couche de SiO_2 de la capacité équivalente, formée par la mise en série de la capacité de polydéplétion ($C_{polydep}$) et de la capacité d'oxyde de grille (C_{ox}). Formellement :

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{polydep}} \Rightarrow \frac{EOT}{\epsilon_{SiO_2}} = \frac{t_{ox}}{\epsilon_{SiO_2}} + \frac{t_{polydep}}{\epsilon_{Si}} \quad \text{Eq. I-54}$$

$$EOT = t_{ox} + \frac{\epsilon_{SiO_2}}{\epsilon_{Si}} t_{polydep} \quad \text{Eq. I-55}$$

[Gautier 03] donne les valeurs usuelles des valeurs de polydéplétion à prendre en compte:

- NMOS : $\frac{\epsilon_{SiO_2}}{\epsilon_{Si}} t_{polydep} = 0.4nm$
- PMOS : $\frac{\epsilon_{SiO_2}}{\epsilon_{Si}} t_{polydep} = 0.6nm$

Naturellement, la polydéplétion n'est à prendre à compte que pour les grilles en polysilicium.

I.C.3.b) Le darkspace

Selon la théorie classique, en régime d'inversion, le maximum de la distribution de porteurs minoritaires est localisé à l'interface oxyde-silicium (courbe bleue sur la Figure I-20). Néanmoins, en réalité, donc avec la prise en compte des effets quantiques, le maximum de la distribution de porteurs minoritaires est localisé à une distance noté DS (pour darkspace) de l'interface oxyde-silicium (courbe rouge sur la Figure I-20). Ce phénomène s'explique par la quantification des niveaux d'énergie dans le puits triangulaire formé par la bande de de conduction [Gautier 03]. La distance entre l'oxyde de grille et le maximum de la distribution de porteurs minoritaires peut être considéré constante, mais dépend du type de transistor. Pour le silicium, on a :

- NMOS : $\frac{\epsilon_{SiO_2}}{\epsilon_{Si}} DS = 0.4nm$
- PMOS : $\frac{\epsilon_{SiO_2}}{\epsilon_{Si}} DS = 0.6nm$

Comme pour la polydéplétion, le darkspace a pour effet d'épaissir l'oxyde de grille du point de vue électrique. Il est alors également commun d'introduire une épaisseur d'oxyde de grille équivalente, notée t_{inv} , donnée par l'expression :

$$t_{inv} = EOT + \frac{\epsilon_{SiO_2}}{\epsilon_{Si}} DS \quad \text{Eq. I-56}$$

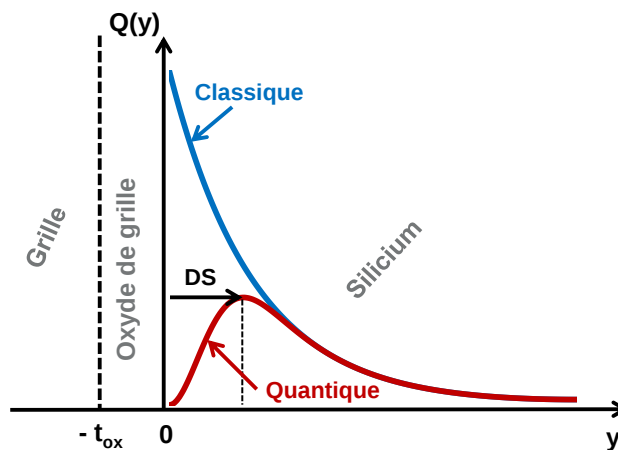


Figure I-20 : Distribution de porteurs donnée par la théorie classique (bleue) puis par la théorie quantique (rouge). Illustration du phénomène dit de darkspace, noté ici DS.

I.C.3.c) Correction du formalisme

Avec l'introduction des épaisseurs équivalentes (EOT et t_{inv}), tenir compte de la polydépétion et du darkspace par l'intermédiaire des équations précédemment développées est très simple. Remarquons d'abord que la polydépétion existe en régime de dépétion et d'inversion (i.e $V_g > 0$ pour un NMOS, $V_g < 0$ pour un PMOS), c'est-à-dire dans les conditions de polarisation classiques d'une technologie CMOS. Remarquons également que le darkspace a un sens uniquement pour le régime d'inversion forte, car sa correction porte sur la distribution des porteurs minoritaires. Par conséquent, dans les équations précédentes, l'épaisseur d'oxyde de grille t_{ox} devra être remplacé par:

- t_{inv} pour les paramètres électriques définis en inversion forte, donc pour $V_g > V_t$, comme I_{on} par exemple.
- EOT pour tous les autres paramètres (V_t , DIBL, $S_{..}$).

I.C.3.d) Fuites de grille

Nous venons de voir que pour les technologies CMOS avancées ($t_{ox} < 2nm$), une réduction de 10% de t_{ox} n'apporte pas un gain de 10% en capacité d'oxyde de grille, et donc sur les paramètres du transistor (DIBL, I_{on} ...) du fait de la polydépétion et du darkspace. De plus, la réduction de l'oxyde se paie par une complexification du procédé de fabrication, mais également par l'apparition de nouvelle composante de fuite, à travers la grille. La théorie classique indique que l'oxyde de grille forme une barrière infranchissable pour les porteurs minoritaires. Cependant, selon la théorie quantique, certains peuvent la franchir par effet tunnel et constitués un courant. Ce courant, dit courant de fuite de grille, devient non négligeable lorsque l'épaisseur d'oxyde de grille devient très faible ($t_{ox} < 2nm$), et peut devenir comparable au courant I_{off} . La fuite de grille peut alors devenir la principale source de fuite du transistor, comme illustrée par la Figure I-21 qui compare la variation de fuite de grille mesurée par [Chen 08] en fonction de l'épaisseur équivalente d'oxyde de grille en inversion à deux valeurs typiques d'application basse consommation de courant I_{off} pour une longueur de grille de 30nm.

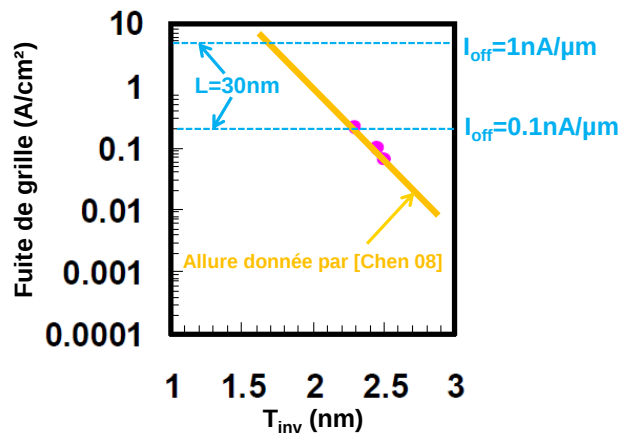
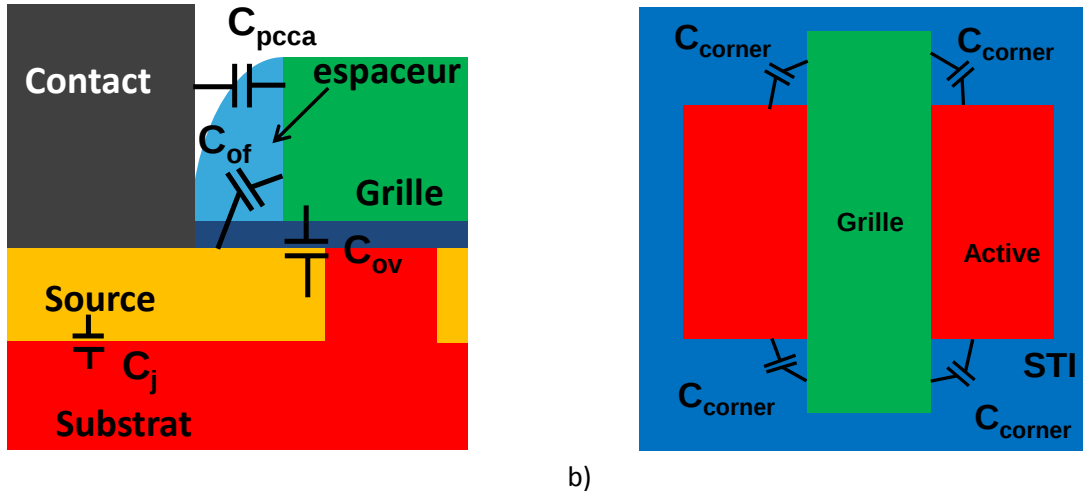


Figure I-21 : Comparaison la variation de fuite de grille mesurée par [Chen 08] en fonction de l'épaisseur équivalente d'oxyde de grille en inversion à deux valeurs typiques d'application basse consommation de courant I_{off} pour une longueur de grille de 30nm.

I.C.4. Réduction de la distance grille-contact

La réduction du pas de répétition d'une grille et d'un contact (CPP) implique une réduction de la distance entre la grille et le contact et donc une augmentation de la capacité entre la grille et le contact. Les capacités parasites ne sont cependant pas limitées à la capacité grille contact, ces dernières sont représentées sur la Figure I-22 (chacune sera détaillée au cours du chapitre III).

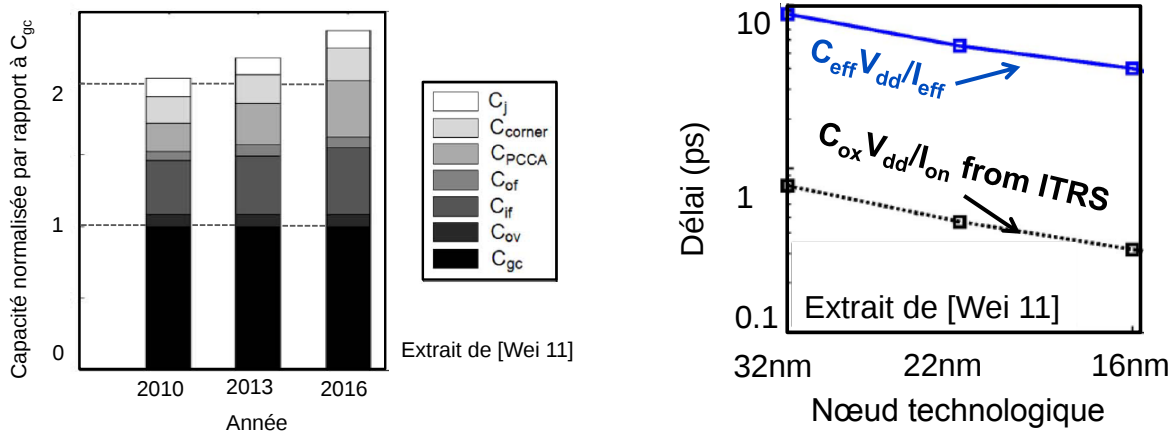


a) b)
Figure I-22 : Représentation des capacités parasites sur une vue en coupe (a) et une vue de dessus (b).

[Wei 11] propose un modèle simple pour évaluer chacune des composantes puis trace leur évolution en fonction de l'année, d'après les données de la feuille de route ITRS [ITRS] (Figure I-23-a). On remarque que le poids des parasites sur la capacité totale du dispositif augmente de nœud en nœud et ne doivent pas être négligés car leur somme est déjà égale à C_{gc} (c'est-à-dire C_{ox}) pour l'année 2010. Afin d'évaluer leur impact sur le délai de l'inverseur (et donc sur la vitesse d'un circuit) [Wei 11] propose une méthode pour évaluer la capacité équivalente C_{eff} du transistor comprenant les capacités parasites. Le délai peut alors être obtenu par l'équation :

$$\tau = \frac{C_{eff} V_{dd}}{I_{eff}} \quad \text{Eq. I-57}$$

La Figure I-23-b montre la variation du délai de l'inverseur avec le nœud technologique évaluée avec l'équation I-40 (donc sans prise en compte des capacités parasites) et avec l'équation I-55 (donc avec prise en compte des capacités parasites). Ce graphe démontre l'importance de tenir compte des capacités parasites dans les estimations de performances circuit d'une technologie CMOS.



a) b)
Figure I-23 : a) Représentation des capacités du transistor normalisées par la capacité grille-canal C_{gc} (i.e. C_{ox}) en fonction de l'année selon les spécifications de l'ITRS. b) tracé de l'évolution du délai avec le nœud technologique avec (courbe bleue) et sans (courbe noire) prise en compte de parasites. Ces deux graphes sont extraits de [Wei 11].

I.D. Solutions technologiques pour poursuivre la course à la miniaturisation

Dans cette partie, nous allons décrire les améliorations technologiques nécessaires à l'architecture conventionnelle sur substrat massif afin de pouvoir poursuivre la réduction de ses dimensions sans être trop impactée par les effets parasites décrits au paragraphe précédent.

I.D.1. Ingénierie de jonction

I.D.1.a) Zone de recouvrement

La source et le drain d'un transistor MOSFET sont réalisés par implantation ionique suivie d'un recuit haute température d'activation des dopants. Ce recuit a également pour effet de faire diffuser les dopants ce qui mène à une zone de recouvrement entre la grille d'une part et la source et le drain d'autre part. La longueur de cette zone est alors notée dL et est considérée identique pour la source et pour le drain. Par cet effet, la distance entre la source et le drain n'est plus égale à la longueur de grille. Le chemin que parcourent les porteurs dans le canal est donc plus court que la longueur de grille. Du point de vue électrique, la longueur du transistor n'est plus la longueur de grille, mais la longueur dite électrique L_{el} (illustrée par la Figure I-24) donnée par l'équation :

$$L_{el} = L_g - 2dL \quad \text{Eq. I-58}$$

Pour tenir compte des zones de recouvrements dans les équations des paramètres du transistor, il suffit de remplacer L_g par L_{el} .

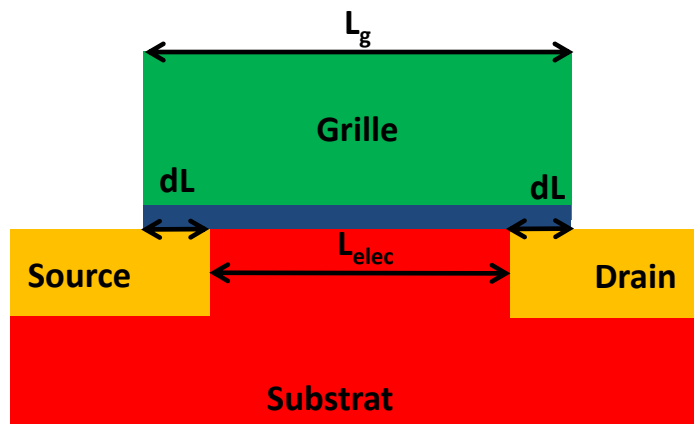


Figure I-24 : Illustration des zones de recouvrement de source et drain et de la longueur électrique L_{elec} .

I.D.1.b) Implantation poches

Au cours du paragraphe I.C.2, nous avons montré que le contrôle électrostatique de la grille sur le canal est amélioré si le niveau de dopage canal N_{ch} est augmenté. Donc, si N_{ch} augmente les paramètres DIBL, et SCE sont réduits tout comme la pente sous le seuil S si la longueur de grille est agressive. On serait donc tenté de doper très fortement le canal, mais cela aurait pour effet d'augmenter la tension de seuil canal long (par le terme Q_{dep}/C_{ox} de l'équation I-12) mais également de réduire la mobilité (I.B.3), donc le courant de saturation I_{on} (I.B.2). L'idéal serait donc d'augmenter la valeur du dopage uniquement pour les longueurs de grille les plus courtes. Afin d'atteindre cet objectif, des zones fortement dopées de même type que le canal (les poches) sont réalisées par implantation ionique. On parle alors d'implantation poche (Figure I-25).

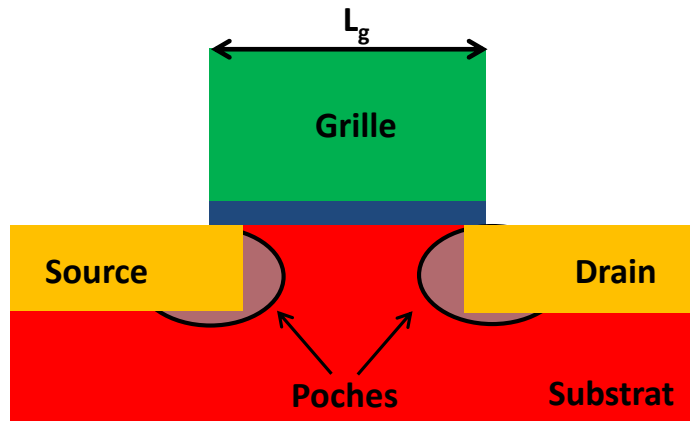


Figure I-25 : illustration des poches sur un transistor court.

Les poches n'ont pas d'effet sur le transistor long, car leurs dimensions sont négligeables devant la longueur de grille. Par contre, lorsque que la longueur de grille diminue, les poches ont pour effet d'augmenter le dopage moyen du canal du transistor, et donc d'améliorer son contrôle électrostatique. Enfin, lorsque la longueur de grille est très agressive, les poches se rejoignent et le dopage canal n'augmente plus lorsque la longueur de grille diminue, mais reste constant à la valeur égale au niveau de dopage des poches (Figure I-26-a). L'effet des poches sur la tension de seuil est illustré par la Figure I-26-b. Comme attendu, lorsque la longueur de grille est très importante, les poches n'ont pas d'effet sur le dopage canal moyen et la tension de seuil est la même avec et sans poche. Puis la longueur de grille est réduite, le dopage moyen commence à augmenter et, comme les effets canaux courts (SCE et DIBL) sont encore faibles, la tension de seuil augmente. Ensuite, lorsque la longueur de grille est agressive, les effets canaux courts deviennent très important et l'augmentation de dopage moyen ne suffit plus à les compenser et la croissance de la tension de seuil avec le dopage moyen commence à s'atténuer. Enfin, lorsque les poches recouvrent l'intégralité du canal, le dopage moyen reste constant si on diminue encore la longueur de grille et la variation de tension de seuil est uniquement régie par les effets canaux courts.

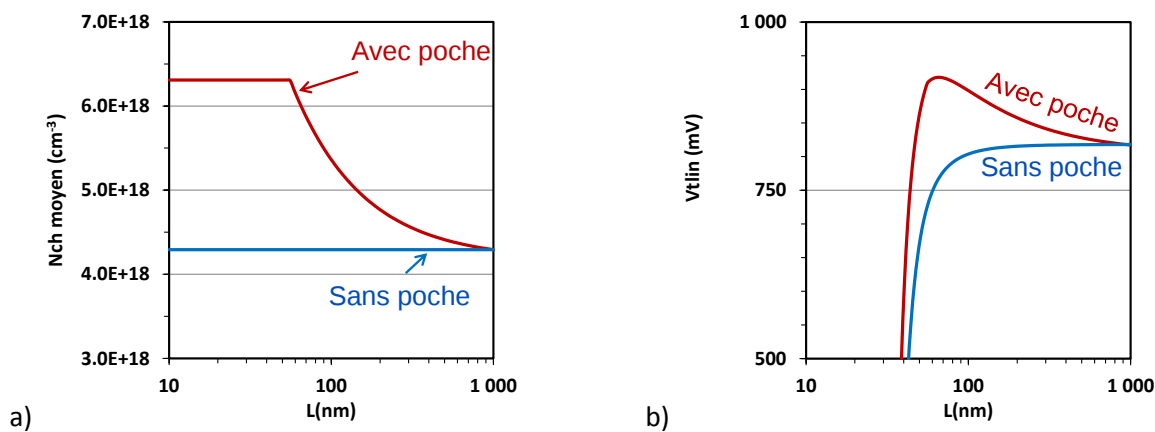


Figure I-26 : a) Variation du dopage moyen avec la longueur de grille avec et sans poches. b) variation de la tension de seuil en régime linéaire avec et sans poche

La Figure I-26-b montre que les implantations poches permettent bien d'améliorer le contrôle des effets canaux courts.

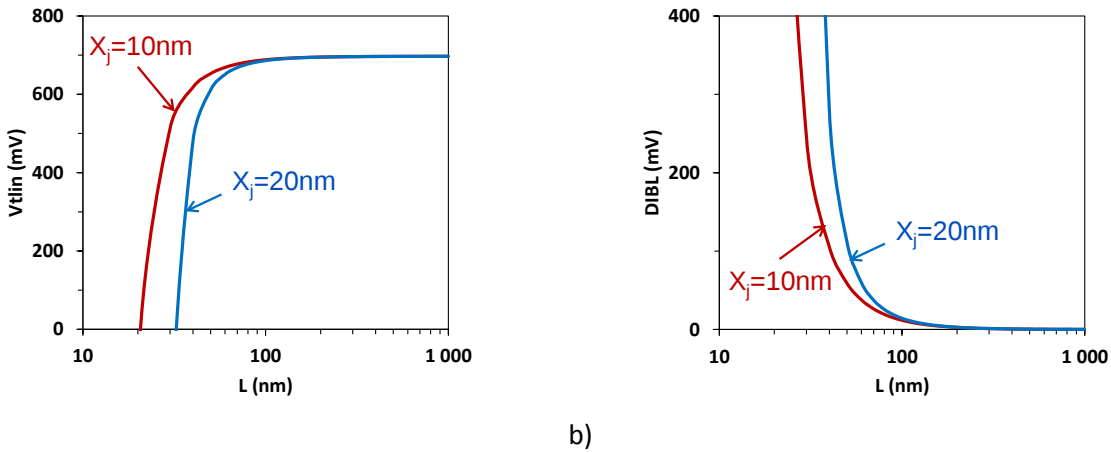
I.D.1.c) Jonctions fines

[Gautier 03] indique que l'analyse des lignes de champs entre source et drain montre que ces dernières sont plus courtes dans le cas de jonctions profondes. On peut donc en déduire que la longueur électrique moyenne est réduite si la jonction est profonde (i.e. X_j fort) et que les effets canaux courts (SCE et DIBL) sont plus importants dans le cas de jonctions profondes. Pour en tenir compte [Gautier 03] propose de corriger les équations I-48 et 49 à partir d'observation géométrique par le terme $(1 + X_j^2/L_{el}^2)$ comme suit :

$$SCE = \frac{\epsilon_{si}}{\epsilon_{SiO_2}} \frac{t_{ox}}{L_{el}} \frac{T_{dep}}{L_{el}} \left(1 + \frac{X_j^2}{L_{el}^2}\right) \phi_d \quad \text{Eq. I-59}$$

$$DIBL = \frac{\epsilon_{si}}{\epsilon_{SiO_2}} \frac{t_{ox}}{L_{el}} \frac{T_{dep}}{L_{el}} \left(1 + \frac{X_j^2}{L_{el}^2}\right) V_d \quad \text{Eq. I-60}$$

La Figure I-41 montre les variations de la tension de seuil en régime linéaire V_{tlin} (a) et du DIBL (b) avec la longueur de grille tracées avec les équations I-59 et I-60. Ces deux graphes démontrent que l'effet de la profondeur de jonction X_j ne doit pas être négligé, et qu'il est intéressant de réduire la profondeur de jonction pour améliorer l'électrostatique du transistor. Cependant, comme nous les verrons dans le paragraphe I.D.3, réduire la profondeur de jonction a également pour effet d'augmenter la résistance d'accès du dispositif et donc de dégrader, notamment le courant de saturation I_{on} .



a) b)
Figure I-27 : Variations de la tension de seuil en régime linéaire V_{tlin} (a) et du DIBL (b) avec la longueur de grille tracées avec les équations I-58 et I-59.

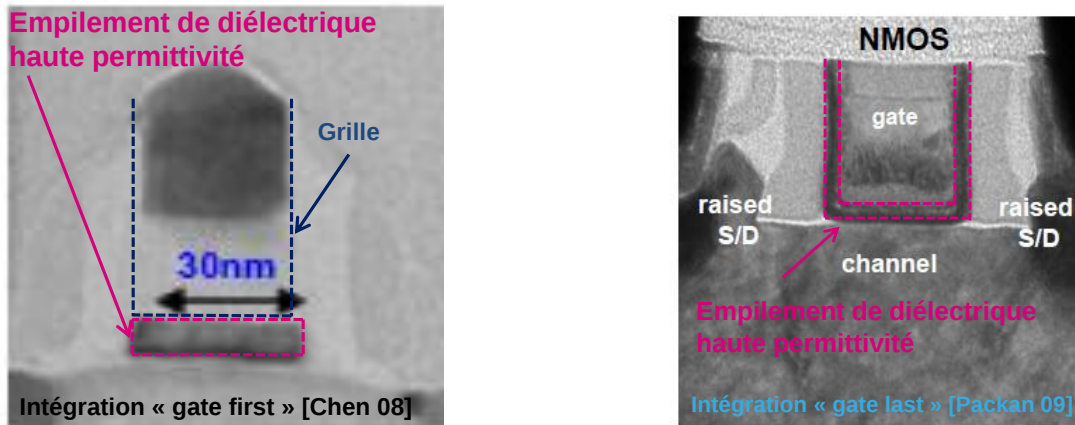
I.D.2. Empilement diélectrique haute permittivité-grille métallique

Comme nous l'avons vu dans le paragraphe I.C.3, la réduction de l'épaisseur d'oxyde de grille est limitée par le darkspace et par la polydéplétion. Le darkspace est intrinsèque à tout semiconducteur et est incompressible. Par contre la polydéplétion n'existe que pour les grilles en polysilicium, et pour s'en affranchir, il « suffit » d'utiliser une grille métallique. L'introduction des grilles métalliques a eu lieu pour le nœud 45nm pour les technologies développées par Intel [Mistry 07] et 32-28nm pour les autres industriels [Chen 08] [Arnaud 09]. Les grilles métalliques peuvent être intégrées de deux manières distinctes :

- L'approche dite « Gate first »: l'intégration reste la même, mais le matériau de grille est désormais un métal. La difficulté repose sur la gravure de ce métal. De plus, la suite du procédé de fabrication comporte des recuits hautes températures qui peuvent modifier le travail de sortie du métal de grille [Westlinder 04]. C'est l'intégration choisie par [Chen 08] [Arnaud 09].

- L'approche dite « Gate last »: l'intégration du transistor est conventionnelle, mais la grille en polysilicium est retirée à la fin du procédé pour être remplacée par la grille métallique. On parle aussi d'approche « damascene » [Guillaumot 02]. De cette manière, on s'affranchit du problème de la gravure du métal. De plus le métal « ne voit pas » tout le procédé de fabrication et n'est pas affecté par les recuits haute température. La difficulté de cette approche réside dans l'étape finale de planarisation de la grille. C'est l'intégration choisie par [Mistry 07].

Une autre différence entre ces deux types d'intégration est la présence de l'empilement de diélectrique haute permittivité sur les flancs de la grille uniquement dans le cas de l'intégration « gate last » (Figure I-28). Ceci peut sembler anecdotique, mais aura son importance dans l'évaluation des capacités parasites (Chapitre III).



a) b)
Figure I-28: Photos prise au microscope électronique par transmission montrant la présence de l'empilement de diélectrique haute permittivité pour l'intégration gate last b), contrairement à l'intégration « gate first » a).

L'intégration de grille métallique crée également un nouveau levier d'ajustement de la tension de seuil, par l'intermédiaire de son travail de sortie ϕ_m différent de celui du polysilicium (Figure I-29).

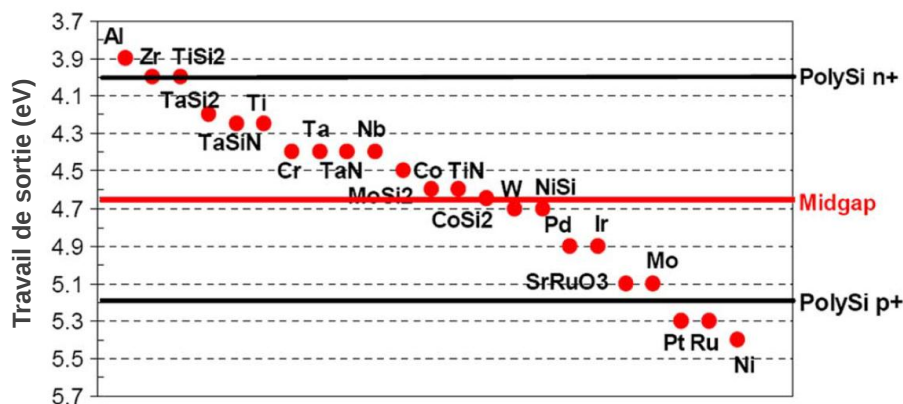


Figure I-29: Travaux de sortie ϕ_m sur silicium de différents métaux et alliages, d'après [Skotnicki 08].

Nous avons également vu dans le paragraphe I.C.3.d) que la réduction de l'épaisseur d'oxyde de grille se heurtait à la limite due à l'augmentation des fuites de grille. L'introduction de diélectrique haute permittivité (High-K en

anglais, $\epsilon_{HK} > 10$) en remplacement du SiO_2 en tant qu'oxyde de grille permet d'obtenir les mêmes valeurs de EOT et de t_{inv} mais avec une épaisseur de diélectrique supérieure. Les fuites de grille dépendant exponentiellement de l'épaisseur de diélectrique (et pas à l'EOT, ni au t_{inv}), un diélectrique haute permittivité permet de garantir un même contrôle électrostatique tout en réduisant le niveau des fuites de grille (Figure I-30).

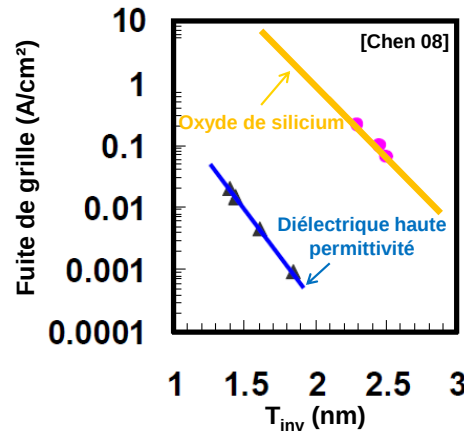


Figure I-30 : Comparaison de la variation de fuite de grille mesurée par [Chen 08] en fonction de l'épaisseur équivalente d'oxyde de grille en inversion pour un oxyde de grille en oxyde de silicium (jaune) puis en diélectrique haute permittivité.

Pour conserver une excellente interface avec le silicium, le diélectrique haute permittivité (son épaisseur est notée t_{HK}) n'est pas déposé directement sur le silicium et on dépose d'abord une couche de SiO_2 très fine ($< \sim \text{nm}$), on parle d'oxyde piédestal et on note son épaisseur $t_{oxpiéd}$.

Enfin, pour conserver le formalisme précédent, on continuera à raisonner en EOT et t_{inv} , qui, dans le cas d'une grille métallique avec diélectrique haute permittivité (HKMG pour High-K Metal Gate), sont donnés par les équations :

$$EOT = t_{ox} + \frac{\epsilon_{\text{SiO}_2}}{\epsilon_{HK}} t_{HK} \quad \text{Eq. I-61}$$

$$t_{inv} = t_{ox} + \frac{\epsilon_{\text{SiO}_2}}{\epsilon_{HK}} t_{HK} + \frac{\epsilon_{\text{SiO}_2}}{\epsilon_{Si}} t_{DS} \quad \text{Eq. I-62}$$

I.D.3. Résistance d'accès

Jusqu'à présent, nous avons considéré que le chemin parcouru par les porteurs entre le canal et le contact était un métal parfait et que les polarisations appliquées sur la source et le drain sont celles qu'on retrouve aux bornes du canal de conduction. Or, en réalité, il faut tenir compte de la résistance d'accès (représentée sur la Figure I-31) qu'on peut décomposer en trois principales composantes :

- La composante dans la zone de silicium dopé formant la source ou le drain, qui dépend donc du niveau de dopage et de sa profondeur X_j .
- La composante due au siliciure, liée à la qualité du procédé de fabrication.
- La composante due au contact.

La résistance d'accès entraîne une chute de potentiel entre le haut du contact (i.e. où sont appliqués V_d et V_s) et le canal. Par conséquent, la polarisation drain-source effectivement aux bornes du canal est inférieure à la

polarisation V_{ds} appliquée, ce qui entraîne une réduction du courant de saturation. Diverses méthodes sont proposées pour en tenir compte dans la littérature [Gautier 03] [Taur 98], nous en détaillerons une au cours du chapitre V.

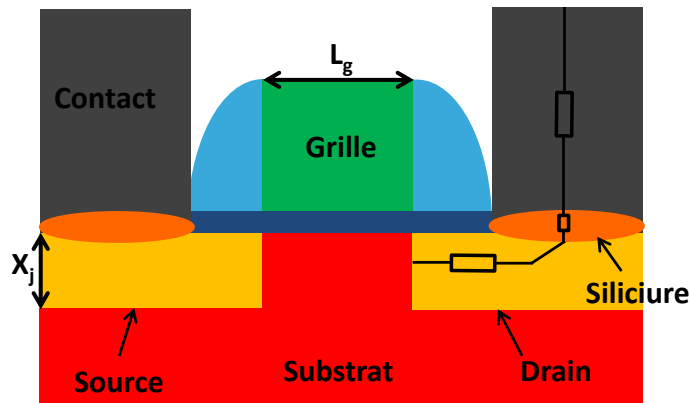


Figure I-31 : représentation schématique des composants de la résistance d'accès d'un transistor MOSFET.

I.D.4. Contact à prise ou Contact ruban ?

Historiquement, les transistors MOS sont contactés par des contacts dits « à prise ». Mais, avec la réduction de la longueur de grille, l'impact de la résistance d'accès sur le courant de saturation I_{on} est de plus en plus important et il devient indispensable de réduire la valeur de cette résistance. Une solution est de changer de schéma de contacts et de passer au contact ruban, où le contact est une barre occupant toute la largeur du transistor. Ces deux schémas de contacts sont représentés schématiquement sur la Figure I-32. Dans les faits, les contacts à prise ont été utilisés par toutes les plateformes CMOS industrielles jusqu'au nœud 45nm. Puis à compter de ce nœud, Intel a choisi de passer au contact ruban [Auth 08] et a naturellement conservé ce choix pour les nœuds suivants 32nm [Packan 09] puis 22nm [Auth 12]. Les autres industriels ont conservé le contact à prise jusqu'au nœud 32nm inclus [Arnaud 09], et sont passés au contact ruban à partir du nœud 20nm [Shang 12].

Le gain en résistance d'accès apporté par les contacts ruban est cependant à pondérer, car ces derniers impliquent une augmentation importante de la capacité entre la grille et le contact et sont donc pénalisant pour la vitesse de la technologie. L'évaluation de cette pénalité capacitive fera l'objet du chapitre III.

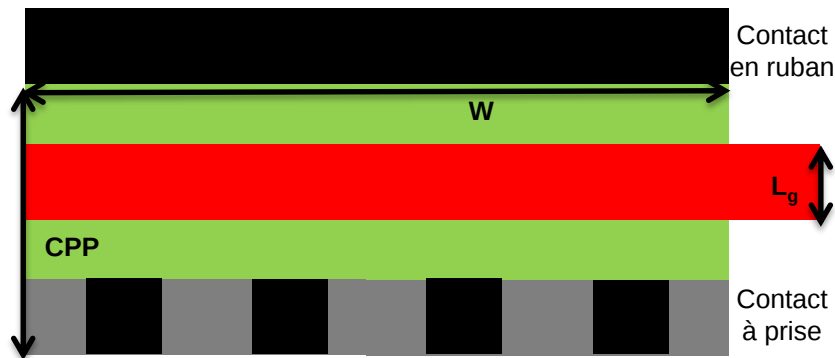


Figure I-32 : illustration des schémas des deux schémas de contact concurrents.

I.D.5. Epitaxie facettée et espaceur faible permittivité

Afin de limiter la résistance d'accès, les source-drains surélevés sont introduit par Intel pour le nœud 90nm [Mistry 04], puis deviennent un standard à partir du nœud 20nm (Figure I-33-a). Cependant, leur introduction génère l'apparition d'une nouvelle composante de capacité parasite (représentée sur la Figure I-33-b). Cette

dernière joue un rôle prépondérant car la distance qui la sépare de la grille est plus faible que la distance grille-contact.

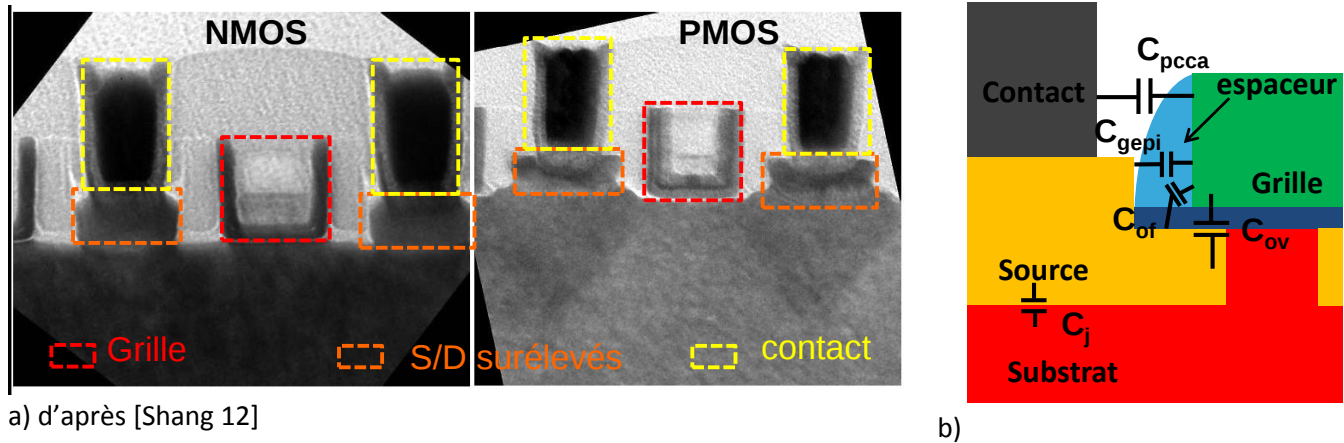


Figure I-33 : a) Photos prises au microscope électronique par transmission montrant les source-drains surélevés de la technologie CMOS 20nm présentée par [Shang 12]. b) Représentation schématique des différentes capacités parasites sur un transistor présentant des source-drains surélevés.

Afin de limiter cette nouvelle capacité parasite, il est proposé d'utiliser des épitaxies avec une facette pour la formation des source-drains dans le but d'augmenter la distance entre la grille et l'épitéxie et donc de réduire la capacité grille-épitaxie (Figure I-34-a). Ce type d'intégration est utilisé par [Cheng 11].

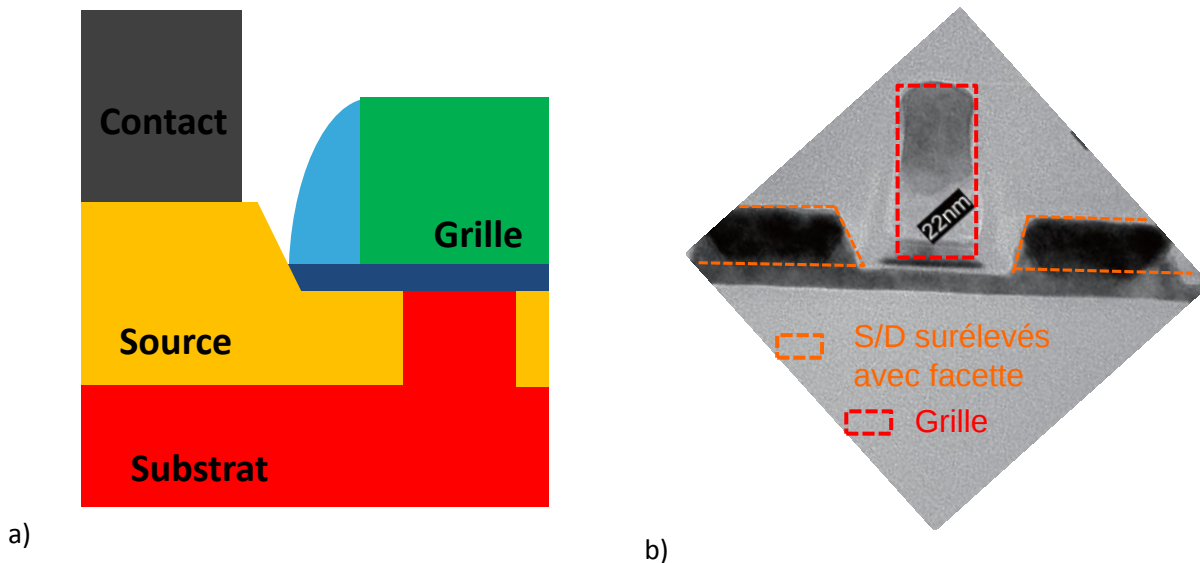


Figure I-34 : a) représentation schématique de source-drains formés par une épitéxie présentant une facette. b) Photo prise au microscope électronique par transmission montrant les source-drains surélevés réalisés avec une épitéxie présentant une facette [Cheng 11]

Une seconde approche pour réduire cette capacité est de changer le matériau de l'espaceur, usuellement en nitrure de silicium (permittivité relative $\epsilon=7$) par un matériau à plus faible permittivité comme le SiCON [Zhang 08] qui présente des permittivités relatives comprises entre 2 et 3.

I.D.6. Amélioration du transport

Ce paragraphe a pour but d'exposer les différentes techniques mises en œuvre afin d'améliorer le transport dans un transistor MOSFET, donc d'augmenter la mobilité effective μ_{eff} et par conséquent le courant de saturation I_{on} (I_{on} est proportionnel à μ_{eff} selon l'équation I-38). Ce gain en mobilité peut être obtenu en déformant la maille cristalline du canal avec application d'une contrainte mécanique, mais également par changement d'orientation cristalline.

I.D.6.a) Application d'une contrainte mécanique pour déformer la maille cristalline

L'application d'une contrainte mécanique provoque une déformation de la maille cristalline du silicium entraînant une modification de la structure de bande et de la masse effective. Les conditions de contrainte favorables aux électrons ne sont pas les mêmes que celles qui sont favorables aux trous. La Figure I-35 illustre l'exemple d'une contrainte uniaxiale suivant L:

- Si celle-ci est compressive, elle améliore la mobilité effective des trous (donc du PMOS) et dégrade celle des électrons (donc du NMOS).
- A l'inverse, si elle est tensile, la contrainte dégrade la mobilité effective des trous (donc du PMOS) alors qu'elle améliore celle des électrons (donc du NMOS).

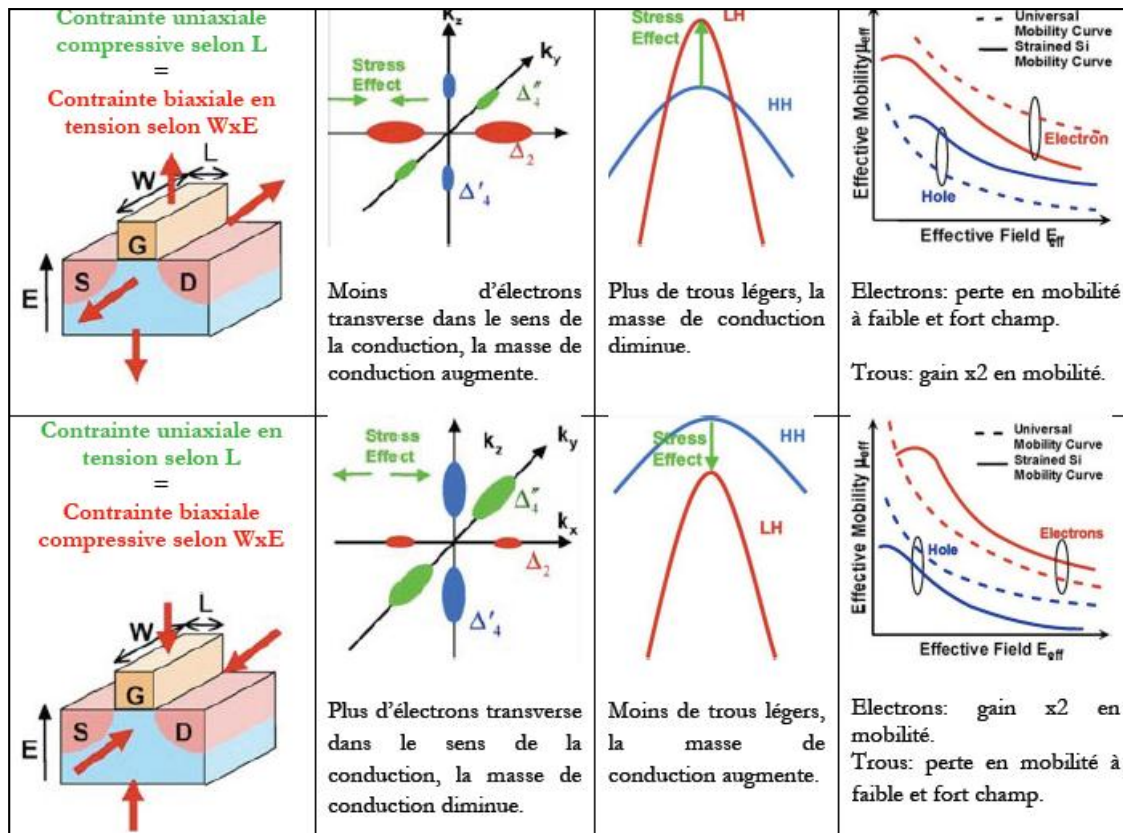


Figure I-35 : Analyse de l'impact d'une contrainte uniaxiale suivant la longueur de grille sur la mobilité des porteurs [Skotnicki08].

Une contrainte uniaxiale peut être obtenue par l'intermédiaire d'une couche contrainte (ou liner) qui enveloppe la grille [Ghani 03]. On citera par exemple la couche d'arrêt de la gravure contact (ou CESL pour Contact Etch Stop

Layer). Sa mise en place est assez simple car cette couche existe déjà dans le procédé de fabrication conventionnel du transistor MOSFET et il « suffit » donc d'utiliser un matériau contraint (tensile pour les NMOS, et compressif pour les PMOS)

Une autre solution est d'utiliser la technique de SMT (Stress Memorization Technique). Une couche de nitrure sacrificielle est déposée avant les recuits d'activation des source-drains et de la grille (dans le cas d'une grille en polysilicium) et vient encapsuler le transistor. On procède ensuite aux recuits et la recristallisation de la grille et des source-drains vient appliquer une contrainte tensile sur le canal, donc favorable au NMOS. Ces deux techniques sont utilisées conjointement pour la technologie 45nm de STMicroelectronics (Figure I-36-a) et [Bœuf 04] reporte un gain de 20% sur le compromis I_{on}/I_{off} (Figure I-36-b).

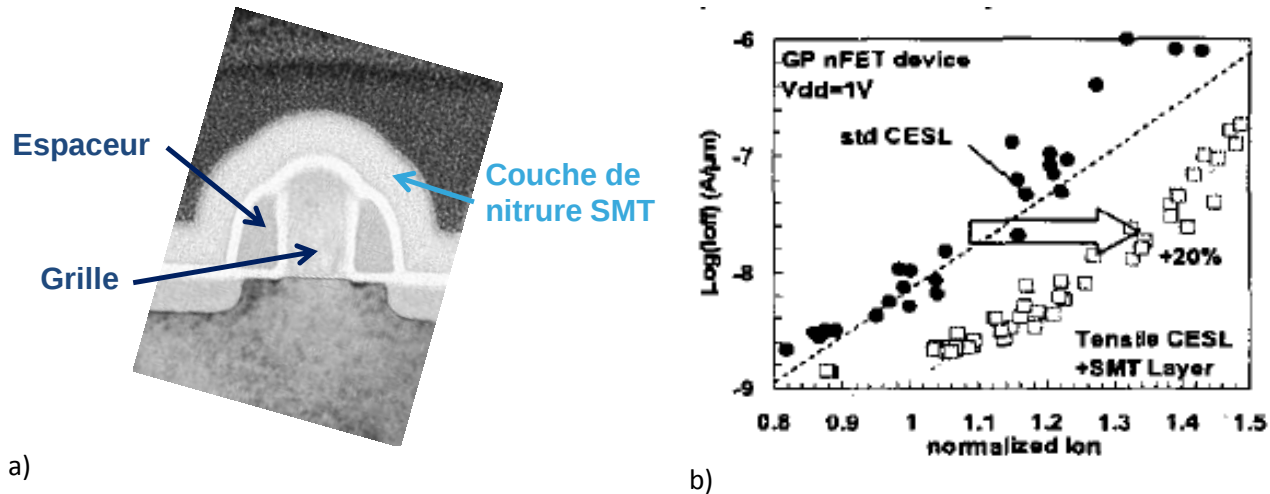
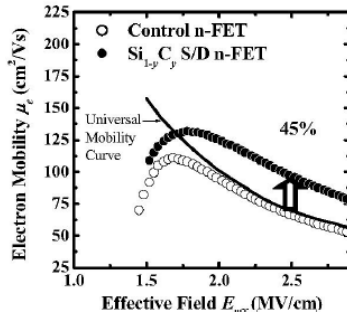
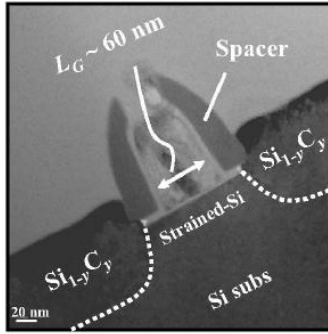
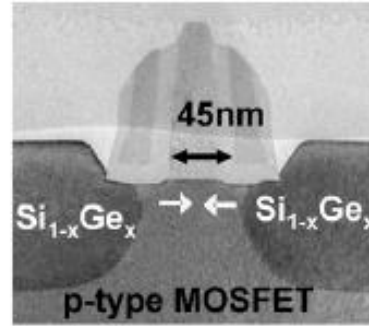


Figure I-36 : a) Photo prise au microscope électronique par transmission montrant l'intégration de la couche de SMT. b) Compromis I_{on}/I_{off} obtenue pour le nœud 45nm avec et sans intégration du CESL contraint et de la SMT [Bœuf 04].

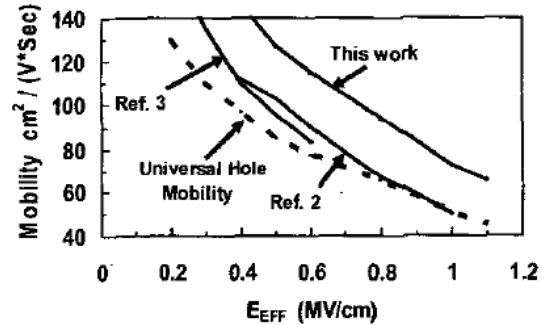
La maille cristalline du silicium du canal peut également être déformée par le remplacement des source-drains en silicium par des source-drains en silicium-germanium (SiGe) pour les PMOS [Ghani03, Mistry04] ou en silicium-carbone (SiC) pour les NMOS [Ang 04]. Le désaccord de maille entre le canal de silicium et les source-drains en SiGe (pour les PMOS) ou SiC (pour les NMOS) induit une contrainte compressive (pour les PMOS) ou tensile (pour les NMOS) sur le canal et améliore donc la mobilité des porteurs minoritaires. Les intégrations de ce type de source-drains et le gain qu'ils apportent sur la mobilité effective μ_{eff} sont illustrés sur la Figure I-37.



a) NMOS : Source-drains SiC [Ang 07]



[Thompson 05]

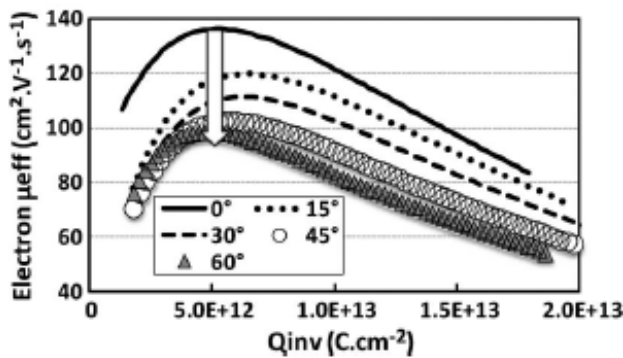


b) PMOS : Source-drains SiGe [Ghani 03]

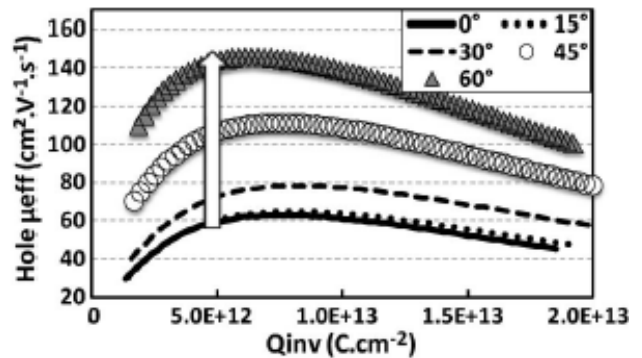
Figure I-37 : Photo prise au microscope électronique par transmission montrant l'intégration des source-drains SiC a) et SiGe b). Graphes μ_{eff}/E_{eff} reportés par la littérature démontrant le gain apporté pour les source-drain SiC et SiGe.

I.D.6.b) Changement d'orientation cristalline

Une autre solution pour améliorer le transport d'un transistor MOSFET est de jouer sur l'orientation cristalline du matériau. Ceci est démontré par caractérisation électrique par [Huguenin 11] (Figure I-38). On remarque que les orientations de canal favorables à la mobilité électrons sont défavorables à la mobilité des trous.



a) NMOS



b) PMOS

Figure I-38 : Variation de mobilité effective mesurée [Huguenin 11] sur substrat (110) pour différentes orientations de canal.

I.E. Vers de nouvelles architectures

Bien que nombreuses, les innovations technologiques apportées sur l'architecture conventionnelle sur substrat massif ne sont plus suffisantes pour permettre d'atteindre les spécifications de performance (I_{on}/I_{off} et I_{eff}/I_{off}), principalement à cause du contrôle électrostatique qui est de plus en plus difficile à obtenir. L'introduction de nouvelles architectures, et par conséquent la fin des plateformes CMOS basée sur l'architecture conventionnelle sur substrat massif, est prévue par l'ITRS [ITRS]. La Figure I-39 montre les prédictions de l'ITRS donnée par l'édition 2011 de sa feuille de route (roadmap en anglais) pour les dispositifs LOP (Low Operating Power, faible tension d'alimentation en français). On constate que la fin du BULK est prévue pour 2015, soit pour le nœud 20nm, alors que le FD SOI est prévu pour 2013 (également pour le 20nm). Cette architecture sera décrite dans le paragraphe I.E.1. Les dispositifs multi-grille (ou multi-gate) sont quant à eux attendus à partir de 2015 (soit le nœud 14nm). Cependant, l'ITRS ne précise pas si la technologie restera planeaire ou non et le nombre de grilles des dispositifs n'est pas non plus spécifié. Le double grille planeaire sera décrit au paragraphe I.E.2, les FinFETs et trigate seront présentés au paragraphe I.E.3. Pour prospecter encore plus loin, c'est-à-dire au-delà du nœud 10nm, nous décrivons succinctement l'architecture à base de nano-fils au paragraphe I.E.4. Enfin, l'ITRS prévoit à compter de 2018 l'introduction de matériaux à haute mobilité : canaux III-V pour les NMOS et Germanium pour les PMOS. Ces matériaux ont une permittivité supérieure à celle du silicium, les effets canaux courts sont donc plus difficiles à contrôler ; c'est la raison pour laquelle l'ITRS ne prévoit leur introduction qu'au travers de dispositifs multi-grille. Il ne faut cependant pas oublier les autres architectures possibles même avec ce type de matériau, à savoir les dispositifs dit « OI » pour On-Insulator (sur isolant) semblables à l'architecture FDSOI, mais également des architectures encore plus innovantes comme les IFQW (Implant Free Quantum Well). Ces derniers feront l'objet du paragraphe I.E.5.

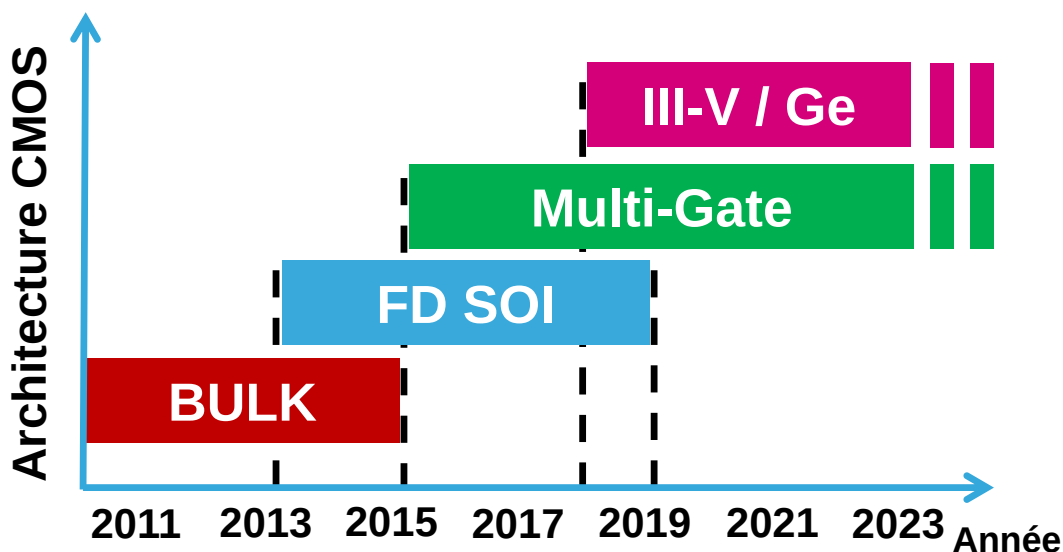


Figure I-39: Prédiction de l'utilisation des différentes architectures CMOS par l'ITRS en fonction de l'année sur sa roadmap 2011, pour les dispositifs LOP (Low Operating Power pour puissance de fonctionnement faible)

Ces prédictions sont confirmées, à quelques années près, par les annonces et différentes publications des industriels. STMicroelectronics a présenté sa plateforme CMOS FDSOI pour le nœud 28nm [Planes 12] et a annoncé le développement du nœud 20nm toujours basé sur l'architecture FDSOI. Enfin, Intel a présenté sa plateforme 22nm utilisant l'architecture Trigate [Auth 12], déjà commercialisée dans les applications hautes performances (micro-processeurs) et l'utilisation de l'architecture est prévue pour les applications mobile, ou SOC (System On Chip pour Système sur Puce).

I.E.1. FDSOI

I.E.1.a) Le substrat SOI

Le substrat SOI (Silicon On Insulator pour silicium sur isolant) se divise en trois couches. La première, la plus épaisse, est composée de silicium et sert de substrat mécanique. La seconde est communément appelée 'BOX' (Buried OXide pour oxyde enterré) et est généralement en SiO_2 . Enfin la dernière est une fine couche de silicium (on parle de film de silicium) qui se doit d'être mono cristalline car destinée à être la zone active, donc le canal, d'un transistor MOSFET. La Figure I-40 représente une vue schématique d'un substrat SOI avec les noms usuels des principales dimensions.

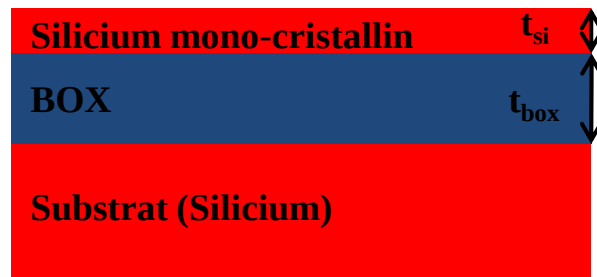
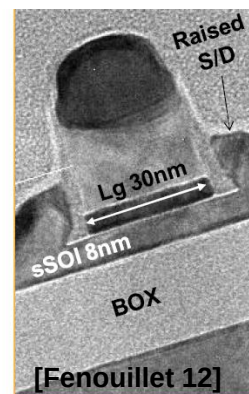
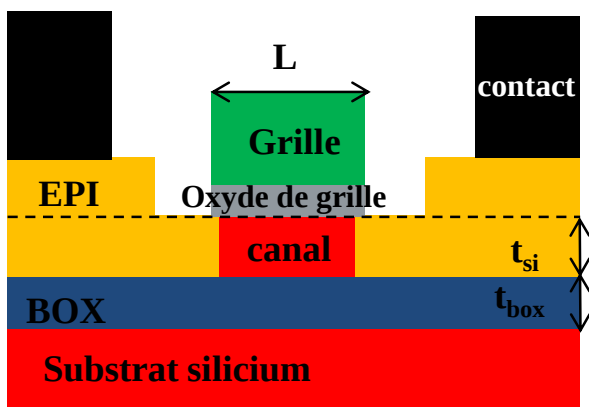


Figure I-40: Vue schématique en coupe d'un substrat SOI.

Les transistors sur substrats SOI ont été utilisés pour la première fois dans une application à visée commerciale dans les domaines militaires et spatiaux [Leray 90] [Hite 92] grâce à leur immunité aux radiations. Cependant, pour les applications mobiles grand public, l'immunité aux radiations n'est pas la préoccupation première et les transistors sur SOI sont utilisés seulement depuis le nœud 28nm [Planes 12] dans la configuration UTBB (Ultra Thin Body and Box pour film et oxyde enterré ultra-fin).

I.E.1.b) Le transistor FDSOI

Pour les applications CMOS, l'architecture FDSOI-UTBB est utilisée et elle est schématiquement représentée sur la Figure I-41. Lorsque le film de silicium est fin, il est indispensable de réaliser les sources-drains par épitaxie pour limiter la résistance série.



a) b)
Figure I-41: a) Vue schématique en coupe d'un transistor FDSOI UTBB. Les espaceurs entre grille et source/drain surélevés ne sont pas représentés pour la clarté du schéma. b) photo prise au microscope électronique à transmission d'un transistor FDSOI [Fenouillet 12].

Dans un tel dispositif, la couche de diélectrique enterré (le BOX) vient limiter physiquement la profondeur de déplétion dans le canal à l'épaisseur du film de silicium. Ceci a deux conséquences, qui constituent deux des principaux avantages de cette structure : les effets canaux courts sont mieux contrôlés et il est possible de travailler avec des canaux non dopés. Le deuxième point permet aux dispositifs FDSOI de s'affranchir de la principale source de variabilité des transistors BULK : le RDF (pour Random Doping fluctuation, fluctuation aléatoire due au dopage) et d'atteindre des records de robustesse à la variabilité ($A_{VT}=0.88\text{mV}$ reporté par [Weber 08]).

Dans une modélisation au premier ordre des effets canaux courts, on peut remplacer les termes T_{dep} (profondeur de déplétion) et X_j (profondeur de jonction) du modèle de l'architecture conventionnelle sur substrat massif (Eq. I-59 et I-60) par l'épaisseur de film de silicium t_{si} [Skotnicki 08] [MASTAR], ce qui donne:

$$SCE = \frac{\epsilon_{si}}{\epsilon_{ox}} \left(1 + \frac{t_{si}^2}{L^2}\right) \frac{t_{ox} t_{si}}{L^2} \phi_d \quad \text{Eq. I-63}$$

$$DIBL = \frac{\epsilon_{si}}{\epsilon_{ox}} \left(1 + \frac{t_{si}^2}{L^2}\right) \frac{t_{ox} t_{si}}{L^2} V_d \quad \text{Eq. I-64}$$

Sachant que pour les dernières technologies conventionnelles sur substrat massif, la profondeur de jonction est d'environ 10nm, et que la profondeur de déplétion est d'environ 20 nm, on voit rapidement grâce aux équations I-63 et I-64 que, pour un transistor FDSOI avec un film de silicium de 10nm (peu agressif) on gagne déjà un facteur 2 sur les paramètres SCE et DIBL.

Il ne faut cependant pas négliger l'effet du BOX. En effet, les lignes de champs se propagent dans cette couche isolante et il en résulte un couplage entre le drain et le canal qui vient aggraver le DIBL. Dans la littérature, on parle de « fringing fields » [Ernst 02] et [Ernst 07]. Pour limiter cet effet on cherche à utiliser des BOX les plus minces possibles, et on parle alors de dispositifs UTBB (Ultra-Thin Body and BOX pour film et oxyde enterré ultra-fin).

Néanmoins, pour des épaisseurs de BOX suffisamment fines (<25nm), le potentiel à l'interface BOX/substrat est suffisamment important pour créer une zone de déplétion à cette interface, et donc à augmenter l'épaisseur effective, d'un point de vue électrostatique, du BOX. Pour lutter contre cet effet, il a été proposé d'introduire un plan de masse, ou « ground plane » [Ernst 99] qui consiste en une couche dopée ($\sim 10^{18}\text{cm}^{-3}$) de type opposé aux sources-drains sous le BOX mince. Cette couche, réalisée par implantation ionique à travers le BOX réduit la déplétion, et limite donc l'épaisseur effective de BOX à son épaisseur physique. On parvient alors à limiter le couplage entre drain et canal par l'intermédiaire du BOX et donc à conserver un bon contrôle du canal par la grille (cf. II.B).

Enfin, l'utilisation de BOX mince permet également de pouvoir contrôler la tension de seuil du dispositif par polarisation de la face arrière, dans la littérature on parle de back-biasing [Liu 11] [Fenouillet 11]. L'efficacité du back-biasing sur la tension de seuil se mesure en mV/V de polarisation de la face arrière (V_b). On peut l'estimer facilement grâce à un simple diviseur capacitif. Les détails seront donnés dans le chapitre II. Pour des dimensions typiques du 20nm pour un transistor FDSOI ($t_{inv}=1.2\text{nm}$, $t_{si}=6\text{nm}$ et $t_{box}=15\text{nm}$) on obtient un décalage de V_t de $\sim 70\text{mV/V}$ de V_b , alors que pour un transistor BULK ($t_{inv}=1.2\text{nm}$, $N_{ch}=4 \times 10^{18}\text{cm}^{-3}$) on a seulement $\sim 25\text{mV}$.

L'efficacité du back-biasing pour les transistors FDSOI constitue un énorme avantage car il peut être utilisé de manière constante pour ajuster la tension de seuil, si le métal de grille ne permet pas d'atteindre le « bon V_t » (c'est-à-dire la tension de seuil permettant d'avoir la valeur de courant I_{off} spécifiée). Par contre, si le métal de grille est disponible le back-biasing peut être utilisé de manière dynamique et devient un plus au niveau système. Lorsque le système est au repos, on l'utilise pour augmenter la tension de seuil et donc réduire la fuite statique, on parle alors de RBB (Reverse Back Biasing : $V_b < 0$ pour un NMOS). De manière duale, lorsque le système nécessite un niveau de performance maximale, la tension de seuil est abaissée et le courant débité augmente ; on parle alors de FBB (Forward Back Biasing : $V_b > 0$ pour un NMOS).

I.E.2. Double grille planaire

L'architecture est très similaire au FDSOI : « il suffit » de remplacer le BOX par une grille, comme le montre le schéma de la Figure I-42.

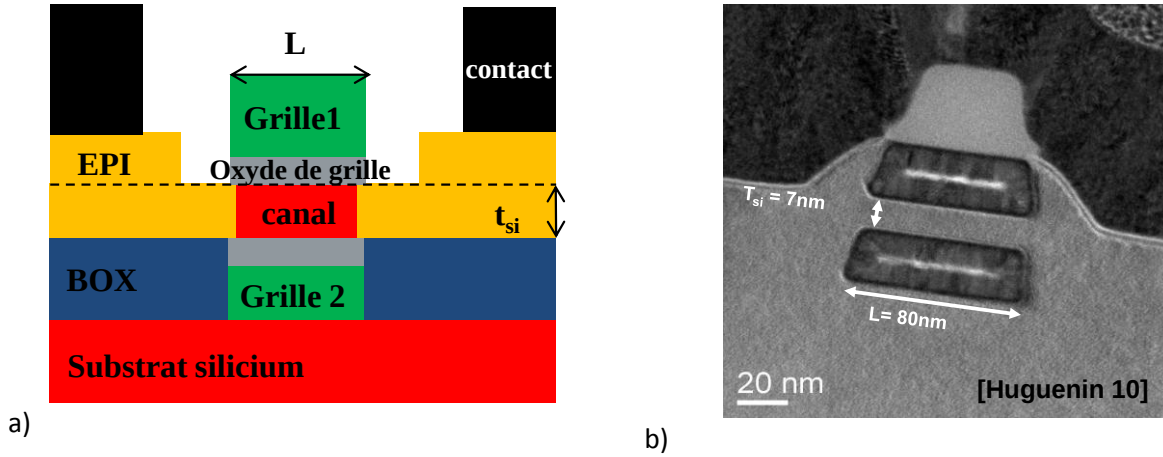


Figure I-42 : a) représentation schématique d'un transistor double grille planaire. b) photo prise au microscope électronique à transmission d'un transistor double grille [Huguenin 10].

En ajoutant cette seconde grille, le courant circule non plus sur une, mais sur deux interfaces, et la largeur électrique double pour un même encombrement. De plus, le contrôle électrostatique du canal est amélioré : par une modélisation au premier ordre, on peut considérer que chaque grille contrôle la moitié du canal, et donc remplacer les profondeurs de déplétion et de jonction du modèle de l'architecture conventionnelle sur substrat massif (Eq. I-59 et I-60) par la moitié de l'épaisseur du film de silicium, soit $t_{si}/2$. On obtient alors les équations suivantes pour les paramètres SCE et DIBL [Skotnicki 08] [MASTAR]:

$$SCE = \frac{\epsilon_{si}}{\epsilon_{ox}} \left(1 + \frac{\left(\frac{t_{si}}{2}\right)^2}{L^2} \right) \frac{t_{ox} \left(\frac{t_{si}}{2}\right)}{L^2} \phi_d \quad \text{Eq. I-65}$$

$$DIBL = \frac{\epsilon_{si}}{\epsilon_{ox}} \left(1 + \frac{\left(\frac{t_{si}}{2}\right)^2}{L^2} \right) \frac{t_{ox} \left(\frac{t_{si}}{2}\right)}{L^2} V_d \quad \text{Eq. I-66}$$

On constate que, en comparaison au FDSOI, DIBL et SCE sont améliorés d'un facteur 2 pour une même épaisseur de canal et une même longueur de grille. La structure présentée ci-dessus est le cas idéal d'un double grille planaire et sa faisabilité technologique a été démontrée par [Monfray 10]. D'un point de vue purement procédé de fabrication, le principal avantage du double grille planaire est l'excellent contrôle de l'épaisseur du film de silicium (car défini par épitaxie) alors que son principal inconvénient est l'alignement entre les deux grilles [Huguenin 10].

I.E.3. FinFET / Trigate

Le FinFET, en configuration double grille, est représenté sur la Figure I-43-a. Contrairement au double grille planaire (I.E.2), les canaux de conduction sont verticaux et ressemblent à un aileron (fin en anglais) de requin, d'où son nom. Pour empêcher la conduction sur le dessus du fin, et assurer une parfaite configuration double grille, un masque dur de nitrure est usuellement déposé [Kawazaki 09]. Le Trigate (Figure I-43-b) est identique au FinFET sauf que le masque dur de nitrure est remplacé par l'empilement d'oxyde de grille, ce qui permet bien au fin d'être contrôlé par trois grilles. Ces dispositifs peuvent être fabriqués sur substrat bulk conventionnel, ou sur SOI (cas représenté sur la Figure I-43-b). Essentiellement pour des raisons économiques, le substrat bulk est choisi

(~ 3-4 fois moins cher qu'un substrat SOI) : la première plateforme CMOS utilisant l'architecture Trigate est réalisée sur substrat massif [Auth 12]. Cependant, contrairement au Trigate sur SOI qui est isolé du substrat par la couche de diélectrique enterrée, le Trigate sur substrat massif (Figure I-43-c) n'est pas isolée naturellement du substrat. Il est alors nécessaire de réaliser une couche dopée dite anti-perçage, pour, comme son nom l'indique, éviter le phénomène de perçage entre source et drain par le substrat [Skotnicki 88-b]. Pour être efficace, cette couche doit être localisée près du fin. Elle est usuellement réalisée par implantation ionique et le recuit d'activation a pour effet de faire diffuser les dopants dans le fin. Il n'est donc pas possible de travailler avec un fin non dopé et ce type de dispositif sera affecté par la variabilité liée au dopage (RDF).

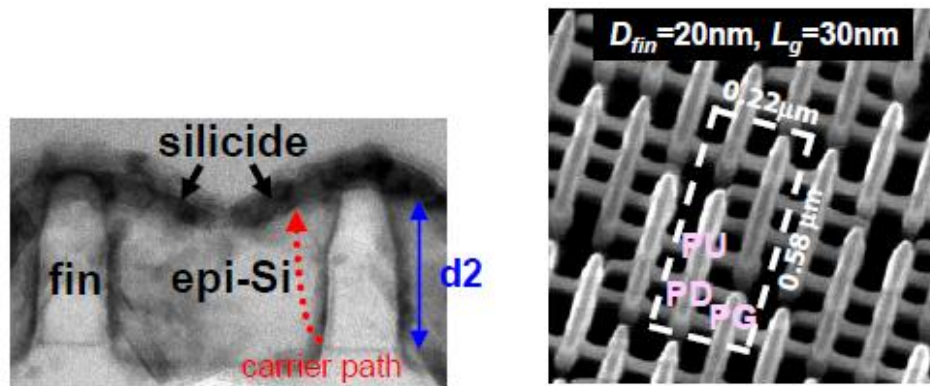
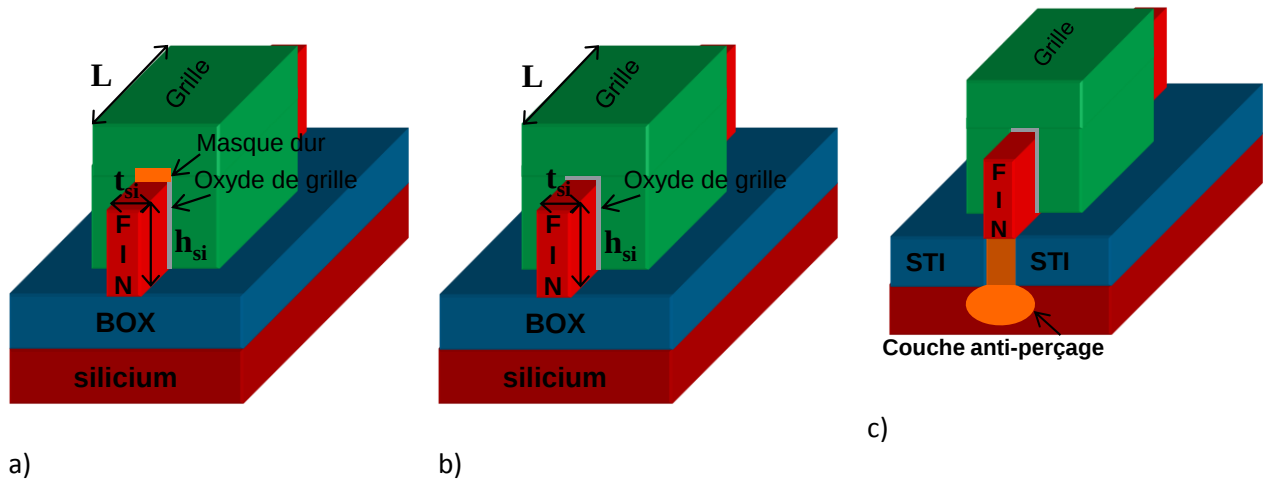


Figure I-43 Représentation schématisée d'un FinFET sur SOI (a), Trigate sur SOI (b) et d'un Trigate sur substrat conventionnel. (c). d) photos prises au microscope électronique par [Kawazaki 09].

Enfin, certaines applications nécessitent des niveaux de courant élevé. En technologie planaire, on répond à ce besoin en augmentant la largeur des zones actives, ce qui augmente, de manière continue, la largeur électrique des dispositifs, et donc le niveau de courant débité. Cet ajustement peut être aussi fin que le procédé de fabrication est précis, sachant qu'aujourd'hui les largeurs de zone active sont contrôlées à ± 1 nm. En ce qui concerne le FinFET ou le Trigate, comme la hauteur et la largeur de Fin sont des constantes technologiques et ont un effet direct sur les paramètres électriques du dispositif comme le DIBL par exemple, la seule solution pour augmenter la largeur électrique est d'ajouter un ou plusieurs fins, et passer en configuration dite multi-doigts (ou « multi-finger », représentée sur la Figure I-44). Dans ce cas, les sources et drains de chaque fin doivent être connectés entre eux. Ceci peut être réalisé de deux façons : avec l'épithaxie source drain qui vient fusionner tous les fins (Figure I-44-a) ou avec le contact (Figure I-44-b). Une nouvelle dimension caractéristique apparaît: le

« Finpitch » ou pas de répétition d'un fin, qui est égal à la somme de l'épaisseur d'un fin et d'un espacement entre deux fins (Figure I-44). Cette dimension est essentielle car elle va traduire la densité de fins et donc, la densité de courant, pour un encombrement donné.

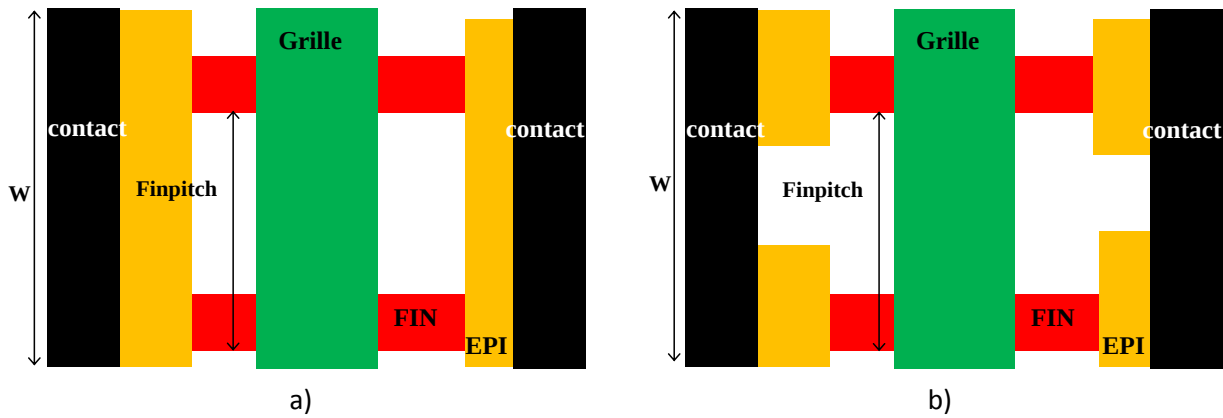


Figure I-44: Représentation schématique d'un FinFET en vue de dessus en configuration multi-doigts. a) cas où l'épitaxie source-drain vient fusionner les Fins. b) cas où c'est le contact qui vient connecter les Fins.

La configuration multi-doigts a pour effet de discrétiser la largeur électrique, et donc de discrétiser la densité de courant par unité de largeur d'active. Par conséquent, contrairement aux architectures planaires, la variation du I_{on} avec l'encombrement W (communément noté $I_{on}(W)$) sera discrète et non plus continue. Ceci est représenté sur la Figure I-45, où le nombre de Fins (N_{fin}) est estimé à partir de la largeur d'active W (i.e. l'encombrement), de la hauteur de Fin (h_{si}) et de son épaisseur (t_{si}), en prenant la partie entière de l'expression suivante :

$$N_{fin} = \text{ent} \left(\frac{W}{2h_{si} + t_{si}} \right) \quad \text{Eq. I-67}$$

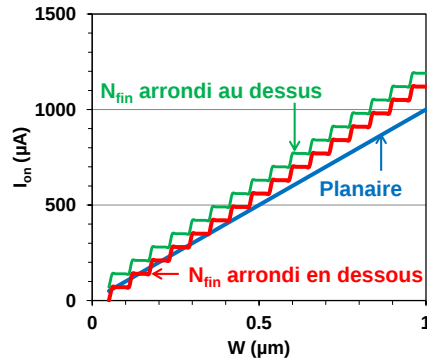


Figure I-45: $I_{on}(W)$ pour un dispositif planaire et pour deux dispositifs multi-doigts, en arrondissant le nombre de fin en dessus et en dessous. En considérant $I_{on}=1000\mu A/\mu m$, $Finpitch=60nm$, $h_{si}=30nm$ et $t_{si}=10nm$.

I.E.4. Nanofils

Pour les nœuds les plus avancés, en dessous de $7nm$ ($L_g < 10nm$), les effets canaux courts seront si importants qu'il sera certainement nécessaire de passer à une architecture permettant un contrôle électrostatique encore meilleur que le Trigate ou le FDSOI. Cette architecture pourrait être le transistor à grille enrobante (GAA pour Gate All Around dans la littérature) [Huguenin 10], ou le nanofil [Coquand 12] représenté sur la Figure I-46.

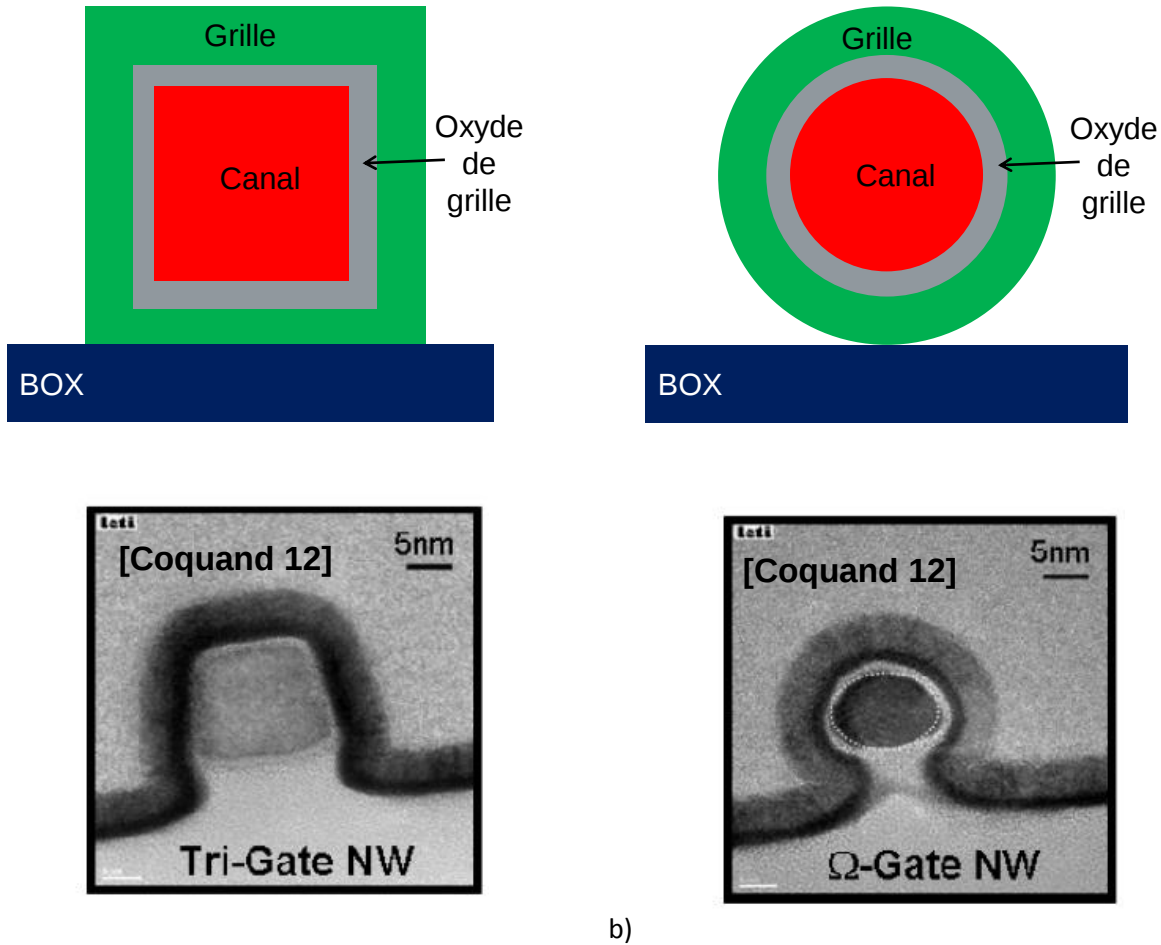


Figure I-46: Vue en coupe schématique et photo prise au microscope électronique à transmission d'un transistor à base de nanofils à section carré (a) et circulaire (b).

Cette architecture offre la meilleure immunité électrostatique car le canal est intégralement entouré par la grille est un canal cylindrique complètement “enrobé” par la grille. La forme carré/rectangulaire ou circulaire du canal dépend du procédé de fabrication, et plus particulièrement des recuits [Coquand 12]. Comme pour le Trigate et le FinFET, les nanofils sont couramment utilisés en mode “multi-doigts”, mais également en matrice pour pouvoir débiter des niveaux de courant important [Dupré 08].

I.E.5. Matériau III-V et germanium

Les matériaux dit haute mobilité (High Mobility Material) sont attrayant car ils présentent une mobilité bien supérieure à celle du silicium grâce à leur faible masse effective (Figure I-47).

	Si	Ge	GaAs	InAs	InP	InSb
Masse effective des électrons (la plus faible) m_e/m_0	0.191	0.08	0.067	0.023	0.073	0.012
Masse effective des trous lourds m_{hh}/m_0	0.53	0.35	0.62	0.6	0.85	0.45
Masse effective des trous légers m_{lh}/m_0	0.16	0.043	0.074	0.027	0.089	0.015
Mobilité des électrons μ_n ($\text{cm}^2.\text{V}^{-1}.\text{s}^{-1}$)	1350	3600	9200	30000	4500	77000
Mobilité des trous μ_p ($\text{cm}^2.\text{V}^{-1}.\text{s}^{-1}$)	480	1800	400	500	200	450
Ratio μ_n/μ_p	2.8	2	23	60	23	170
Gap E_g (eV) à 300K	1.12	0.66	1.43	0.36	1.27	0.17
Concentration de porteur intrinsèque n_i (cm^{-3})	10^{10}	2.10^{13}	3.10^6	8.10^{14}	3.10^7	2.10^{16}
Permittivité relative ϵ	11.7	16	12.9	15.15	12.5	16.8

Figure I-47: Propriété de quelques matériaux semi-conducteurs de type IV et III-V

Cependant, tous ces matériaux alternatifs ont également une permittivité ϵ supérieure à celle du silicium. Les effets canaux courts, et notamment le DIBL, étant directement proportionnels à cette permittivité, le contrôle électrostatique pour un même dispositif sera moins bon s'il est fabriqué sur substrat III-V que sur silicium. De plus, pour conserver le bénéfice apporté par une haute mobilité un bon contrôle électrostatique est indispensable, c'est la raison pour laquelle il est envisagé d'introduire les matériaux III-V au travers d'architecture type « III-V OI » (pour III-V On Insulator, en français III-V sur isolant) [Kim 12], Trigate [Radosavljevic 11] ou encore IFQW (pour « Implant Free Quantum Well ») [Hellings 10] et [Dewey 12]. Cependant les performances mesurées sont souvent moins bonnes qu'attendue, essentiellement du fait de la mauvaise qualité de l'interface oxyde de grille-semi-conducteur qui entraîne des valeurs importantes de D_{it} (Density of Interface State pour densité d'état d'interface), alors que ce problème technologique est inexistant ou presque en technologie silicium grâce à l'excellente interface Si/SiO₂. Enfin, pour obtenir une technologie CMOS utilisable et compétitive, NMOS et PMOS sont nécessaires, et ils doivent avoir des performances (i.e. niveau de courant débité I_{on} , I_{eff}) comparables. Un premier indicateur est la valeur du ratio μ_n/μ_p qui serait idéalement 1 (2.8 pour le silicium). La Figure I-47, montre que le germanium a un ratio μ_n/μ_p comparable à celui du silicium, tout en ayant une mobilité 2 à 3 fois supérieure, ce qui en fait un excellent candidat. Cependant, il a été démontré, entre autre par [Batail 09], que le germanium est très difficile à intégrer, notamment pour les transistors courts. De plus, la faible largeur de sa bande interdite (gap) génère davantage de fuite qu'une technologie silicium. Enfin, les matériaux III-V (4 dernières colonnes de la Figure I-47) présentent des mobilités d'électron impressionnantes, jusqu'à ~30 fois supérieure à celle du silicium. Cependant, la mobilité des trous dans ces matériaux est assez faible, et le ratio μ_n/μ_p est très grand, rendant ce type de matériau inutilisable pour les applications CMOS. La seule solution pour introduire ces matériaux à haute mobilité semble être la co-intégration, de transistors à canaux III-V pour les NMOS, et canaux germanium pour les PMOS. La faisabilité technologique été démontrée par [Yokoyama 11] : la mobilité mesurée est 3.5x et 2.3x supérieur au silicium respectivement pour N et PMOS mais pour des transistors très longs (Longueur de grille supérieure à 20 μm). La faisabilité technologique reste donc à démontrer pour des transistors courts et l'amélioration de transport à quantifier. Enfin, une récente étude [Yuan 12] démontre qu'il est possible d'intégrer N et PMOS sur un même substrat In_xGa_{1-x}Sb et d'obtenir des niveaux de courant comparables pour les deux types de transistor. Néanmoins, cette étude ne présente que des dispositifs de longueur de grille égale à 50 μm , la faisabilité et les performances des transistors courts restent à explorer.

I.F. Conclusion du chapitre

Au cours de ce premier chapitre, nous avons commencé par succinctement décrire le fonctionnement idéal du transistor MOSFET et son architecture conventionnelle sur substrat massif. Afin de tendre vers le fonctionnement

réaliste du transistor, nous avons ensuite exposé les différents effets physiques à prendre en compte pour corriger ce modèle simplifié. Nous avons alors défini puis démontré les principaux paramètres et équations régissant son fonctionnement dans chaque régime. Ensuite, nous avons explicité les règles issues de la loi de Moore qui guident la course à la miniaturisation des technologies CMOS depuis ces dernières décennies : diviser l'encombrement surfacique des circuits par deux entre chaque nœud technologique, tout en améliorant la performance des dispositifs et en réduisant leur consommation. Les conséquences de cette réduction des dimensions de nœud en nœud sur le fonctionnement du transistor sont ensuite décrites. Les effets parasites liés à la réduction deviennent un frein à la performance pour les nœuds technologiques les plus avancées et nous avons par la suite explicité les solutions technologiques apportées à l'architecture conventionnelle sur substrat massif afin de poursuivre la réduction des dimensions avec un gain en performance. Enfin, nous avons vu que les améliorations technologiques apportées n'étaient plus suffisantes pour poursuivre la réduction des dimensions et que, comme prévu par l'ITRS avec MASTAR [MASTAR], l'introduction de nouvelles architecture de transistor est nécessaire à compter du nœud technologique 20nm. La description de ces nouvelles architectures de transistor MOSFET a fait l'objet de la dernière partie de ce premier chapitre.

L'objectif de ce manuscrit de thèse est de d'abord fournir une modélisation précise et prédictive des différentes architectures de transistor dans le but de prédire leur performance statique (chapitre II), puis dynamique (chapitre III). Ces différents modèles seront utilisés au travers d'une méthode numérique et itérative d'évaluation du délai de propagation d'une chaîne d'inverseur (chapitre IV). Enfin, ces modèles seront adaptés et intégrés dans un modèle compact, que nous nommerons MASTAR_VA, afin d'utiliser un simulateur de circuit conventionnel [ELDO] pour prévoir les performances logiques des différentes technologies CMOS concurrentes au nœud 20nm (chapitre V) puis au nœud 16nm en y ajoutant l'estimation des performances SRAM (chapitre VI).

- CHAPITRE II -

MODELISATION ANALYTIQUE DES CARACTERISTIQUES STATIQUES DES DIFFERENTES ARCHITECTURES CMOS.

Dans ce chapitre, nous allons décrire la méthodologie utilisée pour prédire les performances statiques des principales architectures CMOS à partir de leurs paramètres technologiques. Par performance statique, on entend généralement compromis I_{on}/I_{off} et I_{eff}/I_{off} , qui traduisent respectivement le niveau de courant débité à la tension d'alimentation nominale et la valeur du courant effectif (plus représentatif de la vitesse d'un circuit [Na 02]) pour un niveau de fuite donné.

A partir de la valeur du courant drain-source au seuil ($I_{th} = 10^{-7} \frac{W}{L}$) la fuite I_{off} sera estimée grâce au modèle classique de courant de diffusion et à une bonne prédiction des paramètres électrostatiques (i.e. V_t , SCE, DIBL et SS) de chaque architecture. Pour I_{on} et I_{eff} , le modèle classique de courant de dérivation sera utilisé, combiné avec les expressions de mobilité universelle [Takagi 94] auxquelles on applique la dégradation due à la réduction de longueur de grille [Bidal 09], ainsi que l'amélioration due aux contraintes [Payet 08]. Enfin, une estimation précise de la tension de seuil et du DIBL sera également nécessaire.

La modélisation de l'électrostatique des architectures conventionnelle sur substrat massif, FDSOI et Double grille faiblement dopé sera décrite respectivement dans les parties II.A, II.B et II.C. Dans ces trois parties, nous procéderons de la même manière : nous commencerons par expliciter et justifier le critère de seuil (ou d'inversion), puis nous proposerons un modèle de tension de seuil pour un canal long et enfin nous étudierons l'effet de la réduction de la longueur de grille. Pour assurer la compatibilité avec les outils de CAO conventionnels, un modèle de courant drain-source parfaitement continu, ainsi que ses dérivées, sera proposé dans la partie II.D.

II.A. Tension de seuil et électrostatique de l'architecture conventionnelle sur substrat massif

Dans cette partie, nous proposons une modélisation de l'électrostatique d'un transistor conventionnel sur substrat massif à partir de ses paramètres technologiques, à savoir son épaisseur d'oxyde de grille (t_{ox}), sa longueur de grille (L), son niveau de dopage canal (N_{ch}) et sa profondeur de jonction. Ces paramètres sont représentés sur la Figure II-1.

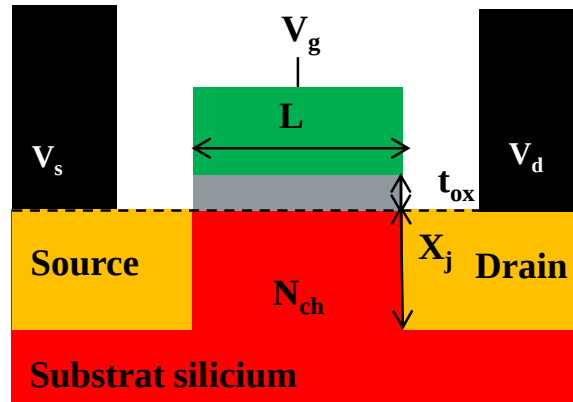


Figure II-1: Schéma de l'architecture conventionnelle sur substrat massif, avec la définition des principaux paramètres technologiques et polarisations.

II.A.1. Critère d'inversion

Usuellement, dans les transistors conventionnels sur substrat massif, la condition de seuil se traduit par une courbure de bande (Figure II-2-a), et donc un potentiel de surface donné par l'expression suivante:

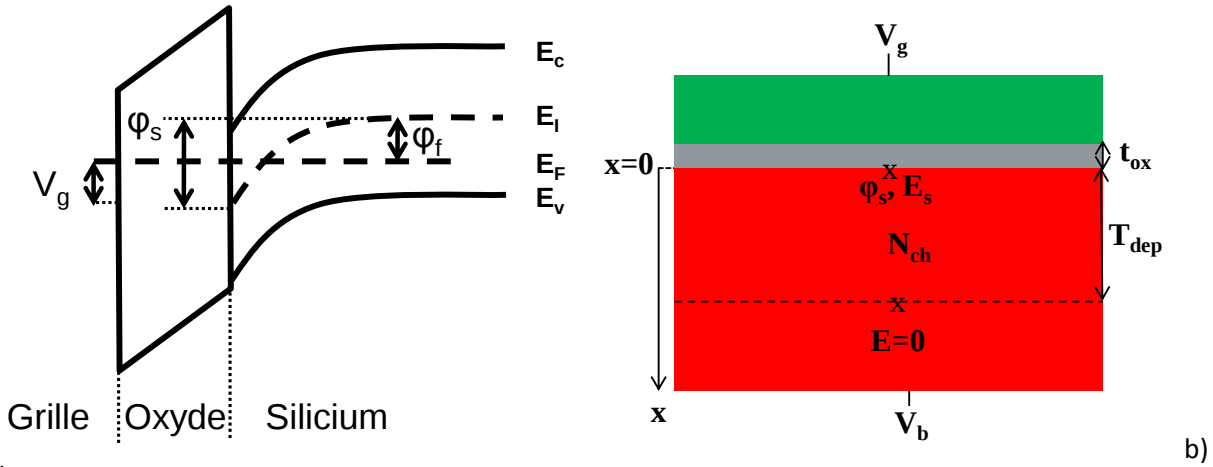
$$\varphi_s = 2\varphi_f \quad \text{Eq. II-1}$$

Comme il est indiqué dans le paragraphe I.A.4, cette égalité se justifie en considérant que le régime d'inversion débute lorsque la concentration en porteur majoritaire est égale à celle des minoritaires [Mathieu 04]. Dans le cas d'un NMOS, c'est la densité d'électrons n_e qui sera égale au dopage canal N_{ch} :

$$N_{ch} = n_i e^{\frac{q\phi_f}{kT}} = n_{e-} = n_i e^{\frac{q(\phi_f - \phi_s)}{kT}}$$

D'où : $\phi_f = \phi_f - \phi_s \quad \rightarrow \quad \phi_s = 2 \cdot \phi_f$ Eq. II-2

Avec ϕ_f potentiel de Fermi, n_i concentration de porteur intrinsèque et kT/q la tension thermodynamique.



a) Figure II-2: (a) Diagramme de bandes d'énergie au seuil avec la représentation des différents potentiels et polarisation. (b) Schéma et dimensions caractéristiques de la capacité MOS.

II.A.2. Tension de seuil canal long

Pour obtenir une expression de la tension de seuil pour un transistor conventionnel sur substrat massif canal long, on commence par écrire la loi de Gauss entre l'interface oxyde-silicium et le bas de la zone de déplétion ($x=T_{dep}$ sur la Figure II-2-b), en négligeant les sources-drains (donc dans une capacité MOS classique) :

$$\epsilon_{si} E_s = C_{ox} (V_g - V_{fb} - \phi_s) = Q_{dep} + Q_{inv} + Q_{ss}$$
Eq. II-3

Avec E_s champ en surface, V_{fb} tension de bandes plates, Q_{dep} charge de déplétion, Q_{inv} charge d'inversion et Q_{ss} charge due aux états d'interface (Surface State, lié au D_{it} qui traduit la qualité de l'interface oxyde de grille/silicium). Lorsqu'on se situe juste avant le seuil, on peut considérer que la charge d'inversion est négligeable devant la charge de déplétion et on choisit de se placer dans le cas d'une interface parfaite, donc avec $Q_{ss}=0$. En se plaçant au seuil (i.e. $\phi_s = 2\phi_f$), l'équation II-4 donne l'expression de la tension de seuil pour un canal long (V_{tlong}) :

$$V_{tlong} = V_{fb} + \frac{Q_{dep}}{C_{ox}} + 2\phi_f$$
Eq. II-4

En remplaçant Q_{dep} par son expression classique [Taur 98] au seuil, on a :

$$V_{tlong} = V_{fb} + \frac{\sqrt{2qN_{ch}\epsilon_{si}(2\phi_f - V_b)}}{C_{ox}} + 2\phi_f$$
Eq. II-5

Qu'on écrit usuellement :

$$V_{tlong} = V_{fb} + \frac{qN_{ch}T_{dep}}{C_{ox}} + 2\phi_f \quad \text{Eq. II-6}$$

Avec $T_{dep} = \sqrt{2\epsilon_{si} \cdot \frac{2\phi_f - V_b}{qN_{ch}}}$ Eq. II-7

Avec V_b polarisation du substrat et T_{dep} profondeur de la zone de déplétion.

II.A.3. Effets canaux courts (SCE) et DIBL

Pour prendre en compte la dépendance avec la longueur de grille, nous avons choisi d'utiliser la Transformation Tension Dopage, ou VDT (Voltage Doping Transformation) [Skotnicki 88-a].

II.A.3.a) La Transformation Tension-Dopage (VDT)

La VDT consiste à transformer l'influence du champ entre la source et le drain sur la barrière de potentiel par une réduction de dopage moyen dans le canal. De cette manière, l'équation de Poisson 2D peut être réduite à une dimension.

La première étape est de définir un repère curviligne ayant pour ordonnée les lignes de courant (ou de champ) entre source et drain et pour abscisse les courbes perpendiculaire aux lignes de champ (Figure II-3-a). Pour simplifier les calculs ultérieurs, on choisit de placer l'origine du repère curviligne sur la cathode virtuelle, c'est-à-dire sur l'axe x où le potentiel est minimum. Il a été démontré dans [Skotnicki 88-a] que l'équation de Poisson 2D dans le repère curviligne (x,y) peut s'écrire comme dans le repère cartésien classique (α, β), à savoir :

$$\frac{d^2\phi}{dx^2} + \frac{d^2\phi}{dy^2} = q \frac{N_{ch}}{\epsilon_{si}} \quad \text{Eq. II-8}$$

En remarquant que $\frac{d^2\phi}{dy^2}$ est une fonction de x (et ne dépend donc pas de y) sur la cathode virtuelle (donc sur la courbe (x,0)), on peut réécrire l'équation de Poisson :

$$\frac{d^2\phi}{dx^2} = q \frac{N_{ch}}{\epsilon_{si}} - \frac{d^2\phi}{dy^2} = q \frac{N_{ch}(x)^*}{\epsilon_{si}} \quad \text{Eq. II-9}$$

Où $N_{ch}(x)^*$ est une fonction de x uniquement et définit le dopage effectif qui va traduire l'abaissement de barrière du champ électrique entre source et drain. Pour déterminer l'expression du dopage effectif, on considère que le potentiel est une parabole entre source et drain, dans le repère curviligne (x,y) comme représenté sur la Figure II-3-b. L'expression du potentiel est alors :

$$\phi(y) = 2 \frac{V_{ds}(x)^*}{L(x)^2} y^2 + by + c \quad \text{Eq. II-10}$$

Ceci constitue la première approximation de cette méthode qui a l'avantage de fournir des expressions finales très simples qui permettent de « sentir » l'impact des paramètres technologiques sur le comportement électrique du transistor. L'inconvénient de cette approximation est qu'elle est de moins en moins vraie quand la longueur de grille diminue, le potentiel ayant dans ce cas plutôt une forme de cuvette à bords abrupts. Mathématiquement, pour des longueurs de grille agressives, le potentiel est mieux reproduit par une fonction à base d'exponentielles [Toyabe 79].

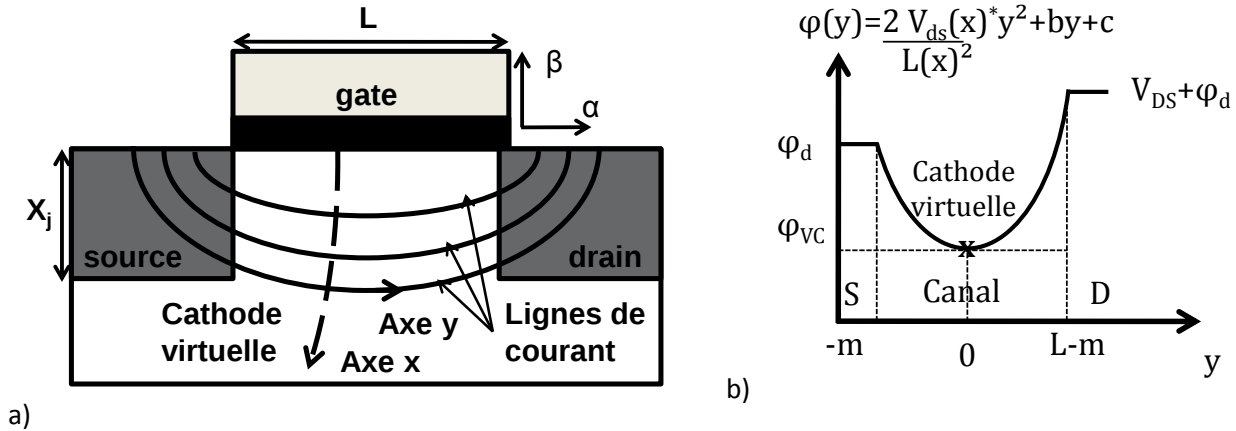


Figure II-3: (a) représentation des deux repères de la VDT. (b) représentation des conditions aux limites de l'approximation de potentiel parabolique pour le calcul du dopage effectif.

En remplaçant ϕ par son expression dans la définition du dopage effectif, on a :

$$N_{ch}(x)^* = N_{ch} - \frac{\epsilon_{si}}{q} \frac{d^2 \phi}{dy^2} = N_{ch} - 2 \frac{\epsilon_{si}}{q L(x)^2} V_{DS}(x)^* \quad \text{Eq. II-11}$$

Pour déterminer $V_{DS}^*(x)$, on utilise les conditions aux limites exposées sur la Figure II-3, soit :

$$\begin{aligned} \text{A la source :} & \quad \phi\left(m - \frac{L(x)}{2}\right) = \phi_d \\ \text{Au drain :} & \quad \phi\left(m + \frac{L(x)}{2}\right) = V_{DS} + \phi_d \\ \text{A la cathode virtuelle :} & \quad \phi(m) = \phi_{VC}(x) \\ \text{Par définition :} & \quad \left(\frac{d\phi}{dy}\right)(y = m) = 0 \end{aligned} \quad \text{Eq. II-12}$$

Avec $L(x)$, longueur de la ligne de courant passant par x , $\phi_{VC}(x)$ potentiel sur la cathode virtuelle en x et m la position de la cathode virtuelle sur l'axe y . Après résolution du système d'équations, on a :

$$V_{DS}^*(x) = V_{DS} + 2(\phi_d - \phi_{VC}(x)) + 2\sqrt{(\phi_d - \phi_{VC}(x))(V_{DS} + \phi_d - \phi_{VC}(x))} \quad \text{Eq. II-13}$$

Par conséquent le dopage effectif s'exprime:

$$N_{ch}(x)^* = N_{ch} - 2 \frac{\epsilon_{si}}{q L(x)^2} (V_{DS} + 2(\phi_d - \phi_{VC}(x)) + 2\sqrt{(\phi_d - \phi_{VC}(x))(V_{DS} + \phi_d - \phi_{VC}(x))}) \quad \text{Eq. II-14}$$

A partir de là, différentes approches peuvent être utilisées pour aboutir à l'expression de la tension de seuil tenant compte des effets canaux courts. Nous allons en présenter deux : la première est utilisée dans le logiciel MASTAR [MASTAR] propose des expressions très simples des paramètres SCE et DIBL, obtenues grâce à de nombreuses approximations (entraînant des imprécisions dans certains cas). La seconde, développée au cours de cette thèse propose des expressions plus complexes, mais obtenues avec le moins d'approximations possibles.

II.A.3.b) Méthode "MASTAR"

Pour simplifier l'expression du dopage effectif afin de supprimer sa dépendance en x , il est proposé dans [Gautier 03] de considérer la longueur des lignes de courant $L(x)$ constante et égale à la longueur électrique du transistor

et de remplacer le potentiel sur la cathode virtuelle ($\varphi_{vc}(x)$) dans l'expression de $V_{ds}(x)^*$ par sa valeur moyenne, estimée à $3/4 \varphi_d$. L'expression du dopage effectif est donc réduite à :

$$N_{ch}^* = N_{ch} - 2 \frac{\varepsilon_{si}}{qL(x)^2} \left(V_{DS} + 2 \left(\frac{1}{4} \varphi_d \right) + 2 \sqrt{\left(\frac{1}{4} \varphi_d \right) \left(V_{DS} + \frac{1}{4} \varphi_d \right)} \right) \quad \text{Eq. II-15}$$

Ensuite, pour obtenir une expression de l'effet SCE seul, on fixe $V_{ds}=0$:

$$N_{ch}^* = N_{ch} - 2 \frac{\varepsilon_{si}}{qL(x)^2} \varphi_d \quad \text{Eq. II-16}$$

En remplaçant le dopage canal N_{ch} par le dopage effectif N_{ch}^* dans l'expression de la tension de seuil canal long (Eq. II-6), on a :

$$V_{th} = V_{fb} + 2\varphi_f + \frac{\sqrt{2qN_{ch}^* \varepsilon_{si} (2\varphi_f - V_b)}}{C_{ox}} \quad \text{Eq. II-17}$$

$$V_{th} = V_{fb} + 2\varphi_f + \frac{\sqrt{2qN_{ch} \varepsilon_{si} (2\varphi_f - V_b)}}{C_{ox}} \sqrt{\left(1 - 2 \frac{\varepsilon_{si}}{qN_{ch} L^2} \varphi_d \right)} \quad \text{Eq. II-18}$$

Par un développement de Taylor de la seconde racine, on parvient à isoler l'expression de la tension de seuil canal long a :

$$V_{th} = V_{fb} + 2\varphi_f + \frac{\sqrt{2qN_{ch} \varepsilon_{si} (2\varphi_f - V_b)}}{C_{ox}} - \frac{\varepsilon_{si}}{\varepsilon_{ox}} \frac{t_{ox}}{L^2} \sqrt{2\varepsilon_{si} \frac{2\varphi_f - V_b}{qN_{ch}}} \varphi_d \quad \text{Eq. II-19}$$

Par définition du paramètre SCE :

$$V_{th} = V_{thlong} - SCE \quad \text{Eq. II-20}$$

$$SCE = \frac{\varepsilon_{si}}{\varepsilon_{ox}} \frac{t_{ox} T_{dep}}{L^2} \varphi_d \quad \text{Eq. II-21}$$

Pour le calcul du DIBL, on procède de la même manière, mais en considérant $V_{ds} \neq 0$. L'expression du dopage effectif est alors donné par l'équation II-14. Pour réduire son expression et pouvoir isoler la contribution du DIBL, il est proposé dans [Gautier 03] de considérer les deux cas extrêmes :

- $V_{DS} \gg \varphi_d$: $V_{DS}(x)^* \approx V_{DS} + \varphi_d$
- $V_{DS} \ll \varphi_d$: $V_{DS}(x)^* \approx \varphi_d \approx V_{DS} + \varphi_d$

Ce qui permet d'avoir une même expression de V_{ds}^* dans tous les cas, donnant l'expression du dopage effectif suivante :

$$N_{ch}^* = N_{ch} - 2 \frac{\varepsilon_{si}}{qL^2} (V_{DS} + \varphi_d) \quad \text{Eq. II-22}$$

Enfin, en remplaçant le dopage canal N_{ch} par le dopage effectif N_{ch}^* dans l'expression de la tension de seuil canal long (Eq. II-6), on a :

$$V_{th} = V_{fb} + 2\varphi_f + \frac{qN_{ch} T_{dep}}{C_{ox}} - \frac{\varepsilon_{si}}{\varepsilon_{ox}} \frac{t_{ox} T_{dep}}{L^2} \varphi_d - \frac{\varepsilon_{si}}{\varepsilon_{ox}} \frac{t_{ox} T_{dep}}{L^2} V_{ds} \quad \text{Eq. II-23}$$

$$V_{th} = V_{thlong} - SCE - DIBL \quad \text{Eq. II-24}$$

Par identification:

$$DIBL = \frac{\epsilon_{si}}{\epsilon_{ox}} \frac{t_{ox} T_{dep}}{L^2} V_{ds} \quad \text{Eq. II-25}$$

Cette approche a l'avantage de proposer des expressions très simples qui donnent une bonne estimation de l'électrostatique d'un transistor conventionnel sur substrat massif. Cependant, les expressions ci-dessus ne tiennent pas compte de l'effet de la profondeur des jonctions source-drain (X_j) sur les paramètres SCE et DIBL. Or il est bien connu que la dimension X_j a un impact important sur le contrôle des effets canaux courts. Pour prendre en compte cet effet, [Gautier 03] remarque que la longueur électrique moyenne est diminuée si la profondeur de jonction augmente. Par observation géométrique, il évalue cette dépendance comme proportionnelle à $(1+X_j^2/L^2)$, d'où les nouvelles expressions de DIBL et SCE :

$$SCE = 0.64 \frac{\epsilon_{si}}{\epsilon_{ox}} \frac{t_{ox} T_{dep}}{L^2} \left(1 + \frac{X_j^2}{L^2}\right) \varphi_d \quad \text{Eq. II-26}$$

$$DIBL = 0.8 \frac{\epsilon_{si}}{\epsilon_{ox}} \frac{t_{ox} T_{dep}}{L^2} \left(1 + \frac{X_j^2}{L^2}\right) V_{ds} \quad \text{Eq. II-27}$$

Où 0.64 et 0.8 sont des paramètres d'ajustement donnés par [Gautier 03] et déterminer d'après simulation numériques. Pour vérifier si cette méthode est bien prédictive pour l'évaluation de la tension de seuil et des effets canaux courts en fonction des principaux paramètres technologiques, nous avons effectué des simulations numérique 2D avec l'outil Synopsys [Synopsys] en faisant varier le dopage canal N_{ch} , l'épaisseur d'oxyde de grille équivalente (EOT) et la profondeur de jonction (X_j).

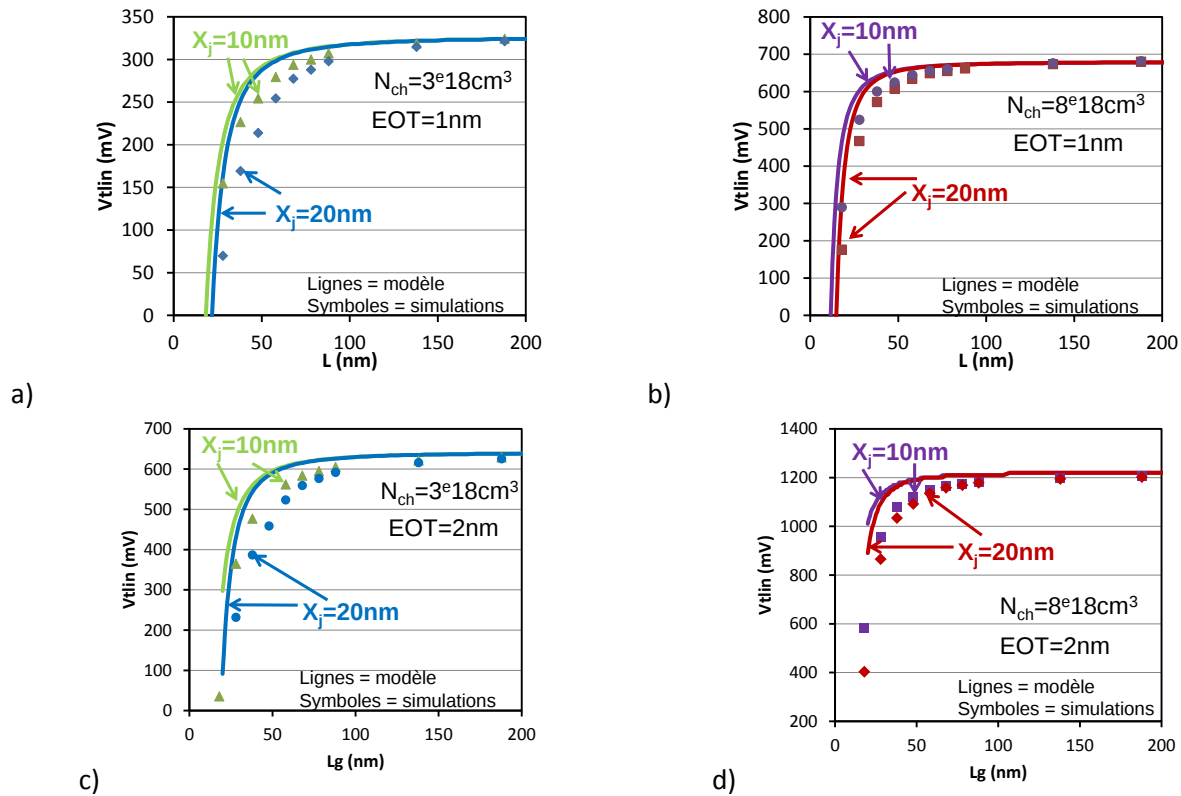
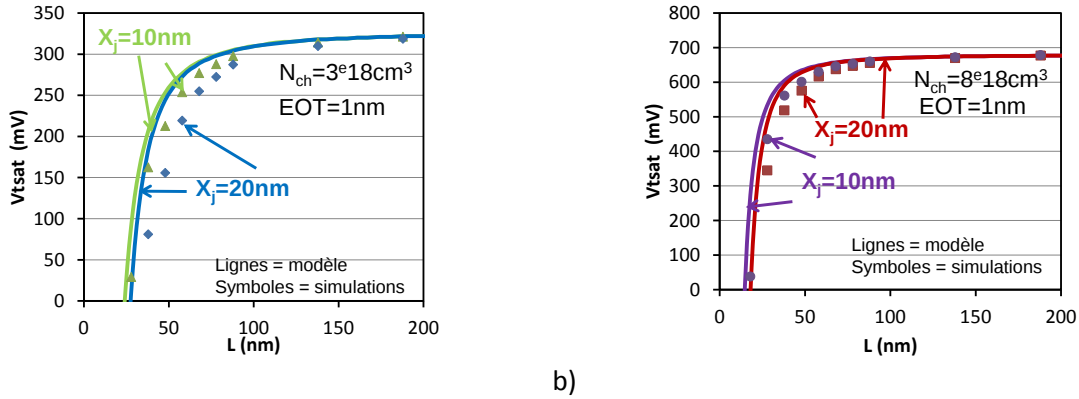


Figure II-4: Tracé de la tension de seuil en régime linéaire V_{tlin} ($V_d=0.1V$) en fonction de la longueur de grille L_g pour différentes profondeurs de jonction et d'EOT obtenue avec le modèle MASTAR et par simulation numérique 2D. a) EOT= 1nm, $N_{ch}=3 \times 10^{18} \text{cm}^{-3}$, a) EOT= 1nm, $N_{ch}=8 \times 10^{18} \text{cm}^{-3}$, a) EOT= 2nm, $N_{ch}=3 \times 10^{18} \text{cm}^{-3}$, a) EOT= 2nm, $N_{ch}=8 \times 10^{18} \text{cm}^{-3}$.

La Figure II-4 représente la tension de seuil en régime linéaire (V_{tlin} extrait à $V_{ds}=0.1V$) en fonction de la longueur de grille pour différentes valeurs de dopage canal, d'EOT et de profondeur de jonction. On constate que la méthode de modélisation précédente ne reproduit pas très bien le coude, c'est-à-dire les valeurs de tensions de seuil où les effets canaux courts commencent à compter. De plus, l'écart entre le modèle et la simulation grandit quand le dopage canal diminue : cela s'explique par les développements de Taylor effectués pour passer de l'équation II-18 à l'équation II-19 qui deviennent de moins en moins idoines quand la valeur de N_{ch} diminue. Enfin, les variations d'effet canal court dues aux variations de profondeur de jonction sont sous-estimées par cette méthode de modélisation.



a) b)
Figure II-5 : Tracé de la tension de seuil en régime saturé V_{tsat} ($V_d=V_{dd}=1V$) en fonction de la longueur de grille L_g pour différentes profondeur de jonction et d'EOT obtenue avec le modèle MASTAR et par simulation numérique 2D. a) EOT=1nm, $N_{ch}=3 \times 10^{18} cm^{-3}$, b) EOT=1nm, $N_{ch}=8 \times 10^{18} cm^{-3}$

La Figure II-5 représente l'évolution de la tension de seuil en régime saturée (V_{tsat}), incluant donc l'effet du DIBL avec la longueur de grille pour deux valeurs de dopage canal et deux valeurs de profondeur de jonction. On voit bien que V_{tsat} est assez bien modélisé pour les jonctions fines. Par contre, la dégradation de DIBL qui a lieu lorsque la jonction est plus profonde est sous-estimée par le modèle.

II.A.3.c) Approche exacte

Afin d'être plus prédictif sur l'électrostatique du transistor conventionnel sur substrat massif, nous proposons dans cette sous-partie une nouvelle manière de résoudre l'équation de Poisson 2D avec la VDT, mais cette fois en effectuant aucun développement de Taylor et le moins d'approximations possible pour aboutir à une expression finale complètement analytique de la tension de seuil.

Dans l'équation de la tension de seuil canal long (Eq II-4), il est clair que le dopage effectif n'a d'impact que sur le terme de charge de déplétion Q_{dep} . Le problème du calcul de la tension de seuil, comprenant les effets canaux courts, est donc réduit au calcul d'une charge de déplétion effective que nous noterons dans la suite Q_{dep}^* . On commence donc par écrire l'équation de Poisson 2D sur la cathode virtuelle ($\phi(x) = \phi_{vc}(x)$) et on remplace le dopage canal par le dopage effectif $N_{ch}^*(x)$ donné par l'équation II-13, 14 :

$$\frac{d^2 \phi}{dx^2} = q \frac{N_{ch}(x)^*}{\epsilon_{si}} = \frac{q}{\epsilon_{si}} \left(N_{ch} - 2 \frac{\epsilon_{si}}{qL(x)^2} (V_{DS} + 2(\phi_d - \phi(x)) + 2\sqrt{(\phi_d - \phi(x))(V_{DS} + \phi_d - \phi(x))}) \right) \quad \text{Eq. II-28}$$

Afin de pouvoir intégrer analytiquement cette expression, il est nécessaire de faire une hypothèse forte sur la longueur des lignes de courant $L(x)$, comme illustré sur la Figure II-6.

- $x \leq X_j \rightarrow L(x) = L$
- $x > X_j \rightarrow L(x) = +\infty$

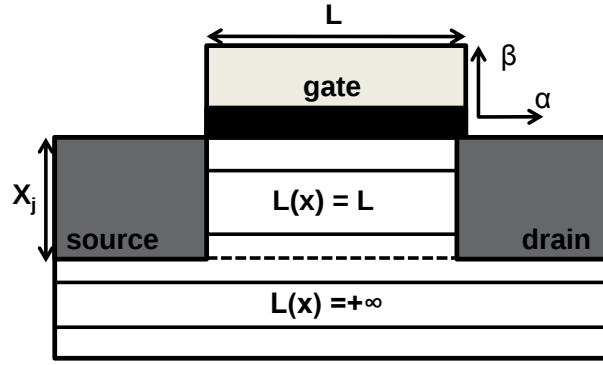


Figure II-6: Illustration de l'approximation faite sur la longueur $L(x)$ dans le processus de modélisation.

Grâce à cette hypothèse, nous sommes désormais capables de mener l'intégration de l'équation de Poisson de manière analytique. De plus, à partir du nœud CMOS 45nm, la profondeur de déplétion est généralement plus grande que la profondeur de jonction car cette dernière est réduite pour permettre un meilleur contrôle électrostatique, nous allons d'abord nous focaliser sur ce cas. En multipliant le premier terme de l'équation de Poisson (Eq II-28) par la dérivée du potentiel par rapport à x , et en intégrant de $x=0$ à $x=X_j$:

$$\int_0^{X_j} \frac{d\varphi}{dx} \frac{d^2\varphi}{dx^2} dx = \frac{1}{2} \left[\left(\frac{d\varphi}{dx} \right)^2 \right]_0^{X_j} = \frac{1}{2} \left[\left(\frac{d\varphi}{dx} \right)^2_{x=X_j} - \left(\frac{d\varphi}{dx} \right)^2_{x=0} \right] = \frac{1}{2} \left[\left(\frac{d\varphi}{dx} \right)^2_{x=X_j} - E_s^2 \right] \quad \text{Eq. II-29}$$

Où E_s est la valeur du champ électrique à l'interface silicium/oxyde de grille. En écrivant la loi de Gauss (Eq II-3) et en négligeant la charge d'inversion (Q_{inv}) car nous travaillons autour du seuil, on peut écrire que $E_s = Q_{dep}^* / \epsilon_{si}$. En remplaçant, dans Eq. II-29, on obtient une expression analytique de la charge de déplétion :

$$Q_{dep}^* = \epsilon_{si} \sqrt{\left(\frac{d\varphi}{dx} \right)^2_{x=X_j} - 2 \int_0^{X_j} \frac{d\varphi}{dx} \frac{d^2\varphi}{dx^2} dx} \quad \text{Eq. II-30}$$

Il reste donc à évaluer les deux termes sous la racine. Le premier est obtenu en intégrant (Eq. II 28) de $x=X_j$ à $x=T_{dep}$. Dans cette zone, on a $L(x)=+\infty$ ce qui implique $N_{ch}^*(x)=N_{ch}$:

$$\int_{X_j}^{T_{dep}} \frac{d^2\varphi}{dx^2} dx = \int_{X_j}^{T_{dep}} q \frac{N_{ch}}{\epsilon_{si}} dx \quad \text{Eq. II-31}$$

$$\frac{d\varphi}{dx}_{x=T_{dep}} - \frac{d\varphi}{dx}_{x=X_j} = q \frac{N_{ch}}{\epsilon_{si}} (T_{dep} - X_j) \quad \text{Eq. II-32}$$

A la limite de la zone de déplétion, le champ est nul, donc $\frac{d\varphi}{dx}_{x=T_{dep}} = 0$. L'équation II-32 donne alors l'expression analytique du premier terme de l'équation II-30 :

$$\left(\frac{d\varphi}{dx} \right)^2_{x=X_j} = \left(\frac{q N_{ch}}{\epsilon_{si}} (T_{dep} - X_j) \right)^2 \quad \text{Eq. II-33}$$

Pour le second terme de l'équation II-30, il faut intégrer analytiquement de $x=0$ à $x=X_j$ (donc $L(x)=L$) l'équation de Poisson transformée avec la VDT:

$$\int_0^{X_j} \frac{d\varphi}{dx} \frac{d^2\varphi}{dx^2} dx = \int_0^{X_j} \frac{d\varphi}{dx} \frac{q}{\epsilon_{si}} \left(N_{ch} - 2 \frac{\epsilon_{si}}{q L^2} (V_{DS} + 2 (\varphi_d - \varphi(x)) + 2 \sqrt{(\varphi_d - \varphi(x))(V_{DS} + \varphi_d - \varphi(x))}) \right) dx \quad \text{Eq. II-34}$$

Le terme de droite de l'égalité II-34 admet pour primitive analytique la fonction $\text{prim}(x)$:

$$\text{prim}(x) = \frac{qN_{\text{ch}}\varphi(x)}{\epsilon_{\text{si}}} + \frac{1}{L^2} \left(\frac{V_d^2}{2} \ln \left(\varphi_d - \varphi(x) + \frac{V_{DS}}{2} - \sqrt{\varphi(x)^2 - (V_{DS} + 2\varphi_d)\varphi(x) + \varphi_d(\varphi_d + V_d)} \right) + \sqrt{(\varphi_d - \varphi(x))(V_d + \varphi_d - \varphi(x))(V_d + 2\varphi_d - 2\varphi(x))} + \varphi(x)(2\varphi(x) - 4\varphi_d - V_d) \right) \quad \text{Eq. II-35}$$

Et donc:

$$\int_0^{X_j} \frac{d\varphi}{dx} \frac{d^2\varphi}{dx^2} dx = \frac{q}{\epsilon_{\text{si}}} (\text{prim}(X_j) - \text{prim}(0)) \quad \text{Eq. II-36}$$

Il ne reste alors à calculer que la valeur du potentiel sur la cathode virtuelle en $x=0$ (donc en surface) et en $x=X_j$. Pour le potentiel de surface, on cherche à calculer la tension de seuil, donc sa valeur est connue et égale à $2\varphi_f$. Pour le second, on intègre deux fois l'équation de Poisson de $x=X_j$ à $x=T_{\text{dep}}$ et on obtient :

$$\varphi(X_j) = \frac{qN_{\text{ch}}}{\epsilon_{\text{si}}} (X_j - T_{\text{dep}})^2 \quad \text{Eq. II-37}$$

Finalement, la charge de déplétion effective au seuil, tenant compte des effets canaux courts est donnée par:

$$Q_{\text{dep}}^* = \epsilon_{\text{si}} \sqrt{\left(\frac{qN_{\text{ch}}}{\epsilon_{\text{si}}} (T_{\text{dep}} - X_j) \right)^2 - 2 (\text{prim}(x = X_j) - \text{prim}(x = 0))} \quad \text{Eq. II-38}$$

Par conséquent, la tension de seuil, incluant SCE et DIBL, dans le cas où $X_j < T_{\text{dep}}$ s'écrit:

$$V_{\text{th}} = V_{\text{fb}} + 2\varphi_f + \frac{\epsilon_{\text{si}}}{C_{\text{ox}}} \sqrt{\left(\frac{qN_{\text{ch}}}{\epsilon_{\text{si}}} (T_{\text{dep}} - X_j) \right)^2 - 2 (\text{prim}(x = X_j) - \text{prim}(x = 0))} \quad \text{Eq. II-39}$$

Maintenant, si nous sommes dans le cas où la profondeur de déplétion est inférieure à la profondeur de jonction ($T_{\text{dep}} < X_j$) le calcul de la charge de déplétion est simplifié. On intègre cette fois de $x=0$ à $x=T_{\text{dep}}$, où on a $L(x)=L$:

$$\int_0^{T_{\text{dep}}} \frac{d\varphi}{dx} \frac{d^2\varphi}{dx^2} dx = E_s^2 \quad \text{Eq. II-40}$$

Comme précédemment, on peut lier champ électrique en surface et charge de déplétion effective:

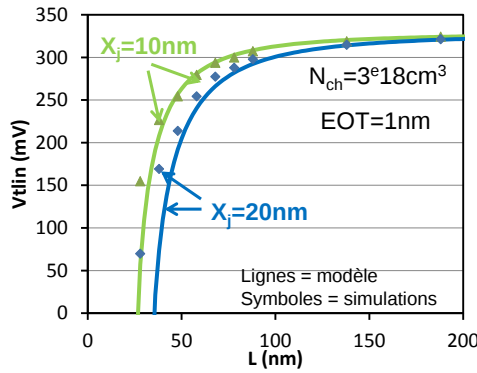
$$Q_{\text{dep}}^* = \epsilon_{\text{si}} \sqrt{2 \int_0^{T_{\text{dep}}} \frac{d\varphi}{dx} \frac{d^2\varphi}{dx^2} dx} = \epsilon_{\text{si}} \sqrt{2 \int_0^{T_{\text{dep}}} \frac{d\varphi}{dx} \frac{q}{\epsilon_{\text{si}}} \left(N_{\text{ch}} - 2 \frac{\epsilon_{\text{si}}}{qL^2} (V_{DS} + 2(\varphi_d - \varphi(x)) + 2\sqrt{(\varphi_d - \varphi(x))(V_{DS} + \varphi_d - \varphi(x))}) \right) dx} \quad \text{Eq. II-41}$$

Comme précédemment, l'intégrale sous la racine est intégrable analytiquement, ce qui permet également d'aboutir à une expression analytique de la tension de seuil :

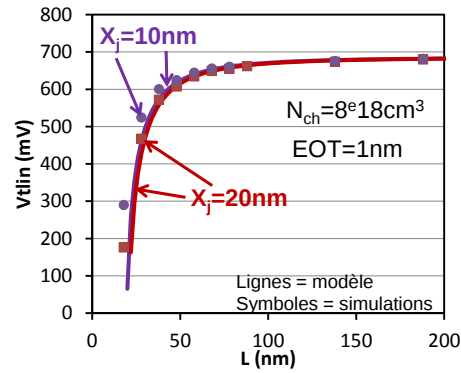
$$V_{th} = V_{fb} + 2\varphi_f + \frac{\varepsilon_{si} \sqrt{2(\text{prim}(0) - \text{prim}(T_{dep}))}}{C_{ox}} \quad \text{Eq. II-42}$$

Avec $\varphi(0) = 2\varphi_f$ and $\varphi(T_{dep}) = 0$

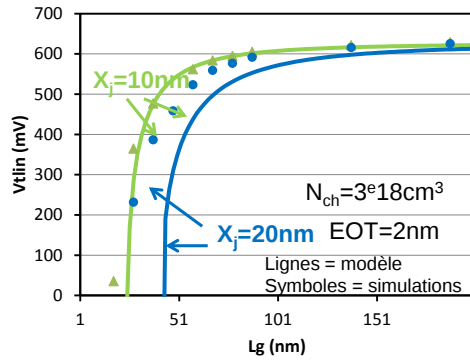
L'expression finale de la tension de seuil, tenant compte des effets canaux courts et du DIBL, est donc complètement analytique (Eq II-39 et II-42). Elle est obtenue avec une seule approximation sur la longueur des lignes de courant et ne nécessite pas de paramètre d'ajustement. Pour valider cette approche, nous avons comparé les résultats du modèle aux mêmes simulations qu'utilisées dans le paragraphe II.A.3.a.1).



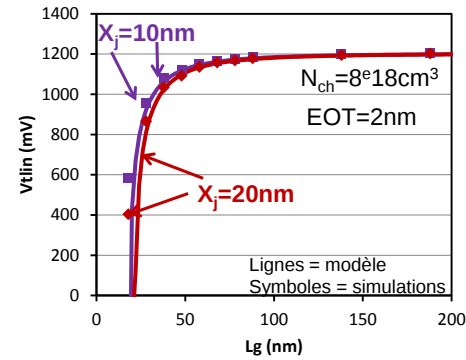
a)



b)



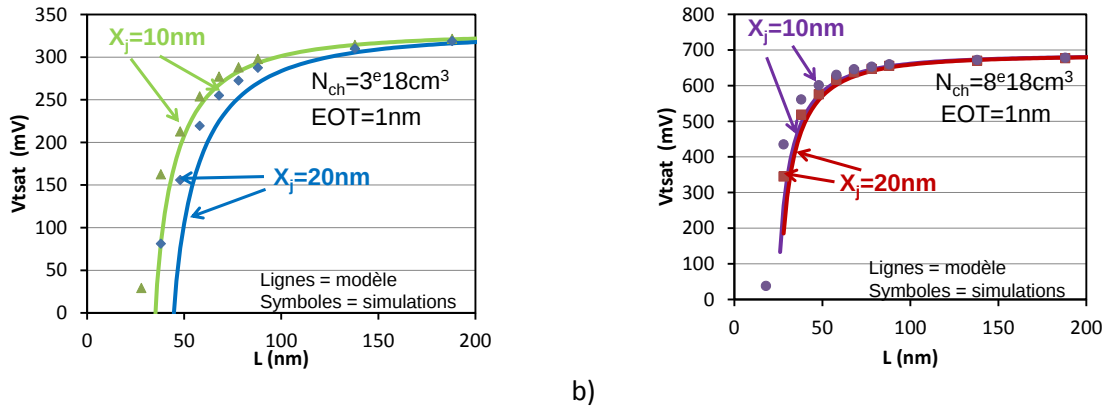
c)



d)

Figure II-7: Tracé de la tension de seuil en régime linéaire ($V_d=0.1V$) en fonction de la longueur de grille L_g pour différentes valeurs de dopage canal N_{ch} , profondeur de jonction X_j et d'EOT. a) EOT=1nm, $N_{ch}=3 \times 10^{18} \text{cm}^{-3}$, b) EOT=1nm, $N_{ch}=8 \times 10^{18} \text{cm}^{-3}$, c) EOT=2nm, $N_{ch}=3 \times 10^{18} \text{cm}^{-3}$, d) EOT=2nm, $N_{ch}=8 \times 10^{18} \text{cm}^{-3}$.

La Figure II-7 représente la tension de seuil en linéaire, V_{tlin} (à $V_{ds}=0.1V$) en fonction de la longueur de grille pour différentes valeurs de dopage canal, d'EOT et de profondeur de jonction. Cette approche permet une meilleure prise en compte de l'effet du dopage et de la profondeur de jonction, même si le modèle n'est pas très précis dans le cas EOT=2nm et $X_j=20\text{nm}$. Ceci s'explique par la valeur relâchée de l'EOT menant à des effets canaux courts très forts, mais également par la valeur assez importante de profondeur de jonction pour laquelle l'approximation parabolique du potentiel n'est plus tout à fait exacte.



a) b)
Figure II-8: Tracé de la tension de seuil en régime saturé ($V_d = V_{dd} = 1V$) en fonction de la longueur de grille L_g pour différentes valeurs de dopage canal N_{ch} , profondeur de jonction X_j et d'EOT. a) $EOT = 1 \text{ nm}$, $N_{ch} = 3 \times 10^{18} \text{ cm}^{-3}$, b) $EOT = 1 \text{ nm}$, $N_{ch} = 8 \times 10^{18} \text{ cm}^{-3}$.

La Figure II-8 montre que cette méthode de résolution de la VDT permet aussi une bonne prédiction de la tension en saturée, incluant donc le DIBL.

Enfin, on peut remarquer sur la Figure II-8 que le modèle ne donne pas de valeur pour les longueurs de grille inférieure à 30nm sur le graphe de a, et à 20nm sur le graphe de b). En effet, la charge de déplétion effective, donnée par les équations II-38 et II-41, n'a pas de valeur numérique car le terme sous la racine est négatif. Cela correspond approximativement à la longueur de grille pour laquelle les effets canaux courts sont forts, et où le dopage effectif est faible et devient négatif. On est alors proche du régime de perçage volumique qui, selon [Skotnicki 88-b] correspond au moment où le dopage effectif de la VDT devient négatif, et où la tension de seuil conventionnelle n'a plus vraiment de sens.

II.A.4. Pente sous le seuil

Comme pour les effets canaux courts sur la tension de seuil, la pente sous le seuil sera calculée par 2 méthodes distinctes : l'une d'ores et déjà implémenté dans Mastar [MASTAR] et l'autre partant de la charge de déplétion calculée dans le paragraphe II.A.3.b), dans les équations II-38 et II-41).

De manière classique [Gwoziecki 02], la pente sous le seuil est définie par :

$$S = \frac{dV_g}{d \log(I_d)} = \frac{kT}{q} \ln 10 \frac{dV_g}{d\phi_s} \quad \text{Eq. II-43}$$

En écrivant la loi de Gauss comme dans l'équation II-3, et en négligeant la charge d'inversion (Q_{inv}) car nous travaillons en régime de faible inversion ou de déplétion, on a :

$$S = \frac{kT}{q} \ln 10 \left(1 + \frac{C_{ss}}{C_{ox}} + \frac{1}{C_{ox}} \frac{dQ_{dep}}{d\phi_s} \right) \quad \text{Eq. II-44}$$

Le problème du calcul de la pente sous le seuil est donc réduit au calcul de la dérivée de la charge de déplétion ; problème que nous aborderons de deux manières :

II.A.4.a) Méthode « Mastar »

Afin d'obtenir une expression simple, il est proposé dans [Gwoziecki 02] d'approximer la dérivée de la charge de déplétion par son taux de variation entre $\phi_s = \phi_f$ et $\phi_s = 2\phi_f$:

$$S = \frac{kT}{q} \ln 10 \left(1 + \frac{C_{ss}}{C_{ox}} + \frac{1}{C_{ox}} \frac{\Delta Q_{dep}}{\phi_f} \right) \quad \text{Eq. II-45}$$

La charge de déplétion est ensuite évaluée à l'aide la VDT, de manière analogue au calcul de la tension de seuil [Pouydebasque 07]:

$$Q_{\text{dep}} = -qN_{\text{ch}}T_{\text{dep}} + 2\epsilon_{\text{si}}V_{\text{ds}} * \frac{X_j^2}{L^2} \quad \text{Eq. II-46}$$

La Figure II-9 nous montre que le modèle de pente sous le seuil donne une bonne première évaluation, mais ne reproduit mal les effets des paramètres technologiques autres que la longueur de grille.

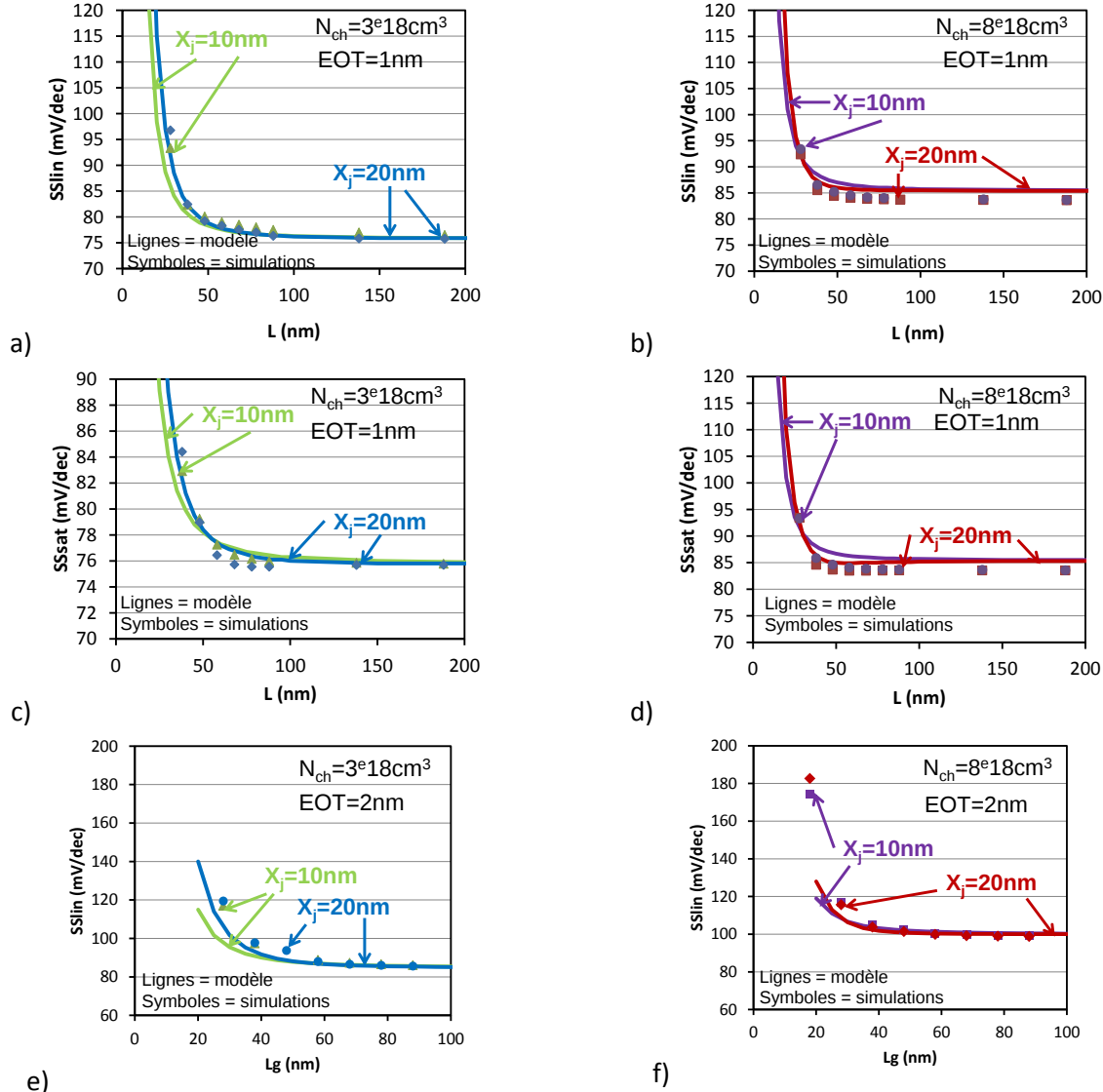


Figure II-9: Comparaison de l'approche "mastar" de modélisation de la pente sous le seuil (SS_{lin} : linéaire à $V_d = 0.1 \text{ V}$ et SS_{sat} : saturé à $V_d = V_{\text{dd}} = 1 \text{ V}$) avec les simulations numériques 2D en fonction de la longueur de grille (L_g), pour différents EOT, profondeur de jonction X_j et dopage canal N_{ch} .

II.A.4.b) Méthode exacte :

Dans cette deuxième approche, nous allons utiliser la charge de déplétion effective calculée grâce à la deuxième approche, donc les équations II-38 ou II-41 (en fonction de la valeur de la profondeur de déplétion), mais en

considérant le potentiel de surface comme une variable et non plus comme une constante égale à $2\phi_f$. Formellement, on a :

- $T_{dep} = \sqrt{2\epsilon_{si} \cdot \frac{\phi_s}{qN_{ch}}} \geq X_j \rightarrow Q_{dep}$ donnée par Eq. II-38
- $T_{dep} = \sqrt{2\epsilon_{si} \cdot \frac{\phi_s}{qN_{ch}}} < X_j \rightarrow Q_{dep}$ donnée par Eq. II-41

En dérivant la charge de déplétion et en réinjectant dans l'équation II-44, on obtient la valeur de la pente sous le seuil en fonction du potentiel de surface. Pour avoir sa valeur dans le régime de faible inversion, on choisit de l'évaluer pour $\phi_s = 1.5\phi_f$. La Figure II-10 représente l'évolution de pente sous le seuil avec la longueur de grille obtenue avec cette méthode de modélisation. Bien que toujours imparfaite, cette formulation permet une meilleure prise en compte de l'impact des paramètres technologiques sur la pente sous le seuil.

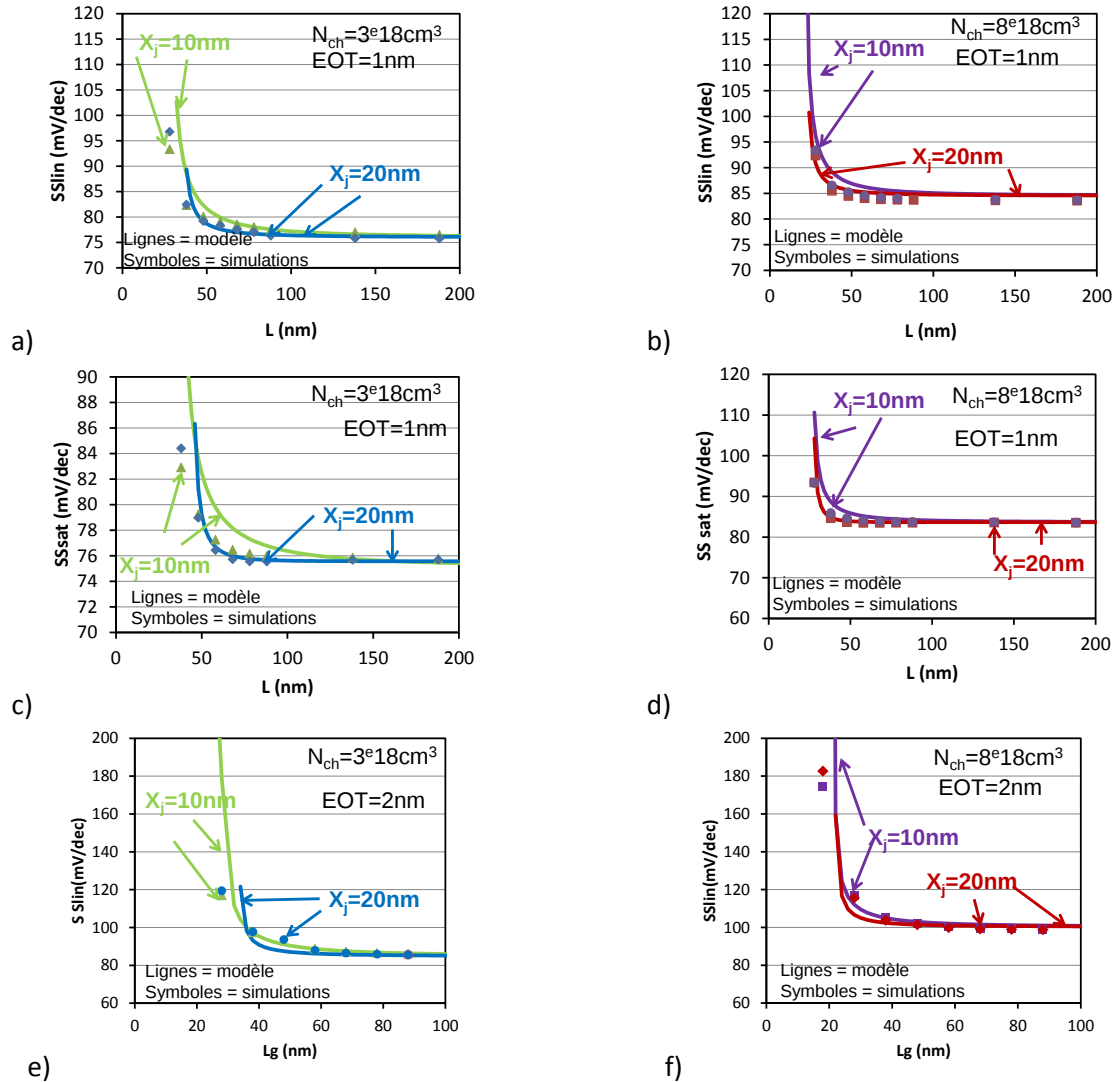
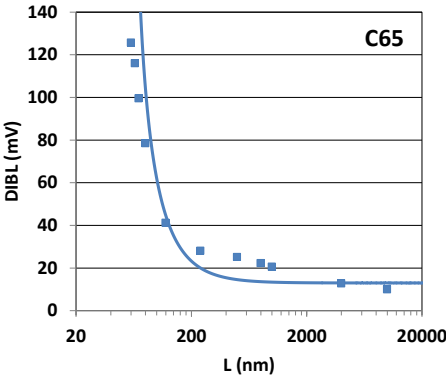
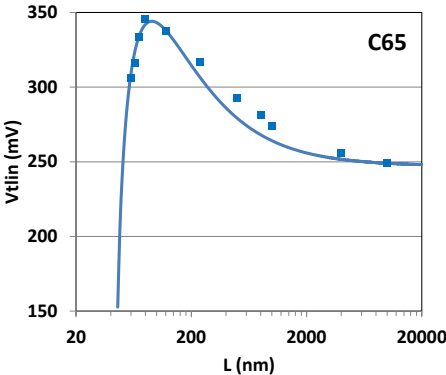


Figure II-10: Comparaison de l'approche "exacte" de modélisation de la pente sous le seuil (linéaire à $V_d = 0.1 \text{V}$ et saturé à $V_d = V_{dd} = 1 \text{V}$) avec les simulations numériques 2D en fonction de la longueur de grille (L_g), pour différents EOT, profondeur de jonction X_j et dopage canal N_{ch} . a) SS_{lin} , $EOT = 1 \text{nm}$, $N_{ch} = 3 \times 10^{18} \text{cm}^{-3}$, b) SS_{lin} , $EOT = 1 \text{nm}$, $N_{ch} = 8 \times 10^{18} \text{cm}^{-3}$, c) SS_{sat} , $EOT = 1 \text{nm}$, $N_{ch} = 3 \times 10^{18} \text{cm}^{-3}$, d) SS_{sat} , $EOT = 1 \text{nm}$, $N_{ch} = 8 \times 10^{18} \text{cm}^{-3}$, e) SS_{sat} , $EOT = 2 \text{nm}$, $N_{ch} = 3 \times 10^{18} \text{cm}^{-3}$, f) SS_{sat} , $EOT = 2 \text{nm}$, $N_{ch} = 8 \times 10^{18} \text{cm}^{-3}$.

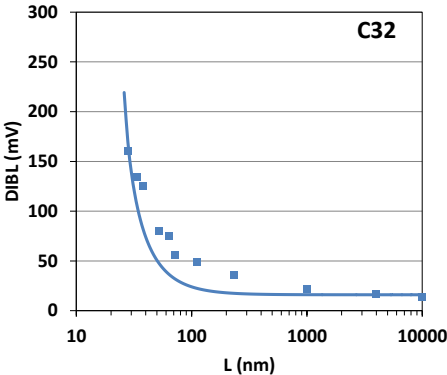
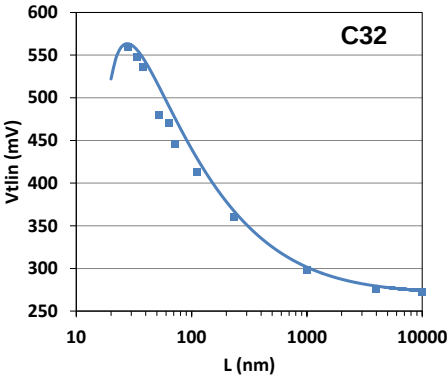
II.A.5. Comparaison aux mesures silicium :

Pour définitivement valider notre modèle de tension de seuil nous allons le comparer à des mesures effectuées sur deux technologies CMOS ST : le 65 et le 32 nm. Le 65nm utilise une grille en polysilicium et le SiO₂ comme matériau d'oxyde de grille : il sera donc affecté par la déplétion de grille (I.C.3.a)). Le 32nm utilise une grille métallique et un diélectrique haute permittivité (empilement dit « HK-metal ») : il ne sera donc pas affecté par la déplétion de grille mais il sera nécessaire de raisonner en EOT (I.C.3). Comme mentionné dans le paragraphe I.D.1.b) les technologies conventionnelles sur substrat massif avancées utilisent des implantations poches pour limiter l'impact des effets canaux courts, ce qui conduit à un dopage canal moyen dépendant de la longueur de grille. Pour tenir compte de cette spécificité technologique, nous remplaçons le dopage canal N_{ch} dans nos équations (Eq. II-42 et II-44) par l'expression du dopage canal moyen en fonction de la longueur de grille donné par le modèle de poches utilisé dans [MASTAR].

	silicium	Modèle
V _{dd}	1.2 V	1.2 V
X _j	/	15 nm
t _{ox}	1.9 nm	1.9 nm
Poly déplétion	oui	0.4 nm
Darkspace	oui	0.4 nm



	silicium	Modèle
V _{dd}	1 V	1 V
X _j	/	15 nm
EOT	1.1 nm	1.1 nm
Poly déplétion	non	0 nm
Darkspace	oui	0.4 nm



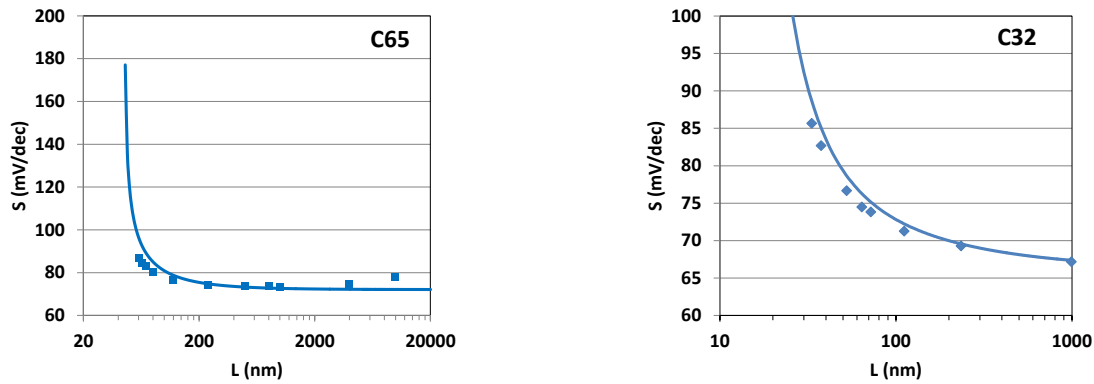
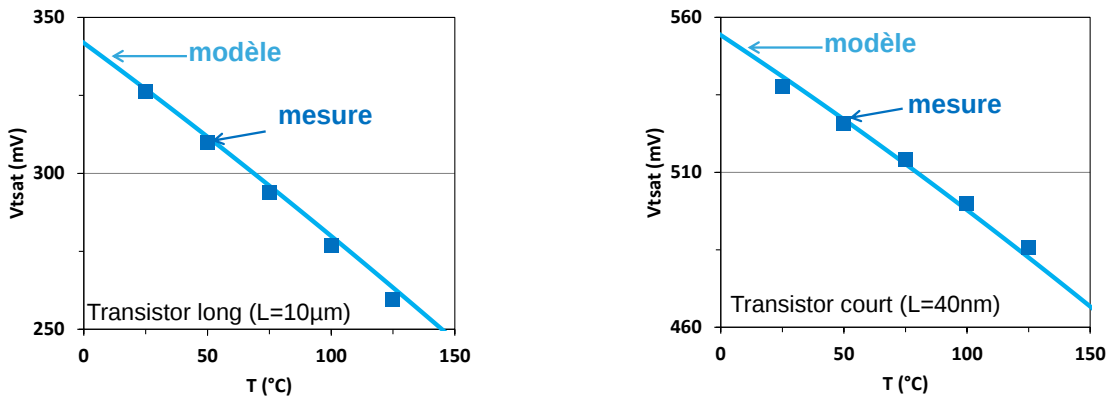


Figure II-11: Comparaison de l'approche "exacte" de modélisation avec les mesures effectuées sur les technologies ST 65 (colonne de gauche) et 32 nm (colonne de droite).

Le modèle proposé dans cette partie permet donc bien de reproduire les caractéristiques électrostatiques d'une plateforme CMOS complète, quel que soit l'empilement de grille utilisée.

II.A.6. Validation du comportement en température :

Nous allons utiliser ce modèle pour des simulations de circuit, le comportement en température de ce modèle est par conséquent primordial. Pour le valider, nous avons effectué des mesures sur un transistor long et un transistor court de technologie CMOS 40nm pour différentes températures (de 0 à 125°C). Notre modèle de tension reproduit bien les mesures, et le comportement en température de ce modèle est donc bien validé (Figure II-12).



a)

b)

Figure II-12 : Tracé de la tension de seuil en fonction de la température obtenu par mesure de transistors de technologie. Notre modèle reproduit les mesures effectuées sur un transistor long (a, $L=10\mu m$) et un transistor court (b, $L=40nm$)

II.B. Tension de seuil et électrostatique de l'architecture FDSOI

II.B.1. Critère d'inversion

Pour une bonne modélisation de tension de seuil, un critère d'inversion adapté à l'architecture est nécessaire. Comme la condition de seuil usuelle de l'architecture conventionnelle sur substrat massif est donnée par le potentiel de surface égal à deux fois le potentiel de Fermi, les premiers travaux de modélisation de tension de

seuil du FDSOI ont été effectués avec cette hypothèse [Lim 83], mais également des plus récents [Suzuki 03]. Or, les transistors FDSOI utilisent des canaux non-dopés ou faiblement dopés, c'est-à-dire avec un dopage inférieur à $5 \times 10^{16} \text{cm}^{-3}$, et dans de telles conditions, il est montré par [Lee 89] que le passage en inversion forte a lieu pour une concentration de porteurs minoritaire supérieure au dopage canal. Comme la condition de seuil $\phi_s = 2\chi\phi_f$ découle de l'égalité entre densité de porteurs minoritaires et majoritaires (cf II.A.1) il est désormais clair que pour des dispositifs à canaux faiblement dopés (dont les transistors FDSOI) cette condition ne sera pas utilisable, et que le seuil a lieu pour un potentiel de surface supérieur à $2\phi_f$.

Pour cette raison, des travaux plus récents proposent d'utiliser d'autre critère d'inversion. Dans [Chen 03], les auteurs ont effectué des simulations numériques pour déterminer la valeur de la charge d'inversion au seuil sur le transistor à canal long et ils l'utilisent pour remonter à la tension de seuil. On peut justifier cette approche en la considérant équivalent à l'extraction de tension de seuil à courant constant. D'autres travaux [Han 08] [Poiroux 05] définissent le seuil par l'égalité entre capacité d'inversion et capacité d'oxyde de grille ($=\epsilon_{ox}/t_{ox}$), ce qui est justifié par [Tsividis 87] et aboutissent alors à un critère de seuil donnant une condition sur la valeur de la capacité d'inversion au seuil.

Dans notre cas, pour obtenir un critère de seuil avec du sens physique, nous choisissons d'utiliser le point d'inflexion sur la courbe capacité grille-tension ($C_g(V)$), qu'on traduit mathématiquement par l'équation :

$$\frac{d^2 C_g}{dV_g^2} = 0 \quad \text{Eq. II-47}$$

Il est montré dans [Flandre 10] que le point d'inflexion de la $C_g(V)$ est équivalent au maximum de la dérivée de la transconductance, qui est également un critère de seuil [Wong 87]. Enfin, pour définitivement valider ce critère de seuil, nous avons procédé à des simulations numériques sur des dispositifs à canaux non-dopés en utilisant l'outil FlexPDE [FlexPDE] qui est un solveur d'équation aux dérivées partielles. Nous avons ensuite tracé la variation de la charge d'inversion (Q_{inv}), la capacité grille (C_g) et de sa dérivée en fonction de la polarisation de grille (V_g) (Figure II-13).

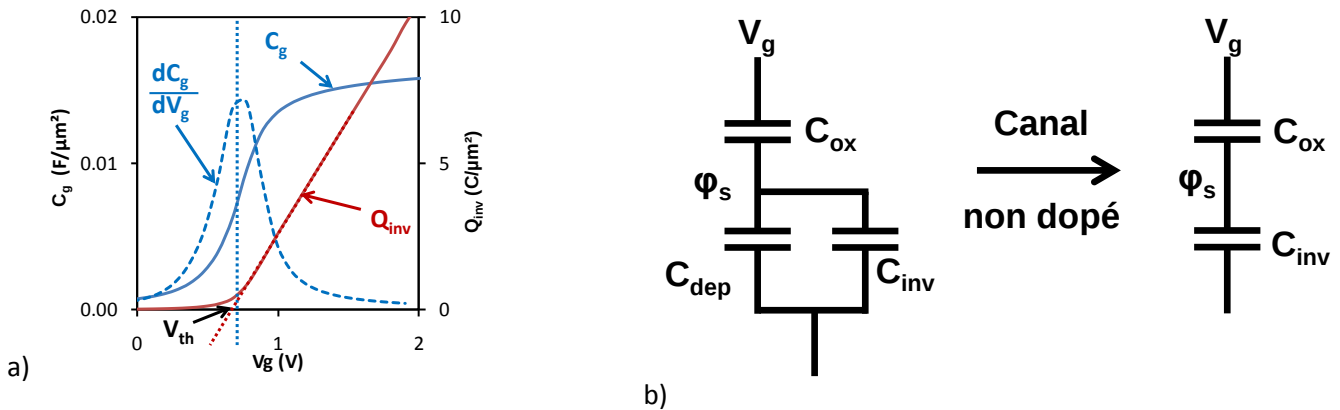


Figure II-13: (a) Variation de la charge d'inversion (Q_{inv}), la capacité grille (C_g) et de sa dérivée en fonction de la polarisation de grille (V_g) pour un transistor à canal non dopé. (b) schéma équivalent de la capacité grille dans le cas d'un transistor non dope, la capacité de déplétion peut être négligée.

La Figure II-13 montre que la tension de grille correspondant au point d'inflexion sur la $C_g(V)$ (donc le maximum de sa dérivée) correspond bien à la tension de seuil obtenue par extrapolation sur la charge d'inversion.

Le point d'inflexion sur la $C_g(V)$, et donc l'équation II-47 sont bien utilisables comme critère de seuil pour un transistor à canal non dopé, ou faiblement dopé comme le FDSOI. Cependant, pour être utilisable dans une modélisation de tension de seuil, ce critère doit être développé. En écrivant la capacité grille (C_g) en fonction de la capacité d'oxyde (C_{ox}) et de la capacité d'inversion (C_{inv}) à l'aide du schéma de la Figure II-13-b) :

$$C_g(\varphi_s) = \frac{C_{ox}C_{inv}(\varphi_s)}{C_{ox}+C_{inv}(\varphi_s)} \quad \text{Eq. II-48}$$

De plus, sachant que l'expression de la charge inversion en fonction du potentiel de surface dans un dispositif à film mince (épaisseur t_{si}) peut s'écrire approximativement :

$$Q_{inv} = q t_{si} n_i e^{\frac{q\varphi_s}{kT}} \quad \text{Eq. II-49}$$

Avec n_i concentration de porteurs intrinsèques. La dérivée de la charge d'inversion par rapport au potentiel de surface, donc la capacité due à la charge d'inversion, vaut alors :

$$C_{inv}(\varphi_s) = \frac{dQ_{inv}}{d\varphi_s} = \frac{q}{kT} Q_{inv} \quad \text{Eq. II-50}$$

En écrivant la loi de Gauss (Eq. II-3) en négligeant la charge de déplétion, on peut exprimer la dérivée du potentiel de surface par rapport à la tension de grille :

$$\frac{dV_g}{d\varphi_s} = 1 + \frac{dQ_{inv}}{d\varphi_s} \frac{1}{C_{ox}} \quad \text{Eq. II-51}$$

Enfin, en décomposant le calcul de la dérivée de la capacité de grille par rapport à la tension de grille comme ci-dessous :

$$\frac{dC_{gc}}{dV_g} = \frac{dC_{gc}}{d\varphi_s} \frac{d\varphi_s}{dV_g} \quad \text{Eq. II-52}$$

La résolution de l'équation II-47 donne alors comme critère de seuil :

$$Q_{ith} = \frac{kT}{q} \frac{C_{ox}}{2} \quad \text{Eq. II-53}$$

Où Q_{ith} est la valeur de la charge d'inversion au seuil.

Ensuite, pour aboutir à l'expression du potentiel de surface au seuil (φ_{sth}), on recherche le potentiel de surface pour lequel la charge d'inversion de l'équation II-48 atteint la valeur donnée par le critère d'inversion (Eq. II-53). Formellement :

$$q n_i t_{si} e^{\left(\frac{q\varphi_{sth}}{kT}\right)} = Q_{ith} = \frac{kT}{q} C_{ox} \quad \text{Eq. II-54}$$

$$\varphi_{sth} = \frac{kT}{q} \ln \left(\frac{\frac{kT}{q} C_{ox}}{q n_i t_{si}} \right) \quad \text{Eq. II-55}$$

II.B.2. Tension de seuil canal long

Pour la modélisation de la tension de seuil dans le cas d'un transistor FDSOI à canal long, nous procédons comme pour l'architecture conventionnelle sur substrat massif et commençons donc à écrire la loi de Gauss dans le canal du transistor :

$$\varepsilon_{si} E_s = C_{ox}(V_g - V_{FB} - \varphi_s) \quad \text{Eq. II-56}$$

Où V_{FB} est la tension de bandes plates. Pour déterminer la tension de seuil, il nous suffit alors de remplacer le potentiel de surface ϕ_s par sa valeur au seuil ϕ_{sth} donnée par l'équation II-54.

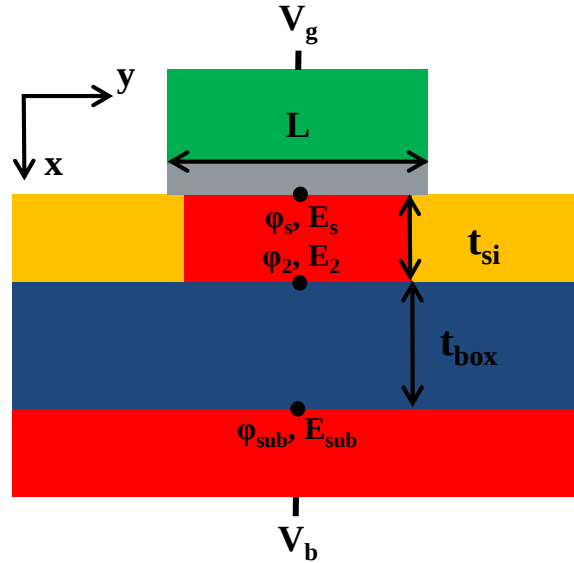


Figure II-14 : Représentation schématique de l'architecture FDSOI avec ses principales dimensions et les valeurs de potentiel et de champ électrique nécessaires à la modélisation.

Il faut cependant déterminer la valeur du champ de surface correspondante. Pour ce faire, résolvons l'équation de Poisson dans la structure de la Figure II-14:

$$\frac{d^2\phi}{dx^2} = -\frac{qN_{ch}}{\epsilon_{si}} \quad \text{Eq. II-57}$$

En intégrant deux fois de $x=0$ à $x=t_{si}$, et connaissant les valeurs du potentiel et du champ en $x=t_{si}$, respectivement ϕ_2 et E_2 , on obtient l'expression du champ électrique en surface puis l'équation de variation du potentiel dans le canal:

$$E_s = E_2 - \frac{qN_{ch}t_{si}}{\epsilon_{si}} \quad \text{Eq. II-58}$$

$$\phi(x) = -\frac{qN_{ch}}{2\epsilon_{si}}(x - t_{si})^2 - E_2(x - t_{si}) + \phi_2 \quad \text{Eq. II-59}$$

Nous considérerons dans un premier temps le cas simple, où le potentiel ϕ_{sub} est égal à la polarisation de la face arrière V_b . L'expression du champ E_2 est obtenu par continuité du vecteur déplacement du champ électrique à une interface entre deux matériaux, et en considérant le champ constant dans l'oxyde enterré (E_{box}). En notant E_2^- la valeur du champ électrique en $x=t_{si}$ dans le silicium, et E_2^+ la valeur du champ électrique en $x=t_{si}$ dans l'oxyde enterré, on a :

$$\epsilon_{si}E_2^- = \epsilon_{siO_2}E_2^+ = \epsilon_{siO_2}E_{box} \quad \text{Eq. II-60}$$

Le champ dans l'oxyde enterré étant constant, on peut écrire :

$$E_{box} = \frac{(\phi_2 - V_b)}{t_{box}} \quad \text{Eq. II-61}$$

En combinant les équations II-60 et II-61, on obtient l'expression de φ_2 :

$$\varphi_2 = \frac{\varepsilon_{si}}{\varepsilon_{ox}} E_2 t_{box} + V_b \quad \text{Eq. II-62}$$

En injectant les expressions de E_2 (Eq. II-58) et de φ_2 (Eq. II-62) dans l'expression du potentiel dans le canal (Eq. II-59) et en l'évaluant en surface en $x=0$, on obtient l'expression analytique du potentiel de surface φ_s :

$$\varphi_s = -\frac{qN_{ch}t_{si}^2}{\varepsilon_{si}} + \left(E_s + \frac{qN_{ch}t_{si}}{\varepsilon_{si}}\right)t_{si} + \left(E_s + \frac{qN_{ch}t_{si}}{\varepsilon_{si}}\right)\frac{\varepsilon_{si}}{\varepsilon_{ox}} t_{box} + V_b \quad \text{Eq. II-63}$$

En ré-arrangeant l'équation II-63, on peut écrire l'expression du produit du champ électrique en surface E_s avec la permittivité du silicium ε_{si} :

$$\varepsilon_{si} E_s = \frac{C_{si}C_{box}}{C_{si}+C_{box}} \left(\varphi_s - V_b + qN_{ch}t_{si} \left(\frac{1}{2C_{si}} + \frac{1}{C_{box}} \right) \right) \quad \text{Eq. II-64}$$

En réinjectant l'équation II-64 dans la loi de Gauss (Eq. II-56), puis en se plaçant au seuil, on obtient l'expression de la tension de seuil ($\varphi_s = \varphi_{sub}$) pour un transistor FDSOI à canal long :

$$V_{thlong} = V_{FB} + \left(1 + \frac{C_{box}C_{si}}{C_{ox}(C_{si}+C_{box})}\right) \varphi_{sth} + \frac{qN_{ch}t_{si}}{C_{ox}} \left(\frac{1}{2C_{si}} + \frac{1}{C_{box}} \right) - \frac{C_{box}C_{si}}{C_{ox}(C_{si}+C_{box})} V_b \quad \text{Eq. II-65}$$

Cependant, la structure considérée dans la Figure II-14 est trop simplifiée pour être représentative du comportement « réel » d'un tel dispositif. Les trois phénomènes suivant sont omis :

- La tension de bandes plates de la face arrière (i.e du plan de masse) due à la différence des travaux de sortie entre le silicium du canal et le silicium du plan de masse.
- L'existence d'une couche de déplétion dans le plan de masse (de dopage N_{GP}), à l'interface avec l'oxyde enterré, qui provoque un épaissement de la couche d'oxyde enterré à considérer d'un point de vue électrostatique. Il s'agit d'un phénomène analogue à la déplétion de grille dans le cas des grilles en polysilicium. Son existence et son épaisseur est conditionnée par le type de dopage du plan de masse, son niveau de dopage ainsi que de la polarisation de la face arrière.
- Le couplage du canal avec la face arrière dépend de la face du canal qui est en conduction (face avant : interface oxyde de grille canal / face arrière : interface canal oxyde enterré) [Noel 11].

Pour tenir compte de l'effet de la tension de bandes plates face arrière, il suffit de remplacer la polarisation de la face arrière V_b par sa valeur V_b' corrigée par la tension de bandes plates face arrière V_{fbGP} :

$$V_b' = V_b - V_{fbGP} \quad \text{Eq. II-66}$$

Où V_{fbGP} est évaluée par analogie avec la tension de bandes plates classique de la grille. En fonction du type de dopage du plan de masse, on a :

Type de transistor	PMOS	NMOS
Plan de masse type N	$V_{fbGP} = \frac{kT}{q} \ln \left(\frac{N_{ch}}{N_{GP}} \right)$	$V_{fbGP} = \frac{kT}{q} \ln \left(\frac{n_i^2}{N_{GP}N_{ch}} \right)$
Plan de masse type P	$V_{fbGP} = \frac{kT}{q} \ln \left(\frac{N_{GP}N_{ch}}{n_i^2} \right)$	$V_{fbGP} = \frac{kT}{q} \ln \left(\frac{N_{GP}}{N_{ch}} \right)$

S'il existe une couche de déplétion (les conditions de son existence seront données plus tard) à l'interface plan de masse-oxyde enterré, la structure à modéliser n'est plus celle de la Figure II-14 et le potentiel ϕ_{sub} est différent de la polarisation de la face arrière V_b' . La structure à modéliser est donc celle représentée ci-dessous (Figure II-15) :

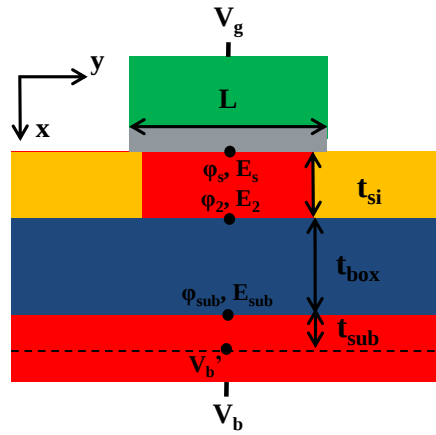


Figure II-15 : Structure à considérer si il existe une couche de déplétion à l'interface plan de masse-oxyde enterré.

La nouvelle inconnue à déterminer est donc l'épaisseur de cette couche de déplétion t_{sub} . Cette dernière est évaluée avec la formule classique d'une couche de déplétion dans un semiconducteur :

$$t_{sub} = \sqrt{\frac{2\epsilon_{si}}{qN_{GP}}(\phi_{sub} - V_b')} \quad \text{Eq. II-67}$$

Enfin, pour connaître t_{sub} il suffit de déterminer ϕ_{sub} . Pour ce faire, il faut résoudre à nouveau l'équation de Poisson, mais cette fois avec $\phi_{sub} \neq V_b'$. Reprenons le raisonnement précédent à partir de l'expression du potentiel dans le canal (Eq. II-59) puis réécrivons l'égalité obtenue à partir de la continuité du vecteur déplacement du champ électrique à l'interface entre deux matériaux. On applique cette loi à l'oxyde enterré, dans lequel le champ électrique E_{box} est constant. On a :

$$\epsilon_{si}E_2 = \epsilon_{ox}E_{box} = \epsilon_{si}E_{sub} \quad \text{Eq. II-68}$$

La valeur du champ E_{sub} est étant donné par l'expression :

$$E_{sub} = \frac{qN_{GP}}{\epsilon_{si}}t_{sub} \quad \text{Eq. II-69}$$

En couplant les équations II-68 et II-69, on obtient l'expression de E_2 :

$$E_2 = E_{sub} = \frac{qN_{GP}}{\epsilon_{si}}t_{sub} \quad \text{Eq. II-70}$$

Le champ dans l'oxyde enterré est considéré constant son expression est connue et permet d'obtenir la formule du potentiel ϕ_2 :

$$E_{box} = \frac{\phi_2 - \phi_{sub}}{t_{box}} \rightarrow \phi_2 = \phi_{sub} + E_{box}t_{box} \quad \text{Eq. II-71}$$

En remplaçant les termes φ_2 et E_2 par leurs expressions (Eq. II-70 et II-71) dans l'expression du potentiel dans le canal (Eq. II-59), puis en évaluant ce potentiel en $x=0$, on obtient l'expression liant le potentiel de surface φ_s et le potentiel φ_{sub} :

$$\varphi_{sub} = \varphi_s - \frac{qN_{ch}t_{si}^2}{2\epsilon_{si}} - \frac{\epsilon_{si}t_{box} + t_{si}}{\epsilon_{ox}} \sqrt{2\epsilon_{si}qN_{GP}(\varphi_{sub} - V_b')} \quad \text{Eq. II-72}$$

Notre objectif est de calculer la tension de seuil, le potentiel de surface est alors connu et vaut φ_{sth} (Eq. II-55). En posant $X^2 = (\varphi_{sub} - V_b')$, l'équation II-72 devient une équation du second degré facilement soluble et l'expression du potentiel φ_{sub} au seuil est donc :

$$\varphi_{sub} = V_b' + \frac{1}{4} \left[- \left(\frac{1}{C_{si}} + \frac{1}{C_{box}} \right) \sqrt{2\epsilon_{si}qN_{GP}} \pm \sqrt{2\epsilon_{si}qN_{GP} \left(\frac{1}{C_{si}} + \frac{1}{C_{box}} \right)^2 - \frac{2qN_{ch}t_{si}^2}{\epsilon_{si}} - 4(V_b' - \varphi_{sth})} \right]^2 \quad \text{Eq. II-73}$$

La valeur de la profondeur de déplétion à l'interface plan de masse oxyde enterré est maintenant connue. Pour prendre en compte son impact sur la tension de seuil, nous allons procéder de la même manière que pour la déplétion de grille, en remplaçant l'épaisseur d'oxyde enterré t_{box} par une épaisseur d'oxyde enterré équivalente, notée t_{boxeq} , dans l'expression de la tension de seuil (Eq. II-65) :

$$t_{boxeq} = t_{box} + \frac{\epsilon_{ox}}{\epsilon_{si}} t_{sub} \quad \text{Eq. II-74}$$

Pour déterminer les conditions de l'existence de cette couche de déplétion, nous avons procédé à des simulations numériques et tracé la variation de la tension de seuil en fonction de la polarisation de la face arrière (Figure II-16), exclusivement pour des NMOS mais pour les deux types de plan de masse (GPN : plan de masse dopé N et GPP plan de masse de type P) et pour deux valeurs des niveaux de dopage ($2 \times 10^{18} \text{ cm}^{-3}$ où une couche de déplétion peut se former et $1 \times 10^{20} \text{ cm}^{-3}$ où le dopage est trop élevé pour permettre la formation d'une couche de déplétion non négligeable).

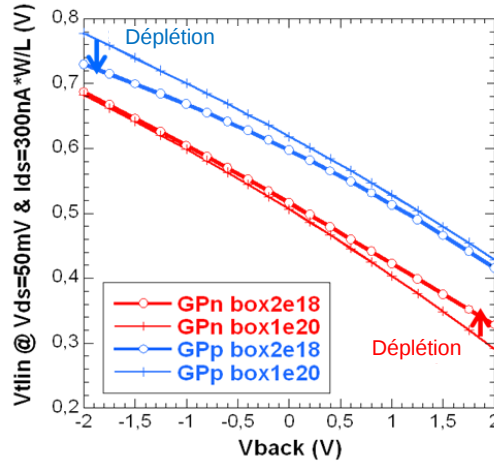


Figure II-16 : tracé la variation de la tension de seuil en fonction de la polarisation de la face arrière, exclusivement pour des NMOS mais pour les deux types de plan de masse (GPN : plan de masse dopé N et GPP plan de masse de type P) et pour deux valeurs des niveaux de dopage ($2 \times 10^{18} \text{ cm}^{-3}$ où une couche de déplétion peut se former et $1 \times 10^{20} \text{ cm}^{-3}$ où le dopage est trop élevé pour permettre la formation d'une couche de déplétion non négligeable).

D'après la Figure II-16, nous pouvons considérer que, dans le cas d'un transistor NMOS, la couche de déplétion d'épaisseur t_{sub} se forme sous l'oxyde enterré pour :

- $V_b' > 0$ pour un plan de masse de type N.
- $V_b' < 0$ pour un plan de masse de type P.

Enfin, il est démontré dans [Noel 11] que le canal de conduction peut se former sur l'interface oxyde enterré-canal (dite interface arrière) alors que nous n'avons considéré jusqu'à présent que le cas de la conduction à l'interface oxyde de grille-canal (dite interface avant). Selon [Noel 11] il faut considérer :

- La conduction a lieu en face arrière si $V_b' > 0$
- La conduction a lieu en face avant si $V_b' < 0$

Cette remarque a un impact sur le comportement de la tension de seuil (V_t) avec la polarisation de la face arrière (V_b). On modélise cet effet en écrivant la tension de seuil comme la somme de la tension de seuil à V_b' nul et d'un terme proportionnel à V_b' :

$$V_t = V_t(V_b' = 0) + \gamma \times V_b' \quad \text{Eq. II-75}$$

Ensuite, [Noel 11] propose d'estimer la pente de la courbe $V_t(V_b')$ (notée γ) à l'aide de deux simples diviseurs capacitifs, représentant chacun un type de conduction. Ces derniers sont représentés sur la Figure II-17.



a)

b)

Figure II-17 : Diviseur capacitif représentatif de la conduction face avant (a) et face arrière (b) utilisé pour calculer l'impact de la polarisation sur la face arrière.

Finalement, en tenant compte de la couche de déplétion dans le plan de masse, le paramètre γ de l'équation II-75 est évalué en distinguant quatre cas, en fonction de la polarisation V_b' et du type du plan de masse. Pour un NMOS, les valeurs du paramètre γ sont données dans la Figure II-18.

	Plan de masse de type N	Plan de masse de type P
$V_b' < 0$	$\gamma = \frac{C_{box} C_{si}}{C_{ox}(C_{si} + C_{box})}$	$\gamma = \frac{C_{boxeq} C_{si}}{C_{ox}(C_{si} + C_{boxeq})}$
$V_b' > 0$	$\gamma = \frac{C_{boxeq}(C_{ox} + C_{si})}{C_{ox} C_{si}}$	$\gamma = \frac{C_{box}(C_{ox} + C_{si})}{C_{ox} C_{si}}$

Figure II-18 : Valeur du paramètre de couplage γ pour un NMOS en fonction du type du plan de masse et de la polarisation de la face arrière. C_{boxeq} est la capacité équivalente de l'oxyde enterré, tenant compte de la déplétion dans le plan de masse et vaut $C_{boxeq} = \epsilon_{ox}/t_{boxeq}$ où t_{boxeq} est donné par l'équation II-73.

Pour tenir compte de cet effet, il faut remplacer le terme proportionnel à la polarisation de la face arrière dans l'équation II-65.

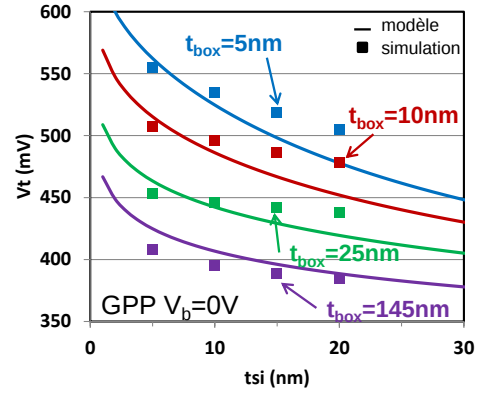
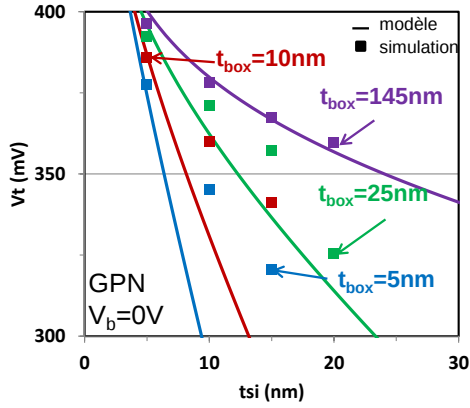
Finalement, la tension de seuil d'un transistor FDSOI dans le cas d'un canal long vaut :

$$V_{thlong} = V_{FB} + \left(1 + \frac{C_{boxeq} C_{si}}{C_{ox}(C_{si} + C_{boxeq})}\right) \varphi_{sth} - \frac{q N_{ch} t_{si}}{C_{ox}} \left(\frac{1}{2 C_{si}} + \frac{1}{C_{boxeq}}\right) - \gamma V_b' \quad \text{Eq. II-76}$$

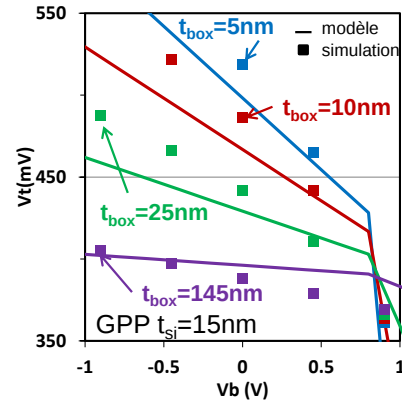
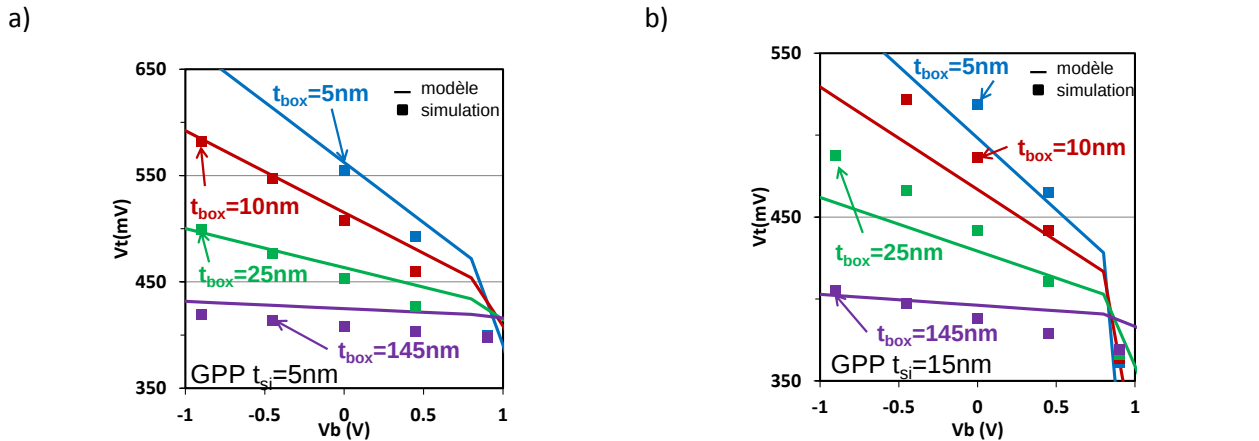
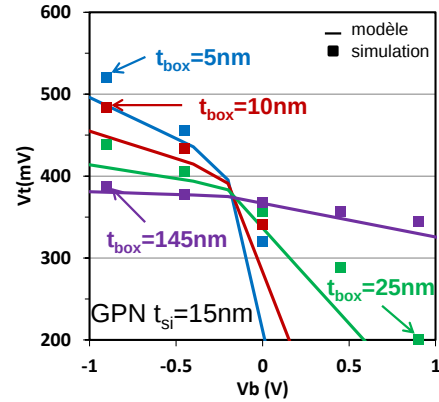
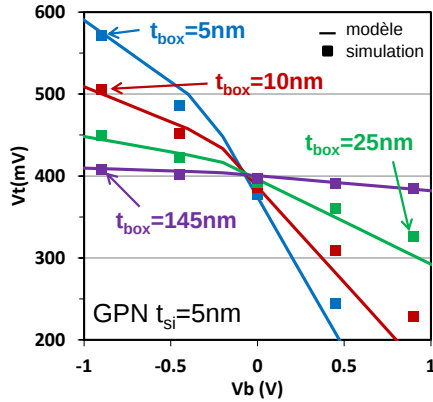
Avec :

- V_b' est donné par l'équation II-66.
- $C_{boxeq} = \epsilon_{ox}/t_{boxeq}$ où t_{boxeq} est donnée par l'équation II-74 si les conditions induisent l'existence d'une couche de déplétion dans le plan.
- $C_{boxeq} = C_{box}$ si les conditions n'induisent pas l'existence d'une couche de déplétion dans le plan de masse.
- γ est donné par la Figure II-18.

Nous avons procédé à des simulations numériques sur un transistor FDSOI à canal long ($L=1\mu m$) pour valider notre modèle en faisant varier l'épaisseur de film de silicium t_{si} , l'épaisseur d'oxyde enterré t_{box} , le type du plan de masse et sa polarisation. Dans chaque cas, l'EOT vaut 1nm et le dopage du plan de masse vaut $2 \times 10^{18} cm^{-3}$ (dopage typique de l'état de l'art [Fenouillet 11]). Les simulations numériques sont effectuées en résolvant uniquement l'équation de Poisson, pour éviter d'impacter les résultats par les effets quantiques, qui ne sont pas pris en compte dans notre modèle, et qui n'ont de réel effet que pour des films de silicium très fin, inférieur à 5nm [Ramey 03]. La Figure II-19 représente la variation de la tension de seuil avec l'épaisseur du film de silicium pour différentes valeurs d'épaisseur d'oxyde enterré et les deux types de plan de masse et la Figure II-20 montre les variations de tension de seuil avec la polarisation de la face arrière pour les mêmes géométries et pour les deux types de plan de masse. Pour un NMOS avec plan de masse de type P (Figure II-19-b), on constate que le modèle reproduit bien les simulations numériques de tension de seuil canal long pour des variations d'épaisseur de film de silicium t_{si} . Cependant, si le plan de masse est de type N, le modèle est moins précis, mais donne tout de même une approximation premier ordre correcte. Enfin, le comportement de la tension de seuil avec la polarisation de la face arrière est bien pris en compte par le modèle, même si l'accord entre valeurs de tension de seuil obtenues par modèle et par simulations numériques n'est pas parfait.



a) b)
Figure II-19 : Variation de la tension de seuil d'un NMOS à canal long avec l'épaisseur du film de silicium pour une polarisation de plan de masse nulle et quatre épaisseurs d'oxyde enterré. a) plan de masse de type N et b) plan de masse de type P. Les lignes continues correspondent aux valeurs obtenues par notre modèle et les symboles aux simulations numériques.



c) d)
Figure II-20 : Variations de la tension de seuil d'un NMOS à canal long avec la polarisation de la face arrière pour quatre épaisseurs d'oxyde enterré. a) plan de masse de type N et $t_{si}=5nm$ b) plan de masse de type N et $t_{si}=15nm$ c) plan de masse de type P et $t_{si}=5nm$ d) plan de masse de type P et $t_{si}=15nm$. Les lignes continues correspondent aux valeurs obtenues par notre modèle et les symboles aux simulations numériques.

II.B.3. Effets canaux courts (SCE) et DIBL

Comme pour l'architecture conventionnelle sur substrat massif, nous allons utiliser la VDT pour modéliser l'effet canal court SCE et le DIBL. Dans un transistor FDSOI, le canal est non dopé et on peut considérer que le potentiel dans le canal, donc le long de la cathode virtuelle, est constant et vaut, au seuil, ϕ_{sth} . On peut écrire l'expression du dopage effectif correspondante :

$$N_{ch}^* = N_{ch} - 2 \frac{\epsilon_{si}}{qL^2} (V_{DS} + 2(\phi_d - \phi_{sth}) + 2\sqrt{(\phi_d - \phi_{sth})(V_{DS} + \phi_d - \phi_{sth})}) \quad \text{Eq. II-77}$$

Dans un transistor FDSOI, le canal est d'épaisseur t_{si} et est totalement déplété. Nous pouvons donc considérer que la profondeur de déplétion vaut t_{si} . En procédant comme dans le cas du transistor sur substrat massif (II.A.3.b), on obtient les expressions de l'effet canal court et du DIBL (noté $DIBL_{VDT}$):

$$SCE = 0.7 \frac{8\epsilon_{si}t_{si}}{C_{ox}L^2} (\phi_d - \phi_{sth}) \quad \text{Eq. II-78}$$

$$DIBL_{VDT} = \frac{4\epsilon_{si}t_{si}}{C_{ox}L^2} V_{ds} \quad \text{Eq. II-79}$$

Où 0.7 est un paramètre d'ajustement, déterminé par simulations numériques. Pour éprouver notre méthode, nous comparons les valeurs des paramètres SCE et DIBL obtenues par notre modèle analytique simple à des simulations numériques 2D [Synopsys].

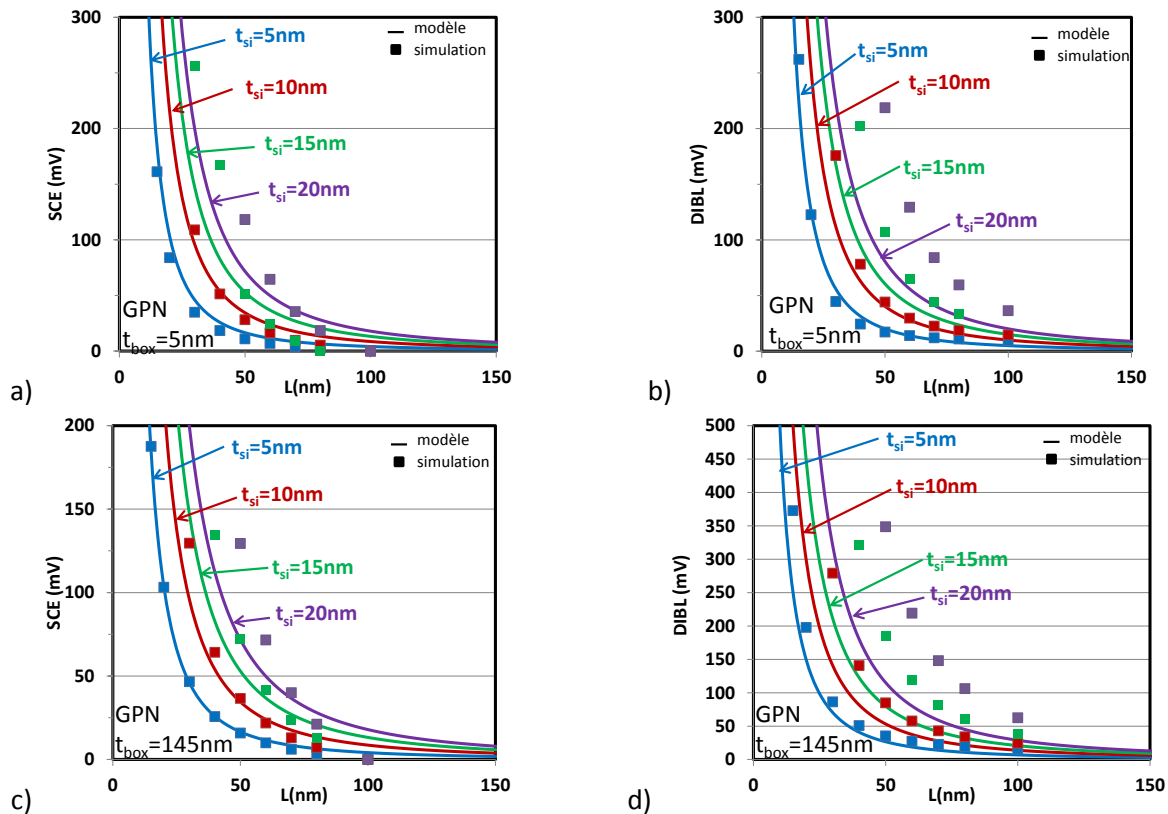


Figure II-21 : Tracé du SCE et du DIBL d'un NMOS avec un plan de masse de type N en fonction de la longueur de grille L pour une EOT de 1nm et pour différentes valeurs de t_{si} (5, 10, 15 et 20nm). a) SCE pour $t_{box}=5nm$, b) SCE pour $t_{box}=145nm$, c) DIBL pour $t_{box}=5nm$ et d) DIBL pour $t_{box}=145nm$

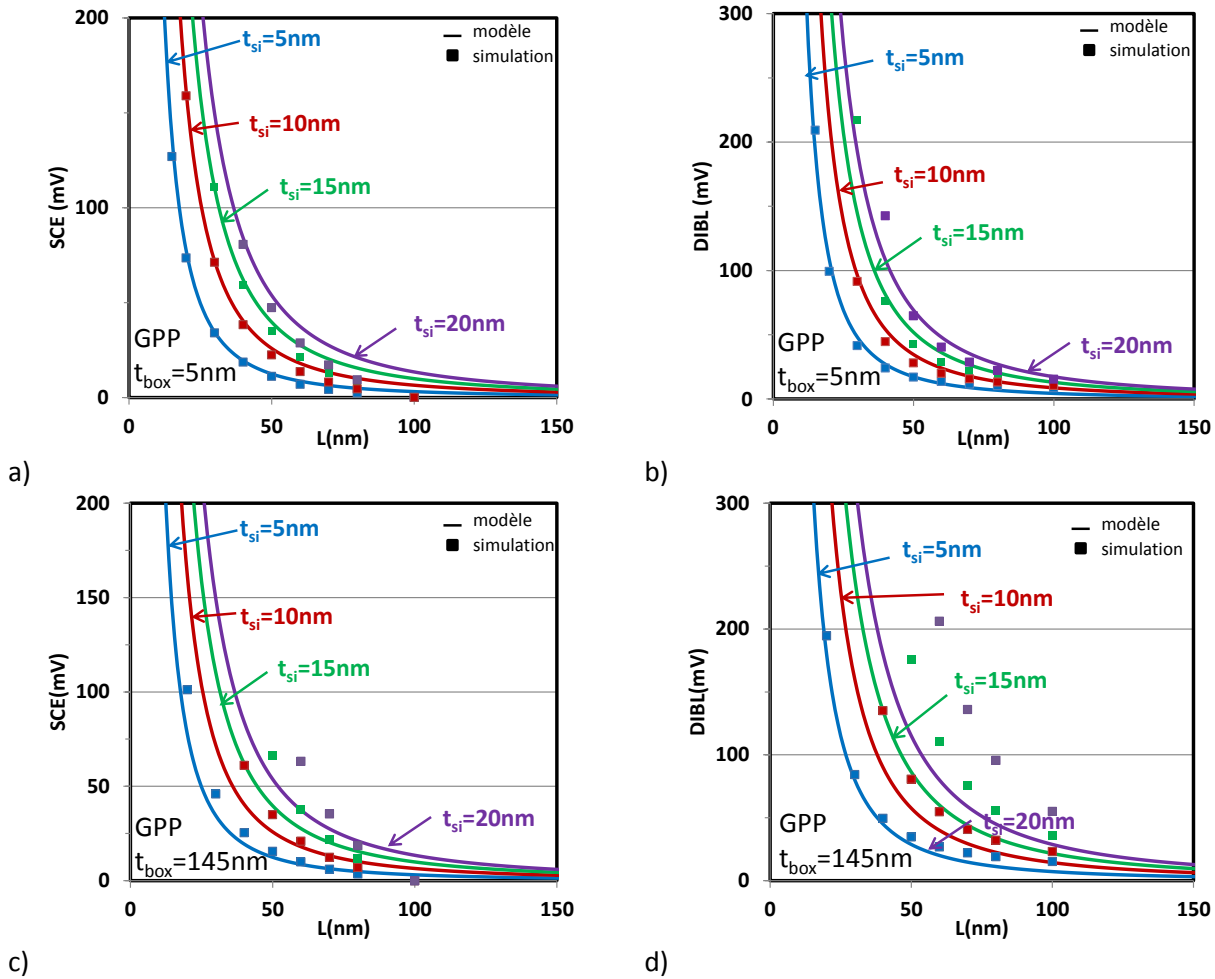


Figure II-22 : Tracé du SCE et du DIBL d'un NMOS avec un plan de masse de type P en fonction de la longueur de grille L pour une EOT de 1nm et pour différentes valeurs de t_{si} (5, 10, 15 et 20nm). a) SCE pour $t_{\text{box}} = 5\text{nm}$, b) SCE pour $t_{\text{box}} = 145\text{nm}$, c) DIBL pour $t_{\text{box}} = 5\text{nm}$ et d) DIBL pour $t_{\text{box}} = 5\text{nm}$

Sur les graphes de les Figure II-21-a et b (NMOS GPN) et Figure II-22-a et b (NMOS GPP), on remarque que le modèle reproduit bien le comportement du paramètre SCE en fonction de la longueur de grille pour différentes valeurs d'épaisseur de film de silicium, dans le cas d'une épaisseur d'oxyde enterré mince (a) et épaisse (b). Ceci valide donc notre modèle du paramètre SCE. Sur les Figure II-21-c et Figure II-23-c, on valide notre modèle de DIBL pour un oxyde enterré mince. Cependant, dans le cas d'un oxyde enterré épais (Figure II-21-d et Figure II-22-d), notre modèle de DIBL n'est plus valable, quel que soit le type du plan de masse.

Les simulations numériques effectuées par [Gallon 07] (Figure II-23), mettent en évidence l'existence d'un couplage électrostatique entre le canal et le drain, à travers l'oxyde enterré. Ce couplage qui est d'autant plus grand que l'oxyde enterré est épais, n'est pas modélisé par la VDT, ce qui explique le comportement de notre modèle pour les oxydes enterrés épais ($t_{\text{box}} > 25\text{nm}$).

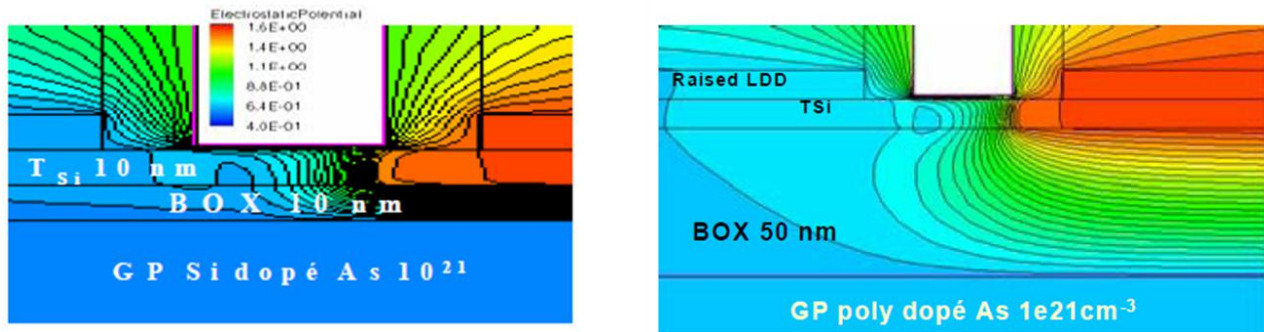


Figure II-23 : Simulations numériques représentant la cartographie des courbes iso-potentielles dans un transistor FDSOI à canal court pour deux épaisseurs d'oxyde enterré (10 et 50nm) extraites de [Gallon 07], mettant en évidence le couplage électrostatique entre le drain et le canal à travers l'oxyde enterré.

Ce couplage électrostatique a déjà été mis en évidence dans la littérature, notamment par [Ernst 07], et on parle de « Fringing Field ». Pour modéliser le DIBL, nous allons considérer qu'il se décompose en deux composantes indépendantes (Figure II-24) :

- La première, modélisant les effets canaux courts à travers le canal, sera modélisée par l'équation II-79.
- La seconde traduira l'effet du couplage électrostatique entre le drain et le canal à travers l'oxyde enterré et son effet sur la tension de seuil sera modélisé par une technique exposée par [Ernst 07]. Ce dernier propose de traduire ce couplage par une capacité.

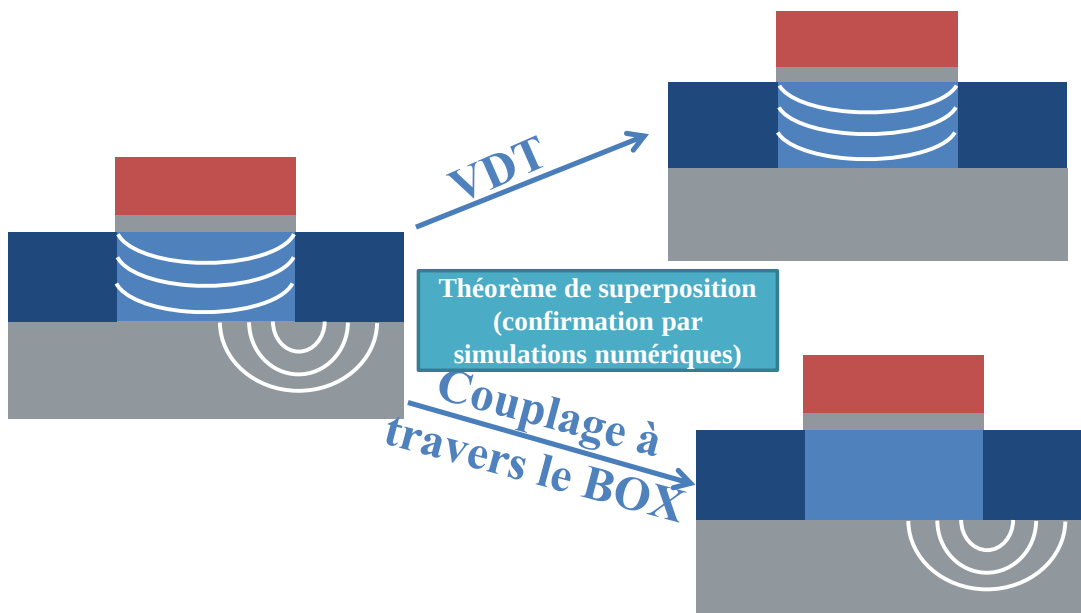


Figure II-24 : Structure équivalente considérée pour la modélisation du DIBL.

Cette capacité, formée par les lignes de champs électriques entre le drain et le canal, ne peut être calculée dans le repère cartésien classique. Le repère doit donc être transformé pour rendre une évaluation analytique possible. [Ernst 07] suggère d'utiliser la transformation de Schwarz-Christoffel, décrite dans [Durand 66], sur une structure simplifiée équivalente :

- Nous cherchons à modéliser une composante du DIBL, nous nous focaliserons donc sur le drain. La structure équivalente sera donc limitée à la moitié du transistor, côté drain.

- La composante du DIBL physiquement due au canal est modélisée par la VDT. Dans le cas du calcul de l'impact du couplage entre le drain et le canal à travers le BOX, nous considérerons que le canal n'aura pas d'effet et nous fixerons donc son potentiel à la masse dans notre structure équivalente.
- L'utilisation de plan de masse dopé confine les lignes de champs dans l'oxyde enterré et ce dernier n'aura que peu d'effet sur la forme des lignes de champs électriques. Nous considérerons donc que ce dernier est à la masse dans notre structure équivalente.

La structure équivalente ainsi obtenue est représentée sur la Figure II-25.

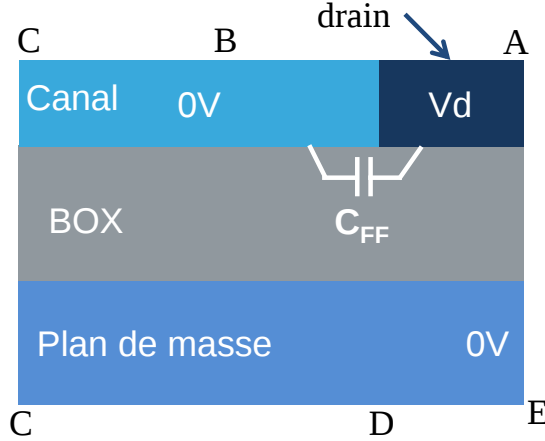


Figure II-25 : Structure équivalente utilisée pour le calcul de la composante du DIBL due au couplage entre drain et canal à travers l'oxyde enterré (BOX).

Appliquons maintenant la transformation de Schwarz-Christoffel à cette structure équivalente. Comme suggérée dans [Ernst 07], considérons le plan complexe cartésien initial (x,y) dont l'origine se situe à l'intersection du bas de l'oxyde enterré (abscisse) et du centre du canal (ordonnée) (Figure II-26). Le repère (ξ, η) obtenu par la transformation de Schwarz-Christoffel indiquée dans [Durand 66] est une bande circulaire infinie (zone bleue sur le schéma de droite de la Figure II-26) dans laquelle le potentiel (complexe) due à la polarisation d'une bande ($V=V_d$ entre A et B, cf Figure II-26) est connu et donné par [Durand 66]. Pour connaître les relations de passage entre les repères, on applique la définition de la fonction de transformation de Schwarz-Christoffel, donnée par [Durand 66] :

$$x + iy = \alpha \int \frac{1}{t} dt + \beta = \alpha \ln(\xi + i\eta) + \beta \quad \text{Eq. II-80}$$

Où α et β sont déterminées par les conditions aux limites :

- Au point B :

$$x + iy = it_{\text{box}} \quad \text{Eq. II-81}$$

$$\xi + i\eta = -1 = \exp(i\pi) \quad \text{Eq. II-82}$$

- Au point D:

$$x + iy = 0 \quad \text{Eq. II-83}$$

$$\xi + i\eta = 1 \quad \text{Eq. II-84}$$

En injectant les conditions aux limites énoncées ci-dessus dans l'équation II-80, on obtient :

$$\alpha = \frac{t_{\text{box}}}{\pi} \quad \text{et} \quad \beta = 0 \quad \text{Eq. II-85}$$

Les relations de passage entre les repères sont donc les suivantes :

$$x + iy = \frac{t_{box}}{\pi} \ln(\xi + i\eta) \quad \text{Eq. II-86}$$

$$\xi + i\eta = \exp\left(\frac{\pi}{t_{box}} (x + iy)\right) \quad \text{Eq. II-87}$$

Comme mentionné ci-dessus, le potentiel complexe W dans le plan complexe (ξ, η) est connu, son expression est donnée par [Durand 66] :

$$W(\xi + i\eta) = \frac{V_d}{i\pi} \ln(\xi + i\eta + 1) \quad \text{Eq. II-88}$$

En appliquant la fonction de transformation à l'expression du potentiel complexe dans le repère transformée, on obtient l'expression du potentiel complexe V dans le repère initial :

$$V(x + iy) = \frac{V_d}{i\pi} \ln\left(1 + \exp\left(\frac{\pi}{t_{box}} (x + iy)\right)\right) \quad \text{Eq. II-89}$$

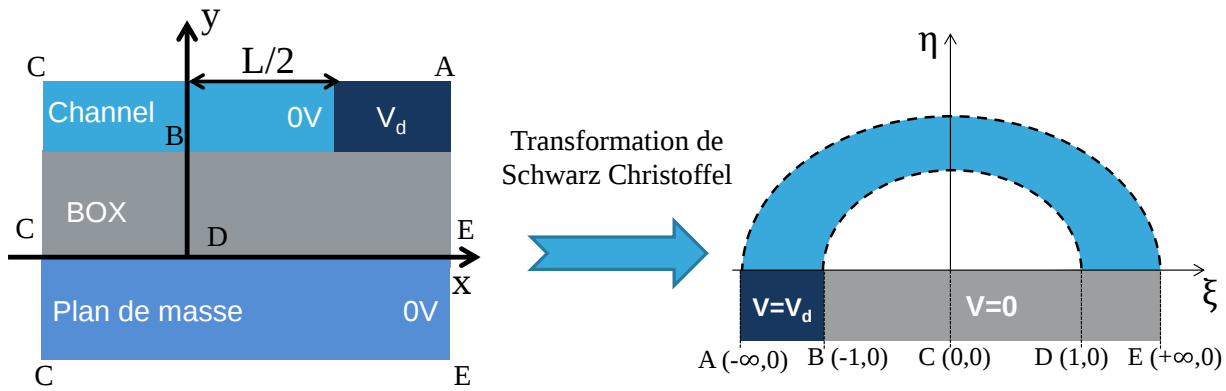


Figure II-26 : représentation schématique de la transformation de Schwarz-Christoffel utilisée pour la modélisation du couplage électrostatique entre le drain et le canal à travers l'oxyde enterré.

Pour aboutir à l'expression de la capacité de couplage C_{FF} , [Ernst 07] propose de prendre la partie réelle du potentiel complexe donnée par l'équation II-89, après y avoir appliquée une translation de $x=L/2$ pour se placer sur le drain. De cette manière, l'expression du potentiel de perturbation due à la polarisation de drain est obtenue. Ensuite, l'expression analytique du champ électrique est obtenue de manière analytique à partir de sa définition. Enfin, la capacité C_{FF} est obtenue par sa définition à partir du champ électrique et du potentiel :

$$C_{FF} = \epsilon \frac{dE_d}{dV_d} \quad \text{Eq. II-90}$$

Où ϵ est la permittivité de l'oxyde enterré, E_d le champ électrique obtenu à partir du potentiel de perturbation, et V_d la polarisation de drain. Les détails de ce calcul sont donnés dans [Ernst 07]. Finalement, on a :

$$C_{FF} = \frac{\epsilon_{ox}}{t_{boxeq} \left(\exp\left(\frac{\pi L}{2t_{boxeq}}\right) - 1 \right)} \quad \text{Eq. II-91}$$

Où t_{boxeq} est l'épaisseur d'oxyde électrique équivalente, tenant compte de la couche de déplétion dans le plan de masse donnée par l'équation II-74.

En remplaçant la capacité d'oxyde enterré par la capacité C_{FF} dans l'expression de la capacité du diviseur capacitif classique, on obtient l'expression du DIBL dû au couplage électrostatique entre drain et canal à travers l'oxyde enterré, noté $DIBL_{FF}$:

$$DIBL_{FF} = \frac{C_{FF}C_{si}}{(C_{ox}(C_{si}+C_{FF}))} V_d \quad \text{Eq. II-92}$$

Finalement, le DIBL total est donnée par la somme des deux composantes, l'une donnée par la VDT, et l'autre donnée par la transformation conforme de Schwarz-Christoffel :

$$DIBL = DIBL_{VDT} + \eta DIBL_{FF} \quad \text{Eq. II-93}$$

Où η est un paramètre d'ajustement, déterminé par simulations numériques, visant à amplifier l'impact du couplage capacitif sur le DIBL pour les films de silicium et oxydes enterrés minces dont l'expression est donnée par :

$$\eta = \frac{t_{si}}{5e^{-9}} \times \frac{t_{box}}{5e^{-9}} \quad \text{Eq. II-94}$$

Pour valider notre modèle plus complet de DIBL, nous avons procédé à des simulations numériques et tracé le DIBL en fonction de la longueur de grille pour un NMOS avec plan de masse de type N (Figure II-27) et de type P (Figure II-28).

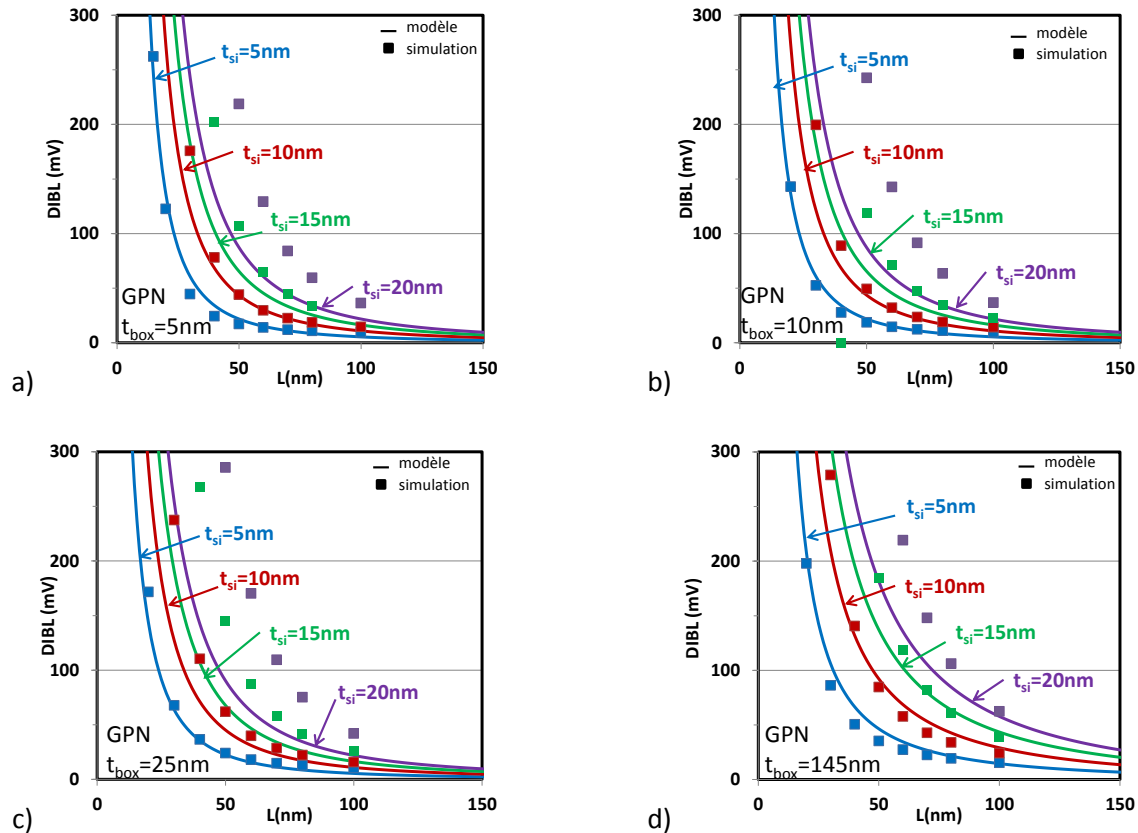


Figure II-27 : Tracé du DIBL d'un NMOS avec un plan de masse de type N en fonction de la longueur de grille L pour une EOT de 1nm et pour différentes valeurs de t_{si} (5, 10, 15 et 20nm). a) DIBL pour $t_{box}=5nm$, b) DIBL pour $t_{box}=10nm$, c) DIBL pour $t_{box}=25nm$ et d) DIBL pour $t_{box}=145nm$.

La modélisation du DIBL est améliorée pour les épaisseurs d'oxyde enterrée importante ($t_{\text{box}} > 25\text{nm}$). Cependant, le modèle n'est toujours pas très précis pour les épaisseurs de film de silicium importantes ($t_{\text{si}} > 15\text{nm}$). Ceci s'explique par notre approximation de potentiel constant dans le film de silicium, et égal à sa valeur en surface qui est vraie pour les films minces, mais devient inexacte lorsque le canal est plus épais.

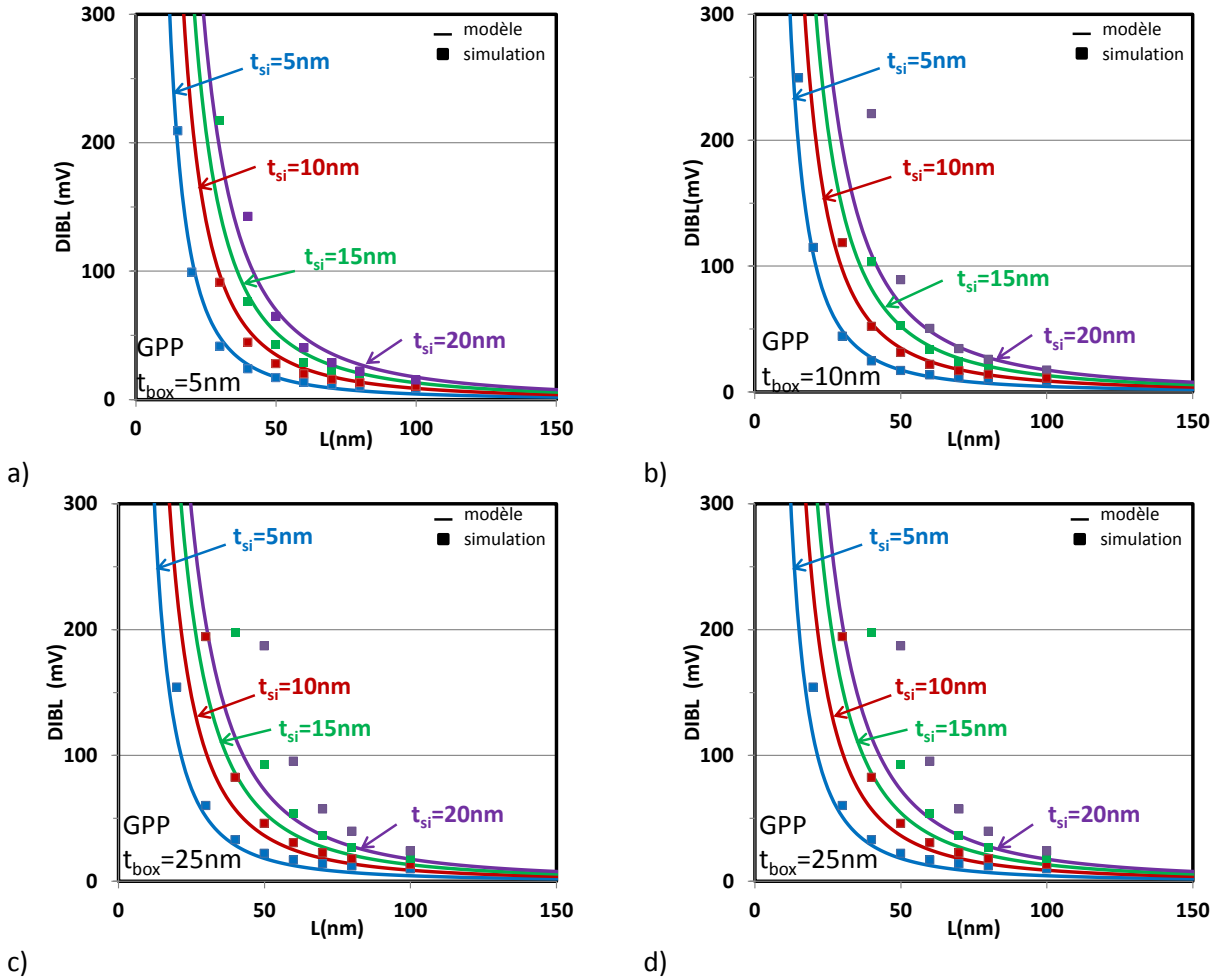


Figure II-28 : Tracé du DIBL d'un NMOS avec un plan de masse de type P en fonction de la longueur de grille L pour une EOT de 1nm et pour différentes valeurs de t_{si} (5, 10, 15 et 20nm). a) DIBL pour $t_{\text{box}} = 5\text{nm}$, b) DIBL pour $t_{\text{box}} = 10\text{nm}$, c) DIBL pour $t_{\text{box}} = 25\text{nm}$ et d) DIBL pour $t_{\text{box}} = 145\text{nm}$

II.B.4. Pente sous le seuil

Comme pour le transistor conventionnel sur substrat massif (II.A.4), nous allons commencer par écrire la loi de Gauss dans le canal. Mais, dans le cas d'un transistor FDSOI, le canal est complètement déplété et la charge de déplétion est négligeable car le dopage canal est faible et physiquement limitée par l'épaisseur du film t_{si} ($Q_{\text{dep}} = qN_{\text{ch}}t_{\text{si}}$). De plus, la pente sous le seuil est évaluée en régime de faible inversion, donc le terme de charge d'inversion est également nul. Ce qui nous mène à une pente sous le seuil constante, égale à sa valeur idéale de 60mV/dec [Sktonicki 08]. Cependant, il est bien connu que la pente sous le seuil n'est, même dans un dispositif FDSOI, pas constante et sera dégradée quand la longueur de grille sera réduite. Pour modéliser cette dégradation, nous allons utiliser de nouveau la VDT qui va traduire la réduction de la barrière de potentiel entre source et drain

due à la réduction de longueur de grille et à la polarisation de drain par une modification de dopage canal effectif, et donc de charge de déplétion effective. Formellement, cela donne:

$$Q_{dep}^* = qN_{ch}t_{si} - 2\frac{t_{si}\epsilon_{si}}{L^2}(V_{DS} + 2(\varphi_d - \varphi_s) + 2\sqrt{(\varphi_d - \varphi_s)(V_{DS} + \varphi_d - \varphi_s)}) \quad \text{Eq. II-95}$$

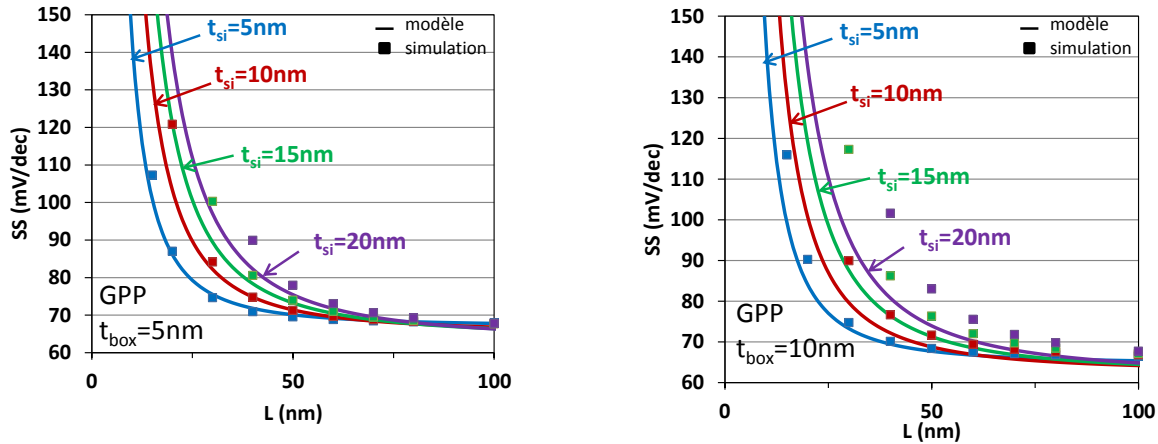
En écrivant la définition de la pente sous le seuil, et en remplaçant le terme de charge de déplétion par celui de déplétion effective donnée par l'équation II-95, on a :

$$SS = \frac{kT}{q} \ln(10) \frac{dV_g}{d\varphi_s} = \frac{kT}{q} \ln(10) \left(1 + \frac{1}{C_{ox}} \left(\frac{dQ_{dep}^*}{d\varphi_s} \right) \right) \quad \text{Eq. II-96}$$

En dérivant l'équation II-96, on obtient l'expression analytique de la pente sous le seuil :

$$SS = \frac{kT}{q} \ln(10) \left(1 + \frac{2t_{si}\epsilon_{si}}{C_{ox}L^2} \left(2 + \sqrt{\frac{\varphi_d - \varphi}{V_{DS} + \varphi_d - \varphi}} + \sqrt{\frac{V_{DS} + \varphi_d - \varphi}{\varphi_d - \varphi}} \right) \right) \quad \text{Eq. II-97}$$

Où φ est le potentiel de surface que nous choisissons de fixer à $\varphi = 0.75\varphi_{sth}$ pour garantir que notre expression est évaluée en régime de faible inversion. La Figure II-29 montre que notre modèle reproduit bien les simulations numériques 2D, pour les oxydes enterrés minces ($t_{box} < 10\text{nm}$).



a) b)
Figure II-29 : Tracé de la pente sous le seuil ne fonction de la longueur de grille pour un NMOS avec un plan de masse de type P, pour différentes épaisseurs de film de silicium t_{si} et d'oxyde enterré t_{box} . a) $t_{box}=5\text{nm}$ et b) $t_{box}=10\text{nm}$.

Cependant, pour les oxydes enterrés plus épais, notre modèle analytique n'est plus précis mais donne tout de même une première approximation (Figure II-30).

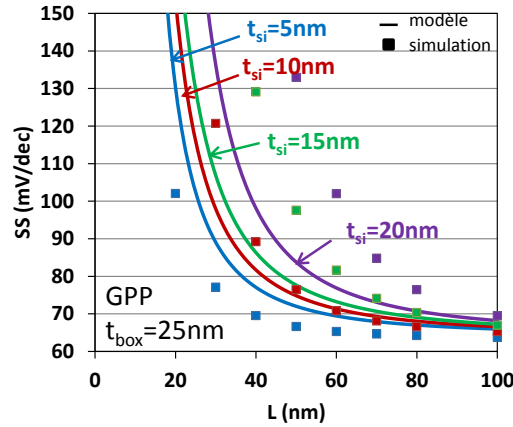
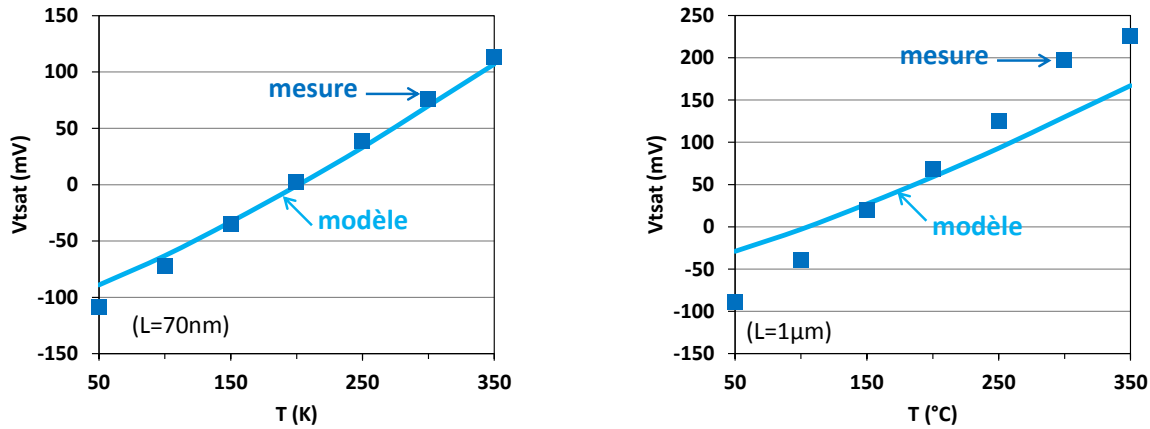


Figure II-30 : Tracé de la pente sous le seuil en fonction de la longueur de grille pour un NMOS avec un plan de masse de type P pour différentes épaisseurs de film de silicium t_{si} pour une épaisseur d'oxyde enterré $t_{box}=25nm$.

II.B.5. Validation du comportement en température :

Comme pour l'architecture sur substrat massif, il faut valider le comportement en température (II.A.6) de ce modèle. Pour le valider, nous avons effectué des mesures sur un transistor de longueur 70nm, puis un de longueur 1 μm typique du nœud technologie CMOS 28nm pour différentes températures (de 50 à 350K). Notre modèle de tension reproduit bien les mesures effectuées sur le transistor de longueur $L=70nm$ (Figure II-31-a), mais est moins précis pour le transistor à canal long $L=1\mu m$ (Figure II-31-b), et le comportement en température de ce modèle est donc bien validé.



a)

b)

Figure II-31 : Tracé de la tension de seuil en fonction de la température obtenu par mesure de transistors typiques du nœud technologique 28nm. Notre modèle reproduit les mesures effectuées sur un transistor de longueur 70nm a) et est un peu moins précis pour un transistor FDSOI à canal long $L=1\mu m$ (b).

II.C. Tension de seuil et électrostatique de l'architecture Double grille faiblement dopée.

II.C.1. Critère d'inversion

Comme pour les transistors FDSOI (II.B.1), le critère de seuil usuel de l'architecture conventionnelle sur substrat massif $\phi_s = 2\phi_f$ ne peut pas être utilisé dans la modélisation de la tension de seuil d'un transistor double grille faiblement dopé. Pour les mêmes raisons que dans le paragraphe II.B.1, nous choisissons d'utiliser comme critère de seuil le point d'inflexion sur la courbe ($C_g(V)$) ce qui nous donne une valeur de charge d'inversion au seuil donnée par l'équation II-53.

II.C.2. Tension de seuil canal long

Pour modéliser la tension de seuil d'un transistor double grille à canal long (Figure II-32) on commence par écrire la loi de Gauss dans la moitié du film, c'est-à-dire de $x=0$ à $x=t_{si}/2$:

$$qn_i \frac{t_{si}}{2} e^{\left(\frac{q\phi_s}{kT}\right)} + qN_{ch} \frac{t_{si}}{2} = C_{ox}(V_g - V_{fb} - \phi_f - \phi_s) \quad \text{Eq. II-98}$$

où n_i est la concentration intrinsèque de porteurs, ϕ_f le potentiel de Fermi et V_{fb} la tension de bande plate et les autres paramètres sont donnés par la Figure II-32. Le premier terme de l'équation II-98 représente la charge d'inversion et le second la charge de déplétion.

Enfin, en réinjectant dans la loi de Gauss de l'équation II-98, on obtient l'équation de la tension de seuil canal long pour un transistor double grille :

$$V_{tlong} = V_{fb} + \phi_f + \frac{2Q_{ith} + qN_{ch}t_{si}}{2C_{ox}} + \phi_{sth} \quad \text{Eq. II-99}$$

Où le potentiel de surface au seuil est donné par l'équation II-55. Ce modèle est ensuite validé par simulation numérique 1D effectuée avec l'outil FlexPDE [FlexPDE], en ne résolvant que l'équation de Poisson, donc en négligeant les effets quantiques.

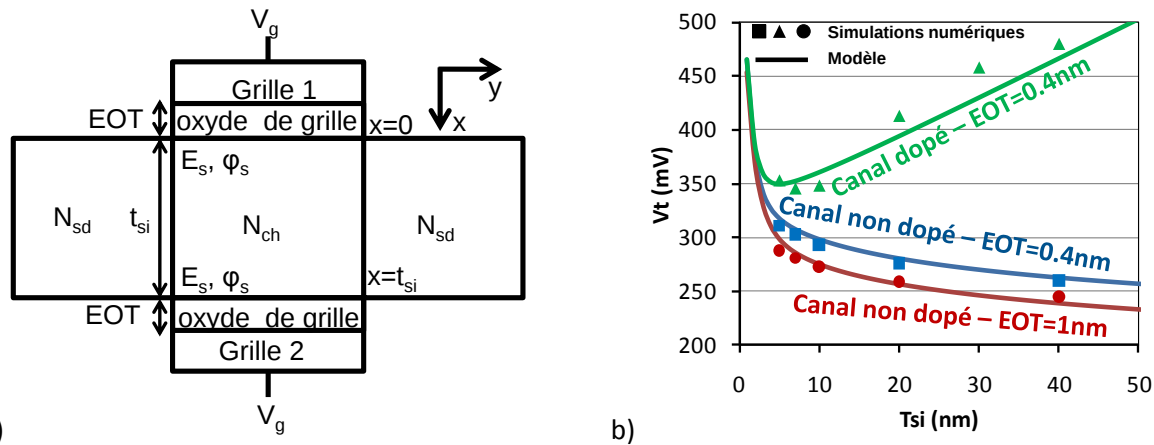


Figure II-32 : (a) schéma et dimension d'un transistor double grille. (b) Evolution de la tension de seuil canal long avec l'épaisseur du film de silicium ; les symboles représentent les valeurs de V_t obtenues par simulation numérique, et les traits pleins les valeurs obtenues avec l'équation II-99.

La Figure II-32-b montre la variation de tension de seuil canal long (V_{tlong}) avec l'épaisseur du film de silicium (t_{si}) pour deux valeurs d'EOT et pour des canaux faiblement dopés. On constate que le modèle reproduit très bien les

simulations numériques dans le cas faiblement dopés et donne une bonne approximation quand le niveau de dopage augmente.

Cependant, négliger les effets de confinement quantique pour des transistors double grille avec un film de silicium d'épaisseur inférieure à 10nm mène à des erreurs croissantes avec la réduction d'épaisseur de canal [Omura 93] [Wong 98]. On trouve dans la littérature différentes manières de corriger la tension de seuil canal long du double grille pour prendre en compte ces effets. Néanmoins, les formulations sont souvent compliquées, ou nécessitent des résolutions numériques [Munteanu 05] [Baccarani 99]. Dans ce modèle, nous choisissons une manière plus pragmatique d'aborder le problème. En effet, dans la théorie classique, la distribution de porteurs minoritaires atteint son maximum à l'interface oxyde de grille/silicium alors que, dans la théorie quantique, le maximum de la distribution est à quelques angströms de l'interface, cette distance dépendant du matériau (~1.2nm pour le silicium) et on parle alors de darkspace (cf I.C.3.b). Dans le transistor conventionnel sur substrat massif, son effet est traduit par une épaisseur d'oxyde électrique (t_{inv}) en inversion plus grande que l'EOT (de 4 Å en équivalent SiO₂). Or dans le transistor double grille, comme le canal est limité par l'épaisseur du film, le confinement quantique a également un effet sur la charge d'inversion (égale à l'intégrale de la distribution de porteurs minoritaires), qui, pour une même polarisation sera plus faible que dans le cas classique. Cela est illustré sur la Figure II-33.

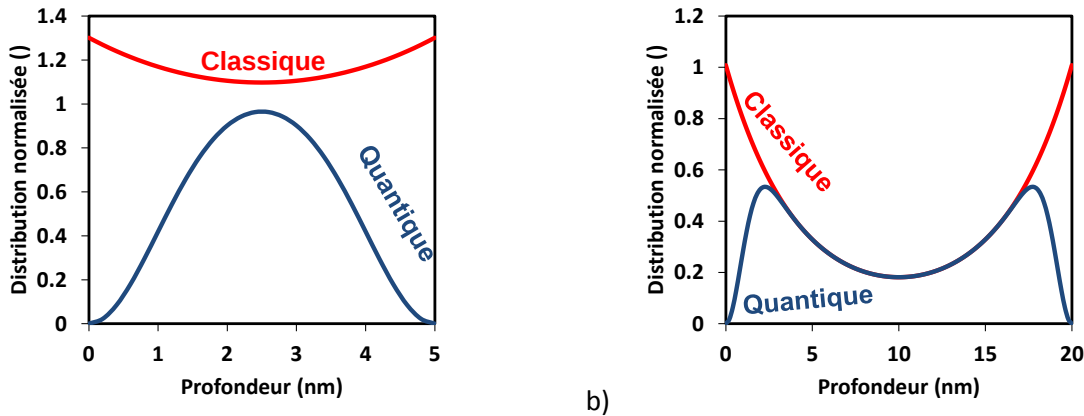


Figure II-33 : distribution de porteurs minoritaires dans un transistor double grille pour deux épaisseurs de canaux : $t_{si}=5\text{nm}$ (a) $t_{si}=20\text{nm}$ (b).

Pour prendre en compte cet effet, nous proposons d'introduire λ la « longueur quantique » (quantum length en anglais), donnée par [Hänsch 89] [Rios 94] qui va nous permettre, à partir de la formulation classique de la distribution de porteurs minoritaires en fonction de la position dans le film de silicium ($n_{class}(x)$), obtenir la formulation quantique ($n_q(x)$):

$$n_{class}(x) = n_0 \exp\left(\frac{q\phi(x)}{kT}\right) \quad \text{Eq. II-100}$$

$$n_q(x) = n_{class}(x) \cdot \left(1 - \exp\left(-\frac{x^2}{\lambda^2}\right)\right) \cdot \left(1 - \exp\left(-\frac{(x-t_{si})^2}{\lambda^2}\right)\right) \quad \text{Eq. II-101}$$

Où le second et le troisième terme de l'équation II-101 sont respectivement la correction quantique pour la grille du haut ($x=0$) et celle du bas ($x=t_{si}$). Pour permettre des expressions complètement analytique, on suppose le champ électrique constant, noté F , la distribution classique devient :

$$n_{class}(x) = n_0 \left(\exp\left(\frac{-qF(x)}{kT}\right) + \exp\left(\frac{-qF(t_{si}-x)}{kT}\right) \right) \quad \text{Eq. II-102}$$

Les expressions de la charge d'inversion en prenant en compte ou non les effets de confinements quantiques sont obtenues en intégrant dans le film de silicium (de $x=0$ à $x=t_{si}$) les expressions des distributions de porteurs, à savoir l'équation II-102 pour la charge d'inversion « classique » (Q_{iclass}) et II-101 pour la charge d'inversion « quantique » (Q_{iq}). Pour être utilisable dans notre modèle de tension de seuil, nous définissons un facteur de correction, noté $CF(t_{si})$, égal au ratio des deux formulations de la charge d'inversion :

$$CF(t_{si}) = \frac{Q_{iq}}{Q_{iclass}} = \frac{\int_0^{t_{si}} n_0 \left(\exp\left(\frac{-qFx}{kT}\right) + \exp\left(\frac{-qF(t_{si}-x)}{kT}\right) \right) \left(1 - \exp\left(\frac{-x^2}{\lambda^2}\right) \right) \left(1 - \exp\left(\frac{-(x-t_{si})^2}{\lambda^2}\right) \right) dx}{\int_0^{t_{si}} n_0 \left(\exp\left(\frac{-qFx}{kT}\right) + \exp\left(\frac{-qF(t_{si}-x)}{kT}\right) \right) dx} \quad \text{Eq. II-103}$$

En se plaçant au seuil, la champ électrique est connu et vaut $2Q_{ith}/\epsilon_{si}$. On peut alors intégrer analytiquement l'équation II-103 :

$$CF(t_{si}) = 1 - \frac{(a(t_{si}, t_{si}) - a(0, t_{si}) - b(t_{si}, t_{si}) + b(0, t_{si}) + c(t_{si}, t_{si}) - c(0, t_{si}))}{d(t_{si}) - d(0)} \quad \text{Eq. II-104}$$

$$a(x, t_{si}) = e^{-\left[\left(\frac{t_{si}}{\lambda}\right)^2 + \left(\frac{qF\lambda}{2kT} - \frac{t_{si}}{\lambda}\right)^2\right]} \frac{\lambda\sqrt{\pi}}{2} \operatorname{erf}\left(\frac{\lambda F}{2kT} - \frac{t_{si}}{\lambda} + \frac{x}{\lambda}\right) \quad \text{Eq. II-105}$$

$$b(x, t_{si}) = e^{-\left[\left(\frac{t_{si}}{\lambda}\right)^2 + \left(\frac{qF\lambda}{2\sqrt{2}kT} - \frac{t_{si}}{\lambda}\right)^2\right]} \frac{\lambda\sqrt{\pi}}{2\sqrt{2}} \operatorname{erf}\left(\frac{\lambda F}{2\sqrt{2}kT} - \frac{t_{si}}{\sqrt{2}\lambda} + \frac{\sqrt{2}x}{\lambda}\right) \quad \text{Eq. II-106}$$

$$c(x, t_{si}) = \frac{\lambda\sqrt{\pi}}{2} \operatorname{erf}\left(\frac{q\lambda F}{2kT} + \frac{x}{\lambda}\right) e^{\left(\frac{qF\lambda}{2kT}\right)^2} \quad \text{Eq. II-107}$$

$$d(x, t_{si}) = e^{\left(\frac{-qFx}{kT}\right)} \frac{kT}{qF} \quad \text{Eq. II-108}$$

Où $\operatorname{erf}(x)$ est la fonction erreur, usuellement définie par :

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz \quad \text{Eq. II-109}$$

L'approximation de champ électrique constant dans le film est valable pour les films minces, mais devient fausse quand le film s'élargit. Pour assurer la validité de notre expression du facteur de correction, nous le normalisons par sa valeur pour un film épais (100nm), là où les effets quantiques sont négligeables. L'allure de notre facteur de correction est représentée sur la Figure II-34-a.

Enfin, pour inclure cette correction quantique dans notre expression de tension de seuil canal long (Eq. II-99), nous recherchons la valeur du potentiel de surface (ϕ_{sthq}) pour laquelle notre charge d'inversion, corrigée par le facteur de correction de l'équation II-104, atteint la valeur donnée par notre critère de seuil (Eq. II-53), qui est naturellement le même avec et sans prise en compte des effets quantiques. On a alors :

$$CF(t_{si}) q n_i \frac{t_{si}}{2} e^{\left(\frac{q\phi_{sthq}}{kT}\right)} = Q_{ith} \quad \text{Eq. II-110}$$

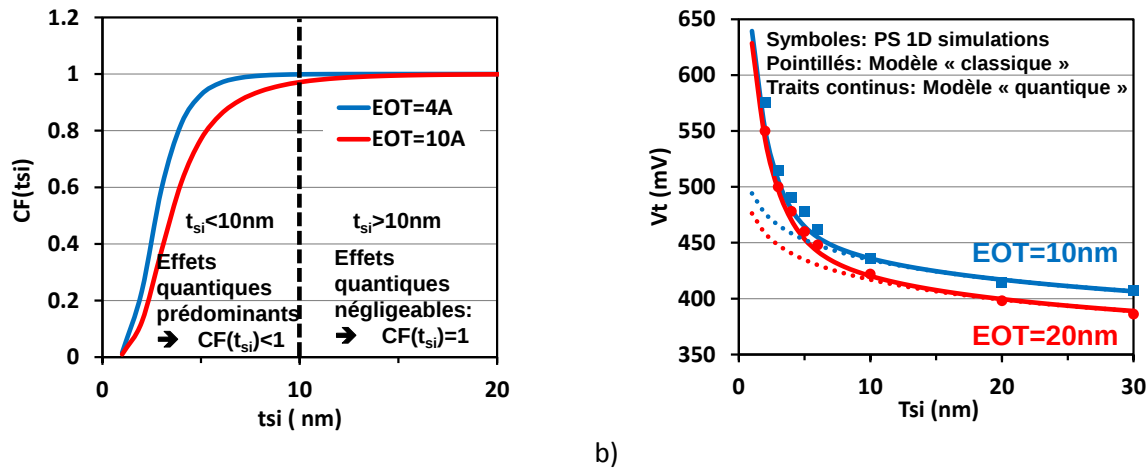
D'où l'expression du potentiel de surface au seuil tenant compte du confinement quantique :

$$\phi_{sthq} = \frac{kT}{q} \ln \left(\frac{\frac{kT}{q} C_{ox}}{CF(t_{si}) q n_i t_{si}} \right) \quad \text{Eq. II-111}$$

Enfin, en réinjectant l'équation II-111 dans l'expression de tension de seuil canal long classique (Eq. II-99), on obtient l'expression de la tension de seuil canal long avec prise en compte des effets de confinement quantique :

$$V_{t_{long}} = V_{fb} + \varphi_f + \frac{2Q_{ith} + qN_{ch}t_{si}}{2C_{ox}} + \varphi_{sthq} \quad \text{Eq. II-112}$$

Pour valider notre approche, nous avons procédé à des simulations Poisson-Schrödinger 1D basée sur les travaux de [Garetto 10] et comparé les résultats obtenus par l'expression analytique (II-112) sur la Figure II-34-b. On peut y voir que notre modèle reproduit bien les simulations, en fixant $\lambda=2\text{nm}$. On peut également constater que négliger les effets quantiques (courbes pointillées sur la Figure II-34-b pour les films minces mènent à de grandes erreurs.



a) b)
Figure II-34 : (a) Allure du facteur de correction pour deux valeurs d'EOT. (b) Comparaison du modèle de tension de seuil canal long tenant compte des effets quantiques avec les simulations Poisson-Schrödinger 1D. Les pointillés représentent le modèle de tension de seuil classique.

II.C.3. Effets canaux courts (SCE) et DIBL

Afin de modéliser les effets canaux courts, nous allons à nouveau recourir à la VDT [Skotnicki 88]. Comme pour le modèle de transistor conventionnel sur substrat massif (II.A.3), la première étape est d'obtenir une expression adéquate du dopage effectif N_{ch}^* , et plus précisément comment estimer le potentiel le long de la cathode virtuelle $\varphi_{vc}(x)$. Comme il a été montré dans [Munteanu 05] que la variation de potentiel entre la surface (i.e l'interface silicium/oxyde de grille, en $x=0$ ou $x=t_{si}$) et le centre du film est faible au seuil, nous choisissons de le considérer constant. Ensuite, bien qu'il ait été montré que le chemin de conduction en faible inversion (conducting path) dans [Tiwari 10] est situé à une distance de l'interface égale à environ $0.1 \sim 0.3x_{t_{si}}$, nous choisissons de considérer le potentiel constant égal à sa valeur en surface, à savoir au seuil à φ_{sth} (Eq. II-55). Enfin, comme dans le modèle pour le transistor conventionnel sur substrat massif, nous considérons la longueur des lignes de courant ($L(x)$) égale à la longueur électrique tant qu'on se situe entre les jonctions (i.e $x < X_j$), c'est-à-dire quel que soit x dans un dispositif double grille. L'expression du dopage effectif est alors très simple :

$$N_{ch}^* = N_{ch} - 2 \frac{\varepsilon_{si}}{qL^2} (V_{DS} + 2(\varphi_d - \varphi_{sth}) + 2\sqrt{(\varphi_d - \varphi_{sth})(V_{DS} + \varphi_d - \varphi_{sth})}) \quad \text{Eq. II-113}$$

Pour simplifier l'expression du dopage effectif, nous nous permettons un développement de Taylor de la racine de l'équation II-113 :

$$N_{ch}^* = N_{ch} - 2 \frac{\varepsilon_{si}}{qL^2} (2V_{DS} + 4(\varphi_d - \varphi_{sth})) \quad \text{Eq. II-114}$$

On peut alors calculer la charge de déplétion effective pour un transistor double grille, en considérant que la profondeur de déplétion est limitée par l'épaisseur du film de silicium t_{si} :

$$Q_{dep}^* = q N_{ch} * t_{si} = q N_{ch} t_{si} - 2 \epsilon_{si} \frac{2 V_{DS} + 4 (\varphi_d - \varphi_{sth})}{L^2} t_{si} \quad \text{Eq. II-115}$$

En réinjectant dans l'équation de tension de seuil canal long (Eq II-99) et par définition des paramètres DIBL et SCE (Eq II-20 et II-24), on obtient :

$$SCE = \frac{\epsilon_{si}}{\epsilon_{ox}} t_{si} t_{ox} \frac{4 (\varphi_d - \varphi_{sth})}{L^2} \quad \text{Eq. II-116}$$

$$DIBL = \frac{\epsilon_{si}}{\epsilon_{ox}} t_{si} t_{ox} \frac{2 V_{DS}}{L^2} \quad \text{Eq. II-117}$$

Pour valider ces expressions, nous avons effectué des simulations numériques 2D avec l'outil Synopsys [Synopsys]. La Figure II-35 montre que les expressions II-116 et II-117 reproduisent bien les simulations numériques, et que notre modèle est bien valide, pour deux valeurs d'EOT distinctes et trois valeurs d'épaisseur de film de silicium t_{si} en fonction de la longueur de grille.

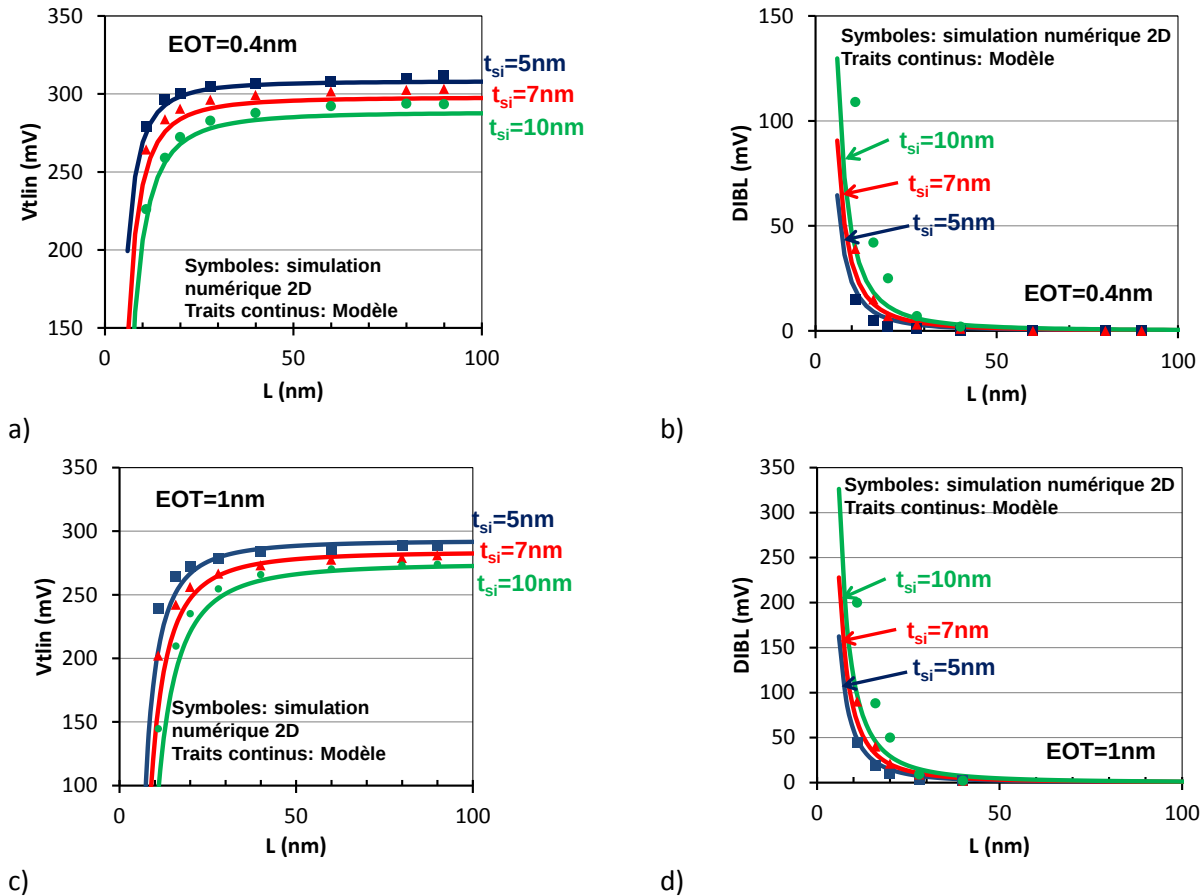
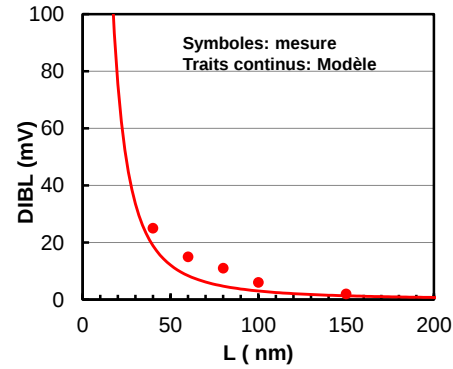
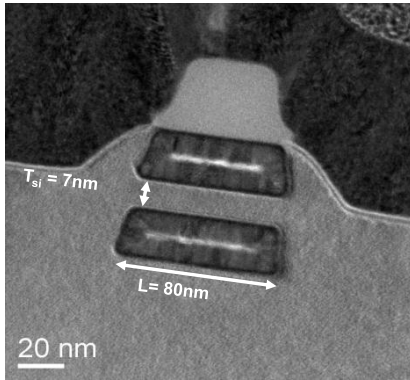


Figure II-35 : (a,c) Tension de seuil en régime linéaire (V_{tlin} extrait à $V_d=0.1V$, donc image du paramètre SCE) et (b,d) DIBL ($V_d=V_{dd}=0.7V$) en fonction de la longueur de grille pour des transistors double grille non dopés et pour deux valeurs d'EOT.

Pour valider définitivement nos expressions, nous les comparons aux mesures effectuées par [Huguenin 10] sur des dispositifs double grille non dopé (Figure II-36). Ce dernier a été capable d'extraire une courbe DIBL(L) pour un dispositif présentant un canal de 7nm et une EOT de 2.2nm, que nous comparons à l'expression analytique donnée par l'Eq. II-117.



a) b)
Figure II-36 : (a) Vue en coupe TEM (Microscopie Electronique à Transmission) d'un des transistors double grille mesuré par [Huguenin 10]. (b) Comparaison des courbes DIBL(L) obtenues par mesures et par notre équation analytique II-117 à $V_d=1V$.

II.C.4. Pente sous le seuil

Pour obtenir une expression de la pente sous le seuil dans un dispositif double grille, nous procéderons comme pour le transistor FDSOI, et nous aboutirons à la même expression, donnée par l'équation II-97. La Figure II-37 montre que notre modèle reproduit bien les simulations numériques 2D.

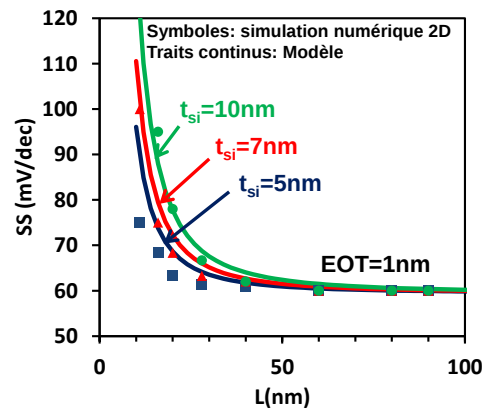
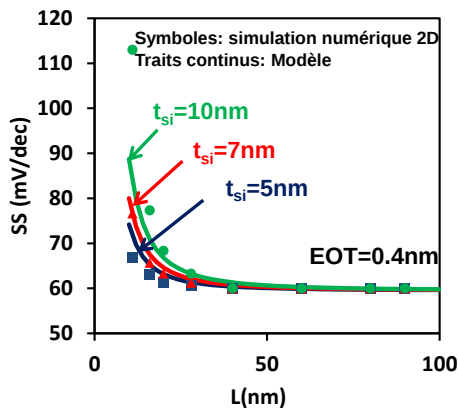


Figure II-37 : Pente sous le seuil en fonction de la longueur de grille pour trois épaisseurs de film de silicium et deux valeurs d'EOT. Les symboles représentent les simulations numériques et les traits continus représentent l'expression analytique II-75.

II.C.5. Validation du comportement en température :

Ne disposant pas de mesure de tension de seuil en fonction de la température pour l'architecture double grille, nous n'avons pas pu valider le comportement de notre modèle en fonction de la température.

II.D. Modélisation du courant de drain

Grace aux modèles électrostatiques proposés dans les trois parties précédentes, il est aisé de prévoir les valeurs de fuite (I_{off}) et de courant débité (I_{on}) à l'aide de la théorie classique de courant de dérive diffusion (drift diffusion en anglais) (I.B.2) pour chaque architecture. Cependant, un raccord propre entre les régimes linéaire et saturé (sur les I_d - V_d) est nécessaire pour une bonne estimation du courant effectif (I_{eff}) et donc pour avoir une première idée du délai. Enfin, notre but final étant d'utiliser ces modèles dans un simulateur de circuit conventionnel (comme ELDO [ELDO]), une continuité parfaite est indispensable entre les régimes de diffusion et de dérive (sur les I_d - V_g) mais également pour les dérivées de ce courant par rapport aux polarisations de grille et de drain, respectivement nommées g_m et g_d . Le modèle proposé dans cette partie répond à cette problématique

II.D.1.Méthodologie

Dans la littérature, il existe deux grandes familles de modèle compact fournissant des caractéristiques courant-tension de transistor MOSFETs : les modèles en potentiel de surface, ou $Q(\psi)$ (type PSP [PSP]) et les modèles basés sur la tension de seuil et le raccord des différents régimes, (type BSIM [BSIM]). La première catégorie, utilisée par [Khakifirooz 09] et [Wei 12] permet d'obtenir les caractéristiques courant-tension ainsi que leur dérivées continues qui reproduisent des mesures silicium publiées (Figure II-38). Dans cette méthode, les $I(V)$ sont construites à partir du courant I_{off} , qui doit être connu ou évalué d'une autre manière, de la pente sous le seuil, du DIBL et de la vitesse de saturation et ne nécessite donc pas de connaître la tension de seuil. Il en résulte que ce modèle ne peut reproduire directement l'impact de certains paramètres technologiques, comme le travail de sortie de la grille par exemple et ne peut donc pas être utilisé comme un outil d'ingénierie inverse (reverse engineering). De plus, dans ce modèle, le courant débité est peu dépendant de la mobilité, mais très dépendant de la vitesse de saturation, ce qui rend la prise en compte de l'effet des contraintes mécaniques sur le courant plus complexe. Par conséquent, il semble difficile d'utiliser cette méthode pour des études de prédiction de performance.

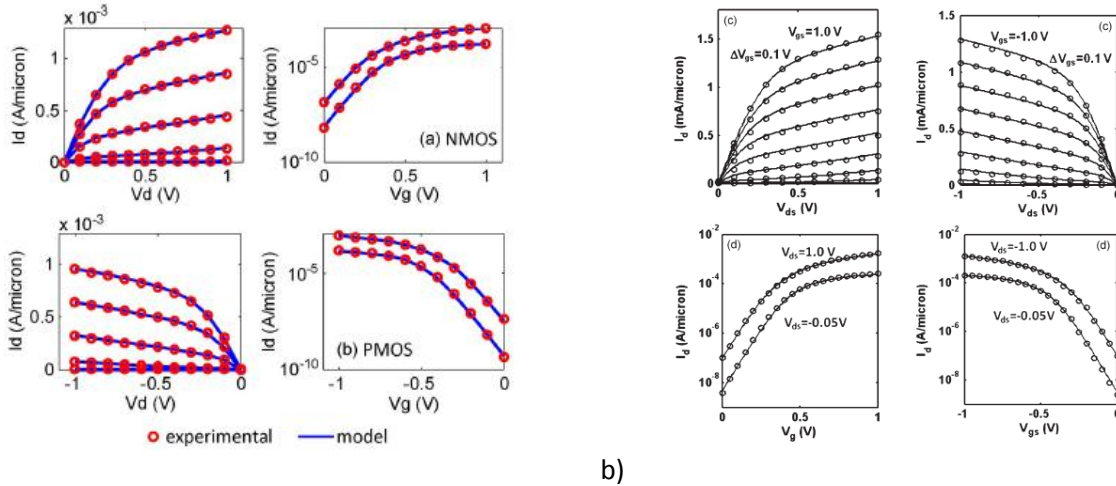


Figure II-38 : (a) figure extraite de [Wei 12] montrant que le modèle reproduit la technologie intel 45HP. (b) figure extraite de [Khakifirooz 09] montrant que le modèle reproduit la technologie intel 32HP

La seconde méthode, utilisée entre autre par [Skotnicki 94], permet d'obtenir les $I(V)$ à partir de la tension de seuil. Au seuil, donc à $V_g = V_t$, le courant est connu et fixé à $10^{-7} \times W/L$ (valeur en ampère, qui peut varier en fonction des nœuds technologiques). Au-delà du seuil, la théorie classique du courant de dérive est utilisée et des équations dépendant de la mobilité sont obtenues. En dessous du seuil, la théorie classique du courant de diffusion est utilisée et on obtient des expressions dépendant de la pente sous le seuil. Ce jeu d'équations est ensuite mathématiquement raccordé à l'aide de fonctions de continuité. On obtient alors les caractéristiques

courant-tension uniquement à partir de la description technologique du transistor (i.e. dimensions, permittivité, dopage...). Cette méthode permet d'obtenir un modèle compact et prédictif, donc idéal pour notre objectif final qui est de faire de la prédiction de performance. Néanmoins, le modèle fournit par [Skotnicki 94] n'assure pas une continuité parfaite, notamment sur les I_d-V_g autour du seuil, et ne peut donc pas être utilisé en tant que tel dans des simulateurs de circuits comme ELDO.

Le modèle proposé dans ce qui suit conserve la même philosophie que [Skotnicki 94] mais en assurant cette fois un parfaite continuité des caractéristiques courant-tension et de leurs dérivées. Le flot de modélisation utilisé est illustré sur la Figure II-39.

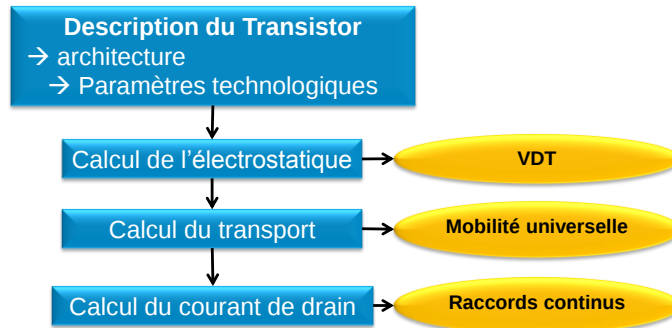


Figure II-39 : flot de modélisation pour la construction des caractéristiques courant-tension.

Grace aux modèles électrostatiques développés pour chaque architecture CMOS (II-A, II-B et II-C), nous sommes capables d'évaluer la tension de seuil, le DIBL et la pente sous le seuil de chaque architecture. Concernant le transport, nous utiliserons les équations de mobilité universelle, données par [Takagi 94], car chaque dispositif utilise un canal silicium et l'impact des contraintes sur la mobilité sera évalué à l'aide de travaux comme [Payet 08] et traduit par un facteur multiplicatif de la mobilité. Enfin, la continuité sera assurée par des fonctions de continuités, décrites ci-dessous.

II.D.2. Transport

Pour obtenir la valeur de la mobilité pour chaque valeur de polarisation, nous utilisons les formules classiques reproduisant les courbes de mobilité universelle, donnée par [Takagi 94]. Comme indiqué sur la Figure II-40-a, la mobilité effective se décompose en trois termes :

- μ_{sr} : (sr pour Surface Roughness) : régime d'interaction des porteurs avec les rugosités de surface

$$\text{NMOS:} \quad \mu_{sr} = 600. E_{eff}^{-3} \quad \text{Eq. II-118}$$

$$\text{PMOS:} \quad \mu_{sr} = 140. E_{eff}^{-1} \quad \text{Eq. II-119}$$

- μ_{ac} : (ac pour Acoustic Phonons) : régime d'interaction des porteurs avec les phonons acoustiques.

$$\text{NMOS:} \quad \mu_{ac} = 460. E_{eff}^{-0.3} \quad \text{Eq. II-120}$$

$$\text{PMOS:} \quad \mu_{ac} = 90. E_{eff}^{-0.3} \quad \text{Eq. II-121}$$

- μ_{cbs} : (Cbs pour Coulomb scattering): régime d'interaction des porteurs avec les centres coulombiens.

$$\text{NMOS:} \quad \mu_{cbs} = \frac{1380}{1 + \left(\frac{N_{ch}}{10^{17}}\right)^{0.6}} \quad \text{Eq. II-122}$$

$$\text{PMOS:} \quad \mu_{cbs} = \frac{500}{1 + \left(\frac{N_{ch}}{10^{17}}\right)^{0.7}} \quad \text{Eq. II-123}$$

Où E_{eff} est le champ effectif donné par :

$$\text{NMOS :} \quad E_{eff} = \frac{V_g - V_t}{6 t_{inv}} + \frac{1}{3} \frac{Q_{dep}}{C_{inv}} \frac{1}{t_{inv}} \quad \text{Eq. II-124}$$

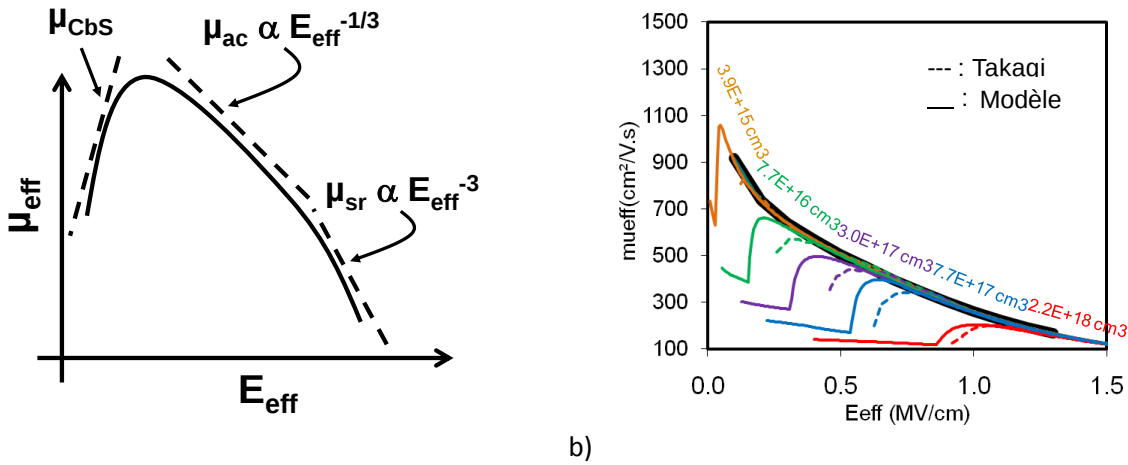
$$\text{PMOS:} \quad E_{eff} = \frac{V_g - V_t}{9 t_{inv}} + \frac{1}{3} \frac{Q_{dep}}{C_{inv}} \frac{1}{t_{inv}} \quad \text{Eq. II-125}$$

Où Q_{dep} est la charge d'inversion, t_{inv} l'épaisseur d'oxyde de grille en inversion (i.e. EOT+darkspace) et C_{inv} la capacité d'oxyde de grille en inversion.

Ces trois composantes sont ensuite liées par une loi de Mathiesen, mais le terme μ_{cbs} doit être corrigé du fait de l'écrantage de la charge de déplétion par la charge d'inversion [Bœuf 09] :

$$\frac{1}{\mu_{eff}} = \frac{1}{\mu_{ac}} + \frac{1}{\mu_{sr}} + \frac{\frac{C_{bscr} \cdot Q_{dep}}{Q_i + C_{bscr} \cdot Q_{dep}}}{\mu_{cbs}} \quad \text{Eq. II-126}$$

Avec C_{bscr} facteur d'écrantage de la charge de déplétion par la charge d'inversion, Q_{dep} charge de déplétion et Q_i charge d'inversion. Avec cette équation, on peut tracer la courbe μ_{eff}/E_{eff} représentée sur la Figure II-40, qui reproduit bien les mesures de [Takagi 94]



a) b)
Figure II-40 : (a) décomposition de la mobilité effective μ_{eff} dans le cas du NMOS. (b) Tracé des courbes de mobilité universelle pour différentes valeurs de dopage canal.

II.D.3. Raccord faible-forte inversion

Le raccord entre les régimes de faible et de forte inversion est la principale source de discontinuité dans une modélisation du courant de drain dans les transistors MOSFET. En effet, il s'agit ici de joindre deux régimes décrit par deux équations distinctes, et n'ayant pour paramètre commun que la tension seuil qui jouera le rôle de frontière supérieure pour le régime sous le seuil, donc du courant de diffusion (I_{diff}) (paragraphe I.B.1) et de frontière inférieure pour le régime au-delà du seuil, donc du courant de dérivation (I_{drift}) (paragraphe I.B.2). Dans notre cas, nous choisissons de fixer la valeur du courant de drain au seuil ($V_g = V_t$) à la valeur $I_{th} = 10^{-7} W/L$, ce qui fixe la composante de diffusion, et permet d'avoir une valeur du courant de fuite I_{off} . Or, la composante de diffusion « s'arrête » brutalement à $V_g = V_t$ et la composante de dérivation devient brutalement non nulle. Il est clair qu'une simple somme des composantes ne permet pas d'obtenir une courbe continue. La Figure II-41 illustre cette problématique.

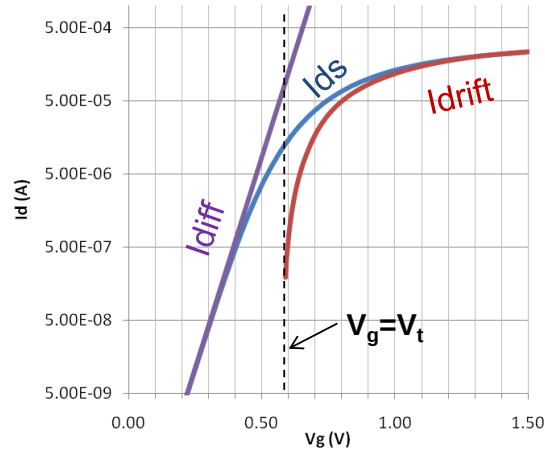


Figure II-41 : Tracé des composantes de dérive et de diffusion avec les équations classique en fonction de la polarisation de grille et comparaison avec la courbe complète du courant drain source, mettant en évidence la problématique du raccord au seuil.

Pour obtenir une courbe continue, nous allons utiliser une méthode proche de celles utilisées par [BSIM] et [Skotnicki 94], c'est-à-dire en utilisant une expression de tension de grille effective, donnée par l'équation :

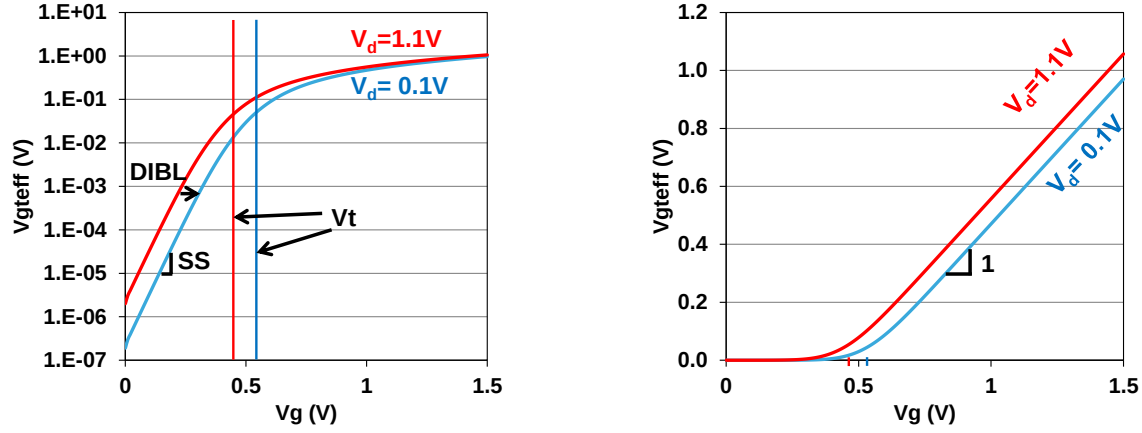
$$V_{g\text{eff}} = \frac{\text{terme1}}{\text{terme2}} \quad \text{Eq. II-127}$$

$$\text{terme1} = \frac{2S \ln \left(\frac{-\ln \left(\frac{1 + (V_g - V_{tsat}) \ln(10)}{2S} \right)}{\ln(10)} \right)}{\ln(10)} \quad \text{Eq. II-128}$$

$$\text{terme2} = 1 + \frac{2 \frac{S}{kT} e^{\left(-\ln(10) \left(\frac{(V_g - V_{tsat}) - 2V_{off}}{2S} \right) \right)}}{\ln(10)} \quad \text{Eq. II-129}$$

$$V_{off} = \frac{S}{\ln(10)} \ln \left(\frac{\mu_{eff} C_{ox} (kT)^2}{(1+d) I_{th}} \right) \quad \text{Eq. II-130}$$

Grace à cette expression de la tension de grille effective, on obtient une tension croissante sous le seuil, qui est linéaire et de pente égale à la pente sous le seuil S , puis, après le seuil, cette tension devient linéaire, de pente égale à 1 (donc même variation que la tension de grille V_g). De plus, en faisant varier la tension de drain, le DIBL est bien visible en échelle logarithmique dans le régime sous le seuil. Ces observations sont illustrées par la Figure II-42.



a)

b)

Figure II-42 : Tracé de la tension de grille effective ($V_{g\text{teff}}$) donné par les équations II-127 à II-130 en échelle logarithmique (a) et en échelle linéaire (b).

En remplaçant la différence entre la tension de grille et la tension de seuil ($V_g - V_t$) par la tension de grille effective ($V_{g\text{teff}}$) dans l'expression du courant de dérivation (I_{drift}), on a :

$$I_{ds} = \mu_{\text{eff}} C_{\text{ox}} \frac{W}{L} \frac{V_{g\text{teff}}}{\left(1 + \frac{V_{ds}}{LE_c}\right)} \left(1 - \frac{(1+d)V_d}{2(V_{g\text{teff}} + 2kT)}\right) V_d \quad \text{Eq. II-131}$$

Avec $V_d = V_{ds}$ si $V_d < V_{dsat}$
 $V_d = V_{dsat}$ sinon

Où V_{dsat} est la tension de saturation de drain donnée par :

$$V_{dsat} = \frac{1}{\frac{1}{LE_c} + \frac{1+d}{V_{gt} + 2kT}} \quad \text{Eq. II-132}$$

On obtient alors des caractéristiques $I_d - V_g$ et $g_m - V_g$ (obtenue par dérivation numérique) parfaitement continues ; ces dernières sont représentées sur la Figure II-43.

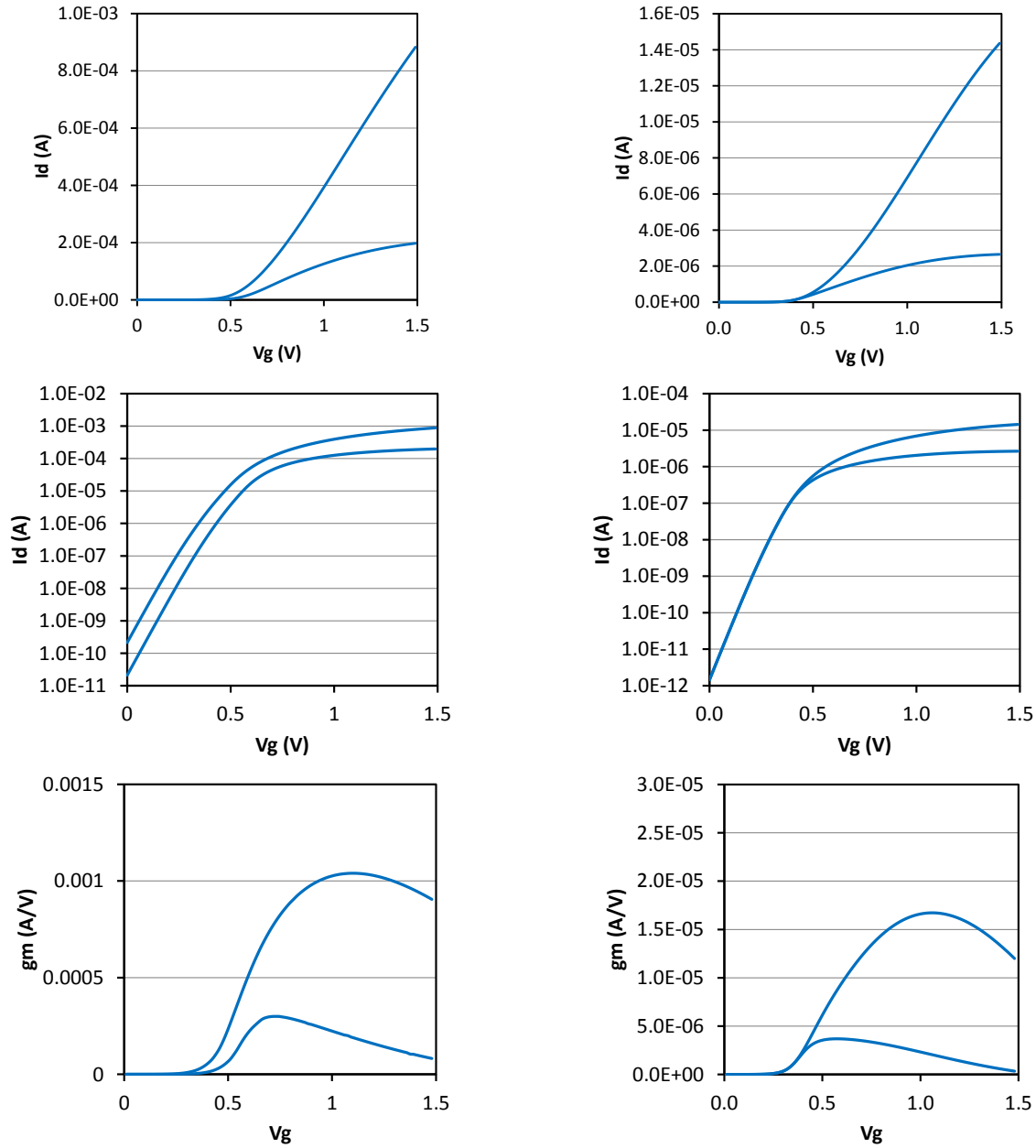
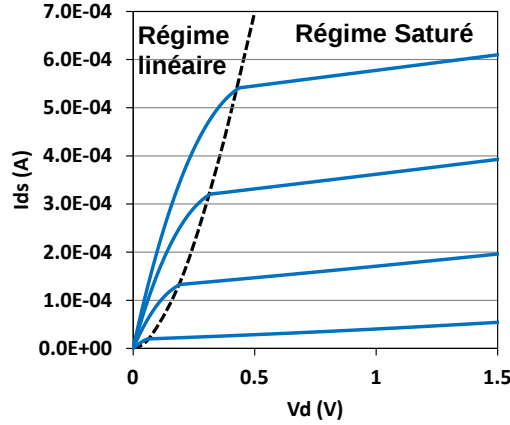


Figure II-43 : I_d - V_g et g_m - V_g tracées pour deux valeurs de V_d (0.1V et 1.1V) démontrant la parfaite continuité de nos équations. Colonne de gauche : transistor court, colonne de droite transistor long.

La continuité des I_d - V_g est assurée de cette manière, il reste maintenant à travailler sur les I_d - V_d .

II.D.4. Raccord régime linéaire-saturé

En traçant les I_d - V_d avec l'équation II-131, il apparaît une discontinuité lors de la transition entre le régime linéaire et le régime saturé, délimités par $V_d = V_{dsat}$ (Figure II-44).

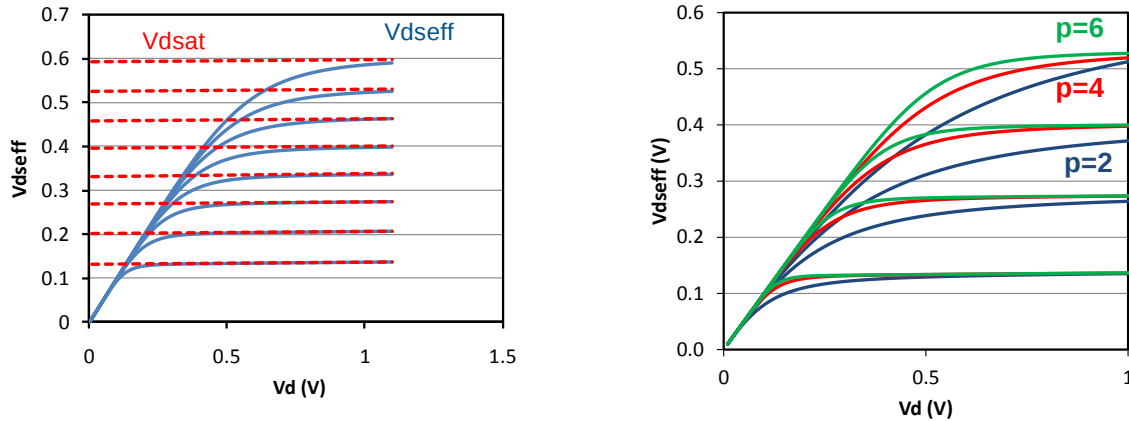

 Figure II-44 : I_d - V_d obtenues à partir de l'équation II-91 démontrant que la continuité n'est pas assurée.

Comme dans le paragraphe II.D.3, nous allons utiliser une tension effective de drain, qui sera égale à la tension de drain quand celle-ci sera inférieure à la tension de saturation de drain, pour tendre vers V_{dsat} pour $V_d > V_{dsat}$, avec une transition continue entre les deux régimes. Cette tension effective est définie par :

$$V_{dseff} = V_d \left(1 + \left(\frac{V_d}{V_{dsat}} \right)^p \right)^{-\frac{1}{p}} \quad \text{Eq. II-133}$$

$$V_{dsat} = \frac{1}{\frac{1}{LE_c} + \frac{1+d}{V_{gseff} + \frac{2kT}{q}}} \quad \text{Eq. II-134}$$

Où E_c est le champ critique, L la longueur de grille, d facteur d'effet substrat et p paramètre permettant d'ajuster la « vitesse » de passage au régime de saturation. Le tracé de V_{dseff} en fonction de V_d pour différentes valeurs de V_g et de p est représenté sur la Figure II-45.



a) b)
Figure II-45: (a) Tracé de V_{dseff} en fonction de V_d pour différentes valeurs de V_g montrant que $V_{dseff} = V_{ds}$ puis V_{dsat} lorsque $V_d > V_{dsat}$ en étant continue. (b) Tracé de V_{dseff} en fonction de V_d pour différentes valeurs de V_g et 3 valeurs distinctes du paramètre d'ajustement p pour illustrer son impact.

En remplaçant V_d par V_{dseff} dans l'équation II-131, on obtient :

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L} \frac{V_{gteff}}{\left(1 + \frac{V_{dseff}}{LE_c}\right)} \left(1 - \frac{(1+d)V_{dseff}}{2(V_{gteff} + 2kT)}\right) V_{dseff} \quad \text{Eq. II-135}$$

La Figure II-46 montre les courbes I_d-V_d et g_d-V_d (obtenue par dérivation numérique) obtenues à partir de l'équation II-135 et prouve que la continuité est assurée entre les deux régimes.

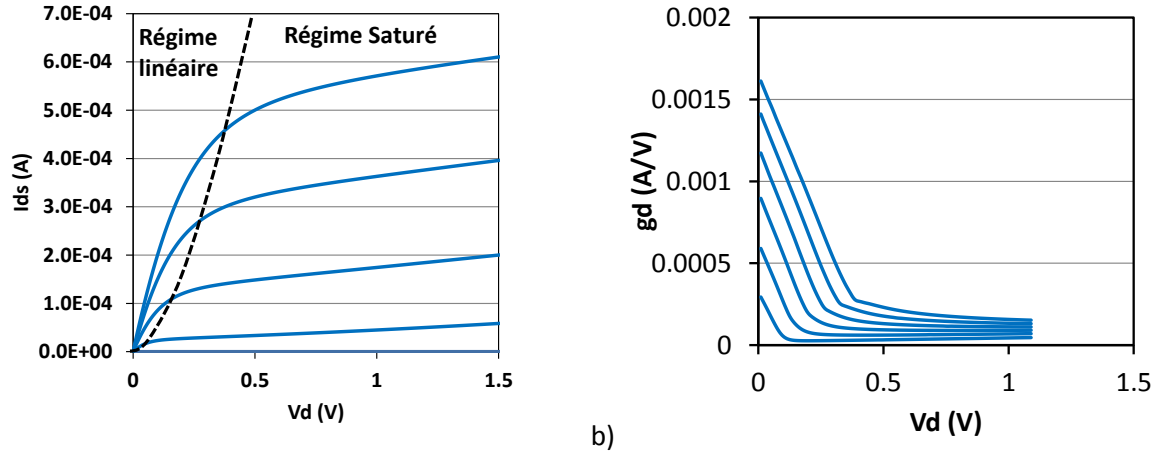


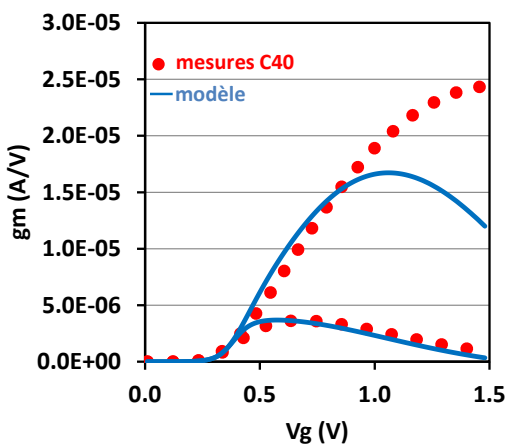
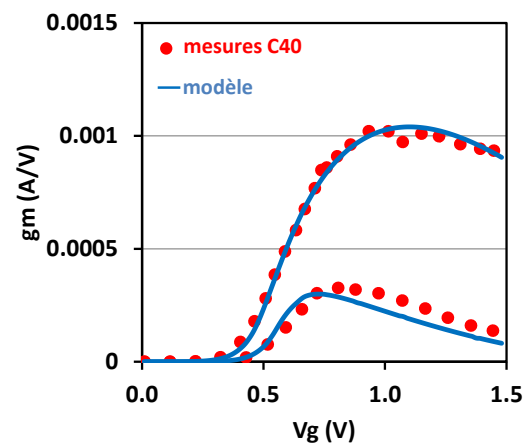
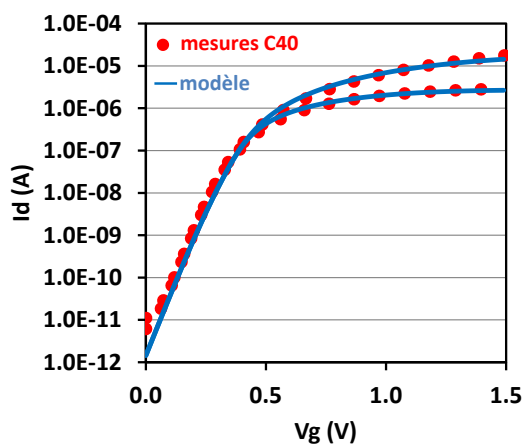
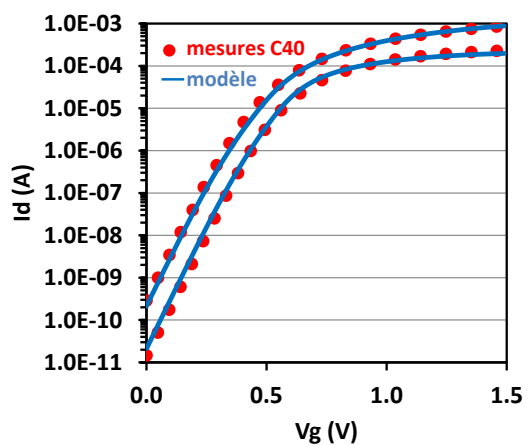
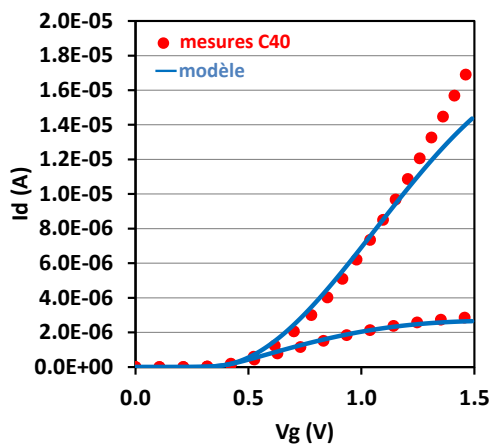
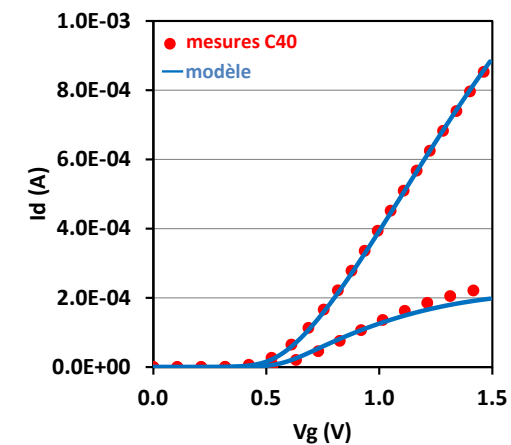
Figure II-46 : I_d-V_d (a) et g_d-V_d (b) obtenues à partir de l'équation II-135 démontrant que la continuité est assurée.

II.D.5. Comparaison au silicium

Pour valider notre modèle nous allons l'utiliser pour reproduire les caractéristiques courant-tension obtenues par des mesures effectuées sur différents nœuds technologiques et sur différentes architecture.

II.D.5.a) Transistor conventionnel sur substrat massif typique du nœud 40nm

Commençons par reproduire les mesures effectuées sur la technologie conventionnelle sur substrat massif CMOS 40nm (shrink de la technologie 45nm [Josse 06] de STMicroelectronics. Cette technologie utilise une grille en polysilicium (donc un travail de sortie de 4.06V) et un oxyde de grille en SiO_2 d'épaisseur 1.7nm, la longueur de grille nominale est de 40nm, la profondeur de jonction X_j est environ de 15nm, la résistance d'accès R_{acc} est de $180\Omega \cdot \mu\text{m}$ (la correction du courant de drain par la résistance d'accès sera explicité dans le chapitre IV). et la tension d'alimentation nominale est de 1.1V. En entrant ces paramètres dans notre modèle et en ajustant le facteur de contrainte devant la mobilité (cette technologie est contrainte par un liner tensile et par la SMT [Josse 06]) nous obtenons les courbes de la Figure II-47.



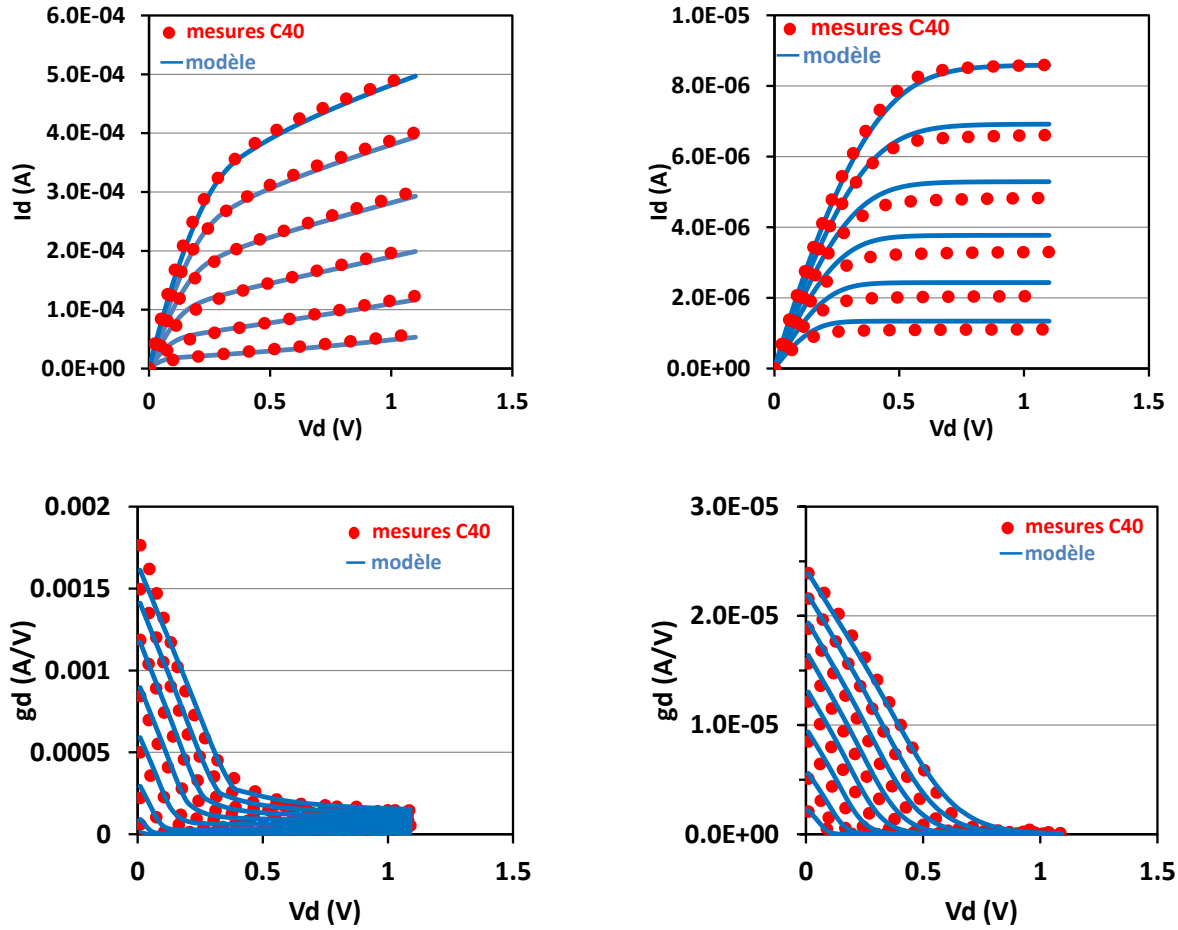


Figure II-47 : Reproduction des caractéristiques courant-tension mesurées sur la technologie CMOS 40nm de STMicroelectronics sur le transistor nominal (longueur de grille à 40nm) et le transistor long ($L=10\mu\text{m}$). Les I_d - V_g sont mesurées à $V_d=0.1\text{V}$ et 1.1V (tension d'alimentation nominale du 40nm) et les I_d - V_d à $V_g=0.4\text{V}$ à 1.1V par pas de 100mV .

On peut donc voir que notre modèle reproduit de manière satisfaisante les mesures effectuées sur la technologie conventionnelle sur substrat massif du noeud 40nm.

II.D.5.b) Transistor FDSOI typique du noeud 28nm

Dans ce paragraphe nous reproduisons les caractéristiques courant-tension obtenues sur la technologie FDSOI 28nm de STMicroelectronics [Planes 12]. Nous nous concentrerons cette fois sur le transistor nominal $L=24\text{nm}$. Cette technologie utilise un empilement grille métallique/diélectrique haute permittivité, dont le t_{inv} vaut 1.5nm , la tension d'alimentation nominale est de 1V , l'épaisseur du film de silicium est $t_{\text{si}}=7\text{nm}$ et celle de l'oxyde enterré est de 25nm . Enfin des plans de masse (ground plane GP) sont réalisés par implantation ionique de même type que les source-drains sous le BOX à un niveau de dopage de 10^{18}cm^{-3} . En entrant ces paramètres dans notre modèle, nous obtenons les tracés de la Figure II-48.

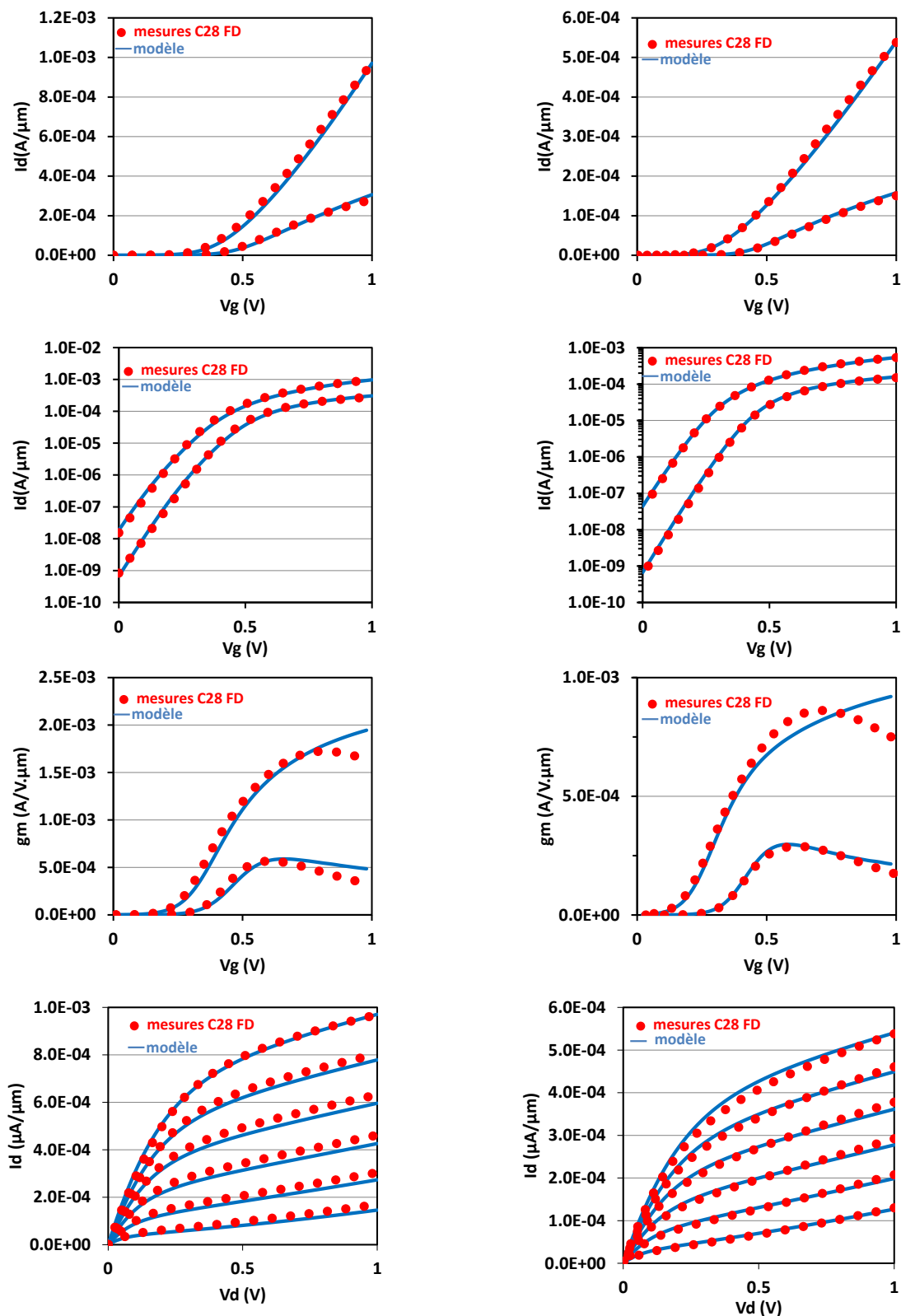


Figure II-48 : Reproduction des caractéristiques courant-tension mesurées sur la technologie CMOS FDSOI 28nm de STMicroelectronics sur le transistor nominal (longueur de grille à 24nm). Les I_d - V_g sont mesurées à $V_d=0.1\text{V}$ et 1V et les I_d - V_d à $V_g=0.4\text{V}$ à 1V par pas de 100mV.

II.D.5.c) Double Grille planaire

Enfin, pour valider notre modèle sur la dernière architecture, à savoir le double grille, nous allons reproduire les mesures effectuées par [Huguenin 10] sur un PMOS de longueur de grille $L=40\text{nm}$, d'épaisseur de film de silicium 7nm et de $t_{\text{inv}}=2.4\text{nm}$. La Figure II-49 montre que le modèle reproduit bien les mesures.

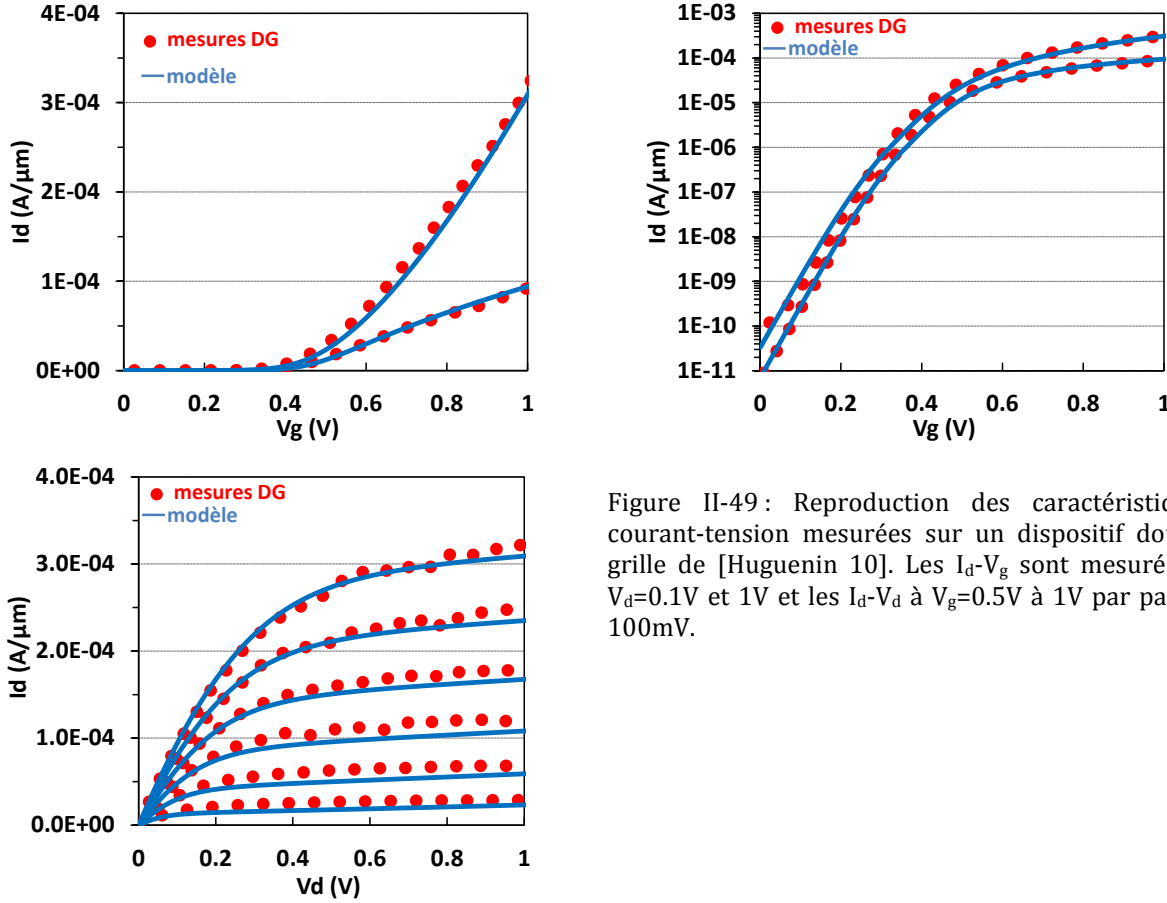


Figure II-49 : Reproduction des caractéristiques courant-tension mesurées sur un dispositif double grille de [Huguenin 10]. Les I_d - V_g sont mesurées à $V_d=0.1\text{V}$ et 1V et les I_d - V_d à $V_g=0.5\text{V}$ à 1V par pas de 100mV .

II.E. Conclusion du chapitre

Dans l'optique d'évaluer les performances au niveau circuit des différentes technologies CMOS, les caractéristiques courant-tension de ces architectures sont préalablement nécessaires. Nous avons proposé dans ce chapitre un modèle compact universel permettant de produire les courbes (I_d - V_g) et (I_d - V_d) pour chaque architecture. Les régimes sous le seuil d'un transistor conventionnel sur substrat massif, FDSOI et double grille faiblement dopé sont décrits par une modélisation précise et prédictive de la tension de seuil, de l'effet canal court SCE, du DIBL et de la pente sous le seuil. Afin de tenir compte de la spécificité de chaque technologie, la détermination du critère de seuil idoine a été le point de départ de chacun des modèles. La tension de seuil canal long, c'est-à-dire sans tenir compte de l'effet de la longueur de grille, est alors obtenue par résolution de l'équation de Poisson avec les conditions aux limites spécifiques à chaque architecture. La polarisation de la face arrière n'a d'effet significatif que pour l'architecture FDSOI et nous avons donc modélisé son effet uniquement pour cette architecture. Les effets de confinement quantique n'ont été pris en compte que pour la tension de seuil canal long du transistor double grille car ils sont significatifs uniquement pour ce type de dispositif, pour une épaisseur de film de silicium $t_{\text{si}} < 10\text{nm}$. Nous avons tenu compte de l'effet de la longueur de grille par l'utilisation de la transformation tension-dopage VDT [Skotnicki 88-a]. Celle-ci a été suffisante pour l'architecture conventionnelle sur substrat massif et cette technique a pu être adaptée pour le transistor double grille. La VDT

ne tient intrinsèquement compte que des lignes de champ électrique incluses dans le canal. Afin de modéliser l'effet de l'épaisseur de l'oxyde enterré sur la tension de seuil du transistor FDSOI, nous avons adapté les travaux de [Ernst 07].

Pour résumer, l'électrostatique de chaque architecture a été modélisée en tenant compte des spécificités de chaque dispositif et de l'effet des différents paramètres technologiques, c'est-à-dire:

- Pour chaque architecture, la longueur de grille L_g , l'épaisseur d'oxyde de grille t_{ox} , la température T et les caractéristiques du métal de grille au travers de l'estimation de la tension de bandes plates V_{FB} .
- Pour le transistor conventionnel sur substrat massif, le dopage canal N_{ch} et la profondeur de jonction X_j .
- Pour le transistor FDSOI, l'épaisseur de film de silicium t_{si} , l'épaisseur d'oxyde enterré t_{box} , le dopage du plan de masse N_{GP} et son type et la polarisation de la face arrière V_b .
- Pour le transistor double grille faiblement dopé, l'épaisseur du film de silicium t_{si} .

Chaque modèle est validé par simulations numériques 2D [synopsis] et par caractérisations électriques quand celles-ci étaient disponibles.

Le régime au-delà du seuil est ensuite décrit par une modélisation classique du courant de dérive où la mobilité effective est estimée par les lois universelles de [Takagi 94]. L'effet d'amélioration du transport par contrainte mécanique est pris en compte avec les travaux de [Payet 08] et la dégradation de mobilité effective due à la longueur de grille est prise en compte par le simple modèle de [Bidal 09]. Les effets de confinement quantique sur la charge d'inversion (i.e. darkspace) sont modélisés par l'intermédiaire du paramètre t_{inv} (I.C.3.c).

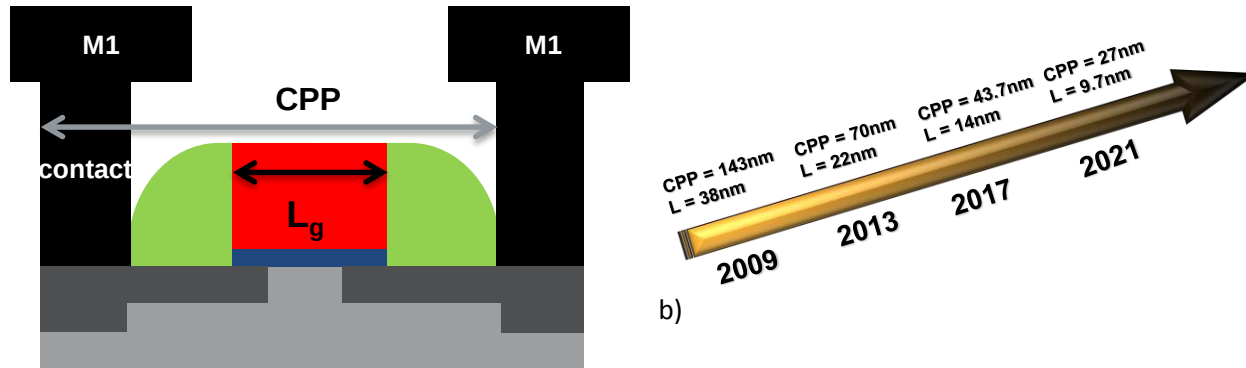
Enfin, l'objectif étant de prévoir les performances circuit par l'utilisation d'un outil de CAO conventionnel, la continuité des caractéristiques courant-tension (I_d-V_g et I_d-V_d) et de leurs dérivées (g_m-V_g et g_d-V_d) est indispensable. Nous avons alors proposé les expressions de tensions effectives assurant les raccords continus entre chaque régime de fonctionnement.

Les caractéristiques courant-tension (I_d-V_g et I_d-V_d) et leurs dérivées (g_m-V_g et g_d-V_d) sont finalement validés par caractérisations électriques pour chaque architecture de transistor, démontrant la pertinence et l'efficacité du modèle proposé.

- CHAPITRE III -

EVALUATION ANALYTIQUE DES CAPACITES PARASITES DANS LES STRUCTURES CMOS.

Avec la poursuite de la réduction des dimensions dans le développement des architectures CMOS avancées, et notamment la réduction de longueur de grille de génération en génération, le contrôle électrostatique est de plus en plus difficile à assurer. La modélisation de ces effets est proposée par le chapitre II. Cependant, cette course à la miniaturisation porte aussi sur le pas de répétition d'une grille et d'un contact (paramètre noté usuellement CPP pour Contacted Poly Pitch). Cette dimension est illustrée sur la Figure III-1-a ainsi que les valeurs attendues par l'ITRS dans l'édition 2010 de sa feuille de route [ITRS] dans les années à venir pour L_g et CPP (Figure III-1-b).



a)
Figure III-1 : (a) Représentation schématique de la longueur de grille (L_g) et du Contacted Poly Pitch (CPP). (b) Evolution de L_g et CPP prévue par la feuille de route ITRS 2010.

Comme le montre la Figure III-1-a, la distance grille-contact est étroitement liée à CPP et L_g . En considérant que cette distance est égale à la largeur d'un contact, on peut l'estimer à $(CPP - L_g)/3$. En se référant aux valeurs attendues par l'ITRS, on voit bien que la distance va également énormément diminuer, pour atteindre environ 6nm en 2021 :

- La diminution de la distance grille-contact implique une augmentation de la capacité parasite grille-drain.
- Dans les technologies CMOS avancées, la longueur de grille est relâchée car les effets canaux courts deviennent difficiles à contrôler (le nœud technologique ne correspond plus à la longueur de grille). Par conséquent, la distance grille-contact est encore plus faible, et la capacité parasite grille-drain encore plus importante.
- Dans la course à la performance, les source-drains sont désormais réalisés par épitaxie (on parle de source-drains surélevés) pour limiter la résistance d'accès. Cette nouvelle architecture de jonction provoque une nouvelle augmentation de la capacité parasite grille-drain.

Par ces remarques, on constate que les capacités parasites vont être de plus en plus importantes, ce qui a été anticipé il y a déjà quelques années (Figure III-2) et vont devenir un des principaux freins à la performance (I.C.4). C'est la raison pour laquelle une modélisation précise des capacités parasites sur chaque architecture est indispensable pour pouvoir effectuer une évaluation de performance idoine.

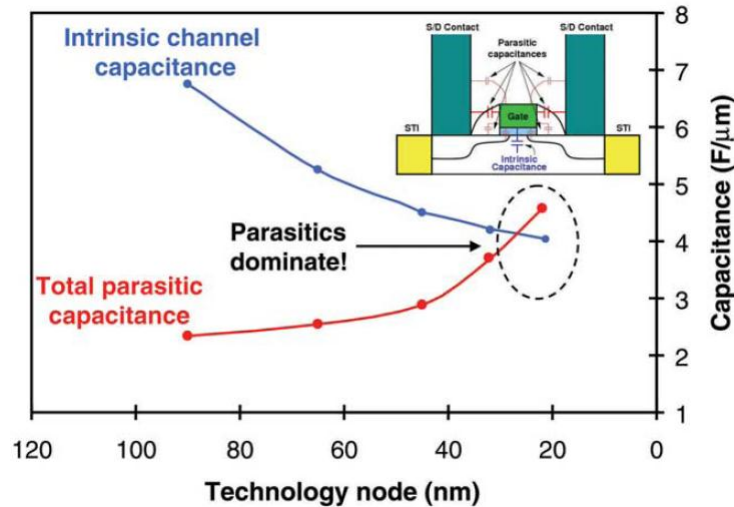


Figure III-2 : Evolution de la capacité parasite totale et de la capacité grille intrinsèque d'un transistor en fonction du nœud technologique, montrant que les capacités parasites vont dépasser la capacité grille autour du nœud 20nm (graphe extrait de S. E. Thompson, Mater. today 2005)

III.A. Méthodologie

Pour modéliser ces capacités parasites, la première étape consiste à déterminer les méthodes de calcul à utiliser dans chaque cas. On commence donc par représenter schématiquement les capacités parasites sur un transistor sur substrat massif (d'après [Wei 11]) sur la Figure III-3.

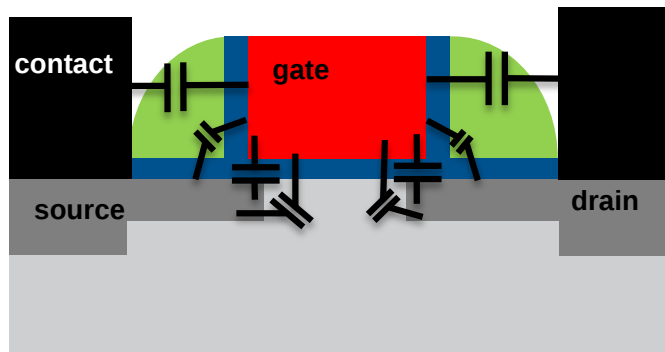


Figure III-3 : représentation des différentes capacités parasites sur un transistor sur substrat massif.

On peut alors constater qu'il existe deux types de capacités parasites :

- La capacité formée par deux électrodes parallèles, comme la capacité entre la grille et le contact.
- La capacité formée par deux électrodes perpendiculaires, comme la capacité entre le flanc de grille et la source ou le drain, à travers l'espaceur.

Les méthodes de calcul pour évaluer de manière précise ces deux types de capacités sont présentés dans les deux paragraphes ci-dessous.

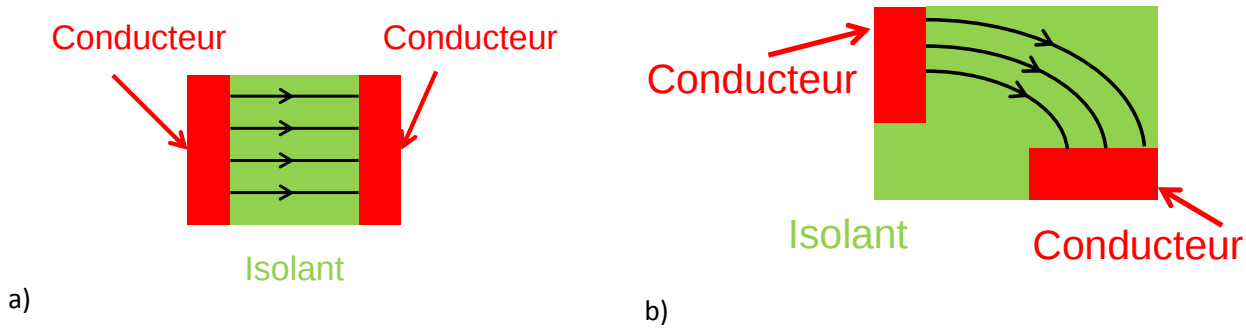


Figure III-4 : représentation schématique des deux cas de capacités parasites : (a) Capacité formée par deux électrodes parallèles (b) Capacité formée par deux électrodes perpendiculaires. Les flèches noires représentent la forme des lignes de champs électriques.

III.A.1. Capacité formée par deux électrodes parallèles

La modélisation est ici très simple car on se trouve dans le cas classique du condensateur plan (Figure III-4-a) où les lignes de champ électrique sont perpendiculaires aux électrodes et rectilignes. La capacité surfacique C (en F/m^2) est donnée par :

$$C = \frac{\epsilon}{t} \quad \text{Eq. III-1}$$

Où ϵ est la permittivité de l'isolant et t l'épaisseur de cet isolant.

III.A.2. Capacité formée par deux électrodes perpendiculaires

Dans ce cas, comme indiqué schématiquement par la Figure III-4-b, les lignes de champs ne sont plus rectilignes, et donc l'équation du condensateur plan ne peut plus être utilisée. De précédents travaux traitent cette question en faisant l'hypothèse que les lignes de champs sont des cercles [Elmasry 82], [Shrivastava 82] [Suzuki 99] (ce qui présuppose que les électrodes sont de même dimension). Par une simple intégration, ils obtiennent des expressions analytiques simples mais qui ne sont pas très précises si on compare ces résultats à des simulations numériques. Pour une meilleure évaluation, de plus récents travaux [Bansal 05] [Wei 12] proposent de considérer que les lignes de champs sont des ellipses, et non plus des cercles. Il n'est alors plus possible de calculer la capacité directement, et ces derniers proposent d'utiliser un changement de repère, basée sur une transformation conforme [Plonsey 61], appelé en anglais conformal mapping. Cependant, ces travaux ont simplifié les expressions obtenues, et ils ont perdu en précision.

Dans le paragraphe suivant, nous proposons de reprendre la transformation conforme depuis le début, pour obtenir une modélisation la plus précise possible.

III.B. La transformation conforme

III.B.1. Définition du problème

Pour évaluer une capacité entre deux électrodes, la première méthode à utiliser est de calculer la capacité par unité de longueur d'électrode (en F/m), à l'aide de la formule du condensateur plan, puis de l'intégrer le long de l'électrode, ce qui donne :

$$C = \int_a^b \frac{\epsilon}{L(x)} dx \quad \text{Eq. III-2}$$

Où a et b sont les coordonnées d'une électrode, ϵ la permittivité de l'isolant et $L(x)$ la longueur de la ligne de champ en x . la Figure III-5 représente schématiquement les cas de la capacité à deux électrodes parallèles (haut) et celui de la capacité à électrode perpendiculaire, en considérant les lignes de champs circulaires. En réinjectant $L(x)$ dans l'équation III-2, on a:

- Electrodes parallèles : $C = \frac{(a-b)\epsilon}{t}$ Eq. III-3

- Electrodes perpendiculaires : $C = \frac{2\epsilon}{\pi} \ln\left(\frac{a}{b}\right)$ Eq. III-4

On remarque que l'équation III-4 correspond aux résultats obtenus dans [Suzuki 99], et ne donnera donc pas satisfaction. Comme nous l'avons mentionné dans le paragraphe III.A.2, nous allons considérer que les lignes de champs sont de forme elliptique. Nous justifions ce choix par la littérature [Bansal 05] et [Wei 11], mais également en effectuant des simulations numériques. Comme le montre la Figure III-5-b, les surfaces iso-potentielles dans une capacité à électrodes perpendiculaires sont très semblables à des hyperboles. Les lignes de champs électriques étant, par définition, perpendiculaires aux surfaces iso-potentielles (le champ électrique est donné par le gradient du potentiel), on peut alors justifier la forme elliptique des lignes de champs électriques. Par conséquent, le $L(x)$ à utiliser dans l'équation III-2 sera le quart du périmètre d'une ellipse. Du fait de la complexité de l'expression du périmètre d'une ellipse, l'équation III-2 ne pourra pas être intégrée analytiquement dans ce cas. C'est la raison pour laquelle il est indispensable de passer par une transformation conforme.

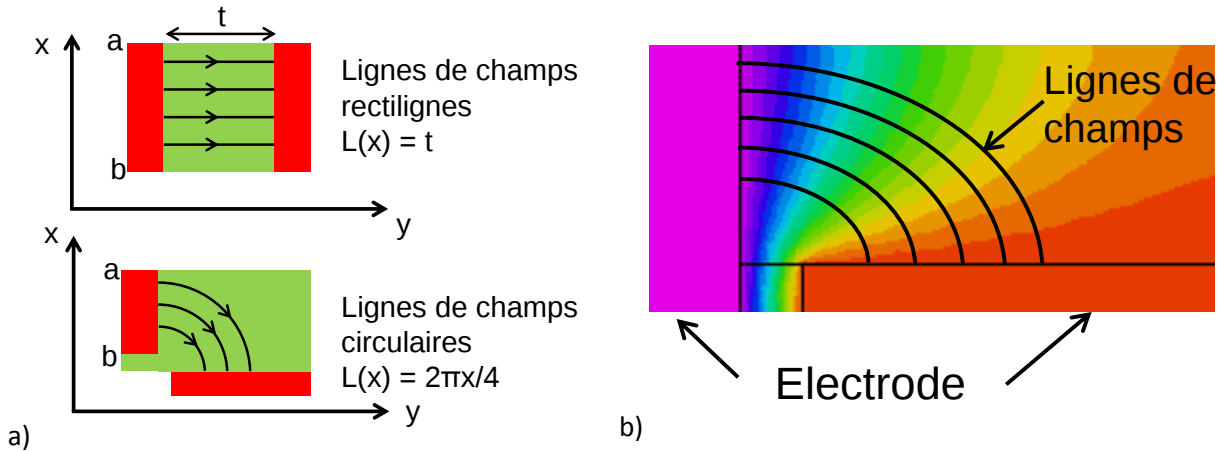


Figure III-5 : (a) Représentation schématique des coordonnées pour le calcul d'une capacité à électrodes parallèles (haut) et d'une capacité à électrodes perpendiculaires en considérant les lignes de champs circulaires (bas). (b) Cartographie des surfaces iso-potentielles dans une capacité à électrodes perpendiculaires obtenue par simulations numériques.

Dans ces travaux de modélisation, nous allons utiliser une transformation conforme pour transformer le repère complexe cartésien initial en un nouveau repère complexe de forme elliptique, où le calcul de la capacité à électrodes perpendiculaires sera réduit au calcul classique d'une capacité à électrodes parallèles (Figure III-6).

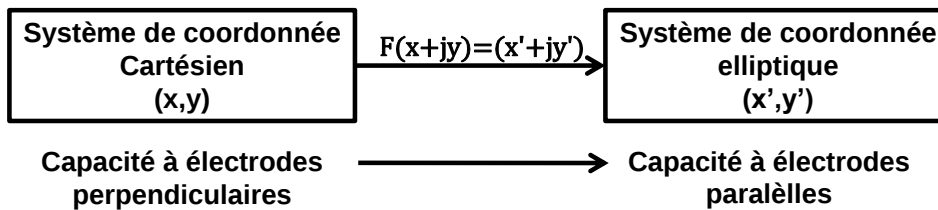


Figure III-6 : Principe de la transformation conforme.

III.B.2. La fonction de transformation

La première étape est donc de choisir la fonction de transformation. [Plonsey 61] propose d'utiliser la fonction (F) cosinus réciproque:

$$F(x + jy) = \arccos(x + jy) = x' + jy' \quad \text{Eq. III-5}$$

$$F(x' + jy')^{-1} = \cos(x' + jy') = x + jy \quad \text{Eq. III-6}$$

Ensuite, il nous faut exprimer les coordonnées du nouveau repère (x' , y') en fonction de celle de l'ancien (x, y). Commençons par isoler x et y à partir de l'équation III-6. Par définition du cosinus pour un nombre complexe Z :

$$\cos(Z) = \frac{e^{jZ} + e^{-jZ}}{2} \quad \text{Eq. III-7}$$

Avec $Z = x' + jy'$, l'équation III-7 devient :

$$\cos(x' + jy') = \frac{e^{j(x' + jy')} + e^{-j(x' + jy')}}{2} \quad \text{Eq. III-8}$$

En développant, on a :

$$\cos(x' + jy') = \frac{e^{-y'} e^{jx'}}{2} + \frac{e^{y'} e^{-jx'}}{2} \quad \text{Eq. III-9}$$

En égalisant la définition exponentielle et la définition trigonométrique d'un nombre complexe, l'équation III-9 peut être réécrite :

$$\cos(x' + jy') = \frac{e^{-y'} (\cos(x') + j \sin(x'))}{2} + \frac{e^{y'} (\cos(-x') + j \sin(-x'))}{2} \quad \text{Eq. III-10}$$

Sachant que pour un nombre réel a , $\cos(-a) = \cos(a)$ et $\sin(-a) = -\sin(a)$, et en factorisant l'équation III-10:

$$\cos(x' + jy') = \cos(x') \frac{(e^{y'} + e^{-y'})}{2} - \sin(x') \frac{(e^{y'} - e^{-y'})}{2} \quad \text{Eq. III-11}$$

En identifiant les définitions des fonctions cosinus et sinus hyperbolique (ch et sh) dans l'équation III-11, on obtient :

$$\cos(x' + jy') = \cos(x') ch(y') - j \sin(x') sh(y') \quad \text{Eq. III-12}$$

On peut alors réécrire l'équation III-6, et isoler x et y et les exprimer en fonction de (x' , y') :

$$x = \cos(x') ch(y') \quad \text{Eq. III-13}$$

$$y = -\sin(x') sh(y') \quad \text{Eq. III-14}$$

Afin de vérifier que la fonction de transformation donne bien un repère elliptique, on élève au carré les équations III-13 et III-14, puis on les divise respectivement par $ch(y')^2$ et $sh(y')^2$. En sommant les termes obtenus, on a :

$$\frac{x^2}{ch(y')^2} + \frac{y^2}{sh(y')^2} = \cos(x')^2 + \sin(x')^2 = 1 \quad \text{Eq. III-15}$$

L'équation III-15 donne bien l'équation paramétrique classique d'une ellipse (en fixant y') et la famille de courbes obtenues (pour différentes valeurs de y') est un système d'ellipses confocales, de foyer 1. De la même manière, on peut démontrer que la famille de courbes obtenues en isolant x' est un système d'hyperboles confocales :

$$\frac{x^2}{\cos(x')^2} - \frac{y^2}{\sin(x')^2} = \text{ch}(y')^2 - \text{sh}(y')^2 = 1 \quad \text{Eq. III-16}$$

La Figure III-7 représente le repère cartésien initial et le repère obtenu par transformation conforme, tracé avec les équations III-15 et III-16.

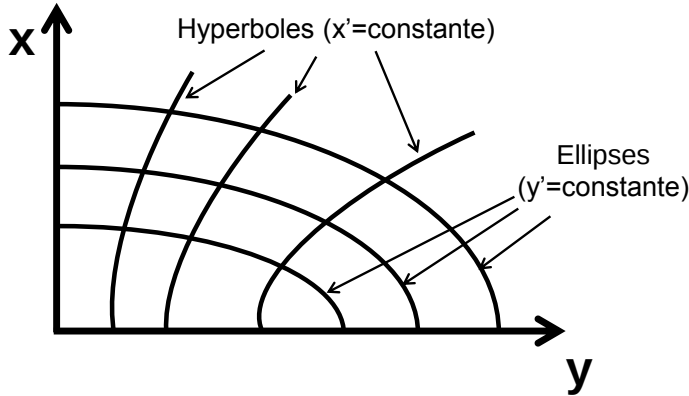


Figure III-7 : Représentation du repère cartésien initial et du repère obtenu par transformation conforme, tracé avec les équations III-15 et III-16.

A partir des deux équations précédentes, nous allons pouvoir isoler x' et y' et les exprimer en fonction de (x,y) [Plonsey 61]. Pour isoler x' , on remarque que :

$$x^2 + y^2 + 1 = \text{sh}(y')^2 + \cos(x')^2 + 1 \quad \text{Eq. III-17}$$

$$(x^2 + y^2 + 1)^2 - 4x^2 = (\text{sh}(y')^2 - \cos(x')^2 + 1)^2 \quad \text{Eq. III-18}$$

En soustrayant à la racine de l'équation III-16 à l'équation III-15, et en appliquant au radical de cette somme la fonction cosinus réciproque, on parvient à exprimer x' en fonction de (x,y) :

$$x' = \cos^{-1} \sqrt{\left(\frac{x^2 + y^2 + 1 - \sqrt{(x^2 + y^2 + 1)^2 - 4x^2}}{2} \right)} \quad \text{Eq. III-19}$$

Par un raisonnement analogue, on parvient à isoler y' :

$$x^2 + y^2 - 1 = \text{ch}(y')^2 - \sin(x')^2 - 1 \quad \text{Eq. III-20}$$

$$(x^2 + y^2 - 1)^2 + 4y^2 = (\text{ch}(y')^2 + \sin(x')^2 - 1)^2 \quad \text{Eq. III-21}$$

En ajoutant l'équation III-15 à la racine de l'équation III-16, et en appliquant au radical de cette somme la fonction sinus hyperbolique réciproque, on parvient à exprimer y' en fonction de (x,y) :

$$y' = \text{sh}^{-1} \sqrt{\frac{x^2 + y^2 - 1 + \sqrt{(x^2 + y^2 - 1)^2 + 4y^2}}{2}} \quad \text{Eq. III-22}$$

III.B.3. Expression de la capacité dans le nouveau repère

Pour calculer la capacité entre deux électrodes perpendiculaires, séparées par un isolant de permittivité ϵ , on commence par définir les dimensions des zones conductrices et des zones isolantes (Figure III-8-a). La fonction de transformation décrite dans le paragraphe III.B.2 n'est utilisable que pour un système d'ellipses confocales [plonsey 61], [Bansal 05], il est alors nécessaire de transformer la structure initiale pour garantir cette condition. Nous choisissons de jouer sur la dimension x_2 (le même raisonnement peut être effectué en modifiant y_2). Sachant que la plus petite ellipse est définie dans le repère initial Cartésien par l'équation paramétrique :

$$\frac{x^2}{x_1^2} + \frac{y^2}{y_1^2} = 1 \quad \text{Eq. III-23}$$

Et la plus grande ellipse par :

$$\frac{x^2}{(x_1 + x_2)^2} + \frac{y^2}{(y_1 + y_2)^2} = 1 \quad \text{Eq. III-24}$$

Le foyer f d'une ellipse étant défini par :

$$f = \sqrt{a^2 - b^2} \quad \text{Eq. III-25}$$

Avec a dimension du grand côté de l'ellipse, et b petit côté.

Il suffit d'égaliser les formules des foyers des deux ellipses extrêmes pour obtenir la valeur de x_2 (x_{2t}) qui assure que toutes les ellipses du système auront bien le même foyer. Formellement, il faut résoudre :

$$\sqrt{x_1^2 - y_1^2} = \sqrt{(x_1 + x_{2t})^2 - (y_1 + y_2)^2} \quad \text{Eq. III-26}$$

Ce qui donne comme valeur transformée x_{2t} :

$$x_{2t} = \sqrt{x_1^2 + 2y_1y_2 + y_2^2} - x_1 \quad \text{Eq. III-27}$$

On applique ensuite la fonction de transformation à notre structure, et on obtient la Figure III-8-b. On constate qu'il suffit alors d'appliquer la formule classique du condensateur plan dans le repère elliptique transformée pour évaluer cette capacité :

$$C = W\epsilon \frac{(y'_2 - y'_1)}{(x'_2 - x'_1)} \quad \text{Eq. III-28}$$

Où W est la largeur des électrodes selon l'axe z , ϵ est la permittivité de l'isolant et (x'_1, y'_1) (x'_2, y'_2) sont obtenus avec les équations III-19 et III-22 et correspondent respectivement à la plus petite et la plus grande ellipse du système, c'est-à-dire pour $(x,y)=(x_1/f,0)$ et $(x,y)=((x_1+x_{2t})/f,0)$, avec x_{2t} donné par l'équation III-27 et f valeur du foyer du système d'ellipses, donnée par l'équation III-26. Les dimensions x_2 et y_2 ayant un rôle dual, et comme la

valeur d'une capacité est limitée par l'électrode la plus petite, on remplace y_2 par la valeur minimum entre x_2 et y_2 dans l'équation III-27. Formellement, on a :

$$y'_2 = \text{sh}^{-1} \left(\sqrt{\frac{x_1^2 + \min(y_2, x_2)^2 + 2 \cdot y_1 \min(y_2, x_2)}{\sqrt{|x_1^2 - y_1^2|}}} \right) \quad \text{Eq. III-29}$$

$$y'_1 = \text{sh}^{-1} \left(\sqrt{\frac{x_1}{\sqrt{|x_1^2 - y_1^2|}}} \right) \quad \text{Eq. III-30}$$

$$x'_2 = \frac{\pi}{2} \quad \text{Eq. III-31}$$

$$x'_1 = 0 \quad \text{Eq. III-32}$$

De cette manière, une capacité à électrodes perpendiculaires dans le repère cartésien est équivalente et est évaluée comme une capacité à électrodes parallèles dans le repère elliptique transformé dont les électrodes sont séparées par un isolant d'épaisseur $\pi/2$ et de largeurs données par la fonction de transformation.

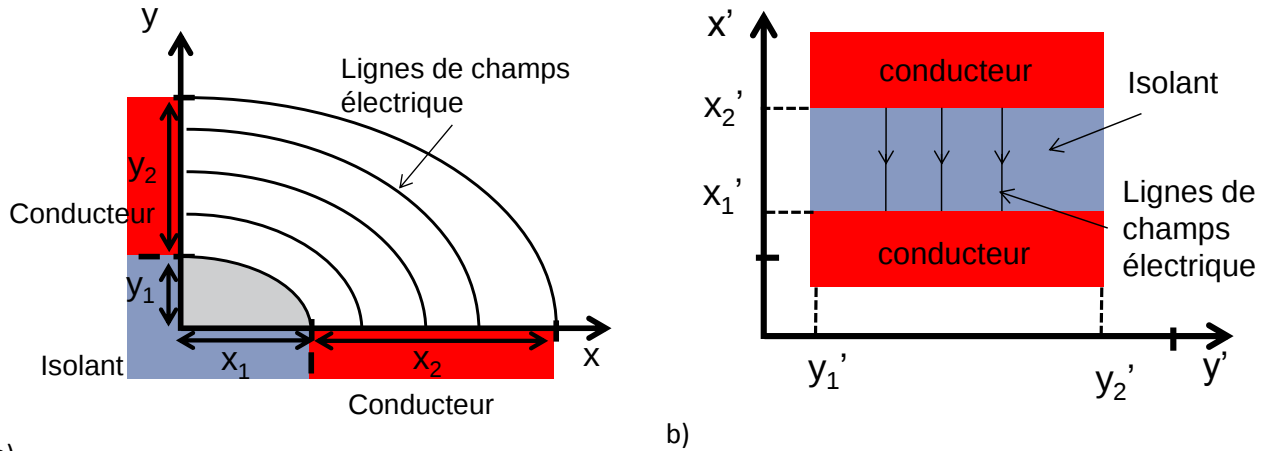


Figure III-8 : Coordonnées pour le calcul d'une capacité à électrode perpendiculaires, dans le repère cartésien initial (a) et dans le repère elliptique transformé (b).

Ce modèle permet de bien modéliser la capacité due aux lignes de champ elliptiques, ce qui est le cas des lignes débutant dans le repère cartésien (Figure III-8-a) pour $x > x_1$ et $y > y_1$. Cependant, il existe des lignes de champs entre les deux électrodes qui ne sont pas elliptiques et qui sont localisées dans la partie grisée de la Figure III-8-a). Ce couplage entre les deux électrodes n'est pas pris en compte par le modèle présenté ci-dessus, et est très difficile à modéliser proprement car il n'est pas possible de faire d'hypothèse sur la forme des lignes de champs dans cette zone. Pour avoir une estimation au premier ordre de cette capacité, nous utilisons [Plonsey 61] :

$$C = W 0.35 \frac{\epsilon}{\pi} \ln \left(\frac{\pi W}{\sqrt{|x_1^2 - y_1^2|}} \right) \quad \text{Eq. III-33}$$

Avec W largeur de la capacité (dimensions selon l'axe z sur Figure III-8-a) et 0.35 un paramètre d'ajustement obtenu par simulation numérique.

En définitive, la capacité de deux électrodes perpendiculaires, de dimensions données par la Figure III-8-a), séparées par un isolant de permittivité ϵ , et de largeur W selon l'axe z est évaluée par la somme des équations III-28 et III-33 :

$$C = 2W \frac{\epsilon}{\pi} \left[\operatorname{sh}^{-1} \left(\frac{\sqrt{x_1^2 + \min(y_2, x_2)^2 + 2y_1 \min(y_2, x_2)}}{\sqrt{|x_1^2 - y_1^2|}} \right) - \operatorname{sh}^{-1} \left(\frac{x_1}{\sqrt{|x_1^2 - y_1^2|}} \right) \right] + W 0.35 \frac{\epsilon}{\pi} \ln \left(\frac{\pi W}{\sqrt{|x_1^2 - y_1^2|}} \right) \quad \text{Eq. III-34}$$

L'équation III-34 n'est valable que si $x_1 \neq y_1$. Si nous sommes dans le cas $x_1 = y_1$ nous sommes dans le cas où les lignes de champs sont circulaires, et les travaux de [Suzuki 99] et des équations semblables à III-4 sont à utiliser.

Enfin, cette équation permet d'obtenir des résultats plus précis que les précédents travaux proposés dans la littérature, notamment par [Wei 11] pour plusieurs raisons :

- Nous n'avons pas cherché à simplifier les expressions, alors que les précédentes études ont toutes cherchées à éliminer le sinus hyperbolique réciproque. Pour ce faire, elles ont utilisé la définition exponentielle de cette fonction et ont « trop » simplifié le résultat [Bansal 05].
- Le terme « $\min(x_2, y_2)$ » traduit le fait que la valeur de la capacité est limitée par l'électrode la plus petite, ce qui n'était pas pris en compte précédemment.

III.C. Capacités parasites sur un dispositif planaire

Nous allons maintenant pouvoir appliquer, puis valider, les modèles définis pour les deux types de capacités identifiées sur les dispositifs CMOS, à savoir les capacités à électrodes parallèles et les capacités à électrodes perpendiculaires (III.A.2 et III.B).

III.C.1. Composantes parasites sur structures planaires

Nous allons commencer par identifier chaque capacité parasite sur les trois principales architectures planaires, à savoir le transistor conventionnel sur substrat massif (noté BULK), le transistor complètement déplété sur substrat sur isolant (noté FDSOI pour Fully Depleted Silicon On Insulator) et le double grille planaire (DG). La Figure III-9 représente schématiquement une vue en coupe de chaque architecture planaire, où :

- C_{ov} est la capacité de recouvrement (overlap en anglais). Il s'agit de la capacité entre la grille et l'extension des jonctions source-drain sous la grille à travers l'oxyde.
- C_{of} est la capacité de bords externes (outer-fringe en anglais). Il s'agit de la capacité entre le flan de grille et les source-drain à travers l'espaceur.
- C_{if} est la capacité de bords internes (inner-fringe en anglais). Il s'agit de la capacité entre la grille et l'extension des jonctions source-drain sous la grille à travers l'oxyde de grille et le canal du dispositif.
- C_{gepi} est la capacité grille-épitaxie. Il s'agit de la capacité entre la grille et l'épitaxie source-drain à travers l'espaceur.
- C_{pcca} est la capacité grille-contact. Il s'agit de la capacité entre la grille et le contact à travers l'espaceur.
- C_j est la capacité de jonction. Il s'agit de la capacité due à la jonction PN formée par les source-drains et le substrat.
- C_{box} est la capacité de BOX. Il s'agit de la capacité entre les source-drains et le substrat à travers l'oxyde enterré.
- C_{corner} est la capacité de coin (corner en anglais) qui est la capacité entre l'extension de grille sur STI et le transistor.

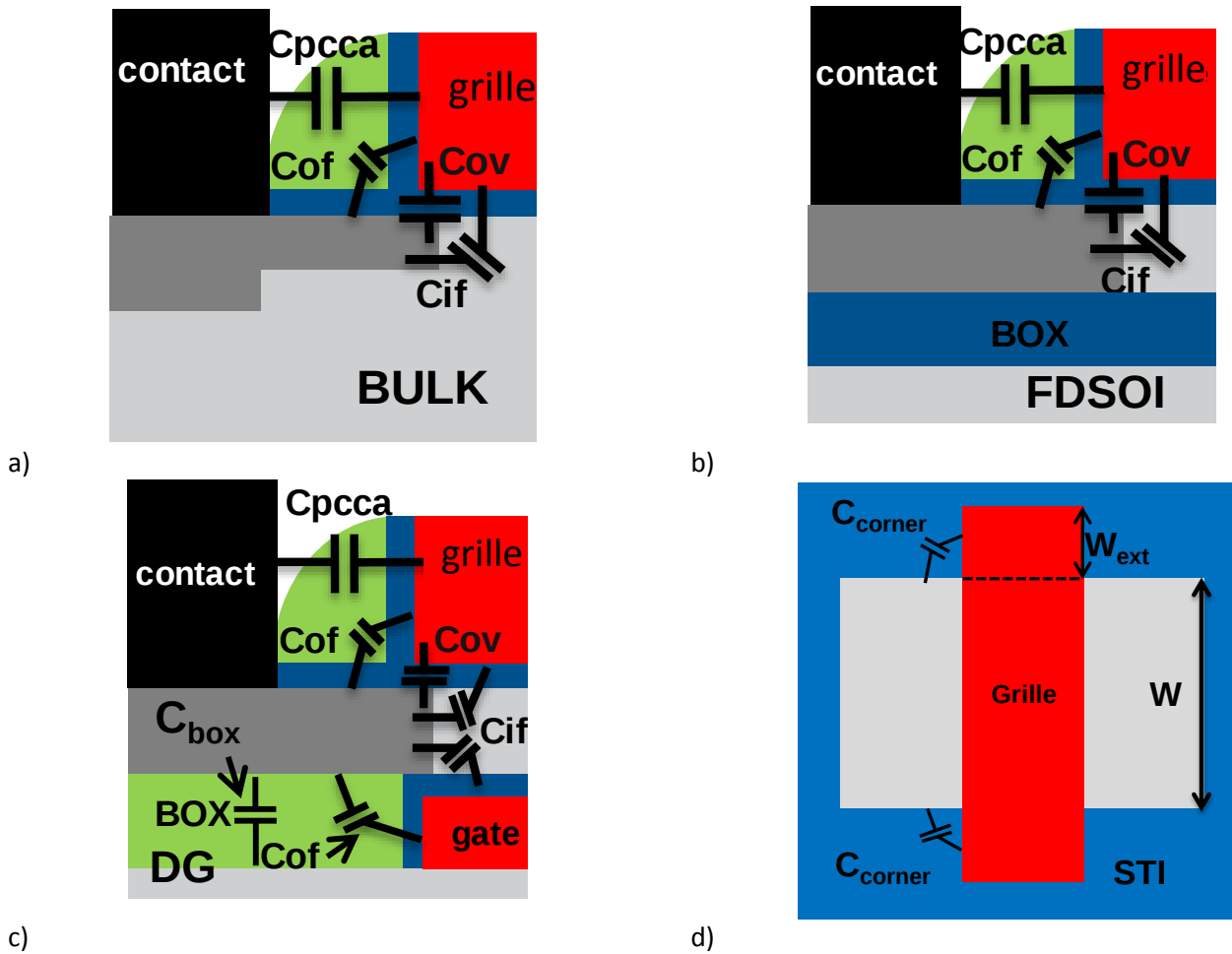


Figure III-9 : Représentation des capacités parasites sur les architectures planaires (a) architecture conventionnelle, b) FDSOI, c) double grille planaire et d) vue de dessus commune à chaque architecture planaire) sans source-drain surélevé pour assurer une lisibilité convenable. Ces derniers sont traités et représentés au paragraphe III.C.6.

La Figure III-10 représente les dimensions nécessaires au calcul des capacités parasites sur les architectures transistor sur substrat massif (a) et FDSOI (b), avec :

- L_g est la longueur de grille.
- CPP le Contacted Poly Pitch.
- dL longueur de recouvrement de la grille et des jonctions (overlap en anglais).
- t_{ox} l'épaisseur d'oxyde de grille.
- t_{sp} l'épaisseur de l'espaceur.
- H_g la hauteur de grille. Dans le cas du double grille on distinguera grille du haut et grille du bas en notant leur hauteur respectivement H_{gt} et H_{gb} .
- H_{M1} distance entre le haut de la grille et le premier niveau de métallisation.
- X_j la profondeur de jonction (pour l'architecture sur substrat massif uniquement).
- t_{si} l'épaisseur de film de silicium (pour le FDSOI et le DG).
- T_{box} l'épaisseur d'oxyde enterré (BOX) (pour le FDSOI et le DG).
- H_{epi} la hauteur d'épitaxie pour les source-drains surélevés.
- W_{ext} la largeur de l'extension de grille sur le STI (cf Figure III-9-d)
- W largeur du transistor.

De plus, les permittivités des matériaux sont nécessaires aux calculs ultérieurs ; nous noterons :

- ϵ_{ox} la permittivité du SiO_2
- ϵ_{spacer} la permittivité de l'espaceur.
- ϵ_{ox} la permittivité du silicium.
- ϵ_{PMD} la permittivité de l'oxyde de remplissage.
- ϵ_{BOX} la permittivité de l'oxyde enterré (BOX).

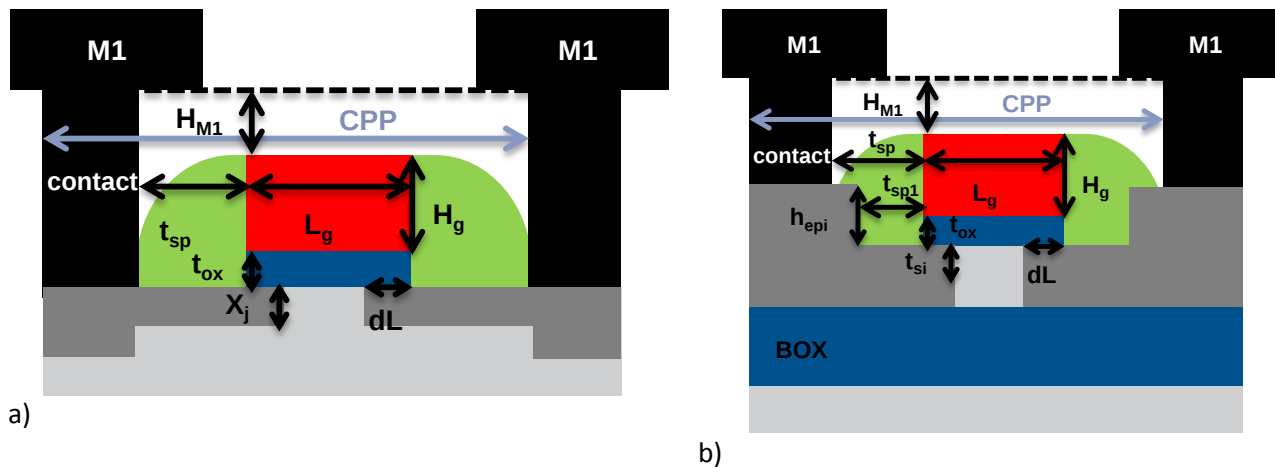


Figure III-10 : Représentation des dimensions sur l'architecture sur substrat massif conventionnelle (a) et sur l'architecture FDSOI (b). Les dimensions sur le double grille planaire sont semblables à celles du FDSOI (il suffit de définir en plus la hauteur de la grille du bas H_{gb}).

III.C.2. Capacité de recouvrement (C_{ov})

La capacité de recouvrement est évaluée comme une capacité à électrodes parallèles métalliques (car le niveau de dopage de l'extension source-drain est très élevé) :

$$C_{ov} = WdL \frac{\epsilon_{ox}}{EOT} \quad \text{Eq. III-35}$$

Où EOT est l'épaisseur équivalente d'oxyde de grille.

Cette expression est valable pour les architectures sur substrat massif et FDSOI. Pour un double grille planaire, il suffit de multiplier cette expression par deux, pour tenir compte de la deuxième grille, et on aura :

$$C_{ov} = 2WdL \frac{\epsilon_{ox}}{EOT} \quad \text{Eq. III-36}$$

Nous validons ce modèle par simulation numérique 2D effectuées avec FlexPDE [FlexPDE] (Figure III-11).

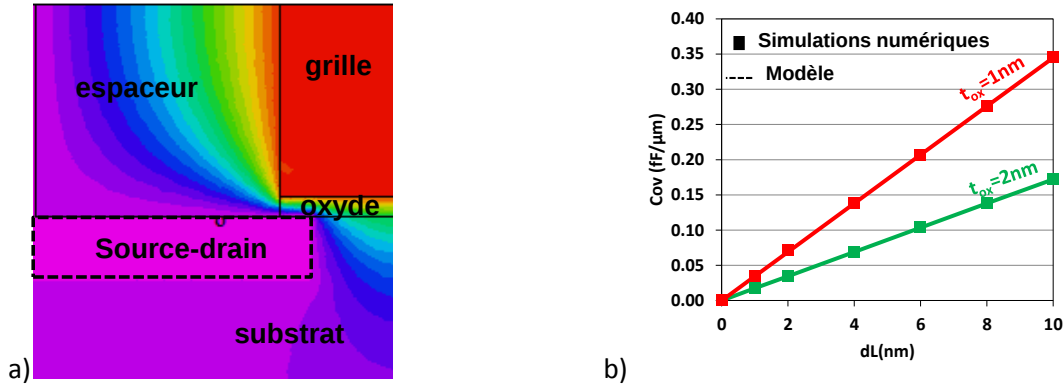


Figure III-11 : (a) Kit de simulation FlexPDE pour la validation du modèle de capacité de recouvrement C_{ov} . (b) Tracé de la variation de capacité de recouvrement en fonction de la longueur de recouvrement dL avec notre modèle et comparaison avec les simulations numériques.

III.C.3. Capacité de bords externes (C_{of})

La capacité de bords externes est une capacité entre deux électrodes perpendiculaires. Pour l'évaluer, il nous suffit donc d'utiliser l'équation III-34 avec $(x_1, x_2, y_1, y_2) = (0, t_{ox}, t_{sp}, H_g)$:

$$C_{of} = \frac{2}{\pi} W \epsilon_{spacer} \text{sh}^{-1} \sqrt{\frac{\min(H_g, t_{sp})^2 + 2 t_{ox} \min(H_g, t_{sp})}{t_{ox}}} + 0.35 \epsilon_{spacer} \frac{W}{\pi} \ln \left(\pi \frac{W}{t_{ox}} \right) \quad \text{Eq. III-37}$$

Nous validons ce modèle par simulations numériques FlexPDE avec le même kit utilisé dans le paragraphe III.C.2 pour la capacité de recouvrement C_{ov} , mais cette fois nous fixons la longueur de recouvrement à 0. La Figure III-12 montre que notre modèle reproduit bien les résultats de simulations en fonction de la hauteur de grille (b). Elle montre cependant également que le modèle reproduit convenablement les simulations pour les épaisseurs d'espaceur supérieures à 10nm (a). En dessous, l'écart entre modèle et simulation augmente du fait de la part de plus en plus importante prise par le terme de capacité due aux lignes de champ non elliptiques (Eq III-33). L'erreur reste tout de même raisonnable et sera négligeable par rapport aux autres capacités parasites lors de calcul de capacité de dispositifs totale ($t_{sp} < 10 \text{ nm} \rightarrow C_{of} \sim 0.05 \sim 0.08 \text{ fF}/\mu\text{m}$). Enfin, le graphe de la Figure III-12-b montre que notre modèle prend bien en compte la dépendance de la capacité de bords externes C_{of} avec la hauteur de grille, ce qui n'était pas le cas des précédentes études [Wei 11]. La cassure observée correspond à la frontière entre les dimensions où c'est la hauteur de grille qui limite la capacité de bords externes C_{of} et les dimensions où c'est l'épaisseur de l'espaceur qui la régit.

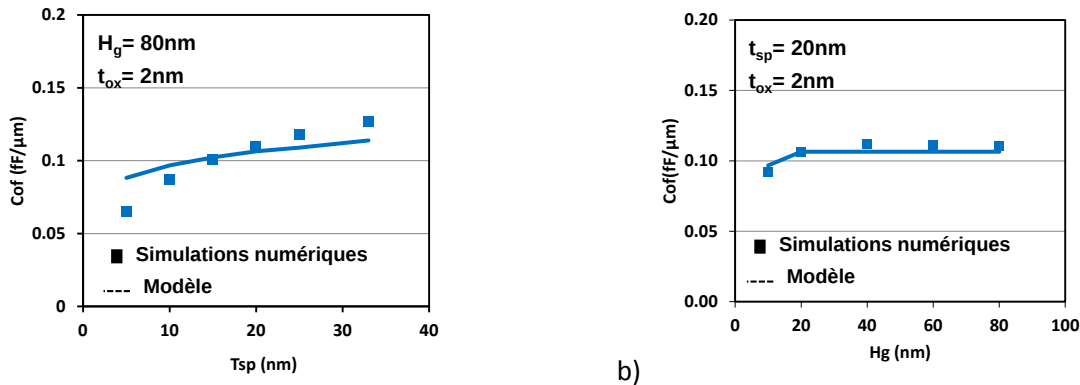


Figure III-12 : tracé de la capacité de bords externes C_{of} en fonction de l'épaisseur d'espaceur (a) et de hauteur de grille (b).

Cette équation est valable pour les architectures conventionnelles sur substrat massif (BULK) et FDSOI. Pour le double grille planaire, il faut tenir compte de la seconde grille, on aura alors :

$$C_{of} = C_{of_{top}} + C_{of_{bottom}} \quad \text{Eq. III-38}$$

Où $C_{of_{top}}$ (pour la grille du haut) est calculé exactement comme dans le cas du transistor sur substrat massif :

$$C_{of_{top}} = \frac{2}{\pi} W \epsilon_{\text{spacer}} \text{sh}^{-1} \left(\sqrt{\frac{\min(H_g, t_{\text{sp}})^2 + 2 t_{\text{ox}} \min(H_g, t_{\text{sp}})}{t_{\text{ox}}}} \right) + 0.35 \epsilon_{\text{spacer}} \frac{W}{\pi} \ln \left(\pi \frac{W}{t_{\text{ox}}} \right) \quad \text{Eq. III-39}$$

Et $C_{of_{bottom}}$ (pour la grille du bas) dépend de la hauteur de la grille du bas notée H_{gb} et de la longueur de jonction (en lieu et place de l'épaisseur d'espaceur) estimée par CPP- L_g :

$$C_{of_{bottom}} = \frac{2}{\pi} W \epsilon_{\text{spacer}} \text{sh}^{-1} \left(\sqrt{\frac{\min\left(H_{gb}, \frac{\text{CPP}-L_g}{2}\right)^2 + 2 t_{\text{ox}} \min\left(H_{gb}, \frac{\text{CPP}-L_g}{2}\right)}{t_{\text{ox}}}} \right) + 0.35 \epsilon_{\text{spacer}} \frac{W}{\pi} \ln \left(\pi \frac{W}{t_{\text{ox}}} \right) \quad \text{Eq. III-40}$$

III.C.4. Capacité de bords internes (C_{if})

La procédure de modélisation de la capacité de bords internes (C_{if}) est exactement la même que pour le calcul de la capacité de bords externes (C_{of}). En utilisant l'équation III-34 avec $(x_1, x_2, y_1, y_2) = (0, t_{\text{ox}}, L_{\text{elec}}/2, X_j)$ et en prenant pour permittivité de diélectrique la permittivité du silicium, on obtient :

$$C_{if} = \frac{2}{\pi} W \epsilon_{\text{si}} \text{sh}^{-1} \left(\sqrt{\frac{\min\left(\frac{L_{\text{elec}}}{2}, X_j\right)^2 + 2 t_{\text{ox}} \min\left(\frac{L_{\text{elec}}}{2}, X_j\right)}{t_{\text{ox}}}} \right) + 0.35 \epsilon_{\text{si}} \frac{W}{\pi} \ln \left(\pi \frac{W}{t_{\text{ox}}} \right) \quad \text{Eq. III-41}$$

Il a cependant été montré par [Pregaldiny 02] que la capacité de bords internes n'est pas constante et dépend fortement de la polarisation de grille. En effet, considérer C_{if} constant revient à considérer le silicium du canal du transistor comme un isolant parfait. En réalité, C_{if} est négligeable en accumulation et en inversion car elle est écrantée respectivement par les charges d'accumulation et d'inversion. Elle atteint ensuite son maximum, donnée par l'équation III-41, en régime de déplétion. Pour prendre en compte cet écrantage, nous utilisons les travaux de [Fleury 09] :

$$C_{if}(V_g) = \frac{C_{if_{\text{max}}} C_{\text{ox}}^2}{(C_{\text{ox}} + C_{\text{gc}}(V_g) - C_{\text{min}})^2 + C_{if_{\text{max}}} (C_{\text{gc}}(V_g) - C_{\text{min}})} \quad \text{Eq. III-42}$$

Où $C_{if_{\text{max}}}$ est donné par l'équation III-41, C_{ox} est la capacité de grille en inversion forte, C_{min} la valeur minimum de la capacité de grille et $C_{\text{gc}}(V_g)$ la capacité grille-canal. Pour pouvoir valider cette approche, nous commençons par simuler et extraire la capacité $C_{\text{gc}}(V_g)$ avec le kit de simulation FlexPDE représenté sur la Figure III-13-a. Ensuite, nous utilisons l'équation III-42 où $C_{if_{\text{max}}}$ est donné par l'équation III-41 et $C_{\text{gc}}(V_g)$ et C_{min} extraits des simulations. Enfin, nous comparons la courbe obtenue aux simulations sur la Figure III-13-b, ce qui démontre que cette prise en compte de l'écrantage de la capacité de bords internes est correcte.

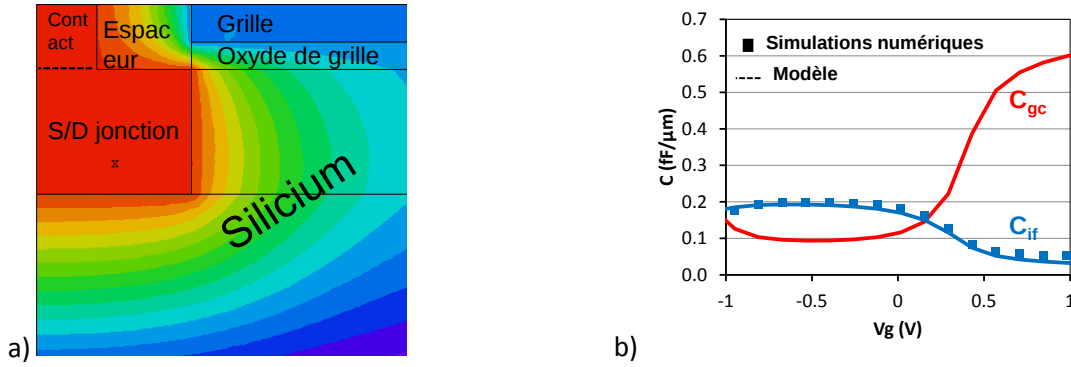


Figure III-13 : (a) Kit de simulation FelxPDE. (b) tracé de la capacité de bords internes (C_{if}) par simulation numérique et avec notre modèle. La capacité grille canal $C_{gc}(V_g)$ obtenue par simulation est également tracé pour démontrer que la valeur max de C_{if} est bien en régime de déplétion.

Cette expression peut être utilisée pour l'architecture FDSOI en remplaçant X_j par t_{si} et pour le double grille planaire en remplaçant X_j par $t_{si}/2$ et en multipliant la valeur obtenue par 2 (pour tenir compte de la seconde grille).

III.C.5. Capacité grille-contact (C_{pcca})

Dans les technologies CMOS actuelles, deux schémas de contact coexistent. En effet jusqu'au nœud technologique 45nm, toutes les plateformes CMOS étaient construites avec des contacts à prise, comme représenté sur la Figure III-14-b. Puis, à compter du 45nm, certains ont fait le choix du contact en ruban (la largeur du contact est égale à la largeur du transistor) pour les technologies hautes performances [Auth 08], contrairement aux technologies basse consommation [Josse 06] qui ont conservé le contact à prise pour le nœud suivant également (32-28nm) [Planes 12] [Arnaud 09]. Au nœud 20nm aura lieu un consensus pour ce second schéma de contact, quelle que soit l'architecture des transistors choisie et l'application visée [Auth 12] [Shang 12]. Les deux paragraphes suivants proposent la modélisation des capacités parasites pour les deux schémas de contacts, alors que les précédents travaux sur les capacités parasites [Wei 11] se sont concentrés sur la modélisation du contact ruban.

Lorsque la grille fait face au contact, la capacité grille contact est divisée en deux composantes (Figure III-14-a) :

- $C_{pccaflat}$ est la capacité entre le flan de grille et le contact. Il s'agit d'une capacité à électrodes parallèles.
- $C_{pccatop}$ est la capacité entre le dessus de la grille et le contact. Il s'agit d'une capacité à électrodes perpendiculaires.

Ces deux composantes sont suffisantes pour modéliser les capacités parasites pour le contact en ruban, pour les contacts à prise il faudra également traiter les zones où il n'y a pas de contact face à la grille.

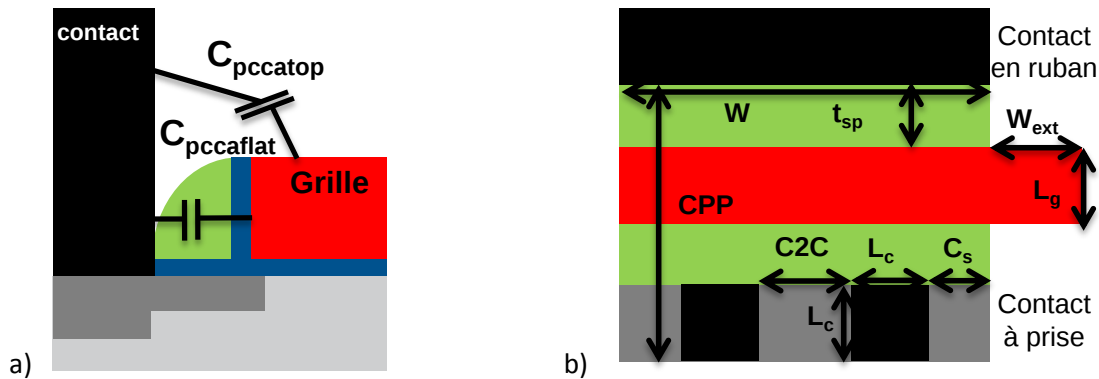


Figure III-14 (a) Vue schématique en coupe dans un contact et représentation des deux composantes de la capacité grille-contact. (b) Représentation en vue de dessus des deux schémas de contact modélisés.

Les expressions développées ci-dessous seront communes aux trois architectures planaires étudiées dans cette partie.

III.C.5.a) Contact en ruban

Pour la composante entre le flan de grille et le contact, on utilise la formule classique du condensateur plan :

$$C_{pccaflat} = WH_g \frac{\epsilon_{spacer}}{t_{sp}} \quad \text{Eq. III-43}$$

Où W est la largeur du transistor, mais également ici la largeur du contact.

Pour la seconde composante, on travaille par analogie avec la capacité de bords externes (C_{of}). En utilisant l'équation III-34 avec $(x_1, x_2, y_1, y_2) = (0, t_{sp}, L_g/2, H_{M1})$:

$$C_{pccatop} = \frac{2}{\pi} W \epsilon_{pmd} \operatorname{sh}^{-1} \left(\sqrt{\frac{\min\left(H_{M1}, \frac{L_g}{2}\right)^2 + 2t_{sp} \min\left(H_{M1}, \frac{L_g}{2}\right)}{t_{sp}}} \right) + 0.35 \cdot \epsilon_{pmd} \frac{W}{\pi} \ln \left(\pi \frac{L_g}{2 t_{sp}} \right) \quad \text{Eq. III-44}$$

Pour valider notre approche, nous utilisons à nouveau FlexPDE avec les kits de simulation représentés sur la Figure III-15. Le premier nous permet d'extraire la somme $C_{of} + C_{pccaflat}$ et le second nous permet d'obtenir la somme $C_{of} + C_{pccaflat} + C_{pccatop}$.

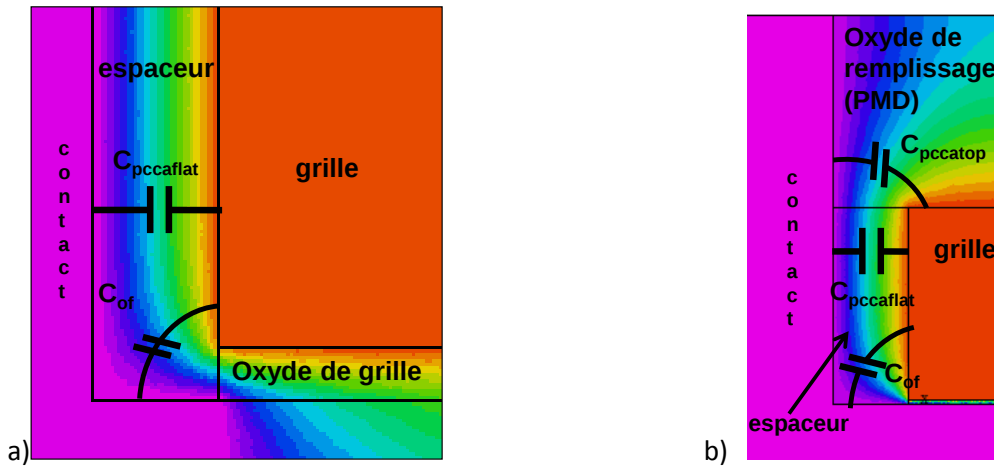


Figure III-15 : (a) kit de simulation pour la validation de la capacité grille-contact, sans la composante $C_{pccatop}$. (b) kit de simulation pour la validation de la capacité grille-contact complète.

On peut alors tracer les variations de ces sommes avec l'épaisseur d'espaceur (Figure III-16-a) et avec la hauteur de grille (Figure III-16-b).

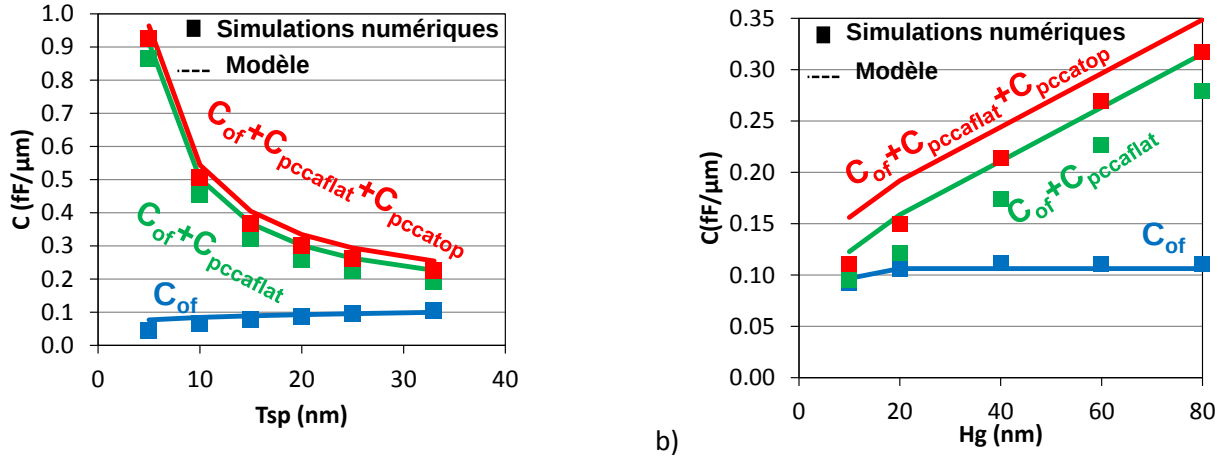


Figure III-16 : variations des capacités parasites entre la grille et le contact en fonction de l'épaisseur d'espaceur (a) pour $t_{ox}=2nm$, $H_g=80nm$ et $L_g=40nm$ puis la hauteur de grille (b) pour $t_{ox}=2nm$, $t_{sp}=20nm$ et $L_g=40nm$ tracé par simulations numériques puis avec notre modèle.

On constate que notre modèle reproduit convenablement la tendance pour les variations d'espaceur et pour une hauteur de grille importante. Cependant, quand on observe le tracé en fonction de la hauteur de grille, on remarque que le modèle reproduit bien la pente, mais qu'il y a un décalage entre les valeurs obtenues par modèle et par simulation. Les précédentes études sur ce sujet [Wei 11] ont dû rencontrer le même problème (car ils utilisent les mêmes expressions), mais les différentes publications ne montrent pas de graphes de variation de la capacité grille-contact en fonction de la hauteur de grille. En observant de plus près les surfaces iso-potentielles dans cette structure (Figure III-15), on remarque que ces dernières ne sont pas parallèles au contact du haut de la grille à l'oxyde de grille. En effet, les surfaces iso-potentielles deviennent curvilignes lorsqu'on est suffisamment proche de l'oxyde de grille, et, dans ce cas, c'est la capacité de bords externes $C_{of}(0)$ qui domine. On estime la distance entre l'oxyde de grille et la frontière entre ces deux zones à la moitié de l'épaisseur de l'espaceur (illustrée sur la Figure III-17-a). Par conséquent, la composante $C_{pccaflat}$ doit être corrigée comme suit :

$$C_{pccaflat} = W(H_g - \frac{t_{sp}}{2}) \frac{\epsilon_{spacer}}{t_{sp}} \quad \text{Eq. III-45}$$

En comparant cette nouvelle expression aux mêmes simulations que dans la Figure III-16, on constate que notre modèle corrigé reproduit bien mieux les simulations.

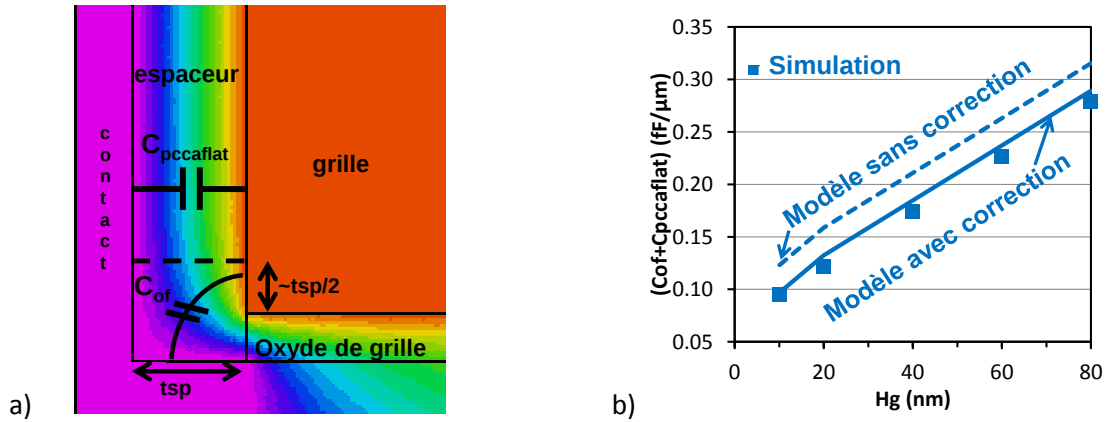


Figure III-17 : (a) illustration de la correction qui doit être appliquée pour le calcul de la composante $C_{pccaflat}$. (b) tracé de la variation de la somme $C_{pccaflat} + C_{of}$ en fonction de la hauteur de grille avec et sans correction due à la capacité de bords externes C_{of} .

Pour valider définitivement ce modèle, nous le comparons à des simulations numériques pour différentes géométries sur la Figure III-18.

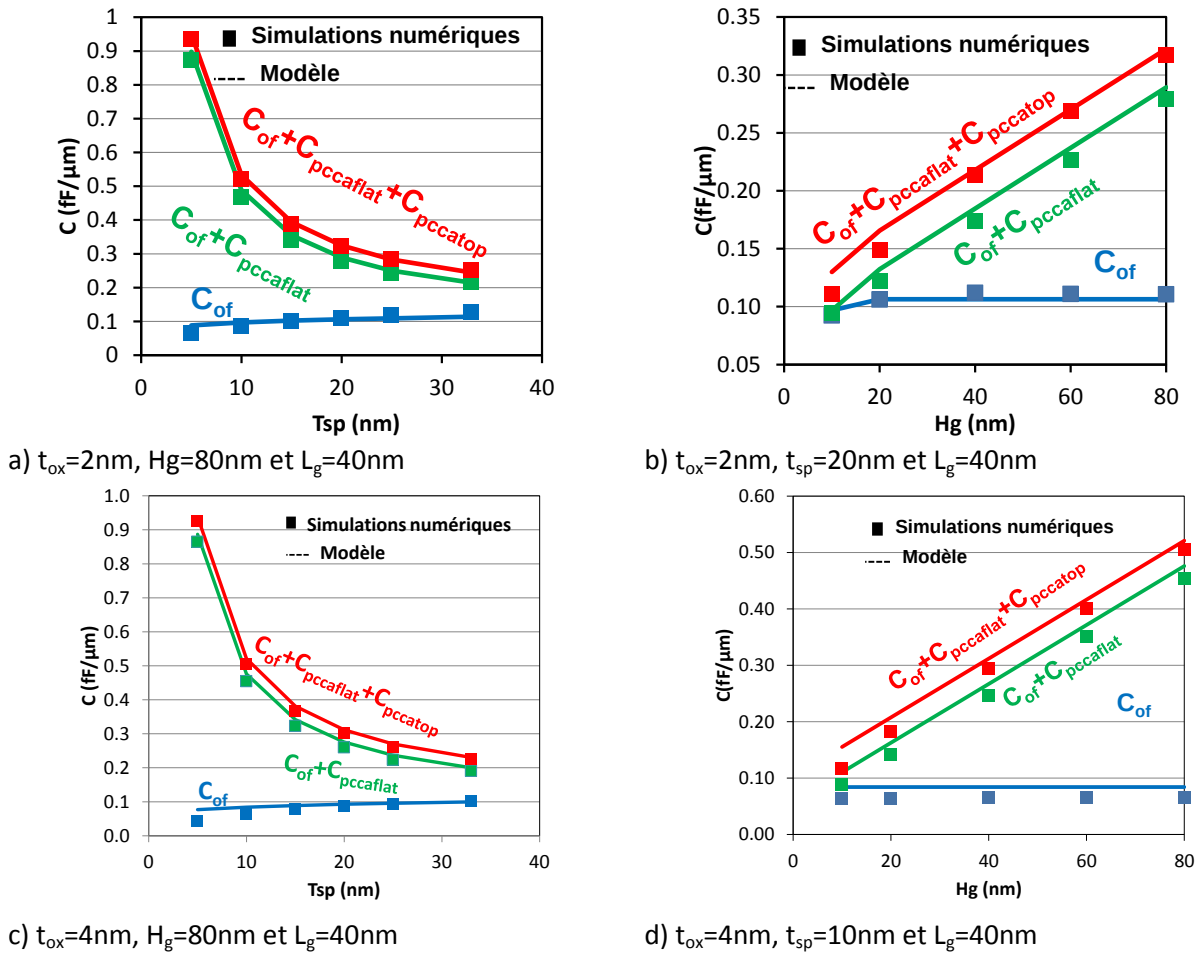


Figure III-18: Variations des capacités parasites entre la grille et le contact en fonction de l'épaisseur d'espaceur (a,c) puis la hauteur de grille (b,d) obtenues par simulations numériques et par le modèle corrigé.

III.C.5.b) Contact à prise

Pour la modélisation du schéma de contact à prise (plug en anglais), nous considérons pour simplifier les calculs que chaque contact est un carré de côté L_c (Figure III-14) alors qu'en réalité ils sont plutôt de forme circulaire ou elliptique. Ensuite, nous distinguons deux régions :

- Région 1 : La grille fait face au contact, la composante $C_{pccaflat}$ donnée par l'équation III-45 peut être utilisé en remplaçant la largeur du transistor W par la largeur du contact L_c (Figure III-14).
- Région 2 : La grille fait face à un vide entre deux contacts et il existe une nouvelle capacité parasite $C_{pccafringe}$ qui est formée de deux électrodes perpendiculaires (la grille et le flanc du contact). Pour le contact situé au bord de l'active, cette capacité sera nommée $C_{pccafringe_edge}$ (Figure III-19, a).

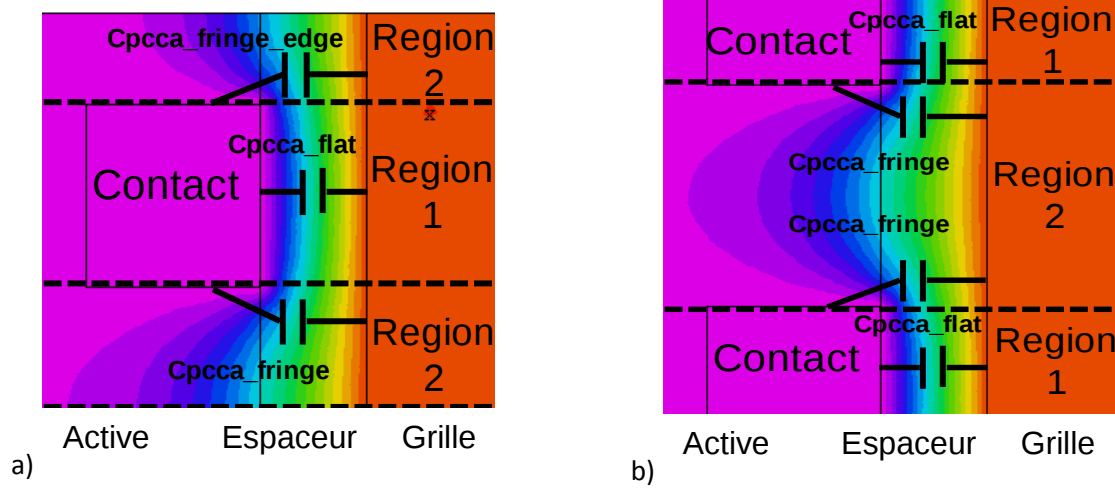


Figure III-19 : Illustration des deux régions (a) centrée sur Région 1, b) centrée sur Région 2) à modéliser en vue de dessus pour les contacts à prise et représentation des composantes de capacités parasites sur une cartographie de surfaces iso-potentielles réalisée avec FlexPDE.

Par conséquent, la capacité pour la région 1 sera divisée en deux composantes, obtenues par analogie avec la modélisation du contact en ruban (0, en remplaçant la largeur du dispositif par la largeur du contact L_c :

$$C_{pccaflat} = L_c \left(H_g - \frac{t_{sp}}{2} \right) \frac{\epsilon_{spacer}}{t_{sp}} \quad \text{Eq. III-46}$$

$$C_{pcca_{top}} = \frac{2}{\pi} L_c \epsilon_{pmd} \text{sh}^{-1} \left(\sqrt{\frac{\min(H_{M1}, \frac{L_g}{2})^2 + 2t_{sp} \min(H_{M1}, \frac{L_g}{2})}{t_{sp}}} \right) + 0.35 \cdot \epsilon_{pmd} \frac{L_c}{\pi} \ln \left(\pi \frac{\frac{L_g}{2}}{t_{sp}} \right) \quad \text{Eq. III-47}$$

Pour la région 2, nous utilisons l'équation III-34 avec, d'après le schéma de la Figure III-14-b), $(x_1, y_1, x_2, y_2) = (0, t_{sp}, C_s, L_c)$ pour $C_{pccafringe_edge}$ et $(x_1, y_1, x_2, y_2) = (0, t_{sp}, C2C/2, L_c)$ pour $C_{pccafringe}$, ce qui donne formellement :

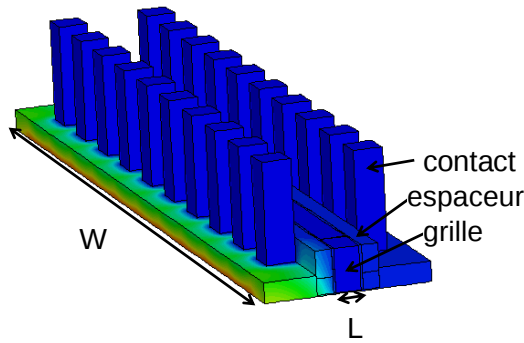
$$C_{pccafringe_edge} = \frac{2}{\pi} H_g \epsilon_{spacer} \cdot \text{sh}^{-1} \left(\sqrt{\frac{\min(C_s, L_c)^2 + 2t_{sp} \min(C_s, L_c)}{t_{sp}}} \right) + 0.35 \cdot \epsilon_{spacer} \frac{H_g}{\pi} \ln \left(\pi \frac{\frac{H_g}{2}}{t_{sp}} \right) \quad \text{Eq. III-48}$$

$$C_{pccafringe} = \frac{2}{\pi} H_g \epsilon_{spacer} \cdot \text{sh}^{-1} \left(\sqrt{\frac{\min(\frac{C_2 C}{2}, L_c)^2 + 2 t_{sp} \min(\frac{C_2 C}{2}, L_c)}{t_{sp}}} \right) + 0.35 \cdot \epsilon_{spacer} \frac{H_g}{\pi} \ln \left(\pi \frac{H_g}{t_{sp}} \right) \quad \text{Eq. III-49}$$

La capacité grille-contact totale est finalement :

$$C_{pcca} = N_c C_{pccaflat} + N_c C_{pccatop} + 2(N_c - 1) C_{pccafringe} + 2 C_{pccafringeedge} \quad \text{Eq. III-50}$$

Où N_c est le nombre de contacts. Pour valider cette approche, la simulation 2D n'est pas suffisante et nous choisissons alors l'outil Raphael [Raphael] pour effectuer des simulations 3D. Comme ce type de simulation est assez lourd, nous n'avons fait des variations que du nombre de contacts (les autres dépendances comme la distance grille-contact peuvent être considérées comme validées dans le paragraphe III.C.5.a)). Le kit de simulations utilisé est représenté sur la : (a) kit de simulation Raphael utilisé pour valider notre modèle de capacité grille contact dans le cas de contacts à prise. (b) Comparaison des valeurs de capacités grille-contact obtenues par simulations Raphael et par notre modèle analytique pour différents nombre de contact. Figure III-20-a et la comparaison entre modèle et simulation est résumée dans le tableau de la Figure III-20-b. On constate que notre modèle reproduit bien les simulations.



$L=46\text{nm} / t_{sp}=33\text{nm}$	C_{pcca} (modèle)	C_{pcca} (simulation)
N=1	1.49e-17 F	1.64e-17 F
N=2	2.98e-17 F	3.28e-17 F
N=3	4.47e-17 F	4.92e-17 F
N=5	7.45e-17 F	8.20e-17 F
N=10	1.49e-16 F	1.64e-16 F

b) ($t_{sp}=33\text{nm}$, $C_2C=72\text{nm}$, $H_g=80\text{nm}$, $C_s=37\text{nm}$ et $L=46\text{nm}$).

a)

Figure III-20 : (a) kit de simulation Raphael utilisé pour valider notre modèle de capacité grille contact dans le cas de contacts à prise. (b) Comparaison des valeurs de capacités grille-contact obtenues par simulations Raphael et par notre modèle analytique pour différents nombre de contact.

III.C.6. Capacité grille-source-drain epitaxiés (C_{gepi})

Cette capacité est due à l'utilisation de source-drain surélevés, nécessaires pour les architectures à film mince pour des raisons de résistance d'accès et de siliciuration mais également utilisées dans l'architecture sur substrat massif pour induire de la contrainte mécanique et améliorer le transport. Comme représenté sur la Figure III-21, cette capacité est divisée en deux composantes :

- C_{gepi1} : capacité entre le flan de grille et le flanc de l'épitaxie, il s'agit donc d'une capacité à électrodes parallèles
- C_{gepi2} : la capacité entre le flan de grille et le dessus de l'épitaxie, il s'agit d'une capacité à électrodes perpendiculaires.

On considère dans cette partie que les deux espaceurs sont constitués du même matériau (donc même permittivité), et que l'épitaxie est droite (ne présente pas de facettes dues aux plans de croissance cristalline). Une modélisation plus réaliste, mais aussi plus complexe, sera proposée au paragraphe III.E. Les expressions de capacité de ce paragraphe sont valables pour les trois architectures planaires.

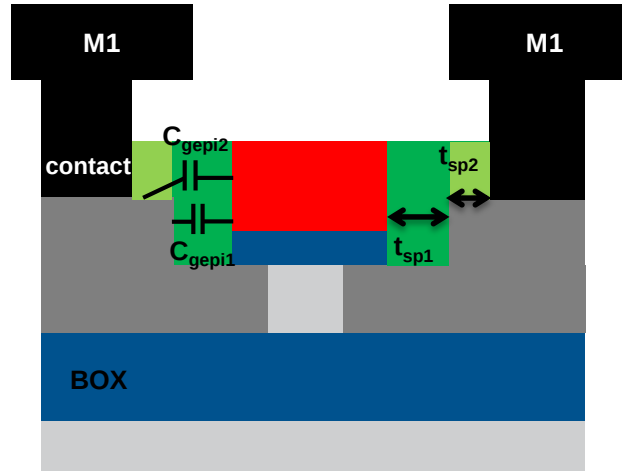


Figure III-21 : Représentation de la capacité grille-épitaxie et de ses dimensions spécifiques.

Par un même raisonnement que pour la capacité grille-contact dans le cas d'un contact en ruban (III.C.5.b), l'épitaxie jouant le rôle de la grille et la grille le rôle de du contact, on obtient facilement :

$$C_{gepi1} = \frac{W \left(H_{epi} - t_{ox} - \frac{t_{sp1}}{2} \right) \epsilon_{spacer}}{t_{sp1}} \quad \text{Eq. III-51}$$

$$C_{gepi2} = \frac{2}{\pi} W \epsilon_{spacer} \operatorname{sh}^{-1} \left(\sqrt{\frac{\min(H_g - H_{epi}, t_{sp2})^2 + 2 t_{sp1} \min(H_g - H_{epi}, t_{sp2})}{t_{sp1}}} \right) + 0.35 \epsilon_{spacer} \frac{W}{\pi} \ln \left(\pi \frac{W}{t_{sp1}} \right) \quad \text{Eq. III-52}$$

La Figure III-22-a représente le kit de simulation FlexPDE utilisé pour valider notre modèle. Il n'est pas possible de tracer la variation de la capacité grille-épitaxie car les capacités C_{pcca} et C_{of} ont un impact sur la forme des lignes de champs et donc sur la valeur de la capacité grille-épitaxie C_{gepi} . C'est la raison pour laquelle nous comparons la capacité totale ($C_{of} + C_{pcca} + C_{gepi}$) obtenue par modélisation et par simulation sur la Figure III-22. Comme les expressions de C_{of} et C_{pcca} sont validées dans le paragraphe précédent (III.C.6, nous validons bien notre modèle de capacité grille-épitaxie pour les variations des épaisseurs des deux espaceurs et de la hauteur d'épitaxie).

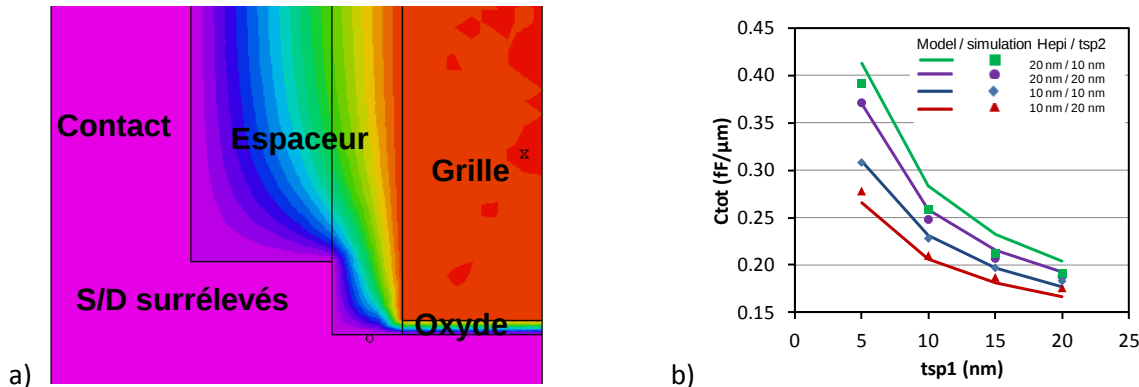


Figure III-22 : (a) Kit de simulation FlexPDE. (b) Tracé de la capacité totale ($C_{of} + C_{pcca} + C_{gepi}$) en fonction de l'épaisseur du premier espaceur pour différentes valeurs de hauteur d'épitaxie et d'épaisseur du second espaceur.

III.C.7. Capacité de coin (C_{corner})

La capacité de coin C_{corner} est la capacité due à l'extension de grille et sur le STI, et est représentée en vue de dessus sur la Figure III-9. Cette dernière est faible et n'a pas d'impact pour les dispositifs larges (i.e. $W > \sim 100\text{nm}$). Cependant, pour les nœuds technologiques avancés, les transistors vont être de plus en plus étroits et la valeur de C_{corner} ne sera plus négligeable. Cette capacité est modélisée dans [Wei 11], mais l'expression proposée tend vers l'infini quand W_{ext} tend vers l'infini, ce qui n'est pas réaliste. Afin de proposer une modélisation fiable et assez simple, nous décomposons C_{corner} en quatre composantes (Figure III-23) :

- $C_{\text{corner_SD}}$ est la capacité entre les source-drains et l'extension de grille. C'est l'unique composante que [Wei 11] tente de modéliser.
- $C_{\text{corner_G}}$ est la capacité entre le bas de l'extension de grille et le canal du transistor à travers le STI.
- $C_{\text{corner_contact}}$ est la capacité entre le flanc de l'extension de grille et le contact.
- $C_{\text{corner_contact_top}}$ est la capacité entre le dessus de l'extension de grille et le contact.

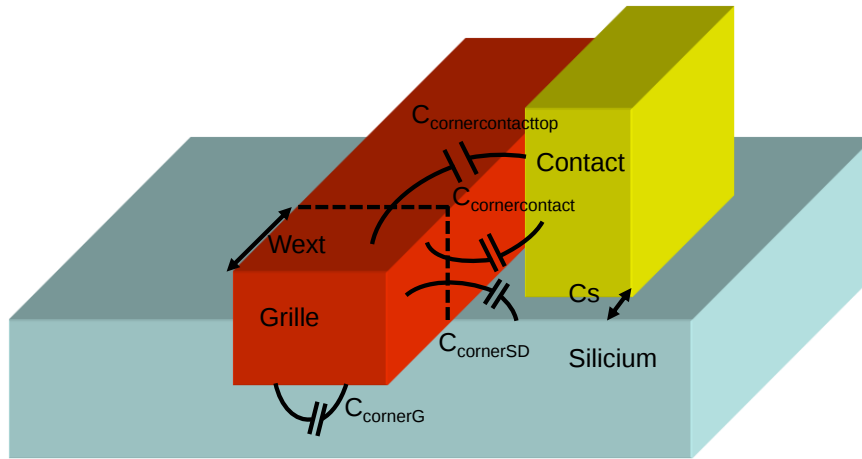


Figure III-23 : représentation schématique des quatre composantes de la capacité de coin C_{corner} .

Nous commençons par modéliser la composante $C_{\text{corner_contact}}$ où l'équation III-34 peut être utilisée directement avec $(x_1, y_1, x_2, y_2) = (t_{\text{sp}}, C_s, W_{\text{ext}}, C_s + W_{\text{ext}})$, ce qui donne :

$$C_{\text{corner_contact}} = \frac{2}{\pi} H_g \epsilon_{\text{spacer}} \left(\text{sh}^{-1} \left(\sqrt{\frac{t_{\text{sp}}^2 + 2(C_s + W_{\text{ext}})C_s + (C_s + W_{\text{ext}})^2}{t_{\text{sp}}^2 - C_s^2}} \right) - \text{sh}^{-1} \left(\sqrt{\frac{t_{\text{sp}}^2}{t_{\text{sp}}^2 - C_s^2}} \right) \right) \quad \text{Eq. III-53}$$

Pour la composante $C_{\text{corner_G}}$ nous procédons de la même manière, mais cette fois avec $(x_1, y_1, x_2, y_2) = (0, t_{\text{ox}}, W_{\text{ext}}, W_{\text{ext}})$:

$$C_{\text{corner_G}} = \frac{2}{\pi} L \epsilon_{\text{STI}} \text{sh}^{-1} \left(\sqrt{\frac{W_{\text{ext}}^2 + 2W_{\text{ext}}t_{\text{ox}}}{t_{\text{ox}}}} \right) + 0.35 \cdot \epsilon_{\text{STI}} \frac{W}{\pi} \ln \left(\pi \frac{W}{t_{\text{ox}}} \right) \quad \text{Eq. III-54}$$

Pour les deux dernières composantes, la transformation conforme ne peut être utilisée directement car les électrodes ne sont pas dans le même plan. Nous choisissons alors d'adapter cette méthode. Pour la composante C_{cornerSD} , nous divisons l'extension de grilles en n tranches (découpage illustré sur la Figure III-24-a) et nous appliquons la transformation conforme sur chaque tranche. Il nous suffit ensuite d'ajouter les valeurs de capacité obtenue sur chaque tranche. Avec H_g/n étant l'épaisseur d'une tranche élémentaire, et en adaptant la méthode décrite dans la partie III.B.3 en utilisant l'équation III-34 avec $((x_1, y_1, x_2, y_2) = (iH_g/n, t_{\text{ox}}, W_{\text{ext}}, W_{\text{ext}}))$, on obtient :

$$C_{\text{cornerSD}} = \sum_i \frac{2}{\pi} \frac{H_g}{n} \epsilon_{\text{spacer}} \cdot \text{sh}^{-1} \left(\sqrt{\frac{\left(i \frac{H_g}{n}\right)^2 + 2 t_{\text{ox}} W_{\text{ext}} + W_{\text{ext}}^2}{\left(i \frac{H_g}{n}\right)^2 - t_{\text{ox}}^2}} \right) \quad \text{Eq. III-55}$$

Où i est un nombre entier, plus i est grand, meilleure est la précision.

$C_{\text{cornercontacttop}}$ peut être modélisé par une méthode similaire au calcul de C_{cornerSD} . Nous choisissons de couper l'extension de grilles également en n tranches mais cette fois dans le sens de l'extension de grille (Figure III-24-b), donc l'épaisseur d'une tranche sera (W_{ext}/n) . Ensuite, nous utilisons le théorème de Pythagore pour définir $H_{\text{mineff}}(i)$ la distance la plus courte, pour une tranche donnée du contact au-dessus de la grille et $H_{\text{maxeff}}(i)$ la dimension effective d'une tranche. Nous utilisons alors l'équation III-34 avec $(x_1, y_1, x_2, y_2) = (0, H_{\text{mineff}}(i), H_{\text{maxeff}}(i), H_{\text{maxeff}}(i))$:

$$C_{\text{cornercontacttop}} = \sum_i 2 \frac{2}{\pi} \frac{W_{\text{ext}}}{n} \epsilon_{\text{spacer}} \text{sh}^{-1} \left(\sqrt{\frac{H_{\text{maxeff}}^2 + 2 H_{\text{maxeff}} H_{\text{mineff}}}{H_{\text{mineff}}}} \right) \quad \text{Eq. III-56}$$

$$H_{\text{mineff}}(i) = \sqrt{\left(C_s + i \frac{W_{\text{ext}}}{n}\right)^2 + t_{\text{sp}}^2} \quad \text{Eq. III-57}$$

$$H_{\text{maxeff}}(i) = \sqrt{\left(C_s + i W_{\text{ext}}/n\right)^2 + (t_{\text{sp}} + L/2)^2} - H_{\text{mineff}}(i) \quad \text{Eq. III-58}$$

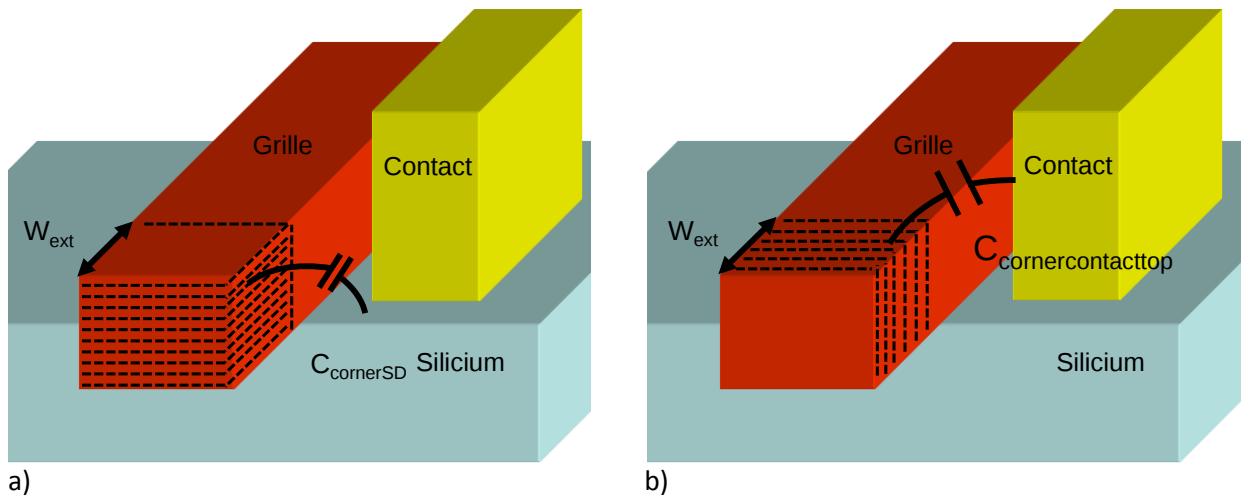
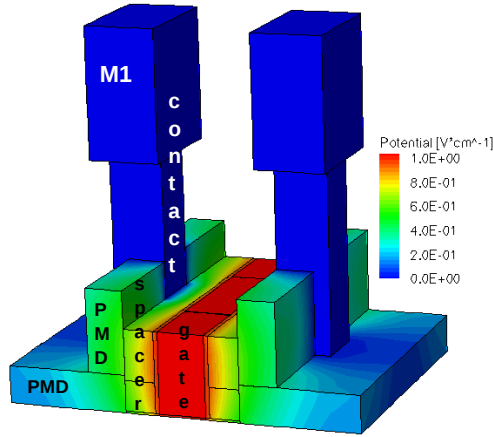


Figure III-24: découpage de l'extension de grille pour le calcul de la composante C_{cornerSD} (a) et de la composante $C_{\text{cornercontacttop}}$ (b).

La capacité de coin C_{corner} totale est donc donnée par la somme des quatre composantes décrites ci-dessus :

$$C_{\text{corner}} = C_{\text{corner}_{\text{SD}}} + C_{\text{corner}_{\text{G}}} + C_{\text{corner}_{\text{contact}}} + C_{\text{corner}_{\text{contacttop}}} \quad \text{Eq. III-59}$$

Cette équation est utilisable pour les trois architectures planaires étudiée. Nous comparons les résultats obtenus avec cette équation à ceux obtenus par simulations numériques 3D Raphael [Raphael] (Figure III-25). Le tableau de cette même figure nous montre que le modèle proposé dans cette partie donne une bonne estimation de la capacité de coin C_{corner} , avec un écart entre valeurs obtenues par modèle et par simulation inférieur à 10%.



$W_{\text{ext}}=70\text{nm}$ $C_s = 30 \text{ nm}$	C_{corner} (modèle)	C_{corner} (simulation)	Erreur
$L=46\text{nm}$	$1.30\text{e-}17 \text{ F}$	$1.37\text{e-}17 \text{ F}$	5%
$L=406\text{nm}$	$5.21\text{e-}17 \text{ F}$	$5.21\text{e-}17 \text{ F}$	0%
$L=790\text{nm}$	$8.63\text{e-}17 \text{ F}$	$9.66\text{e-}17 \text{ F}$	10%

b)

a)

Figure III-25 : (a) Kit de simulation 3D Raphael. (b) Comparaison des valeurs de C_{corner} obtenues par simulations numériques et par notre modèle analytique.

III.C.8. Capacité de jonction ou d'oxyde enterré (C_j ou C_{box})

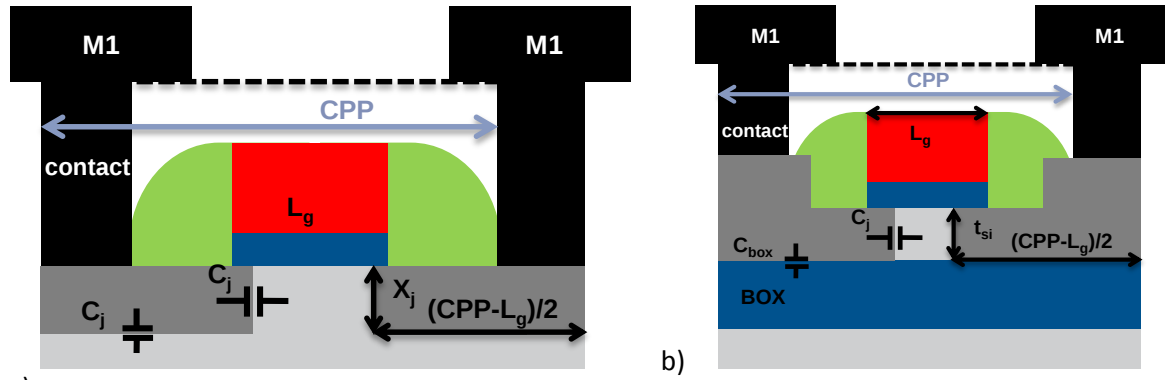
Dans le cas du transistor sur substrat massif, la capacité de jonction est évaluée avec l'équation classique d'une jonction PN, donnée dans [Wei 11]. Elle se décompose en deux composantes (représentées sur la Figure III-26-a) ; ce qui donne formellement :

$$C_j = W \left(\left(X_j + \frac{1}{2} (CPP - L_g) \right) \sqrt{\frac{qE_{si}N_{sd}N_b}{(N_{sd}+N_b)}} \frac{1}{\sqrt{\varphi_d + V_{ds} - 2\frac{kT}{q}}} \right) \quad \text{Eq. III-60}$$

N_{sd} est la valeur du dopage dans les LDD et N_b le dopage du substrat.

Pour les transistors FDSOI et double grille (Figure III-26-b), les canaux sont non dopés, donc la capacité de jonction entre le canal et la source ou le drain est très faible. La capacité de jonction dans ces architectures est alors réduite une capacité d'oxyde enterré (BOX) car les source-drains sont isolés du substrat par le BOX. On a alors :

$$C_{\text{box}} = W \frac{(CPP - L_g)}{2} \frac{\epsilon_{\text{BOX}}}{T_{\text{BOX}}} \quad \text{Eq. III-61}$$



a)
Figure III-26 : représentation des composantes de capacité de jonction pour le transistor conventionnel sur substrat massif (a) et pour le FDSOI (b).

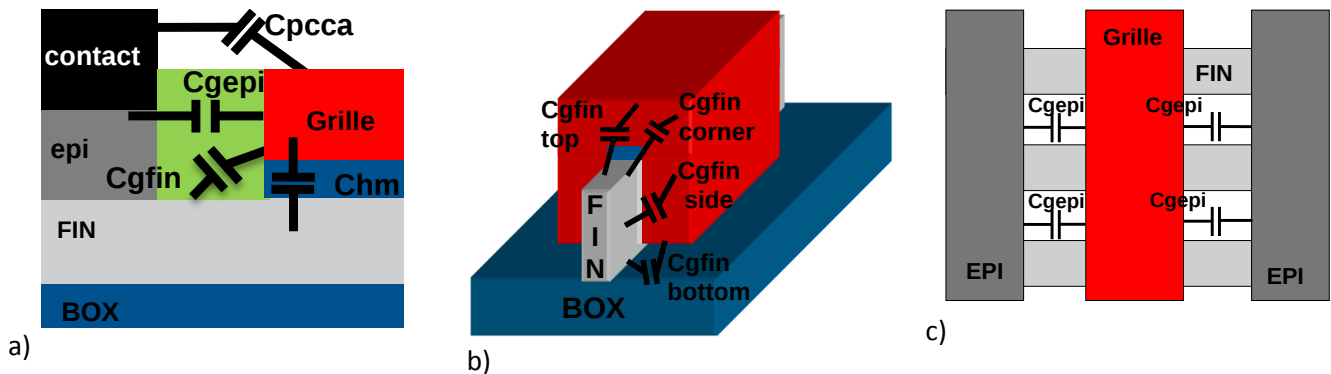
III.D. Capacités parasites sur un dispositif non-planaire

III.D.1. Composantes parasites sur structures 3D

La question de la modélisation des capacités parasites sur les dispositifs non plans a d'ores et déjà été traitée notamment par [Wu 07]. Cependant, les auteurs ont utilisé la même méthode que [Bansal 05] et [Wei 11], leurs formulations sont donc imprécises. Enfin, dans cette publication toutes les capacités parasites nécessaires à une bonne évaluation de performance dynamique n'ont pas été modélisées.

Pour modéliser les capacités parasites sur les dispositifs non plans, nous allons utiliser la même méthode que pour les dispositifs plans, en commençant par identifier les différentes composantes. La Figure III-27 représente les différentes capacités sur un FinFET sur SOI (cf paragraphe I.E.3 si le dispositif est sur substrat massif, la seule différence en termes de capacités parasites résidera dans les capacités de jonction) :

- C_{gfin} est la capacité entre la grille et le fin à travers l'espaceur. Elle est semblable à la capacité de bords externes C_{of} dans les transistors plans, mais est divisée en quatre composantes distinctes.
- C_{gepi} est la capacité entre la grille et l'épithésie source-drain à travers l'espaceur.
- C_{pcca} est la capacité entre la grille et le contact.
- C_{hm} est la capacité entre la grille et le fin à travers les masque dur (HM : Hard Mask en anglais). Cette capacité est propre au FinFET (double grille). Pour le Trigate, le dessus du fin est recouvert de l'empilement diélectrique de grille métal, et la capacité compte donc comme de la capacité grille classique.
- C_{ov} est la capacité de recouvrement (overlap en anglais). Elle est semblable au cas planaire et n'est pas représentée sur la Figure III-27 pour ne pas surcharger le schéma.



a)
b)
c)
Figure III-27 : Représentation des différentes capacités parasites dans un FinFET à l'aide d'une vue en coupe dans un fin (a), vue en perspective (b) et d'une vue de dessus (c).

La Figure III-28 définit les dimensions nécessaires à l'évaluation des capacités parasites sur ces dispositifs non-planaires :

- L_g est la longueur de grille.
- CPP le Contacted Poly Pitch.
- dL la longueur de recouvrement de la grille et des jonctions (overlap en anglais).
- t_{ox} l'épaisseur d'oxyde de grille.
- t_{sp} l'épaisseur de l'espaceur.
- H_g la hauteur de grille. Dans le cas du double grille on distinguera grille du haut et grille du bas en notant leur hauteur respectivement H_{gt} et H_{gb} .
- H_{M1} distance entre le haut de la grille et le premier niveau de métallisation.
- t_{si} l'épaisseur du fin.
- h_{si} la hauteur du fin.
- FP le finpitch, égal à la somme de l'épaisseur d'un fin et d'un espace entre deux fins.
- T_{box} l'épaisseur d'oxyde enterré (BOX).
- H_{epi} la hauteur d'épitaxie pour les source-drains surélevés à partir du haut du fin.
- W_{ext} la largeur de l'extension de grille sur le STI (cf Figure III-9-d)
- N_{fin} nombre de fin du dispositif.

De plus, les permittivités des matériaux sont nécessaires aux calculs ultérieurs ; nous noterons :

- ϵ_{ox} la permittivité du SiO_2
- ϵ_{spacer} la permittivité de l'espaceur.
- ϵ_{ox} la permittivité du silicium.
- ϵ_{PMD} la permittivité de l'oxyde de remplissage.
- ϵ_{BOX} la permittivité de l'oxyde enterré (BOX).

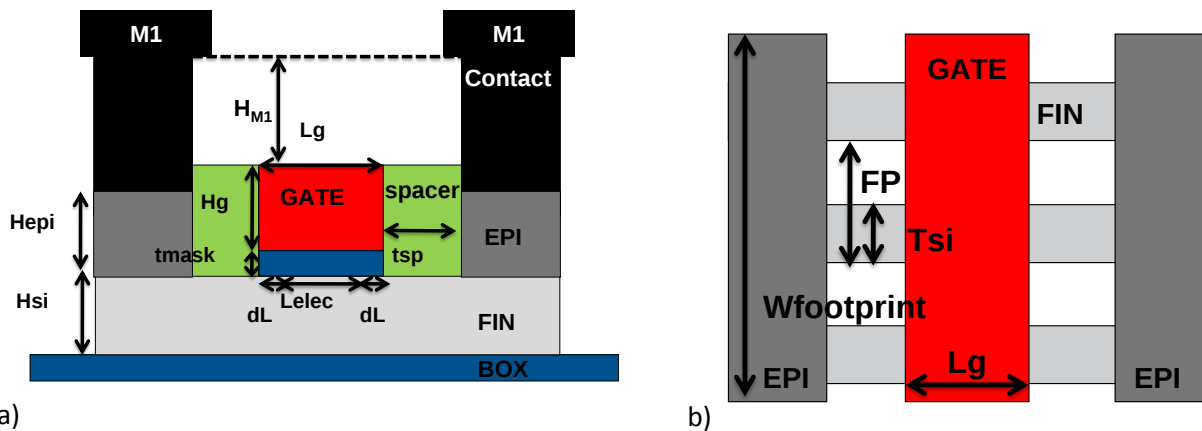


Figure III-28 : Représentation des dimensions sur un dispositif non planaire sur une vue en coupe (a) et sur une vue de dessus (b).

Enfin, la Figure III-29 illustre la différence entre un FinFET et un Trigate. Pour assurer la configuration double grille d'un FinFET (Figure III-29-a), un masque dur, souvent en nitrure, est déposé sur le fin. Pour le Trigate (Figure III-29-b), on souhaite que les trois faces du fin soient directement contrôlées par la grille ; c'est la raison pour laquelle on trouve l'empilement diélectrique métal sur le fin.

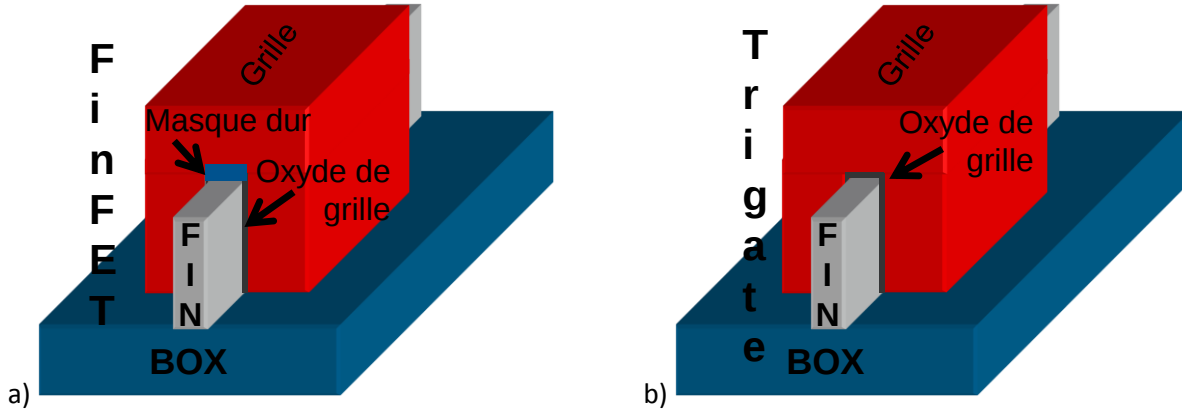


Figure III-29 : Illustration de la distinction entre FinFET (a) et Trigate (b).

III.D.2. Capacité de recouvrement (C_{ov})

Pour la capacité de recouvrement C_{ov} , dans le cas du FinFET, on procède exactement de la même manière que pour le double grille planaire (0) :

$$C_{ov} = 2N_{fin}H_{si}dL\frac{\epsilon_{ox}}{\epsilon_{OT}} \quad \text{Eq. III-62}$$

Pour le Trigate, l'influence de la troisième grille doit être prise en compte:

$$C_{ov} = (2N_{fin}H_{si} + N_{fin}T_{si})dL\frac{\epsilon_{ox}}{\epsilon_{OT}} \quad \text{Eq. III-63}$$

Par conséquent, la capacité de recouvrement sera supérieure dans le cas du Trigate, comparé au FinFET. Cependant, la capacité due au masque dur sera nulle pour un Trigate, alors que pour le FinFET elle sera :

$$C_{HM} = N_{fin}T_{si}L_g\frac{\epsilon_{HM}}{T_{mask}} \quad \text{Eq. III-64}$$

III.D.3. Capacité grille-fin (C_{gfin})

Comme représenté sur la Figure III-27, la capacité entre la grille et le fin est divisée en quatre composantes. $C_{gfin\text{side}}$ est identique pour le FinFET et pour le Trigate. Comme la capacité de bords externes C_{of} dans le cas planaire (0), elle est évaluée grâce à l'équation III-34, avec $(x_1, y_1, x_2, y_2) = (0, t_{ox}, t_{sp}, (FP-T_{si})/2)$:

$$C_{gfin\text{side}} = \frac{2}{\pi}N_{fin}H_{si}\epsilon_{spacer}\text{sh}^{-1}\left(\sqrt{\frac{\min(\frac{(FP-T_{si})}{2}, t_{sp})^2 + 2t_{ox}\min(\frac{(FP-T_{si})}{2}, t_{sp})}{t_{ox}}}\right) + 2N_{fin}0.35\frac{H_{si}}{\pi}\epsilon_{spacer}\ln\left(\pi\frac{H_{si}}{t_{ox}}\right) \quad \text{Eq. III-65}$$

L'estimation de $C_{gfin\text{top}}$ est quant à elle différente dans le cas du FinFET et celui du Trigate. Pour le FinFET, on utilise l'équation III-34, avec $(x_1, y_1, x_2, y_2) = (0, t_{mask}, t_{sp}, H_g)$:

$$C_{gfin\text{top}} = \frac{2}{\pi}N_{fin}T_{si}\epsilon_{spacer}\text{sh}^{-1}\left(\sqrt{\frac{\min(t_{sp}, H_g)^2 + 2\min(t_{sp}, H_g)t_{mask}}{t_{mask}}}\right) + N_{fin}0.35\frac{T_{si}}{\pi}\epsilon_{spacer}\ln\left(\pi\frac{T_{si}}{t_{mask}}\right) \quad \text{Eq. III-66}$$

Pour le Trigate, on utilise à nouveau III-34, mais cette fois avec $(x_1, y_1, x_2, y_2) = (0, t_{ox}, t_{sp}, H_g)$:

$$C_{gfin_top} = \frac{2}{\pi} N_{fin} T_{si} \epsilon_{spacer} \operatorname{sh}^{-1} \left(\sqrt{\frac{\min(t_{sp}, H_g)^2 + 2 \min(t_{sp}, H_g) t_{ox}}{t_{ox}}} \right) + N_{fin} 0.35 \frac{T_{si}}{\pi} \epsilon_{spacer} \ln \left(\pi \frac{T_{si}}{t_{ox}} \right) \quad \text{Eq. III-67}$$

Ensuite, pour la composante C_{gfin_corner} , on utilise une méthode similaire à celle utilisée pour la modélisation de $C_{corner_contacttop}$ (III.C.7) dans le cas des dispositifs planaires :

$$C_{gfin_corner} = \sum_i 2 N_{fin} \frac{2}{\pi} \frac{it_{sp}}{\operatorname{div}} \epsilon_{spacer} \operatorname{sh}^{-1} \left(\sqrt{\frac{H_{maxeff}^2 + 2 H_{maxeff} H_{mineff}}{H_{mineff}}} \right) \quad \text{Eq. III-68}$$

$$H_{mineff}(i) = \sqrt{(i t_{sp}/n)^2 + t_{ox}^2} \quad \text{Eq. III-69}$$

$$H_{maxeff}(i) = \sqrt{(i t_{sp}/n)^2 + (t_{ox} + T_{si})^2} - H_{mineff}(i) \quad \text{Eq. III-70}$$

La dernière composante C_{gfin_bottom} est modélisée comme $C_{cornerG}$ (III.C.7):

$$C_{gfin_bottom} = \frac{2}{\pi} 2 N_{fin} \frac{T_{si}}{2} \epsilon_{box} \operatorname{sh}^{-1} \left(\sqrt{\frac{t_{sp}^2 + 2 t_{sp} t_{ox}}{t_{ox}}} \right) + 2 N_{fin} 0.35 \frac{T_{si}}{2 \pi} \epsilon_{spacer} \ln \left(\pi \frac{T_{si}}{2 t_{ox}} \right) \quad \text{Eq. III-71}$$

Finalement, la capacité entre la grille et le fin est donnée par la somme :

$$C_{gfin} = C_{gfin_top} + C_{gfin_corner} + 2 C_{gfin_side} + C_{gfin_bottom} \quad \text{Eq. III-72}$$

III.D.4. Capacité de bords internes (C_{if})

Comme la capacité de recouvrement C_{ov} , la capacité de bords internes C_{if} est modélisée de manière similaire au cas planaire et seulement sa valeur maximum C_{ifmax} dépend de l'architecture. En adaptant l'équation III-41 au FinFET, on obtient :

$$C_{ifmax} = \frac{2}{\pi} 2 N_{fin} H_{si} \epsilon_{si} \operatorname{sh}^{-1} \left(\sqrt{\frac{\min\left(\frac{L_g}{2}, T_{si}\right)^2 + 2 t_{ox} \min\left(\frac{L_g}{2}, T_{si}\right)}{t_{ox}}} \right) + 0.35 \epsilon_{si} \frac{2 N_{fin} H_{si}}{\pi} \ln \left(\pi \frac{H_{si}}{t_{ox}} \right) \quad \text{Eq. III-73}$$

Pour le Trigate, la troisième grille doit être prise en compte :

$$C_{ifmax} = \frac{2}{\pi} 2N_{fin} H_{si} \epsilon_{si} \operatorname{sh}^{-1} \left(\sqrt{\frac{\min\left(\frac{L_g}{2}, T_{si}\right)^2 + 2 t_{ox} \min\left(\frac{L_g}{2}, T_{si}\right)}{t_{ox}}} \right) + 0.35 \epsilon_{si} \frac{2N_{fin} H_{si}}{\pi} \ln \left(\pi \frac{H_{si}}{t_{ox}} \right) \\ + \frac{2}{\pi} N_{fin} T_{si} \epsilon_{si} \operatorname{sh}^{-1} \left(\sqrt{\frac{\min\left(\frac{L_g}{2}, H_{si}\right)^2 + 2 t_{ox} \min\left(\frac{L_g}{2}, H_{si}\right)}{t_{ox}}} \right) + 0.35 \epsilon_{si} \frac{N_{fin} T_{si}}{\pi} \ln \left(\pi \frac{T_{si}}{t_{ox}} \right)$$

Eq. III-74

III.D.5. Capacité grille-contact (C_{pcca})

Nous nous contenterons ici de modéliser la capacité grille-contact dans le cas du schéma de contact en ruban car les dispositifs 3D sont attendus pour le nœud 22nm, pour lequel plus aucune plateforme CMOS n'utilisera de contacts à prise. Si nécessaire, la modélisation des capacités parasites dues aux contacts à prise pourra être effectuée en couplant ce paragraphe à celui traitant le cas des dispositifs planaires (III.C.5).

La difficulté est ici de déterminer la zone où les lignes de champs sont rectilignes entre le contact et la grille. Par un raisonnement analogue à au paragraphe III.C.5.a), la capacité grille-contact est donnée par :

$$C_{pcca} = N_{fin} C_{pcca_{flat}} + N_{fin} C_{pcca_{top}} \quad \text{Eq. III-75}$$

$$C_{pcca_{flat}} = \frac{FP \left(\frac{\min\left(t_{sp}, \frac{H_{epi} - t_{mask}}{2}\right)}{H_g - H_{epi} + t_{mask}} \right) \epsilon_{spacer}}{t_{sp}} \quad \text{Eq. III-76}$$

$$C_{pcca_{top}} = \frac{2}{\pi} FP \epsilon_{pmd} \operatorname{sh}^{-1} \left(\sqrt{\frac{\min\left(H_{M1}, \frac{L_g}{2}\right)^2 + 2 t_{sp} \min\left(H_{M1}, \frac{L_g}{2}\right)}{t_{sp}}} \right) + 0.35 \frac{FP}{\pi} \epsilon_{pmd} \ln \left(\pi \frac{\frac{L_g}{2}}{t_{sp}} \right) \quad \text{Eq. III-77}$$

Les équations de capacité grille-contact pour le Trigate sont obtenues en remplaçant simplement l'épaisseur du masque dur t_{mask} par l'épaisseur d'oxyde de grille dans les équations III-76 et III-77.

III.D.6. Capacité grille-épitaxie (C_{gepi})

Comme cela peut être vu sur la Figure III-27, la capacité entre la grille et l'épitaxie doit être divisée en trois composantes :

- $C_{gepi_betweenfin1}$ est la capacité entre la grille et l'épitaxie, comprise entre deux fins, du BOX jusqu'au haut du fin (Figure III-27-b).
- $C_{gepi_betweenfin2}$ est la capacité entre la grille et l'épitaxie, comprise entre deux fins, du haut du fin jusqu'au haut de l'épitaxie (Figure III-27-b).
- C_{gepi_onfin} est la capacité entre la grille et l'épitaxie, au-dessus du fin (Figure III-27-a).

La capacité $C_{gepi_betweenfin1}$ est donc évaluée à l'aide de la formule classique du condensateur plan. Similairement au cas planaire où $C_{pccafiat}$ a été corrigée en fonction de C_{of} (III.C.5.a) par $t_{sp}/2$, on corrige la largeur des électrodes $C_{gepi_betweenfin1}$ par le minimum entre $t_{sp}/2$ et $(FP-t_{si})/4$ (cf expression de $C_{gfininside}$ Eq-III-65). Formellement, pour un fin, on a :

$$C_{gepi_betweenfin1} = \frac{H_{si} \left(FP - T_{si} - 2 t_{ox} - \frac{\min(t_{sp}, \frac{FP - T_{si}}{2} - t_{ox})}{2} \right) \epsilon_{spacer}}{t_{sp}} \quad \text{Eq. III-78}$$

Pour la seconde composante de la capacité entre la grille et l'épithaxie entre les fins, la modélisation est plus simple car cette composante n'est affectée par aucune autre capacité. On obtient alors aisément pour un fin :

$$C_{gepi_betweenfin2} = \frac{H_{epi} (FP - T_{si} - 2 t_{ox}) \epsilon_{spacer}}{t_{sp}} \quad \text{Eq. III-79}$$

Enfin, la composante C_{gepi_onfin} est évaluée similairement à $C_{gepi_betweenfin1}$, mais la correction de largeur d'électrode est cette liée à C_{gfin2} (Eq-III-66). On obtient pour un fin :

$$C_{gate_epi_onfin} = T_{si} (H_{epi} - t_{mask} - \frac{\min(t_{sp}, H_g - t_{mask})}{2}) \frac{\epsilon_{spacer}}{t_{sp}} \quad \text{Eq. III-80}$$

Naturellement, on remarque que l'équation III-80 n'est plus valable si $(H_{epi} - t_{mask} - \frac{\min(t_{sp}, H_g - t_{mask})}{2}) < 0$. Dans ce cas, on aura $C_{gate_epi_onfin} = 0$.

Finalement, la capacité grille-épithaxie totale est donnée par la somme :

$$C_{gepi} = N_{fin} C_{gate_epi_onfin} + N_{fin} C_{gepi_betweenfin1} + N_{fin} C_{gepi_betweenfin2} \quad \text{Eq. III-81}$$

Pour simplifier les schémas et les expressions, nous avons traité le cas où l'épithaxie et le contact sont alignés (c'est-à-dire à la même distance de la grille). Si la situation étudiée ne respecte pas cette condition, il suffit d'adapter la composante décrite dans le cas planaire par l'équation III-52 :

$$C_{gate_epi_onfin2} = \frac{2}{\pi} FP \epsilon_{spacer} \text{sh}^{-1} \left(\sqrt{\frac{\min(H_g - H_{epi}, t_{sp2})^2 + 2 t_{sp1} \min(H_g - H_{epi}, t_{sp2})}{t_{sp1}}} \right) + 0.35 \epsilon_{spacer} \frac{FP}{\pi} \ln \left(\pi \frac{FP}{t_{sp1}} \right) \quad \text{Eq. III-82}$$

Où t_{sp1} est la distance entre la grille et l'épithaxie, et la somme $t_{sp1} + t_{sp2}$ la distance entre la grille et le contact (cf Figure III-21). La capacité entre grille et épithaxie totale sera alors :

$$C_{gepi} = N_{fin} C_{gate_epi_onfin} + N_{fin} C_{gate_epi_onfin2} + N_{fin} C_{gepi_betweenfin1} + N_{fin} C_{gepi_betweenfin2} \quad \text{Eq. III-83}$$

Pour évaluer la capacité grille-épithaxie pour un Trigate, il suffit de remplacer t_{mask} par t_{ox} dans l'équation III-80. Enfin, nous avons traité ici le cas où l'épithaxie source-drain vient connecter tous les fins entre eux : la largeur de l'épithaxie est alors égale à la largeur du dispositif. Il est cependant possible d'avoir un FinFET ou un Trigate dont les fins ne seraient pas connectées par l'épithaxie mais par le contact en ruban. La modélisation des capacités

parasites dues à l'épithaxie serait alors différentes (et leurs valeurs plus faibles). Cette estimation pourrait être effectuée en calculant la capacité entre grille et épithaxie entre les fins (la composante sur le fin ne serait pas affectée) par analogie avec la modélisation du schéma de contact à prise dans le cas planaire (III.C.5.b).Contact

III.D.7. Validation

La plupart des équations sont déjà validées dans la partie concernant les dispositifs planaires (III.C), nous choisissons donc d'effectuer des simulations 3D avec l'outil Raphael [Raphael] et d'en extraire uniquement la capacité entre la grille et le fin (C_{gfin}) et la capacité entre la grille et l'épithaxie (C_{gepi}).

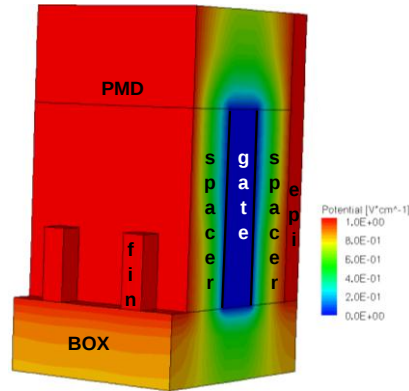


Figure III-30 : kit de simulation 3D Raphael utilisé pour valider le modèle de capacité parasites sur les dispositifs non-planaires.

Afin de simuler des structures réalistes, les dimensions utilisées sont déduites de [Kawazaki 09]. La Figure III-31 représente la variation des capacités C_{gepi} et C_{gfin} en fonction de l'épaisseur de l'espaceur pour deux FinFETs, dont les dimensions sont précisées dans la Figure III-31. On constate que notre modèle parvient à bien reproduire les résultats de simulations.

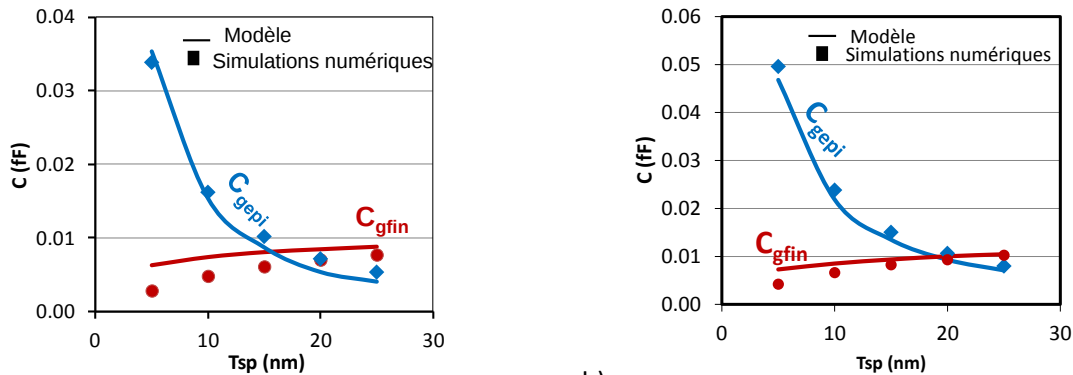


Figure III-31 : Tracé de la variation des capacités C_{gfin} et C_{gepi} pour deux FinFETs distincts.

III.D.8. Capacité corner (C_{corner})

La capacité de coin C_{corner} est identique au cas planaire et sera évaluée directement avec les équations III-53 à III-59 (III.C.7).

III.D.9. Capacité de jonction ou d'oxyde enterré (C_j ou C_{box})

Pour les évaluer les capacités de jonctions (substrat massif classique) ou de BOX (substrat SOI), il suffira d'utiliser directement les résultats donnés dans le paragraphe (III.C.8).

III.E. Raffinement du modèle

Les modèles décrits ci-dessus pour l'évaluation des capacités parasites pour les dispositifs planaires (III.C) et non planaires (III.D) donnent une bonne première estimation, cependant, certaines hypothèses peuvent être revues afin de mieux représenter la réalité morphologique des dispositifs, imposée par les procédés d'intégration. Dans cette partie nous allons raffiner notre modèle pour prendre en compte des situations qu'il est probable de rencontrer dans les technologies CMOS avancées. Nous nous concentrerons ici sur les dispositifs planaires, sachant que les équations obtenues pourront aisément être adaptées au cas des FinFETs ou des Trigate.

III.E.1. Source-drain épitaxiés avec facettes

Dans les technologies CMOS récentes, les source-drains sont réalisées par des épitaxies qui peuvent présenter des facettes. Ce type d'épitaxie est en train de devenir un standard à compter du nœud technologique 28nm, ce qui est confirmé par de récentes publications pour des dispositifs FDSOI [Cheng 11] ou Trigate [Auth 12]. Les capacités parasites dues à ce type d'épitaxie sont naturellement différentes, et leurs valeurs plus faibles. Nous proposons dans ce paragraphe de corriger le modèle pour les dispositifs planaires et la Figure III-32 représente les dimensions et capacités parasites pour une épitaxie facetée d'angle α .

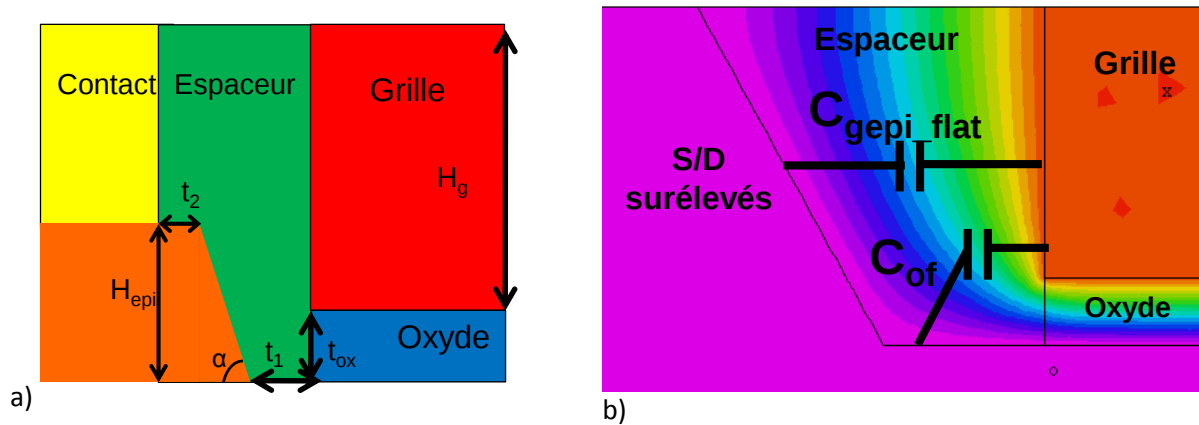


Figure III-32 : (a) représentation des dimensions nécessaires à la description et à la modélisation des capacités parasites pour des source-drains présentant une facette d'angle α . (b) Capacités parasites à modéliser pour une épitaxie facetée d'angle α représentée sur une cartographie d'iso-potentiels obtenue par FlexPDE.

Comme l'angle de la facette n'impacte que la composante à électrodes parallèles de la capacité grille-épitaxie (Figure III-32-b) nous allons corriger uniquement C_{gepi_flat} et conserver les mêmes équations pour les autres composantes (C_{gepi_top} and C_{of}).

Nous considérons alors que C_{gepi_flat} est composée de capacités à électrodes parallèles élémentaires, égales à $\frac{\epsilon_{spacer}}{t_1 + \frac{y}{\tan(\alpha)}} dy$ pour une hauteur y . On somme ensuite ces composantes élémentaires en intégrant ces capacités élémentaires de $y = t_1/2 + t_{ox}$ à $y = H_{epi}$ pour obtenir la capacité totale. Il est à noter que l'intégration ne débute pas de 0 pour prendre en compte la correction due à C_{of} (cf correction de $C_{pccflat}$ en III.C.5). Pour une facette d'angle α , on obtient:

$$C_{gate_epi_flat} = \int_{\frac{t_1}{2} + t_{ox}}^{H_{epi}} \frac{\epsilon_{spacer}}{t_1 + \frac{y}{\tan(\alpha)}} dy \quad \text{Eq. III-84}$$

Cette expression est intégrable analytiquement:

$$C_{\text{gateepi_flat}} = W \epsilon_{\text{spacer}} \tan(\alpha) \ln \left(\frac{t_1 + \frac{H_{\text{epi}}}{\tan(\alpha)}}{t_{\text{ox}} + \frac{t_1}{2}} \right) \quad \text{Eq. III-85}$$

Nous étudions uniquement l'intervalle d'angle de facettes compris entre 45° et 90° (90° correspondant au cas traité en III.C.6), typique des épitaxies silicium effectuées sur substrat (100), en procédant à des simulations 2D avec des variations de hauteur d'épitaxie (H_{epi}) et d'angle de facette (α) par pas de 15°. La Figure III-33 montre la somme des capacités grille-épitaxie facettée et de bords externes. On peut remarquer que le modèle reproduit bien l'impact des variations d'épaisseur d'espaceur et d'angle de facette.

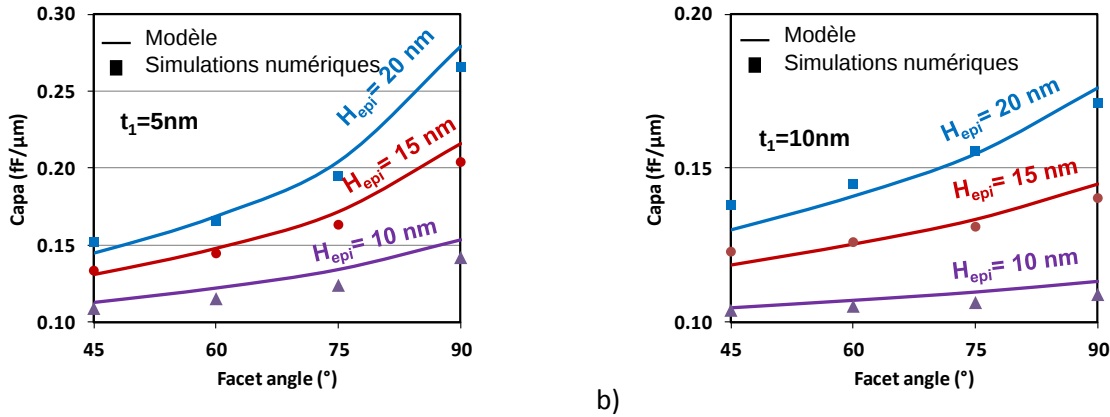


Figure III-33 : Comparaison des valeurs de la somme $C_{\text{of}} + C_{\text{gepiflat}}$, obtenue par notre modèle et par simulation numériques FlexPDE pour des variations d'angle de facette, d'épaisseur d'espaceur et de hauteur d'épitaxie.

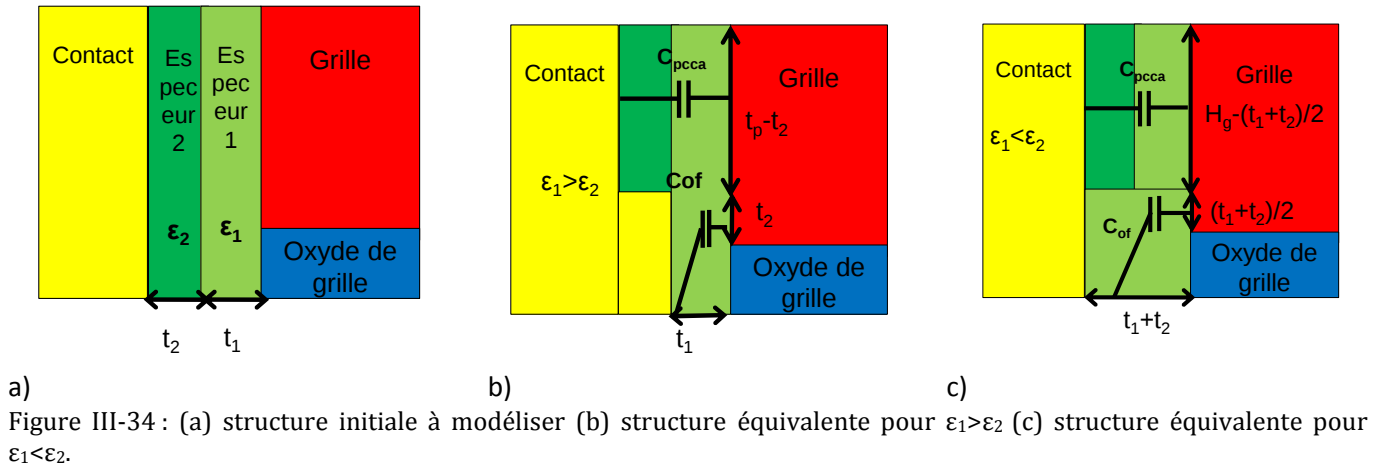
Cette expression (Eq. III-85) peut être aisément adaptée au cas des dispositifs non planaires.

III.E.2. Espaceur multi-couche

Nous avons considéré précédemment que l'espaceur entre grille et contact était composé d'un unique matériau, ce qui n'est pas le cas pour un transistor « réel ». En effet, il y a habituellement une couche d'arrêt de gravure en SiO_2 entre la grille et l'espaceur nitrure. De plus, dans une approche « gate last » il y a une couche de diélectrique haute permittivité sur les flancs de grille, entre le métal de la grille et l'espaceur. La modélisation de la structure d'espaceur à deux couches est effectuée en définissant des structures équivalentes, qui vont dépendre des valeurs des permittivités des couches ϵ_1 et ϵ_2 .

- $\epsilon_1 > \epsilon_2$, la composante à électrode perpendiculaire (C_{of}) est confinée près de la grille et est négligeable dans le second espaceur.
- $\epsilon_1 < \epsilon_2$, la composante à électrode perpendiculaire (C_{of}) est incluse dans les deux couches.

Ceci conduit aux deux structures équivalentes représentées sur la Figure III-34.



Avec ces deux structures équivalentes, la capacité de bords externes peut être évaluée avec l'équation III-37 en utilisant dans les deux cas la permittivité du premier espaceur ϵ_1 . La capacité grille-contact C_{pcca} est équivalente dans les deux cas à deux capacités à électrodes parallèles, respectivement définie par ϵ_1 , t_1 et ϵ_2 , t_2 , connectées en série. Enfin, la correction de hauteur de grille pour C_{pcca} est estimée à partir des simulations numériques 2D FlexPDE et est indiquée sur la Figure III-34. Le tableau de la Figure III-35 montre la comparaison entre notre modèle analytique 2D et les simulations numériques 2D FlexPDE et démontre la bonne précision de notre modélisation par structure équivalente.

t_1	ϵ_1	t_2	ϵ_2	$C_{of} + C_{pcca}$ (modèle)	$C_{of} + C_{pcca}$ (simulation)	Erreur
15 nm	20.ε0	15 nm	7.ε0	4.23E-16 F/μm	4.35E-16 F/μm	3 %
10 nm	20.ε0	20 nm	7.ε0	3.81E-16 F/μm	4.01E-16 F/μm	5 %
5 nm	20.ε0	25 nm	7.ε0	3.34E-16 F/μm	3.45E-16 F/μm	3 %
15 nm	4.ε0	15 nm	7.ε0	1.14E-16 F/μm	1.10E-16 F/μm	-3 %
10 nm	4.ε0	20 nm	7.ε0	1.18E-16 F/μm	1.17E-16 F/μm	0.5 %
5 nm	4.ε0	25 nm	7.ε0	1.22E-16 F/μm	1.30E-16 F/μm	6 %

Figure III-35 : comparaison des valeurs de C_{of} et C_{pcca} obtenues par modèle analytique et par simulations numériques 2D.

III.F. Applications

Dans cette dernière partie, nous allons utiliser les modèles développés au cours de ce chapitre dans deux courtes études : l'estimation des capacités parasites suivant la feuille de route ITRS et la comparaison des capacités parasites sur dispositifs double grille : le double grille planaire et le FinFET.

III.F.1. Estimation des parasites suivant la feuille de route ITRS

En suivant les projections de la feuille de route ITRS [ITRS], nous avons évalué les capacités parasites pour les architectures sur substrat massif (BULK), FDSOI, double grille planaire (DG) et FinFET jusqu'en 2021 pour les technologies Low Standby Power. Il doit être mentionné que l'ITRS ne distingue pas les différents types de dispositifs multi-grilles et, par conséquent, FinFET et DG planaire sont inclus dans ce même groupe. Pour chaque année, nous utilisons les règles de dessin fournies par l'ITRS, à savoir la longueur de grille (L_g), l'EOT, la profondeur de jonction (X_j) pour l'architecture conventionnelle sur substrat massif (BULK) ou l'épaisseur du film (t_{si}) de silicium pour les architectures à films minces et la moitié du pas de répétition du premier niveau de métallisation (en anglais, M1 contacted half pitch). Leur évolution est montrée sur la Figure III-36:

	Année	2009	2011	2013	2015	2017	2019	2021
	L (nm)	38	29	22	17	14	11.7	9.7
	M1 contacted ½ pitch (nm)	54	38	27	21	16.9	13.4	10.6
Bulk	t_{inv_bulk} (nm)	1.83	1.54	1.27				
	X_{j_bulk} (nm)	15	11.5	9				
FDSOI	t_{inv_fdsoi} (nm)			1.4	1.3	1.2		
	t_{si_fdsoi} (nm)			7.5	5.5	4.7		
Multi grille	t_{inv_DG} (nm)				1.5	1.4	1.3	1.2
	t_{si_DG} (nm)				8	7	5.8	4.8
	W_n (μm)	0.43	0.30	0.21	0.16	0.13	0.10	0.08
	W_p (μm)	0.43	0.30	0.21	0.16	0.13	0.10	0.08

Figure III-36 : évolution des données de l'ITRS avec l'année pour chaque architecture. Les cases rouges signifient que l'architecture ne sera pas utilisée pour cette année.

Pour une évaluation simple des capacités parasites, nous considérons pour chaque architecture :

- Contacts en ruban.
- Epitaxies alignées avec le contact.
- Espaceurs constitués d'un unique matériau.
- Epitaxies non facettées.

Ensuite, nous estimons les dimensions qui ne sont pas inscrites sur la feuille de route ITRS mais nécessaire à l'évaluation des capacités parasites avec les règles simples résumées dans la Figure III-37.

Paramètres	Evaluation	Paramètres	Evaluation
t_{sp}	$(CPP-L_g)/3$	t_{ox}	$2t_{inv}$
H_g	$2L$	H_{si}	$3xt_{si}$
W_{ext}	L	FP	$t_{si} + H_{si}$
W	$3xCPP$	N_{fin}	W/FP
dL	$L_g/4$	t_{mask}	t_{si}

Figure III-37 : Règles simples pour l'estimation des dimensions pour chaque architecture à partir des données de l'ITRS.

Enfin, nous pouvons estimer les capacités parasites pour chaque architecture pour chaque année. Nous représentons sur la Figure III-38 l'évolution du ratio C_{tot}/C_{inv} , avec capacité totale du transistor $C_{tot} = C_{par} + C_{inv}$, ($C_{par} = C_{ov} + C_{of} + C_{if} + C_{pcca} + C_{corner}$ pour les dispositifs planaires, $C_{par} = C_{ov} + C_{gfin} + C_{if} + C_{pcca} + C_{gepi} + C_{corner}$ pour le FinFET) et C_{inv} la capacité de grille intrinsèque. On constate que le poids des parasites sur la capacité totale d'un transistor est sous-estimé par l'ITRS. De plus, il semble que le double grille planaire présente moins de capacités parasites que le FinFET pour une même année. Comme les capacités parasites d'un FinFET sont très dépendantes du dessin du transistor (h_{si} , t_{si} , FP essentiellement) le paragraphe suivant propose une comparaison un peu plus poussée de ces deux architectures, en termes de capacités parasites.

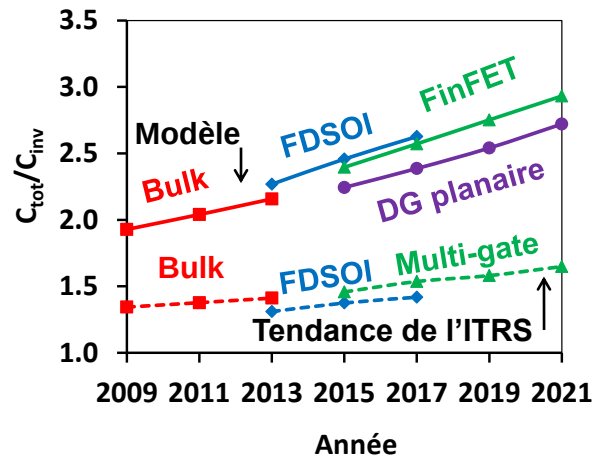


Figure III-38 : Evolution du ratio C_{tot}/C_{inv} avec l'année évalué avec notre modèle analytique puis par l'ITRS. Nous montrons ici que les capacités parasites sont sous-estimées par l'ITRS.

III.F.2. Comparaison FinFET / Double grille planaire

Dans ce paragraphe, nous comparerons les capacités parasites entre double grille planaire et FinFET pour un nœud technologique sub-10nm en utilisant comme donnée de base les dimensions données par l'ITRS pour l'année 2017 (cf Figure III-36). Comme au paragraphe III.F.1, les autres dimensions sont estimées avec les règles résumées dans la Figure III-37. Nous choisissons pour largeur (encombrement sur la puce) nominale des dispositifs $W=110\text{nm}$ pour le NMOS et $W=150\text{nm}$ pour le PMOS, puis nous définissons trois dessins de FinFET :

- Le cas A vise à comparer les deux types d'approches (planaires et non planaires) à encombrement équivalent. En estimant que l'encombrement d'un FinFET vaut $N_{fin} \cdot FP$, on obtient facilement le nombre de fin du transistor.
- Le cas B cherche à comparer les deux types de double grille mais cette fois à largeur électrique équivalente. Sachant que la largeur électrique vaut deux fois l'encombrement pour un dispositif double grille planaire et $2 \cdot N_{fin} \cdot H_{si}$ pour un FinFET, on parvient aisément à déterminer le nombre de fins de chaque FinFET.
- Le cas C a pour objectif de fournir des règles de dessin d'un FinFET permettant d'avoir une empreinte et une largeur électrique équivalente à celle du double grille planaire. Pour ce faire, il est indispensable de jouer sur autre paramètre que le nombre de fin, et nous choisissons de réduire le FinPitch FP pour pouvoir placer un plus grand nombre de fins dans une même largeur. Il a été nécessaire de réduire le FinPitch jusqu'à 17nm, ce qui constitue une configuration très agressive et certainement très difficile à réaliser sur silicium.

La Figure III-39 propose de résumer les dimensions de chaque cas étudié.

	DG Planaire	FinFET case A	FinFET case B	FinFET case C
T_{Si} (nm)	7	7	7	7
H_{Si} (nm)	N/A	19	19	19
N_{fin} (nMOS/pMOS)	N/A	3 / 5	6 / 8	6 / 8
FP (nm)	N/A	26	26	17
nMOS W_{FP} (nm)	110	97	175	112
pMOS W_{FP} (nm)	150	149	227	146
nMOS W_{Elec} (nm)	220	114	228	228
pMOS W_{Elec} (nm)	300	190	304	304

Figure III-39 : dimensions de chaque dispositif étudié.

Nous pouvons alors évaluer les capacités parasites pour ces quatre dispositifs et représentons sur la Figure III-40 la variation du ratio C_{tot_inv}/C_{inv} où C_{tot_inv} est la capacité totale d'un inverseur, donnée dans [Wei 11] et C_{inv} la somme de la capacité de grille intrinsèque en inversion du NMOS et du PMOS.

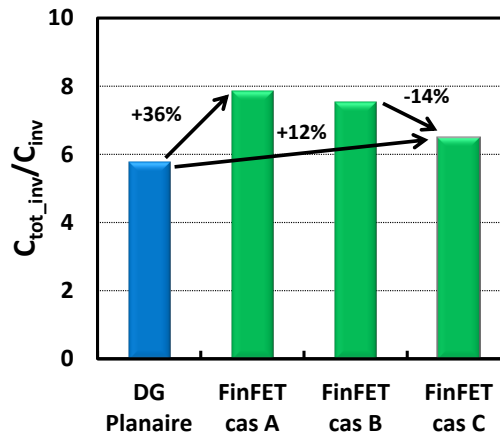


Figure III-40 : Variation du ratio de la capacité totale d'un inverseur sur la somme de la capacité de grille intrinsèque du NMOS et du PMOS.

On constate que, même avec des règles de dessin très agressives (cas C), le FinFET est toujours plus affecté par les capacités parasites que le double grille planaire. Par conséquent, un FinFET devra fournir plus de courant en mode d'inversion qu'un double grille planaire pour garantir une même performance au niveau circuit (délai).

III.G. Conclusion du chapitre

Poursuivant notre objectif d'estimer les performances au niveau circuit, nous proposons dans ce chapitre un modèle complètement analytique de toutes les capacités parasites, de la zone active du transistor jusqu'au premier niveau de métallisation des interconnexions pour les principales architectures CMOS, à savoir l'architecture conventionnelle sur substrat massif, le FDSOI, le double grille planaire, le FinFET et le Trigate. Nous avons commencé par distinguer les deux types de capacités parasites existants pour chaque technologie CMOS:

- Les capacités dont les électrodes sont parallèles, estimées par la formule classique du condensateur plan.

- Les capacités dont les électrodes sont perpendiculaires, estimées par la transformation de Schwarz-Christoffel [Plonsey 61].

Nous avons ainsi pu proposer une méthodologie générique permettant l'évaluation de l'intégralité des capacités parasites des technologies CMOS planaires et non-planaires. Cette même méthodologie pourra être utilisée afin d'estimer les capacités parasites de dispositifs plus « exotiques », comme des IFQW [Hellings 10].

Afin d'être plus représentatif de la réalité des technologies CMOS et de leur évolution, nous avons raffiné nos modèles pour tenir compte :

- Des deux schémas de contact concurrent, le contact à prise et le contact ruban.
- Des source-drains réalisés par une épitaxie présentant une facette.
- Des espaceurs à plusieurs couches.

Nous avons ensuite utilisé notre modèle pour estimer l'évolution du poids des parasites sur la capacité totale d'un transistor selon la feuille de route ITRS et démontré que l'ITRS sous-estime les capacités parasites. Enfin, nous avons comparé le poids des parasites sur deux architectures de double grille : planaire et FinFET et démontré que le FinFET est plus affecté par les parasites que le double grille planaire.

- CHAPITRE IV -

APPLICATION DES MODELES DE CAPACITES PARASITES : ESTIMATION DU DELAI D'UN CIRCUIT SIMPLE.

Avec la prédiction des performances statiques (Chapitre II), et l'estimation des capacités parasites (Chapitre III), il est maintenant possible de s'intéresser aux performances dynamiques. La méthode la plus simple et la plus rapide est d'utiliser les travaux de [Na 02], qui propose d'estimer le délai de commutation d'un inverseur à partir du courant effectif (I.B.4). Afin de gagner en précision sur l'estimation de ce délai, nous utilisons une résolution numérique et itérative des équations différentielles régissant le fonctionnement d'un inverseur (IV.A.1). Cette méthode est d'ores et déjà implémentée dans l'outil MASTAR [MASTAR]. Toutefois, pour ces deux méthodes, il reste une même inconnue à déterminer : la capacité équivalente de l'inverseur. Son calcul, tiré de [Wei 11] sera présenté au paragraphe IV.A.2. Enfin, la fin de ce chapitre sera dédiée à deux études comparatives de performance dynamique menées avec cette méthode. La première est consacrée à l'étude de performances de dispositifs III-V comparée à des architectures silicium standards (IV.B). La seconde portera sur une comparaison des performances d'un inverseur 3D (circuit construit par intégration séquentielle 3D monolithique [Batude 11]) avec un inverseur classique.

IV.A. Méthode de résolution

IV.A.1. Détermination du délai pour $C=1fF$

Le principe de cette méthode est de tracer les variations de tension de sortie de chaque inverseur constituant une chaîne d'inverseur en fonction du temps et d'en extraire le délai de propagation. Ce tracé est effectué de manière itérative, menant à une évaluation du délai est un peu plus longue que la méthode CV_{dd}/I_{eff} donnée par [Na 02]. Cependant, on considère ici la trajectoire complète de commutation de l'inverseur et n'effectuons pas d'approximation. Nous gagnerons donc en précision. Le schéma d'un inverseur est indiqué sur la Figure IV-1-a. En appliquant la loi de Kirchhoff sur ce circuit, on arrive à l'équation différentielle suivante [Wei 11]:

$$C \frac{dV_{out}}{dt} = I_p(V_{in}, V_{out}) - I_n(V_{in}, V_{out}) \quad \text{Eq. IV-1}$$

Où C est la capacité équivalente de l'inverseur, considérée ici égal à 1fF.

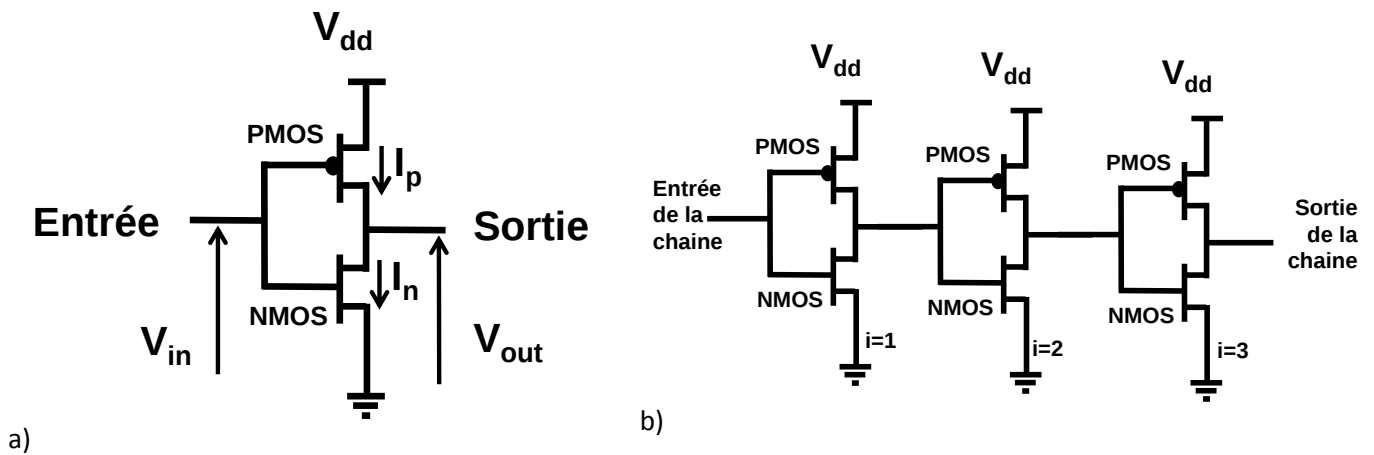


Figure IV-1 : (a) schéma d'un inverseur et ses paramètres électriques. (b) schéma de principe d'une chaîne d'inverseur (à 3 étages).

Pour pouvoir extraire un délai de propagation, il est nécessaire de monter les inverseurs en chaîne où la sortie d'un inverseur d'un étage donné est connectée à l'entrée de l'inverseur de l'étage suivant (illustrée sur la

Figure IV-1-b pour une chaîne à trois étages). Ensuite, on considère que la chaîne d'inverseurs est dans un état stable, avec la tension d'entrée de la chaîne à 0V, donc :

- $V_{out}(i, t=0)=V_{dd}$ si i est impair.
- $V_{out}(i, t=0)=0$ si i est pair.

En discrétisant le temps dans l'équation IV-1 (la pas temporel sera noté Δt), on peut facilement tracer la variation de la tension de sortie de chaque étage de la chaîne en fonction du temps. En initialisant la tension d'entrée du premier étage à V_{dd} , pour mettre la chaîne dans un état instable, on obtient de manière itérative la tension de sortie de l'étage i , pour le temps t avec :

$$V_{out}(i, t + \Delta t) = V_{out}(i, t) + \frac{1}{C} \left[I_p \left(V_g = V_{dd} - V_{out}((i-1), t), V_d = V_{dd} - V_{out}(i, t) \right) - I_n(V_g = V_{out}((i-1), t), V_d = V_{out}(i, t)) \right] \Delta t \quad \text{Eq. IV-2}$$

Où I_n et I_p sont les courants drain-source traversant respectivement le NMOS et le PMOS (calculés grâce au paragraphe II.D) et $V_{out}(i, t)$ la tension de sortie de l'étage i pour le temps t . On remarquera également que $V_{out}(i-1, t)=V_{in}(i, t)$.

On utilise alors l'équation IV-2 pour tracer la variation des tensions de sorties de chacun des i étages (ici, $i=6$) d'une chaîne d'inverseur (Figure IV-2). Sur cette même figure est représenté le délai de propagation (égal à deux fois le délai de commutation) de la chaîne d'inverseur qui est extrait numériquement à partir de ce jeu de courbes.

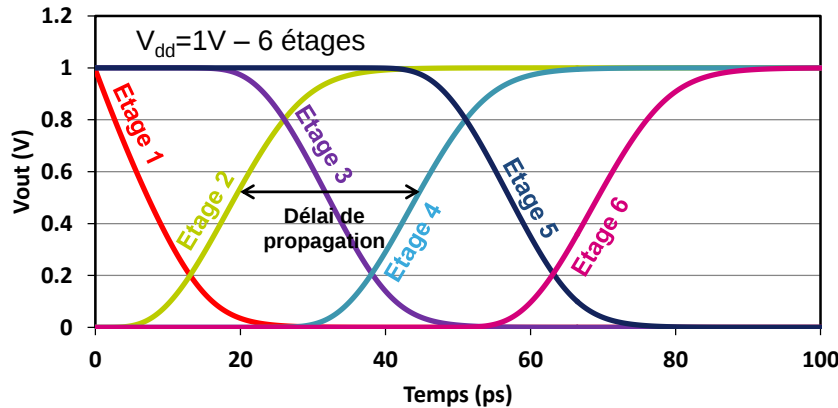


Figure IV-2 : Tracé des tensions de sorties de chacun des six étages d'une chaîne d'inverseur obtenu avec l'équation IV-2 et illustration du délai de propagation.

Cette méthode permet donc de connaître le délai de propagation d'une chaîne d'inverseurs à partir des $I(V)$ du NMOS et du PMOS. Elle ne tient cependant pas compte de la capacité équivalente de l'inverseur, qui varie si on change de dessin du circuit (largeur ou longueur des transistors) ou si différentes architectures de dispositifs sont utilisées. Le paragraphe suivant propose une méthode pour évaluer la capacité équivalente d'un inverseur, à partir de la capacité de grille et des capacités parasites de chaque dispositif (estimées dans le chapitre III).

IV.A.2. Calcul de la capacité équivalente

Cette capacité est évaluée à partir des travaux exposés dans [Wei 11]. La capacité totale qui est chargée et déchargée au nœud de sortie est donnée par la somme de trois composantes :

- C_d est la capacité de drain de l'étage en conduction, égale à la somme de la capacité grille-drain (C_{gd}) et de la capacité de jonction (C_j).
- C_g la capacité de grille de l'étage à charger (donc le suivant), en incluant les parasites, c'est-à-dire donnée par la somme $C_{gc} + 2C_{gd}$ où C_{gc} est la capacité grille-canal.
- C_{int} est la capacité d'interconnexion entre les 2 étages.

Afin d'effectuer de meilleures comparaisons de performance dynamique à partir d'inverseurs, il est courant d'utiliser différentes valeurs de FanOut (FO). Pour une chaîne d'inverseurs, le FanOut représente le nombre d'inverseur connecté à la sortie d'un étage. Pour FO=1, le schéma de la chaîne d'inverseur est celui de la Figure IV-1. Pour FO=3, trois inverseurs sont connectés à la sortie de chaque étage, un qui compose la chaîne (son entrée est connectée à la sortie du précédent, et sa sortie à l'entrée du suivant) et deux qui sont connectés à un nœud flottant. Un inverseur FO=3 est représenté sur la Figure IV-3.

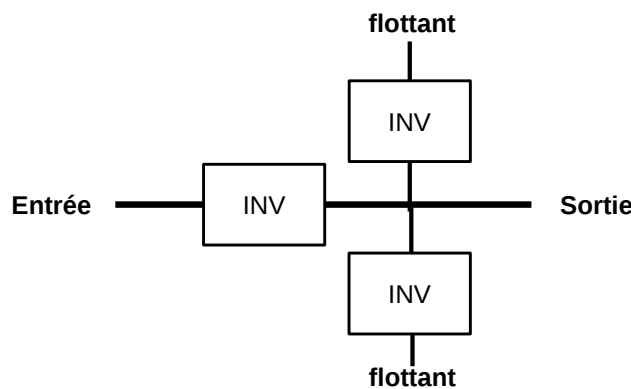


Figure IV-3 : schéma de principe d'une chaîne d'inverseur avec une valeur de FanOut FO=3.

L'intérêt d'augmenter le FanOut est d'être plus représentatif de la performance de circuit plus complexe et donc de la performance d'une architecture de dispositif dans un circuit réel. La valeur de FanOut sera prise en compte dans l'expression de la capacité totale de l'inverseur.

Comme les tensions appliquées aux bornes de C_{gd} (la grille et le drain, respectivement l'entrée et la sortie de l'inverseur) varient dans une direction opposée pendant la commutation, il faut inclure un multiplicateur $M=2$ pour tenir compte de l'effet Miller :

$$C_d = MC_{gd} + C_j \quad \text{Eq. IV-3}$$

Pour la commutation de l'état haut à l'état bas, V_{out} passe de V_{dd} à $V_{dd}/2$ alors que pour la commutation de l'état bas à l'état haut, V_{out} augmente de $V_{dd}/2$ à V_{dd} . Donc, au premier ordre, la capacité de jonction C_j est évaluée comme la moyenne entre les valeurs $C_j(V_{out}=V_d=0V)$ et $C_j(V_{out}=V_d=V_{dd})$ [Wei 11].

La capacité de grille C_g est également dépendante de la tension de sortie de l'inverseur. Pour la commutation de l'état haut à l'état bas le PMOS de l'étage à charger passe de l'état bloqué à l'état passant alors que le NMOS de l'étage à charger reste à l'état passant. A l'opposé, pour la commutation de l'état bas à l'état haut, le PMOS reste à l'état passant alors que le NMOS passe de l'état bloqué à l'état passant. Pour moyenner la contribution de NMOS et PMOS pour les deux types de commutation, la capacité de grille de chaque transistor est donnée par la somme de $\frac{1}{4}$ de la capacité de grille à l'état bloqué (C_{goff}), et de $\frac{3}{4}$ de la capacité de grille à l'état passant (C_{gon}) [Wei 11] :

$$C_g = \frac{1}{4}C_{goff} + \frac{3}{4}C_{gon} \quad \text{Eq. IV-4}$$

En considérant que :

- Lorsque le transistor est à l'état passant, la capacité d'inner-fringe C_{if} est négligeable (III.C.4) et la capacité grille-canal C_{gc} est égale à son maximum, c'est-à-dire C_{ox} .
- Lorsque le transistor est à l'état bloqué, la capacité d'inner-fringe C_{if} est maximale (III.C.4) et la capacité grille-canal C_{gc} est négligeable.

On obtient alors les formulations suivantes :

$$C_{goff} = C_{gb_off} + 2C_{ov} + 2C_{if} + 2C_{of} + 2C_{pcca} + 2C_{corner} \quad \text{Eq. IV-5}$$

$$C_{gon} = C_{gc} + 2C_{ov} + 2C_{of} + 2C_{pcca} + 2C_{corner} \quad \text{Eq. IV-6}$$

Les valeurs des composantes des équations IV-5 et 6 sont données par le chapitre III et C_{gb_off} est la capacité grille-substrat à l'état OFF [Wei 11]. Par conséquent, en tenant compte de la valeur du FanOut, la capacité totale équivalente pour un type de transistor est donnée par :

$$C_{tot} = C_d + FO \cdot C_g = (C_{ov} + C_{of} + C_{pcca} + C_{corner}) \cdot M + C_j + FO \cdot [0.25 (C_{gb_off} + 2C_{ov} + 2C_{if} + 2C_{of} + 2C_{pcca} + 2C_{corner}) + 0.75 (C_{gc} + 2C_{ov} + 2C_{of} + 2C_{pcca} + 2C_{corner})] \quad \text{Eq. IV-7}$$

Enfin, en notant C_{totN} et C_{totP} les capacités totales, respectivement, du NMOS et du PMOS évaluées par l'équation IV-7, on arrive à l'expression de la capacité totale de l'inverseur :

$$C_{TOT} = C_{totn} + C_{totp} + C_{int} \cdot FO \quad \text{Eq. IV-8}$$

Pour un circuit très simple comme un inverseur, les interconnexions entre transistors sont très courtes et nous pourrions considérer que $C_{int}=0$.

Le délai de propagation τ d'une chaîne d'inverseur peut alors être simplement évalué par :

$$\tau = \tau_{1ff} \cdot \frac{C_{TOT}}{1ff} \quad \text{Eq. IV-9}$$

Où τ_{1ff} est le délai évalué pour une capacité équivalente égale à 1ff, avec la méthode indiquée dans le paragraphe IV.A.1 et C_{TOT} est la capacité équivalente de l'inverseur, donnée par l'équation IV-8.

IV.B. Evaluation de performance dispositifs III-V/Ge: IFQW

IV.B.1. Contexte

Les NMOS utilisant des canaux en matériau III-V, combinés à des PMOS utilisant des canaux germanium sont de sérieux candidats pour les derniers nœuds technologiques de la feuille de route CMOS ITRS. Cependant, comme mentionné dans le paragraphe I.E.5, certains paramètres viennent modérer le gain attendu sur les performances, comme le DIBL, la pente sous le seuil et la capacité de grille qui sont supérieurs au cas du canal silicium.

Dans ce paragraphe, nous allons effectuer une comparaison de performance dynamique (délai d'un inverseur) entre des dispositifs à puits quantiques (IFQW pour Implant Free Quantum Well en anglais) utilisant des canaux III-V/Ge et des dispositifs conventionnel sur substrat massif et FDSOI en silicium.

Nous avons choisi les dispositifs IFQW car nous avons eu accès à des simulations Monte-Carlo de ces dispositifs à travers [Duallogic D4-2]. Les dispositifs simulés sont représentés sur la Figure IV-4.

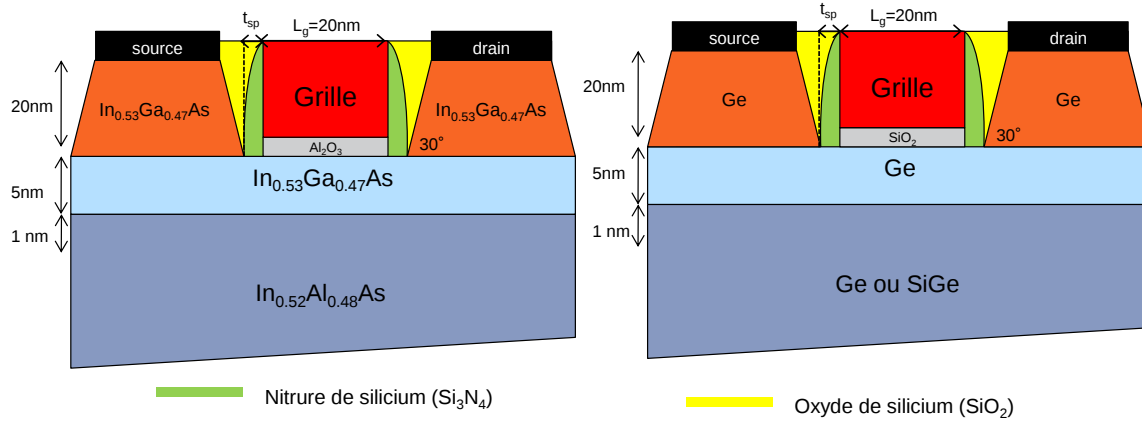


Figure IV-4 : Représentation schématique des IFQW simulés dans [Duallogic D4-2]

On remarque que ces IFQW ne présentent pas de zone de recouvrement entre les jonctions et la grille (en anglais, on parle d'underlap). Par conséquent, ce type de dispositif est très affecté par la résistance d'accès (I.D.3) [Chan 11], et sa valeur est directement liée à la distance entre la jonction et le flan de grille, donc à l'épaisseur de l'espaceur. Les simulations de caractéristiques courant-tension ont donc été effectuées sur des dispositifs avec un espaceur de 1nm (faible résistance, courant de saturation élevé) puis avec un espaceur de 5nm (haute résistance, courant de saturation réduit). La Figure IV-5 résume les paramètres électriques extraits de ces simulations.

	III-V nIFQW		Ge pIFQW	
	Espaceur 1 nm	Espaceur 5 nm	Espaceur 1 nm	Espaceur 5 nm
DIBL	35 mV/V	27 mV/V	74 mV/V	65 mV/V
SCE	0.1 V	0.087 V	0.16 V	0.20 V
SS	71 mV/Dec	76 mV/Dec	91 mV/Dec	95 mV/Dec
Résistance d'accès	15 $\Omega \cdot \mu\text{m}$	28 $\Omega \cdot \mu\text{m}$	100 $\Omega \cdot \mu\text{m}$	375 $\Omega \cdot \mu\text{m}$
Capacité grille-canal	1.85 $\mu\text{F} \cdot \text{cm}^2$	1.85 $\mu\text{F} \cdot \text{cm}^2$	4.147 $\mu\text{F} \cdot \text{cm}^2$	4.163 $\mu\text{F} \cdot \text{cm}^2$
T_{inv}	1.8 nm	1.8 nm	0.83 nm	0.83 nm
Capacité d'overlap	N/A	N/A	N/A	N/A
Capacité grille-drain	1.50 $\mu\text{F} \cdot \text{cm}^2$	1.30 $\mu\text{F} \cdot \text{cm}^2$	2.635 $\mu\text{F} \cdot \text{cm}^2$	1.767 $\mu\text{F} \cdot \text{cm}^2$
Capacité de jonction	0.90 $\mu\text{F} \cdot \text{cm}^2$	0.82 $\mu\text{F} \cdot \text{cm}^2$	$1.63 \times 10^{-8} \mu\text{F} \cdot \text{cm}^2$	$1.85 \times 10^{-8} \mu\text{F} \cdot \text{cm}^2$
μ_{eff}	472 $\text{cm}^2/\text{V} \cdot \text{s}$	276 $\text{cm}^2/\text{V} \cdot \text{s}$	125.4 $\text{cm}^2/\text{V} \cdot \text{s}$	71.83 $\text{cm}^2/\text{V} \cdot \text{s}$

Figure IV-5 : Résumé des paramètres électriques extraits par simulation Monte Carlo [Duallogic D4-2]

IV.B.2. Adaptation du modèle CMOS silicium pour reproduire les simulations Monte-Carlo IFQW

Pour évaluer les performances au niveau circuit des IFQW simulés par [Duallogic D4-2], il faut en premier lieu reproduire les I_d-V_g et I_d-V_d extraits des simulations avec notre modèle de courant, présenté au cours du paragraphe II.D. Pour ce faire, nous commençons par injecter dans le modèle les dimensions connues, c'est-à-dire la longueur de grille L_g à 20nm dans chaque cas et l'épaisseur d'oxyde effective en inversion t_{inv} . Cette dernière n'est pas donnée directement mais on peut y remonter aisément à partir de la valeur de la capacité grille-canal donnée dans la Figure IV-5 en remarquant que $C_{\text{gc}} = \epsilon_{\text{SiO}_2} / t_{\text{inv}}$. Comme nous ne disposons pas de modèle de l'électrostatique de ce type de transistor, nous entrons directement les valeurs de DIBL et de pente sous le seuil, extraites des simulations, dans le modèle. Pour obtenir les bonnes valeurs de I_{on} pour chaque dispositif, nous nous donnons des libertés sur les valeurs de résistance d'accès et de mobilité effective.

De cette manière, nous pouvons reproduire approximativement les I_d-V_g . Cependant, les I_d-V_d obtenues par notre modèle (courbes pointillées violettes sur la Figure IV-6) ne reproduisent absolument pas les simulations Monte

Carlo (points bleus sur la Figure IV-6). Ceci s'explique par le fait qu'il n'y a pas de zone de recouvrement entre les jonctions source-drain et la grille (underlap). Ceci a pour effet d'augmenter la résistance d'accès des dispositifs et explique la forme non-classique des I_d-V_d . Pour tout de même parvenir à les reproduire, nous avons dû adapter le modèle MOSFETs classique de courant de la manière suivante :

- Pour les NMOS (Canal InGaAs), nous avons introduit une résistance d'accès variable avec la polarisation de drain, dont les allures sont présentées sur la Figure IV-6. Pour les PMOS (Canal Ge), la méthode de correction de la résistance n'était pas efficace.
- Nous avons alors choisi de modifier les valeurs de DIBL et de vitesse de saturation (k_{vs}). Ceci a pour effet de dégrader la correspondance entre simulation et modèle sur les I_d-V_g .
-

La Figure IV-6 montre que notre modèle adapté (courbes rouges) reproduit convenablement les I_d-V_d obtenues par simulation (points bleus).

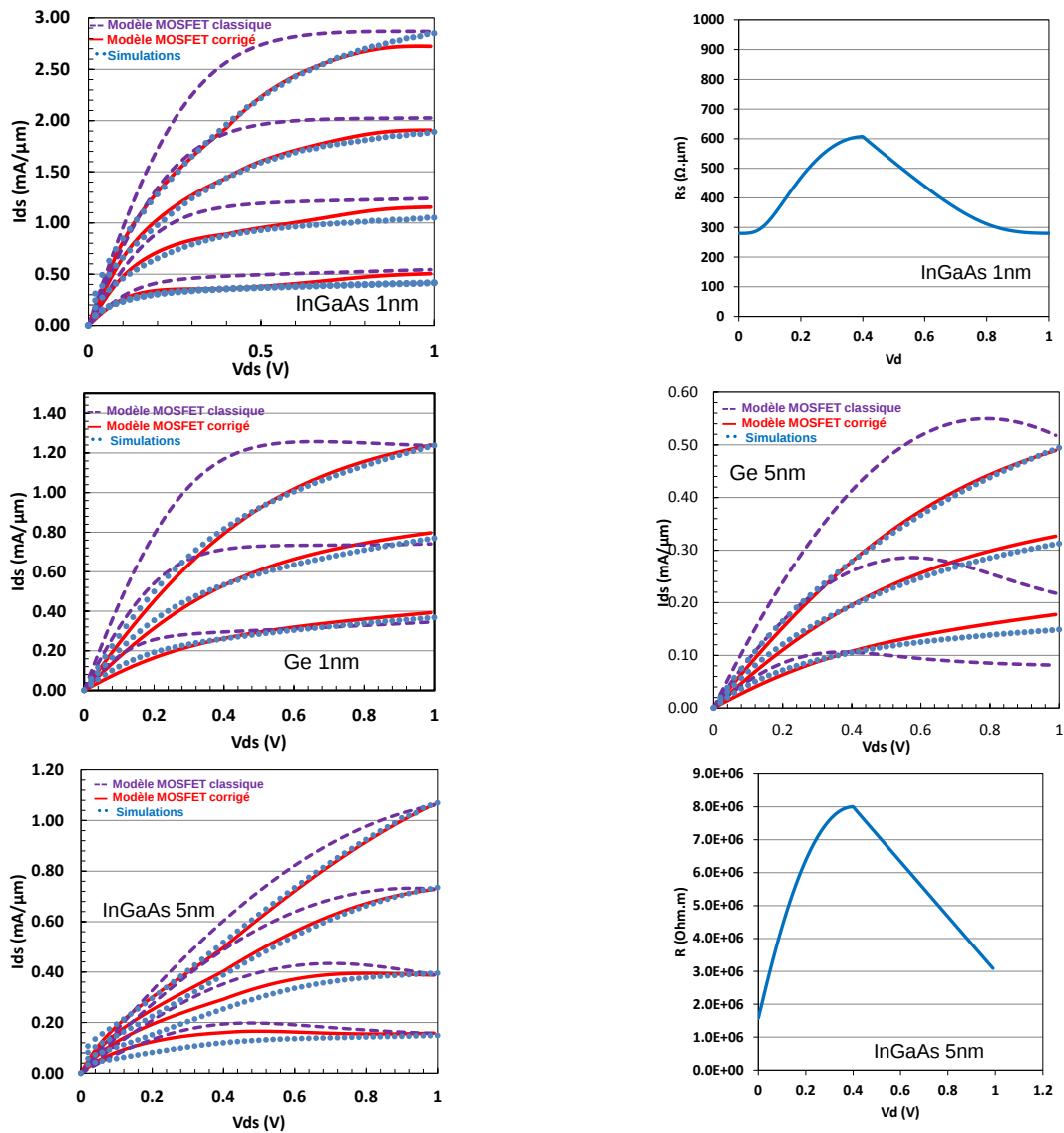


Figure IV-6: I_d-V_d obtenues par simulations numériques (points bleus), modèle MOSFETs classique (pointillés violets) et modèle MOSFETs adapté (traits pleins rouges) et allures de la variation de la résistance d'accès avec la polarisation de drain utilisées pour corriger le modèle de courant MOSFETs classique.

La Figure IV-7 démontre que le modèle, même adapté, reproduit bien les I_d - V_g extraites des simulations Monte Carlo.

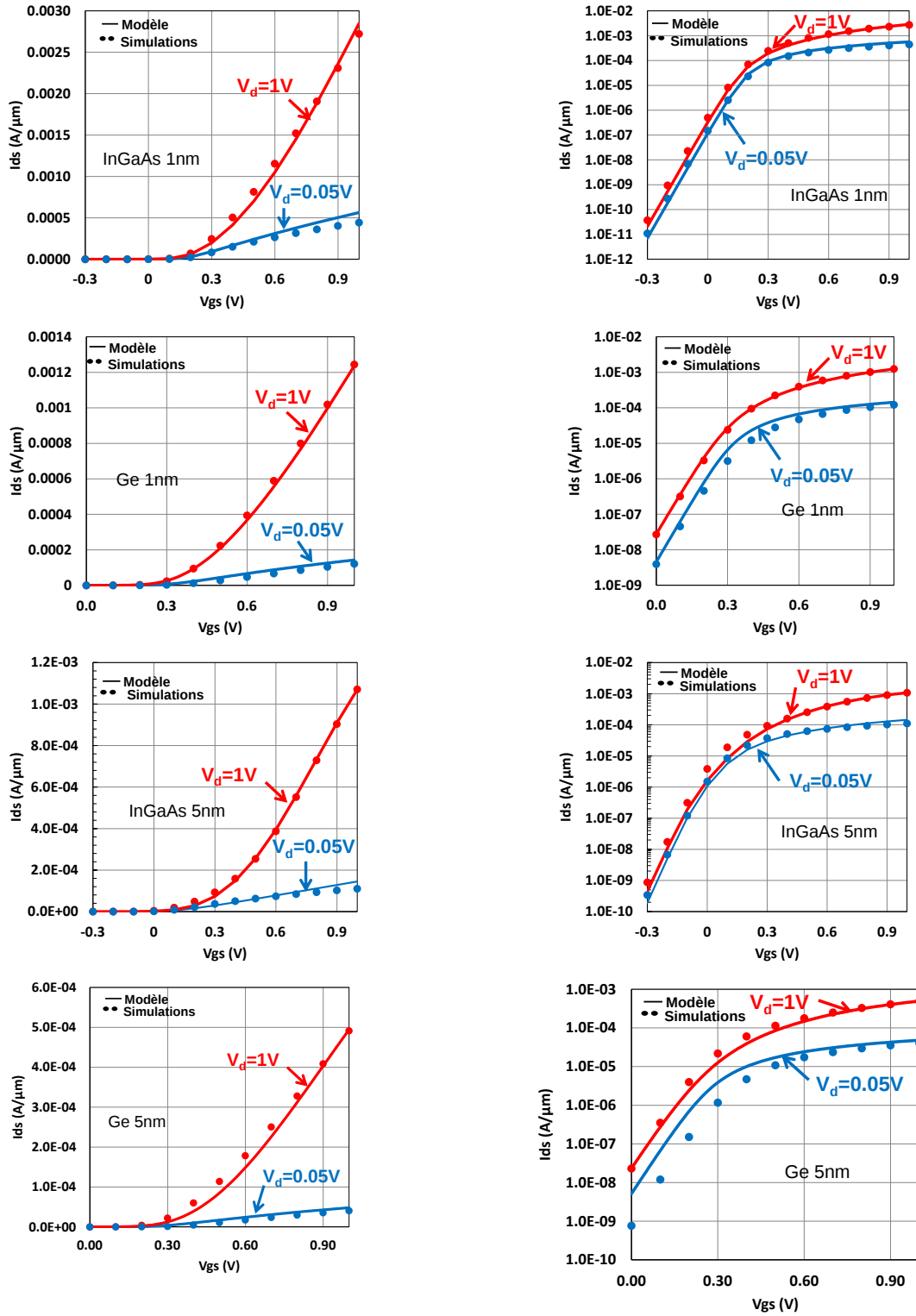


Figure IV-7 : I_d - V_g des quatres IFQW obtenues par simulation Monte Carlo et par notre modèle analytique.

Enfin, nous démontrons avec la Figure IV-8 qu'il est bien indispensable d'adapter le modèle de courant pour ne pas surestimer le courant effectif, et donc la vitesse de commutation d'un inverseur.

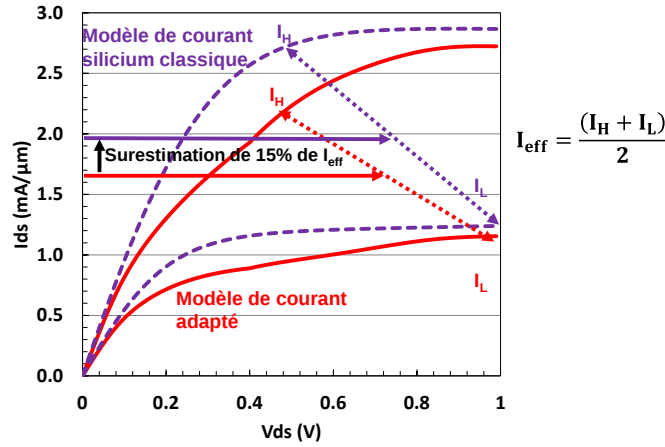


Figure IV-8 : Comparaison des valeurs de courant effectif obtenu avec le modèle de courant classique et le modèle adapté.

IV.B.3. Construction des références silicium

Le but de ce paragraphe est de construire deux références silicium pour comparer les performances dynamiques des IFQW à des architectures classiques silicium. Les dimensions des IFQW simulés sont typiques du nœud 20nm, nous choisissons donc de les confronter à un inverseur dessiné avec des dispositifs MOSFETs sur substrat massif, et à un inverseur dessiné avec des transistors FDSOI, tous typiques du nœud 20nm. Le tableau de la Figure IV-9 résume les dimensions utilisées pour ces deux architectures ainsi que les paramètres de transport et les valeurs des paramètres électrostatiques obtenus par nos modèles (II.A et II.B).

	nFDSOI	pFDSOI	nBULK	pBULK
L (nm)	20	20	20	20
T_{inv} (Å)	11	13	10	10
DIBL (mV/V)	75	75	140	180
SS (mV/dec)	75	75	95	95
R_s (Ω.μm)	200	200	220	220
μ (cm ² /V.s)	100	60	100	150
k_{vs}	0.8	0.6	1.17	1.15
T_{si} or X_j (nm)	5	5	10	10
T_{box} (nm)	25	25	/	/
N_{ch} (cm ³)	1e16	1e16	8e17	8e17

Figure IV-9 : Résumé des dimensions utilisées pour les architectures transistor sur substrat massif et FDSOI ainsi que les paramètres de transport utilisés et les valeurs des paramètres électrostatiques obtenus par nos modèles.

En premier lieu, nous comparons les courants effectifs de tous les transistors NMOS, puis de tous les transistors PMOS. La Figure IV-10 représente les variations de courant effectif I_{eff} avec le courant de fuite I_{off} . Le dispositif IFQW NMOS (Canal InGaAs) avec un espaceur de 1nm présente un niveau de courant effectif très élevés par rapport aux trois autres architectures, et cela quelle que soit la fuite statique I_{off} . Par contre, le courant effectif de

l'IFQW PMOS (Canal Ge) avec un espaceur de 1nm est très proche des valeurs obtenues avec les transistors sur substrat massif et FDSOI. Sachant qu'une meilleure image du délai est donnée par un « courant effectif équivalent » donnée par cette équation :

$$\frac{1}{I_{eff_{eq}}} = \frac{1}{I_{eff_N}} + \frac{1}{I_{eff_P}} \quad \text{Eq. IV-10}$$

Par conséquent, nous pouvons d'ores et déjà prédire que l'écart vu sur les courants effectifs des NMOS sera réduit lorsque le délai sera calculé.

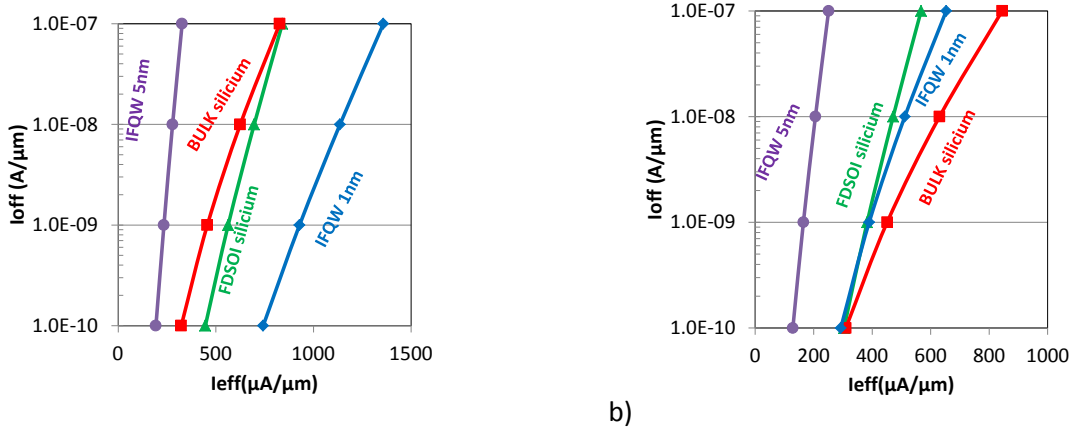


Figure IV-10 : Représentation du courant effectif I_{eff} en fonction du courant de fuite I_{off} pour chaque architecture: a) NMOS, b) PMOS.

Le tableau de la Figure IV-11 représente les écarts des valeurs de courant effectifs, en prenant pour référence le transistor sur substrat silicium massif (BULK). On constate bien que l'écart entre le courant effectif équivalent de l'IFQW NMOS avec un espaceur de 1nm et celui des références silicium est bien réduit, comparé à l'écart qu'il y a entre les courants effectifs des NMOS. On peut aussi constater que l'IFQW avec un espaceur de 5nm n'est absolument pas compétitif, quelle que soit la métrique observée.

Architectures	ΔI_{eff_N} (%) vs Bulk Silicium				ΔI_{eff_P} (%) Bulk Silicium				$\Delta I_{eff_{eq}}$ Bulk Silicium			
I_{off} (nA/μm)	0.1	1	10	100	0.1	1	10	100	0.1	1	10	100
IFQW sp 1nm	x 2,3	x 1,0	+82%	+64%	-6%	-14%	-19%	-23%	+43%	+30%	+20%	+12%
FDSOI silicium	+38%	+23%	+11%	+2%	-2%	-15%	-25%	-33%	+18%	+4%	-7%	-16%
IFQW sp 5nm	-40%	-49%	-56%	-60%	-59%	-64%	-67%	-70%	-50%	-56%	-61%	-65%

Figure IV-11 : Représentation de la variation (%) du courant effectif des NMOS, des PMOS et du courant effectif équivalent donnée par l'équation IV-10, pour chaque architecture, en prenant pour référence le transistor sur substrat silicium massif (BULK).

IV.B.4. Evaluation des capacités parasites

La Figure IV-12 est un résumé des capacités évaluées pour chaque dispositif. Il est important de remarquer la valeur très élevée de la capacité grille de l'IFQW PMOS (canal Ge) qui correspond à une épaisseur très fine de t_{inv} (0.83nm, indiqué sur la Figure IV-5). Cette valeur était nécessaire pour garantir un niveau de performance compétitif (i.e. I_{eff} et I_{on}), mais sera aussi, comme nous le verrons dans le paragraphe IV.B.6 une grande limitation de ces dispositifs en terme de fuite. A contrario, le IFQW NMOS (canal InGaAs) a une capacité de grille relativement faible ($t_{inv} = 18.5\text{Å}$), ce qui peut être expliqué par la valeur très élevée de la distance entre la couche d'inversion et l'interface oxyde-semi-conducteur (darkspace) dans les matériaux III-V [Skotnicki 10] [Rafhay 10].

La capacité grille-drain (C_{gd}) représente approximativement les capacités parasites sur chaque dispositif. Pour les IFQW, elle a été extraite des simulations Monte Carlo de [Duallogic D4.2] et estimée grâce aux expressions du chapitre III pour les références silicium. On peut alors déduire la capacité totale de chaque transistor C_{tot} par l'équation IV-7 (paragraphe IV.A.2). Ces trois valeurs de capacités sont résumées dans le tableau de la Figure IV-12 pour chaque architecture.

Dispositif	Cgc (fF)	Cgd (fF)	Ctot (fF)
III-V espaceur 1 nm	8.57E-02	3.46E-02	2.24E-01
III-V espaceur 5 nm	8.57E-02	3.00E-02	2.04E-01
N Bulk	1.45E-01	6.08E-02	4.49E-01
N Fdsoi	1.32E-01	6.83E-02	3.86E-01
Ge espaceur 1 nm	2.74E-01	8.70E-02	5.54E-01
Ge espaceur 5 nm	2.74E-01	5.83E-02	4.40E-01
P Bulk	2.07E-01	8.47E-02	6.38E-01
P Fdsoi	1.59E-01	9.21E-02	5.10E-01

Figure IV-12 : Résumé des capacités de chaque dispositif pour chaque architecture

IV.B.5. Estimation du délai, et variation avec I_{off} et V_{dd}

Avec les caractéristiques courant-tension de chaque dispositif de chaque architecture, et les valeurs de toutes les capacités, nous pouvons évaluer le délai de propagation d'un inverseur dans chaque cas à partir de la méthode décrite au cours des paragraphes (IV.A.1 et IV.A.2). De cette manière, on peut tracer les courbes de la Figure IV-13 qui représente le délai de propagation d'un inverseur en fonction du courant de fuite I_{off} pour une tension d'alimentation $V_{dd}=1V$, puis celles de la Figure IV-14 qui représentent le délai en fonction de la tension d'alimentation pour deux valeurs de fuite statique I_{off} 1nA/ μm (a) et I_{off} 1 $\mu A/\mu m$ (b) à $V_{dd}=1V$.

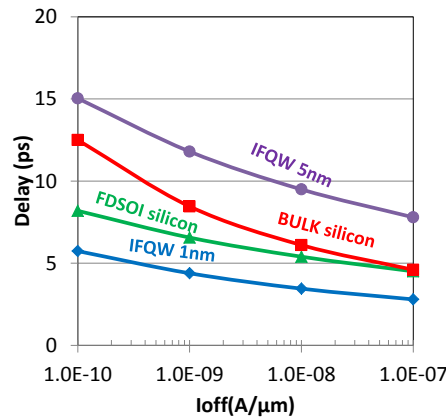


Figure IV-13 : délai de propagation d'un inverseur en fonction du courant de fuite I_{off} , à $V_{dd}=1V$. et en fonction de la tension d'alimentation V_{dd} (b).

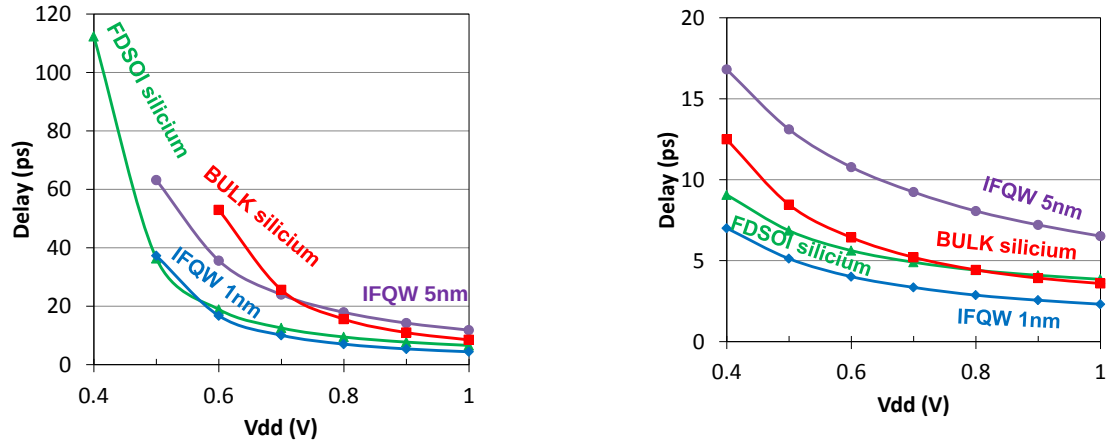


Figure IV-14 : délai de propagation d'un inverseur en fonction de la tension d'alimentation V_{dd} pour une fuite $I_{off} = 1 \text{ nA}/\mu\text{m}$ (a) et $I_{off} = 1 \mu\text{A}/\mu\text{m}$ (b).

Il est important de remarquer la grande différence entre le délai de l'IFQW avec espaceur à 1nm et celui à 5nm. Ceci pouvait être attendu après la détermination du courant effectif équivalent (Figure IV-11). Cette dernière confirme que, pour que l'architecture IFQW soit compétitive, son intégration doit être très bien contrôlée en termes de variabilité du procédé de fabrication. En effet, une variation de seulement 4nm de l'épaisseur de l'espaceur (de 1nm à 5nm) a un impact direct sur le délai de propagation de l'inverseur, qui peut être divisé par 1.5 dans le meilleur cas, et par 3 dans le pire cas. La

Figure IV-15 représente les variations de délai de propagation pour chaque architecture en prenant pour références l'inverseur construit avec l'architecture conventionnelle sur substrat silicium massif (BULK), pour une tension d'alimentation $V_{dd} = 1 \text{ V}$ et deux valeurs de fuite I_{off} .

$V_{dd} = 1 \text{ V}$	$I_{offN} = I_{offP} = 1 \text{ nA}/\mu\text{m}$	$I_{offN} = I_{offP} = 1 \mu\text{A}/\mu\text{m}$
IFQW espaceur 1nm	-48 %	-36 %
FDSOI silicium	-22%	+7 %
IFQW espaceur 5nm	+39%	+82%

Figure IV-15 : Comparaison des délais de propagation pour chaque architecture, en prenant pour référence les performances du transistor sur substrat massif, pour deux valeurs de fuite statique I_{off} et une tension d'alimentation à $V_{dd} = 1 \text{ V}$.

IV.B.6. Estimation de la fuite totale

Comme nous l'avons vu précédemment en comparant les valeurs de capacité de chaque architecture (Figure IV-12), le t_{inv} du PMOS IFQW est très faible (0.83 nm). Cela était nécessaire pour garantir un courant de saturation I_{on} suffisant, mais l'impact de cette faible valeur de t_{inv} sur toutes les composantes de fuite n'a pas été pris en compte dans les simulations Monte Carlo. Afin de fournir une analyse plus réaliste, nous allons estimer les composantes de fuite « omises » dans les simulations Monte Carlo :

- Fuite de substrat par effet tunnel bande à bande (BTBT pour Band To Band Tunneling).

Celui-ci est évalué à partir des valeurs extraites par caractérisation de transistor sur substrat germanium massif [Diouf 11].

- Fuite de grille, lorsque le transistor est en conduction (J_g), lorsque le transistor est bloqué (I_{goff}).

Pour estimer ces composantes, nous partons des valeurs ciblées de fuite dans les SRAM du nœud 28nm. Nous utilisons ces valeurs pour tracer les abaques de variations de la fuite en fonction de l'épaisseur d'oxyde de grille (EOT, donc sans darkspace), avec l'outil MASTAR 5. Nous faisons donc l'approximation que l'empilement métal de grille/oxyde enterré (même matériau) est le même que celui du nœud 28nm. La Figure IV-16 représente les abaques ainsi obtenues pour la fuite de grille en fonction de l'épaisseur électrique d'oxyde de grille (EOT) lorsque le transistor est en conduction (J_g : a et c) et pour la fuite de grille lorsque le transistor est bloqué (I_{goff} : b et d). Sur ces graphes, l'étoile rouge désigne la valeur de fuite à prendre en compte pour le dispositif concerné ; sachant que pour les NMOS (a, b) $EOT=1.4nm$ et que pour les PMOS (c, d) $EOT=0.45nm$.

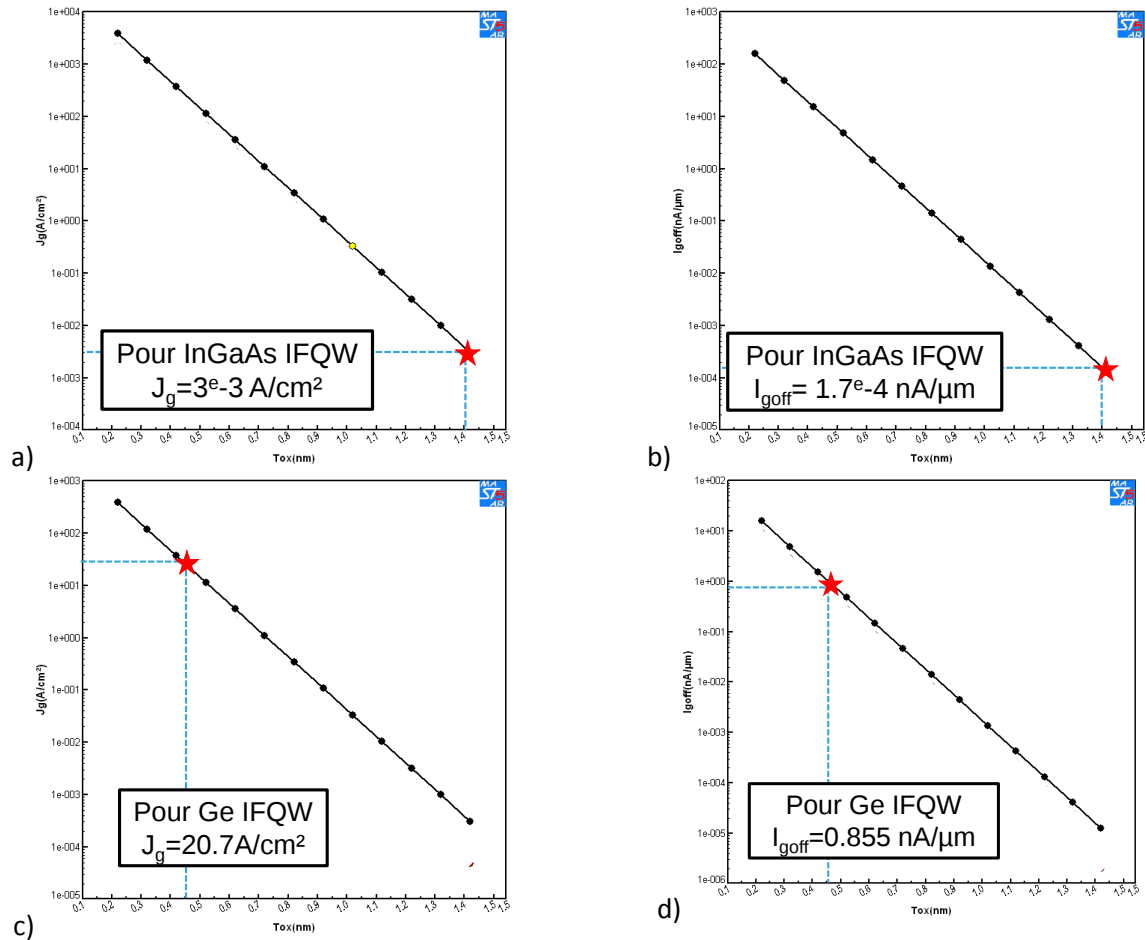
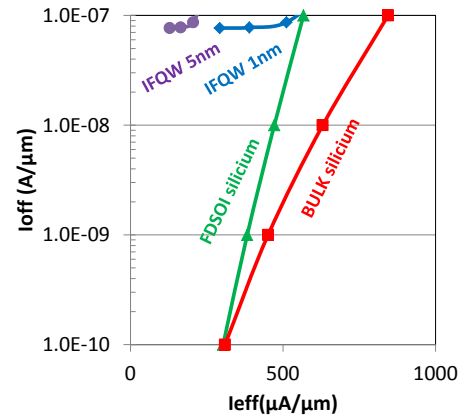


Figure IV-16 : Abaques calibrées sur les spécifications de fuite des cellules SRAMs 28nm (courbes noires) et extraction des valeurs de fuite de grille pour chaque IFQW étudié (étoiles rouges).

Les valeurs de composantes de fuite supplémentaires à prendre en compte sont résumées dans la Figure IV-17-a. La Figure IV-17-b montre que le domaine des applications basse consommation ne semble plus atteignable par les architectures IFQW à cause des composantes de fuite supplémentaires.

	InGaAs	Ge
Jg(A)	7.95E-14	1.37E-09
Igoff(A)	1.66E-14	2.57E-10
BTBT(nA/μm)	~ 0	250
BTBT(A)	0	7.50E-08
Fuite à ajouter (A)	9.61E-14	7.66E-08

a)



b)

Figure IV-17 : (a) valeurs des composantes de fuite supplémentaires pour les dispositifs IFQW. (b) Tracé du courant effectif I_{eff} en fonction du courant de fuite I_{off} pour chaque PMOS où I_{off} est ici la somme du courant de fuite du dispositif et la fuite extrapolée inscrite dans la dernière ligne du tableau de a).

On retrace alors les variations de délai de propagation en fonction de la fuite totale sur la Figure IV-18. On constate bien que l'architecture IFQW avec un espaceur de 1nm fournit de bonnes performances dynamiques, si la fuite n'est pas une contrainte trop forte (comme pour les applications haute performance). Cependant, cette architecture est, quelle que soit l'épaisseur de l'espaceur, inutilisable pour les applications basse consommation (comme les applications mobiles).

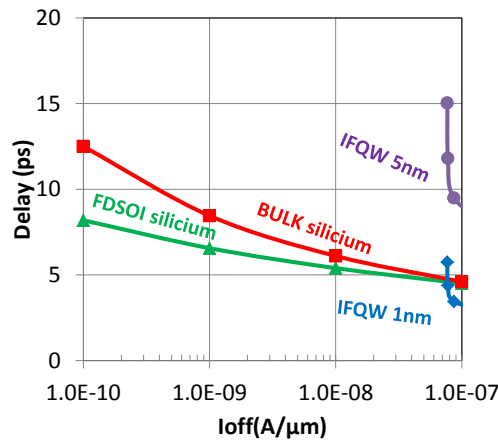


Figure IV-18 : Tracé du délai de propagation d'un inverseur en fonction de la fuite totale corrigée.

IV.C. Evaluation de performance d'un inverseur 3D monolithique.

IV.C.1. Contexte de l'étude, réalisation technologique.

L'intégration 3D génère beaucoup d'intérêt pour résoudre les limites fondamentales de la réduction des dimensions des interconnexions dans les technologies CMOS avancées, qui conduisent à une augmentation du délai RC [Banerjee 01]. L'intégration 3D monolithique, par opposition avec l'intégration parallèle (TSV 3D), est la seule solution technologique permettant de bénéficier complètement du potentiel de la troisième dimension à l'échelle du transistor grâce à sa grande précision d'alignement [Fukushima 09] [Batude 08].

L'intégration 3D monolithique peut être utilisée de deux manières différentes (Figure IV-19) :

- Dans une approche de miniaturisation (« More Moore »), où elle permettra d'intégrer d'avantage de transistor par unité de surface tout en relâchant la contrainte sur la réduction latérale des dimensions du transistor, comme la longueur de grille par exemple. Par conséquent, les effets néfastes de la réduction des dimensions de nœud en nœud pour suivre la loi de Moore, comme le contrôle électrostatique de la grille sur le canal seraient réduits [Batude 11].
- Dans une approche de diversification (« More than Moore »), elle permettrait de co-intégrer des transistors utilisant des matériaux pour le canal différents, comme démontrer par [Batude 09-b] qui a co-intégré NMOS en silicium et PMOS en germanium. Cette approche ouvre la voie à une nouvelle manière d'envisager la co-intégration III-V/Ge.

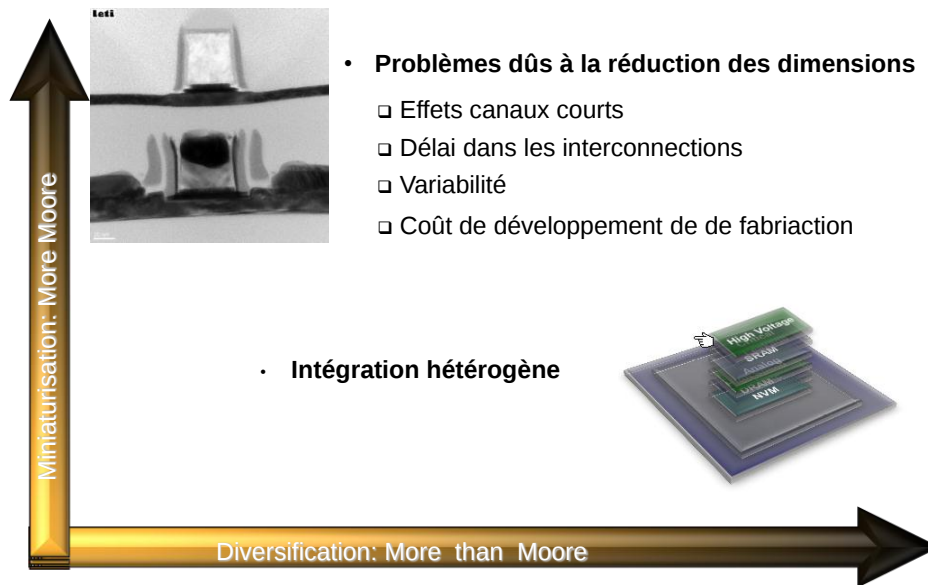


Figure IV-19 : Illustration des deux principales manières de tirer avantage de l'intégration 3D monolithique [Batude 11].

La faisabilité technologique est d'ores et déjà démontrée, mais deux questions restent en suspens :

- L'intégration 3D monolithique est-elle affectée par des problèmes d'auto-échauffement ?
- L'intégration 3D monolithique est-elle pénalisée par les capacités parasites entre les deux niveaux de transistor ?

Dans ce paragraphe, nous allons répondre à la seconde interrogation, en évaluant la vitesse d'un inverseur construit avec l'intégration 3D monolithique et la comparer à un inverseur classique planaire 2D.

IV.C.2. Présentation et définition des inverseurs 2D et 3D

IV.C.2.a) L'inverseur classique 2D

Dans cette étude, nous avons choisi pour référence un inverseur classique, planaire. Le dessin (layout en anglais) classique de son circuit est représenté sur la Figure IV-20. Les grilles des transistors sont connectées pour donner l'entrée de l'inverseur et les drains sont aussi connectés pour donner la sortie.

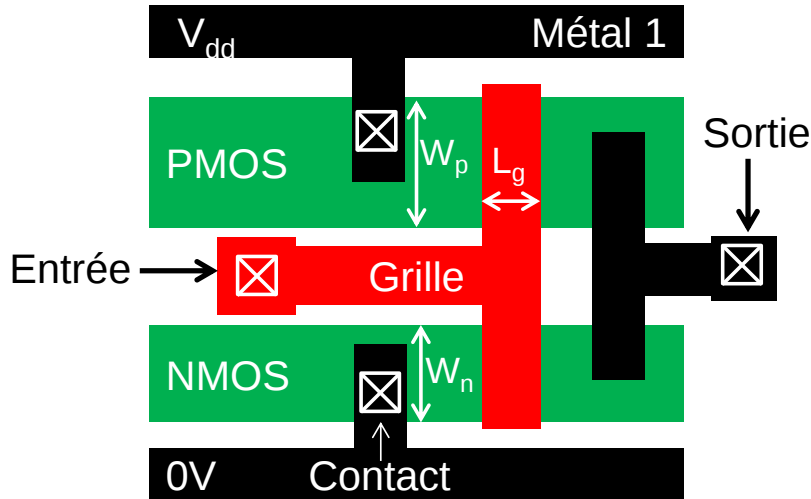


Figure IV-20 : Dessin classique du circuit d'un inverseur

IV.C.2.b) L'inverseur 3D

Pour dessiner un inverseur avec un dessin 3D, typique de l'intégration 3D monolithique, nous basons notre étude sur [Batude 09-a] qui a démontré la faisabilité technologique de ce type d'intégration 3D avec l'architecture FDSOI. Nous définissons alors un « inverseur 3D », schématiquement représenté sur la Figure IV-21 avec ses principales dimensions. Sur la Figure IV-21, le NMOS est le transistor du haut, et sa source est connectée à la masse (0V). A contrario, la source du PMOS est connectée à la tension d'alimentation V_{dd} . Les deux drains (à gauche sur le schéma) et les deux grilles sont connectés et donnent respectivement la sortie et l'entrée de l'inverseur. Un circuit similaire peut être obtenu en intervertissant NMOS et PMOS. On peut constater sur la Figure IV-21-a que le transistor du haut est dessiné avec les règles classiques et agressives utilisées pour un transistor placé dans un circuit classique. Par contre, comme les contacts sont déportés, le transistor du bas a des règles de dessin relâchées (distance grille-contact) comparées à un circuit 2D classique.

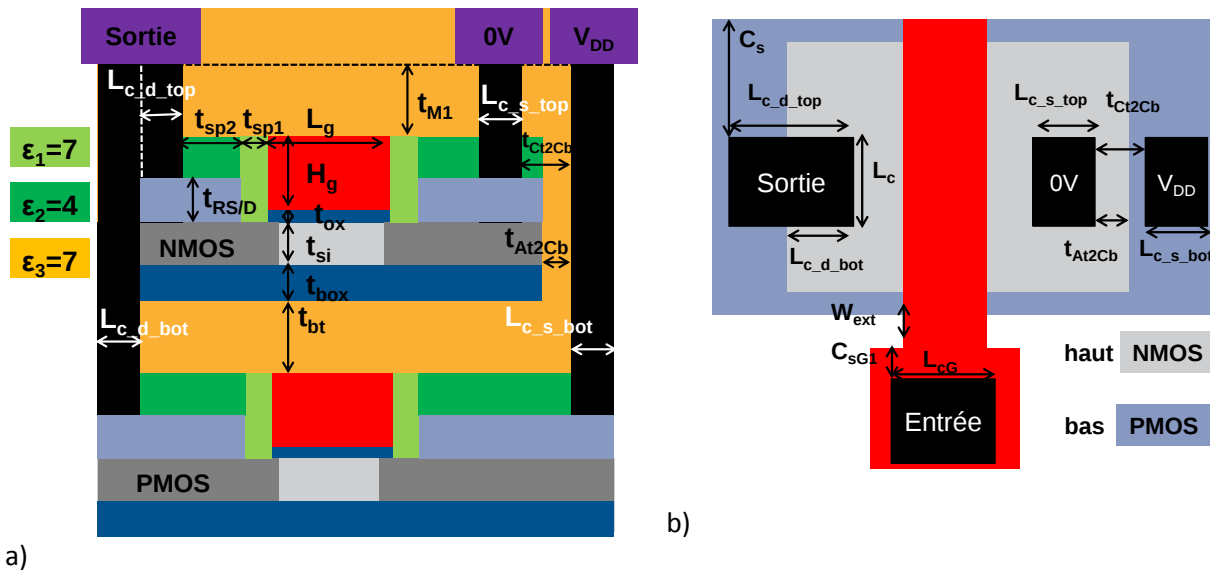


Figure IV-21 : Représentation de l'inverseur 3D dans le cas où le NMOS est le transistor du haut. a) vue en coupe et b) vue de dessus.

Le tableau de la Figure IV-22 donne les valeurs des dimensions, spécifiques à l'inverseur 3D, utilisées dans cette étude.

Dimension	Valeur
t_{At2Cb} (nm)	$t_{Ct2Cb}-10$
t_{Ct2Cb} (nm)	variable
$L_{c_s_bot}$ (nm)	45
$L_{c_s_top}$ (nm)	40
$L_{c_d_bot}$ (nm)	45
$L_{c_d_top}$ (nm)	40
t_{bt} (nm)	variable

Figure IV-22 : valeurs des dimensions en nanomètre spécifiques à l'inverseur 3D.

IV.C.3. Performance statique

Les différentes démonstrations technologiques de ce type d'intégration ont été réalisées avec l'architecture FDSOI [Batude 11]. De plus, il a été démontré dans [Batude 09-a] que l'intégration 3D monolithique n'altère pas (ou très peu) les performances statiques des transistors, en comparaison des performances obtenus par les même dispositifs dans une intégration 2D classique. Les dernières performances reportées pour l'architecture FDSOI l'ont été par [Planes 12] pour le nœud 28nm et ont d'ores et déjà été reproduites avec notre modèle de courant au cours du chapitre II.

Par conséquent, pour les deux types d'inverseur, nous utiliserons les caractéristiques courant-tension données dans le chapitre II (II.D.5.b).

IV.C.4. Capacités des dispositifs dans les deux types d'inverseur

Pour l'inverseur 2D, nous utilisons directement la méthode décrite dans le chapitre III (III.A), en considérant les dimensions typiques du nœud 28nm, données dans [Planes 12]. Le nœud 28nm étant compris dans la période de transition entre contact à prise et contact en ruban, nous étudierons les deux cas. La Figure IV-23 rappelle le schéma en coupe d'un transistor FDSOI (a) et les règles utilisées pour connaître chaque dimension (b).

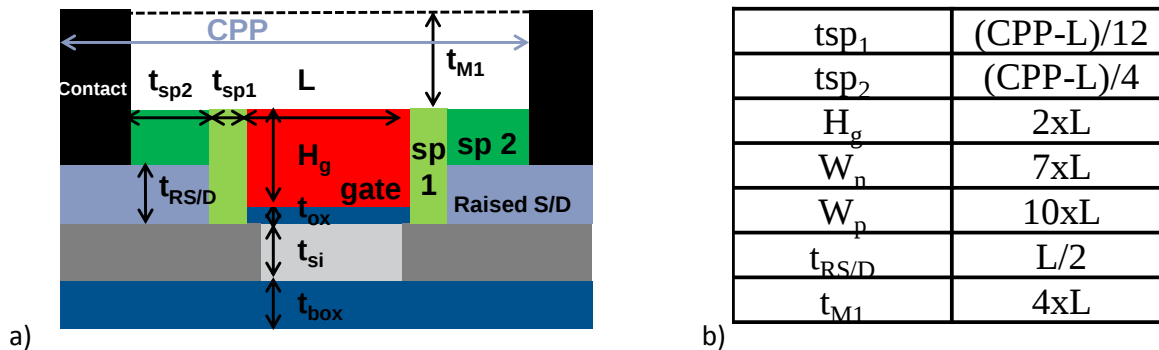


Figure IV-23 : (a) vue en coupe d'un transistor et de ses principales dimensions. b) règles utilisées pour déterminer chaque dimension.

Comme mentionné dans le paragraphe IV.C.2.b), dans l'inverseur 3D, le transistor du haut suit les mêmes règles de dessin que dans le cas planaire. Ses capacités parasites « intrinsèques », c'est-à-dire celles estimées dans le paragraphe III.F, ont alors la même valeur que dans le cas classique 2D. Pour le transistor du bas, la distance grille-contact est relâchée et les capacités parasites seront moins importantes. Pour les évaluer, il suffit d'utiliser les équations développées dans le chapitre III, en remplaçant la distance grille-contact par la somme $t_{sp1} + t_{sp2} + L_{c_d_top}$ (cf Figure IV-23). La Figure IV-24 compare les valeurs de capacité grille-épitaxie et grille-contact d'un inverseur 3D pour deux valeurs extrêmes de la distance entre les deux contacts de sources ($t_{Ct2Cb} = 20$ et 100 nm) et d'un inverseur classique 2D. Nous avons choisi de distinguer coté source et coté drain car seulement les capacités parasites du coté source dépendent de t_{Ct2Cb} et nous avons effectué les calculs pour un contact à prise (a) et en ruban (b). On constate que la capacité grille-épitaxie ne dépend pas du schéma de contact utilisé, qu'elle est supérieure pour l'inverseur 3D comparé à l'inverseur classique, et qu'elle augmente avec t_{Ct2Cb} . Ce dernier point était attendu, car lorsque t_{Ct2Cb} augmente, la surface de l'électrode qui forme le dessus de l'épitaxie augmente. A contrario, la capacité grille-contact est inférieure pour l'inverseur 3D en comparaison de l'inverseur 2D, et elle diminue lorsque t_{Ct2Cb} augmente (car la distance grille-contact augmente).

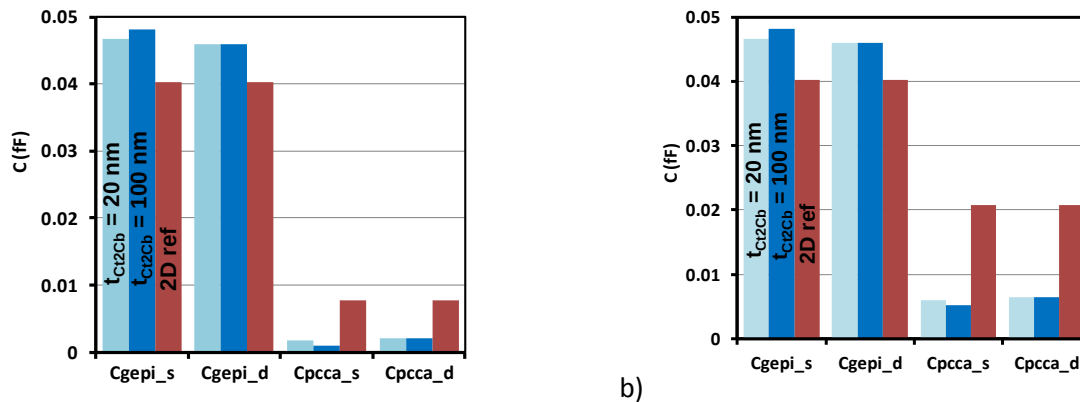


Figure IV-24 : Représentation des capacités grille-épitaxie (C_{gepi}) et grille-contact (C_{pcca}) en distinguant le coté source (indice s) et le coté drain (indice d). a) cas contact à prise b) cas contact en ruban.

IV.C.5. Capacités additionnelles dues à l'intégration 3D

Pour l'inverseur 3D, de nouvelles composantes de capacités parasites sont à prendre en compte. Ces dernières sont représentées sur la Figure IV-25.

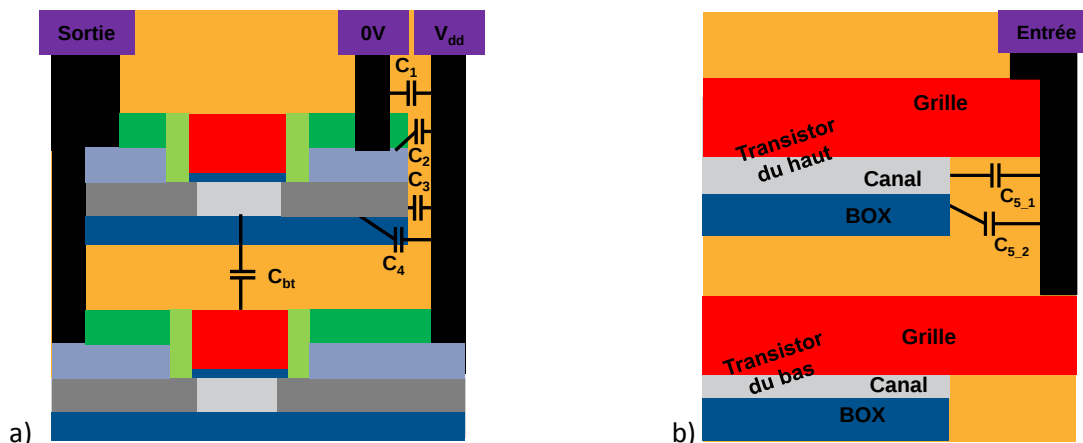


Figure IV-25 : représentation des nouvelles composantes de capacités parasites à prendre en compte pour l'inverseur 3D. a) vue en coupe perpendiculairement à la grille b) vue en coupe dans le sens de la grille.

Ces nouvelles composantes se divisent en trois groupes :

- Les capacités dues à la proximité entre le contact de source du transistor du haut et celui du transistor du bas. Cette capacité se décompose en quatre composantes (C_1 à C_4), représentées sur la Figure IV-25-a.
- La capacité due à la proximité entre la grille du transistor du bas et le canal du transistor du haut (C_{bt}), représentée sur la Figure IV-25-a.
- La capacité due à la proximité entre le contact qui connecte les deux grilles et le canal du transistor du haut (C_{5_1} à C_{5_3}), représentée sur Figure IV-25-b.

Ces capacités sont modélisées grâce à la méthode proposée dans le chapitre III (III.A), et nous nous concentrerons ici sur le schéma de contacts à prise.

C_1 est la capacité à électrodes parallèles entre les contacts de source, à partir du haut de l'épithaxie jusqu'au premier niveau de métallisation :

$$C_1 = \frac{\varepsilon_2 t_{M1} L_c}{t_{Ct2Cb}} + \frac{\varepsilon_3 \varepsilon_2 (H_g + t_{ox} - H_{epi} - \frac{t_{Ct2Cb}}{2}) L_c}{(\varepsilon_3 t_{At2Cb} + \varepsilon_2 (t_{Ct2Cb} - t_{At2Cb}))} \quad \text{Eq. IV-11}$$

C_2 est la capacité à électrodes perpendiculaires entre le contact de la source du transistor du bas et le haut de l'épithaxie formant la source du transistor du haut. Par analogie avec la capacité entre le haut de la grille et le contact (III.C.5), on a :

$$C_2 = \frac{2}{\pi} \varepsilon_2 L_c \operatorname{sh}^{-1} \left(\sqrt{\frac{\min^2 + 2(t_{Ct2Cb} - t_{At2Cb}) \min}{t_{At2Cb}}} \right) + 0.35 \varepsilon_2 \frac{L_c}{\pi} \ln \left(\pi \frac{W_c}{t_{At2Cb}} \right) \quad \text{Eq. IV-12}$$

$$\min = \min(t_{M1} + H_g + t_{ox} - H_{epi}, t_{At2Cb})$$

C_3 et $C_{3\text{fringe}}$ (non représentés sur la Figure IV-25) sont les capacités entre le contact de source du transistor du bas et le bord d'active et de l'épithaxie formant la source du transistor du haut. Par analogie avec la capacité grille-contact à prise (III.C.5.b):

$$C_3 = \frac{\varepsilon_2 L_c (t_{si} + H_{epi})}{t_{At2Cb}} \quad \text{Eq. IV-13}$$

$$C_{3\text{fringe}} = \frac{4}{\pi} (T_{si} + H_{epi}) \varepsilon_3 \cdot \operatorname{sh}^{-1} \left(\sqrt{\frac{\min(C_s, L_c)^2 + 2 t_{At2Cb} \min(C_s, L_c)}{t_{At2Cb}}} \right) + 0.35 \cdot \varepsilon_3 \frac{(T_{si} + H_{epi})}{\pi} \ln \left(\pi \frac{(T_{si} + H_{epi})}{t_{At2Cb}} \right) \quad \text{Eq. IV-14}$$

C_4 est la capacité entre le bas de la zone active du transistor du haut, et le contact sur la source du transistor du bas. Par la même analogie que dans le calcul de C_2 , on a :

$$C_4 = \frac{2}{\pi} L_c \varepsilon_2 \operatorname{sh}^{-1} \left(\sqrt{\frac{\min^2 + 2 t_{At2Cb} \min}{t_{At2Cb}}} \right) + 0.35 \varepsilon_2 \frac{L_c}{\pi} \ln \left(\pi \frac{L_c}{t_{At2Cb}} \right) \quad \text{Eq. IV-15}$$

$$\min = \min(t_{box} + t_{bt}, C2G + L_{cdtop} + t_{Ct2Cb} - t_{At2Cb})$$

Enfin, la somme de ces composantes donne une « pénalité capacitive » fixe pour le délai de l'inverseur, car les polarisations des électrodes sont constantes (l'une à V_{dd} l'autre à 0V) quel que soit l'état du circuit. Cette pénalité est donnée par:

$$C_{\text{wiring}} = C_1 + C_2 + C_3 + C_{3\text{fringe}} + C_4 \quad \text{Eq. IV-16}$$

Ces équations (IV-11 à IV-15) donnent la valeur des capacités parasites additionnelles pour l'inverseur 3D dans le cas où les transistors utilisent des contacts à prise. Dans le cas des contacts en ruban, il suffira d'utiliser ces même équations, en remplaçant la largeur des contacts L_c par la largeur du transistor du haut, et de fixer $C_{3fringe}=0$.

La capacité due au contact connectant les deux grilles se décompose en trois composantes :

- C_{5_1} est la capacité à électrodes parallèles entre le contact et la zone active du transistor du haut.
- C_{5_2} est la capacité à électrodes perpendiculaires entre le contact et la zone active du transistor du haut, travers le BOX.
- C_{5_3} est la capacité à électrodes perpendiculaires entre le contact et la zone active (analogue à C_{pcca_fringe} , III.C.5).

Ces trois composantes s'évaluent de manière similaire à la capacité grille-contact dans le cas contact à prise (III.C.5.b). Formellement, cela donne :

$$C_{5_1} = L_{cg} \left(T_{si} - \frac{C_{sg1} + W_{ext}}{2} \right) \frac{\epsilon_3}{C_{sg1} + W_{ext}} \quad \text{Eq. IV-17}$$

$$C_{5_2} = \frac{2}{\pi} L_{cg} \epsilon_3 \operatorname{sh}^{-1} \left(\sqrt{\frac{\min^2 + 2(C_{sg1} + W_{ext})\min}{C_{sg1} + W_{ext}}} \right) + 0.35. \epsilon_3 \frac{L_{cg}}{\pi} \ln \left(\pi \frac{\min}{C_{sg1} + W_{ext}} \right) \quad \text{Eq. IV-18}$$

Avec $\min = \min(t_{box} + t_{bt}, W_{top})$

$$C_{5_3} = \frac{2}{\pi} T_{si} \epsilon_3. \operatorname{sh}^{-1} \left(\sqrt{\frac{\min^2 + 2(C_{sg1} + W_{ext})\min}{C_{sg1} + W_{ext}}} \right) + 0.35. \epsilon_3 \frac{T_{si}}{\pi} \ln \left(\pi \frac{T_{si}}{C_{sg1} + W_{ext}} \right) \quad \text{Eq. IV-19}$$

Avec $\min = \min((CPP - L_g)/2, L_c)$

Enfin, la dernière capacité parasite supplémentaire dans l'inverseur 3D est la capacité entre la grille du transistor du bas et la zone active (i.e son canal) du transistor du haut, notée C_{bt} . Elle est évaluée simplement, comme une capacité d'oxyde enterré (III.C.8):

$$C_{bt} = \frac{0.5(CPP - L)W\epsilon_2\epsilon_3}{\epsilon_2 t_{bt} + \epsilon_3 t_{box}} \quad \text{Eq. IV-20}$$

On peut alors tracer la variation des capacités parasites additionnelles dans l'inverseur 3D avec les deux principaux paramètres qui vont influencer sur leurs valeurs (t_{bt} et t_{ct2cb}), mais également sur la densité d'intégration de l'inverseur 3D (t_{ct2cb}). Comme le NMOS et le PMOS ont un rôle dual, nous étudierons les cas où le NMOS est le transistor du bas puis le transistor du haut (et réciproquement pour le PMOS). La Figure IV-26-a représente la variation de C_{wiring} avec t_{ct2cb} (sa valeur de dépend pas de t_{bt}), pour les deux schémas de contacts et en modifiant le type du transistor du haut. On constate que le cas contact en ruban avec le PMOS en transistor est le plus impacté, ce qui s'explique par la plus grande largeur du PMOS à comparer au NMOS et par le fait que le schéma contact en ruban est le plus capacitif. On peut aussi remarquer que l'inverseur 3D avec le schéma de contact à prise ne semble pas impacté par le type du transistor du haut, et que la capacité supplémentaire due à la 3D est assez faible. La Figure IV-26-b représente les variations de C_{bt} et de C_5 ($=C_{5_1} + C_{5_2} + C_{5_3}$) avec la distance entre les deux niveaux de transistor t_{bt} . A nouveau, le cas où le PMOS est le transistor du haut est le plus impacté, à cause de largeur plus importante du PMOS, pour la composante C_{bt} . La capacité C_5 n'est pas impactée par le type du transistor du haut, et varie très peu en fonction de t_{bt} . Ces composantes ne dépendent pas de la distance entre les

deux contacts de source (t_{ct2Cb}) et leurs valeurs sont très faibles. Elles n'auront que peu de poids dans le calcul de la vitesse de l'inverseur.

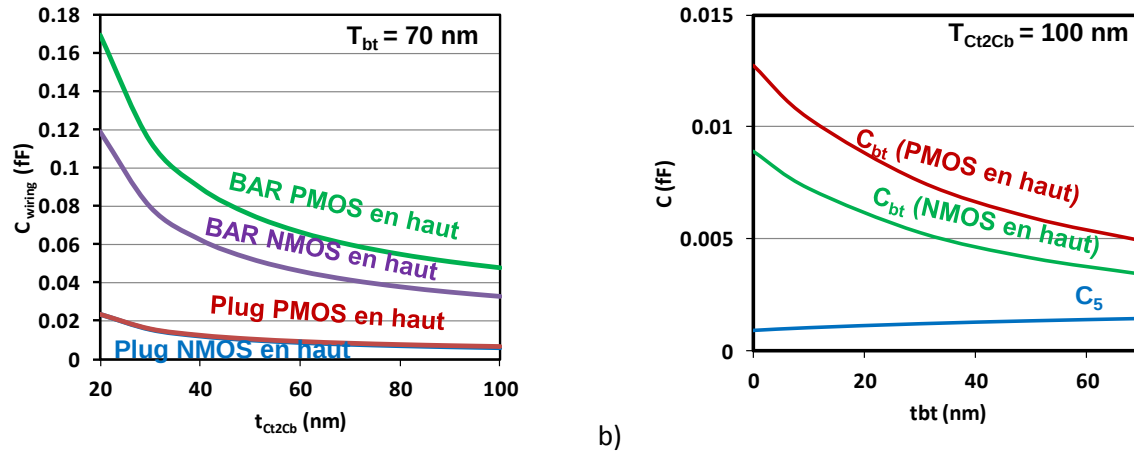


Figure IV-26 : Variation de C_{wiring} avec t_{ct2Cb} pour l'inverseur 3D, dans les schémas de contact à prise (plug) et en ruban (BAR) (a) et des autres composantes avec t_{bt} (b).

IV.C.6. Estimation du délai

Pour estimer le délai, nous disposons déjà des caractéristiques courant-tension de chaque transistor (paragraphe IV.C.3). Il ne manque alors que la capacité totale équivalente de chaque inverseur. Pour l'inverseur 2D, nous utilisons directement l'expression donnée par le paragraphe IV.A.2. Pour l'inverseur 3D, il nous faut adapter cette expression. La capacité C_{wiring} joue le rôle d'une capacité fixe (i.e. quelle que soit les valeurs de polarisation) supplémentaire. Cette capacité existe pour chaque étage, le terme à ajouter à la capacité totale de l'inverseur 2D est la valeur de C_{wiring} multipliée par la valeur du FanOut. C_{bt} joue le rôle d'une capacité de BOX, et remplacera donc la capacité de BOX dans l'expression de la capacité équivalente du transistor du haut. Enfin, C_5 vient en complément de la capacité de coin C_{corner} du transistor du haut et doit être prise en compte de la même manière. En distinguant bien coté source et coté drain pour le transistor du bas, la capacité totale équivalente de l'inverseur est donnée par :

$$C_{TOT} = C_{tottop} + C_{totbot} + C_{wiring} \cdot FO \quad \text{Eq. IV-21}$$

$$C_{totbot} = (C_{ov} + C_{of} + C_{gepi_d} + C_{gepi_s} + C_{pcca_d} + C_{pcca_s} + C_{corner}) \cdot 2 + C_{box} + FO \cdot [0.25(2C_{ov} + 2C_{if} + 2C_{of} + C_{gepi_d} + C_{gepi_s} + C_{pcca_d} + C_{pcca_s} + C_{corner_d} + C_{corner_s}) + 0.75(C_{gc} + 2C_{ov} + 2C_{of} + C_{gepi_d} + C_{gepi_s} + C_{pcca_d} + C_{pcca_s} + C_{corner_d} + C_{corner_s})] \quad \text{Eq. IV-22}$$

$$C_{tottop} = (C_{ov} + C_{of} + C_{gepi} + C_{pcca} + C_{corner}) \cdot 2 + C_{bt} + FO \cdot [0.25(2C_{ov} + 2C_{if} + 2C_{of} + 2C_{gepi} + 2C_{pcca} + 2C_{corner}) + 0.75(C_{gc} + 2C_{ov} + 2C_{of} + 2C_{gepi} + 2C_{pcca} + 2C_{corner})] \quad \text{Eq. IV-23}$$

Où C_{tottop} et C_{totbot} sont respectivement les capacités équivalentes du transistor du haut et de celui du bas. Chaque terme des équations IV-21 à IV-23 est décrit dans le chapitre III.

Nous pouvons alors tracer la variation des capacités effectives de chaque inverseur 3D, et la comparer aux références 2D (contact à prise et contact en ruban). La Figure IV-27 représente la variation de la capacité effective de chaque inverseur (C_{TOT}) normalisée par la capacité effective de l'inverseur 2D utilisant des contacts à prise.

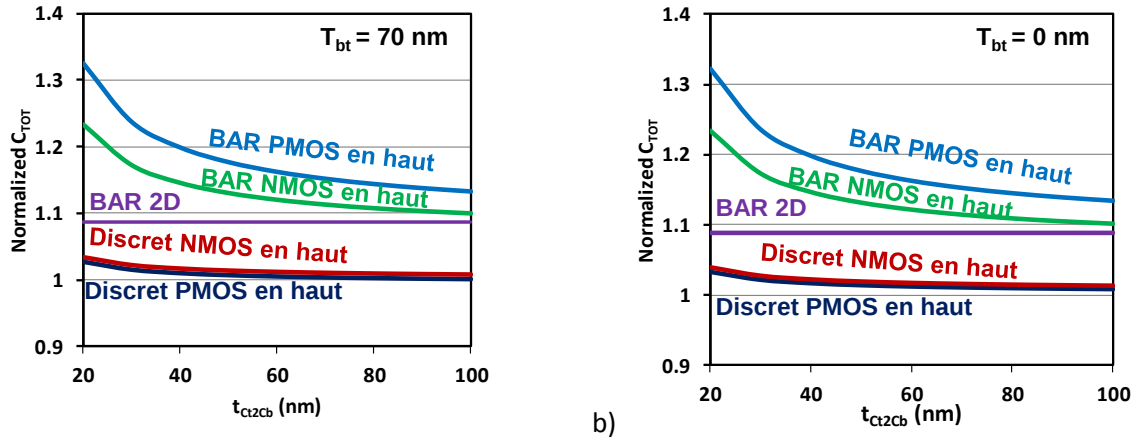


Figure IV-27 : Tracé de la variation de la capacité effective de chaque inverseur normalisée par celle de l'inverseur 2D utilisant les contacts à prise en fonction de la distance entre deux contacts sur les deux sources (t_{Ct2Cb}) pour deux valeurs distinctes entre les deux niveaux de transistors (t_{bt}) : a) $t_{bt}=70$ nm et b) $t_{bt}=0$ nm. Les deux schémas de contact sont considérés (contact à prise (plug) et contact en ruban (BAR)).

Comme nous l'avons constaté précédemment, l'inverseur 3D avec des contacts en ruban et le PMOS en haut est le plus pénalisé. Cependant, si t_{Ct2Cb} est relâché (>60 nm), les capacités parasites dues à l'intégration 3D uniquement (donc en le comparant à l'inverseur 2D avec des contacts en ruban) reste inférieure à 10%. Les inverseurs 3D avec des contacts à prise sont très peu pénalisés par les capacités parasites et semblent être compétitifs à comparer à l'inverseur 2D. Enfin, si nous comparons uniquement les inverseurs 2D entre eux, on peut constater que la pénalité capacitive du passage du contact à prise au contact en ruban est d'approximativement 10%.

Enfin, nous pouvons tracer la variation du délai de chaque inverseur à l'aide de la méthode décrite au paragraphe (IV.A.1), normalisé par celui de l'inverseur 2D avec des contacts à prise, dans les mêmes conditions que dans la Figure IV-27. On obtient alors la Figure IV-28. Comme attendu, la variation du délai est identique à la variation de la capacité effective de chaque inverseur car nous avons considéré que les caractéristiques courant-tension sont les même dans chaque cas.

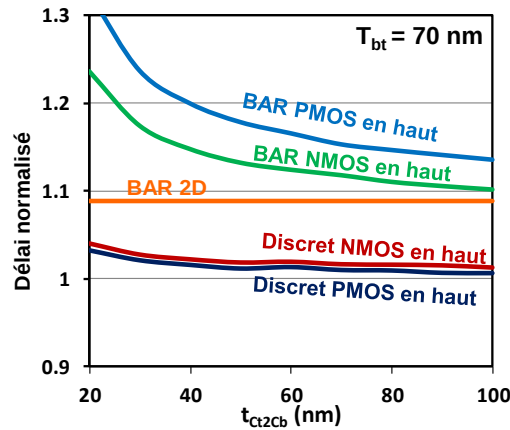


Figure IV-28 : Tracé de la variation du délai de chaque inverseur, normalisé par celui de l'inverseur 2D avec des contacts à prise en fonction de t_{Ct2Cb} , pour les deux schémas de contact (à prise (plug) et en ruban (BAR)).

IV.C.7. Impact du couplage dynamique entre transistor du haut et transistor du bas dans l'inverseur 3D.

Dans le paragraphe précédent, nous avons considéré que les variations de dimensions de l'inverseur 3D avaient d'impact uniquement sur le délai au travers des capacités parasites. Il a cependant été démontré par [Batude 11] qu'il existe un couplage électrostatique entre le canal du transistor du haut et la grille du transistor du bas, à travers l'oxyde enterré. Ce couplage se traduit par un décalage de la tension de seuil du transistor du haut, proportionnel à la polarisation de grille. Il est à noter que comme nous travaillons dans un inverseur, les grilles sont connectées et nous pouvons donc bien parler de « la » polarisation de grille car elle est commune aux deux transistors.

Ce décalage de tension de seuil dépend de la distance entre la grille du transistor du bas et le canal du transistor du haut ($t_{bt}+t_{box}$), mais également du type du transistor du haut. Afin d'intégrer cet effet dans nos simulations d'inverseurs, nous estimons ce décalage de V_t par simulation numériques [synopsis]. Nous représentons sur la Figure IV-29 la variation de ce décalage de tension de seuil en fonction de la polarisation de grille.

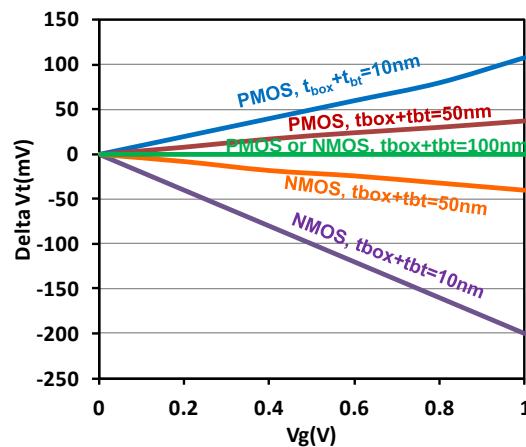


Figure IV-29 : variation du décalage de tension de seuil (ΔV_t) en fonction de la polarisation de grille V_g avec différentes valeurs de t_{bt} . La mention NMOS ou PMOS indique la nature du transistor du haut.

Ensuite, nous incorporons cette dépendance de la tension de seuil avec la polarisation de grille et relançons les simulations de délai, toujours avec la méthode décrite au paragraphe IV.A. La Figure IV-30 représente les résultats de ces simulations pour des variations de t_{bt} (a) et t_{ct2cb} (b). Les traits pleins et pointillés indiquent respectivement les valeurs du délai avec et sans prise en compte du couplage dynamique de la tension de seuil avec la polarisation de grille. Comme nous pouvions nous y attendre au vu de la Figure IV-29, si le PMOS est le transistor du haut, le délai augmente à cause du couplage dynamique car sa tension de seuil augmente avec la tension grille, et donc le niveau de courant débité diminue. Si le NMOS est le transistor du haut, le délai diminue car la tension de seuil baisse lorsque la polarisation de grille augmente, et donc le niveau de courant débité augmente.

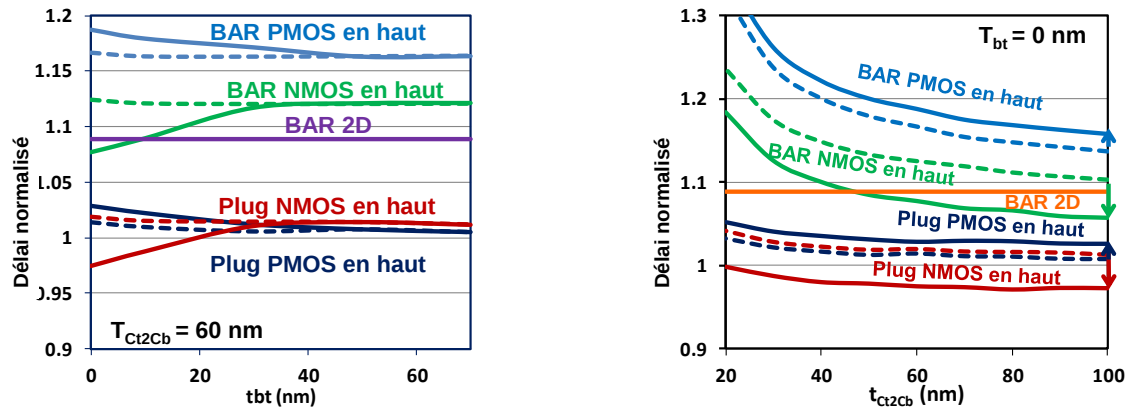


Figure IV-30 : Variation du délai des inverseurs normalisé par le délai de l'inverseur 2D avec contacts à prise avec t_{bt} (a) et t_{Ct2Cb} (b), avec prise en compte du couplage dynamique de la tension de seuil du transistor du haut avec la polarisation de grille. Les pointillés représentent les simulations sans prise en compte du couplage dynamique. Les deux schémas de contact sont considérés (à prise (plug) et en ruban (BAR)).

La Figure IV-30 indique que, grâce à ce couplage électrostatique entre les deux niveaux de transistors, l'inverseur peut être plus rapide que la référence 2D classique. Cependant, ce gain en vitesse est dû à un décalage de tension de seuil, ce qui signifie que le niveau de fuite doit également augmenter. Sa valeur ne pouvant être évaluée précisément et simplement avec notre outil actuel et elle devra d'être estimée lors de prochaines études, effectuées avec d'autres méthodes d'analyse. L'autre point négligé par notre approche est d'avoir omis le gain apporté par la réduction de longueur des lignes d'interconnexion sur le délai d'un inverseur 3D. Ce point ne pouvait également pas être analysé avec notre méthode d'étude, mais son évaluation semble possible avec l'utilisation d'outils de CAO conventionnels, qu'il faudra d'abord adapter à l'intégration 3D monolithique (notamment la prise en compte de la troisième dimension).

IV.D. Conclusion du chapitre

Au cours de ce chapitre, nous avons utilisé les modèles développés dans les chapitre II et III afin d'estimer la vitesse d'une chaîne d'inverseurs. Nous avons d'abord décrit une méthode itérative et numérique qui permet l'estimation du délai d'une chaîne d'inverseur pour une capacité équivalente de 1fF à partir des caractéristiques courant-tension d'un couple NMOS/PMOS d'une technologie CMOS donnée, obtenues par le modèle analytique du chapitre II. Cette méthode est plus précise que la méthode 'CV/ I_{eff} ' et reste suffisamment rapide (quelques secondes de calcul). Nous avons ensuite explicité la méthode d'estimation de la capacité équivalente d'un inverseur [Wei 11]. Sa valeur est obtenue à partir des expressions des capacités parasites données dans le chapitre III.

Avec cette méthode d'évaluation de performances, nous avons mené deux études comparatives bien distinctes. La première a pour objectif d'estimer la performance circuit dans le cadre du nœud technologique 20nm, obtenue par co-intégration de dispositif IFQW III-V/Ge [Hellings 10] et de la comparer à des références silicium (transistor conventionnel sur substrat massif et FDSOI). Les modèles des chapitre II et III ont été utilisés tel quel pour les références silicium. Pour les dispositifs IFQW, le modèle de courant drain-source, et plus particulièrement la correction due à la résistance d'accès a dû être modifiée pour pouvoir reproduire les caractéristiques courant-tension obtenues par simulations numériques Monte-Carlo. Les dispositifs IFQW ne sont pas des transistors MOS classiques et ne présentent pas de zone de recouvrement entre la source ou le drain et la grille. La longueur électrique de ces dispositifs est supérieure à la longueur de grille (on parle de dispositif présentant un « underlap » en anglais), ce qui justifie d'une part la valeur élevée de résistance d'accès sur ces dispositifs et d'autre part la dépendance de sa valeur avec la tension de drain. Nous avons ensuite pu démontrer que la co-intégration de dispositifs IFQW III-V/Ge pouvait permettre des délais de propagation d'inverseur très

courts, mais ce délai est très impacté par la variabilité du procédé de fabrication, notamment celle de l'épaisseur de l'espaceur. Enfin, ce type de dispositif présente un niveau de fuite trop élevé, causé par le t_{inv} très faible nécessaire aux PMOS, pour les applications systèmes sur puce.

La seconde étude a pour but d'évaluer l'impact, positif ou négatif, des capacités parasites sur la vitesse d'un inverseur construit par intégration 3D monolithique [Batude 10] et de la comparer à une référence planaire classique dans le cadre du nœud technologique 28nm. Les performances de la référence 2D ont été simplement estimées à partir des modèles des chapitre II et III. Pour tenir compte de la spécificité de l'intégration 3D monolithique, nous avons d'abord identifié les nouvelles composantes de capacités parasites dues à cette intégration. Nous les avons évaluées à partir de la méthodologie générique d'estimation de capacités donnée dans le chapitre III. Nous avons ensuite intégré ces nouvelles capacités parasites dans le calcul de la capacité équivalente de l'inverseur 3D et avons enfin estimé son délai. Afin d'obtenir une évaluation de délai plus pertinente et réaliste, nous avons inclus dans nos simulations la dépendance dynamique de la tension de seuil du transistor du haut avec la polarisation de grille due au couplage capacitif entre les deux niveaux de transistors. L'impact de cet effet sur la fuite de l'inverseur n'a pas été estimé, mais devrait l'être dans une étude plus approfondie, tout comme l'effet de la réduction de la longueur des lignes métalliques d'interconnexion. Il serait alors plus pertinent d'adapter un outil de CAO conventionnel.

- CHAPITRE V -

EVALUTATION DES PERFORMANCES AVEC DES OUTILS DE CAO CONVENTIONNELS.

Dans ce chapitre, nous allons commencer par expliquer comment utiliser les modèles présentés dans les chapitres II et III pour définir un modèle compact, pour leur intégration dans un simulateur de circuit conventionnel comme [ELDO]. On peut alors se demander pourquoi ne pas plutôt utiliser les modèles SPICE industriels classiques, au lieu de développer un nouveau modèle. La réponse est que le modèle compact que nous allons présenter dans ce chapitre, nommé MASTAR VA, est utilisable pour toutes les architectures de transistor CMOS, contrairement au modèle SPICE industriel et qu'il est prédictif pour les paramètres électrostatiques de chaque architecture, mais également pour les capacités parasites ce qui n'est pas le cas non plus pour les modèles SPICE industriels. De plus, MASTAR VA est beaucoup plus flexible, et peut être adapté pour la simulation de transistor plus « exotiques » comme les IFQW III-V/Ge présentés dans le chapitre IV. Enfin, MASTAR VA est plus facile à calibrer car il ne comporte que très peu de paramètres pour décrire une architecture (pour l'essentiel des paramètres technologiques) contrairement aux modèles SPICE industriels qui nécessitent l'entrée de plusieurs centaines de paramètres. Cependant, MASTAR VA n'a pas vocation à remplacer un modèle industriel, car il ne tient notamment pas compte de l'effet de la largeur des dispositifs et des effets de proximité liés aux autres transistors, tous deux indispensables à la conception de circuit. Par conséquent, MASTAR VA sera seulement utilisé pour de l'évaluation de performance au niveau circuit.

Nous utiliserons ensuite cet outil pour comparer les performances dynamiques des trois architectures concurrentes au nœud 20nm :

- L'architecture conventionnelle sur substrat massif (BULK) [Shang 12]
- L'architecture FDSOI [Khakifirooz 12]
- L'architecture Trigate [Auth 12]

V.A. Méthodologie et implémentation du modèle MASTAR VA

V.A.1. Définir un dispositif dans un simulateur CAO

Pour implémenter un modèle compact décrivant le comportement d'un transistor MOSFET dans un simulateur de circuit conventionnel comme [ELDO], les premières questions à se poser sont les suivantes :

- Comment le simulateur considère-t-il un transistor ?
- De quelles grandeurs décrivant le comportement du transistor le simulateur a-t-il besoin ?

Les réponses à ces questions se trouvent dans la documentation du simulateur ELDO, concernant les modèles définis par l'utilisateur (UDM pour User Defined Model) [ELDO UDM manual]. Pour le simulateur, le transistor MOSFET est un dispositif à quatre électrodes : le drain (D), la source (S), la grille (G) et le substrat (B). A chaque électrode est associé un potentiel (V_d , V_s , V_g , V_b) qui est une variable d'entrée. A partir de ces potentiels, le modèle compact doit fournir une description du comportement statique au simulateur par l'intermédiaire de trois courants :

- Le courant drain-source (I_{ds} , évalué dans le chapitre II).
- Les courants substrat-source et substrat-drain (considéré nul dans notre cas).
- Les courants de grille (considérés nul, leur prise en compte peut être réalisée similairement à [MASTAR]).

Pour prendre en compte l'impact de la résistance d'accès coté source et coté drain sur le courant, il est commun d'introduire deux nœuds internes (D' ($V_{d'}$) et S' ($V_{s'}$)), directement aux bornes du générateur de courant drain-source (cf Figure V-1). Ce générateur de courant symbolise le transistor MOSFET idéal (i.e résistance d'accès nulle) et le courant qu'il débite est donc évalué à l'aide du modèle décrit dans le paragraphe II.D, en prenant pour tension drain-source $V_{ds}=V_{d'}-V_{s'}$. En plaçant une résistance entre l'électrode liant le dispositif au reste du circuit (S et D) et l'électrode interne (S' et D'), on prend en compte la chute de tension due à la résistance d'accès.

Le modèle doit aussi procurer au simulateur une description du comportement dynamique du transistor. Il y a ici le choix entre deux solutions :

- Cinq capacités :
 - La capacité grille-source C_{gs} .
 - La capacité grille-drain C_{gd} .
 - La capacité grille-substrat C_{gb} (considérée nulle dans notre étude).
 - La capacité substrat-source C_{bs} (considérée nulle dans notre étude).
 - La capacité substrat-drain C_{bd} (considérée nulle dans notre étude).
- Trois charges :
 - Charge de source Q_s .
 - Charge de drain Q_d .
 - Charge de grille Q_g .

Sachant que ces deux descriptions sont liées par les relations suivantes :

$$C_{ij} = -\frac{dQ_i}{dV_j} \quad \text{si } i \neq j \quad \text{Eq. V-1}$$

$$C_{ij} = \frac{dQ_i}{dV_j} \quad \text{si } i = j \quad \text{Eq. V-2}$$

Où les indices i et j représentent les électrodes concernées.

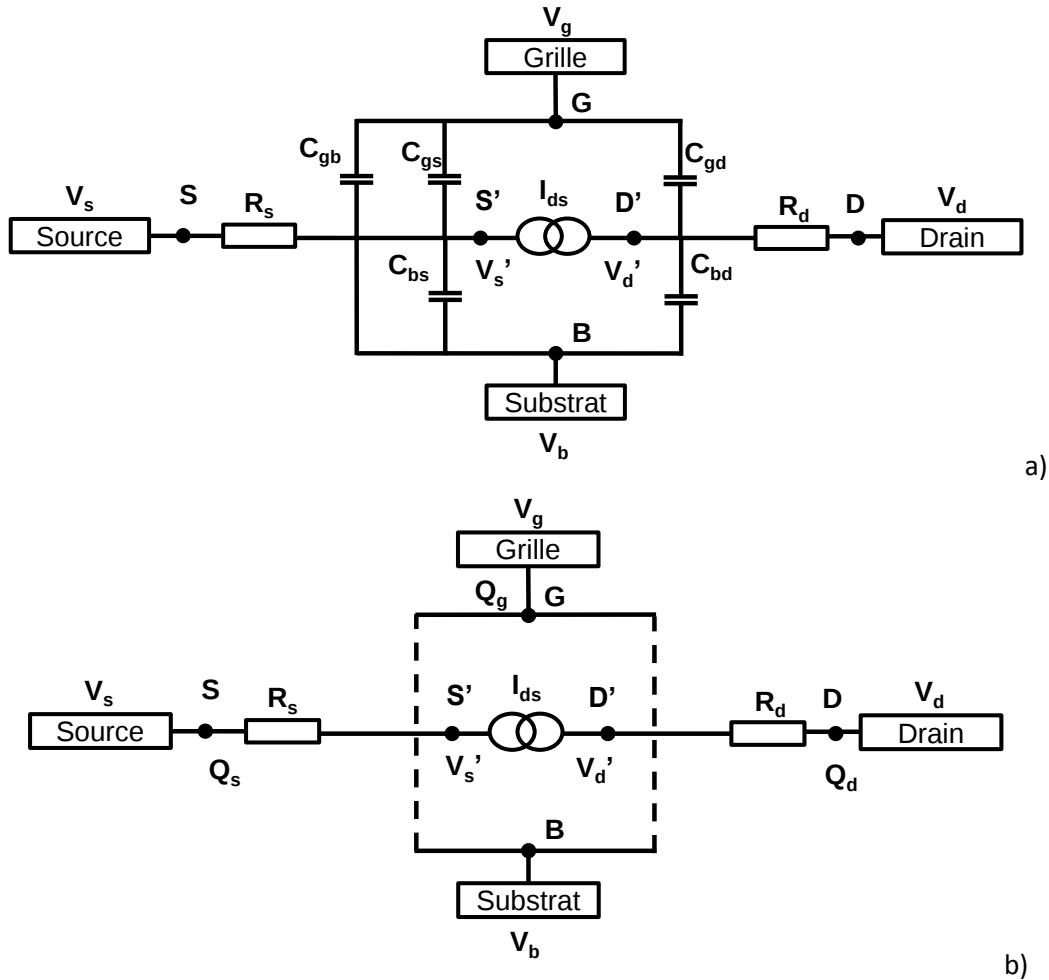


Figure V-1: schéma équivalente du transistor MOSFET pour le simulateur dans le cas où le régime dynamique est décrit par les cinq capacités (a) puis par les trois charges (b).

Pour être utilisable dans le simulateur de circuit conventionnel, la description du régime dynamique doit satisfaire les contraintes suivantes :

- Les capacités doivent être symétriques et réciproques
- La conservation de la charge doit être respectée.
- Les expressions doivent être continues, ainsi que leurs dérivées.

Le modèle de Meyer [Meyer 71] répond à cette problématique en donnant des expressions analytiques assez simples de chacune des seize capacités intrinsèques du transistor MOSFET [Arora 93]. Cependant, cette approche ne garantit pas la conservation de la charge [Ward 73] [Yang 83] et les expressions ne sont pas continues entre tous les régimes. [Mac Andrew 02] propose un modèle de charge et de capacités qui respectent toutes les contraintes énoncées ci-dessus. Cependant, les équations sont très complexes et difficiles à adapter à notre modèle de courant (courant et charge doivent être bien corrélés pour assurer les convergences des simulations) puis à implémenter au sein d'un modèle compact. Dans notre cas, nous avons choisi la méthode de [BSIM] qui consiste à extraire l'expression de la charge d'inversion de l'expression du courant source-drain. Le régime dynamique est donc décrit dans ce cas par les trois charges. Dans [BSIM], des expressions de charge de déplétion et d'accumulation sont ajoutées pour décrire tous les régimes classiques d'une capacité MOS. La charge totale de grille est alors obtenue par la somme des expressions de charge d'inversion Q_{inv} , d'accumulation Q_{acc} et de déplétion Q_{dep} (chaque charge nécessite des fonctions de raccord pour assurer la continuité). Les charges de source et de drain sont ensuite obtenues par la méthode du partage de charge (charge partitioning en anglais) :

$$Q_g = Q_{inv} + Q_{acc} + Q_{dep} \quad \text{Eq. V-3}$$

$$Q_{inv} = Q_s + Q_d \quad \text{Eq. V-4}$$

$$Q_s = \alpha Q_{inv} \quad \text{Eq. V-5}$$

$$Q_d = (1 - \alpha) Q_{inv} \quad \text{Eq. V-6}$$

Où α est le coefficient de partage de charge, compris en 0 et 1. L'équation V-3 assure le respect de loi la conservation de la charge.

Nous allons utiliser notre modèle compact pour de l'évaluation de performances logique et SRAM des technologies CMOS. Par conséquent, pour le NMOS, nous travaillerons exclusivement avec des tensions positives (négatives pour le PMOS). On peut donc négliger la charge d'accumulation ($V_g < 0$) et même la charge de déplétion qui aura un faible impact sur les performances de l'architecture conventionnelle sur substrat massif, et pas d'impact du tout sur les architectures à films minces car elles utilisent des canaux non dopés. Pour simplifier les expressions, et assurer la convergence des simulations, nous choisirons $\alpha = 0.5$ (donc $Q_s = Q_d = 0.5 Q_{inv}$). Il reste cependant à déterminer une expression continue de la charge d'inversion à partir de notre expression du courant drain-source (II.D). En considérant que le courant de drain-source respecte l'égalité :

$$I_{ds} = \mu_{eff} \frac{W}{L} Q_{inv} V_{ds} \quad \text{Eq. V-7}$$

En utilisant l'équation II-131 on obtient la formule continue de la charge d'inversion :

$$Q_{inv} = C_{ox} \frac{V_{gteff}}{(1 + \frac{V_{dseff}}{L E_c})} \left(1 - \frac{(1+d)V_{dseff}}{2(V_{gteff} + 2kT)} \right) \quad \text{Eq. V-8}$$

Où tous les termes sont définis dans le chapitre II.

En implémentant ce modèle de charge pour l'utiliser dans le simulateur de circuit conventionnel [ELDO], on peut tracer les allures des capacités intrinsèques du transistor, en fonction de la polarisation de grille puis de la polarisation drain-source pour les deux types de transistor (NMOS et PMOS).

Cependant, cette description ne tient compte que des charges intrinsèques du transistor, c'est-à-dire entre les nœuds virtuels D' et S' et la grille G. Il faut également décrire les charges dues aux nœuds « réels », c'est-à-dire le drain et la source (D et S). On parle alors de charges extrinsèques et celle-ci sont déterminées par les capacités parasites, dont toutes les composantes sont modélisées dans le chapitre III, quelle que soit l'architecture. Nous aurons alors deux charges additionnelles, données par les équations :

- Pour les architectures planaires :

$$Q_{gse} = (C_{of} + C_{ov} + C_{gepi} + C_{pcca} + C_{corner})V_{gs} \quad \text{Eq. V-9}$$

$$Q_{gde} = (C_{of} + C_{ov} + C_{gepi} + C_{pcca} + C_{corner})V_{gd} \quad \text{Eq. V-10}$$

- Pour les architectures non planaires :

$$Q_{gse} = (C_{gfin} + C_{ov} + C_{gepi} + C_{pcca} + C_{corner})V_{gs} \quad \text{Eq. V-11}$$

$$Q_{gde} = (C_{gfin} + C_{ov} + C_{gepi} + C_{pcca} + C_{corner})V_{gd} \quad \text{Eq. V-12}$$

Où Q_{gse} et Q_{gde} sont les charges extrinsèques entre respectivement la grille puis la source et la grille et le drain. La Figure V-2 représente la variation des capacités totales (extrinsèques et intrinsèques) en fonction des polarisations de grille et de drain. On constate bien que les courbes obtenues sont continues et seront donc utilisables dans un simulateur de circuit conventionnel.

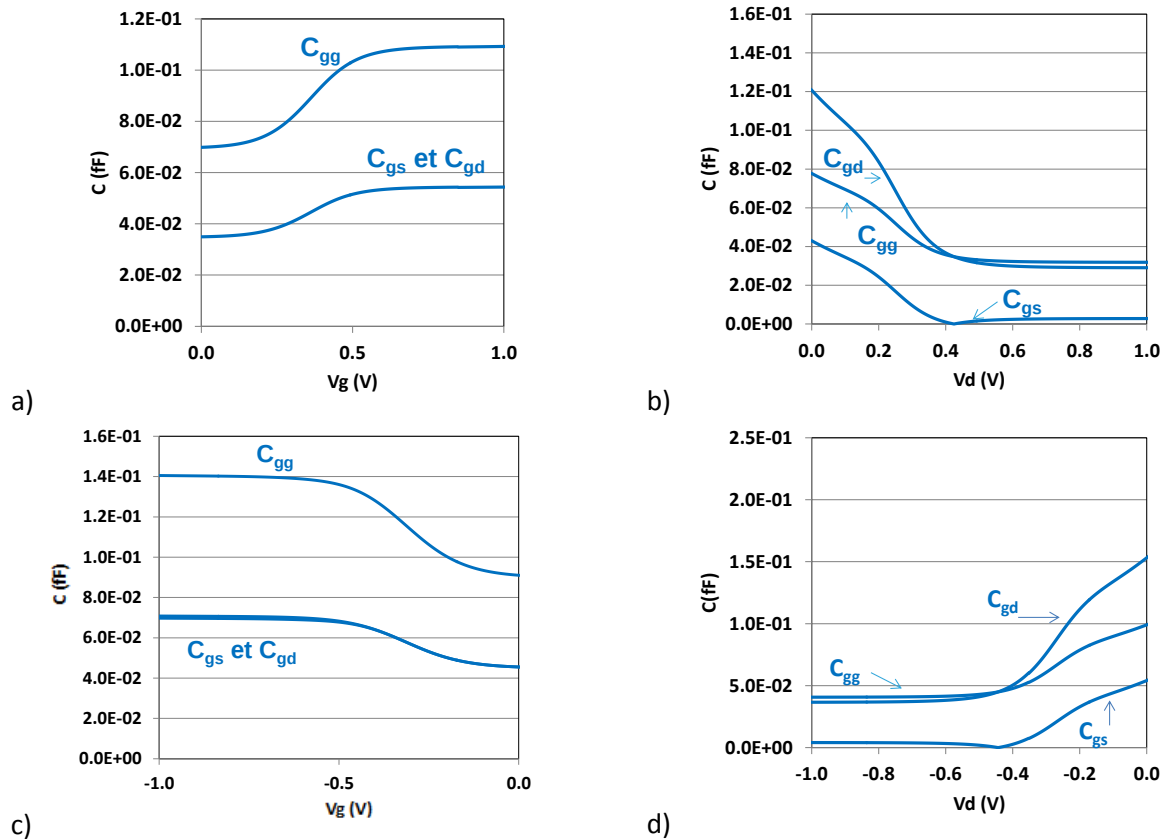


Figure V-2 : Capacités intrinsèques d'un transistor MOSFET pour le NMOS (a,b) et pour le PMOS (c,d).

Le modèle compact ainsi obtenu est nommé MASTAR VA.

V.A.2. Flot de simulation

A partir de la description de l'architecture d'une technologie CMOS, nous pouvons désormais donner les caractéristiques statiques et dynamiques du couple NMOS/PMOS de cette technologie, puis d'évaluer les performances circuit au travers de simulations effectuées avec un simulateur de circuit conventionnel [ELDO]. Le flot de simulation de MASTAR VA est illustré par la Figure V-3.

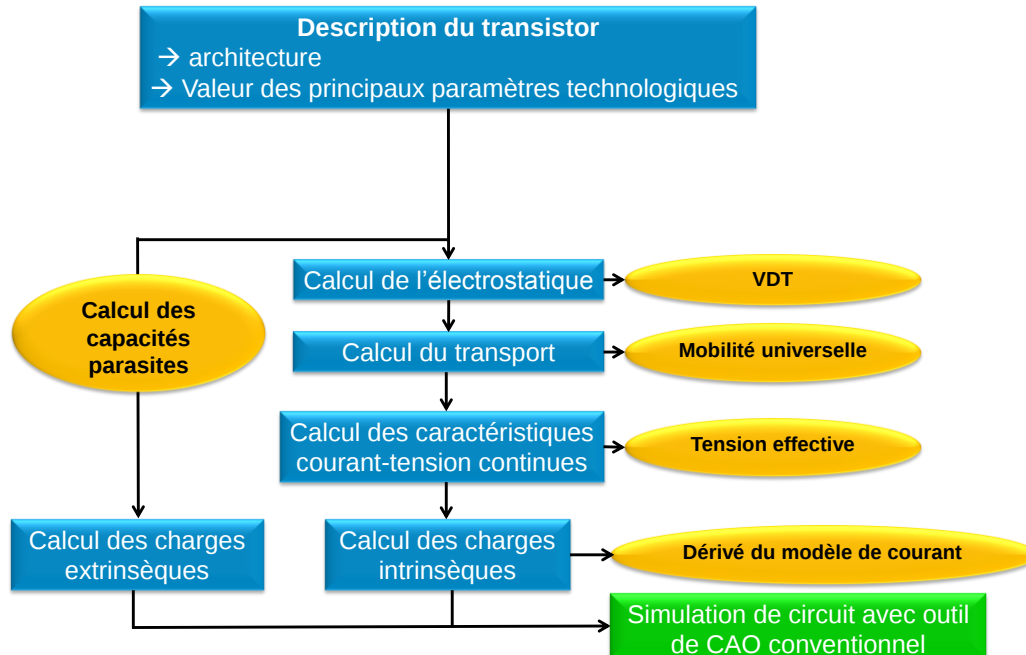


Figure V-3 : flot de simulation de MASTAR VA.

V.A.3. Circuits disponibles

Avec MASTAR VA, nous pouvons simuler de nombreux circuits :

- Les chaînes d'inverseurs, décrits au cours du chapitre IV, pour différents FanOut. Pour être encore plus représentatif des circuits « réels », nous avons placé une charge fixe à la sortie de chaque étage. Ce circuit est illustré sur la Figure V-4-d.
- Les anneaux résonnants d'inverseurs, qui n'apportent guère plus d'informations concernant la vitesse d'une technologie, mais qui permet de tester la robustesse du modèle et sa convergence. Ce type de circuit est illustré sur Figure V-4-e.
- Des chaînes constituées de portes logiques un peu plus complexes : des portes « non et » (en anglais NAND Figure V-4-b) et des portes « non ou » (en anglais NOR Figure V-4-c). Les résultats de ces simulations n'apportent peu, voire pas d'information supplémentaire comparés à ceux obtenus avec les inverseurs. Comme ces simulations sont bien plus longues à effectuer par le simulateur (temps multiplié par ~5 car chaque étage compte un plus grand nombre de transistors que pour une chaîne d'inverseur) nous avons choisi d'exploiter uniquement les résultats des chaînes d'inverseurs. Ces simulations ont cependant permis d'éprouver la robustesse du modèle.
- Des cellules mémoires SRAM (Static Random Access Memory Figure V-4-f.), constituées de deux inverseurs tête-bêche et de deux transistor d'accès, soit au total six transistors :
 - 2 PG (Pass Gate) : les transistors d'accès qui sont des NMOS.
 - 2 PD (Pull Down) : le PMOS de chaque inverseur
 - 2 PU (Pull Up) : le NMOS de chaque inverseur

- Des bascules FLIP-FLOP, qui ont permis d'évaluer la robustesse du modèle de charges intrinsèques de MASTAR VA.
- Un chemin critique d'un contrôleur d'accès mémoire au format DDR3 [Arnaud 11] [Choi 11], illustré sur la Figure V-4-g. Ce circuit est composé d'inverseur, de bascules... qui est représentatif de la performance circuit réelle d'une architecture à un nœud technologique donné.

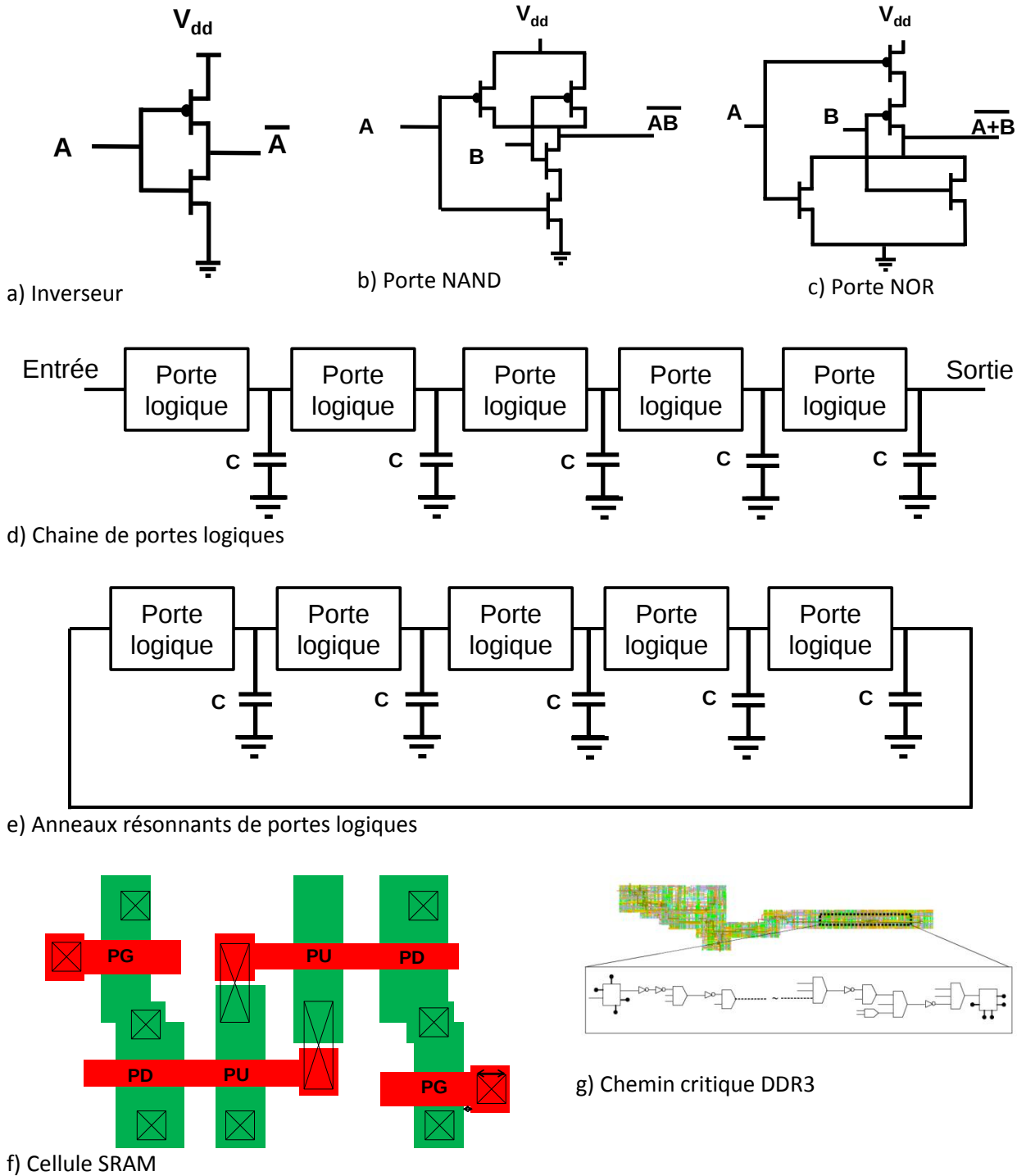


Figure V-4 : Schéma des différents circuits simulés avec de MASTAR VA.

Ces circuits ont été simulés, et la Figure V-5 représente les résultats de simulations, obtenus pour des transistors conventionnels sur substrat massif quelconques :

- Figure V-6-a : résultats de simulations d'une chaîne d'inverseurs : la première courbe (verte, en haut du graphe) est le signal d'entrée créneau et les courbes suivantes (en dessous) sont les signaux obtenus en sortie de chaque étage.
- Figure V-6-b : résultats de simulations d'une cellule SRAM et tracé de la courbe papillon.
- Figure V-6-c : résultats de simulations d'un anneau résonnant constitué d'inverseurs et tracé de la tension de sortie d'un étage.

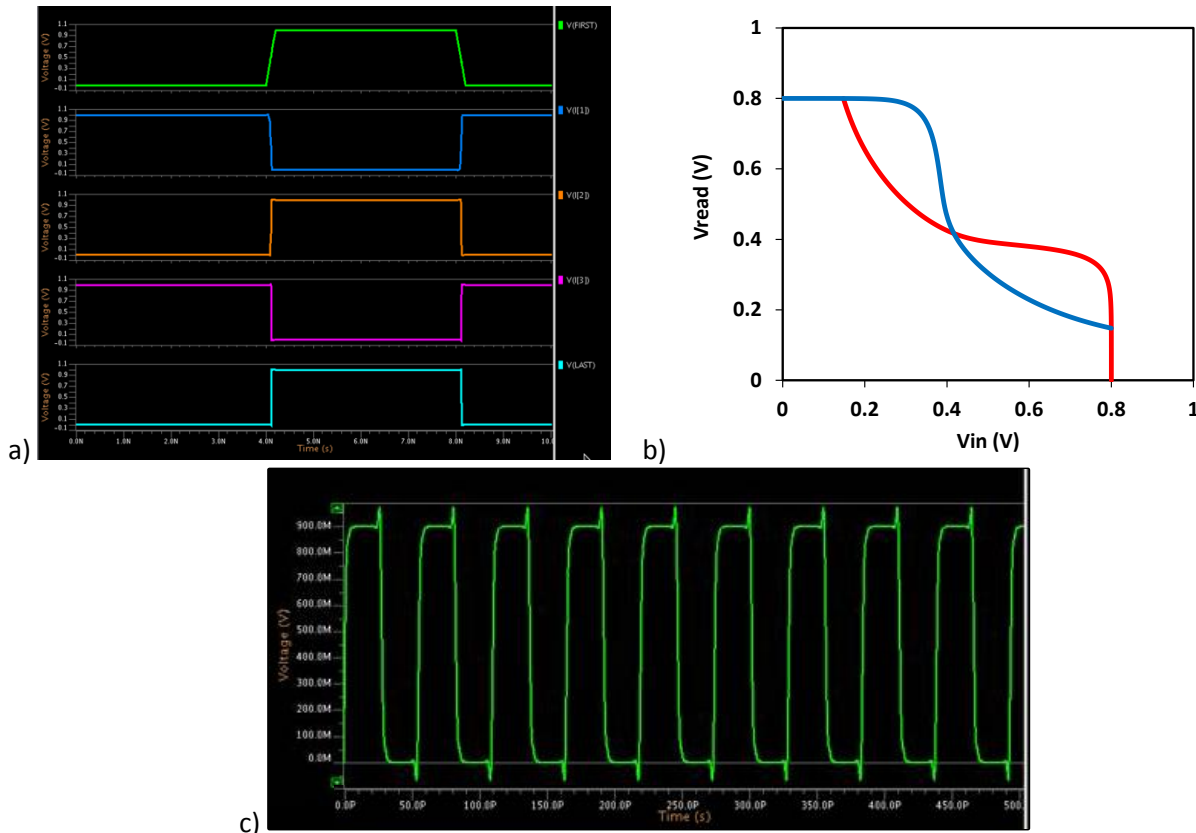


Figure V-5 : Capture d'écran de résultats de simulations effectuées sur certains circuits mentionnés ci-dessus avec ELDO: a) chaîne d'inverseur, b) cellules SRAM et c) Anneau résonnant.

V.A.4. Prise en compte des parasites dans le back-end

La méthodologie de simulation de circuits décrite au cours des paragraphes précédents négligent l'impact des interconnexions sur les performances dynamiques. Pour en tenir compte, il faut évaluer leur résistance et leur capacité, dont les valeurs sont très dépendantes :

- du nœud technologique : les règles de dessin des différents niveaux de métallisations des interconnexions (BEOL pour Back End Of Line) sont réduites de nœud en nœud et les matériaux (métaux et isolants) peuvent varier d'un nœud à l'autre.
- de l'architecture du transistor : pas de répétition d'un contact (CPP), distance grille-contact..
- du circuit étudié et de sa complexité : le nombre de dispositifs qui le compose, le nombre de niveau de métallisation nécessaire...
- du dessin du circuit : largeur des dispositifs, distance entre chaque transistor, surface totale du circuit.

Leur évaluation est très complexe à effectuer de manière analytique et nous choisirons alors d'utiliser un outil d'extraction automatique des parasites dans les interconnexions : le PEX (pour Parasitics EXtraction). A partir du dessin du circuit, et des règles de dessin des interconnexions, le PEX fournit les valeurs d'un circuit équivalent RC, qui va permettre la prise en compte de la perte de vitesse due aux interconnexions, et donc permettre une évaluation plus réaliste, tant sur la vitesse que sur la consommation dynamique d'un circuit.

Ce type d'évaluation étant très complexe, nous avons dû utiliser ce qui était à disposition au moment de l'étude, à savoir le PEX développé pour l'architecture conventionnelle sur substrat massif du nœud 20nm. Nous avons considéré que les parasites liés aux interconnexions étaient les mêmes, quelle que soit l'architecture des transistors composant les circuits.

V.A.5. Définition de source de variation aléatoire des paramètres technologiques pour prise en compte de la variabilité.

L'utilisation d'un simulateur de circuit conventionnel rend possible simplement l'estimation de l'impact de la variabilité des dimensions du transistor due au procédé de fabrication (longueur de grille, dopage...) au niveau des performances électriques du transistor (tension de seuil, courant de fuite, de saturation..) mais également au niveau des performances du circuit (fréquence...). En effet, le simulateur peut générer des valeurs de paramètres pseudo-aléatoires qui suivront la distribution statistique prédéfinie par l'utilisateur. Il suffit ensuite de lancer un nombre de simulations suffisant (~500) et d'extraire les distributions des paramètres de sortie estimés par MASTAR VA: tension de seuil (V_t), courant de saturation (I_{on}), fréquence d'une chaîne d'inverseurs... Le flot de simulation (Figure V-6) est alors un peu différent de celui illustré par la Figure V-3.

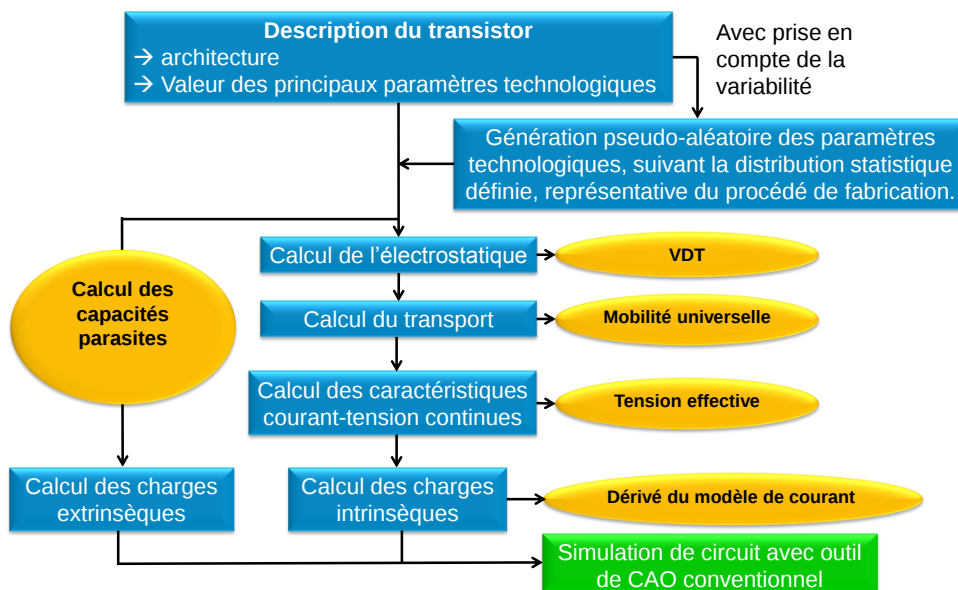


Figure V-6 : flot de simulation de MASTAR VA, avec prise en compte de la variabilité due au procédé de fabrication.

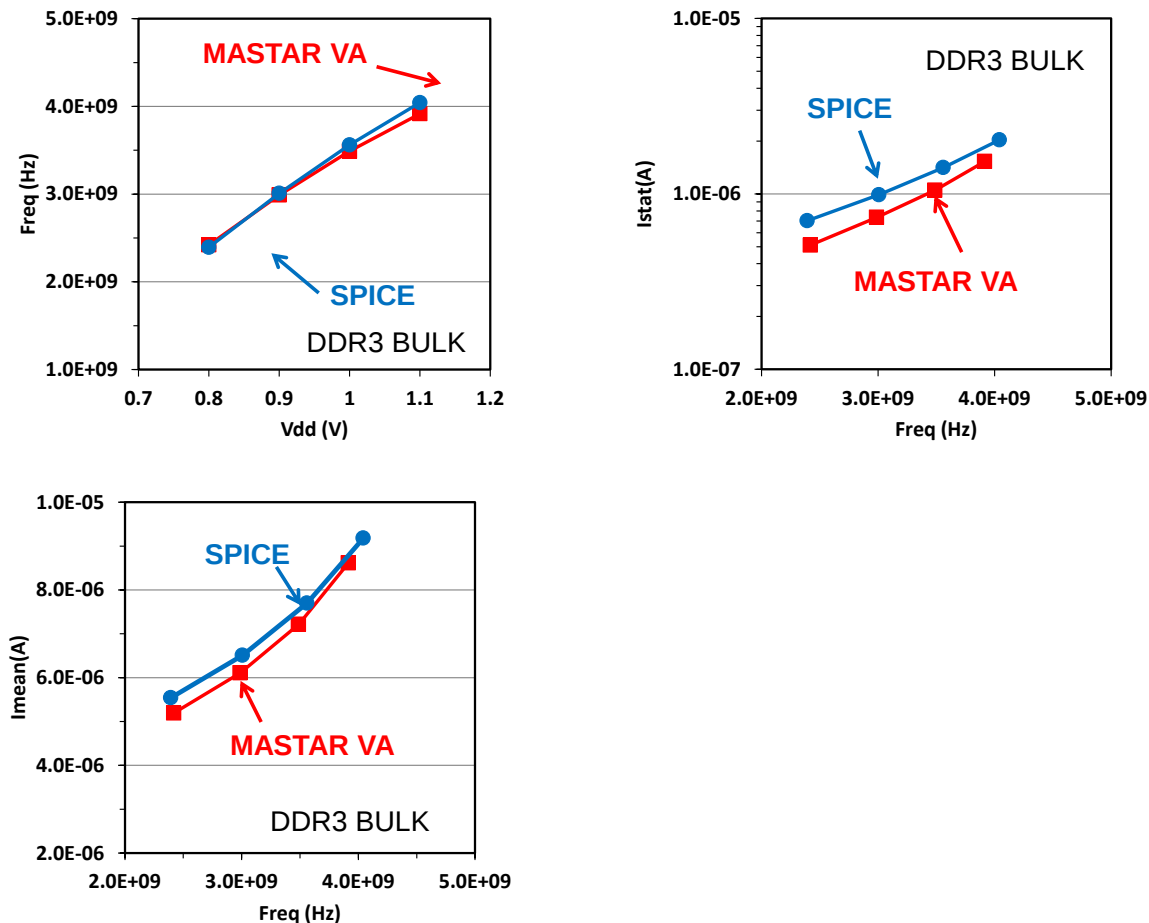
Par conséquent, connaissant la distribution statistique suivie par les dimensions du transistor due à la variabilité du procédé de fabrication il est possible d'estimer la distribution résultante des paramètres électriques du transistor, ou des performances d'un circuit donné. Nous pouvons alors évaluer la robustesse à la variabilité du procédé de fabrication d'une architecture donnée.

V.A.6. Validation de la méthode par comparaison avec un modèle SPICE

Nous avons montré que les simulations convergent dans le paragraphe V.A.3. Il faut maintenant valider que les performances dynamiques qu'il prévoit sont justes. Pour ce faire, nous reproduisons les caractéristiques courant-tension extraites d'un modèle industriel ajusté sur des mesures silicium représentatives du nœud 20nm pour deux architectures (conventionnel sur substrat massif et FDSOI). Ensuite, nous évaluons les capacités parasites en entrant les dimensions de l'architecture considérées dans le modèle industriel. Enfin, nous effectuons des simulations pour différentes tension d'alimentation de chemin critiques DDR3 pour l'architecture conventionnelle sur substrat massif et d'anneaux résonnants d'inverseur FanOut 1 avec MASTAR VA et le modèle industriel SPICE. Nous extrayons de ces simulations :

- La fréquence (notée f_{req}) qui traduit la vitesse de l'architecture pour un circuit donné.
- Le courant de fuite statique (I_{stat}). On trace plutôt la puissance statique ($P_{stat}=I_{stat}\times V_{dd}$) qui est représentative de la consommation statique.
- Le courant dynamique (noté I_{dyn}). On préfère souvent tracer la puissance dynamique ($P_{dyn}=I_{dyn}\times V_{dd}$) qui est représentative de la consommation dynamique.
- Pour le chemin critique DDR3, un courant moyen (I_{mean}) est extrait. Il est corrélé à la fois au courant statique et au courant dynamique, mais un lien clair entre ces trois courants n'existe pas, il sera alors difficile d'extraire une information des valeurs de courant moyen sur un chemin critique DDR3.

La Figure V-7 montre que MASTAR VA reproduit bien les variations de fréquence, courant dynamique ou moyen et courant statique en fonction de la tension d'alimentation V_{dd} obtenues avec le modèle SPICE industriel.



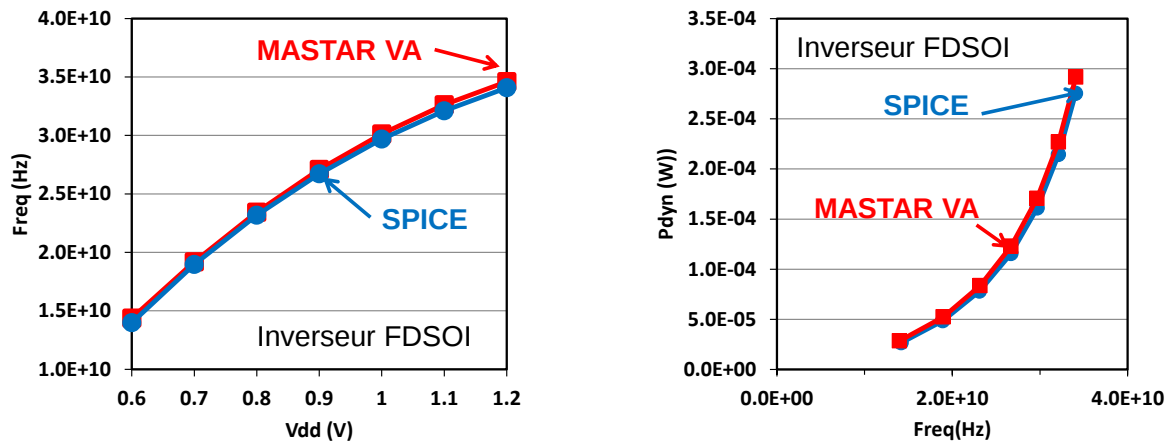


Figure V-7 : comparaison des valeurs de fréquence, puissance dynamique et de courant statique obtenues par MASTAR VA et par le modèle industriel (SPICE).

V.B. Comparaison des performances au nœud 20nm

V.B.1. Contexte

Le nœud technologique 20nm sera le seul nœud où trois architectures différentes seront développées industriellement. STMicroelectronics, IBM, Samsung et Global Foundries ont conjointement présenté leur plateforme CMOS utilisant l'architecture conventionnelle sur substrat massif [Shang 12]. STMicroelectronics a également présenté sa plateforme FDSOI pour le nœud 28nm [Planes 12] et a annoncé le développement d'une plateforme CMOS 20nm avec l'architecture FDSOI. Enfin, Intel a présenté sa plateforme 22nm utilisant l'architecture Trigate [Auth 12].

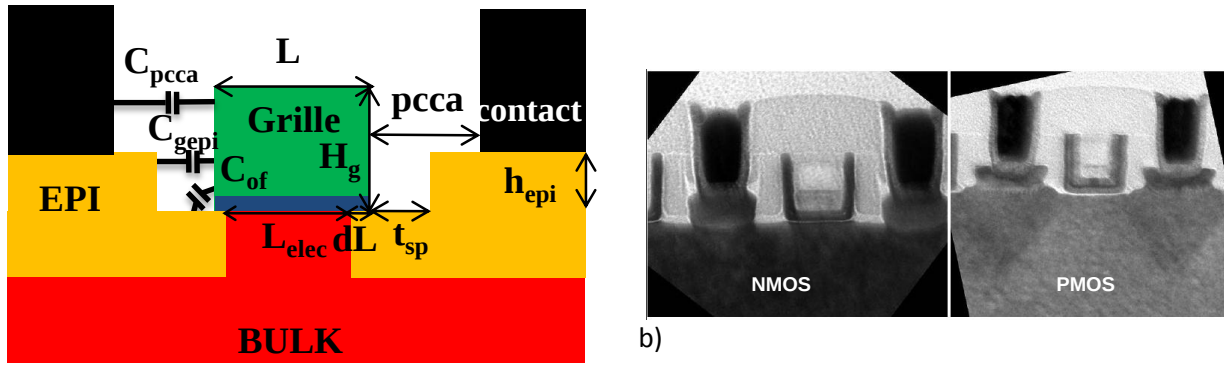
Il est donc parfaitement justifié d'une part, et intéressant d'autre part de comparer les performances de ces trois architectures pour des circuits typique du nœud 20nm. Pour garantir une comparaison équitable, les trois architectures seront alignées au même niveau de fuite statique à leur tension d'alimentation nominale. Pour plutôt étudier l'aspect performance (i.e vitesse), nous nous fixons un niveau de fuite statique élevé : $I_{off}=60nA/\mu m$. Enfin, pour être représentatif de l'état de l'art des performances pour chaque architecture, nous définirons nos structures de manière à reproduire les dernières (et meilleures) performances reportées dans la littérature, à savoir la pente sous le seuil(SS), le DIBL et le compromis I_{on}/I_{off} .

V.B.2. Définition des dispositifs

Dans ce paragraphe, nous proposons de définir les trois architectures que nous allons comparer, en donnant les dimensions et les performances statiques de chacune. Pour une comparaison juste et typique du nœud technologique 20nm, nous fixerons le pas de répétition d'une grille et d'un contact à CPP=90nm.

V.B.2.a) L'architecture conventionnelle sur substrat massif

Pour définir ce type de dispositifs, nous nous baserons sur les travaux de [Cho 11] et [Shang 12] qui présentent des réalisations technologiques de cette architecture au nœud 20nm ainsi que des mesures de caractéristiques statiques. La Figure V-8-a présente une vue en coupe schématique de l'architecture que nous allons considérer pour la modélisation. Cette dernière est semblable à celle utilisée dans [Shang 12], comme le montre la vue en coupe prise au microscope électronique à transmission de la Figure V-8-b.



a) Figure V-8 : a) Schéma de l'architecture conventionnelle sur substrat massif avec ses dimensions caractéristiques. b) Vue au microscope électronique à transmission (TEM en anglais) de l'architecture utilisée dans [Shang 12].

Le tableau de la Figure V-9 résume les paramètres technologiques que nous avons choisis pour définir notre architecture conventionnelle sur substrat massif ainsi que ses performances statiques. Ces derniers sont ceux indiqués dans [Shang 12], ou, si l'information n'était pas présente dans la publication, estimés à l'aide des règles simples définies dans le chapitre III. Nous nous sommes tout de même accordés la liberté d'augmenter la longueur de grille par rapport à [Shang 12] car nous ne parvenions pas à obtenir la valeur de DIBL mesurée dans la publication avec notre modèle analytique. Ce dernier point est également justifié par le fait que la longueur dite « design » (donc dessinée par le concepteur de circuit) est très souvent différente, et inférieure, à la valeur physique, réellement présente sur le silicium.

	BULK		
type	NMOS	PMOS	Hypothèses
Vdd (V)	0.9	0.9	[Shang 12]
L (nm)	26	26	L est supérieur à [Shang 12] pour obtenir le même DIBL.
CPP (nm)	90	90	[Shang 12]
EOT (nm)	0.9	0.9	[Shang 12]
h_{epi} (nm)	13	13	$L/2$
t_{sp} (nm)	10	10	$pcca/2$
pcca (nm)	21	21	$(CPP-L)/3$
H_g (nm)	35	35	Estimé sur [Shang 12]
DIBL (mV)	110	115	[Shang 12]
SS (mV/dec)	90	90	[Shang 12]
I_{on} ($\mu A/\mu m$)	1178	1248	[Shang 12]
I_{eff} ($\mu A/\mu m$)	570	630	[Shang 12]
I_{off} (nA/ μm)	60	60	[Shang 12]

Figure V-9 : Résumé des paramètres technologiques considérés pour la modélisation de l'architecture conventionnelle sur substrat massif.

On entre alors ces paramètres dans MASTAR VA, puis on ajuste le travail de sortie de la grille (ϕ_m) pour obtenir la valeur de fuite statique visée ($I_{off}=60nA/\mu m$). Pour régler les paramètres de contraintes, nous considérons qu'ils sont utilisés au maximum de leur potentiel. On détermine alors le facteur d'amélioration de la mobilité correspondant à l'aide de [Payet 08], tout en incluant la dégradation de mobilité due à la réduction de longueur de grille donné par [Bidal 09]. De cette manière, les niveaux de courant de saturation obtenus avec MASTAR VA sont proches de ceux reportés par [Shang 12]. Enfin, on trace les caractéristiques courant-tension du NMOS et du PMOS de l'architecture conventionnelle sur substrat massif (Figure V-10)

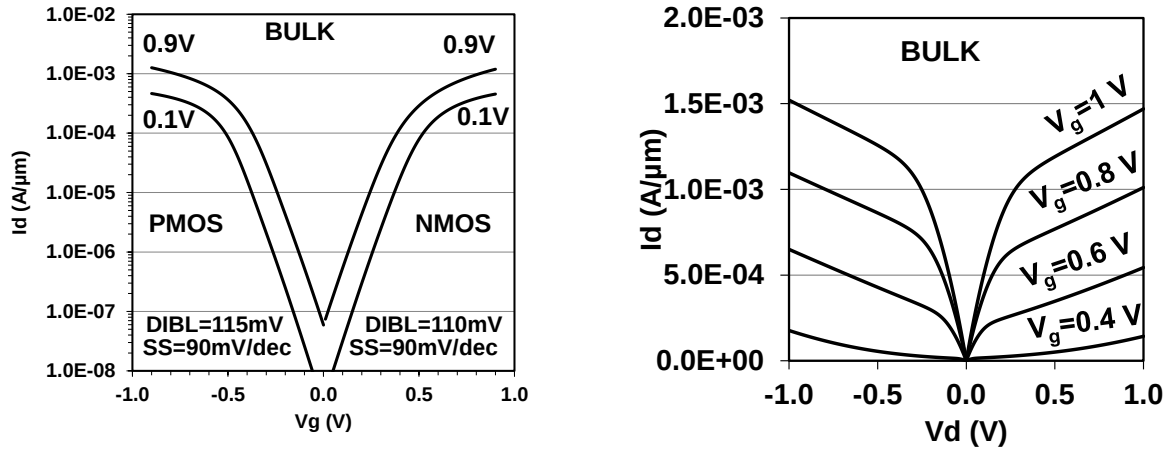


Figure V-10 : Caractéristiques courant-tension de l'architecture conventionnelle sur substrat massif obtenues par MASTAR VA.

V.B.2.b) L'architecture FDSOI

Pour définir l'architecture FDSOI (Figure V-11-b), nous utiliserons les performances statiques données dans [Khakifirooz 12] ainsi que les principales dimensions. Les source-drains fabriqués par épitaxies présentant une facette sont communément utilisés pour les technologies planaires avancées et nous utiliserons les dimensions de [Cheng 11] pour la décrire (Figure V-11-b). Enfin, l'épaisseur d'oxyde enterré n'étant pas précisée dans [Khakifirooz 12], nous utiliserons l'épaisseur donnée dans [Liu 11]. Enfin, les dimensions non fournies par les différentes publications seront estimées, comme dans le paragraphe V.B.2.a), à partir des règles simples utilisées dans le chapitre III.

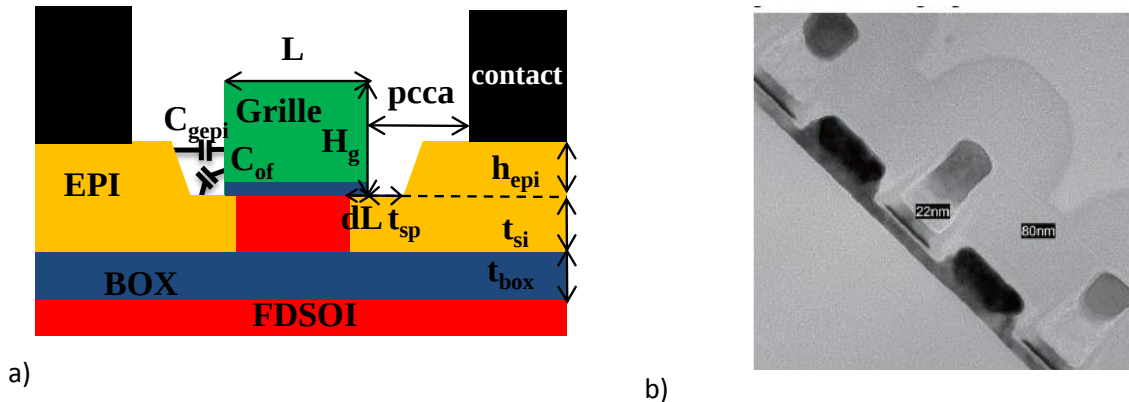


Figure V-11 :a) Schéma de l'architecture FDSOI avec ses dimensions caractéristiques. b) Vue au microscope électronique à transmission (TEM en anglais) de l'architecture utilisée dans [Cheng 11].

La Figure V-12 donne les paramètres technologiques décrivant complètement l'architecture FDSOI.

	FDSOI		
type	NMOS	PMOS	Hypothèses
V_{dd} (V)	1	1	[Khakifirooz 12]
L (nm)	22	22	[Khakifirooz 12]
CPP (nm)	90	90	[Khakifirooz 12]
t_{si} (nm)	6	6	[Khakifirooz 12]
t_{box} (nm)	25	25	[Liu11]
EOT (nm)	0.9	0.9	[Khakifirooz 12]
h_{epi} (nm)	15	15	$\sim 2xt_{si}$
t_{sp} (nm)	8	8	Epitaxie facettée [Cheng 11] $\sim pcca/3$
pcca (nm)	22	22	(CPP-L)/3
H_g (nm)	37	37	Estimée sur [Khakifirooz 12]
DIBL (mV)	75	130	[Khakifirooz 12]
SS (mV/dec)	80	95	[Khakifirooz 12]
I_{on} ($\mu A/\mu m$)	1555	1210	[Khakifirooz 12]
I_{eff} ($\mu A/\mu m$)	942	660	[Khakifirooz 12]
I_{off} (nA/ μm)	60	60	[Khakifirooz 12]

Figure V-12 : Résumé des paramètres technologiques considérés pour la modélisation de l'architecture FDSOI.

En entrant ces paramètres dans MASTAR VA, et en ajustant le travail de sortie de la grille (ϕ_m) pour obtenir la valeur de fuite statique visée ($I_{off}=60nA/\mu m$) puis les paramètres de transport et de résistance d'accès pour garantir le niveau de courant de saturation donné dans [Khakifirooz 12], on peut tracer les caractéristiques courant-tension du NMOS et du PMOS de l'architecture FDSOI (Figure V-12).

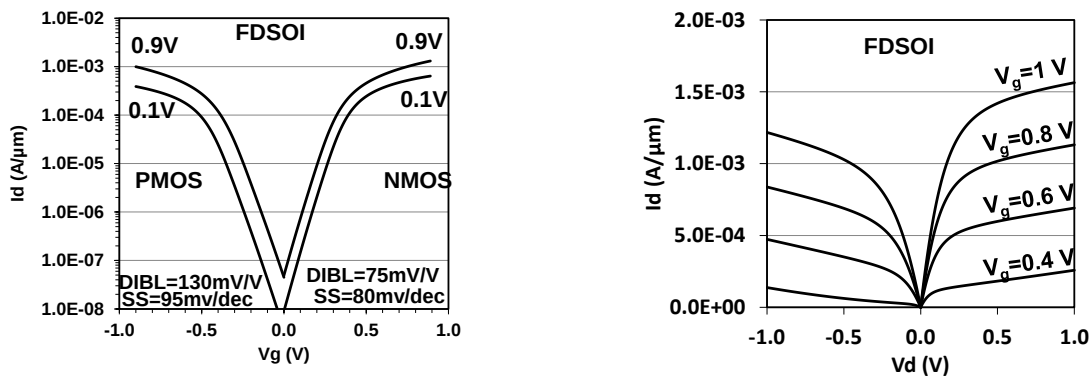
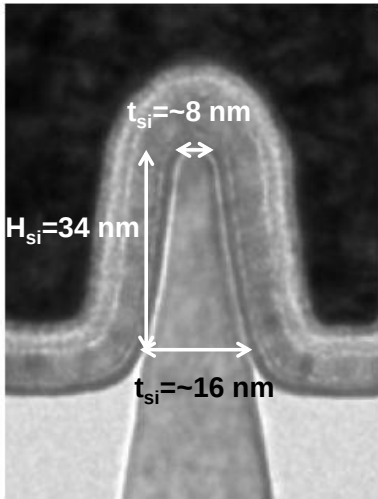


Figure V-13 : Caractéristiques courant-tension de l'architecture FDSOI obtenues par MASTAR VA.

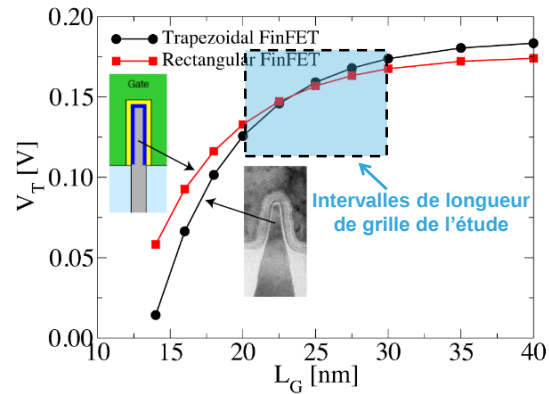
V.B.2.c) L'architecture Trigate

Pour définir l'architecture Trigate, nous utiliserons les performances statiques données dans [Auth 12] ainsi que les principales dimensions. Cependant, l'estimation des dimensions de cette architecture est moins immédiate que pour les architectures planaires, du fait de la nature « 3D » du Trigate. Nous allons l'effectuer en analysant les photos prises au microscope électronique à transmission données dans [Auth 12] ainsi que celles données par

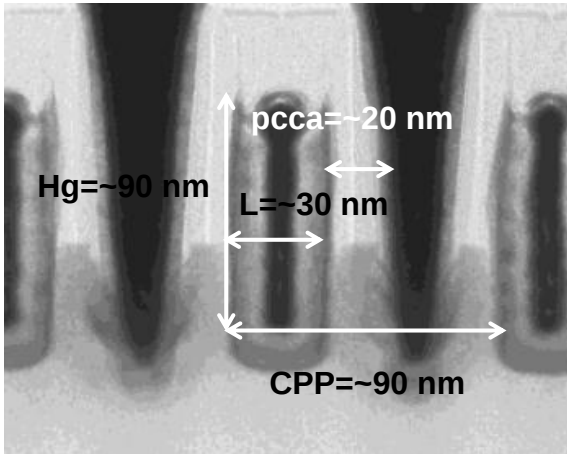
[Chipworks]. La Figure V-4-a montre une vue en coupe dans le fin donnée par [Auth 12]. Les dimensions indiquées sur la photo sont données par [Auth 12]. Or notre modèle de tension de seuil (II.C) n'est valable que pour un transistor double grille et il y en a trois dans le cas du Trigate. Néanmoins, au vue de du rapport de forme du fin ($H_{si}/t_{si} \sim 3$), on peut estimer que la troisième grille (celle qui est sur le fin) n'aura pas d'effet sur l'électrostatique du dispositif. On constate également que la section du fin n'est pas rectangulaire, mais plutôt trapézoïdale. Cependant, il a été démontré dans [GSS-a] que l'électrostatique est à peu près la même pour un Trigate à section rectangulaire et pour un Trigate à section trapézoïdale, dans l'intervalle de longueur de grille compris entre 20 et 30 nm (Figure V-14-b), qui est celui sur lequel nous allons focaliser cette étude. Les Figure V-14-c et Figure V-14-d présentent toutes deux une vue en coupe prise dans le sens du fin. Comme [Auth 12] donne la valeur de la longueur de grille ($L_g=30nm$) et du pas de répétition d'une grille et d'un contact (CPP=90nm), on parvient à estimer les dimensions de l'architecture Trigate (notamment dimensions des espaceurs et distance grille-contact). Enfin au vue de ces photos, il est clair que l'épithaxie ne présente pas de facette.



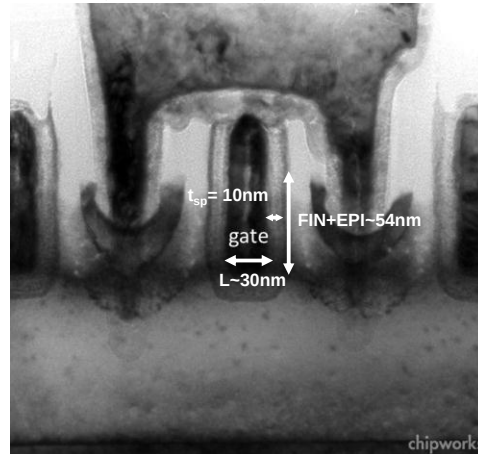
a) Vue en coupe dans le fin [Auth 12]



b) Tracé de variation de la tension de seuil en fonction de la longueur de grille pour un Trigate à section rectangulaire (courbe rouge) et à section trapézoïdale (courbe noire).



c) Vue en coupe entre deux fins [Auth 12]



d) Vue en coupe entre deux fins [Chipworks]

Figure V-14 : Vue du Trigate par microscopie électronique à transmission et courbe extraite de simulation.

Pour simplifier la modélisation, sans perdre en précision, nous considérerons que le fin présente une section rectangulaire. Le calcul du DIBL et de pente sous le seuil effectué avec notre modèle pour $t_{si}=10\text{nm}$ et $EOT=0.9\text{nm}$ (donnée par [Auth 12]) donne les mêmes valeurs que celle donnée dans [Auth 12]. En nous imposant le facteur de forme du fin de [Auth 12] (~ 3), on arrive à un fin rectangulaire d'épaisseur $t_{si}=10\text{nm}$ et de hauteur de fin $h_{si}=30\text{nm}$. De cette manière, la surface de conduction pour un fin est à peu près la même que celle qu'on peut estimer sur la Figure V-14-a. L'architecture Trigate proposée par [Auth 12] est réalisée sur substrat conventionnel massif. Or nous ne disposons pas de modélisation analytique des mécanismes de fuites par le bas du fin à travers le substrat. Nous considérerons donc que le Trigate est réalisé sur substrat SOI, et serons donc un peu optimiste concernant la consommation du dispositif.

De plus, au vue du Finpitch (égal à la somme de l'épaisseur d'un fin et d'un espace entre deux fins) donné par [Auth 12] (60 nm), et la hauteur d'épitaxie estimée sur la Figure V-14-d (20nm) il semble peu probable que l'épitaxie formant les source-drains vienne contacter les fins entre eux. En effet, en considérant que la croissance de l'épitaxie est isotrope et connaissant sa hauteur à partir du haut du fin ($t_{epi}=20\text{nm}$), on arrive à une largeur de 40nm, soit un espace entre les source-drains de chaque fin de $\text{Space_epi}=20\text{nm}$. Cette architecture est illustrée sur la Figure V-15.

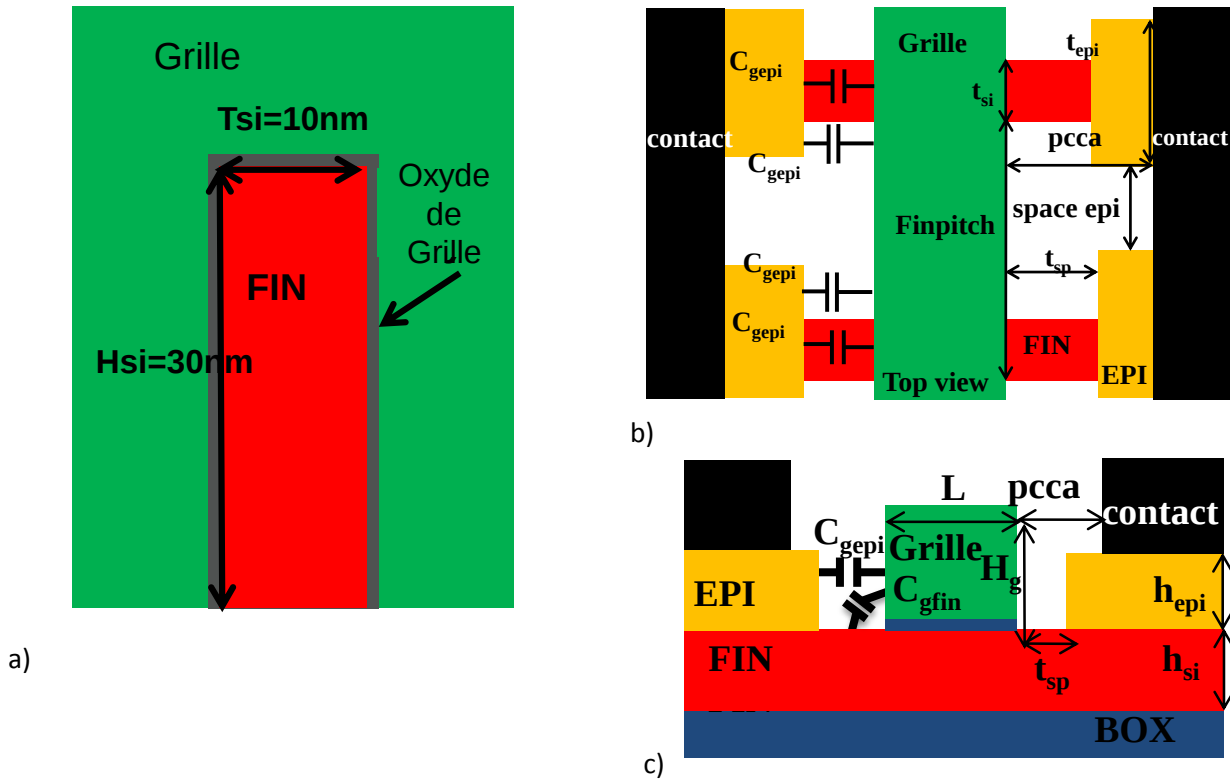


Figure V-15 :a) Schéma en vue en coupe perpendiculairement au fin, à travers la grille b) Schéma en vue en coupe dans le sens du fin de l'architecture Trigate avec ses dimensions caractéristiques. b) Schéma en vue de dessus de l'architecture Trigate avec ses dimensions caractéristiques.

Les dimensions de l'architecture Trigate sont résumées dans la Figure V-16.

	Trigate A		
type	NMOS	PMOS	Hypothèses
Vdd (V)	0.8	0.8	[Auth 12]
L (nm)	30	30	[Auth 12]
CPP (nm)	90	90	[Auth 12]
Finpitch (nm)	60	60	[Auth 12]
h_{si} (nm)	30	30	[Auth 12]
t_{si} (nm)	10	10	[Auth 12]
EOT (nm)	0.9	0.9	[Auth 12]
h_{epi} (nm)	20	20	$2xt_{si}$
t_{epi} (nm)	40	40	$2xh_{epi}$
Space epi (nm)	20	20	Finpitch- t_{epi}
t_{sp} (nm)	7	7	Estimée sur [Auth 12]
pcca (nm)	20	20	Estimée sur [Auth 12]
H_g (nm)	60	60	Estimée sur [Auth 12]
DIBL (mV)	45	45	[Auth 12]
SS (mV/dec)	72	72	[Auth 12]
I_{on} ($\mu A/\mu m$)	1190	1044	[Auth 12]
I_{eff} ($\mu A/\mu m$)	688	584	[Auth 12]
I_{off} (nA/ μm)	60	60	[Auth 12]

Figure V-16 : Résumé des paramètres technologiques considérés pour la modélisation de l'architecture Trigate-A.

Avant de générer les caractéristiques courant-tension de cette architecture à partir de MASTAR VA, il faut savoir comment sont normalisées les valeurs de courant visées (données dans [Auth 12]). Pour les dispositifs multi-grilles, il est usuel dans la littérature de normaliser les courants par le périmètre (noté W_{elec}) qui vaut dans notre cas pour un fin $t_{si}+2xh_{si}=70nm$. Or par une analyse interne à STMicroelectronics [ST Rapport Interne] il a été démontré que les courants donnés dans la publication de [Auth 12] sont normalisés par la largeur de l'empreinte physique sur la puce (footprint en anglais, noté $W_{Footprint}$). Pour un seul fin cela correspond au finpitch, donc à 60nm.

Dans le cas d'une évaluation de performance circuit, ceci est extrêmement important. Pour le démontrer, évaluons le courant de saturation (I_{on}) absolu débité par un seul fin (I_{on1fin} en μA) dans le cas du NMOS ($I_{on}=1190\mu A/\mu m$), pour chaque méthode de normalisation de la valeur de courant de saturation ciblée :

- Normalisation par la largeur électrique :

$$I_{on1fin} = I_{on} \cdot (t_{si} + 2h_{si}) = 1190 \cdot (10 \cdot 10^{-3} + 2 \cdot 30 \cdot 10^{-3}) = 83.3\mu A$$

- Normalisation par l'empreinte physique :

$$I_{on1fin} = I_{on} \cdot (Finpitch) = 1190 \cdot (60 \cdot 10^{-3}) = 71.4\mu A$$

En faisant le ratio des deux valeurs absolues de courant de saturation obtenues pour un fin, on constate que l'écart est de 17%. Donc, sachant par [ST rapport interne] que les courants sont normalisés par l'empreinte physique (i.e par le Finpitch) dans [Auth 12], nous aurions surestimé les performances circuit de l'architecture d'environ 17% si nous avons considéré que les courants visés étaient normalisés par la largeur électrique.

On peut alors tracer les caractéristiques courant-tension de cette architecture (Trigate-A) en considérant que les courants sont normalisés par l'empreinte physique (donc par $W_{Footprint}=N_{fin} \cdot Finpitch$ avec N_{fin} nombre de fin du

dispositif). Ces courbes sont représentées sur la Figure V-17. Le graphe au bas de cette figure montre que MASTAR VA reproduit bien les I_d - V_g reportées dans [Auth 12].

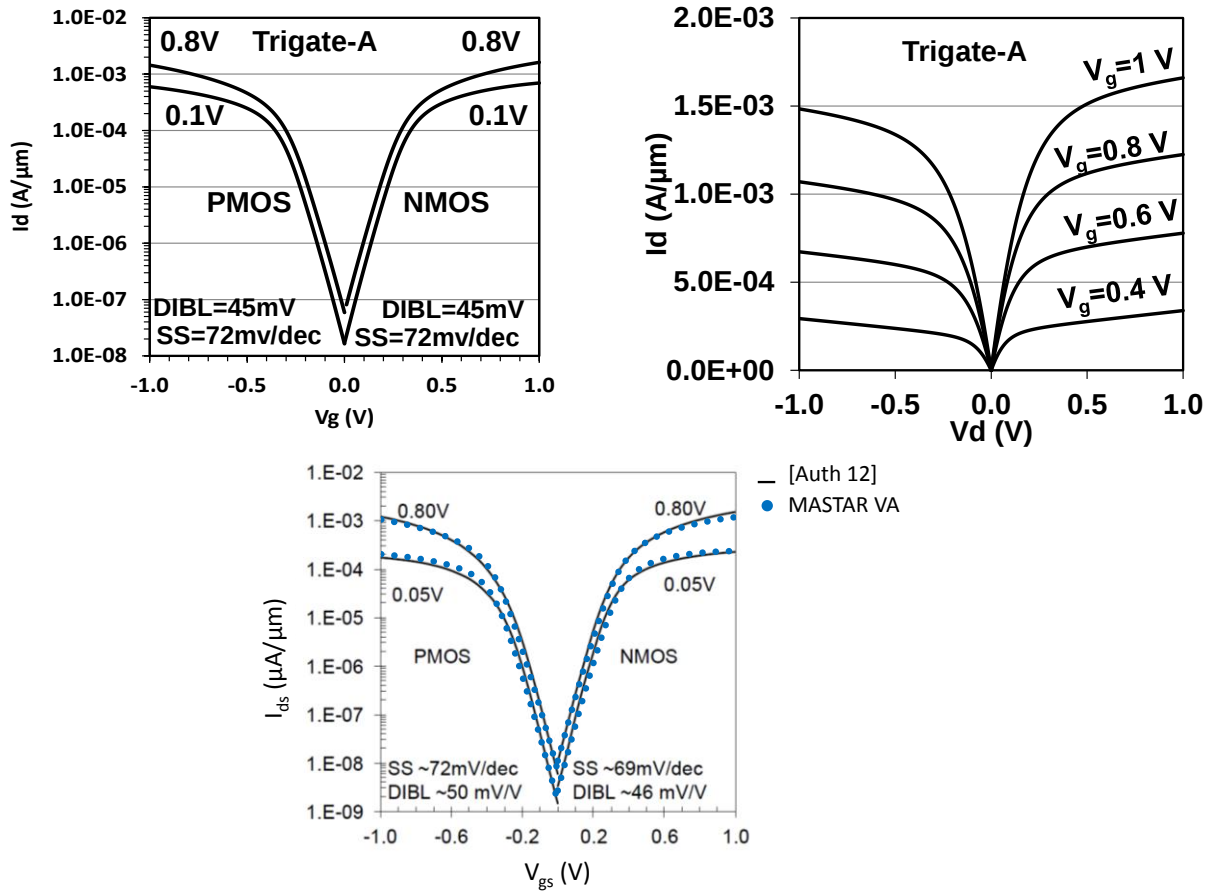


Figure V-17 : Caractéristiques courant-tension de l'architecture Trigate-A obtenues par notre MASTAR VA. Le graphe du bas montre que MASTAR VA reproduit bien les I_d - V_g reportées dans [Auth 12].

Enfin, il faut remarquer la longueur de grille importante que nous avons choisie et qui est la même que [Auth 12]. Cette longueur de grille relâchée pour le nœud 20 nm s'explique par deux points :

- La grille est construite par une intégration de type « gate last ». Ceci permet d'induire de la contrainte dans le canal (donc d'améliorer le transport) et le courant de saturation mais il est plus difficile d'obtenir des longueurs de grilles très agressives.
- La morphologie de cette architecture due à l'utilisation de plusieurs fins pour un même dispositif rend également les longueurs de grille agressives difficiles à obtenir.

L'utilisation d'une grille relâchée permet un très bon contrôle des effets canaux courts (DIBL et pente sous le seuil très faibles). Cependant, la capacité de grille intrinsèque sera plus élevée (car directement proportionnelle à la longueur de grille) tout comme les capacités parasites car la distance grille-contact sera réduite. Cette augmentation importante de capacité aura un effet important sur la vitesse d'un circuit, c'est la raison pour laquelle nous étudierons une autre architecture de Trigate (Trigate-B) sur laquelle nous allons réduire la longueur de grille. Pour déterminer cette longueur, nous allons chercher la valeur qui permet d'avoir un DIBL comparable à celui du FDSOI, soit environ 100mV/V. De cette manière, on obtient $L_g=24\text{nm}$. En utilisant la prédictivité de

MASTAR VA pour l'électrostatique, en conservant le niveau de contrainte déterminé pour le Trigate-A et en incluant la dégradation de transport due à la réduction de la longueur de grille (type $\mu(L)$ [Bidal 09]), on obtient les performances statiques de cette seconde architecture Trigate (Trigate-B). Ces dernières, ainsi que les dimensions de cette architecture, sont renseignées sur la Figure V-18.

	Trigate B		
type	N	P	Hypothèses
Vdd (V)	0.8	0.8	Estimée sur [Auth 12]
L (nm)	24	24	Pour obtenir le même DIBL que le FDSOI [Khakifirooz 12]
CPP (nm)	90	90	[Auth 12]
Finpitch (nm)	60	60	[Auth 12]
h_{si} (nm)	30	30	[Auth 12]
t_{si} (nm)	10	10	[Auth 12]
EOT (nm)	0.9	0.9	[Auth 12]
h_{epi} (nm)	20	20	$2 \times t_{si}$
t_{epi} (nm)	40	40	$2 \times h_{epi}$
Space epi (nm)	20	20	$Finpitch - t_{epi}$
t_{sp} (nm)	12	12	$pcca/2$
pcca (nm)	25	25	$(CPP - L)/3$
H_g (nm)	60	60	Estimé sur [Auth 12]
DIBL (mV)	80	80	Projection du modèle
SS (mV/dec)	83	84	Projection du modèle
I_{on} ($\mu A/\mu m$)	1204	1067	Dégradation type $\mu(L)$ incluse
I_{eff} ($\mu A/\mu m$)	655	572	Dégradation type $\mu(L)$ incluse
I_{off} (nA/ μm)	60	60	Ajusté avec ϕ_{ms}

Figure V-18 : Résumé des paramètres technologiques considérés pour la modélisation de l'architecture Trigate-B.

En entrant ces paramètres dans MASTAR VA, on trace les caractéristiques courant-tension de cette architecture (Trigate-B), toujours en considérant que les courants sont normalisés par l'empreinte physique (donc par $W_{Footprint} = N_{fin} \cdot Finpitch$ avec N_{fin} nombre de fin du dispositif). Ces courbes sont représentées sur la Figure V-19.

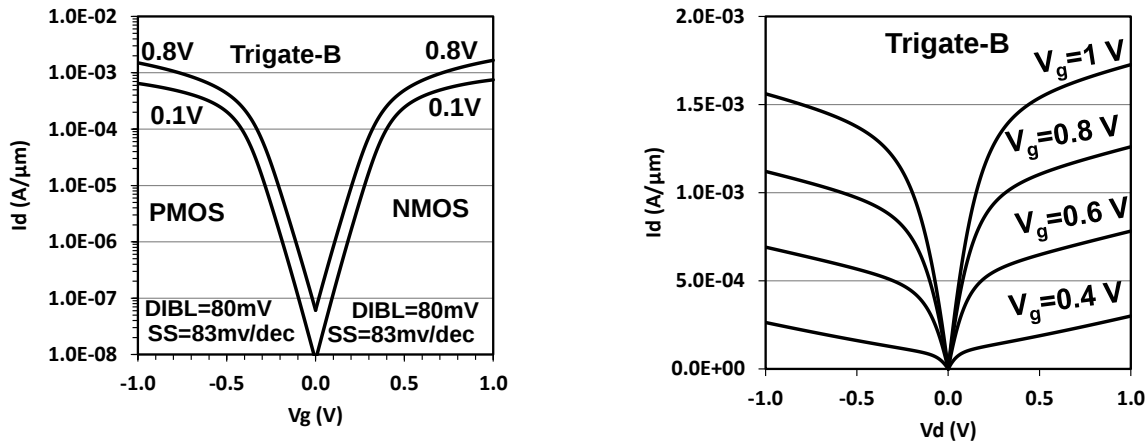


Figure V-19 : Caractéristiques courant-tension de l'architecture Trigate-A obtenues par MASTAR VA.

V.B.3. Comparaison des performances statiques

Nous pouvons commencer à comparer les architectures du point de vue performance statique. Sur la Figure V-20, nous avons représenté les compromis I_{on}/I_{off} et I_{eff}/I_{off} pour les quatre architectures décrites au paragraphe V.B.2 pour une même tension d'alimentation (V_{dd}), fixée à 0.8V pour être alignée sur la publication de [Auth 12]. On constate que le courant de saturation I_{on} du NMOS FDSOI est proche de celui des deux Trigate, contrairement au PMOS. En ce qui concerne le courant effectif I_{eff} le Trigate-A est bien meilleur grâce à son excellent DIBL alors que le Trigate-B est pénalisé par sa valeur plus forte de DIBL, mais reste tout de même meilleur que le FDSOI. Bien que le PMOS de l'architecture conventionnelle sur substrat massif (BULK) soit compétitif, son NMOS l'est trop peu pour envisager de bonnes performances dynamiques.

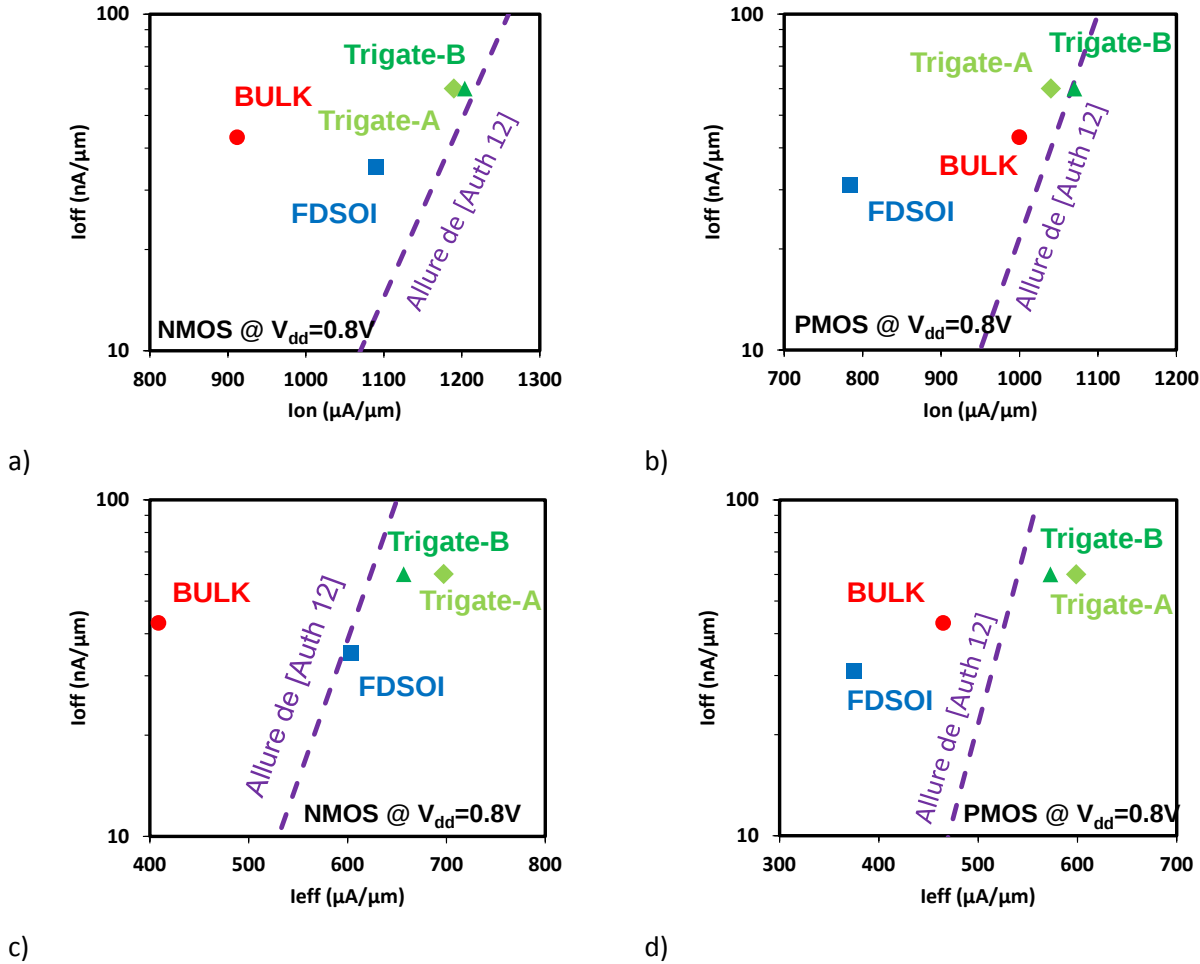


Figure V-20 : Compromis I_{on}/I_{off} (a)NMOS et b) PMOS)et I_{eff}/I_{off} (c)NMOS et d) PMOS)

V.B.4. Robustesse à la variabilité

Pour terminer l'analyse statique, nous allons comparer la robustesse des trois architectures à la variabilité des paramètres technologiques dues au procédé de fabrication en comparant les distributions de tension de seuil obtenues. Il faut commencer par définir de manière réaliste les distributions statistiques de variations aléatoires des paramètres technologiques. Pour ne pas avoir une étude trop complexe, nous définirons des variations uniquement pour les paramètres technologiques clés, à savoir la longueur de grille (L_g), l'épaisseur d'oxyde de grille (t_{ox}) l'épaisseur du film de silicium d'un transistor FDSOI ou d'un fin (t_{si}), la hauteur d'un fin uniquement dans le cas du Trigate (h_{si}), la largeur du dispositif uniquement dans le cas planaire (W), le travail de sortie de la

grille (ϕ_{ms}) et le dopage canal uniquement dans le cas de transistor sur substrat massif (N_{ch}). Nous considérerons chaque distribution Gaussienne centrée autour de la valeur visée et d'écart type:

- Pour les dimensions de paramètres technologiques définies par gravure, nous considérerons que l'écart type σ respecte la loi [Bœuf 08] :

$$3\sigma = 12\%CD$$

Où CD (Critical Dimension en anglais) est la dimension visée.

Ceci s'applique pour :

- La longueur de grille L_g
- L'épaisseur de fin t_{si}
- La hauteur de fin h_{si}
- La largeur des dispositifs W
- Pour l'épaisseur de film de silicium dans le cas de technologie FDSOI, nous utiliserons la valeur reportée par [Weber 08] qui vaut $\sigma_{tsi}=0.2nm$.
- Pour l'épaisseur d'oxyde de grille (EOT), nous choisissons un écart type de $\sigma_{EOT}=0.1nm$.
- Pour le dopage canal (N_{ch}), uniquement dans le cas de l'architecture conventionnelle sur substrat massif, nous considérerons un écart type $\sigma_{NCH}=2^{e17}cm^{-3}$. Cette valeur nous permet d'obtenir le même écart type de la distribution de tension de seuil que celle obtenue par simulations numériques [GSS-b].
- Pour le travail de sortie de grille (ϕ_m), nous considérerons un écart type de $\sigma_{\phi m}=20mV$ pour chaque architecture, ce qui nous permet d'obtenir les mêmes distributions statistiques de tension que celle obtenues par simulations numériques pour l'architecture conventionnelle sur substrat massif [GSS-b] et pour le Trigate [GSS-c].

Les valeurs des écarts types de chaque paramètre technologiques pour chaque architecture sont indiquées dans le tableau de la Figure V-21.

	Bulk	FDSOI	Trigate-A	Trigate-B
σ_L (nm)	1	1	1.2	1
σ_{tox} (nm)	0.1	0.1	0.1	0.1
σ_{tsi} (nm)	/	0.2	0.4	0.4
$\sigma_{h_{si}}$ (m)	/	/	1.2	1.2
σ_W (nm)	1	1	/	/
$\sigma_{\phi m}$ (mV)	20	20	20	20
$\sigma_{N_{ch}}$ (cm-3)	2e17	/	/	/
σ_{vt} (mV)	41	20	22	28
$\langle Vt \rangle$ (mV)	390	287	222	257

Figure V-21 : Valeur des écarts types des paramètres technologiques de chaque architecture.

En entrant ces paramètres de variabilité dans MASTAR VA, et en lançant pour chaque architecture 500 simulations avec génération aléatoire des paramètres technologiques, on extrait les distributions statistiques de tension de seuil, représentées sur la Figure V-22. Comme attendu, l'architecture conventionnelle sur substrat massif est la plus impactée par la variabilité, notamment à cause de son canal dopé. FDSOI et Trigate présentent quant à eux une robustesse à la variabilité comparable, bien que le Trigate-B soit pénalisé par sa longueur de grille plus agressive. Il faut cependant remarquer que nous avons considéré un Trigate fabriqué sur substrat SOI, donc avec un canal non dopé. Or, [Auth 12] utilise des Trigates fabriqués sur substrat massif conventionnel et ne pourra présenter des canaux non dopés à cause de la diffusion des dopants de la couche de dopage anti-perçage

dans le fin. Notre estimation de distribution de tension de seuil est donc optimiste pour l'architecture Trigate, car nous ne pouvons pas prendre en compte la variabilité du dopage canal.

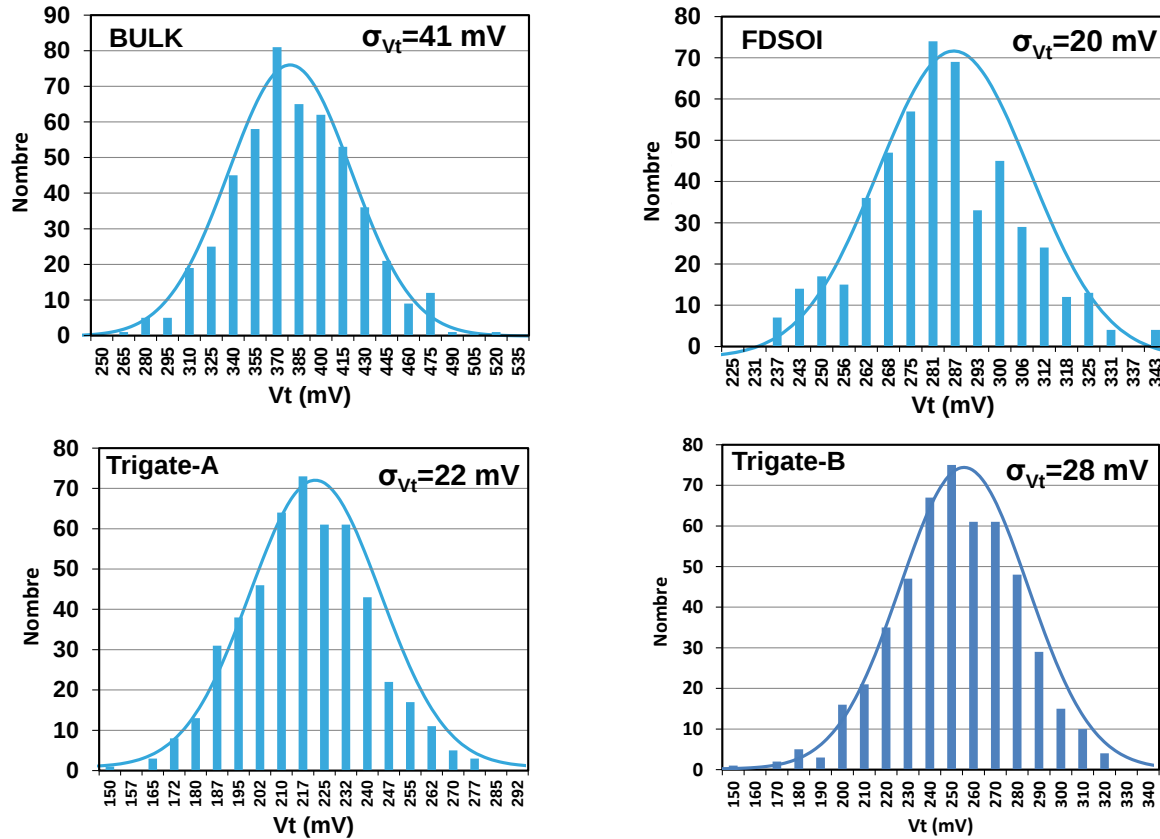


Figure V-22 : Distributions de tension de seuil pour chaque architecture obtenue avec MASTAR VA pour 500 tirages aléatoires des valeurs des paramètres technologiques, respectant les distributions données dans la Figure V-21.

V.B.5. Evaluation des capacités parasites

La dernière étape avant d'effectuer les simulations de circuit est d'estimer les capacités parasites de chaque architecture. Nous utilisons alors les modèles décrits dans le chapitre III en considérant l'intégration « gate last » pour le Trigate et l'architecture conventionnelle sur substrat massif (il y a donc une couche de diélectrique haute permittivité en plus de l'espaceur entre le métal de la grille et l'épitaxie) et les épitaxies présentant des facettes uniquement pour le FDSOI. Ces considérations sont en accord avec les dernières publications sur chacune des trois architectures étudiées [Shang 12], [Khakifirooz 12] et [Auth 12]. Nous représentons sur la Figure V-23 la somme de la capacité de grille intrinsèque en inversion (C_{inv}) et de la capacité parasite grille-drain (C_{gd}) pour chaque architecture, en distinguant chaque composante. Cette somme est normalisée par l'empreinte sur silicium, donc par la largeur du transistor pour les technologies planaires, et par $W_{FP} = N_{fin} \times FinPitch$ pour l'architecture Trigate. On constate que le Trigate-A est très pénalisé du point de vue capacitif par sa longueur de grille relâchée à 30nm. Cependant, même avec une longueur de grille plus agressive, cette architecture souffre de capacités parasites grille-drain plus importantes que les technologies planaires, comme le montre les valeurs de capacités du Trigate-B.

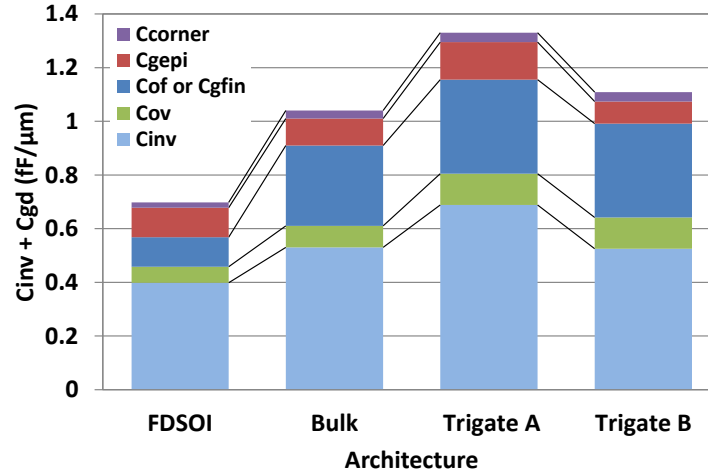


Figure V-23 : Somme de la capacité de grille intrinsèque en inversion (C_{inv}) et de la capacité parasite grille-drain (C_{gd}) pour chaque architecture, en distinguant chaque composante. Les capacités sont normalisées par la largeur de l’empreinte sur silicium.

V.B.6. Conditions de simulations :

Pour être représentatif du nœud 20nm, nous considérerons que les inverseurs seront constitués de NMOS et de PMOS de même largeur, à savoir 169nm. Pour les Trigates, nous devons déterminer le nombre de fin correspondant à cette largeur. Pour ne pas désavantager l’architecture Trigate, nous arrondirons à l’entier supérieur le ratio de la largeur de dispositif W visée (ici 169nm) et du finpitch, formellement :

$$N_{fin} = \text{ent} \left(\frac{W}{\text{Finpitch}} \right) + 1 \quad \text{Eq. V-13}$$

Nous travaillerons donc avec des Trigates composés de trois fin dans les inverseurs. Pour le chemin critique DDR3, qui est constitué de nombreux transistors de largeurs différentes, nous utiliserons l’équation V-13 pour déterminer le nombre de fin de chaque dispositif.

Dans les technologies planaires, la polarisation de la face arrière est couramment utilisée pour accélérer un circuit (polarisation $V_b > 0$ pour diminuer la tension de seuil) et on parle de FBB pour (Forward Back Bias) ou réduire sa consommation statique (polarisation $V_b < 0$ pour augmenter la tension de seuil) et on parle de RBB pour (Reverse Back Bias). Cependant cette technique est peu efficace sur l’architecture conventionnelle sur substrat massif car la polarisation applicable sur la face arrière est limitée à 300mV, pour ne pas placer la jonction PN entre les caissons et les source-drains en conduction. Nous étudierons donc l’effet de la polarisation de la face arrière uniquement pour l’architecture FDSOI, et uniquement dans le cas FBB, car nous cherchons à comparer ici les performances dynamiques. Pour l’architecture définie dans la Figure V-12, le décalage de tension de seuil obtenue par la polarisation de la face arrière sera de 70mV/V.

Enfin, pour l’architecture Trigate, il a été démontré [Lee 11] que la polarisation de la face arrière n’a pas d’effet sur la tension de seuil, et ce quel que soit le substrat utilisé (substrat massif (BULK) ou SOI). La

Figure V-24 présente les simulations effectuées par [Lee 11] qui démontre que la polarisation de la face arrière n’a pas d’effet sur la tension de seuil d’un Trigate.

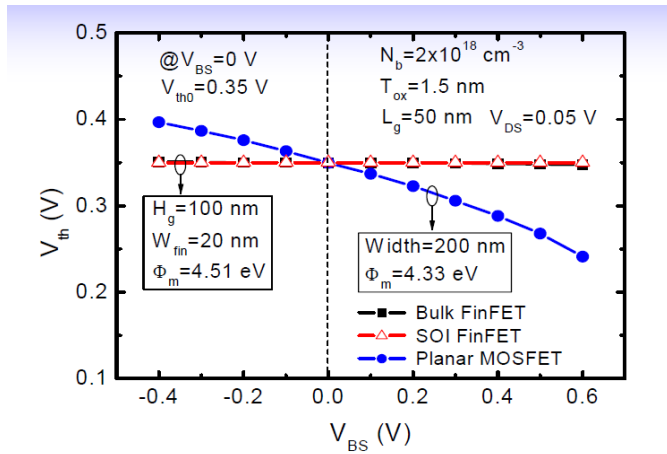


Figure V-24 : Tracé de la variation de la tension de seuil en fonction de la polarisation de la face arrière pour un Trigate sur substrat massif (noir), un Trigate sur SOI (rouge) et un transistor planaire (bleu) obtenu par simulations numériques dans [Lee 11]

Par conséquent, l'option d'amélioration de la vitesse d'un circuit par polarisation de la face arrière ne sera étudiée que pour l'architecture FDSOI, avec un décalage de tension de seuil de 70mV/V.

V.B.7. Ring d'inverseur FO1 avec charge additionnelle de sortie variable

Nous commençons par effectuer des simulations d'anneaux résonnants d'inverseur FanOut 1 (Figure V-4-e), avec une charge de sortie fixe, dont la valeur sera comprise entre 0 et 10fF. De ces simulations on extrait d'abord la fréquence (représentatif de la vitesse) en fonction de la tension d'alimentation et de la valeur de la charge en sortie.

Sur la Figure V-25-a représentant la variation de la fréquence en fonction de la tension d'alimentation pour l'anneau résonnant d'inverseurs FanOut1, on constate que l'architecture conventionnelle sur substrat massif est la moins compétitive. De plus, l'architecture FDSOI est la plus performante en termes de fréquence, et ce même sans utilisation du FBB. Ceci s'explique par la faible valeur de capacité du FDSOI et de son niveau de courant comparable aux deux architectures Trigate étudiées. Concernant l'architecture Trigate, on remarque que le Trigate-A est plus rapide que le Trigate-B pour les faibles valeurs de V_{dd} , ce qui s'explique par l'excellente électrostatique du Trigate-A. Cependant, si nous nous tournons vers les valeurs plus élevées de V_{dd} , le faible DIBL du Trigate-A devient pénalisant (car une augmentation de tension d'alimentation provoquera une faible baisse de la tension de seuil du Trigate-A) et le Trigate-B devient plus rapide, ici pour $V_{dd} = 0.65 \text{ V}$.

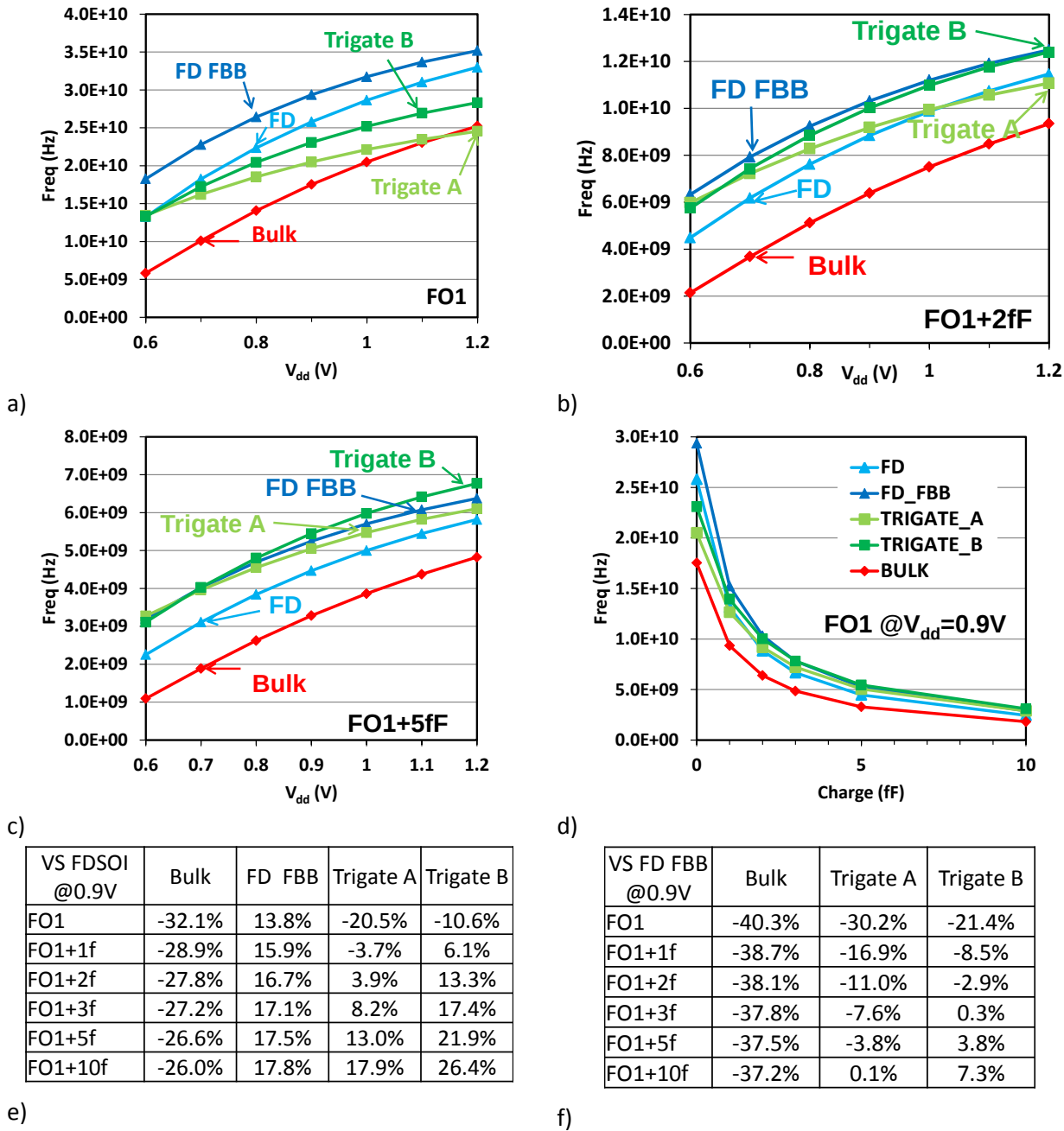


Figure V-25 : Tracé la variation de fréquence avec la tension d'alimentation V_{dd} des anneaux résonnants d'inverseur FanOut 1 pour trois valeurs de charge de sortie : 0fF (a), 2fF (b) et 5fF (c). d) Tracé de la variation de fréquence avec la valeur de la charge de sortie, à la tension d'alimentation nominale du nœud 20nm, soit $V_{dd}=0.9V$. Ecart en fréquence (%) entre chaque architecture, en comparaison du FDSOI (e) et du FDSOI avec FBB (f) en fonction de la charge de sortie, pour la tension d'alimentation nominale du nœud 20nm, soit $V_{dd}=0.9V$.

Mais, quand la charge de sortie augmente (2fF (b) et 5fF (c)) l'écart de fréquence entre FDSOI et Trigate se réduit et le Trigate peut même devenir plus rapide que le FDSOI. Ceci s'explique par la valeur importante de la charge fixe qui vient écranter l'effet des capacités du transistor (i.e capacités grille et parasites) sur la fréquence du circuit. Pour mettre en évidence ce phénomène d'écrantage des capacités du transistor par la valeur de la charge en sortie de chaque étage par un calcul analytique simple, nous utilisons l'équation (IV.A.2) pour estimer la

capacité totale de chaque inverseur puis nous traçons sur la Figure V-26 la variation de l'écart (%) entre les capacités totales des inverseurs à base de Trigate-A et B, et celle d'un inverseur FDSOI, en fonction de la valeur de la charge en sortie de l'inverseur. On constate bien que l'écart entre les capacités totales des inverseurs Trigate et FDSOI passe de $\sim 150\%$ (i.e $C_{tot_Trigate} \sim 2.5 \times C_{tot_FDSOI}$) pour une charge de sortie nulle, à $\sim 15\%$ pour une charge de 10fF.

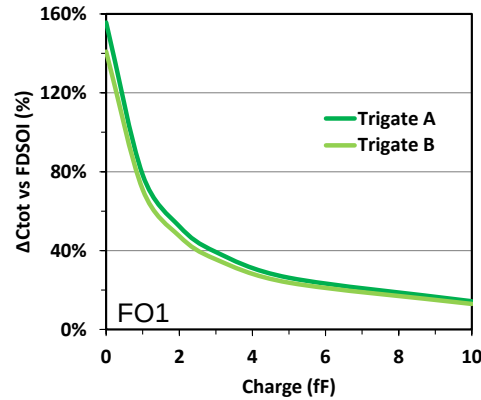


Figure V-26 : Variation de l'écart (%) entre les capacités totales des inverseurs à base de Trigate-A et B, et celle d'un inverseur FDSOI, en fonction de la valeur de la charge en sortie de l'inverseur obtenu par un calcul analytique simple (IV.A.2).

Par conséquent, pour les valeurs de charge de sortie élevée, c'est la quantité de courant débitée qui va le plus influencer sur la vitesse du circuit, ce qui justifie que l'architecture Trigate devient plus performante. Cependant l'utilisation du FBB permet à l'architecture FDSOI de rester comparable aux deux architectures Trigate, même pour des valeurs de charge de sortie élevée. Cela est confirmé par la Figure V-25-d qui montre la variation de fréquence avec la valeur de la charge de sortie, à la tension d'alimentation nominale du nœud 20nm, soit $V_{dd}=0.9V$, mais également par les tableaux de la Figure V-25-e et -f qui montrent les écarts de fréquence entre les architectures Trigate et le FDSOI (Figure V-25-e) puis avec le FDSOI avec l'utilisation du FBB (Figure V-25-f). Néanmoins, l'utilisation du FBB sur l'architecture FDSOI a un impact sur la fuite statique du circuit, comme le montre la Figure V-27. La fuite statique sera cinq fois supérieure pour le FDSOI avec utilisation du FBB que pour le FDSOI sans FBB, quelle que soit la tension d'alimentation (l'écart est constant car ils ont le même DIBL). Si on le compare au Trigate-B, le FDSOI avec FBB aura une fuite trois fois supérieure et ce quel que soit la tension d'alimentation, toujours parce qu'ils ont le même DIBL. Par contre, la comparaison de la fuite statique du FDSOI avec FBB avec le Trigate-A dépend de la tension d'alimentation utilisée. En effet, la fuite est 3 fois supérieure pour le FDSOI avec FBB à $V_{dd}=0.9V$ et 5 fois supérieure à $V_{dd}=1.2V$, ce qui s'explique par l'excellent DIBL du Trigate-A.

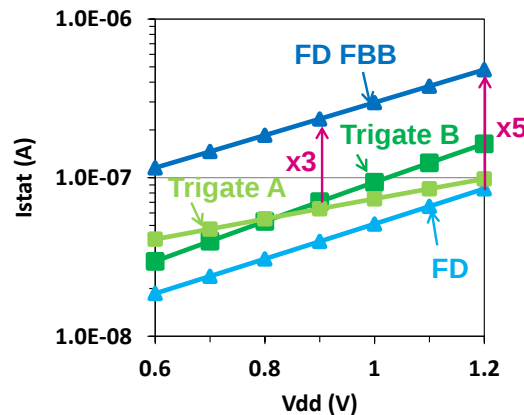
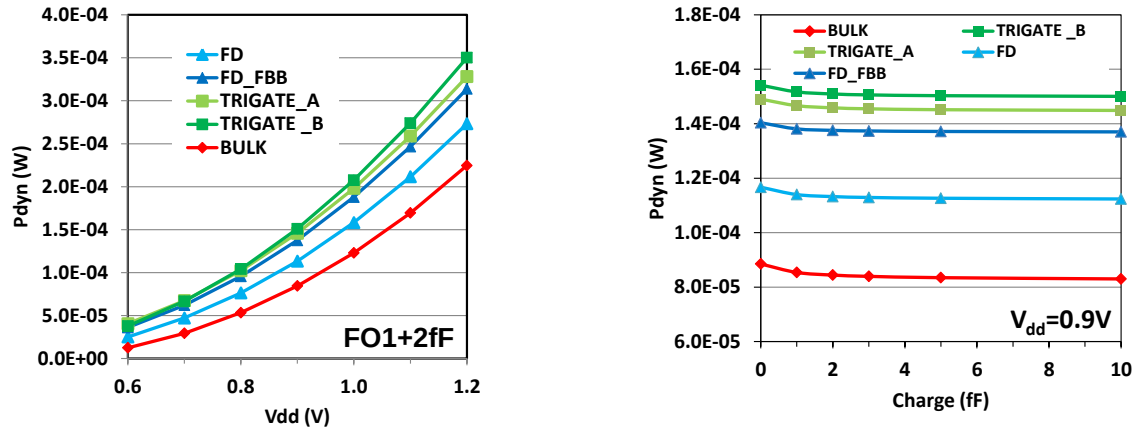


Figure V-27 : Variation de la fuite statique de l'anneau résonnant d'inverseur FanOut 1 avec V_{dd} .

Pour que l'analyse soit complète, il faut examiner la consommation dynamique à travers l'extraction du courant dynamique. La Figure V-28 montre que la puissance dynamique du Trigate est supérieure quelle que soit la tension d'alimentation (a) et la charge en sortie de chaque étage (b). Ceci est dû à la plus grande valeur de capacité des Trigate (A et B) comparée aux technologies planaires.



a) b)
Figure V-28 : a) tracé de la puissance dynamique (P_{dyn}) en fonction de la tension d'alimentation V_{dd} pour un anneau résonnant d'inverseur FanOut 1 avec une charge fixe en sortie de valeur 2fF. b) Tracé de la puissance dynamique (P_{dyn}) pour un anneau résonnant d'inverseur FanOut 1 en fonction de la valeur de la charge fixe en sortie à tension d'alimentation constante $V_{dd}=0.9V$.

On remarque aussi sur la Figure V-28 que l'architecture conventionnelle sur substrat massif semble être celle qui consomme le moins. Cependant, une bonne approximation de la puissance dynamique est donnée par la formule :

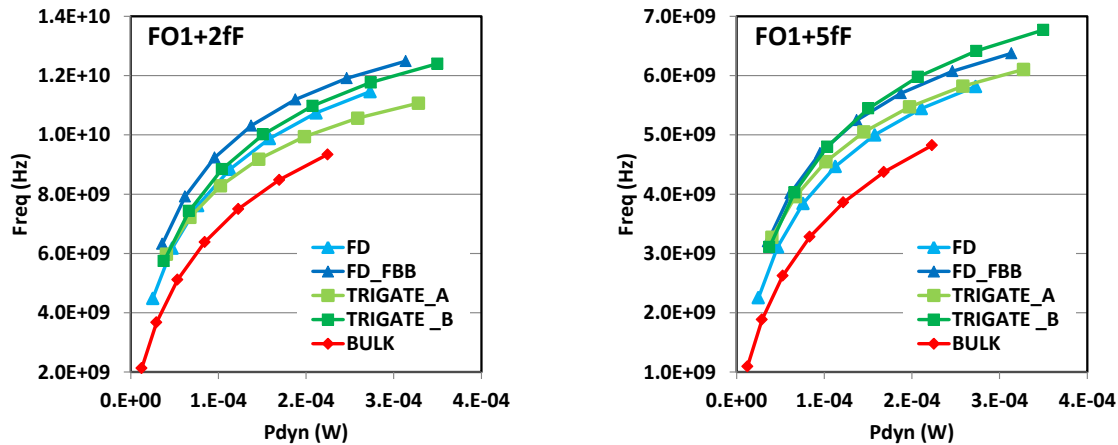
$$P_{dyn} = CV_{dd}^2f \quad \text{Eq. V-14}$$

Où C est la capacité équivalente du circuit, V_{dd} la tension d'alimentation et f la fréquence du circuit. Nous avons vu précédemment (Figure V-25) que l'architecture conventionnelle sur substrat massif n'est pas compétitive en terme de vitesse (i.e fréquence) et donc le terme « f » de l'équation V-14 est beaucoup plus faible pour cette architecture, ce qui explique sa puissance dynamique inférieure.

Il est donc difficile de comparer les puissances dynamiques sans prendre en compte les différences de fréquence. C'est la raison pour laquelle il est usuel de tracer la fréquence du circuit en fonction de la puissance dynamique. Ce type de graphique va traduire l'efficacité énergétique d'une architecture de transistor, c'est-à-dire :

- Quelle puissance dynamique est consommée pour atteindre une certaine fréquence.
- Quelle fréquence maximale est-il possible d'atteindre si la puissance dynamique maximum est limitée.

De cette manière, les architectures ne sont plus comparées à tension d'alimentation identique, mais à consommation dynamique identique, ce qui mène à une comparaison plus juste et plus réaliste.



a) b)
Figure V-29 : Tracé de la fréquence en fonction de la puissance dynamique pour un anneau résonnant d'inverseur FanOut 1 avec une charge en sortie de 2fF (a) et 5fF (b).

La Figure V-29-a montre que dans un anneau résonnant FanOut 1 avec une charge en sortie de 2fF, le FDSOI avec FBB permet d'atteindre la fréquence la plus haute quelle que soit la puissance dynamique et, de manière duale, présente la plus faible consommation dynamique à une fréquence donnée. On remarque également que le Trigate-B présente une meilleure efficacité énergétique que le Trigate-A, ce qui s'explique aisément par la longueur de grille plus agressive du Trigate-B (donc sa valeur de capacité plus faible). De plus, le Trigate-B présente pratiquement le même compromis fréquence/puissance dynamique que le FDSOI, même si l'égalité des fréquences pour les deux architectures (et donc des puissances dynamiques) est obtenue pour des tensions d'alimentation différentes ($V_{ddFD} \sim V_{ddTrigate-B} + 100\text{mV}$). Enfin, ce graphe confirme que l'architecture conventionnelle sur substrat massif (BULK) présente des performances en retrait.

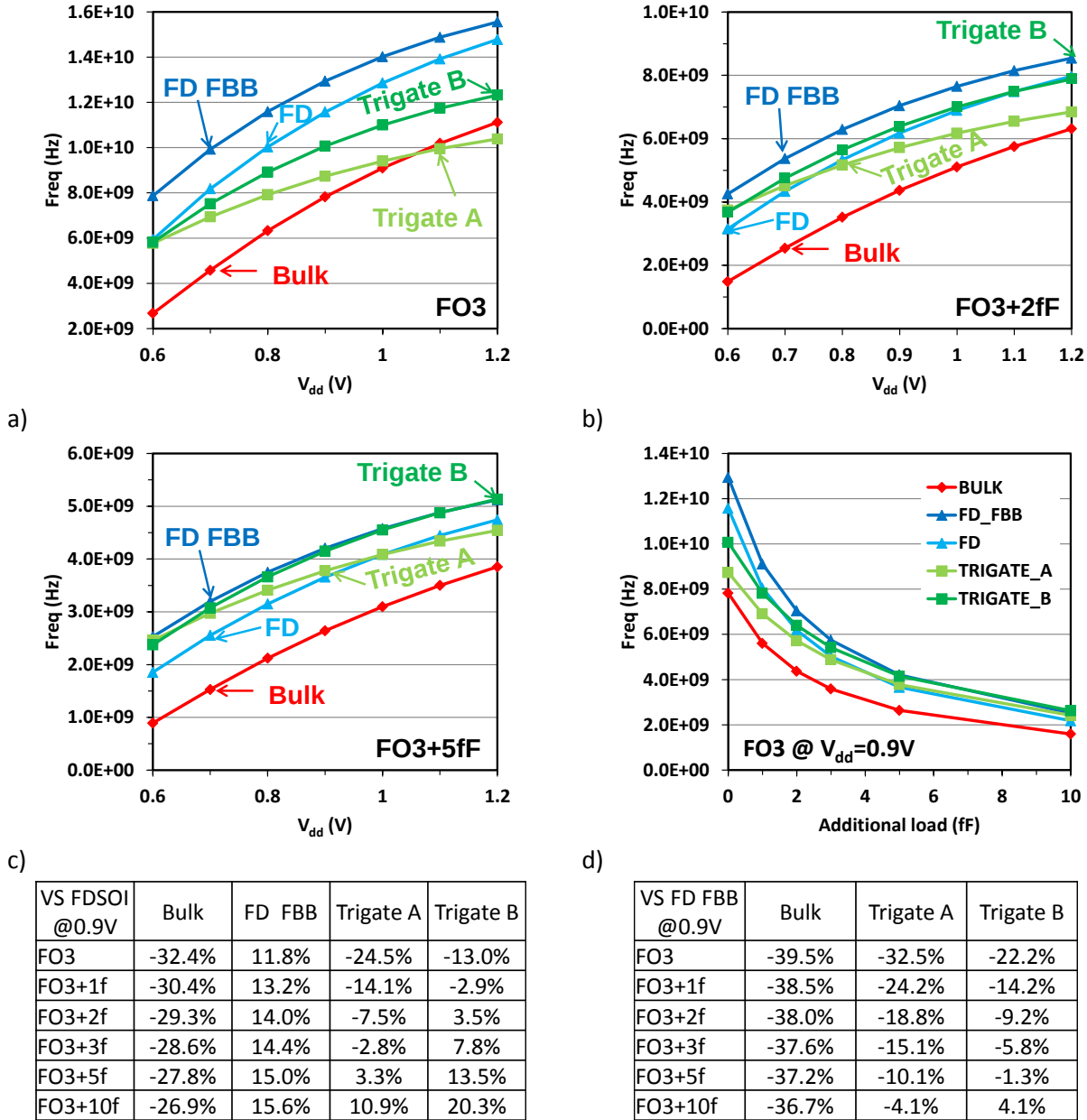
Comme pour la vitesse du circuit (Figure V-25), lorsque la valeur de la charge de sortie augmente, la situation devient favorable à l'architecture Trigate car l'impact des capacités grille et parasites est écranté, et la performance du circuit est alors davantage corrélée au niveau de courant débité par le transistor. Cette remarque explique la Figure V-29-b où l'on constate que le Trigate-B devient l'architecture la plus performante, même si il se situe sur le même compromis fréquence/puissance dynamique que le FDSOI avec FBB pour les puissances dynamiques inférieures à $2 \times 10^{-4}\text{W}$. Cependant, le FDSOI avec FBB présente toujours une meilleure efficacité énergétique que le Trigate-A, qui est sur un compromis semblable au FDSOI. Ceci s'explique toujours par la longueur de grille relâchée du Trigate-A qui est responsable d'une valeur de capacité très importante.

V.B.8. Ring d'inverseur FO3 avec charge additionnelle de sortie variable

L'anneau résonnant FanOut 1 présente l'avantage d'être rapide à simuler et de fournir des résultats assez simple à interpréter. Néanmoins, les performances de ce type de circuit ne sont pas représentatives d'un circuit réaliste. En effet, dans un « vrai » circuit, la sortie d'un inverseur (ou de toute autre cellule standard) en charge plusieurs autres, et pas une seule comme dans un anneau résonnant d'inverseur FanOut 1. Il est donc plus adapté d'analyser les performances circuits d'un anneau résonnant d'inverseur FanOut 3 (FO3), toujours avec une charge fixe en sortie de chaque étage. Dans ce type de circuit, chaque étage « verra » trois inverseurs. La capacité équivalente de chaque étage sera donc plus grande, et directement proportionnelle aux valeurs de capacités au niveau du transistor (i.e grille et parasite). La performance de ce type de circuit sera donc davantage corrélée à l'architecture du dispositif, et à ces capacités, ce qui est plus représentatif de la véritable performance circuit d'une architecture de transistor.

Comme au cours du paragraphe précédent (V.B.7), nous commençons par étudier le comportement en fréquence des anneaux résonnants en fonction de la tension d'alimentation V_{dd} et de la charge de sortie de chaque étage (Figure V-30). Comme précédemment (Figure V-30-a), lorsque la charge de sortie est nulle l'architecture Trigate

est pénalisée par ses capacités et le FDSOI est plus rapide (même sans FBB) quelle que soit la tension d'alimentation. Il faut tout de même mentionner que dans cette configuration, l'écart en fréquence entre les architectures Trigate et FDSOI est plus important que pour les inverseurs FanOut 1. Ceci s'explique par le poids plus important des capacités des transistors dans la vitesse d'un inverseur FanOut 3 que pour un inverseur FanOut 1.



e) Figure V-30 : Tracé la variation de fréquence avec la tension d'alimentation V_{dd} des anneaux résonnants d'inverseur FanOut 3 pour trois valeurs de charge de sortie : 0fF (a), 2fF (b) et 5fF (c). (d) Tracé de la variation de fréquence avec la valeur de la charge de sortie, à la tension d'alimentation nominale du nœud 20nm, soit $V_{dd}=0.9V$. Ecart en fréquence (%) entre chaque architecture, en comparaison du FDSOI (e) et du FDSOI avec FBB (f) en fonction de la charge de sortie, pour la tension d'alimentation nominale du nœud 20nm, soit $V_{dd}=0.9V$.

Lorsque la charge de sortie augmente (2fF Figure V-30-b et 5fF Figure V-30-c), l'architecture Trigate devient plus performante mais, contrairement au cas FanOut 1, reste moins rapide que l'architecture FDSOI avec utilisation du FBB. Le faible DIBL du Trigate-A lui permet toujours d'être plus rapide que le Trigate-B pour les faibles tensions d'alimentation. Cependant, sa grille de longueur 30nm le pénalise du point de vue capacitif et son faible DIBL n'induit pas une chute de tension de seuil assez importante quand la tension d'alimentation augmente pour conserver cet avantage, et le Trigate-B est plus rapide pour les tensions d'alimentation supérieures à 0.6V.

La Figure V-30-d représente la variation de fréquence de chaque architecture (à $V_{dd}=0.9V$) en fonction de la valeur de la charge en sortie et les Figure V-30-e et f représentent l'écart (en %) de fréquence entre chaque architecture et le FDSOI (e) et le FDSOI avec FBB (f). On constate que, à cette tension d'alimentation ($V_{dd}=0.9V$, nominale du nœud technologique 20nm), le FDSOI avec FBB est l'architecture la plus rapide, sauf si la charge de sortie vaut 10fF où il s'agit du Trigate-B. Cela s'explique par la pénalité capacitive de l'architecture Trigate, qui a plus de poids sur la fréquence quand la valeur du FanOut augmente. Cette remarque explique également pourquoi le Trigate-A n'est pas compétitif, hormis pour les tensions d'alimentation faibles (<0.6V) où sa faible valeur de DIBL constitue un avantage décisif. Enfin, chaque graphe de la Figure V-30 confirme que l'architecture conventionnelle sur substrat massif (BULK) n'est pas compétitive pour ce nœud technologique. Comme dans le paragraphe précédent, nous devons analyser l'impact de l'utilisation des techniques de polarisation de la face arrière pour améliorer la performance du circuit sur la fuite statique du circuit. Les remarques sont les mêmes que pour l'anneau résonnant d'inverseur FO1, car la fuite statique est directement proportionnelle à la valeur du FanOut et ce, quelle que soit l'architecture des transistors. Par conséquent, la fuite statique de l'inverseur FO3 est trois fois supérieure à celle de l'inverseur FO1 (Figure V-27), simplement parce qu'il y a trois fois plus de transistors qui fuient au repos. Cependant, dans un circuit, le FBB n'est pas utilisé en permanence mais uniquement lorsque le système a besoin de performance. Par conséquent, l'augmentation de fuite statique due au FBB n'est pas aussi forte que celle représentée sur la Figure V-31 qui considère que le FBB est utilisé en permanence. L'augmentation réelle est donc très dépendante du circuit considérée, et le graphique ci-dessous indique en réalité la valeur maximale d'augmentation de fuite statique due au FBB pour l'architecture FDSOI.

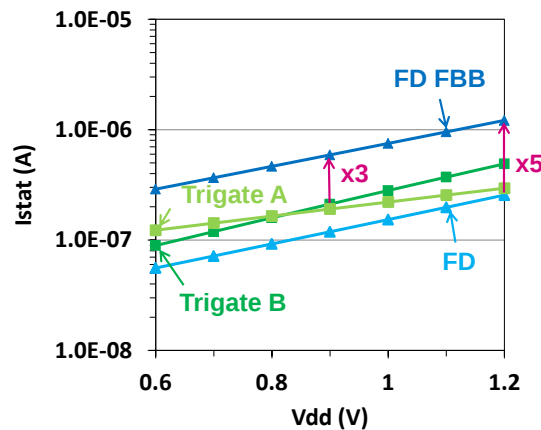


Figure V-31 : Variation de la fuite statique de l'anneau résonnant d'inverseur FanOut 3 avec V_{dd} .

Examinons maintenant la consommation dynamique de chaque architecture. La Figure V-32-a montre la variation de puissance dynamique en fonction de la tension d'alimentation pour chaque architecture dans le cas d'un anneau résonnant d'inverseurs FO3 avec une charge de sortie de 2fF à chaque étage. On remarque à nouveau que l'architecture Trigate est celle qui consomme le plus d'énergie en régime dynamique alors que l'architecture conventionnelle sur substrat massif est celle qui consomme le moins, du fait de sa fréquence bien plus faible que les autres architectures. Dans ce type d'anneau résonnant, les capacités des transistors ont plus d'impact et on remarque que les deux Trigate présentent à peu près les mêmes valeurs de puissance dynamique, et que le FDSOI avec FBB présente une consommation dynamique plus faible que les Trigate A et B. Ces deux constatations

n'étaient pas valables pour les inverseurs FO1 (Figure V-29). Enfin, le gain en fréquence compris entre 10 et 15% obtenu avec l'utilisation du FBB sur l'architecture FDSOI se paie par une augmentation de puissance dynamique d'environ 30% à $V_{dd}=0.9V$. La Figure V-32-b présente la variation de puissance dynamique avec la valeur de la charge en sortie de chaque étage pour chaque architecture à la tension d'alimentation nominale du nœud technologique 20nm ($V_{dd}=0.9V$). Ce graphe démontre à nouveau que l'architecture Trigate est la plus consommatrice d'énergie en régime dynamique, et ce, quelle que soit la charge en sortie de chaque étage.

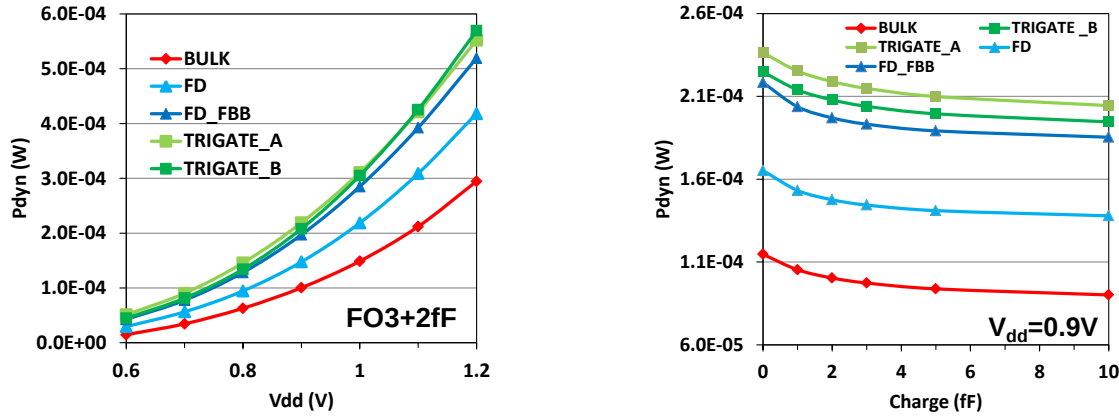


Figure V-32 : a) tracé de la puissance dynamique (P_{dyn}) en fonction de la tension d'alimentation V_{dd} pour un anneau résonnant d'inverseur FanOut 1 avec une charge fixe en sortie de valeur 2fF. b) Tracé de la puissance dynamique (P_{dyn}) pour un anneau résonnant d'inverseur FanOut 1 en fonction de la valeur de la charge fixe en sortie à tension d'alimentation constante $V_{dd}=0.9V$.

Pour une analyse plus fine, il faut lier fréquence et puissance dynamique. La Figure V-33 représente la fréquence en fonction de la puissance dynamique pour chaque architecture pour des anneaux résonnants FO3 avec une charge de sortie de 2fF (a) et 5fF (b). Ce type de graphique va traduire l'efficacité énergétique de chaque architecture. Dans les deux cas, le FDSOI avec FBB est l'architecture la plus efficace énergétiquement, même pour une charge de sortie de 5fF. Cependant, quand la charge de sortie augmente (Figure V-33-b) l'écart entre FDSOI avec FBB et les deux architectures Trigate se réduit et le Trigate-B permet d'atteindre des niveaux de fréquence plus élevés que le FDSOI pour une même consommation dynamique. Cependant le Trigate-A reste moins performant que le FDSOI, car la pénalité capacitive due à sa longueur de grille relâchée est trop importante, et n'est toujours pas suffisamment atténuée par la charge de sortie de l'inverseur.

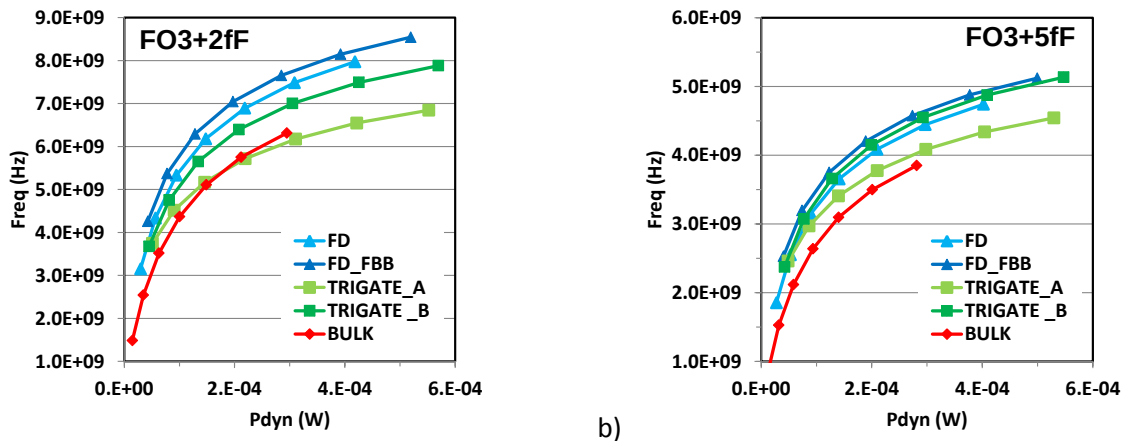


Figure V-33: Tracé de la fréquence en fonction de la puissance dynamique pour un anneau résonnant d'inverseur FanOut 3 avec une charge en sortie de 2fF (a) et 5fF (b).

Si on se focalise sur la comparaison entre les architectures FDSOI et Trigate, les graphes de la Figure V-33 peuvent se lire différemment. En nous focalisant sur l'anneau résonnant FanOut 3 avec une charge de sortie de 2fF, la Figure V-30 nous indique que, à $V_{dd}=0.9V$, le FDSOI avec FBB est l'architecture la plus rapide, et que le FDSOI est plus rapide que le Trigate-A, mais moins que le Trigate-B. Les écarts sont rappelés dans la Figure V-34. Mais, pour un concepteur de circuit, la tension d'alimentation est une variable, et c'est plutôt la consommation totale du circuit qui va limiter la performance. En effet, ce dernier peut ajuster la tension d'alimentation pour que la consommation soit à sa limite maximum, et, de la sorte, à ce que la fréquence soit maximale.

La puissance dynamique est liée à la capacité équivalente (Eq.V-14). Par conséquent, si on fixe la valeur de la puissance dynamique, cette dernière est atteinte pour une valeur de tension d'alimentation plus faible dans le cas de l'architecture Trigate. Cela explique pourquoi les écarts en vitesse entre les architectures FDSOI et Trigate vus à tension d'alimentation constante (Figure V-34-a) sont exacerbés par la comparaison à puissance dynamique constante (Figure V-34-b). On constate même par cette comparaison que l'architecture FDSOI est plus performante que les deux architectures Trigate, car elle présente un meilleur compromis puissance dynamique/fréquence (la même puissance dynamique est atteinte pour une tension d'alimentation 100mV plus faible pour le FDSOI que pour le Trigate).

vs FD	0.6	0.7	0.8	0.9	1	1.1	1.2
FD_FBB	35.2%	24.0%	18.0%	14.0%	11.0%	8.8%	7.2%
Trigate A	18.9%	4.2%	-3.1%	-7.5%	-10.4%	-12.5%	-14.1%
Trigate B	16.7%	9.7%	6.0%	3.5%	1.6%	0.1%	-1.1%

vs FD_FBB	0.6	0.7	0.8	0.9	1	1.1	1.2
FD	-26.0%	-19.3%	-15.2%	-12.3%	-9.9%	-8.1%	-6.7%
Trigate A	-12.0%	-16.0%	-17.8%	-18.8%	-19.3%	-19.6%	-19.9%
Trigate B	-13.6%	-11.5%	-10.2%	-9.2%	-8.5%	-8.0%	-7.7%

a)

vs FD	0.1mW		0.2mW		0.3mW		0.4mW	
	Freq	Vdd	Freq	Vdd	Freq	Vdd	Freq	Vdd
FD_FBB	13.2%	0.79	7.1%	0.93	5.7%	1.04	4.2%	1.13
Trigate A	-11.8%	0.74	-15.7%	0.89	-15.9%	1	-16.9%	1.1
Trigate B	-5.0%	0.75	-5.7%	0.9	-6.2%	1	-6.3%	1.09

vs FD_FBB	0.1mW		0.2mW		0.3mW		0.4mW	
	Freq	Vdd	Freq	Vdd	Freq	Vdd	Freq	Vdd
FD	-11.7%	0.81	-6.6%	0.97	-5.4%	1.08	-4.1%	1.18
Trigate A	-22.1%	0.74	-21.2%	0.89	-21.0%	1	-20.3%	1.1
Trigate B	-16.1%	0.75	-12.0%	0.9	-11.2%	1	-10.1%	1.09

b)

Figure V-34: Ecart (%) en fréquence entre chaque architecture pour un anneau résonnant d'inverseurs FanOut 3 avec une charge de sortie de 2fF à même tension d'alimentation (a) et à même puissance dynamique (b) en prenant pour référence le FDSOI puis le FDSOI avec FBB.

Pour terminer l'analyse, en prenant en compte vitesse et efficacité énergétique, nous avons représenté les écarts de fréquence entre l'architecture FDSOI et le Trigate-A (Figure V-35-a), puis l'écart du ratio P_{dyn}/f entre ces deux même architectures (Figure V-35-b) en fonction de la tension d'alimentation et de la valeur de la charge en sortie de chaque étage. Les mêmes graphes sont représentés pour comparer le Trigate-A et le FDSOI avec utilisation du FBB (Figure V-35-c et d). La Figure V-36 représente la comparaison entre le Trigate-B et le FDSOI avec et sans l'utilisation de FBB par le même type de graphe.

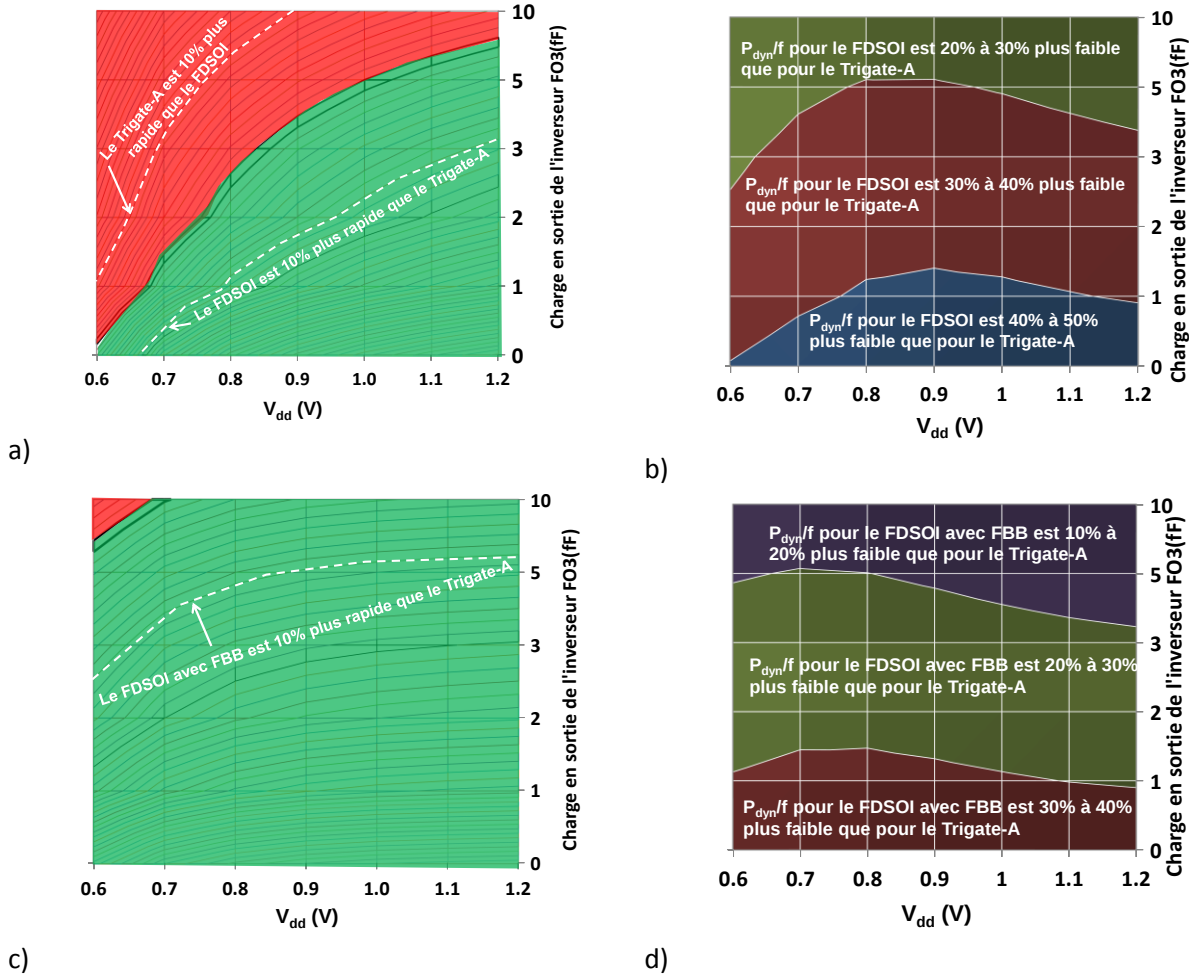


Figure V-35 : Cartographie de la comparaison de fréquence d'anneaux résonnants d'inverseur FanOut 3 en fonction de la tension d'alimentation et de la charge en sortie de chaque étage (chaque bande correspond à 1%), entre le FDSOI et le Trigate-A (a) et le FDSOI avec FBB et le Trigate-A (c). Cartographie de la comparaison d'efficacité énergétique (P_{dyn}/f) d'anneaux résonnants d'inverseur FanOut 3 en fonction de la tension d'alimentation et de la charge en sortie de chaque étage, entre le FDSOI et le Trigate-A (b) et le FDSOI avec FBB et le Trigate-A (d).

La Figure V-35 montre que le Trigate-A est plus rapide que l'architecture FDSOI, quand la tension d'alimentation est faible (grâce au DIBL faible) et quand la capacité en sortie de chaque étage augmente (ce qui écrante la pénalité capacitive du Trigate-A, cf Figure V-26). Cependant, quelles que soient les conditions de simulation, l'architecture FDSOI présente une efficacité énergétique toujours meilleure que celle du Trigate-A : de 20% quand le couple V_{dd} -charge de sortie est favorable à l'architecture Trigate, et jusqu'à 50% quand les conditions sont favorables au FDSOI (V_{dd} fort, capacité en sortie faible).

Avec l'utilisation du FBB, l'architecture FDSOI est plus rapide que l'architecture Trigate-A, sauf pour une tension d'alimentation égale à 0.6V et une charge de sortie supérieure à 9fF. Néanmoins, ce gain en vitesse obtenu par l'utilisation du FBB se paie en consommation dynamique et donc en efficacité énergétique. Au final, l'efficacité énergétique du FDSOI avec FBB est toujours meilleure que celle du Trigate-A, mais l'écart est réduit à un intervalle compris entre 10 à 40%, les zones favorables au Trigate étant naturellement les mêmes que lors de la comparaison avec le FDSOI sans FBB, et ce pour les mêmes raisons.

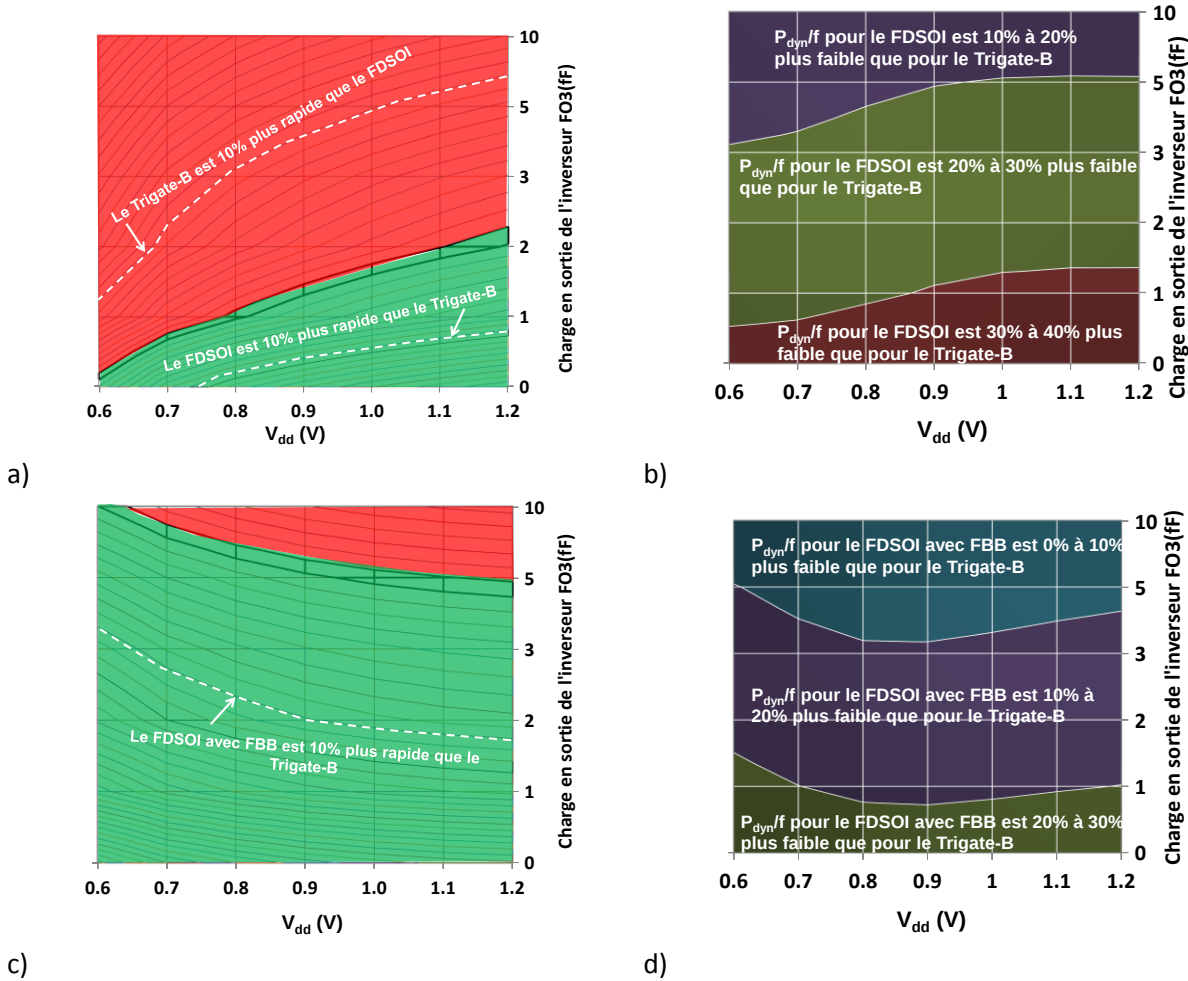


Figure V-36 : Cartographie de la comparaison de fréquence d'anneaux résonnants d'inverseur FanOut 3 en fonction de la tension d'alimentation et de la charge en sortie de chaque étage (chaque bande correspond à 1%), entre le FDSOI et le Trigate-B (a) et le FDSOI avec FBB et le Trigate-B (c). Cartographie de la comparaison d'efficacité énergétique (P_{dyn}/f) d'anneaux résonnants d'inverseur FanOut 3 en fonction de la tension d'alimentation et de la charge en sortie de chaque étage, entre le FDSOI et le Trigate-B (b) et le FDSOI avec FBB et le Trigate-B (d).

La comparaison entre le Trigate-B et le FDSOI est moins favorable au FDSOI (Figure V-36). En effet la longueur de grille plus agressive du Trigate-B (24nm alors qu'elle valait 30nm pour le Trigate-A) a deux effets :

- La pénalité capacitive est réduite en comparaison du FDSOI.
- Le DIBL est plus fort. Le Trigate-B sera alors plus compétitif pour les tensions d'alimentation élevées. Par conséquent, le domaine des faibles tensions d'alimentation ne sera plus une zone favorable au Trigate-B.

Ces deux remarques expliquent les quatre graphes de la Figure V-36. Bien que plus compétitif, le Trigate-B est toujours pénalisé par ces capacités parasites et, quand les conditions lui sont favorables le gain en fréquence

qu'apportent cette architecture à un coût en terme de consommation dynamique. Ce dernier est compris entre 20 et 40% en comparaison au FDSOI et compris entre 0 et 20% en comparaison du FDSOI avec FBB.

V.B.9. Chemin critique DDR3

Nous avons également pu effectuer des simulations de chemin critique DDR3 avec MASTAR VA pour chaque architecture. Ce circuit comprend des dispositifs de différentes largeurs (le nombre de fins sera obtenu en arrondissant à l'entier supérieur le ratio $W/\text{Finpitch}$), ainsi que différentes cellules standards (inverseur, porte NAND, bascule FLIP FLOP...). L'intérêt de ce circuit est d'être représentatif d'un circuit réaliste, et va valider les comportements que nous avons pu identifier sur les anneaux résonnants. Cependant, le signal de sortie de ce circuit ne permet pas une bonne extraction de la puissance dynamique consommée, et nous pourrions seulement comparer ici vitesse (i.e. fréquence) et consommation statique (P_{stat}). La Figure V-37 montre les variations de la fréquence (a) et du courant de fuite statique (b) extraites de simulations de chemin critique DDR3 pour chaque architecture. La comparaison en vitesse donne des résultats analogues à ceux obtenus sur les anneaux résonnants FanOut 3. Le Trigate-A est avantagé par son faible DIBL pour les basses valeurs de tension d'alimentation, mais cette faible valeur de DIBL devient un inconvénient pour les tensions d'alimentation plus forte ($>0.9\text{V}$). Du fait de la pénalité capacitive de l'architecture Trigate, le FDSOI avec FBB est l'architecture la plus rapide, quel que soit la tension d'alimentation alors que le FDSOI (sans FBB) est plus rapide que le Trigate-A pour $V_{\text{dd}} > 0.8\text{V}$ et que le Trigate-B pour $V_{\text{dd}} > 0.9\text{V}$. Enfin, on peut remarquer que les écarts en fréquence sont plus faibles que ceux obtenus pour les anneaux résonnants. Cela s'explique aisément par notre estimation du nombre de fin correspondant à la largeur de chaque dispositif du chemin critique (en arrondissant à l'entier supérieur), qui provoque une légère surestimation du courant débité par chaque dispositif Trigate.

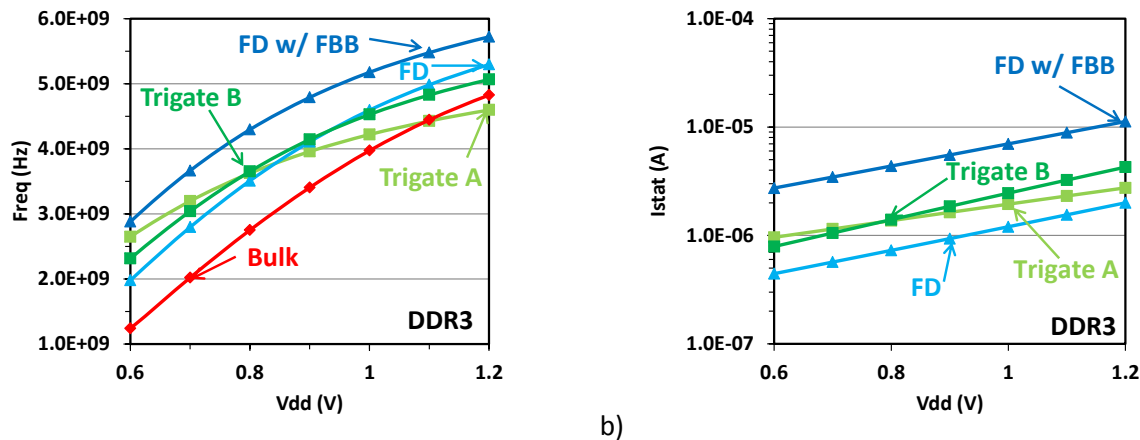


Figure V-37 : Variation de la fréquence (a) et du courant de fuite statique (b) extraites de simulations de chemin critique DDR3 pour chaque architecture.

V.C. Conclusion du chapitre

Dans ce chapitre, nous avons évalué puis comparé les performances logiques des trois architectures CMOS concurrentes pour le nœud technologique 20nm: l'architecture conventionnelle sur substrat massif, le FDSOI et le Trigate par l'utilisation d'un simulateur de circuit conventionnel [ELDO]. Pour être intégrable dans ce type de simulateur, un modèle de transistor doit fournir des caractéristiques courant-tension parfaitement continues (tout comme ses dérivées), mais également un modèle de charges intrinsèques devant être lui aussi continu (ainsi que ces dérivées) en fonction des polarisations. Le modèle de courant drain-source décrit au cours du chapitre II répond parfaitement à cette première exigence. Comme suggéré dans [BSIM], nous avons dérivé notre modèle de charge à partir du modèle de courant drain-source. Pour que le modèle de charge soit complet, nous lui avons ajouté les expressions des charges extrinsèques, c'est-à-dire les capacités parasites, données par le chapitre III. Ce modèle compact, que nous nommons MASTAR_VA, est ensuite implémenté en VerilogA et peut être utilisé avec

un simulateur de circuit comme [ELDO]. MASTAR_VA est un outil prédictif, universel pour chaque technologie CMOS et flexible. Il permet l'évaluation des performances statiques et dynamiques des circuits logiques, des performances de cellules SRAM et de la robustesse à la variabilité des architectures CMOS.

Pour l'évaluation des performances logiques au nœud technologique 20nm, nous avons défini chaque architecture à partir de l'état de l'art, donné par de récentes publications d'industriels (Intel, STMicroelectronics, Samsung et IBM), tant pour les performances des transistors (I_{on}/I_{off} et I_{eff}/I_{off}) que pour le choix des paramètres technologiques. Remarquant que l'architecture Trigate de [Auth 12] présente une longueur de grille relâchée en comparaison des autres architectures, ce qui lui procure un avantage du point de vue électrostatique (DIBL et pente sous le seuil), mais un inconvénient du point de vue capacité (intrinsèque et extrinsèque), nous avons choisi de définir une seconde architecture Trigate avec une longueur de grille plus agressive qui reproduit approximativement l'électrostatique de l'architecture FDSOI. Enfin, la polarisation de la face arrière n'ayant pas d'effet sur les deux autres architectures étudiées, nous n'avons évalué son effet que sur l'architecture FDSOI.

Nous avons ensuite effectué des simulations d'anneaux résonnants d'inverseur de FanOut 1 puis FanOut 3 avec une charge fixe en sortie comprise entre 0 et 10fF et en faisant varier la tension d'alimentation V_{dd} . Nous avons alors démontré que l'architecture conventionnelle sur substrat massif présente des performances en retrait. Nous avons également montré que l'architecture Trigate avec la longueur de grille relâchée est la plus rapide pour les faibles tensions d'alimentation grâce à son faible DIBL. Mais, quand la tension d'alimentation augmente, le faible DIBL devient un inconvénient pour la performance et cette architecture est de moins en moins compétitive. De plus, elle est affectée par un niveau de capacité supérieure à toutes les autres architectures ce qui est pénalisant pour les faibles valeurs de charge de sortie. Cette pénalité est réduite lorsque la charge en sortie augmente car le poids des capacités parasites intrinsèques au transistor est écranté et la vitesse est davantage liée au niveau de courant débité. Ceci explique pourquoi le Trigate avec la longueur de grille plus agressive est plus rapide que la première architecture Trigate dans la plupart des configurations (DIBL plus élevé et capacités parasites plus faibles pour le Trigate agressif). Pour les mêmes raisons, l'architecture FDSOI est plus rapide que les deux architectures Trigate quand la charge de sortie est faible et la tension d'alimentation est élevée. Avec l'utilisation du FBB (i.e. polarisation du substrat en direct, s'est à dire $V_b > 0$ pour un NMOS), l'architecture FDSOI est la plus rapide, sauf pour les charges de sortie très élevées ($> 5fF$) car les architectures Trigate tirent avantage du niveau de courant débité supérieur au FDSOI, même avec FBB, grâce à leurs largeurs électriques supérieures. Le gain en fréquence apporté par l'utilisation du FBB, ~15%, se paie par une augmentation de fuite statique de l'inverseur d'un facteur 5 (limite maximum car le FBB n'est utilisé que lorsque le système a besoin de performance) quelle que soit la tension d'alimentation V_{dd} . En comparaison aux architectures Trigate, l'augmentation de fuite statique du FDSOI avec FBB dépend de la tension d'alimentation car les valeurs de DIBL sont différentes. Celle-ci est augmentée d'un facteur 3 à $V_{dd}=0.8V$.

Les comparaisons de vitesse d'inverseur énoncées ci-dessus ont été effectuées à tension d'alimentation constante. Elles ne reflètent donc pas les différences de consommation dynamique et d'efficacité énergétique vues entre chaque architecture, qui sont très dépendante des valeurs de capacités. En extrayant les valeurs de fréquence pour chaque architecture à même puissance dynamique, nous avons montré que les écarts vus en faveur du FDSOI (avec et sans FBB) comparés aux architectures Trigate à même tension d'alimentation sont exacerbés. Ceci était attendu car le FDSOI présente de faibles valeurs de capacités, donc pour que le FDSOI atteigne une même puissance dynamique que les architectures Trigate, il faut lui appliquer une tension d'alimentation supérieure, ce qui procure un gain supplémentaire en fréquence. Cette comparaison est réaliste et justifiée car les circuits sont aujourd'hui conçus pour être utilisés à leur maximum de performance, et ce maximum est souvent fixé par la dissipation thermique limite du boîtier. La tension d'alimentation est alors fixée pour atteindre cette limite.

Enfin, n'oublions pas les quelques limitations de notre méthodologie: MASTAR_VA ne permet pas de tenir compte des effets de proximité (par exemple la contrainte induite par le transistor voisin qui peut modifier la mobilité), ni des spécificités de chaque architecture concernant les interconnexions métalliques. Ces aspects sont pris en

compte par les modèles SPICE industriels, que MASTAR_VA n'a pas vocation à remplacer, mais pour des technologies dont le développement est déjà avancé.

- CHAPITRE VI -

EVALUATION DES PERFORMANCES AU NŒUD 14-16 NM.

VI.A. Contexte

Dans ce dernier chapitre, nous allons utiliser la prédictivité de MASTAR VA pour comparer les performances logiques des architectures conventionnelles sur substrat massif, FDSOI et Trigate (VI.B) de manière analogue au chapitre précédent. Nous incluons l'impact de la variabilité des paramètres technologiques due au procédé de fabrication de ces architectures sur les performances de cellules SRAM à six transistors (0).

VI.B. Estimation des performances logiques au nœud 14nm

Dans cette partie, nous allons commencer par définir les paramètres technologiques de chaque architecture de dispositif, correspondant au nœud 14nm. Nous comparerons ensuite les performances statiques obtenues, puis évaluerons les capacités parasites de chaque architecture. Enfin, nous comparerons les performances dynamiques obtenues en nous focalisant uniquement sur des simulations d'anneaux résonnants de FanOut égal à 3, avec une charge de sortie fixe (car il s'agit du circuit le plus représentatif de la performance réelle d'un circuit).

VI.B.1. Définition des dispositifs

Cette étape est cruciale pour fournir une estimation de performance réaliste. Pour l'effectuer, nous choisissons comme point de départ les cartes modèles réalisées pour l'évaluation de performance au nœud 20nm (Chapitre V). Ensuite, nous réduisons les dimensions, en accord avec une miniaturisation suivant la loi de Moore, c'est-à-dire qu'elles subissent une réduction de 30% en comparaison du nœud précédent:

- Pas de répétition d'une grille et d'un contact CPP=56nm. Sa valeur est bien égale à la moitié de celle reportée pour le nœud 28nm où CPP=113nm [Arnaud 09]. Il faut cependant remarquer que, si on applique un facteur de 0.7 au CPP reporté pour la technologie Intel 22nm par [Auth 12], on obtient CPP=64nm, ce qui correspond davantage au nœud 16nm.
- Epaisseur électrique d'oxyde de grille en régime d'inversion $t_{inv}=1.2\text{nm}$.
- Tension d'alimentation nominale $V_{dd}=0.8\text{V}$.

Les autres paramètres technologiques sont propres à chaque architecture, et seront déterminés comme suit :

- L'architecture conventionnelle sur substrat massif :
 - La longueur de grille est fixée à $L=20\text{nm}$ (elle valait 26nm pour le nœud précédent, cf chapitre V).
 - Le dopage est ajusté de manière à avoir un contrôle des effets canaux courts corrects, tout en évitant d'aller vers des valeurs trop élevées afin de ne pas trop dégrader la mobilité.
 - Les facteurs d'amélioration du transport sont les mêmes que pour le nœud 20nm, car nous avons considéré que les techniques de contrainte étaient déjà au maximum de leur possibilité au nœud 20nm.
- L'architecture FDSOI :
 - La longueur de grille est fixée à $L=16\text{nm}$ (elle valait 22nm au nœud précédent, cf chapitre V).
 - L'épaisseur de film de silicium est jugée constante à $t_{si}=6\text{nm}$. Dans ces gammes d'épaisseur, nous avons estimé qu'il était plus simple du point de vue du procédé de fabrication, et tout autant efficace pour ce qui est du contrôle électrostatique de réduire l'épaisseur de l'oxyde enterré en conservant la même épaisseur de canal.
 - L'épaisseur d'oxyde enterré est réduite à $t_{box}=10\text{nm}$.
 - La mobilité est améliorée d'un facteur 2 pour le NMOS à canal long grâce à l'introduction des substrats SOI contraints (SSOI pour « Strained Silicon On Insulator ») [Fenouillet 12], ce qui mène à un gain de 50% pour la mobilité du NMOS à canal court. Pour le PMOS, nous considérerons que la mobilité canal court est la même que pour le nœud précédent (on fait donc l'hypothèse que les

améliorations technologiques permettront de compenser la perte en mobilité effective sur le PMOS due au substrat SSOI).

- L'architecture Trigate (extrapolée à partir du Trigate-A du chapitre V) :
 - La longueur de grille est fixée à $L=20\text{nm}$ (elle valait 30nm au nœud précédent, cf chapitre V).
 - L'épaisseur du fin est réduite à $t_{\text{si}}=8\text{nm}$ (elle valait 10nm au nœud précédent)
 - Nous considérons que le rapport de forme du fin reste le même, c'est-à-dire que le ratio $h_{\text{si}}/t_{\text{si}}=3$. Donc la hauteur de fin vaut $h_{\text{si}}=24\text{nm}$.
 - Le finpitch est réduit à 40nm .
 - La mobilité des deux types de transistor est la même que pour le nœud 20nm car ils semblent que plusieurs techniques de contraintes ont déjà été mise en œuvre dans [Auth 12], notamment par l'introduction d'épitaxie SiGe qui viennent contraindre le canal, et par le dépôt du couche de nitrure ou d'oxyde qui vient également exercer une contrainte sur le canal.

Les informations énoncées ci-dessus sont résumées dans le tableau de la Figure VI-1, ainsi que les paramètres électrostatiques et les courants caractéristiques (normalisés par l'empreinte physique) obtenus avec MASTAR VA.

	BULK		FDSOI		Trigate	
Type	N	P	N	P	N	P
V_{dd} (V)	0.8					
CPP (nm)	56					
t_{inv} (nm)	1.2					
L_{nom} (nm)	20	20	16	16	20	20
t_{si} (nm)	/	/	6	6	8	8
h_{si} (nm)	/	/	/	/	20	20
t_{box} (nm)	/	/	10	10	/	/
FinPitch (nm)	/	/	/	/	40	40
DIBL (mV)	162	162	109	109	52	52
SS (mV/dec)	101	100	90	90	75	75
I_{on} ($\mu\text{A}/\mu\text{m}$)	1090	1233	1224	1257	1368	1252
I_{off} (nA/ μm)	60	60	60	60	60	60
I_{eff} ($\mu\text{A}/\mu\text{m}$)	476	569	632	641	789	718

Figure VI-1: Tableau résumant les paramètres technologiques de chaque architecture de transistor, puis résumé des performances statiques ainsi obtenus, en fixant la fuite $I_{\text{off}}=60\text{nA}/\mu\text{m}$.

VI.B.2. Performance statique

Dans ce paragraphe, nous comparons les performances statiques ainsi obtenues. La Figure VI-2 représente les compromis $I_{\text{on}}/I_{\text{off}}$ et $I_{\text{eff}}/I_{\text{off}}$ estimés pour chaque architecture de dispositifs et pour les deux types de transistor (a,b NMOS et c,d PMOS). Tout d'abord, en ce qui concerne les NMOS, on remarque que l'architecture conventionnelle sur substrat massif fournit un courant de saturation I_{on} faible en comparaison des deux autres structures (~15% de moins que le FDSOI et ~30% que le Trigate à $I_{\text{off}}=60\text{nA}/\mu\text{m}$). Ces écarts sont exacerbés si on compare les courants effectifs, du fait du fort DIBL de l'architecture conventionnelle sur substrat massif. Cette

seconde remarque est également valable pour la comparaison entre FDSOI et Trigate, car le DIBL du FDSOI est bien supérieur à celui de l'architecture Trigate (car la longueur de grille est plus agressive pour le FDSOI). Pour les transistors PMOS, les courants de saturation sont approximativement tous les mêmes à $I_{off}=60\text{nA}/\mu\text{m}$. Cependant, si on se concentre plutôt sur les courants effectifs I_{eff} , les écarts se creusent entre chaque architecture du fait des DIBL différents de chacune.

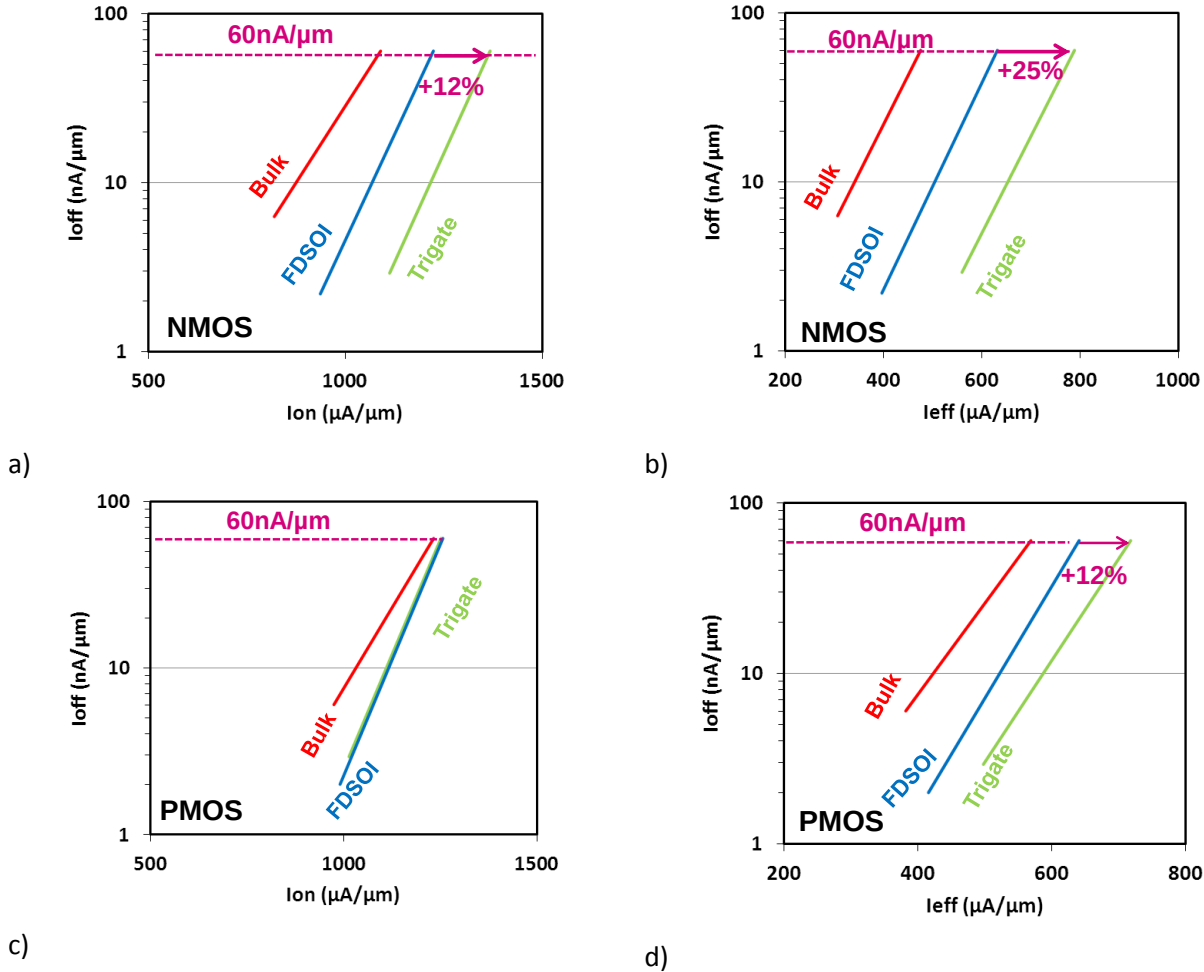


Figure VI-2 : Comparaison des compromis I_{on}/I_{off} (a,c) et I_{eff}/I_{off} (b,d) pour chaque architecture de dispositif.

VI.B.3. Evaluation des capacités parasites

Pour estimer les capacités parasites de chaque architecture, il faut en premier lieu connaître chaque dimension. Pour ce nœud technologique, nous considérons que l'architecture conventionnelle sur substrat massif et l'architecture FDSOI sont les mêmes que pour le nœud précédent. Par contre, pour l'architecture Trigate, nous considérerons que l'épithaxie source-drain vient joindre tous les fins car les dimensions sont trop agressives, notamment l'espace entre deux fins (32nm), pour permettre une discrétisation des source-drains. Comme pour le chapitre précédent, les dimensions sont estimées à partir des règles simples, énoncées dans le chapitre III et résumées dans la figure III-37. Les dimensions ainsi obtenues sont résumées dans le tableau de la Figure VI-4.

(nm)	BULK	FDSOI	Trigate
L_{nom}	20	16	20
t_{inv}	1.2	1.2	1.2
T_{si}	/	6	8
H_{si}	/	/	24
T_{box}	/	10	/
H_g	40	32	40
T_{sp1}	6	6.66	6
T_{sp2}	6	6.66	6
H_{epi}	20	12	20
Finpitch	/	/	40

Figure VI-3: Résumé des dimensions considérées pour l'évaluation des capacités parasites de chaque architecture.

En utilisant les modèles analytiques développés dans le chapitre III, on peut estimer chaque capacité parasite, pour chaque architecture. Ces dernières sont représentées sur la Figure VI-4. Comme pour le nœud 20nm, on constate que l'architecture Trigate est pénalisée par ses capacités parasites. Cependant, la pénalité capacitive de l'architecture Trigate due à la longueur de grille relâchée vue pour le nœud 20nm est très réduite dans cette nouvelle étude (car les écarts entre longueurs de grille sont plus faibles). Par conséquent, la pénalité capacitive totale de l'architecture Trigate, comparée au FDSOI, sera plus faible dans la comparaison de performance au nœud 14nm que celle vue dans le chapitre précédent pour le nœud 20nm.

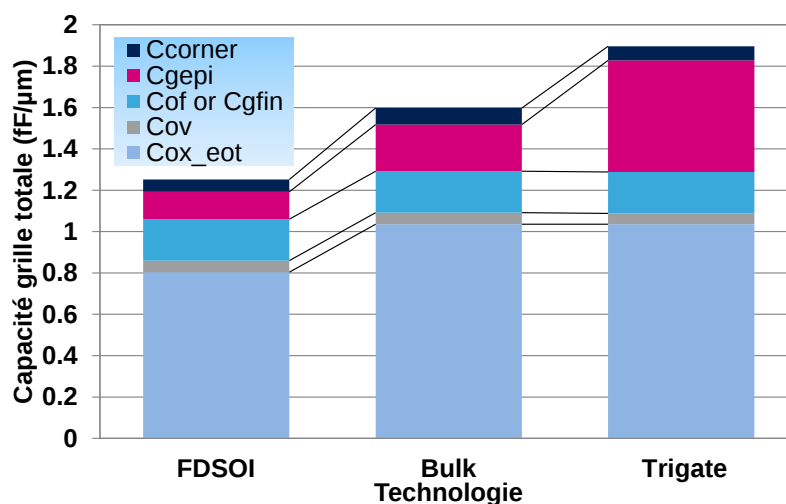


Figure VI-4: Somme de la capacité de grille intrinsèque en inversion (C_{inv}) et de la capacité parasite grille-drain (C_{gd}) pour chaque architecture, en distinguant chaque composante. Les capacités sont normalisées par la largeur de l'empreinte sur silicium.

VI.B.4. Anneaux résonnants d'inverseur FO3 avec charge additionnelle de sortie variable

Pour l'évaluation de performance dynamique au nœud 16nm, nous nous focaliserons sur les anneaux résonnants constitués d'inverseurs FO3 avec une charge en sortie de chaque étage comprise entre 0 et 10fF. Les inverseurs sont constitués de NMOS et de PMOS de même largeur, fixée à $W=0.12\mu\text{m}$ en considérant une réduction de 30% par rapport au nœud 20nm. Les inverseurs utilisant l'architecture Trigate seront constitués de trois fins ce qui mène à une empreinte physique sur silicium $W_{\text{FootPrint}}=120\text{nm}$, et donc une largeur électrique $W_{\text{elec}}=168\text{nm}$. Nous choisissons donc de ne pas étudier le comportement des inverseurs FO1 car ces derniers sont moins représentatifs de la performance réelle d'un circuit. De plus, pour ce nœud technologique, nous ne disposerons pas du chemin critique DDR3 car la conception de ce dernier doit être effectuée pour chaque nœud technologique et n'a pas encore été faite par les équipes de conception. Enfin, pour la prise en compte du délai RC dû aux interconnexions des niveaux de métallisation, nous utiliserons le même PEX (Parasitic EXtraction) que celui utilisé et développé pour le nœud 20nm (le PEX 14nm n'est pas encore disponible car les règles de dessin et les matériaux des niveaux de métallisation ne sont pas encore connus). Comme dans le chapitre précédent, et pour les mêmes raisons, nous effectuerons des simulations avec polarisation de la face arrière à V_{dd} uniquement pour l'architecture FDSOI. Au vue des dimensions de cette architecture pour le nœud 14nm, le décalage de tension de seuil sera de 100mV/V de polarisation de face arrière.

Commençons par examiner le comportement en vitesse des inverseurs FO3 en fonction de la tension d'alimentation V_{dd} . La Figure VI-5-a représente la variation de fréquence d'un anneau résonnant FO3 sans charge en sortie. Comme dans le chapitre précédent, on remarque que l'architecture est pénalisée par ces capacités parasites. Cependant, pour les faibles tensions d'alimentation ($V_{\text{dd}} < 0.65\text{V}$) l'excellente électrostatique de cette architecture, notamment son faible DIBL, lui permet d'être plus rapide que le FDSOI. Enfin, quelle que soit la tension d'alimentation, l'architecture FDSOI avec FBB est la plus rapide grâce à sa faible capacité totale et à son niveau de courant débité élevé. Comme dans le chapitre précédent, lorsque la charge de sortie augmente la pénalité capacitive de l'architecture Trigate est écrantée (Figure V-26) et le Trigate devient plus rapide que le FDSOI pour une charge de sortie de 2fF (Figure VI-5-b) puis à peu près équivalent au FDSOI avec FBB pour 5fF (Figure VI-5-c). Enfin, la Figure VI-5-d représente la fréquence d'un anneau résonnant FO3 en fonction de la valeur de la charge en sortie de chaque étage pour la tension d'alimentation nominale du nœud 14nm, c'est-à-dire $V_{\text{dd}}=0.8\text{V}$. Ce graphe confirme les remarques précédentes : l'architecture FDSOI avec FBB est la plus performante en terme de vitesse, et que l'architecture Trigate est à peu près équivalente à celle-ci uniquement pour les charge de sortie élevée (>5fF).

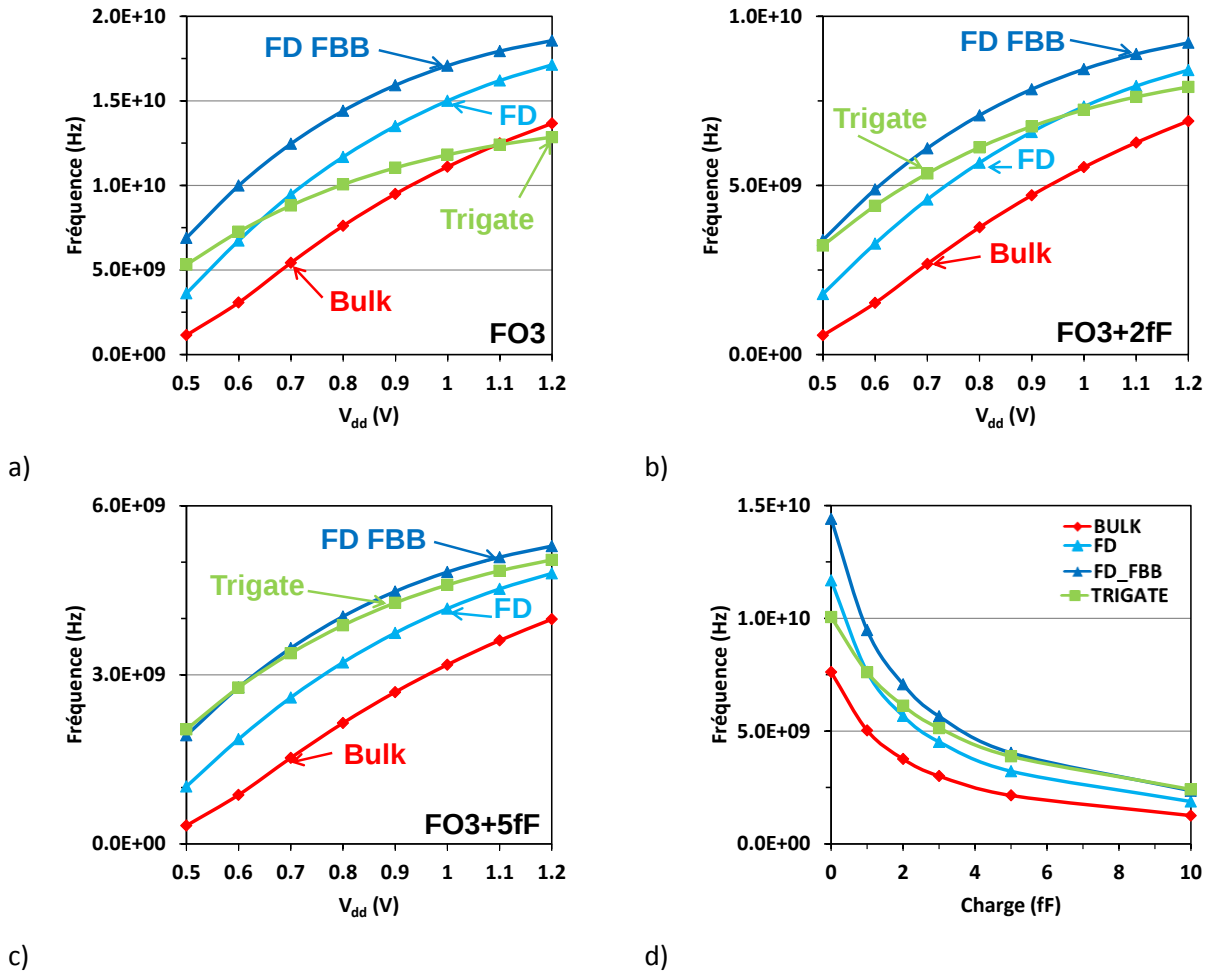
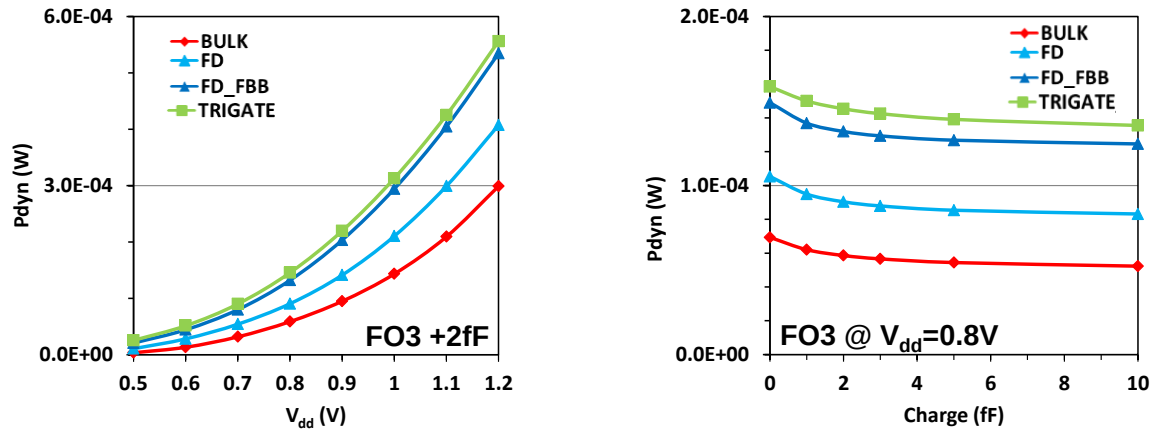


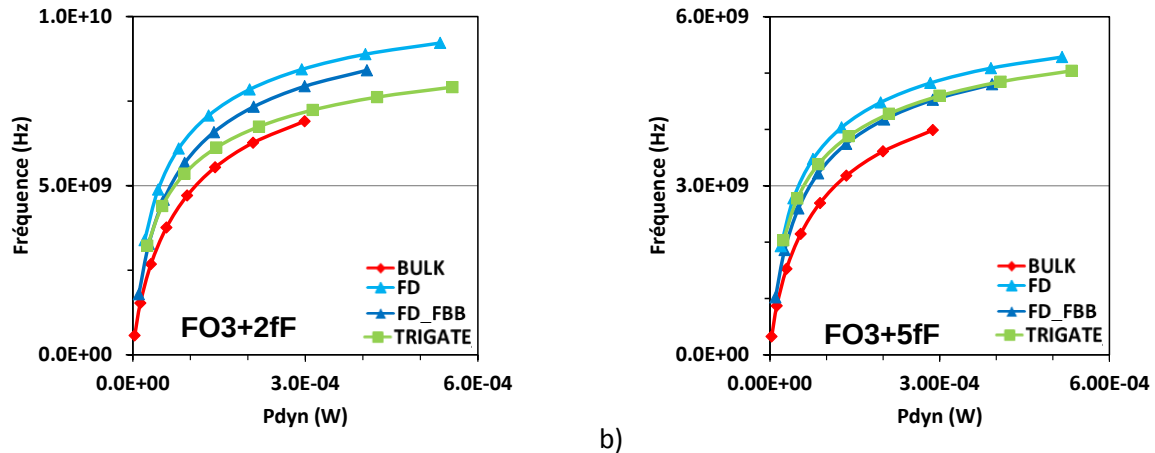
Figure VI-5 : Tracé de la fréquence d'un anneau résonnant FO3 en fonction de la tension d'alimentation V_{dd} : a) 0fF, b) 2fF et c) 5fF. d) Tracé de la fréquence d'un anneau résonnant FO3 en fonction de la valeur de la charge en sortie de chaque étage pour la tension d'alimentation nominale du nœud 14nm, c'est-à-dire $V_{dd}=0.8V$.

La Figure VI-6-a représente la variation de puissance dynamique d'un anneau raisonnant FO3 avec une charge de sortie à 2fF en fonction de la tension d'alimentation V_{dd} . On constate que l'architecture Trigate est celle qui consomme le plus du fait de sa valeur élevée de capacité. La Figure VI-6-b montre la variation de puissance dynamique pour un anneau résonnant d'inverseur FanOut 3 en fonction de la valeur de la charge fixe en sortie à tension d'alimentation constante $V_{dd}=0.8V$. On remarque également que l'architecture Trigate est toujours la plus consommatrice. Enfin, sur ces deux graphiques, l'architecture conventionnelle sur substrat massif (BULK) est celle qui consomme le moins. Comme dans le chapitre précédent, ceci s'explique par sa faible valeur de fréquence en comparaison des architectures concurrentes. Pour analyser la consommation dynamique, il faut donc la représenter en fonction de la fréquence.



a) b)
Figure VI-6 : a) Tracé de la puissance dynamique (P_{dyn}) en fonction de la tension d'alimentation V_{dd} pour un anneau résonnant d'inverseur FanOut 3 avec une charge fixe en sortie de valeur 2fF. b) Tracé de la puissance dynamique (P_{dyn}) pour un anneau résonnant d'inverseur FanOut 3 en fonction de la valeur de la charge fixe en sortie à tension d'alimentation constante $V_{dd}=0.8V$.

La Figure VI-7 représente la fréquence en fonction de la puissance dynamique pour chaque architecture pour des anneaux résonnants FO3 pour une charge de sortie de 2fF (a) et 5fF (b). Ces graphes vont traduire l'efficacité énergétique de chaque technologie CMOS en indiquant quelle fréquence peut être atteinte à une consommation donnée et de manière duale, donnée la consommation dynamique à une fréquence donnée. Pour les deux valeurs de charge de sortie étudiée ici, le FDSOI avec utilisation du FBB est l'architecture la plus efficace énergétiquement. Pour une charge de sortie de 2fF (a), le Trigate est moins efficace énergétiquement que le FDSOI, alors que pour une charge de sortie de 5fF (b), ces deux architectures présentent le même compromis puissance dynamique/fréquence. Ce comportement s'explique toujours par la pénalité capacitive de l'architecture Trigate qui est de plus en plus écartée par la charge de sortie quand celle-ci augmente.



a) b)
Figure VI-7 : Tracé de la fréquence en fonction de la puissance dynamique pour un anneau résonnant d'inverseur FanOut 3 avec une charge en sortie de 2fF (a) et 5fF (b).

La Figure VI-8 montre la variation de la fuite statique d'un anneau résonnant d'inverseur FanOut 3 avec la tension d'alimentation V_{dd} . Ce graphe montre que l'utilisation du FBB le FDSOI a un impact sur la fuite statique du circuit : elle sera sept fois supérieure à celle du FDSOI sans utilisation du FBB quelle que soit la tension d'alimentation (l'écart est constant car ils ont le même DIBL). Si on la compare à celle de l'architecture Trigate, le niveau de fuite

statique sera quatre fois supérieur à $V_{dd}=0.5V$ et neuf fois supérieur à $V_{dd}=1.2V$, ce qui s'explique par l'excellent DIBL de l'architecture Trigate. Cependant, comme dans le chapitre précédent, il faut mentionner que le FBB n'est utilisé que lorsque que la performance (i.e la vitesse) est nécessaire au circuit. Par conséquent, l'augmentation de fuite statique due au FBB reportée dans la Figure VI-8 est en fait une limite maximum et la valeur réelle d'augmentation de fuite est très dépendante du circuit.

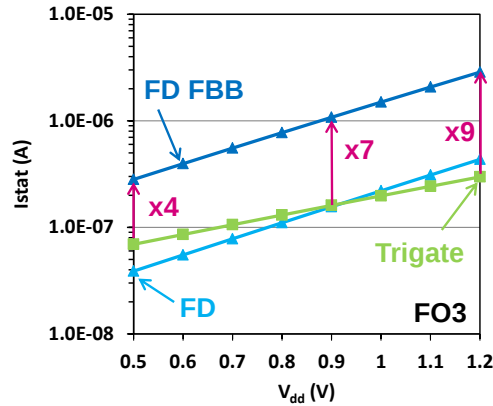


Figure VI-8 : Variation de la fuite statique de l'anneau résonnant d'inverseur FanOut 3 avec la tension d'alimentation V_{dd} .

Enfin, comme dans le chapitre précédent, pour une analyse plus fine sur l'espace complet défini par les tensions d'alimentation étudiées d'une part, et des charges en sortie de chaque étage d'autre part, nous traçons les cartographies des écarts en fréquence et en efficacité énergétique (P_{dyn}/f) entre le FDSOI et le Trigate (Figure VI-9) puis entre le FDSOI avec FBB et le Trigate (Figure VI-10). Sur la Figure VI-9-a, on remarque que l'espace favorable au Trigate se situe pour les tensions d'alimentation faibles et pour les charges en sortie de chaque étage élevées. Ceci s'explique par le faible DIBL du Trigate (pour les faibles V_{dd}) et par la pénalité capacitive du Trigate (pour les charges en sortie élevées). De plus, quelles que soient les conditions, l'architecture FDSOI est la plus efficace énergétiquement (Figure VI-9-b). L'écart est compris entre 30 et 50% lorsque le FDSOI est l'architecture la plus rapide, et entre 10 et 30% lorsqu'il s'agit du Trigate qui est plus rapide. Par conséquent, une comparaison à même puissance dynamique, et non plus à même tension d'alimentation élargirait l'espace tension d'alimentation/charge favorable à l'architecture FDSOI.

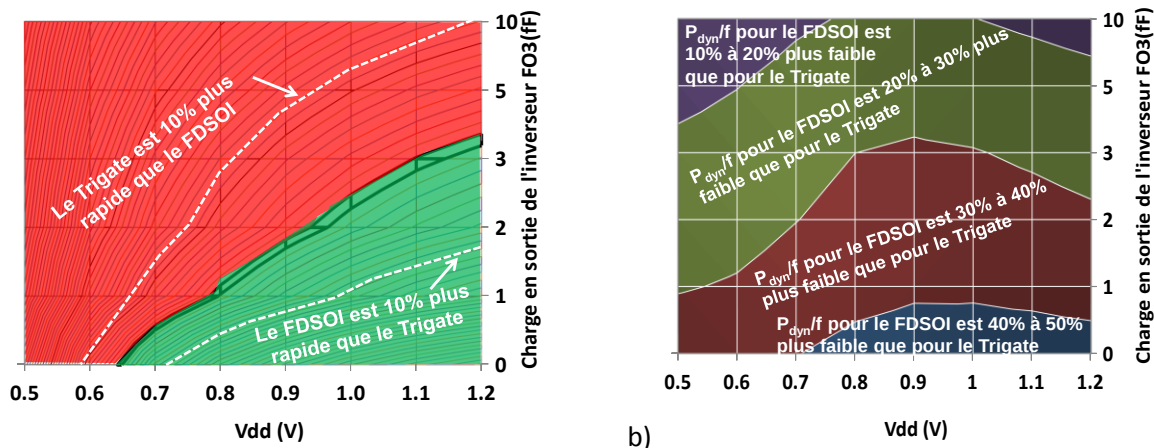
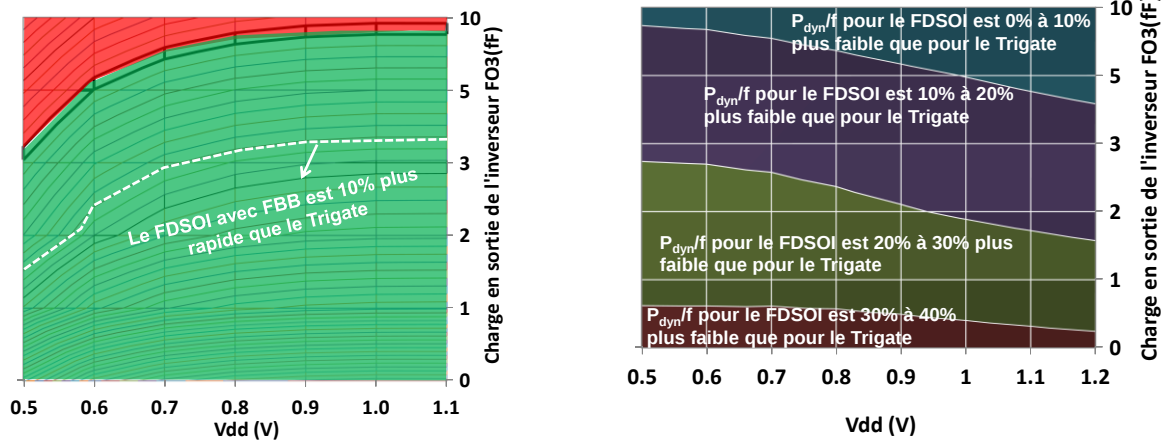


Figure VI-9 : a) Cartographie de la comparaison de fréquence d'anneaux résonnants d'inverseur FanOut 3 en fonction de la tension d'alimentation et de la charge en sortie de chaque étage, entre le FDSOI et le Trigate (chaque bande correspond à 1%). b) Cartographie de la comparaison d'efficacité énergétique (P_{dyn}/f) d'anneaux résonnants d'inverseur FanOut 3 en fonction de la tension d'alimentation et de la charge en sortie de chaque étage, entre le FDSOI et le Trigate.

Enfin, si l'option FBB est utilisée sur l'architecture FDSOI, l'espace favorable à l'architecture Trigate en terme de vitesse se situe toujours pour les faibles tensions d'alimentation et les charges de sortie élevées mais ce dernier est considérablement réduit (Figure VI-10-a). Cependant, le gain en vitesse apporté par le FBB pour le FDSOI se paie par une augmentation de la consommation dynamique, et l'efficacité énergétique est un peu moins bonne. Elle reste tout de même meilleure que celle du Trigate (Figure VI-10-b) et l'écart est compris entre 10 et 40% lorsque le FDSOI est l'architecture la plus rapide, et entre 0 et 15% lorsque le Trigate est l'architecture la plus rapide.



a) Figure VI-10 : a) Cartographie de la comparaison de fréquence d'anneaux résonnants d'inverseur FanOut 3 en fonction de la tension d'alimentation et de la charge en sortie de chaque étage, entre le FDSOI avec FBB et le Trigate (chaque bande correspond à 1%). b) Cartographie de la comparaison d'efficacité énergétique (P_{dyn}/f) d'anneaux résonnants d'inverseur FanOut 3 en fonction de la tension d'alimentation et de la charge en sortie de chaque étage, entre le FDSOI avec FBB et le Trigate.

VI.C. Performance SRAM

Nous proposons dans cette partie de comparer les performances SRAM des trois architectures au nœud 14nm. Dans une plateforme CMOS, les transistors utilisés dans les parties logiques et mémoire SRAM d'un même circuit ne subissent pas exactement le même procédé de fabrication. Ils ne présentent donc pas les mêmes performances statiques et nous commencerons donc par définir les dispositifs qui seront comparés dans cette partie. Nous déterminerons ensuite les règles de dessin de chaque cellule SRAM au nœud 14nm pour chaque architecture de dispositif, puis définirons les sources de variabilité dans chaque cas. Enfin, nous procéderons à des simulations des cellules SRAM définies, en tenant compte de la variabilité, et comparerons leur performance.

VI.C.1. Définition des dispositifs

Pour définir les dispositifs des cellules SRAM, la première étape est de se donner les principaux paramètres technologiques de chaque architecture. Commençons par les paramètres que nous considérerons commun à chaque architecture :

- Le pas de répétition d'une grille et d'un contact sera égal à CPP=64nm. Nous obtenons cette valeur en considérant que, en accord avec la loi de Moore, cette dernière est divisée par deux tous les deux nœuds (CPP=128nm pour le nœud 32nm [Chen 08]).
- La tension d'alimentation nominale du nœud 16nm est fixée à 0.8V, en considérant une réduction de 100mV par rapport au nœud 20nm [Cho 11].

- L'épaisseur d'oxyde équivalente en régime d'inversion est fixée à $t_{inv}=1.2nm$. La réduction de cette dimension est faible par rapport au nœud 20nm ($t_{inv}=1.3nm$) car un affinement trop important de l'oxyde de grille aurait des conséquences trop importantes sur les fuites de grille.
- La longueur de grille est fixée à 20nm. Pour les nœuds technologiques CMOS avancés, la longueur de grille ne correspond plus à la valeur du nœud technologique (i.e nœud 22nm $L=30nm$ [Auth 12] nœud 20nm $L=26nm$ [Shang 12].)

Pour les dimensions spécifiques à chaque architecture, nous procédons de la manière suivante :

- L'épaisseur de jonction pour l'architecture conventionnelle sur substrat massif X_j , l'épaisseur du film de silicium t_{si} pour l'architecture FDSOI et l'épaisseur du fin pour l'architecture Trigate t_{si} , nous utiliserons [Skotnicki 08] qui donne des règles simples qui permettent de déterminer ces dimensions à partir de t_{inv} et L en assurant un contrôle électrostatique de la grille sur le canal correct, et donc des dispositifs fonctionnels.
- La hauteur de fin pour l'architecture Trigate est déterminée à partir de l'épaisseur du fin t_{si} et du rapport de forme $h_{si}/t_{si}=2.5$ donné par [Kawazaki 09].

Les paramètres technologiques énoncés ci-dessus sont résumées dans le tableau de la Figure VI-11.

	BULK	FDSOI	FinFET
CPP (nm)	64		
L_{nom} (nm)	20		
t_{inv} (nm)	1.2		
V_{dd} (V)	0.8		
N_{ch} (cm³)	6.8e18	1e16	1e16
T_{si}/X_j (nm)	10	5	10
H_{si} (nm)	/	/	25
t_{box} (nm)	/	10	100

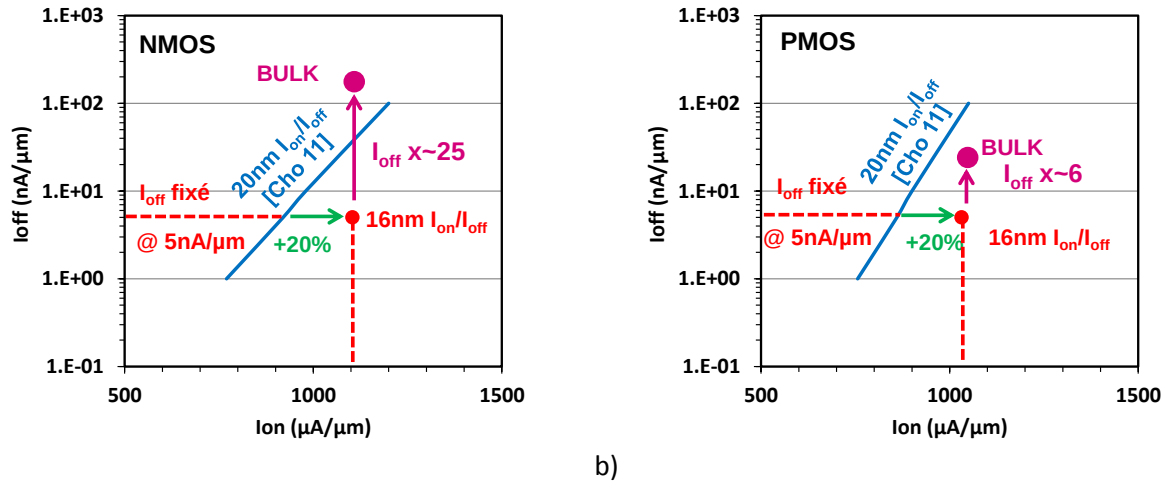
Figure VI-11 : Tableau résumant les paramètres technologiques des trois architectures utilisés pour l'évaluation des performances SRAM.

Pour ensuite ajuster les paramètres de transport et de résistance d'accès de chaque structure, il faut nous donner une spécification de performance statique, c'est-à-dire un compromis I_{on}/I_{off} typique du nœud 16nm. Ne disposant d'aucune publication pour ce nœud technologique, nous avons choisi d'extrapoler sa valeur à partir de performances statiques typiques du nœud 20nm reportées dans la littérature par [Cho 11] pour l'architecture conventionnelle sur substrat massif :

- Nous fixons le courant de fuite $I_{off}=5nA/\mu m$.
- Nous considérons une amélioration du courant de saturation I_{on} de 20% entre le nœud 20nm et le nœud 16nm, à fuite constante.
- On obtient alors les valeurs de courant de saturation I_{on} ciblées pour le nœud 16nm à $I_{off}=5nA/\mu m$ et $V_{dd}=0.8V$ pour les deux types de transistors :
 - NMOS : $I_{on}= 1104 \mu A/\mu m$
 - PMOS : $I_{on}= 1032 \mu A/\mu m$

Avec les paramètres technologiques de la Figure VI-11 insérés dans MASTAR VA, les architectes FDSOI et Trigate parviennent à atteindre nos spécifications I_{on}/I_{off} pour le nœud 16nm. Par contre, pour que l'architecture

conventionnelle sur substrat massif atteint la valeur ciblée de I_{on} , la contrainte sur le courant de fuite doit être relâchée, d'un facteur 20 pour le NMOS, et 6 pour le PMOS. Ceci est illustré sur la Figure VI-12, ainsi que la méthode d'extrapolation des spécifications de performances statiques du nœud 16nm à partir de celles du nœud 20nm.



a) b) Figure VI-12: Compromis I_{on}/I_{off} illustrant la méthodologie d'extrapolation des spécifications de performances statiques pour le nœud 16nm, à partir de celles du nœud 20nm. Ces graphes illustrent également que l'architecture conventionnelle sur substrat massif ne peut pas atteindre les valeurs de courant de saturation spécifiée sans relâcher la contrainte sur le courant de fuite I_{off} . a) NMOS b) PMOS

VI.C.2. Définition des dessins de cellule SRAM

La seconde étape est de définir les dessins de cellules SRAM construites à partir de chaque architectures de transistor, afin de garantir une comparaison équitable d'une part, et avec du sens d'autre part. La première contrainte que nous nous fixons porte sur la surface du point mémoire SRAM. En effet, comme pour la partie logique, l'encombrement de la partie mémoire d'un circuit doit être réduit de nœuds technologiques en nœuds technologiques. Pour obtenir une valeur cible de surface d'une cellule SRAM typique du nœud 16nm, nous l'avons extrapolée sur la courbe représentant la surface des cellules SRAM publiée par les industriels pour chaque nœud technologique (Figure VI-13). Par extrapolation linéaire sur cette courbe, on obtient une surface de $0.04 \sim 0.045 \mu\text{m}^2$ pour un point mémoire SRAM au nœud 16nm.

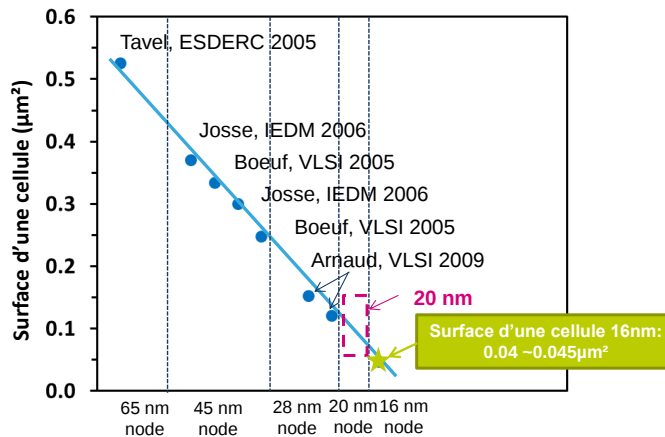


Figure VI-13 : Variation de la surface des cellules SRAM publiée par les industriels en fonction du nœud technologique. Par extrapolation linéaire sur cette courbe, on obtient une surface de $0.04 \sim 0.045 \mu\text{m}^2$ pour un point mémoire SRAM au nœud 16nm.

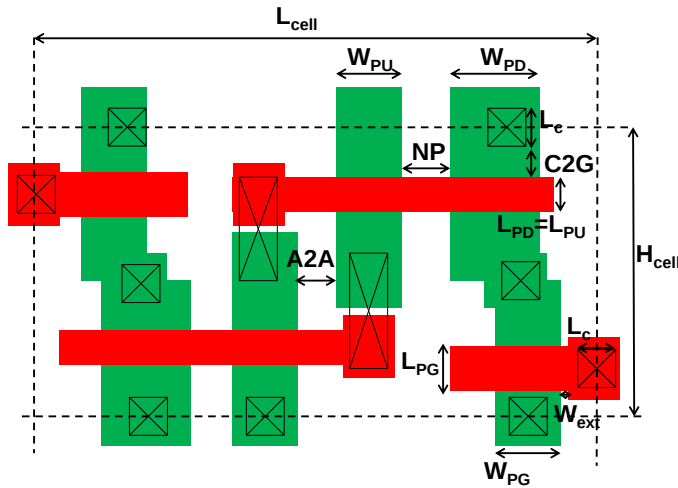
Ensuite, pour estimer l'encombrement surfacique S des cellules SRAM que nous allons définir pour chaque architecture, nous utiliserons l'équation suivante :

$$S = H_{\text{cell}} \times L_{\text{cell}} \quad \text{Eq. VI-1}$$

$$H_{\text{cell}} = 2(L_{\text{PD}} + \frac{L_c}{2} + C2G) \quad \text{Eq. VI-2}$$

$$L_{\text{cell}} = 2(W_{\text{PD}} + W_{\text{PU}} + L_c + \frac{A2A}{2} + \text{polyendcap} + \frac{\text{polyspace}}{2} + NP) \quad \text{Eq. VI-3}$$

Où L_{cell} et H_{cell} sont représentés sur le dessin classique d'une cellule SRAM (Figure VI-14-a) ainsi que chaque dimension nécessaire à leur évaluation. Les dimensions ne dépendant pas du dessin de la cellule sont estimées à partir des valeurs de longueur de grille L et de pas de répétition d'un contact CPP, suivant les règles simples énoncées dans la Figure VI-14-b.



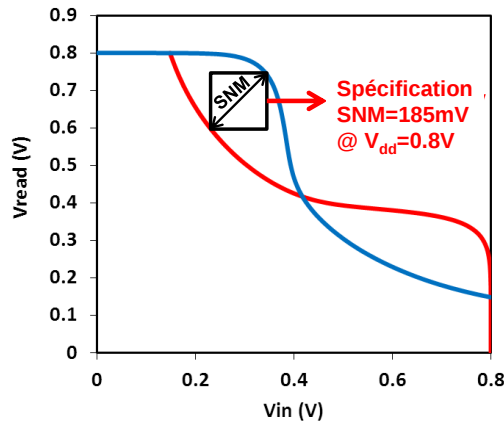
	Equation
C2G	$(\text{CPP} - L_{\text{nom}})/3$
L_c	$(\text{CPP} - L_{\text{nom}})/3$
Polyendcap	L_{nom}
A2A	$(\text{CPP} - L_{\text{nom}})/3$
Polyspace	$(\text{CPP} - L_{\text{nom}})/3$
NP	$2.(\text{CPP} - L_{\text{nom}})/3$

b)

a)

Figure VI-14 : a) Dessin classique d'une cellule SRAM avec ces principales dimensions, nécessaires à l'évaluation de son encombrement surfacique. b) Règles simples données au cours du chapitre III pour l'estimation des dimensions à partir des valeurs de longueur de grille L et de pas de répétition d'un contact CPP.

Enfin, pour ajuster chaque dessin de cellule SRAM, nous choisissons de cibler une valeur de SNM (Static Noise Margin, marge de bruit statique en français, définie sur la Figure VI-15-a) à la tension d'alimentation nominale du nœud 16nm, soit $V_{\text{dd}}=0.8\text{V}$. Nous la fixons à une valeur typique garantissant le fonctionnement de la cellule, soit $\text{SNM}=185\text{mV}$, comme démontré par [Planes 08] pour le nœud 45nm. Nous choisirons donc d'ajuster nos cellules SRAM de manière à atteindre $\text{SNM}=185\text{mV}$ à $V_{\text{dd}}=0.8\text{V}$, tout en respectant la contrainte d'encombrement surfacique.



	BULK	FDSOI	Trigate
WPU	50 nm	55 nm	60 nm
LPU	25 nm	20 nm	20 nm
WPD	50 nm	55 nm	60 nm
LPD	25 nm	20 nm	20 nm
WPG	42 nm	40 nm	60 nm
LPG	30 nm	26 nm	30 nm
Area	0.047 μm^2	0.046 μm^2	0.049 μm^2
SNM	186 mV	182 mV	186 mV

b)

a)

Figure VI-15 : a) définition de la SNM (Static Noise Margin) sur une courbe papillon (tension d'entrée en fonction de la tension de sortie) typique d'une cellule SRAM. b) Dessins de cellules SRAM obtenus pour chaque architecture de transistor.

Il faut remarquer ici que, du fait de notre contrainte portant sur la surface de la cellule SRAM, la cellule SRAM construite à partir de transistor Trigate est limitée à des transistors à un seul fin. Dans ce cas, l'ajustement de la cellule pourra être effectué seulement au travers de la longueur de grille des transistors pour atteindre la valeur de SNM spécifiée (contrairement au cas planaire où chacune des largeurs des trois transistors de la cellule pourra être librement modifiée). Nous pouvons alors effectuer des simulations des cellules SRAM données par la Figure VI-15-b en utilisant MASTAR VA au travers d'un simulateur de circuit conventionnel [ELDO]. On peut alors tracer la variation de la SNM en fonction de la tension d'alimentation V_{dd} dans chaque cas (Figure VI-16).

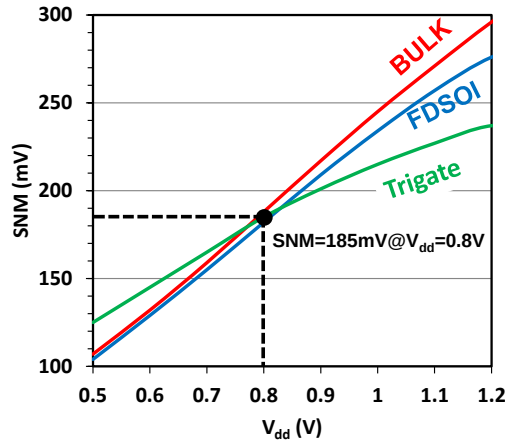


Figure VI-16 : Variation de la SNM en fonction de la tension d'alimentation V_{dd} pour chaque architecture.

VI.C.3. Définition des sources de variabilité

Pour définir les sources de variabilité des paramètres technologiques due au procédé de fabrication, nous allons considérer que chaque paramètre suit une distribution gaussienne, définie par :

- Sa valeur moyenne M , égale à la valeur ciblée du paramètre en question.
- Son écart type σ , donné par :
 - L'équation : $3\sigma=12\%M$ pour les paramètres technologiques obtenus par gravure (L , t_{si} pour le Trigate uniquement, W , h_{si}).
 - Par la littérature pour les autres paramètres [Weber 08] [GSS-b] [GSS-c].

Le tableau de la Figure VI-17 résume les écarts types obtenus pour les paramètres technologiques que nous avons jugés comme étant les principales sources de variabilité pour les performances SRAM et pour chaque architecture de dispositif. Pour vérifier que nos distributions statistiques de variations sont bien réalistes, nous avons effectué, pour chaque architecture, 500 simulations avec génération aléatoire des paramètres technologiques, suivant les distributions gaussiennes prédéfinies. Ce nombre de simulations est un bon compromis, car il est suffisamment petit pour permettre un temps de simulations et de traitement des données acceptable et suffisamment grand pour fournir des résultats statistiques représentatifs. Nous avons ensuite extrait dans chaque cas l'écart type des distributions de tension de seuil obtenues (avant dernière ligne de la Figure VI-17). Pour l'architecture conventionnelle sur substrat massif et l'architecture Trigate, ces valeurs sont en lignes avec les valeurs obtenues par simulations numériques atomistiques reportées par [GSS-b] et [GSS-c] (dernière ligne de la Figure VI-17).

	BULK	FDSOI	FinFET
σ_L (nm)	0.8	0.8	0.8
σ_{ϕ_m} (mV)	20	20	20
σ_{tsi} (nm)	/	0.2[1]	0.4
$\sigma_{h_{si}}$ (nm)	/	/	1
σ_w (nm)	1.6	1.6	1.6
σ_{VT} (mV)	68	30	34
MASTAR VA			
σ_{VT} (mV)	67 [2]	/	32 [3]
Simulations			

Figure VI-17 : Tableau résumant les écarts types des paramètres technologiques dont nous avons considéré des variations aléatoires.

Les distributions de tension de seuil pour chaque architecture sont représentées sur la Figure VI-18. On constate déjà que l'architecture conventionnelle sur substrat massif semble être très affectée par la variabilité due au procédé de fabrication, contrairement au FDSOI et au Trigate.

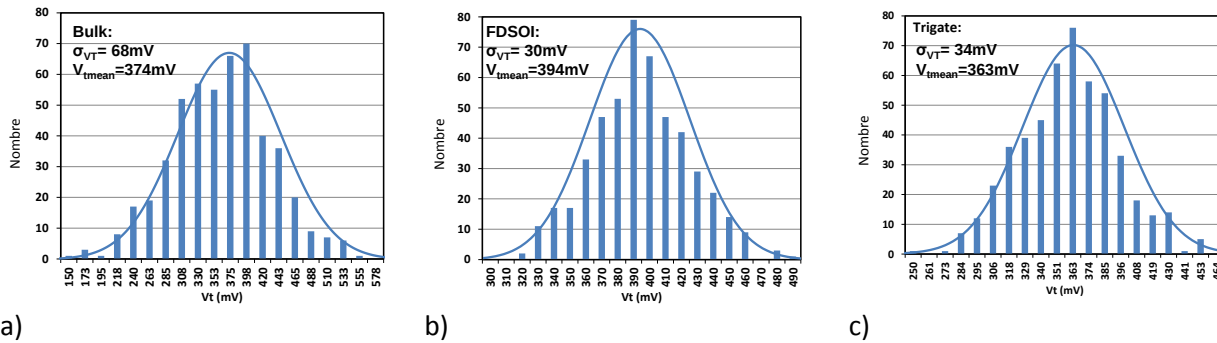


Figure VI-18 : Distributions de tension de seuil obtenues après 500 simulations avec génération pseudo-aléatoire des principaux paramètres technologiques, suivant les distributions gaussiennes prédéfinies. a) architecture conventionnelle sur substrat massif, b) architecture FDSOI et c) architecture Trigate

VI.C.4. Simulations de cellules SRAM, incluant la variabilité

Comme pour extraire l'écart type de la distribution de tension de seuil (Figure VI-18), nous procédons à 500 simulations à la tension d'alimentation nominale du noeud 16nm, soit $V_{dd}=0.8V$, avec génération aléatoire des paramètres technologiques, suivant les distributions gaussiennes prédéfinies afin de tracer les courbes papillons

pour chaque architecture (Figure VI-19), puis d'extraire les écarts type des distributions de SNM (σ_{SNM} , reportée dans le tableau de la Figure VI-20). Ces derniers sont des marqueurs de la robustesse à la variabilité d'une architecture de transistor.

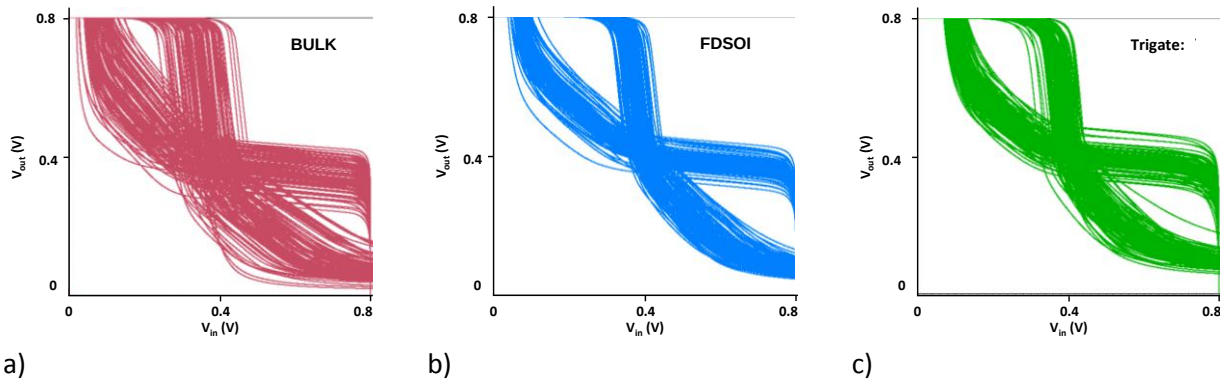


Figure VI-19: Distributions de courbes papillon obtenues à $V_{dd}=0.8V$ après 500 simulations avec génération pseudo-aléatoire des principaux paramètres technologiques, suivant les distributions gaussiennes prédéfinies. a) Architecture conventionnelle sur substrat massif, b) architecture FDSOI et c) architecture Trigate.

D'après les courbes papillons (Figure VI-19), on constate que l'architecture conventionnelle sur substrat massif est très impactée par la variabilité, alors que les architectures FDSOI et Trigate semblent toutes deux présenter des performances correctes et équivalentes. Ces remarques sont confirmées par les valeurs d'écart type de la SNM σ_{SNM} (Figure VI-20) qui sont pratiquement identiques pour le FDSOI et le Trigate, tout en étant deux fois plus faibles que celle de l'architecture conventionnelle sur substrat massif.

	BULK	FDSOI	Trigate
V_{dd} (V)	0.8	0.8	0.8
σ_{Vt} (mV)	68	30	34
A_{VT} (mV. μm)	2.4	1.03	1.18
SNM (mV)	186	182	186
σ_{SNM} (mV)	50	23	26
Area (μm^2)	0.047	0.046	0.049

Figure VI-20 : Résumé des performances SRAM, en incluant la variabilité.

VI.C.5. Extraction de la tension minimale d'alimentation (V_{MIN})

Afin de déterminer si les cellules SRAM décrites ci-dessus sont fonctionnelles, nous allons proposer dans ce dernier paragraphe une méthode pour extraire la tension d'alimentation minimale pour laquelle un réseau de cellules reste fonctionnel (V_{MIN}).

Considérons tout d'abord que les valeurs de SNM suivent une distribution gaussienne, dont la valeur moyenne sera la valeur obtenue sans prise en compte de la variabilité (notée SNM_{mean}), et l'écart type sera de valeur σ_{SNM} (Figure VI-20). La probabilité que la valeur de marge au bruit statique soit égale à SNM est alors donnée par l'équation :

$$P(SNM) = \frac{1}{\sigma_{SNM}\sqrt{2\pi}} e^{-\frac{(SNM-SNM_{mean})^2}{2\sigma_{SNM}^2}} \quad \text{Eq. VI-4}$$

Par conséquent, la probabilité qu'il y ait une erreur dans la cellule, c'est-à-dire $SNM < 0$, est donnée par l'intégrale :

$$P_{\text{bitfail}} = \frac{1}{\sigma_{SNM}\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{(SNM-SNM_{\text{mean}})^2}{2\sigma_{SNM}^2}} dSNM \quad \text{Eq. VI-5}$$

Avec le changement de variable suivant :

$$x = \frac{(SNM-SNM_{\text{mean}})}{\sqrt{2}\sigma_{SNM}}, \quad dx = \frac{1}{\sqrt{2}\sigma_{SNM}} \quad \text{Eq. VI-6}$$

L'intégrale de l'équation VI-5 peut se simplifier comme :

$$P_{\text{bitfail}} = \frac{1}{\sqrt{2}} \int_{-\infty}^{\frac{-SNM_{\text{mean}}}{\sqrt{2}\sigma_{SNM}}} e^{-z^2} dz \quad \text{Eq. VI-7}$$

L'équation VI-7 est intégrable analytiquement par définition de la fonction erreur et est donnée par l'équation :

$$P_{\text{bitfail}} = \frac{1}{2} \left(\text{erf}\left(\frac{-SNM_{\text{mean}}}{\sqrt{2}\sigma_{SNM}}\right) - \text{erf}(-\infty) \right) = \frac{1}{2} \left(1 - \text{erf}\left(\frac{Z}{\sqrt{2}}\right) \right) \quad \text{Eq. VI-8}$$

Où $Z = \sigma_{SNM}/SNM$.

La fonctionnalité du réseau de cellules SRAM est donc liée au paramètre Z défini ci-dessus. Il nous faut désormais estimer sa valeur, à la limite de fonctionnalité du réseau de cellules, et ce en fonction de la taille de ce réseau.

Si nous considérons un réseau de cellules SRAM complètement fonctionnel, le nombre d'erreur N_{bitfail} est donc strictement inférieur à 1. De plus, le nombre d'erreur est donné par le produit du nombre de points mémoire N_{bit} du réseau de cellule SRAM par la probabilité d'avoir une erreur P_{bitfail} . Ceci nous donne l'inéquation :

$$N_{\text{bitfail}} = N_{\text{bits}} P_{\text{bitfail}} < 1 \rightarrow P_{\text{bitfail}} < \frac{1}{N_{\text{bits}}} \quad \text{Eq. VI-9}$$

En remplaçant P_{bitfail} par son expression (Eq. VI-8), et en résolvant l'équation ainsi obtenue à la limite de la fonctionnalité (i.e on remplace le symbole $>$ par $=$ dans Eq. VI-9), on a alors l'expression de Z en fonction de la taille du réseau de cellules SRAM :

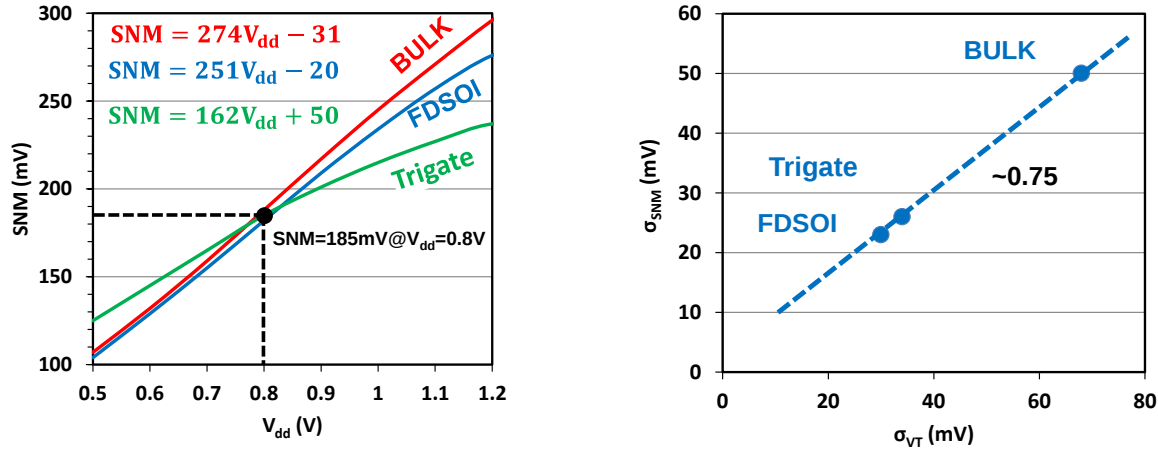
$$Z = \sqrt{2} \text{erf}^{-1}\left(1 - \frac{2}{N_{\text{bits}}}\right) \quad \text{Eq. VI-10}$$

Dans cette étude, nous considérons des réseaux de cellules SRAM de taille $N_{\text{bit}}=100\text{Mbits}$ (valeur typique pour les applications systèmes sur puce), ce qui mène à une valeur de $Z=5.9$.

Enfin, pour aboutir à la valeur de la tension d'alimentation minimale garantissant la fonctionnalité du réseau de cellule SRAM, considérons que la variation de la SNM en fonction de la tension d'alimentation V_{dd} est linéaire. On peut alors écrire :

$$SNM(V_{dd}) = \alpha \times V_{dd} + \beta \quad \text{Eq. VI-11}$$

Où α et β sont extraits des courbes $SNM(V_{dd})$ obtenues par simulations sans prise en compte de la variabilité pour chaque architecture (Figure VI-16, reproduit sur la Figure VI-21-a avec les valeurs de α et β). Cette méthode a déjà été proposée par [Bœuf 11], mais nous procurons ici un meilleur développement. La Figure VI-21-a montre également que la variation de SNM avec la tension d'alimentation V_{dd} est bien linéaire, tant que $V_{dd} < 1V$.



a) b) Figure VI-21 : a) Variation de la SNM en fonction de la tension d'alimentation V_{dd} pour chacun des trois dessins de cellules SRAM. b) Tracé des écarts types de la SNM σ_{SNM} en fonction de ceux de la tension de seuil σ_{VT} démontrant la proportionnalité des deux paramètres.

Ensuite à la limite de fonctionnalité, on a:

- $Z=5.9$
- $V_{dd}=V_{MIN}$

En écrivant la définition de Z à la limite de fonctionnalité, et en remplaçant la valeur moyenne de la SNM SNM_{mean} par son approximation linéaire donnée par l'équation VI-11, on obtient l'équation :

$$\frac{SNM}{\sigma_{SNM}} = \frac{(\alpha V_{MIN} + \beta)}{\sigma_{SNM}} = 5.9 \quad \text{Eq. VI-12}$$

En la résolvant, on obtient l'expression de la tension d'alimentation minimale garantissant la fonctionnalité d'un réseau de cellules SRAM de 100Mbits :

$$V_{MIN} = \frac{(5.9\sigma_{SNM} - \beta)}{\alpha} \quad \text{Eq. VI-13}$$

On peut également exprimer V_{MIN} en fonction de l'écart type de la tension de seuil en considérant que $\sigma_{VT}=0.75\sigma_{SNM}$ (démontré dans notre étude sur la Figure VI-21-b). On a alors :

$$V_{MIN} = \frac{(5.9 \times 0.75 \times \sigma_{VT} - \beta)}{\alpha} \quad \text{Eq. VI-14}$$

L'équation VI-14 permet donc d'évaluer la valeur V_{MIN} d'une cellule SRAM d'une technologie donnée sans effectuer de simulations SRAM, mais en connaissant uniquement l'écart type de la distribution de tension de seuil σ_{VT} de cette technologie. Le tableau de la Figure VI-22 résume les performances SRAM de chaque architecture de transistor. L'extraction du paramètre V_{MIN} démontre deux points :

- L'architecture conventionnelle sur substrat massif présente un V_{MIN} largement supérieur à la tension d'alimentation nominale du nœud 16nm ($V_{MIN}=1.19V > 0.8V=V_{ddnom}$). Elle ne pourra donc pas fournir de SRAM fonctionnelles pour le nœud 16nm du fait d'une sensibilité trop importante à la variabilité.
- L'architecture FDSOI et l'architecture Trigate peuvent fournir des SRAM fonctionnelles pour le nœud 16nm (les V_{MIN} extraits pour chaque architecture sont bien inférieurs à 0.8V). De plus, ces deux architectures présentent une robustesse à la variabilité très proche, avec des valeurs de σ_{SNM} d'une part et de V_{MIN} d'autre part très proches.

	BULK	FDSOI	Trigate
V_{dd} (V)	0.8	0.8	0.8
σ_{V_t} (V)	68	30	34
A_{VT} (mV.μm)	2.4	1.03	1.18
SNM (V)	185.5	182	186
σ_{SNM} (mV)	50	23	26
Area (μm²)	0.047	0.046	0.049
V_{MIN} (V)	1.19	0.62	0.64

Figure VI-22 : Résumé des performances SRAM pour chaque architecture.

VI.D. Conclusion du chapitre

Dans ce dernier chapitre, nous avons utilisé la prédictivité de MASTAR_VA pour évaluer les performances logiques dans une première partie, mais également SRAM dans une seconde partie de l'architecture conventionnelle sur substrat massif, FDSOI et Trigate au nœud 16nm.

Comme dans le chapitre précédent, la première étape a été de définir les paramètres technologiques de chaque dispositif. Ne disposant d'aucune publication présentant des dispositifs du nœud technologique 16nm, nous avons estimé les dimensions de chaque architecture en appliquant globalement la loi de réduction de dimension de Moore sur les architectures définies pour le nœud technologique 20nm, au chapitre V. Les paramètres électrostatiques sont alors déterminés par MASTAR_VA et les paramètres de transport et de résistance d'accès sont extrapolés à partir de ceux du nœud 20nm. Nous avons ensuite effectué des simulations d'anneaux résonnants d'inverseur FanOut 3 avec une charge en sortie fixe comprise entre 0 et 10fF et en faisant varier la tension d'alimentation. Nous avons ainsi démontré que l'architecture FDSOI avec FBB tire à nouveau avantage de sa faible valeur de capacité pour être la plus rapide dans la plupart des configurations. L'architecture Trigate est la plus rapide uniquement lorsque la tension d'alimentation est faible du fait de son faible DIBL et lorsque la charge en sortie est forte car elle écrante le poids des capacités propres à l'architecture. Cependant, la puissance dynamique, et donc la consommation dynamique de l'architecture Trigate est toujours bien plus importante que celle du FDSOI, même avec FBB, à cause des capacités parasites importantes. La comparaison à puissance dynamique constante est à nouveau à l'avantage de l'architecture FDSOI car une tension d'alimentation plus importante doit lui être appliquée pour atteindre la même puissance dynamique que l'architecture Trigate. Enfin, comme pour le nœud précédent, l'architecture conventionnelle sur substrat massif présente des performances bien en retrait, en comparaison des architectures FDSOI et Trigate.

Pour l'estimation de performance SRAM, nous avons défini de nouveaux dispositifs pour chaque architecture car les transistors des parties SRAM et logique d'un circuit ne subissent pas exactement les mêmes procédés de fabrication, donc n'ont pas exactement les mêmes paramètres technologiques ni les mêmes caractéristiques courant-tension. Les transistors de chaque architecture sont alors définis en visant une valeur de compromis I_{on}/I_{off} extrapolée des performances reportées pour le nœud 20nm par [Cho 11]. Les dessins de chaque cellule SRAM sont ensuite déterminés en visant simultanément une surface de cellule typique du nœud 16nm, qui est extrapolée sur l'évolution des cellules SRAM industrielles avec le nœud technologique, et une valeur de marge de bruit statique (SNM Static Noise Margin) à 185mV à la tension d'alimentation nominale du nœud 16nm, soit $V_{dd}=0.8V$, typique des technologies CMOS [Planes 08]. Pour définir les sources de variabilité dues au procédé de fabrication, nous avons utilisé les données de la littérature puis vérifié que MASTAR_VA donnait bien les mêmes distributions de tension de seuil que celles reportées par [GSS-b] et [GSS-c], obtenues par simulations atomistiques. Nous avons ensuite pu effectuer 500 simulations de cellules SRAM en incluant la variabilité et

extrait les distributions de SNM. Nous avons alors démontré que l'architecture conventionnelle était très affectée par la variabilité et présente un écart type de SNM deux fois supérieur à ceux extraits pour l'architecture FDSOI et Trigate. Enfin, nous avons proposé une méthode pour extraire la tension minimale assurant le fonctionnement d'un réseau de cellules SRAM, notée V_{MIN} , et démontré que l'architecture conventionnelle sur substrat massif présente une valeur de V_{MIN} supérieure à la tension d'alimentation nominale ($V_{MIN}=1.19V > 0.8V = V_{dd}$). Cette architecture ne peut donc pas assurer le fonctionnement d'un réseau de cellules SRAM pour le nœud technologique 16nm car elle est trop impactée par la variabilité due au procédé de fabrication. Les valeurs de V_{MIN} extraites pour les architectures FDSOI et Trigate sont quant à elle très proches ($\sim 0.6V$). Elles fournissent donc des cellules SRAM fonctionnelles au nœud 16nm présentant une robustesse à la variabilité similaire.

CONCLUSION GENERALE

Depuis maintenant plus de quatre décennies, le monde de la microélectronique vit dans le rythme effréné de la loi de Moore. Celle-ci dicte une réduction de la densité d'intégration un facteur deux entre chaque nœud technologique de l'élément central des circuits intégrés : le transistor MOSFET. Ceci permet soit de réduire la surface des puces, soit d'accroître la puissance de calcul ou le nombre de fonctionnalité d'une puce en conservant une surface constante. La miniaturisation du transistor MOSFET se traduit par un gain en performance [Dennard 74] et une réduction de la consommation. Jusqu'aux années 2000 environ, ce gain en performance était fourni uniquement par la réduction des dimensions de l'architecture conventionnelle sur substrat massif du transistor MOSFET. Ensuite, la réduction des dimensions du transistor a impliqué la naissance d'effets parasites comme les effets canaux courts liés à la diminution de longueur de grille et des améliorations technologiques ont été nécessaires. Ces dernières ont permis de poursuivre l'évolution dictée par la loi de Moore pendant encore quelques nœuds technologiques. Cependant, à compter du nœud 20nm, les améliorations technologiques mises en œuvre ne sont plus suffisantes pour contrôler notamment les effets canaux courts et il devient nécessaire d'introduire de nouvelles architectures. Cela a été prévu par l'ITRS et confirmé par les publications de différents industriels : cette année (2012), STMicroelectronics a présenté sa plateforme CMOS du nœud 28nm utilisant le transistor FDSOI [Planes 12], alors que Intel a la sienne basée sur l'architecture Trigate pour le nœud technologique 22nm [Auth 12]. N'oublions tout de même pas que l'alliance d'industriels ISDA (principalement IBM, STMicroelectronics et Samsung) a présenté cette année une plateforme CMOS utilisant l'architecture conventionnelle sur substrat massif [Shang 12], qui sera sans doute la dernière. On peut donc remarquer que pour le nœud technologique 20nm, trois architectures de transistor seront en concurrence, ce qui constituera un fait unique dans l'histoire de la microélectronique.

La question qui se pose alors est de savoir quelle architecture sera la plus performante sur le plan applicatif et économique. La performance d'une technologie CMOS ne se réduit plus au compromis I_{on}/I_{off} , ni même au compromis I_{eff}/I_{off} mais à la performance circuit, traduite par les métriques fréquence, puissance dynamique et statique. Ceci constitue l'objectif de cette thèse dans laquelle nous avons proposé une méthodologie permettant, à partir des paramètres technologiques de cette architecture, d'évaluer la performance circuit d'une architecture CMOS par l'intermédiaire d'un simulateur de circuit conventionnel [ELDO].

Au cours du premier chapitre, nous avons décrit le fonctionnement idéal du transistor MOSFET et son architecture conventionnelle sur substrat massif. Nous avons ensuite exposé les différents effets physiques à prendre en compte pour corriger ce modèle simplifié afin de tendre vers le fonctionnement réaliste du transistor MOSFET. Nous avons alors défini puis démontré les principaux paramètres et équations régissant son fonctionnement dans chaque régime. Nous avons par la suite décrit l'effet de la réduction des dimensions sur le comportement électrique du transistor puis les solutions technologiques apportées pour limiter l'impact de ces effets parasites sur la performance. Malgré ces diverses améliorations technologiques l'architecture conventionnelle sur substrat massif atteint ses limites de terme de performance et l'introduction de nouvelles architectures devient indispensable. La description de ces nouvelles architectures de transistor MOSFET a fait l'objet de la dernière partie de ce premier chapitre.

Dans un second chapitre, nous avons développé des modèles analytiques afin d'évaluer les caractéristiques courant-tension des architectures conventionnelles sur substrat massif, FDSOI et double grille faiblement dopé. Les régimes sous le seuil sont décrits à partir d'une modélisation précise et prédictive de la tension de seuil, de l'effet canal court SCE, du DIBL et de la pente sous le seuil, en tenant compte de la spécificité de chaque technologie. L'effet de la longueur de grille a été modélisé par l'utilisation de la transformation tension-dopage VDT [Skotnicki 88-a]. Celle-ci a été suffisante pour l'architecture conventionnelle sur substrat massif et a dû être

adaptée pour les architectures double grille et FDSOI. Pour compléter le modèle de tension de seuil du transistor FDSOI et tenir compte de l'effet de l'épaisseur de l'oxyde enterré, nous avons adapté les travaux de [Ernst 07]. Chaque modèle est validé par simulations numériques 2D [synopsis] et par caractérisations électriques quand celles-ci étaient disponibles.

Le régime au-delà du seuil est décrit par une modélisation classique du courant de dérive où la mobilité effective est estimée par les lois universelles de [Takagi 94]. L'effet d'amélioration du transport par contrainte mécanique est pris en compte avec les travaux de [Payet 08] et la dégradation de mobilité effective due à la longueur de grille est prise en compte par le simple modèle de [Bidal 09].

Enfin, la continuité entre chaque régime des caractéristiques courant-tension (I_d - V_g et I_d - V_d) et de leurs dérivées (g_m - V_g et g_d - V_d), indispensable pour la compatibilité avec les outils de CAO classique est assurée par les expressions de tensions effectives.

Les caractéristiques courant-tension (I_d - V_g et I_d - V_d) et leurs dérivées (g_m - V_g et g_d - V_d) sont finalement validés par caractérisations électriques pour chaque architecture de transistor, démontrant la pertinence et l'efficacité du modèle proposé.

Dans notre troisième chapitre, nous avons proposé un modèle complètement analytique de toutes les capacités parasites, de la zone active du transistor jusqu'au premier niveau de métallisation des interconnexions pour les principales architectures CMOS, à savoir l'architecture conventionnelle sur substrat massif, le FDSOI, le double grille planaire, le FinFET et le Trigate. Nous avons alors commencé par mettre en place une méthode générique d'évaluation de capacités, en distinguant les deux types de capacité rencontrés dans les architectures CMOS :

- Les capacités dont les électrodes sont parallèles, estimées par la formule classique du condensateur plan.
- Les capacités dont les électrodes sont perpendiculaires, estimées par la transformation de Schwarz-Christoffel [Plonsey 61].

Les expressions de chaque capacité parasite de chaque structure ont été obtenues par cette méthode. Afin d'être plus représentatif de la réalité des technologies CMOS et de leur évolution, les modèles ont été raffinés pour tenir compte :

- Des deux schémas de contact concurrents, le contact à prise et le contact ruban.
- Des source-drains réalisés par une épitaxie présentant une facette.
- Des espaceurs à plusieurs couches.

Nous avons ensuite utilisé notre modèle pour démontrer que l'ITRS sous-estime le poids des capacités parasites. Enfin, nous avons comparé le poids des parasites sur deux architectures de double grille : planaire et FinFET et démontré que le FinFET est plus affecté par les parasites que le double grille planaire.

Au cours du quatrième chapitre, nous avons utilisé les modèles développés dans les second et troisième chapitres afin d'estimer la vitesse d'une chaîne d'inverseurs. Nous avons d'abord décrit une méthode itérative et numérique qui permet l'estimation du délai d'une chaîne d'inverseur pour une capacité équivalente de 1fF à partir des caractéristiques courant-tension d'un couple NMOS/PMOS. Cette méthode est plus précise que la méthode ' CV/I_{eff} ' et reste suffisamment rapide (quelques secondes de calcul). Nous avons ensuite explicité la méthode d'estimation de la capacité équivalente d'un inverseur [Wei 11]. Sa valeur est obtenue à partir des expressions des capacités parasites données dans le chapitre III. Nous avons alors mené deux études comparatives bien distinctes. La première a pour objectif d'estimer la performance circuit dans le cadre du nœud technologique 20nm, obtenue par co-intégration de dispositif IFQW III-V/Ge [Hellings 10] et de la comparer à des références silicium (transistor conventionnel sur substrat massif et FDSOI). Nous avons ainsi montré que la co-intégration de dispositifs IFQW III-V/Ge pouvait permettre des délais de propagation d'inverseur très courts, mais ce délai est très impacté par la variabilité du procédé de fabrication, notamment celle de l'épaisseur de

l'espaceur. Enfin, ce type de dispositif présente un niveau de fuite trop élevé pour les applications systèmes sur puce.

Dans la seconde étude, nous avons montré qu'un inverseur construit par intégration 3D monolithique [Batude 10] pouvait être plus rapide que la référence classique planaire au nœud 28nm. Nous avons dû au préalable évaluer les nouvelles composantes de capacités parasites spécifiques à l'intégration 3D. Nous avons également intégré la dépendance dynamique de la tension de seuil du transistor de l'étage supérieur avec la polarisation de grille. L'impact de cet effet sur la fuite de l'inverseur n'a pas été estimé, mais devrait l'être dans une étude plus approfondie, tout comme l'effet de la réduction de la longueur des lignes métalliques d'interconnexion. Il serait alors plus pertinent d'adapter un outil de CAO conventionnel.

Dans ce cinquième chapitre, nous avons évalué puis comparé les performances logiques des trois architectures CMOS concurrentes pour le nœud technologique 20nm: l'architecture conventionnelle sur substrat massif, le FDSOI et le Trigate par l'utilisation d'un simulateur de circuit conventionnel [ELDO]. Nous avons adapté et complété les modèles des chapitre II et III afin d'être intégrables dans un simulateur de circuit classique [ELDO]. La continuité indispensable des caractéristiques courant-tension et de leurs dérivées est assurée par le modèle du second chapitre et le modèle de charge est dérivé à partir de ce modèle de courant drain-source. Les expressions de capacités parasites du troisième chapitre donnent les valeurs des charges extrinsèques. Ceci fournit un modèle compact, que nous nommons MASTAR_VA, qui est ensuite implémenté en VerilogA et peut être utilisé dans un simulateur de circuit comme [ELDO]. Les performances statiques et dynamiques des circuits logiques, les performances de cellules SRAM et la robustesse à la variabilité des architectures CMOS peuvent alors être évaluées.

Pour l'évaluation des performances logiques au nœud technologique 20nm, nous avons défini chaque architecture à partir de l'état de l'art pour les performances des transistors (I_{on}/I_{off} et I_{eff}/I_{off}) mais aussi pour fixer les paramètres technologiques. Nous avons ensuite effectué des simulations d'anneaux résonnants d'inverseur de FanOut 1 puis FanOut 3 avec une charge fixe en sortie comprise entre 0 et 10fF et en faisant varier la tension d'alimentation V_{dd} . Nous avons alors démontré que l'architecture conventionnelle sur substrat massif présente des performances en retrait. Nous avons également montré que l'architecture Trigate avec la longueur de grille relâchée est la plus rapide pour les faibles tensions d'alimentation grâce à son faible DIBL. Mais, lorsque la tension d'alimentation augmente, le faible DIBL devient un inconvénient pour la performance et cette architecture est de moins en moins compétitive. De plus, elle est affectée par un niveau de capacité supérieure à toutes les autres architectures ce qui est pénalisant pour les faibles valeurs de charge de sortie. Cette pénalité est réduite lorsque la charge en sortie augmente car le poids des capacités parasites intrinsèques au transistor est écranté et la vitesse est davantage liée au niveau de courant débité. Ceci explique pourquoi l'architecture FDSOI est plus rapide que les deux architectures Trigate quand la charge de sortie est faible et la tension d'alimentation est élevée. Avec l'utilisation du FBB, l'architecture FDSOI est la plus rapide, sauf pour les charges de sortie très élevées ($> \sim 5\text{fF}$) car les architectures Trigate tirent avantage du niveau de courant débité supérieur au FDSOI, même avec FBB, grâce à leurs largeurs électriques supérieures. Le gain en fréquence apporté par l'utilisation du FBB, $\sim 15\%$, se paie par une augmentation de fuite statique de l'inverseur d'un facteur 5 (limite maximum car le FBB n'est utilisé que lorsque le système a besoin de performance) quelle que soit la tension d'alimentation V_{dd} . En comparant les vitesses d'inverseur non plus à tension d'alimentation constante, mais à puissance dynamique constante, nous avons constaté que les écarts vus en faveur du FDSOI (avec et sans FBB) comparés aux architecture Trigate à même tension d'alimentation sont exacerbés. Ceci était attendu car le FDSOI présente de faibles valeurs de capacités, donc pour que le FDSOI atteigne une même puissance dynamique que les architectures Trigate, il faut lui appliquer une tension d'alimentation supérieure, ce qui procure un gain supplémentaire en fréquence. Cette comparaison est réaliste et justifiée car les circuits sont aujourd'hui conçus pour être utilisés à leur maximum de performance, et ce maximum est souvent fixé par la dissipation thermique limite du boîtier. La tension d'alimentation est alors fixée pour atteindre cette limite.

Enfin, n'oublions pas les quelques limitations de notre méthodologie: MASTAR_VA ne permet pas de tenir compte des effets de proximité (par exemple la contrainte induite par le transistor voisin qui peut modifier la mobilité), ni des spécificités de chaque architecture concernant les interconnexions métalliques.

Dans ce sixième et dernier chapitre, nous avons utilisé la prédictivité de MASTAR_VA pour évaluer les performances logiques puis SRAM de l'architecture conventionnelle sur substrat massif, FDSOI et Trigate au nœud 16nm. Comme dans le chapitre précédent, la première étape a été de définir les paramètres technologiques de chaque dispositif. Celles-ci ont été fixées en appliquant globalement la loi de réduction de dimension de Moore sur les architectures définies pour le nœud technologique 20nm (chapitre V). Les paramètres électrostatiques sont alors déterminés par MASTAR_VA et les paramètres de transport et de résistance d'accès sont extrapolés à partir de ceux du nœud 20nm. Nous avons ensuite effectué des simulations d'anneaux résonnants d'inverseur FanOut 3 avec une charge en sortie fixe comprise entre 0 et 10fF et en faisant varier la tension d'alimentation. Nous avons ainsi démontré que l'architecture FDSOI avec FBB tire à nouveau avantage de sa faible valeur de capacité pour être la plus rapide dans la plupart des configurations. L'architecture Trigate est la plus rapide uniquement lorsque la tension d'alimentation est faible du fait de son faible DIBL et lorsque la charge en sortie est forte car elle écrante le poids des capacités propres à l'architecture. Cependant, la puissance dynamique, et donc la consommation dynamique de l'architecture Trigate est toujours bien plus importante que celle du FDSOI, même avec FBB, à cause des capacités parasites importantes. La comparaison à puissance dynamique constante est à nouveau à l'avantage de l'architecture FDSOI car une tension d'alimentation plus importante doit lui être appliquée pour atteindre la même puissance dynamique que l'architecture Trigate. Enfin, comme pour le nœud précédent, l'architecture conventionnelle sur substrat massif présente des performances bien en retrait, en comparaison des architectures FDSOI et Trigate.

Pour l'estimation de performance SRAM, nous avons défini de nouveaux dispositifs pour chaque architecture car les transistors des parties SRAM et logique d'un circuit ne subissent pas exactement les mêmes procédés de fabrication, donc n'ont ni les mêmes paramètres technologiques ni les mêmes caractéristiques courant-tension. Les transistors de chaque architecture sont alors définis en visant une valeur de compromis I_{on}/I_{off} extrapolée des performances reportées pour le nœud 20nm par [Cho 11]. Les dessins de chaque cellule SRAM sont ensuite déterminés en visant simultanément une surface de cellule typique du nœud 16nm et une valeur de marge de bruit statique (SNM Static Noise Margin) à 185mV à la tension d'alimentation nominale du nœud 16nm, soit $V_{dd}=0.8V$. Pour définir les sources de variabilité dues au procédé de fabrication, nous avons utilisé les données de la littérature puis vérifié que MASTAR_VA donnait bien les mêmes distributions de tension de seuil que celles reportées par [GSS-b] et [GSS-c], obtenues par simulations atomistiques. Nous avons ensuite effectué 500 simulations de cellules SRAM en incluant la variabilité et extrait les distributions de SNM. Nous avons alors démontré que l'architecture conventionnelle était très affectée par la variabilité et présente un écart type de SNM deux fois supérieur à ceux extraits pour l'architecture FDSOI et Trigate. Enfin, nous avons proposé une méthode pour extraire la tension minimale assurant le fonctionnement d'un réseau de cellules SRAM, notée V_{MIN} . Nous avons ainsi démontré que l'architecture conventionnelle sur substrat massif présente une valeur de V_{MIN} supérieure à la tension d'alimentation nominale ($V_{MIN}=1.19V>0.8V=V_{dd}$) et qu'elle ne peut donc pas assurer le fonctionnement d'un réseau de cellules SRAM pour le nœud technologique 16nm du fait d'une trop grande sensibilité à la variabilité due au procédé de fabrication. Les valeurs de V_{MIN} extraites pour les architectures FDSOI et Trigate sont quant à elle très proches ($\sim 0.6V$). Elles fournissent donc des cellules SRAM fonctionnelles au nœud 16nm présentant une robustesse à la variabilité similaire.

Afin de poursuivre la miniaturisation après le nœud 20nm, la question du choix de l'architecture sera primordiale. Nous savons d'ores et déjà que l'architecture conventionnelle ne sera plus utilisée mais la question reste ouverte pour l'architecture FDSOI ou l'architecture Trigate.

Le passage à l'architecture FDSOI est une rupture technologique, mais une rupture que l'on peut qualifier de légère. En effet, si on fait omission du changement de substrat, un transistor FDSOI est très semblable à un

transistor conventionnel sur substrat massif et beaucoup de briques technologiques du procédé de fabrication sont similaires. De plus, la manière de concevoir un circuit à base de transistors FDSOI est pratiquement identique à celle d'un circuit à base de transistors conventionnels sur substrat massif, notamment par la continuité de la largeur électrique (égale à la largeur des zones actives) pour ces deux types de dispositifs. Cette remarque est également valable pour la définition des largeurs des dispositifs des cellules SRAM qui peuvent être ajustées au nanomètre prêt. Enfin, le décalage de tension de seuil provoquée par la polarisation de la face arrière d'un transistor FDSOI, on parle de back-biasing, est un dernier avantage de cette architecture. En l'utilisant de manière statique, cela permet d'ajuster la tension de seuil en fonction de l'application visée (i.e le courant de fuite I_{off}) s'il n'existe pas de solution technologique. Utiliser de manière dynamique, cela permet d'abaisser la tension de seuil et donc d'augmenter la vitesse quand le système nécessite un niveau élevé de performance ou d'augmenter la tension de seuil et donc de réduire la consommation du circuit.

Le passage à l'architecture Trigate est une rupture technologique plus violente car on rompt avec le modèle planaire utilisé jusqu'à présent en passant à une technologie 3D. Ceci implique l'existence de nouveaux challenges technologiques comme la formation du fin, mais également une nouvelle façon d'appréhender la conception de circuits logiques et SRAM du fait de la discrétisation de la largeur électrique. Cette architecture permet un meilleur contrôle électrostatique que le FDSOI et une plus grande largeur électrique pour un même encombrement qu'une technologie planaire en général, dont le FDSOI. Contrairement au FDSOI, les techniques de back-biasing ne sont pas efficaces pour le Trigate. Enfin, le choix du substrat reste une question ouverte pour l'architecture Trigate: l'utilisation d'un substrat SOI est plus onéreuse qu'un substrat massif mais permet la réduction des fuites et de conserver un fin non dopé.

Par la méthodologie décrite dans ce manuscrit, nous avons démontré que pour les nœuds 20 et 16nm, l'architecture Trigate est très impactée par ses capacités parasites et présente des performances légèrement en retrait par rapport à l'architecture FDSOI pour les circuits faiblement chargés, donc pour les applications type système sur puce. Pour les circuits plus chargés, davantage représentatif des circuits type microprocesseurs, le niveau de courant plus élevé débité par l'architecture Trigate, notamment grâce au développement de la largeur électrique, lui donne l'avantage sur l'architecture FDSOI.

MASTAR_VA ne permet cependant pas de tenir compte des spécificités de chaque architecture concernant les niveaux métalliques des interconnexions. Les effets de proximité d'un transistor avec ses voisins sont difficiles à prévoir sans caractérisation électrique et ne sont également pas pris en compte par MASTAR_VA.

La feuille de route ITRS et la course à la miniaturisation ne s'arrêteront pas pour le nœud 16nm. En deçà, la question de l'architecture restera ouverte. On peut se demander si l'architecture FDSOI garantira toujours un contrôle électrostatique suffisant, mais également si l'architecture Trigate ne sera pas trop impactée par les capacités parasites. Si les transistors FDSOI sont trop impactés par les effets canaux courts, on peut envisager l'introduction de l'architecture double grille planaire qui permettra d'améliorer le contrôle électrostatique tout en conservant une technologie planaire. La réduction des capacités parasites sur l'architecture Trigate pourra se faire au travers d'un passage au matériau faible permittivité pour les espaceurs, ou par un travail sur le procédé de fabrication afin de réduire les dimensions du Trigate liées aux capacités parasites, comme la hauteur de grille, mais surtout le Finpitch.

La poursuite de la miniaturisation des technologies CMOS après le nœud 7nm verra sans doute l'introduction de matériaux haute mobilité pour garantir une amélioration des courants caractéristiques I_{on} et I_{eff} par l'intermédiaire d'architecture 'OI' (On Insulator en anglais pour sur isolant) ou d'architecture Trigate afin de conserver un contrôle électrostatique de la grille sur le canal suffisant. Il sera donc intéressant d'adapter MASTAR_VA afin de pouvoir évaluer les performances de circuits basés sur une technologie non-silicium. La prédictivité de MASTAR_VA ne pourra être assurée que par une modélisation précise des paramètres électrostatiques associée à une bonne évaluation analytique des paramètres de transport. Ces deux aspects devront naturellement tenir compte des spécificités de chaque matériau haute-mobilité, comme la modification de la permittivité, l'augmentation du darkspace puis les changements de masse effective et de structures de bandes.

Une solution, en rupture avec la méthode habituelle de l'industrie de la microélectronique d'assurer la croissance de la densité d'intégration des transistors, serait de commencer à tirer avantage de la troisième dimension avec, par exemple, l'intégration 3D monolithique. Les transistors pourraient alors être répartis sur plusieurs niveaux, relâchant la contrainte sur leurs dimensions comme la longueur de grille.

Satisfaire Gordon Moore et sa fameuse loi semble possible pour encore quelques nœuds technologiques, avec l'introduction de nouvelles architectures et de nouveaux matériaux. Mais la question qui, même avec MASTAR VA, reste en suspens est « jusqu'à quand ? ».

BIBLIOGRAPHIE

- [Ang 07] K.-W. Ang, K.-J. Chui, C.-H. Tung, G. Samudra, N. Balasubramanian and Y.-C. Yeo "Enhanced Carrier Transport in Strained Bulk N-MOSFETs with Silicon-Carbon Source/Drain Stressors" IEEE VLSI-TSA conference proceedings 2007.
- [Antoniadis01] D. A. Antoniadis, I. J. Djomehri, and A. Lochtefeld "Electron velocity in sub-50-nm channel mosfets". In IEEE International Conference on Simulation of Semiconductor Processes and Devices 2001, pages 156 – 161.
- [Arnaud 04] F. Arnaud, B.Durie, B.Tavel, L.Pain, J.Todeachini, M.Jurdi, Y.Laplanche, F.Boeuf, F.Salvettio, D.Lenable, J.P.Reynard, F.Wacquan, P.Morin, N.Emonet, D.Barge, M.Bidaud, D.Ceccarelli, P.Vannier, Y.Loque, H.Leninger, F.Judong, C.Perrot, I.Guilmesu, R.Palla, A.Beverina, V.DeJonghe, M.Broekaart, V.Vachellerie, R.A.Bianchi, B.Borot, T.Devoivre, N.Bicair, D.Roy, M.Denuis, K.Rochereau, R.Difrenra, N.Planes, H.Brut, L.Vishnobulta "Low Cost 65nm CMOS Platform for Low Power & General Purpose Applications" in VLSI Symp. Tech. Dig., 2004, pp. 10–11.
- [Arnaud 09] F.Arnaud, A.Thean, M.Eller, M.Lipinski, Y.W.Teh, M.Ostermayr, K.Kang, N.S.Kim, K.Ohuchi, J-P.Han, D.R.Nair, J.Lian, S.Uchimura, S.Kohler, S.Miyaki, P.Ferreira, J-H. Park, M.Hamaguchi, K.Miyashita, R.Augur, Q.Zhang, K.Strahrenberg, S.ElGhouli, J.Bonnouvrier, F.Matsuoka, R.Lindsay, J.Sudijono, F.S.Johnson, J.H.Ku, M.Sekine, A.Steege, R.Sampson "Competitive and Cost Effective high-k based 28nm CMOS Technology for Low Power Applications" IEDM Tech. Dig., pp.651-654, 2009.
- [Arnaud 11] F.Arnaud, S.Colquhoun, A.L.Mareau, S.Kohler, S.Jeannot, F.Hasbani, R. Paulin, S.Cremer, C.Charbuillet, G.Druais, P.Scheer "Technology-Circuit Convergence for Full-SOC Platform in 28 nm and Beyond" IEDM Tech. Dig., pp.374-377, 2011.
- [Arora 93] N. Arora "MOSFET Models for VLSI circuit Simulation – Theory and Practice ", Springer-Verlag Wien New York, 1993.
- [Auth 08] C. Auth, A. Cappellani, J.-S. Chun, A. Dalis, A. Davis, T. Ghani, G. Glass, T. Glassman, M. Harper, M. Hattendorf, P. Hentges, S. Jaloviar, S. Joshi, J. Klaus, K. Kuhn, D. Lavric, M. Lu, H. Mariappan, K. Mistry, B. Norris, N. Rahhal-orabi, P. Ranade, J. Sandford, L. Shifren, V. Souw, K. Tone, F. Tambwe, A. Thompson, D. Towner, T. Troeger, P. Vandervoorn, C. Wallace, J. Wiedemer, C. Wiegand "45nm High-k + Metal Gate Strain-Enhanced Transistors" in VLSI Symp. Tech. Dig., 2008, pp. 128–129.
- [Auth 12] C. Auth, C. Allen, A. Blattner, D. Bergstrom, M. Brazier, M. Bost, M. Buehler, V. Chikarmane, T. Ghani, T. Glassman, R. Grover, W. Han, D. Hanken, M. Hattendorf, P. Hentges, R. Heussner, J. Hicks, D. Ingerly, P. Jain, S. Jaloviar, R. James, D. Jones, J. Jopling, S. Joshi, C. Kenyon, H. Liu, R. McFadden, B. McIntyre, J. Neiryneck, C. Parker, L. Pipes, I. Post, S. Pradhan, M. Prince, S. Ramey, T. Reynolds, J. Roesler, J. Sandford, J. Seiple, P.

- Smith, C. Thomas, D. Towner, T. Troeger, C. Weber, P. Yashar, K. Zawadzki and K. Mistry "A 22nm High Performance and Low-Power CMOS Technology Featuring Fully-Depleted Tri-Gate Transistors, Self-Aligned Contacts and High Density MIM Capacitors" VLSI Tech. Dig., 131-132, 2012.
- [Bacarani 99] G. Bacarani, S. Reggiani "A Compact Double-Gate MOSFET Model Comprising Quantum-Mechanical and Nonstatic Effects" in IEEE Trans. Electron Devices, vol. 46, no. 8, Aug. 1999 pp. 1656-1666.
- [Balestra 87] F. Balestra, S. Cristoloveanu et al., "Double gate silicon on insulator transistor with volume inversion: a new device with greatly enhanced performance", Electron Devices Letter, Vol.8, N°9, pp. 410, 1987.
- [Banerjee 01] Banerjee, S.-J. Souri, P. Kapur, K.-C. Saraswat, "3-D ICs: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," Proceedings of the IEEE, vol.89, no.5, pp.602-633, May 2001.
- [Bansal 05] A. Bansal, B. C. Paul, and K. Roy, "Modeling and optimization of fringe capacitance of nanoscale DGMOS devices," IEEE Trans. Electron Devices, vol. 52, no. 2, pp. 256–262, Feb. 2005.
- [Batail 09] E. Batail "Etude et Intégration de Dispositifs MOS à canal de conduction haute mobilité obtenus en technologie Silicon-On-Nothing / Germanium-On-Nothing » Thèse de doctorat, de l'institut polytechnique de grenoble, soutenue publiquement en 2009.
- [Batude 08] P. Batude, et al. "Enabling 3D monolithic integration"; Proceedings of the Electro-Chemical Society (ECS) spring meeting VOL 16 p 47 2008.
- [Batude 09-a] P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, J.-M. Hartmann, L. Sanchez, L. Baud, V. Carron, A. Toffoli, F. Allain, V. Mazzocchi, D. Lafond, O. Thomas, O. Cueto, N. Bouzaida, D. Fleury, A. Amara, S. Deleonibus and O. Faynot. "Advances in 3D CMOS Sequential Integration" in IEDM Tech. Dig., 2009, pp: 345–348.
- [Batude 09-b] P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, L. Clavelier, S. Michaud, A. Valentian, O. Thomas, O. Rozeau, P. Coudrain, C. Leyris, K. Romanjek, X. Garros, L. Sanchez, L. Baud, A. Roman, V. Carron, H. Grampeix, E. Augendre, A. Toffoli, F. Allain, P. Grosgeorges, V. Mazzochi, L. Tosti, F. Andrieu, J.-M. Hartmann, D. Lafond, S. Deleonibus and O. Faynot « GeOI and SOI 3D monolithic cell integrations for high density applications » VLSI symp Tech. Dig., 2009 p166-167.
- [Batude 11] P. Batude, M. Vinet, C. Xu, B. Previtali, C. Tabone, C. Le Royer, L. Sanchez, L. Baud, L. Brunet, A. Toffoli, F. Allain, D. Lafond, F. Aussenac, O. Thomas, T. Poiroux and O. Faynot. "Demonstration of low temperature 3D sequential FDSOI integration down to 50 nm gate length" VLSI symp Tech. Dig., 2011 p158-159.

- [Bidal 09] G. Bidal, D. Fleury, G. Ghibaudo, F. Boeuf et T. Skotnicki "Guidelines for MOSFET Device optimization accounting for L-dependent mobility degradation" In Silicon Nano Workshop 2009.
- [Boeuf 04] F.Boeuf, F.Amaud, B.Tavel, B. Duriez, M.Bidaud, P. Gouraud, C. Chaton, P. Morin, J. Todeschini, M. Jurdit, L. Pain, V. De-Jonghe, M.T. Basso, D. Sotta, F. Wacquant, J. Rosa, R. El-Farhane ,S. Jullian, "A Conventional 45nm CMOS node Low-Cost Platform for General Purpose and Low Power Applications" IEDM Tech. Dig., pp.425-428, 2004.
- [Boeuf 08] F. Boeuf, M. Sellier, A. Farcy, and T. Skotnicki « An Evaluation of the CMOS Technology Roadmap From the Point of View of Variability, Interconnects, and Power Dissipation » in IEEE Trans. Electron Devices, vol. 55, no6, june 2008, pp 1433-1440.
- [Boeuf 09] F. Boeuf, G. Ghibaudo et T.Skotnicki "Impact of Coulomb Scattering on the Characteristics of Nanoscale Devices" dans International Conference on SSDM 2009.
- [Boeuf 11] F. Boeuf "FDSOI, an innovative technology for low-Vdd / high-performance logic" in a shortcourse given in ISSCC conference 2011.
- [BSIM] BSIM manual, <http://www-device.eecs.berkeley.edu/bsim/?page=BSIM4>.
- [Chan 11] K. Chan, B. Benbakhti, C. Riddet, J. Watling and A. Asenov "Monte Carlo Simulation of a 20 nm Gate Length Implant Free Quantum Well Ge pMOSFET with Different Lateral Spacer width" in proceedings of ULIS conference 2011.
- [Chen 03] Q. Chen; E.M. Harrell and J.D. Meindl "A physical short-channel threshold voltage model for undoped symmetric double-gate MOSFETs " in IEEE Trans. Electron Devices, vol. 50, no7, jul. 2003, pp 1631-1637.
- [Chen 08] X. Chen, S. Samavedam, V. Narayanan, K. Stein, C. Hobbs, C. Baiocco, W. Li, D. Jaeger, M. Zaleski, H. S. Yang, N. Kim, Y. Lee, D. Zhang, L. Kang, J. Chen H. Zhuang, A. Sheikh, J. Wallner, M. Aquilino, J. Han, Z. Jin, J. Li, G. Massey, S. Kalpat, R. Jha, N. Moumen, R. Mo, S. Kirshnan, X. Wang, M. Chudzik, M. Chowdhury, D. Nair, C. Reddy, Y. W. Teh, C. Kothandaraman, D. Coolbaugh, S. Pandey, D. Tekleab, A. Thean, M. Sherony, C. Lage, J. Sudijono, R. Lindsay, J. H. Ku, M. Khare, A. Steegen "A Cost Effective 32nm High-K/ Metal Gate CMOS Technology for Low Power Applications with Single-Metal/Gate-First Process" in VLSI Symp. Tech. Dig., 2008, pp. 88–89.
- [Cheng 11] K. Cheng, A. Khakifirooz, P. Kulkarni, S. Ponothe, B. Haran, A. Kumar, T. Adam, A. Reznicek, N. Loubet, H. He, J. Kuss, M. Wang, T. M. Levin, F. Monsieur, Q. Liu, R. Sreenivasan, J. Cai, A. Kimball, S. Mehta, S. Luning, Y. Zhu, Z. Zhu, T. Yamamoto, A. Bryant, C. –H. Lin, S. Naczas, H. Jagannathan, L. F. Edge, S. Allegret-Maret, A. Dube, S. Kanakasabapathy, S. Schmitz, A. Inada, S. Seo, M. Raymond, Z. Zhang, A. Yagishita, J. Demarest, J. Li, M. Hopstaken, N. Berliner, A. Upham, R. Johnson, S. Holmes, T. Standaert, M. Smalley, N. Zamdmer, Z. Ren, T. Wu, H. Bu, V. Paruchuri, D. Sadana, V. Narayanan, W. Haensch, J. O'Neill, T. Hook, M. Khare, B. Doris « ETSOI CMOS for System-on-Chip Applications Featuring 22nm Gate Length, Sub-100nm Gate Pitch, and 0.08 μ m² SRAM Cell » in VLSI Symp. Tech. Dig., 2011, pp. 128–129.

- [Chipworks] "Intel's 22-nm Tri-gate Transistors Exposed" accessible en ligne via <http://www.chipworks.com/en/technical-competitive-analysis/resources/technology-blog/2012/04/intel%e2%80%99s-22-nm-tri-gate-transistors-exposed/>
- [Cho 11] H.-J. Cho, K.-I. Seo, W.C. Jeong, Y.-H. Kim, Y.D. Lim, W.W. Jang, J.G. Hong, S.D. Suk, M. Li, C. Ryou, H.S. Rhee, J.G. Lee, H.S. Kang, Y.S. Son, C.L. Cheng, S.H. Hong, W.S. Yang, S.W. Nam, J.H. Ahn, D.H. Lee, S. Park, M. Sadaaki, D.H. Cha, D.W. Kim, S.P. Sim, S. Hyun, C.G. Koh, B.C. Lee, S.G. Lee, M.C. Kim, Y.K. Bae, B. Yoon, S.B. Kang, J.S. Hong, S. Choi, D.K. Sohn, J. S. Yoon and C. Chung "Bulk Planar 20nm High-K/Metal Gate CMOS Technology Platform for Low Power and High Performance Applications" ", IEDM Tech. Dig., pp.350,353 2011.
- [Choi 11] H. Choi, J. Lee and W. Sung "Memory Access Pattern-Aware DRAM Performance Model for Multi-Core Systems" in ISPASS conference proceedings, pp. 66-75, 2011.
- [Colinge 90] J.P. Colinge, M.H. Gao, A. Romano-Rodriguez, H. Maes, C. Claeys, "Silicon-on-insulator gate all around device", IEDM Tech. Dig., pp.595, 1990.
- [Coquand 12] R. Coquand, M. Cassé S. Barraud, P. Leroux, D. Cooper, C. Vizios, C. Comboroure, P. Perreau, V. Maffini-Alvaro, C. Tabone, L. Tosti, F. Allain, S. Barnola, V. Delaye, F. Aussenac, G. Reimbold, G. Ghibaudo, D. Munteanu, S. Monfray, F. Boeuf, O. Faynot, and T. Poiroux "Strain-Induced Performance Enhancement of Tri-Gate and Omega-Gate Nanowire FETs Scaled Down to 10nm Width" VLSI Tech. Dig., 13-14, 2012.
- [Dennard 74] R.H.Dennard, F. H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassous et A. R. LeBlanc "Design of ion-implanted mosfet's with very small physical dimensions." IEEE Journal of Solid-State Circuits, sc-9(5):256 – 268, 1974.
- [Dewey 12] G. Dewey, M. Radosavljevic, and N. Mukherjee "III-V Quantum Well Field Effect Transistors on Silicon for Future High Performance and Low Power Logic Applications" ", IEDM Tech. Dig., pp.714-718, 2011.
- [Duallogic D4.2] "Device simulations of III-V and Ge MOSFETs", Deliverable D4.2 du projet Duallogic.
- [Diouf 11] C. Diouf, A. Cros, S. Monfray, J. Mitard, J. Rosa, F. Boeuf et G. Ghibaudo "Transport characterization of Ge pMOSFETs in saturation regime" In proceedings of ESDERC conference 2011.
- [Dupré 08] C. Dupré, A. Hubert, S. Becu, M. Jublot, V. Maffini-Alvaro, C. Vizios, F. Aussenac, C. Arvet, S. Barnola, J.-M. Hartmann, G. Garnier, F. Allain, J.-P. Colonna, M. Rivoire, L. Baud, S. Pauliac, V. Loup, T. Chevolleau, P. Rivallin, B. Guillaumot, G. Ghibaudo, O. Faynot, T. Ernst et S. Deleonibus. "15nm-diameter 3d stacked nanowires with independent gates operation : PhiFET". In International Electron Devices Meeting. Technical Digest, pages 749 –752, 2008.
- [Durand 66] E. Durand "Electrostatique, Tome 2: problèmes généraux conducteurs" Editions Masson et Cie, 1966.

- [ELDO] ELDO, www.mentor.com
- [ELDO UDM manual] Eldo UDM/GUDM User's Manual, www.mentor.com
- [Elmasry 82] M. I. Elmasry "Capacitance Evaluation in MOSFET VLSI" IEEE Electron Device Lett., vol.3, no. 1, jan 82, pp.6-7.
- [Ernst 99] T. Ernst, D. Munteanu, S. Cristoloveanu, T. Ouisse, S. Horiguchi, Y. Ono. "Investigation of SOI MOSFETs with ultimate thickness", Microelectronics Engineering, Vol.48, pp.339-342, 1999.
- [Ernst 99] T. Ernst and S. Cristoloveanu, "The GP concept for the reduction of short-channel effect in fully depleted SOI devcies", SOI Technology and Device IX, Electrochem. Soc., Pennington, pp.329, 1999.
- [Ernst 02] T. Ernst, C. Tinella, C. Raynaud, S. Cristoloveanu "Fringing fields in sub-0.1 μ m fully depleted SOI MOSFETs: optimization of the device architecture" Solid State Electronics Journal 46, pages 373-378, 2002.
- [Ernst 07] T. Ernst, R. Ritzenthaler, O. Faynot et S. Cristoloveanu "A Model of Fringing Fields in Short-Channel Planar and Triple-Gate SOI MOSFETs" IEEE Trans. On Electron Devices, Vol. 54, n°6, pp.1366-1375 , juin 2007.
- [Fenouillet 11] C. Fenouillet-Beranger, P. Perreau, L. Tosti, O. Thomas, J-P. Noel, T. Benoist, O. Weber, F. Andrieu, A. Bajolet, S. Haendler, M. Cassé, X. Garros, K.K. Bourdelle, F. Boedt, O. Faynot, F. Boeuf "Low power UTBOX and Back Plane (BP) FDSOI technology for 32nm node and below" in poceedings of ICICDT 2011.
- [Fenouillet 12] C.Fenouillet-Beranger, P. Perreau, O. Weber, I. Ben-Akkez, A. Cros, A. Bajolet, S. Haendler, P. Fonteneau, P. Gouraud, E. Richard, F. Abbate, D. Barge, D. Pellissier-Tanon, B. Dumont, F. Andrieu, J. Passieux, R. Bon, V. Barral, D. Golanski, D. Petit, N. Planes, O. Bonin, W. Schwarzenbach, T. Poiroux, O. Faynot, M. Haond, F. Boeuf "Enhancement of Devices Performance of hybrid FDSOI/Bulk Technology by using UTBOX sSOI substrates" in VLSI Tech. Dig., 115-116, 2012.
- [Fischetti 01] M.V. Fischetti, D.A. Neumayer et E.A. Cartier. "Effective electron mobility in Si inversion layers in metal-oxide-semiconductor systems with a high-k insulator : The role of remote phonon scattering". Journal of Applied Physics, vol. 90, no. 9, pages 4587 – 608, 2001.
- [Flandre 10] D. Flandre, V. Kilchytska, and T.Rudenko "gm/Id Method for Threshold Voltage Extraction Applicable in Advanced MOSFETs With Nonlinear Behavior Above Threshold" IEEE Electron Device Lett., Vol. 31, NO. 9, Sept. 2010 pp 930-932.
- [Fleury 09] D. Fleury « Contribution à l'étude expérimentale du transport dans les transistors de dimensions déca-nanométriques des technologies CMOS sub-45nm. » Thèse de doctorat, de l'institut polytechnique de Grenoble, soutenue publiquement en 2009.
- [FlexPDE] www.pdesolutions.com

- [Fukushima 09] T. Fukushima, et al. "Three-dimensional integration technology based on reconfigured wafer-to-wafer and multichip-to-wafer stacking using self-assembly method" in IEDM Tech. Dig., 2009, pp: 349–352.
- [Gallon 07] C. Gallon "Architectures avancées de transistors CMOS SOI pour le noeud 32 nm et en deçà: films ultra-fins, contraintes mécaniques, BOX mince et plan de masse". Thèse de doctorat, INP Grenoble, soutenue publiquement en 2007.
- [Garetto 10] D. Garetto, D. Rideau, E. Dornel, W. F. Clark, C. Tavernier, Y. Leblebici, A. Schmid, and H. Jaouen, " Modeling study of capacitance characteristics in strained High-K Metal gate technology: impact of Si/SiO₂/HK interfacial layer and band structure model" 13th International Nanotech Conference and Expo 2010, 2010.
- [Gautier 03] J. Gautier "Physique des dispositifs pour circuits intégrés sur silicium", chapitre III rédigé par T. Skotnicki et F. Boeuf, edition Lavoisier 2003.
- [Ghani 03] Y. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson et M. Bohr. « A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors ». In International Electron Devices Meeting. Technical Digest, pages 978 – 981, 2003.
- [Guillaumot 02] B Guillaumot, X. Garros, F. Lime, K. Oshima, B. Tavel, J.A. Chroboczek, P. Masson, R. Truche, A. M. Papon, F. Martin, J. F. Damlencourt, S. Maitrejean, M. Rivoire, C. Leroux, S. Cristoloveanu, G. Ghibaudo, J. L. Autran, T. Skotnicki S. Deleonibus « Damascene metal gate and high-k integration for advanced CMOS devices » In: International Electron Devices Meeting Technical Digest 2002, pp:353–356.
- [GSS-a] Rapport en accès libre via http://www.goldstandardsimulations.com/index.php/news/blog_search/simulation-analysis-of-the-intel-22nm-finfet/
- [GSS-b] 25nm template bulk MOSFET Comprehensive simulation study of statistical variability and corresponding advanced statistical compact model parameters extraction and generation. Rapport en accès libre via <http://www.goldstandardsimulations.com/reports/examples/>
- [GSS-c] Statistical Variability in an Example 22nm FinFET, Rapport en accès libre via <http://www.goldstandardsimulations.com/reports/examples/>
- [Gwoziecki 02] R. Gwoziecki and T. Skotnicki, "Physics of the subthreshold slope—Initial improvement and final degradation in short CMOS devices," in Proc. 32nd Eur. Solid-State Device Res. Conf (ESDERC)., 2002, pp. 639–642.
- [Han 08] J. Han, C. Kim, Y. Choi "Universal Potential Model in Tied and Separated Double Gate MOSFETs with Consideration of Symmetric and Asymmetric Structure" IEEE Trans. Electron Devices, vol. 55, NO. 6, JUNE 2008, pp 1472 – 1479.

- [Hänsch 89] W. Hänsch, TH. Vogelsang, R. Kircher and M. Orlowski "Carrier Transport near Si/SiO₂ Interface of a MOSFET" Solid-State Electronics Vol. 32, No. 10, pp. 839-849, 1989.
- [Hellings 10] G. Hellings, L. Witters, R. Krom, J. Mitard, A. Hikavy, R. Loo, A. Schulze, G. Eneman, C. Kerner, J. Franco, T. Chiarella, S. Takeoka, J Tseng, W.-E Wang, W. Vandervorst, P. Absil, S. Biesemans, M. Heyns, K. De Meyer, M. Meuris, T. Hoffmann. "Implant-Free SiGe Quantum Well pFET: A novel, highly scalable and low thermal budget device, featuring raised source/drain and high-mobility channel" In: International Electron Devices Meeting Technical Digest 2010, pp:241–244.
- [Hisamoto 89] D. Hisamoto, T.Kaga et al., "A fully-depleted channel transistor (DELTA) – a novel vertical ultra-thin SOI MOSFET", IEDM Tech. Dig., pp.833, 1989.
- [Hisamoto 91] D. Hisamoto, T. Kaga, E. Takeda, "Impact of the vertical SOI DELTA structure on planar device technology", IEEE Trans. on Electron Devices, Vol.38, N°6, 1991.
- [Hite 92] L.R Hite, H. Lu, T.W. Houston, D.S. Hurta, W.E. Bailey, "An SEU resistant 256K SOI SRAM", IEEE Trans. Nucl. Sci., Vol.39, pp.2121, 1992.
- [Huguenin 10] J.-L. Huguenin, S. Monfray, G. Bidal, S. Denorme, P. Perreau, N. Loubet, Y. Campidelli, M.-P. Samson, C. Arvet, K. Benotmane, F. Leverd, P. Gouraud, B. Le-Gratiet, C. De-Butet, L. Pinzelli, R. Beneyton, S. Barnola, T. Morel, A. Hali-maoui, F. Boeuf, G. Ghibaudo, T. Skotnicki "Ultra-Thin (4nm) Gate-All-Around CMOS devices with High-k/Metal for Low Power Multimedia Applications" International Conference on SSDM 2010.
- [Huguenin 11] J.-L. Huguenin, S. Monfray, J.-M. Hartmann, V. Destefanis, V. Delaye, M.-P. Samson, P. Boulitreau, Y. Morand, P. Brianceau, C. Arvet, P. Gautier, T. Skotnicki, G. Ghibaudo and F. Boeuf « Performance of Localized-SOI MOS Devices on (110) substrates: Impact of Channel Direction ». IEEE Electron Device Letters, 32(8) pp. 996 – 999, 2011.
- [ITRS] www.itrs.net
- [Josse 99] E. Josse and T. Skotnicki "Polysilicon gate with depletion – or – metallic gate with buried channel: what evil worse ?" In: International Electron Devices Meeting Technical Digest 1999.
- [Josse 06] E. Josse, S. Parihar, O. Callen, P. Ferreira, C. Monget, A. Farcy, M. Zaleski, D. Villanueva, R. Ranica, M. Bidaud, D. Barge, C. Laviron, N. Auriac, C. Le Cam, S. Harrison, S. Warrick, F. Leverd, P. Gouraud, S. Zoll, F. Guyader, E. Perrin, E. Baylac, J. Belledent, B. Icard, B. Minghetti, S. Manakli, L. Pain, V. Huard, G. Ribes, K. Rochereau, S. Bordez, C. Blanc, A. Margain, D. Delille, R. Pantel, K. Barla, N. Cave, M. Haond "A Cost-Effective Low Power Platform for the 45-nm Technology Node" In: International Electron Devices Meeting Technical Digest 2006 pp:1-4.
- [Kawasaki 09] H. Kawasaki, V.S. Basker, T. Yamashita, C.-H. Lin, Y. Zhu, J. Faltermeier, S. Schmitz, J. Cummings, S. Kanakasabapathy, H. Adhikari, H. Jagannathan, A. Kumar, K. Maitra, J. Wang, C.-C. Yeh, C. Wang, M. Khater, M. Guillorn, N. Fuller, J. Chang, L. Chang, R.

- Muralidhar, A. Yagishita, R. Miller, Q. Ouyang, Y. Zhang, V.K. Paruchuri, H. Bu, B. Doris, M. Takayanagi, W. Haensch, D. McHerron, J. O'Neill, K. Ishimaru. "Challenges and solutions of FINFET integration in an SRAMP cell and a logic circuit for 22nm node and beyond." In: International Electron Devices Meeting Technical Digest 2009, pp:1–4.
- [Kim 12] S. H. Kim, M. Yokoyama, N. Taoka, R. Nakane, T. Yasuda, O. Ichikawa, N. Fukuhara, M. Hata, M. Takenaka and S. Takagi "Sub-60 nm Deeply-Scaled Channel Length Extremely-Thin Body InxGa1-xAs-On-Insulator MOSFETs on Si with Ni-InGaAs Metal S/D and MOS Interface Buffer Engineering" VLSI Tech. Dig., 177-178, 2012.
- [Khakifirooz 09] A. Khakifirooz, O. M. Nayfeh, and D. Antoniadis, "A simple semiempirical short-channel MOSFET current-voltage model continuous across all regions of operation and employing only physical parameters," IEEE Trans. Electron Devices, vol. 56, no. 8, pp. 1674–1680, Aug. 2009.
- [Khakifirooz 12] A. Khakifirooz, K. Cheng, T. Nagumo, N. Loubet, T. Adam, A. Reznicek, J. Kuss, D. Shahrjerdi, R. Sreenivasan, S. Ponothe, H. He, P. Kulkarni, Q. Liu, P. Hashemi, P. Khare, S. Luning, S. Mehta, J. Gimbert, Y. Zhu, Z. Zhu, J. Li, A. Madan, T. Levin, F. Monsieur, T. Yamamoto, S. Naczas, S. Schmitz, S. Holmes, C. Aulnette, N. Daval, W. Schwarzenbach, B.-Y. Nguyen, V. Paruchuri, M. Khare, G. Shahidi, and B. Doris "Strain Engineered Extremely Thin SOI (ETSOI) for High-Performance CMOS" VLSI Tech. Dig., 117-118, 2012.
- [Lee 89] C.-T. Lee, K. K. Young "Submicrometer near-intrinsic thin-film SOI complementary MOSFETs" in IEEE Trans. Electron Devices, Vol. 36, no. 11, Nov. 1989, pp 2537-2547.
- [Lee 11] J-H Lee "Bulk FinFETs: Fundamentals, Modeling, and Application" in Sematech Symposium 2011, available online on www.sematech.org/meetings/archives/symposia/10202.
- [Leray 90] J.L. Leray, E. Dupont-Nivet, J.F. Peret, Y.M. Coïc, M. Raffaelli, A.J. Auberton-Hervé et al., "CMOS/SOI hardening at 100 Mrad (SiO2)", IEEE Trans. Nucl. Sci., Vol.37, pp.2013, 1990.
- [Lim 83] H.-K. Lim et J. G. Fossum "Threshold Voltage of thin film Silicon-on-insulator (SOI) MOSFETs" IEEE Trans. on Electron Devices, Vol.30, N°10, 1983, p1244-1251.
- [Liu 11] Q. Liu, F. Monsieur, A. Kumar, T. Yamamoto, A. Yagishita, P. Kulkarni, S. Ponothe, N. Loubet, K. Cheng, A. Khakifirooz, B. Haran, M. Vinet, J. Cai, J. Kuss, B. Linder, L. Grenouillet, S. Mehta, P. Khare, N. Berliner, T. Levin, S. Kanakasabapathy, A. Upham, R. Sreenivasan, Y. Le Tiec, N. Posseme, J. Li, J. Demarest, M. Smalley, E. Leobandung, S. Monfray, F. Boeuf, T. Skotnicki, K. Ishimaru, M. Takayanagi, W. Kleemeier, H. Bu, S. Luning, T. Hook, M. Khare, G. Shahidi, B. Doris, R. Sampson "Impact of Back Bias on Ultra-Thin Body and BOX (UTBB) Devices" VLSI Tech. Dig., 160-161, 2011.
- [Mac Andrew 02] C. C. McAndrew, J. J. Victory "Accuracy of Approximations in MOSFET Charge Models" IEEE Transaction On Electron Devices, Vol. 49, No. 1, Jan. 2002
- [MASTAR] MASTAR, disponible en ligne sur <http://www.itrs.net/models.htm>

- [Mathieu 04] H. Mathieu « Physique des semiconducteurs et des composants électroniques » , éditions Dunod, 2004.
- [Meyer 71] J. Meyer “MOS models and circuit simulation” RCA Review, 32, pp.42-63, 1971.
- [Mistry 04] K. Misty, M. Armstrong, C. Auth, S. Cea, T. Coan, T. Ghani, T. Hoffmann, A. Murthy, J. Sandford, R. Shaheed, K. Zawadzki, K. Zhang, S. Thompson and M. Bohr “Delaying Forever: Uniaxial Strained Silicon Transistors in a 90nm CMOS Technology” in Symposium on VLSI Technology, Digest of Technical Papers, pages 50 – 51, 2004.
- [Mistry 07] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. McIntyre, P. Moon, J. Neiryck, S. Pae, C. Parker, D. Parsons, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren°, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, K. Zawadzki “A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging” In: International Electron Devices Meeting Technical Digest 2007, pp: 247–250
- [Monfray 10] S. Monfray, J-L. Huguenin, M. Martin, M. Samson, C. Borowiak, C. Arvet, J. Dalemcourt, P. Perreau, S. Barnola, G. Bidal, S. Denorme, Y. Campidelli, K. Benotmane, F. Leverd, P. Gouraud, B. Le-Gratiet, C. De-Buttet, L. Pinzelli, R. Beneyton, T. Morel, R. Wacquez, J. Bustos, B. Icard, L. Pain, S. Barraud, T. Ernst, F. Boeuf, O. Faynot, T. Skotnicki, “A solution for an ideal planar multi-gates process for ultimate CMOS?” In: International Electron Devices Meeting Technical Digest 2010, pp: 257–260.
- [Munteanu 05] D. Munteanu, J. L. Autran, S. Harrison, K. Nehari, O. Tintori and T. Skotnicki “Compact model of the quantum short-channel threshold voltage in symmetric Double-Gate MOSFET” Molecular Simulation, Vol. 31, No. 12, 15 Oct. 2005, pp 831–837.
- [Na 02] M. H. Na, E. J. Nowak, W. Haensch, J. Cai “The effective drive current in CMOS inverters” In: International Electron Devices Meeting Technical Digest 2002, pp: 121–124.
- [Noel 11] J-P. Noel, O. Thomas, M-A. Jaud, O. Weber, T. Poiroux, C. Fenouillet-Beranger, P. Rivallin, P. Scheiblin, F. Andrieu, M. Vinet, O. Rozeau, F. Boeuf, O. Faynot, and A. Amara “Multi-VT UTBB FDSOI Device Architectures for Low-Power CMOS Circuit” IEEE Trans. Electron Devices, vol. 58, no. 8, pp. 2473–2482, Aug. 2011.
- [Omura 93] Y. Omura, S. Horiguchi, M. Tabe and K. Kishi “Quantum-Mechanical Effects on the Threshold Voltage of Ultrathin-SOI nMOSFET’s” Electron Devices Letter, Vol.14, N°12, pp.569-571, 1993.
- [Packan 09] P. Packan, S. Akbar, M. Armstrong, D. Bergstrom, M. Brazier, H. Deshpande, K. Dev, G. Ding, T. Ghani, O. Golonzka, W. Han, J. He, R. Heussner, R. James, J. Jopling, C. Kenyon, S-H. Lee, M. Liu, S. Lodha, B. Mattis, A. Murthy, L. Neiberg, J. Neiryck, S. Pae, C. Parker, L.

- Pipes, J. Sebastian, J. Seiple, B. Sell, A. Sharma, S. Sivakumar, B. Song, A. St. Amour, K. Tone, T. Troeger, C. Weber, K. Zhang, Y. Luo, S. Natarajan "High Performance 32nm Logic Technology Featuring 2nd Generation High-k + Metal Gate Transistors" In: International Electron Devices Meeting Technical Digest 2009, pp: 659–663.
- [Pao 66] H. C. Pao, C. T. Sah "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors" Solid State Electronics Journal, Vol. 9, Issue 10, pp. 927-937, October 66.
- [Payet 08] F. Payet, F. Boeuf, C. Ortolland, and T. Skotnicki " Nonuniform Mobility-Enhancement Techniques and Their Impact on Device Performance" IEEE Trans. Electron Devices, vol. 55, no. 4, pp. 1050–1058, Apr. 2008.
- [Planes 08] N. Planes, V.Huard, C.Laviron, O.Callen, J.Bonnouvrier, O.Menut, S.Haendler, M. Haond and F.Boeuf "Process Architecture for Spatial and Temporal Variability Improvement of SRAM Circuits at the 45nm node" in International Conference on SSDM 2008.
- [Planes 12] N. Planes, O. Weber, V. Barral, S. Haendler, D. Noblet, D. Croain, M. Bocat, P.O. Sassoulas, X. Federspiel, A. Cros, A. Bajolet, E. Richard, B. Dumont, P. Perreau, D. Petit, D. Golanski, C. Fenouillet-Béranger, N. Guillot, M. Rafik, V. Huard, S. Puget, X. Montagner, M.-A. Jaud, O. Rozeau, O. Saxod, F. Wacquant, F. Monsieur, D. Barge, L. Pinzelli, M. Mellier, F. Boeuf, F. Arnaud and M. Haond "28nm FDSOI Technology Platform for High-Speed Low-Voltage Digital Applications" VLSI Tech. Dig., 133-134, 2012.
- [Plonsey 61] R. Plonsey and R. E. Collin, "Principles and Applications of Electromagnetic Fields". New York: McGraw-Hill, 1961.
- [Poiroux 05] T. Poiroux, M. Vinet, O. Faynot, J. Widiez, J. Lolivier, T. Ernst, B. Previtali, S. Deleonibus "Multiple gate devices: advantages and challenges" in Microelectronic Engineering 80, 2005, pp 378-385.
- [Pouydebasque 07] A. Pouydebasque, C. Charbuillet, R. Gwoziecki, and T. Skotnicki, "Refinement of the Subthreshold Slope Modeling for Advanced Bulk CMOS Devices" in IEEE Trans. Electron Devices, vol. 54, no. 10, Oct 2007, pp2723-2729.
- [PSP] <http://pspmodel.asu.edu/spc.htm>
- [Pregaldiny 02] F. Pregaldiny, C. Lallement, A. Mathiot, "A simple efficient model of parasitic capacitances of deep-submicron LDD MOSFETs" Solid-State Electronics, v 46, n 12, p 2191-8, Dec. 2002.
- [Radosavljevic 11] M. Radosavljevic, G. Dewey, D. Basu, J. Boardman, B. Chu-Kung, J. M. Fastenau, S. Kabehie, J. Kavalieros, V. Le, W. K. Liu, D. Lubyshev, M. Metz, K. Millard, N. Mukherjee, L. Pan, R. Pillarisetty, W. Rachmady, U. Shah, H. W. Then and R. Chau "Electrostatics Improvement in 3-D Tri-gate Over Ultra-Thin Body Planar InGaAs Quantum Well Field Effect Transistors with High-K Gate Dielectric and Scaled Gate-to-Drain/Gate-to-Source Separation" In: International Electron Devices Meeting Technical Digest 2011, pp: 765–768.

- [Rafhay 10] Q. Rafhay, R. Clerc, J. Coignus, G. Pananakakis and G. Ghibaudo "Dark Space, Quantum Capacitance and Inversion Capacitance in Si, Ge, GaAs and In_{0.53}Ga_{0.47}As nMOS Capacitors" in ULIS conference 2010.
- [Ramey 03] S. M. Ramey and D. K. Ferry "Threshold Voltage Calculation in Ultrathin-Film SOI MOSFETs Using the Effective Potential" IEEE Trans. Electron Devices, vol. 2, no. 3, pp. 121–125, Sep. 2003.
- [Raphael] Raphael®, www.synopsys.com
- [Rios 94] R. Rios, N.D. Arora "Determination of ultra-thin gate oxide thicknesses for CMOS structures using quantum effects" IEDM Tech. Dig., 1994, pp: 613–616.
- [Shang 12] H. Shang, S. Jain, E. Josse, E. Alptekin, M.H. Nam, S.W. Kim, K.H. Cho, I. Kim, Y. Liu, X. Yang, X. Wu, J. Ciavatti, N.S. Kim, R. Vega, L. Kang, H.V. Meer, S. Samavedam, M. Celik, S. Soss, H. Utomo, R. Ramachandran, W. Lai, V. Sardesai, C. Tran, J.Y. Kim, Y.H. Park, W.L. Tan, T. Shimizu, R. Joy, J. Strane, K. Tabakman, F. Lalanne, P. Montanini, K. Babich, J. B. Kim, L. Economikos, W. Cote, C. Reddy, M. Belyansky, R. Arndt, U. Kwon, K. Wong, D. Koli, D. Levedakis, J.W. Lee, J. Muncy, S. Krishnan, D. Schepis, X. Chen, B.D. Kim, C. Tian, B.P. Linder, E. Cartier, V. Narayanan, G. Northrop, O. Menut, J. Meiring, A. Thomas, M. Aminpur, S.H. Park, K.Y. Lee, B.Y. Kim, S.H. Rhee, B. Hamieh, R. Srivastava, R. Koshy, C. Goldberg, M. Pallachalil, M. Chae, A. Ogino, T. Watanabe, M. Oh, H. Mallela, D. Codi, P. Malinge, M. Weybright, R. Mann, A. Mittal, M. Eller, S. Lian, R. Divakaruni, S. Bukofsky, J.D. Kim, J. Sudijono, W. Neumueller, F. Matsuoka, R. Sampson. "High Performance Bulk Planar 20nm CMOS Technology for Low Power Mobile Applications" VLSI Tech. Dig., 129–130, 2012.
- [Shrivastava 82] R. Shrivastava and K. Fitzpatrick, "A simple model for the overlap capacitance of a VLSIMOS device," IEEE Trans. Electron Devices, vol. ED-29, no. 12, pp. 1870–1875, Dec. 1982.
- [Skotnicki 88-a] T. Skotnicki, G. Merckel, and T. Pedron, "The voltage–doping transformation: A new approach to the modeling of MOSFET short-channel effects," IEEE Electron Device Lett., vol. 9, no. 3, pp. 109–112, Mar. 1988.
- [Skotnicki 88-b] T. Skotnicki, G. Merkel and T. Pedron "A New Punchthrough Current Model Based on the Voltage-Doping Transformation" IEEE Trans. Electron Devices, vol. 35, no. 7, pp. 1076–1086, Jan. 1988.
- [Skotnicki 94] T. Skotnicki, C. Denat, P. Senn, G. Merckel, and B. Hennion, "A New Analog/Digital CAD Model for sub-halfmicron MOSFETs" IEDM Tech. Dig., 1994, pp: 165–168.
- [Skotnicki 00] T. Skotnicki "Circuits intégrés CMOS sur silicium" Technique de l'ingénieur, E2 432, 2000.
- [Skotnicki 08] T. Skotnicki, C. Fenouillet-Beranger, C. Gallon, F. Boeuf, S. Monfray, F. Payet, A. Pouydebasque, M. Szczap, A. Farcy, F. Arnaud, S. Clerc, M. Sellier, A. Cathignol, J.-P. Schoellkopf, E. Perea, R. Ferrant, and H. Mingam, "Innovative materials, devices, and

- CMOS technologies for low-power mobile multimedia,” IEEE Trans. Electron Devices, vol. 55, no. 1, pp. 96–130, Jan. 2008.
- [Skotnicki 09] T. Skotnicki « Low Power Logic and Mixed-Signal Technologies » in a shortcourse given in IEDM conference 2009.
- [Skotnicki 10] T. Skotnicki et F. Boeuf “How can high mobility channel materials boost or degrade performance in advanced CMOS?” Invited paper in VLSI tech. Dig. 2010.
- [Suzuki 99] Suzuki K. “Parasitic capacitance of submicrometer MOSFET’s.” IEEE Trans Electron Dev, vol. 46, no 9:1895–900, sept 1999.
- [Suzuki 03] K. Suzuki et S. Pidin “Short Channel Single-Gate SOI MOSFET Model” IEEE Trans. Electron Devices, vol. 50, no. 5, pp. 1297–1305, May 2003.
- [Synopsys] www.synopsys.com
- [Sze 81] S. M. Sze “Physics of Semiconductor Devices” 2nd Edition, 1981.
- [Takagi 94] S. Takagi, A. Toriumi, M. Iwase, H. Tango “On the Universality of inversion layer mobility in Si MOSFETs” IEEE Trans. Electron Devices, vol. 41, no. 12, pp. 2357–2362, Dec. 1994.
- [Taur 98] Y. Taur and T. H. Ning, “Fundamentals of Modern VLSI Devices”. Cambridge, U.K.: Cambridge Univ. Press, 1998.
- [Tiwari 10] P.K. Tiwari C.R. Panda A. Agarwal P. Sharma S. Jit “Modelling of doping-dependent subthreshold swing of symmetric double-gate MOSFETs” IET Circuits, Devices and Systems, vol.4, pp.337-345 2010.
- [Toyabe 79] T. Toyabe, S. Asai “Analytical Models of Threshold Voltage and Breakdown Voltage of Short-Channel MOSFET’s Derived from Two-Dimensional Analysis” IEEE Trans. Electron Devices, vol. 26, no.4, pp. 453–461, Dec. 1979.
- [Tsividis 87] Y. P. Tsividis « Operation and modeling of the MOS transistor », MacGraw-Hill Edition, 1987.
- [Ueno 05] T. Ueno, H. S. Rhee, S. H. Lee, H. Lee, D. S. Shin, Y.-S. Jin, S. Maeda, N.-I. Lee “Dramatically Enhanced Performance of Recessed SiGe Source-Drain PMOS by In-Situ Etch and Regrowth Technique (InSERT)” In Symposium on VLSI Technology, Digest of Technical Papers, pages 24 – 25, 2005.
- [Ward 78] D.E. Ward and R. W. Dutton “A Charge-Oriented Model for MOS Transistor Capacitances” in IEEE Journal of Solid-State Circuits Vol. SC-13, No. 5, Oct. 1978
- [Weber 08] O. Weber, O. Faynot, F. Andrieu, C. Buj-Dufournet, F. Allain, P. Scheiblin, J. Foucher, N. Daval, D. Lafond, L. Tosti, L. Brevard, O. Rozeau, C. Fenouillet- Beranger, M. Marin, F. Boeuf, D. Delprat, K. Bourdelle, B.-Y. Nguyen and S. Deleonibus “High Immunity to Threshold Voltage Variability in Undoped Ultra-Thin FDSOI MOSFETs and its Physical

- Understanding". In 2008 International Electron Devices Meeting. Technical Digest, pages 245 – 248.
- [Wei 09] L. Wei, F. Boeuf, D. Antoniadis, T. Skotnicki, and H.-S. P. Wong "Exploration of Device Design Space to Meet Circuit Speed Targeting 22nm and Beyond" in International SSDM conference 2009.
- [Wei 11] L. Wei, F. Boeuf, T. Skotnicki, and H.-S. P. Wong, "Parasitic Capacitances: Analytical Models and Impact on Circuit-Level Performance" IEEE Trans. Electron Devices, vol. 58, no. 5, May. 2011, pp. 1361–1370.
- [Wei 12] L. Wei, O. Mysore, D. Antoniadis "Virtual-Source-Based Self-Consistent Current and Charge FET Models: From Ballistic to Drift-Diffusion Velocity-Saturation Operation" IEEE Trans. Electron Devices, vol. 59, no. 5, may 2012, pp. 1263-1272.
- [Westlinder 03] J. Westlinder, T. Schram, L. Pantisano, E. Cartier, A. Kerber, G. S. Lujan, J. Olsson and G. Groeseneken "On the Thermal Stability of Atomic Layer Deposited TiN as Gate Electrode in MOS Devices" in IEEE Electron Device Letters, Vol. 24, No. 9, Sep. 2003.
- [Wong 87] H.S.P. Wong, M.H. White, T.J. Krutsick, R.V. Booth "Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFETs" , Solide-State Electronics, v 30, no 10, Sept. 1987, pp 953-968.
- [Wong 98] H.S.P. Wong, D. J. Frank, and P. M. Solomon "Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SO1 MOSFET's at the 25 nm Channell Length Generation" In 1988 International Electron Devices Meeting. Technical Digest, pages 407 – 411.
- [Wu 07] W. Wu and M. Chan, "Analysis of geometry-dependent parasitics in multifin double-gate FinFETs," IEEE Trans. Electron Devices, vol. 54, no. 4, pp. 692–698, Apr. 2007.
- [Yang 83] P. Yang, B. D. Epler et P. K. Chatterjee "An investigation of the charge conservation problem for mosfet circuit simulation" in IEEE Journal of Solid-State Circuits Vol. SC-18, no 1, Feb. 1983.
- [Yokoyama 12] M. Yokoyama, S. H. Kim, R. Zhang, N. Taoka, Y. Urabe, T. Maeda, H. Takagi, T. Yasuda, H. Yamada, O. Ichikawa, N. Fukuhara, M. Hata, M. Sugiyama, Y. Nakano, M. Takenaka, and S. Takagi "CMOS integration of InGaAs nMOSFETs and Ge pMOSFETs with self-align Ni-based metal S/D using direct wafer bonding" VLSI Tech. Dig., 60-61, 2011.
- [Yuan 12] Z. Yuan, A. Nainani, A. Kumar, X. Guan, B. R. Bennett, J. B. Boos, M. G. Ancona and K. C. Saraswat "InGaSb: Single Channel Solution for Realizing III-V CMOS" VLSI Tech. Dig., 185-186, 2012.
- [Zhang 08] L. Zhang, H.-W. Guo, C. Zhang, W. Zhang, S.-J. Ding "Characterization of low-dielectric-constant SiCON films grown PECVD under different RF power" in ICSICT international conference 2008 proceedings, pp. 788-791.

PUBLICATIONS DE L'AUTEUR

Articles de journaux :

- **J. Lacord**, J.-L. Huguenin, S. Monfray, R. Coquand, T. Skotnicki, G. Ghibaudo and F. Boeuf « Comparative Study of Circuit Perspectives for Multi-gate Structures at sub-10nm Node” in Solid State Electronics Journal 2012.
- **J. Lacord**, J.-L. Huguenin, T. Skotnicki, G. Ghibaudo, and F. Boeuf “Simple and Efficient MASTAR Threshold Voltage and Subthreshold Slope Models for Low-Doped Double-Gate MOSFET” in Transaction on Electron Devices 2012.
- **J. Lacord**, G. Ghibaudo, and F. Boeuf “Comprehensive and Accurate Parasitic Capacitance Models for Two- and Three-Dimensional CMOS Device Structures” in Transaction on Electron Devices 2012.

Articles de conférences internationales :

- J.-L. Huguenin, **J. Lacord**, S. Monfray, R. Coquand, T. Skotnicki, G. Ghibaudo et F. Boeuf (2011). Comparative study of circuit perspectives for multi-gate structures at sub-10nm node. In European Solid-State Device Research Conference (ESDERC) 2011.
- **J. Lacord**, J.-L. Huguenin, G. Ghibaudo, T. Skotnicki and F. Boeuf “Simple and Efficient MASTAR Threshold Voltage and Subthreshold Slope Models for Double Gate Structures” in SSDM conference 2011.
- **J. Lacord**, D. Hoguet, D. Rideau, G. Ghibaudo, and F. Boeuf “Accurate and Ready-to-use Parasitic Capacitances Models for Advanced 2D/3DCMOS Device Structure Comparison” in SSDM conference 2011.
- T. Dutta, Q. Raffay, R. Clerc, **J. Lacord**, S. Monfray, G. Pananakakis, F. Boeuf and G. Ghibaudo “Origins of the Short Channel Effects Increase in III-V nMOSFET Technologies” in ULIS conference 2012.
- **J. Lacord**, P. Batude, G. Ghibaudo and F. Boeuf “Analytical Modeling of Parasitics in Monolithically Integrated 3D Inverters” in ICICDT conference 2012.
- **J. Lacord**, G. Ghibaudo, and F. Boeuf “A Comparative Study of Minimal Supply Voltage of 6T-SRAM at the 16nm node using MASTAR into a Conventional CAD Environment” in SSDM conference 2012.

Annexe: Utilisation de MASTAR VA

1. Séquence de modélisation dans un outil de CAO conventionnel comme ELDO

L'utilisation de MASTAR VA (et de tout modèle compact) pour une simulation avec un simulateur de CAD conventionnel comme ELDO s'effectue selon le flot représenté ci-dessous :

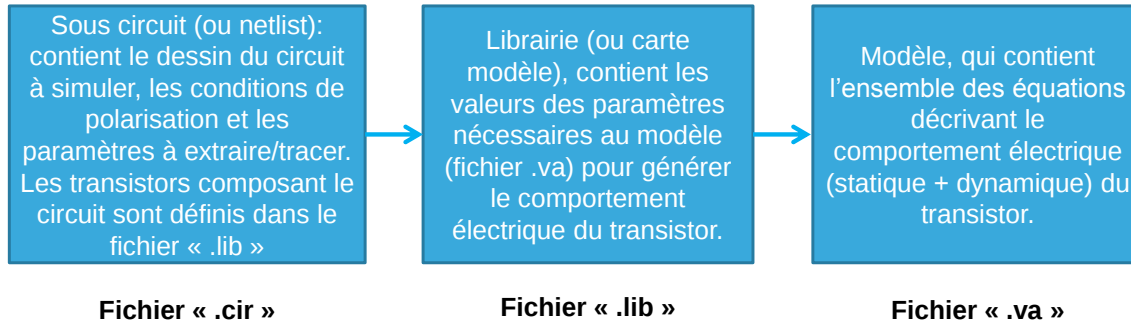


Figure 23: Flot d'utilisation de MASTAR VA

2. Exemple d'utilisation de MASTAR VA : Tracé d'une $I_d(V_g)$

Ci-dessous est représentée une copie d'écran d'une netlist (fichier « .cir ») permettant de tracer une courbe I_d-V_g :

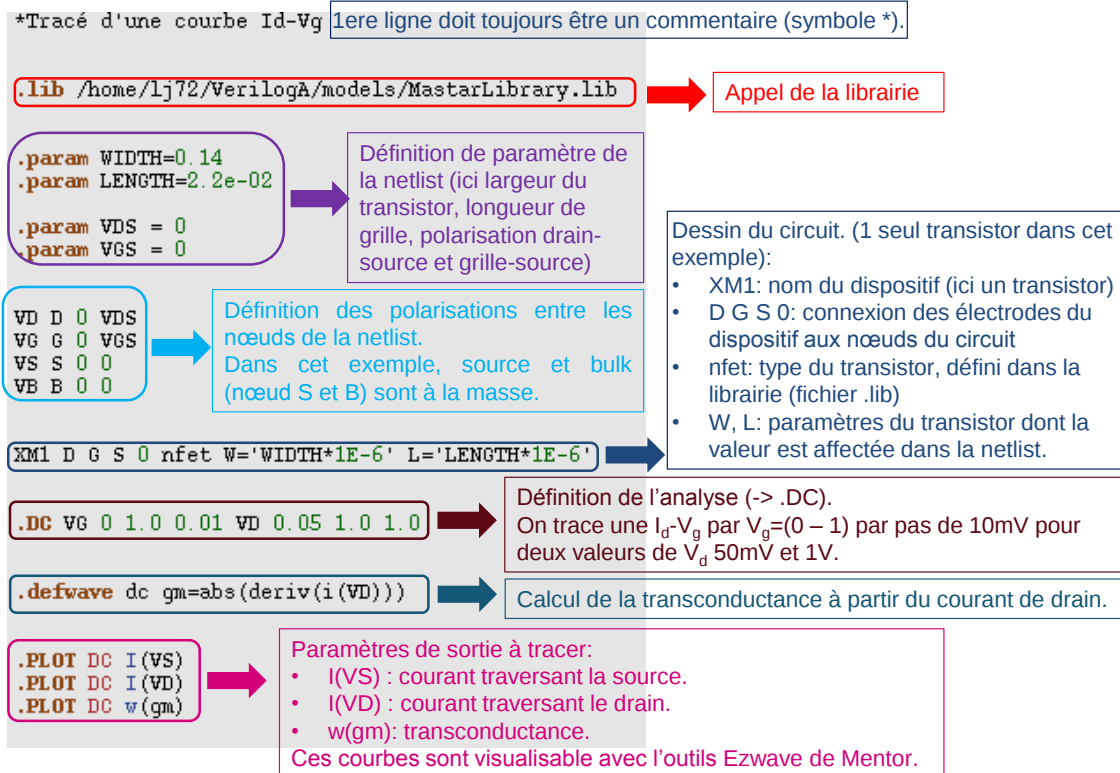


Figure 24: Copie d'écran d'une netlist (fichier « .cir ») permettant de tracer une courbe I_d-V_g

En effectuant une simulation avec ELDO de cette netlist (nommée « IdVg.cir »), par la commande :

Eldo IdVg.cir –ezwave

Les courbes I_d-V_g et $g_m(V_g)$ tracées à partir de la librairie MastarLibrary.lib s'affichent, par l'intermédiaire de l'outil de visualisation de courbe de Mentor Ezwave. On obtient alors les courbes représentées sur la capture d'écran ci-dessous.

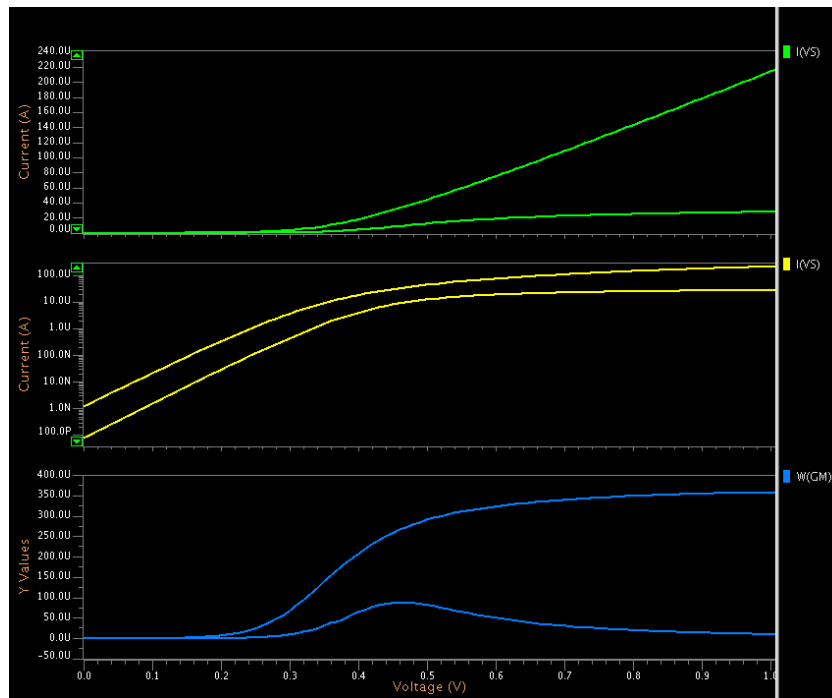


Figure 25: Capture d'écran représentant les courbes tracées par l'intermédiaire du fichier IdVg.cir.

Sur la Figure 3, on a :

- $I(VS)$ est le courant qui traverse l'électrode de source,
- $I(VD)$ (non représenté ici) le courant qui traverse l'électrode de drain (donc $I(VS) = -I(VD)$)
- $W(GM)$ est la transconductance.

3. Définition d'un inverseur dans l'environnement de CAO

Si on souhaite effectuer des simulations d'inverseurs, montés en chaînes ou en anneaux résonnants, il est préférable de définir des sous-circuits dans la netlist, comme illustré ci-dessous :

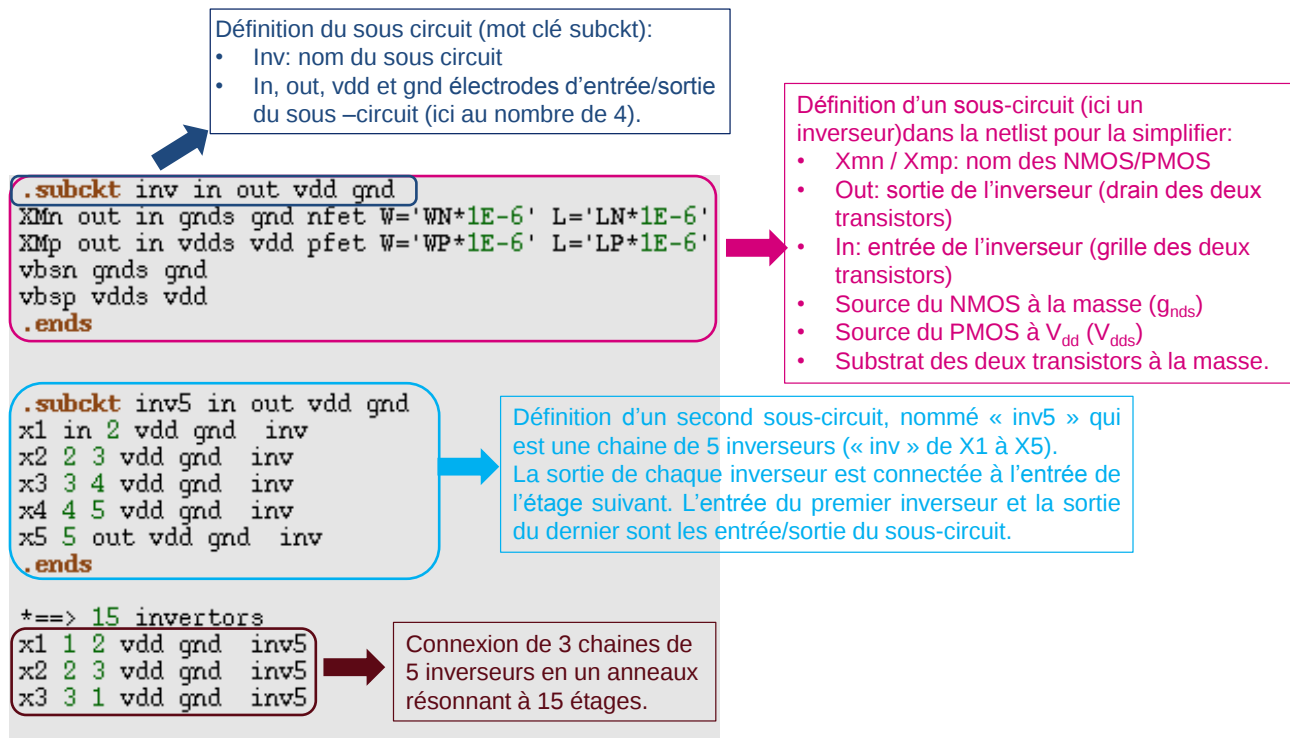


Figure 26: Copie d'écran d'une netlist (fichier « .cir ») avec définition de sous-circuits.

4. Description d'une librairie de MASTAR VA

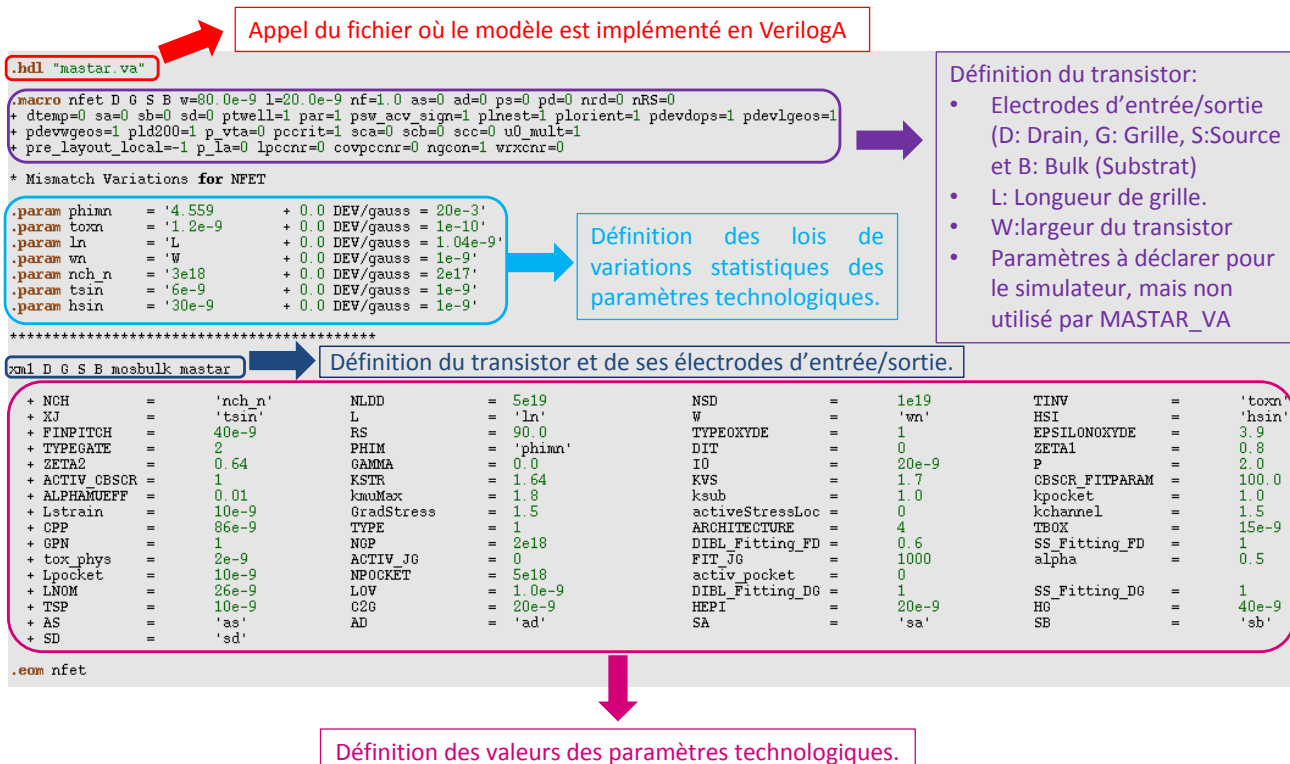


Figure 27: Capture d'écran d'une librairie appelant le modèle MASTAR VA.

5. Description du corps du modèle

Il est maintenant temps de décrire le corps du modèle, c'est-à-dire le contenu du fichier mastar.va. Cette partie se décompose en quatre paragraphes :

5.a. Schéma équivalent du transistor, vu par le simulateur

Le schéma équivalent du transistor est donné ci-dessous.

Les électrodes d'entrées/sortie du transistor sont :

- G: Grille
- D: Drain
- S: Source
- B : Substrat

Les paramètres de sortie, calculés par le modèle, sont divisés en deux groupes :

- Les paramètres statiques :
 - I_{ds} : courant drain-source
 - I_{gcs} : courant grille-source
 - I_{gcd} : courant grille-drain
- Les paramètres dynamiques :
 - Q_g : Charge de grille
 - Q_s : Charge de source
 - Q_d : charge de drain
 - Q_b : Charge de substrat

Les valeurs de ces 4 charges sont utilisées par le simulateur pour déterminer les valeurs des capacités représentées sur le schéma équivalent.

La prise en compte des résistances d'accès coté source et coté drain s'effectue par l'introduction de nœuds virtuels (D' et S') qui sont directement aux bornes du canal et reliés aux électrodes de source (S) et de drain (D) par une résistance, égale à la résistance d'accès.

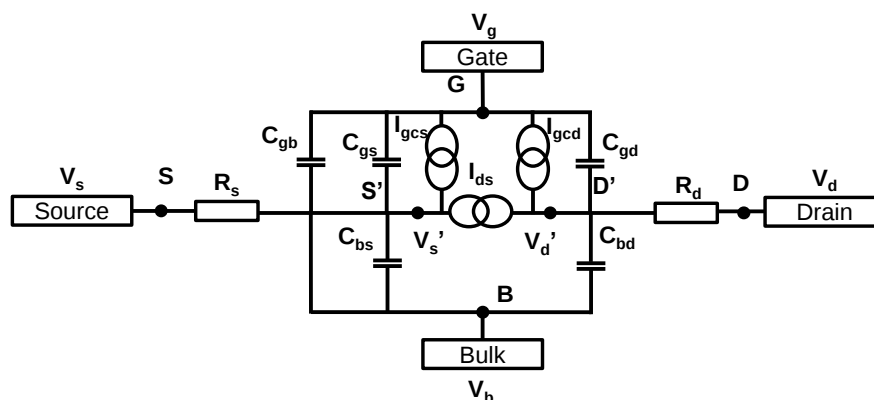


Figure 28: Le schéma équivalent du transistor, vu par le simulateur.

5.b. Structure du modèle

Le modèle se décompose en 6 sections :

5.b.1) Définition des constantes

Définition des constantes physiques utilisées par le modèle :

Nom	Description	Valeur
eps0	Permittivité du vide	$8.854214871 \times 10^{-12} \text{ F.m}^{-1}$
epssi	Permittivité du silicium	$1.053655 \times 10^{-10} \text{ F.m}^{-1}$
epssio2	Permittivité du SiO_2	$3.453157 \times 10^{-11} \text{ F.m}^{-1}$
epssi3n4	Permittivité du Si_3N_4	$6.197974 \times 10^{-11} \text{ F.m}^{-1}$
q	Charge de l'électron	$1.6 \times 10^{-19} \text{ C}$
k	Ratio de la constante de Boltzman et de la charge de l'électron q	$0.0000861 \text{ V.K}^{-1}$
pi	pi	3.14159265358979323846
m0	Masse de l'électron	$9.10956 \times 10^{-31} \text{ kg}$
mc	Masse relative effective de l'électron	1.084
mv	Masse relative effective d'un trou	0.549
h	Constante de Planck	6.6261×10^{-34}
REFTEMP	Temperature de reference (27°C)	300.15 K

Tableau 1: Description des constantes.

5.b.2) Initialisation du modèle

Cette section a pour objectif de calculer les différentes constantes qui sont fonction des conditions de simulations et des paramètres technologiques du transistor, donnés dans le fichier « .lib ». Cette section débute par le mot clé « begin : initializeModel ».

Un échantillon des constantes calculées est listé ci-dessous :

- Valeur de la largeur de la bande interdite (E_g)
- Concentration en porteurs intrinsèque (n_i)
- Longueur électrique du transistor (L_{elec})
- EOT
- Potentiel de Fermi (ϕ_f)
- Potentiel de diode (ϕ_d)
- ...

5.b.3) Evaluation des paramètres statiques

Cette section a pour objectif de fournir la description du comportement statique du transistor, c'est-à-dire les valeurs des différents courants (I_{ds} , I_{gsc} et I_{gsd}) pour un jeu de polarisation donné, mais également des capacités parasites. Elle débute par le mot clé « begin : evaluateStatic ».

La principale, et première, partie du code de cette section est consacrée à l'évaluation du courant drain-source I_{ds} est organisée de la manière suivante :

- Récupération des différentes polarisations.

- Calcul des paramètres électrostatiques par des modèles analytiques spécifiques à chaque architecture (BULK, FDSOI ou Double/Triple grille) :
 - Tension de seuil
 - DIBL
 - Pente sous le seuil
- Calcul des paramètres de transport, avec pour résultat final la mobilité effective, en tenant compte des effets de contraintes et de la longueur de grille.
- Calcul du courant drain-source

La seconde partie propose d'évaluer le courant de fuite de grille, donné par les deux composantes suivantes :

- I_{gsc} : fuite de grille coté source.
- I_{gsd} : fuite de grille coté drain.

La dernière partie a pour objectif d'évaluer les capacités parasites du transistor par les expressions analytiques de chaque composante, spécifiques à chaque architecture.

5.b.4) Evaluation des paramètres dynamiques

Cette section a pour objectif de fournir la description du comportement dynamique du transistor, c'est-à-dire les valeurs des différentes charges (Q_g , Q_s , Q_d et Q_b) pour un jeu de polarisation donné, qui seront utilisées par le simulateur pour obtenir les valeurs des différentes capacités intrinsèques du transistor. Elle débute par le mot clé « *begin : evaluateDynamic* ».

5.b.5) Chargements des paramètres statiques à chaque nœud du transistor

Cette section a pour objectif d'indiquer au simulateur les valeurs de chaque paramètre statique pour chaque nœud (i.e chaque électrode réelle : D G S B et virtuelle D' S'), c'est-à-dire les valeurs des courants circulant entre chaque nœud. Elle débute par le mot clé « *begin : loadStatic* ».

5.b.6) Chargements des paramètres dynamiques à chaque nœud du transistor

Cette section a pour objectif d'indiquer au simulateur les valeurs de chaque paramètre dynamique pour chaque nœud (i.e chaque électrode réelle : D G S B et virtuelle D' S'), c'est-à-dire la valeur de la charge (intrinsèque + extrinsèque (->capacités parasites)) sur chaque nœud. Elle débute par le mot clé « *begin : loadDynamic* ».

5.c. Paramètres du modèle

Nom + Unité	Description
NCH (m^{-3})	Dopage canal.
NLDD (m^{-3})	Dopage des LDD.
NSD (m^{-3})	Dopage S/D.
TINV (m)	Épaisseur d'oxyde de grille équivalente en inversion.
XJ (m)	Profondeur de jonction pour le BULK, de film de silicium pour les architectures à film

	mince.
L (m)	Longueur de grille.
W (m)	Largeur physique du transistor.
HSI (m)	Hauteur du fin (uniquement pour architecture FinFET/Trigate).
FINPITCH (m)	Pas de répétition d'un fin (uniquement pour architecture FinFET/Trigate).
RS ($\Omega \cdot \mu\text{m}$)	Résistance d'accès S/D.
TYPEOXYDE ()	Test si oxyde de grille en SiO_2 ou en matériau HK.
EPSILONOXYDE ()	Permittivité de l'oxyde de grille.
TYPEGATE ()	Test si grille en polysilicium ou en métal.
PHIM (V)	Travail de sortie de la grille.
DIT (cm^2/eV)	Densité d'état d'interface.
ZETA1 ()	Paramètre d'ajustement du DIBL pour le modèle BULK.
ZETA2 ()	Paramètre d'ajustement du SCE pour le modèle BULK.
GAMMA ()	Evaluation de la longueur de la zone de recouvrement des jonctions sous la grille en fonction de la longueur de grille.
P ()	Paramètre ajustant le passage en saturation des I_d - V_d .
IO (A)	Proportionnel à la valeur du courant au seuil.
KVS ()	Paramètre de vitesse de saturation.
CBSCR_FITPARAM ()	Valeur du paramètre de l'effet d'écrantage de Coulomb sur la mobilité.
ACTIV_CBSCR ()	Activation de l'effet d'écrantage de Coulomb sur la mobilité.
KSTR ()	Paramètre de dégradation/amélioration de la mobilité effective du à la contrainte.
ALPHAMUEFF ()	Paramètre de dégradation de la mobilité due à la réduction de longueur de grille.
kmuMax ()	Facteur d'amélioration maximale de la mobilité, par défaut 1.8 pour un NMOS, 2.5 pour un PMOS.
ksub ()	Facteur d'amélioration de la mobilité à faible champ.
kpocket ()	Facteur d'amélioration de la mobilité due aux poches de contrainte (i.e au liner contraint).
Lstrain (m)	Longueur de la poche de contrainte résultante.
GradStress ()	Gradient de contrainte du au liner contraint.
activeStressLoc ()	Activation (=1) de l'effet du liner contraint sur la mobilité.
kchannel ()	Facteur d'amélioration de la mobilité dû au canal.
CPP (m)	Contacted Poly Pitch.
TYPE ()	Type du transistor: NMOS ou PMOS.
TBOX (m)	Epaisseur de l'oxyde enterré.
ARCHITECTURE ()	Identifie l'architecture considérée: • 1: BULK without Raised S/D. • 2: BULK with Raised S/D. • 3: FDSOI_no_Rs/d • 4: FDSOI_with_Rs/d • 5: DG_no_Rs/d • 6: DG_with_Rs/d • 7: FINFET_with_Rs/d • 8: Trigate_with_Rs/d
GPN ()	Type d'implantation du Ground Plane (pour FDSOI uniquement) : • 1 : GPN • 0 : GPP
NGP (m^{-3})	Dopage du GP.

DIBL_FITTING_FD ()	Paramètre d'ajustement du DIBL pour le modèle électrostatique FDSOI.
SS_FITTING_FD ()	Paramètre d'ajustement de la pente sous le seuil pour le modèle électrostatique FDSOI.
tox_phys (m)	Epaisseur d'oxyde de grille totale déposée pour le calcul de la fuite de grille.
ACTIV_JG ()	Activation (=1, 0 sinon) de la prise en compte des fuites de grilles.
FIT_JG ()	Paramètre d'ajustement du modèle de fuite de grille (valeur absolue).
alpha ()	Paramètre d'ajustement du modèle de fuite de grille (dépendance en V_d).
Lpocket (m)	Longueur de la poche de dopage.
NPOCKET (m^{-3})	Dopage de la poche.
activ_pocket ()	Activation de la prise en compte du dopage poche.
LNOM (m)	Longueur de grille nominale de la technologie.
LOV (m)	Entrée manuelle de la longueur de la zone de recouvrement des jonctions sous la grille.
DIBL_Fitting_DG ()	Paramètre d'ajustement du DIBL pour le modèle électrostatique Double grille.
SS_Fitting_DG ()	Paramètre d'ajustement de la pente sous le seuil pour le modèle électrostatique Double grille.
TSP (m)	Epaisseur de l'espaceur.
C2G (m)	Distance grille-contact.
HEPI (m)	Hauteur de l'épitaxie.
HG (m)	Hauteur de grille.
AS (m^2)	Surface de la source (non utilisé par MASTAR VA mais doit être déclaré pour le simulateur).
AD (m^2)	Surface du drain (non utilisé par MASTAR VA mais doit être déclaré pour le simulateur).
SA (m)	Longueur de la source (non utilisé par MASTAR VA mais doit être déclaré pour le simulateur).
SB (m)	Longueur du drain (non utilisé par MASTAR VA mais doit être déclaré pour le simulateur).
SD (m)	Distance entre chaque doigts (non utilisé par MASTAR VA mais doit être déclaré pour le simulateur).
Tableau 2 : Description des paramètres modèle de MASTAR VA	

TITRE: Développement de modèles pour l'évaluation des performances circuit des technologies CMOS avancées sub-20nm.

RESUME

Depuis la commercialisation du premier circuit intégré en 1971, l'industrie de la microélectronique s'est fixée comme leitmotiv de réduire les dimensions des transistors MOSFETs, en suivant la loi de Moore. Comme indiqué par Dennard, cette miniaturisation améliore automatiquement les performances des transistors.

A partir des nœuds 28-22nm, les effets canaux courts sont trop difficiles à contrôler et de nouvelles architectures de transistors sont introduites: FDSOI pour STMicroelectronics, Trigate pour Intel.

Dans ce contexte, l'évaluation des performances des technologies CMOS est clé et les travaux de cette thèse proposent de les évaluer au niveau circuit. Des modèles spécifiques d'estimation des paramètres électrostatiques et des capacités parasites sont développés. Ceux-ci sont d'abord utilisés sur des technologies amonts (co-intégration III-V/Ge et intégration 3D) puis sont implémentés en VerilogA pour être utilisés avec les outils conventionnel de CAO. Ceci fournit un modèle compact prédictif et utilisable pour toutes les architectures CMOS, qui est utilisé pour évaluer les performances logiques et SRAM des architectures BULK, FDSOI et Trigate aux nœuds 20nm et 16nm.

Mots clés :

CMOS, architecture, évaluation de performance, modélisation analytique, FDSOI, FinFET, Trigate, double grille, capacité parasite, électrostatique.

TITLE: Models development for power performance assessment of advanced CMOS technologies (sub-20nm).

ABSTRACT

Since the commercialization of the first integrated circuit in 1971, the microelectronic industry has fixed as an objective to reduce MOSFET transistor dimensions, following Moore's law. As indicated by Dennard, this miniaturization automatically improves device performances.

Starting from the 28-22nm technological node, short channel effects are too strong and industrial companies choose to introduce new device structure: FDSOI for STMicroelectronics and Trigate for Intel.

In such a context, CMOS technology performance evaluation is key and this thesis proposes to evaluate them at circuit level. Specific models for electrostatic parameters and parasitic capacitances for each device structure are developed for each device structure. Those models have first been used to evaluate performances of advanced technologies, such as III-V/Ge co-integration and 3D monolithic integration and have then been implemented in VerilogA to ensure compatibility with conventional CAD tools such as ELDO. This provides a compact model, predictive and usable for each device structure, which has been used to evaluate logic and SRAM performances of BULK, FDSOI and Trigate devices for the 20nm and 16nm technology node.

Key words :

CMOS, structure, performance assessment, analytical modeling, FDSOI, FinFET, Trigate, double gate, parasitic capacitance, electrostatic.