N° d'ordre NNT : 2023ECDL0026

THÈSE de DOCTORAT DE L'ÉCOLE CENTRALE DE LYON
opérée au sein de l'Université de Lyon

École Doctorale n°160
Électronique Électrotechnique et Automatique (EEA)

Discipline : Électronique, Micro et Nanoélectronique, Optique et Laser

Soutenue publiquement le 5 juillet 2023, par :

Mayeul Cantan

Calcul éco-énergétique avec matériaux ferroélectriques intégrés pour les systèmes embarqués et à la périphérie de réseau

---

Devant le jury composé de :

Maneux, Cristell	Professeur des Universités[a]	Université de Bordeaux / IMS Bordeaux	Présidente
Niemier, Michael	Full Professor	University of Notre Dame, Indiana, USA	Rapporteur
Portal, Jean-Michel	Professeur des Universités[a]	Aix-Marseille Université / IM2NP	Rapporteur
O'Connor, Ian	Professeur des Universités[a]	École Centrale de Lyon / INL	Directeur de thèse
Deleruyelle, Damien	Professeur des Universités[a]	INSA Lyon / INL	Co-directeur de thèse
Marchand, Cédric	Maître de Conférences[a]	École Centrale de Lyon / INL	Co-directeur de thèse
Slesazeck, Stefan	Senior Researcher	NaMLab, Dresden, Allemagne	Invité
Giraud, Bastien	Ingénieur Chercheur	CEA-LIST	Invité

[a] 63ème section

UNIVERSITÉ DE LYON

Résumé

École Centrale de Lyon
Institut des Nanotechnologies de Lyon

PhD

Calcul éco-énergétique avec matériaux ferroélectriques intégrés pour les systèmes embarqués et à la périphérie de réseau

by Mayeul Cantan

L'inclusion de matériaux ferroélectriques dans les circuits intégrés suscite de plus en plus d'intérêt, notamment grâce à l'introduction d'oxydes de hafnium dopés au zirconium, compatibles avec les matériaux semi-conducteurs et les procédés de fabrication les plus modernes. Leurs propriétés ferroélectriques, combinées aux technologies CMOS conventionnelles, permettent de créer de nouvelles architectures de circuits innovantes. Le rapprochement physique de Mémoires non volatiles aux éléments de calcul ouvre de nombreuses possibilités d'amélioration de l'efficacité énergétique en réduisant les transferts de données, en diminuant la consommation d'énergie statique et en permettant l'utilisation de paradigmes de calcul normalement-éteint (normally-off computing).

Dans cette thèse, les matériaux ferroélectriques sont abordés du point de vue de la conception de circuits, offrant une explication élémentaire de leurs propriétés et des approches utilisées pour leur modélisation. Plusieurs architectures de circuits utilisant des matériaux ferroélectriques sont également présentées, utilisant des technologies de dépôt en bout de ligne, ainsi que de transistors à effet de champ ferroélectriques (FeFETs, Ferroelectric Field-Effect Transistor), avec les résultats de caractérisation électrique obtenus, le cas échéant. Enfin, des techniques d'exploration de l'espace de conception (DSE, Design-Space Exploration)et un outil interne d'évaluation des performances au niveau système sont combinés pour extraire des chiffres de performance projetés, afin de permettre une comparaison avec des technologies plus matures, tant au niveau du circuit que du système.

Les résultats obtenus correspondent à la conception de nouveaux circuits, dont certains ont été fabriqués en 130 nm et 28 nm, aux résultats des simulations DSE pour des paramètres tels que la fenêtre de mémoire et la consommation d'énergie, ainsi qu'à de multiples outils logiciels créés au long du projet.

Remerciements

Ce document porte peut-être mon nom sur la première page, mais je n'aurais pas pu le terminer sans les contributions directes et indirectes des personnes qui m'ont aidé et encouragé au long de ce périple. Je ne peux pas lister tout le monde sur cette page, mais à celles et ceux dont le nom n'est pas inscrit : merci pour votre aide et votre soutien tout au long de mon doctorat.

Je suis avant tout reconnaissant à mon directeur de thèse, Ian O'Connor, qui m'a accompagné tout au long de cette thèse, m'apportant soutien et conseils, et se montrant réactif lorsque je rencontrais des difficultés. Mes co-encadrants Cédric Marchand et Damien Deleruyelle ont été d'une grande aide pour discuter des simulations et de la physique des dispositifs ferroélectriques, pour la révision de ce document, ainsi que pour les contributions plus directes de Damien à la sous-section 2.2.2.

Mes remerciements s'étendent à mes estimés collègues et amis pour leur soutien au fil des ans : Clément Zrounba et Arnaud Poittevin m'ont aidé à réaliser le tracé du layout des circuits MAD200 discutés dans chapitre 3, et ont été source de nombreuses discussions scientifiques fructueuses. Etienne Dupuis, pour sa motivation contagieuse et son cœur sur la main, Adil Brik, Lucien Del Bosque, suivis par les générations suivantes de doctorants, étaient toujours prêts à aider, que ce soit dans le cadre du travail ou pour des questions personnelles. Il en est résulté un environnement de travail accueillant, dans lequel je garde certains de mes meilleurs souvenirs.

Pour leur aide et leur soutien, je voudrais remercier le reste des membres de mon équipe à INL, y compris Alberto Bosio, pour son aide avec la synthèse de logique du du filtre d'image présenté dans section 4.6. D'autres membres du projet 3εFERRO ont également été d'une grande aide, y compris à INL, où Jordan Bouaziz et d'autres membres de l'équipe Matériaux, notamment Ingrid Cañero Infante, Bertrand Vilquin et Pedro Rojo Romeo, m'ont patiemment enseigné encore et encore les aspects physiques et de fabrication des matériaux ferroélectriques. Je dois également une partie de ma compréhension des matériaux ferroélectriques au personnel de NaMLab, en particulier Stefan Slesazeck et Evelyn Breyer, qui ont réalisé une grande partie du travail de conception et de mise en œuvre du filtre d'image convolutif présenté dans la section 4.6. Evelyn a joué un rôle déterminant dans le travail présenté dans ce document en fournissant le modèle de simulation ferroélectrique « Preisach » utilisé pour la plupart des simulations, bien que les premières discussions de modélisation et données expérimentales obtenues m'aient été gracieusement fournies par Carlotta Gastaldi de l'EPFL, ce qui a permis mes premières simulations basées sur le modèle « Landau ».

Pour m'avoir successivement aidé à développer la plate-forme d'évaluation des performances au niveau du système décrite dans la section 5.4, je voudrais remercier, dans l'ordre chronologique, les étudiants en master Pierre-Etienne Polet et Luca Mozzone, et le chercheur postdoctoral Marcello Traiolla, pour avoir successivement fourni l'essentiel de l'effort de développement pour la mise en œuvre des architectures que je proposais, et m'avoir fourni d'excellentes critiques à leur sujet. Leurs efforts de documentation ont été extrêmement utiles pour la rédaction de cette section.

Pour leur soutien et leurs encouragements constants, je voudrais également remercier ma famille, y compris mes parents et mes frères et sœurs, qui m'ont constamment taquiné sur l'avancement de mon manuscrit, tout en faisant de leur mieux pour me soutenir pendant ces moments difficiles. L'écriture de ce manuscrit n'a pas été une tâche facile, ni pour moi ni pour les personnes qui m'entourent, c'est pourquoi je suis aussi très reconnaissant envers Liz pour sa patience, sa compréhension et son soutien.

J'ai également trouvé un soutien inconditionnel auprès de mes amis, que je n'ai pas oubliés, et qui me manquent après ce temps passé à travailler de manière isolée. Vous vous reconnaîtrez, et j'espère vous revoir bientôt ! Je tiens à remercier tout particulièrement Claire Segovia pour avoir pris de son temps pour m'aider à travailler sur les premières étapes de la rédaction de ce manuscrit, et pour m'avoir aidée à surmonter le syndrome de la page blanche et à rejeter les distractions, ce qui m'a été extrêmement utile.

De manière moins conventionnelle, j'aimerais également remercier Romano Giannetti, le responsable de la bibliothèque d'illustrations de circuits circuitikz que j'ai largement utilisée tout au long de ce document, pour avoir répondu à ma demande de création de symboles

pour les condensateurs et transistors ferroélectriques[1]. Ces remerciements s'étendent aux mainteneurs des innombrables outils et logiciels libres utilisés au cours de cette thèse, ainsi qu'à Laurent Carrel pour l'administration de nos ressources informatiques locales.

Enfin, je remercie la Commission européenne pour le financement de projets intéressants et pertinents, qui m'a permis de réaliser ce travail dans le cadre du projet 3εFERRO qui a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de l'accord de subvention N°780302.

[1] https ://github.com/circuitikz/circuitikz/issues/515

Table des matières

Table des figures

Liste des Circuits

Liste des extraits de code

Chapitre 1

Introduction

Contents

1.1 À propos de ce document

1.1.1 Licence

1.1.2 Liens internes et code couleur

Ce document contient des liens internes et externes qui peuvent être suivis par un lecteur compatible. Ces liens sont codés par couleur comme suit :

1. Liens internes vers les sections, les figures, les extraits de code et les différentes parties du document
2. Liens internes vers la section Bibliographie
3. Liens internes vers le Glossaire et la liste des Acronymes
4. Liens externes vers des sites web

1.1.3 Code source du document, errata et matériel supplémentaire

Une révision à jour de ce document, ainsi que son code source LaTeX et des documents supplémentaires (extraits de codes et données additionnelles) sont disponibles à l'adresse `https://these.mayeul.net` [1].

1.1.4 Objectif du présent document

Ce document a pour but de résumer le travail accompli et les résultats obtenus au cours de ce doctorat, ainsi que de servir de document d'introduction à la conception de circuits ferroélectriques.

1.2 Contexte

Pendant des décennies, l'industrie de la microélectronique a connu une croissance exceptionnelle tant en termes de taille que de fonctionnalités. Des améliorations technologiques constantes ont été réalisées grâce à la miniaturisation, apportant à la fois des gains continus de performance et d'efficacité. L'industrie est aujourd'hui structurée autour de ces gains escomptés, mais s'il y avait initialement « beaucoup de place en bas » (pour citer Feynman) pour la miniaturisation, ce n'est plus le cas, et la poursuite de l'augmentation de la densité des circuits intégrés entraînera maintenant davantage de changements de géométrie, puisque les Fin FETs (FinFETs) sont désormais courants dans la plupart des nœuds avancés, et que des nanofils verticaux[Poi22] sont envisagés.

Une autre voie possible pour accroître les performances et l'efficacité énergétique n'implique pas d'augmenter la densité, mais plutôt de permettre de nouvelles fonctionnalités à un niveau de miniaturisation similaire. C'est le cas des modifications apportées aux dispositifs tels que les transistors silicium entièrement appauvri sur isolant (FDSOI, Fully Depleted Silicon on Insulator) et transistors à effet de champ ferroélectriques, mais aussi des modifications apportées aux architectures informatiques, qui peuvent apporter des gains de performance tangibles.

Outre les considérations économiques (les nouvelles technologies stimulent les ventes en rendant obsolètes les générations précédentes d'appareils), les gains d'efficacité énergétique sont importants pour compenser l'empreinte écologique et énergétique croissante des appareils électroniques. Ces gains sont compensés dans une certaine mesure par un effet de rebond, car des appareils moins chers, plus puissants et plus économes en énergie permettent de nouveaux cas d'utilisation. Cela entraîne toutefois une diversification des utilisation possibles qui est difficilement quantifiable, et rend la technologie financièrement plus abordable.

Dans cette introduction, les tendances actuelles en microélectronique sont résumées, en soulignant pourquoi l'approche historique n'est plus suffisante pour l'obtention de gains d'efficacité. De nouveaux cas d'utilisation tels que l'Internet des Objets (IoT, Internet of Things) et le calcul de périphérie sont décrits, ainsi qu'une technologie émergente spécifique, basée sur les $HfZrO_2$ ferroélectriques, ainsi que le projet européen 3εFERRO dont ce travail fait partie.

[1] miroirs disponibles à `https://thesis.mayeul.net` `https://mayeulc.gitlab.io/thesis`, `https://thesis.mayeul.cantan.eu` et `https://mayeulc.github.io/thesis`

1.2.1 Internet des Objets et calcul de périphérie

Internet des Objets

L'IoT comprend une multitude d'appareils et de catégories d'appareils. Il s'agit généralement de capteurs mesurant des quantités physiques telles que la température, l'humidité ou la pression atmosphérique, ou des valeurs plus spécifiques telles que la consommation énergétique, le nombre de voitures dans un parking, les capteurs de mouvement dans le cadre d'un système d'alarme, etc. Des dispositifs plus complexes peuvent également être considérés comme faisant partie de l'IoT, tels que les voitures modernes connectées à Internet, ou les dispositifs « intelligents » de plus en plus courants dans les maisons, tels que les thermostats, bouilloires, lumières, rideaux, litières pour chats, etc... connectés au réseau. Ces dispositifs comprennent parfois des actionneurs ainsi que des capteurs, dont certains peuvent être essentiels à la sécurité des personnes, comme une barrière contrôlant l'accès à une rampe de sortie d'urgence sur une autoroute.

L'IoT est un marché en pleine expansion, car la collecte de données et l'accès à des données supplémentaires depuis Internet permettent des comportements beaucoup plus complexes. Par exemple, l'utilisation de capteurs de température et d'humidité du sol peuvent aider à surveiller les cultures ; et la connexion d'un système d'irrigation aux prévisions météorologiques peut permettre d'économiser de l'eau. Cependant, contrairement au cloud, qui est composé de serveurs puissants et coûteux situés dans des centres de données connectés à un réseau électrique, l'IoT est généralement bon marché, non réutilisable, avec des contraintes d'énergie et de connectivité. Le coût et l'efficacité énergétique signifient que les dispositifs IoT ont généralement des quantités très limitées de mémoire et de puissance de calcul. En effet, qu'il s'agisse d'un capteur d'humidité placé au milieu d'un champ durant plusieurs années, d'un capteur de pression sur la jante d'une roue de voiture ou d'un capteur d'ouverture/fermeture en haut d'une fenêtre, l'alimentation électrique de ces dispositifs est souvent difficile, ce qui conduit à un grand nombre de dispositifs fonctionnant sur batterie, parfois complétés par des mécanismes de récupération d'énergie : panneaux solaires, triboélectricité ou collecteurs d'énergie vibratoire, modules peltier ou thermoélectriques pour récupérer la chaleur perdue, etc. Cela signifie également que ces appareils s'appuient souvent sur une connexion sans fil pour transmettre des données, ce qui nécessite une quantité d'énergie relativement importante : selon sa fiche technique[Esp22], un ESP8266EX très répandu consomme généralement environ 50 mW avec son modem radio éteint, tandis que la transmission de données à 15 dBm utilise un ordre de grandeur supplémentaire d'énergie. Lorsque son processeur est éteint, cet appareil ne consomme que 0.5 mW à 20 mW.

Calcul de périphérie

L'IoT utilise largement le cloud comme source et destination de données. Toutefois, comme indiqué précédemment, la transmission de données a un coût énergétique non négligeable pour les appareils alimentés par batterie. Cela devient également un problème de bande passante lorsque les données atteignent les serveurs : les coûts de transit sur Internet peuvent être considérables, et les coûts de traitement augmentent avec la quantité de données.

L'calcul de périphérie vise à relocaliser certaines tâches de traitement en dehors du cloud, à sa périphérie. Il s'agit du traitement de données qui est trop intensif pour être effectué sur les appareils IoT eux-mêmes en raison de contraintes de ressources, mais pour lequel l'application d'une certaine quantité de prétraitement peut réduire la latence et la bande passante envoyées au cloud.

En pratique, le traitement des données au plus près de la source permet généralement de réaliser des gains d'efficacité. Par exemple, dans un système de vidéosurveillance, une caméra sans fil (dispositif IoT) peut éviter de transmettre des données lorsqu'aucun mouvement n'est détecté, et compresser le flux vidéo avant de l'envoyer pour économiser de l'énergie lors de la transmission. Ce flux serait ensuite envoyé à un dispositif intermédiaire plus puissant (*dispositif de périphérie*) pour un traitement ultérieur. Il peut s'agir d'une tâche de reconnaissance d'images, d'identification d'un renard dans un élevage de poulets ou d'une personne armée dans une foule. Ce dispositif périphérique effectue donc des tâches de traitement relativement intensives pour pré-filtrer les données avant de les envoyer à des serveurs distants

pour un traitement ultérieur. Dans l'exemple ci-dessus, il peut s'agir d'archiver la vidéo ou de la transmettre à un autre appareil.

Un autre exemple est un assistant vocal pouvant effectuer une reconnaissance vocale locale basique, pour répondre plus rapidement aux commandes, mais aussi pour transmettre du texte plutôt que de la voix au cloud ; cela présente également des avantages en termes de disponibilité et de respect de la vie privée.

1.2.2 Fin de la mise à l'échelle de Dennard et de la loi de Moore

Loi de Moore

La loi de Moore a été formulée en 1965[Moo65] par Gordon Moore (cofondateur d'Intel, 1929–2023), selon laquelle le nombre d'élément (c'est-à-dire de transistors) sur un circuit intégré doublait tous les deux ans pour les puces les plus économiques (au coût le plus faible par élément). Cette observation est généralement résumée par un doublement de densité de transistor tous les deux ans, et s'est maintenue bien plus longtemps qu'initialement prévu. Cela est en partie dû au fait que l'industrie l'ait utilisée comme feuille de route pour le développement, comprenant les investissements dans de nouvelles usines, les dépenses de R&D et la planification des produits.

Mise à l'échelle de Dennard

Après la transition de l'industrie vers les transistors Transistor à Effet de Champ Métal-Oxide-Semiconducteur (MOSFET), la mise à l'échelle de Dennard a été identifiée en 1974[Den+74] comme un moyen faisable de continuer à augmenter la densité de transistors sur les puces électroniques, en réduisant leurs dimensions tel que spécifié dans le tableau 1.1. Ces valeurs sont établies en réduisant la taille des transistors tout en maintenant le champ électrique constant.

Paramètre du circuit ou dispositif	Facteur d'échelle
Dimensions des dispositifs t_{ox}, L, W	$1/\kappa$
Concentration en dopants N_a	κ
Tension V	$1/\kappa$
Courant I	$1/\kappa$
Délai par circuit $V \cdot C/I$	$1/\kappa$
Dissipation de puissance par circuit $V \cdot I$	$1/\kappa^2$
Densité de puissance $V \cdot I/A$	1

Tab. 1.1 : Paramètres de mise à l'échelle de Dennard et impact sur les performances, d'après [Den+74]. La densité de puissance est maintenue constante.

Cette méthode de mise à l'échelle a conduit à une constante augmentation de la densité au cours des années suivantes, la période de 1980 à 1995 étant surnommée l'ère de la « mise à l'échelle facile » (happy scaling). Grâce à l'amélioration des méthodes de fabrication, les dimensions des dispositifs et des circuits ont pu être réduites à un rythme régulier. Cette réduction de la taille des composants et des interconnexions a permis de diminuer les capacités de grille, augmentant ainsi simultanément la fréquence maximale de fonctionnement et l'efficacité énergétique en réduisant les pertes par commutation.

Afin de respecter la trajectoire de la loi de Moore, κ a été choisi pour réduire de moitié la surface des transistors $A \approx L \cdot W$ à chaque génération : $1/\kappa^2 = 1/2 \implies \kappa = \sqrt{2} \approx 1.4$[Boh07], approximativement tous les deux ans.

Conséquences d'une « mise à l'échelle facile »

Les gains de performance constants eurent un effet profond sur l'ensemble de l'industrie : il était souvent possible de réutiliser des modèles de puces antérieurement conçues, de les redimensionner, puis de les fabriquer avec le processus amélioré afin d'obtenir des gains de performance. En outre, les algorithmes conçus pour des processeurs non spécialisés pouvaient

également bénéficier de ces gains de performance, à condition que la rétrocompatibilité soit maintenue d'une génération de puces à l'autre. Le maintien de la rétrocompatibilité avec les précédentes générations de matériel et de logiciels s'est avéré bénéfique, permettant des améliorations itératives et répercutant l'impact des investissements sur plusieurs générations.

En conséquence, le développement des architectures de processeurs non spécialisées et les algorithmes, outils et compilateurs associés ont dépassé celui du matériel à usage spécifique : dans certains cas, des circuits ASICs et architectures de processeurs personnalisées pouvaient offrir des avantages en termes de performances et d'efficacité par rapport aux implémentations sur processeur (CPU, Central Processing Unit). Cependant, la longue et coûteuse phase de conception combinée à l'augmentation constante des performances des CPU pouvait rendre ces derniers aussi performants qu'une solution spécifique avant que celle-ci ne soit disponible.

Notre écosystème matériel et logiciel actuel doit être observé sous cette perspective historique, et ce travail reflète le récent regain d'intérêt pour les approches alternatives.

Fin de la mise à l'échelle de Dennard

Il est impossible de continuer indéfiniment la mise à l'échelle de Dennard. En effet, des limites techniques, de fabrication et physiques interviennent : les tensions ne peuvent pas descendre en dessous d'un certain niveau, car cela dégrade le rapport signal/bruit des niveaux logiques et de leur amplification par les transistors. Les concentrations de dopants deviennent difficiles à augmenter à mesure que la taille diminue, ce qui entraîne des problèmes tels qu'une distribution atomique inégale et un effet tunnel direct de bande à bande à travers les jonctions PN (source et drain du transistor)[Boh07]. Les dimensions ont également des limites fondamentales, car le courant d'effet tunnel devient non négligeable en dessous d'une certaine taille (en fonction de la tension et d'autres facteurs, environ 20 nm).

Alors que les dimensions de fabrications descendaient sous 65 nm, les courants de fuite continuaient d'augmenter, particulièrement au niveau des isolants de grille, alors d'une épaisseur de 1.2 nm, soit environ cinq couches atomiques de SiO_2 en 2005[Boh07]. Cela conduisit à l'introduction de diélectriques « high-k » afin de conserver l'intensité du champ électrique dans le canal du transistor malgré l'utilisation d'oxydes de grille plus épais. Cela permit de réduire le courant de fuite, mais rompu avec la mise à l'échelle de Dennard. Par la suite, la mise à l'échelle à champ constant rencontra davantage de problèmes qui nécessitèrent de repenser l'approche.

La fin de la mise à l'échelle de Dennard eut un effet notable sur la dissipation de puissance des circuits intégrés : les tensions ne pouvant plus être abaissées davantage, et les courants de fuite augmentant, la densité de puissance ne pouvait plus rester constante. Cela entraîna une augmentation considérable de la consommation d'énergie des puces et processeurs modernes, bien que l'efficacité énergétique puisse encore être améliorée grâce à l'augmentation des performances. Les courant de fuite peuvent également être atténués en désactivant l'alimentation des zones inutilisées des puces, une approche intitulée « silicium sombre » (dark silicon).

Fin de la loi de Moore

Bien que la réduction des dimensions des transistors put continuer plusieurs années après la fin de la mise à l'échelle de Dennard, les limites de la technologie planaire MOSFET étaient proches. En l'absence d'une feuille de route claire pour la mise à l'échelle, les investissements nécessaires pour suivre la feuille de route tracée par Moore ont continué à croître rapidement, accélérant la consolidation de l'industrie des semi-conducteurs.

Plus récemment, les augmentations de densité n'ont pas été à la hauteur des prédictions de Moore et ont été obtenues grâce à des modifications plus importantes de l'architecture des transistor à effet de champ (FET, Field Effect Transistor), avec la généralisation de FinFET et le développement de FET à grille englobante (GAAFET, Gate-All-Around FET) et de FETs à nanofils verticaux. Ces modifications permettent d'augmenter la densité de transistors en exploitant la dimension verticale, sans réduction majeure de dimensions.

L'augmentation de la taille des puces est un autre moyen de continuer à augmenter le nombre d'éléments par puce, avec des progrès récemment effectués dans le domaine des puces à l'échelle de la plaquette. Toutefois, il est peu probable que cela compense l'augmentation des coûts des dernières générations de procédés de fabrication de semi-conducteurs, qui reflète les investissements importants nécessaires à l'augmentation continue de la densité de transistors.

Les futures améliorations de performance pourraient continuer à provenir de l'augmentation de la densité et du nombre de structures (à une cadence ralentie), mais également de nouvelles architectures de systèmes et de nouveaux dispositifs, possiblement non électriques (photonique, stockage moléculaire). Diverses appellations ont été utilisées pour décrire ces approches respectives, notamment « plus de Moore » et « plus que Moore ».

L'amélioration de l'efficacité énergétique est peut-être un défi plus difficile à relever que l'augmentation de densité, sans l'assistance de la mise à l'échelle de Dennard, qui s'effectue à densité de puissance constante. Les problèmes de dissipation thermiques sont aggravés par l'abandon des transistors planaires, qui permettaient de maximiser la surface de contact avec le substrat de silicium.

Reflétant ce changement d'orientation, l'industrie cherche à se réinventer, et étudie de nombreuses technologies émergentes, y compris certaines pistes abandonnées durant l'ère de la « mise à l'échelle heureuse ». La cadence réduite de l'amélioration des performances des transistors conventionnels pourrait ainsi permettre aux alternatives de rattraper leur retard en termes de performances. Ce travail présente l'un de ces axes de recherche sur les matériaux ferroélectriques, afin d'apporter de nouvelles fonctionnalités aux technologies de semi-conducteurs existantes.

L'International Technology Roadmap for Semiconductors (ITRS), qui coordonne le développement des prochaines génération de dispositifs à base de silicium dans l'industrie, a également illustré cette tendance en 2016, changeant de nom pour l'International Roadmap for Devices and Systems (IRDS). Les rapports de cet organisme offrent une vue d'ensemble des technologies en cours de développement et de leurs applications potentielles[IRDS22].

1.2.3 Architecture Von Neumann

Les architectures d'ordinateurs sont aujourd'hui sont majoritairement conçus autour du concept de programmes stockés : en stockant les programmes dans la mémoire de travail, ceux-ci peuvent facilement être chargés et modifiés, ainsi que copiés et transférés. Cela rend nos ordinateurs bien plus polyvalents, car il n'est ainsi pas nécessaire de leur apporter de modifications physiques afin d'exécuter une nouvelle fonction, contrairement aux premières architectures.

L'une des architectures informatiques suivant ce paradigme les plus utilisées est l'architecture Von Neumann, illustrée par la figure 1.1, et ses dérivées[Paw22]. Portant le nom de John von Neumann, celle-ci lit instructions et données depuis la même mémoire, via d'un bus mémoire. Cette architecture peut être généralisée en plaçant plusieurs périphériques, y compris des dispositifs d'Entrée/Sortie (I/O, Input/Output), sur ce bus mémoire (ou bus système) à des adresses préalablement définies, et communiquer avec grâce à la même interface composée d'un bus de données et du bus d'adresses, effectuant des opérations de lecture et d'écriture. Cette généralisation a permis de rendre les architectures modernes très modulaires, utilisant une abstraction commune pour la plupart des cas d'usage. Les « périphériques » tels que les adaptateurs d'I/O supplémentaires, les coprocesseurs ou les dispositifs spécialisés peuvent communiquer avec le CPU et le programme qui s'y exécute en s'interfaçant avec ce bus mémoire unifié.

1.2.4 Goulot d'étranglement de Von Neumann

En raison de la position centrale du bus système dans les architectures Von Neumann et dérivées, celui-ci opère un rôle majeur dans la plupart des opérations des processeurs, limitant les performances du système par son empreinte physique à la conception, sa vitesse de transfert et sa consommation d'énergie. À mesure que la vitesse des circuits augmente, la quantité de données à transférer pour ne pas affamer le processeur s'accroît, ce qui sollicite encore davantage la mémoire et les bus système.

Les processeurs modernes utilisent une hiérarchie de caches (ou mémoire tampon) à plusieurs niveaux afin de réduire les accès à la mémoire externe pour diminuer la latence et augmenter la disponibilité des bus de mémoire, tout en réduisant la consommation d'énergie. Cependant, l'utilisation de caches est insuffisante, car leur efficacité dépend fortement de l'algorithme les utilisant. En outre, l'augmentation de la densité des mémoire vive statique (SRAM, Static Random Access Memory) semble ralentir[Sch22] malgré la croissance continue

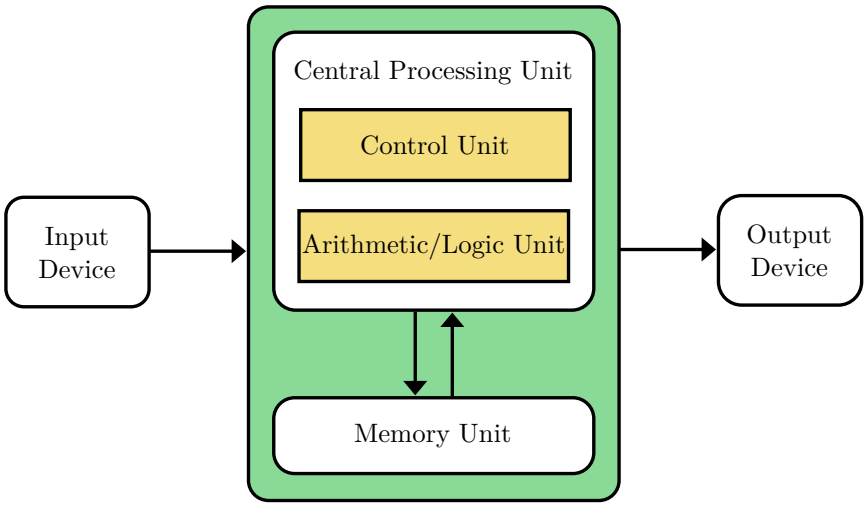


FIG. 1.1 : Diagramme de l'architecture Von Neumann, adaptation CC-BY-SA de [Kap13].

des tailles de mémoire cache, qui atteint plusieurs gigaoctets sur les modèles commerciaux récents grâce à des technologies telles que l'empilement 3D[Wuu+22]. Le nombre de bus système et de contrôleurs de mémoire augmente également, ce qui rend l'acheminement des signaux plus complexe et nécessite une plus grande quantité de connecteurs d'I/O sur les puces[Bec+18].

Une autre préoccupation majeure est la fraction croissante du budget de puissance utilisée pour la transmission de données. Les besoins en énergie, mesurés en $\mathrm{J\,bit^{-1}}$, augmentent linéairement avec la bande passante. Cela fait de la transmission de données l'une des plus grandes sources de dissipation thermique par les microprocesseurs[BI13], malgré une gestion approfondie du budget énergétique[Bec+18].

De nombreux autres moyens de diminuer l'impact de ce goulot d'étranglement sont à l'étude, notamment la transmission non électrique de données par radiofréquence[Cha+08] ou interconnexion optique[Liu+14], ainsi que des architectures alternatives fusionnant traitement et mémoire pour réduire le besoin de transmission de données. Cette thèse se concentre sur cette dernière approche.

1.2.5 $HfZrO_2$ ferroélectrique

Les oxydes de hafnium et de zirconium sont utilisés depuis de nombreuses années[WSW00] comme diélectriques à haute permittivité (high-k) dans les applications de semi-conducteurs, en particulier en dessous de 45 nm, comme utilisé par Intel depuis 2007 avec sa ligne de processeurs « Penryn » comportant des oxydes de grille high-k à base de hafnium.

En 2006, lors du développement de matériaux diélectriques pour des applications de condensateurs mémoire vive dynamique (DRAM, Dynamic Random Access Memory) chez Infineon/Qimonda (Dresde/Allemagne), une variété de films minces dopés HfO_2 et ZrO_2 a été analysée, révélant un comportement de commutation non standard dans le HfO_2 dopé au silicium, sous certaines concentrations ferrobook$_{2019\mathrm{P}}$refaces$_{\mathrm{S}}$CHROEDER$_{2019}$xvii. Une analyse plus approfondie a révélé que sous certaines conditions de dopage et de recuit, le matériau pouvait cristalliser dans une structure non centrosymétrique, compatible avec la ferroélectricité. Cela conduisit à une première série de publications en 2011[Bös+11a ; Bös+11b ; Mül+11], révolutionnant le domaine des dispositifs ferroélectriques à échelle nanométrique.

Divers dopants peuvent être utilisés afin de permettre aux oxydes de hafnium de cristalliser en une phase ferroélectrique, bien que le zirconium soit généralement privilégié en raison de sa faible température de recuit de 500 °C[Bou20, p. 144].

En effet, les oxydes de hafnium ferroélectriques présentent de multiples caractéristiques avantageuses par rapport aux matériaux ferroélectriques précédents :

- Compatibilité CMOS : il s'agit de la propriété la plus intéressante, car l'oxyde de hafnium était déjà largement utilisé dans les processus CMOS industriels comme matériau d'oxyde de grille.
- Faible champ électrique coercitif : de 0.5 MV cm^{-1} à 2.5 MV cm^{-1}[Bou20, p. 145], typiquement 1.2 MV cm^{-1}, cela permet des tensions coercitives proches 1.2 V pour une épaisseur d'oxyde typique de 10 nm, ce qui est compatible avec les niveaux logiques des CMOS.
- Faible température de recuit : à 500 °C, $HfZrO_2$ peut être déposé au-dessus des dispositifs CMOS sans les endommager.
- Stable en température : la température de Curie peut facilement dépasser 200 °C[Bou20, p. 143], en fonction de la concentration de dopants et de la taille du transistor. Cette caractéristique est intéressante pour les applications de Mémoire non volatile.

1.2.6 Conclusion

Depuis la fin de la mise à l'échelle de Dennard vers 2005, la densité de puissance des circuits intégrés a augmenté, ce qui conduisit à la recherche de circuits et d'architectures plus éco-énergétiques. Cette situation est accélérée par la fin de la « loi » de Moore, où les transistors conventionnels « planaires » ont atteint des limites physiques de mise à l'échelle. De nouvelles pistes sont explorées pour augmenter la densité, avec des excursions dans la 3^{e} dimension pour la conception de circuits et de transistors. Cependant, ces approches augmentent également la densité de puissance, ce qui entraîne des problèmes de dissipation thermique. Parallèlement, l'augmentation des performances continue à mettre en évidence le goulot d'étranglement de l'architecture Von Neumann et de ses dérivées.

De nouvelles applications présentant des exigences extrêmement élevées en matière de performances ou d'efficacité énergétique ont également fait leur apparition : génétique, applications « big data », y compris l'apprentissage automatique (machine learning), ainsi que l'IoT, les réseaux de capteurs intelligents et intégrés, etc. Ces applications font de plus en plus appel à des architectures et à des dispositifs spécifiques plutôt qu'à des CPUs à usage général, inversant ainsi une tendance amorcée dans les années 1980. Les besoins en matière de calcul hautes performances sont désormais satisfaits par les processeur graphiques (GPUs, Graphics Processing Unit) et les processeur de tenseurss (TPUs, Tensor Processing Unit), ainsi que par d'autres ASIC conçus spécifiquement. Les besoins en efficacité énergétique sont de plus satisfaits par des ASICs et FPGAs, et des transistors FDSOI.

En conséquence, les alternatives à l'architecture de Von Neumann et aux circuits CMOS connaissent aujourd'hui un regain d'intérêt.

1.3 Projet Européen 3εFERRO

Les travaux présentés dans ce document s'inscrivent dans le cadre d'un projet européen plus vaste : 3εFERRO (pour *Energy Efficient Embedded Non-volatile Memory Logic based on Ferroelectric Hf(Zr)O_2*). Dans le contexte de ses activités de diffusion, le projet héberge un site Web à l'adresse `https://www.3eferro.eu` et a commandité une vidéo d'introduction disponible à l'adresse `https://youtu.be/M8tL-nN7G-A`, résumant efficacement les objectifs et le contexte du projet.

1.3.1 Partenaires du projet

Le projet compte huit participants, avec des domaines d'expertise variés. Notre équipe a donc interagi davantage avec certains partenaires. Nos plus étroites collaborations figurent au-dessus de la liste des partenaires présentée ci-dessous :

1. INL, en tant que ECL, Lyon, France
2. NaMLab, Dresden, Allemagne
3. STMicroelectronics, Grenoble, France

4. CEA-LETI, Grenoble, France
5. EPFL, Lausanne, Suisse
6. Demokritos, Athènes, Grèce
7. NIMP, Bucharest, Roumanie
8. FZJ, Jülich, Allemagne

1.3.2 Objectifs du projet

Ce projet a une large portée, reflétée par les partenaires susmentionnés. Son objectif principal est de développer des technologies liées à l'intégration des $HfZrO_2$ ferroélectriques dans les processus technologiques de la microélectronique.

L'approche peut être considérée comme ascendante, commençant par la science des matériaux et la caractérisation des oxydes déposés, ainsi que l'optimisation du processus de fabrication et des propriétés des matériaux ; puis par la conception, la caractérisation et la modélisation des dispositifs et des circuits, jusqu'aux démonstrateurs technologiques et aux prévisions de performance des architectures conçues.

Accomplissements

Ce projet a permis de faire progresser l'état de l'art en matière d'intégration des matériaux $HfZrO_2$ ferroélectriques.

Les méthodes de dépôt[2] ont été améliorées[Fra+19b] afin d'optimiser les performances de la couche ferroélectrique, y compris l'optimisation des cycles de recuit, de la température et de la pression[Bou+19], ainsi que la composition du matériau et du substrat[Zac+22]. Beaucoup d'efforts ont été consacrés à l'amélioration de la fiabilité et du rendement, à l'obtention de caractéristiques uniformes sur les dispositifs et les wafer, en utilisant des méthodes de caractérisation électrique et microscopique avancées, notamment la microscopie à force de réponse piézoélectrique, ainsi que la spectroscopie photoélectronique à rayons X durs et mous, perettant l'obtention de meilleurs modèles. Les propriétés des dispositifs ont été étudiées, particulièrement celles des Condensateur Ferroélectriques (FeCaps) et transistors à effet de champ ferroélectriques (FeFETs, Ferroelectric Field-Effect Transistor), ainsi que la capacité négative[Gas+19] pour une utilisation dans les transistor à effet de champ à capacité négatives (NCFETs, Negative-Capacitance Field-Effect Transistor). Plusieurs circuits et démonstrateurs ont également été réalisés, y compris des tableaux mémoire de 16 kbit[Fra+19a ; Fra+21].

Le projet a démontré la viabilité des dispositifs de mémoire à base d'hafnie dopée et leur compétitivité par rapport aux mémoires flash en termes de vitesse, d'endurance, de rétention, de consommation d'énergie, de densité et de facilité d'intégration[Gre+20 ; Oku+21 ; Alc+22].

Contributions

Nos contributions, telles que décrites dans le présent document, se concentrent principalement au niveau d'abstraction du circuit, et plus élevé. Celles-ci comprennent la conception de circuits, détaillée dans chapitre 3, ainsi que la validation d'un démonstrateur de filtre d'image plus complexe, discutée dans le chapitre 4. Au niveau le plus élevé, une plateforme d'évaluation de performances au niveau du système a été réalisée. Cette plate-forme a pour vocation d'aider à prédire les performances du système sur la base des données de simulation et de caractérisation disponibles au niveau du dispositif, comme décrit dans chapitre 5.

[2] Incluant dépôt en couches atomiques (ALD, Atomic Layer Deposition), ablation laser pulsé (PLD, Pulsed Laser Deposition), dépôt physique par phase vapeur (PVD, Physical Vapor Deposition), dépôt par jet moléculaire (MBD, Molecular Beam Deposition)

Chapitre 2

Ferroélectriques : comportement et modélisation

Contents

2.1 Ferroélectricité

La ferroélectricité est aux champs électriques ce que le ferromagnétisme est aux champs magnétiques : lorsqu'elle est soumise à un champ électrique suffisamment fort ($E > E_C$), la structure interne d'un matériau ferroélectrique se réarrange, faisant de lui un dipôle électrique orienté suivant le champ électrique externe. Cette polarisation est conservée après la disparition du champ électrique, et peut être inversée ultérieurement en appliquant un champ électrique d'orientation opposée.

2.1.1 Cristaux ferroélectriques

En réponse à une contrainte externe, la géométrie ou l'organisation interne d'un réseau cristallin est modifiée, ce qui change la densité de charge ρ à la surface du matériau, s'il est composé de dipôles électriques. Cette modification de la densité de charge affecte à son tour

le potentiel électrostatique mesuré à la surface du cristal, en fonction du type de stimulus et de la réponse du matériau.

Plus précisément, selon le type de stimulus capable de changer le potentiel de surface ou la densité de charge de surface d'un cristal, celui-ci peut être classé en trois catégories imbriquées, comme illustré dans le tableau 2.1 :

1. les cristaux piézoélectriques modifient leur potentiel de surface en réponse à une contrainte mécanique ;
2. les cristaux pyroélectriques sont piézoélectriques et les changements de température induisent également un changement mesurable de leur potentiel de surface ;
3. les cristaux ferroélectriques sont pyroélectriques ; en outre, le changement de potentiel de surface persiste après l'application d'un champ électrique puissant. La principale différence par rapport aux pyroélectriques est la capacité d'inverser leur polarisation sans de subir de claquage[Ihl19, p. 7].

La variation de la densité de charge superficielle peut être mesurée en formant une structure de type condensateur avec le matériau placé entre deux électrodes : une variation de tension devrait être observée en réponse au stimulus externe. Intuitivement, on peut comprendre que des charges sont présentes dans la structure cristalline et que divers stimuli externes provoquent un changement de la densité de charge (soit un déplacement des charges) près des électrodes, via un changement de géométrie. Ce changement de géométrie peut être compris comme une réponse à une contrainte mécanique (appliquée directement, pour la piézoélectricité, et indirectement par dilatation/contraction thermique pour la pyroélectricité). En outre, dans le cas d'un cristal ferroélectrique, ces charges peuvent atteindre plusieurs emplacements stables possibles dans le réseau cristallin, ce qui engendre un comportement bistable.

Il convient de noter que les ferroélectriques conservent la nouvelle polarisation de manière stable, mais réversible. Si la polarisation n'est pas conservée de manière stable, il s'agit de paraélectriques, comme indiqué dans la sous-section 2.1.3 et illustré par la figure 2.6c.

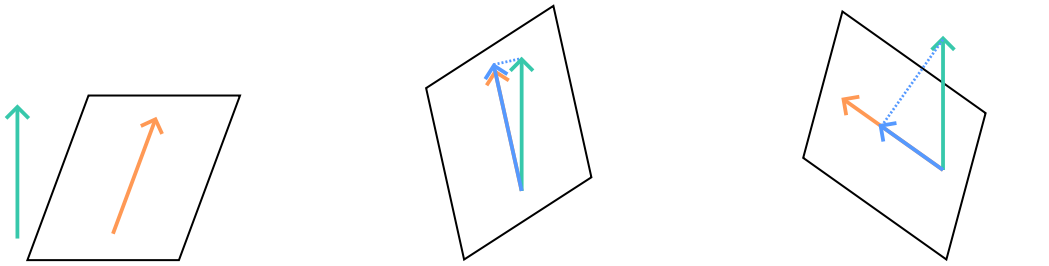
Stimulation	Piezzoélectrique	Pyroélectrique	Ferroélectriques
Stress mécanique	✓	✓	✓
Changement de température		✓	✓
Champ électrique			changement non-volatile

Tab. 2.1 : Classification cristalline en fonction du stimulus provoquant un changement de potentiel électrique à la surface. Notez que si les champs électriques induisent un changement de densité de charge de surface dans tous ces matériaux (comme pour un condensateur), les matériaux ferroélectriques conservent ce changement après la dissipation du champ s'il est supérieur à Champ Électrique Coercitif (E_C). De plus, la polarité peut être inversée en changeant l'orientation du champ électrique externe.

Champ coercif

La modification de la polarisation électrique d'un matériau ferroélectrique, nécessite l'application d'un champ électrique suffisamment intense, dépassant le seuil du Champ Électrique Coercitif (E_C), et associé à une Tension Coercitive (V_C). Cette valeur dépend des propriétés du matériau ferroélectrique et correspond à la force nécessaire pour déplacer les charges internes d'une position d'équilibre à l'autre, en surmontant les forces de cohésion internes. Ces forces résultent d'une variété de mécanismes, dont des effets électrostatiques et mécaniques, et liaisons chimiques. Ceci est illustré en figure 2.2 par un système équivalent charge-ressort.

La valeur du champ coercitif est une propriété intrinsèque du matériau, bien que divers facteurs puissent l'affecter. Une variation importante est l'orientation du domaine : la plupart des matériaux ferroélectriques étudiés étant polycristallins, ceux-ci sont constitués de multiples domaines cristallins orientés différemment. Comme l'illustre la figure 2.1, un champ électrique externe aura un impact moindre sur les domaines cristallins non alignés, augmentant leur valeur de E_C. Il en résulte une distribution de valeurs de E_C réparties sur chaque domaine plutôt qu'une valeur E_C unique par dispositif ferroélectrique.

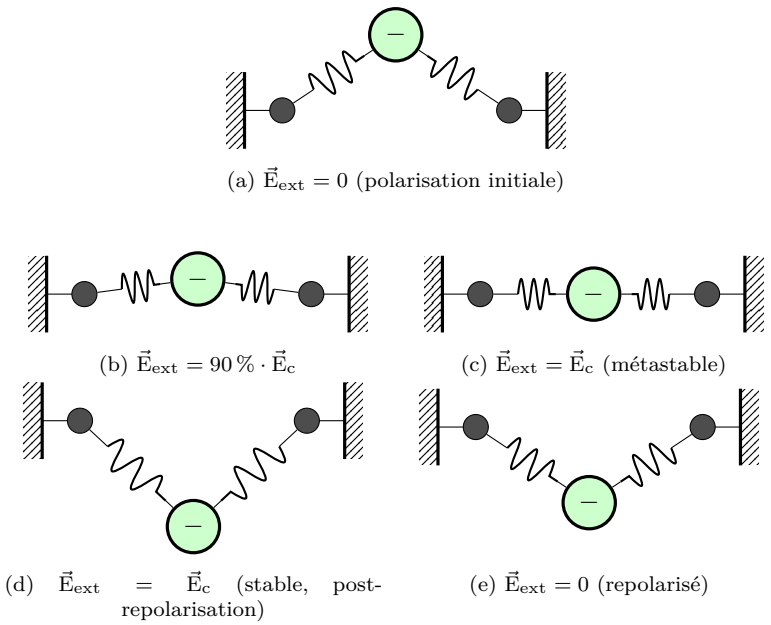


(a) Domaine ferroélectrique avec E_C associé, en présence d'un champ électrique externe.

(b) Domaine ferroélectrique orienté de manière quasi optimale.

(c) Domaine ferroélectrique avec une orientation sous-optimale.

FIG. 2.1 : Effet du décalage d'orientation entre le champ électrique externe et le cristal ferroélectrique. 2.1a, 2.1b et 2.1c montrent le même domaine ferroélectrique et le même champ externe, avec des orientations relatives différentes. Seul le cristal 2.1b, qui a un Champ Électrique Coercitif presque aligné sur le champ externe, peut être repolarisé par celui-ci. 2.1c nécessiterait un champ électrique quasiment deux fois plus intense.



(a) $\vec{E}_{ext} = 0$ (polarisation initiale)

(b) $\vec{E}_{ext} = 90\,\% \cdot \vec{E}_c$

(c) $\vec{E}_{ext} = \vec{E}_c$ (métastable)

(d) $\vec{E}_{ext} = \vec{E}_c$ (stable, post-repolarisation)

(e) $\vec{E}_{ext} = 0$ (repolarisé)

FIG. 2.2 : Illustration de la ferroélectricité comme système bistable charge-ressort. Lorsqu'ils sont comprimés ou trop tendus, les ressorts ramènent la charge à l'une des deux positions d'équilibre (figure 2.2a ou 2.2e). Un champ électrique suffisamment intense $\vec{E}_{ext} \geq \vec{E}_c$ (progressivement appliqué de la figure 2.2b à la figure 2.2d) permet de surmonter la force répulsive et de « faire basculer » la charge vers l'autre position d'équilibre, repolarisant ainsi le matériau ferroélectrique. Ceci illustre également comment les charges de surface d'un matériau ferroélectrique se déplacent vers le côté opposé pendant la repolarisation.

Polycristaux et domaines

En fonction du matériau et de la méthode de croissance cristalline, de multiples domaines cristallins peuvent se former. C'est le cas des méthodes de dépôt de $HfZrO_2$ actuelles, particulièrement au-delà d'une certaine taille.

Les technologies actuelles de dépôt de $HfZrO_2$, ne sont pas adaptées à la croissance de cristaux à domaine unique, pour de multiples raisons :

- Plus le cristal ferroélectrique est grand, plus le risque de division en plusieurs domaines est élevé. C'est pourquoi les monocristaux sont relativement petits, de l'ordre de 10 nm de diamètre.
- Il est difficile de contrôler l'orientation des cristaux ferroélectriques ; par conséquent, de multiples dispositifs auraient des caractéristiques différentes, certains pouvant être défectueux.
- Même si l'objectif est la production de monocristaux de petite taille, il est extrêmement probable d'obtenir de multiples domaines dans quelques dispositifs, particulièrement lorsque le nombre de dispositifs augmente. Cette variabilité est préjudiciable au rendement.

L'objectif de produire uniquement des monocristaux est donc actuellement irréalisable, principalement en raison de la difficulté de rendre le processus reproductible sur plusieurs dispositifs : le dépôt de monocristaux de même taille et de même orientation sur un grand nombre (des centaines à des milliards) de composants électroniques afin de maintenir des caractéristiques électriques similaires sur l'ensemble d'entre eux n'est actuellement pas réalisable.

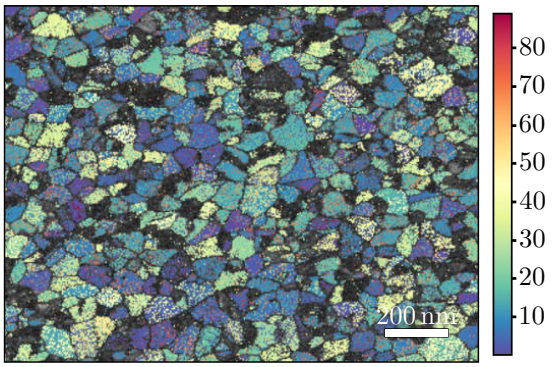


Fig. 2.3 : Mesure expérimentale de l'orientation des domaines ferroélectriques dans un matériau polycristallin HfO_2 dopé au Si. L'orientation est donnée en degrés, par rapport au plan de l'image. Image CC-BY [Led+20], citant un diamètre équivalent moyen de 28.5 nm et 33.9 nm pour le HfO_2 dopé avec Zr et Si, respectivement.

Il est actuellement préférable de déposer sur chaque dispositif une population de domaines ferroélectriques suffisamment importante pour que le comportement soit contrôlé par une statistique de distribution, qui peut être rendue similaire d'un dispositif à l'autre. La figure 2.3 illustre le problème : pour que la distribution des orientations des domaines soit semblable dans de multiples échantillons aléatoires, ceux-ci doivent être choisis suffisamment grands pour contenir une variété de domaines représentative de l'ensemble. L'utilisation de dispositifs polycristallins de plus grandes dimensions implique ce qui suit :

- Afin que la distribution soit similaire dans tous les dispositifs, ceux-ci doivent dépasser une certaine taille afin d'inclure plusieurs domaines ferroélectriques, actuellement de l'ordre de 200 nm de diamètre pour $HfZrO_2$[Led+20].
- La valeur du champ coercitif devient une distribution de valeurs multiples, car elle dépend de l'orientation relative de chaque domaine et du champ électrique externe, comme illustré par la figure 2.1. Une telle distribution simulée est illustrée en figure 2.14. L'effet de l'augmentation de la taille de l'échantillon est équivalent à celui de la distribution à grain de plus en plus fin illustrée par la figure 2.16.

- Ayant des E_C différents, les domaines ferroélectriques basculeront à des tensions différentes. La polarisation sera donc progressive, ce qui permet de polariser partiellement la population, comme le montrent les boucles mineures dans l'hystérésis P-V de la figure 2.15.
- Ces boucles mineures et la possibilité d'une polarisation partielle permettent d'envisager la conception de cellules cellule multi-niveaux (MLC, Multi-Level Cell).
- Les domaines étant sont plus ou moins réceptifs aux champs externes, cela peut entraîner des effets supplémentaires de réveil et de fatigue.

Note sur la nomenclature utilisée pour les opérations d'écriture et d'effacement

Les termes écriture (programmation) et effacement (déprogrammation) sont parfois utilisés dans la littérature pour désigner la direction d'un champ coercitif appliqué, et donc la direction de la polarisation ferroélectrique qui en résulte.

Le présent document désignera explicitement la polarisation résultante, utilisant ces termes ainsi :

1. Écriture ou programmation : opération consistant à appliquer un champ Champ Électrique Coercitif sur le matériau ferroélectrique. La direction du champ n'est pas précisée, à moins que la valeur enregistrée soit importante.
2. Effacement : les opérations d'effacement placent la polarisation dans une direction connue, soit non spécifiée, soit considérée comme « par défaut » pour le tableau de mémoire. La principale différence avec l'opération d'écriture est l'objectif, qui est ici d'effacer l'information précédemment stockée.
3. Repolarisation : se produit lorsque le matériau ferroélectrique change de direction de polarisation. Ceci peut se produire après une opération d'écriture ou d'effacement, mais pas nécessairement : aucune repolarisation n'a lieu si la polarisation était déjà alignée sur le Champ Électrique Coercitif appliqué.

2.1.2 Courbe *P-V*

L'un des outils les plus utilisés pour caractériser les ferroélectriques est la courbe $P = f(V)$, ci-après simplement appelée (P-V). Cette courbe représente la densité de charges électriques à la surface du matériau (potentiel de surface P, généralement exprimé en $\mu C\,cm^{-2}$) en fonction de la tension d'entrée. Expérimentalement, elle est souvent obtenue en intégrant le courant mesuré en soumettant le matériau à une tension déterminée, comme dans les mesures Positif-Haut, Negatif-Bas (PUND, Positive-Up, Negative-Down) décrites dans la sous-section 2.1.4. Celle-ci peut également être obtenue par d'autres moyens, tels que la microscopie à force atomique (AFM, Atomic Force Microscopy)[Hon+01].

En supposant qu'il n'y a aucun courant de fuite, l'intégration du courant d'alimentation permet d'obtenir la quantité de charges stockées sur une plaque de condensateur, et donc la polarisation de surface du diélectrique. Si le diélectrique du condensateur est constitué d'un matériau ferroélectrique, un déséquilibre peut être observé pendant la décharge du condensateur, si la polarisation a été modifiée : lorsque le potentiel de surface augmente d'un côté et diminue de l'autre, les électrons sont respectivement libérés et piégés sur les plaques, ce qui entraîne un pic de courant équivalent à un courant de fuite temporaire lors de la repolarisation. La mesure du nombre de charges libérées permet de déduire la polarisation résiduelle (Pr), c'est-à-dire la quantité de charges piégées à la surface, une valeur importante pour la comparaison de matériaux ferroélectriques.

Comme indiqué sur la figure 2.4, plusieurs données peuvent être extraites de la courbe P-V :

- Tension Coercitives, et par extension E_C.
- Polarisation résiduelle, généralement lu comme $2 \cdot Pr$ puisqu'il s'agit de la hauteur de la courbe

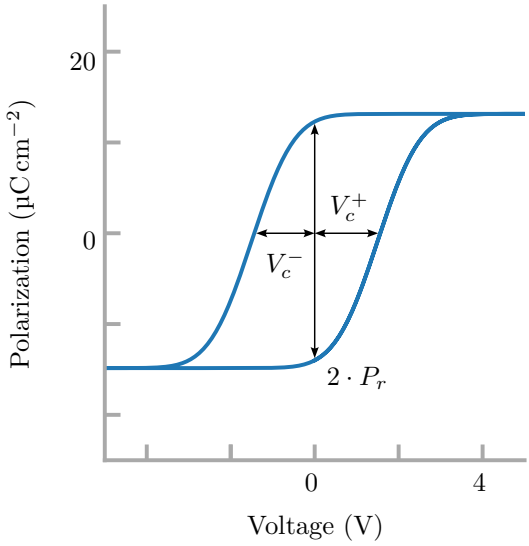


FIG. 2.4 : Lecture de Pr et V_C sur une courbe P-V. Notez que les valeurs de V_C et Pr obtenues dépendent de la tension appliquée si plusieurs domaines de différentes caractéristiques sont présents. Le cycle ici représenté est celui de la figure 2.9b, réutilisé à des fins d'illustration. Ce cycle n'est pas systématiquement centré horizontalement sur zéro pour des raisons d'empreinte.

- les contributions capacitives, résistives et antiferroélectriques, dans une certaine mesure, comme indiqué par la figure 2.5, et extraites à l'aide des formes d'onde PUND, comme décrit plus loin dans la sous-section 2.1.4.

Les limites des courbes P-V comprennent le fait qu'elles sont générées pour une tension donnée et que de multiples valeurs de tension peuvent induire des Pr différents, manifestés par des boucles concentriques. Toutefois, il s'agit d'une limitation mineure, car la tension utilisée est visible sur le graphique. Inversement, ces boucles mineures peuvent être exploitées, lorsqu'une tension plus faible est délibérément utilisée afin de ne polariser qu'une fraction des domaines ferroélectriques. Cela est généralement employé dans le cadre de MLC ou pour le contrôle analogique de la tension de seuil d'un FeFET. Une telle boucle mineure est illustrée dans la figure 2.15.

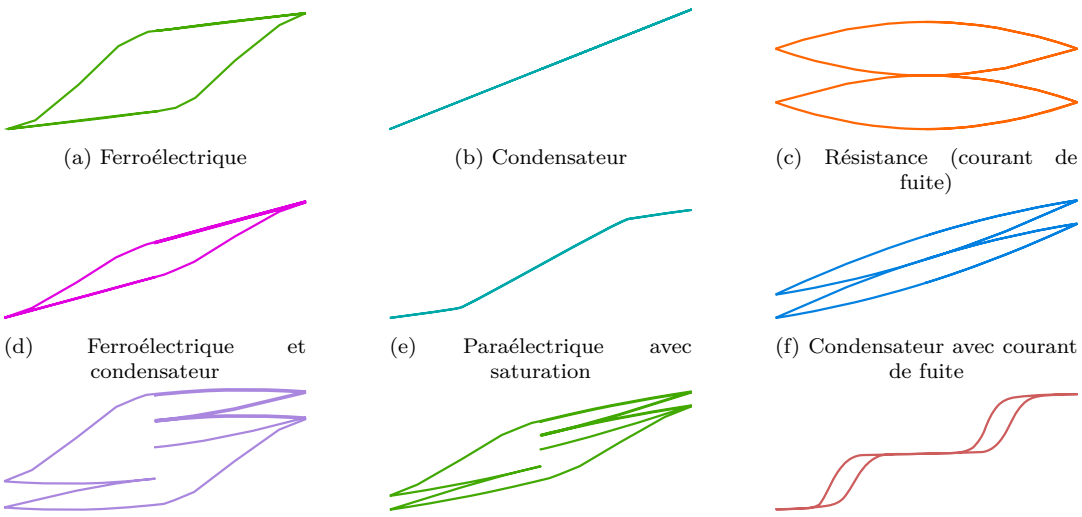
2.1.3 Relation avec la capacité et la paraélectricité

Les oxydes ferroélectriques peuvent être considérés comme des diélectriques dont la permittivité électrique ε_r varie en fonction du champ électrique actuel et de l'historique du champ électrique.

Un ferroélectrique réagit à un champ électrique externe en modifiant sa polarisation électrique interne. Cette propriété est étroitement liés à la paraélectricité et à l'antiferroélectricité, tous deux illustrés par la figure 2.6 et la figure 2.5. Toutefois, contrairement à ces derniers, un matériau ferroélectrique conserve une polarisation non nulle en l'absence de champ électrique externe.

Également diélectriques, toutes ces catégories de matériaux peuvent être utilisées afin d'isoler des plaques de condensateur. Leur réponse interne au champ électrique appliqué extérieurement renforce l'accumulation de charges à la surface des électrodes, ce qui augmente la capacité et accélère la vitesse de chargement. Le comportement exact dépend de la classe de matériau :

- Les matériaux paraélectriques ont un comportement est semblable à celui des condensateurs simples, la mobilité des charges internes ayant pour effet d'augmenter la capacité, ce qui donne un ε_r effectif plus élevé. La mobilité des charges peut être limitée à l'intérieur de la structure cristalline du matériau, ce qui restreint l'effet paraélectrique et abaisse



(a) Ferroélectrique (b) Condensateur (c) Résistance (courant de fuite)

(d) Ferroélectrique et condensateur (e) Paraélectrique avec saturation (f) Condensateur avec courant de fuite

(g) Ferroélectrique avec courant de fuite (h) Ferroélectrique avec courant de fuite et condensateur (i) Antiferroélectrique

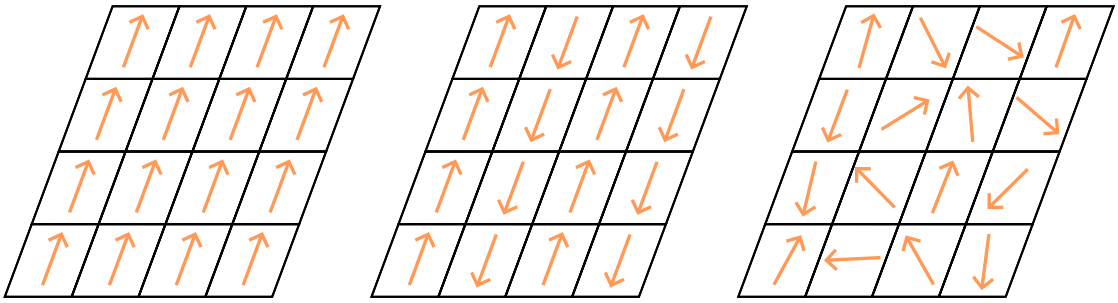
FIG. 2.5 : Sélection de courbes P-V pouvant être obtenues, centrées autour de 0 V. Ces courbes ont majoritairement été obtenues en simulation, avec C =20 fF de capacité, R =100 GΩ de résistance de courant de fuite, et une alimentation électrique de 50 Ω d'impédance. Il convient de préciser que l'échelle de temps de ces simulations est beaucoup plus importante que les différents τ, ce qui permet d'obtenir des courbes montrant le régime quasi stationnaire.

la valeur de la capacité après une certaine intensité de champ (tension d'alimentation), jusqu'à ce que le diélectrique ne claque.

- Dans le cas de matériaux ferroélectriques, tant que E_C n'est pas atteint ni dépassé, le condensateur se comporte comme un condensateur paraélectrique normal. Cependant, lorsque E_C est atteint :
  - si les domaines étaient déjà polarisés dans la même direction que le champ électrique appliqué, aucune repolarisation n'a lieu. Le condensateur continue à fonctionner comme un condensateur normal en dessous et au-dessus de E_C.
  - si les domaines n'étaient pas déjà polarisés dans la même direction que le champ électrique appliqué, ceux-ci commencent à se réorienter suivant le champ externe, en commençant par ceux dont le champ coercitif est plus faible (c'est-à-dire les domaines les mieux alignés sur le champ externe). Ce processus libère une grande quantité de charges, ce qui correspond à une courbe P-V beaucoup plus raide, comme on peut l'observer sur les côtés de la figure 2.4. Cela se poursuit tant que la tension continue d'augmenter et jusqu'à ce que la plupart des domaines soient polarisés.

  Comme dans le cas des paraélectriques, la mobilité des charges, permettant d'améliorer la capacité, peut atteindre ses limites, ce qui diminue la capacité jusqu'au claquage.

- Dans le cas des antiferroélectriques, l'effet paraélectrique est naturellement compensé par un dipôle électrique secondaire, qui annule la polarisation globale du matériau. Cela se traduit par une capacité théoriquement nulle (0 F) et, par conséquent, aucun changement de charge à la surface. Ce dipôle compensateur peut toutefois être renversé sous l'effet de champs électriques suffisamment intenses, à condition que le diélectrique ne claque pas. Cela s'observe sous la forme de cycles ferroélectriques plus petits de part et d'autre de la courbe P-V, comme le montre la figure 2.5i.



(a) Ferroélectrique (b) Antiferroélectrique (c) Paraelélectrique

FIG. 2.6 : Orientation de la polarisation électrique des domaines après l'application d'un champ externe intense (coercitif). La figure représente plusieurs domaines cristallins de trois types de matériaux différents. Le matériau ferroélectrique (2.6a) est aligné sur le champ externe, le matériau paraélectrique (2.6c) a rétabli ses orientations initiales aléatoires, et le matériau antiferroélectrique (2.6b) a spontanément compensé la nouvelle orientation.

Compensation des charges et champ de dépolarisation

La modification de la polarisation ferroélectrique rapproche des particules chargées de la surface de l'une des électrodes et les éloigne de la seconde, ce qui modifie la densité de charge P à la surface du matériau. Cela signifie que les deux électrodes ont une charge électrique non nulle (et de valeur opposée) après un changement de polarisation.

En conséquence :

1. Un champ électrique existe à travers le FeCap après la polarisation : le champ de dépolarisation E_{dep} ;
2. Les charges mobiles proches des électrodes seront attirées ou repoussées pour compenser, ou « faire écran » (charge screening), à la nouvelle charge, comme le montre la figure 2.7 ;
3. Les charges « capturées » pour compenser la polarisation précédente sont libérées, ce qui crée un courant de repolarisation mesurable.

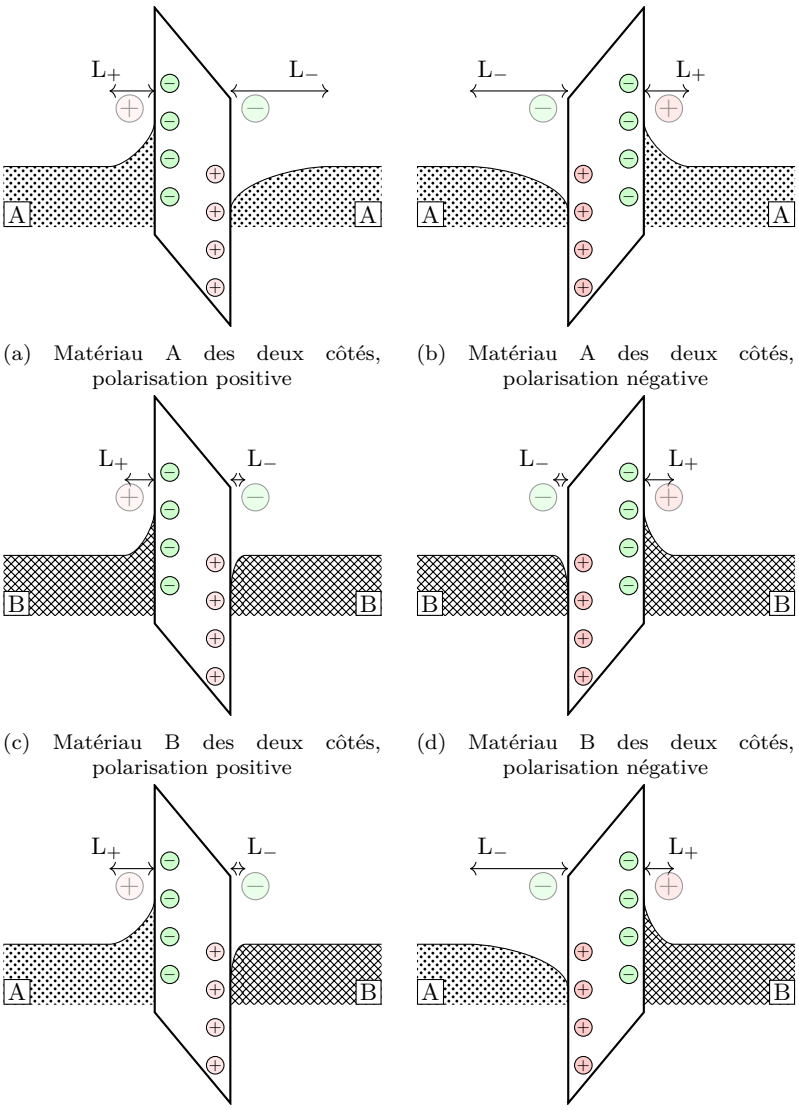
Le champ de dépolarisation qui en résulte est, comme son nom l'indique, orienté dans la direction opposée au champ récemment appliqué, et a donc pour effet de s'opposer à la nouvelle polarisation. Il peut s'agir d'une source importante de perte de rétention, en particulier lorsque E_{dep} est beaucoup plus grand que E_C. Les dispositifs ferroélectriques à base de $HfZrO_2$ ayant un champ coercitif plus important que les matériaux ferroélectriques plus traditionnels $Pb(Zr, Ti)O_3$ (PZT) ou $SrBi_2Ta_2O_9$ (SBT), la perte de rétention due au champ de dépolarisation est beaucoup plus faible[MG19], bien que E_{dep} soit plus important dans les films plus minces[Maj22].

Le second mécanisme empêche de mesurer la ferroélectricité à l'échelle macroscopique. Il a été suggéré[CL07] que cet effet de compensation de charges est responsable de la découverte beaucoup plus tardive de la ferroélectricité par rapport au ferromagnétisme. En effet, la distance de compensation est relativement faible et varie de 0.1 nm dans les métaux à quelques nm dans les matériaux moins conducteurs, selon [PLH21].

Cette compensation de charges a des effets très locaux et joue donc un rôle dans les structures FeFETs décrites plus loin dans la section 2.4. Les charges compensant la polarisation ne sont pas mobiles, et ne peuvent donc pas modifier la conductivité ou les tension de seuil (V_{th}) de MOSFETs si elles sont situées dans le canal[PLH21]. Les électrons temporairement piégés dans l'oxyde peuvent également jouer ce rôle de compensation, dégradant ainsi les performances des FeFETs. Ces charges créent par ailleurs des problèmes de lecture après écriture, signifiant que la nouvelle polarisation des FeFETs ne peut pas être lue immédiatement après l'écriture[Kle+21]. Ce piégeage est plus probable lorsque des porteurs chauds (en raison des tensions plus élevées utilisées pendant la programmation) traversent le ferroélectrique par effet tunnel et sont bloqués au niveau de la couche interfaciale par le champ de dépolarisation.

Une impulsion de dépiégeage peut être appliquée si une lecture rapide après écriture est nécessaire[Mul+21].

Le dernier effet important est la libération des charges qui compensant la polarisation (opposée) précédente : ce courant de repolarisation permet d'effectuer des mesures PUND comme présenté dans la sous-section 2.1.4, et permet de multiples mécanismes de lecture destructive généralement utilisés dans les circuits à base de FeCap, détaillés dans le chapitre 3.



(a) Matériau A des deux côtés, polarisation positive

(b) Matériau A des deux côtés, polarisation négative

(c) Matériau B des deux côtés, polarisation positive

(d) Matériau B des deux côtés, polarisation négative

(e) Matériau A sur la gauche, B sur la droite, polarisation positive

(f) Matériau A sur la gauche, B sur la droite, polarisation négative

FIG. 2.7 : Diagrammes de bandes avec des longueurs de compensation de charges L_+ et L_- des deux côtés du ferroélectrique, correspondant à deux matériaux A et B. Le diagramme est affiché au repos (0 V) après polarisation avec $+V_C$ et $-V_C$ dans la colonne de gauche et de droite, respectivement (la référence de tension est prise du côté droit). Il est important de noter que $L_+ + L_-$ est constant quelle que soit la polarisation lorsque les deux côtés sont constitués du même matériau, et que cette longueur devient dépendante de la polarisation dans un assemblage asymétrique.

Jonction à effet Tunnel Ferroelectrique

Bien que cet effet ne soit pas l'objet principal de ce document, un courant de fuite existe à travers la couche ferroélectrique, particulièrement dans le cas de faibles épaisseurs, en raison de l'effet tunnel[GB14]. La polarisation ferroélectrique modifie la densité des charges

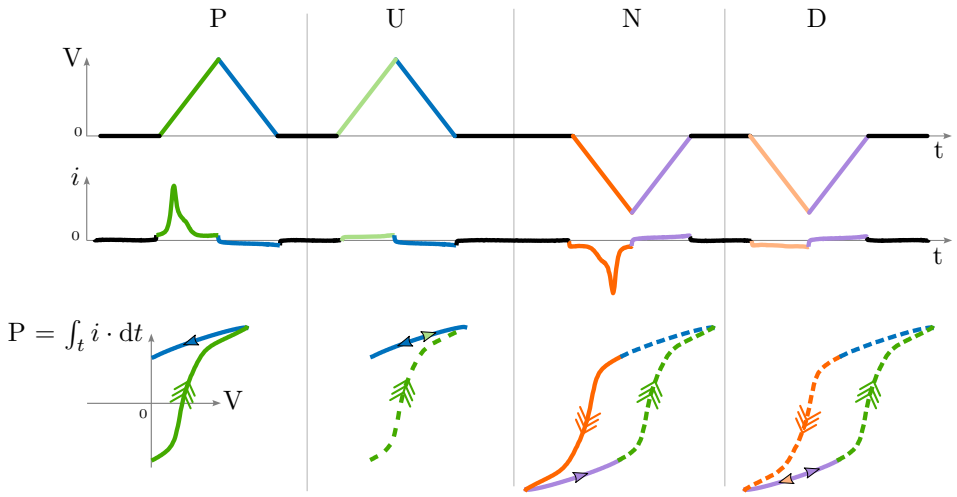


FIG. 2.8 : Forme d'onde PUND *P-V* (haut), avec le courant correspondant $I = f(t)$ (au milieu), et le cycle d'hystérésis $P = f(V)$ non corrigé associé simultanément construit. Différentes couleurs ont été utilisées pour représenter :

- la phase de polarisation avec une tension positive (**P**)
- la seconde phase de montée en tension (« up », **U**), reflétant uniquement les courants de fuite et de charge du condensateur
- la deuxième phase de polarisation avec une rampe de tension négative (**N**)
- la dernière phase, où la tension est abaissée (« down », **D**) à nouveau afin de mesurer les courants de fuite et de chargement des condensateurs dans cette direction
- ainsi que les phases de « retour à 0 V » (respectivement par le haut et par le bas), identiques pour P et U, comme pour N et D, quel que soit le nombre de phases, la polarisation ne changeant pas durant celles-ci.

à la surface, compensées par le déplacement des charges du matériau de l'électrode. Cette compensation peut être très localisée ou se produire sur de plus grandes distances, ce qui réduit considérablement le courant d'effet tunnel, car les matériaux normalement conducteurs sont localement privés de porteurs de charge mobiles. Par conséquent, un assemblage tel que celui illustré dans la figure 2.7e et la figure 2.7f, comportant des électrodes asymétriques (et donc différentes longueurs de compensation des charges) peut moduler de manière significative le courant d'effet tunnel en fonction de la polarisation du ferroélectrique, ce qui permet de lire l'état de polarisation sans repolariser le ferroélectrique. Ce principe est exploité dans le mode de fonctionnement Jonction à effet Tunnel Ferroelectrique (FTJ) du circuit 2T1C présenté dans la section 3.5. Les électrodes composées de métal et de semi-conducteur semblent produire la dépendance courant-polarisation la plus forte [GB14 ; Maj+18 ; Maj22, p. 10].

2.1.4 Mesures PUND

Afin de répondre au besoin de mesure indépendante des caractéristiques électriques des parties paraélectriques et ferroélectriques d'un condensateur ferroélectrique, une forme d'onde PUND peut être utilisée[Mül+11 ; SG96]. Comme le suggère l'acronyme, cette stratégie de mesure est divisée en quatre parties. Une séquence spécifique de rampes de tension est appliquée pendant que le courant est mesuré :

1. Une rampe de tension positive change la polarité du condensateur ferroélectrique vers l'état « positif » (opération parfois qualifiée de « programmation » dans la littérature, comme indiqué dans la section 2.1.1). La vitesse de balayage est suffisamment lente pour que la polarisation du ferroélectrique soit complète. La tension est progressivement redescendue à zéro.

2. Une rampe identique est de nouveau appliquée. Cependant, aucune repolarisation n'a lieu cette fois-ci, car celle-ci s'est produite lors de l'étape précédente : les domaines dont la polarisation changerait sous l'effet d'un tel champ électrique l'ont déjà fait lors de la première rampe. Cette étape supplémentaire permet de mesurer la capacité normale, (paraélectrique), ainsi que les courants de fuite. Seuls les domaines paraélectriques s'alignent à nouveau sur la polarisation « montante » (« up ») du champ électrique.

3. Ce processus est répété afin de mesurer la partie négative du diagramme de commutation : la même rampe de tension est appliquée, sa polarité inversée, pour renverser la polarisation des domaines ferroélectriques précédemment commutés. Une polarisation « négative » est inscrite dans l'oxyde ferroélectrique pendant cette phase.

4. Une fois de plus, la même rampe de polarité négative est appliquée dans le but de mesurer la contribution des domaines paraélectriques et des courants de fuite aux mesures. Les domaines paraélectriques s'alignent sur la direction « vers le bas » (« down ») du champ électrique, et retournent à leur position de repos une fois que la tension appliquée extérieurement revient à zéro.

La forme d'onde résultante est représentée sur la figure 2.8. La séquence ci-dessus est généralement répétée deux fois, ou au moins précédée d'une impulsion négative, afin de placer chaque domaine ferroélectrique dans un état déterminé avant d'effectuer un cycle de mesure. Cela rend le cycle extrait continu, et permet ainsi un bouclage parfait de la courbe *P-V*.

Le courant est mesuré lors de l'application des rampes de tension décrites ci-dessus. Les mesures de courant obtenues contiennent :

1. le courant de repolarisation pendant les étapes 1 et 3,

2. le courant de fuite, qui est présent pendant tout le cycle,

3. le courant de charge/décharge du condensateur, également présent au long du cycle.

Le courant de repolarisation transporte les charges stockées dans le condensateur ferroélectrique, ce qui implique que la polarisation de surface peut être trouvée en intégrant cette quantité : $P(t=T) = {}^{1}/_{A_{C_{FE}}} \cdot \int_{t=0}^{T} i(t) \cdot \mathrm{d}t$. Celle-ci est généralement exprimée en $\mu\mathrm{C\,cm^{-2}}$.

Si la procédure ci-dessus est effectuée avec le courant mesuré pendant la première impulsion de chaque polarité, la réponse ferroélectrique *P-V* résultante sera faussée par la contribution du condensateur paraélectrique (avec une relation *P-V* linéaire), ainsi que par le courant de fuite. Les secondes impulsions ne contiennent à leur tour que la mesure du courant de fuite et la réponse paraélectrique, car la réponse ferroélectrique ne peut être obtenue qu'après une dépolarisant les domaines ferroélectriques. Cela est effectué lors de la seconde moitié du cycle PUND en renversant la polarité.

En soustrayant la seconde phase de la première, il est possible d'observer le cycle ferroélectrique pur, non incliné, tel que représenté sur la figure 2.9. Cette différence peut être observée en comparant la figure 2.9a et la figure 2.9b. Le premier graphique est obtenu avec :

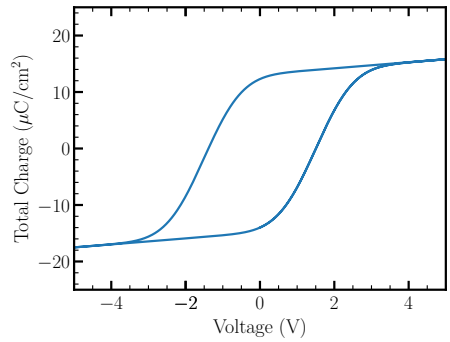
$$P(t) = \int_{t=0}^{t=N} i(t) \cdot \mathrm{d}t$$

en notant Δt la séparation des cycles P et U, ainsi que des cycles N et D, la figure 2.9b est généré à partir de :

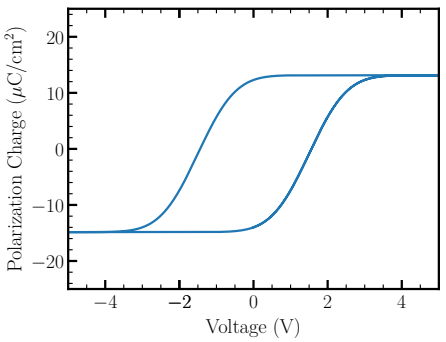
$$P(t) = \int_{t=0}^{t=\Delta t} i(t) - i(t+\Delta t) \cdot \mathrm{d}t \qquad (t \notin U, D)$$

ou plus simplement, avec i_X le courant mesuré pendant la phase X :

$$\begin{cases} P(V) = \int_{V} i_P(V) - i_U(V) \cdot \mathrm{d}V, & \text{phase PU} \\ P(V) = \int_{V} i_N(V) - i_D(V) \cdot \mathrm{d}V, & \text{phase ND} \end{cases}$$



(a) Réponse *P-V* obtenue en intégrant simplement le courant mesuré pendant un balayage de tension.



(b) Réponse *P-V* obtenue en soustrayant la phase U et D avant l'intégration du courant pendant un balayage PUND.

FIG. 2.9 : Réponse *P-V* affectée par la capacité (2.9a), et réponse ferroélectrique pure correspondante (2.9b) extraite par la stimulation PUND. Simulation à partir de la distribution ferroélectrique idéalisée présentée dans la figure 2.14, où V_C commence à 0 V.

2.2 Modélisation

Toutes les simulations ont été effectuées avec Cadence® Spectre®.

Il existe différentes approches de modélisation, en fonction de la précision requise. Les modèles plus précis sont utiles afin de valider de petits circuits tels que des portes logiques, et garantissent que les matériaux ferroélectriques peuvent être programmés de manière fiable. Un modèle simplifié peut ensuite être utilisé à plus grande échelle pour valider un circuit complet, comme décrit dans la sous-section 4.6.4.

Dans cette section, trois approches de modélisation sont détaillées :

1. Dans la sous-section 2.2.1, le modèle idéalisé de transition de phase de Landau-Kalathnikov donne un aperçu des mécanismes de repolarisation ferroélectrique. Celui-ci n'est utilisable que pour des ferroélectriques à domaine unique, et ne doit donc être utilisé que pour des dispositifs ferroélectriques de faible taille.

2. Dans la sous-section 2.2.2, l'approche de modélisation de Preisach, qui s'applique davantage à une distribution statistique des domaines ferroélectriques, est présentée. Comme telle, cette approche devrait être favorisée lorsque des domaines ferroélectriques de taille plus importante sont considérés, de préférence ceux comprenant une population suffisamment grande et homogène d'un dispositif à l'autre.

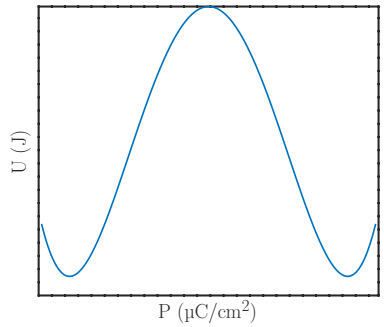
3. Enfin, dans la sous-section 2.2.3, quelques approches de simplification de modèles pour simulation à grande échelle seront données.

2.2.1 Modèle de Landau

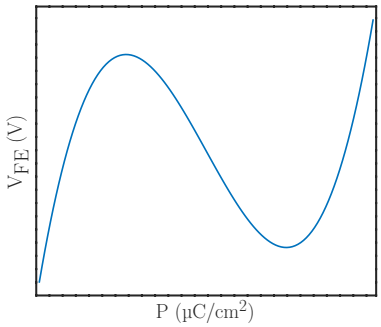
Description

Le modèle de Landau permet de décrire le comportement de repolarisation de cristaux ferroélectriques individuels. Il s'agit d'une approche établie à partir de la symétrie de l'équation de l'énergie libre de Gibbs, appliquée aux problèmes de transition de phase.

Appliquée pour la première fois aux ferroélectriques par Devonshire[CL07 ; Dev49], l'existence de deux positions d'équilibre visibles sur la figure 2.10a rend cette théorie adaptée aux simulations mono-domaine. L'équation mathématique est obtenue avec une expansion de Taylor de l'équation de l'énergie libre de Gibbs. Celle-ci est ensuite simplifiée grâce à des considérations de symétrie et en supposant certains paramètres statiques (indépendants du temps et de la fréquence d'entrée), ce qui limite le nombre de degrés de liberté[CL07].



(a) Équation de l'énergie libre de Gibbs appliquée au problème de transition de phase de la repolarisation d'un matériau ferroélectrique. Les deux puits de potentiel correspondent aux positions d'équilibre des deux orientations ferroélectriques possibles.



(b) Dérivée du tracé de la figure 2.10a. Ce tracé est plus directement lié à la conception de circuits, car celui-ci représente le champ électrique ou le potentiel électrique à travers le matériau ferroélectrique.

FIG. 2.10 : Équation de l'énergie libre de Gibbs simulée (2.10a) et dérivée (2.10b) pour les coefficients publiés dans la littérature[Yin+16] pour $HfZrO_2$: $\alpha = -7 \times 10^9\,\mathrm{m\,F^{-1}}$, $\beta = 3.3 \times 10^{10}\,\mathrm{m^5/F/C^2}$, $\gamma = -0.2 \times 10^{10}\,\mathrm{m^9/F/C^4}$, épaisseur $t_{FE} = 5.7\,\mathrm{nm}$.

Équation

L'équation de l'énergie libre de Gibbs $U = \mathrm{f}(P)$ donne l'énergie de la surface en fonction de la polarisation de la surface. Une expansion de Taylor du sixième degré est généralement considérée comme suffisante pour décrire le système, la symétrie du système permettant de supprimer la moitié des coefficients[CL07 ; SD08 ; WA17]. Pour un système unidimensionnel, cette expression peut être formulée ainsi :

$$U = \alpha P^2 + \beta P^4 + \gamma P^6 - \vec{E}_{ext} \cdot \vec{P}$$

$$\frac{\mathrm{d}U}{\mathrm{d}P} = \nabla_{\vec{P}} U = 2\alpha P + 4\beta P^3 + 6\gamma P^5 - E_{ext}$$

L'application de la théorie phénoménologique de Landau-Ginzburg-Devonshire pour décrire les propriétés dynamiques des ferroélectriques conduit à l'équation de Landau-Khalatnikov suivante [FKK12 ; Mas+21] :

$$\rho \frac{\mathrm{d}\vec{P}}{\mathrm{dt}} + \nabla_{\vec{P}} U = 0 \tag{2.1}$$

En utilisant l'équation 2.1 de Landau-Khalatnikov :

$$0 = 2\alpha P + 4\beta P^3 + 6\gamma P^5 - E_{ext} + \rho \frac{\mathrm{d}P}{\mathrm{dt}}$$

$$E_{ext} = 2\alpha P + 4\beta P^3 + 6\gamma P^5 + \rho \frac{\mathrm{d}P}{\mathrm{dt}}$$

En première approche, le cas statique peut être étudié (ce qui permet également d'ignorer l'équation 2.1 en supposant une polarisation uniforme, auquel cas $\nabla_{\vec{P}} U = 0$) :

$$E_{ext} = 2\alpha P + 4\beta P^3 + 6\gamma P^5 \tag{2.2}$$

t_{FE} est l'épaisseur du ferroélectrique, donc $V_{ext} = E_{ext} \cdot t_{FE}$:

$$V_{ext} = 2\alpha t_{FE} P + 4\beta t_{FE} P^3 + 6\gamma t_{FE} P^5$$

Avec A l'aire du condensateur et Q le nombre de charges, tel que $P = Q/A$:

$$V_{ext} = 2\alpha t_{FE} \frac{Q}{A} + 4\beta t_{FE} \frac{Q^3}{A^3} + 6\gamma t_{FE} \frac{Q^5}{A^5}$$

$$V_{ext} = \frac{1}{A}\left(2\alpha t_{FE} Q + 4\beta t_{FE}\frac{Q^3}{A^2} + 6\gamma t_{FE}\frac{Q^5}{A^4}\right)$$

$$\frac{1}{C_{FE}} = \frac{\mathrm{d}V_{ext}}{\mathrm{d}Q} = \frac{1}{A}\left(2\alpha t_{FE} + 12\beta t_{FE}\frac{Q^2}{A^2} + 30\gamma t_{FE}\frac{Q^4}{A^4}\right)$$

$$\frac{1}{C_{FE}} = \frac{1}{A}\left(2\alpha t_{FE} + 12\beta t_{FE}P^2 + 30\gamma t_{FE}P^4\right)$$

Il en résulte une expression de la capacité ferroélectrique :

$$\frac{A_{C_{FE}}}{C_{FE}} = 2\alpha t_{FE} + 12\beta t_{FE}P^2 + 30\gamma t_{FE}P^4 \tag{2.3}$$

Utilisation

Cette expression être simplement utilisée en conjonction avec un modèle de condensateur, et peut être considérée comme équivalente à un condensateur en série avec un générateur de tension, dont la tension dépend de la quantité de charges accumulées dans le condensateur. Cette approche s'applique également aux FeFET[WA17], où la couche ferroélectrique est utilisée soit directement comme oxyde de grille, soit empilée au-dessus, auquel cas elle est en série avec le condensateur de grille, comme décrit dans la sous-section 2.4.3.

Dans un dispositif réel, la couche ferroélectrique est partiellement ferroélectrique et partiellement paraélectrique (purement capacitive), en raison de l'existence de deux électrodes et de domaines non ferroélectriques, comme le montrent les formes d'onde PUND, détaillées dans la sous-section 2.1.4. Ce modèle de capacité ferroélectrique est donc plus précis lorsque utilisé en parallèle avec un condensateur classique.

Obtention des paramètres

L'ajustement du modèle aux données expérimentales est relativement facile, car celui-ci consiste en un simple ajustement polynomial, avec certains coefficients contraints à zéro. Il convient toutefois de souligner que ce modèle n'est pas conçu pour modéliser précisément le comportement d'un condensateur, ce qui implique que les données expérimentales doivent être obtenues avec un stimulus PUND afin d'éliminer la réponse du condensateur paraélectrique et des courants de fuite de la réponse purement ferroélectrique : bien que le premier terme du polynôme puisse modéliser un condensateur idéal avec une réponse linéaire $^1/_{C_{FE}} = {}^{\mathrm{d}V}/_{\mathrm{d}Q}$, il est préférable de confier cette tâche à un modèle spécialisé, en calquant le modèle ferroélectrique sur les données expérimentales exemptes de parasites provenant d'un condensateur réel non linéaire. Un exemple d'ajustement sur des données expérimentales obtenues à l'EPFL est présenté en figure 2.11. Le code GNU Octave (compatible MATLAB®) effectuant cet ajustement est fourni dans l'extrait de code A.7. Les numéros de ligne indiqués dans le paragraphe suivant renvoient à cet extrait de code.

Afin d'automatiser le processus d'ajustement, et puisque la Courbe en S n'est pas directement observable depuis les données de caractérisation, les parties verticales (représentées par des points sur la figure 2.12) du cycle d'hystérésis doivent être éliminées. Ceci est illustré sur la figure 2.11a, où des régions d'intérêt (ROIs, regions of interest) sont identifiées. Ces ROIs contiennent les seuls points sur lesquels le polynôme est ajusté, et sont sélectionnés comme suit :

1. Les deux extrémités du cycle sont sélectionnées. Il s'agit des coordonnées correspondant aux valeurs de tension les plus élevées (ligne 34).
2. Les deux points d'inflexion du cycle sont sélectionnés :
   (a) La bissectrice du cycle est d'abord calculée, c'est-à-dire la ligne qui relie les deux extrémités définies précédemment (ligne 53)
   (b) Les points du cycle les plus éloignés de cette ligne sont identifiés, par la distance à leur projeté orthogonal sur la bissectrice (ligne 72).
   (c) Le point le plus éloigné de chaque moitié du cycle est identifié comme point d'inflexion (ligne 85).

3. Les deux ROIs sont celles qui relient les points d'intérêt susmentionnés de la même moitié de cycle, par le chemin le plus court suivant l'axe des tensions (ligne 110).

La courbe polynomiale est ensuite ajustée, avec les axes inversés sous la forme $V = f(P)$, par opposition au tracé $P = f(V)$ *P-V* plus généralement représenté. Cet ajustement est effectué en contraignant les coefficients pairs à zéro afin de satisfaire à l'équation 2.2, comme le fait le code fourni dans l'extrait de code A.9.

Conclusion

Bien que l'utilisation de ce modèle soit limitée, car restreint aux cristaux mono-domaine, celui-ci permet l'exploration du comportement à capacité négative des dispositifs ferroélectriques, et est généralement utilisé pour simuler les NCFETs[SR17]. Celui-ci est également utilisable comme première approche pour la simulation de circuits ferroélectriques polycristallins, à condition que ces circuits ne reposent pas sur un contrôle analogique précis de la polarisation électrique[Azi+18].

Ce modèle contient également une composante dynamique qui n'a pas été étudiée ici, mais qui pourrait être intéressante pour les simulations dynamiques. Les deux équations importantes du modèle sont numérotées 2.2 et 2.3. Celles-ci introduisent les coefficients de *Landau* α, β et γ, spécifiques à chaque matériau, ceux-ci décrivant sa polarisation (et donc les charges libérées lors du changement de polarisation) en fonction du champ (ou tension) externe appliqué.

2.2.2 Modèle de Preisach

Le modèle de Preisach[Pre35] utilise une analyse mathématique d'un cycle d'hystérésis obtenu expérimentalement, tel que celui illustré sur la figure 2.11 ou sur la figure 2.12. Ces cycles peuvent être estimés à l'aide d'une superposition d'« hystérons » : des fonctions mathématiques hystérétiques à l'expression simple.

Conçue à l'origine pour le ferromagnétisme, cette approche peut être appliquée à la ferroélectricité et permet la modélisation de plusieurs domaines ferroélectriques. Cela permet l'utilisation de boucles mineures dans le cycle ferroélectrique, par exemple dans les applications MLC[KN03].

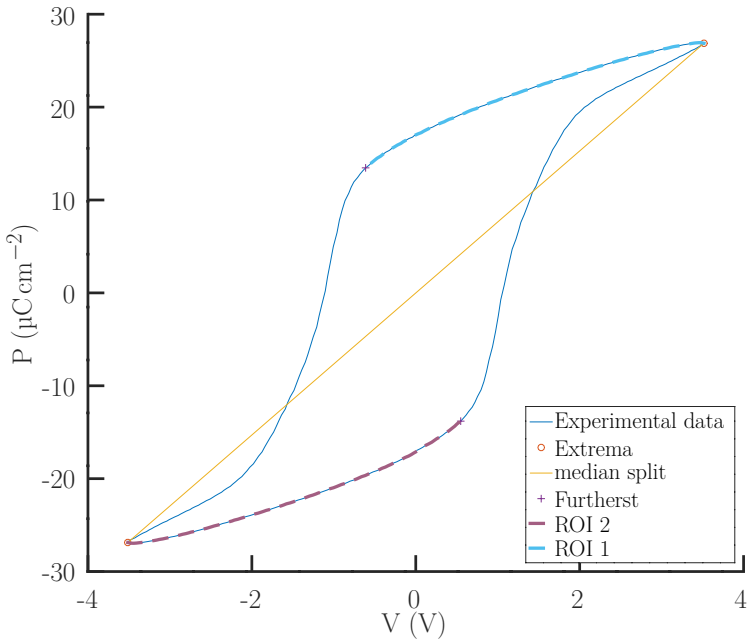
Toutefois, ce modèle devient moins précis lorsqu'un nombre réduit de domaines ferroélectriques interagissent, ce qui est le cas pour des dispositifs de taille réduite. Selon NaMLab, leur implémentation ne devrait pas être utilisée avec des condensateurs ferroélectriques de diamètres inférieurs à 150 nm. La limite de validité de ce modèle est donc estimée à environ 150 nm, ce qui correspond à la limite en dessous de laquelle les domaines ferroélectriques individuels ont un effet mesurable.

Remerciements

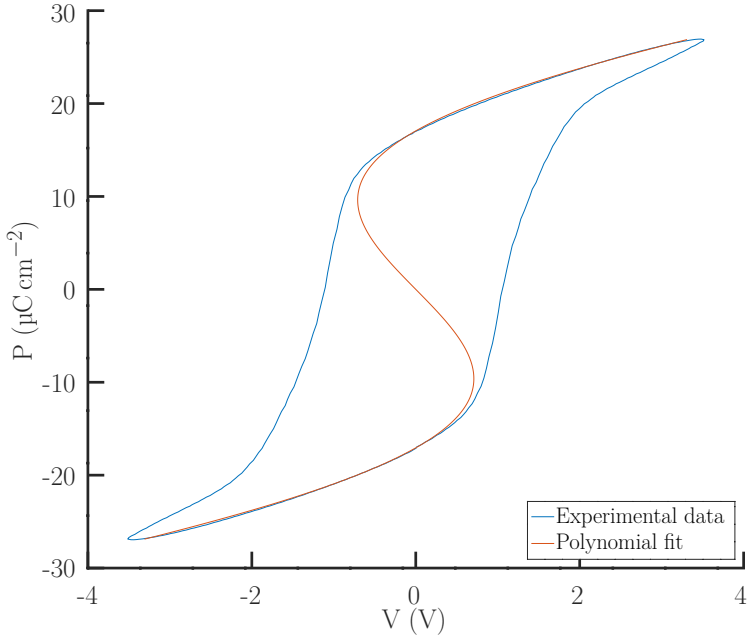
Damien Deleruyelle a apporté des contributions majeures à cette section, notamment via la réalisation d'un modèle de Preisach utilisé pour générer les illustrations.

Hystérons

Le principe fondamental du modèle de Preisach est la considération du matériau ferroélectrique comme un ensemble de dipôles électriques individuels, contribuant à la polarisation totale. Chaque dipôle a deux tensions (champs) coercitives individuelles (V_c^+ et V_c^-), devant être atteintes pour le repolariser dans la direction opposée. Cette inversion de polarisation est supposée rapide et continue, et ininterruptible. Ainsi, chaque dipôle possède une boucle d'hystérésis quasi-rectangulaire (hystéron), comme le montre la figure 2.13. En partant de l'hypothèse que les dipôles n'interagissent pas entre eux, la boucle d'hystérésis du système macroscopique est considérée comme une superposition de ces boucles unitaires. Les tensions coercitives de chaque dipôle du système sont supposées suivre une distribution statistique : cette condition limite l'application du modèle de Preisach aux dispositifs de faibles dimensions. La distribution des états *haut* et *bas* parmi les dipôles (notés par la suite μ) dépend de la tension appliquée, ainsi que de l'historique de la tension, et des tensions coercitives pour le



(a) Identification de ROIs pour l'ajustement des données expérimentales



(b) Ajustement polynomial de degré cinq, contraint à des puissances impaires, comparé aux données expérimentales

FIG. 2.11 : Sélection de ROIs à partir des données expérimentales (2.11a). Seules ces régions seront utilisées pour l'ajustement du polynôme représenté dans 2.11b, dont l'équation est $V = -0.111695P + 0.000423615P^3 - 1.36673e-07P^5$, ou, de manière équivalente, les coefficients de Landau extraits comme $\alpha = -3.490\,47 \times 10^8$ m/F, $\beta = 6.618\,98 \times 10^9$ m^5/F/C^2, $\gamma = -1.423\,68 \times 10^{10}$ m^9/F/C^4.
L'ajustement n'a pas ici été effectué sur une courbe *P-V* PUND, en raison de l'absence de données expérimentales. Ces tracés ont été directement générés par l'extrait de code A.7.

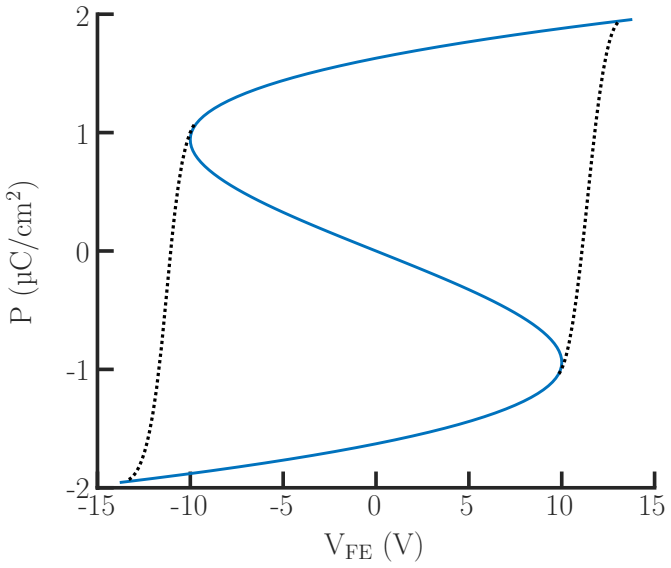


FIG. 2.12 : Exemple de cycle d'hystérésis ferroélectrique, utilisant les mêmes paramètres que la figure 2.10. Il s'agit du même graphique que celui de la figure 2.10b, mais avec les axes permutés pour afficher la « Courbe en S » révélatrice d'un dispositif ferroélectrique. La trajectoire suivie par l'oxyde ferroélectrique lorsque soumis à une tension croissante (ou décroissante) de manière monotone est également indiquée en pointillés.

dipôle considéré $\mu_{V_c^+,V_c^-}$. Cela signifie que les éventuels « points tournants » (inversion du sens de l'évolution) de la courbe de tension ont un impact direct sur la distribution de l'état des dipôles μ.

Le comportement de l'hystéron de chaque dipôle peut être modélisé mathématiquement par l'équation suivante :

$$\mu_{V_c^+,V_c^-}\left(V(t)\right) = \begin{cases} -1, & V(t) \leq V_c^- \\ \gamma, & V_c^- \leq V(t) \leq V_c^+ \\ +1, & V(t) \geq V_c^+ \end{cases}$$

où la valeur de γ est déterminée par l'historique des tensions précédemment appliquées. Sa valeur est changée à $\gamma = 1$ lorsque $V(t)$ dépasse la borne supérieure (respectivement à -1 sous la borne inférieure) de la plage $[V_c^+, V_c^-]$. Dans cette équation, il est supposé qu'il n'existe pas de délai de commutation, et chaque hystéron est pondéré par la distribution (V_c^+ et V_c^-). Bien que l'équation ci-dessus suppose une fonction échelon carrée, il ne s'agit pas forcément de la plus physiquement réaliste, ni de la plus mathématiquement simple. De plus, la discontinuité introduite par ce créneau peut être source d'instabilité dans la simulation. Pour ces raisons, la variante Miller[Mil+90] du modèle utilise une fonction échelon tanh.

En désignant V_m la l'amplidude maximale de la tension appliquée au condensateur ferroélectrique macroscopique, et en supposant une distribution normalisée (V_c^+ et V_c^-), notée ρ, nous pouvons écrire :

$$\int_{-V_m}^{V_m} \int_{-V_m}^{V_m} \rho(V_c^+, V_c^-) \cdot \mathrm{d}V_c^+ \cdot \mathrm{d}V_c^- = 1 \quad [1]$$

Par conséquent, des distributions gaussiennes bidimensionnelles de (V_c^+ et V_c^-), comme illustré dans la figure 2.14, sont de bonnes candidates pour décrire la distribution des hystérons.

[1] Les limites de ces intégrales peuvent être égales à 0 dans le cas typique où $V_c^- < 0$ et $V_c^+ > 0$.

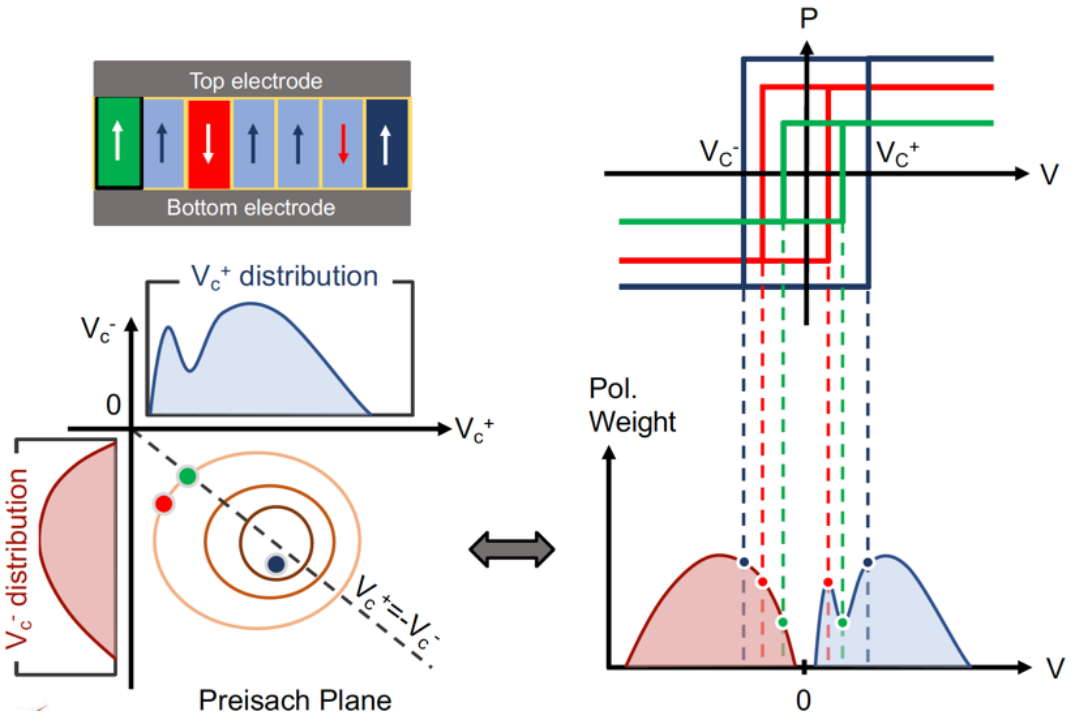


FIG. 2.13 : Modèle de Preisach – hystérons et plan

Comportement cumulatif des hystérons

La variation de la polarisation est calculée en intégrant $\rho \cdot \mu$ sur le plan de Preisach, c'est-à-dire en intégrant l'état de chaque hystéron (ou plutôt, chaque partie de la distribution) sur l'historique de la tension :

$$P\left(V(t)\right) = \iint_D \rho(V_c^+, V_c^-) \cdot \mu_{V_c^+, V_c^-} \cdot V(t) \cdot \mathrm{d}V_c^+ \cdot \mathrm{d}V_c^-$$

où D dépend de la trajectoire de $V(t)$ de telle sorte que :

$$\begin{cases} D^+ = [V_i; V_f] \times]-\infty, +\infty[\,, & \text{si } \frac{\mathrm{d}V}{\mathrm{d}t} > 0 \text{ (tension croissante)} \\ D^- =]-\infty, +\infty[\times [V_i; V_f]\,, & \text{si } \frac{\mathrm{d}V}{\mathrm{d}t} < 0 \text{ (tension décroissante)} \end{cases}$$

Dans ces équations, V_i (respectivement V_f) est la valeur initiale (respectivement finale) de $V(t)$. Il est également possible de calculer des boucles internes, dues à des changements de direction de la tension d'entrée, dans l'intervalle $[-V_m, +V_m]$, et qui conduisent également à des points d'inflexion de la tension, comme illustré sur la figure 2.15.

Il est intéressant de noter que la distribution $\rho\left(V_c^+, V_c^-\right)$ peut théoriquement être extraite de cycles de polarisation incrémentaux δP (c'est-à-dire $\rho \cdot \mu \cdot \delta V_c$), qui peuvent être déduits en dérivant les courbes P-V issues de plusieurs balayages rétrogrades consécutifs, une mesure dite de courbe d'inversion du premier ordre (FORC, First Order Reversal Curve)[Peš+17]. Cette méthode est illustrée dans la figure 2.16 et peut être utilisée pour déterminer expérimentalement la distribution de tensions coercitives des domaines ferroélectriques.

Grâce à la flexibilité de cette approche numérique, des distributions arbitraires (non gaussiennes) $\{V_c^+, V_c^-\}$ peuvent également être employées ou extraites, comme le montre la figure 2.17.

Limites

Ce modèle est actuellement l'un des plus utilisés pour la simulation de circuits. Cependant, bien qu'il soit capable de représenter différentes variabilités de domaines ferroélectriques avec de multiples distributions, celui-ci est incapable de modéliser les FTJ, la capacité négative,

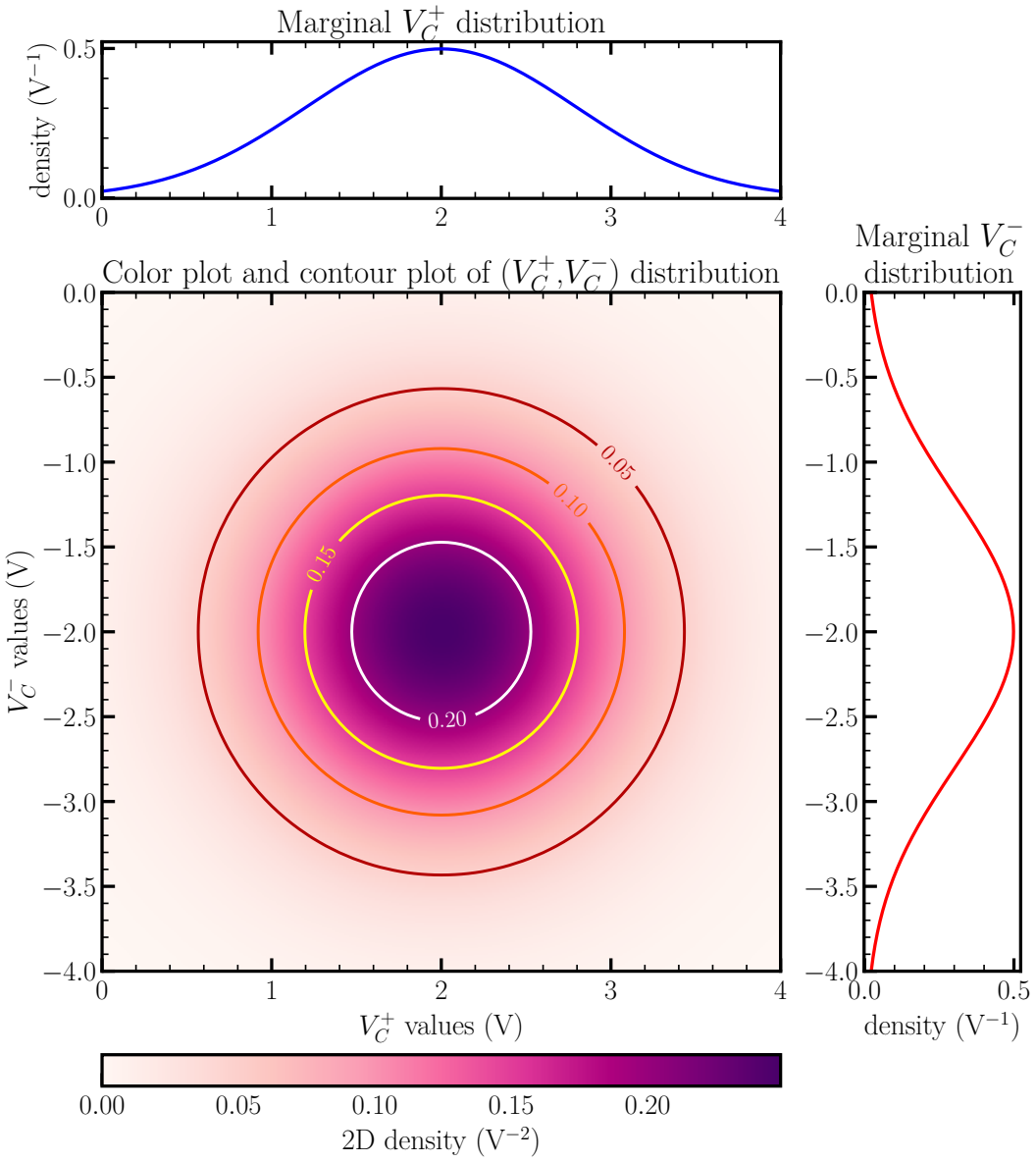


FIG. 2.14 : Distribution Gaussienne 2D (V_c^+, V_c^-)

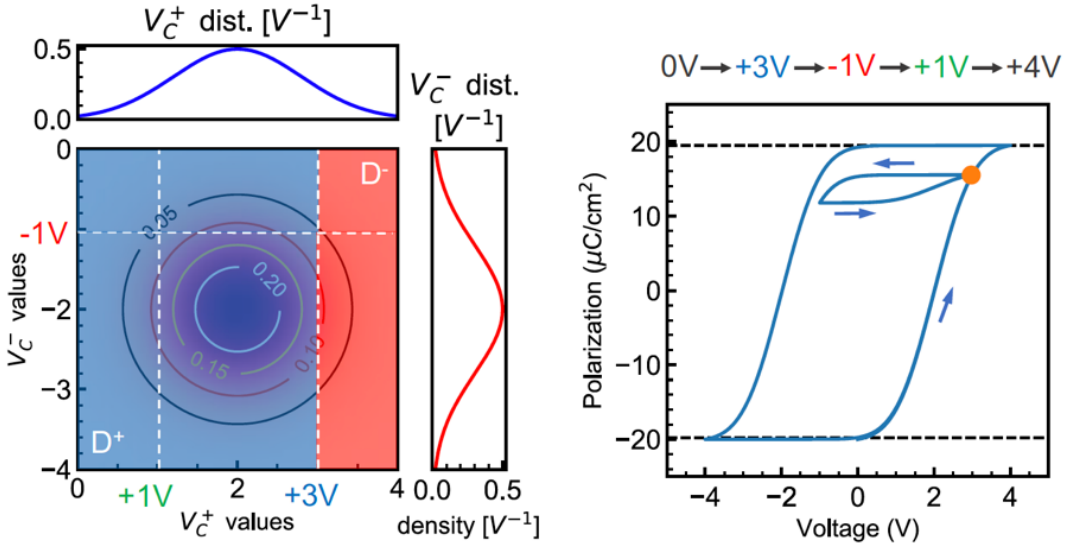


FIG. 2.15 : Boucles internes et points d'inflexion.

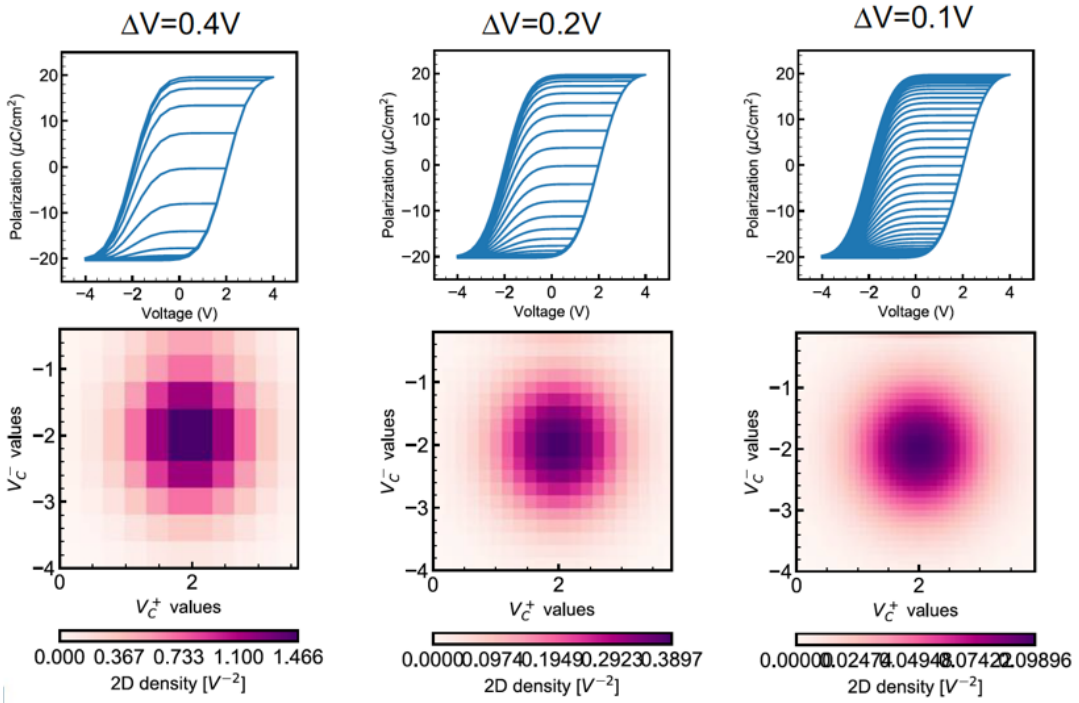


FIG. 2.16 : Exemple d'extraction de distribution gaussienne 2D V_c^+/V_c^- avec une résolution de tension V_c de plus en plus fine. L'augmentation du nombre de domaines cristallins a un effet similaire sur la distribution.

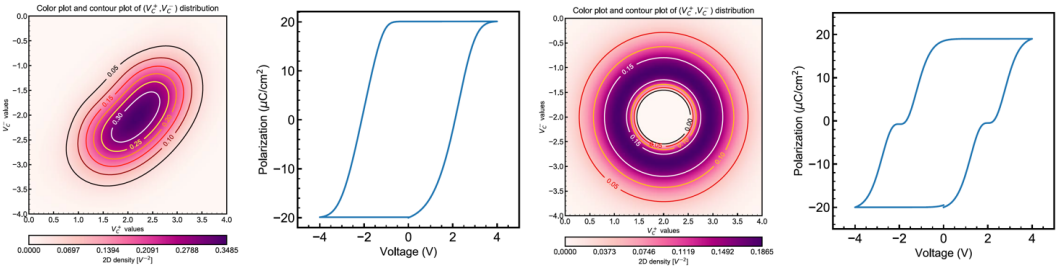


FIG. 2.17 : Gestion de distributions V_c^+/V_c^- arbitraires.

Extrait de code 2.1 : Modèle ferroélectrique minimal écrit en Verilog-A pour utilisation en simulation de FeFET, comme alternative plus légère à calculer.

```
`include "constants.vams"
`include "disciplines.vams"

module miniferro(in, out);
input in;
output out;
electrical in, out;

parameter real vc = 2; // volts
parameter real vth = 0.4; // volts
parameter real vout = 1.0; //volts

integer is_on = 1;

analog begin

    @(cross(V(in) - vc, 1)) // Goes above vc
        is_on = 1;

    @(cross(V(in) + vc, -1)) // Goes below -vc
        is_on = 0;

    V(out) <+ transition(is_on*vout*(V(in) > vth),0,100p);

end

endmodule
```

ainsi que les comportements de commutation accumulatifs et stochastiques[Den+20]. En outre, ce modèle, du moins en première approche, doit maintenir en mémoire l'état de la population d'hystérons, ce qui entraîne des contraintes de performance, en particulier pour les circuits de grande taille contenant plusieurs dispositifs ferroélectriques.

2.2.3 Modèle simplifié pour les simulations à grande échelle

Afin de réduire le coût de calcul de la simulation de grands circuits avec des modèles complexes nécessitant plusieurs octets de mémoire pour stocker leur état, ainsi que des interactions complexes qui pourraient causer des problèmes de convergence dans des cas spécifiques, un modèle complet de Preisach ou de Landau peut être remplacé par un modèle simplifié. Nous avons implémenté un tel modèle simplifié en Verilog-A, décrit dans l'extrait de code 2.1.

Dans ce cas, de multiples considérations doivent être prises en compte :

- Les circuits individuels devraient être validés à l'aide d'un modèle physiquement plus réaliste avant l'utilisation d'un modèle simplifié pour la simulation d'architectures à moyenne et grande échelle.
- Les paramètres du modèle simplifié, tels que les tensions de sortie et de seuil, devraient être ajustés afin que le comportement du circuit corresponde à celui obtenu avec le modèle plus réaliste.
- Un modèle simplifié tel que celui présenté dans l'extrait de code 2.1 ne reflète pas (et en général n'est pas destiné à refléter) les comportements plus complexes, tel que les cycles mineurs ou une polarisation partielle.

Le modèle simplifié présenté dans l'extrait de code 2.1 est construit autour du principe d'un générateur de tension (`V(out)`) contrôlé par la tension (`V(in)`), comportant une mémoire (`is_on`).

Afin d'alléger le modèle autant que possible, le code proposé inclut plusieurs hypothèses et simplifications :

- le modèle est asymétrique. Celui-ci possède une entrée d'impédance infinie et une sortie d'impédance nulle. Par conséquent, l'entrée est supposée commandée en tension.
- le courant et l'énergie de sortie ne sont pas limités. On suppose la sortie connectée à un nœud flottant, dont la tension en régime permanent a été mesurée à l'aide d'un modèle plus complet.
- V_C est supposé symétrique : $V_C^- = -V_C^+$
- le changement de polarisation est instantané dès le franchissement de V_C, bien que la sortie ait une période de transition arbitraire de 100 ps, facilitant la convergence. Cette faible valeur a été choisie pour éviter l'introduction d'une latence supplémentaire, mais celle-ci pourrait être ajustée en fonction des mesures de latence obtenues lors de la caractérisation.
- il est supposé que le modèle sera utilisé dans un FeFET, comme décrit dans la sous-section 2.4.3 : celui-ci inclut une tension de seuil au lieu d'ajouter une tension fixe à la tension de sortie comme le ferait `V(out) <+ transition(V(in)+is_on*deltaV,0,100p)`.

Néanmoins, grâce à sa simplicité et à sa stabilité, ce modèle est tout à fait adapté à l'exploration d'architectures plus complexes. L'utilisation d'une fonction `transition` standard garantit la continuité du signal de sortie pour la stabilité de la simulation. Ce modèle a été validé par la phase de vérification du circuit à grande échelle discuté dans la sous-section 4.6.4, contenant plusieurs milliers de FeFETs. Avec l'utilisation de composants passifs externes tels qu'un condensateur en série, ou de petits ajustements au modèle comme décrit ci-dessus, celui-ci peut également être utilisé dans une plus large gamme de cas d'utilisation. extrait de code 2.1 est particulièrement adapté à l'utilisation en série avec une grille de transistor, comme décrit dans la sous-section 2.4.3.

2.3 Condensateurs ferroélectriques

Comme nous l'avons vu dans la sous-section 2.1.3, les matériaux ferroélectriques peuvent être utilisés comme matériaux diélectriques de haute performance dans une structure de condensateur ordinaire. Le dispositif qui en résulte est nommé FeCaps, et peut être considéré comme un condensateur ordinaire avec une capacité de stockage de charge supplémentaire contrôlée par l'historique de la tension appliquée.

2.3.1 Condensateur ordinaire

Le régime du « condensateur ordinaire » peut être exploité de 0 V à $\pm V_C$. Dans cet intervalle de tension, la tension appliquée est insuffisante pour repolariser le cristal ferroélectrique, qui agit alors comme un diélectrique de condensateur ordinaire. Ce condensateur se comporte de la même manière, quel que soit l'historique de la tension ou l'orientation du ferroélectrique, à l'exception des pertes par effet tunnel généralement négligeables décrites dans la section 2.1.3. Cela les rend adaptés à des applications de condensateur à faible tension, y compris pour les cellules DRAM. Une inversion de polarisation peut précharger le condensateur, mais l'historique de la polarisation n'affecte pas sensiblement la courbe P-V à proximité de zéro volt.

2.3.2 Non-volatilité

Contrairement aux mémoires *volatiles*, les mémoires Mémoire non volatile n'ont pas besoin d'une alimentation électrique externe pour conserver l'information mémorisée. Cette propriété les rend adaptées au stockage de données à long terme, car leurs performances souvent

moindres (par rapport aux mémoires volatiles) sont compensées par des économies d'énergie, et une plus grande fiabilité en cas de coupure de l'alimentation. Les mémoires volatiles comprennent les DRAM et les SRAM. Les mémoires non volatiles comprennent les mémoires flash (courantes dans les clés USB, les cartes SD et les stockage électroniques (SSDs, Solid-State Disk)), la RAM magnétorésistive, les mémoires à changement de phase ainsi que les mémoires à noyau magnétique (obsolètes). Une comparaison plus détaillée est fournie dans la sous-section 2.5.4.

Tant que le champ électrique dans le cristal ferroélectrique n'atteint pas $\pm E_C$, la polarisation interne des domaines ferroélectriques reste stable, ce qui en fait un bon candidat pour une Mémoire non volatile.

La polarisation interne ne peut être modifiée que lorsque la tension d'alimentation dépasse une certaine valeur, portant l'intensité du champ électrique traversant le cristal ferroélectrique au-delà du seuil $\pm E_C$. Si la polarisation est renversée, les charges stockées sont libérées, ce qui entraîne un pic de courant. Ce comportement est similaire à une augmentation de la capacité lors du franchissement du seuil. À mesure que la tension d'entrée continue d'augmenter, d'autres domaines peuvent être repolarisés, ce qui poursuit l'augmentation virtuelle de la capacité, jusqu'à ce que tous les domaines soient polarisés ou jusqu'à ce que la tension soit de nouveau abaissée.

Puisque le pic de courant apparaît uniquement si la polarisation est renversée, comme illustré dans la figure 2.8, sa présence peut être utilisée pour déduire l'état antérieur du cristal ferroélectrique. C'est le principe de lecture utilisé par les mémoires 1T1C, étudiées dans la section 3.2. L'opération de lecture plaçant la mémoire dans un nouvel état connu, celle-ci est dite *destructrice*, et la valeur précédente doit être réécrite si celle-ci doit être conservée pour une lecture ultérieure.

Les Mémoires non volatiles sont l'une des applications les plus prometteuses des matériaux ferroélectriques, et la fenêtre $]-V_c; +V_c[$ (dans laquelle l'information stockée n'est pas modifiée) laisse la porte ouverte à de futures mémoires hybrides volatiles, à condensateurs (de type DRAM), et non volatiles, utilisant les propriétés ferroélectriques.

2.3.3 Capacité négative

Comme le montre la figure 2.12, chaque domaine ferroélectrique comporte une zone où la courbe P-V est inversée par rapport à un condensateur normal : la zone centrale diminue de façon monotone au lieu d'augmenter.

La pente de la courbe P-V correspondant à la capacité ($C = {}^{Q}/_{V} = {}^{A \cdot P}/_{V}$), cette zone se traduit par une capacité négative, où l'augmentation de la tension diminue la quantité de charges. Une telle capacité négative peut être utilisée en série avec un condensateur ordinaire, pour augmenter la vitesse à laquelle celui-ci accumule les charges. Celle-ci a donc été proposée comme mécanisme pour augmenter la vitesse de commutation des transistors Métal-Oxide-Semiconducteur (MOS) en abaissant la pente sous le seuil en deçà de la limite théorique de 60 mV/dec pour les FETs conventionnels, le dispositif obtenu étant un NCFET.

Toutefois, le maintien de ce régime est extrêmement difficile en pratique :

1. Comme le montre la figure 2.12, le condensateur ferroélectrique peut nécessiter d'être placé dans un état prédéterminé avant de pouvoir fonctionner dans la bonne région.

2. Aucun autre domaine ne doit changer de polarisation pendant le fonctionnement du dispositif, car l'effet inverse serait obtenu : cela limite l'utilisation au domaine ayant le plus faible E_C dans le dispositif.

3. La plage de fonctionnement et la pente varient d'un dispositif à l'autre en raison de la variabilité.

4. Un condensateur paraélectrique existe toujours en parallèle, leur rapport doit donc être ajusté, tout en équilibrant la taille et la variabilité.

2.4 Transistors ferroélectriques

2.4.1 Dispositifs FeFET

Un transistor à effet de champ ferroélectrique (FeFET, Ferroelectric Field-Effect Transistor) est un transistor à effet de champ (FET, Field Effect Transistor) dans lequel une couche de matériau ferroélectrique aide à contrôler le champ électrique du canal. Le plus souvent, il s'agit de transistors MOS comprenant un matériau ferroélectrique dans l'empilement de la grille.

(a) Symbole du FeFET (b) Empilement FeFET MFMOS

FIG. 2.18 : Symbole du FeFET à côté d'un empilement de grille possible, ajoutant une couche de métal et d'oxyde ferroélectrique à l'empilement d'un MOS.

Le principal avantage comparatif des FeFETs par rapport aux condensateurs ferroélectriques est la possibilité de lecture non destructive offerte par l'utilisation du transistor comme mécanisme de lecture intégré. Après le renversement de polarisation, des charges sont libérées de la surface de la couche ferroélectrique. Plutôt que de retourner à l'alimentation, celles-ci sont piégées entre la grille du transistor et la couche ferroélectrique, les deux agissant comme des condensateurs en série. Ces charges piégées influencent la création d'un canal dans le transistor : en d'autres termes, elles déplacent sa tension de seuil, comme détaillé dans la sous-section 4.1.2.

À son tour, ce déplacement de la tension de seuil permet à la valeur stockée d'être lue de manière non destructrice, tandis que les charges restent piégées, ce qui rend la mémoire non volatile.

Il convient de souligner que la conductance du transistor dépend du nombre de charges piégées au niveau de la grille du transistor, et non directement de la polarisation ferroélectrique : le mécanisme de rétention est différent, et proche de celui des mémoires flash. En effet, la couche ferroélectrique sert essentiellement de mécanisme de contrôle pour le chargement du nœud flottant, par opposition au courant de tunnel utilisé pour les mémoires flash. La proximité du condensateur ferroélectrique avec la grille du transistor améliore la rétention des charges piégées, mais celles-ci peuvent encore se dissiper avec le temps. Dans ce cas, il serait peut-être possible de les restaurer en inversant de nouveau la polarisation ou grâce à un courant tunnel contrôlé par la FTJ, bien que cela n'ait pas encore été étudié.

La caractéristique I_{DS}—V_{GS} du transistor étant modulée par l'information stockée dans la couche ferroélectrique, cela peut être exploité pour effectuer des opérations logiques, comme détaillé dans le chapitre 4.

Désavantages

Malgré l'aspect très intéressant de la lecture non destructive, l'utilisation de structures début de ligne (FEoL, Front-End of Line) FeFET, directement intégrées sur les transistors, présente de multiples inconvénients :

- Le problème le plus critique est l'augmentation de la tension requise pour appliquer le champ E_C nécessaire à travers la couche ferroélectrique. Ces tensions peuvent facilement dépasser 3 V, ce qui nécessite des transistors compatibles capables de supporter ces tensions relativement élevées. Cela peut s'avérer impossible avec certaines technologies de fabrication. Des transistors spéciaux peuvent être nécessaires, car l'augmentation de la tension accroît l'intensité du champ électrique sur leur propre pile de grille, ce qui risque de faire claquer l'oxyde. Les circuits contrôlant les FeFETs nécessitent donc des transistors intégrant des oxydes de grille plus épais.

- Des champs électriques intenses peuvent claquer l'oxyde à l'intérieur de l'empilement de grille du FeFET lui-même, ce qui entraîne une usure supplémentaire et une endurance moindre par rapport aux condensateurs ferroélectriques. La rupture se produit généralement après 10^4 à 10^6 cycles. Ce phénomène est aggravé par la permittivité habituellement plus faible de l'oxyde interfacial, donc soumis à une plus grande fraction de l'intensité du champ, tout en étant plus mince[LHS22].

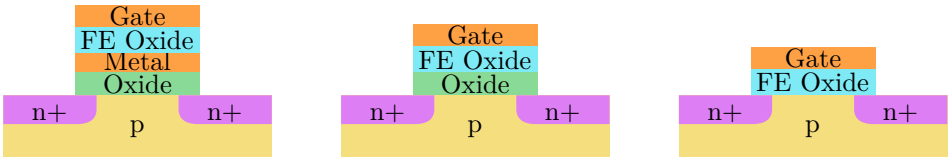
- Le rapport de surface et donc de capacité entre la couche ferroélectrique et le transistor est fixe, ce qui rend la conception moins flexible, contrairement à l'intégration bout de ligne (BEoL, Back-End of Line) de FeCaps.

L'impact de ces problèmes peut être réduit par l'utilisation de condensateurs ferroélectriques BEoL connectés à la grille d'un FET conventionnel, comme décrit dans section 3.3. L'utilisation d'un empilement de grille optimisé peut également atténuer certains des problèmes susmentionnés.

2.4.2 Empilements de grille

La couche ferroélectrique d'un FeFET étant elle-même un diélectrique, il existe de multiples façons de l'intégrer au-dessus d'une porte de transistor MOS, comme illustré par la figure 2.19 :

1. Métal-Ferroélectrique-Métal-Oxyde-Semiconducteur (MFMOS, MFMIS pour Isolant, ou plus communément MFM) illustré en figure 2.19a, dont la structure est similaire à celle du Pseudo-FeFET (PsFeFET) décrit dans la section 3.3
2. Métal-Ferroélectrique-Oxyde-Semiconducteur (MFOS), illustré en figure 2.19b
3. Métal-Ferroélectrique-Semiconducteur (MFS), représenté sur la figure 2.19c



(a) MFM (b) MFOS (c) MFS

Fig. 2.19 : Représentation en coupe des empilements de grille de FeFET possibles, en retirant progressivement la couche métallique intermédiaire (2.19b), puis la couche d'oxyde (2.19b) de l'empilement métal-ferroélectrique-métal-oxyde-semiconducteur (2.19a).

Le premier de ces empilements de grille contient une couche métallique ce qui rend le champ électrique plus homogène à travers la couche ferroélectrique. En effet, une différence de potentiel peut être appliquée entre la grille du transistor et la masse (substrat p), la source et le drain (zones dopées n), ou l'ensemble de ces éléments. Selon les connexions possibles à ces bornes, un gradient du champ électrique peut apparaître à travers l'oxyde ferroélectrique, ou le champ peut ne pas être orienté verticalement, ce qui pourrait nécessiter des tensions plus élevées pour polariser le dispositif[Ni+18]. En outre, les charges libérées peuvent ne pas être mobiles sur la surface du ferroélectrique, ce qui crée un champ de dépolarisation local menaçant la rétention[CL07, p. 29] et ayant un effet local sur le canal du transistor. Il a également été démontré que la couche métallique supplémentaire réduisait la taille des grains (domaines ferroélectriques)[Led+21 ; Leh+21], ce qui est souhaitable pour réduire la taille des dispositifs. Les empilements de grille MFMOS sont donc généralement préférables aux MFOS, bien qu'ils ne soient pas toujours concrètement réalisables, rendant les empilements MFOS plus courants.

Une seconde différence notable entre les empilements de grille MFMOS et MFOS est la position et les dimensions de la zone où se produit la compensation de charges. Alors que la compensation des charges de la surface ferroélectrique inférieure se produit principalement dans le canal du transistor dans le cas les empilements sans métal, celui-ci se produit dans la couche métallique intermédiaire pour l'empilement MFM, en raison de l'excellente conductivité du métal. Ce déplacement de charge modifie la charge électrique locale de la couche métallique.

Cependant, les courants de fuite peuvent entraîner une perte progressive de cette polarisation, plus importante que dans les autres empilements de grille. À leur tour, les charges polarisant la couche métallique sont compensées dans le canal du transistor ; plutôt que celui-ci ne compense directement les charges à la surface de la couche ferroélectrique.

Des développements récents sont survenus concernant la fabrication de MFS, notamment grâce à la déposition en BEoL de FeFETs[Dut+22]. Ces transistors ont un empilement de grille plus fin, ce qui se traduit par des champs plus élevés à des tensions équivalentes, réduisant ainsi V_C pour le même E_C. En outre, moins de problèmes de piégeage de charges entre la couche ferroélectrique et la couche d'oxyde intermédiaire sont rencontrés, et la couche d'oxyde intermédiaire (généralement de plus faible κ, donc exposée à des champs plus intenses[LHS22]) est moins sujette au claquage[Dut+22]. Cependant, les oxydes de grille plus fins se décomposent aussi plus facilement sous l'effet de champs électriques élevés, ce qui limite les gains de tension réalisables.

2.4.3 Modélisation

Un transistor ferroélectrique peut être modélisé par un condensateur ferroélectrique connecté en série avec la grille d'un transistor MOS. Cependant, une telle approche directe de la modélisation suppose que le matériau ferroélectrique est déposé sur le métal de la grille du MOS. Bien que cette approche puisse fonctionner pour divers empilements de grilles, le modèle doit idéalement être ajusté pour considérer ou compenser l'orientation du champ, ce qui modifie la distribution apparente de E_C sur les domaines.

Une autre approche possible consiste à utiliser un modèle agissant comme un générateur de tension en série avec le transistor MOS, contrôlé par l'état de la capacité MOS[WA17]. Cette approche peut mieux refléter les empilements MFS, mais nécessite l'ajustement du modèle de transistor MOS, contrairement à un modèle indépendant. En outre, un modèle en série avec le transistor MOS est capable de mesurer le courant chargeant la capacité de grille, montant qu'une telle approche intégrée n'est pas forcément nécessaire.

Un modèle plus précis, particulièrement pour les empilements non MFM, devrait idéalement être complètement intégré aux modèles de transistors MOS, ayant ainsi accès aux intensités et orientations des champs électriques internes. Les approches existantes ont été jugées satisfaisantes pour les besoins actuels de conception de circuits.

2.5 État de l'art des circuits ferroélectriques

2.5.1 Hafnie ferroélectrique

Bien que ce ne soit pas l'objet du présent document, les techniques de dépôt ferroélectrique ont grandement progressé au cours du projet 3εFERRO[Bou+19 ; Bou20 ; Fra+19b], en s'appuyant sur des travaux antérieurs, notamment sur les projets effectués à NaMLab et GlobalFoundries[SHF19]. Cela s'est accompagné d'une meilleure compréhension des effets physiques conduisant à la création de la structure orthorhombique ferroélectrique dans les oxydes de hafnium dopés, permettant ainsi l'amélioration les rendements, de Pr, et d'autres critères de performance tels que l'endurance ou la réduction des effets de réveil (wake-up) [Mue+13b ; Alc+22].

Modélisation

La première étape de la conception d'un circuit consiste à obtenir des modèles fiables des dispositifs utilisés. Les circuits décrits dans le présent document ont d'abord été étudiés à l'aide d'un modèle de Landau[WA17]. Bien que suffisant pour valider le fonctionnement général des circuits, ce modèle simple ne tient pas compte des multiples phénomènes physiques susceptibles de perturber leur fonctionnement. En outre, ce modèle n'était pas calibré sur les procédés de fabrication visés.

Davantage de données expérimentales et une meilleure compréhension du comportement des dispositifs ferroélectriques ont également conduit à l'adoption de modèles améliorés, des modèles de Preisach [Yin+19], aux modèles plus complets, représentant une plus large gamme de comportements, y compris celui d'accumulation [Den+20].

2.5.2 Conception de circuits Condensateur Ferroélectriques et bout de ligne

Les partenaires du projet au CEA-LETI ont conçu puis fabriqué un tableau mémoire 1T-1C de 16 kbit[Fra+19a ; Fra+21] utilisant des FeCaps BEoL, montrant des performances prometteuses[Gre+20 ; Oku+21].

Parmi les autres utilisations de FeCaps, on peut citer l'informatique neuromorphique, où leur comportement de commutation accumulative peut émuler le mécanisme d'intégration des synapses[Maj22].

Récemment, de multiples architectures de FeFET BEoL ont été proposés, allant de l'intégration BEoL de dispositifs FeFET complets[Dut+22], à des structures semblables aux FeFET, telle que celle présentée dans la section 3.3[Ni+18 ; Leh+21]. Le présent document couvre uniquement ces dernières.

2.5.3 Circuits utilisant des transistors à effet de champ ferroélectriques

Les FeFETs sont une application extrêmement prometteuse des matériaux ferroélectriques, réalisant essentiellement une cellule de mémoire MOSFET à grille flottantes (FGMOSs) ou flash, où l'accumulation de charges sur la grille flottante est contrôlée par l'inversion de la polarisation.

Cela permet de contrôler la tension de seuil des transistors, comme décrit dans la sous-section 4.1.2 et la section 3.3. À son tour, ce système peut être utilisé pour concevoir un ensemble de portes logiques élémentaires[Bre+18 ; Mar+21].

Divers circuits utilisant des FeFETsont été proposées récemment. Les circuits les plus prometteurs incluent des mémoires adressable par contenu (CAMs, Content-Addressable Memory) et mémoires ternaires adressables par contenu (TCAMs, Ternary Content-Addressable Memory), ainsi que des tableaux de FeFETs pour le calcul hyperdimensionnel. Des gains importants en termes de performances et de surface sont anticipés pour ces applications, par rapport aux architectures CMOS classiques : par exemple, la surface peut être réduite à un dixième et le produit énergie-délai peut être amélioré d'un facteur quatre pour un circuit TCAM[Yin+19].

Les FeFETs sont adaptés à la conception de CAMs, en raison de leur lecture non destructrice et de leur sortie en conductance. Cela permet de réaliser des tableaux de CAM extrêmement compacts, avec un bit par FeFET, en connectant leurs sorties en série. Le circuit TCAM présenté dans [Yin+19 ; Ni+19] utilise ce principe, avec l'utilisation de deux FeFET la réalisation des opérations ternaires. Ce circuit peut également être modifié pour servir de mémoire vive (RAM, Random-Access Memory)[Mar+22], permettant l'accès en lecture et en écriture à des bits individuels. L'opération multiniveau peut par ailleurs être exploitée pour stocker plusieurs bits par FeFET, créant ainsi des mémoires multi-bit adressables par contenu (MCAMs, Multi-bit Content-Addressable Memories)[Kaz+21a ; Nie+23], avec des applications possibles en calcul hyperdimensionnel[Kaz+21b]. La précision peut rivaliser avec celle des implémentations CMOS conventionnelles, tandis que les performances et la densité sont sensiblement améliorées[Kaz+22].

Les FeFET ont également démontré un comportement neuromorphique pour une utilisation dans des synapses artificielles, en tirant parti de la commutation progressive de plusieurs domaines[Mul+17 ; Jer+17 ; Maj22].

Par ailleurs, ceux-ci sont également adaptés aux architectures conventionnelles de réseaux neuronaux profonds, utilisant intensivement des opérations matrice-vecteur et multiplication-accumulation. Les FeFETs sont des processeurs de flux naturels, ce qui est exploité dans la section 4.6 pour réaliser un multiplicateur. D'autres architectures ont été proposées, notamment un multiplicateur matrice-vecteur dans [Yoo+19], exploitant une sortie de courant différentielle pour effectuer l'accumulation, bien que cette approche nécessite l'utilisation de convertisseurs analogiques-numérique.

Transistors à effet de champ ferroélectriques canal P

Des progrès récents ont été réalisés dans la fabrication de FeFETs à canal p, notamment sur des substrats en germanium [Zac+22] et en silicium[Kle+21].

Cela ouvre la voie à des circuits de type CMOS, qui éliminent la nécessité d'une alimentation dynamique ou d'un circuit de transrésistance, comme détaillé dans la section 4.3, à condition que leur courant de fuite soit faible.

2.5.4 Comparaison avec d'autres Mémoires non volatiles

Aujourd'hui, les mémoires flash sont la solution Mémoire non volatile omniprésente, utilisée pour des applications allant des cartes SD aux SSDs grand public et d'entreprise, mais également aux flash intégrés (eFlash) dans les microcontrôleurs. Cela est principalement dû à la maturité de la technologie, à sa haute densité et à son faible coût, ainsi qu'à sa facilité de fabrication. Cependant, celles-ci souffrent de multiples inconvénients limitant leur utilité : faible vitesse d'écriture, forte consommation d'énergie, faible endurance et vulnérabilité aux radiations. La technologie flash a de plus été fortement optimisée, et il n'est pas certain que sa mise à l'échelle puisse continuer bien en deçà des dimensions actuelles.

Cela a entraîné le développement d'une gamme de Mémoire non volatile « émergentes », commençant à concurrencer les mémoires flash dans des applications spécifiques. De multiples candidats présentant des caractéristiques de vitesse élevée et de faible consommation énergétique sont apparus : MRAM, PCM, ReRAM et FeRAM. Parmi ceux-ci, les FeRAM construites à partir de $HfZrO_2$ devraient atteindre une excellente endurance et une faible consommation d'énergie. Cela en fait un candidat prometteur pour remplacer la technologie flash dans les applications embarquées, en particulier dans les cas d'utilisation de normalement-éteint, afin de réduire l'énergie dépensée lors du stockage et de la récupération de l'état du processeur.

Les caractéristiques souhaitables pour les Mémoires non volatiles embarquées (eNVM) sont des temps d'accès rapides, une faible consommation énergétique et une adaptabilité autant aux microcontrôleurs à faible coût, qu'aux produits haute performance dans un environnement exigeant (de type automobile). Une grande fiabilité et un faible coût sont requis, ce qui comprend la compatibilité avec les étapes de fabrication telles que les opérations de refusion, où les éléments de mémoire préprogrammés sont soudés sur des cartes de circuits imprimés. Les applications embarquées nécessitent généralement de plus petites quantités de mémoire, ce qui réduit l'avantage de densité de la technologie flash. Deux problèmes majeurs ont jusqu'à présent retardé l'adoption généralisée des mémoires non volatiles intégrées émergentes :

1. le manque de confiance dans leur maturité, leur fabricabilité et leur fiabilité,
2. capacité de fonctionnement à haute température limitée.

Si les FeRAM $HfZrO_2$ présentaient d'ores et déjà une compatibilité prometteuse avec les environnements à haute température, le projet 3εFERRO a permis de démontrer leur facilité d'intégration et leur fiabilité. La tableau 2.2 résume les caractéristiques de performance de la génération actuelle de FeRAM utilisant $HfZrO_2$ (1T1C ainsi que FeFET), et les compare à la technologie flash plus mature ainsi qu'à d'autres technologies émergentes. Les deux principaux avantages des dispositifs ferroélectriques utilisant $HfZrO_2$ par rapport aux autres technologies de Mémoire non volatile émergentes pour les applications embarquées sont leur très faible consommation d'énergie et leur facilité de co-intégration. Les mémoires ferroélectriques réalisées à partir de $HfZrO_2$ ont démontré un fonctionnement très économe en énergie, une endurance élevée (bien que réduite par les opérations de lecture destructives dans les structures 1T1C), une vitesse élevée, un faible coût et une stabilité en température satisfaisante[Fra+21]. Les problèmes restants concernent le piégeage de charges, la dérive de la boucle ferroélectrique (fatigue, réveil, empreinte) et de multiples phénomènes liés au cyclage et à la rétention des données.

2.5.5 Exploration de l'espace de conception

Le travail présenté dans le chapitre 5 se concentre sur la exploration de l'espace de conception (DSE, Design-Space Exploration), et s'appuie sur des outils internes précédemment développés[Bri21].

La DSE des circuits ferroélectriques n'est pas souvent mentionnée dans la littérature. Au niveau du dispositif, une étude[LHS22] a tenté d'optimiser la géométrie des FeFET en évaluant l'impact d'entretoises high-k à l'aide de simulations simulation technologique physique (TCAD, Technology Computer-Aided Design).

	Flash (mature)	MRAM	PCM	ReRAM	$HfZrO_2$ FeRAM FeFET	$HfZrO_2$ FeRAM 1T1C
Energie d'écriture (pJ/bit)	<200[a]	0.1 à 20[b]	~90	~100	<20	<0.1 [Fra+21]
Taille de cellule (μm^2)	0.05	<0.01 [Ike+20]	<0.04	0.1	0.05	0.04
Temps d'accès lecture (ns)	~15[c]	~1[d]	~5 (Pas de haute tension)			~4[e] [Fra+21]
Granularité d'effacement	Page	Bitcell[f]				
Endurance	~10^6, note[g]	10^{15} [Gir+21]	5×10^5	10^5, note[h]	10^5, note[i]	10^{11}, note[j]
Rétention	~10 ans à 55 °C[k]	Compromis énergie[l]	150 °C, note[m]	Compromis énergétique	>10 ans [Mül+15]	Élevé[n]
Brasage par refusion	Mature	Difficile[o]	Prouvé	Possible	Probable[p]	Prouvé [Fra+21]
Masques supplémentaire	Nombreux (>10)	Limités (3 à 5)			Peu (1 à 3)	
Procédé de fabrication	Complexe		Simple			
Nouveaux éléments vs CMOS	Aucun	Nouveaux (fabriquables)	Matériaux high-k[q]			

[a] ~100 pJ/bit pour les dispositifs Embedded Select in Trench Memory (eSTM).

[b] Démonstration d'un taux d'erreur inférieur à 100 fJ/bit pour les dispositifs dont la taille est réduite à 11 nm, mais avec des taux d'erreur accrus[Now+16]. Compromis avec la rétention.

[c] La flash utilise des dispositifs à haute tension, ce qui entraîne des coûts énergétiques élevés

[d] Entièrement compatible avec les éléments logiques (tant en termes de vitesse que d'intégration/voltage), pourrait constituer une alternative aux cellules SRAM rapides[Gal+19b].

[e] Lecture destructive : la FeRAM 1T1C a besoin de Réécriture (WB, Write Back) après lecture, mais pas de niveaux de tension élevés

[f] Granularité au niveau de la bitcell, bien que dépendante de l'architecture

[g] Pour la technologie flash, la fréquence des cycles réduit proportionnellement la rétention des données. L'endurance visée semble être de 10^6 cycles, jusqu'à 10^7 dans certains cas. [AG21] énonce 10 ans de rétention des données à 55 °C après 10^6 écritures réparties sur 10 années, ou 2 ans si 10^6 écritures sont réparties sur 18 mois.

[h] compromis avec le taux d'erreur binaire

[i] Les grilles des transistors sont exposées à des champs électriques de forte intensité, ce qui dégrade l'oxyde de la grille. Les charges piégées et d'autres mécanismes limitent également l'endurance à 10^5 à 10^6 cycles.

[j] moitié pour la lecture à cause du WB

[k] La durée de rétention est un compromis avec la fréquence d'écriture, et diminue avec la température. [AG21] cite 20 ans après 10^5 cycles répartis sur 18 mois à 55 °C

[l] La principale faiblesse des MRAM est le compromis énergie-rétention, bien que celui-ci soit ajustable pour favoriser une faible consommation d'énergie ou une longue durée de rétention. [Now+16] montre une valeur proche de 10 ans à température ambiante pour des cellules miniaturisées.

[m] conforme aux applications automobiles

[n] Élevé en théorie, les études expérimentales[Fra+19a ; Fra+21] n'ont pas pu extrapoler la durée de rétention, la fenêtre mémoire restant ouverte après 10^4 s à 125 °C

[o] L'état stocké est très sensible au processus d'assemblage par refusion, bien que cela puisse être compensé par l'utilisation de courants d'écriture plus élevés, du blindage, l'utilisation de codes de correction d'erreur ou des dispositifs de taille supérieure [Gal+19b ; Gal+19a].

[p] Non démontré expérimentalement

[q] FEoL pour les FeFET, BEoL dans les autres cas.

Tab. 2.2 : Comparaison de mémoires FeRAM et autres technologies Mémoire non volatile émergentes, incluant la flash comme technologie mature de référence. Cette dernière est très répandue, compte de nombreux fournisseurs, et ses caractéristiques de performance sont connues. Ce tableau met en évidence les caractéristiques problematiques, ainsi que les problèmes majeurs et mineurs d'industrialisationn ainsi que et les performances visées.
Dérivé de documentation interne au projet par STMicroelectronics.

D'autres études se sont concentrées sur le cas d'utilisation NCFET[SR17 ; YS17 ; Pal+18], qui n'est pas évalué dans ce document. Néanmoins, il est intéressant de noter qu'elles s'appuient toutes sur le modèle de Landau décrit dans la sous-section 2.2.1.

Une DSE des tableaux de mémoire FTJ a été réalisée dans [Jao+21], avec une attention particulièrement portée sur les sélecteurs de mémoire et une comparaison avec les jonctions tunnel magnétiques. Les résultats sont compétitifs, mais soulignent que l'augmentation du courant tunnel entraîne des pertes d'efficacité énergétique pendant la programmation.

Une étude plus proche du présent travail s'est concentrée sur l'impact de la transconductance des FeFET pour les réseaux neuronaux profonds[Yoo+19]. Toutefois, la portée de cette étude est réduite par l'utilisation d'une fonction sigmoïde comme modèle ad hoc pour la linéarité de la transconductance, ce qui découple l'étude des paramètres physiques du dispositif.

2.5.6 Évaluation des performances au niveau du système

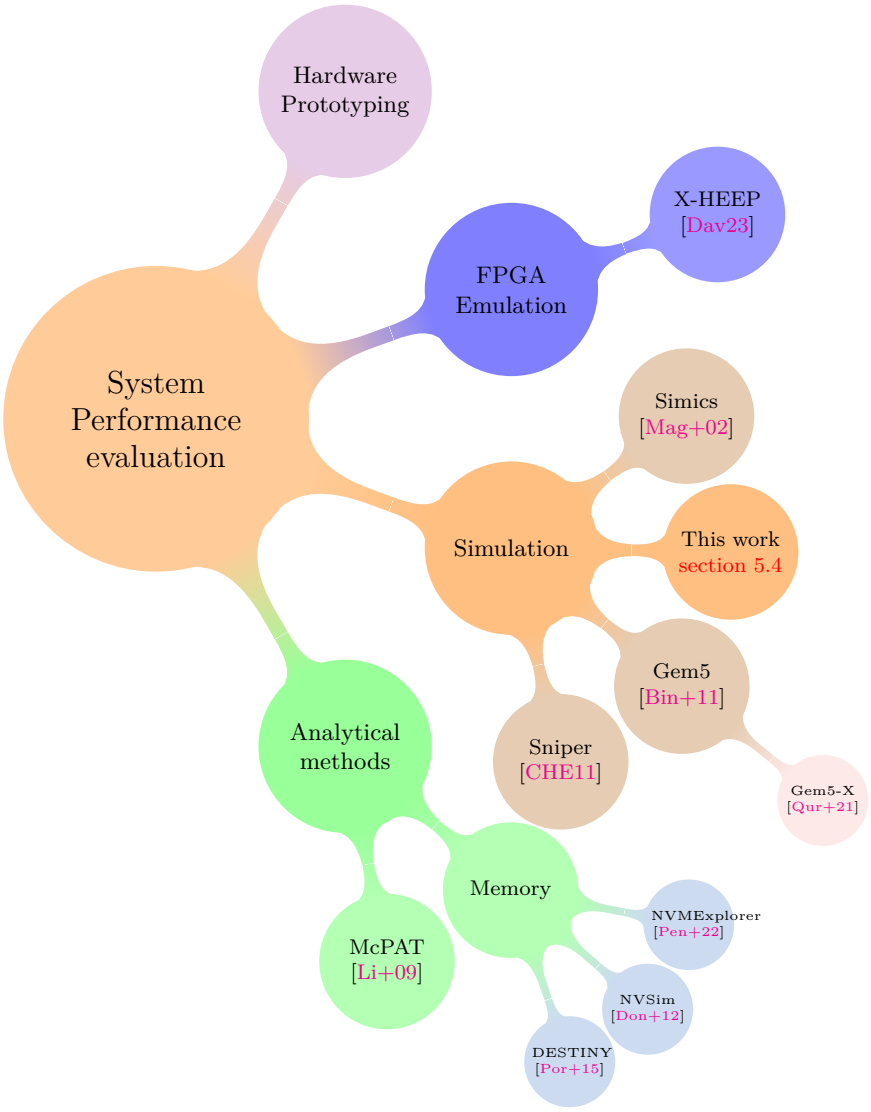


Fig. 2.20 : Environnement de l'évaluation des performances au niveau du système. Bien que ne soit pas un aperçu exhaustif, celui-ci met en contexte les travaux présentés dans la section 5.4.

L'évaluation des performances au niveau du système est un aspect crucial de l'évaluation des technologies émergentes, et les résultats peuvent être exploités pour orienter plus efficacement les efforts de recherche au niveau des dispositifs et des matériaux[Nie+23]. Si la section 5.4

présente une approche focalisée sur un simulateur, il existe de multiples solutions alternatives, comme illustré dans la figure 2.20, et décrites dans cette section.

Prototypage matériel

L'évaluation de performances consiste idéalement en la fabrication d'un prototype et la mesure de ses performances. Cependant, cette stratégie n'est pas économiquement viable pour tester un large éventail de stratégies et de configurations de dispositifs. Cette approche de prototypage peut être approximée par des modèles moins coûteux, matériels ou simulés.

Émulateurs FPGA

De récentes initiatives en matière de matériel libre, notamment le jeu d'instructions RISC-V[Wat16] sous licence libre, permettent d'accélérer les cycles de développement en tirant parti de la réutilisation des composants. La plateforme X-HEEP[2][Dav23] vise ce cas d'utilisation en fournissant des mécanismes pour l'intégration d'accélérateurs nouvellement conçus avec un processeur RISC-V. L'architecture système résultante peut ensuite être fabriquée, mise en œuvre sur un FPGA ou simulée. Les émulateurs FPGA offrent un compromis intéressant : bien que moins flexibles que les simulateurs, leur performance peut être largement supérieure, en particulier dans le cas d'architectures massivement parallèles telles que des tableau reconfigurable à gros grains (CGRAs, Coarse-Grained Reconfigurable Array)[Den+22].

Simulateurs logiciels

Les simulateurs dédiés au niveau système tels que Gem5[Bin+11] et ses dérivés tels que Gem5-X[Qur+21] offrent une plus grande flexibilité, au détriment de la vitesse d'exécution. Il existe également d'autres simulateurs, tels que Sniper[CHE11] et Simics[Mag+02].

Enfin, des méthodes et outils analytiques permettent d'estimer les performances d'une architecture donnée sans recourir à des simulations. Ces estimations peuvent être moins précises, mais sont généralement plus rapides que les simulations et peuvent ne pas nécessiter une conception architecturale complète, mais uniquement des spécifications de haut niveau. Des outils de modélisation tels que McPAT[Li+09] peuvent estimer les caractéristiques de performance au niveau du système à partir de bibliothèques de technologies. Il existe davantage d'outils orientés vers les performances de la mémoire, tels que DESTINY[Por+15], NVSim[Don+12], ou NVMExplorer[Pen+22].

Prise en charge par les compilateurs et instrumentation du code

La prise en charge par les compilateurs est une étape nécessaire à l'utilisation facile et généralisée des architectures non-Von Neumann. Bien que les compilateurs soient capables, dans une certaine mesure, de cibler les accélérateurs à partir de code générique, par exemple grâce à l'auto-vectorisation, cette approche est intrinsèquement limitée. Une utilisation plus optimale du matériel peut être obtenue en le ciblant explicitement. Cela nécessite cependant de réécrire le code et les algorithmes pour tirer parti des avancées architecturales, ce qui rend leur évaluation plus longue et complexe. Cependant, si le code utilise déjà un langage de programmation ou une extension ciblant les coprocesseurs, tels que Halide[Rag+17], ROCm/HIP[HIP23], CUDA[CUDA17], SYCL[SYCL14], OpenMP[OMP], etc., le back-end du compilateur peut être modifié pour tirer parti de la nouvelle architecture.

L'approche décrite dans la sous-section 5.4.3 est différente : le code source des algorithmes existants est d'abord modifié afficher des traces d'exécution ; ces traces sont converties manuellement vers l'architecture cible et rejouées sur le simulateur. Des efforts similaires pour générer automatiquement de telles traces à partir de code instrumenté ou de compilateurs sont en cours au CEA-LETI[Koo+18 ; Mam+21].

[2] eXtendable Heterogeneous Energy-Efficient Platform

Chapitre 3

Circuits à condensateurs ferroélectriques

Contents

3.1 Introduction

3.1.1 Technologie bout de ligne

Grâce à leur compatibilité avec les processus de conception et fabrication CMOS, ainsi qu'à leur température de recuit relativement basse d'environ 450 °C[Bou20, p. 44], les oxydes d'hafnium ferroélectriques dopés (hafnie) tels que $HfZrO_2$ peuvent être déposés sur les couches

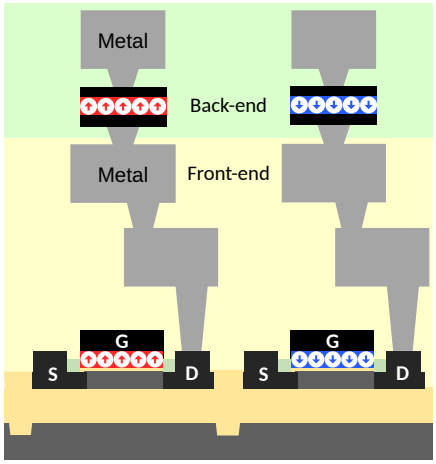


FIG. 3.1 : Illustration de la technologie ferroélectrique frontale (tête de ligne, jaune,◯, bas) et dorsale (bout de ligne, vert,◯, haut). Le bout de ligne (BEoL, Back-End of Line) est déposé après début de ligne (FEoL, Front-End of Line)

métalliques inférieures sans faire fondre les transistors et les interconnexions déjà déposées, comme illustré dans la figure 3.1. Il s'agit de dépôt BEoL, dont l'utilisation actuelle se limite aux Condensateur Ferroélectriques (FeCaps).

Le dépôt BEoL peut être économiquement avantageux dans le cas des FeCaps, car les structures de condensateurs sont généralement plus grandes que les circuits logiques des couches inférieures.

Les Condensateur Ferroélectriques BEoL :

- peuvent être conçues avec une finesse de gravure moindre, ce qui permet d'utiliser des masques et procédés moins coûteux
- peuvent disposer d'une plus grande capacité (physiquement plus grandes) sans réduire la densité des circuits logiques
- ont une taille découplées de celle des transistors, ce qui permet d'élargir l'espace de conception, comme expliqué dans la section 3.3 et la sous-section 3.6.4
- sont généralement de meilleure qualité, ce qui permet d'améliorer l'endurance
- sont plus simples à intégrer, car présentent moins d'interfaces entre oxydes

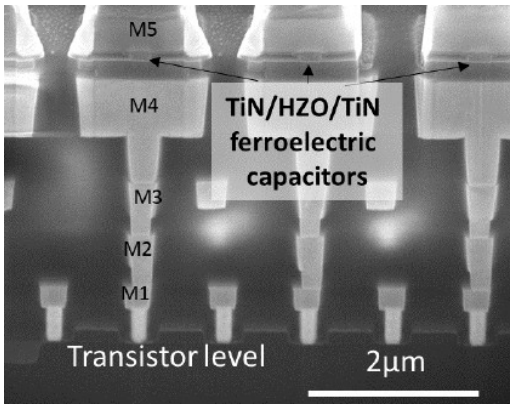
Dans le cadre de cette thèse, des circuits logiques exploratoires utilisant des FeCap conçus en technologie « full-custom » ont été soumis pour la réalisation d'une plaquette multi-projets. Ces circuits ont été fabriqués avec une gamme de condensateurs ferroélectriques de diamètre 300 nm, 400 nm et 550 nm : 2T1C (3.5), « Pseudo-FeFET » (3.3) et TCAM (3.4.1), et seront détaillés dans leur section respective.

3.1.2 Technologie MAD200

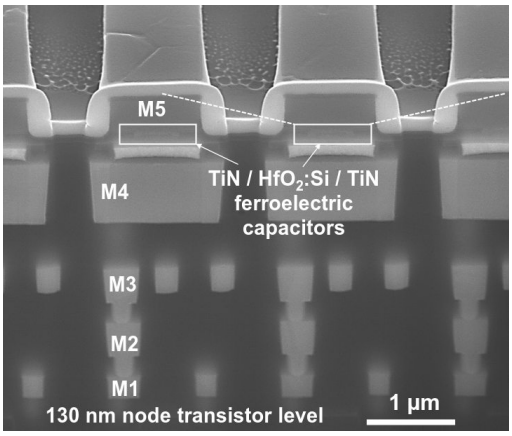
Le processus de fabrication utilisé pour le travail présenté est connu sous le code MAD200, il s'agit d'une variante du processus HCMOS9A de STMicroelectronics, lui-même dérivé du processus technologique 130 nm HCMOS9GP.

Le processus HCMOS9A s'arrête après le dépôt de la quatrième couche métallique (*M4*) de HCMOS9GP. Une couche d'OxRAM est ensuite déposée au-dessus de *M4* par le CEA-LETI, avec une couche métallique supplémentaire (*M5*) pour établir un contact avec l'arrière de la couche OxRAM ainsi qu'avec les plots de contact. La variante MAD200 remplace cette couche OxRAM par une couche de $HfZrO_2$ déposée au CEA-LETI.

La structure résultante est visible sur la figure 3.2, illustrée par la figure 3.5, avec la couche de $HfZrO_2$ visible entre les couches métalliques *M4* et *M5*.



(a) Couche $HfZrO_2$ dans [Fra+19a]



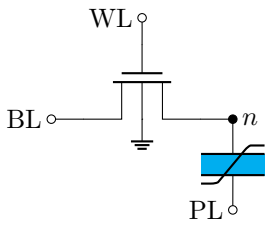
(b) Couche HfO_2 dopée avec Si dans [Fra+21]

FIG. 3.2 : Visualisation au microscope électronique des couches supérieures de MAD200, y compris la couche ferroélectrique BEoL. Images fournies par le CEA-LETI.

3.2 Cellule de mémoire 1T1C

1T1C signifie « un transistor, un condensateur » et décrit succinctement la cellule mémoire illustrée par le Circuit 3.1. Bien que cette structure soit couramment utilisée dans les tableaux de DRAM, un FeCap peut être utilisé à la place d'un condensateur normal dans la cellule mémoire, rendant ainsi la mémoire non volatile. La séparation entre le transistor et le condensateur rend ce circuit particulièrement adapté à l'intégration BEoL, car des condensateurs plus grands peuvent être déposés au-dessus de la logique de contrôle et d'adressage.

Cette cellule mémoire est l'un des circuits ferroélectriques les plus simples. Bien qu'elle ne fasse pas partie des cellules contribuées au démonstrateur MAD200, son mécanisme de fonctionnement est semblable à celui d'autres circuits, et a été étudié par simulation. En outre, le CEA-LETI a conçu un tableau de 16 kbit de RAM 1T1C dans le cadre du projet 3εFERRO, dont le développement a entraîné celui du processus MAD200 [Fra+21].



CIRCUIT 3.1 : Cellule mémoire 1T1C. n est le nœud flottant entre le transistor d'accès et la FeCap. Contrairement aux DRAM ordinaires, le diélectrique du condensateur est composé d'un matériau ferroélectrique, ce qui permet un fonctionnement non volatil.

3.2.1 Opération

Sélection et programmation des cellules mémoire

Comme les cellules DRAM, cette structure est adaptée à une intégration dense dans un tableau de mémoire tel que celui illustré par le Circuit 3.2. Dans ces réseaux de condensateurs[Mik+19], une cellule unique est sélectionnée à l'intersection d'une Word Line (WL) et d'une Bit Line (BL), en activant tous les transistors d'accès sur une WL donnée et en portant la BL voulue à la tension nécessaire.

Comme le montre le Circuit 3.2, la cellule située à l'intersection est ainsi sélectionnée, le condensateur étant connecté à la BL active. La seconde électrode de chaque FeCap est

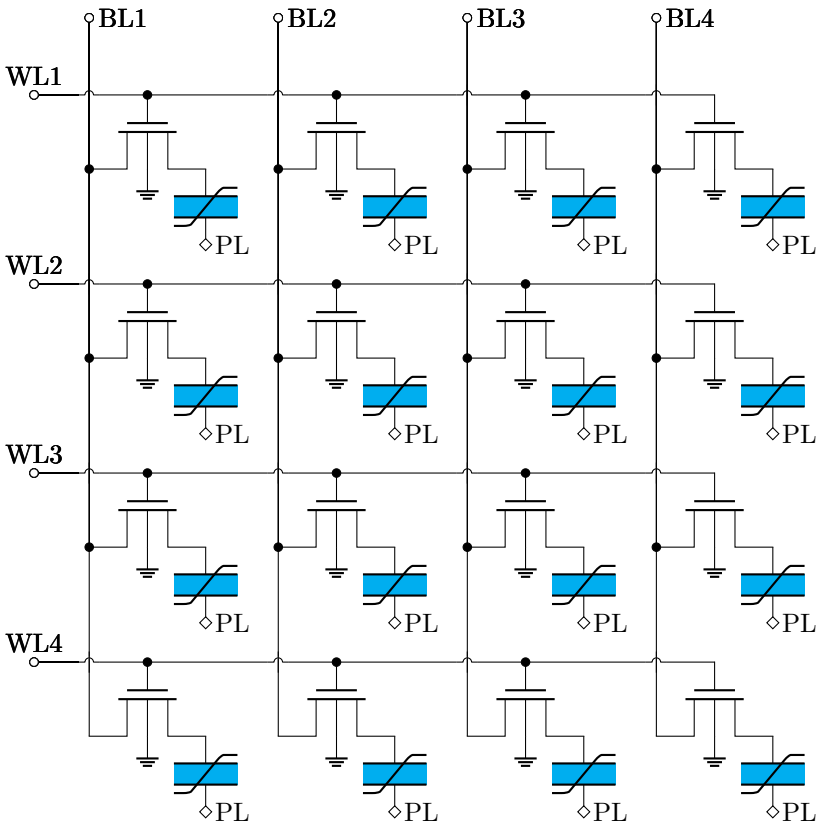
connectée à la Plate Line (PL) globale. Une tension peut donc être appliquée aux bornes du condensateur en contrôlant $V(BL - PL)$.

Pour modifier la polarisation du FeCap, l'intensité du champ électrique traversant le diélectrique doit dépasser E_C ; la tension appliquée entre BL et PL doit donc être supérieure à $\pm V_C$. Deux approches sont possibles pour contrôler la polarité de la tension appliquée :

1. l'application directe de tensions positives ou négatives à la BL ou PL tandis que l'autre est conservée à une tension fixe, ou

2. l'application d'une tension positive soit sur la BL, soit sur la PL tandis que l'autre reste au potentiel de la masse.

La première approche nécessite l'application de tensions négatives, qui peuvent être complexes à générer et à distribuer sur la puce, mais simplifie la conception des circuits, en permettant par exemple de relier PL à la masse. La seconde approche nécessite davantage de circuits de contrôle, afin de fournir une tension positive à BL ou PL, mais ne nécessite pas l'utilisation de tensions négatives.



Circuit 3.2 : tableau 1T1C 4×4 montrant quatre WL différentes, quatre BL, et la PL commune.

Lecture de la cellule mémoire

Afin de lire le contenu stocké de la cellule mémoire, et de manière semblable à la DRAMs, la procédure de programmation est répétée avec une tension et polarité connues. Cette tension est choisie suffisamment élevée pour permettre l'inversion de la polarisation des FeCap ($|V| > V_C$), tandis qu'un amplificateur de détection mesure le courant de la BL ou PL. Si la polarisation a été inversée lors de la procédure d'écriture, un pic de courant correspondant à la libération des charges piégées est enregistré, et la valeur précédemment enregistrée peut être déduite : si un courant de repolarisation est détecté, la valeur précédemment stockée était de

polarité opposée ; si aucun courant n'est détecté (outre le courant de charge du condensateur paraélectrique), la polarité appliquée correspond à celle déjà mémorisée par la FeCap.

Cependant, la détection d'un pic de courant signifie que la valeur précédente a été effacée et qu'elle doit donc être réinscrite en mémoire si celle-ci doit être utilisée ultérieurement. La procédure de lecture est de ce fait destructive, comme dans le cas de la DRAMs.

La quantité de charges libérées est simplement :

$$\Delta Q = 2 \cdot Pr \cdot \mathrm{A_{FE}} \tag{3.1}$$

Le facteur 2 existe, car la repolarisation a inversé la polarité de surface de $\pm Pr$ et changé celle-ci pour la valeur opposée.

Mémoire à plusieurs niveaux

Le stockage multiniveau est largement adopté dans l'industrie dans le cas d'autres technologies de mémoire, communément nommées MLC. L'utilisation d'un seul élément de mémoire pour le stockage de plusieurs bits de données en tirant parti d'états intermédiaires permet d'augmenter considérablement la densité de stockage. En contrepartie, cela entraîne une complexification des circuits, et une diminution des performances en termes de temps de rétention, de vitesse de lecture et d'écriture, ainsi que d'endurance : il n'est plus possible d'accéder directement aux bits individuels, qui doivent être lus et écrits par lots correspondant au nombre de niveaux de chaque cellule mémoire. Cela augmente le nombre de cycles d'écriture par cellule, car la lecture d'un seul bit nécessite généralement la réécriture de l'ensemble de la cellule. Inversement, l'écriture d'un seul bit nécessite la lecture (destructive) préalable de la cellule. Cela incite fortement à éviter les applications nécessitant des accès aléatoires au niveau de granularité du bit, qui réduiraient considérablement l'endurance des cellules multiniveaux. Ce cas d'utilisation est heureusement peu fréquent, et les vitesses de lecture et d'écriture bénéficient également des opérations multibits.

Le facteur limitant est la fenêtre mémoire : à mesure que le nombre d'états distincts par cellule binaire augmente, la fenêtre mémoire disponible pour chaque état diminue. Cela augmente le taux d'erreur, réduit le temps de rétention et nécessite des circuits de correction d'erreur supplémentaires. La réduction de la fenêtre mémoire rend également les cellules MLC plus sujettes à la fatigue, la fenêtre se rétrécissant avec l'augmentation du nombre de domaines défaillants. Les cellules à niveau unique pourraient accroître la tension de programmation en réponse, afin d'exploiter des domaines vierges, tandis que les cellules MLC sont plus susceptibles d'utiliser la gamme de tension complète dès le début de leur utilisation.

Il est possible de réaliser des mémoires multiniveaux avec des matériaux ferroélectriques, ce qui les rend plus compétitifs en termes de densité de stockage : le matériau ferroélectrique polycristallin présente une gamme de champs coercitifs distribués sur les différents domaines ferroélectriques, en raison de leurs orientations aléatoires, ainsi que de la variabilité du processus de fabrication. Si la distribution est suffisamment large, une partie de la population peut être polarisée de manière sélective : les domaines cristallins présentant un faible champ coercitif peuvent être programmés à l'aide d'une tension de programmation plus faible ou d'impulsions de programmation plus courtes.

Un niveau unique (un bit de données par cellule) ne possède que deux orientations possibles, l'une d'entre elles étant forcée lors de la lecture. Par conséquent, la probabilité qu'une valeur doive être réécrite est seulement $P(WB)_{SLC} = 1/2$: un protocole de lecture optimisé peut essayer de prédire la valeur actuelle afin d'éviter les réécritures. Pour les cellules à n niveaux avec 2^n états possibles, la probabilité de réécriture augmente à $P(WB)_{MLC} = {}^{2^n-1}/_{2^n}$, ce qui rend un tel mécanisme de prédiction moins utile, à moins qu'il ne soit très précis.

Cet état intermédiaire partiellement polarisé se traduit par une quantité proportionnellement plus faible de charges détectées lors de la phase de lecture, forçant les domaines dans un état connu. La quantité de charges électriques détectées étant directement liée à la largeur de la fenêtre de mémoire, celle-ci doit être suffisamment grande pour distinguer une polarisation partielle d'une inversion totale de la polarisation. Cela implique l'utilisation de condensateurs plus grands pour les cellules à plusieurs niveaux, afin d'augmenter la population de domaines ferroélectriques. Une large distribution est plus adaptée aux cellules à plusieurs niveaux, car cela diminue également la variabilité entre les cellules.

En effet, dans le cas d'une cellule MLC à $n-$ niveaux partiellement polarisée, en supposant que l'impulsion de lecture vise une polarisation complète (tension maximale, conduisant à $\pm Pr$) à partir du $s^{\text{ième}}$ niveau, un facteur supplémentaire est simplement introduit dans l'équation 3.1 :

$$\Delta Q = \frac{s}{2^n} \cdot 2 \cdot Pr \cdot \mathrm{A_{FE}} \tag{3.2}$$

Cela montre que la distinction entre les niveaux individuels s_i devient plus difficile à mesure que le nombre n de niveaux augmente, à moins que $\mathrm{A_{FE}}$ ne soit augmentée à son tour, ou que des améliorations technologiques ne soient apportées pour augmenter Pr.

Pour faciliter la lecture de cellules MLC, il est également possible d'augmenter progressivement la tension appliquée et de détecter l'apparition des premières repolarisations. Cela permet de réduire les contraintes sur l'amplificateur de détection et l'Convertisseur Analogique-Numérique (CAN), permettant également l'utilisation d'un CANs 1 bit, mais ralentit l'opération de lecture et nécessite également un contrôle précis de la tension appliquée.

3.2.2 Simulation

Étant donné la simplicité du circuit, celui-ci a été choisi pour valider la fonctionnalité et la stabilité des modèles de simulation. Associées à des outils d'exploration de l'espace de conception, les simulations ont permis de prédire les performances réalisables, comme discuté dans la section 5.3.1, et d'affiner les modèles en les comparant aux valeurs expérimentales.

La cellule 1T1C a été simulée avec deux *design kit* :

- STMicroelectronics 130 nm MAD200
- GlobalFoundries 28 nm 28SLP

Simulations en technologie MAD200

Bien qu'aucune cellule 1T1C n'ait été spécifiquement conçue pour la technologie MAD200, celle-ci est proche de la cellule 2T1C décrite dans la section 3.5. La seule différence consiste en un transistor de lecture supplémentaire, dont l'utilisation est facultative, et qui équivaut à une capacité parasite supplémentaire s'il n'est pas utilisé. En effet, la majeure partie du transistor reste alors connectée au rail d'alimentation dans l'implémentation choisie.

Les simulations ont été effectuées avec des paramètres choisis pour évaluer la performance des cellules mémoire, afin de fournir des mesures de performance aux simulations de plus haut niveau, comme décrit dans la section 5.4.2.

L'espace des paramètres a été exploré de manière plus approfondie dans la sous-section 5.3.1 afin de déterminer les compromis de conception.

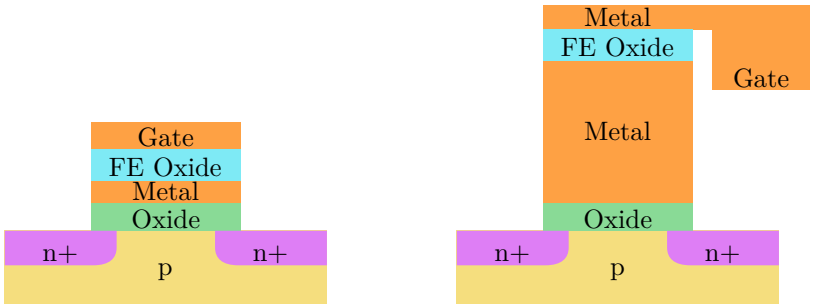
3.3 Structure de type FeFET

3.3.1 Description

Les FeFETs, comme détaillé dans le chapitre 4, ne sont pas disponibles en technologie BEoL, car l'oxyde ferroélectrique n'est pas déposé directement sur l'oxyde de grille, contrairement à un procédé FEoL. Il s'agit ici d'une couche supérieure (entre les couches M4 et M5 dans le processus MAD200). Cela peut être considéré comme un empilement métal-ferroélectrique-métal-oxyde-semiconducteur (MFM)[Leh+21], avec la seconde couche de métal remplacée par plusieurs via connectant la grille du transistor à la couche ferroélectrique.

L'objectif de cette structure expérimentale est d'explorer la faisabilité de reproduire le fonctionnement FeFET en technologie BEoL, en connectant directement le contact de grille à la couche ferroélectrique par l'intermédiaire de via métalliques, comme illustré dans les figures 3.3, 3.5a et 3.4. Les noms « Pseudo-FeFET » (PsFeFET), « », « FeFET à via métallique » et « FeFET de deuxième génération » ont été proposés, ainsi que « FeFETBEoL 1T1C » ([Leh+21]). Cette structure est également parfois simplement appelée FeFET MFM ou FeMFET.

Bien que cela aille à l'encontre des premiers développements visant une intégration étroite[Mue+13a], cette configuration n'avait pas été étudiée précédemment. Celle-ci ouvre de nouvelles possibilités de conception, puisque l'aire de la capacité ferroélectrique peut être choisie très différente de



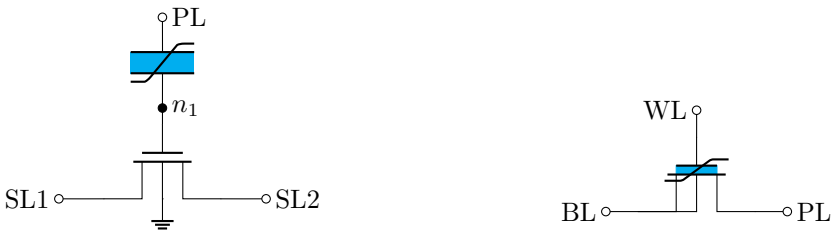
(a) Empilement conventionnel FeFET (b) FeFET ferroélectrique BEoL

FIG. 3.3 : Vue en coupe d'un FeFET conventionnel (3.3a) et d'un empilement proposé « Pseudo-FeFET » reliant un transistor n-MOS conventionnel à un condensateur ferroélectrique BEoL (3.3b).

celle de la grille du transistor. L'endurance peut également être améliorée, car celle observée pour les FeCap BEoL (environ 10^{12} cycles) dépasse de loin celle des FeFET FEoL (environ 10^6 cycles). Des structures similaires ont par la suite fait l'objet d'études dans la littérature, et été comparées aux empilements de grille MFM et de [Leh+21].

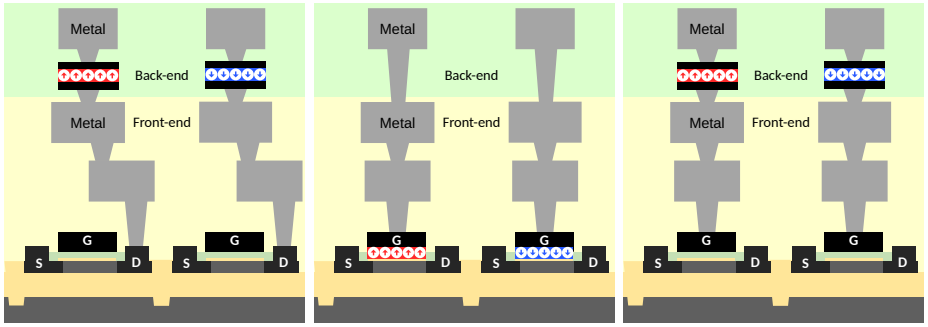
Cependant, ces avantages se font au prix de risques de piégeage de charges dans la couche intermédiaire, qui peuvent compenser les charges de polarisation ferroélectriques ou même, dans certains cas, détruire l'oxyde. Un autre problème est celui des courants de fuite plus importants du nœud flottant n_1 indiqué sur le Circuit 3.3a. Ces courants de fuite réduisent la durée pendant laquelle une valeur peut être conservée dans le FeFET, c'est-à-dire sa rétention. De plus, les capacités parasites introduites par les longues lignes métalliques doivent également être compensées. Une fois que les charges stockées dans le nœud flottant dissipées, la valeur stockée doit être lue depuis la couche ferroélectrique, ce qui est une opération destructive, et doit donc être réécrite. Cela diffère des FeFETs ordinaires qui reposent sur l'attraction des charges de compensation par le canal du transistor, et qui sont donc moins susceptibles d'être compensées dans la couche intermédiaire. Cette compensation peut aussi se produire dans un FeFET ordinaire en raison des charges piégées entre les couches ferroélectrique et diélectrique dans un empilement métal-ferroélectrique-isolant-semiconducteur. Cependant, une plus grande structure conductrice et chargée est plus susceptible d'attirer et de retenir les charges libres.

Enfin, le PsFeFET reste soumis aux mêmes inconvénients que le FeFET, tels que des tensions de programmation plus élevées que le FeCaps. Il présente également un risque plus élevé de piégeage de charges sur le nœud flottant pendant la fabrication, ce qui pourrait menacer l'intégrité de l'oxyde de la grille du transistor. Les problèmes de tension de programmation peuvent être réduits en permettant l'accès au nœud flottant (n_1 sur le Circuit 3.3a), ce que fait la structure 2T1C décrite dans la section 3.5.



(a) Schéma électrique de « Pseudo-FeFET », similaire à un FeFET (b) Symbole deFeFET pour référence

CIRCUIT 3.3 : Structure « Pseudo-FeFET » comparée à un symbole de FeFET. Les PsFeFET devraient pouvoir remplacer ces derniers dans la plupart des applications, au coût possible de caractéristiques de rétention dégradées.



(a) Intégration BEoL de condensateurs ferroélectriques (b) Intégration FEoL d'un transistor ferroélectrique (FeFET) (c) Utilisation d'un condensateur BEoL pour la réalisation d'unFeFET

FIG. 3.4 : Vue en coupe d'un PsFeFET (3.4c), comparé à un FeCap intégré en BEoL (3.4a) et à un FeFET ordinaire (3.4b)

3.3.2 Conception

Les objectifs principaux du démonstrateur de PsFeFET étaient de vérifier la faisabilité de la fabrication d'un circuit fonctionnel malgré les risques encourus par l'oxyde de grille, de montrer que le fonctionnement du dispositif était semblable à celui d'un FeFET, et d'évaluer son utilisation dans des circuits. Les objectifs secondaires comprenaient la mesure des performances de rétention et d'endurance, le courant de fuite et l'étude de la fonctionnalité logique en mémoire (LiM, Logic-in-Memory) de la variante 2T1C.

Ainsi, ce dispositif a été abordé comme une preuve de concept. Les paramètres ont été choisis pour maximiser les marges opérationnelles, malgré des incertitudes telles que la polarisation ferroélectrique rémanente, qui dépend de la proportion d'$HfZrO_2$ cristallisant en une phase orthorhombique. Des structures d'essai de différentes dimensions ont été conçues, comme listé dans la tableau 3.1, afin de compenser la variabilité possible du processus de fabrication, ainsi que les inconnues liées à la simulation.

Les itérations futures pourraient se concentrer sur l'abaissement des tensions de fonctionnement requises (en mode écriture et lecture) et sur la réduction de la taille du transistor pour le rendre plus rapide.

Égalisation de la capacité

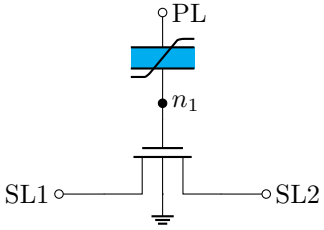
Permettre autant la programmation que la lecture du condensateur ferroélectrique nécessite un examen attentif des différentes capacités présentes dans le circuit. Les valeurs optimales des capacités présentes dans le circuit diffèrent selon l'opération effectuée, celles-ci doivent donc être considérées simultanément lors du choix des capacités relatives :

- Pour permettre la lecture d'une valeur après l'avoir programmée, les charges libérées par le FeCap doivent avoir un impact mesurable sur le potentiel du nœud flottant, ce qui suggère l'utilisation d'un condensateur ferroélectrique plus grand. Cela correspond à la maximisation de ΔV dans l'équation 3.5
- Pour permettre la programmation du FeCap, celui-ci doit être soumis à un champ électrique plus fort que le transistor, le transistor doit être plus grand. Cela correspond à la maximisation de V_{fe} dans l'équation 3.5.

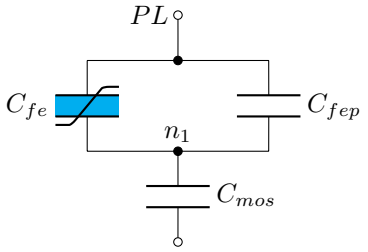
Plus généralement, il est important que le potentiel du nœud flottant ait un impact significatif sur le champ à travers le transistor lors de la lecture, et sur le même champ à travers FeCap lors de la programmation. L'égalisation des capacités est également nécessaire lors de la conception de FeFETs, bien que dans ce cas, les capacités sont ajustées via les épaisseurs d'oxyde relatives et les paramètres technologiques plutôt qu'à travers les aires relatives du condensateur et du transistor.

Pour permettre la lecture du condensateur, les charges piégées près de la surface du matériau ferroélectrique (polarisation résiduelle) pendant la programmation doivent avoir un effet suffisant sur le transistor de lecture lorsqu'elles sont libérées. Cela afin que la tension

de grille devienne supérieure ou inférieure à la tension de seuil : $\Delta V > 2 \cdot V_{th}$ (en supposant qu'il n'y ait pas d'empreinte et que l'état initial, non polarisé, soit 0 V, ce qui fait osciller la tension du nœud flottant entre $\pm^{\Delta V}/_{2}$).



(a) Schéma électrique d'un FeFET « Pseudo-FeFET »



(b) Schéma équivalent montrant uniquement les condensateurs, incluant la fraction paraélectrique du condensateur ferroélectrique

CIRCUIT 3.4 : PsFeFET (3.4a), et schéma équivalent (3.4b) montrant les condensateurs, dont la capacité MOS, et décomposant le condensateur ferroélectrique en fractions paraélectrique pures (C_{fep}) et ferroélectrique (C_{fe})

Pour égaliser ces valeurs, le diamètre du condensateur a d'abord été fixé à l'une des valeurs choisies de $D_{Cfe} = 300$, 400 ou 550 nm. La fraction paraélectrique des capacités ferroélectriques (comme illustré sur le Circuit 3.4b) a été mesurée expérimentalement lors des campagnes de fabrication précédentes comme étant comprise entre 7 fF et 30 fF pour une taille de condensateur de $2.38 \times 10^{-13}\,\mathrm{m}^2$, ce qui donne de 0,03 à $0.13\,\mathrm{F\,m^{-2}}$. Cela permet de calculer la capacité paraélectrique totale C_{fep}. Le dispositif est dimensionné pour être opérationnel dans cette plage en ajustant la taille du transistor de lecture Q_R pour garantir que la grille atteigne la tension de seuil lorsque les charges sont libérées du matériau ferroélectrique.

En considérant une polarisation résiduelle P_r de $0.19\,\mathrm{C\,m^{-2}}$, et avec une aire de condensateur $A_{Cfe} = \pi \cdot (\frac{D_{Cfe}}{2})^2$, la quantité de charges libérées est :

$$Q_{fe} = 2 \cdot P_r \cdot A_{Cfe}$$

Le facteur 2 existe, car la polarisation résiduelle bascule de P_r à $-P_r$. Les charges Q_{fe} précédemment piégées à la surface du ferroélectrique sont libérées simultanément dans la fraction paraélectrique C_{fep} du condensateur ferroélectrique, et dans la capacité de grille du transistor C_{mos}, produisant une différence de tension ΔV au nœud flottant.

Cette différence de tension étant appliquée sur la grille du transistor, celle-ci doit être suffisamment importante pour engendrer un changement mesurable de I_{DS}, par exemple en franchissant la tension de seuil. La capacité de grille C_{mos} peut être ajustée en conséquence, étant proportionnelle à l'aire transistor :

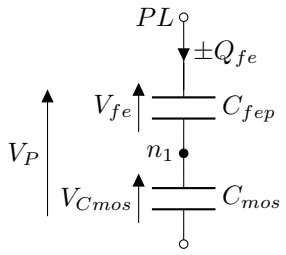
$$\Delta V = \frac{Q_{fe}}{C_{mos} + C_{fep}} \tag{3.3}$$

$$C_{mos} = \frac{Q_{fe}}{\Delta V} - C_{fep}$$

$$C_{mos} = \frac{2 \cdot P_r \cdot A_{Cfe}}{\Delta V} - C_{fep}$$

Plus la taille du transistor (donc C_{mos}) est réduite, plus la tension est élevée, donc meilleure est la fonctionnalité. Cependant, pour que la programmation soit possible, la tension du condensateur ferroélectrique V_{fe} doit dépasser la valeur coercitive V_c. Il faut pour cela que la capacité du MOSFET atteigne une certaine valeur, car lorsqu'une tension V_P est appliquée aux deux condensateurs, V_{fe} est donnée par l'expression :

$$V_{fe} = V_P \cdot \frac{C_{mos}}{C_{fep} + C_{mos}} \tag{3.4}$$



Circuit 3.5 : Circuit équivalent au PsFeFET du Circuit 3.4b, avec tensions indiquées, et en omettant le condensateur purement ferroélectrique, celui-ci servant uniquement de source pour une quantité de charges Q_{fe}, qui est la même dans les deux condensateurs en série.

Preuve, avec la notation utilisée sur le Circuit 3.5 :

$$V_{fe} = V_P - V_{Cmos} = V_P - \frac{Q_{fe}}{C_{mos}} = V_P - \frac{V_P \cdot C_{eq}}{C_{mos}}$$

$$V_{fe} = V_P - \frac{V_P}{C_{mos}} \cdot \frac{1}{\frac{1}{C_{mos}} + \frac{1}{C_{fep}}}$$

$$V_{fe} = V_P - \frac{V_P \cdot C_{fep}}{C_{mos} + C_{fep}}$$

$$V_{fe} = V_P \cdot (1 - \frac{C_{fep}}{C_{mos} + C_{fep}})$$

$$V_{fe} = V_P \cdot \frac{C_{mos}}{C_{fep} + C_{mos}}$$

Les quantités à maximiser sont à la fois ΔV et V_{fe} :

$$\begin{cases} V_{fe} = V_P \cdot \frac{C_{mos}}{C_{fep}+C_{mos}} \\ \Delta V = \frac{2 \cdot P_r \cdot A_{Cfe}}{C_{fep}+C_{mos}} \end{cases} \tag{3.5}$$

Le layout du PsFeFET tel que réalisé est montré sur le Circuit 3.6, ainsi qu'une coupe transversale et une représentation tridimensionnelle de l'empilement des couches dans les figures 3.5a et 3.5b respectivement, bien que ces représentations ne soient pas à l'échelle dans la dimension verticale, et ne montrent pas la géométrie exacte. Les paramètres géométriques correspondants sont indiqués dans tableau 3.1.

L'inconvénient de ces structures par rapport aux FeFETs ordinaires est la taille supérieure du nœud flottant, qui est une source de fuite de charges et de capacité parasite. Ceci est particulièrement vrai dans cette implémentation non optimisée : comme montré sur la figure 3.5, la couche métallique supérieure est reliée au nœud flottant, alors que le chemin pourrait être raccourci en inversant la polarité du condensateur (en reflétant verticalement le condensateur dans la layout du Circuit 3.6). La variante de diamètre 400 nm du circuit 2T1C est différente à cet égard, et la comparaison de ses performances est une future piste intéressante.

En outre, bien que l'intégration d'un circuit suiveur était initialement prévue sur certaines variantes de circuit afin de permettre l'observation directe de la tension du nœud flottant, ce fut abandonné en raison du temps nécessaire pour corriger les erreurs des règles d'antenne. C'est la raison pour laquelle une connexion au nœud flottant inutilisée est visible dans le layout du Circuit 3.6 et de la figure 3.5, ce qui détériore probablement davantage les performances.

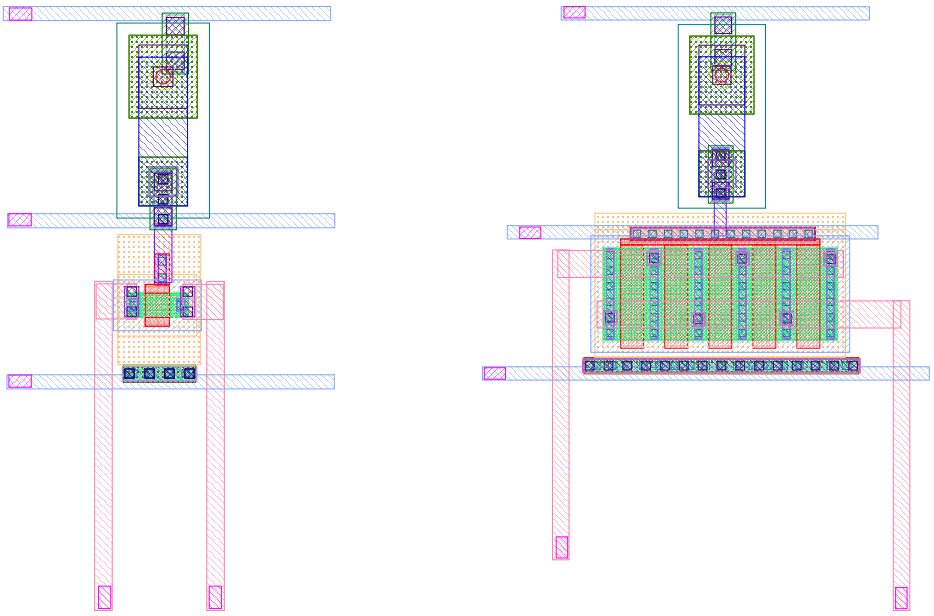
3.3.3 Caractérisation

Protocole et résultats

Seules des mesures élémentaires ont été effectuées jusqu'à présent sur le dispositif. Celles-ci ont été effectuées comme suit :

Égalisé	TW (µm)	FeCap ∅ (nm)	TL (nm)	A_{MOS}/A_{FE}	C_{MOS}/C_{FE}
Oui	10	300	500	70.7	2.2
	20	400		79.6	2.4
	40	550		84.2	2.6
Non	0.5	300		3.5	0.11
	0.5	400		2.0	0.061
	0.5	550		1.1	0.032

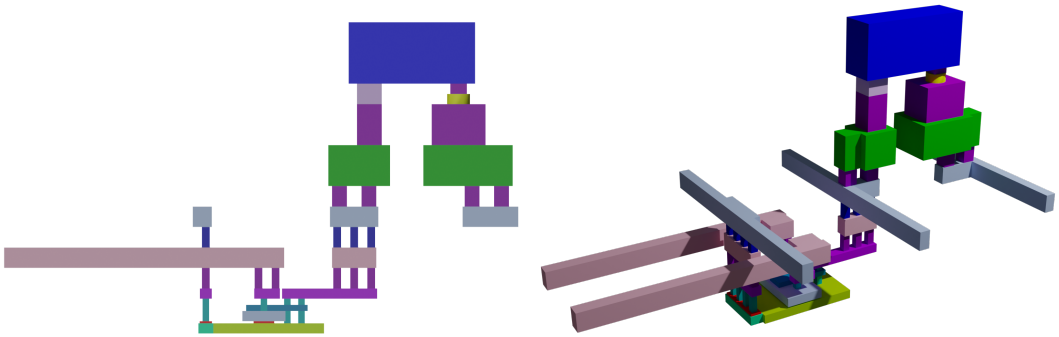
TAB. 3.1 : Dimensions choisies pour les variantes égalisées et non égalisées de la cellule PsFeFET.
Ce tableau liste la surface calculée et le ratio des capacités équivalentes du FeCap et du MOSFET vues depuis le nœud flottant, ainsi que le diamètre du FeCap cylindrique, la largeur (TW) et la longueur (TL) du transistor. Valeurs calculées pour $\varepsilon_0 \cdot \varepsilon_{r,MOS}/t_{MOS} = 3.86\,\mathrm{mF\,m^{-2}}$ et $\varepsilon_0 \cdot \varepsilon_{r,FE}/t_{FE} = 126.3\,\mathrm{mF\,m^{-2}}$, bien que la compatibilité ait été vérifiée pour $\varepsilon_0 \cdot \varepsilon_{r,FE}/t_{FE}$ =126.3 mF m^{-2} et 23.1 mF m^{-2} en cas de problème de fabrication. Les lignes vides de ce tableau ont la même valeur que la ligne supérieure la plus proche. Ce tableau est similaire à la tableau 3.4 car les mêmes dimensions ont été utilisées pour la cellule 2T1C.



(a) PsFeFET à taille de transistor minimale.

(b) Variante égalisée : le transistor égalisant la capacité selon la section 3.3.2 est beaucoup plus large (20×).

CIRCUIT 3.6 : Layout du PsFeFET pour la production MAD200, avec le condensateur à gauche et le transistor à droite. Les lignes horizontales représentent respectivement les connexions à la grille, au nœud flottant et au substrat du transistor FeFET. Les lignes verticales représentent celles de la source et du drain. Le diamètre du condensateur ferroélectrique est de 300 nm. La largeur de la grille est de 500 nm et 10 µm pour les variantes appariées et non appariées, respectivement, avec une longueur de grille de 500 nm pour les deux, comme indiqué dans la tableau 3.1.



(a) Vue de côté.

(b) Vue de trois quarts depuis le haut, avec ombres.

FIG. 3.5 : Représentation 3D de l'agencement illustré dans le Circuit 3.6a. Il s'agit ici d'une simple extrusion, et bien que l'ordre d'empilement soit exact, les bords verticaux ne représentent pas fidèlement les processus physiques de gravure et de dépôt. De plus, les hauteurs ne sont pas à l'échelle. Le $HfZrO_2$ ferroélectrique est représenté par la partie cylindrique jaune (●) au sommet de l'empilement, connectée électriquement entre les deux couches métalliques supérieures (*M4*,● et *M5*,●), avec des vias (●) utilisées comme électrodes de surface.

1. Le condensateur est d'abord effacé par l'application d'une tension négative sur la borne PL. Cela abaisse le potentiel du nœud flottant, ce qui augmente la tension de seuil du PsFeFET.

2. La grille du PsFeFET est ensuite balayée de 0.0 V à 1.8 V pour extraire la caractéristique $I_D = f(V_G)$ du transistor. L'oxyde ferroélectrique est repolarisé pendant ce balayage.

3. Le même balayage est ensuite effectué en sens inverse (de 1.8 V à 0.0 V) pour fournir une référence.

L'opération ci-dessus a été répétée pour plusieurs valeurs de tension d'impulsions négatives, de 1.0 V, 1.5 V, 2.0 V, 2.5 V et 3.0 V. Les résultats sont présentés sur la figure 3.6, validant le fonctionnement du PsFeFET.

Lors du balayage croissant (représenté par une ligne continue sur la figure 3.6), l'oxyde ferroélectrique est repolarisé, ce qui libère des charges supplémentaires dans le nœud flottant. Cela peut être considéré comme un abaissement de la tension de seuil du PsFeFET pendant le balayage, ce qui se traduit par une pente sous le seuil virtuellement plus raide.

La tension de seuil qui résulte de la tension la plus importante appliquée sur la grille du PsFeFET lors du balayage en tension croissante, peut être observée lors du balayage en tension décroissante (représentée par une ligne pointillée sur figure 3.6), car la polarisation de l'oxyde ferroélectrique n'est pas modifiée lors de celui-ci.

La fenêtre mémoire est définie comme la différence entre les tensions de seuil associées aux deux états de polarisation extrêmes (illustrée par une flèche bleue sur figure 3.6). Le graphique montre que la fenêtre mémoire augmente avec l'amplitude de la tension de Plate Line (PL) utilisée pour repolariser le dispositif, ce qui était attendu. Une fenêtre de mémoire allant jusqu'à 0.8 V a été extraite lors de cette série de mesures expérimentales.

3.3.4 Extension aux circuits à transistors multiples

Bien que cette variante n'ait pas été fabriquée, il devrait être possible d'utiliser la même structure pour modifier la tension de seuil de plusieurs transistors simultanément, lorsque les données d'entrée sont identiques.

Cela peut être particulièrement intéressant dans les circuits CMOS, où les transistors partagent généralement la même entrée pour les transistors à canal n et p. Le processus d'égalisation de capacité est le même quel que soit le nombre de transistors ; C_{mos} correspond à la capacité équivalente des multiples transistors MOS connectés à la même capacité. Ceci est illustré dans le Circuit 3.7 avec un circuit inverseur. Cette structure présente les avantages de :

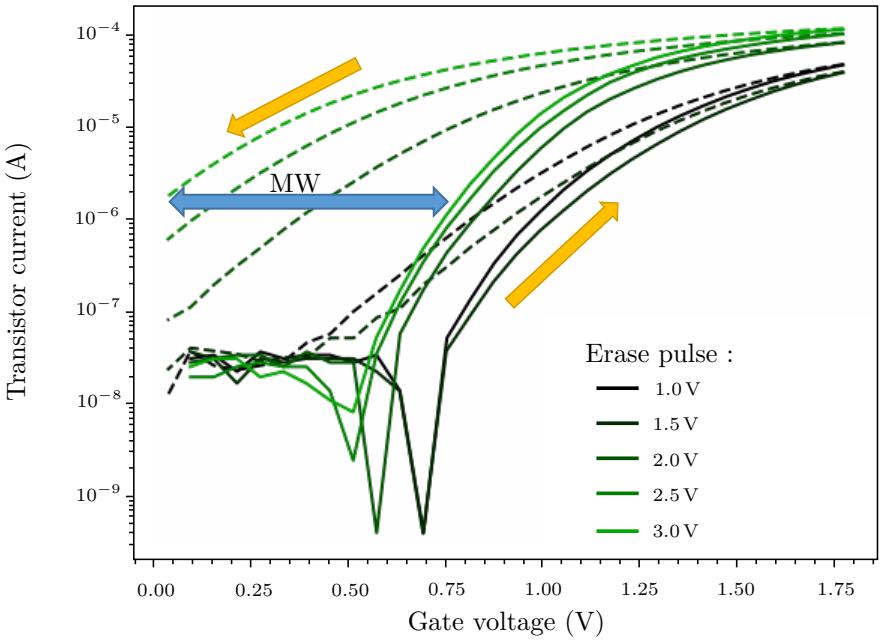
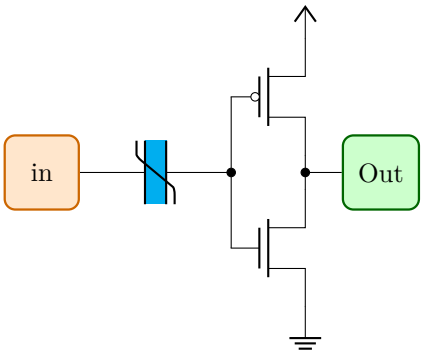


FIG. 3.6 : Caractéristique $I_D = f(V_G)$ du PsFeFET après application via PL d'impulsions de polarité négative de différentes tensions. Le balayage en tension croissante (ligne continue) et en tension décroissante (ligne pointillée) sont indiqués par des flèches orange. La fenêtre de mémoire est indiquée par une flèche bleue. Comme rapporté dans [Ni+18], V_C semble être plus bas que sur les dispositifs FEoL, proche de 2 V.

- réduire le nombre de FeCaps à programmer, ce qui simplifie le protocole de programmation ;
- réduire la surface du circuit en réduisant le nombre de FeCaps et en simplifiant le circuit de programmation ;
- augmenter davantage la densité du circuit en divisant la surface nécessaire à l'égalisation de la capacité entre plusieurs transistors.

Lors de la repolarisation, le FeCap libère des charges Q_{fe} qui affecteront tous les transistors, la tension du nœud flottant augmentant (ou diminuant) en fonction de l'équation 3.3, de manière inversement proportionnelle à la capacité équivalente. Ce ΔV déplace les tensions de seuil des p-MOS et n-MOS dans la même direction. Ce décalage peut être considéré comme une tension additionnée au signal d'entrée, ce qui permet d'augmenter ou d'abaisser la tension de celui-ci au-dessus ou au-dessous du seuil du circuit CMOS normal.

Il toujours possible de dimensionner les transistors différemment (les p-MOS étant généralement plus grand pour compenser leur conductivité plus faible), bien que l'égalisation de la capacité doive être effectuée après le choix des tailles relatives des transistors. La phase de programmation est particulièrement sensible, car une tension supérieure à V_C doit être appliquée à travers FeCap. La tension appliquée l'est à travers un diviseur de tension capacitif, comme détaillé dans sous-section 3.3.2. Toutefois, contrairement au PsFeFET, cette tension peut être appliquée à travers un plus grand nombre de dispositifs. Deux effets sont à considérer lors de la sélection des dispositifs à travers lesquels la tension de programmation est présentée : l'impact néfaste du champ électrique sur les diélectriques de grille, et l'égalisation de la capacité, car une C_{fe} comparativement plus faible facilite la programmation, d'après l'équation 3.4. Par conséquent, le choix pourrait être effectué d'épargner à certains dispositifs l'application du champ électrique plus intense, ou, ce qui est peut-être plus intéressant, un condensateur unique, séparé et plus important pourrait être utilisé pour la programmation, comme le transistor d'accès du circuit 2T1C décrit dans la section 3.5.



Circuit 3.7 : Circuit inverseur CMOS construit à partir d'un FeCap partagé, formant un inverseur à V_{th} variable. En considérant la polarisation ferroélectrique comme une entrée secondaire, et en fonction du choix des tensions de seuil, de la taille du condensateur et de la polarisation, ce circuit peut être considéré comme un NAND2 tel que décrit dans la sous-section 4.4.1 (sortie logique haut ou inverseur), un NOR2 (inverseur ou sortie niveau logique bas), ignorant le signal d'entrée (toujours haut ou bas), ou une variation de ceux-ci (y compris une modulation analogique par polarisation partielle).

3.4 TCAM à lecture destructive

3.4.1 Description

mémoire ternaire adressable par contenu

Les mémoires ordinaires stockent les informations sous forme d'un tableau, chaque valeur stockée se trouvant à une position déterminée dans le tableau. La mémoire est ensuite interrogée avec l'indice de position comme adresse afin de restituer la valeur sauvegardée.

Une mémoire adressable par contenu (CAM, Content-Addressable Memory) est une mémoire qui peut être interrogée avec une des valeurs stockées au lieu d'une adresse. Lors de l'interrogation, la position du contenu correspondant à la valeur demandée est renvoyée. Si plusieurs positions contiennent cette valeur, tous les indices correspondants peuvent être renvoyés, en fonction de l'implémentation.

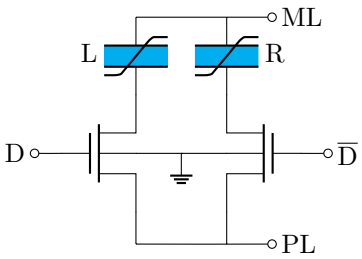
Une mémoire ternaire adressable par contenu (TCAM, Ternary Content-Addressable Memory) permet de spécifier des états de type « joker » qui correspondront à n'importe quelle valeur demandée. Ces états peuvent être spécifiés avec une granularité au niveau du bit, et seront représentés par X pour les valeurs « peu importe ». Ce troisième état possible donne son nom aux CAM « ternaires » (ou TCAM).

Ces structures sont souvent conçues pour être chaînées afin de permettre des recherches sur plusieurs bits pour des mots de différentes tailles, et leurs implémentations CMOS traditionnelles présentent un nombre de transistors relativement élevé. Elles sont principalement utilisées dans des applications réseau telles que le stockage de tables de routage, mais également de plus en plus dans des applications de réseaux de neurones.

Principe de fonctionnement

L'implémentation présentée sur le Circuit 3.8 est le résultat d'une tentative de réalisation d'une mémoire ternaire adressable par contenu conçue à partir de FeCaps. Celle-ci est constituée de deux condensateurs ferroélectriques, chacun connecté en série à deux transistors. L'objectif du circuit est d'accepter un signal d'entrée et de déterminer s'il correspond au contenu stocké dans les condensateurs.

Cette implémentation particulière fonctionne en stockant initialement une valeur complémentaire dans les deux condensateurs. Pendant la phase de recherche, une différence de potentiel supérieure à V_C est appliquée entre Match Line (ML) et PL. Une valeur complémentée est alors présentée sur les entrées D et $\overline{\text{D}}$. Cela active l'un des deux transistors, ce qui peut provoquer l'inversion de la polarisation de l'un des FeCaps, en fonction de sa polarisation initiale.



Circuit 3.8 : TCAM à lecture destructive. Il convient de noter que les entrées D et $\overline{D}$ sont complémentaires en mode de fonctionnement normal, mais qu'elles peuvent prendre la même valeur pour les opérations ternaires

L'inversion de polarisation peut être interprétée selon la convention choisie (correspondance ou non correspondance). Si plusieurs dispositifs sont reliés en parallèle, il peut être préférable de choisir l'absence d'inversion de polarisation (et donc de courant de polarisation) comme une correspondance. En effet, la détection d'une correspondance complète revient ainsi à ne détecter aucun courant, plutôt qu'à l'identification d'une valeur analogique. Dans ce cas, la séquence complète est la suivante (des tensions négatives sont utilisées pour simplifier la séquence) :

- Programmation initiale : écriture de A dans le FeCap L, et $\overline{A}$ dans le FeCap R :
- V_C est appliqué sur ML, et 0 V sur PL, puis A et $\overline{A}$ sur D et $\overline{D}$, respectivement : si A est à l'état logique haut, L est polarisé. Si A est au niveau logique bas, R est polarisé. D et $\overline{D}$ sont ensuite retournés à 0.
- La même opération est effectuée avec $-V_C$ sur la ML, et les valeurs précédemment appliquées à D et $\overline{D}$ sont inversées : le second FeCap sera ainsi polarisé dans la direction opposée.
- Pour la lecture, la première opération d'écriture est effectuée de nouveau avec une valeur B : V_C est appliqué à la ML, puis B et $\overline{B}$ sont appliqués à D et $\overline{D}$, respectivement. Une repolarisation ne se produit que si B est différent de A. Si un niveau logique bas es présenté simultanément sur D et *overline*D, aucune repolarisation ne se produira, ce qui correspond à l'état « indifférent », correspondant systématiquement.

Un état « indifférent » peut aussi bien être stocké dans la cellule TCAM qu'être fourni comme donnée d'entrée. Dans les deux cas, une correspondance sera toujours indiquée par la bitcell pour ce bit. Un tel bit est simplement indiqué en s'assurant qu'aucun des FeCap n'est repolarisé. Cela peut être effectué en stockant la même valeur dans les deux condensateurs, ou en fournissant la même valeur d'entrée sur D et $\overline{D}$, sous la tension de seuil.

Une utilisation alternative de structures semblables est celle, plus générique, de calcul en mémoire (IMC, In-Memory Computing), sous forme numérique ou analogique : la distance de Hamming entre les valeurs d'entrée et les valeurs stockées peut être mesurée, par exemple, en comptant le nombre de FeCap repolarisés. Pour ce faire, le courant de repolarisation total ou le nombre de charges transportées pendant la repolarisation est mesuré. Une pondération supplémentaire pourrait être effectuée en modulant l'amplitude de l'impulsion de programmation initiale dans chaque condensateur. Le stockage multiniveaux dans les FeCaps pourrait aussi théoriquement augmenter le nombre de bits de correspondance pour chaque FeCap.

Il est également possible d'échanger le mode de fonctionnement ternaire contre un bit de correspondance supplémentaire, en évitant l'utilisation complémentée du stockage et des accès.

Le tableau de vérité complet de la fonction TCAM présentée sur le Circuit 3.8 est donné dans la tableau 3.2. Celle-ci peut être utilisée pour implémenter la fonctionnalité TCAM, ou d'autres fonctions IMC génériques.

Limites

Cette structure étant essentiellement dérivée de deux circuits 1T1C en parallèle, celle-ci présente l'inconvénient d'effacer l'information lors de la lecture (lecture destructrice), ce qui

Valeur sauvegardée			Valeur d'entrée			Correspondance	Commentaire
L	R	Eq	D	$\overline{D}$	Eq		
0	1	H	0	1	H	1	Fonctionnement normal. Correspondance.
0	1	H	1	0	L	0	Fonctionnement normal. Pas de correspondance.
0	1	H	1	1	$\overline{X}$	0	Ne correspond jamais.
0	1	H	0	0	X	1	Opération ternaire (recherche « X »). Correspondance.
1	0	L	0	1	H	0	Fonctionnement normal. Pas de correspondance.
1	0	L	1	0	L	1	Fonctionnement normal. Correspondance.
1	0	L	1	1	$\overline{X}$	0	Ne correspond jamais.
1	0	L	0	0	X	1	Opération ternaire (recherche « X »). Correspondance.
0	0	$\overline{X}$	0	1	H	0	Ne correspond jamais, sauf pour l'entrée « X »
0	0	$\overline{X}$	1	0	L	0	
0	0	$\overline{X}$	1	1	$\overline{X}$	0	
0	0	$\overline{X}$	0	0	X	1	
1	1	X	0	1	H	1	Correspond toujours (« X » stocké) : les deux condensateurs sont déjà polarisés
1	1	X	1	0	L	1	
1	1	X	1	1	$\overline{X}$	1	
1	1	X	0	0	X	1	

TAB. 3.2 : États logiques du circuit de TCAM destructive. L'état « indifférent » (X) peut soit être fourni sous la forme d'un 0 pour les deux entrées, soit stocké sous la forme d'un 1 dans les deux condensateurs. Un état supplémentaire « ne correspond jamais » ($\overline{X}$) existe également, bien qu'il puisse être d'une utilité limitée. Dans ce tableau, une correspondance se produit (« 1 ») si aucune inversion de polarisation ne s'est produite, donc $\{\text{Left} = 0; \text{Right} = 1\}$ est équivalent (comme indiqué dans les colonnes « Eq ») à la mémorisation ou à l'entrée d'un niveau logique haut. La zone d'opération normale de la TCAM est mise en évidence.

limite fortement son utilité par rapport à une TCAMs utilisant des FeFET[Yin+19 ; Ni+19]. Deux bits d'information étant mémorisés par bit de sortie, celle-ci ne peut pas facilement être réécrite dans son état précédent, si la fonctionnalité de mémorisation de l'état « X » est utilisée.

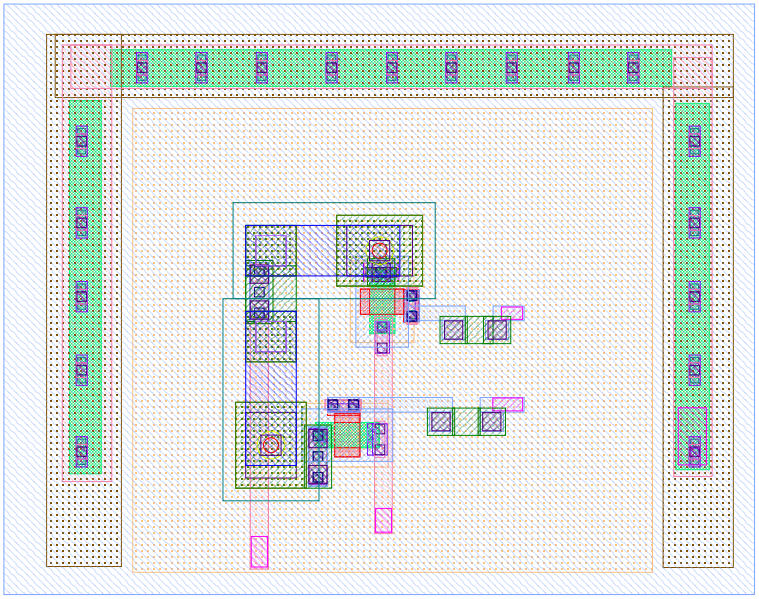
Ce circuit peut néanmoins s'avérer utile dans certaines situations, par exemple lorsque des données sont accumulées et qu'une correspondance qu'une correspondance n'est recherchée qu'une unique fois, ou dans des applications de calcul en mémoire.

3.4.2 Conception

Le Circuit 3.9 montre une structure TCAM telle que conçue pour le procédé de fabrication MAD200, avec les paramètres énumérés dans la tableau 3.3. Les transistors ont été maintenus aux dimensions physiques minimales (pour les dispositifs à oxyde de grille épais : $L = W = 500\,\text{nm}$), combinés avec plusieurs diamètres de FeCap, ainsi que des variantes reliant ensemble les MLs de deux et trois cellules TCAM, respectivement. Cela permet de faire correspondre jusqu'à trois bits en mode TCAM, ou six en mode CAM (lorsqu'un seul niveau de polarisation est utilisé par condensateur).

FeCap ∅	WL	Bits (TCAM)
300	2	1
400	2	1
550	2	1
400	4	2
400	6	3

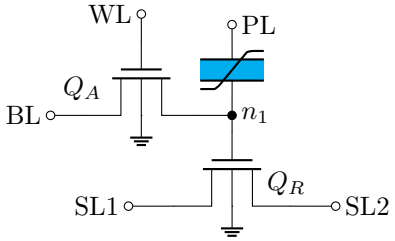
Tab. 3.3 : Diamètres de FeCap employés pour la TCAM destructrice, nombre de connecteurs ML, et bits de corresponsance disponibles dans le mode de fonctionnement TCAM. Les dimensions des transistors sont gardées constantes à $L = W = 500\,\text{nm}$. Sur les trois premières lignes figurent des TCAMs à un bit, pour lesquels les PLs sont également connectées ensemble en raison du nombre limité de connecteurs disponibles. Les deux dernières lignes sont des TCAMs multibits, réalisés en connectant les ML de deux et trois TCAMs monobit, respectivement. Celles-ci utilisent des FeCap d'un diamètre de 400 nm.



Circuit 3.9 : Layout de TCAM. Les deux grandes structures sont les condensateurs ferroélectriques, avec une ML et une PL descendant verticalement. Les transistors d'accès disposent de connexions horizontales sur la droite, franchissant un niveau de métallisation plus élevé pour respecter les règles d'antenne. Les connexions au substrat sont visibles autour de la structure.

3.5 Bitcell polyvalente 2T1C

3.5.1 Description



Circuit 3.10 : Bitcell 2T1C telle que conçue. Un transistor d'accès Q_A relie le FeCap à la BL et est contrôlé par la WL. Un second transistor, le transistor de lecture Q_R, est utilisé pour observer le potentiel du nœud flottant n_1 pendant l'opération de lecture en mesurant son courant source-drain via SL1. Enfin, la PL est connectée à la seconde électrode du FeCap.

Cette structure a été proposée pour la première fois par NaMLab dans [Sle+19b ; SP21], avec un FeFET à la source et au drain court-circuités afin d'obtenir un FeCap. Toutefois, l'utilisation d'un FeCap dédié est plus adaptée, il était donc intéressant de réaliser cette structure avec une technologie FeCap dédiée, telle que le MAD200. Cette structure est également adaptée à l'exploration des possibilités de conception offertes par une technologie de fabrication donnée. En effet, celle-ci permet plusieurs modes de fonctionnement pour un seul dispositif : fonctionnement comparable à un FeFET, mode de fonctionnement utilisant la FTJ, et mode de fonctionnement de type DRAM 1T1C : il s'agit de différents mécanismes de lecture de la valeur inscrite dans le FeCap. De plus, cette structure contourne les difficultés habituelles pouvant survenir lors de la programmation d'un FeCap dans certains dispositifs tels que les FeFET, en fournissant un accès direct aux deux bornes du condensateur.

Programmation

La programmation du 2T1C s'effectue de manière identique à celle du 1T1C, comme décrit dans la sous-section 3.2.1. Les opérations de programmation et d'effacement commencent par l'activation du transistor d'accès étiqueté Q_A sur le Circuit 3.10, ce qui est effectué en appliquant une tension suffisamment élevée à la WL. Des tensions de programmation sont ensuite appliquées au FeCap via BL et PL afin de changer sa polarisation interne. Dans le mode de fonctionnement FeFET, n_1 doit être laissé flottant en supprimant la tension appliquée à BL avant la repolarisation, comme décrit dans les sous-sections suivantes.

Comme dans le cas de la structure 1T1C, il existe plusieurs façons de contrôler les tensions de programmation, en fonction des contraintes. La programmation peut être effectuée par :

1. la tension de la BL sélectionnée ($\pm V_C$) alors que la PL est mise à la masse
2. la PL (en mettant celle-ci à $\pm V_C$), tandis que seule la BL sélectionnée est reliée au potentiel de masse
3. s'il est impossible d'utiliser des tensions négatives, la BL sélectionné peut être portée à V_C, tandis que PL est reliée à la masse, et inversement pour obtenir une polarisation opposée

Dans les trois cas, les BLs non sélectionnées sont laissés flottantes. Le transistor de lecture Q_R peut également être activé par le protocole de programmation. La protection de l'oxyde de grille de celui-ci contre les tensions élevées peut nécessiter une attention particulière si des dispositifs à oxyde fin sont utilisés.

La structure de test a été conçue pour connecter directement PL à un plot, ce qui facilite l'application de tensions négatives à cette borne (deuxième cas), alors que d'autres modes restent possibles. Ceci a été exploité lors de la caractérisation, où des impulsions positives et négatives ont été appliquées à la PL. Dans un circuit plus complexe, aucune tension négative

ne peut généralement être adoptée. Dans ces conditions, il est nécessaire d'appliquer une impulsion positive à la PL ou BL afin de modifier la polarité effective de l'impulsion perçue par le FeCap (troisième cas dans la liste ci-dessus). L'application de potentiels positifs importants sur la BL nécessite également une certaine attention : Q_A étant un n-MOS, la tension de la WL doit augmenter, ce qui peut endommager le transistor d'accès. Il est possible d'y remédier en employant un transistor p-MOS supplémentaire pour former une porte de transmission complète, au prix d'une augmentation de la surface.

Lecture FTJ – Lecture non destructrice

Après avoir préchargé le nœud flottant à l'aide du condensateur d'accès, les charges commencent à fuir à travers le FeCap par effet tunnel, comme indiqué dans la section 2.1.3. L'intensité de ce courant de fuite dépend de la distance de tunnelage, qui dépend de la longueur de la zone de compensation des charges dans les deux électrodes du FeCap. Si cette longueur de compensation est asymétrique, comme cela peut être le cas avec différents matériaux, la distance d'effet tunnel est modulée par la polarisation ferroélectrique [FS19 ; Jao+21 ; Maj22]. Cela affecte à son tour le taux de décroissance de la tension du nœud flottant, qui est directement mesurée par le condensateur de lecture Q_R, puisque la tension n_1 du nœud flottant affecte $I_{\mathrm{SL1-SL2}}$. L'égalisation de la capacité entre le transistor de lecture Q_R et le FeCap n'est pas cruciale dans ce cas, car la capacité affecte principalement le taux de décroissance, de sorte qu'elle doit seulement rester stable sur l'ensemble d'un tableau de mémoire, ce qui est plus aisément réalisé avec des condensateurs de plus grande taille.

Les dispositifs FTJ utilisent généralement des FeCaps comprenant une couche mince d'oxyde ferroélectrique, ainsi qu'une couche paraélectrique supplémentaire sur un des deux côtés (en partie pour compenser les problèmes de piégeage des charges[PLH21, p. 5]). Deux métaux différents peuvent également être utilisés pour les électrodes afin de maximiser l'asymétrie du courant d'effet tunnel. Il a été démontré que l'utilisation d'un matériau semi-conducteur pour l'une des électrodes pouvait encore accroître l'asymétrie de la réponse[GB14 ; Maj+18 ; Maj22, p. 10]. L'augmentation de la surface de Q_R permet au nœud flottant n_1 de conserver son potentiel pendant une période plus longue, ce qui permet un temps de mesure plus long, ce qui ralentit également la mémoire : la détection de l'état de polarisation peut être effectuée soit en mesurant le taux de décroissance du courant I_{DS} de Q_R, soit en détectant la conductivité après un délai prédéterminé.

Lecture en mode DRAM, ou 1T1C – Lecture destructrice

Le mode de fonctionnement DRAM est presque identique à celui de la cellule 1T1C, tel que décrit dans la section 3.2, avec l'ajout d'un transistor de lecture connecté au nœud intermédiaire. Dans ce mode de fonctionnement, une tension est préchargée sur le nœud flottant avant d'appliquer une tension différente à la borne opposée du condensateur. La différence de potentiel est supérieure à la tension coercitive, ce qui modifie la polarisation du ferroélectrique si celle-ci était différente. Cette inversion de polarisation est matérialisée par un changement de la quantité de charges dans le nœud flottant, ce qui crée une différence de tension détectée par le transistor de lecture. L'équilibrage de la capacité du nœud flottant avec la quantité de charges libérées lors de l'inversion de la polarisation du condensateur est absolument critique, car le ratio influence directement la tension détectée par le transistor de lecture. Le nombre de charges ajoutées ou soustraites lors de l'inversion de la polarisation ferroélectrique doit permettre de franchir la tension de seuil du transistor de lecture, ou au minimum de provoquer un changement mesurable de son courant de drain.

Cette procédure de *lecture* est identique à la procédure d'*écriture* du mode de fonctionnement FeFET.

Lecture en mode FeFET – Lecture non destructive

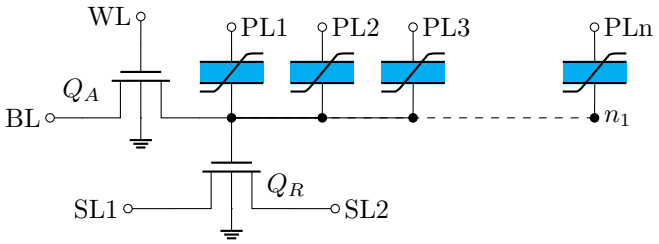
Le mode de fonctionnement FeFET du 2T1C est semblable à celui du dispositif PsFeFET de la section 3.3 ou à celui d'un FeFET normal. Le transistor d'accès permet d'améliorer le contrôle lors de la phase de programmation en permettant le préchargement du nœud flottant, fournissant ainsi un contrôle plus précis de la tension appliquée au condensateur lors de la phase de programmation. Ce contrôle direct réduit les exigences en matière de tension,

et diminue l'intensité des champs électriques appliqués à l'oxyde de la grille du transistor, autorisant l'utilisation d'oxydes plus fins et préservant les diélectriques du claquage.

Cette cellule peut également être programmée sans utiliser le transistor d'accès, à condition que l'équilibrage des capacités permette au nœud flottant d'atteindre une tension suffisamment élevée pour reprogrammer l'oxyde ferroélectrique. C'est le cas des circuits ici conçus en technologie MAD200.

Il est important de noter que pour permettre une lecture en mode FeFET, l'inversion de la polarisation de la phase de programmation doit se produire lorsque le nœud est flottant : à défaut, la tension de seuil du FeFET ne serait pas affectée par les charges libérées. Dans ce mode de fonctionnement, l'état non volatile mémorisé et lu est le potentiel du nœud flottant n_1. Ce mode de fonctionnement nécessite donc une procédure d'écriture différente.

2T-nC



Circuit 3.11 : Cellule 2TnC, une variante multicondensateurs de la cellule 2T1C présentée sur le Circuit 3.10.

Une variante 2TnC a également été étudiée, connectant plusieurs FeCaps au nœud flottant au lieu d'un seul, comme illustré sur le Circuit 3.11. Cela permet un meilleur contrôle de la capacité du ferroélectrique vis-à-vis de la grille du transistor, ce qui est essentiel pour certains modes de fonctionnement, par exemple pour effectuer les opérations d'écriture dans le mode FeFET (lecture non destructive). Cette variante permet de régler le ratio des capacités après la fabrication, en décidant d'utiliser un nombre plus ou moins élevé de condensateurs et en connectant le nombre requis de bornes PL ensemble. De plus, en programmant différentes valeurs dans les différents FeCaps, il est possible de réaliser des calculs en mémoire[Sle+19a ; Rav+19]. En effet, une mesure de distance de Hamming peut être effectuée en écrivant simultanément dans plusieurs condensateurs en mode FeFET ou DRAM. Des opérations booléennes peuvent également être réalisées avec les données stockées dans chaque FeCap, soit en mode FTJ, soit en mode DRAM.

Cette structure permet également l'étude de bitcells ferroélectriques multiniveaux sans nécessiter un contrôle précis de la durée ou tension de programmation : la programmation d'un seul condensateur équivaut en effet à la polarisation d'un sous-ensemble des domaines du FeCap équivalent. Comme détaillé dans la tableau C.1 et la tableau C.2, la sélection simultanée de plusieurs FeCaps permet d'adresser et de repolariser une fraction de la surface totale (éventuellement en plusieurs phases de programmation). Cela rend possible la programmation de plusieurs FeCap pour que ceux-ci agissent comme une unique MLC virtuelle pendant les opérations de lecture, ce qui permet un contrôle beaucoup plus déterministe de l'état stocké dans la MLC virtuelle. Cela pourrait également favoriser l'utilisation de FeCaps plus petits, tendant vers des dispositifs à domaine unique. L'utilisation de FeCaps monodomaines serait particulièrement faisable si un mécanisme est mis en œuvre pour remplacer les condensateurs défectueux, ou corriger la sortie d'un bloc de données si un condensateur serait trouvé défectueux. La fonction de transfert (ou transconductance) de la bitcell, y compris sa linéarité, peut être modulée en modifiant la surface totale des condensateurs sélectionnés. Cela pourrait avoir des applications dans le domaine de l'informatique neuromorphique[Yoo+19]. Un contrôle aussi fin nécessite cependant un circuit d'adressage supplémentaire, ce qui peut le rendre inintéressant, à moins qu'il ne soit associé à des structures peu encombrantes telles que des tableaux de type crossbar. De même, s'il est théoriquement possible d'écrire (et donc de lire) simultanément sur différents condensateurs, la complexité de contrôle supplémentaire peut rendre cette fonctionnalité inintéressante.

Enfin, il est intéressant de noter que cette structure est une variante de la cellule PsFeFET détailléee dans la section 3.3 (avec un transistor d'accès supplémentaire), où le ratio des capacités peut être réglé de manière dynamique. Cela facilite la reprogrammation du ferroélectrique (lors de la sélection d'un seul FeCap, ce qui entraîne l'application de la majeure partie de la tension à travers celui-ci), tout en étant suffisamment sensible à la polarisation du nœud flottant et au signal d'entrée (lors de l'application simultanée de celui-ci sur plusieurs FeCaps).

Émulation de la TCAM destructrice

Le circuit 2TnC peut être substitué à celui de la TCAM décrite dans la sous-section 3.4.1. Dans ce cas, ML est connectée au nœud flottant n_1 et est utilisée comme BL lorsque Q_A est passant. Contrairement au Circuit 3.8, PL n'est pas partagée, avec un transistor d'accès pour adresser les condensateurs individuels : plutôt que de générer des signaux pour D et $\overline{\text{D}}$, chaque condensateur est accessible par une PL dédiée pendant les phases de programmation et de recherche. Cette possibilité n'avait pas été envisagée avant la fabrication du circuit, et le 2TnC tel qu'il a été conçu n'est pas exposé à la capacité parasite du plot connecté à la ML, qui correspond au nœud flottant n_1 dans le circuit 2TnC.

3.5.2 Conception

Le circuit a été réalisé avec le Design Kit (DK) MAD200. L'objectif étant de valider la fonctionnalité du circuit et d'évaluer la compatibilité de chaque mode de fonctionnement avec cette technologie de fabrication.

Égalisation de la capacité pour le fonctionnement DRAM avec lecture destructrice

L'égalisation de la capacité est essentielle au fonctionnement de la cellule de mémoire en mode DRAM et, plus généralement, à l'utilisation du transistor de lecture.

La cellule est très semblable à la structure de type FeFET présentée en section 3.3, la capacité doit donc être adaptée comme détaillé dans la section 3.3.2. Il convient de mentionner que l'existence du transistor d'accès assouplit les contraintes relatives à l'égalisation de la capacité. En effet, le nœud flottant peut être préchargé à un certain potentiel, et il n'est plus nécessaire d'appliquer une tension à travers la grille du transistor.

Néanmoins, plus la variation de tension sur le nœud flottant n_1 est importante lors de l'inversion de la polarisation, plus la lecture est aisée, d'autant plus que BL est directement connectée à plot dont la capacité est élevée pour faciliter son accès. De plus, le fait d'avoir deux circuits similaires facilite les comparaisons et les diagnostics. Il a donc été décidé d'utiliser les mêmes tailles de condensateurs et de transistors que dans la section 3.3.

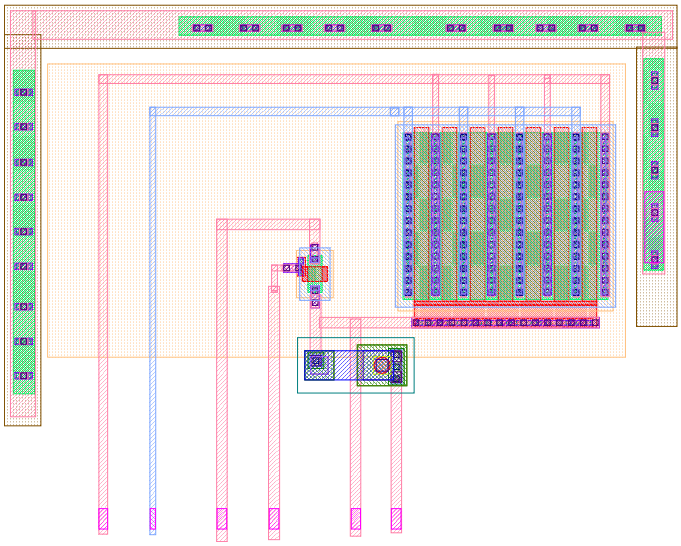
Le transistor d'accès Q_A a un impact minimal sur la capacité du nœud flottant et a donc été choisi à $W = 500\,\text{nm}$ et $L = 500\,\text{nm}$, ce qui correspond à la taille minimale d'un transistor à hauteV_{th}. En effet, celui-ci doit pouvoir supporter les opérations à haute tension requises par le matériau ferroélectrique. Ces dimensions sont la taille minimale fournie par le DK pour les transistors pouvant supporter des tensions coercitives sur leur grille. Bien certains modes d'opérations évitent d'exposer le transistor d'accès à de telles tensions, celles-ci sont nécessaires à l'opération de programmation en mode FeFET, et la longueur relativement élevée de ces transistors devrait contribuer à réduire les fuites de courant au niveau du nœud flottant.

Le Circuit 3.12 montre le layout d'une structure 2T1C telle qu'elle a été construite, pour un diamètre de condensateur de 550 nm, et une géométrie de transistor de $W = 40\,\mu\text{m} \times L = 500\,\text{nm}$, et un FeCap avec environ 40% de la capacité de grille du transistor, selon la tableau 3.4.

Le transistor d'accès Q_A peut également être utilisé pour précharger le nœud flottant, compensant un ratio de capacité défavorable par des charges supplémentaires.

2T-nC

Afin d'éviter d'effectuer l'égalisation de capacité et un nouveau vérification des règles de dessin (DRC, Design Rules Check) sur les structures 2TnC, les structures précédemment conçues ont été réutilisées en connectant leurs nœuds flottants, comme le montre le Circuit 3.13.

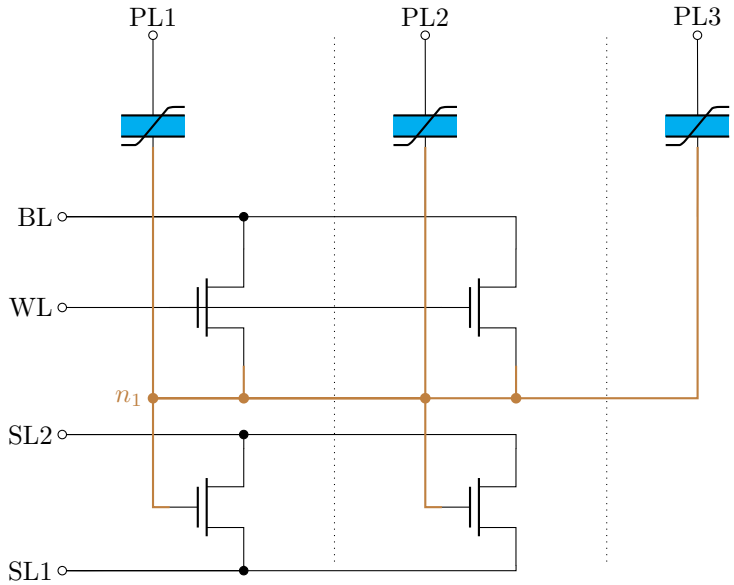


Circuit 3.12 : Layout du 2T1C. Le transistor d'accès est visible à gauche, le transistor de lecture à capacité ajustée à droite, et le FeCap en bas (cercle rouge à l'intérieur des carrés). Les connexions verticales sont, dans l'ordre : Source et drain, BL, WL, nœud flottant (utilisé pour le 2TnC), et PL. La connexion au substrat est visible autour de la structure. Le FeCap présenté ci-dessus a un diamètre de 550 nm, et est associé à un transistor de 40 µm par 500 nm, ce qui correspond à un rapport de surface de 84.2 pour un ratio de capacité de 2.6, tous deux en faveur du transistor, d'après la tableau 3.4.

Circuit	TW (µm)	FeCap ∅ (nm)	TL (nm)	A_{MOS}/A_{FE}	C_{MOS}/C_{FE}
2T1C	10	300	500	70.7	2.2
	20	400		79.6	2.4
	40	550		84.2	2.6
2T4C	70	924		52.1	1.6
		300		495.1	15.1
2T5C		934		51.1	1.6
		300		495.1	15.1

Tab. 3.4 : Dimensions choisies pour égaliser la capacité de la cellule 2T1C et des variantes 4&5T. Les variantes multi-FeCap sont représentées ici avec leurs rapports de capacité minimum et maximum disponibles, avec des diamètres équivalents correspondant aux surfaces maximum et minimum (0 exclu) de la tableau C.1 et la tableau C.2.
Ce tableau affiche la surface calculée et le ratio des capacités entre la capacité équivalente du FeCap et celle des MOSFET vues depuis le nœud flottant, ainsi que le diamètre du FeCap circulaire, et la largeur (TW) et la longueur (TL) du transistor de lecture Q_R. Valeurs calculées pour $\varepsilon_0 \cdot \varepsilon_{r,MOS}/t_{MOS} = 3.86\,\mathrm{mF\,m^{-2}}$ et $\varepsilon_0 \cdot \varepsilon_{r,FE}/t_{FE} = 126.3\,\mathrm{mF\,m^{-2}}$. La compatibilité a également été vérifiée pour $\varepsilon_0 \cdot \varepsilon_{r,FE}/t_{FE} = 126.3\,\mathrm{mF\,m^{-2}}$ et $23.1\,\mathrm{mF\,m^{-2}}$. Les lignes vides de ce tableau ont la même valeur que la ligne la plus proche au-dessus. Les premières lignes sont semblables à la tableau 3.1 car les dimensions du PsFeFET ont été réutilisées pour la cellule 2T1C.

Cette approche augmente néanmoins la capacité et les fuites de courant du nœud flottant. Le fait d'avoir plusieurs condensateurs connectés en parallèle à plusieurs plots permet d'utiliser un sous-ensemble si la capacité ferroélectrique est trop élevée par rapport à celle des transistors. Pour cette raison, un grand condensateur (d'un diamètre d'environ 550 nm) a été ajouté à chaque structure 2TnC pour permettre l'ajustement dans la direction opposée. Les plages disponibles pour l'égalisation de la capacité sont listées dans la tableau C.1 et la tableau C.2 pour les structures 2T4C et 2T5C, respectivement.



Circuit 3.13 : Exemple d'implémentation d'une structure 2T-3C, comparable aux versions fabriquées 2T4C et 2T5C. Deux structures 2T-1C sont réunies, plus un FeCap. Le nœud flottant n_1 a été coloré en marron pour faciliter la lecture.

3.5.3 Résultats de caractérisation

Une étude préliminaire des performances du dispositif a été réalisée par NaMLab, et d'autres études sont prévues en fonction de la disponibilité des équipes et du matériel. Les résultats présentés ici se focalisent sur le mode de fonctionnement 1T1C (tel que présenté dans la section 3.5.1), qui est moins sensible aux fuites de charge.

Tracé I_{DS}—V_{GS} de référence pour Q_R

Un balayage de tension est d'abord effectué sur le transistor de lecture Q_R pour tracer sa caractéristique de référence : le transistor d'accès Q_A est activé en appliquant une tension importante (plus élevée que les tensions BL ultérieures) sur WL. Cela permet un accès direct à la grille du transistor Q_R par BL. Une tension de 100 mV est appliquée entre SL1 et SL2, pendant que le courant entre SL1 et SL2 est mesuré durant le balayage de la BL de 0 V à 1.75 V.

La figure 3.7 montre la caractéristique $I_D = f(V_G)$ du transistor de lecture Q_R, après de multiples impulsions de programmation et d'effacement. Ce tracé de I_{DS}—V_{GS} a deux objectifs :

1. Permettre de déduire la tension du nœud flottant n_1 à partir du courant mesuré
2. Montrer que le comportement du transistor de lecture est indépendant de l'état du FeCap.

Protocole de caractérisation

Le protocole suivant est répété plusieurs fois afin d'étudier l'impact des différentes tensions de programmation et de valider la fonctionnalité de mémorisation du circuit :

1. Une impulsion d'effacement (e, ici choisie négative) est effectuée pour placer le transistor ferroélectrique dans un état connu.
2. Une impulsion de programmation (p, choisie positive) d'une tension donnée est appliquée.
3. La même mesure de I_D que pour le tracé de la courbe de référence est ensuite effectuée pendant l'application d'une dernière impulsion de lecture, le transistor d'accès Q_A étant désactivé.

L'évolution du courant drain-source de Q_R pendant l'impulsion de lecture est représentée sur le côté droit de la figure 3.7, jusqu'à 500 µs avant et après l'impulsion.

Une impulsion de programmation est définie par $V_{PL} - V_{n_1} > V_C$, ou de manière équivalente, *uniquement lorsque* Q_A *est activé* comme $V_{PL} - V_{BL} > V_C$. De même, une impulsion d'effacement est effectuée lorsque $V_{PL} - V_{BL} < -V_C$, V_C représentant la Tension Coercitive.

Cependant, lorsque Q_A est désactivé, outre la tension PL, la chute de tension est principalement déterminée par le pont diviseur de tension capacitif entre FeCap et la capacité de grille du transistor de lecture Q_R, d'où découlent les considérations relatives à l'adaptation de la capacité de la section 3.3.2.

Pour atteindre la tension coercitive lors de la phase de lecture, le nœud interne n_1 est préchargé en activant le transistor d'accès Q_A via la WL, en appliquant une tension adaptée sur la BL, puis en désactivant le transistor d'accès Q_A. Une impulsion est ensuite appliquée entre BL et PL.

Il en résulte un changement de tension appliquée au FeCap. Si cette tension est supérieure à la valeur coercitive alors que la valeur précédemment stockée était de polarisation opposée, celle-ci est renversée, ce qui modifie le potentiel électrique à la surface du condensateur. À son tour, cela libère ou capture davantage de charges, ce qui entraîne un changement de potentiel au niveau du nœud flottant n_1.

La tension $V_{BL} - V_{PL}$ appliquée doit être choisie pour que la différence de potentiel aux bornes du FeCap soit suffisamment importante pour le commuter ($> V_c$)

Résultats et interprétation

La figure 3.7 montre les résultats des mesures. Sur la partie gauche du tracé, la caractéristique $I_d = f(V_g)$ (I_{DS}—V_{GS}) de Q_R est montrée telle qu'elle a été caractérisée au préalable en balayant la tension V_g de grille de Q_R via la BL alors que le transistor d'accès Q_A était passant. La partie droite montre l'évolution du courant source-drain sur une fenêtre de 1 µs centrée sur l'impulsion de lecture appliquée à la PL. Les valeurs de courant, associées à la caractéristique I_{DS}—V_{GS}, permettent de déterminer la tension du nœud flottant interne n_1, comme l'illustrent les lignes pointillées rouges. Plusieurs niveaux de courant peuvent être observés sur le graphique après l'impulsion de programmation, chacun correspondant à une tension d'impulsion de programmation différente (0 V, 1 V, 2 V, 3 V et 4 V). Cela démontre une lecture correcte de l'information mémorisée, ainsi que la capacité de distinguer entre plusieurs valeurs de Pr, permettant ainsi le stockage de plusieurs bits (cellule multi-niveaux (MLC, Multi-Level Cell)).

Une légère augmentation du courant a été mesurée après chaque impulsion de lecture, y compris après la phase de programmation utilisant une tension de 0 V. Cette dernière tension de programmation ne devrait pas modifier la polarisation interne du FeCap, laissant donc celle-ci dans le même état qu'après l'impulsion d'effacement. Cette impulsion de lecture, qui a le même effet sur le FeCap qu'une impulsion d'effacement, ne devrait ainsi pas avoir d'impact mesurable sur le courant mesuré.

Cela indique un impact possible des courants de fuite à travers la grille de Q_R, ou à travers le FeCap pendant les opérations de lecture, ou l'influence d'effets de rétention à court terme. Cette question fera l'objet d'études plus approfondies.

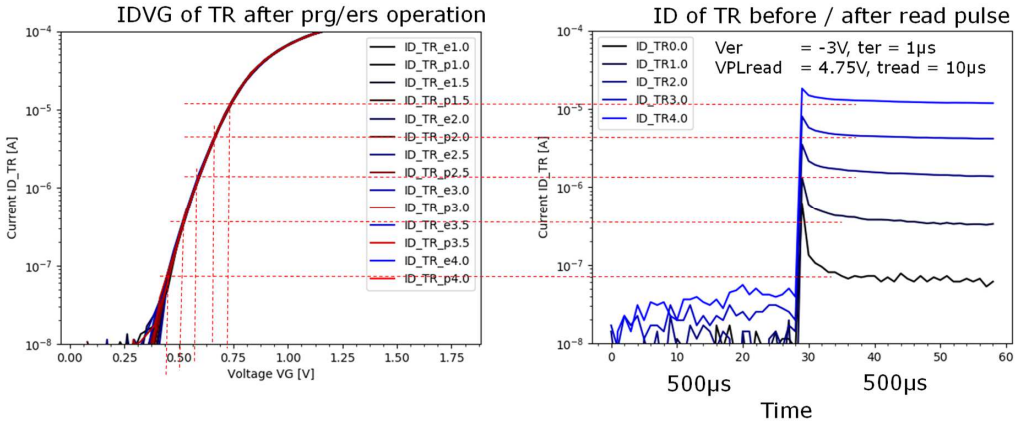


FIG. 3.7 : Caractéristique I_{DS}—V_{GS} du transistor de lecture 2T1C et courant mesuré lors de l'impulsion de lecture après application préalable d'une impulsion d'effacement de $V_{er} = -3\,V$. Le graphique de droite montre plusieurs courants de lecture distincts après application d'impulsions de programmation de 0 V, 1 V, 2 V, 3 V et 4 V, qui peuvent être attribuées à plusieurs potentiels au nœud flottant. Cela montre également qu'une mémorisation multiniveau est possible.
Les mesures ont été effectuées sur un FeCap circulaire unique de 300 nm de diamètre. La caractéristique du transistor est mesurée avec des impulsions de programmation (ID_TR_p*) et d'effacement (ID_TR_e*) de 1 µs et 1 V à 4 V ; et des impulsions de lecture de 10 µs, et 4.75 V. Le courant est mesuré sous un potentiel de 100 mV.

Enquête sur la dynamique de commutation

Afin de mieux comprendre la relation entre une repolarisation commandée en tension et une commande en largeur d'impulsion, des mesures cinétiques ont été effectuées. En effet, les deux paramètres d'impulsion (tension, largeur) ont un impact sur la repolarisation du ferroélectrique, et l'objectif de cette étude était de quantifier cela plus en détail. Les résultats sont résumés dans figure 3.8. Comme prévu, Pr dépend de la longueur de l'impulsion de programmation et de la tension, et l'augmentation de l'un ou l'autre de ces paramètres entraîne une augmentation de polarisation résiduelle. Une analyse quantitative complète de la relation temps-tension fera l'objet d'études ultérieures, y compris la comparaison de ce circuit avec un condensateur seul, afin de déterminer si la cellule 2T1C modifie la cinétique de repolarisation.

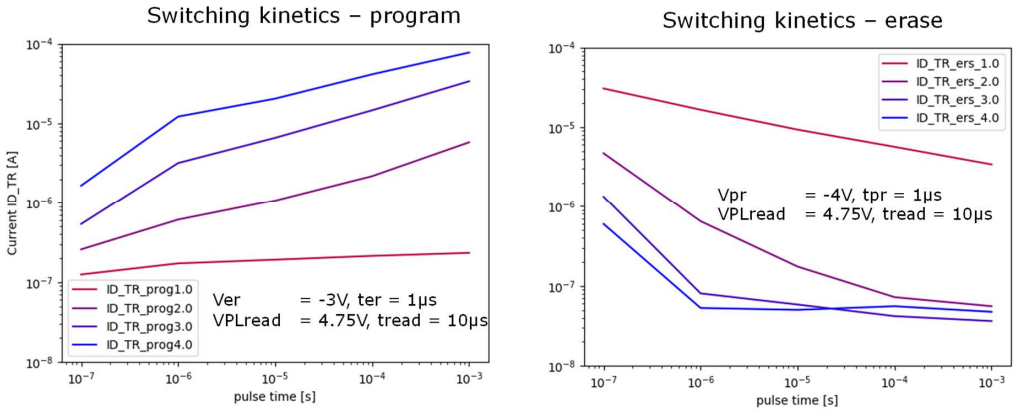


FIG. 3.8 : Courant du transistor de lecture Q_R mesuré après l'application de différentes impulsions d'effacement et de programmation, avec des largeurs (abscisse) et amplitudes (légende) d'impulsion variables, après l'utilisation d'impulsions de lecture de 10 µs et 4.75 V identiques à celles de la figure 3.7.

Afin d'éliminer la variabilité d'une cellule à l'autre, particulièrement pour le stockage de plusieurs bits dans une mémorisation multiniveaux, des algorithmes de programmation plus sophistiqués sont généralement utilisés. De multiples étapes de programmation et de vérification nécessitant des schémas d'impulsion complexes[Zho+20] sont nécessaires, variant l'amplitude et la durée des impulsions de programmation. Cela peut augmenter la durée des opérations de programmation, la consommation d'énergie et la complexité des circuits.

3.6 Conclusion

3.6.1 Cellule de mémoire 1T1C

Bien que les cellules 1T1C ne puissent être lues que de manière destructive, celles-ci promettent des caractéristiques de performance prometteuses, avec des opérations de lecture et d'écriture à faible consommation d'énergie et à faible latence. Il a été démontré que de tels tableaux de mémoire pouvaient fonctionner jusqu'à 4 ns et 100 $\mathrm{fJ\,bit^{-1}}$ [Fra+21]. Cela est dû en partie à l'accès direct aux deux électrodes du FeCap pour les opérations de lecture et de programmation, ce qui réduit les exigences en matière de tension par rapport à un FeFET. Le fait qu'aucune grille de transistor ne soit soumise à des tensions élevées pendant le fonctionnement augmente également l'endurance du dispositif, bien que l'endurance puisse être sacrifiée pour des performances plus élevées, de 10^{15} cycles à 2 V[Oku+21] à 10^{7} cycles à 4 V[Fra+21].

Ces cellules présentent également d'excellentes caractéristiques de rétention, grâce à l'utilisation directe de la polarisation du ferroélectrique comme mécanisme de mémorisation.

En raison de leur structure simple, ces circuits constituent cible intéressante pour valider les modèles ferroélectriques et les approches de simulation. Cela en fait également un objet d'étude fréquent dans la littérature scientifique, les données expérimentales associées constituant une source d'étalonnage possible pour les modèles.

Ces avantages font des tableaux RAM 1T1C $HfZrO_2$ une technologie de mémoire prometteuse, et une alternative possible à la flash.

3.6.2 Structure FeFET bout de ligne

La structure PsFeFET présente un potentiel prometteur, permettant l'utilisation de FeFETs sans modification des processus FEoL, rendant ainsi les FeFETs disponibles comme technologie complémentaire pour les processus et les nœuds technologiques existants. Des activités de caractérisation plus approfondies seront menées sur les dispositifs fabriqués, afin de mieux comprendre leurs performances vis-à-vis des FeFETs FEoL. Comme le 2T1C en mode de fonctionnement FeFET, la cellule devrait comporter des caractéristiques de rétention dégradées, perdant l'état stocké sur la grille du transistor flottant plus rapidement que les FeFETs FEoL en raison de sources de fuite de courant supplémentaires.

Néanmoins, les PsFeFETs sont plus flexibles que les FeFETs FEoL, tant du point de vue de la fabrication, car ils ne nécessitent pas de modifications importantes des processus existants, que de celui de la conception, puisque ceux-ci peuvent être dimensionnés indépendamment de la grille du transistor. En outre, en tant qu'empilement de grille MFM, ceux-ci héritent de meilleures caractéristiques de performance, telles que des tensions de programmation plus faibles, et d'une plus grande endurance grâce à la réduction des problèmes de piégeage des charges au niveau de l'interface. En effet, les charges sont plus mobiles à la surface des électrodes métalliques qu'à l'interface ferroélectrique-isolant.

Ces avantages ont également conduit à l'étude d'une structure similaire par plusieurs groupes[Ni+18 ; Leh+21], sous différents noms, comme décrit dans la section 3.3.

3.6.3 TCAM à lecture destructive

Ce circuit a originellement été conçu comme TCAM, bien que son mode de fonctionnement destructif le rende peu adapté au cas d'utilisation traditionnel des TCAM : des opérations de recherche fréquentes dans un tableau de mémoire. Un routeur effectuerait, par exemple, une telle recherche pour chaque paquet entrant, afin de trouver l'adresse de destination dans la table de routage. Ce cas d'utilisation nécessiterait de rafraîchir une partie du tableau

CAM avec des données provenant d'une mémoire externe à chaque accès : en raison de l'état ternaire, un mode d'auto-rafraîchissement (ou WB) ne peut être implémenté, car le bit de sortie ne reflète pas nécessairement le contenu stocké. Il est également impossible de reconstruire plusieurs bits à partir d'un seul bit de sortie dans le cadre de recherches multibits.

Ce cas d'usage est mieux servi par des TCAMs pouvant être lues de manière non destructive, comme des architectures construites à partir de FeFET[Yin+19 ; Ni+19].

Cette cellule peut également être utile pour le calcul, y compris celui de distances de hamming, bien que la cellule 2T1C puisse également être utilisée à cette fin.

3.6.4 2T1C

Une cellule 2T1C entièrement fonctionnelle comprenant un condensateur ferroélectrique $HfZrO_2$ intégré en BEoL a été démontrée pour la première fois. Les travaux de caractérisation ont été menés sur la variante de condensateur de diamètre 300 nm, la plus petite fabriquée dans le cadre de ces travaux. Les cellules étudiées présentent le comportement attendu, en dehors de quelques différences intéressantes pouvant être attribuées aux courants de fuite. De plus, la cellule 2T1C permet la caractérisation expérimentale des condensateurs individuels utilisés.

Des cellules 2TnC plus complexes comprenant jusqu'à quatre FeCaps ont également été réalisées, permettant des opérations simultanées de lecture et d'écriture de différents condensateurs, ce qui permet d'effectuer des opérations logiques élémentaires pour de futures preuves de concept.

Les cellules 2T1C et 2TnC sont des structures polyvalentes pouvant émuler d'autres circuits, y compris la reproduction partielle des fonctionnalités de la TCAM destructrice et du PsFeFET. Ces cellules feront ainsi l'objet d'études ultérieures et serviront probablement de preuve de concept pour de futurs circuits. Un important travail de caractérisation reste à effectuer sur ces cellules, notamment l'étude du comportement FTJ.

Bien que la cellule 2T1C soit intéressante comme objet de caractérisation et circuit polyvalent, son utilisation dans des dispositifs commerciaux est moins évidente, en raison de sa densité réduite par rapport aux cellules 1T1C, du moins pour le mode de fonctionnement DRAM (les autres modes ont l'avantage d'un accès direct aux deux pôles du condensateur pour la programmation). Cela peut être compensé par des architectures exploitant le potentiel de calcul de la cellule 2TnC. Il reste également à étudier comment la réduction de la taille des condensateurs affecte leur fiabilité et la variabilité d'un circuit à l'autre[Den+20].

La faible densité des cellules pourrait être compensée par une augmentation du nombre de niveaux, dans une certaine mesure : des condensateurs plus grands induisent des courants de lecture plus importants, plus faciles à distinguer. Cela s'accompagne également de temps de programmation plus longs, ce qui peut être un avantage pour la lecture de cellules multiniveaux, mais peut nécessiter des tensions plus élevées pour atteindre des performances acceptables. Le choix optimal est probablement spécifique à l'application, notamment en fonction du nœud technologique utilisé pour la logique co-intégrée : des circuits de contrôle plus complexes peuvent compenser une variabilité accrue. Une approche de type DSE serait ainsi appropriée pour déterminer les compromis optimaux entre la taille des condensateurs et le nombre de niveaux.

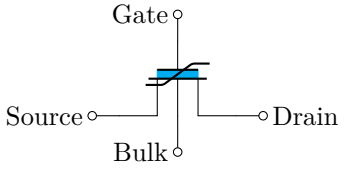
Il est également important de souligner que la perte de densité due à l'ajout d'un transistor est faible : le transistor d'accès Q_A est relativement petit par rapport au transistor de lecture. L'architecture BEoL découple aussi la taille des transistors de celle des condensateurs, contrairement aux FeFETs, comme détaillé dans le chapitre 4. Des circuits hybrides plus proches du concept 2T1C peuvent aussi être envisagés, partageant un transistor de lecture entre plusieurs cellules 1T1C.

Chapitre 4

Circuits à transistors ferroélectriques

Contents

4.1 Introduction aux circuits FeFET



Circuit 4.1 : Symbole et connexions du FeFET, tels que présentés dans la section 2.4

Dans ce chapitre, les avantages de l'utilisation de FeFETs comme mémoires à transistor unique (1T) pour les calculs normalement-éteint et LiM sont étudiés. Pour ce faire, la spécificité

des FeFETs est exploitée, plus précisément leur combinaison directe, au niveau du dispositif, de commutation logique (transistor) et du stockage d'informations non volatiles (couche ferroélectrique).

En préambule, ce chapitre commence par une discussion sur les principes de fonctionnement de circuits à FeFET, en particulier sur la manière de programmer l'oxyde ferroélectrique lorsque celui-ci est intégré dans l'empilement de grille du transistor. Sont également explorés les principes entourant l'effet physique souhaité, c'est-à-dire le déplacement de la tension de seuil du transistor par l'intermédiaire de l'état de polarisation de l'oxyde ferroélectrique.

4.1.1 Programmation de l'oxyde ferroélectrique

Pour polariser l'oxyde ferroélectrique, un champ électrique doit être appliqué à celui-ci. Un état logique haut « 1 » est enregistré lorsqu'une impulsion de haute tension ($V > V_C$, pour que $E > E_C$) est appliquée à la grille du transistor. Un état logique bas « 0 » est obtenu en inversant le sens du champ électrique appliqué.

Cela peut être difficile à réaliser, car la fourniture de tensions négatives nécessite des circuits d'alimentation et de gestion plus complexes. Une autre approche consiste à appliquer le potentiel de masse à la grille et une tension positive à la masse, à la source et au drain.

Bien que l'application d'une tension à une seule de ces trois bornes permette de programmer l'oxyde ferroélectrique dans la plupart des cas, le champ électrique est plus homogène lorsqu'il est appliqué simultanément aux trois bornes. Comme indiqué dans la sous-section 2.4.2, ceci est particulièrement vrai dans le cas d'un empilement de grille sans couche métallique entre l'oxyde ferroélectrique et l'oxyde de grille, car cette électrode métallique force la verticalité du champ électrique à travers l'oxyde ferroélectrique.

Dans ce chapitre, la polarité des impulsions de programmation est définie comme étant la même que celle du potentiel de grille, en prenant le côté opposé de l'empilement de grille du transistor comme référence de tension. Par exemple, lorsque l'on programme la couche ferroélectrique en appliquant une tension entre le substrat du transistor (V_{BLK}) et sa grille (V_G), on obtient :

$$\begin{cases} V_G - V_{BLK} > V_C^+, & \text{Impulsion Positive} \quad (4.1) \\ V_G - V_{BLK} < V_C^-, & \text{Impulsion négative} \quad (4.2) \end{cases}$$

4.1.2 Décalage du V_{th}

Le principe de fonctionnement de la plupart des circuits à FeFET est le déplacement de V_{th} : en fonction de l'état de polarisation de la couche ferroélectrique, la tension de seuil du transistor peut être augmentée ou diminuée.

Les FETs à canal N sont considérés comme non conducteurs lorsque la tension de grille appliquée est inférieure à leur tension de seuil, qui dépend de l'épaisseur de l'oxyde et de la longueur de la grille, ainsi que de paramètres technologiques.

La repolarisation de la couche ferroélectrique libère des charges dans la grille du transistor, qui est un nœud flottant. Ces charges contribuent à rendre le canal conducteur, nécessitant une tension sur la grille flottante du transistor plus faible pour atteindre tension de seuil si l'impulsion de polarisation précédemment appliquée était positive, ou une tension plus élevée si celle-ci était négative.

La modulation de V_{th} peut affecter le comportement des transistors de manière non volatile. Pour concevoir des circuits utilisant des FeFET, il est nécessaire de sélectionner la tension de seuil initiale, ainsi que la quantité de décalage désirée.

Contrôle analogique du décalage de V_{th}

Si le champ appliqué sur l'oxyde ferroélectrique de la grille peut être contrôlé assez finement, des états de polarisation intermédiaires peuvent être atteints, permettant un contrôle analogique de la nouvelle valeur, décalée, de V_{th}. La gamme des états de polarisation possibles s'étend avec le nombre de domaines ferroélectriques. En outre, un décalage plus important de V_{th} est possible si les charges libérées par l'oxyde ferroélectrique ont un impact plus élevé sur la grille du transistor. Cela est obtenu en choisissant une capacité ferroélectrique supérieure à la capacité MOS, comme indiqué dans la section 3.3.2.

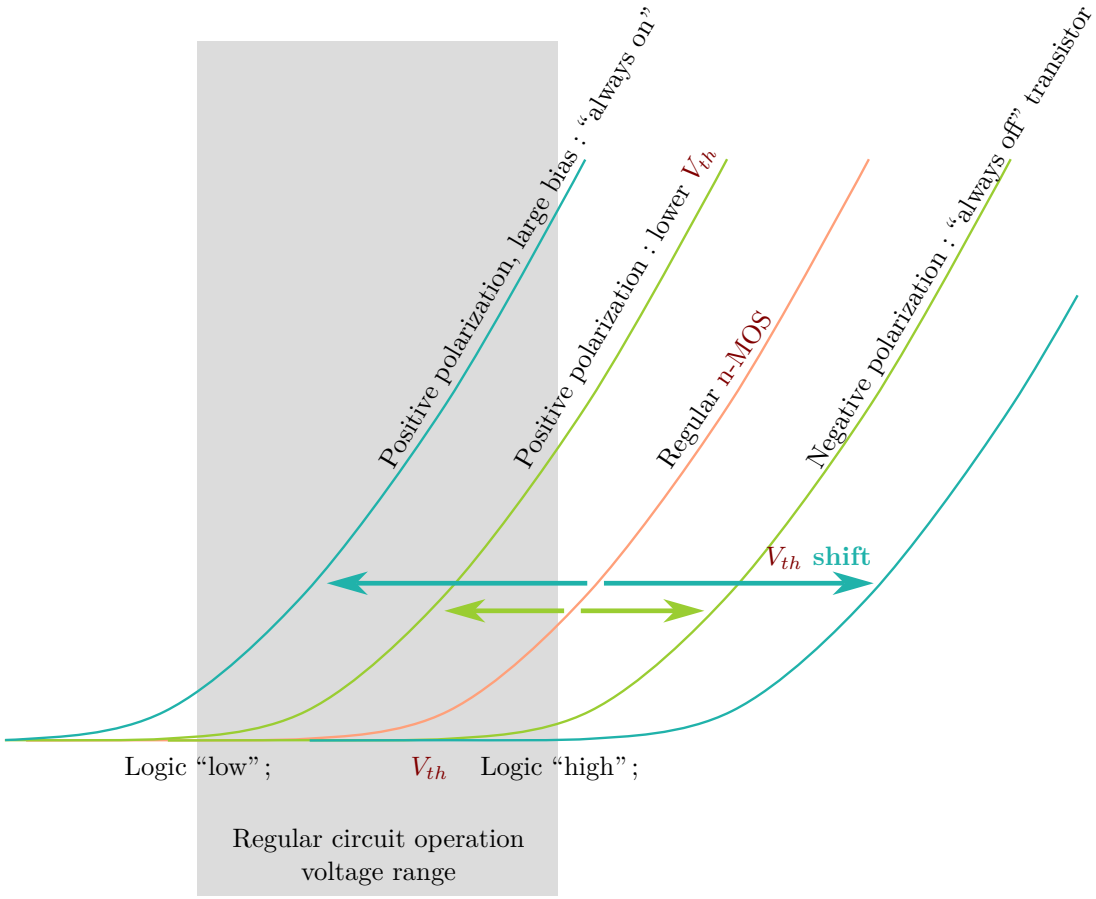


FIG. 4.1 : Illustration du décalage de V_{th} : la caractéristique I_{DS}—V_{GS} d'un n-MOS classique (🟠) est montrée à côté du décalage induit par deux couches ferroélectriques différentes, capables de libérer une quantité de charges relativement faible (🟢) ou importante (🔵), respectivement. Dans un cas concret, la valeur de la tension de seuil du transistor sera probablement plus élevée que celle représentée ci-dessus, afin d'améliorer les caractéristiques de commutation après la programmation de l'oxyde ferroélectrique. En choisissant la tension de seuil initiale et en ajustant le rapport de capacité du transistor et du condensateur ferroélectrique, le FeFET résultant peut basculer entre deux des trois modes de fonctionnement possibles visibles ci-dessus : n-MOS normal, toujours passant ou toujours ouvert.

Des états de polarisation intermédiaires peuvent être atteints en contrôlant la hauteur (tension) et la longueur (durée) de l'impulsion. Les domaines ferroélectriques étant sensibles à l'intensité du champ électrique, la tension d'impulsion a un impact direct sur leur polarisation. Cependant, les capacités relativement importantes de l'oxyde ferroélectrique et de la grille du transistor MOS limitent la rapidité de l'augmentation de la tension, ce qui les rend également sensibles à la durée de l'impulsion. La repolarisation n'étant pas instantanée, cela peut aussi être un facteur contribuant à la sensibilité temporelle.

Enfin, il est possible de repolariser un ferroélectrique en appliquant un train d'impulsions de même tension (commutation accumulative), bien que le modèle de Preisach soit incapable de modéliser ce comportement[Den+20].

Le contrôle analogique peut être intéressant pour l'informatique neuromorphique ([Yoo+19]), ainsi que pour les mémoires multiniveaux (MLC).

Décalage de V_{th} entre deux positions

Dans d'autres cas, lorsque la polarisation ne peut être contrôlée qu'à l'aide de deux tensions de programmation correspondant à deux états ferroélectriques (haut et bas), les positions souhaitées des nouvelles tensions de seuil doivent être choisies à l'avance. Comme illustré sur la figure 4.1, les possibilités pour la tension de seuil décalée V'_{th} sont les suivantes :

- $V'_{th} < L < V_{th} < H$ Toujours passant
- $L < V'_{th} < V_{th} < H$ V_{th} réduit (fonctionnement MOS normal)
- $L < V_{th} < V'_{th} < H$ V_{th} augmenté (fonctionnement MOS normal)
- $L < V_{th} < H < V'_{th}$ Toujours éteint

Dans la liste ci-dessus, H et L représentent respectivement les niveaux de tension associés à un état logique haut et bas. V_{th} représente la tension de seuil d'un transistor normal. Dans un cas idéal, la tension de seuil non décalée du FeFET (V_{th}^F) devrait être choisie indépendamment de celui-ci, et fixée au point médian des deux tensions de seuil décalées souhaitées $V_{th}^{FL} < V_{th}^{FH}$. Le décalage réel de la tension de seuil dépend de la tension coercitive utilisée pour la programmation, et peut être utilisé de manière analogique ou multiniveau[Yoo+19] ; ce cas d'utilisation ne sera toutefois pas détaillé.

Trois cas peuvent donc être considérés :

- Toujours passant / MOS normal : $V_{th}^{FL} < L$ et $L < V_{th}^{FH} < H$
- MOS normal/ toujours éteint : $L < V_{th}^{FL} < H$ et $H < V_{th}^{FH}$ et
- Toujours passant / Toujours éteint : $V_{th}^{FL} < L$ et $H < V_{th}^{FH}$, ce qui correspond au décalage de V_{th} le plus grand

Ce chapitre se concentre sur le deuxième cas d'utilisation, qui peut être considéré comme un transistor n-MOS désactivable. Ce choix a été effectué car il s'agissait du type de dispositif disponible pour fabrication. Différents circuits sont réalisables dans les autres cas. Le troisième cas, correspondant à un transistor n-MOS toujours allumé ou toujours éteint, pourrait être utilisé dans les interconnexions de circuits FPGA, en particulier car le ratio $R_{\mathrm{DS,on}}/R_{\mathrm{DS,off}}$ peut être conçu pour être relativement important dans ce cas. Les configurations réalisables dépendent de paramètres technologiques et de conception (V_{th}^F et décalage maximal de V_{th}), ainsi que de la tension appliquée au moment de la programmation. Par conséquent, les trois configurations (toujours activé, toujours désactivé, transistor normal) peuvent être disponibles sur le même dispositif.

4.1.3 Comparaison avec la logique CMOS

Avantages par rapport à la logique CMOS

La possibilité de stocker de l'information en déplaçant la tension de seuil du transistor est une alternative à la transmission d'opérandes logiques à partir de la mémoire. Cette mémorisation étant non volatile, cela réduit également le besoin de Mémoire non volatile dédiée, car cette fonctionnalité est alors intégrée au circuit. Cela peut permettre de :

1. simplifier la conception de la carte, les Mémoires non volatiles étant souvent des puces séparées
2. réduire la consommation d'énergie, car des mémoires externes supplémentaires n'ont pas besoin d'être alimentées ni d'être lues
3. augmenter les performances, car l'unité de calcul n'a pas besoin d'attendre les accès à la mémoire, ce qui réduit également davantage la consommation d'énergie

Comme détaillé dans ce chapitre, les circuits utilisant des FeFET peuvent également réduire leur nombre de transistors face à leurs homologues CMOS : par exemple, un circuit TCAM [Yin+19] peut être réduit de 16 MOSFETs par bit à 2 FeFETs par bit, plus trois MOSFET.

Inconvénients par rapport à la logique CMOS

La logique complémentaire FeFET n'était pas une option disponible pour ce travail, car la fabrication de p-FeFET n'était pas mature, comme détaillé ci-dessous. Cela a fortement limité l'efficacité énergétique statique réalisable, comme expliqué dans la section suivante. Les dimensions minimales du condensateur requises en raison de la taille des domaines est également importante (de l'ordre de 200 nm de diamètre, comme indiqué dans la section 2.1.1), ce qui produit des transistors de grande taille lorsque la capacité est égalisée. Ces transistors de grande taille, comparés aux transistors aux dimensions minimales autorisées par la technologie, sont donc plus lents et nécessitent plus d'énergie lors de la commutation. Enfin, la programmation de l'oxyde ferroélectrique nécessite des tensions relativement élevées (d'environ 3 V à 4 V), qui augmentent avec l'épaisseur de l'empilement de grille, car le Champ Électrique Coercitif du matériau ferroélectrique doit être atteint. Cela nécessite des circuits compatibles, autant pour générer que pour supporter l'application de telles tensions, généralement incompatibles avec les nœuds technologiques avancés.

Procédé technologique et disponibilité de p-FeFET

Les travaux réalisés dans ce chapitre utilisent le kit de conception standard de la technologie GlobalFoundries 28SLP avec l'ajout d'un module FeFET [Bey+20]. Ce kit a été employé pour la conception, la simulation au niveau du transistor et la fabrication. La couche ferroélectrique est simulée à l'aide de l'approche Preisach, comme décrite dans la sous-section 2.2.2.

Pour ce procédé technologique, le V_{th} du transistor est d'environ 1 V, avec un Champ Électrique Coercitif du ferroélectrique $E_C \approx 1.2\,\mathrm{MV\,cm^{-1}}$, ce qui donne des tensions de programmation entre la grille et le substrat du FeFET d'environ ± 3 V.

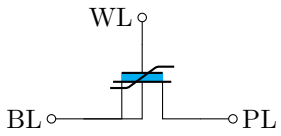
Lors des premières tentatives d'intégration d'$HfZrO_2$ ferroélectrique à l'empilement de grille, l'accent a été mis sur la réalisation de dispositifs FeFET à canal n destinés à être utilisés dans des tableaux de mémoire[Mül+21] tels que ceux décrits dans la section 4.2. Bien que des progrès aient été récemment réalisés vers la fabrication de p-FeFETs entièrement intégrés, fabriqués dans la même technologie et présentant des caractéristiques semblables à celles des n-FeFETs, ceux-ci ne constituaient pas une option disponible au moment de la conception. Cela a orienté les conceptions suivantes vers des architectures alternatives telles que la logique résistive ou dynamique, comme détaillé dans la section 4.3.

4.2 Mémoire 1T-FeFET

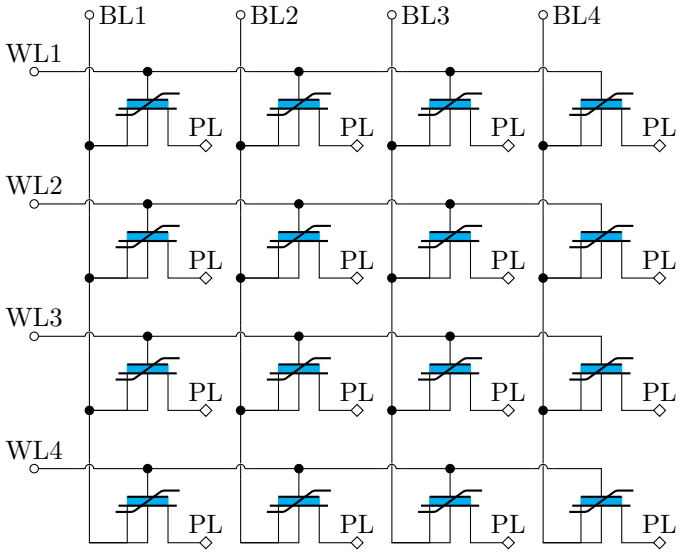
L'un des circuits les plus simples utilisant un FeFET est la mémoire 1T-FeFET, dont la bitcell est constituée d'un seul FeFET. Son principe de fonctionnement est très semblable à celui de la mémoire flash, comme détaillé dans la sous-section 4.2.2.

4.2.1 Principe de fonctionnement

Le Circuit 4.2 montre la cellule unitaire, tandis que le Circuit 4.3 montre comment celle-ci peut être intégrée dans un tableau. D'autres méthodes de programmation et de connexion sont possibles ([Rei+19]), mais ne diffèrent pas substantiellement de celle présentée dans ces schémas.



Circuit 4.2 : Bitcell 1T-FeFET montrant les connexions WL, BL et PL.
Le substrat est connecté à BL afin de faciliter la programmation.



Circuit 4.3 : Tableau de mémoire 4×4 de bitcells provenant du Circuit 4.2.

Opération de lecture

En supposant que les paramètres de fabrication aient été choisis de telle sorte que les FeFET du tableau de mémoire fonctionnent dans le mode n-MOS normal, ou toujours ouvert, leur état passant est obtenu lorsque les conditions suivantes sont toutes deux réunies :

- Le ferroélectrique est polarisé dans l'état « haut »
- Le signal WL est à l'état logique haut.

Par conséquent, comme dans le cas d'un tableau 1T1C, il est possible d'adresser les bitcells individuelles à l'intersection d'une WL et d'une BL actifves : une WL entière est activée, ce qui rend tous les FeFETs programmés à l'état logique haut de cette WL passants. Dans cette implémentation, une tension positive est appliquée à la PL, tandis que la BL, connecté au substrat, est mise à la masse.

Le transistor situé à l'intersection de la BL et WL devient passant si un état logique haut a été stocké, ce qui permet la détection d'un courant entre la PL et la BL active.

Si les BLs non sélectionnées sont mises au niveau logique bas, d'autres transistors de la même WL peuvent conduire. Bien que cela ne pose pas de problème pour détecter l'état de la cellule sélectionnée, cela entraîne une plus grande consommation d'énergie : ces lignes peuvent être mises à l'état logique haut, ou laissées flottantes.

Pour obtenir un meilleur rendement énergétique, il est possible d'effectuer une lecture en tension au lieu d'une lecture en courant. Cela nécessite de connecter le substrat séparément pour permettre des tensions de sortie plus élevées. En outre, le schéma de lecture décrit ci-dessus suppose que la capacité de la PL est supérieure à celle de la BL : Pour obtenir une vitesse de lecture optimale, il peut être nécessaire d'appliquer une tension à BL et de lire la sortie sur PL, en fonction des capacités respectives de la BL et PL.

Opération d'écriture

Pour effectuer une opération d'écriture, un champ électrique doit être appliqué à travers l'oxyde ferroélectrique. C'est-à-dire entre le contact de la WL et la source, le drain ou le substrat du FeFET, avec une préférence pour le substrat. Il s'agit de la raison pour laquelle le substrat a été connecté à la BL dans le Circuit 4.2.

La PL est commune à tous les dispositifs du tableau, et, pour cette raison, doit être laissé flottante lorsque que V_C est appliqué entre les BL et WL sélectionnées. Pour programmer un état « logique haut », le potentiel de la WL est monté jusqu'à V_C, tandis que celui de la BL est abaissé jusqu'à celui de la masse. L'inverse est effectué pour mémoriser un état « logique bas », évitant ainsi l'utilisation de tensions négatives.

Seule l'une des opérations d'écriture nécessitant la modification du potentiel de substrat (si contrôlé séparément) et de la PL de celui utilisé pour la lecture, les opérations de programmation des niveaux « logique bas » dans des cellules adjacentes devraient être groupées pour améliorer l'efficacité énergétique et la latence, de manière comparable aux opérations « effaçage par bloc » d'une mémoire flash. Pré-écrire une valeur connue dans les zones mémoires non utilisées peut également améliorer ces critères de performance, en fonctionnant de manière similaire aux opérations « TRIM » des mémoires flash.

4.2.2 Comparaison avec technologies de mémoires à transistors à grille flottante

La principale différence avec les autres technologies de mémoire FGMOS (comprenant les flash, UVPROM, EEPROM) réside dans le mécanisme utilisé pour modifier la quantité de charges électriques sur la grille flottante du transistor : les mémoires susmentionnées utilisent généralement l'effet tunnel Fowler-Nordheim ou l'injection de porteurs chauds pour stocker ou vider progressivement des charges vers et depuis la grille flottante. Les FeFETs utilisent l'inversion de polarisation comme mécanisme, échangeant des charges entre la surface du ferroélectrique et la grille métallique flottante.

Les avantages comprennent des tensions de programmation plus faibles, et la compatibilité avec les circuits CMOS dans le cas de $HfZrO_2$, éventuellement associée à une plus grande endurance, une vitesse accrue et une consommation d'énergie réduite.

Parmi les inconvénients, l'incapacité à compenser l'accumulation ou la perte de charges dans la porte flottante peut être citée. Alors que d'autres technologies peuvent complètement vider la grille flottante de ses charges, les mémoires à FeFET ne peuvent pas contrôler la quantité absolue de charges, que ce soit pour éliminer un excès de charges ou pour compenser un manque de charges. Cela peut entraîner un déplacement indésirable de V_{th} et menacer les oxydes de grille en cas d'accumulation excessive de charges. Une repolarisation régulière de l'oxyde ferroélectrique (par exemple, en modifiant la localisation ou la signification des bits) peut réduire cet effet.

Des tableaux de type NAND et NOR similaires à de la mémoire flash sont également possibles, à condition que le décalage de V_{th} soit suffisamment faible pour permettre au transistor d'être mis dans un état conducteur sans effacer la polarisation ferroélectrique.

4.2.3 Mode de fonctionnement hybride

En lumière des problèmes mentionnés ci-dessus, des approches hybrides peuvent également être explorées pour améliorer les performances. L'effet tunnel Fowler-Nordheim ou celui de la jonction ferroélectrique pourraient, par exemple, précharger la grille flottante à l'état souhaité. Une autre possibilité consiste à ajouter un transistor d'accès à la porte à transistor flottant, comme dans le cas du circuit 2T1C de la section 3.5.

Bien que cela soit impossible avec les mémoires flash traditionnelles ou d'autres mémoires de type FGMOS, car le courant de fuite supplémentaire aurait un impact sévère sur la rétention, les mémoires ferroélectriques stockent également l'état sous forme de polarisation de l'oxyde ferroélectrique, qui n'est pas menacée par les courants de fuite. Par conséquent, même si le mode de lecture normal 1T-FeFET n'est plus possible après dissipation des charges de la grille flottante, la valeur précédemment stockée peut être mesurée grâce au mode de lecture 1T1C, en effectuant une opération de lecture destructive sur l'oxyde ferroélectrique comme décrit dans la section 3.5.1 et la section 3.2.

Cela permettrait à la mémoire ferroélectrique de fonctionner rapidement, comme une DRAM, avec la possibilité de stocker des données dans la partie paraélectrique volatile ou la partie ferroélectrique non volatile du tableau de FeCap en contrôlant la tension d'écriture. Le transistor devient dans ce cas un mécanisme de lecture non destructif pour le potentiel de grille flottante, comme dans le cas des DRAM-1T[Giu21 ; LFZ11] (Floating Body 1T-DRAMs).

Les performances d'une telle mémoire hybride restent à déterminer, notamment la rétention maximale pour un fonctionnement de type DRAM, la vitesse de fonctionnement et la consommation d'énergie. La durée de rétention dicte la fréquence à laquelle les charges de la grille flottante doivent être rafraîchies, il s'agit donc du paramètre le plus crucial. Selon le cas d'utilisation, le rafraîchissement peut être géré :

- comme pour une DRAM, en compensant la fuite de charges du nœud flottant ;
- en effectuant une lecture de type 1T1C, destructrice, du ferroélectrique ;
- en laissant les charges s'évaporer complètement et en effectuant uniquement des lectures destructrices de type 1T1C, suivies d'un possible rafraîchissement.

Le mode de rafraîchissement le plus approprié peut être choisi en fonction de l'utilisation, et plusieurs combinaisons peuvent être utilisées sur un même tableau de mémoire hybride. Le fonctionnement DRAM convient mieux aux zones de mémoire nécessitant un accès à faible latence ou lues et écrites fréquemment, comme le haut de la pile d'un programme informatique. Ces zones pourraient renoncer à la non-volatilité grâce à des tensions de programmation plus faibles, ne reprogrammant ainsi pas l'oxyde ferroélectrique, économisant de l'énergie et accroissant le débit. Certaines données peuvent également être conservées de manière non volatile (checkpointing) : les valeurs stockées à l'intérieur de l'oxyde ferroélectrique et de la grille flottante peuvent être différentes, ce qui permet de stocker deux bits par cellule, dont l'un de manière non volatile à l'intérieur de la polarisation ferroélectrique.

Le rafraîchissement destructif peut être utile pour « pre-warming » une zone de mémoire pouvant devenir sensible à la latence, tandis que le mode de fonctionnement 1T1C peut être réservé aux données archivées, où la latence ou l'endurance ne sont pas critiques. Comme indiqué ci-dessus, la grille flottante de ces zones de mémoire pourrait être réutilisée comme DRAM en fonctionnant en deçà de V_C, ce qui laisserait les données persistantes non modifiées, mais augmenterait la quantité de mémoire volatile disponible sur le système.

4.3 Circuits de transrésistance

Les FETs, y compris les MOSFETs et FeFETs, sont des dispositifs transconducteurs : leur courant de sortie dépend de la tension d'entrée (transconductance). Pour cascader plusieurs circuits FETs, le courant de sortie doit être converti en une tension d'entrée pour le circuit suivant (transrésistance).

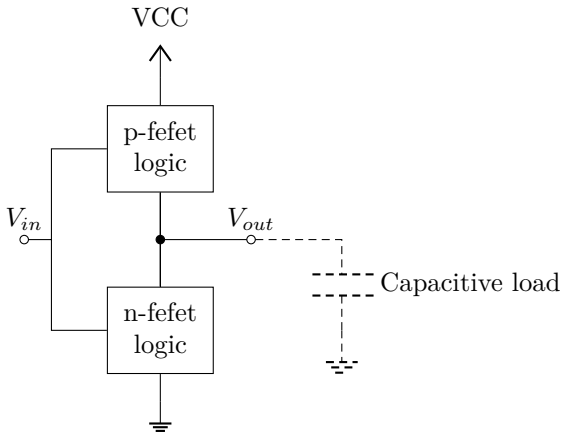
Cette section donne un bref aperçu des stratégies disponibles pour de sortir les résultats de calcul sous forme de tension logique.

4.3.1 Logique complémentaire avec p-FeFET

Bien que les p-FeFET nétaient pas une option disponible pour cette thèse, comme expliqué dans la section 4.1.3, ceux-ci sont théoriquement réalisables, et ont été récemment démontrés[Kle+21]. Cela ouvre la voie à une logique c-FeFET semblable aux CMOS, telle que représentée sur le Circuit 4.4, utilisant simultanément des n-FeFET et p-FeFET pour réduire la consommation statique d'énergie.

Pour que la logique complémentaire fonctionne correctement, les p-FeFET doivent être conçus comme dispositifs complémentaires des n-FeFET. La tableau 4.1 résume les états logiques nécessaires. Comme lisible dans ce tableau, les deux circuits doivent être activés après réception de la même impulsion de programmation.

Comme décrit dans la sous-section 4.1.2, l'application d'une telle impulsion augmente le potentiel de la grille flottante, abaissant la tension de seuil des FeFETs. L'augmentation de la tension de seuil d'un p-FeFET signifie qu'il restera dans un état conducteur quelle que soit la valeur d'entrée. Par conséquent, la même approche est compatible avec les FeFET à canal



CIRCUIT 4.4 : Circuit de transrésistance avec FeFET complémentaire. La charge capacitive nécessaire au fonctionnement est représentée en pointillés.

B	A	etat de résistance
0	0	Haut
0	1	
1	0	
1	1	Bas

(a) n-FeFET

B	A	etat de résistance
0	0	Bas
0	1	
1	0	
1	1	Haut

(b) p-FeFET

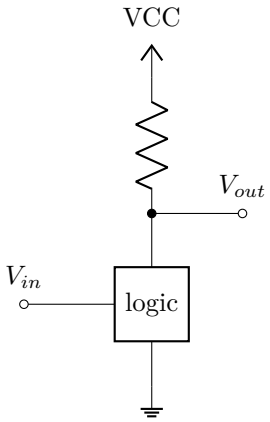
TAB. 4.1 : n-FeFET (4.1a) et p-FeFET (4.1b) avec comportement complémentaire : le comportement CMOS normal est observé lorsque l'oxyde ferroélectrique est polarisé dans l'état logique « haut » (ce qui signifie que le FeFET a reçu une impulsion de tension positive supérieure à V_C sur la grille). B, A, etat de résistance sont respectivement l'état de l'oxyde ferroélectrique, le signal reçu sur la grille du FeFET et la conductivité correspondante du canal.

n ou p : pour des transistors ayant le même V_{th} initial, l'ajout d'une couche ferroélectrique désactivera le transistor ferroélectrique après une impulsion de programmation négative. Les p-FeFET resteront dans un état conducteur, tandis que les n-FeFET resteront dans un état ouvert.

La compatibilité de ces dispositifs est due à leur plage de fonctionnement dans les circuits CMOS ; en réalité, le comportement ferroélectrique d'augmentation de la tension peut être partagé entre plusieurs transistors. Cela permet de partager un seul condensateur ferroélectrique entre les p- et n-PsFeFET, y compris pendant la phase de programmation, ce qui peut radicalement simplifier les circuits périphériques et de contrôle. Ceci est détaillé dans la sous-section 3.3.4, bien que les mêmes considérations de rétention s'appliquent.

4.3.2 Logique résistive

Une simple résistance de tirage peut être utilisée pour produire une sortie logique haute par défaut. Lorsque le circuit logique est conducteur, la sortie est rappelée vers le bas. Il s'agit de la façon la plus simple d'obtenir une tension de sortie à partir d'un changement de conductivité, au prix de courants de fuite relativement élevés. Cette architecture correspond à la famille RTL (*Resistor-Transistor Logic*).



Circuit 4.5 : Circuit FeFET avec résistance de tirage (*pull-up*) pour effectuer la transrésistance.

4.3.3 Logique dynamique

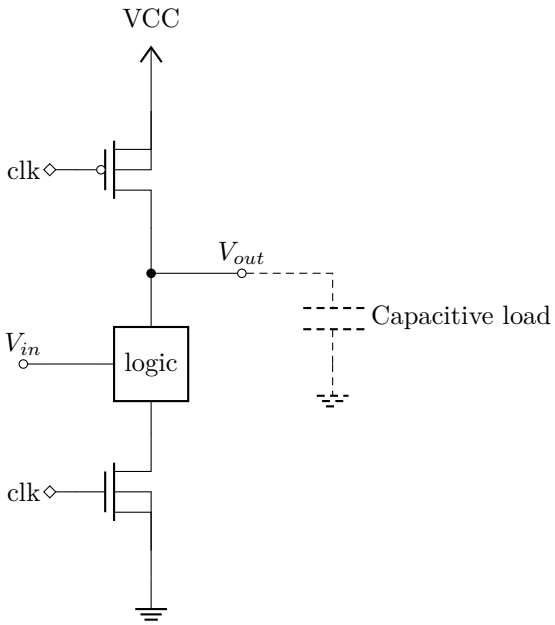
En l'absence de p-FeFET, la consommation d'énergie statique peut être réduite par le séquençage de l'alimentation, en s'appuyant sur la capacité de grille des transistors suivants afin de maintenir la valeur logique de sortie.

Comme indiqué dans le Circuit 4.6, le signal d'horloge permet de charger dans un premier temps la capacité d'entrée de l'étage suivant (phase de précharge). Ensuite, en fonction de la valeur de conductance du réseau logique, la tension de sortie peut être rappelée à une valeur basse (phase d'évaluation).

La logique dynamique est prometteuse pour l'efficacité énergétique, car celle-ci limite fortement la consommation d'énergie statique. En revanche, des problèmes de synchronisation apparaissent pour les portes logiques cascadées en aval, puisque le signal d'horloge doit se propager à la même vitesse que le signal de sortie. La synchronisation des signaux d'entrée pour mettre en cascade des portes logiques comportant des délais de propagation variables peut également conduire à choisir la période d'horloge sur la base du délai de propagation le plus défavorable, réduisant la performance.

Logique dynamique hybride avec étage CMOS

La mise en cascade de plusieurs blocs de logique dynamique est possible grâce à l'utilisation d'étages de synchronisation. Cela peut être effectué en intégrant un pipeline dans le chemin

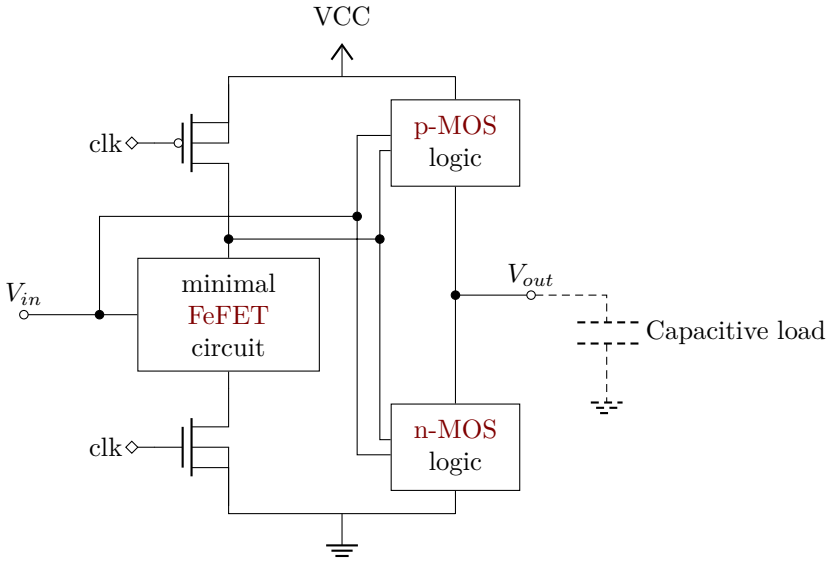


Circuit 4.6 : Transrésistance par logique dynamique.

de données, ajoutant ainsi des verrous (*latches*) pour stocker les opérandes et synchroniser les entrées.

Cependant, les étages de synchronisation augmentent la latence et ne sont pas toujours justifiés lorsque utilisés pour synchroniser de petits retards, de l'ordre d'une fraction d'un signal d'horloge.

Au lieu de cela, une approche hybride illustrée dans le Circuit 4.7 peut être adoptée, en réduisant le nombre de FeFET au minimum et en préférant l'utilisation de CMOS lorsque les avantages comparatifs des FeFETs ne sont pas exploités. Ceci est particulièrement utile lors de la conception de circuits logiques séquentiels asynchrones.



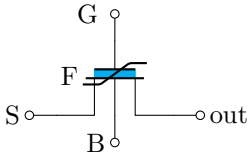
Circuit 4.7 : Architecture hybride CMOS et logique dynamique

4.3.4 Logique à transistors ballast

Transistors ballast à FeFET

Les FeFETs peuvent être utilisés pour contrôler des portes de transmission (*pass-gates*), ou directement comme transistors ballast (*pass transistors*). Comme l'illustre le Circuit 4.8, la logique à porte de transmission met en œuvre une approche radicalement différente de la transmission du signal. Cette approche évite le recours à la logique complémentaire et réduit fréquemment le nombre de transistors par rapport à d'autres techniques de réduction de la puissance statique. La logique à transistors ne reposant pas sur la charge et la décharge périodiques d'un nœud, celle-ci n'est pas sujette à des problèmes de dépendance.

Cela s'accompagne des compromis habituels relatifs à la logique à porte de transmission : des fan-out plus faibles et une meilleure transmission des niveaux logiques bas que des niveaux logiques haut lorsque les p-FET ne sont pas utilisés. Ces problèmes nécessitent généralement des techniques d'atténuation telles que des rétablisseurs de niveau logique ou des tampons de sortie après un faible nombre de portes cascadées.



(a) Schéma

B	F	G	S	out
0	1	1	0	0
			1	h
		0	-	Z
	0	-		
1	-			

(b) Table de vérité

Circuit 4.8 : Schéma électrique d'une porte de transmission utilisant un FeFET comme transistor de ballast, et tableau d'état associé. Dans la table de vérité, « - » représente les états sans importance, « Z » est une haute impédance, et « h » est un faible niveau logique haut.

Le Circuit 4.8 conduit lorsque la porte et le ferroélectrique ont un état logique haut : $\text{out} = G \cdot F$. Par souci d'exhaustivité, les potentiels de source et de substrat sont inclus dans la table de vérité, car ces signaux peuvent être combinés pour créer d'autres fonctions logiques. Ce circuit est généralement associé à d'autres circuits forçant la valeur de sortie lorsque celui-ci est en état de haute impédance « Z ».

4.4 Portes logiques non volatiles à FeFET

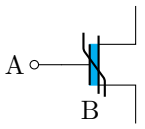
Le principe de décalage de V_{th} détaillé dans la sous-section 4.1.2 peut être exploité pour effectuer une opération logique entre la valeur d'entrée du FeFET et celle stockée dans la polarisation du ferroélectrique[Mar+21 ; Bre+18 ; OCo+18]. Par convention, une valeur logique stockée « 1 » (ou haute) correspond à la valeur programmée par l'application d'un signal de niveau logique haut, amplifié en tension, sur la grille d'entrée. Une valeur stockée « 0 » correspond à l'application d'une différence de potentiel de signe opposé. En d'autres termes, lors de la programmation d'un FeFET avec des tensions positives ou négatives, la mémorisation de « 1 » est effectuée en appliquant une tension V_C positive sur la grille du FeFET, et la mémorisation de « 0 » est effectuée en appliquant une tension V_C négative.

Cette section décrit les réseaux de rappel de multiples portes logiques : un état de conductivité élevé se traduira par une sortie logique basse. Un circuit de transrésistance, ou le circuit dual p-FeFET peut être employé dans la branche de tirage, comme indiqué dans la section 4.3 et la sous-section 4.3.1 respectivement. Ces concepts sont présentés séparément, car les p-FeFET en technologie 28SLP n'étaient pas commercialisés lors de la conception de ces circuits.

En conséquence, les tables de vérité suivantes présentent l'état de résistance (RS, Resistance State) : haute (H) ou basse (L) résistance de l'étage de rappel, ainsi que la tension de sortie (VO, Voltage Output) après transconduction de l'étage de tirage.

4.4.1 NV-NAND2

La porte logique à FeFET la plus simple est également une porte universelle : la porte NAND peut être construite à partir d'un seul transistor ferroélectrique, comme le montre le Circuit 4.9.



(a) Circuit NV-NAND2

A	B	RS	VO
1	1	L	0
0	1	H	1
1	0	H	1
0	0	H	1

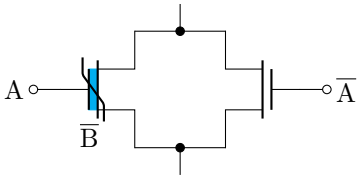
(b) Table de vérité NV-NAND2

Circuit 4.9 : Circuit de rappel de porte logique FeFET NAND. « B » est l'opérande non-volatile stockée à l'avance dans le FeFET.

Le FeFET ne conduit que si V_{th} a été déplacée dans la plage correspondant aux transistors normaux, et que l'entrée du transistor est à un niveau logique haut. Le circuit ne rappelle la sortie vers le bas que dans ce cas, ce qui correspond à une opération logique NAND, car la sortie est tirée vers le haut dans tous les autres cas.

4.4.2 NV-AND2

Le transistor de droite du Circuit 4.10 conduit si l'entrée A est au niveau logique bas, excluant tous ces cas. Le FeFET de gauche conduit si B est au niveau logique bas (stockage d'un niveau logique $\overline{\text{B}}$ haut), uniquement si A est au niveau logique haut. Cela crée une porte logique ET, puisque le seul cas non-conducteur (pour lequel la sortie restera tirée vers le haut) est celui où les deux entrées sont au niveau logique haut.



(a) Circuit NV-AND2

A	B	Left	Right	RS	VO
1	1	H	H	H	1
0	1	H	L	L	0
1	0	L	H	L	0
0	0	H	L	L	0

(b) Table de vérité NV-AND2, incluant l'etat de résistance des branches gauche et droite.

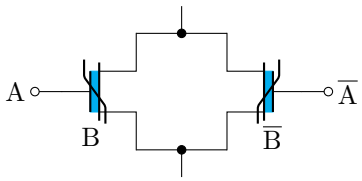
Circuit 4.10 : Circuit de rappel de la porte logique ET à FeFET. Notez la valeur complémentaire $\overline{\text{B}}$ stockée dans le transistor ferroélectrique de gauche, et l'entrée complémentée $\overline{\text{A}}$ de la branche de droite. La sortie dépend de l'etat de résistance des deux branches : si celui-ci est faible sur l'une d'elles, la résistance équivalente totale RS sera faible.

Selon le circuit contrôlant les entrées et le décalage de tension, ce circuit peut être plus ou moins pratique à réaliser. Il convient de noter la valeur inversée stockée dans le FeFET. Si un convertisseur de niveau logique est dédié au circuit, ou si la phase de programmation accepte des valeurs d'entrée différentes du cas de fonctionnement normal du circuit, le stockage d'une valeur complémentée ne devrait pas être problématique. Cependant, fournir une valeur complémentée sur le transistor de droite peut nécessiter un inverseur supplémentaire, augmentant le nombre total de transistors.

4.4.3 NV-XOR2

Le circuit XOR présenté dans le Circuit 4.11 est semblable à la porte logique ET présentée précédemment, avec l'ajout d'un second transistor ferroélectrique dans la branche parallèle. Cete seconde partie du circuit comporte deux entrées (l'entrée FeFET et la valeur stockée) inversées, et conduit donc dans le cas opposé où les deux entrées sont au niveau logique bas.

Le réseau de rappel conduit lorsque l'une des deux branches comporte un etat de résistance faible, comme illustré dans le Circuit 4.11b.



(a) Circuit NV-XOR2

A	B	Left	Right	RS	VO
1	1	L	H	L	0
0	1	H	H	H	1
1	0	H	H	H	1
0	0	H	L	L	0

(b) Table de vérité NV-XOR2, comprenant le RS des branches gauche et droite

Circuit 4.11 : Circuit de rappel XOR à FeFET. L'opération effectuée est $\overline{A \cdot B + \overline{A} \cdot \overline{B}}$, équivalente au XOR :$A \cdot \overline{B} + \overline{A} \cdot B$.

Une porte logique XOR peut donc être construite avec deux FeFET, plus un inverseur. Comme pour les autres circuits à FeFET, des multiplexeurs ou des convertisseurs de niveau logique supplémentaires sont probablement nécessaires, et l'inverseur peut avoir besoin de délivrer des tensions plus élevées pendant la phase de programmation.

4.5 FeFETs comme technologie d'appoint

Grâce à leur simplicité, les mémoires à FeFET, qui consistent en un seul transistor ferroélectrique, peuvent être utilisées pour améliorer les architectures de mémoire classiques afin d'offrir des fonctionnalités supplémentaires, telles que la non-volatilité.

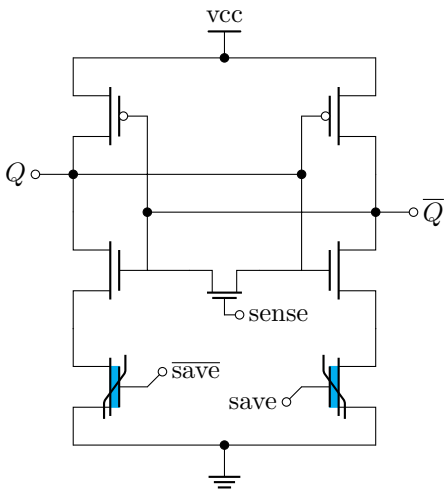
4.5.1 Cellule mémoire Black & Das comme mécanisme de checkpointing

À titre d'illustration de ce concept, la structure de type « Black & Das »[BD00] présentée dans [OCo+18] est une cellule SRAM modifiée pour stocker un bit de données supplémentaire dans le circuit d'alimentation de la cellule, construit à partir de FeFETs, comme illustré sur le Circuit 4.12a.

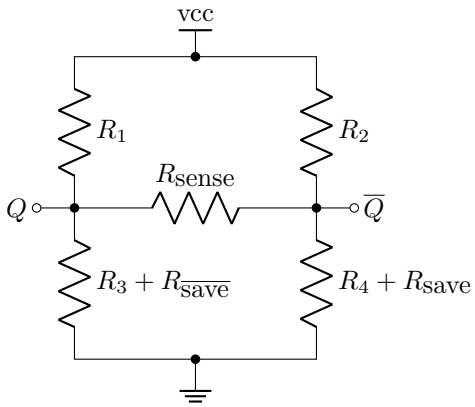
Cette cellule fonctionne en déséquilibrant la symétrie des cellules SRAM : les deux transistors n-MOS sont positionnés en série avec un FeFET pouvant stocker un état différent, ce qui entraîne des résistances équivalentes différentes, comme le montre le Circuit 4.12b. Cela affecte la capacité de rappel des inverseurs croisés, comme le montre la figure 4.2b. Le côté ayant la résistance la plus faible dans le circuit de rappel sera donc biaisé vers cet état au démarrage, lorsque $Q = \overline{Q}$. Par exemple, lorsque la branche gauche est dans un état de faible etat de résistance (correspondant à une tension de programmation positive appliquée à gauche, sur l'entrée $\overline{save}$), la sortie Q sera rappelée vers le bas et $\overline{Q}$ sera donc tirée à l'état haut.

Cette « restauration à l'allumage » peut être déclenchée manuellement sans couper l'alimentation (et ainsi devoir attendre que Q et $\overline{Q}$ se déchargent) en court-circuitant les deux sorties avec un transistor de « mesure » supplémentaire, visible au milieu du Circuit 4.12a et représenté par R_{sense} dans le Circuit 4.12b. Activer ce transistor force le circuit à entrer dans l'état métastable. Une fois le désactivé, le déséquilibre dans les branches d'alimentation des n-MOS oriente la sortie vers l'état stocké, comme le montre la figure 4.2a.

Le Circuit 4.12a peut être optimisé, conduisant à des améliorations en termes d'encombrement et de vitesse : le transistor de mesure peut être supprimé si l'opération de rappel n'est nécessaire qu'au démarrage, lorsque suffisamment de temps s'est écoulé pour permettre aux deux sorties de se décharger. Un seul FeFET est également suffisant, à condition que $R_3 + R_{\text{FeFET DS}}(0) > R_4 > R_4 + R_{\text{FeFET DS}}(1)$, avec la même nomenclature que sur le Circuit 4.12b, et $R_{\text{FeFET DS}}\binom{1}{0}$ la résistance équivalente du FeFET après l'avoir programmé avec une tension de grille de $\begin{smallmatrix}\text{HAUTE}\\ \text{BASSE}\end{smallmatrix}$. Les performances du circuit peuvent être encore améliorées si les paramètres sont choisis pour maintenir une faible résistance équivalente, car une résistance plus élevée du circuit de connexion à la masse ralentit le fonctionnement de la SRAM. Cela peut être effectué en ajustant la grille du FeFET, en la dopant davantage, en la rendant plus courte ou plus large, ou en fournissant constamment une tension positive sur la borne de la grille pendant le fonctionnement. La fuite de courant à travers le FeFET à l'état bloqué n'est pas un

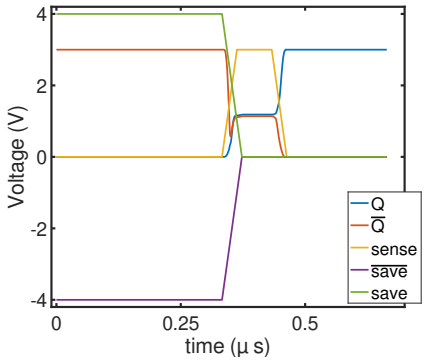


(a) Circuit Black & Das

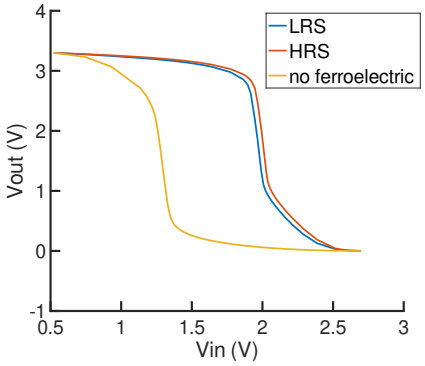


(b) Circuit équivalent pendant la restauration

CIRCUIT 4.12 : SRAM Black & Das à FeFETs, et circuit résistif équivalent durant la mesure



(a) Signaux lors de la programmation et de la mesure de la polarisation des FeFETs.



(b) Effet de la polarisation des FeFETs sur la fonction de transfert des inverseurs de la cellule SRAM, avec un inverseur sans ferroélectrique pour référence.

FIG. 4.2 : Comportement de la cellule Black & Das : Un signal de programmation positif met le FeFET dans etat de résistance bas, un signal négatif dans un etat de résistance haut. L'état programmé peut être restauré à tout moment en court-circuitant la cellule à l'aide du signal de *mesure*. Ces simulations ont été réalisées avec un modèle de Landau.

problème pour cette application, tant que la condition ci-dessus est respectée : de meilleures performances peuvent être obtenues grâce à un FeFET à canal court et fortement dopé.

Un cas d'utilisation de ce type d'architecture consiste à stocker des valeurs par défaut dans la mémoire SRAM, comme pourrait l'être un chargeur d'amorçage. Celle-ci peut également être utilisée pour le calcul normalement-éteint : en cas de coupure de courant, les valeurs contenues dans la SRAM sont effacées, mais pas celles qui sont stockées dans les FeFET. Cela peut fournir un mécanisme de checkpointing en cas de perte d'alimentation, ou si des opérations intermédiaires doivent être annulées et la SRAM restaurée à un état antérieur. Comme architecture hybride, les SRAM apportent leur vitesse et leur endurance inégalées, tandis que les FeFET assurent la non-volatilité. L'utilisation de PsFeFET peut par ailleurs contribuer à réduire l'empreinte supplémentaire des FeFET en plaçant des condensateurs ferroélectriques au-dessus des transistors de la cellule SRAM.

4.6 Filtre d'image convolutif avec logique en mémoire FeFET

Dans le cadre du projet 3εFERRO, un démonstrateur technologique a été conçu comme exemple pratique d'application de logique FeFET.

Afin de démontrer la performance potentielle des FeFETs pour des applications LiM, la réalisation d'un circuit de filtre d'image en temps réel a été ciblée. Dans cette application, le filtre d'image peut être reconfiguré pour effectuer, par exemple, une détection de contours, mais également d'autres opérations arbitraires telles que du floutage ou de l'accentuation de netteté, en fonction du cas d'utilisation.

L'effort étant dirigé par NaMLab, l'ECL-INL a contribué des blocs logiques spécifiques et a assuré la validation finale du circuit, en identifiant de multiples problèmes critiques qui auraient empêché celui-ci de fonctionner.

4.6.1 Choix d'un filtre d'image convolutif

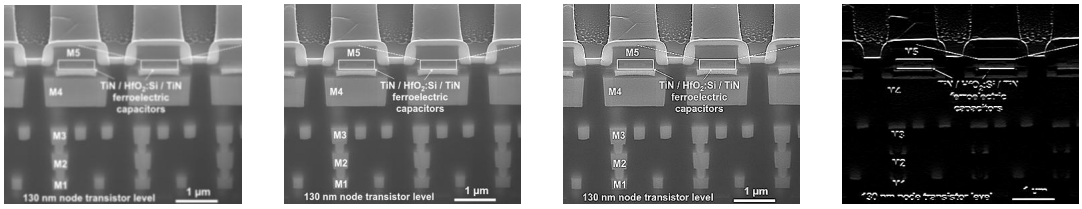
Après avoir étudié de multiples options pouvant faire l'objet d'un prototype fonctionnel pour démontrer un cas d'utilisation crédible de circuits FeFET, le choix s'est porté sur un filtre d'image convolutif. Les circuits à FeFET acceptent deux opérandes : une « statique » stockée en Mémoire non volatile sous forme de la polarisation de l'oxyde et de la grille flottante, ainsi qu'une « dynamique » fournie en entrée sur la grille du FeFET.

Cela oriente le choix du circuit vers le traitement de signal. En effet, dans ce domaine, les opérations sont généralement déterminées à l'avance et restent inchangées pendant la durée du traitement, tandis qu'un signal d'entrée dynamique est acquis (généralement par un capteur), traité et restitué. La non-volatilité des FeFET rend le normalement-éteint attrayant pour le stockage des paramètres de traitement, ceux-ci n'ayant pas besoin d'être ensuite lus à partir d'une mémoire externe pendant la séquence de démarrage.

Les processeur de signal numérique (DSP, Digital Signal Processor) sont couramment utilisés dans l'industrie pour effectuer des opérations de traitement de signaux. Parmi les algorithmes possibles, les filtres réponse impulsionnelle finie (FIR, Finite Impulse Response) sont l'un des plus courants, et nécessitent seulement de retarder le signal d'entrée, de multiplier les échantillons avec des coefficients fixes, et d'effectuer la sommer des valeurs de sortie. Ces filtres sont suffisamment polyvalents pour être utilisés dans une grande variété de contextes, du traitement audio (annulation d'écho, filtrage du bruit, égalisation) aux applications de contrôle (contrôle PID), ainsi qu'au traitement d'images.

Le choix s'est porté sur la conception d'un filtre convolutif adapté au traitement d'images, car la convolution est une opération nécessitant de nombreux accès mémoire. Une simple modification des coefficients permet de générer une large variété de résultats, comme l'illustre la figure 4.3. C'est pourquoi ceux-ci sont couramment utilisés dans le traitement d'images, ainsi que dans l'apprentissage automatique avec des réseau de neurones convolutionnels (CNNs, Convolutional Neural Network). En définissant à l'avance la topologie du filtre, la complexité du circuit est réduite, tout en démontrant l'intégration d'un algorithme non trivial. Le filtre résultant peut également être utilisé pour d'autres tâches que le traitement d'images, à condition que les données d'entrée et les coefficients puissent être adaptés à la

nouvelle application : par exemple, un tel filtre est directement utilisable comme perceptron monocouche, ce qui suggère d'autres utilisations potentielles dans les applications d'apprentissage automatique et de réseaux de neurones artificiels.



(a) Filtre de flou gaussien : $\begin{bmatrix} 0 & 2 & 3 & 2 & 0 \\ 2 & 8 & 13 & 8 & 2 \\ 3 & 13 & 21 & 13 & 3 \\ 2 & 8 & 13 & 8 & 2 \\ 0 & 2 & 3 & 2 & 0 \end{bmatrix}$

(b) Image originale, réduction de la figure 3.2b

(c) Filtre de netteté : $\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$

(d) Filtre de Sobel (détection des bords horizontaux) : $\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$

FIG. 4.3 : Exemples d'opérations de filtrage et noyaux correspondants, générés avec le code de l'extrait de code A.10.

Opération de convolution à une dimension

Comme les autres filtres FIR, l'opération de convolution discrète opère sur une série d'échantillons d'entrée et les multiplie par des coefficients, puis les additionne. Cela correspond à la catégorie de filtres FIR la plus simple, la forme directe, où aucune opération de multiplication n'est effectuée sur les résultats intermédiaires.

L'expression continue de la convolution entre deux signaux temporels $f(t)$ et $g(t)$ peut être écrite ainsi :

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) \cdot g(t - \tau)\, \mathrm{d}\tau$$

Dans le domaine temporel discret, f_{k+n} représente le$(\mathrm{k} + \mathrm{n})^{\text{ème}}$ échantillon :

$$(f * g)_k = \sum_{n=-\infty}^{\infty} f_n \cdot g_{k-n} = \sum_{n=-\infty}^{\infty} f_{k-n} \cdot g_n$$

Concrètement, seul le signal d'entrée f a une durée infinie, car il s'agit souvent d'un flux de données émis par un capteur. g est le signal de filtrage et est de longueur finie : $g_k = 0$ pour $k > l$, avec l la longueur (nombre d'échantillons) du signal de filtrage.

$$(f * g)_k = \sum_{n=1}^{l} f_{k-n} \cdot g_n = f_{k-1} \cdot g_1 + \cdots + f_{k-l} \cdot g_l$$

l échantillons du signal d'entrée f doivent donc être temporairement mémorisés, au plus (car il est courant de combler les délais entre les signaux avec des zéros), entraînant l multiplications et $l - 1$ additions.

Convolution d'une image bidimensionnelle

L'opération de convolution entre deux images est similaire, comme illustré par la figure 4.4. L'objectif d'une telle technique de filtrage étant d'appliquer une transformation de manière uniforme à une image d'entrée, la transformation appliquée est souvent une image de dimensions très réduites appelée noyau.

Le calcul est identique à celui de la convolution 1D, effectué indépendamment sur plusieurs lignes de l'image avec la ligne de filtre correspondante du noyau. Les résultats sont ensuite additionnés, donnant une nouvelle valeur de pixel dépendant de celle des pixels voisins. Il convient toutefois de souligner les différences liées au stockage en mémoire et à la transmission

des images. Traditionnellement, en raison de l'héritage des moniteurs à tube cathodique, les valeurs individuelles des pixels (luminosité) sont envoyées ligne par ligne, du haut à gauche au bas à droite de l'image. Pour une image de W pixels de largeur, W pixels doivent donc être transmis entre un pixel donné et son voisin inférieur immédiat. Un filtre d'image 2D correspond ainsi à un filtre convolutif 1D avec mise à zéro de ces W pixels intermédiaires. Bien que les calculs comportant ces coefficients ne soient pas nécessaires, les valeurs des pixels d'entrée précédemment reçus doivent soit être stockées temporairement, soit retransmises ultérieurement.

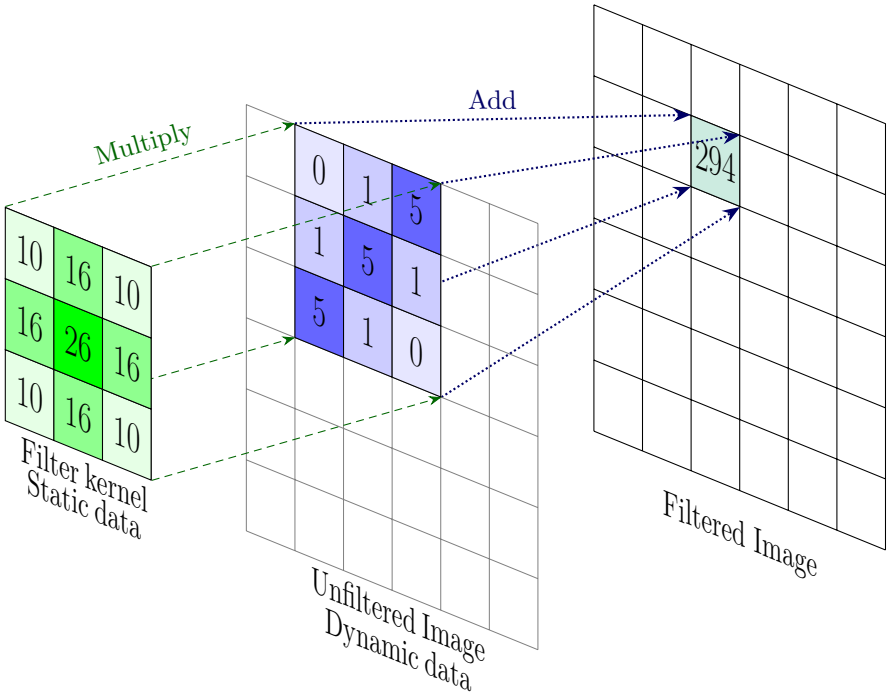


FIG. 4.4 : Représentation schématique de l'application du filtre à FeFET. Le noyau du filtre est stocké de manière non volatile dans le FeFET. Une image externe est fournie au circuit, donnant une image filtrée. Dans cette illustration, les valeurs du noyau de filtre correspondent à un filtre de flou gaussien, et la saturation des couleurs dépend de l'intensité lumineuse du pixel.

Post-traitement nécessaire

Une seconde différence entre la convolution d'images 2D et le traitement de signaux 1D est l'existence possible de canaux de couleur : les canaux rouge (R), vert (V) et bleu (B) des images RVB classiques sont généralement calculés individuellement comme trois images distinctes avec un noyaux pouvant être identique ou différent. Un filtre d'image convolutif permet également d'obtenir des résultats plus complexes en mélangeant les composantes de couleur.

Pour une convolution 1D de M échantillons par un second vecteur de taille N, le vecteur résultant est de taille $M+N-1$, bien que le résultat soit souvent considéré comme inintéressant lorsque le noyau couvre partiellement le signal d'entrée. En éliminant ces valeurs, la sortie est de dimension $M-N+1$. C'est également le cas pour le filtrage 2D, selon des deux dimensions de l'image. Les données d'entrée peuvent également nécessiter des échantillons de remplissage, ou être réordonnées si le filtre matériel n'est pas informé des tailles d'image, ou possède des capacités de stockage insuffisantes.

Enfin, la sortie de la convolution doit être normalisée pour qu'elle se situe dans le même intervalle que les valeurs de pixel de l'image d'entrée : chaque pixel de sortie contient les contributions de $K_w \cdot K_h$ pixels, avec K_w et K_h représentant respectivement la largeur et la

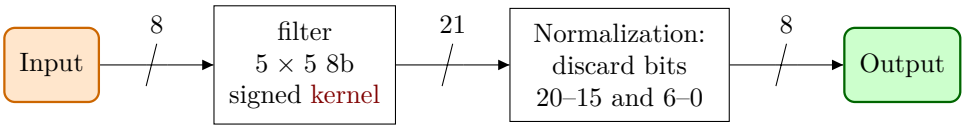


FIG. 4.5 : Schéma haut niveau du filtre d'image proposé.

hauteur du noyau (généralement carré) du filtre. Pour une image dont les valeurs des pixels se situent dans l'intervalle $[0; I]$ avec un noyau dans l'intervalle $[0; K_r]$, l'intervalle de l'image de sortie est $[0; K_w \cdot K_h \cdot K_r \cdot I]$. Par exemple, pour une image et un noyau 5×5 8-bit non signé [1] dans l'intervalle $[0; 255]$, la plage de sortie est $[0; 1625625]$, nécessitant 21 bits non signés pour conserver la précision.

Par conséquent, la renormalisation est souvent intégrée aux coefficients du noyau lors de calculs en virgule flottante, ou effectuée avant le stockage et la transmission de l'image, parfois après l'application d'étapes de traitement supplémentaires telles que des transformées gamma.

4.6.2 Architecture du filtre

L'objectif du démonstrateur est d'appliquer unnoyau de convolution à chaque image d'un flux vidéo d'entrée, en temps réel. Cela impose des contraintes à l'architecture du filtre, en termes de chemin de données (car les données d'entrée doivent être transmises en continu) et de temps de traitement (la durée des opérations affectant directement la fréquence d'images).

La figure 4.5 fournit une vue haut niveau de l'architecture du filtre.

Échantillons intermédiaires

Avec une image filtrée de largeur W et un noyau carré de taille K, l'opération de filtrage traite K lignes simultanément, avec $(W - K) \cdot (K - 1)$ pixels intermédiaires devant être soit stockés temporairement, soit retransmis ultérieurement. Bien que le stockage des échantillons intermédiaires ait été envisagé dans un premier temps, il a été décidé de simplifier la conception en renonçant au stockage de ces $(W-K)\cdot(K-1) \approx W \cdot K$ valeurs de pixels intermédiaires[2]. Ce choix permet de réduire la complexité et l'aire du circuit au prix d'un besoin supplémentaire de bande passante, chaque pixel d'entrée devant être retransmis K fois pour le filtrage de chaque image.

Le Circuit 4.13 montre le chemin complet des données pour les pixels filtrés. Un pixel de chaque ligne de l'image d'entrée est fourni séquentiellement et stocké dans le premier étage de Bascule D D. Ensuite, la deuxième colonne de Bascule D D ($K_w = 5$ par ligne d'image) est activée, faisant avancer le noyau latéralement sur l'image. Les multiplicateurs M peuvent ensuite être activés simultanément, car leur entrée est à jour.

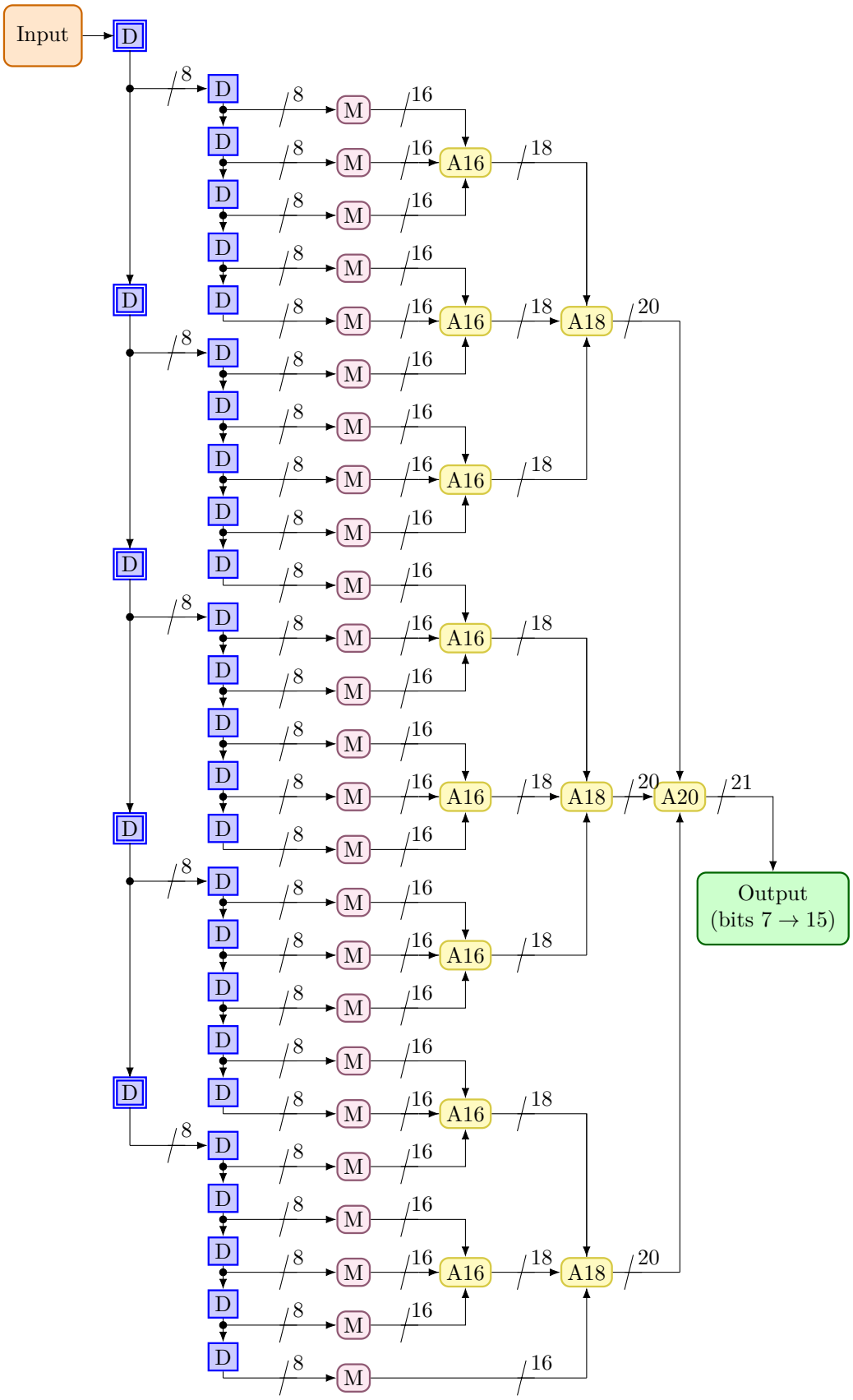
L'horloge des cellules d'entrée fonctionne donc à une fréquence cinq fois supérieure à celle du multiplicateur, ce qui confirme le facteur de surcharge présenté ci-dessus : chaque pixel individuel est transmis cinq fois lors du filtrage d'une image, le circuit n'ayant aucune mémoire de la ligne précédemment transmise. Cette situation est acceptable pour un démonstrateur, mais peut également constituer un compromis viable en termes d'efficacité, en fonction de la consommation d'énergie de la mémoire. Concernant la vitesse de traitement, celle du multiplicateur est ici le facteur limitant, comme détaillé dans section 4.6.3.

Chaîne de test

Les registres à décalage conçus pour fournir les données d'image d'entrée aux multiplicateurs ont été modifiés pour inclure un chemin de données alternatif de « débogage », où les registres à décalage sont tous connectés en série. Des registres à décalage supplémentaires ont également

[1] Certains noyaux utilisent des coefficients négatifs, tels que l'opérateur Sobel. La plage de sortie devient $[-816000; 809625]$, nécessitant également 21 bits ; il est habituellement choisi de préserver uniquement la magnitude

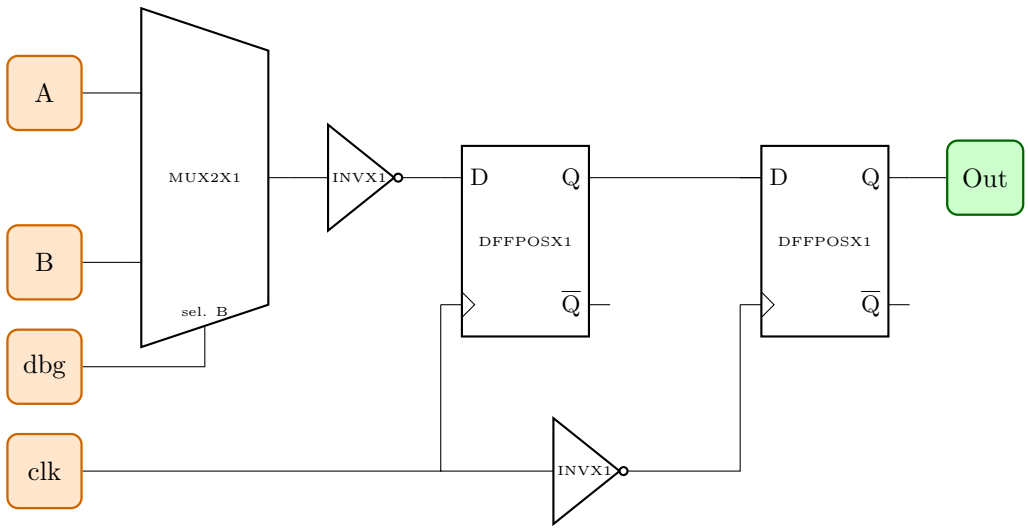
[2] Pour une image haute définition complète, cela représente $\approx 1920 \times 5 = 9.6\,\text{kB}$ de données, soit une quantité non négligeable.



Circuit 4.13 : Schéma de l'architecture du filtre montrant le chemin des données d'entrée sans chaîne de test. Les Bascule Ds sont marqués D, les multiplicateurs M, et les additionneurs Axx, où « xx » est la taille du mot d'entrée. Les Bascule Ds D de la colonne la plus à gauche fonctionnent cinq fois plus vite que le reste du circuit.

été placés dans le circuit pour permettre l'échantillonnage de la plupart des signaux à des fins de débogage. Lorsque le signal de débogage est activé, les registres à décalage lisent les données d'entrée à partir de leur port secondaire de « débogage », comme le montre le Circuit 4.14. Initialement conçus pour être déclenchés sur front descendant, comme décrit dans l'extrait de code A.11, les simulations ont montré que des problèmes survenaient alors avec la distribution manuelle du signal d'horloge, et le circuit multiplicateur hautement sensible aux délais. Plutôt que d'échantillonner le signal d'entrée sur front descendant et de le refléter immédiatement sur la sortie, le circuit a été modifié pour changer la sortie sur un front montant d'horloge, comme décrit dans l'extrait de code A.12.

Une variante 1-bit de circuit a ensuite été synthétisée dans l'extrait de code A.13, car cela permettait de diminuer la charge de travail de layout manuel, effectué en raison de l'indisponibilité des bibliothèques de cellules standard. La variante 1-bit permet la simplification du layout, puis l'utilisation de huit structures en parallèle pour obtenir un registre à décalage 8 bit. Une optimisation supplémentaire est la réutilisation d'une bascule à front descendant avec un signal d'horloge inversé, rendu possible par la non-criticité des délais de propagation de l'horloge. Cela évite la duplication du travail de layout de la bascule, qui est la partie la plus complexe. Le circuit résultant est montré sur le Circuit 4.14.



Circuit 4.14 : Schéma interne du registre à décalage. Le signal dbg d'« activation du débogage » sélectionne l'entrée secondaire du multiplexeur, généralement connectée à la sortie du registre à décalage précédent, comme montré sur le Circuit 4.17.

Précision binaire

Le circuit du multiplieur a été conçu pour être aussi générique que possible. La compatibilité avec les noyaux de filtrage courants, tels que ceux présentés dans la figure 4.3, a donc été considérée lors de la phase de conception.

Les exigences étaient les suivantes :

- Données d'entrée : Image avec 8 bit/pixel
- Données de sortie : Image avec la même profondeur de bits (8 bit/pixel)
- Le noyaux doit pouvoir contenir :
  - Des valeurs négatives (Sobel, unsharp)
  - Des valeurs relativement élevées, car les masques unsharp de 5×5 ont une valeur maximale de 476 (9 bits)
  - Un facteur de mise à l'échelle d'au moins 256, pour permettre l'utilisation de coefficients aussi élevés

Comme compromis entre ces exigences et la complexité des efforts de conception avec une capacité d'I/O limitée, un noyau 5×5 de 8 bit signés a été sélectionné.

4.6.3 Multiplicateur logique en mémoire à FeFET

Circuit multiplicateur et circuit additionneur

Les portes logiques (non-)ET peuvent être conçues à l'aide d'un unique FeFET, comme détaillé dans la sous-section 4.4.1. Cela peut directement être exploité dans un circuit multiplicateur : comme illustré dans la tableau 4.2, une multiplication binaire utilise plusieurs opérations ET. Toutefois, la véritable difficulté du circuit multiplicateur réside dans l'ajout des bits de retenue.

Noyau					I_1	I_0
K_0					$K_0 \cdot I_1$	$K_0 \cdot I_0$
K_1				$K_1 \cdot I_1$	$K_1 \cdot I_0$	
K_s			$K_s \cdot I_1$	$K_s \cdot I_0$		
$K_{s,ext}$		$K_s \cdot I_1$	$K_s \cdot I_0$			
	$K_s \cdot I_1$	$K_s \cdot I_0$				
Bit sortie	O_5	O_4	O_3	O_2	O_1	O_0
Valeur	$K_s \cdot I_1$ $+K_s \cdot I_0$	$K_s \cdot I_1$ $+K_s \cdot I_0$	$K_s \cdot I_1$ $+K_s \cdot I_0$	$K_1 \cdot I_1$ $+K_s \cdot I_0$	$K_1 \cdot I_1$ $+K_1 \cdot I_0$	$K_0 \cdot I_0$

TAB. 4.2 : Exemple concis de multiplication d'un entier non signé de 2 bits (I, horizontal) par un entier signé de 3 bits (K, vertical). Le bit de signe K_s doit être étendu du dernier bit du noyau (3^{rd} bit K_3) à la longueur maximale de sortie, 5 bits. La double ligne horizontale délimite la partie du multiplicateur ne nécessitant pas d'extension de signe, placée dans le premier étage du pipeline. Les bits de retenue ne sont pas indiqués dans les expressions de la ligne « valeur ».

Comme illustré dans la tableau 4.2, un multiplicateur peut être considéré comme un additionneur dont les opérandes d'entrée sont des nombres progressivement décalés, multipliés bit-à-bit. La multiplication bit-à-bit est un ET logique, et peut donc être réalisée par un seul FeFET, entre la valeur stockée et la valeur d'entrée. Un additionneur standard a ainsi été conçu, et l'opération logique ET intégrée dans un FeFET.

Additionneur à propagation de retenue

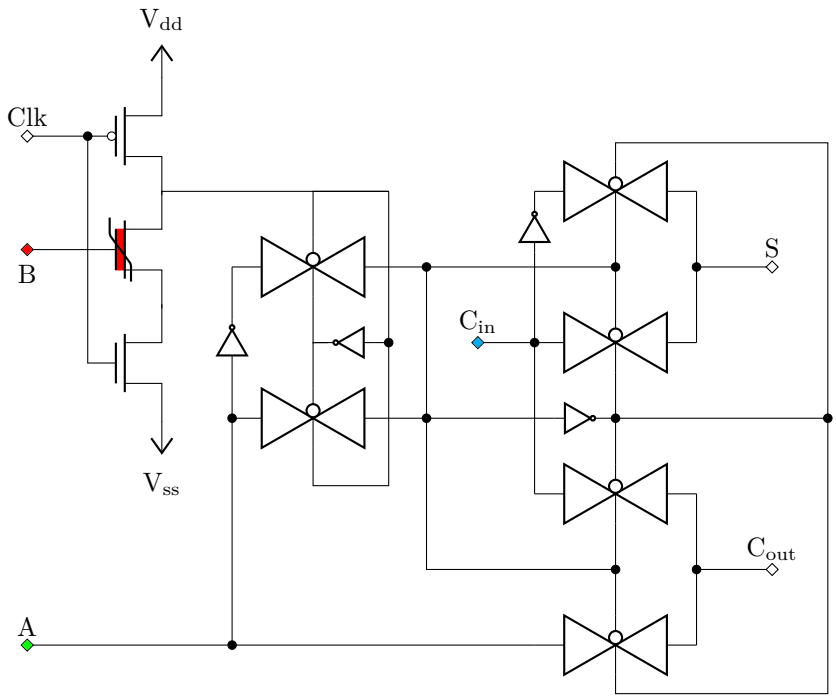
Une architecture d'additionneur à propagation de retenue a été choisie en raison de sa régularité, afin de réduire le temps et la complexité de la phase de conception. Les multiplicateurs à tableau de propagation de retenue sont très réguliers et nécessitent uniquement des circuits additionneurs complets et demi-additionneurs, au prix de performance généralement plus faible et d'un nombre de transistors plus élevé. Un demi-additionneur étant également un additionneur complet dont l'entrée de retenue est fixée à 0, la seule cellule devant être conçue est celle d'un additionneur complet dont les sorties sont données par l'équation 4.3 et l'équation 4.4.

$$\begin{cases} S_{out} = C_{in} \oplus A \oplus B & (4.3) \\ C_{out} = (A \cdot B) + (C_{in} \cdot (A \oplus B)) & (4.4) \end{cases}$$

Les Circuit B.1, Circuit B.2 et Circuit 4.15 montrent plusieurs itérations de l'évolution du circuit de l'additionneur complet.

Architecture pipelinée

Le délai le plus défavorable du multiplicateur en simulation ne permettait pas de maintenir un flux vidéo Full-HD à 25 images par seconde (FPS, Frames Per Second) en niveaux de gris



CIRCUIT 4.15 : Circuit final d'un additionneur complet comportant une porte logique ET à un FeFET intégré (en rouge), utilisé dans le filtre. La porte ET effectue une multiplication binaire de l'entrée B avec la valeur stockée dans l'oxyde ferroélectrique, avant de l'ajouter aux entrées A et C_{in}. C_{out} et S sont les sorties de l'additionneur complet, correspondant respectivement à la retenue et à la somme.

(clk $= 1920 \times 1080 \times 25 \approx 52\,\text{MHz} \approx 49\,\text{MiB}\,\text{s}^{-1}$, ou 19 ns/pixel). Il a donc été choisi d'utiliser une architecture pipelinée, car le facteur limitant de la première version du multiplicateur, présenté sur le Circuit B.1, était le taux de décharge du nœud pour propager la retenue. L'architecture pipelinée introduit une plus grande complexité de conception, ainsi qu'une latence augmentée d'un cycle, mais double approximativement le débit en permettant à la décharge du nœud de se produire en parallèle dans les deux étages.

Le multiplicateur a donc été divisé en deux moitiés, comme contré sur le Circuit 4.16 et le Circuit 4.17, le long de la délimitation de l'extension de signe représentée sur la tableau 4.2.

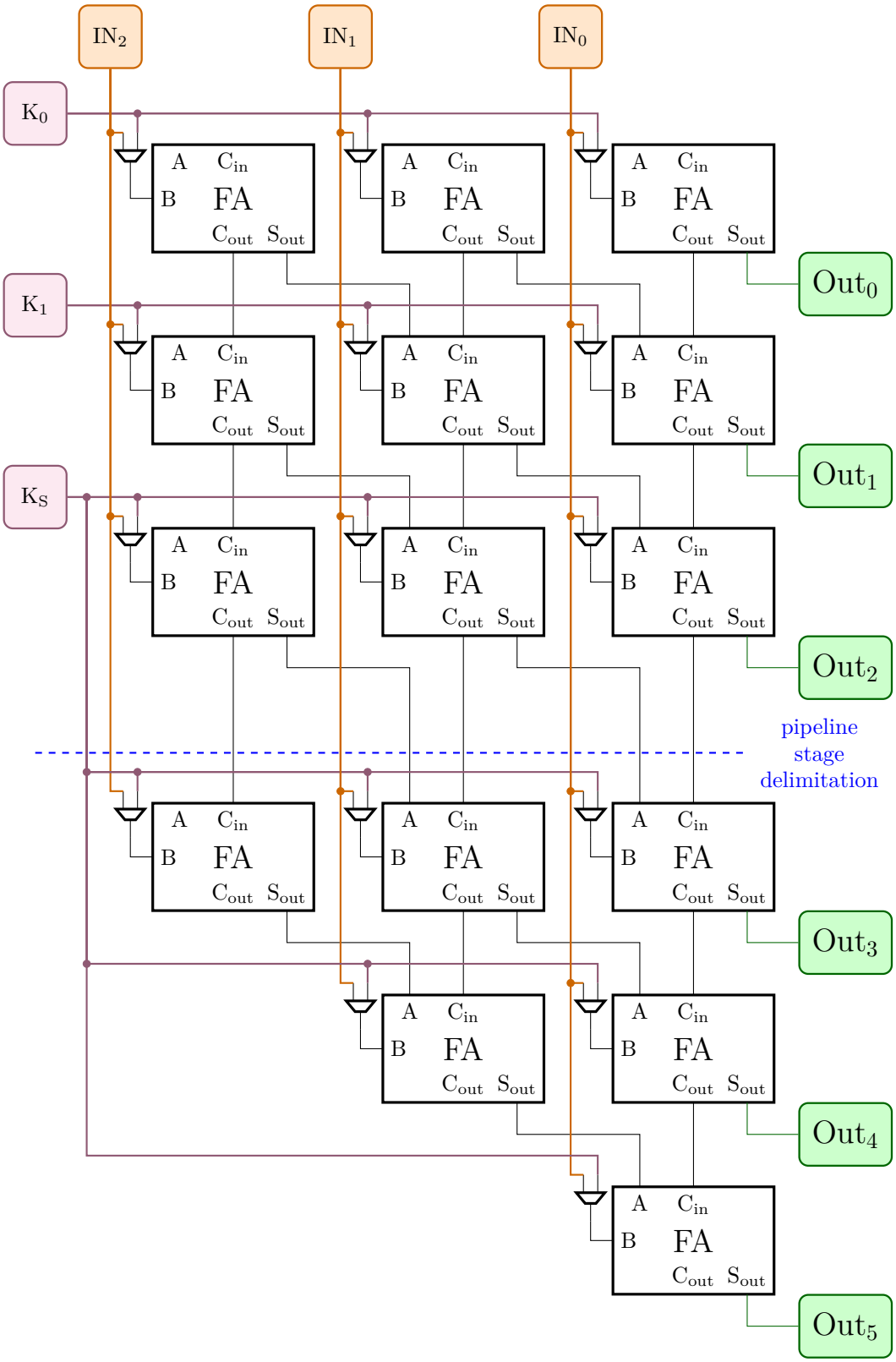
Le second étage est comporte moins d'éléments, car les bits de retenue ne sont pas propagés au-delà. L'implémentation choisie utilise deux fois le même circuit, ignorant les signaux de sortie C_{out} et S superflus, comme montré sur le Circuit 4.17. Des additionneurs complets ne sont pas nécessaires pour la première ligne (et colonne) du multiplicateur, car leurs entrées C_{in} et A sont inutiles, comme visible sur le Circuit 4.16. Des demi-additionneurs auraient pu être conçus, mais la surface de circuit n'étant pas une limitation, le même circuit a été réutilisé à la place, avec des entrées connectées à V_{ss}, comme le montre le Circuit 4.17.

En conséquence du choix de séparer les étages du pipeline du multiplicateur selon l'extension de signe :

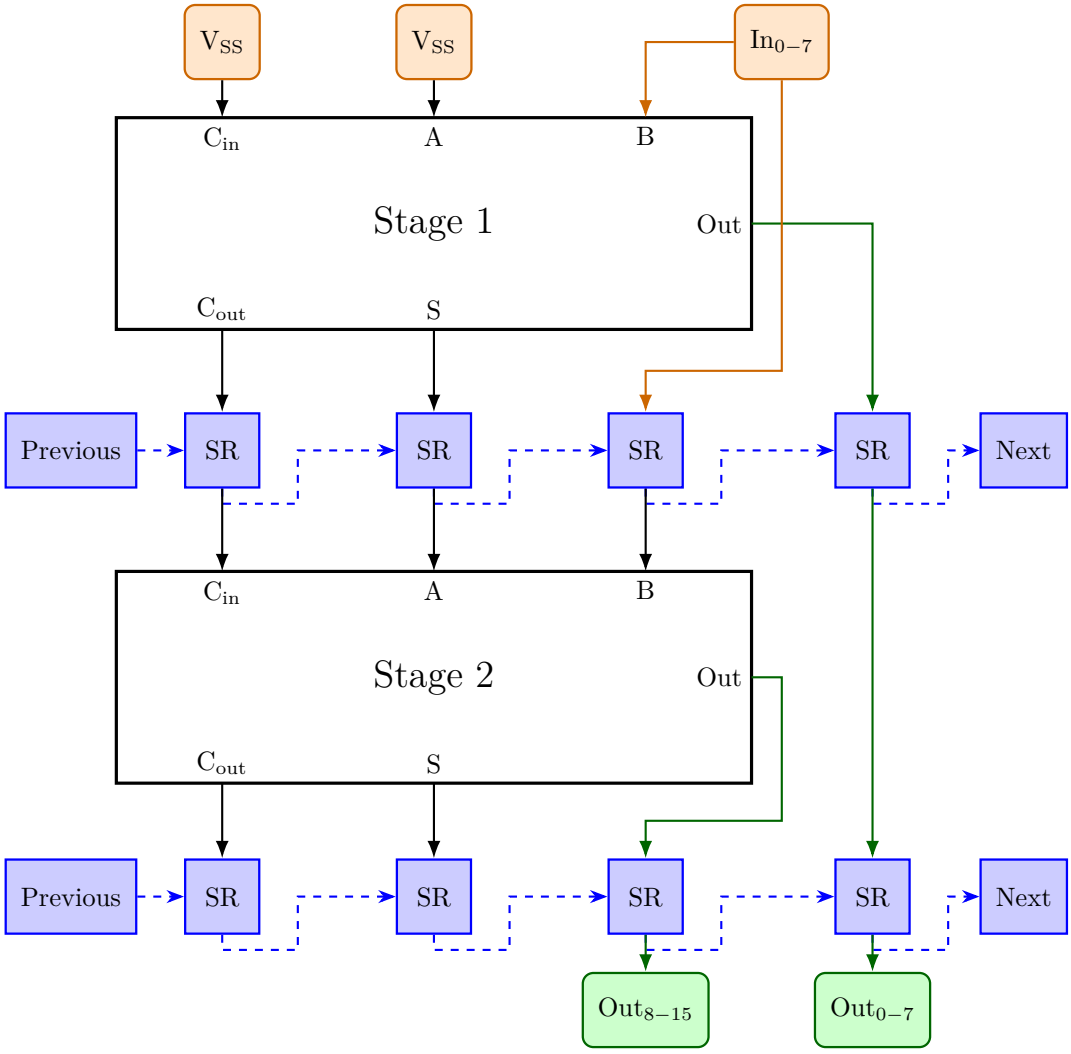
1. Chaque étage du multiplicateur produit 8 bits (l'amplitude maximale de la valeur de sortie est $-2^7 \cdot (2^8 - 1) = -32640$, ce qui nécessite 16 bits)

2. Seule la seconde moitié du pipeline nécessite l'extension de signe. Plus précisément, seule l'extension de signe est requise par les multiplicateurs du second étage de pipeline, les autres bits du noyau n'étant pas utilisés après le premier étage.

Les mêmes registres à décalage de 8 bits peuvent donc être utilisés entre les deux étages.



Circuit 4.16 : Schéma de l'additionneur à propagation de retenue 3×3, conçu à partir d'additionneurs complets (la version fabriquée étant de dimension 8×8). Les I/Os indiquées sont les bits d'entrée de l'image IN_n, les bits de sortie Out_n ainsi que les bits du noyau du filtre K_n. Les multiplexeurs sont contrôlés par le signal de programmation, pour connecter chaque entrée sur les bits correspondants du noyau, à tension amplifiée, permettantd'écrire les coefficients du noyau dans chaque additionneur complet. La chaîne critique suit les additionneur complet les plus à gauche sur chaque ligne, les entrées non connectées sont rappelées à un niveau logique bas.



CIRCUIT 4.17 : Étages de pipeline du multiplicateur, avec leurs entrées et sorties. Ce schéma montre l'architecture complète du pipeline du multiplicateur avec ses 8 bits d'entrée et 16 de sortie, ainsi que les registres à décalage utilisés pour le pipeline, et la chaîne de test utilisée pour le débogage et la programmation. Le dernier multiplicateur du premier étage de pipeline voit sa chaîne de test connectée à l'entrée de celle du second étage du premier multiplicateur. Chaque ligne de données représente un bus 8 bit. Ce schéma représente le bloc M du Circuit 4.13.

Programmation des coefficients poids du noyau

Un avantage supplémentaire de l'architecture pipelinée est que le noyau 8 bits du filtre peut être transmis en série sur les mêmes registres à décalage 8 bits, grâce à la chaîne de test : le multiplicateur du second étage utilise comme coefficients les 8 bits d'extension de signe. Les poids peuvent donc être transmis jusqu'aux mêmes entrées du multiplicateur (plus précisément, le FeFET des circuits additionneur complet) que celles qui recevront plus tard les données de pixels de l'image d'entrée.

L'extension du signe peut donc être envoyée en premier le long de la chaîne de test, répétée dans chaque bit de données, avant d'envoyer les 8 bits restants des coefficients du noyau. La chaîne de test relie également les registres à décalage entre les étages du multiplicateur, qui sont plus difficiles d'accès pour les données de pixels en fonctionnement normal. Cette opération de transmission des données du filtre est présentée sur le banc d'essai Verilog de l'extrait de code A.14. Les poids étant stockés de manière non-volatile, cette opération n'est pas critique en termes de temps d'exécution, et peut donc être mise en œuvre de manière plus efficace en termes d'espace, comme c'est ici le cas avec l'approche de la chaîne de test.

Il convient de noter que cette approche stocke les bits d'extension de signe séparément, au lieu d'étendre automatiquement le bit de signe du noyau. Cette approche permet également d'effectuer des opérations non signées (en programmant des extensions de signe de valeur zéro), ou de programmer des valeurs de noyau plus importantes pour des opérations non signées : malgré le risque de débordement, cette approche peut être exploitée pour le calcul.

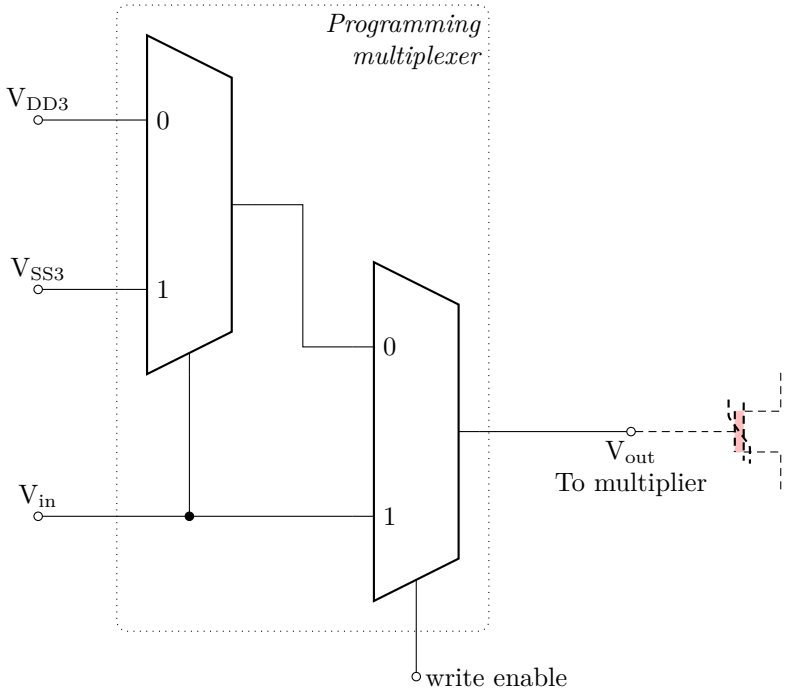
Une fois que les poids sont « alignés » avec le multiplicateur correspondant, c'est-à-dire que chaque bit du noyau est positionné là où les bits de l'image se trouvent pendant le fonctionnement du filtre, les bits doivent d'abord être redistribués : comme le montre le Circuit 4.16, chaque bit de l'image est distribué sur une colonne d'additionneurs complets, tandis que les bits du noyau sont distribués sur les lignes. Ceci est réalisé en activant un multiplexeur devant chaque additionneur complet, contrôlé par le signal *écriture*. Afin de programmer l'entrée des FeFETs avec les poids de noyau présentés, les niveaux de tension doivent également être élevés au-dessus de V_C pour les polariser. Un convertisseur de niveau de tension logique est utilisé à cet effet, placé après le multiplexeur. Comme le montre le Circuit 4.18, ce circuit est aussi composé de deux multiplexeurs, afin d'élever ou d'abaisser la valeur d'entrée à la valeur haute tension correcte. Lorsque le signal de programmation est actif, la tension d'entrée est portée à l'une ou l'autre des tensions de programmation externes V_{DD3} et V_{SS3}, comme le montre le diagramme temporel de la figure 4.6. Pour des raisons détaillées dans la section 4.6.4, le signal *écriture* est effectivement le même que le signal « débogage » qui active la chaîne de test. Cela ne pose pas de problème lors du fonctionnement normal de la chaîne de test, car l'oxyde ferroélectrique n'est pas repolarisé tant que les tensions d'alimentation V_{DD3} et V_{SS3} restent inférieures à V_C.

Bien qu'il soit possible pour V_{DD3} et V_{SS3} de fournir simultanément leurs valeurs respectives de $\pm V_C$, cela créerait des différences potentielles allant jusqu'à 6 V. Cela risquerait d'endommager les grilles de transistor, y compris celles des cellules d'entrée/sortie. La rapidité de l'opération de programmation n'étant pas critiques, l'application des deux niveaux de tension est décalée, comme le montre la figure 4.6 : les FeFETs qui devant être programmés avec un état « logique haut » sont programmés en premier, et les FeFETs « logique bas » sont programmés lors de l'application d'une seconde impulsion. Cette approche peut également réduire la hauteur du pic de consommation d'énergie pendant l'écriture. La durée des impulsions de programmation est détaillée dans la section 4.6.5.

4.6.4 Validation en simulation et problèmes identifiés

La principale contribution au circuit conçu par NaMLab fut la vérification en simulation : les instabilités du modèle, ainsi que la taille et la complexité du circuit, ont empêché de simuler intégralement le circuit.

La simplification du modèle ferroélectrique, la création d'une représentation niveau transfert de registre (RTL, Register Transfer Level) et une implémentation logicielle du circuit, puis la comparaison des résultats à chaque étape ont contribué à l'identification de deux défauts majeurs dans la conception initiale, qui auraient empêché le circuit de fonctionner.



Circuit 4.18 : Multiplexeur de tension utilisé pour programmer le FeFET du multiplicateur. En fonction des entrées V_{in} et *écriture*, le multiplicateur recevra soit le signal V_{in} (en fonctionnement normal), soit l'un des niveaux de tension V_{DD3} ou V_{SS3} pendant la programmation. Ce circuit a été conçu par NaMLab, le second multiplexeur réalisé avec deux portes à transmission.

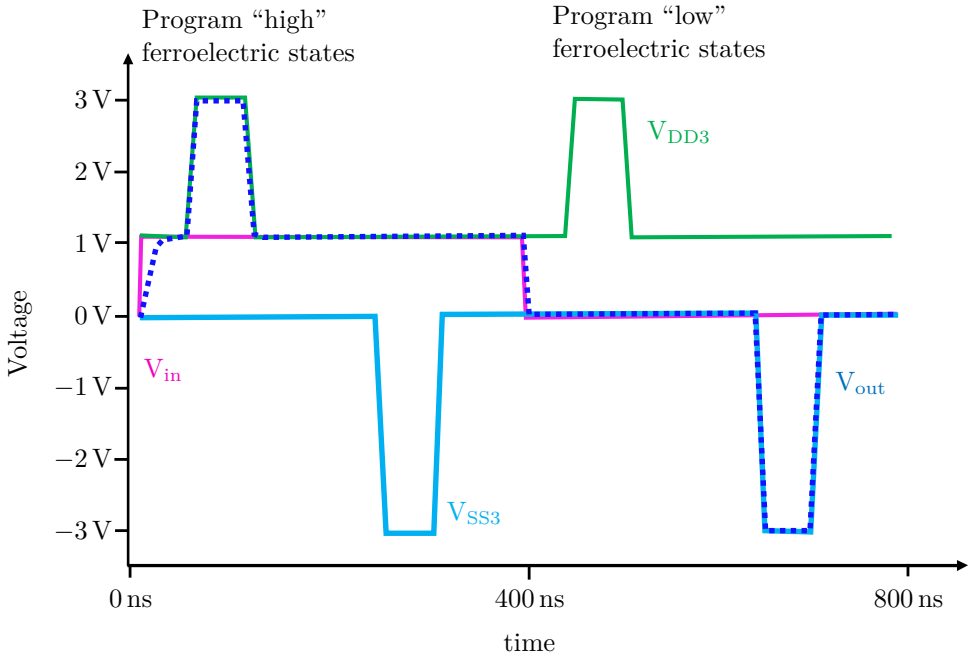


Fig. 4.6 : Entrée et sortie du multiplexeur décrit dans le Circuit 4.18 et connecté à l'entrée du FeFET du additionneur complet, noté « B » sur le Circuit 4.15. Lorsque la programmation est activée, le multiplexeur amplifie le signal V_{in} à V_{DD3} ou V_{SS3}, comme observable sur la sortie V_{out}.

Simplification de la simulation au niveau circuit

Le filtre proposé contient 25 multiplicateurs afin d'opérer avec un noyaux 5×5, chacun stockant un pixel complet du noyau de filtrage. Chaque multiplicateur possède deux étages de pipeline, chacun avec une matrice de 8×8 additionneurs complets contenant 1 FeFET chacun. Le nombre total est donc de $8 \cdot 8 \cdot 2 \cdot 25 = 1600$ FeFETs, et le nombre de transistors est plusieurs ordres de grandeur au-delà.

Il est possible de simuler un tel circuit, toutefois :

- la consommation mémoire de la simulation est relativement élevée, entre 10 GB à 20 GB ;
- la vitesse de simulation est un problème, prenant quelques jours par kilopixel sur une machine 8 cœurs ;
- les instabilités du modèle rendent la convergence difficile.

Les deux premiers points peuvent être améliorés grâce à des calculateurs plus puissants. Cependant, le modèle ferroélectrique expérimental utilisé pour modéliser les FeFET présente de rares problèmes de convergence, exacerbés par le nombre de dispositifs présents dans le circuit, empêchant ainsi une simulation complète dès les premiers cycles d'horloge. Ce modèle étant relativement lent et exigeant beaucoup de mémoire en raison du suivi de l'historique du ferroélectrique et des points d'inflexion décrits dans la sous-section 2.2.2, son remplacement par le modèle simplifié de l'extrait de code 2.1 décrit dans la sous-section 2.2.3 a amélioré tous les points susmentionnés, permettant une simulation réussie du filtre d'image traitant une image complète de 30 pixel × 30 pixel en quelques heures.

Génération de signaux d'entrée et de sortie de référence

Pour permettre la vérification des signaux de sortie, une référence de correcte doit être établie. Une approche descendante (« top-down ») a été utilisée pour générer des données de sortie correspondant à des données d'entrée arbitraires (pseudo-aléatoires). Les objectifs étaient les suivants :

1. valider que l'opération logique effectuée par le circuit est exactement la même que celle définie dans les bibliothèques de traitement d'images
2. aider à la création de signaux de test complexes pour programmer le noyau du filtre, puis fournir les données d'entrée, tout en respectant la synchronisation et l'alignement relatif des différents signaux d'horloge et de données
3. obtenir des signaux de sortie de référence à comparer aux signaux de sortie du circuit simulé.

Pour ce faire, une approche de test en plusieurs étapes a été élaborée, telle qu'illustrée dans la figure 4.7.

Un modèle de très haut niveau du filtre a d'abord été créé avec la fonction de convolution 2D intégrée `conv2` de GNU Octave (compatible avec MATLAB), ainsi que des données aléatoires de test. La sortie est considérée comme « réalité de terrain », la sortie des différents modèles doit donc y correspondre pour que ceux-ci soient considérés comme fonctionnels.

Ensuite, un modèle Verilog de haut niveau a été réalisé, afin d'établir une référence précise des données d'entrée et sortie, cadencées à la fréquence du filtre. Ce modèle a été développé en comparant systématiquement les résultats obtenus au modèle de niveau supérieur, pour s'assurer de sa fiabilité. Cependant, lors de la réalisation simultanée du schéma électrique et de la représentation Verilog, il est apparu qu'un modèle « boîte noire » du filtre n'était ni utile pour le débogage, ni facile à ajuster pour correspondre au schéma : lorsque des différences sont apparues dans la sortie, la modélisation des sous-composants est devenue une nécessité afin d'isoler les défauts. La hiérarchie du modèle Verilog final est donc très proche de celle du circuit.

Grâce à la disponibilité d'un modèle Verilog détaillé, il est possible de comparer les signaux internes à divers points. En outre, la disponibilité d'une chaîne de test permet une approche de *génération automatique de motifs de test*, en générant une entrée pseudo-aléatoire, en la filtrant par les différents modèles, puis en comparant les résultats. Les formes d'onde « vcd »

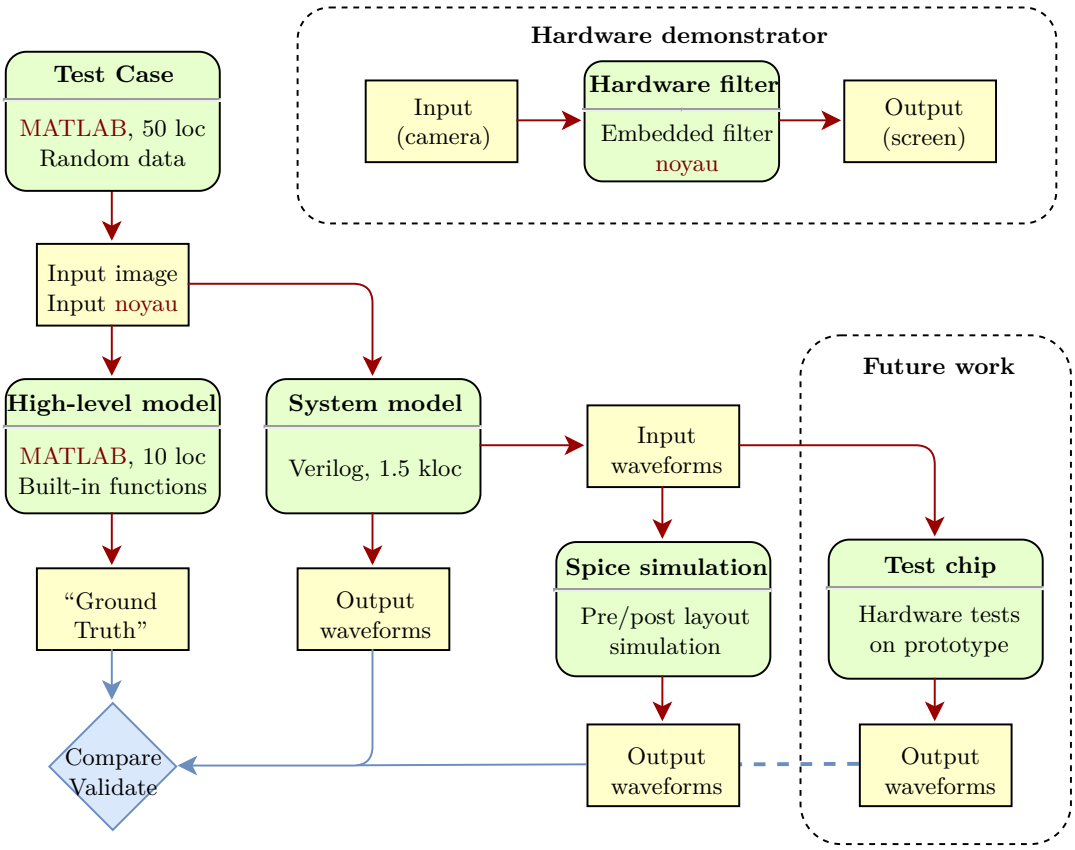


FIG. 4.7 : Flot de vérification utilisé pour valider le circuit du filtre d'image

générées à partir du banc d'essai Verilog simulé sous Icarus Verilog[Wil] ont été directement importées dans le simulateur Spectre de Cadence[3]. Les blocs Verilog-A fournis dans l'extrait de code A.5 et l'extrait de code A.6 ont été utilisés pour sérialiser les données transitant sur les bus internes vers des fichiers binaires qui pouvant directement être comparés avec ceux émis par leurs homologues Verilog. L'automatisation de la comparaison ci-dessus s'est avérée cruciale pour diagnostiquer deux problèmes détaillés dans les sections suivantes.

Il est important de noter que les problèmes identifiés sont des problèmes d'intégration, et que ceux-ci ne peuvent être découverts en effectuant des simulations isolées. Ces résultats illustrent l'utilité des modèles simplifiés pour les simulations à grande échelle.

Mauvaise synchronisation pour déclencher le multiplicateur

Ce premier problème est relativement simple, mais seule une simulation englobant un multiplicateur ainsi que ses registres à décalage d'entrée peut le mettre en évidence. Dans la première version du filtre, les multiplicateurs partageaient la même horloge que les registres à décalage mémorisant leur sortie. Il en résultait un problème de synchronisation où les registres à décalage recevaient un front montant pour stocker les valeurs de sortie du multiplicateur avant que celui-ci n'ait fini de calculer. Ce problème ne se produisait que pour les multiplicateurs les plus lents (c'est-à-dire dans les cas où le bit de retenue doit être propagé à travers plusieurs additionneurs complets).

La solution évidente aurait été l'ajustement de la phase entre ces deux horloges, mais une mauvaise valeur aurait également compromis la fonctionnalité du circuit. Il était donc préférable d'utiliser un signal d'horloge dédié aux multiplicateurs. Cependant, le circuit étant limité par le nombre d'I/O disponibles, l'ajout d'un signal supplémentaire était impossible, car le budget de 25 plots était entièrement utilisé.

[3]Dans ADE L, Configuration -> Fichiers de simulation

Le protocole d'écriture alternatif utilisé afin de libérer le pad write_enable pour un signal d'horloge supplémentaire est intéressant à détailler. L'affectation des plots d'I/O, en plus des 16 plots occupés par les bus parallèles 8 bit d'entrée et de sortie, était auparavant :

- alimentations standard : $V_{DD} = 1.0\,V$, $V_{SS} = 0.0\,V$;
- alimentations pour programmer les ferroélectriques :$V_{DD3} = V_P$, $V_{SS3} = -V_P$, avec V_P la tension de programmation choisie, supérieure à V_C ;
- une tension supplémentaire V_{shift} pour piloter les transistors du convertisseur de niveau logique du circuit de programmation représenté par le multiplexeur gauche du Circuit 4.18 ;
- Clk_1, horloge utilisée pour les cinq premiers registres à décalage du chemin des données d'entrée montré sur le Circuit 4.13, normalement cinq fois plus rapide que Clk_2 ;
- Clk_2, utilisée pour les multiplicateurs et registres à décalage ;
- $Write_{Enable}$, signal de commande pour amplifier le niveau des tensions d'entrée des transistors ferroélectriques depuis V_{DD} et V_{SS} vers V_{DD3} et V_{SS3} respectivement, afin de permettre la repolarisation de la couche ferroélectrique ;
- $Debug_{Enable}$, signal de contrôle pour activer la chaîne de test de « débogage », avec laquelle les registres à décalage sont connectés en série.

Le signal $Write_{Enable}$ a été regroupé avec $Debug_{Enable}$, et les tensions de programmation fournies depuis l'extérieur ont été rendues dynamiques, car celles-ci doivent être suffisamment élevées pour permettre l'écriture :

- Inchangé : V_{DD}, V_{SS}, V_{shift}, Clk_1, Clk_2 ;
- Les tensions V_{DD3} et V_{SS3} sont désormais pilotées dynamiquement depuis l'extérieur, et portées soit à V_{DD} et V_{SS}, soit à $\pm V_P$ pendant la phase de programmation.
- Un nouveau signal d'horloge Clk_3 a été ajouté pour les multiplicateurs, utilisant le plot d'I/O précédemment utilisé pour $Write_{Enable}$;
- $Debug_{Enable}$ est donc utilisé tant pour activer le mode de débogage que pour changer le niveau des signaux d'entrée des FeFETs depuis V_{DD} et V_{SS} vers V_{DD3} et V_{SS3}, respectivement.

Les tensions de programmation étant fournies par une source externe, celles-ci peuvent être abaissées aux niveaux logiques normaux si la programmation de l'oxyde ferroélectrique n'est pas souhaitée. Ainsi, même si le signal $Debug_{Enable}$ active les mêmes circuits de pilotage de la tension de programmation que le signal $Write_{Enable}$ ne le faisait précédement, la repolarisation de l'oxyde ferroélectrique est évitée en abaissant les niveaux de V_{DD3} et V_{SS3} sous V_C.

Signaux dépendant de la décharge du nœud flottant

Le second problème critique identifié a conduit à une révision majeure du circuit additionneur complet à FeFET. Il s'agit également d'un problème ne pouvant être identifié qu'en simulation post-intégration.

La source de ce problème provient de l'utilisation de logique dynamique en raison de l'indisponibilité de p-FeFET : un nœud capacitif est préchargé, puis l'expression est évaluée et le nœud est déchargé si nécessaire. Cependant, celui-ci ne peut pas être rechargé avant le prochain cycle d'horloge, même si les signaux d'entrée sont modifiés. Un tel problème de dépendance est apparu en simulation et a conduit à la refonte complète du multiplicateur, migrant d'une architecture entièrement conçue en logique dynamique vers une architecture mixte dynamique et statique.

Le Circuit 4.16 illustre le problème : le bit de retenue est propagé à travers les additionneurs, ce qui entraîne une latence importante dans le pire cas. Lorsque la latence est supérieure au temps restant pour l'évaluation, le bit de retenue est ignoré lors du calcul de la valeur de sortie, entraînant des erreurs dans les bits de poids fort.

Le circuit initial présenté sur le Circuit B.2 est composé de deux moitiés dynamiques : la première, à gauche, calcule la retenue de sortie C_{out} à partir de trois entrées, comprenant la retenue de sortie C_{in} du bit précédent. La retenue obtenue est propagée vers le bloc de logique dynamique de l'additionneur complet suivant par l'intermédiaire du nœud flottant NOT(C_{out}). Si ce nœud se décharge prématurément pendant que l'entrée n'est pas encore stabilisée (la propagation de la retenue d'entrée n'étant pas terminée), celui-ci ne peut pas être rehaussé. Les valeurs de la retenue et du résultat de l'addition générées par l'étage seront donc toutes deux erronées.

Il est possible de remédier à ce problème en retardant précisément le signal d'horloge afin de ne commencer l'évaluation des bits de poids fort qu'après avoir attendu le pire délai de propagation de la retenue. Cependant, cela plafonne la performance du circuit en introduisant une valeur de latence minimale par additionneur complet. Outre la complexité de conception et d'intégration du circuit de synchronisation, la latence précise du circuit n'étant pas connue pour cette technologie exploratoire, l'introduction d'une latence minimale aurait également compromis la mesure de la performance du circuit.

Le circuit a donc été redessiné avec une approche hybride CMOS et logique dynamique, comme décrit dans la section 4.3.3. Le circuit final présenté sur le Circuit 4.15 utilise un unique transistor ferroélectrique alimenté dynamiquement, dont la sortie contrôle plusieurs portes de transmission. Ces signaux de contrôle des portes de transmission sont les seuls nœuds flottants. Leur état ne dépend pas de la retenue d'entrée, car cette partie du calcul est contrôlée à l'aide de logique CMOS conventionnelle. Cela élimine la dépendance à l'égard de l'étage multiplicateur précédent, assouplissant les contraintes temporelles et réduisant la latence minimale. Le nombre total de transistors est augmenté, mais le nombre de FeFETs est réduit. Les FeFETs étant plus grands (500 nm × 500 nm dans ce cas) que les transistors ordinaires, cela conduit à des surfaces de circuit comparables. Les contraintes sur la capacité du nœud flottant préchargé par l'alimentation peuvent également être détendues, ce qui permet de réduire les dimensions des portes de transmission connectées au nœud flottant, accélérant également la vitesse de calcul.

En conséquence, le circuit a été amélioré en réduisant le nombre de FeFETs, et en réduisant auttant que possible la proportion de logique dynamique. Cela a également permis d'améliorer le délai de propagation maximal dans le multiplicateur, en le réduisant à 1.52 ns en simulation, ce qui rend l'architecture en pipeline redondante. Les additionneurs présents en sortie nécessitent uniquement 535 ps, 595 ps et 661 ps pour fonctionner, respectivement pour les versions 16 bit, 18 bit et 20 bit.

4.6.5 Résultats

Programmation des coefficients

Lors de la programmation des poids selon le protocole décrit dans la section 4.6.3, il est possible de moduler deux paramètres influençant la tension de seuil : la hauteur (tension) de l'impulsion de programmation et sa largeur (durée). Ces paramètres influencent le nombre de domaines repolarisés : plus ceux-ci sont élevés, plus le décalage de V_{th} est important.

Les données expérimentales sont affichées sur la figure 4.8. Celles-ci ont mené au choix d'une tension de programmation de ±3 V, afin de préserver les transistors, et d'une durée d'impulsion de 10 µs. En effet, la programmation est une opération ponctuelle, effectuée lors de la mise à jour des coefficients du noyau, qui n'est pas critique en termes de performances. En outre, les tensions de programmation positive et négative ne sont pas appliquées simultanément, afin d'éviter des différences de potentiel de 6 V pouvant endommager les transistors, y compris ceux d'entrée/sortie supportant habituellement des tensions plus élevées.

Point de fonctionnement

Comme visualisable sur le Circuit 4.15, la résistance drain-source du transistor ferroélectrique contrôle le taux de décharge du nœud flottant. Cette résistance doit donc être réglée avec soin, car celui-ci ne doit pas se décharger avant que la retenue n'ait été propagée à tous les additionneur complet : R_{DS}(off) doit être suffisamment élevée. Le même nœud doit également se décharger rapidement dans le cas inverse, cette entrée étant nécessaire pour calculer la retenue ensuite propagée. Quatre cas sont donc possibles :

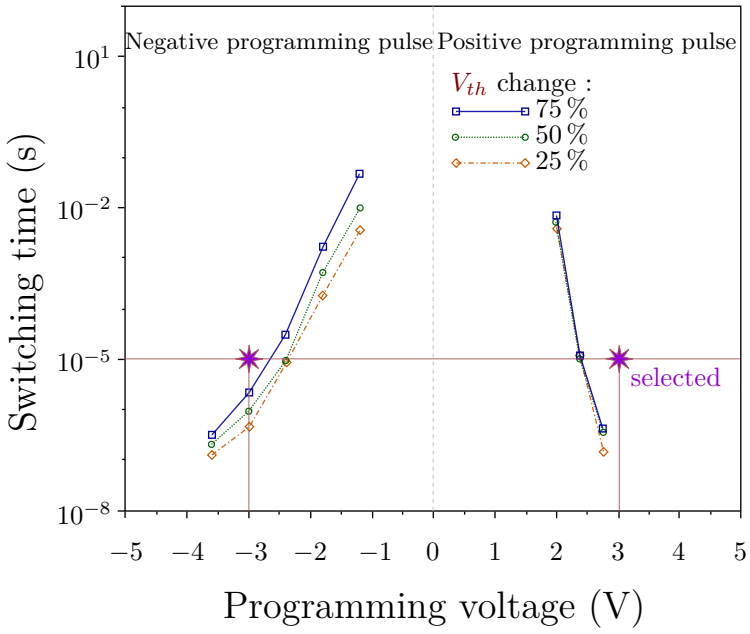


FIG. 4.8 : Temps d'impulsion mesuré nécessaire pour obtenir un changement de V_{th} du FeFET de 25 %, 50 % et 75 % entre le V_{th} bas (V_{th}^{FL}) et le V_{th} haut (V_{th}^{FH}), en fonction de la tension appliquée. Les étoiles indiquent la position des durées de programmation et d'effacement choisis (10 µs) pour le fonctionnement sous 3 V. Ces points de fonctionnement ont été choisis pour obtenir la plus grande variation possible de V_{th} tout en préservant le circuit de l'application de tensions élevées.

A (V_{in})	État ferroélectrique	R_{DS}	Contrainte de décalage de V_{th}
Haut	Élevé	Bas	$V_{in}^{H} + \Delta V^{+} > V_{th}$
Bas	Élevé	Élevé	$V_{in}^{L} + \Delta V^{+} < V_{th}$
Élevé	Bas	Élevé	$V_{in}^{H} - \Delta V^{-} < V_{th}$
Bas	Bas	Élevé	$V_{in}^{L} - \Delta V^{-} < V_{th}$

D'après le tableau ci-dessus, cinq paramètres peuvent être ajustés pour respecter les conditions :

- V_{th} : Dépend du processus de fabrication, intrinsèque aux FeFET
- V^{+} et V^{-} : décalage de V_{th} dans les deux directions, ajusté lors de la procédure de programmation des coefficients
- V_{in}^{H} et V_{in}^{L} : tensions logiques (V_{DD} et V_{SS}) pour l'I/O du circuit

Le premier point correspond à une étape d'optimisation du processus technologique, plus qu'un paramètre réglable. Celui-ci peut être ajusté par dopage ou ingénierie de la fonction de travail de la grille[Rei+19]. Le second point a été choisi pour maximiser le décalage ΔV de V_{th}, comme détaillé dans la partie précédente. Enfin, le troisième point peut être ajusté librement, bien que cela n'affecte également les performances des CMOS.

La figure 4.9 montre les données recueillies pour le choix du point de fonctionnement. Avec des entrées de filtre et des coefficients de noyau connus (tous deux entièrement remplis de zéros ou de un), la sortie est observée et comparée aux résultats attendus.

Une décharge précoce, plus susceptible de se produire lorsque l'entrée est proche de la tension seuil choisie pour le FeFET, $V_{th} + \Delta V_{th}^{x}$, se produit lorsque la sortie passe à « 1 » de manière inattendue. Cet événement est représenté en bleu clair sur la figure 4.9. Un tel événement est plus probable quand l'oxyde ferroélectrique est dans l'état « haut » (faible V_{th}), ou lorqu'un niveau logique haut est présent sur l'entrée. Au contraire, les zones bleu foncé montrent les endroits où cette décharge est attendue. Les zones rouge clair et foncé indiquent respectivement des zéros inattendus et attendus dans la valeur de sortie. Entre les deux, on trouve des zones marquées en gris où seuls certains bits de sortie sont à zéro ou à un. Ces zones correspondent également à des valeurs non désirées.

Le filtre d'image doit donc fonctionner correctement dans une région où seuls les résultats attendus sont présents. Cette région correspond à l'union des zones colorées teintes plus foncées, et est représentée en vert sur la droite du diagramme. Un point de fonctionnement à 0.6 V et autour de 22 ± 0.5 ns a été identifié, juste en au-delà des 19.2 ns requises pour le traitement Full-HD à 25 FPS, ce qui met en évidence la difficulté du réglage de tels circuits. Le circuit reste utilisable à un taux de rafraîchissement inférieur de 23.5 FPS à 24.7 FPS.

Ce graphique montre également que la zone grise « invalide » est généralement beaucoup plus large pour le bit de poids fort (MSB, Most-Significant Bit) (chiffre n° 0) que pour le bit de poids faible (LSB, Least-Significant Bit) (chiffre n° 7). Cela est dû au fait que plusieurs additionneurs complets contribuent au MSB (chemin critique plus long pour la retenue) alors qu'un seul contribue au LSB, comme visible sur le Circuit 4.16. La variabilité d'un dispositif à l'autre implique que certains multiplicateurs ont des nœuds flottants qui se déchargent plus rapidement que d'autres, la combinaison d'un plus grand nombre de multiplicateurs se traduit donc par une plus large gamme de fréquences produisant une sortie invalide.

Ce graphique montre également que des tensions de fonctionnement plus basses entraînent un fonctionnement plus lent du dispositif, en raison de la tension plus faible appliquée sur le FeFET et le n-MOS d'évaluation. Cela ralentit de la même façon la vitesse de fonctionnement des circuits CMOS via l'augmentation du R_{DS}(ON) des n-MOS.

Démonstrateur interactif final

La fonctionnalité du filtre d'image a été démontrée à l'aide d'une plate-forme hybride combinant l'ASIC du filtre avec un FPGA contenant de la logique additionnelle. Le FPGA décode les images reçues sur une de ses entrées HDMI, les transmet au filtre d'image via son GPIO, et récupère les données filtrées par le biais de la même interface. Celui-ci effectue donc la mise en mémoire tampon nécessaire, qui n'a pas été implémentée sur l'ASIC, conférant

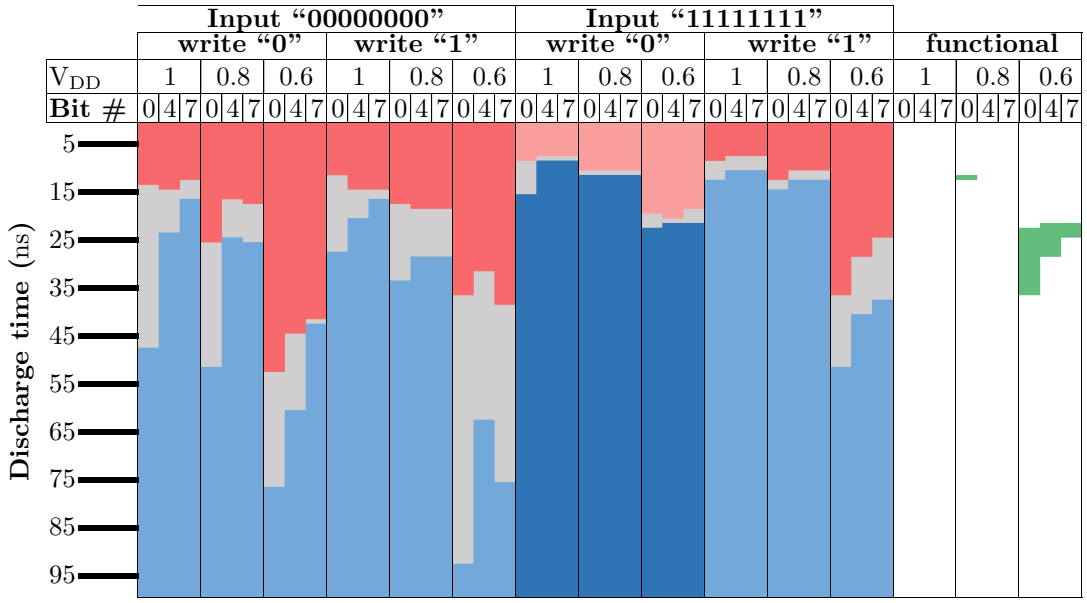


FIG. 4.9 : Résultats de la caractérisation dynamique du filtre d'image. Deux données d'entrée représentant un pixel entièrement noir « 00000000 » ou blanc « 11111111 » sont introduites à l'entrée du multiplicateur. Dans ces deux cas, le cas où l'oxyde ferroélectrique a été programmé par une impulsion positive ou négative est examiné, et la valeur de sortie de chaque bit est reportée. Les changements dans la sortie attendue du multiplicateur permettent de mesurer le temps de décharge du nœud flottant. La sortie est observée : les sorties « tout à un » sont représentées en bleu, tandis que les sorties « tout à zéro » le sont au-dessus, en rouge. Les sorties mixtes sont au milieu, en gris. Les régions sombres indiquent la sortie attendue, le point de fonctionnement doit donc être choisi à leur intersection pour un fonctionnement correct du filtre. Les valeurs de tension et de durée correspondantes sont affichées en vert, sur la droite. Chaque point de données représente le résultat statistique de 100 opérations de multiplication.

également une plus grande flexibilité à ce dernier. Le FPGA est aussi chargé de la génération du signal d'horloge et de la programmation des différents noyaux de filtre.

4.7 Conclusion

4.7.1 Logique à FeFET

La logique à FeFET a le potentiel d'accélérer les opérations logiques en stockant l'un des opérandes in-situ : à l'intérieur du circuit logique lui-même, en Mémoire non volatile. Cela peut s'avérer utile dans des cas spécifiques, et s'accompagne d'un ensemble de contraintes, principalement dues à la logique supplémentaire nécessaire à la programmation des transistors.

En effet, les transistors ferroélectriques ont besoin d'un circuit d'adressage séparé leur fournissant les valeurs à stocker. La programmation des oxydes ferroélectriques est commandée en la tension, ce qui nécessite un convertisseur de niveau logique. Des progrès ont été réalisés afin de réduire les tensions requises, bien que des valeurs plus élevées restent intrinsèquement nécessaires pour différencier les signaux de programmation des valeurs logiques. En conséquence, le reste du circuit doit être compatible avec ces tensions de programmation plus élevées.

Le besoin de niveaux de tension plus élevés peut être réduit en ajoutant des transistors d'accès pour programmer la couche ferroélectrique, comme détaillé dans la section 3.5 avec le circuit 2T1C. En outre, l'utilisation de structures 2T1C ou PsFeFET permet de partager l'oxyde ferroélectrique entre plusieurs transistors, ce qui est particulièrement intéressant dans le cadre de circuits CMOS, car les p- et n-FeFET partagent leurs signaux d'entrée.

Toutefois, le fait de séparer la couche ferroélectrique de la grille du transistor augmente les courants de fuite du nœud flottant, réduisant les capacités de rétention du dispositif. Si des périodes de rétention plus longues ne sont pas nécessaires, la plupart des circuits décrits dans ce chapitre peuvent également être réalisés avec un condensateur paraélectrique ordinaire en série avec la grille du transistor. Une structure de type FGMOS est ainsi obtenue, puisque l'inversion de la polarisation ferroélectrique n'est pas exploitée en utilisation normale. Le nœud flottant doit cette fois être préalablement chargé, ce qui peut se faire par l'intermédiaire d'un second transistor, comme présent dans la structure 2T1C, ou par d'autres moyens (la mémoire flash utilise des porteurs chauds, nécessitant des tensions élevées).

Dans l'état actuel de la technologie, une étude au cas par est nécessaire pour évaluer si la complexité supplémentaire liée à l'inclusion de FeFETs dans un circuit de calcul vaut le gain d'efficacité apporté.

4.7.2 Filtre d'image

Comme étude d'un cas pratique, une application prometteuse a été examinée, dont les besoins sont adaptés aux atouts des FeFET (lecture non destructive, non-volatilité), tout en évitant leurs faiblesses (écriture plus complexe).

En se focalisant sur un circuit acceptant un flux d'entrée séquentiel tel que des données vidéo, audio ou d'autres données sensorielles et effectuant des opérations LiM directes avec des données stockées dans des FeFET, la longue capacité de rétention de ceux-ci est exploitée, tout en évitant les écritures fréquentes avec leurs effets parasites associés, et éventuellement dégradants, de piégeage charges. La phase de programmation n'étant pas critique en termes de performances, ce cas d'utilisation permet d'éviter le problème de la lecture après écriture, tout en continuant d'utiliser les propriétés de reconfigurabilité. Cela simplifie également la logique de programmation, mais ne permet pas de tirer parti des écritures à vitesse élevée et à faible consommation d'énergie dont sont capables les FeFET.

Ce démonstrateur de filtre d'image montre la faisabilité de l'utilisation de FeFET dans des circuits de calcul réalistes, tirant parti de leur performance et de leur efficacité énergétique.

Celui-ci met également en évidence de nombreux points problématiques, notamment l'étroite plage de fonctionnement du dispositif, la variabilité relativement élevée des FeFET, et les limitations apportées par l'indisponibilité de p-FeFET. Néanmoins, cela constitue un point de référence intéressant pour la technologie FeFET actuelle, et un objectif d'optimisation permettant de mesurer les progrès des processus technologiques.

Un filtre classique stockerait probablement les poids dans des registres ou des bascules volatiles de type SRAM, et ceux-ci devraient être reprogrammés avec les coefficients souhaités

après chaque perte d'alimentation. Au contraire, la non-volatilité de ce circuit permet de réduire les délais et coûts énergétiques associés au démarrage, ainsi que sa consommation d'énergie statique. En supposant que les coefficients du filtre soient rarement modifiés, les contraintes sur le mécanisme d'écriture peuvent être réduites (un chemin de données plus lent et des temps de programmation plus longs sont tolérables), pouvant réduire son impact sur le reste du circuit.

4.7.3 Mémoires à FeFET

Les mêmes considérations s'appliquent aux mémoires, que les caractéristiques de rétention des oxydes ferroélectriques rendent attrayantes. En effet, ceux-ci promettent des périodes de rétention supérieures à 10 ans, ainsi que des opérations de lecture et d'écriture rapides, à faible consommation d'énergie et à tension modérée. Le point de référence pour cette application est celui des mémoires flash. Leurs avantages comparatifs incluent une meilleure rétention, une endurance vraisemblablement supérieure, des opérations plus rapides, des tensions de programmation et une consommation d'énergie plus faibles, ainsi que la compatibilité avec les circuits CMOS. Les mémoires à FeFET paraissent donc avantageuses vis-à-vis des mémoires flash.

Il est toutefois possible d'obtenir des gains de performance supplémentaires grâce à des architectures hybrides, car les dispositifs FeFET combinent deux mécanismes de rétention : ferroélectrique (exploité dans le circuit 1T1C) et capacitif (utilisé par les mémoires flash et DRAM). L'ajout d'un transistor d'accès à la grille flottante pourrait permettre d'utiliser les deux mécanismes de manière semi-indépendante, augmentant ainsi la densité de stockage et améliorant les performances en tirant parti des points forts de chacun. Comme abordé dans la section 3.5, cela permettrait également d'améliorer les caractéristiques de programmation, en abaissant les niveaux de tension nécessaires, et d'améliorer leur endurance d'écriture.

Des gains de densité supplémentaires peuvent également être obtenus grâce à des structures plus denses, comme dans le cas des cellules de NAND flash.

Chapitre 5

Exploration et optimisation de l'espace de conception

Contents

Pour explorer les possibilités offertes par les circuits à FeFET et les architectures LiM d'une manière rapide et efficace, une étude plus approfondie que ce que ne permettent les prototypes actuels est nécessaire. Cela comprend les caractéristiques des dispositifs, ainsi que la complexité des architectures. C'est pourquoi une approche de simulation a été adoptée, dans le cadre de laquelle une plateforme d'évaluation des performances a été conçue pour être utilisée en DSE des circuits 1T-1C et FeFET.

L'objectif est de cibler une approche co-optimisation circuit et technologie (DTCO, Design-Technology Co-Optimization) afin de permettre d'obtenir des mesures de performance au niveau du système depuis les valeurs mesurée au niveau des dispositifs, tout en considérant les particularités architecturales.

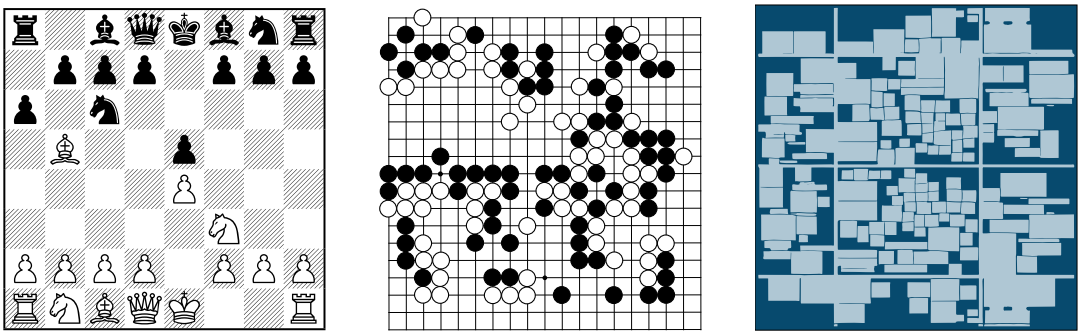
Les objectifs de cette exploration sont :

- de comprendre comment les approches de calcul normalement-éteint peuvent bénéficier des tableaux de mémoire ferroélectrique par rapport à d'autres technologies de mémoire non volatile, et évaluer l'impact des paramètres du système (cycles d'utilisation, complexité de la logique intégrée, activité de la source du capteur, etc.) ainsi que celui des dispositifs (endurance, puissance/temps de lecture/écriture, etc.) sur la performance globale.
- d'explorer les applications LiM à gros grain, y compris la mise en œuvre de fonctions logiques de bas niveau (additionneur complet), une bibliothèque de fonctions de traitement du signal optimisées pour la vitesse et l'efficacité énergétique, et la mise en œuvre d'une application complète de prétraitement d'images.
- d'explorer les concepts LiM à grain fin par le développement d'une bibliothèque de portes logiques non volatiles, pouvant faire l'objet d'une évaluation comparative complète des performances par rapport à d'autres technologies de dispositifs non volatiles.

5.1 Introduction à l'exploration d'espace de conception

La majorité des dispositifs utilisés dans les circuits nécessitent des modèles complexes pour obtenir des résultats de simulation réalistes. Ces modèles comprennent de multiples degrés de liberté liés à leur nombre de paramètres. Cela se retrouve particulièrement dans le cas de modèles ferroélectriques, en raison de leur comportement hystérétique. L'exploration exhaustive de l'espace de paramètres devient ainsi infaisable lorsqu'une grande quantité de dispositifs interagit au sein d'un circuit plus complexe. Il s'agit cependant de ces interactions qui apportent sa fonctionnalité au circuit.

Certains paramètres sont contrôlés par le concepteur, tandis que d'autres sont sujets à la variabilité, d'une puce à l'autre, ou au sein de la même puce. Les outils de DSE visent à faciliter la tâche de choix des paramètres en automatisant la sélection et la simulation d'ensembles de paramètres, afin d'obtenir les résultats de performance correspondants. Une exploration exhaustive de cet espace de paramètres et des résultats associés, qui forment l'espace de conception, est impossible en raison de l'explosion combinatoire, comme l'illustre la figure 5.1 : alors que le nombre de dispositifs augmente, le nombre de combinaisons de paramètres possibles s'accroît de manière polynomiale, de même que les possibilités d'interconnexion [Poi22, p. 73]. Les outils de DSE doivent donc utiliser des heuristiques pour explorer un sous-ensemble de l'espace de conception qui soit le plus utile possible au concepteur. En disposant de ces points de données, les concepteurs peuvent faire des choix plus optimaux, explorer les compromis et mieux prédire le comportement du système.



(a) Échiquier. États : $\approx 10^{123}$ (b) Plateau de go. États : $\approx 10^{360}$ (c) Placement de blocs de CI. États : $\approx 10^{90\,000}$

FIG. 5.1 : Illustration de l'augmentation de la complexité due à l'explosion combinatoire[Syn21].

5.1.1 Espace des paramètres et espace des performances, optimal de Pareto

Espace de paramètres et de performance

Afin de caractériser le comportement d'un circuit, d'un dispositif ou d'un système, ses propriétés doivent être déterminées. Certaines peuvent être contrôlées par le concepteur, comme la géométrie des transistors, ou celles-ci peuvent correspondre à une caractéristique intrinsèque du matériau ou du processus de fabrication. Une certaine variabilité se produit inévitablement lors de la fabrication et peut également être considérée.

L'objectif d'un concepteur est de choisir un ensemble de paramètres contrôlés qui permettront à l'appareil de fonctionner malgré d'éventuels facteurs externes, tout en maximisant les performances. Les propriétés contrôlées par le concepteur forment l'espace des paramètres $\mathcal{X}$. La gamme des performances réalisables constitue l'espace de performance $\mathcal{F}$. Cela est illustré sur la figure 5.2.

Un ensemble de paramètres x correspond à un ensemble de performances f. Il est important que le concepteur définisse des *métriques de performance*, valeurs numériques utilisées pour quantifier et comparer les divers critères de performance. La recherche des paramètres maximisant les performances peut alors être exprimée comme un problème d'optimisation visant à minimiser ces valeurs numériques.

La traduction des paramètres en métriques s'effectue par le biais de la fonction de transfert $F : \mathcal{X} \mapsto \mathcal{F}$. Cette fonction représente le comportement du dispositif fabriqué, ou son approximation à l'aide d'un modèle de simulation.

Optimisation multi-objectifs

L'optimisation multi-objectifs est généralement exprimée comme suit :

$$\min f = F(x) = \begin{cases} f_1(x) \\ \vdots \\ f_n(x) \end{cases} \quad \text{contraint par} \begin{cases} c_{eq}(x) = 0 \\ c_{ineq}(x) \leqslant 0 \end{cases} \tag{5.1}$$

où c_{eq} et c_{ineq} représentent toutes les contraintes de l'application, formulées sous forme d'égalités et d'inégalités, x est un ensemble de paramètres et f_i est l'une des n mesures de performance. La variabilité et les paramètres intrinsèques peuvent être rattachés à l'espace des paramètres $\mathcal{X}$, intégrés dans la fonction de transfert F, ou considérés comme faisant partie des contraintes.

Front de Pareto

Un ensemble de paramètres x_1 est dit dominant un autre ensemble x_2 si le premier est au moins aussi performant dans l'espace de performance, et strictement meilleur pour au moins une métrique de performance. En termes mathématiques[BG15] :

$$f_i(x_1) \leq f_i(x_2) \forall i \in \{1, \cdots, m\} \qquad \text{et} \qquad \exists j \in \{1, \cdots, m\} \, | f_j(x_1) < f_j(x_2) \tag{5.2}$$

Le processus de conception standard peut être amélioré en observant que les solutions de l'équation 5.1.1 conduisent à des configurations dites optimales de Pareto, où l'amélioration d'une performance ne peut se faire qu'au détriment d'une autre : celles-ci sont non dominées, selon l'équation 5.2. L'ensemble de ces points optimaux est appelé front de Pareto dans l'espace des performances ($\mathcal{F}$), tandis que les points correspondants dans l'espace des paramètres ($\mathcal{X}$) constituent l'ensemble de Pareto. Les deux ensembles sont liés par la fonction de transfert $F(x)$ modélisant le système, comme illustré dans la figure 5.2.

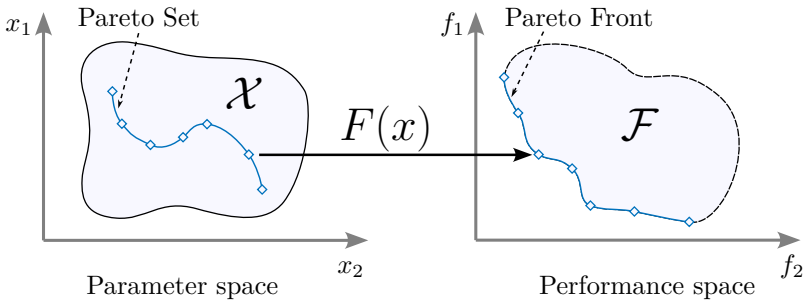
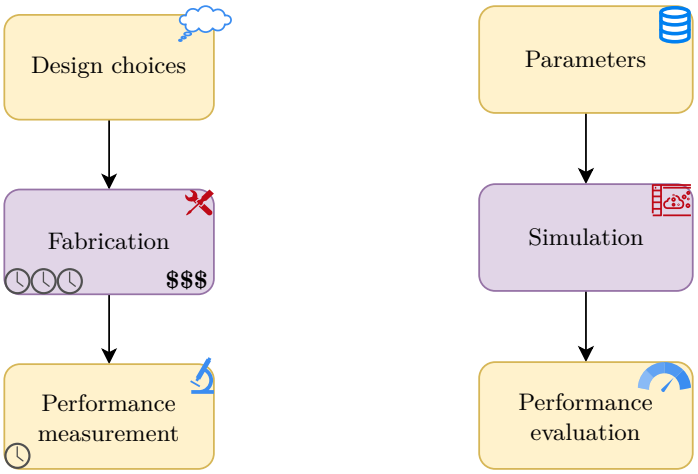


FIG. 5.2 : Espace des paramètres $\mathcal{X}$ et espace des performances $\mathcal{F}$, reliés par la fonction de modélisation F. L'objectif étant de minimiser les métriques de performance f_1 et f_2, le front de Pareto est l'ensemble des valeurs optimales (non dominées) dans $\mathcal{F}$, et l'ensemble de Pareto est l'ensemble des valeurs associées dans $\mathcal{X}$.

Les configurations optimales de Pareto correspondent à des doublets paramètres/résultats de performances $\{f, x\}$. En cas d'objectifs non contradictoires, la solution optimale peut être la même pour plusieurs métriques de performance. Dans le cas extrême où aucun des objectifs n'est contradictoire, l'ensemble et le front de Pareto peuvent être réduits à un seul point[Deb01, p. 23].



(a) *Caractérisation* expérimentale des performances

(b) *Estimation* des performances par simulation

FIG. 5.3 : Approches de mesure de performances par voie expérimentale et en simulation. Les approches de simulation sont plus rapides et nécessitent moins de capitaux, permettant des cycles d'itération beaucoup plus rapides. Cependant, celles-ci sont moins fiables.

5.1.2 Exploration automatisée

Une exploration exhaustive de l'espace des paramètres est généralement irréalisable, et une sélection manuelle des valeurs intéressantes pour les paramètres est complexe à obtenir en raison du nombre de simulations nécessaires pour les identifier. Des outils automatisés de DSE sont donc utilisés. D'un point de vue haut niveau, ces outils sont assez simples : à l'aide de métriques pour évaluer les performances du circuit, et en partant d'un ensemble de paramètres pouvant être ajustés dans des limites spécifiées, l'objectif est de tirer le meilleur parti du temps de simulation disponible, en convergeant rapidement vers des optimums de Pareto. Pour ce faire, des algorithmes d'optimisation multi-objectifs et des heuristiques sont utilisés pour sélectionner des populations de paramètres.

Procédure

Un circuit doit être spécifié comme un problème d'optimisation, en définissant explicitement les métriques devant être maximisées et minimisées, ainsi que les contraintes rendant une solution faisable ou utilisable. Un algorithme d'optimisation multi-objectifs est ensuite appliqué au problème, afin de sélectionner les points d'intérêt pour la prochaine série de simulations. Compte tenu du résultat de ces simulations, le prochain lot de paramètres peut être identifié.

Le processus est semblable à une boucle de conception à itérations manuelles telle qu'illustrée dans sur la figure 5.3 : le concepteur sélectionne les paramètres, évalue les performances expérimentalement ou en simulation, et utilise le résultat pour affiner les paramètres de l'évaluation suivante. Les flots de conception par simulation sont plus facilement automatisables, et les itérations peuvent être effectuées automatiquement jusqu'à ce qu'un ensemble optimal de paramètres ait été trouvé.

Comme illustré dans la figure 5.4, dans le cas de DSE assisté par des outils, le concepteur n'intervient plus dans la boucle de conception, car il limiterait la vitesse d'itération. L'objectif d'un tel processus est plutôt d'identifier pour le concepteur des points d'intérêt dans l'espace de paramètres : les algorithmes d'optimisation multi-objectifs visent à identifier pour chaque métrique un ensemble de paramètres optimaux, et permettre au concepteur de choisir le compromis final en fonction des spécifications. Par exemple, la latence peut être sacrifiée au profit de l'efficacité énergétique. Les paramètres résultant du processus d'optimisation peuvent être exploités directement ou guider l'amélioration des circuits conçus. Idéalement, les ensembles de paramètres obtenus doivent appartenir à l'ensemble de Pareto, informant ainsi le concepteur que le compromis choisi maximise tous les critères de performance.

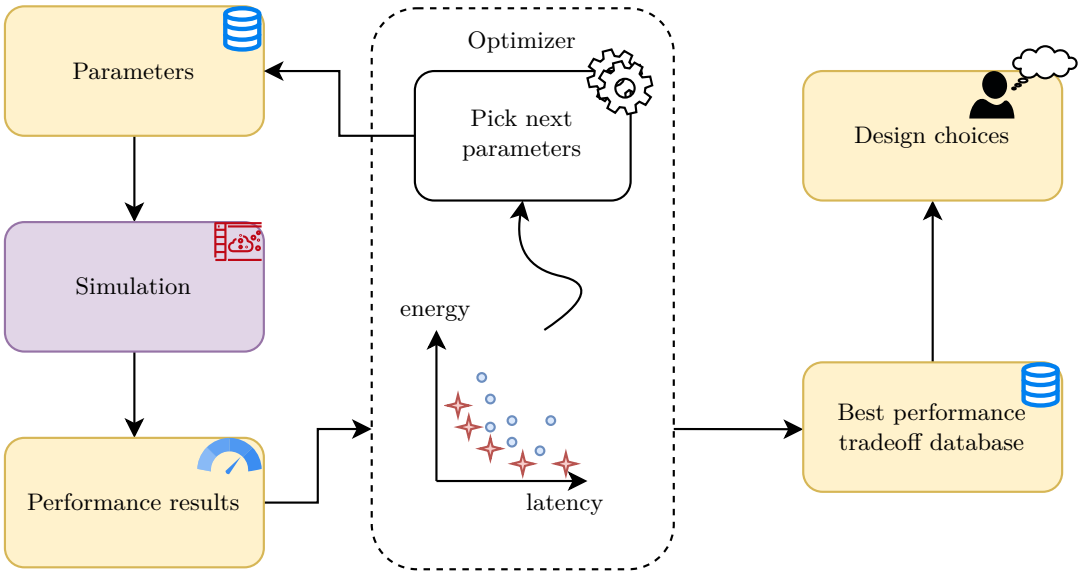


FIG. 5.4 : DSE automatisée

Une exploration exhaustive est cependant nécessaire pour prouver que les points ne sont pas dominés par d'autres solutions. Il est toutefois possible de sélectionner les solutions les plus optimales rencontrées lors de l'exploration de l'espace de conception, permettant ainsi de maintenir la complexité à un niveau atteignable, et de fournir des solutions aux performances adéquates et rerésentatives de l'espace de performance[Dup22, p. 56]. Cela nécessite un algorithme d'optimisation multi-objectifs capable de sortir des minimums locaux, explorant l'espace de performance de manière efficace, et un nombre suffisant d'itérations.

5.1.3 Étude des performances au niveau système

Prédire l'impact des circuits nouvellement conçus sur la performance du système est un exercice difficile, car ces circuits ne représentent qu'une partie réduite de l'ensemble du système. À ce niveau, les performances dépendent fortement de la charge de travail des circuits : différents algorithmes solliciteront différemment le circuit. En outre, de nouveaux circuits permettent des architectures radicalement différentes et innovantes, devant être comparées aux architectures existantes, mieux connues et plus optimisées.

Pour quantifier les gains pouvant être attendus de diverses architectures de circuits et de leurs possibles applications, une plate-forme polyvalente d'évaluation des performances au niveau du système a été élaborée, afin d'éclaircir l'impact des choix au niveau du dispositif ou de l'architecture sur la performance.

5.2 Outils d'exploration de l'espace de conception

5.2.1 Optimiseur LIFT

La structure utilisée pour générer le front de Pareto au niveau du circuit présentée sur la figure 5.5 a été mise en œuvre à l'aide d'un outil d'optimisation interne à l'ECL–INL[Fra12 ; Bri21]. Écrit dans le langage de programmation MATLAB, cet outil baptisé LIFT s'appuie sur plusieurs algorithmes d'optimisation pour produire un ensemble de paramètres depuis les résultats des simulations précédentes.

Les prérequis sont les suivants :

1. Spécification des paramètres et de leurs plages
2. Spécification des métriques, si celles-ci doivent être maximisées ou minimisées
3. Spécification des contraintes opérationnelles permettant de valider la fonctionnalité d'un circuit

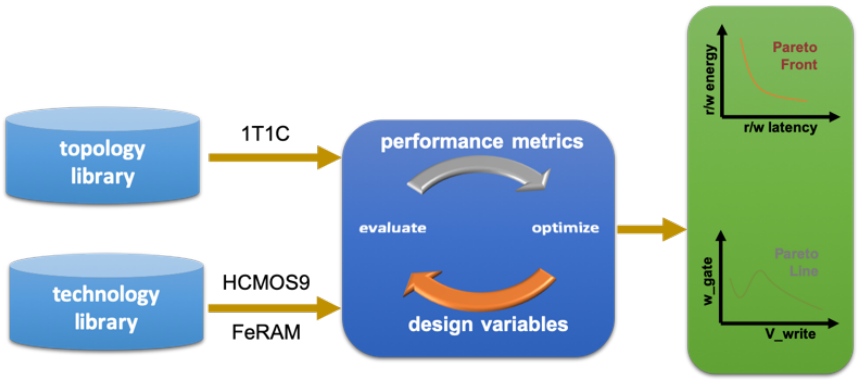


FIG. 5.5 : Flot de génération de l'ensemble et du front de Pareto

4. Exécution de simulations par l'outil et mesure des performances du circuit.

Compte tenu des conditions susmentionnées, deux points problématiques ont été relevés :

- La stabilité du système
- La communication avec le simulateur (point 4 ci-dessus)

Le premier point correspond à une limitation technique de l'outil d'exploration, ainsi que des circuits conçus et des modèles utilisés. Compte tenu de l'éventail des paramètres possibles, certaines combinaisons entraînent des simulations défectueuses, mettant en évidence des comportements inattendus au sein des modèles ou des problèmes de convergence. Les problèmes peuvent également être plus subtils, avec des scores anormalement élevés sur certaines métriques, faussant les résultats du processus d'exploration.

Plus généralement, le processus pourrait être fortement amélioré grâce à une base de données des simulations précédentes, utilisée pour initialiser l'espace de conception avec les résultats précédents plutôt que recommencer complètement l'exploration à chaque nouvelle campagne de simulation. Une fonctionnalité permettant au concepteur de spécifier des zones ne devant pas être explorées permettrait de contourner les zones d'instabilité du modèle ; les contraintes opérationnelles peuvent remplir ce rôle, mais leur spécification est complexe.

Le second point est détaillé dans la sous-section 5.2.2 ci-dessous.

5.2.2 IPC Cadence

L'outil d'exploration de l'espace de conception spécifie le prochain ensemble de paramètres à simuler lors du processus d'exploration de l'espace de conception. Ces paramètres doivent être communiqués au simulateur. Cependant, le langage de script OCEAN® du simulateur est conçu autour de l'hypothèse inverse, où celui-ci contrôle le flot d'optimisation. Par conséquent, une interface de communication inter-processus (IPC, Inter-Process Communication) a été conçue pour contrôler le simulateur depuis l'optimiseur.

Comme illustré sur la figure 5.6, un processus du simulateur OCEAN est d'abord lancé, puis exécute des scripts d'initialisation. Ces scripts ouvrent des descripteurs de fichiers et se mettent en attente de réception d'instructions par ces canaux.

L'optimiseur même est implémenté en langage MATLAB, qui a des capacités limitées pour la communication inter-processus, telle l'absence d'opérations asynchrones ou non bloquantes sur des descripteurs de fichiers. L'ouverture et la fermeture de ceux-ci bloquant le processus de contrôle, l'IPC réécrite en Python, ce langage interopérant facilement avec MATLAB.

5.3 Résultats de l'exploration de l'espace de conception

Au niveau du circuit, l'objectif est de générer des données et des modèles mathématiques exprimant les compromis entre les métriques de performance, et d'extraire les paramètres de conception associés. Ces données peuvent ensuite être utilisées pour une évaluation des performances au niveau du système, comme détaillé dans la section 5.4. Dans cette section,

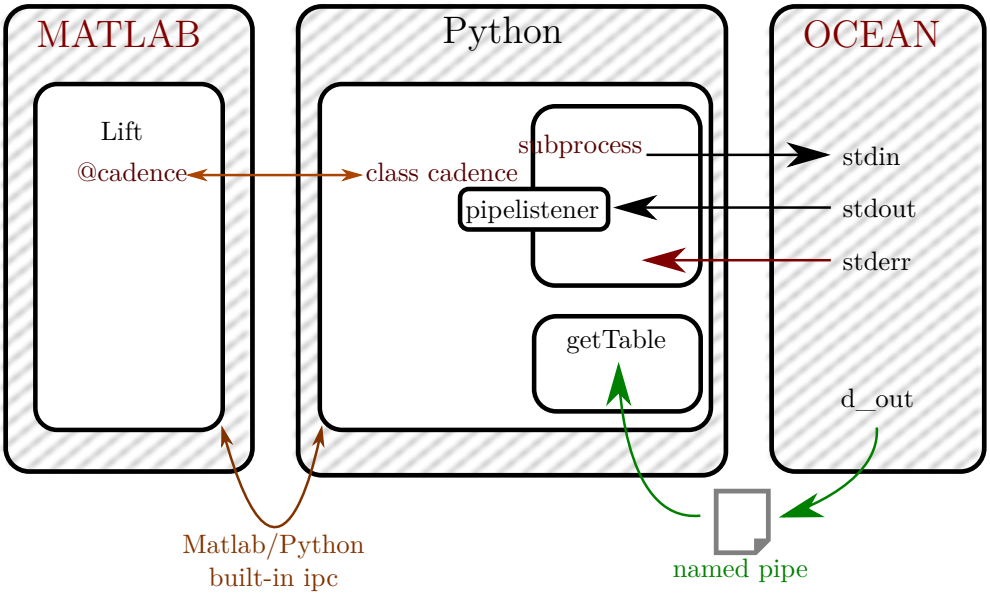


FIG. 5.6 : Architecture de l'IPC Cadence.

des cellules 1T-1C à FeCap telles que décrites dans la section 3.2 sont étudiées, ainsi que les portes logiques non volatiles à FeFET présentées dans la section 4.4.

En première approche, des cartes modèles représentant des mesures de performance typiques sont extraites pour alimenter le le flot d'évaluation au niveau système.

Les résultats de la campagne de DSE automatisée ont été décevants, en partie à cause de l'instabilité du modèle ferroélectrique, qui a souvent entraîné des interruptions du simulateur, ou des signaux de sortie oscillatoires ayant perturbé le script d'extraction de résultats des métriques (s'agissant d'une situation familièrement connue sous le nom de « Garbage In ; Garbage Out »). Cela a entraîné des mesures de performance complètement faussées, sabotant ainsi l'exploration automatisé. Il est notable que les algorithmes d'explorations soient facilement capables de converger vers ces points de fonctionnement problématiques, pouvant en faire un outil adapté au développement, test et à la validation des modèles.

En attendant le développement de modèles améliorés, il a été décidé de procéder dans un premier temps à une DSE manuelle, en explorant une grille de combinaisons selon plusieurs dimensions. Cette méthode donne une vue d'ensemble de l'espace de conception et permet de confirmer les tendances générales et les influences des paramètres sur la performance. En revanche, elle est beaucoup plus proche du processus manuel de DSE présenté dans la figure 5.3b ; ce qui rendant fastidieuse l'extraction des ensembles de Pareto. En outre, certains points de l'espace de conception sont redondants et, compte tenu de la dimensionnalité élevée du problème, l'explosion combinatoire limite le processus d'exploration à quelques points par métrique de performance.

5.3.1 Échantillonnage de l'espace de conception de la bitcell 1T1C

Description du problème et résultats attendus

Les performances des mémoires sont évaluées selon trois critères principaux : leur capacité, leur consommation d'énergie et leur rapidité. Le premier critère est directement lié à l'empreinte de bitcell et des circuits associés : plus la bitcell est petite, plus la mémoire est dense et plus la capacité de stockage est élevée (hormis les considérations relatives aux MLC). La consommation d'énergie dépend généralement de la capacité du circuit et devrait donc diminuer proportionnellement avec la surface du condensateur ferroélectrique. De même, lorsque la surface du condensateur décroît, son temps de charge est réduit, ce qui entraîne une augmentation de la vitesse de fonctionnement et une diminution de la latence.

Ces caractéristiques devraient donc s'améliorer avec la réduction de la surface du condensateur. Toutefois, cette réduction diminue la différence observable entre la lecture d'un état logique

	Efficacité énergétique	Rapidité	Densité	Fenêtre Mémoire
tW	−	++	−	=
tL	−	−	−	=
Ac	−−	−−	− − −	++

Tab. 5.1 : Impact prédit des paramètres utilisés pour le balayage sur la performance de la cellule bitcell 1T1C

haut et d'un état logique bas (fenêtre mémoire), imposant davantage de contraintes aux circuits périphériques, notamment aux mécanismes d'adressage et de décodage.

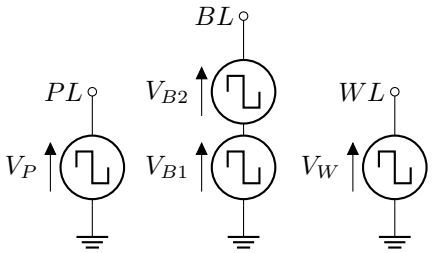
Outre la surface du condensateur, il est possible de modifier la géométrie du transistor d'accès : un transistor plus large augmente le courant, accroissant simultanément la vitesse et la consommation d'énergie. Un transistor plus long diminue le courant de fuite ainsi que la vitesse de fonctionnement, et nécessite plus d'énergie pour être piloté. Le courant de fuite n'est pas une métrique significative dans ce cas, car, contrairement à la DRAM, le condensateur n'a pas besoin de conserver de charge ; le transistor d'accès sert simplement à ouvrir le circuit lors de l'accès aux cellules voisines.

L'impact de ces paramètres sur la fenêtre de mémoire est plus difficile à estimer. Comme indiqué plus loin dans la section 5.3.1, il a été choisi d'intégrer le courant de la BL pour mesurer la fenêtre de mémoire, la liant à l'énergie de la BL. Les prédictions sont résumées dans la tableau 5.1.

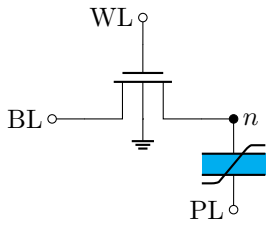
Comme exprimé dans les paragraphes ci-dessus, ce problème est suffisamment simple pour permettre d'anticiper la plupart des résultats, tout en mettant en jeu suffisamment de paramètres pour que leurs interactions ne soient pas triviales. Il s'agit donc d'un cas d'étude intéressant pour mettre au point un système de DSE pour les circuits ferroélectriques, tout en générant des résultats pertinents pour l'évaluation des performances au niveau du système.

Circuit de test

Le circuit de test est une bitcell isolée, telle qu'illustrée sur le Circuit 5.1b, fonctionnant comme précédemment décrit dans la sous-section 3.2.1. Cette exploration utilise un modèle de FeCap Preisach fourni par NaMLab.



(a) Tensions d'alimentation.



(b) Circuit 1T1C de test.

Circuit 5.1 : Bitcell 1T1C avec générateurs de tension. Les tensions fournies à la bitcell mesurée sont PL, WL, BL et gnd.

Les tensions d'entrée sont générées par plusieurs générateurs de signaux carrés, comme indiqué sur le Circuit 5.1a. Fonctionnant à des fréquences différentes, ceux-ci rendent la forme d'onde d'entrée périodique, permettant ainsi de faire subir plusieurs de cycles au ferroélectrique avant d'effectuer les mesures. Les signaux résultants sont représentés dans la figure 5.7.

Les données d'entrée choisies font subir deux cycles complets de repolarisation au condensateur ferroélectrique. Plus concrètement, le banc d'essai exécute deux fois la séquence d'opérations suivante :

1. Écriture de « 0 » dans le ferroélectrique
2. Lecture du ferroélectrique en écrivant « 1 » (courant de commutation détecté).

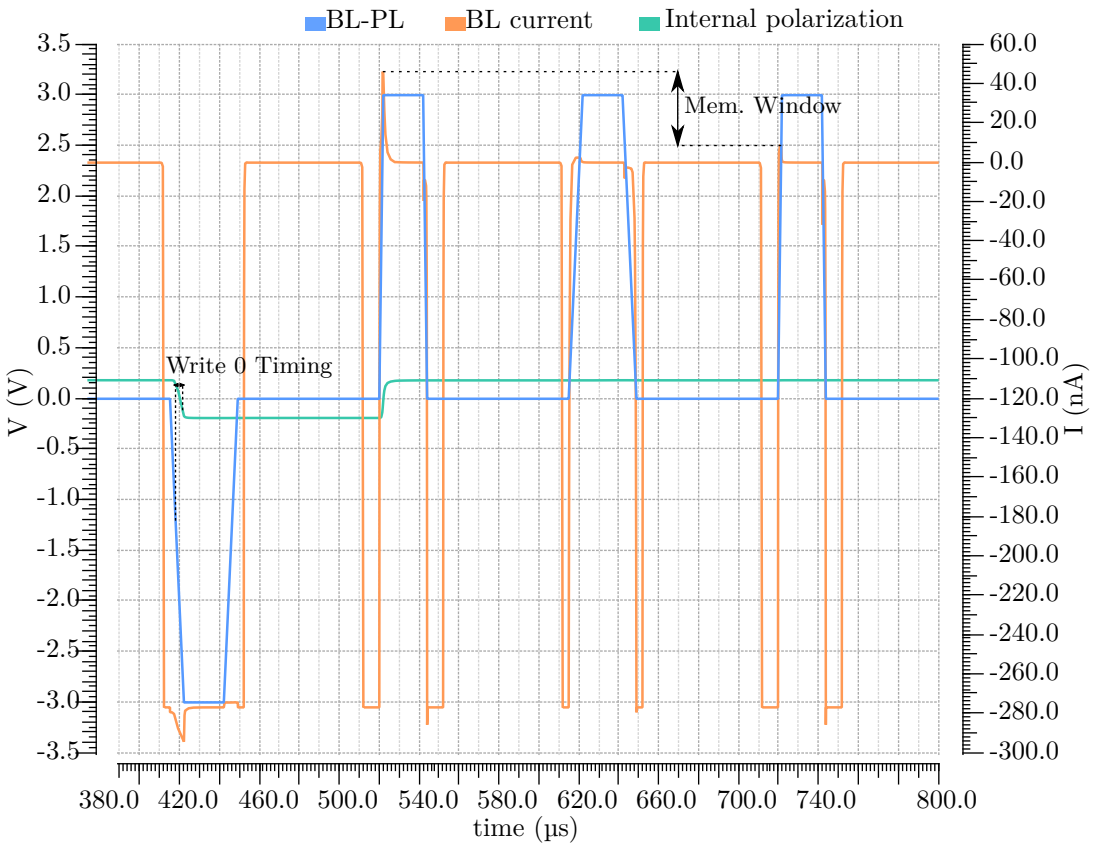


FIG. 5.7 : Formes d'ondes de test pour DSE de la cellule 1T1C. La polarisation interne ● est surveillée, et le temps est mesuré à partir de l'instant où la différence de potentiel $BL-PL$ ● à travers le FeCap dépasse V_C, jusqu'à ce que la polarisation interne Pr atteigne 50 %. La fenêtre mémoire du courant de crête est définie comme la différence entre le courant de crête de BL current ● après la lecture d'un 0 ou d'un 1.
Les pics de courant visibles sur ce diagramme correspondent au courant de repolarisation.

3. Écriture de « 1 » dans le ferroélectrique
4. Lecture en écrivant « 1 » (pas de courant de commutation)

Le premier cycle permet de s'assurer que le modèle est correctement initialisé et rend le second cycle plus représentatif du fonctionnement en régime permanent. Les mesures de performance sont effectuées lors du second cycle. La BL comporte deux générateurs en série, ce qui permet de spécifier différentes caractéristiques pour les signaux de lecture et de programmation. La dernière étape peut être considérée comme redondante, mais celle-ci est représentative des opérations réelles, et fournit un point de données supplémentaire.

Les paramètres sont indiqués dans la tableau 5.2. Comme on indiqué dans ce tableau, les impulsions de tension sont relativement longues, et les signaux de sortie sont observés pour déterminer la vitesse de fonctionnement théorique, plutôt que de raccourcir la durée des signaux d'entrée.

Afin d'améliorer la mesure de la vitesse de commutation ainsi que l'état du matériau ferroélectrique, le modèle est modifié pour exposer la polarisation interne Pr sous forme d'une source de tension exposée sur une borne ajoutée au symbole du condensateur, visible sur le Circuit B.3. Le signal de sortie est représenté sur la figure 5.7, et permet d'observer l'état de l'oxyde ferroélectrique directement, sans reposer sur des indicateurs indirects.

Définition du problème

La géométrie de la bitcell est entièrement paramétrée, comme indiqué dans les tables 5.2 et 5.3, afin de permettre l'étude de l'impact des choix de conception. Les paramètres technologiques

tels que Pr et la fraction paraélectrique, peuvent également être paramétrés pour caractériser l'impact de la variabilité sur les performances de la bitcell, bien que cela n'ait pas été l'objet de la présente étude.

L'espace de conception a été échantillonné avec le script présenté dans l'extrait de code A.3, pour chaque combinaison de paramètre dynamique énumérée dans la tableau 5.4 : cette exploration s'est principalement concentrée sur la géométrie, en gardant les autres paramètres constants.

Extraction des métriques

Les mesures sont extraites des formes d'ondes obtenues par simulation. Plus précisément, le script de l'extrait de code A.1 mesure les quantités suivantes :

- Surface du condensateur (paramètre)
- Largeur du transistor d'accès (paramètre)
- Longueur du transistor d'accès
- Temps d'écriture, tel que mesuré sur la figure 5.7
- Fenêtre mémoire, en quantité de charges
- Énergie, pour l'écriture et la lecture[1]
- Courants de crête, comme représenté sur la figure 5.7

La figure 5.7 montre la différence de courant de crête, mais une autre façon d'évaluer la fenêtre mémoire est d'intégrer la valeur du courant sur le temps de simulation, afin d'extraire la quantité de charges transférées. Le déséquilibre de charge dû au courant de repolarisation peut être mesuré ainsi, et cette valeur peut être utilisée pour détecter la valeur mémorisée.

Il est important de noter qu'il s'agit d'un scénario « optimal », la quantité de charges étant comparée à une lecture de référence de la même cellule effectuée quelques microsecondes auparavant. Néanmoins, les améliorations relatives de cette fenêtre de mémoire devraient se refléter dans les cas réels d'utilisation du circuit.

Une approche similaire a été adoptée pour mesurer la vitesse de fonctionnement : la polarisation interne est observée, et le temps de commutation est mesuré après que le potentiel aux bornes de FeCap a franchi V_C. Plus précisément, la durée est mesurée lors de l'inversion de la polarisation depuis $V > V_C$ jusqu'à $P > \frac{1}{2}\mathrm{Pr}$. Bien qu'il ne s'agisse pas d'un indicateur direct de la rapidité de la mémoire, les lignes d'alimentation étant pilotées par des sources d'impédance non nulle, cette mesure fournit une estimation de la de vitesse maximale théorique de la bitcell. Il serait difficile d'obtenir des estimations plus précises de la vitesse et de la consommation d'énergie en simulant une unique bitcell. La mesure utilisée est jugée suffisante pour fournir un objectif d'optimisation et comparer plusieurs variantes de la même bitcell.

Résultats

Bien qu'il n'ait pas été possible d'automatiser complètement l'exploration, un aperçu de l'espace de conception a été obtenu en sélectionnant les points de simulation en amont, et en validant plusieurs étapes du pipeline de DSE automatisée. Les résultats de cette exploration sont représentés sur la figure 5.8, et le lien identifié entre les paramètres et les métriques de performance est rapporté dans la tableau 5.5. Le temps d'écriture est affecté comme prévu, et s'améliore à mesure que la largeur du transistor augmente et que la longueur du transistor diminue : cela suggère que la vitesse d'écriture est limitée par le courant ou, du moins, que des courants plus élevés permettent des écritures plus rapides, jusqu'à un certain point. Avec une surface de condensateur de $9.0 \times 10^4\,\mathrm{nm}^2$ (correspondant à un diamètre de condensateur de 339 nm), le temps d'écriture est compris entre 88.9 ns et 117 ns, permettant des fréquences de fonctionnement théoriques de 8 MHz à 11 MHz. Il est intéressant de noter que l'écriture avec une impulsion positive d'un FeCap précédemment polarisé négativement conduit à des temps

[1] Énergie pour écrire un 0 sur un 1 (ew_10), et pour lire les états 0 et 1 (er1_w0 et er1_w1 respectivement) avec une impulsion « écriture 1 ».

Paramètre	VP	VB2	VB1	VW
Voltage1	0			
Voltage2	vprog		vread	vwl
Delay	20 µs	220 µs	120 µs	10 µs
Temps de montée	triseprogPL	triseprogBL	trisereadBL	triseWL
Temps de descente	tfallprogPL	tfallprogBL	tfallreadBL	tfallWL
Largeur d'impulsion	20 µs			40 µs
Periode	400 µs		200 µs	100 µs

TAB. 5.2 : Paramètres pour les générateurs de tension. Certaines valeurs ne sont pas spécifiées numériquement, car il s'agit de variables pour la simulation paramétrique.

Transistor	
Longueur	L
Largeur	W
Condensateur	
Superficie	atot_fe
Fraction paraélectrique	0.3
ε_r ferroelectic	30
ε_r dielectric	17
Tfe	10 ns
Psat	2 µC cm^{-2}
Pr	19 µC cm^{-2}
Vc	1.5 V
τ ferroélectrique	10 ns
τ_f RC	50 ns
τ paraélectrique	1 ns
Résistance de fuite	1 MΩ

TAB. 5.3 : Paramètres des transistors et des condensateurs pour la simulation. Les valeurs numériques sont extraites de mesures typiques pour cette technologie, tandis que la géométrie est paramétrisée pour permettre un balayage pendant la simulation.

Paramètres dynamiques			Paramètres statiques		
Superficie FeCap (nm^2)	TL (nm)	TW (nm)	Taux de montée	Tensions[a]	Pr
1.00×10^4	150	130	5 ns V^{-1}	3.6 V	0.19 µC cm^{-2}
2.25×10^4	170	140			
3.24×10^4	200	160			
9.00×10^4		180			

[a]Tension de programmation, lecture et de Word Line

TAB. 5.4 : Paramètres pour DSE de la bitcell 1T1C, dont des paramètres statiques (inchangés à chaque exécution) et dynamiques. Cette exploration couvre toutes les combinaisons possibles de l'ensemble des paramètres dynamiques, soit $4 \times 4 \times 3 = 48$ points de données.

simulés inférieurs, de 49 ns à 59 ns. Ces temps correspondent à la vitesse de commutation du FeCap même. Dans un circuit complet, les délais seront augmentés par la capacité du décodeur d'adresse, pour WL ainsi que pour BL, et par l'amplificateur de détection et l'CAN. Comme attendu, les valeurs mesurées ci-dessus sont du même ordre de grandeur que le paramètre τ_f RC du modèle, listé dans la tableau 5.3.

Paramètre	Efficacité énergétique	Rapidité	Densité	Fenêtre Mémoire
largeur de transistor	$-^a$	$+^b$	$-^a$	$+^a$
Longueur de transistor	$-^a$	$--^c$	$-^a$	$+^a$
Aire du condensateur	$=^d$	$=^d$	$---^e$	$=^d$

[a] Linéaire : coefficient de Pearson $\rho = \pm 1$; avec des pentes similaires pour la longueur et la largeur du transistor
[b] L'impact augmente avec la longueur des transistors
[c] L'impact diminue lorsque la longueur des transistors augmente
[d] Inattendu : un impact important était anticipé
[e] Les condensateurs étant plus grands que les transistors d'accès, leur contribution en termes d'aire est plus importante

Tab. 5.5 : Impact mesuré des paramètres de balayage sur les performances de la bitcell 1T1C, selon la figure 5.8 et la figure 5.9. Lorsque les résultats sont compares avec la tableau 5.1, l'impact de l'aire du condensateur est inattendu, de même que les résultats de la fenêtre mémoire.

La fenêtre mémoire mesurée est comprise entre 26.2 nC à 36.1 nC.

Une comparaison avec la DRAM, où un amplificateur de détection est capable de lire un condensateur de $C = 15\,\text{nF}$, donnant $Q = C \cdot V = 3\,\text{fC}$ à $V_{cc}/2 = 0.3\,\text{V}$, cela suggère que suffisamment de marge existe pour réduire la fenêtre mémoire au profit de la vitesse et de la surface, ceci jusqu'à six ordres de grandeur. Il s'agit d'une voie d'amélioration pour de futurs travaux.

Une fenêtre mémoire aussi large invite également à étudier les mémoires MLC, car la surface du condensateur ne peut pas être réduite en dessous d'une taille d'environ 100 nm de diamètre en raison de la taille des grains ferroélectriques, comme détaillé dans section 2.1.1.

Il est important de noter que les résultats ci-dessus sont issus de simulations ; ceux-ci peuvent être inexacts.

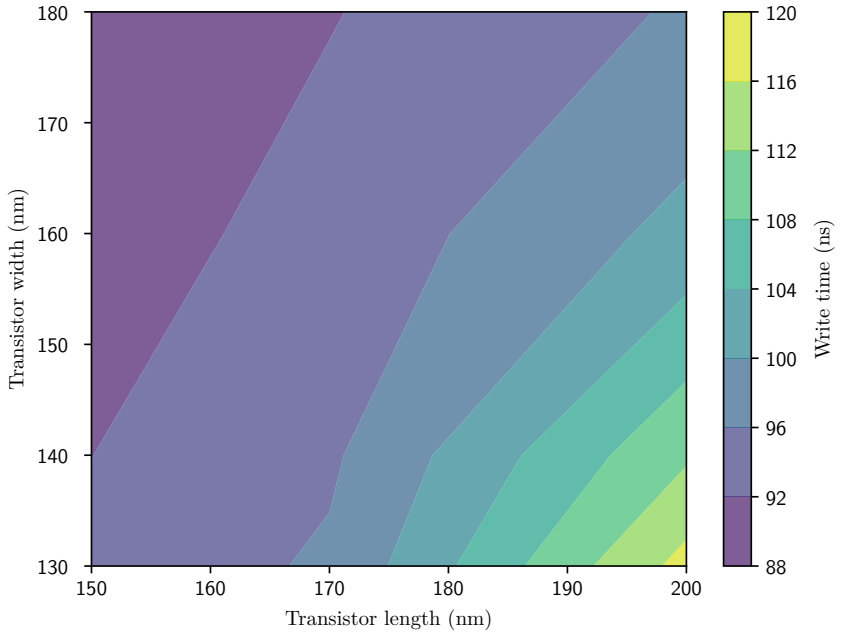
L'un des résultats inattendus en cours d'étude est illustré sur la figure 5.9 : une relation quasi-linéaire entre l'aire du condensateur et la fenêtre mémoire, ainsi qu'avec les temps d'écriture était attendue, particulièrement lorsque la rapidité est limitée par le courant. Dans la même exploration, la figure 5.9 montre que l'aire du FeCap n'a pratiquement aucun impact sur ces valeurs. Cela suggère que d'autres paramètres ont un impact beaucoup plus prononcé, ce qui était inattendu et reste à étudier. Le résultat peut être réaliste, ou peut indiquer un problème de modèle ou de configuration de la simulation, comprenant le script d'extraction des métriques. Il est intéressant de noter que, comme visible sur la figure 5.8b, la fenêtre mémoire s'est avérée fortement corrélée à la géométrie du transistor d'accès, ce qui peut être lié à l'impact réduit sur les performances de la zone FeCap, ainsi qu'au fait que la fenêtre mémoire est déterminée à partir de l'intégration du courant traversant le transistor d'accès. L'exploration sera répétée sur une plus large gamme de zones de condensateurs pour confirmer ces résultats.

5.3.2 Porte logique non volatile NAND à FeFET (NV-NAND2)

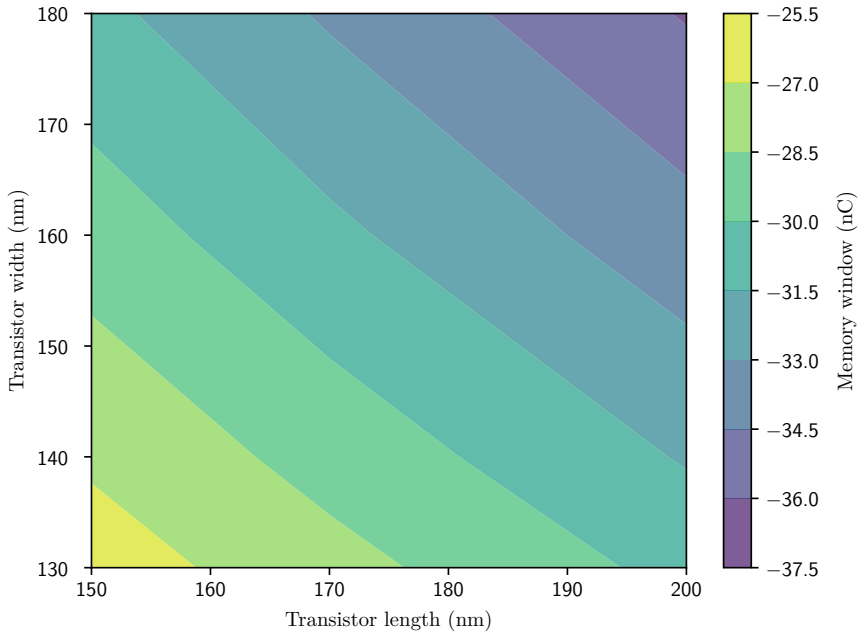
Des travaux préliminaires ont été entrepris pour l'étude d'une porte logique NAND à FeFET, telle que décrite dans la sous-section 4.4.1, avec la technologie 28SLP de GlobalFoundries et le même modèle Preisach fourni par NaMLab, utilisé comme FeCap supplémentaire dans l'empilement de la grille du transistor afin de modéliser un FeFET.

Circuit de test et espace de paramètres

Le problème d'optimisation a été défini conformément aux paramètres présentés dans la tableau 5.6, dans une configuration logique dynamique telle que présentée dans le Circuit 5.2.

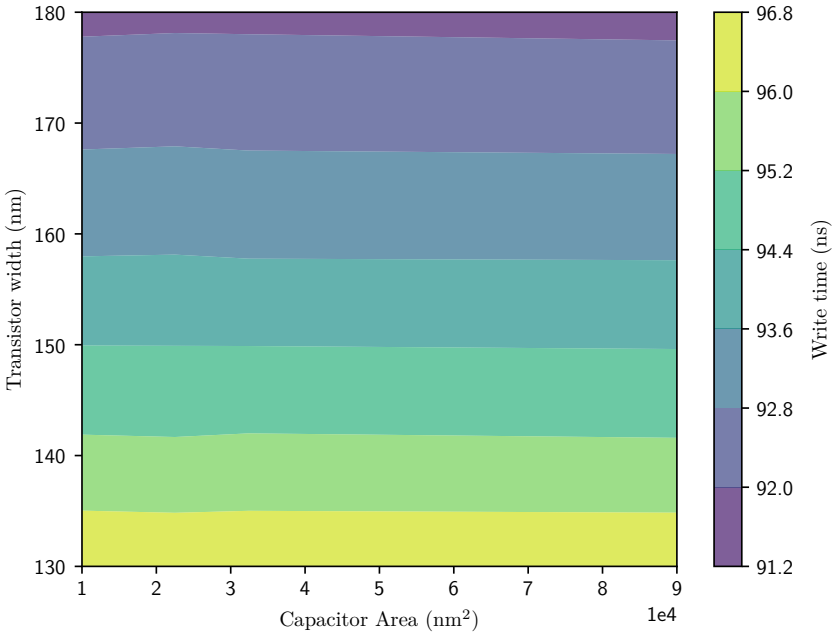


(a) Temps d'écriture pour une aire de condensateur de 9×10^4 nm^2

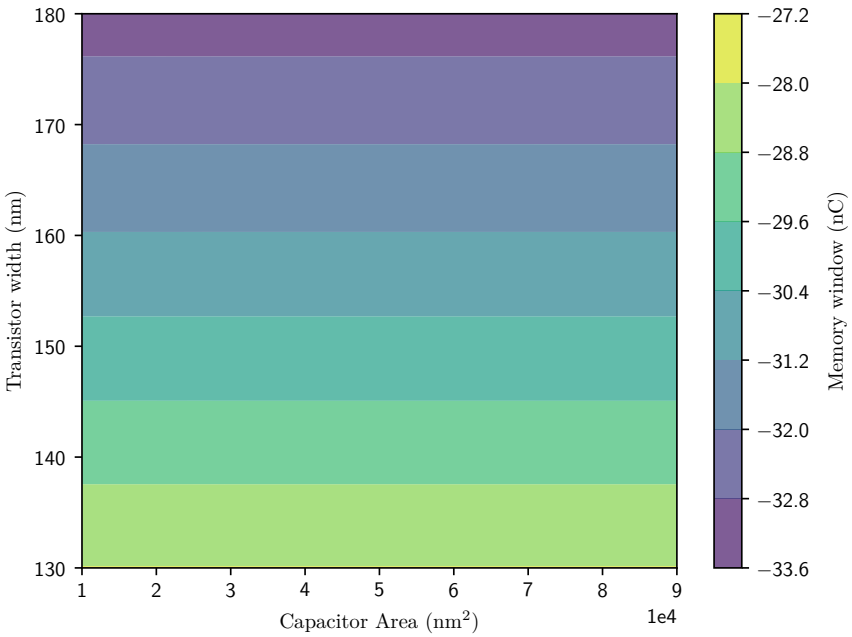


(b) Fenêtre mémoire pour une aire de condensateur de 9×10^4 nm^2

FIG. 5.8 : Résultats de DSE pour circuit 1T1C : temps et énergie vs géométrie du transistor. Une nette amélioration du temps d'écriture est constatée lorsque la largeur du transistor augmente, permettant d'utiliser des courants plus importants. De même, lorsque la longueur du transistor augmente, le courant diminue.
La fenêtre mémoire s'agrandit avec la largeur et la longueur de celui-ci. Sa forte corrélation (coefficient de Pearson $\rho = 1$) à la consommation d'énergie a également été constatée. Les plus petites dimensions des transistors se trouvent dans le coin inférieur gauche, représentant les valeurs idéales pour l'utilisation de la surface.

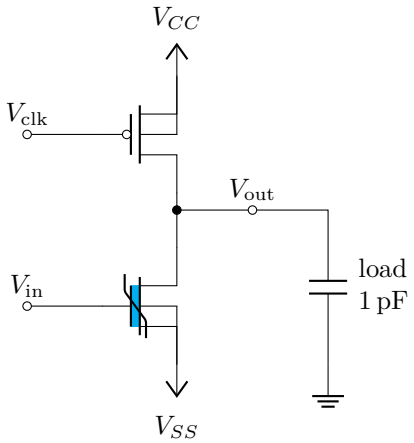


(a) Temps d'écriture pour une longueur de transistor de 170 nm



(b) Fenêtre mémoire pour une longueur de transistor de 170 nm

Fig. 5.9 : résultats de DSE pour l'exploration 1T1C : temps et énergie en fonction de l'aire du condensateur et de la largeur du transistor (à longueur constante), montrant une dépendance quasi inexistante avec l'aire (*Ce résultat a depuis été confirmé être dû à un problème de mesure, se référer à la présentation de la soutenance*). Les temps d'écriture indiqués ici correspondent à la repolarisation du ferroélectrique avec une impulsion négative ; les résultats de simulation montrent des temps réduits de moitié pour des impulsions positives.



Circuit 5.2 : Circuit NAND non-volatile dynamique à deux entrées (NV-NAND2) utilisé pour la DSE. V_{SS} est mis à la masse (0 V) pendant le fonctionnement normal, ou à la tension de programmation (choisie à $V = 5\,\text{V}$) pour programmer l'état « logique bas ». L'état « logique haut » est programmé en élevant V_{in} à la même tension de programmation.

Conception		NV-NAND	
technologie CMOS		GlobalFoundries 28SLP	
Modèle de ferroélectrique		Placé sur la grille du transistor	
Paramètre du circuit		Valeur minimum	Valeur maximum
Niveau logique/programmation	V	1.5 V/5 V	
Largeur de transistor	tW	100 nm	1500 nm
Longueur de transistor	tL	80 nm	1400 nm
Superficie condensateur ferroélectrique	Atot	=tW*tL	

Tab. 5.6 : Paramètres de DSE pour l'exploration des performances de la porte logique NAND non volatile.

Pour réduire la complexité de cette première exploration, les paramètres ont été limités à un nombre de deux : la largeur et la longueur du FeFET. L'espace de conception a d'abord été exploré selon une grille prédéterminée, visible sur la figure 5.10a. L'intégralité du code utilisé pour cette exploration est fourni dans l'extrait de code A.4, montrant l'utilisation de l'IPC présentée en sous-section 5.2.2. Les résultats préliminaires de cette DSE sont présentés sur la figure 5.10.

Espace de performance

Plusieurs critères ont été examinés :

- Fonctionnement de la porte logique : celle-ci est considérée comme opérationnelle si la sortie est conforme à la table logique d'une porte NAND, la tension de sortie étant supérieure à V_{th} uniquement lorsque les deux entrées sont au niveau logique bas.
- Énergie de « préchargement » : l'énergie nécessaire pour précharger le nœud flottant, qui varie en fonction des dimensions du transistor et dicte la consommation énergétique de la porte logique.
- Fenêtre de tension : différence de tension exploitable entre un niveau logique de sortie haut et bas, pour une utilisation en logique transistor-transistor.

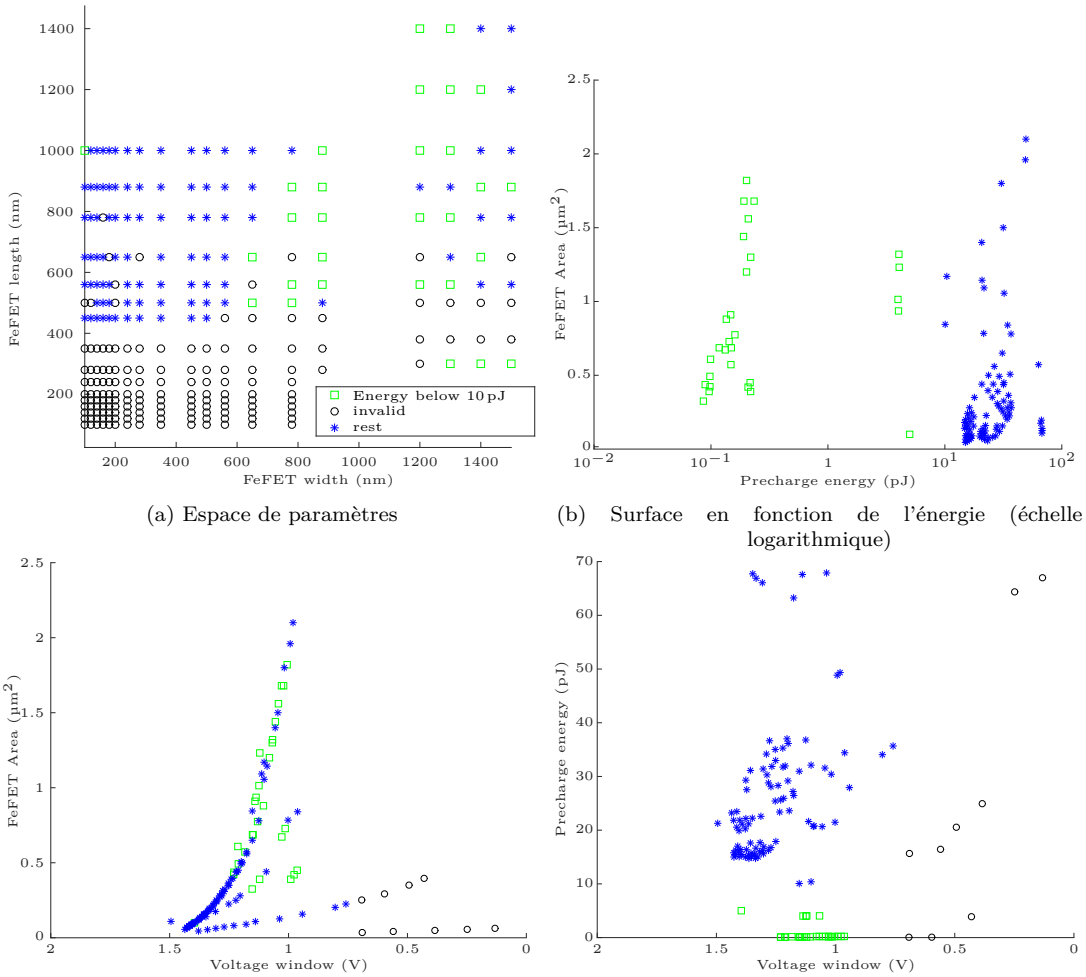
Le code utilisé pour l'extraction des métriques est présenté dans l'extrait de code A.2, écrit dans le langage de programmation SKILL.

Résultats

Même avec un ensemble réduit de paramètres et de points, quelques observations intéressantes peuvent être effectuées :

- L'énergie de préchargement semble être davantage affectée par la largeur du transistor ; cependant, les simulations indiquent que le coût énergétique diminue lorsque la largeur augmente, ce qui était inattendu
- La fenêtre de tension semble être plus optimale pour des dimensions réduites de FeFET ; ces objectifs peuvent être non contradictoires
- De fenêtres de tension plus étroites semblent conduire à des opérations non valides, ce qui est attendu

Cette exploration étant assez restreinte, celle-ci peut ne pas être représentative de la distribution réelle de l'espace de performance. Il est également important de noter que les données affichées sur ces graphiques dépendent fortement de la précision du script d'extraction des métriques, qui n'a pas fait l'objet d'un examen approfondi. Certains résultats sont intrigants et nécessitent une analyse plus approfondie pour être compris. Un problème aussi simple montre déjà une dispersion relativement large des points de fonctionnement dans l'espace de performance, justifiant l'approche DSE.



(a) Espace de paramètres

(b) Surface en fonction de l'énergie (échelle logarithmique)

(c) Surface en fonction de la fenêtre de tension

(d) Énergie en fonction de la fenêtre de tension

FIG. 5.10 : DSE préliminaire de la porte logique non volatile NAND à FeFET. 5.10a montre les points simulés dans l'espace des paramètres, tandis que les autres figures montrent les points correspondants dans l'espace des performances. Pour tous les graphiques de l'espace des performances, le point optimal est situé dans le coin inférieur gauche (faible surface, faible consommation et large fenêtre). Les carrés verts indiquent les correspondant à une énergie de précharge inférieure à 10 pJ. Les cercles noirs représentent les simulations menant à un circuit non fonctionnel, comme défini dans la section 5.3.2. 5.10c semble indiquer la présence d'objectifs non contradictoires.

5.4 Plate-forme d'analyse des performances au niveau système

5.4.1 Introduction

Objectifs

L'objectif principal de cette plateforme est d'évaluer les performances d'une architecture de calcul sur un ensemble de tests de référence, afin :

1. d'évaluer l'impact de changements au niveau du circuit sur les performances du système
2. de comparer les performances de différentes architectures pour la même application

Les métriques de performance peuvent inclure la rapidité, la latence, ou la consommation d'énergie.

Cela suggère une architecture de plate-forme modulaire, capable de simuler différentes architectures de calcul et définitions d'algorithmes de tests.

Cas d'utilisation

L'objectif de cet outil est de guider les choix du concepteur, et de fournir un retour quantitatif sur les gains potentiel liés au calcul normalement-éteint et LiM à gros grain, tout en permettant une exploration des paramètres système tels que le rapport cyclique de la mémoire et les stratégies d'extinction, en fonction de la charge de travail de l'application et des stratégies d'accès mémoire. Ce retour d'information permet également une cooptimisation au niveau du dispositif, les paramètres de conception tels que les dimensions des transistors étant guidés par des considérations de performance au niveau du système.

Remerciements

La conception et la mise en œuvre de ce travail ont été réalisées avec des contributions successives des étudiants de master Pierre-Etienne Polet, Luca Mozzone[Moz21] et du chercheur postdoctoral Marcello Traiola.

5.4.2 Champ d'application de la plateforme d'évaluation des performances

Architectures système cibles

La plateforme de simulation vise à combler le fossé entre les caractéristiques de performance au niveau du dispositif et les mesures de performance au niveau du système, tout en tenant compte des particularités architecturales. Par conséquent, celle-ci doit être suffisamment générique pour accommoder diverses architectures informatiques[OCo+18] : les architectures LiM où de petits circuits sont placés à proximité de la mémoire, et IMC où les éléments mémoire même effectuent les calculs. Différents sous-ensembles peuvent être identifiés parmi ces paradigmes : plus généralement, une distinction est faite entre les LiM à gros grain et à grain fin. Bien que les deux associent logique et mémoire, le premier utilise des tableaux de mémoire denses et de la logique périphérique pour le calcul, tandis que le second exploite des cellules mémoire plus petites intégrées au sein de circuits logiques, tels que des coefficients de filtragenon volatiles, comme illustré sur la figure 5.11. Par conséquent, la granularité fait référence à la taille des éléments mémoire intégrés à la logique.

Cette exploration s'est concentrée sur la LiM à gros grain en première approche.

Métriques extraites

Plusieurs technologies de mémoire doivent être évaluées : 1T-FeFET, structures 1T-1C semblables à de la DRAM avec des condensateurs ferroélectriques, ainsi que d'autres mémoires conventionnelles et émergentes afin de fournir un point de comparaison pour référence. Les valeurs de performances d'entrée peuvent être obtenues à partir de la littérature, de simulations ou de mesures expérimentales, puis introduites dans la plateforme comme paramètres pour les cellules mémoire bas niveau.

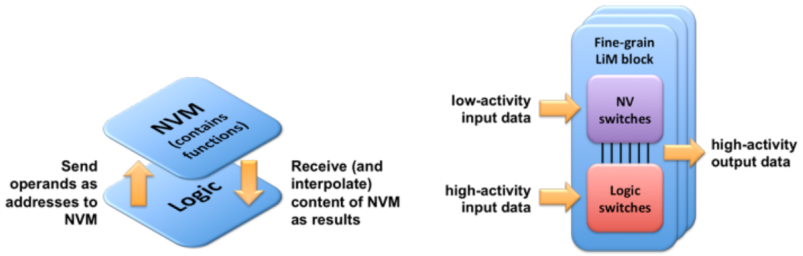


FIG. 5.11 : Concepts de LiM incluant des éléments de mémoire non volatile : (a) Approche à gros grain utilisant des tableaux de mémoire denses avec une logique périphérique pour le calcul, (b) Approche à grain fin utilisant de petites cellules ou éléments de mémoire dans les circuits logiques.

Le rôle de la plateforme est alors de quantifier le coût énergétique des opérations, ainsi que la latence réalisable, en suivant l'utilisation de chaque cellule. Les résultats simulés sont également comparés aux résultats attendus afin de suivre la perte de précision résultant des circuits de calcul approximé.

Algorithmes de test

La partie la plus importante d'une plate-forme d'évaluation est probablement l'ensemble des algorithmes de référence compatibles. L'objectif étant d'obtenir des valeurs de performance pour des cas d'utilisation réalistes, la plateforme doit donc être capable de s'interfacer avec tests de référence dans l'industrie. Comme illustré sur la figure 5.12, l'approche actuelle se concentre sur la simulation d'une architecture « mémoire améliorée », connectée à un CPU via un bus d'adresse et de données conventionnel. Comme coprocesseur, la mémoire peut recevoir

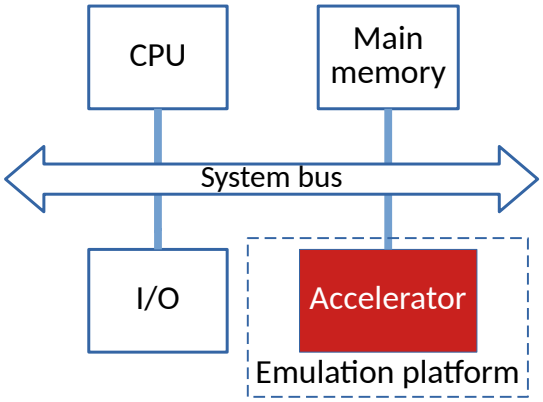


FIG. 5.12 : Localisation d'un accélérateur LiM sur le bus système.

des instructions par ce bus, les tests de référence peuvent donc être évalués en rejouant des traces d'exécution sur le bus. Cela permet une grande liberté pour leur génération : les traces peuvent être des tests synthétiques écrits manuellement, un programme en cours d'exécution interfacé avec la plate-forme, une trace d'exécution générée par un programme, ou extraite par un compilateur instrumanté[Koo+18 ; Mam+21].

5.4.3 Mise en œuvre

La méthodologie d'implémentation de la plate-forme vise à créer un cadre dans lequel les performances des architectures LiM peuvent être estimées de manière objective et fiable, puis comparées. Pour permettre un tel degré de flexibilité, les simulations sont effectuées à

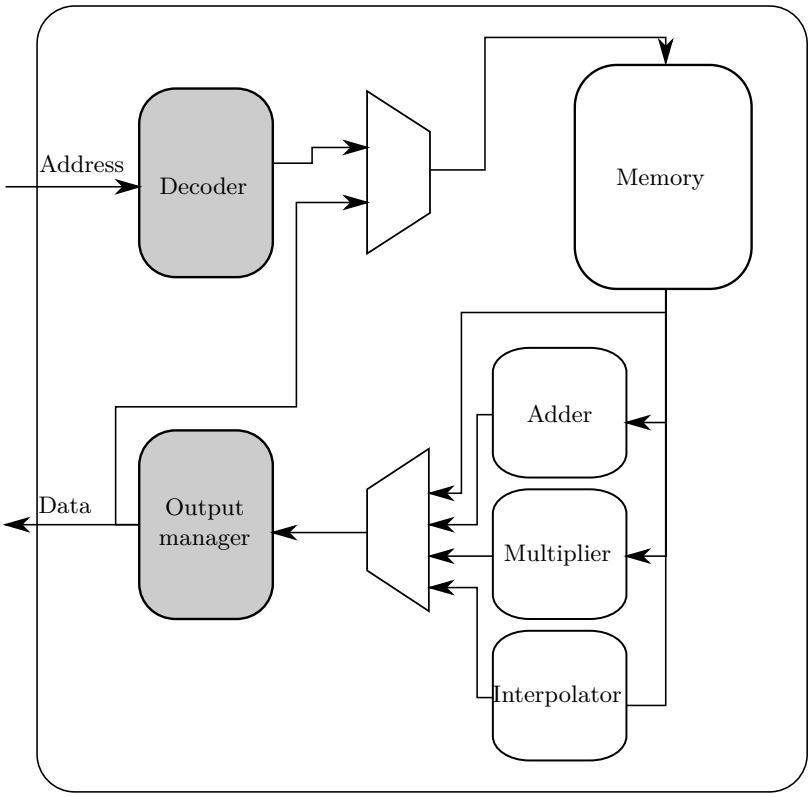


Fig. 5.13 : Représentation schématique de l'architecture actuelle de la plate-forme. Les modules de gestion sont colorés en gris, les modules d'opération en blanc.

un niveau d'abstraction élevé afin de les conserver indépendantes du matériel, permettant différentes implémentations du même élément dans de multiples technologies et avec diverses architectures.

En conséquence, l'architecture est nécessairement modulaire, ce qui permet de facilement remplacer certains composants, et ouvrant la voie à de possibles futures améliorations.

Le simulateur est développé avec les bibliothèques SystemC, qui ajoutent des fonctions de description du matériel au langage de programmation C++. Ce choix a en partie été effectué pour l'interopérabilité possible avec des tests de haut niveau existants, par le biais d'une interface en langage C, ainsi qu'avec des modèles bas niveau (RTL) de certains composants. Cela impose une architecture modulaire, qui facilite son évolution (par exemple, la simulation d'un Unité Arithmétique et Logique (ALU, Arithmetic and Logic Unit) pour obtenir des mesures et des comparaisons de performances plus précises), grâce à la nature orientée objet du langage. SystemC offre également des fonctions d'enregistrement de l'historique des signaux, qui peuvent ensuite être utilisés pour analyser les métriques de performance ainsi que l'état interne.

Architecture

L'architecture de la plate-forme imite celle d'une mémoire instrumentée pour reproduire, à plus grande échelle, les mesures d'énergie et de latence effectuées lors de la simulation des tableaux de bitcell de dimensions réduites. Il est ensuite possible d'utiliser ces mesures de performance au niveau du système, avec une quantité de mémoire adressable appropriée. Plus précisément, cela permet de comparer les performances d'une charge de travail réaliste exécutée avec différentes architectures de mémoire normalement éteinte, de LiM à gros grain ou DRAM classique, lorsque la mémoire simulée fonctionne dans ces différents modes.

Le diagramme de structure interne présenté sur la figure 5.13 montre deux types de modules différents : les modules de gestion et les modules d'opération. Le premier type n'est pas directement lié à l'estimation des paramètres, mais pilote les signaux de contrôle et

aiguille les données ; le décodeur et le gestionnaire de sortie sont ses principaux composants. Les modules d'opération sont liés aux implémentations matérielles : leur description est faite de cartes modèles détaillant leurs caractéristiques de performance, extraites de simulations ou de la littérature, et peuvent être changées ou mises à jour au moment de l'exécution.

Une distinction est également faite entre deux schémas d'utilisation de LiM : WB et sans réécriture (NWB, Non Write Back), qui diffèrent par la manière dont les résultats des calculs sont traités : le premier les stocke en mémoire pour une utilisation ultérieure sans les envoyer sur le bus de données, tandis que le second les renvoie sur le bus de données pour un traitement immédiat par le CPU. Ainsi celles-ci s'apparentent respectivement à des opérations d'écriture et de lecture normales de la perspective du bus de données. L'architecture reflète cela via le module de gestion de la sortie, qui lui permet de fonctionner dans l'un ou l'autre mode.

Pipeline de simulation

La plate-forme a été conçue pour accepter comme entrée une série d'instructions pouvant être générées dynamiquement ou fournies à l'avance sous la forme d'une trace d'exécution. Comme le montre l'extrait de code 5.3, ces instructions sont très semblables aux opcodes des processeurs ordinaires, et consistent en une série d'accès à la mémoire et d'opérations LiM. Ces instructions sont proches de celles qu'une implémentation matérielle pourrait utiliser, bien que leur codage soit modifiable. Leur exécution est simulée simultanément avec la mise à jour des indicateurs de performance. Les indicateurs suivants sont disponibles à chaque instant : la latence globale de l'exécution, la consommation totale d'énergie, et l'erreur entre le résultat de la plate-forme et le résultat attendu. Cette erreur étant causée par l'arithmétique de précision finie – ou approximative, selon l'architecture – utilisée pour le calcul.

À la fin de l'exécution, des traces de simulation sont produites, mémorisant l'évolution des signaux dans le temps, ainsi que la consommation d'énergie estimée.

Une autre approche possible (n'ayant pas été retenue ici) consisterait à compter les instructions et à additionner leurs coûts énergétiques et de latence. Bien que cette approche mène à des résultats similaires dans les cas simples, et que cela puisse considérablement simplifier l'architecture, les comportements plus complexes ne pourraient pas être reflétés, tels que les dépendances entre les opérations, et les mesures de performance dépendant de l'état. Par exemple, il est moins coûteux en énergie de lire un 0 depuis une DRAM, ou l'un des deux états pour les mémoires ferroélectriques 1T-1C. Ne pas en tenir compte empêcherait l'étude comparative d'architectures optimisées pour tirer parti de ces asymétries, comme un système prédictif permettant de limiter le recours au WB lors de la lecture d'une mémoire ferroélectrique.

Module de contrôle décodeur

Ce module analyse un fichier de trace d'exécution en entrée pour simuler une entrée de données sur le bus d'adresse, décode les différentes opérations et génère des signaux de commande pour les autres composants. Celui-ci est également chargé de terminer la simulation à la fin du fichier d'entrée.

Module de gestion de la sortie

Ce module de contrôle sélectionne la destination des données. Dans le cas d'une opération NWB, le bus de sortie de la plate-forme est piloté et le résultat est envoyé au CPU. Dans le cas d'une opération WB, l'adresse d'écriture est chargée depuis le décodeur, puis le module de gestion de la sortie renvoie le résultat à la mémoire pour qu'il y soit écrit, évitant ainsi le coût énergétique associé à la transmission des données vers le CPU. Ce module calcule également l'erreur entre le résultat exact attendu et le résultat approximatif renvoyé sur le bus. Ce calcul est nécessaire, car un débordement peut se produire, ou l'opération peut utiliser de la logique approximée. Le module de contrôle est par ailleurs responsable de la saturation de la sortie lorsqu'un tel dépassement est détecté, si l'arithmétique saturée est souhaitée.

Calcul et suivi des performances

Les modules restants de la figure 5.13 sont les modules opérationnels, effectuant les opérations de calcul : ceux-ci permettent à la plateforme de simuler diverses opérations, en fonction de la taille du mot mémoire, et maintiennent des compteurs de performance pour suivre la consommation d'énergie et la perte de précision. Ces modules sont décrits plus en détail dans la section 5.4.4.

Afin de mesurer la rapidité maximale réalisable et d'identifier les goulots d'étranglement, la plateforme est entièrement asynchrone : les modules d'opération communiquent la latence estimée pour chaque opération, ensuite additionnée au sein d'un compteur de performance global, dictant également les délais pour l'opération suivante.

5.4.4 Modules opérationnels et cartes modèles

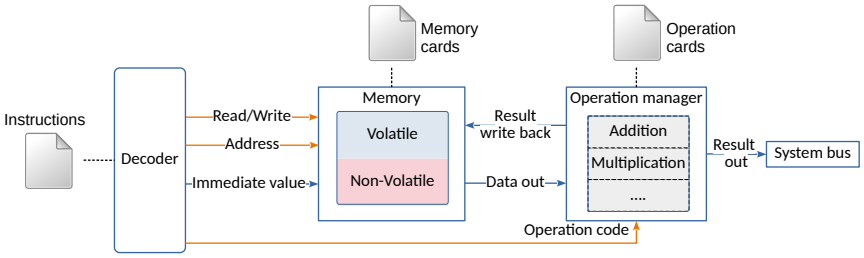


Fig. 5.14 : Architecture de la plate-forme de simulation, montrant les cartes de modèle et le gestionnaire d'opérations.

Module mémoire

Bien que la mémoire soit considérée comme un module d'opération dans la taxonomie de la plate-forme, il s'agit d'un cas particulier, compte tenu de la place centrale que celle-ci occupe. Ce module se trouve en effet au cœur de la plate-forme, fournissant une interface aux autres modules pour lire et écrire des données dans la matrice mémoire. Son modèle comportemental est très haut niveau et générique. En revanche, celui-ci peut émuler le comportement de plusieurs technologies différentes, telles que décrites dans différentes cartes modèles, comme illustré dans la figure 5.14, qu'il s'agisse de mémoires volatiles telles que des DRAMs, ou de mémoires non volatiles telles que les mémoires ferroélectriques 1T-1C ou 1T-FeFET. Les paramètres indiqués sur chaque carte comprennent les coûts énergétiques de lecture/écriture, la latence, la taille de la mémoire, la longueur des mots, le taux de rafraîchissement, ainsi que l'énergie et la latence de redémarrage.

Modules opérationnels

Comme le montre la figure 5.13, les autres modules opérationnels sont des implémentations haut niveau de diverses opérations, l'addition et la multiplication étant actuellement disponibles. Deux variantes de multiplicateur ont été conçues : l'une avec un multiplicateur à transistors standard, et l'autre utilisant une tableau de correspondances (LUT, Lookup Table) creuse (*sparse*), combinée à un interpolateur bilinéaire. Ces modules s'appuient sur des valeurs de performance bas niveau, telles que la consommation d'énergie par bit calculé et la latence du chemin critique. Ces paramètres sont stockés dans des cartes modèles, et peuvent être chargées au moment de l'exécution, comme le montre la figure 5.14. Il est ainsi possible de permuter facilement les architectures et technologies pour un bloc fonctionnel donné, sans modifier la structure interne ou le code de la plateforme elle-même. Les valeurs peuvent être déterminées à l'aide de mesures expérimentales, de simulations bas niveau, ou tirées de la littérature[JGL16 ; Vog10 ; Sin+].

Implémentation des modules opérationnels

Le module de gestion des opérations comprend une interface générique permettant une opération entre deux opérandes et un contrôleur réalisant ces opérations, comme illustré sur la figure 5.15. Cette interface générique est implémentée via le mécanisme d'héritage fourni par le langage de programmation C++. Le polymorphisme permet d'ajouter aisément de nouvelles opérations, tout en assurant leur conformité à l'interface : une classe générique parente operation définit la méthode virtuelle run, qui est ensuite implémentée par chaque opération.

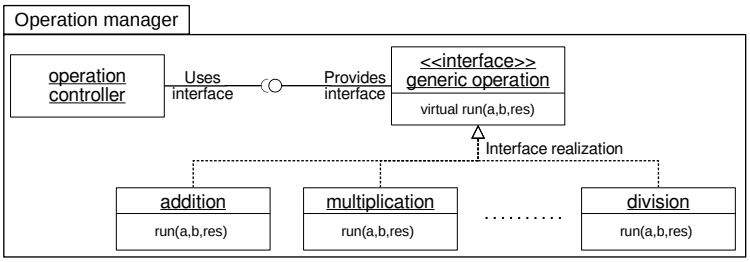


Fig. 5.15 : Illustration de l'interface d'opération commune.

Le module décodeur reçoit la trace d'exécution ou les opcodes de contrôle de l'accélérateur sous la forme d'une liste d'opérations à traiter (instructions). Celui-ci interagit ensuite avec les modules de mémoire et de gestion opérationnelle selon l'instruction, comme illustré en figure 5.16. Le contrôleur d'opérations maintient un tableau de pointeurs vers chaque opération disponible, et appelle l'opération appropriée (les opcodes étant actuellement des indices de tableau). Cela permet de calculer le résultat approprié, tandis que le contrôleur d'opération calcule les coûts de délai et d'énergie pour l'opération en lisant la carte modèle correspondante. Par conséquent, en fournissant des jeux de cartes modèles distincts, différentes implémentations de la même opération peuvent être évaluées avec le même code de test, sans recompilation de la plateforme. Différents modèles de mémoire et d'opérations peuvent également être utilisés, bien qu'un tel changement ne nécessite une recompilation, ou l'utilisation d'opcodes distincts pour chaque modèle dans les traces d'exécution. Cela permet de comparer la précision et les résultats entre différents modèles.

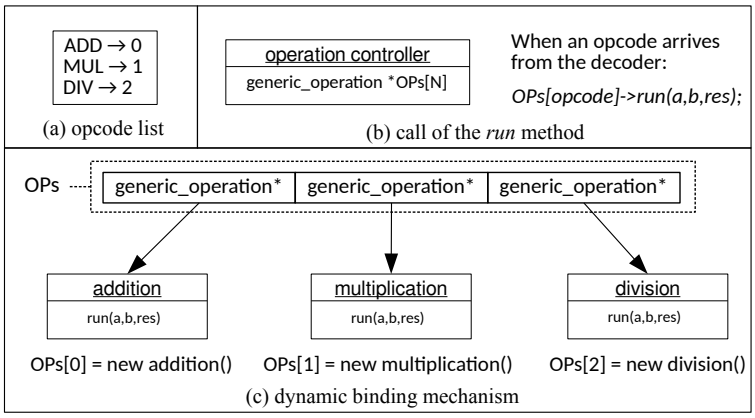


Fig. 5.16 : Diagramme d'exécution du module opérationnel.

Structure de la carte modèle

Il est essentiel de définir correctement les paramètres des modules mémoire et opérationnels (via les cartes modèles) pour obtenir des estimations de performances précises pour l'architecture simulée. L'extrait de code 5.1 illustre la structure d'une carte modèle mémoire, tandis que l'extrait de code 5.2 montre la structure d'une carte modèle de module d'opération.

Extrait de code 5.1 : Structure de carte modèle mémoire

```
v1      énergie pour lire une cellule = 0 (pJ)
v2      énergie pour lire une cellule = 1 (pJ)
v3      énergie pour écrire 0 sur 0 (pJ)
v4      énergie pour écrire 0 sur 1 (pJ)
v5      énergie pour écrire 1 sur 0 (pJ)
v6      énergie pour écrire 1 sur 1 (pJ)
v7      puissance nécessaire pour conserver 0 (pW)
v8      Puissance nécessaire pour conserver 1 (pW)
v9      latence de lecture (ns)
v10     latence d'écriture (ns)
v11     temps de rétention (ns, 0 = inf)
```

Extrait de code 5.2 : Structure de carte modèle de module d'opération

```
v1      nombre de bits de l'opération
v2      énergie par bit (pJ)
v3      latence (ns)
```

Dans les deux cas, les valeurs (v1, v2, etc.) sont lues par les modules *Mémoire* et *Gestionnaire d'opérations* et utilisées dans la simulation pour suivre les coûts associés aux opérations de lecture/écriture de la mémoire ainsi que ceux des calculs, respectivement.

5.4.5 Exemple pratique : Additionneur

Cet exemple simple vise à présenter les capacités de la plate-forme de simulation et à illustrer son utilisation en modélisant un accélérateur calculant la somme de deux valeurs, comme illustré sur la figure 5.17. L'une des valeurs d'entrée est stockée une seule fois dans un

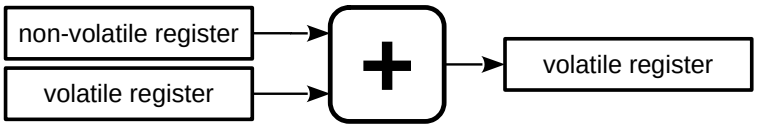


Fig. 5.17 : Opération d'addition effectuée par l'accélérateur choisi pour exemple.

registre non volatile et est ensuite utilisée comme constante. L'autre valeur est stockée dans un registre volatil, et évolue au cours du temps. Le résultat est stocké dans un registre volatil. Pour simplifier, la quantité d'énergie et la latence nécessaire pour la lecture/écriture d'un bit sont supposées indépendantes des valeurs précédentes et actuelles. Le code nécessaire pour mettre en œuvre cette opération dans le décodeur est simplement :

Extrait de code 5.3 : Trace d'exécution d'une addition

```
wv 0 64   #écriture de la valeur volatile 64 à l'adresse 0
wnv 1 32  #écriture de la valeur non volatile 32 à l'adresse 1
ADD 0 1 2 #somme des valeurs aux adresses 0 et 1 et écriture du résultat à l'adresse 2
```

Ce code utilise deux registres (adresses mémoire), l'un dans en mémoire volatile, l'autre en mémoire non volatile. Le code effectuant l'opération est ensuite appelé par l'instruction ADD, qui lit les paramètres tels que la latence et le coût énergétique de l'opération dans la carte modèle correspondante. Comme cette instruction spécifiant un registre dans lequel écrire la valeur du résultat de l'opération, le gestionnaire de sortie fonctionne en mode WB et le coût énergétique associé est enregistré.

Comme illustré sur la figure 5.18, la plateforme peut simuler une combinaison de traces d'exécution, de cartes modèles de mémoire et de calcul, et extraire une estimation des valeurs de performances correspondantes. La tableau 5.7 énumère les résultats de simulation en termes d'énergie et de délai pour l'opération d'addition réalisée avec des additionneurs de différentes largeurs de mot d'entrée (8, 16 et 32 bits), ainsi que différentes implémentations (exprimées en termes d'énergie et de délai nécessaires pour effectuer l'addition).

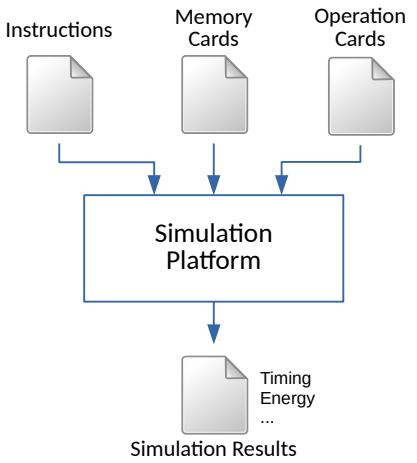


FIG. 5.18 : Entrées et sorties de la plateforme d'évaluation des performances, montrant les cartes modèles en entrée et le résultat d'évaluation des performances en sortie

Paramètre				**Unité**
Opérandes additionneur (bit)	8	16	32	bit
lecture NV	0.01	0.01	0.01	Énergie (pJ/bit)
Écriture NV	0.05	0.05	0.05	
Lecture volatile	0.1	0.1	0.1	
Écriture volatile	0.2	0.2	0.2	
Addition NV	0.5	0.5	0.5	
lecture NV	0.05	0.05	0.05	Délai (ns)
Écriture NV	0.1	0.1	0.1	
Lecture volatile	0.1	0.1	0.1	
Écriture volatile	0.15	0.15	0.15	
Addition NV	8	16	32	
Énergie totale	**8.48**	**16.96**	**33.92**	pJ
Délai total	**8.55**	**16.56**	**32.56**	ns

TAB. 5.7 : Paramètres et résultats de simulation pour l'exemple de l'additionneur. Ce tableau présente trois ensembles de paramètres, avec différentes largeurs de mots pour l'additionneur, ainsi que l'énergie totale simulée et le délai par opération. Les paramètres sont fournis par les cartes modèles et modifiés entre chaque simulation, sans recompilation de l'exécutable de la plate-forme de simulation.

Les simulations ont été réalisées en changeant simplement les cartes modèles de la mémoire et de l'opération selon les paramètres listés dans le tableau entre chaque simulation, sans recompiler l'exécutable de la plateforme. Cela démontre la flexibilité de l'architecture choisie, celle-ci ayant été conçue pour être intégrée dans une boucle DTCO (comme détaillé dans section 5.6.2), et nécessitant donc une friction minimale pour changer les paramètres.

5.5 Résultats de l'exploration au niveau système

Afin d'évaluer et de valider la plate-forme pendant son développement, de simples modules d'opération et traces d'exécution ont été préparés pour démontrer sa capacité d'évaluation d'architectures et de technologies. Cet outil peut être utilisé pour identifier les paramètres critiques durant le développement des dispositifs LiM, et pour acquérir une meilleure compréhension de la manière dont ceux-ci affectent les performances de l'ensemble du système. Les résultats de simulation présentés ici utilisent les paramètres listés dans le tableau 5.8 pour les modules.

Ceux-ci ont été extraits de la littérature disponible[JGL16 ; Vog10 ; Sin+], et estimées si nécessaire. Ces résultats doivent donc être considérés comme une démonstration des fonctionnalités de la plate-forme, plutôt que comme une analyse quantitative.

Paramètre	Valeur	Unité	Source
énergie pour lire '0'	1	nJ	Simulation
énergie pour lire '1'	2.5		
énergie pour écrire '0' sur '0'	0.5		
énergie pour écrire '0' sur '1'	2		
énergie pour écrire '1' sur '0'	2		
énergie pour écrire '1' sur '1'	0.5		
puissance pour enregistrer '0'	0.1	mW	Estimation
puissance pour enregistrer '1'	0.1	mW	
énergie de démarrage	5	nJ	
latence d'extinction	200	ns	
latence de lecture	20	ns	Expérimental
latence d'écriture	20		

Tab. 5.8 : Paramètres de consommation d'énergie tels que figurant sur la carte modèle de la cellule mémoire ferroélectrique 1T-1C utilisée pour la simulation. Les paramètres ont été obtenus par simulation lorsque cela était possible. Dans les autres cas, ces paramètres ont été extraits de la littérature ou estimés.

5.5.1 Cas d'utilisation normalement-éteint

Dans les cas où les accès à la mémoire ou les calculs sont peu fréquents, il peut être avantageux d'éteindre le système. Cela peut être le cas même si un travail supplémentaire, donc une consommation d'énergie supplémentaire, est nécessaire pour sauvegarder et restaurer l'état vers et depuis la Mémoire non volatile[Win+20]. Les gains d'efficacité énergétique sont alors obtenus en réduisant la consommation d'énergie statique du système. Dans de tels cas, il est primordial d'analyser les compromis entre la consommation d'énergie nécessaire à l'arrêt et la consommation d'énergie statique. Il est particulièrement crucial de quantifier le rapport cyclique limite (maximum) en dessous duquel de l'énergie peut être économisée grâce à l'arrêt du système.

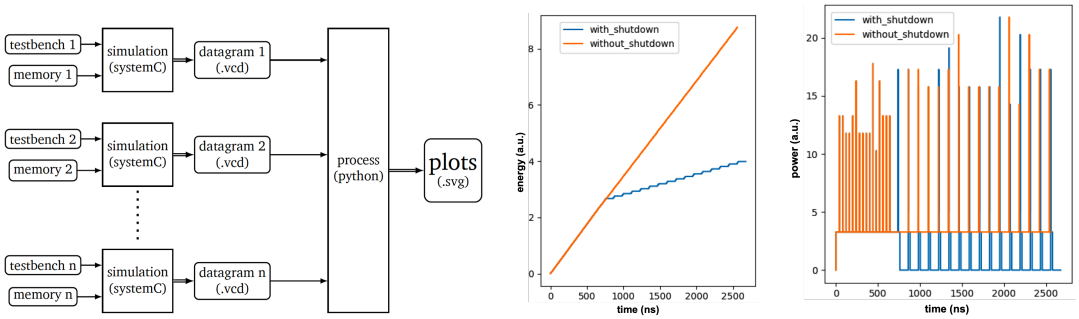
La plateforme d'évaluation peut également être utilisée pour estimer comment la consommation énergétique varie d'une technologie de mémoire à une autre.

La figure 5.19a montre comment la plateforme d'évaluation peut être exploitée pour produire des graphiques de consommation d'énergie à partir d'un ensemble de tests et de cartes modèles. La figure 5.19b et la figure 5.19c montrent respectivement la consommation d'énergie et la puissance nécessaire lors de l'exécution d'une trace de multiplication matricielle dans deux scénarios : l'un avec arrêt et l'autre sans arrêt. Le but étant ici la démonstration des capacités de la plateforme, cette analyse utilise des valeurs arbitraires pour la consommation d'énergie. Néanmoins, les gains réalisables avec le calcul normalement-éteint sont identifiables sur les graphiques obtenus.

Cette évaluation a été conservée comme démonstrateur de fonctionnalité minimale lors de l'évolution de la plate-forme.

5.5.2 Simulations de circuit interpolateur

Un interpolateur est l'une des implémentations possibles de LiM, dans laquelle les données de sortie d'une fonction à n entrées sont stockées dans un tableau à n dimensions (ressemblant ainsi à une LUT). La mémorisation de l'intégralité des combinaisons étant infaisable lorsque le nombre de dimensions et d'états augmente, les résultats sont échantillonnés. Un tableau « creux » (*sparse*) est utilisé, celui-ci mémorisant uniquement quelques points de données, et le résultat est interpolé à partir des points disponibles. Cette approche permet d'équilibrer les



(a) Flot de génération de tracés

(b) Consommation d'énergie en fonction du temps

(c) Puissance utilisée en fonction du temps

FIG. 5.19 : Simulation normalement-éteinte montrant la consommation d'énergie cumulée pendant 16 opérations d'écriture et 16 opérations de lecture avec et sans stratégie d'arrêt, juxtaposé à une illustration du flot utilisé pour générer les graphiques.

exigences du circuit en matière de surface et de performance entre la logique et la mémoire. Selon la fonction d'interpolation (par exemple, linéaire ou bilinéaire) et la densité du tableau, le résultat peut être plus ou moins précis, permettant également de sacrifier de la précision au profit de complexité ou de gains de performance.

L'objectif de cette simulation est d'étudier la compétitivité d'une architecture utilisant un multiplicateur à LUT et un d'interpolateur avec une architecture à logique traditionnelle. La fonction de l'interpolateur a été choisie pour produire des résultats non approximés, comme point de comparaison initial : une interpolation bilinéaire est effectuée entre les valeurs précalculées[Moz21, p. 49]. Cette implémentation a été soumise à des tests approfondis pour comprendre comment ses performances varient en fonction des paramètres physiques et technologiques : celle-ci a été testée avec des opérandes pseudo-aléatoires, et une combinaison exhaustive des entiers signés 8 bits, tout en mesurant la consommation d'énergie moyenne par opération pour chaque taille de LUT.

La figure 5.20 montre l'énergie utilisée par le calcul en fonction de la surface normalisée du circuit de la LUT, équivalente à la taille de la mémoire en bits :

$$A = \frac{A_{tableau}}{A_{bitcell}}$$

La même analyse a été effectuée avec pour paramètres de la carte mémoire ceux de la technologie actuelle, et avec des valeurs améliorées, afin d'identifier les performances requises pour rendre cette architecture compétitive. Ces résultats montrent que l'efficacité énergétique de cette architecture d'interpolateur n'est pas compétitive avec les multiplicateurs classiques, avec la technologie de mémoire actuelle (dont les paramètres sont donnés dans la tableau 5.8), quelle que soit la taille du tableau de mémoire. Seule une réduction de 80 l'énergie de lecture permet d'inverser la tendance, diminuant l'énergie nécessaire pour un calcul lorsque que la taille de la matrice augmente (l'interpolateur ayant moins de calculs à effectuer). Avec cette hypothétique technologie de mémoire améliorée, l'augmentation de consommation d'énergie liée à un tableau de mémoire plus grand est suffisamment faible pour être compensée par l'utilisation réduite de l'interpolateur. Dans ce cas, une multiplication serait effectuée plus efficacement qu'avec une approche à portes logique à partir d'une LUT de 1024 mots, où la consommation d'énergie simulée devient plus faible qu'en l'absence de mémoire. Au-delà de cette valeur, le choix de la taille de LUT devient un compromis entre l'énergie et la surface.

En fonction des progrès nécessaires pour atteindre ce niveau de performance, cette architecture devrait également se maintenir compétitive vis-à-vis des circuits logiques améliorés contemporains. Cela est peu probable, à moins que la technologie des mémoires ne progresse plus rapidement que celle de la logique (ce qui n'était pas le cas historiquement), ou si l'application choisie exclut l'utilisation de nœuds logiques avancés (ce qui peut être le cas dans une puce mémoire). Les mémoires à FeFET, par l'absence de mécanisme WB, peuvent aussi être plus adaptées à cette utilisation, car les valeurs de la LUT changent uniquement lorsque la fonction exécutée

est modifiée. Une autre voie d'amélioration des performances consiste à réduire la complexité de l'interpolateur en diminuant la précision des résultats, rejoignant le paradigme du calcul approximé. Bien que l'évaluation de ce mode de fonctionnement ait été un objectif de conception de la plateforme d'évaluation, celui-ci n'a pas été étudié.

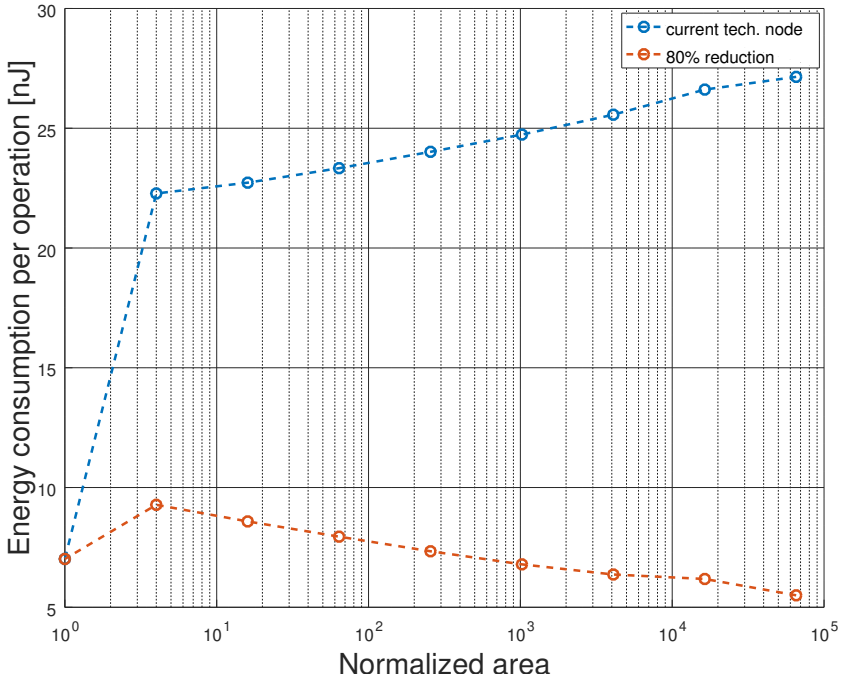


FIG. 5.20 : Consommation moyenne d'énergie de l'interpolateur par opération en fonction de la surface normalisée de la LUT. La première valeur correspond à l'interpolateur sans LUT (A = 0). Le graphique montre les gains possibles en termes d'efficacité énergétique par rapport aux multiplicateurs classiques, si le coût énergétique de la lecture de Mémoire non volatile peut être amélioré d'au moins 80%.

5.5.3 Mesure de performance sur multiplication matricielle

Le comportement d'une multiplication matricielle a été étudié comme point de référence moins synthétique, et comme étape de transition vers des algorithmes plus complexes. Ce test consiste à multiplier deux matrices 5×5 d'entiers signés de 8 bits à l'aide d'un algorithme naïf de multiplication matricielle. La figure 5.21 montre la consommation totale d'énergie de l'accélérateur LiM en fonction de la taille de mot utilisée pour stocker et transmettre le résultat de la multiplication matricielle, pour des cas WB et NWB. Une différence énergétique faible, mais croissante, peut être observée entre les deux scénarios, en raison de la contribution du bus de sortie envoyant les résultats des calculs à l'unité centrale de traitement. Ce coût est toujours présent dans le cas NWB, alors qu'il est absent du cas WB. Il est également important de noter qu'une spécificité du mécanisme d'évaluation de la précision forçait la sauvegarde des données en mémoire dans le cas NWB, contribuant à la consommation d'énergie. Cet exemple a mis en évidence le besoin de mécanismes plus précis d'évaluation de la consommation d'énergie du bus de communication avec l'unité centrale de traitement, celui-ci étant actuellement une estimation.

5.6 Conclusion

5.6.1 Exploration de l'espace de conception

Dans ce chapitre, un flot de DSE entièrement automatisé a été présenté. Bien que les instabilités des modèles aient empêché son utilisation prolongée comme plateforme d'optimisation

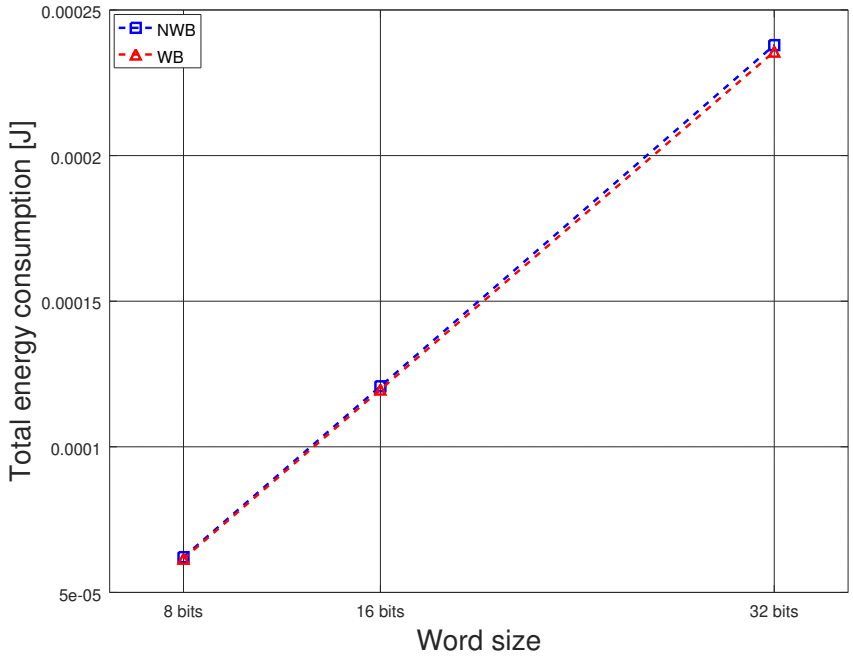


FIG. 5.21 : Consommation totale d'énergie en fonction de la taille du mot pour l'exécution d'une multiplication matricielle.

autonome, la modularité de l'architecture a permis une exploration manuelle de l'espace de conception, sans boucle de rétroaction automatisée susceptible de converger vers des cas limites. Cela a révélé des améliorations potentielles de la méthodologie, ainsi que des résultats inattendus devant encore être étudiés en profondeur, comme le faible impact de l'aire du condensateur sur la fenêtre mémoire[2]. Peut-être plus important encore, de multiples composants, tels que l'IPC décrite dans la sous-section 5.2.2 se sont déjà révélés utiles dans d'autres projets[Poi22, p. 103] [3] .

Un autre résultat de cette exploration de l'espace de conception est la cartographie partielle de l'espace de performance pour les circuits ferroélectriques. Bien que ces résultats n'aient pas été comparés à des données expérimentales, ceux-ci constituent un point de départ raisonnable pour évaluer l'impact de ces nouveaux circuits sur les performances au niveau système.

Instabilité des modèles et précision

S'agissant d'une approche par simulation, la DSE automatisée reflète uniquement la manière dont les modèles compacts interagissent en différents points de l'espace de conception. Il est donc essentiel que ces modèles représentent précisément la performance des dispositifs et des circuits. Des problèmes de stabilité des modèles ont empêché l'exploration de certaines zones de l'espace de conception. En outre, les instabilités étaient souvent accompagnées d'imprécisions, telles que des oscillations de grande amplitude créées par des problèmes de convergence, empêchant ainsi l'automatisation de l'extraction des métriques de performance. Le manque de données expérimentales pour les circuits ferroélectriques a également empêché l'évaluation de la précision des modèles, réduisant la confiance portée aux résultats présentés dans ce chapitre.

Bien que des progrès n'aient été réalisés sur l'automatisation de la DSE, ceux-ci ont mis en évidence les limitations des modèles actuels. Lorsque plus de recul et davantage de données expérimentales seront disponibles sur les circuits et dispositifs ferroélectriques utilisant $HfZrO_2$, de meilleurs modèles devraient permettre des explorations plus approfondies. Des modèles plus simples, tels que celui présenté dans l'extrait de code 2.1, peuvent également fournir des informations intéressantes, en échangeant de la précision contre une simulation

[2] Ce résultat a depuis été confirmé être dû à un problème de mesure. Se référer à la soutenance.
[3] La majorité de ce travail a également pu être réutilisée en substituant pymoo[BD20] à LIFT.

accélérée. D'autres approches de modélisation peuvent aussi refléter des comportements ferroélectriques plus complexes[Den+20].

5.6.2 Plate-forme d'évaluation des performances au niveau système

Statut actuel

Des travaux ont été menés en vue de la réalisation d'une plate-forme d'évaluation des performances au niveau système, prenant comme paramètres les résultats de performance des simulations effectuées au niveau du dispositif et du circuit. L'outil d'évaluation principal a été réalisé en SystemC, permettant un interfaçage direct avec les tests de performance de référence de l'industrie, pour des applications telles que la reconnaissance d'images.

Bien que cette plateforme n'ait commencé à produire des résultats que récemment, celle-ci est prometteuse pour explorer les variations architecturales, ainsi que l'impact au niveau système de l'amélioration de performance des dispositifs.

Celle-ci pourrait également devenir un outil utile pour le développement et l'évaluation de compilateurs pour LiM. Cependant, des outils concurrents tels que Gem5-X[Qur+21] sont apparus depuis, et une réévaluation de leurs capacités est nécessaire.

Approfondissement et analyses complémentaires

Les résultats actuels montrent uniquement que des gains mineurs peuvent être obtenus avec l'architecture évaluée. Cela est en partie dû à une bibliothèque d'architectures et de tests incomplète. En outre, l'incertitude de certains paramètres des cartes modèles est relativement élevée en raison des résultats de DSE, et des estimations conservatrices ont été utilisées à plusieurs endroits, pour accélérer le temps de développement. Par conséquent, ces résultats montrent des tendances intéressantes et des pistes à approfondir, mais les valeurs mêmes de performance sont très incertaines, rendant celles-ci difficiles à comparer de manière absolue.

Le problème réside en partie dans le fait que la plateforme ne prend actuellement en compte que les contributions en énergie et en latence des coprocesseurs LiM, sans les comparer aux calculs effectués sur CPU. L'extension de la plate-forme pour couvrir cette partie du système est probablement nécessaire pour une analyse complète de l'utilisation de l'énergie au niveau système : l'architecture actuelle est effectivement capable de comparer plusieurs implémentations de coprocesseurs, mais la comparaison est incomplète sans l'inclusion de chiffres de référence sur CPU.

Co-optimisation circuit et technologie

Les fronts de Pareto au niveau du circuit générés par DSE comme décrit dans la section 5.1 aident le concepteur à choisir des compromis de performance adaptés au cahier des charges. Cependant, l'estimation des performances au niveau système sur des tests réalistes est l'objectif de la plateforme d'évaluation des performances, et ne peut pas être réalisée précisément au niveau du circuit. Par conséquent, l'ultime objectif est une approche DTCO, utilisant les outils de DSE, avec les valeurs de performance au niveau système extraites de la plateforme.

La figure 5.22 montre l'architecture envisagée, la partie droite effectuant une boucle d'optimisation décrite dans la section 5.1 avec les outils de la section 5.2, extrayant les fronts de Pareto des résultats de performance obtenus avec la plateforme représentée sur le côté gauche.

Ce projet n'a pas abouti en raison de contraintes de temps, bien que de nombreuses étapes aient été franchies vers cet objectif. En effet, pour permettre une telle boucle d'optimisation, chaque élément doit parfaitement fonctionner, du flot de DSE (comprenant les modèles), jusqu'à la plateforme d'évaluation des performances au niveau système. Cette approche étagée du paradigme DTCO a permis la création d'outils modulaires et génériques, autorisant ainsi leur réutilisation dans d'autres projets.

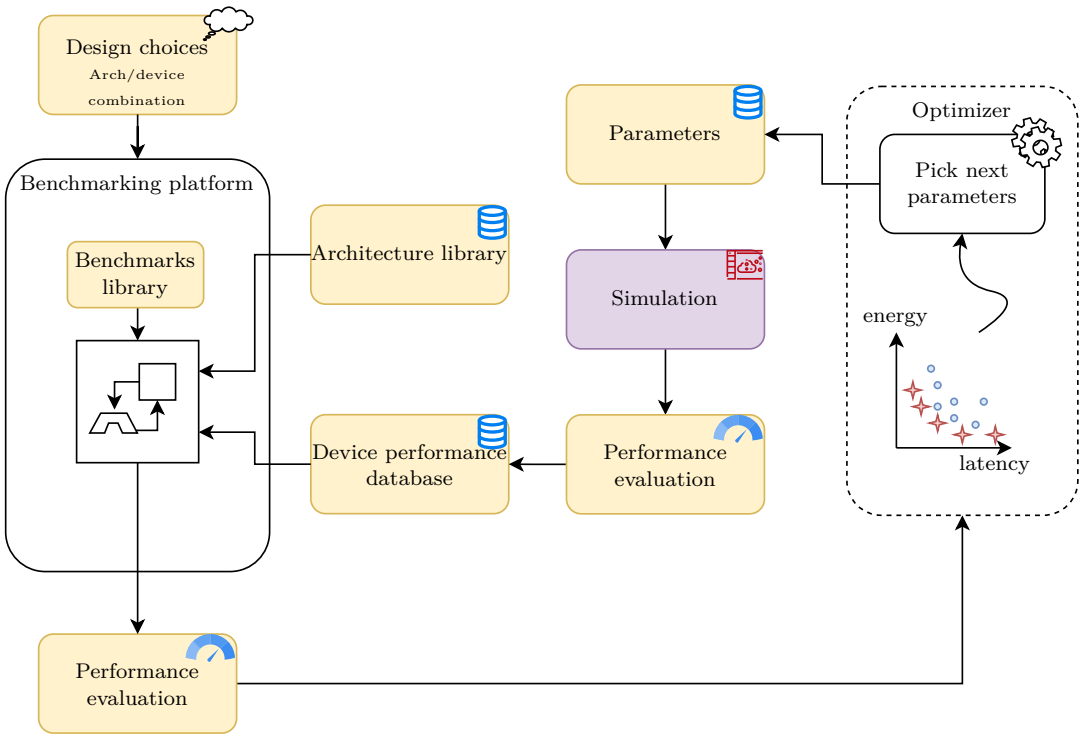


FIG. 5.22 : Flot de DTCO automatisée complet tel qu'initialement envisagé, combinant l'évaluation des performances au niveau système, à gauche, et DSE avec sélection des paramètres des dispositifs, à droite.

Chapitre 6

Conclusion

6.1 Technologie ferroélectrique bout de ligne

Le chapitre 3 se focalise sur les condensateurs ferroélectriques BEoL. Grâce à leur température de recuit relativement basse d'environ 450 °C pendant 30 s[Bou20, p. 44], les transistors ferroélectriques peuvent être déposés au-dessus des couches métalliques sans endommager les structures précédentes. Cela réduit les exigences de précision lors du dépôt de FeCaps, car les couches supérieures et les structures de condensateur ont généralement des empreintes relativement grandes, et découple la taille des condensateurs de celle des transistors.

Par rapport aux FeFETs, la technologie BEoL accorde davantage de liberté au concepteur, lui permettant simultanément beaucoup plus de flexibilité pour égaliser la capacité, et de réaliser une plus grande variété d'architectures de circuits. En effet, alors que la fonctionnalité des FeFET peut être reproduite à l'aide de FeCaps discrets (résultant en PsFeFET ou FeMFET), y compris BEoL comme détaillé dans la section 3.3, l'inverse donne des résultats médiocres[Sle+19b].

Bien que les résultats d'une comparaison approfondie entre PsFeFET et FeFET ne soient pas encore disponibles, les résultats préliminaires sont prometteurs et valident le fonctionnement FeFET. Les résultats expérimentaux montrent que la cellule PsFeFET est capable de mémoriser et de restituer des données, avec une fenêtre mémoire suffisamment importante pour pouvoir être subdivisée en plusieurs niveaux dans le cadre d'un fonctionnement MLC. Les résultats de la caractérisation du circuit 2T1C confirment également ces conclusions, la différence portant sur l'utilisation d'un transistor dédié à l'accès au nœud flottant, permettant de réduire les tensions de fonctionnement requises et de simplifier la séquence de programmation. Une analyse plus détaillée du courant de fuite et la rétention de ces structures serait intéressante, bien que la valeur stockée puisse toujours être lue de manière destructive afin de rafraîchir la mémoire, comme le ferait une cellule DRAM.

Une cellule TCAM à lecture destructrice a également été conçue et fabriquée. Bien que des circuits plus récents[Yin+19] dont l'opération est non destructrice rendent celle-ci partiellement obsolète, cette cellule peut être utilisée pour les applications LiM. Cependant, la plupart de ses fonctionnalités pouvant être reproduites avec la cellule 2TnC, les efforts de caractérisation sont d'abord focalisés sur cette dernière.

6.2 Avantages et limitations actuelles des FeFETs

Les FeFETs présentent des caractéristiques attrayantes pour les implémentations matérielles : non-volatilité avec un temps de rétention allant jusqu'à 10 ans (extrapolé), compatibilité CMOS, protocoles connus pour l'écriture de tableaux, tensions d'écriture modérées, stabilité au brasage par refusion, ainsi qu'une lecture et écriture rapides. En outre, ceux-ci combinent naturellement des éléments logiques (transistors) et de mémorisation (oxyde ferroélectrique).

Cependant, les FeFETs sont confrontés à des effets de piégeage de charges, qui empêchent les opérations de lecture après écriture. Les charges piégées peuvent également causer des problèmes d'endurance, compenser la fenêtre mémoire, et changer le V_{th}, particulièrement dans le cas des petits dispositifs. Les champs électriques intenses endommagent aussi l'oxyde de grille du transistor, limitant leur endurance à environ 10^4 à 10^6 cycles. Des empilements de grille plus épais nécessitent des tensions plus élevées, complexifiant leur intégration avec des transistors CMOS délicats.

Par conséquent, l'intégration de FeFETs dans de nouvelles applications LiM est particulièrement attrayante lorsqu'une valeur interne, sauvegardée dans la couche ferroélectrique du FeFET, est utilisée pour traiter un flux de données venant de l'extérieur. Les FeFET sont plus efficaces lorsque rarement repolarisés, qu'il n'est pas nécessaire d'effectuer une lecture immédiate après l'écriture, et quand des opérations de lecture et d'écriture rapides sont nécessaires.

6.2.1 Avenir de la technologie FeFET

Selon l'analyse présentée, les FeFET sont relativement difficiles à utiliser, car ceux-ci nécessitent des tensions plus élevées, des transistors de grande taille, ainsi qu'un circuit de programmation dédié et son mécanisme d'adressage associé. Ces points peuvent être compensés par l'utilisation de FeFETs conçus à partir de FeCaps BEoL, comme décrit dans la section 3.3 sous le nom de PsFeFET. Le découplage du condensateur et du transistor permet également l'insertion d'un transistor d'accès supplémentaire, comme détaillé dans la section 3.5. Ce transistor d'accès peut être intégré au mécanisme d'adressage pour programmer l'oxyde ferroélectrique, réduisant les tensions de programmation et leur effet néfaste sur l'oxyde de grille du transistor. En outre, ce découplage permet de partager un seul condensateur entre plusieurs transistors, ce qui pourrait s'avérer particulièrement bénéfique dans le cas de circuits CMOS, comme décrit dans la sous-section 4.3.1, mais pourrait aussi faciliter un contrôle plus précis de la polarisation, et d'autres opérations LiM, comme illustré avec le circuit 2TnC présenté dans la section 3.5.1.

Les circuits PsFeFET et 2T1C pourraient donc devenir plus répandus à l'avenir, si leurs caractéristiques de rétention sont compatibles avec l'application. En outre, certaines applications décrites dans le chapitre 4 peuvent également être mises en œuvre avec des condensateurs ordinaires en série avec les grilles des transistors, dans une structure semblable aux FGMOS.

6.3 DSE automatisée et modélisation

La exploration de l'espace de conception automatisée de circuits ferroélectriques a été rapidement limitée par la maturité des modèles disponibles. Néanmoins, une exploration manuelle a permis de concevoir les éléments individuels du flot automatisé. Cela a permis de mettre en évidence des comportements inattendus, tels que l'effet moins important que prévu de la surface du condensateur sur la fenêtre mémoire, détaillé dans la section 5.3.1. Ces résultats peuvent provenir des modèles eux-mêmes, ou être des artefacts de simulation provenant de l'interaction entre les différentes étapes du flot. Cette question est actuellement étudiée, et sera comparée aux résultats expérimentaux lorsque ceux-ci seront disponibles.

Les outils produits pour le pipeline d'automatisation, parmi lesquels une IPC *open source* pour communiquer avec le simulateur Spectre, ont été réutilisés dans d'autres projets[Poi22, p. 103], et permettront d'explorer davantage l'espace de conception lorsque des modèles améliorés seront disponibles.

6.3.1 Problèmes de modélisation

L'instabilité des modèles était un problème récurrent lors de ce projet. Les problèmes rencontrés comprennent des difficultés de convergence dues à des non-linéarités, ainsi que la divergence de métriques de performance, ne reflétant pas le comportement physique et induisant en erreur les algorithmes de DSE. Des dégradations de performance importantes ont également été identifiées dans des cas extrêmes, dues en partie à la quantité d'états mémorisés par le modèle de Landau, s'agissant des points d'inflexion rencontrés pendant les balayages de tension. L'exploration automatisée tend à converger systématiquement vers des configurations problématiques, en raison de leur interaction avec les méthodes d'extraction de métriques. Cela a permis d'identifier et de corriger de multiples problèmes au sein des modèles, ce qui pourrait s'avérer être une approche utile durant le développement de modèles, éventuellement comme outil complémentaire aux méthodes de fuzzing.

Malheureusement, l'instabilité des modèles a empêché l'utilisation de flots DSE entièrement automatisés. Des modèles moins précis, mais plus stables, pourraient être utilisés à l'avenir, bien que l'impact de la perte de précision sur les résultats de DSE doive être quantifié.

6.4 Évaluation des performances au niveau du système

L'évaluation des performances au niveau système est l'un des buts fixés pour ce travail, bien que l'objectif de prédire la performance au niveau système à partir de mesures au niveau du circuit n'ait pas été atteint. Cela peut être attribué à deux facteurs : l'absence de telles mesures et l'étendue du problème.

Les données de performance au niveau du circuit devaient provenir des activités de DSE. Cependant, le manque de confiance dans la précision et l'optimalité de ces résultats a conduit à fréquemment utiliser des valeurs tirées de la littérature ainsi que des données expérimentales, même avant la caractérisation complète des échantillons.

Le développement de la plateforme d'évaluation des performances au niveau système comprend la définition de nouvelles architectures système, la sélection de tests pertinents, et le portage de ces tests vers les architectures sélectionnées. Les résultats sont ensuite comparés à des performances de référence provenant d'architectures classiques.

L'approche choisie pour la plateforme d'évaluation des performances décrite dans la section 5.4 considère uniquement l'équivalent de la mémoire unifiée d'une architecture Von Neumann, avec d'éventuels accélérateurs compris dans celle-ci. Ce choix a complexifié l'obtention de points de comparaison pour les architectures traditionnelles, car la latence et la consommation d'énergie du CPU ne sont pas mesurées. Bien que ces mesures de performances puissent être extraites de systèmes commerciaux contemporains ou de la littérature, la comparaison n'est alors pas effectuée au même niveau d'abstraction et d'optimisation, réduisant la confiance accordable aux résultats comparatifs.

L'utilisation d'autres outils d'évaluation de performances et simulateurs tels que Gem5-X[Qur+21] pourrait contribuer à surmonter cette limitation, en complétant ou en remplaçant l'implémentation actuelle.

Néanmoins, cette plateforme peut être utilisée pour l'évaluation comparative de plusieurs architectures d'accélérateurs et de technologies mémoire. À titre d'exemple, une architecture d'interpolateur a été étudiée afin de déterminer si les performances actuelles des mémoires ferroélectriques lui permettrait de concurrencer les multiplieurs actuels. Il a ainsi été constaté que cette architecture nécessiterait une mémoire consommant 80 % moins d'énergie pour que cette architecture LiM soit intéressante. De telles améliorations de performances ont été démontrées expérimentalement[Fra+21], ce qui pourrait permettre à cette implémentation, et à d'autres, de concurrencer les implémentations actuelles sur certaines métriques.

6.5 Perspectives à court terme

Les perspectives sont nombreuses, qu'il s'agisse du développement de modèles améliorés facilitant la DSE automatisée, ou de la suite du travail de caractérisation des échantillons fabriqués.

6.5.1 Travaux de caractérisation restants

La bitcell PsFeFET réalisée est intéressante lorsque les FeFET ne sont pas disponibles, et qu'une cellule 2T1C ne fournit pas les performances requises. Une caractérisation plus poussée de cette cellule est attendue, qui comparera notamment les deux orientations FeCap, permettant ainsi de mieux comprendre l'impact des vias métalliques sur la performance.

La cellule 2T1C a été étudiée de manière plus approfondie, et devrait faire l'objet de travaux de caractérisation supplémentaires, notamment concernant ses performances dans le mode de fonctionnement 1T1C. Celle-ci fera également l'objet d'une étude plus détaillée du courant de lecture identifié sous 0 V, comme indiqué dans la figure 3.7.

Le filtre d'image décrit dans la section 4.6 sera également étudié plus en détail, ainsi que la variante multi-étage sans pipeline, afin d'évaluer leur potentielle application aux applications neuromorphiques telles que les CNNs.

6.5.2 Simulations futures

Les résultats de DSE présentés dans la section 5.3.1 doivent faire l'objet d'études plus approfondies, bien que ceux-ci soient probablement dus à des paramètres mal appliqués, ou

à des problèmes liés au modèle de FeCap. Une analyse comparative des différents modèles compacts de FeCaps et de leur impact sur la précision des résultats de DSE déterminerait les compromis acceptables entre performance et précision pour ce cas d'utilisation.

Les évaluations de la performance du système peuvent être revisitées avec des valeurs de performance à jour. Une implémentation d'un filtre convolutif doit également être achevée, celle-ci devant permettre l'exploration de variations architecturales et la comparaison avec les résultats expérimentaux du filtre d'images.

6.6 Considérations sur l'avenir de la technologie ferroélectrique

6.6.1 Compacité

De plus grands FeCap pourraient être étudiés, notamment en vue d'application MLC, tout en réduisant l'empreinte des circuits existants grâce à l'emploi de géométries optimisées. Par exemple, l'emploi de condensateurs à tranchée profonde rend la couche d'oxyde et les électrodes associées verticales plutôt qu'horizontales, réduisant la surface horizontale du dispositif.

D'autres géométries compactes sont également étudiées, dont des nanofils gainés d'oxyde ferroélectrique. Cette approche pourrait par ailleurs être combinée à des transistors à nanofils verticaux[Poi22] pour former des FeFET à nanofils verticaux.

La tension coercitive dépendant fortement de l'orientation du champ, le contrôle de sa direction à l'aide de plusieurs électrodes pourrait ouvrir des alternatives au contrôle fin de la tension de programmation pour le fonctionnement MLC. En outre, les domaines cristallins alignés sur le champ électrique ayant une tension coercitive apparente plus faible, cette approche pourrait permettre la mémorisation de plusieurs bits de données par condensateur selon des axes différents. La polarisation des $HfZrO_2$ ferroélectriques nécessite toutefois des champs électriques d'une intensité relativement élevée de $\sim 1\,\mathrm{MV\,m^{-1}}$[Mul+21 ; KCJ21], généralement obtenus en appliquant une faible tension à travers une fine (environ 10 nm) couche d'oxyde. Ces géométries réduites sont plus faciles à fabriquer verticalement, car l'épaisseur d'oxyde est mieux contrôlée dans cette direction avec les techniques de dépôt que latéralement avec les techniques de lithographie et de traçage. Cet effet pourrait néanmoins être mesuré pour de petits angles de déviation.

6.6.2 Signaux de contrôle

Les dispositifs ferroélectriques de plus grande taille ont des distributions d'E_C plus importantes. En supposant qu'un dispositif ne rencontre pas de défaillances catastrophiques telles que des courts-circuits électriques, et en fonction des mécanismes de fatigue, la fenêtre mémoire peut être maintenue ouverte en augmentant progressivement la tension de fonctionnement lorsque le dispositif vieillit. Cette augmentation de tension permet d'accéder aux domaines ferroélectriques ayant des valeurs V_C plus élevées.

Comme décrit dans la sous-section 2.3.1, il est possible d'utiliser les FeCaps comme des condensateurs ordinaires en maintenant la tension appliquée en dessous de V_C. Lorsque la tension s'approche de la zone d'inversion de polarisation, la capacité devient non linéaire et augmente brusquement. Cela peut constituer un mécanisme de lecture non destructif supplémentaire pour les structures de type 1T1C. Cet effet peut être moins mesurable dans les FeCaps multidomaines, plus grands, où la repolarisation est progressive, ainsi que dans les dispositifs dont le rapport entre la capacité ferroélectrique et paraélectrique est faible.

Le mode de fonctionnement en lecture destructive des FeCap peut être contraignant. En revanche, celui-ci offre également de nouvelles perspectives : les opérations de lecture peuvent être combinées avec une opération d'écriture sans coût supplémentaire, minimisant la latence et l'usure si les deux opérations doivent être effectuées simultanément. Cette propriété peut aussi fournir des garanties de confidentialité supplémentaires dans des contextes sensibles en fournissant un mécanisme de désactivation du circuit WBs, inversant la propriété des mémoires écriture unique, lecture multiple (WORM, Write Once, Read Many) : plusieurs écritures, mais une seule lecture possible. Enfin, des mécanismes de prédiction pourraient réduire le nombre de cycles causés par la lecture destructive des cellules 1T1C. En effet, un cycle WB n'est nécessaire que dans les cas où un FeCap est repolarisé lorsque sa polarisation est mesurée. Une prédiction correcte de la valeur stockée, suivie de la confirmation que

l'application de la tension correspondante ne provoque pas de repolarisation, rend redondante la réécriture de la valeur originale. Cela n'est possible que si les données peuvent être prédites, et un tel mécanisme serait moins efficace avec des cellules MLC. Cependant, une réduction du nombre d'écritures serait bénéfique pour la vitesse de lecture, ainsi que pour l'endurance de la mémoire.

Bibliographie

[AG21] Infineon Technologies AG. *Endurance and Data Retention Characterization of Infineon Flash Memory, Application note AN217979*. 19 avr. 2021. URL : https://www.infineon.com/dgdl/Infineon-AN217979_Endurance_and_Data_Retention_Characterization_of_Infineon_Flash_Memory-ApplicationNotes-v03_00-EN.pdf?fileId=8ac78c8c7cdc391c017d0d30d6b064f5 (visité le 14/04/2023).

[Alc+22] R. Alcala et al. « BEOL Integrated Ferroelectric HfO -Based Capacitors for FeRAM : Extrapolation of Reliability Performance to Use Conditions ». In : *IEEE Journal of the Electron Devices Society* 10 (2022). Conference Name : IEEE Journal of the Electron Devices Society, p. 907-912. ISSN : 2168-6734. DOI : 10.1109/JEDS.2022.3198138.

[Azi+18] A. Aziz et al. « Computing with ferroelectric FETs : Devices, models, systems, and applications ». In : *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*. Mars 2018, p. 1289-1298. DOI : 10.23919/DATE.2018.8342213.

[BD00] William C. Black et Bodhisattva Das. « Programmable logic using giant-magnetoresistance and spin-dependent tunneling devices (invited) ». In : *Journal of Applied Physics* 87 (mai 2000), p. 6674-6679. DOI : 10.1063/1.372806.

[BD20] Julian Blank et Kalyanmoy Deb. « pymoo : Multi-Objective Optimization in Python ». In : *IEEE Access* 8 (2020). Conference Name : IEEE Access, p. 89497-89509. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.2990567. URL : https://ieeexplore.ieee.org/document/9078759/?arnumber=9078759 (visité le 02/02/2025).

[Bec+18] Noah Beck et al. « 'Zeppelin' : An SoC for multichip architectures ». In : *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*. 2018 IEEE International Solid - State Circuits Conference - (ISSCC). ISSN : 2376-8606. Fév. 2018, p. 40-42. DOI : 10.1109/ISSCC.2018.8310173.

[Bey+20] Sven Beyer et al. « FeFET : A versatile CMOS compatible device with game-changing potential ». en. In : *2020 IEEE International Memory Workshop (IMW)*. Dresden, Germany : IEEE, mai 2020, p. 1-4. ISBN : 978-1-72816-306-2. DOI : 10.1109/IMW48823.2020.9108150. URL : https://ieeexplore.ieee.org/document/9108150/ (visité le 21/06/2021).

[BG15] S. Brisset et F. Gillon. « 4 - Approaches for multi-objective optimization in the ecodesign of electric systems ». In : *Eco-Friendly Innovation in Electricity Transmission and Distribution Networks*. Sous la dir. de Jean-Luc Bessède. Oxford : Woodhead Publishing, 1er jan. 2015, p. 83-97. ISBN : 978-1-78242-010-1. DOI : 10.1016/B978-1-78242-010-1.00004-5. URL : https://www.sciencedirect.com/science/article/pii/B9781782420101000045 (visité le 09/10/2022).

[BI13] Mahdi Nazm Bojnordi et Engin Ipek. « DESC : energy-efficient data exchange using synchronized counters ». In : *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture*. MICRO-46. New York, NY, USA : Association for Computing Machinery, 7 déc. 2013, p. 234-246. ISBN : 978-1-4503-2638-4. DOI : 10.1145/2540708.2540729. URL : https://doi.org/10.1145/2540708.2540729 (visité le 09/01/2023).

[Bin+11] Nathan Binkert et al. « The gem5 simulator ». In : *ACM SIGARCH Computer Architecture News* 39.2 (31 août 2011), p. 1-7. ISSN : 0163-5964. DOI : 10.1145/2024716.2024718. URL : https://doi.org/10.1145/2024716.2024718 (visité le 11/02/2023).

[Boh07] Mark Bohr. « A 30 Year Retrospective on Dennard's MOSFET Scaling Paper ». In : *IEEE Solid-State Circuits Newsletter* 12.1 (2007), p. 11-13. ISSN : 1098-4232. DOI : 10.1109/N-SSC.2007.4785534. URL : http://ieeexplore.ieee.org/document/4785534/ (visité le 12/12/2022).

[Bös+11a] T. S. Böscke et al. « Phase transitions in ferroelectric silicon doped hafnium oxide ». In : *Applied Physics Letters* 99.11 (12 sept. 2011). Publisher : American Institute of Physics, p. 112904. ISSN : 0003-6951. DOI : 10.1063/1.3636434. URL : https://aip.scitation.org/doi/10.1063/1.3636434 (visité le 17/11/2020).

[Bös+11b] Tim Böscke et al. « Ferroelectricity in Hafnium Oxide Thin Films ». In : *Applied Physics Letters* 99 (5 sept. 2011), p. 102903-102903. DOI : 10.1063/1.3634052.

[Bou+19] Jordan Bouaziz et al. « Dramatic impact of pressure and annealing temperature on the properties of sputtered ferroelectric HZO layers ». In : *APL Materials* 7.8 (août 2019). Publisher : American Institute of Physics, p. 081109. DOI : 10.1063/1.5110894. URL : https://aip.scitation.org/doi/10.1063/1.5110894 (visité le 17/01/2023).

[Bou20] Jordan Bouaziz. « Mémoires ferroélectriques non-volatiles à base de (Hf,Zr)O2 pour la nanoélectronique basse consommation ». These de doctorat. Lyon, 15 juill. 2020. URL : https://www.theses.fr/2020LYSEI057 (visité le 25/09/2022).

[Bre+18] E. T. Breyer et al. « Demonstration of versatile nonvolatile logic gates in 28nm HKMG FeFET technology ». In : *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2018 IEEE International Symposium on Circuits and Systems (ISCAS). ISSN : 2379-447X. Mai 2018, p. 1-5. DOI : 10.1109/ISCAS.2018.8351408.

[Bri21] Adil Brik. « Méthode de conception des systèmes intégrés multi-physique et continues-discrets ». These de doctorat. Lyon, 21 sept. 2021. URL : https://www.theses.fr/2021LYSEC036 (visité le 25/09/2022).

[Cha+08] Mau-Chung Frank Chang et al. « RF interconnects for communications on-chip ». In : *Proceedings of the 2008 international symposium on Physical design*. ISPD '08. New York, NY, USA : Association for Computing Machinery, 13 avr. 2008, p. 78-83. ISBN : 978-1-60558-048-7. DOI : 10.1145/1353629.1353649. URL : https://doi.org/10.1145/1353629.1353649 (visité le 09/01/2023).

[CHE11] Trevor E. Carlson, Wim Heirman et Lieven Eeckhout. « Sniper : Exploring the level of abstraction for scalable and accurate parallel multi-core simulation ». In : *SC '11 : Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '11 : Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. ISSN : 2167-4337. Nov. 2011, p. 1-12. DOI : 10.1145/2063384.2063454.

[CL07] Premi Chandra et Peter B. Littlewood. « A Landau Primer for Ferroelectrics ». In : *Physics of Ferroelectrics : A Modern Perspective*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 69-116. ISBN : 978-3-540-34591-6. DOI : 10.1007/978-3-540-34591-6_3. URL : https://doi.org/10.1007/978-3-540-34591-6_3.

[CUDA17] *Nvidia CUDA Home Page*. NVIDIA Developer. 18 juill. 2017. URL : https://developer.nvidia.com/cuda-zone (visité le 12/02/2023).

[Dav23] Schiavone Davide. *X-HEEP Github repository*. original-date : 2022-01-07T18 :19 :11Z. 11 fév. 2023. URL : https://github.com/esl-epfl/x-heep (visité le 11/02/2023).

[Deb01] Kalyanmoy Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. Google-Books-ID : OSTn4GSy2uQC. John Wiley & Sons, 5 juill. 2001. 540 p. ISBN : 978-0-471-87339-6.

[Den+20] Shan Deng et al. « A Comprehensive Model for Ferroelectric FET Capturing the Key Behaviors : Scalability, Variation, Stochasticity, and Accumulation ». In : *2020 IEEE Symposium on VLSI Technology*. 2020 IEEE Symposium on VLSI Technology. ISSN : 2158-9682. Juin 2020, p. 1-2. DOI : 10.1109/VLSITechnology18217.2020.9265014.

[Den+22] Benoît W. Denkinger et al. « VWR2A : a very-wide-register reconfigurable-array architecture for low-power embedded devices ». In : *Proceedings of the 59th ACM/IEEE Design Automation Conference*. DAC '22. New York, NY, USA : Association for Computing Machinery, 23 août 2022, p. 895-900. ISBN : 978-1-4503-9142-9. DOI : 10.1145/3489517.3530980. URL : https://doi.org/10.1145/3489517.3530980 (visité le 11/02/2023).

[Den+74] R.H. Dennard et al. « Design of ion-implanted MOSFET's with very small physical dimensions ». In : *IEEE Journal of Solid-State Circuits* 9.5 (oct. 1974). Conference Name : IEEE Journal of Solid-State Circuits, p. 256-268. ISSN : 1558-173X. DOI : 10.1109/JSSC.1974.1050511.

[Dev49] A.F. Devonshire. « XCVI. Theory of barium titanate ». In : *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 40.309 (1949), p. 1040-1063. DOI : 10.1080/14786444908561372. eprint : https://doi.org/10.1080/14786444908561372. URL : https://doi.org/10.1080/14786444908561372.

[Don+12] Xiangyu Dong et al. « NVSim : A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory ». In : *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 31.7 (juill. 2012). Conference Name : IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, p. 994-1007. ISSN : 1937-4151. DOI : 10.1109/TCAD.2012.2185930.

[Dup22] Etienne Dupuis. « Weight-sharing methods for retraining-free CNN compression ». These de doctorat. Université de Lyon, 19 mai 2022. URL : https://www.theses.fr/2022LYSEC017 (visité le 25/09/2022).

[Dut+22] Sourav Dutta et al. « Logic Compatible High-Performance Ferroelectric Transistor Memory ». In : *IEEE Electron Device Letters* 43.3 (mars 2022). Conference Name : IEEE Electron Device Letters, p. 382-385. ISSN : 1558-0563. DOI : 10.1109/LED.2022.3148669.

[Esp22] Espressif. *ESP8266EX Datasheet, v6.8*. Oct. 2022. URL : https://www.espressif.com/sites/default/files/documentation/0a-esp8266ex_datasheet_en.pdf (visité le 11/12/2022).

[FKK12] V. Fridkin, M. Kuehn et H. Kliem. « The Weiss model and the Landau–Khalatnikov model for the switching of ferroelectrics ». In : *Physica B : Condensed Matter* 407.12 (15 juin 2012), p. 2211-2214. ISSN : 0921-4526. DOI : 10.1016/j.physb.2012.02.043. URL : https://www.sciencedirect.com/science/article/pii/S0921452612002311 (visité le 25/02/2023).

[Fra+19a] T. Francois et al. « Demonstration of BEOL-compatible ferroelectric Hf0.5Zr0.5O2 scaled FeRAM co-integrated with 130nm CMOS for embedded NVM applications ». In : *2019 IEEE International Electron Devices Meeting (IEDM)*. 2019 IEEE International Electron Devices Meeting (IEDM). ISSN : 2156-017X. Déc. 2019, p. 15.7.1-15.7.4. DOI : 10.1109/IEDM19573.2019.8993485.

[Fra+19b] T. Francois et al. « Ferroelectric HfO2 for Memory Applications : Impact of Si Doping Technique and Bias Pulse Engineering on Switching Performance ». In : *2019 IEEE 11th International Memory Workshop (IMW)*. 2019 IEEE 11th International Memory Workshop (IMW). ISSN : 2573-7503. Mai 2019, p. 1-4. DOI : 10.1109/IMW.2019.8739664.

[Fra+21] T. Francois et al. « 16kbit HfO $_2$:Si-based 1T-1C FeRAM Arrays Demonstrating High Performance Operation and Solder Reflow Compatibility ». en. In : *2021 IEEE International Electron Devices Meeting (IEDM)*. San Francisco, CA, USA : IEEE, déc. 2021, p. 33.1.1-33.1.4. ISBN : 978-1-66542-572-8. DOI : 10.1109/IEDM19574.2021.9720640. URL : https://ieeexplore.ieee.org/document/9720640/ (visité le 28/07/2022).

[Fra12] Felipe Frantz Ferreira. « Architectural exploration methods and tools for heterogeneous 3D-IC ». These de doctorat. Ecully, Ecole centrale de Lyon, 26 oct. 2012. URL : https://www.theses.fr/2012ECDL0033 (visité le 09/10/2022).

[FS19] Shosuke Fujii et Masumi Saitoh. « Chapter 10.3 - Ferroelectric Tunnel Junction ». In : *Ferroelectricity in Doped Hafnium Oxide : Materials, Properties and Devices*. Sous la dir. d'Uwe Schroeder, Cheol Seong Hwang et Hiroshi Funakubo. Woodhead Publishing Series in Electronic and Optical Materials. Woodhead Publishing, 2019, p. 437-449. ISBN : 978-0-08-102430-0. DOI : https://doi.org/10.1016/B978-0-08-102430-0.00021-8. URL : https://www.sciencedirect.com/science/article/pii/B9780081024300000218.

[Gal+19a] W.J. Gallagher et al. « 22nm STT-MRAM for Reflow and Automotive Uses with High Yield, Reliability, and Magnetic Immunity and with Performance and Shielding Options ». In : *2019 IEEE International Electron Devices Meeting (IEDM)*. 2019 IEEE International Electron Devices Meeting (IEDM). ISSN : 2156-017X. Déc. 2019, p. 2.7.1-2.7.4. DOI : 10.1109/IEDM19573.2019.8993469.

[Gal+19b] W.J. Gallagher et al. « Recent Progress and Next Directions for Embedded MRAM Technology ». In : *2019 Symposium on VLSI Technology*. 2019 Symposium on VLSI Technology. ISSN : 2158-9682. Juin 2019, T190-T191. DOI : 10.23919/VLSIT.2019.8776547.

[Gas+19] C. Gastaldi et al. « Transient Negative Capacitance of Silicon-doped HfO2 in MFMIS and MFIS structures : experimental insights for hysteresis-free steep slope NC FETs ». In : *2019 IEEE International Electron Devices Meeting (IEDM)*. 2019 IEEE International Electron Devices Meeting (IEDM). ISSN : 2156-017X. Déc. 2019, p. 23.5.1-23.5.4. DOI : 10.1109/IEDM19573.2019.8993523.

[GB14] Vincent Garcia et Manuel Bibes. « Ferroelectric tunnel junctions for information storage and processing ». In : *Nature Communications* 5.1 (24 juill. 2014). Number : 1 Publisher : Nature Publishing Group, p. 4289. ISSN : 2041-1723. DOI : 10.1038/ncomms5289. URL : https://www.nature.com/articles/ncomms5289/ (visité le 25/02/2023).

[Gir+21] Patrick Girard et al. « A Survey of Test and Reliability Solutions for Magnetic Random Access Memories ». In : *Proceedings of the IEEE* 109.2 (fév. 2021). Conference Name : Proceedings of the IEEE, p. 149-169. ISSN : 1558-2256. DOI : 10.1109/JPROC.2020.3029600.

[Giu21] Gino Giusi. « Floating Body DRAM with Body Raised and Source/Drain Separation ». In : *Electronics* 10.6 (jan. 2021). Number : 6 Publisher : Multidisciplinary Digital Publishing Institute, p. 706. ISSN : 2079-9292. DOI : 10.3390/electronics10060706. URL : https://www.mdpi.com/2079-9292/10/6/706 (visité le 15/09/2022).

[Gre+20] L. Grenouillet et al. « Performance assessment of BEOL-integrated HfO2-based ferroelectric capacitors for FeRAM memory arrays ». In : *2020 IEEE Silicon Nanoelectronics Workshop (SNW)*. 2020 IEEE Silicon Nanoelectronics Workshop (SNW). ISSN : 2161-4644. Juin 2020, p. 5-6. DOI : 10.1109/SNW50361.2020.9131648.

[HCMOS] *EUROPRACTICE | STMicroelectronics*. URL : https://europractice-ic.com/technologies/asics/stmicroelectronics/ (visité le 24/02/2023).

[HIP23] *HIP : C++ Heterogeneous-Compute Interface for Portability*. original-date : 2016-01-07T17 :41 :56Z. 10 fév. 2023. URL : https://github.com/ROCm-Developer-Tools/HIP (visité le 12/02/2023).

[HKMG20] *28nm HKMG Technologies | GLOBALFOUNDRIES*. 9 déc. 2020. URL : https://web.archive.org/web/20201209211036/https://www.globalfoundries.com/technology-solutions/cmos/fdx/28nm-hkmg-technologies (visité le 24/02/2023).

[Hon+01] Seungbum HONG et al. « Principle of ferroelectric domain imaging using atomic force microscope ». en. In : *Journal of Applied Physics* 89.2 (jan. 2001), p. 1377-1386. ISSN : 0021-8979, 1089-7550. DOI : 10.1063/1.1331654. URL : http://aip.scitation.org/doi/10.1063/1.1331654 (visité le 28/07/2022).

[Ihl19] Jon F. IHLEFELD. « Chapter 1 - Fundamentals of Ferroelectric and Piezoelectric Properties ». In : *Ferroelectricity in Doped Hafnium Oxide : Materials, Properties and Devices*. Sous la dir. d'Uwe SCHROEDER, Cheol Seong HWANG et Hiroshi FUNAKUBO. Woodhead Publishing Series in Electronic and Optical Materials. Woodhead Publishing, 2019, p. 1-24. ISBN : 978-0-08-102430-0. DOI : https://doi.org/10.1016/B978-0-08-102430-0.00001-2. URL : https://www.sciencedirect.com/science/article/pii/B9780081024300000012.

[Ike+20] Sumio IKEGAWA et al. « Magnetoresistive Random Access Memory : Present and Future ». In : *IEEE Transactions on Electron Devices* 67.4 (avr. 2020). Conference Name : IEEE Transactions on Electron Devices, p. 1407-1419. ISSN : 1557-9646. DOI : 10.1109/TED.2020.2965403.

[IRDS22] *International Roadmap for Devices and Systems (IRDS™) 2022 Edition - IEEE IRDS™*. 2022. URL : https://irds.ieee.org/editions/2022 (visité le 25/02/2023).

[J14] Matt J. *Simple 1D polynomial fitting with particular coefficients constrained to zero*. Mars 2014. URL : https://www.mathworks.com/matlabcentral/answers/123072-curve-fitting-tool-with-custom-equation-odd-power-polynomial#answer_130371 (visité le 21/06/2021).

[Jao+21] Nicholas JAO et al. « Design Space Exploration of Ferroelectric Tunnel Junction Toward Crossbar Memories ». In : *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* 7.2 (1er déc. 2021). Publisher : Institute of Electrical and Electronics Engineers. ISSN : 2329-9231. DOI : 10.1109/JXCDC.2021.3117566. URL : https://www.osti.gov/pages/biblio/1829766 (visité le 22/02/2023).

[Jer+17] Matthew JERRY et al. « Ferroelectric FET analog synapse for acceleration of deep neural network training ». In : *2017 IEEE International Electron Devices Meeting (IEDM)*. 2017 IEEE International Electron Devices Meeting (IEDM). ISSN : 2156-017X. Déc. 2017, p. 6.2.1-6.2.4. DOI : 10.1109/IEDM.2017.8268338.

[JGL16] Mansi JHAMB, GARIMA et Himanshu LOHANI. « Design, implementation and performance comparison of multiplier topologies in power-delay space ». en. In : *Engineering Science and Technology, an International Journal* 19.1 (mars 2016), p. 355-363. ISSN : 22150986. DOI : 10.1016/j.jestch.2015.08.006. URL : https://linkinghub.elsevier.com/retrieve/pii/S2215098615001287 (visité le 31/01/2020).

[Kap13] KAPOOHT. *Von Neumann Architecture Diagramm*. 28 avr. 2013. URL : https://commons.wikimedia.org/wiki/File:Von_Neumann_Architecture.svg (visité le 22/12/2022).

[Kaz+21a] Arman KAZEMI et al. « In-Memory Nearest Neighbor Search with FeFET Multi-Bit Content-Addressable Memories ». In : *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE). ISSN : 1558-1101. Fév. 2021, p. 1084-1089. DOI : 10.23919/DATE51398.2021.9474025.

[Kaz+21b] Arman KAZEMI et al. « MIMHD : Accurate and Efficient Hyperdimensional Inference Using Multi-Bit In-Memory Computing ». In : *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. 2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED). Juill. 2021, p. 1-6. DOI : 10.1109/ISLPED52811.2021.9502498.

[Kaz+22] Arman KAZEMI et al. « Achieving software-equivalent accuracy for hyperdimensional computing with ferroelectric-based in-memory computing ». In : *Scientific Reports* 12.1 (10 nov. 2022). Number : 1 Publisher : Nature Publishing Group, p. 19201. ISSN : 2045-2322. DOI : 10.1038/s41598-022-23116-w. URL : https://www.nature.com/articles/s41598-022-23116-w (visité le 11/02/2023).

[KCJ21] Jae Young KIM, Min-Ju CHOI et Ho Won JANG. « Ferroelectric field effect transistors : Progress and perspective ». In : *APL Materials* 9.2 (fév. 2021). Publisher : American Institute of Physics, p. 021102. DOI : 10.1063/5.0035515. URL : https://aip.scitation.org/doi/10.1063/5.0035515 (visité le 16/04/2023).

[Kle+21] Dominik KLEIMAIER et al. « Demonstration of a p-Type Ferroelectric FET With Immediate Read-After-Write Capability ». In : *IEEE Electron Device Letters* 42.12 (déc. 2021). Conference Name : IEEE Electron Device Letters, p. 1774-1777. ISSN : 1558-0563. DOI : 10.1109/LED.2021.3118645.

[KN03] Hiroaki KATO et Hiroshi NOZAWA. « Proposal for 1T/1C Ferroelectric Random Access Memory with Multiple Storage and Application to Functional Memory ». In : *Japanese Journal of Applied Physics* 42 (sept. 2003), p. 5998-6002.

[Koo+18] Maha KOOLI et al. « Smart instruction codes for in-memory computing architectures compatible with standard SRAM interfaces ». In : *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE). ISSN : 1558-1101. Mars 2018, p. 1634-1639. DOI : 10.23919/DATE.2018.8342276.

[Led+20] Maximilian LEDERER et al. « Structural and Electrical Comparison of Si and Zr Doped Hafnium Oxide Thin Films and Integrated FeFETs Utilizing Transmission Kikuchi Diffraction ». In : *Nanomaterials* 10.2 (fév. 2020). Number : 2 Publisher : Multidisciplinary Digital Publishing Institute, p. 384. ISSN : 2079-4991. DOI : 10.3390/nano10020384. URL : https://www.mdpi.com/2079-4991/10/2/384 (visité le 22/02/2023).

[Led+21] M. LEDERER et al. « Impact of the SiO $_2$ interface layer on the crystallographic texture of ferroelectric hafnium oxide ». In : *Applied Physics Letters* 118.1 (4 jan. 2021), p. 012901. ISSN : 0003-6951, 1077-3118. DOI : 10.1063/5.0029635. URL : https://aip.scitation.org/doi/10.1063/5.0029635 (visité le 22/02/2023).

[Leh+21] David LEHNINGER et al. « Enabling Ferroelectric Memories in BEoL - towards advanced neuromorphic computing architectures ». In : *2021 IEEE International Interconnect Technology Conference (IITC)*. 2021 IEEE International Interconnect Technology Conference (IITC). Kyoto, Japan : IEEE, 6 juill. 2021, p. 1-4. ISBN : 978-1-72817-632-1. DOI : 10.1109/IITC51362.2021.9537346. URL : https://ieeexplore.ieee.org/document/9537346/ (visité le 05/07/2022).

[LFZ11] Zhichao LU, Jerry G. FOSSUM et Zhenming ZHOU. « A Floating-Body/Gate DRAM Cell Upgraded for Long Retention Time ». In : *IEEE Electron Device Letters* 32.6 (juin 2011). Conference Name : IEEE Electron Device Letters, p. 731-733. ISSN : 1558-0563. DOI : 10.1109/LED.2011.2134065.

[LHS22] You-Sheng LIU, Yuan-Yu HUANG et Pin SU. « Design Space Exploration for Scaled FeFET Nonvolatile Memories : High-k Spacer as a Powerful Aid ». In : *2022 6th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*. 2022 6th IEEE Electron Devices Technology & Manufacturing Conference (EDTM). Mars 2022, p. 70-72. DOI : 10.1109/EDTM53872.2022.9798076.

[Li+09] Sheng LI et al. « McPAT : an integrated power, area, and timing modeling framework for multicore and manycore architectures ». In : *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*. MICRO 42. New York, NY, USA : Association for Computing Machinery, 12 déc. 2009, p. 469-480. ISBN : 978-1-60558-798-1. DOI : 10.1145/1669112.1669172. URL : https://doi.org/10.1145/1669112.1669172 (visité le 11/02/2023).

[Liu+14] Wulong LIU et al. « Exploration of Electrical and Novel Optical Chip-to-Chip Interconnects ». In : *IEEE Design & Test* 31.5 (oct. 2014), p. 28-35. ISSN : 2168-2356, 2168-2364. DOI : 10.1109/MDAT.2014.2336217. URL : http://ieeexplore.ieee.org/document/6849444/ (visité le 25/02/2023).

[Mag+02] P.S. MAGNUSSON et al. « Simics : A full system simulation platform ». In : *Computer* 35.2 (fév. 2002). Conference Name : Computer, p. 50-58. ISSN : 1558-0814. DOI : 10.1109/2.982916.

[Maj+18] Sayani MAJUMDAR et al. « Electrode Dependence of Tunneling Electroresistance and Switching Stability in Organic Ferroelectric P(VDF-TrFE)-Based Tunnel Junctions ». In : *Advanced Functional Materials* 28.15 (2018). _eprint : https ://onlinelibrary.wiley.com/do p. 1703273. ISSN : 1616-3028. DOI : 10.1002/adfm.201703273. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201703273 (visité le 25/02/2023).

[Maj22] Sayani MAJUMDAR. « Back-End CMOS Compatible and Flexible Ferroelectric Memories for Neuromorphic Computing and Adaptive Sensing ». In : *Advanced Intelligent Systems* 4.4 (avr. 2022), p. 2100175. ISSN : 2640-4567, 2640-4567. DOI : 10.1002/aisy.202100175. URL : https://onlinelibrary.wiley.com/doi/10.1002/aisy.202100175 (visité le 05/07/2022).

[Mam+21] Kévin MAMBU et al. « Instruction Set Design Methodology for In-Memory Computing through QEMU-based System Emulator ». In : *2021 IEEE International Workshop on Rapid System Prototyping (RSP)*. 2021 IEEE International Workshop on Rapid System Prototyping (RSP). ISSN : 2150-5519. Oct. 2021, p. 43-49. DOI : 10.1109/RSP53691.2021.9806255.

[Mar+21] Cédric MARCHAND et al. « FeFET based Logic-in-Memory : an overview ». In : *2021 16th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS)*. 2021 16th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS). Juin 2021, p. 1-6. DOI : 10.1109/DTIS53253.2021.9505078.

[Mar+22] Cédric MARCHAND et al. « A FeFET-Based Hybrid Memory Accessible by Content and by Address ». In : *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* 8 (1er juin 2022), p. 1-1. DOI : 10.1109/JXCDC.2022.3168057.

[Mas+21] A. G. MASLOVSKAYA et al. « Theoretical and numerical analysis of the Landau–Khalatnikov model of ferroelectric hysteresis ». In : *Communications in Nonlinear Science and Numerical Simulation* 93 (1er fév. 2021), p. 105524. ISSN : 1007-5704. DOI : 10.1016/j.cnsns.2020.105524. URL : https://www.sciencedirect.com/science/article/pii/S1007570420303543 (visité le 25/02/2023).

[MG19] T.P. MA et Nanbo GONG. « Retention and Endurance of FeFET Memory Cells ». In : *2019 IEEE 11th International Memory Workshop (IMW)*. 2019 IEEE 11th International Memory Workshop (IMW). ISSN : 2573-7503. Mai 2019, p. 1-4. DOI : 10.1109/IMW.2019.8739726.

[Mik+19] T. MIKOLAJICK et al. « Next Generation Ferroelectric Memories enabled by Hafnium Oxide ». In : *2019 IEEE International Electron Devices Meeting (IEDM)*. 2019 IEEE International Electron Devices Meeting (IEDM). ISSN : 2156-017X. Déc. 2019, p. 15.5.1-15.5.4. DOI : 10.1109/IEDM19573.2019.8993447.

[Mil+90] S. L. MILLER et al. « Device modeling of ferroelectric capacitors ». In : *Journal of Applied Physics* 68.12 (15 déc. 1990). Publisher : American Institute of Physics, p. 6463-6471. ISSN : 0021-8979. DOI : 10.1063/1.346845. URL : https://aip.scitation.org/doi/10.1063/1.346845 (visité le 26/02/2023).

[Moo65] Gordon E. MOORE. « Cramming more components onto integrated circuits ». In : *Electronics* 38.8 (1965), p. 114-117. URL : https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf.

[Moz21] Luca MOZZONE. « Design of a benchmarking platform for Logic-In-Memory architectures based on ferroelectric HfO2 ». laurea. Politecnico di Torino, 16 avr. 2021. 69 p. URL : https://webthesis.biblio.polito.it/17853/ (visité le 25/09/2022).

[Mue+13a] Stefan MUELLER et al. « From MFM Capacitors Toward Ferroelectric Transistors : Endurance and Disturb Characteristics of HfO_2-Based FeFET Devices ». In : *IEEE Transactions on Electron Devices* 60.12 (déc. 2013). Conference Name : IEEE Transactions on Electron Devices, p. 4199-4205. ISSN : 1557-9646. DOI : 10.1109/TED.2013.2283465.

[Mue+13b] Stefan MUELLER et al. « Reliability Characteristics of Ferroelectric Si : HfO_2 Thin Films for Memory Applications ». In : *IEEE Transactions on Device and Materials Reliability* 13.1 (mars 2013). Conference Name : IEEE Transactions on Device and Materials Reliability, p. 93-97. ISSN : 1558-2574. DOI : 10.1109/TDMR.2012.2216269.

[Mül+11] Johannes MÜLLER et al. « Ferroelectric Zr0.5Hf0.5O2 thin films for nonvolatile memory applications ». In : *Applied Physics Letters* 99 (1er sept. 2011), p. 112901-112901. DOI : 10.1063/1.3636417.

[Mül+15] J. MÜLLER et al. « Ferroelectric Hafnium Oxide Based Materials and Devices : Assessment of Current Status and Future Prospects ». In : *ECS Journal of Solid State Science and Technology* 4.5 (2015), N30-N35. ISSN : 2162-8769, 2162-8777. DOI : 10.1149/2.0081505jss. URL : http://jss.ecsdl.org/lookup/doi/10.1149/2.0081505jss (visité le 01/04/2019).

[Mul+17] H. MULAOSMANOVIC et al. « Novel ferroelectric FET based synapse for neuromorphic systems ». In : *2017 Symposium on VLSI Technology*. 2017 Symposium on VLSI Technology. ISSN : 2158-9682. Juin 2017, T176-T177. DOI : 10.23919/VLSIT.2017.7998165.

[Mul+21] Halid MULAOSMANOVIC et al. « Ferroelectric field-effect transistors based on HfO_2 : a review ». In : *Nanotechnology* 32.50 (10 déc. 2021), p. 502002. ISSN : 0957-4484, 1361-6528. DOI : 10.1088/1361-6528/ac189f. URL : https://iopscience.iop.org/article/10.1088/1361-6528/ac189f (visité le 14/02/2023).

[Mül+21] S. MÜLLER et al. « Development status of gate-first FeFET technology ». In : *2021 Symposium on VLSI Technology*. 2021 Symposium on VLSI Technology. ISSN : 2158-9682. Juin 2021, p. 1-2.

[Ni+18] K. NI et al. « SoC Logic Compatible Multi-Bit FeMFET Weight Cell for Neuromorphic Applications ». In : *2018 IEEE International Electron Devices Meeting (IEDM)*. 2018 IEEE International Electron Devices Meeting (IEDM). ISSN : 2156-017X. Déc. 2018, p. 13.2.1-13.2.4. DOI : 10.1109/IEDM.2018.8614496.

[Ni+19] Kai NI et al. « Ferroelectric ternary content-addressable memory for one-shot learning ». In : *Nature Electronics* 2.11 (nov. 2019). Number : 11 Publisher : Nature Publishing Group, p. 521-529. ISSN : 2520-1131. DOI : 10.1038/s41928-019-0321-3. URL : https://www.nature.com/articles/s41928-019-0321-3 (visité le 12/02/2023).

[Nie+23] Michael NIEMIER et al. « Cross Layer Design for the Predictive Assessment of Technology-Enabled Architectures.pdf ». In : *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). 18 avr. 2023.

[Now+16] Janusz J. NOWAK et al. « Dependence of Voltage and Size on Write Error Rates in Spin-Transfer Torque Magnetic Random-Access Memory ». In : *IEEE Magnetics Letters* 7 (2016). Conference Name : IEEE Magnetics Letters, p. 1-4. ISSN : 1949-3088. DOI : 10.1109/LMAG.2016.2539256.

[OCo+18] Ian O'CONNOR et al. « Prospects for energy-efficient edge computing with integrated HfO2-based ferroelectric devices ». In : *2018 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*. Oct. 2018, p. 180-183. DOI : 10.1109/VLSI-SoC.2018.8644809.

[Oku+21] Jun OKUNO et al. « High-Endurance and Low-Voltage operation of 1T1C FeRAM Arrays for Nonvolatile Memory Application ». In : *2021 IEEE International Memory Workshop (IMW)*. 2021 IEEE International Memory Workshop (IMW). ISSN : 2573-7503. Mai 2021, p. 1-3. DOI : 10.1109/IMW51353.2021.9439595.

[OMP] *The OpenMP API specification for parallel programming Home Page*. OpenMP. URL : https://www.openmp.org/ (visité le 12/02/2023).

[Pal+18] Ashish PAL et al. « Scaling NC-FinFET to Sub-3 nm Nodes ». In : *2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*. 2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S). ISSN : 2573-5926. Oct. 2018, p. 1-2. DOI : 10.1109/S3S.2018.8640184.

[Paw22] Richard PAWSON. « The Myth of the Harvard Architecture ». In : *IEEE Annals of the History of Computing* 44.3 (juill. 2022). Conference Name : IEEE Annals of the History of Computing, p. 59-69. ISSN : 1934-1547. DOI : 10.1109/MAHC.2022.3175612.

[Pen+22] Lillian PENTECOST et al. « NVMExplorer : A Framework for Cross-Stack Comparisons of Embedded Non-Volatile Memories ». In : *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA). ISSN : 2378-203X. Avr. 2022, p. 938-956. DOI : 10.1109/HPCA53966.2022.00073.

[Peš+17] Milan PEŠIĆ et al. « A computational study of hafnia-based ferroelectric memories : from ab initio via physical modeling to circuit models of ferroelectric device ». In : *Journal of Computational Electronics* 16.4 (1er déc. 2017), p. 1236-1256. ISSN : 1572-8137. DOI : 10.1007/s10825-017-1053-0. URL : https://doi.org/10.1007/s10825-017-1053-0 (visité le 19/09/2022).

[PLH21] Hyeon Woo PARK, Jae-Gil LEE et Cheol Seong HWANG. « Review of ferroelectric field-effect transistors for three-dimensional storage applications ». In : *Nano Select* 2.6 (2021). _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/nano.202000281, p. 1187-1207. ISSN : 2688-4011. DOI : 10.1002/nano.202000281. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/nano.202000281 (visité le 25/02/2023).

[Poi22] Arnaud POITTEVIN. « Logic circuits based on Vertical nanowire FETs : Physical and circuit design challenges and opportunities ». These de doctorat. Lyon, 23 juin 2022. URL : https://www.theses.fr/2022LYSEC024 (visité le 23/02/2023).

[Por+15] Matt POREMBA et al. « DESTINY : A tool for modeling emerging 3D NVM and eDRAM caches ». In : *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE). ISSN : 1558-1101. Mars 2015, p. 1543-1546. DOI : 10.7873/DATE.2015.0733.

[Pre35] F. PREISACH. « Über die magnetische Nachwirkung ». In : *Zeitschrift für Physik* 94.5 (1er mai 1935), p. 277-302. ISSN : 0044-3328. DOI : 10.1007/BF01349418. URL : https://doi.org/10.1007/BF01349418 (visité le 20/09/2022).

[Qur+21] Yasir Mahmood QURESHI et al. « Gem5-X : A Many-core Heterogeneous Simulation Platform for Architectural Exploration and Optimization ». In : *ACM Transactions on Architecture and Code Optimization* 18.4 (17 juill. 2021), 44 :1-44 :27. ISSN : 1544-3566. DOI : 10.1145/3461662. URL : https://doi.org/10.1145/3461662 (visité le 11/02/2023).

[Rag+17] Jonathan RAGAN-KELLEY et al. « Halide : decoupling algorithms from schedules for high-performance image processing ». In : *Communications of the ACM* 61.1 (27 déc. 2017), p. 106-115. ISSN : 0001-0782, 1557-7317. DOI : 10.1145/3150211. URL : https://dl.acm.org/doi/10.1145/3150211 (visité le 12/02/2023).

[Rav+19] Taras Ravsher et al. « Adoption of 2T2C ferroelectric memory cells for logic operation ». In : *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. 2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS). Genoa, Italy : IEEE, nov. 2019, p. 791-794. ISBN : 978-1-72810-996-1. DOI : 10.1109/ICECS46596.2019.8965155. URL : https://ieeexplore.ieee.org/document/8965155/ (visité le 21/06/2021).

[Rei+19] Dayane Reis et al. « Design and Analysis of an Ultra-Dense, Low-Leakage, and Fast FeFET-Based Random Access Memory Array ». In : *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* PP (22 juill. 2019), p. 1-1. DOI : 10.1109/JXCDC.2019.2930284.

[Sch22] David Schor. *IEDM 2022 : Did We Just Witness The Death Of SRAM ?* WikiChip Fuse. Section : Foundries. 14 déc. 2022. URL : https://fuse.wikichip.org/news/7343/iedm-2022-did-we-just-witness-the-death-of-sram/ (visité le 09/01/2023).

[SD08] Sayeef Salahuddin et Supriyo Datta. « Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices ». In : *Nano Letters* 8.2 (fév. 2008), p. 405-410. ISSN : 1530-6984. DOI : 10.1021/nl071804g. URL : https://doi.org/10.1021/nl071804g.

[SG96] A. Sheikholeslami et P.G. Gulak. « Transient modeling of ferroelectric capacitors for nonvolatile memories ». In : *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 43.3 (mai 1996). Conference Name : IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, p. 450-456. ISSN : 1525-8955. DOI : 10.1109/58.489404.

[SHF19] Uwe Schroeder, Cheol Seong Hwang et Hiroshi Funakubo. « Preface ». In : *Ferroelectricity in Doped Hafnium Oxide : Materials, Properties and Devices*. Sous la dir. d'Uwe Schroeder, Cheol Seong Hwang et Hiroshi Funakubo. Woodhead Publishing Series in Electronic and Optical Materials. Woodhead Publishing, 2019, p. xvii-xviii. ISBN : 978-0-08-102430-0. DOI : https://doi.org/10.1016/B978-0-08-102430-0.09987-3. URL : https://www.sciencedirect.com/science/article/pii/B9780081024300099873.

[Sin+] Deepa Sinha et al. « New Design for Low Power High Performance 8T Full Adder ». en. In : (), p. 4.

[Sle+19a] Stefan Slesazeck et al. « A 2TnC ferroelectric memory gain cell suitable for compute-in-memory and neuromorphic application ». In : *2019 IEEE International Electron Devices Meeting (IEDM)*. 2019 IEEE International Electron Devices Meeting (IEDM). San Francisco, CA, USA : IEEE, déc. 2019, p. 38.6.1-38.6.4. ISBN : 978-1-72814-032-2. DOI : 10.1109/IEDM19573.2019.8993663. URL : https://ieeexplore.ieee.org/document/8993663/ (visité le 21/06/2021).

[Sle+19b] Stefan Slesazeck et al. « Uniting The Trinity of Ferroelectric HfO_2 Memory Devices in a Single Memory Cell ». In : *2019 IEEE 11th International Memory Workshop (IMW)*. Monterey, CA, USA : IEEE, mai 2019, p. 1-4. ISBN : 978-1-72810-981-7. DOI : 10.1109/IMW.2019.8739742. URL : https://ieeexplore.ieee.org/document/8739742/ (visité le 12/01/2020).

[SP21] Stefan Slesazeck et Milan PESIC. « Ferroelectric memory and logic cell and operation method ». Brev. amér. 11205467B2. Namlab GmbH. 21 déc. 2021. URL : https://patents.google.com/patent/US11205467B2/en (visité le 28/04/2023).

[SR17] Ankit Sharma et Kaushik Roy. « Design Space Exploration of Hysteresis-Free HfZrOx-Based Negative Capacitance FETs ». In : *IEEE Electron Device Letters* 38.8 (août 2017). Conference Name : IEEE Electron Device Letters, p. 1165-1167. ISSN : 1558-0563. DOI : 10.1109/LED.2017.2714659.

[SYCL14] *SYCL - C++ Single-source Heterogeneous Programming for Acceleration Offload*. The Khronos Group. Section : API. 20 jan. 2014. URL : https://www.khronos.org/sycl/ (visité le 12/02/2023).

[Syn21] SYNOPSYS. *What is Design Space Optimization (DSO) ? – How It Works ? | Synopsys*. 12 oct. 2021. URL : https://www.synopsys.com/glossary/what-is-design-space-optimization.html (visité le 26/09/2022).

[Vog10] Thomas VOGELSANG. « Understanding the Energy Consumption of Dynamic Random Access Memories ». In : *2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*. Atlanta, GA, USA : IEEE, déc. 2010, p. 363-374. ISBN : 978-1-4244-9071-4. DOI : 10.1109/MICRO.2010.42. URL : http://ieeexplore.ieee.org/document/5695550/ (visité le 31/01/2020).

[WA17] Muhammad Abdul WAHAB et Muhammad A. ALAM. *A Verilog-A Compact Model for Negative Capacitance FET*. Nov. 2017. DOI : doi:/10.4231/D3QZ22K3Z. URL : https://nanohub.org/publications/95/5.

[Wat16] Andrew WATERMAN. « Design of the RISC-V Instruction Set Architecture ». Thèse de doct. EECS Department, University of California, Berkeley, jan. 2016. URL : http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-1.html.

[Wil] Stephen WILLIAMS. *Icarus Verilog Home Page*. URL : http://iverilog.icarus.com/ (visité le 12/02/2023).

[Win+20] Jasper de WINKEL et al. « Battery-Free Game Boy ». In : *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.3 (4 sept. 2020), p. 1-34. ISSN : 2474-9567. DOI : 10.1145/3411839. URL : https://dl.acm.org/doi/10.1145/3411839 (visité le 10/03/2023).

[WSW00] Robert M. WALLACE, Richard A. STOLTZ et Glen D. WILK. « Zirconium and/or hafnium oxynitride gate dielectric ». Brev. amér. 6013553A. Texas Instruments INC. 11 jan. 2000. URL : https://patents.google.com/patent/US6013553A/en (visité le 21/11/2022).

[Wuu+22] John WUU et al. « 3D V-Cache : the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU ». In : *2022 IEEE International Solid- State Circuits Conference (ISSCC)*. 2022 IEEE International Solid- State Circuits Conference (ISSCC). T. 65. ISSN : 2376-8606. Fév. 2022, p. 428-429. DOI : 10.1109/ISSCC42614.2022.9731565.

[Yin+16] Xunzhao YIN et al. « Exploiting ferroelectric FETs for low-power non-volatile logic-in-memory circuits ». en. In : *Proceedings of the 35th International Conference on Computer-Aided Design - ICCAD '16*. Austin, Texas : ACM Press, 2016, p. 1-8. ISBN : 978-1-4503-4466-1. DOI : 10.1145/2966986.2967037. URL : http://dl.acm.org/citation.cfm?doid=2966986.2967037 (visité le 01/04/2019).

[Yin+19] Xunzhao YIN et al. « An Ultra-Dense 2FeFET TCAM Design Based on a Multi-Domain FeFET Model ». In : *IEEE Transactions on Circuits and Systems II : Express Briefs* 66.9 (sept. 2019). Conference Name : IEEE Transactions on Circuits and Systems II : Express Briefs, p. 1577-1581. ISSN : 1558-3791. DOI : 10.1109/TCSII.2018.2889225.

[Yoo+19] Insik YOON et al. « Design space exploration of Ferroelectric FET based Processing-in-Memory DNN Accelerator ». In : *ArXiv* (12 août 2019). URL : https://www.semanticscholar.org/paper/Design-space-exploration-of-Ferroelectric-FET-based-Yoon-Jerry/fd7d0de5ed04be930f5a4afb3c71d51a1001f967 (visité le 22/02/2023).

[YS17] Wei-Xiang YOU et Pin SU. « Design Space Exploration Considering Back-Gate Biasing Effects for 2D Negative-Capacitance Field-Effect Transistors ». In : *IEEE Transactions on Electron Devices* 64.8 (août 2017). Conference Name : IEEE Transactions on Electron Devices, p. 3476-3481. ISSN : 1557-9646. DOI : 10.1109/TED.2017.2714687.

[Zac+22] Christina ZACHARAKI et al. « Hf0.5Zr0.5O2-Based Germanium Ferroelectric p-FETs for Nonvolatile Memory Applications ». In : *ACS Applied Electronic Materials* 4.6 (28 juin 2022). Publisher : American Chemical Society, p. 2815-2821. DOI : 10.1021/acsaelm.2c00324. URL : https://doi.org/10.1021/acsaelm.2c00324 (visité le 17/01/2023).

[Zho+20] Haidi ZHOU et al. « Endurance and targeted programming behavior of HfO2-FeFETs ». In : *2020 IEEE International Memory Workshop (IMW)*. 2020 IEEE International Memory Workshop (IMW). ISSN : 2573-7503. Mai 2020, p. 1-4. DOI : 10.1109/IMW48823.2020.9108131.

Glossary

28SLP GlobalFoundries 28 nm « Super Low Power technology » (technologie à très faible consommation) basée sur des transistors à grilles métalliques et diélectriques high-k[HKMG20] 64, 91, 98, 134, 137

3εFERRO Projet européen financé par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de la convention de subvention numéro 780302, et finançant le présent travail. 5–7, 18, 24, 25, 52, 54, 61, 102, 173–176

additionneur complet Circuit logique capable d'additionner trois entrées binaires, deux d'entre elles étant généralement les chiffres (bits) des mots à ajouter, et la troisième étant la retenue calculée après l'addition des chiffres de l'étage précédent. Voir section 4.6.3. 108–110, 112–117, 119, 123, 195, 196

ASIC Circuit intégré spécifique à une application, circuit intégré conçu dans un but précis, offrant généralement les meilleures performances et le meilleur rendement énergétique, au prix d'une certaine flexibilité et d'une conception coûteuse. 21, 24, 119

Bascule D Une bascule D est un circuit qui échantillonne son entrée sur les fronts d'horloge montants et la maintient comme valeur de sortie (jusqu'au prochain front d'horloge montant). Celle-ci est fréquemment utilisée pour créer des registres à décalage ou retarder un signal d'entrée d'une période d'horloge. 105, 106

Cadence Cadence est une marque déposée de Cadence Design Systems, Inc. 38, 115, 176, 196

calcul de périphérie Edge Computing. Une extension du paradigme de l'informatique en nuage (cloud computing), le calcul de périphérie désigne le calcul préalable qui a lieu à la périphérie, ou à la lisière, de l'informatique en nuage, avant d'envoyer les données aux serveurs. Ce pré-calcul permet généralement d'économiser de la bande passante et de la latence. 18, 19, *voir* cloud

CEA-LETI Institut de recherche en électronique et technologies de l'information, basé à Grenoble, France. Partenaire du projet 3εFERRO. 25, 53, 57, 60, 61

checkpointing Action de stocker l'état d'un circuit, d'un programme ou d'une mémoire en anticipant le besoin ultérieur de restauration, en réponse à une perte d'alimentation électrique ou à un autre besoin de restauration de l'état précédent. Cela peut par exemple être utilisé pour stocker un état par défaut ou connu comme correct, afin d'accélérer le retour en arrière ou les changements de contexte pendant le calcul 94, 102

cloud Le « cloud », ou informatique en nuage, est constitué de serveurs loués à un tiers. Généralement, ces serveurs sont des ordinateurs puissants situés dans un centre de données (datacenter), et l'utilisation d'un unique serveur peut être divisé entre plusieurs utilisateurs. Cela permet aux entreprises de bénéficier d'une plus grande flexibilité concernant l'évolution de leur capacité informatique et de sa localisation géographique. 19, 20, *voir* calcul de périphérie

CMOS MOS complémentaire, combinant un étage n-MOS et un étage p-MOS pour réduire la consommation de courant statique. 3, 24, 53–55, 59, 70–72, 91, 93–97, 117, 119, 121, 122, 137, 155, 156, *voir* n-MOS, p-MOS & MOS

Courbe en S Courbe *P*-*V* d'un ferroélectrique, affichant une seule courbe continue en forme de S au lieu d'une boucle d'hystérésis, comme sur la figure 2.12. 40, 43, *voir* *P*-*V*

Demokritos Centre national de recherche scientifique (CNRPS) « Demokritos », Grèce, Athènes (grec : Εθνικό Κέντρο Έρευνας Φυσικών Επιστημών (Ε.Κ.Ε.Φ.Ε.) « Δημόκριτος »). Partenaire du projet 3εFERRO. 25

ECL École centrale de Lyon, école d'ingénieurs de l'agglomération lyonnaise, France. Partenaire du projet 3εFERRO. 24, 102, 127

empreinte Imprint. Décrit la réponse asymétrique d'un matériau ferroélectrique à des champs électriques de même amplitude, mais de polarisation opposée[Mue+13b]. La courbe *P-V* est décalée vers la droite ou la gauche sur l'axe de tension en fonction de l'empreinte. 32, 54, 67

EPFL École Polytechnique Fédérale de Lausanne, université technique publique située à Lausanne, en Suisse. Partenaire du projet 3εFERRO. 5, 25, 40

etat de résistance Utilisé pour décrire l'état des dispositifs à conductivité électrique variable ou contrôlée, tels que les transistors. Souvent utilisé pour distinguer deux états, *état de résistance élevée* et *état de résistance faible*, mais peut également être utilisé avec plus de deux niveaux. 95, 99–101, 175, *voir* RS, OxRAM & FeFET

fan-out Nombre d'entrées de portes logiques (grilles de transistor) pouvant être pilotées par une seule sortie de porte logique 98

fenêtre mémoire Distance séparant plusieurs états de la mémoire lors de la lecture d'une mémoire. Plus la fenêtre mémoire est large, plus il est facile de différencier plusieurs états. Si la fenêtre mémoire devient trop petite, la mémoire peut ne pas fonctionner correctement car il devient impossible de distinguer l'état de la mémoire du bruit du capteur. 130

FeRAM RAM ferroélectrique, généralement 1T1C, peut également utiliser des FeFETs. 54, 55, *voir* RAM

flash Type de mémoire non volatile utilisant des portes de transistor flottantes, où l'information est stockée en piégeant des charges sur la grille du transistor, qui est flottante, car connectée en série avec un condensateur. 25, 49, 50, 53–55, 85, 91, 93, 121, 122

FPGA Field-Programmable Gate Array : circuit numérique reconfigurable couramment utilisé pour le prototypage ou pour remplacer les ASIC dans des productions de faible volume 24, 57, 90, 119, 121, *voir* ASIC

fréquence de rafraîchissement également appelé images par seconde, framerate, FPS : vitesse à laquelle les images sont capturées ou affichées, a un effet direct sur la bande passante requise pour la transmission et le traitement de la vidéo.

fuzzing Le test-fuzz (fuzz-testing) est une pratique de développement qui consiste à appliquer des données ou des valeurs aléatoires aux entrées des fonctions afin de découvrir bogues et cas limites imprévus. 156

FZJ Centre de recherche de Jülich, Jülich, Allemagne (en allemand : Forschungszentrum Jülich). Partenaire du projet 3εFERRO. 25

GlobalFoundries Société multinationale de fabrication et de conception de semi-conducteurs et partenaire de NaMLab. 52, 64, 91, 134, 137, 173

GPIO Entrées/sorties à usage général, connecteurs sur un microcontrôleur ou une plate-forme FPGA dont le comportement peut être spécifié par l'utilisateur 119

HDMI High Definition Media Interface, une interface média couramment utilisée par les appareils photo numériques et les téléviseurs 119

$HfZrO_2$ Oxyde de zirconium et de hafnium, dont la phase orthorhombique est ferroélectrique 18, 24, 25, 30, 34, 39, 54, 55, 59–61, 66, 70, 85, 86, 91, 93, 151, 158, 174

high-k Les diélectriques à haute permittivité (κ ou ε_r) se repolarisent davantage que les matériaux à faible κ en réponse à un champ électrique externe, ce qui leur permet de mieux guider et propager les champs électriques. 21, 23, 54, 55, 173

I_{DS} Courant entre le drain et la source d'un MOSFET, qui dépend de V_{GS}. 67, 78

I_{DS}—V_{GS} Représentation classique de la transconductance d'un transistor MOSFET, utilisée pour décrire les caractéristiques de performance. 9, 50, 82–84, 89, *voir* V_{GS} & I_{DS}

INL Institut des Nanotechnologies de Lyon, Lyon, France. 5, 24, 102, 127

LIFT Plate-forme interne d'optimisation multi-objectifs, décrite dans [Bri21]. 127, 151

MAD200 Procédé de démonstration pour $HfZrO_2$ ferroélectrique au CEA-LETI et STMicroelectronics basé sur le procédé HCMOS9A LP/HV OxRAM[HCMOS]. MAD signifie Memory

Advanced Demonstrator (démonstrateur avancé de mémoire), sur wafers 200 mm. 5, 8, 12, 60, 61, 64, 69, 75, 77, 79, 80

MATLAB Langage de programmation propriétaire et environnement de calcul numérique développé par MathWorks. 40, 114, 115, 127–129, 175, 182, 189, 190

MRAM RAM magnétorésistive, éventuellement programmée à l'aide du couple de transmission de spin (STT, Spin-Transfer Torque). Stocke de l'information dans un élément ferromagnétique, lu à l'aide d'une jonction magnétique à effet tunnel. 54, 55

Mémoire non volatile Mémoire conservant les informations stockées sans nécessiter d'alimentation externe. Voir sous-section 2.3.2. 3, 8, 17, 24, 48, 49, 54, 55, 90, 91, 102, 121, 148, 150

n-MOS MOSFET à canal n 65, 71, 78, 89, 90, 92, 100, 119, 173, *voir* MOS

NaMLab Nanoelectronics Materials Laboratory gGmbH, organisme de recherche axé à l'origine sur la recherche de matériaux pour les futurs dispositifs de mémoire, basé à Dresde, en Allemagne. Partenaire du projet 3εFERRO. 5, 24, 41, 52, 77, 82, 102, 112, 113, 130, 134, 174

NAND Porte logique Et (And) complémentée, symbole $\overline{\wedge}$; une porte logique universelle, et l'une des plus courantes. 93, 99, 122, 134, 137–139

NIMP Institut national de physique des matériaux, Bucarest, Roumanie (en roumain : Institutul Național de Cercetare-Dezvoltare pentru Fizica Materialelor). Partenaire du projet 3εFERRO. 25

NOR Porte logique OU (OR) complémentée, symbole $\overline{\vee}$; porte logique universelle commune. 93

normalement-éteint Le calcul normalement éteint (normally-off computing) est un paradigme d'opération dans lequel les éléments de calcul sont conservés hors tension la majorité du temps, afin d'économiser de l'énergie, et effectuent leurs calculs durant de courtes périodes d'activité. Les délais de mise sous tension et hors tension sont critiques pour cette application. 3, 54, 87, 102, 123, 140, 148

noyau également appelé masque ou matrice de convolution, matrice de petite taille (généralement 5×5 ou moins) utilisée dans une convolution bidimensionnelle avec une image à des fins de traitement. Voir la section 4.6.1. 103–105, 107, 108, 110, 112, 114, 115, 119, 121, 193

OCEAN Environnement de commande ouvert pour analyse, scriptage et contrôle de l'environnement de simulation utilisé dans Virtuoso® Analog Design Environment, établi sur le langage de programmation SKILL. 128, 129, *voir* SKILL

GNU Octave Langage de programmation de haut niveau pour le calcul scientifique et le calcul numérique, logiciel libre sous licence GPL et largement compatible avec MATLAB. 15, 40, 114, 182, 189, 190

OxRAM ReRAM utilisant des oxydes, où la résistance d'une couche d'oxyde est modulée par création de filaments conducteurs. 60, 174, *voir* ReRAM

p-MOS MOSFET à canal p 71, 78, 173, *voir* MOS

PCM Mémoire à changement de phase, stockant l'information sous forme de conductivité d'un matériau, pouvant être changée en changeant sa phase par un chauffage et une cristallisation contrôlés. 54, 55

pre-warming Action de déplacer une copie de données dans un niveau de mémoire plus rapide, afin d'accélérer un besoin anticipé d'accès à faible latence ou à haut débit aux données normalement stockées dans un niveau de mémoire plus lent. Il est directement lié au concept de cache, avec des caches « chauds » (récemment accédés, à jour) et des caches « froids ». 94

P-V courbe $P = f(V)$, utilisée pour caractériser les propriétés ferroélectriques comme décrit dans la sous-section 2.1.2. 12, 31–33, 36–38, 41, 42, 44, 48, 49, 173, 174, 182

ReRAM RAM résistive, stockant l'information sous forme d'etat de résistance. La résistance peut être modulée par l'application de courants électriques. 54, 55, 175, *voir* etat de résistance & OxRAM

SKILL Langage de programmation interactif dérivant du langage Lisp et utilisé dans les environnements de Cadence Design Systems. 15, 138, 179, 181

Spectre Simulateur de circuit SPICE propriétaire avec prise en charge du langage de modélisation Verilog-A, détenu et distribué par Cadence Design Systems. 38, 115, 156

STMicroelectronics Multinationale franco-italienne d'électronique et de semi-conducteurs. Partenaire du projet 3εFERRO. 24, 55, 60, 64, 174

SystemC Langage de simulation au niveau du système et de description d'architectures établi à partir du langage C++. 142, 152

Verilog language de description matérielle (HDL, Hardware Description Language) niveau RTL normalisé sous IEEE 1364. Est utilisé pour décrire les systèmes électroniques à des fins de simulation et de synthèse logique. 15, 112, 114, 115, 190–193

Verilog-A Langage de modélisation à temps continu, dérivé du langage Verilog. 15, 47, 115, 176, 185, 186, *voir* Verilog

V_{GS} Potentiel entre la grille et la source d'un MOSFET, également appelé tension de grille. 174

XOR Porte logique OU (OR) exclusive, symbole $\oplus$ ou $\veebar$; cas le plus complexe de la table de Karnaught, et souvent la porte logique nécessitant le plus de transistors. 99, 100

Acronyms

AFM microscopie à force atomique 31
ALD dépôt en couches atomiques 25
ALU Unité Arithmétique et Logique 142

BEoL bout de ligne 3, 12, 51–53, 55, 60, 61, 64–66, 86, 155, 156
BL Bit Line 61, 62, 77, 78, 80–83, 92, 93, 130, 131, 134

CAM mémoire adressable par contenu 53, 72, 75, 86
CAN Convertisseur Analogique-Numérique 64, 134
CGRA tableau reconfigurable à gros grain 57
CNN réseau de neurones convolutionnel 102, 157
CPU processeur 21, 22, 24, 141, 143, 152, 157

DK Design Kit 80
DRAM mémoire vive dynamique 23, 48, 49, 61–63, 77–80, 86, 94, 122, 130, 134, 140, 142–144, 155
DRC vérification des règles de dessin 80
DSE exploration de l'espace de conception 3, 54, 56, 86, 123, 124, 126, 127, 129–133, 135–139, 150–153, 156–158
DSP processeur de signal numérique 102
DTCO co-optimisation circuit et technologie 123, 147, 152, 153

E_C Champ Électrique Coercitif 27–29, 31, 33, 34, 49, 50, 52, 62, 88, 91, 158

FDSOI silicium entièrement appauvri sur isolant 18, 24
FeCap Condensateur Ferroélectrique 25, 34, 35, 48, 51, 53, 60–63, 65, 66, 69, 71–73, 75, 77–86, 94, 129–134, 155–158, 198
FeFET transistor à effet de champ ferroélectrique 3, 12, 14, 18, 25, 32, 34, 40, 47, 48, 50–56, 64–66, 68, 69, 75, 77–80, 85–102, 104, 108, 109, 112–114, 116–119, 121–123, 129, 134, 138–140, 144, 149, 155–158, 174, 181, 195, 196
FEoL début de ligne 12, 50, 55, 60, 64–66, 71, 85
FET transistor à effet de champ 21, 49–51, 88, 94, 98
FGMOS MOSFET à grille flottante 53, 93, 121, 156
FinFET Fin FET 18, 21
FIR réponse impulsionnelle finie 102, 103
FORC courbe d'inversion du premier ordre 44
FPS images par seconde 108, 119, *voir* fréquence de rafraîchissement
FTJ Jonction à effet Tunnel Ferroelectrique 36, 44, 50, 56, 77–79, 86

GAAFET FET à grille englobante 21
GPU processeur graphique 24

HDL language de description matérielle 176

I/O Entrée/Sortie 22, 23, 108, 110, 115, 116, 119
IMC calcul en mémoire 73, 140
IoT Internet des Objets 18, 19, 24
IPC communication inter-processus 12, 128, 129, 138, 151, 156, 183, 184
IRDS International Roadmap for Devices and Systems 22
ITRS International Technology Roadmap for Semiconductors 22

LiM logique en mémoire 13, 66, 87, 102, 121, 123, 140–143, 147, 148, 150, 152, 155–157
LSB bit de poids faible 119
LUT tableau de correspondances 13, 144, 148–150

MBD dépôt par jet moléculaire 25
MCAM mémoire multi-bit adressable par contenu 53
MFM Métal-Ferroélectrique-Métal 51, 52, 64, 65, 85
MFS Métal-Ferroélectrique-Semiconducteur 51, 52
ML Match Line 72, 73, 75, 76, 80
MLC cellule multi-niveaux 31, 32, 41, 63, 64, 79, 83, 90, 129, 134, 155, 158, 159
MOS Métal-Oxide-Semiconducteur 49–52, 70, 88, 90
MOSFET Transistor à Effet de Champ Métal-Oxide-Semiconducteur 20, 21, 34, 53, 67, 69, 81, 91, 94, 174–177
MSB bit de poids fort 119

NCFET transistor à effet de champ à capacité négative 25, 41, 49, 56
NWB sans réécriture 13, 143, 150

PL Plate Line 62, 70–73, 75–78, 80, 81, 83, 92, 93, 130, 131
PLD ablation laser pulsé 25
Pr polarisation résiduelle 12, 31, 32, 52, 63, 64, 66, 67, 83, 84, 131–133, 196
PsFeFET Pseudo-FeFET 12, 14, 51, 60, 64–71, 78, 80, 81, 85, 86, 96, 102, 121, 155–157
PUND Positif-Haut, Négatif-bas – Forme d'onde spécifique pour la caractérisation des condensateurs ferroélectriques, décrite dans la sous-section 2.1.4. 12, 31, 32, 35–38, 40, 42, 182
PVD dépôt physique par phase vapeur 25

RAM mémoire vive 53, 85, 174, 175
ROI région d'intérêt 40–42
RS état de résistance 98–100, *voir* etat de résistance
RTL niveau transfert de registre 112, 142, 176

SRAM mémoire vive statique 22, 49, 55, 100–102, 121
SSD stockage électronique 49, 54

TCAD simulation technologique physique 54
TCAM mémoire ternaire adressable par contenu 14, 53, 60, 72–76, 80, 85, 86, 91, 155
TPU processeur de tenseurs 24

V_C Tension Coercitive 12, 28, 31, 32, 35, 38, 48, 52, 62, 71–73, 77, 83, 88, 93–95, 98, 112, 116, 131, 132, 158
VO tension de sortie 98–100
V_{th} tension de seuil 9, 12, 34, 59, 72, 80, 88–91, 93, 96, 98, 99, 117–119, 138, 155

WB Réécriture 13, 55, 86, 143, 146, 149, 150, 158
WL Word Line 61, 62, 75, 77, 78, 81–83, 92, 93, 130, 134
WORM écriture unique, lecture multiple 158

Annexe A

Extraits de code

Extrait de code A.1 : Code SKILL pour l'extraction de métriques à partir des formes d'ondes de simulation de la bitcell 1T1C

```
analysis('tran ?stop "800u"  ?method "gear2"  )
desVar(   "low_noise_option" 1  )
desVar(   "L" tL    )
desVar(   "W" tW    )
desVar(   "atot_fe" Ac  )
desVar(   "vwl" vwl )
desVar(   "vprog" vprog )
desVar(   "vread" vread )

; slewr in s/V
desVar( "tfallprogBL" (slewr * vprog) )
desVar( "triseprogBL" (slewr * vprog) )
desVar( "tfallreadBL" (slewr * vread) )
desVar( "triseradBL" (slewr * vread) )
desVar( "tfallprogPL" (slewr * vprog) )
desVar( "triseprogPL" (slewr * vprog) )
desVar( "tfallWL" (slewr * vwl) )
desVar( "triseWL" (slewr * vwl) )


envOption(
    'cmd64bit  t
    'analysisOrder  list("tran")
)
save( 'i "/N0/D" )
temp( 27 )
run()
selectResult( 'tran )


(let
  (
   (Vc 1.2) ; For now, look at BL-PL, later look at n
   (P_switch 100mv) ; Internal polarization threshold to measure
    switching time.
   (Vfe ((v "/BL" ?result "tran") - (v "/PL" ?result "tran")))  ; BL-PL
   (Vp (v "/pint" ?result "tran")) ; Pint

   ; First: write 1<-0
   (tb_w10 420us) ; time begin write 0 over 1
   (te_w10 444us) ; time end ; Those two declarations might become
    redundant (unfortunately, no "let*" variant)
   (iw10 (clip (i "/N0/D" ?result "tran") 420us 444us)) ; current
    waveform for above region
   ew_10 ; energy for writing a 0 over a 1, assumes clip works
   ipk_w10 ; peak current for writing a 0 over a 1 (relies on clip)

```

```
45      ; read with 1 after write 0
46      (tb_r1_w0 520us) ; read with a 1 after writing 0
47      (te_r1_w0 544us)
48      (i_r1_w0 (clip (i "/N0/D" ?result "tran") 520us 544us))
49      er1_w0
50      ipk_r1_w0
51
52      ; write with 1 after reading (writing) with 1
53      (tb_w11 620us)
54      (te_w11 644us)
55      (iw11 (clip (i "/N0/D" ?result "tran") 620us 644us))
56      ew_11
57      ipk_w11
58
59      ; read with 1 after writing with 1
60      (tb_r1_w1 720us) ; read with a 1 after writing 1
61      (te_r1_w1 744us)
62      (i_r1_w1 (clip (i "/N0/D" ?result "tran") 720us 744us))
63      er1_w1
64      ipk_r1_w1
65
66      P_win ; max p. int difference, giving a window
67      tw_11 ; time to write 1 from 1 TODO
68      tw_10
69      tw_01
70      )
71
72    ;;
73      ew_10 = (integ ((abs iw10) * (abs Vfe)))
74      ipk_w10 = (ymax -iw10)
75    ;;
76      er1_w0 = (integ ((abs i_r1_w0) * (abs Vfe)))
77      ipk_r1_w0 = (ymax i_r1_w0)
78    ;;
79      ew_11 = (integ ((abs iw11) * (abs Vfe)))
80      ipk_w11 = (ymax iw11)
81    ;;
82      er1_w1 = (integ ((abs i_r1_w1) * (abs Vfe)))
83      ipk_r1_w1 = (ymax i_r1_w1)
84
85      tw_10 = (delay
86            ?wf1 Vfe  ?value1 -Vc  ?edge1 'falling ?td1 tb_w10
87            ?wf2 Vp ?value2 -P_switch ?edge2 'falling ?td2 0 ?stop te_w10)
88
89      tw_01 = (delay
90            ?wf1 Vfe  ?value1 Vc  ?edge1 'rising ?td1 tb_r1_w0
91            ?wf2 Vp ?value2 P_switch ?edge2 'rising ?td2 0 ?stop te_r1_w0)
92
93    ;;
94    chrg_win = (value
95      iinteg( (clip (i "/N0/D" ?result "tran") 415us 444us) )
96      444u) ; note: WL still active
97
98      restable["ipk_r1_w0"] = ipk_r1_w0
99      restable["ipk_r1_w1"] = ipk_r1_w1
100     restable["ew_10"] = ew_10
101     restable["er1_w0"] = er1_w0
102     restable["er1_w1"] = er1_w1
103     restable["tw_10"] = tw_10
104     restable["tw_01"] = tw_01
105     restable["chrg_win"] = chrg_win
106
107 )
```

```
108
109 (inl_ipcWriteTable restable)
```

Extrait de code A.2 : Code SKILL pour l'extraction de métriques à partir des formes d'ondes de la porte logique non volatile NAND à FeFET, comme décrit dans la sous-section 5.3.2

```
1  tpulse=20u
2  analysis('tran ?stop "600u"  )
3  desVar(   "tpulse" tpulse   )
4  desVar(   "trise" 0.2u  )
5  desVar(   "VDC" 1.5 )
6  desVar(   "VPROG" 5 )
7  desVar(   "Wfe" Wfe ) ;500n, inserted above automatically
8  desVar(   "Lfe" Lfe ) ;500n, ditto
9  envOption(
10   'cmd64bit  t 'userCmdLineOption  "+lite" 'analysisOrder  list("tran") )
11 save( 'i "/V3_vdd/PLUS" )
12 save( 'v "/vclk" "/pint" "vout" )
13 temp( 27 )
14 run()
15 selectResult( 'tran )
16
17 (let
18   (
19    (Vc 1.2) ; For now, look at BL-PL, later look at n
20    (Vt 0.7) ; threshold voltage, true
21    (Vtn 0.4) ; threshold voltage, false
22    (P_switch 100m) ; Internal polarization threshold to measure sw time
23    (Vp (v "/pint" ?result "tran")) ; Internal polarization terminal
24    (Vo (v "/vout" ?result "tran"))
25    (Vclk (v "/vclk" ?result "tran"))
26
27    ; Write 0
28    (tw_1 480u)
29
30    ; Write 1
31    (tw_0 340u)
32
33    ; 11
34    (ttt 400u) ott_valid vtt iprch
35
36    ; 00
37    (tff 560u) off_valid vff
38
39    ; 10
40    (ttf 440u) otf_valid vtf
41
42    ; 01
43    (tft 520u) oft_valid vft min_av
44 )
45
46    iprch = (clip (i "/V3_vdd/PLUS" ?result "tran")
47          (ttt + tpulse + tpulse / 2 )
48          (ttt + tpulse + tpulse + tpulse / 2))
49    eprch = (integ ((abs iprch) * (abs Vclk)))
50
51    vtt = (value Vo (ttt + tpulse / 2 ))
52    vff = (value Vo (tff + tpulse / 2 ))
53    vtf = (value Vo (ttf + tpulse / 2 ))
54    vft = (value Vo (tft + tpulse / 2 ))
55
56    ott_valid = ( vtt < Vt)
```

```
    off_valid = ( vff > Vt)
    otf_valid = ( vtf > Vt)
    oft_valid = ( vft > Vt)

    ; send valid=1 if every check is valid. Needs to be fp for the IPC
    restable["valid"] = (if (and ott_valid off_valid otf_valid oft_valid)
                1.0 0.0)
    restable["vtt"] = vtt
    restable["vff"] = vff
    restable["vtf"] = vtf
    restable["vft"] = vft
    restable["eprch"] = eprch
)

(inl_ipcWriteTable restable)
```

Extrait de code A.7 : Code GNU Octave (compatible MATLAB) d'ajustement des coefficients de Landau aux courbes expérimentales *P-V* obtenues par PUND

```
clear all

%%%%%% Dataset-specific data loading %%%%%%
% Note that the code could be simplified if the dataset was prepared by
% splitting the hysteresis code in a top and bottom half.
fields = {'V+ [V]', 'P1 [uC/cm2]'};

load dataset;
x = dataset.(fields{1}); % Voltage values
y = dataset.(fields{2}); % Polarization values in µC cm^-2
tfe = dataset.('Thickness [nm]')*1e-9;

%%%%%% Plot raw data %%%%%%

% Legends
prettyfields = {'V (V)', 'P ({\mu}C{\cdot}cm^{-2})'};
latexfields = {'V (\unit{\volt})',...
 'P (\unit{\micro\coulomb\per\square\centi\meter})'}

figure(1); clf; hold on;
title('Identification of Regions of interest for fitting experimental data')
plot(x,y, 'displayname','Experimental data');
legend show; legend location southeast

xlabel(prettyfields{1}); ylabel(prettyfields{2});

%%%%%%%%%%%%%%%%%% Look for points of interest %%%%%%%%%%%%%%%%%%
% Specifically, both ends of the hysteresis, and the furthest points from the
% line that passes trough both

values = [x'; y'];
[value1, pos1] = min(values');
[value2, pos2] = max(values');
extrema = [pos1; pos2]'; % These are the furthest points from the center

% We choose to take the min and max X, trace a line between them
% (the 'median line'), take the furthest points from it on each side,
% and start grouping closest points together

start_hyst = extrema(1,1);
x1 = [x(start_hyst); y(start_hyst)];
stop_hyst = extrema(1,2);
x2 = [x(stop_hyst); y(stop_hyst)];
```

EXTRAIT DE CODE A.3 : Script python d'exploration de l'espace de conception de la cellule 1T1C comme décrit dans la sous-section 5.3.1, utilisant l'IPC présentée dans la sous-section 5.2.2.

```python
#!/usr/bin/env python3

import sys
import json

sys.path.append("@cadenceipc/")

from ipc import oceanConnector

def sim_iteration(loopbody, parameters):
    # Construct params str
    pstr = ""
    for key in parameters:
        pstr = pstr + f"{key}={parameters[key]:0.10g}\n"
    print("Sending params", parameters)
    i.sendAndWaitCommand(pstr)
    print("Sending loop")
    i.sendAndWaitCommand(loop, 10*60)  # Timeout 10 minutes
    res = i.getTable()
    print("Got results ", res)
    return res


i = oceanConnector(["newoceanmad", "-nograph"], "/tmpfolder")
i.connect()

print("Sending preamble")
with open('preamble.ocn') as pre:
    i.sendAndWaitCommand(pre.read())

with open('loop.ocn') as lfile:
    loop = lfile.read()

simFamily = []
for tW in [130e-9, 140e-9, 160e-9, 180e-9]:
    for tL in [150e-9, 170e-9, 200e-9]:
        for sqrtAc in [100e-9, 150e-9, 180e-9, 300e-9]:

            simFamily.append({"parameters" : {
                "tW" : tW,
                "tL" : tL,
                "Ac" : sqrtAc**2,
                "slewr" : 5e-9,    # 5ns/V
                "vprog" : 3.6,
                "vread" : 3.6,
                "vwl" : 3.6
                }})


for sim in simFamily:
    sim["result"] = sim_iteration(loop, sim["parameters"])

with open("results_explo.json", "w") as resfile:
    json.dump(simFamily, resfile)
```

EXTRAIT DE CODE A.4 : Script d'exploration de la porte logique NAND non volatile à FeFET, telle que décrit dans la sous-section 5.3.2, utilisant l'IPC présentée dans la sous-section 5.2.2. loop.ocn correspond à l'extrait de code A.2.

```python
#!/usr/bin/env python3

import csv
from itertools import product
import sys
import os

sys.path.append(os.path.dirname(os.path.dirname(os.path.abspath(__file__))))

from ipc import oceanConnector

i = oceanConnector(["ocean", "-nograph"],
                   "/tmp/runtimeoceanfolder")

i.connect()

with open('preamble.ocn') as pre:
    i.sendAndWaitCommand(pre.read())

with open('loop.ocn') as lfile:
    loop = lfile.read()

tW = [1.2e-6, 1.3e-6, 1.4e-6, 1.5e-6]

tL = [300e-9, 380e-9, 500e-9, 560e-9, 650e-9, 780e-9, 880e-9, 1000e-9,
      1.2e-6, 1.4e-6]

combinations = product(tW, tL)

with open('results.csv', 'a') as csvfile:
    writer = csv.writer(csvfile)
#  writer.writerow(['Wfe','Lfe','valid','vtt','vff','vtf','vft','eprch'])
# The above line is the csv header. Simulate each parameter set:
    for params in combinations:
        print(f'trying {params}')
        ctext = f'Wfe={params[0]:0.10g}\nLfe={params[1]:0.10g}\n'
        i.sendcommand(ctext)
        i.sendAndWaitCommand(loop, 10*60)
        t = i.getTable()
        writer.writerow([
          params[0], params[1],
          t["valid"],
          t["vtt"], t["vff"], t["vtf"], t["vft"],
          t["eprch"]
          ])
```

EXTRAIT DE CODE A.5 : Sérialiseur de données en Verilog-A

```
`include "constants.vams"
`include "disciplines.vams"

module va_save_8b(clk, data);
input clk;
electrical clk;
input [7:0] data;
electrical [7:0] data;

integer fd_out_data;
integer bin_out, i;

parameter real vth = 0.3;
parameter string out_filename = "save8.bin";

analog begin

  @(initial_step) begin: openfile
          fd_out_data = $fopen(out_filename, "w"); // WB
          if (fd_out_data == 0) begin
            $display("Could not open file");
            $finish;
          end
  end

  @(final_step)
    $fclose(fd_out_data);

  @(cross ( V(clk)-vth, 1)) begin: serialize
          begin
                bin_out = 0;
                  generate i(7,0)
                        bin_out = bin_out + ((V(data[i]) > vth) << i);
                $display("read %h", bin_out);
        $fwrite(fd_out_data, "%c", bin_out);
          end
  end

end // analog

endmodule
```

Extrait de code A.6 : Sérialiseur de données en Verilog-A avec signal d'activation

```
1  `include "constants.vams"
2  `include "disciplines.vams"
3
4  module va_save_out_data(clk, out_data, out_data_valid);
5  input clk;
6  electrical clk;
7  input [7:0] out_data;
8  electrical [7:0] out_data;
9  input out_data_valid;
10 electrical out_data_valid;
11
12 integer fd_out_data;
13 integer bin_out, i;
14
15 parameter real vth = 0.3;
16 parameter string out_filename = "filter_output.bin";
17
18 analog begin
19
20   @(initial_step) begin: openfile
21           fd_out_data = $fopen(out_filename, "w"); // WB
22           if (fd_out_data == 0) begin
23             $display("Could not open file");
24             $finish;
25           end
26   end
27
28   @(final_step)
29     $fclose(fd_out_data);
30
31   @(cross ( V(clk)-vth, 1)) begin: serialize
32           if (V(out_data_valid) > vth)
33           begin
34                 bin_out = 0;
35                   generate i(7,0)
36                         bin_out = bin_out + ((V(out_data[i]) > vth) << i);
37                 $display("read %h", bin_out);
38         $fwrite(fd_out_data, "%c", bin_out);
39           end
40   end
41
42 end // analog
43
44 endmodule
```

```
44
45 % x1 and x2 are the start and stop points of the curve when looking at x values
46 plot([x1(1),x2(1)],[x1(2),x2(2)],'o', 'displayname', 'Extrema')
47
48 plot([x(start_hyst),x(stop_hyst)], [y(start_hyst), y(stop_hyst)],...
49      'displayname','median split'); % Plot the median, or bisector line
50
51 diagonal_slope = (y(stop_hyst)-y(start_hyst))/(x(stop_hyst)-x(start_hyst));
52 diagonal_y_at_origin = y(start_hyst) - diagonal_slope*x(start_hyst); % 0 usual.
53 y_on_diagonal = diagonal_slope*x + diagonal_y_at_origin;
54
55 % note: does not work if the top/bottom curves don't wrap,
56 % and edge points could belong both to the top and bottom part
57 a = min(start_hyst, stop_hyst); b = max(start_hyst, stop_hyst);
58 part_direct = a:b;
59 part_wrap = [b:size(values, 2), 1:a];
60
61 if sum(y(part_direct) > y_on_diagonal(part_direct)) > ...
62    sum(y(part_wrap) > y_on_diagonal(part_wrap))
63   part_top = part_direct; % More points above the diagonal
64   part_bottom = part_wrap;
65 else
66   part_top = part_wrap;
67   part_bottom = part_direct;
68 end
69
70 % Compute the orthogonal projection on the diagonal for every point of the loop
71 % This is to find the point furthest from the median split
72 % See extraitdecode A.8 for the code.
73
74 ortho = orthoproj([x,y],
75                   [x(stop_hyst)-x(start_hyst), y(stop_hyst)-y(start_hyst)],...
76                   [x(start_hyst), y(start_hyst)]);
77
78 sqdist2proj = ortho-values';
79 sqdist2proj = sqdist2proj(:,1).^2 + sqdist2proj(:,2).^2;
80
81 sqdist2proj_top = sqdist2proj; sqdist2proj_top(part_top) = 0;
82 [~, pos_furthest_top] = max(sqdist2proj_top);
83
84 sqdist2proj_bot = sqdist2proj; sqdist2proj_bot(part_bottom) = 0;
85 [~, pos_furthest_bot] = max(sqdist2proj_bot);
86
87 u = [x(pos_furthest_bot);x(pos_furthest_top)];
88 v = [y(pos_furthest_bot);y(pos_furthest_top)];
89
90 plot(u,v,'+', 'displayname', 'Furtherst')
91
92 %%%%%%%%%% Use the POIs to compute regions of interest %%%%%%%%%%
93 % Regions of interest are the regions that are not discarded to compute
94 % the s-shaped curve
95 %[u,v]=ginput(2); % We can ask the user to manually select POIs with this
96
97 % The following two lines are only useful if manually entering points, to select
98 % The points on the curve closest to the user-selected regions (norm 2)
99 [~,pos_user1] =min( ((x-u(1)).^2 + (y-v(1)).^2));
100 [~,pos_user2] =min( ((x-u(2)).^2 + (y-v(2)).^2));
101
102 a = min(stop_hyst, pos_user1); % just make sure indexes a<b for the calculations
103 b = max(stop_hyst, pos_user1);
104
105 direct = a:b;
106 modulo = [a:size(values, 2), 1:b];
```

```
if abs(min(direct) - max(direct)) < abs(min(modulo) - max(modulo))
  range1 = direct;
else
  range1 = modulo; % Take the shortest (x-wise) path
end

a = max(start_hyst, pos_user2);
b = min(stop_hyst, pos_user2);

direct = a:b;
modulo = [a:size(values, 2), 1:b];
if abs(min(direct) - max(direct)) < abs(min(modulo) - max(modulo))
  range2 = direct;
else
  range2 = modulo; % Take the shortest (x-wise) path
end
x_partial = [x(range1); x(range2)];
y_partial = [y(range1); y(range2)];


plot(x(range2),y(range2), '--', 'linewidth',3, 'displayname',...
     'ROI 2', 'color', [.64 .37 .51])
plot(x(range1),y(range1), '--', 'linewidth',3, 'displayname', 'ROI 1')
set(gca,'linewidth',2.5,'Fontsize',22,'ticklength',[0.025 0.025],...
    'PlotBoxAspectRatio',[1 0.85 1]);

%%%%%%%%%%%%%%%%%%%%%%%%%%%% Polynomial fit %%%%%%%%%%%%%%%%%%%%%%%%%%%%
n=6; % fit degree
% Odd numbers up to n, polynomial coefficients to keep
odds = find(mod(1:n,2) ~= 0);
p = polyfitc(y_partial,x_partial, odds); % Odd polynomial fit, see extraitdecode A.9

% Note how the polynomial fit is performed on flipped axes.
% boundary Y values are taken as a new X axis, and the plot is inverted to
% compensate.
X_sim_flip = linspace(y(start_hyst), y(stop_hyst), 1000);
Y_sim_flip = polyval(p, X_sim_flip);

polysign = {'','+'}; % to avoid displaying 2x +-3x^3 etc (see +-)
polysign = {polysign{1+(p>0)}};
polyexp = sprintf('%gP%s%gP^3%s%gP^5',p(5),polysign{3},p(3),polysign{1},p(1))

% Compute Landau coefficients
% -> V = 2·tfe·p + 4·tfe·p^3 +6·tfe·p
pconv = 1e-2; % 1 µC cm^-2 = 10^-2 coulomb/m^2
alpha = p(5)/(2*tfe*pconv^1)
beta  = p(3)/(4*tfe*pconv^3)
gamma = p(1)/(6*tfe*pconv^5)

fprintf('Polynomial coefficients:\nV=%s\nalpha=%d, beta=%d, gamma=%d\n',...
    polyexp, alpha, beta, gamma);
xlabel(prettyfields{1}); ylabel(prettyfields{2});


% Plot Final Fit
figure(2); clf; hold on;
plot(x,y, 'displayname', 'Experimental data')
plot(Y_sim_flip, X_sim_flip, 'displayname', 'Polynomial fit');
title('Fit to experimental data')
legend show
legend location southeast
set(gca,'linewidth',2.5,'Fontsize',22,'ticklength',[0.025 0.025],...
    'PlotBoxAspectRatio',[1 0.85 1]);
```

Extrait de code A.8 : Code GNU Octave (compatible MATLAB) pour la projection orthogonale, dépendance de l'extrait de code A.7.

```
% Orthogonally projects a list of points [x,y] (size n,2)
% On a line defined by a unitary vector [x,y] and a passing point [x,y].
% Returns the coordinates of every point.
function ortho = orthoproj(points, unitary, passing)
  unitary = unitary/norm(unitary); % Ensure

  OB = passing;
  BA = -OB + points; % Points are OA: BA = BO + OA = -OB + OA
  BH = (BA(:,1) .* unitary(1) + BA(:,2) .* unitary(2)) * unitary; % Dot product
  OH = OB + BH;

  ortho = OH;
end
```

Extrait de code A.9 : Code GNU Octave (compatible MATLAB) pour une régression polynomiale avec sélection de coefficients contraints à 0. Adapté de [J14], dépendance de l'extrait de code A.7.

```
function p=polyfitc(x,y,nvec)
%% from https://mathworks.com/matlabcentral/answers/
%       123072-curve-fitting-tool-with-custom-equation-odd-power-polynomial
%
% Simple 1D polynomial fitting with particular coefficients constrained to zero
%
%    p=polyfitc(x,y,nvec)
%
%in:
%
%  nvec: A vector of integer exponents present in the polynomial.
%  x: x data
%  y: y data
%
%out:
%
%   p: vector of polynomial coefficients in decreasing order from max(nvec) to 0

  A=bsxfun(@power,x(:),nvec(:).');
  [QQ,RR]=qr(A,0);
  coeffs = RR\( QQ'*y(:) );

  p=zeros(1,max(nvec)+1);
  p(nvec+1)=coeffs;
  p=p(end:-1:1);
```

Extrait de code A.10 : Code GNU Octave (compatible avec MATLAB) montrant l'utilisation de filtres convolutifs, utilisés pour générer la figure 4.3

```
clear variables; figure(1); clf;
original = imread('MAD200-microscope.jpg');
original = imresize (original, [300 356]); % Downsize for more obvious effect

%% Filter generation
identity = [1]; % One-by-one kernel with factor 1 as the identity
gauss = int8(fspecial('gaussian', [5 5], 1)*128);
sobel = int8(fspecial('sobel'));
sharp = int8(fspecial('unsharp', 0));

map = { % kernel, factor, name
  identity,   1, 'Original';
  gauss,    128, 'Blurred';
  sobel,      1, 'Sobel';
  sharp,      1, 'Sharpened'
};

%% Filtering and plotting
for i=1:size(map,1)
  subplot(2,2,i);
  kernel = map{i,1}; factor = map{i,2}; name = map{i,3};
  filtered = uint8(conv2(original, kernel, 'valid')/factor); % Filter and scale
  imshow(filtered); title(name);
  imwrite(filtered, horzcat(name, '.jpg'));
end
```

Extrait de code A.11 : Description Verilog du registre à décalage à déclenchement sur front descendant, avec entrée alternative comme décrit dans la section 4.6.2. Cette version échantillonne son entrée et modifie sa sortie sur les fronts d'horloge descendants, ce qui peut entraîner des problèmes avec les registres cascadés si les signaux d'horloge ne sont pas réceptionnés simultanément.

```
/* Muxed flip-flops, otherwise known as "SR" in the synoptic document.
 *
 * This circuit selects between two input busses each 8 bits wide, and acts as
 * a D-flip-flop that holds the selected value on positive clock edges.
 */

module muxed_ff(clk, A, B, select_B, O);

        parameter bus_size=8;
    input wire [bus_size-1:0] A,B;
    input wire select_B, clk;

    output reg [bus_size-1:0] O;


    always @(negedge clk)
        O <= select_B ? B: A;

endmodule
```

EXTRAIT DE CODE A.12 : Description Verilog du registre à décalage avec entrée alternative, comme décrit dans la section 4.6.2. Cette version échantillonne le signal d'entrée sur les fronts descendants et met à jour les signaux de sortie sur les fronts montants, permettant à chaque registre cascadé cascade d'échantillonner son entrée avant de modifier la sortie.

```
1  /* Muxed flip-flops, otherwise known as "SR" in the synoptic document.
2   *
3   * This variant is sensitive to both edges of the clock signal:
4   * The circuit still selects between two input busses each 8 bits wide,
5   * with the "select_B" signal, but samples the selected bus on positive clock
6   * edges, and changes its output on negative edges.
7   *
8   * This is done to mitigate possible inconsistent delays in the clock tree.
9   *
10  */
11
12 module muxed_ff_2edges(clk, A, B, select_B, O);
13
14     parameter bus_size=8;
15     input wire [bus_size-1:0] A,B;
16     input wire select_B, clk;
17
18     output reg [bus_size-1:0] O;
19     reg [bus_size-1:0] sampler;
20
21
22     always @(posedge clk)
23         sampler <= select_B ? B: A;
24
25     always @(negedge clk)
26         O <= sampler;
27
28 endmodule
```

Extrait de code A.13 : Description Verilog synthétisée d'une version à 1 bit du registre à décalage de l'extrait de code A.12, telle que décrite dans la section 4.6.2. Cette variante à 1 bit était moins complexe à réaliser et pouvait être placée 8 fois en parallèle pour fournir une version 8 bits. La bascule à déclenchement sur front montant a été réalisée avec une bascule à déclenchement sur front descendant et un signal d'horloge inversé, comme illustré sur le Circuit 4.14, permettant ainsi de réduire l'effort de conception à trois cellules différentes.

```
1  ////////////////////////////////////////////////////////////
2  // Created by: Synopsys DC Expert(TM) in wire load mode
3  // Version   : O-2018.06-SP3
4  // Date      : Fri Mar 20 09:10:19 2020
5  ////////////////////////////////////////////////////////////
6
7
8  module muxed_ff_2edges ( clk, A, B, select_B, O );
9    input [0:0] A;
10   input [0:0] B;
11   output [0:0] O;
12   input clk, select_B;
13   wire  \sampler[0] , n1, n4;
14
15   DFFPOSX1 \sampler_reg[0]  ( .D(n1), .CLK(clk), .Q(\sampler[0] ) );
16   DFFNEGX1 \O_reg[0]  ( .D(\sampler[0] ), .CLK(clk), .Q(O[0]) );
17   INVX1 U6 ( .A(n4), .Y(n1) );
18   MUX2X1 U7 ( .B(A[0]), .A(B[0]), .S(select_B), .Y(n4) );
19 endmodule
```

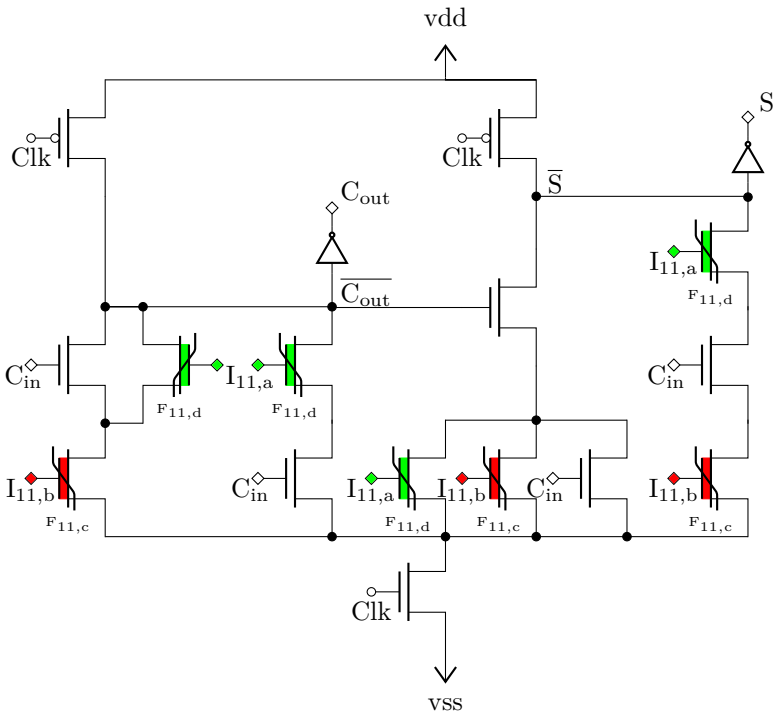
Extrait de code A.14 :] Extrait du banc d'essai Verilog qui alimente initialement le circuit multiplicateur avec les données du noyau.

```
task feed_kernel_data(input reg only_sign_ext);
begin : feed
    integer i, clk_per_mult;
    reg [7:0] data;
    clk_per_mult = only_sign_ext? sr_cycles_per_multiplier : 1;
    for(i=0; i < kernel_width*kernel_width; i = i+1)
    begin
        // We need to feed the kernel head first to have the right
        // values pass by first.
        data = kernel_data[24-i];
        in_data = only_sign_ext? {8{data[7]}} : data;
        // *2 because the loop only has one edge
        for ( i_sr = 0; i_sr< clk_per_mult*2; i_sr=i_sr+1)
        begin
            #(serial_SR_clk_period/2) clk1 = !clk1; clk2 = !clk2;
        end
    end
end
endtask


/* Load filter coefficients trough the "debug" circuit.
*
* The goal is to have each multiplier coefficient line up with its kernel_in
* input.  Note that we feed the kernel coefficients in the "correct" order,
* since the SR chain is a fifo, and the first in are the furthest in the
* image. However, since the image columns are going to be fed from multiplier
* [0,5,10,15,20] to [4,9,14,19,24] in that order, the pixels in the leftmost
* column of the kernel need to be burned into the last column.
*/
task load_kernel;
begin
    debug_enabled = 1; // Filter programming happens in debug mode
    /*
    * First feed the 25 coefficient trough the debug circuit, but there is
    * a catch: each needs to be fed sr_cycles_per_multiplier (normally 8),
    * as that is the space between two multipliers.  Those coefficients
    * are going to be stored in the second half of the multipliers, so
    * they just consist of the sign extension bits, repeated.
    */
    begin : feed_kernel
        // due to the input topology, loop in that order:
        feed_kernel_data(1); // Only feed the sign extension bits
        $display("Finished feeding the sign extension values");
        feed_kernel_data(0);
    end

    /* Now that kernel has been feed, we need to align it with the
    * multipliers by waiting for the initial latency
    */
    begin : fast_forward_kernel
        in_data = 8'hxx;
        for ( i_sr = 0; i_sr< sr_to_multiplier_latency; i_sr=i_sr+1)
        begin
            #(serial_SR_clk_period/2) clk2 = !clk2; clk1 = !clk1;
            #(serial_SR_clk_period/2) clk2 = !clk2; clk1 = !clk1;
        end
        $display("finished catching up to latency, about to write");
    end
    vdd3_on = 1;
    write_enable = 1;
```

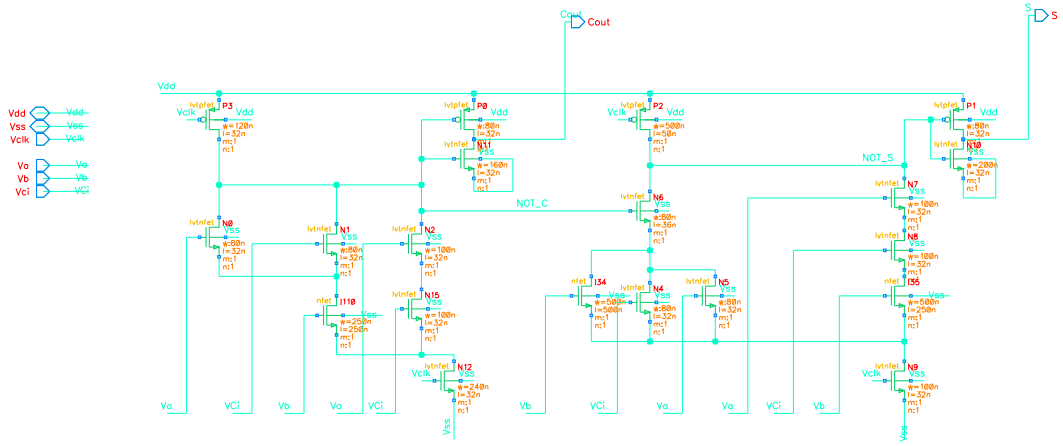
```
62      #(kernel_write_time)
63      write_enable = 0;
64      debug_enabled = 0;
65      vdd3_on = 0;
66      #2
67      $display("Wrote kernel");
68  end
69  endtask
```

Annexe B

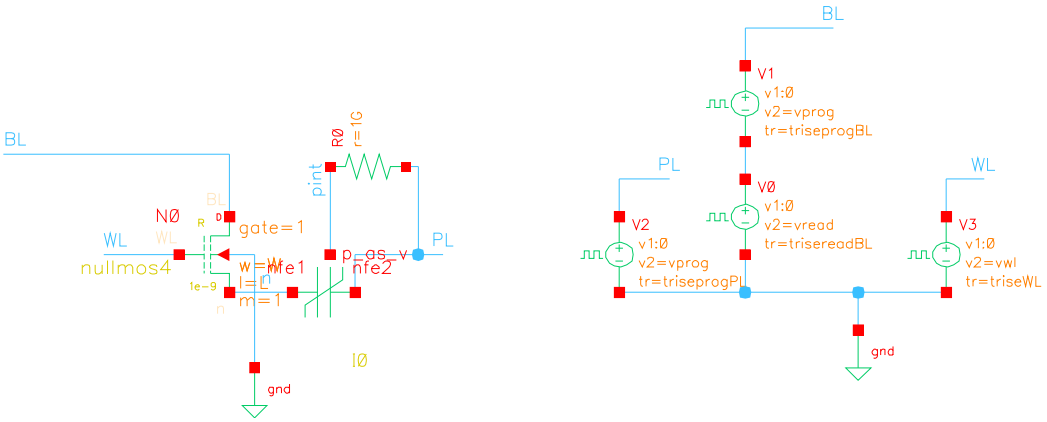
Circuits additionnels



CIRCUIT B.1 : Première version de l'additionneur complet avec des multiplicateurs 1-bit à FeFET, telle qu'initialement proposée. Cette version effectue deux multiplications, contrairement aux suivantes.



Circuit B.2 : Deuxième itération du circuit de l'additionneur complet, avec multiplicateur bit à bit à FeFET. Cet additionneur complet présente un défaut dû à l'utilisation de la logique dynamique, comme décrit dans la section 4.6.4



Circuit B.3 : Circuit Cadence pour le banc d'essai 1T1C. Le nœud *pint* émet une tension correspondant à la valeur interne de Pr, en $\mathrm{C\,cm^{-2}}$, inspiré de [WA17]. Cette approche permet de mesurer plus précisément l'inversion de la polarisation.

Annexe C

Tableaux additionnels

300 71×10^{-15}	400 126×10^{-15}	550 238×10^{-15}	550 238×10^{-15}	←Diamètre Condensateur (nm) ←↓Superficie équivalente (m^2)
				0
X				71×10^{-15}
	X			126×10^{-15}
X	X			196×10^{-15}
			X	238×10^{-15}
X			X	308×10^{-15}
	X		X	363×10^{-15}
X	X		X	434×10^{-15}
		X	X	475×10^{-15}
X		X	X	546×10^{-15}
	X	X	X	601×10^{-15}
X	X	X	X	672×10^{-15}

TAB. C.1 : Combinaisons possibles du circuit 2T4C, telles que fabriquées, classées par surface équivalente totale. Les croix indiquent la combinaison sélectionnée, la colonne de droite indique la surface totale de FeCap, obtenue en additionnant les surfaces des condensateurs sélectionnés. Les doublons conduisant à la même surface totale sont omis. Voir également tableau C.2 pour la variante 2T5C.

300 71×10^{-15}	400 126×10^{-15}	400 126×10^{-15}	400 126×10^{-15}	550 238×10^{-15}	←Diamètre Condensateur (nm) ←↓Superficie équivalente (m^2)
					0
X					71×10^{-15}
			X		126×10^{-15}
X			X		196×10^{-15}
				X	238×10^{-15}
		X	X		251×10^{-15}
X				X	308×10^{-15}
X		X	X		322×10^{-15}
			X	X	363×10^{-15}
	X	X	X		377×10^{-15}
X			X	X	434×10^{-15}
X	X	X	X		448×10^{-15}
		X	X	X	489×10^{-15}
X		X	X	X	560×10^{-15}
	X	X	X	X	615×10^{-15}
X	X	X	X	X	685×10^{-15}

TAB. C.2 : Combinaisons 2T5C possibles, telles que fabriquées, avec la même méthodologie que tableau C.1 pour la variante 2T4C.