Development of microfabrication processes for BEOL integration of ReRAM circuits on CMOS chips

Raphaël Dawant

UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

Développement de procédés de microfabrication pour l'intégration BEOL de circuits ReRAM sur des puces CMOS

Thèse de doctorat
Spécialité : génie électrique

Raphaël DAWANT

Sherbrooke (Québec) Canada

Mai 2024

MEMBRES DU JURY

Dominique DROUIN

Directeur

Serge ECOFFEY

Codirecteur

Malek ZEGAOUI

Évaluateur

Marc GUILMAIN

Évaluateur

Fabien ALIBART

Rapporteur

RÉSUMÉ

L'intérêt croissant pour les applications d'intelligence artificielle, notamment celles basées sur les réseaux de neurones artificiels (ANN), est bridé par les limitations matérielles des processeurs. Ces limitations sont dues principalement aux volumineux transferts de données requis par les opérations de multiplication vecteur-matrice (VMM), qui sont centrales dans les ANN. Pour surmonter ces défis, l'approche du calcul en mémoire a été introduite afin de réduire considérablement le mouvement des données. Cette technique a l'avantage de pouvoir utiliser des mémoires résistives (ReRAM), où la résistance des dispositifs peut changer pour stocker des informations, facilitant ainsi les opérations de VMM directement sur le lieu de stockage des données.

Les ReRAM offrent l'avantage de pouvoir être fabriquées directement dans le Back-End-of-Line (BEOL) des circuits intégrés CMOS, où des progrès significatifs ont été réalisés, notamment avec l'architecture 1T1R (un transistor par mémoire). Toutefois, les architectures passives, où les transistors ne se situent qu'en sortie de réseau, restent encore peu matures bien qu'elles offrent des avantages substantiels en termes de densité d'intégration et d'efficacité énergétique.

L'intégration BEOL de structures passives de ReRAM présente des défis spécifiques liés à la nécessité d'interconnexions denses à proximité immédiate des cellules de mémoire. Contrairement aux architectures 1T1R, qui s'appuient largement sur la circuiterie classique étendue sur plusieurs niveaux du BEOL, l'approche passive permet une augmentation de la densité d'information mais exige une niveau d'interconnexion supplémentaire.

Cette thèse se concentre sur l'utilisation stratégique des techniques de fabrication du BEOL pour intégrer efficacement des circuits ReRAM passifs, exploitant des procédés et des matériaux déjà établis, tout en ajustant ces méthodes pour répondre aux exigences spécifiques des circuits passifs. Cette démarche vise à maintenir la compatibilité avec les standards industriels des procédés CMOS, facilitant ainsi le transfert technologique vers des applications pratiques de l'IA.

Plusieurs procédés de fabrication utilisés dans le BEOL, basés sur des techniques telles que le polissage mécano-chimique (CMP), tels que le procédé Damascene pour les interconnexions de cuivre ou le procédé soustractif pour les interconnexions d'aluminium, seront développés et caractérisés dans la première partie du document. L'intégration de ces procédés de fabrication sera réalisée sur des puces CMOS fabriquées pour un fournisseur industriel. Les résultats de l'intégration permettront de montrer le potentiel d'un transfert industriel pour des circuits ReRAM passifs fabriqués dans le BEOL, et une discussion sur les schémas de fabrication pour une perspective future permettra d'identifier les procédés avec le plus de potentiel.

Mots-clés : Calcul en mémoire, mémoire résistive, ReRAM, circuit RS, intégration CMOS, technologie BEOL

ABSTRACT

Development of microfabrication processes for BEOL integration of ReRAM circuits on CMOS chips

The growing interest in artificial intelligence applications, especially those based on artificial neural networks (ANN), is hindered by the hardware limitations of processors. These limitations are primarily due to the large data transfers required by vector-matrix multiplication (VMM) operations, which are central in ANNs. To overcome these challenges, the in-memory computing approach has been introduced to significantly reduce data movement. This technique has the advantage of using resistive memories (ReRAM), where the resistance of devices can change to store information, thus facilitating VMM operations directly at the data storage site.

ReRAMs offer the advantage of being manufactured directly in the Back-End-of-Line (BEOL) of CMOS integrated circuits, where significant progress has been made, particularly with the 1T1R architecture (one transistor per memory). However, passive architectures, where transistors are only located at the network output, remain immature despite offering substantial advantages in terms of integration density and energy efficiency.

The BEOL integration of passive ReRAM structures presents specific challenges related to the need for dense interconnections in close proximity to the memory cells. Unlike 1T1R architectures, which rely largely on traditional circuitry extended over several levels of the BEOL, the passive approach allows an increase in information density but requires an additional level of interconnection.

This thesis focuses on the strategic use of BEOL fabrication techniques to efficiently integrate passive ReRAM circuits, leveraging established processes and materials while adjusting these methods to meet the specific requirements of passive circuits. This approach aims to maintain compatibility with CMOS process industry standards, thus facilitating technological transfer to practical AI applications.

Several fabrication processes used in the BEOL, based on techniques such as chemical-mechanical polishing (CMP), including the Damascene process for copper interconnections or the subtractive process for aluminum interconnections, will be developed and characterized in the first part of the document. The integration of these manufacturing processes will be carried out on CMOS chips manufactured for an industrial supplier. The integration results will demonstrate the potential for industrial transfer for passive ReRAM circuits manufactured in the BEOL, and a discussion on manufacturing schemes for a future perspective will identify the processes with the most potential.

Keywords: In-memory computing, resistive memory, ReRAM, RS circuit, CMOS integration, BEOL technology

À Deb'

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ACRONYMES

- **1T1R :** 1 Transistor 1 Resistor - 1 Transistor 1 Résistance
- **3D-FC :** 3D Fully Connected - 3D Entièrement Connecté
- **3IT :** Institut Interdisciplinaire d'Innovation Technologique
- **AFM :** Atomic Force Microscopy - Microscopie à Force Atomique
- **AI :** Artificial Intelligence - Intelligence Artificielle
- **ANN :** Artificial Neural Network - Réseau de Neurones Artificiels
- **ASIC :** *Application-specific integrated circuit*
- **BE :** Bottom Electrode - Électrode inférieur
- **BEOL :** Back-End Of Line
- **CBRAM :** Conductive Bridging Random Access Memory - Mémoire à accès aléatoire à pontage conducteur
- **CC :** Contrast Curve - Courbe de Contraste
- **CMOS :** *Complementary metal oxide semi-conductor*
- **CMOS :** Complementary Metal-Oxide Semiconductor - Semi-conducteur à Oxyde Métallique Complémentaire
- **CMP :** Chemical Mechanical Polishing - Polissage Mécanochimique
- **DAC :** Digital-to-Analog Converter - Convertisseur Numérique-Analogique
- **DD :** Dual-Damascene
- **DRAM :** Dynamic Random Access Memory - Mémoire Dynamique à Accès Aléatoire
- **DRIE :** Deep Reactive Ion Etching - Gravure Ionique Réactive Profonde
- **DTC :** Distance to Center - Distance au Centre
- **EBL :** Electron Beam Lithography - Lithographie par Faisceau d'Électrons
- **EBPG :** Electron Beam Pattern Generator - Générateur de Motifs par Faisceau d'Électrons
- **ECC :** Etching Calibration Curve - Courbe de Calibrage de Gravure
- **EHT :** Extra High Tension - Tension Extra Haute
- **EPD :** End Point Detection
- **FEOL :** Front-End Of Line
- **FIB :** Focused Ion Beam - Faisceau d'Ions Focalisé
- **FPGA :** Field Programmable Gate Array - Réseau de Portes Programmables sur Site
- **Fi-BEOL :** Fully Integrated Back-End of Line - Entièrement Intégré Arrière Fin de Ligne
- **HRS :** High Resistance State - État de Haute Résistance
- **I-V :** Courant-tension (caractéristiques I-V)
- **IC :** Integrated Circuit - Circuit Intégré
- **ICP :** Inductively Coupled Plasma - Plasma Couplé Inductivement
- **IDL :** Inter-Dielectric Layer - Couche Inter-Diélectrique
- **INPAQT :** Integrated Nanoelectronics and Packaging for AI and Quantum Technologies

- **IoT :** Internet of Things - Internet des Objets
- **LN2 :** Laboratoire Nanotechnologies Nanosystèmes
- **LRS :** Low Resistance State - État de Basse Résistance
- **M8 :** 8th Metal Level BEOL - 8e Niveau de Métal BEOL
- **MAC :** Multiply and Accumulate - Multiplier et Accumuler
- **MEMS :** Microelectromechanical Systems - Systèmes Microélectromécaniques
- **MIM :** Metal-Insulator-Metal - Métal-Isolant-Métal
- **MOSFET :** Metal-Oxide-Semiconductor Field-Effect Transistor - Transistor à Effet de Champ à Oxyde Métallique-Semiconducteur
- **MRR :** Material Removal Rate - Taux d'Enlèvement de Matériau
- **NN :** Neural Network - Réseau Neuronal
- **NVM :** Non-Volatile Memory - Mémoire Non Volatile
- **OxRAM :** Oxide-based Random Access Memory - Mémoire à Accès Aléatoire à Base d'Oxyde
- **PCB :** *Printed circuit board* - Circuit imprimé
- **PCM :** Phase Change Memory - Mémoire à Changement de Phase
- **PECVD :** Plasma Enhanced Chemical Vapor Deposition - Dépôt Chimique en Phase Vapeur Assisté par Plasma
- **PVD :** Physical Vapor Deposition - Dépôt Physique en Phase Vapeur
- **RAM :** Random Access Memory
- **RS :** Resistive Switching - Commutation Résistive
- **ReRAM :** Resistive Random Access Memory
- **SA :** Stand-Alone - Autonome
- **SD :** single damascene
- **SEM-MEB :** Scanning Electron Microscope - Microscope Électronique à Balayage
- **SL :** Switching Layer - Couche de Commutation
- **SRAM :** Static Random Access Memory - Mémoire à Accès Aléatoire Statique
- **STT-RAM :** Spin-Transfer Torque Random Access Memory - Mémoire à Accès Aléatoire à Couple de Transfert de Spin
- **TE :** Top Electrode - Électrode Supérieure
- **TPU :** Tensor Processing Unit - Unité de Traitement de Tenseur
- **TSMC :** Taiwan Semiconductor Manufacturing Company - Compagnie de Fabrication de Semi-conducteurs de Taiwan
- **UTM :** Ultra Thick Metal - Métal Ultra Épais
- **UdeS :** Université de Sherbrooke
- **VMM :** Vector Matrix Multiplication - Multiplication Matricielle de Vecteurs
- **iMC :** In-Memory Computing - Calcul en Mémoire

CHAPITRE 1

Introduction

Ce chapitre établit le cadre général de cette thèse, abordant la problématique de recherche centrée sur les systèmes matériels conçus pour l'exécution des algorithmes d'intelligence artificielle (IA) ainsi que les approches proposées pour relever ces défis. Cette introduction mène à la formulation de la question de recherche de ce manuscrit. Les objectifs du projet de recherche sont exposés, suivis d'une descriptions détaillée de la structure du document en fin de chapitre.

1.1 Mise en contexte

Les applications d'IA, en particulier celles basées sur les réseaux de neurones artificiels (ANN) tels que l'apprentissage profond, ont révolutionné de nombreux domaines. Par exemple, elles sont devenues incontournables en reconnaissance d'images [3] et de sons [4], en imagerie médicale [5, 6], dans l'analyse de l'ADN [7], les outils de traduction linguistique [8], ainsi que dans les secteurs de la finance [9] et de l'art [10]. Le potentiel économique de l'IA est substantiel. Selon une étude de McKinsey, l'IA pourrait augmenter le PIB mondial de 1,2 % annuellement jusqu'en 2030 [11]. Accenture prédit quant à elle que l'IA pourrait accroître la productivité mondiale de 40 % d'ici 2035 [12]. Des géants technologiques tels que les GAFA, Microsoft, Intel et IBM ont massivement investi dans ces technologies, témoignant de leur importance stratégique.

Le développement des algorithmes d'apprentissage profond est entravé par les limites matérielles (hardware) des ordinateurs actuels. Naveen Rao d'Intel souligne que la croissance rapide des réseaux neuronaux dépasse la capacité du matériel actuel, nécessitant une révision globale de la structure des systèmes matériels [13]. Dans cette section, l'écart entre les besoins logiciels en apprentissage profond et les capacités matérielles sera présenté. Les solutions existantes, incluant les innovations en architecture IA, seront discutées.

Xiaowei Xu et al. ont montré qu'il existait un écart entre les ANNs à l'état de l'art en termes de taille et de précision requises par rapport aux capacités des plateformes matérielles disponibles [14]. La figure 1.1(a) montre que la demande en nombre d'opérations exigées par les meilleurs ANNs augmente exponentiellement, tandis que la figure 1.1(b) indique une saturation de la densité de performance des plateformes matérielles.

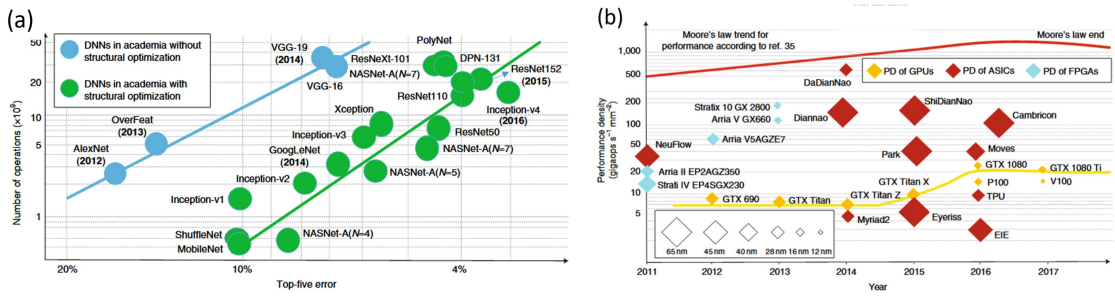


FIGURE 1.1 (a) Croissance exponentielle des opérations requises par les ANNs de pointe et (b) saturation de la densité de performance matérielle. [14]

1.2 Problématique

1.2.1 Goulot d'étranglement de Von Neumann

Les limites matérielles s'expliquent par la structure même des ANNs et de leur exécution sur des systèmes matériels. Ces réseaux de neurones artificiels sont principalement développés pour les algorithmes d'apprentissage machine. Les ANNs sont composés d'un réseau de neurones formels, où chaque neurone fonctionne comme une fonction mathématique avec plusieurs entrées et une seule sortie [15]. Comme illustré sur la figure 1.2, chaque entrée x_i est multipliée par un poids w_{ij}, spécifique à l'entrée i et au neurone j. L'ensemble des entrées pondérées $x_i w_{ij}$ est ensuite sommé dans une opération appelée MAC (*multiply and accumulate*) ou VMM (*vector matrix multiplication*). Après cette somme pondérée, le résultat est traité par une fonction d'activation non linéaire, souvent produisant une valeur non nulle au-delà d'un certain seuil θ_j, appelé aussi fonction neuronale.

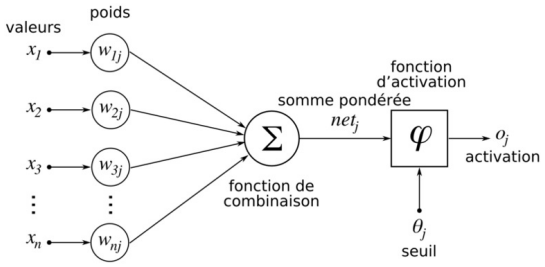


FIGURE 1.2 Représentation schématique du fonctionnement d'un neurone formel [1]

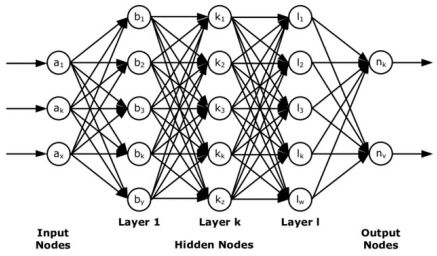


FIGURE 1.3 Réseau de neurones artificiels pour l'apprentissage machine profond [2]

Le perceptron est le plus simple des algorithmes d'apprentissage machine. Il est composé d'un seul neurone et est capable de classer un problème à n entrées en deux catégories. Les algorithmes d'apprentissage machine plus complexes, comme l'apprentissage profond, emploient des ANNs plus élaborés avec plusieurs niveaux de couches, chaque neurone

d'une couche étant connecté à tous les neurones des couches précédentes et suivantes comme illustré sur la figure 1.3.

Dans le cas de réseaux DNN (*deep neural networks*), l'opération de MAC pour chaque couche successive correspond à un produit vectoriel entre un vecteur n (avec n, le nombre d'entrées ou le nombre de neurones de la couche précédente) et une matrice $n \times m$ (où m est le nombre de neurones).

Sur des architectures de CPU (*central processing unit*) conventionnelles, l'opération de MAC se produit de façon séquentielle. Cette opération devient rapidement coûteuse en calculs, notamment pour des algorithmes d'apprentissage profond nécessitant un nombre important d'entrées. De plus, chaque opération requiert un aller-retour entre le CPU et les unités de mémoire où sont stockés les poids synaptiques. Ce flux important de données, limité par le bus de communication entre les unités de mémoire et le CPU, crée un goulot d'étranglement (*bottleneck*), comme le montre la figure 1.4.

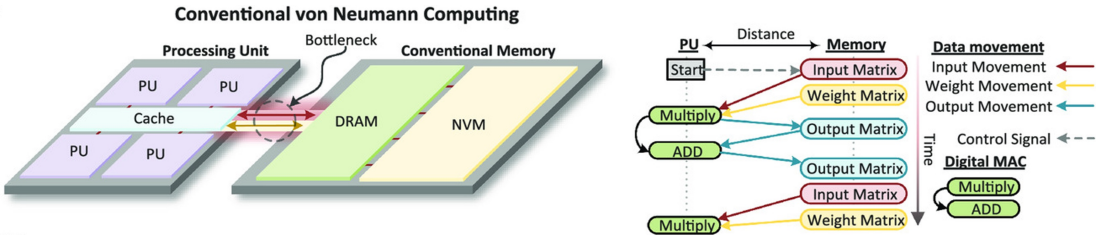


FIGURE 1.4 L'architecture de Von Neumann présente un goulot d'étranglement causé par le transfert de données entre les unités de calcul et la mémoire, ce qui limite la vitesse de traitement [16].

1.2.2 Accélérateur d'IA

Pour répondre aux exigences actuelles, il est nécessaire d'adapter les architectures de processeurs afin de surmonter les limitations causées par les transferts de données précédemment décrits. Avant d'explorer les structures matérielles, il est pertinent de mentionner que de nombreuses optimisations ont été réalisées au niveau algorithmique. Notamment, la modification de la structure des DNN pour limiter les mouvements de données est notable. On peut le retrouver comme dans l'utilisation des CNN (*Convolutional Neural Networks*) [17], qui exploitent la réutilisation des données pour en réduire le transfert, ou le dropout, qui consiste à retirer aléatoirement des neurones pendant l'entraînement afin de diminuer le nombre d'opérations [18]. Cependant, ces solutions sont insuffisantes et ne parviennent pas à résoudre entièrement le problème d'architecture physique, comme le montre la persistance de l'écart de performance de la figure 1.1. Pour cette raison, des architectures

matérielles spécifiques, appelées accélérateurs d'IA [19], ont été développées pour pallier à cette problématique.

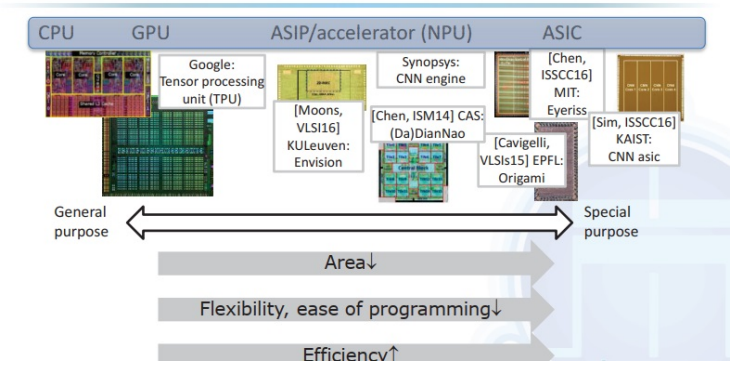


FIGURE 1.5 Illustration de différents types d'accélérateurs numériques pour l'IA. La spécialisation des systèmes offre un gain de performance mais au coût d'une flexibilité moindre [16].

Architecture numérique

Les premiers accélérateurs d'IA exploités sont les unités de traitement graphique (GPU), initialement conçues pour le rendu graphique. Elles ont été largement adoptées pour l'IA en raison de leur haut débit et de leurs capacités de traitement parallèle, ce qui les rend idéales pour l'entraînement de réseaux neuronaux complexes et la gestion de grands ensembles de données [20]. Les réseaux logiques programmables (FPGA) sont des systèmes reconfigurables qui peuvent être programmés pour effectuer des calculs spécifiques, offrant une flexibilité et une meilleure efficacité énergétique, ce qui les rend adaptés au déploiement de modèles d'IA sur des dispositifs nécessitant une faible consommation d'énergie [21]. Des circuits intégrés spécifiques tels que l'unité de traitement tensoriel (TPU) de Google [22], sont des puces conçues sur mesure et optimisées pour des applications particulières. Ces circuits offrent une efficacité et des performances inégalées pour les tâches d'apprentissage profond [23]. Les unités de traitement neuronal (NPU) [24] sont des accélérateurs matériels spécialisés conçus explicitement pour les calculs de réseaux neuronaux, optimisés à la fois pour de hautes performances et une faible consommation d'énergie. Comme le montre la figure 1.5, le choix entre ces différentes architectures représente un compromis entre flexibilité de programmation et efficacité énergétique. Pour minimiser les mouvements de données, les accélérateurs d'IA numériques exploitent :

- La réutilisation des données (*Data reuse*) [25],
- Un stockage local plus proche pour réduire le déplacement des données (*near-memory computing*) [26],
- L'exploitation de la parcimonie dans les réseaux de neurones (*Sparsity*) [27],
- L'utilisation de précisions réduites et/ou variables [28].

Architecture hybride analogique/numérique et calcul en mémoire

Les optimisations obtenues par les architectures numériques pour réduire les mouvements de données sont fondamentalement limitées par l'aspect séquentielle de l'opération de MAC dans un processeur numérique comme indiqué sur la figure 1.6(a). Dans ce cadre,

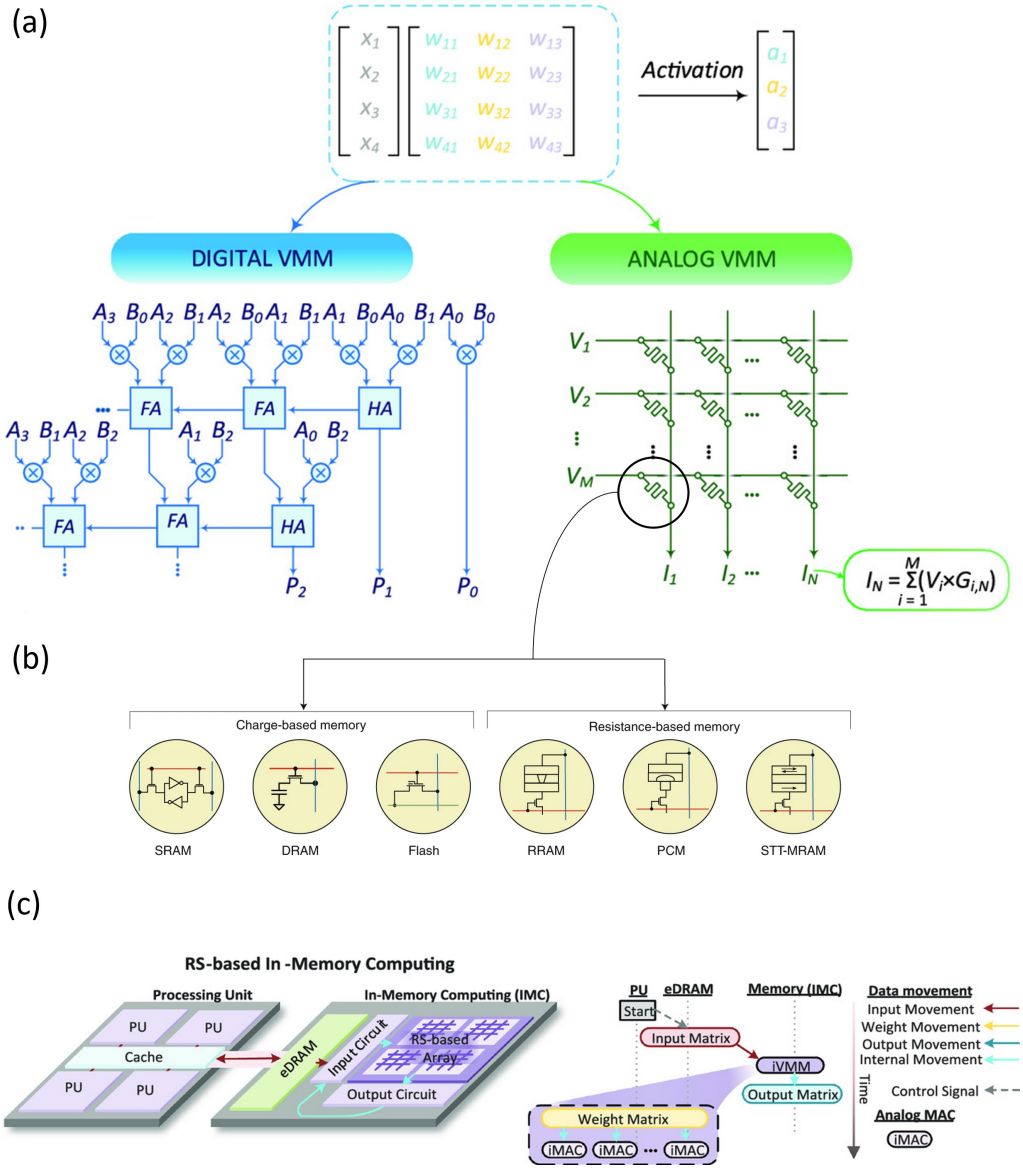


FIGURE 1.6 Architecture de calcul en mémoire pour l'IA. (a) Implémentation d'un calcul VMM effectué par les algorithmes basés sur des ANNs. À gauche, implémentation sur les systèmes classiques effectuée de façon séquentielle. À droite, implémentation du calcul en mémoire sous la forme d'un réseau crossbar. (b) Le calcul en mémoire peut s'implémenter avec différents types de mémoire basés sur la charge ou sur un changement de résistance. (c) L'implémentation du calcul en mémoire dans les systèmes matériels permet de limiter largement le mouvement des données. [16, 29]

l'approche hybride analogique/numérique ouvre de nouvelles voies en permettant l'intégration de l'opération de MAC au sein même des unités de mémoire, grâce à l'exploitation des propriétés physiques des dispositifs de mémoire [29]. Cette méthode est connue sous le nom de calcul en mémoire (iMC, *In-memory computing*), une technique où les dispositifs de mémoire sont intégrés dans une structure transversale dénommée crossbar. Comme il sera détaillé dans la section 2.2, le calcul VMM est effectué de façon direct et parallèle dans cette configuration. En plus de réaliser l'opération de façon parallèle et analogique, le mouvement de données est drastiquement réduit comme l'illustre la Figure 1.6(c) sur le schéma de droite.

Deux classes peuvent être distinguées, comme indiqué sur la Figure 1.6 : (i) les mémoires utilisant la charge pour stocker l'information, basées sur des technologies semiconductrices CMOS ; (ii) les mémoires basées sur des mécanismes de commutation résistive ou RS (*resistance switch*), qui utilisent la résistance pour stocker l'information, également appelées memristors. Ces dernières seront présentées en détail dans la section 2.3. Il a été démontré que les systèmes analogiques utilisant des mémoires à base de charge peuvent théoriquement atteindre des performances supérieures de 2 à 3 ordres de grandeur comparées à celles des systèmes numériques, et de 3 à 4 ordres pour les systèmes analogiques utilisant des mémoires à base de résistance, que ce soit en termes de vitesse ou de consommation énergétique [30]. En outre, les systèmes basés sur les mémoires résistives, qui ont l'avantage d'être non-volatiles, montrent de meilleures performances que les systèmes analogiques MOSFET, ce qui est crucial pour des applications nécessitant une faible consommation énergétique. À noter que la technologie flash, bien que non-volatile, possède un temps de commutation 10^6 fois plus élevé que les mémoires RS, comme démontré dans le tableau 2.1 de la section 2.3.4, rendant cette technologie moins efficace pour des opérations de VMM.

	Digital				Analog			
	CPU 2.66 GHz	GPU 1 GHz	FPGA 200 MHz	ASIC 400 MHz	NOR 180 nm	NOR 55 nm	Memristors 200 nm	3D memristors 10 nm
Time (s)	~8×10^{-3}	~3×10^{-4}	~1.5×10^{-4}	~5×10^{-5}	~2×10^{-6}	~7×10^{-7}	~5×10^{-8}	~1×10^{-8}
Power (W)	~30 to 40	~40	~10	~3	~1	~1	~1	~0.1
Energy (J)	~3×10^{-1}	~1×10^{-2}	~1×10^{-3}	~1×10^{-4}	~2×10^{-6}	~7×10^{-7}	~5×10^{-8}	~1×10^{-9}

TABLEAU 1.1 Efficacité énergétique et comparaison des performances pour un algorithme d'apprentissage profond dédié à la classification d'images de 64 × 64 pixels [30]

1.2.3 Systèmes VMM basés sur des mémoires RS

Pour les raisons évoquées précédemment, les systèmes RS destinés aux opérations de VMM (*RS-Based VMM Engines*) ont suscité un intérêt considérable pour leur potentiel dans le développement d'accélérateurs d'ANN [31, 32, 33]. Ces systèmes exploitent les mémoires RS pour réaliser l'opération VMM en combinant un circuit CMOS qui gère les signaux d'entrée/sortie (I/O) et les autres types d'opérations nécessaires dans un DNN. Les propriétés des mémoires RS incluent ; une haute densité d'intégration, une capacité de mise à l'échelle grâce à leurs dimensions nanométriques, un coût réduit, une faible consommation énergétique, la non-volatilité, des comportement analogique, ainsi que la compatibilité avec une intégration monolithique dans le BEOL (*back-end of line*) des circuits CMOS. Toutes ces propriétés en font un choix privilégié pour le calcul VMM dans une approche iMC [16, 29]. Les différents types de mémoire RS et leurs avantages pour ces systèmes seront explorés plus en détail dans la section 2.3. La fabrication des mémoires RS dans le BEOL offre plusieurs avantages. Premièrement, en termes de densité, les mémoires peuvent bénéficier de l'empilement 3D des niveaux d'interconnexion inhérent au BEOL. Cela permet de réduire significativement l'empreinte spatiale des mémoires RS et de diminuer le délai RC (résistance-capacité) en réduisant la longueur et donc la résistance des interconnexions.

Circuits RS passifs versus 1T1R

Les circuits RS peuvent être divisés en deux configurations principales : (i) l'approche 1T1R, avec un transistor par mémoire adressée ou (ii) l'approche passive avec un transistor qui adresse plusieurs mémoires à la sortie du réseau crossbar. L'approche 1T1R permet un contrôle individuel sur chaque mémoire et évite l'apparition des chemins de courant parasitaires qui apparaissent dans les crossbars passifs. Cela facilite la gestion de la programmation et de la précision des états de lecture mais augmente la complexité et l'empreinte spatiale de l'ensemble, chaque mémoire nécessitant deux connexions dans le BEOL pour se connecter à un transistor. En revanche, l'approche passive, où plusieurs mémoires partagent un même transistor d'accès, permet de réduire considérablement le nombre de transistors nécessaires. Cette configuration 1T1R, est particulièrement avantageuse dans un contexte d'intégration monolithique où l'espace disponible pour les composants CMOS est un facteur limitant. Dans cette configuration, le nombre d'interconnexions et de transistors par unité de mémoire diminue potentiellement au carré à mesure que la taille du réseau augmente, optimisant ainsi l'espace et permettant une densification plus poussée des mémoires RS au sein du circuit.

Intégration Monolithique BEOL de Circuits RS Passifs

Des avancées significatives ont été réalisées dans le développement de réseaux de mémoire 1T1R, avec de multiples produits commercialisés, démontrant l'emploi de ces mémoires dans des applications telles que le calcul en mémoire [34, 35, 36, 37, 38, 39]. Néanmoins, l'évolution des circuits RS analogiques passifs a progressé à un rythme plus modéré [40]. Cette situation s'explique principalement par les défis liés aux non-idéalités inhérentes aux mémoires RS, ainsi qu'aux chemins de courant parasitaires. Même les technologies les plus avancées commercialisées qui exploitent des configurations passives comme la technologie XPoint de Intel opèrent les mémoires en mode numérique [41]. L'intégration BEOL de crossbars passifs dans la technologie des mémoires présente des défis spécifiques, principalement dus à la nécessité de réaliser des interconnexions supplémentaires entre les cellules mémoire. Contrairement aux architectures 1T1R, où la complexité de l'interconnexion repose largement sur la circuiterie classique BEOL étendue sur plusieurs niveaux, les crossbars passifs exigent des schémas d'intégration plus complexes pour connecter directement les cellules mémoire entre elles.

1.2.4 Questions de recherche

Les différentes problématiques introduites mettent en avant la nécessité de développer des procédés spécifiques pour l'intégration des circuits RS passifs en BEOL, compatibles avec les technologies CMOS. Ces procédés doivent être adaptés aux critères d'intégration BEOL, comme la planarité nécessaire pour l'empilement 3D des niveaux d'interconnexion. Pour ces raisons, ce projet de thèse vise à répondre à la question suivante :

> ***Comment adapter les techniques de microfabrication utilisées dans l'industrie CMOS pour intégrer des circuits RS passifs dans le BEOL ?***

1.3 Objectifs du projet de thèse

Cette thèse de doctorat s'inscrit dans le cadre d'un projet de recherche multidisciplinaire au sein du groupe INPAQT dirigé par le Professeur Dominique Drouin à l'Institut interdisciplinaire d'innovation technologique (3IT) nommé HiData pour Heterogeneous Integration of High-Density analog Crossbar for Advanced Data Processing financé par le CRSNG (Conseil de recherches en sciences naturelles et en génie du Canada).

Le projet HiData a pour but l'intégration de mémoire résistive à base de TiOx dans le BEOL de puces CMOS spécialement conçues pour réaliser des opérations VMM directement en mémoire. L'objectif est de créer un système entièrement intégré sur une même puce CMOS, incluant les circuits périphériques tels que les circuits de conversion DAC/ADC et les circuits de routage.

Dans le cadre d ece projet global, l'objectif de cette thèse se focalise sur l'intégration des crossbars passifs sur les puces CMOS TSCM du projet en utilisant les ReRAM à base de TiOx développés dans le groupe de recherche. Des techniques de fabrication utilisées dans le BEOL du CMOS seront exploitées dans le but de faciliter un transfert technologique. Les objectifs peuvent être définis de la manière suivante :

1. A partir du travail réalisé précédemment, le premier objectif consiste à prendre la technologie ReRAM à base de TiOx et de développer un procédé de microfabrication de circuit de mémoire ReRAM dans l'optique d'une intégration CMOS de cette technologie. L'objectif mettra l'accent sur :
   - L'utilisation de procédés et matériaux compatibles avec les lignes de production BEOL.
   - La validation des performances électriques des mémoires, en particulier en ce qui concerne le maintien des résistances d'interconnexion adéquates.
   - La validation des tensions maximales requises et les plages de résistance de commutation des dispositifs pour garantir leur compatibilité avec les tensions d'alimentation et les contraintes électriques des puces CMOS de TSMC.
2. À partir des développements étudiés précédemment, le deuxième objectif consiste à réaliser une intégration dans le dernier niveau BEOL de puces CMOS. Pour cela, il sera nécessaire de :
   - Développer une technique d'alignement adaptée aux puces CMOS pour débloquer une intégration monolithique.
   - Adapter le procédé de fabrication et intégrer des crossbars passifs dans le BEOL de puces CMOS 130nm.

Ces objectifs spécifiques sont conçus pour relever les défis techniques et optimiser les performances des systèmes intégrés. En réussissant à intégrer efficacement les technologies ReRAM dans les BEOL des puces CMOS, ce travail de recherche contribuera à améliorer significativement la capacité de traitement des données de manière plus compacte et énergétiquement efficace.

1.4 Plan du manuscrit

Ce manuscrit est organisé en six chapitres. Le **chapitre 1** a établi le contexte et la problématique du sujet pour introduire la question de recherche. Il a été mis en évidence que les systèmes VMM utilisant de la mémoire RS sont une approche prometteuse pour résoudre les contraintes de mouvement de données imposées par les accélérateurs d'IA numériques. La nécessité d'une intégration dans le BEOL pour répondre au besoin d'interconnectivité dense a été montrée. L'approche passive, bien qu'étant celle qui est la plus prometteuse pour une intégration à haute densité, est également la moins développée et nécessite encore des développements importants pour un transfert industriel. Pour répondre à cette problématique, ce projet de thèse vise à développer un procédé de microfabrication dans le BEOL de puces CMOS en utilisant des techniques de fabrication utilisées en industrie.

Le **chapitre 2** débutera par une revue de la littérature, mettant l'accent sur les principales technologies BEOL à l'état de l'art. L'objectif est de s'inspirer de ces techniques pour le développement du procédé de fabrication et favoriser un transfert industriel. Les circuits RS qui réalisent l'opération VMM seront introduits, et les défis de leur implémentation seront abordés. Ensuite, les différentes classes de mémoire utilisables dans les circuits RS seront présentées et comparées. Un focus particulier sera mis sur les OxRAMs, un type de mémoire ReRAM spécifiquement utilisé dans ce projet. Par la suite, une revue systématique des démonstrations de systèmes VMM intégrés des dernières années sera réalisée. Basée sur la compatibilité CMOS et leur intégration dans le BEOL, une analyse de la maturité technologique de ces systèmes sera effectuée. Cette analyse révélera que les technologies basées sur des ReRAM numériques et des architectures 1T1R sont les plus avancées. En revanche, les approches passives, surtout celles opérant en régime analogique, nécessitent encore d'importants développements pour un transfert technologique.

Dans le **chapitre 3**, trois schémas d'intégration seront introduits en s'inspirant des schémas d'intégration utilisés pour les interconnexions métalliques dans le BEOL. Ceux-ci sont basés sur des procédés de damascène et des procédés soustractifs. Les procédés de fabrication et le développement de l'étape de CMP seront présentés. Les résultats seront analysés d'un point de vue morphologique et leur compatibilité avec une intégration CMOS sera discutée. Les performances électriques des dispositifs mémoires seront comparées pour les différentes approches. Ce chapitre est écrit sous le format d'un article scientifique publié dans le journal *Micro and Nano Engineering* intitulé : *Damascene Versus Subtractive Line CMP Process for Resistive Memory Crossbars BEOL Integration* [42].

Dans le but de développer des schémas d'intégration soustractive avec des morphologies plus abouties, le **chapitre 4** introduira une méthode de fabrication pour graver des structures en 3D. Cette méthode permet de graver un empilement de plusieurs matériaux en utilisant une technique de lithographie en niveau gris. Grâce à cette méthode, un schéma d'intégration de crossbar, largement inspiré des approches TopVia développées pour les interconnexions BEOL au nœud avancé, a pu être réalisé. Ce chapitre est écrit sous le format d'un article scientifique publié dans le journal *Journal of Vacuum Science & Technology B*, intitulé : *Multiple material stack grayscale patterning using electron-beam lithography and a single plasma etching step* [43].

Un des objectifs de ce projet consiste à intégrer les procédés de fabrication développés dans les chapitres précédents dans le BEOL d'une puce CMOS TSCM spécialement conçue pour réaliser un système VMM. Afin de réaliser les étapes de lithographies nécessaires, il est nécessaire de s'aligner avec l'équipement de lithographie utilisé au 3IT. Les règles de conception des puces CMOS engendrent des défis spécifiques. Pour résoudre ces défis, le **chapitre 5** étudiera une méthode d'alignement non-conventionnelle écrite sous le format d'un article scientifique publié dans le journal *Journal of Vacuum Science & Technology B* en collaboration avec la compagnie qui fabrique l'équipement de lithographie, intitulé : *Hybrid cross correlation and line-scan alignment strategy for CMOS chips electron-beam lithography processing* [44].

Pour finir, le **chapitre 6** commencera par une synthèse des procédés de fabrication développés précédemment. En se basant sur plusieurs critères qui seront détaillés, deux de ces schémas seront utilisés pour réaliser l'intégration des crossbar sur le BEOL des puces TSCM. Le procédé pour connecter le crossbar ReRAM aux interconnexions du BEOL sera présenté ainsi que les résultats morphologiques de cette intégration et les potentielles optimisations du procédé de fabrication. Le chapitre finira par une discussion sur des changements dans la séquence des étapes de fabrication pour optimiser le nombre d'états. Ces changements permettront également non pas une intégration après le dernier niveau de métal, mais une intégration dans les différents niveaux de BEOL pour profiter de l'empilement 3D inhérent au BEOL, dans le but d'une intégration massivement dense.

1.5 Contributions scientifiques

Cette section liste les contributions principales de ce projet qui ont été des études publiées dans des revues scientifiques avec la liste des auteurs qui y ont contribué. Les apports des éléments développés dans ce projet aux autres projets menés par le groupe de recherche du Pr. Dominique Drouin seront également brièvement exposés.

Contribution principale

Ce projet a abouti à trois articles scientifiques qui sont respectivement le sujet des chapitres 3, 4 et 5 qui ont été décrits plus haut :

• **R. Dawant**, M. Gaudreau, M.-A. Roy, P.-A. Mouny, M. Valdenaire, P. Gliech, J. Arias Zapata, M. Zegaoui, F. Alibart, D. Drouin et S. Ecoffey, *Damascene Versus Subtractive Line CMP Process for Resistive Memory Crossbars BEOL Integration*
Micro and Nano Engineering, pp. 100251 (2024) [42]
• **R. Dawant**, S. Ecoffey, D. Drouin, *Multiple material stack grayscale patterning using electron-beam lithography and a single plasma etching step*
Journal of Vacuum Science & Technology B, vol. 40, no. 6, pp. 062603 (2022) [43]
• **R. Dawant**, R. Seils, S. Ecoffey, R. Schmid, D. Drouin, *Hybrid cross correlation and line-scan alignment strategy for CMOS chips electron-beam lithography processing*
Journal of Vacuum Science & Technology B, vol. 40, no. 1, pp. 012601 (2021) [44]

Contribution comme co-auteur

Les échantillons développés et fabriqués tout au long de ce projet de recherche ont également pu contribuer à l'aboutissement de plusieurs autres études réalisées dans le groupe INPAQT.

La caractérisation par A. El Mesoudy de mémoires ReRAMs fabriquées avec la première version du procédé de damascène a été publiée dans le journal *Microelectronic Engineering* :

• A. El Mesoudy, G. Lamri, **R. Dawant**, J. Arias-Zapata, P. Gliech, Y. Beilliard, S. Ecoffey, A. Ruediger, F. Alibart, D. Drouin, *Fully CMOS-compatible passive TiO2-based memristor crossbars for in-memory computing*
Microelectronic Engineering, vol. 255, pp. 111706 (2022)

Des réseaux crossbar fabriqués avec le procédé de fabrication damascène, décrits au chapitre 3, ont permis de réaliser une étude sur la programmation de crossbar ReRAM par P. Drolet. Cette étude propose des techniques de programmation des crossbars passifs qui ex-

ploitent les non-idéalités des ReRAM et des circuits crossbar pour améliorer la robustesse du calcul :

• P. Drolet, **R. Dawant**, V. Yon, P.-A. Mouny, M. Valdenaire, J. Arias Zapata, P. Gliech, S. Wood, S. Ecoffey, F. Alibart, Y. Beilliard et D. Drouin, *Hardware-aware Training Techniques for Improving Robustness of Ex-Situ Neural Network Transfer onto Passive TiO_2 ReRAM Crossbars*
arXiv :2305.18495 (2023)

Les échantillons fabriqués avec l'approche soustractive double développée dans le chapitre 3 ont permis la réalisation de trois études par P.-A. Mouny, qui étudie et utilise les dispositifs ReRAM pour des applications cryogéniques :

• P.-A. Mouny, Y. Beilliard, S. Graveline, M.-A. Roux, A. El Mesoudy, **R. Dawant**, P. Gliech, S. Ecoffey, F. Alibart, M. Pioro-Ladrière et D. Drouin, *Memristor-based cryogenic programmable DC sources for scalable in-situ quantum-dot control* IEEE Transactions on Electron Devices, vol. 70, no. 4, pp. 1989-1995 (2023)
• P.-A. Mouny, **R. Dawant**, B. Galaup, S. Ecoffey, M. Pioro-Ladrière, Y. Beilliard et D. Drouin, *Analog programming of CMOS-compatible $AlO2/TiO_{2-x}$ memristor at 4.2 K after metal-insulator transition suppression by cryogenic reforming*
Applied Physics Letters 123, 163505 (2023)
• P.-A. Mouny, **R. Dawant**, P. Dufour, M. Valdenaire, S. Ecoffey, M. Pioro-Ladrière, Y. Beilliard et D. Drouin, *Towards scalable cryogenic quantum dot biasing using memristor-based DC sources*
(Soumis à Elsevier Cryogenics)

Finalement, le procédé damascène développé pour fabriquer une électrode inférieure de TiN a également été utilisé pour la fabrication de mémoire ferroélectrique, un autre type de mémoire RS étudié dans le groupe de recherche. Ces mémoires développées par D. Coffineau ont mené à la soumission de deux publications :

• D. Coffineau, N. Gariépy, B. Manchon, **R. Dawant**, A. Jaouad, É. Grondin, S. Ecoffey, F. Alibart, Y. Beilliard, A. Ruediger, D. Drouin, *CMOS-compatible $Hf_{0.5}Zr_{0.5}O_2$-based ferroelectric memory crosspoints fabricated with damascene process*
(Soumis à IOP Nanotechnology)

• D. Coffineau, N. Gariépy, **R. Dawant**, S. Ecoffey, F. Alibart, Y. Beilliard, A. Ruediger, D. Drouin, *Ultra-thin ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ with Damascene bottom electrodes for in-memory computing*
(Soumis à IEEE Transactions on Electron Devices)

Conférences

• **R. Dawant**, M. Gaudreau, M.-A. Roy, P.-A. Mouny, M. Valdenaire, P. Gliech, J. Arias Zapata, M. Zegaoui, F. Alibart, D. Drouin et S. Ecoffey, *Damascene Versus Etch-Back Chemical Mechanical Planarization for Resistive Memory Crossbars Back-End-Of-Line Integration*
Canadian Semiconductor Technology Conference, *Montréal, Canada* (2023)

• **R. Dawant**, M.-A. Roy, P.-A. Mouny, M. Valdenaire, P. Gliech, J. Arias Zapata, F. Alibart, D. Drouin et S. Ecoffey, *Chemical Mechanical Planarization process for pillar shaped Resistive Memory Crossbars Integration*
Micro and Nano Engineering Conference, *Berlin, Allemagne* (2023)

• **R. Dawant**, S. Ecoffey, D. Drouin, *3D shaping of multi-layers stack using a single plasma etching step and greyscaleelectron-beam lithography*
65th International Conference on Electron, Ion and Photon Beam Technology and Nanofabrication-EIPBN, *Nouvelle-Orléans, USA* (2022)

• **R. Dawant**, R. Seils, S. Ecoffey, R. Schmid, D. Drouin, *Comparison of alignment markers and method for electron-beam lithography on CMOS dies*
64th International Conference on Electron, Ion and Photon Beam Technology and Nanofabrication-EIPBN, *en ligne* (2021)

CHAPITRE 2

État de l'art

Le chapitre précédent a mis en évidence le besoin de développer des systèmes VMM basés sur des mémoires RS pour répondre aux exigences des algorithmes exploitant des ANN. Alors que les approches utilisant une structure 1T1R ont déjà été démontrées à un niveau avancé, les approches passives nécessitent encore des développements pour une éventuelle industrialisation.

Dans le contexte de l'intégration de crossbars passives dans le BEOL de circuits CMOS, il est nécessaire de développer un schéma de fabrication compatible avec la structure du BEOL, spécifiquement pour les interconnexions reliant les mémoires. Par la suite, ces interconnexions seront appelées les électrodes inférieure et supérieure du crossbar (BE et TE). La première partie de ce chapitre présentera une revue de la littérature sur la technologie BEOL, y compris les procédés standards soustractifs et damascènes, ainsi que les schémas d'intégration proposés pour pallier les défis rencontrés dans les nœuds avancés. Cette revue permettra de cerner les besoins et les solutions envisageables pour répondre à la question de recherche.

Ensuite, le fonctionnement des circuits RS sera exposé, avec une analyse des différences techniques entre une architecture 1T1R et passive. Les défis spécifiques aux circuits passifs seront identifiés pour mieux comprendre les enjeux de leur intégration. La présentation et la comparaison des différentes classes de mémoires RS, ainsi que leur performance en termes de vitesse de fonctionnement, de maturité technologique, d'empreinte spatiale et de fonctionnement analogique, permettront de sélectionner le type de mémoire RS le plus adapté.

Finalement, une revue complète des démonstrations de systèmes VMM intégrés réalisées récemment sera présentée. Cette analyse comparative prendra en compte l'architecture passive ou 1T1R, le fonctionnement analogique ou numérique des mémoires, l'intégration sur le BEOL, et la compatibilité des procédés de fabrication pour une intégration BEOL industrielle. Elle mettra en lumière le niveau de maturité des approches passives par rapport aux approches 1T1R et la distinction entre mémoires analogiques et numériques.

2.1 Technologie BEOL

Le terme BEOL désigne les étapes postérieure à la création des dispositifs fondamentaux, tels que les transistors, dans le processus de fabrication des IC (*integrated circuit*). Cette phase se concentre sur l'interconnexion des composants préalablement établis et implique un agencement alterné de matériaux diélectriques et conducteurs, formant ainsi les pistes d'interconnexion et les vias, essentiels pour la continuité électrique à travers les multiples niveaux du circuit. Dans le contexte actuel de la microélectronique, jusqu'à dix niveaux d'interconnexion peuvent être intégrés dans un seul IC.

Les technologies du BEOL jouent un rôle crucial dans la performance des dispositifs semiconducteurs, particulièrement avec la poursuite de la miniaturisation. Les interconnexions, liaisons physiques entre les transistors, sont devenues un facteur critique pour la vitesse et l'efficacité énergétique dans les circuits intégrés. Cette section explore l'évolution des matériaux et techniques utilisés dans le BEOL, passant de l'aluminium (Al) au cuivre (Cu) et les techniques de fabrication associées, ainsi que les options envisagées pour résoudre les problématiques engendrées par la réduction des dimensions dans les nœuds avancés $<$ 5nm.

2.1.1 Procédés standards

Procédé soustractif (Al) vs Procédé Damascene (Cu)

Historiquement, les interconnexions métalliques était faites d'aluminium, qui possède une conductivité électrique relativement basse, une mise en œuvre simple en termes de déposition et de gravure, un coût faible, ainsi qu'une capacité à former une couche de passivation en créant un oxyde natif protecteur [45]. La figure 2.1(b) illustre le procédé de fabrication d'un niveau d'interconnexion à base d'Al, appelé approche classique ou soustractive. L'aluminium est déposé, puis les lignes métalliques sont gravées. Un oxyde de SiO_2 est ensuite déposé, suivi par une étape de CMP (Chemical Mechanical Polishing) pour planariser la surface.

Comme indiqué sur la figure 2.1(a), le délai a longtemps été dominé par celui induit par la grille du transistor. À partir des nœuds de 250 et 180 nm, les délais des interconnexions sont devenus dominants, nécessitant des changements de matériaux métalliques et diélectriques pour diminuer le délai RC [46]. Le Cu a été choisi, étant le meilleur conducteur électrique après l'argent, étant également plus résistant à l'électromigration que l'Al [47]. Cependant, le Cu possède certains désavantages, comme une mauvaise adhésion, une mauvaise résistance à la corrosion et à l'oxydation, ainsi qu'une tendance rapide à diffuser dans les diélectriques et le Si, ce qui peut gravement impacter les performances [48].

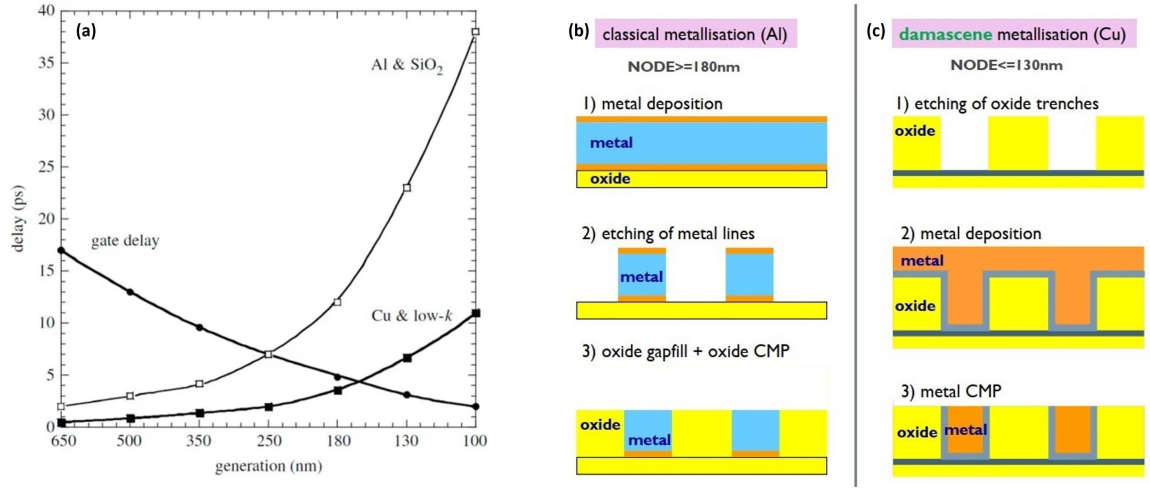


FIGURE 2.1 Interconnexion d'Al soustractif versus interconnexion Cu damascène. (a) Délais RC créés par la gate et par les interconnexions. À faible densité de noeuds, le délai RC des interconnexions devient dominant. Pour cette raison, les interconnexions en Cu ont été introduites à partir du noeud de 130 nm. (b) Procédé de fabrication pour les interconnexions d'Al : le métal est gravé et le diélectrique est planarisé par CMP. (c) Procédé de fabrication pour les interconnexions de Cu : le diélectrique est gravé et le métal est planarisé par CMP. [46]

De plus, le Cu se grave très difficilement par RIE (*reactive ion etching*) a cause de la faible volatilité des ion $CuCl_2$ et CuF a faible température [49]. Pour cette raison, le procédé damascène a été introduit comme remplacent a l'approche soustractive [50]. Comme illustré sur la figure 2.1(c), les lignes électriques sont définies puis gravées dans la couche diélectrique, ensuite les couches d'adhésion et de barrière de diffusion sont déposées. Le Cu est ensuite déposé, généralement en déposant une fine couche d'amorce (*Seed layer*) par PVD (Physical Vapor Deposition), puis le reste par électroplacage. Le Cu est ensuite poli pour enlever le surplus jusqu'à l'obtention de lignes électriques encastrées dans l'oxyde.

Dual-Damascene

Les interconnexions damascènes ont deux variantes : les structures SD (*single damascene*) et DD (*Dual-Damascene*). Le processus SD produit le via et la tranchée séparément, tandis que le processus DD réalise le motif du via et de la tranchée séparément mais leur métallisation ensemble. En raison des avantages économiques de réduire le nombre d'étapes dans un processus, le processus DD est largement préféré. Cependant, le processus SD est toujours utilisé à des fins particulières telles que M1 ou avec des couches métalliques épaisses utilisées pour les interconnexions globales [51].

2.1.2 Diélectriques

Le SiO_2 a été utilisé jusqu'au nœud 180 nm. Il possède une très bonne stabilité thermique et chimique, une bonne rigidité mécanique, est imperméable à l'humidité et peut être déposé par PECVD, ce qui permet de garantir des couches de très bonne qualité [52]. Cependant, sa constante diélectrique est relativement élevée et de nombreux développements ont dû être faits pour diminuer la constante diélectrique des matériaux isolants du BEOL afin de mitiger l'augmentation du délai RC à faible nœud. Ces matériaux sont appelés diélectriques *Low-k*. Des diélectriques dopés au fluor, FSG (*fluorosilicate glass*) avec une constante $k = 3,5$ et des propriétés thermiques, mécaniques et chimiques similaires au SiO_2, ont d'abord été introduits au nœud 130 nm [53]. Afin de réduire à nouveau la constante diélectrique, du SiO_2 dopé au CH_3, appelé SiCOH, a été introduit au nœud 90 nm [54]. Le dopage au CH_3 diminue la densité, ce qui a pour effet de diminuer la constante k ; cet effet peut être amplifié en porosifiant la couche. Le SiOCH poreux (p-SiOCH) a été introduit au nœud 45 nm [55]. Cependant, le dopage au carbone et la porosification ont pour effet de diminuer l'adhésion des barrières, la conduction thermique et la stabilité mécanique (voir figure 2.2) [56].

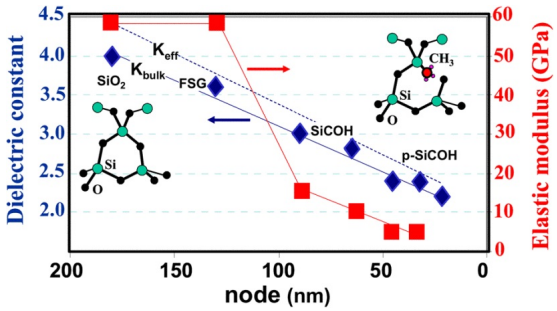


FIGURE 2.2 Évolution des matériaux diélectriques dans le BEOL pour réduire la constante diélectrique, affectant cependant la stabilité mécanique [57].

En raison des problèmes de fiabilité introduits par les matériaux *Low-k*, un intérêt croissant pour la technologie basée sur des cavités d'air (*air gap*) est apparu pour les nœuds avancés [58].

2.1.3 Métallisation Cu

La métallisation du Cu est plus complexe que celle de l'Al, car le Cu nécessite l'utilisation de couches spécifiques pour son encapsulation en raison de sa mauvaise adhésion, sa faible résistance à la corrosion et à l'oxydation, ainsi que sa tendance à diffuser dans le diélectrique [48]. Comme le montre la figure 2.3(a), une interconnexion métallique de Cu est composée d'une barrière, d'une couche de liaison (*liner*) et, à partir du nœud 14 nm, d'un métal d'encapsulation pour améliorer les problèmes d'électromigration [59]. La barrière favorise l'adhésion du métal au diélectrique, protège le Cu de l'oxydation et sert de couche de nucléation pour les métaux de la couche de liaison. La couche de liaison facilite l'amorçage du Cu et le processus de placage, et améliore l'interface du Cu pour la suppression de l'électromigration (EM) [60]. Pour les interconnexions Cu, la barrière et la couche de liaison sont généralement de TaN/Ta

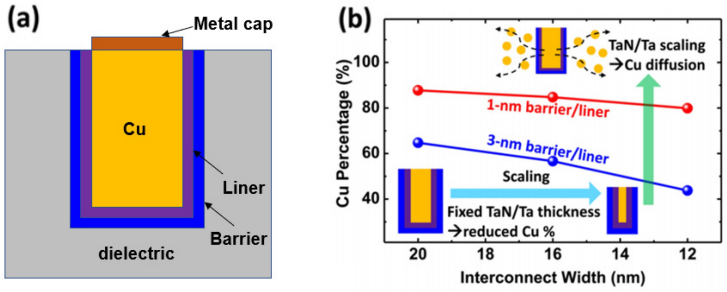


FIGURE 2.3 Interconnexions métalliques. (a) Représentation schématique d'une interconnexion de Cu avec la barrière et la couche de liaison (*liner*). (b) À faibles dimensions, la fraction volumique des couches barrière et de liaison devient significative, ce qui a pour effet d'augmenter les résistances [51].

Comme montré sur la figure 2.3(b), dans les nœuds avancés, l'augmentation de la proportion volumique des matériaux de barrière et de liaison dans les interconnexions en cuivre accroît leur résistivité. Pour minimiser l'épaisseur de ces couches sans compromettre la fiabilité, l'utilisation du Co et du Ru a permis de réduire l'épaisseur de la barrière en TaN jusqu'à 0,8 nm [61]. Avec des dimensions inférieur à 10 nm, l'absence de barrière devient nécessaire [51], poussant vers l'utilisation de matériaux alternatifs au Cu comme le W, Ir, Rh, Mo [62, 63], mais seuls le Co [64] et le Ru ont déjà été démontrés expérimentalement, le Ru étant le plus souvent cité pour des dimensions inférieurs à 10 nm [65, 66, 67]. Une analyse a également montré une meilleure conductance du Ru à faible dimension [68].

2.1.4 Procédé avancé

Les procédés SD [69] et DD [70] ont été examinés pour le Ru, envisagé comme une alternative au Cu. Toutefois, dans le procédé de Damascène de petites dimensions, une ondulation des lignes électriques, ou *wiggling lines*, résulte de la déformation des tranchées d'oxyde lors du dépôt métallique, entraînée par un effet de fermeture éclair (*zipping effect*) comme illustré sur la figure 2.4. Dans des interconnexions denses et à faible pas, l'énergie de surface élevée du Ru peut provoquer la déformation de ces tranchées, modifiant les dimensions critiques et augmentant la résistance en certains points [69]. Le Ru, ayant une énergie de surface relativement élevée comparée à celle du Co ou du TaN, rend plus complexe l'application du procédé damascène pour les interconnexions en Ru.

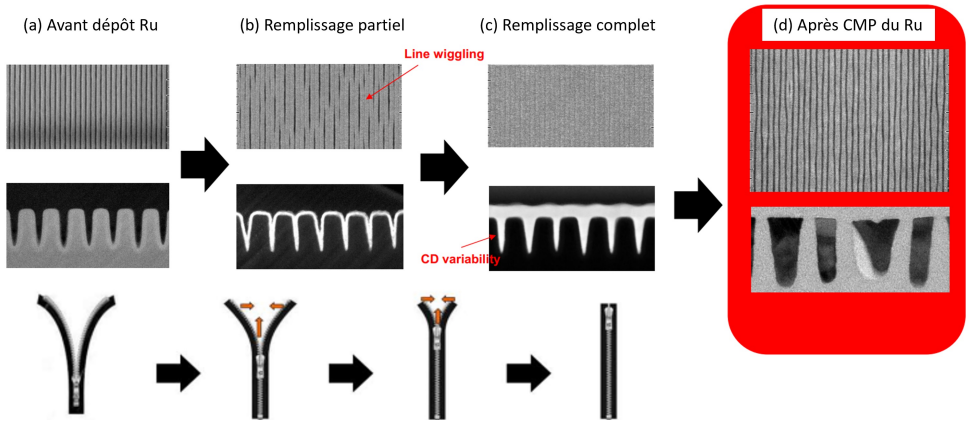


FIGURE 2.4 Effet d'ondulation dans le procédé de damascène à faible dimension. (a) Tranchée gravée dans le diélectrique avant le dépôt de Ru. (b) Pendant le dépôt, l'énergie de surface du Ru est suffisante pour déformer les tranchées de diélectrique et les rabattre sur elles-mêmes. (c) Après le dépôt du Ru, on observe que la forme des tranchées est déformée. (d) Après CMP, l'effet d'ondulation des lignes électriques peut être observé [69].

Pour ces raisons, l'utilisation des techniques soustractives pour le Ru a été réévaluée. Le dépôt du film de Ru permet d'obtenir des grains de taille supérieure, non contraints par la largeur des tranchées, contrairement au procédé damascène. Ceci réduit significativement l'augmentation de la résistance due à la diffusion aux joints de grains. Contrairement au polissage mécano-chimique (CMP), c'est le processus de dépôt qui définit l'épaisseur du métal, permettant ainsi d'atteindre des rapports d'aspect plus élevés sans les contraintes de remplissage des tranchées. Enfin, l'intégration stratégique de cavités d'air, facilitée par un rapport d'aspect accru, peut améliorer les performances. Deux nouvelles méthodes, l'approche semi-damascène et le TopVia soustractif, se présentent comme des alternatives au procédé damascène pour les composants de moins de 1 nm [71].

Procédé Semi-Damascène

La figure 2.5(c) illustre un procédé de fabrication d'une approche semi-damascène où les vias sont auto-alignés. (a) Les lignes métalliques sont gravées puis l'espace est rempli de diélectrique et planarisé par CMP (b). Les vias sont gravés (c) puis le métal est redéposé pour remplir le via et sera ensuite gravé pour former la ligne métallique suivante (d). Avec cette approche, un rapport d'aspect de 4 pour les interconnexions métalliques a été rapporté [72], ainsi que l'intégration de cavités d'air entre les électrodes, comme montré sur la figure 2.5(b).

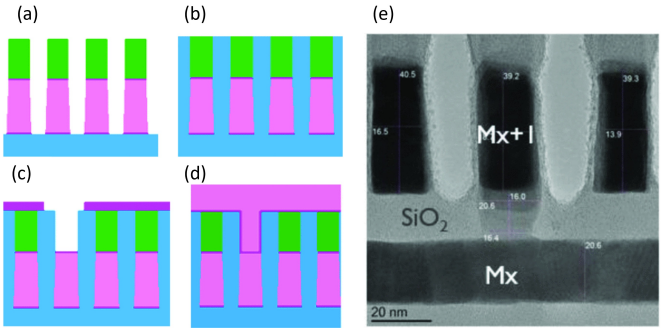


FIGURE 2.5 Procédé Semi-Damascène. (a) Gravure de lignes métalliques. (b) Dépôt de diélectrique et CMP. (c) Ouverture du Via. (d) Dépôt et planarisation du métal puis gravure des lignes métalliques. (e) Image MEB d'une implémentation semi-damascène [72].

Procédé soustractif *TopVia*

Comme illustré sur la figure 2.6, l'approche soustractive TopVia grave les lignes d'interconnexion, puis réalise une gravure partielle dans certaines zones spécifiques pour ne laisser que le via à pleine hauteur. Ensuite, le diélectrique est déposé entre les lignes et les vias, puis planarisé afin de ne révéler que le sommet des vias. Ce procédé peut ensuite être répété. Avec cette méthode, un rapport d'aspect de 4 a également été atteint [73], et l'intégration de cavités d'air entre les électrodes a été réalisée, comme le montre la figure 2.6(b).

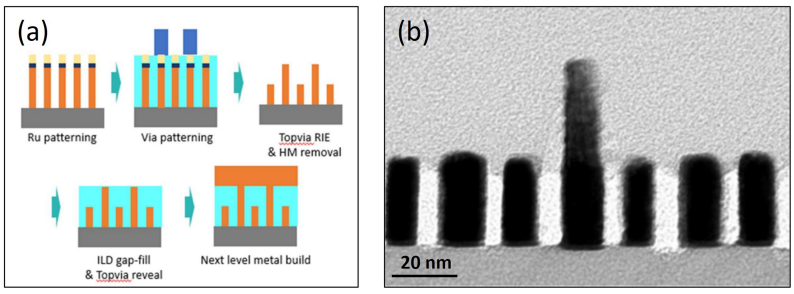


FIGURE 2.6 Procédé TopVia : (a) Procédé de fabrication. (b) Image MEB d'une implémentation TopVia [73].

Comparaison

La figure 2.7 compare les techniques de double damascène (DD) les plus avancées. Elle met en parallèle les approches avec des couches barrière/liaison de Co/TaN (c) et sans barrière au fond du via (d), avec les méthodes soustractives semi-damascène (e) et TopVia (f). Cette illustration résume l'évolution de l'optimisation des interconnexions BEOL, visant à augmenter la proportion volumique entre le matériau conducteur et les couches d'encapsulation afin de diminuer la résistance, critique à faible dimension. L'optimisation a permis de minimiser l'épaisseur des couches barrière/liaison en utilisant une combinaison de TaN/Co. Par la suite, l'élimination de la barrière au fond des vias a permis de réduire la résistance de contact.

L'adoption du Ru est de plus en plus envisagée pour les très hautes densités d'interconnexion (pas < 10 nm). Cette option permet d'augmenter à nouveau la proportion volumique, car elle ne nécessite pas de barrière, seulement une fine couche de liaison. La méthode soustractive semi-damascène a réduit la proportion de volume entre le Ru et sa couche de liaison dans les lignes, tandis que l'approche TopVia, en plus de diminuer la proportion volumique dans les lignes, permet de réduire encore cette proportion dans les vias. En outre, l'approche soustractive permet une augmentation du rapport d'aspect ; on rapporte un rapport d'aspect de 3,8 pour l'approche semi-damascène [72] et de 4 pour l'approche TopVia [73]. Enfin, l'intégration de cavités d'air auto-alignées a également été démontrée pour ces deux approches [72, 73].

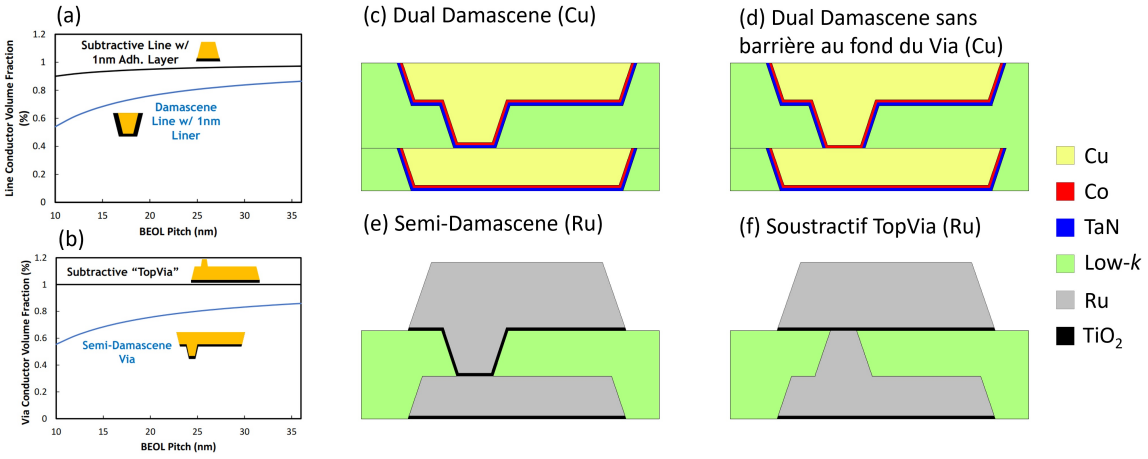


FIGURE 2.7 Procédés de fabrication d'interconnexion BEOL avancées. Comparaison entre la fraction volumique conductrice entre (a) une approche soustractive et damascène, et entre (b) l'approche TopVia et semi-damascène. Illustration des principaux schémas d'intégration : (c) dual-damascène, (d) dual-damascène sans barrière au fond du via, (e) semi-damascène et (f) TopVia [71].

2.2 Circuit RS

Les circuits RS ou réseaux crossbar servent d'unités de calcul pour réaliser l'opération VMM en utilisant des mémoires RS décrites en détails dans la section 2.3. La modulation de la résistance des mémoires est contrôlée par une tension d'écriture V_{write}, tandis que la lecture de la valeur de résistance s'effectue avec une tension plus faible, V_{read}. Comme le montre la figure 2.8(a), lorsqu'une tension est appliquée, la résistance passe d'un état de haute résistance *HRS* (*High Resistance State*) à un état de basse résistance *LRS* (*Low Resistance State*). Selon la loi d'Ohm, le courant de sortie I est égal au produit de la conductance du memristor G par la tension appliquée, comme le démontre la figure 2.8(b).

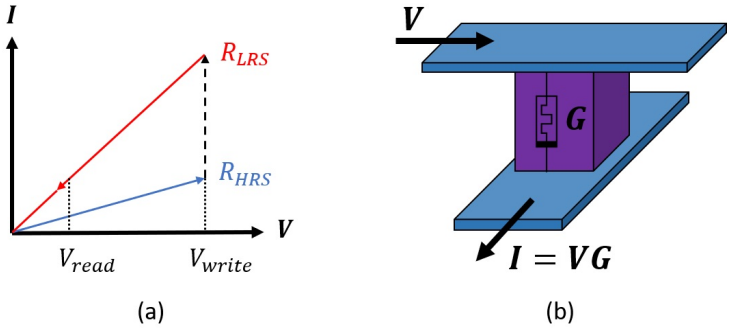


FIGURE 2.8 Illustration du comportement d'une mémoire RS. (a) Courbe I-V d'un memristor : à tension élevée V_{write}, la résistance passe d'une haute résistance R_{HRS} à une basse résistance R_{LRS}. (b) Schéma d'un mémoire RS : en bleu, les électrodes métalliques et en violet, la couche active.

2.2.1 Implémentation

À partir des propriétés de commutation de mémoire RS, il est possible de concevoir un circuit où chaque mémoire RS correspond à un poids w_{ij} du réseau de neurones par sa conductance G_{ij}, en reliant chaque croisement d'un réseau de lignes d'électrodes inférieures et supérieures transversales (voir figure 2.9(a)). Ainsi, l'opération VMM peut être réalisée directement en appliquant les lois d'Ohm et de Kirchhoff. Si la tension appliquée à chaque ligne représente une valeur d'entrée V_i et le courant traversant chaque ligne perpendiculaire est associé à une valeur de sortie I_i, l'opération liant les valeurs d'entrée et de sortie est identique à celle d'une opération VMM dans un ANN, comme démontré sur la figure 2.9 [31]. Comme mentionné dans l'introduction, cela permet d'optimiser considérablement le temps de calcul et l'énergie consommée en parallélisant les opérations d'accumulation du VMM et en éliminant le goulot d'étranglement créé par l'architecture de von Neumann.

Il existe deux méthodes pour effectuer l'étape d'apprentissage avec des circuits RS. L'apprentissage *ex situ*, où les valeurs des poids sont calculées sur un processeur externe puis

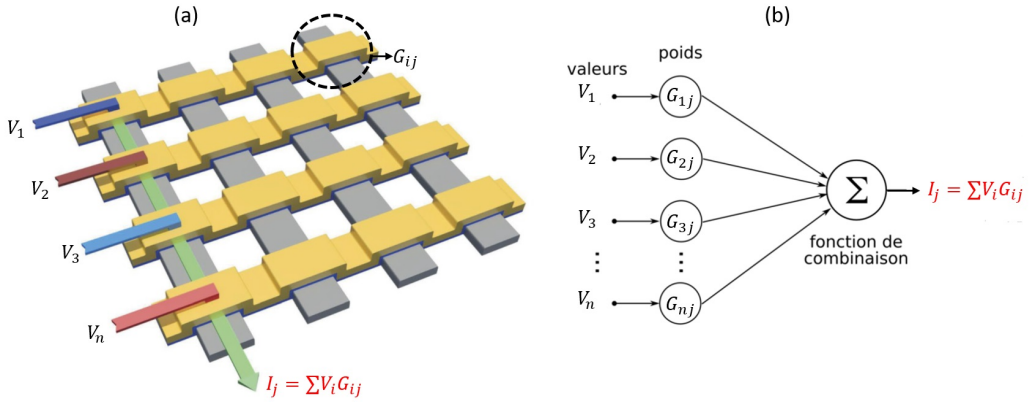


FIGURE 2.9 Correspondance entre (a) un réseau de crossbar constitué de mémoires RS et (b) l'opération de MAC dans un ANN [31, 74]

téléversées dans le circuit. L'apprentissage *in situ* se déroule quant à lui directement dans le circuit RS. L'approche hors ligne est plus simple à mettre en œuvre, est compatible avec tous les types de circuits RS [75], et a été démontrée pour des matrices passives jusqu'à 64x64 mémoires [40]. Toutefois, cette méthode est moins performante en termes de vitesse et d'efficacité énergétique, car elle nécessite un calcul externe. L'approche en ligne, ou *On-chip*, est plus intéressante pour réduire les coûts énergétiques de la phase d'apprentissage, en intégrant l'apprentissage directement dans le circuit.

2.2.2 Circuit RS passif

Dans les réseaux crossbar RS, deux architectures principales se distinguent : les réseaux passifs et les configurations 1T1R. Les réseaux passifs, également connus sous les appellations 0T1R ou simplement 1R, ne disposent pas de transistors associés à chaque mémoire RS. À l'inverse, l'architecture 1T1R incorpore un transistor par mémoire RS, comme le montre la figure 2.16. L'approche 1T1R, souvent adoptée dans les démonstrations de systèmes VMM intégrés [76, 77, 78], améliore la précision d'adressage des cellules et réduit les interférences, ce qui accroît la fiabilité du réseau mais diminue la densité. Cependant, cette technologie est contrainte par la taille des cellules et la haute conductance des dispositifs, nécessitant des circuits périphériques volumineux et consommateurs d'énergie, en opposition aux objectifs de miniaturisation et d'amélioration de l'efficacité énergétique [40]. En parallèle, les circuits RS passifs offrent un net avantage en termes de densité lorsqu'intégrés de manière monolithique [79], mais ils peuvent induire des non-idéalités, comme des problèmes de sélection de cellules et de courants parasitaires.

Courants parasitaires

Un problème inhérent aux circuits RS passifs est la présence de courants parasitaires (*Sneak path*), dû au fait que le circuit est composé uniquement d'éléments passifs, sans transistor pour contrôler le courant. Idéalement, le courant ne doit passer que par le chemin désiré (en vert sur la figure 2.10(a)). En réalité, le courant peut emprunter plusieurs chemins (en vert sur la figure 2.10(b)) du fait que la conductance des autres mémoires peut être non-nulle. Comme indiqué sur la figure 2.10(b), une résistance additionnelle non-désirée s'ajoute en parallèle. Cette résistance additionnelle peut provoquer des erreurs pendant la phase d'écriture et augmente la consommation d'énergie nécessaire pendant l'écriture [80].

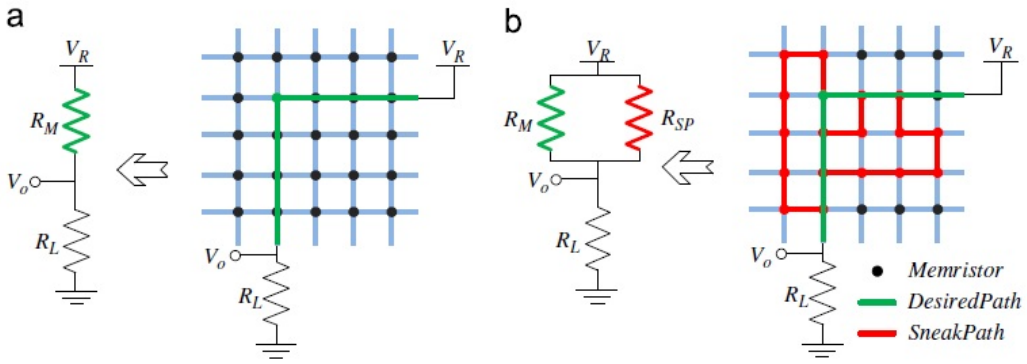


FIGURE 2.10 Illustration de l'effet des courants parasitaires. Opération de lecture dans un circuit RS et son circuit équivalent. (a) Cas idéal où le courant ne passe que par la mémoire désirée. (b) Un exemple où le courant passe par plusieurs chemins non désirés (en rouge), ce qui rajoute une résistance supplémentaire en série [81].

Le problème est d'autant plus difficile à maîtriser que la valeur de résistance due au courant de fuite dépend des valeurs de résistances des mémoires, ce qui rend cette dernière difficiles à estimer. Cependant, des composants passifs comme des diodes ou des sélecteurs peuvent être utilisés en imposant un chemin de courant à sens unique entre les électrodes supérieures et inférieures. Les courants indésirables sont ainsi éliminés, mais au prix d'un procédé plus complexe et d'une tension d'opération plus élevée [82].

Mise à l'échelle

Dans un réseau crossbar RS de $n \times n$, $2n$ transistors sont requis pour contrôler les signaux d'entrée et de sortie. Du fait que les ReRAM situés dans le BEOL ont généralement une empreinte spatiale plus petite par rapport aux transistors situés dans le FEOL, l'empreinte des $2n$ transistors nécessaires pour adresser n^2 ReRAM dépasse souvent celle des éléments de mémoire eux-mêmes dans les configurations passives [16]. Il est donc nécessaire d'agrandir la taille du réseau pour maximiser la densité d'intégration.

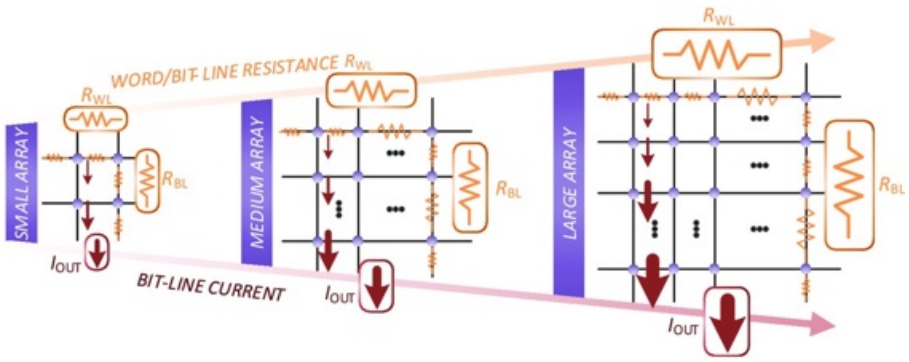


FIGURE 2.11 La mise à l'échelle des réseaux RS est contrainte par l'augmentation des résistances d'accès et l'accumulation de courants durant l'opération MAC. [16].

Cependant, comme le montre la figure 2.11, l'agrandissement d'un réseau crossbar passif présente des défis. La mise à l'échelle du réseau entraîne une augmentation des résistances d'accès aux mémoires en raison de la résistance des interconnexions. L'augmentation de la chute de potentiel aux électrodes augmente les tensions de SET/RESET requises, pouvant les rendre trop élevées pour les transistors CMOS qui les contrôlent. Finalement, plus le réseau est grand, plus le courant accumulé pendant l'opération VMM dans les électrodes de sortie augmente, ce qui peut nécessiter une empreinte spatiale plus grande pour la circuiterie CMOS associée.

Variabilité

La variabilité des dispositifs est un problème important qui peut entraîner des erreurs de lecture et diminuer le nombre maximum d'états analogiques distinguables dans les dispositifs. Les mémoires RS présentent une variabilité, que ce soit de cycle en cycle sur un même dispositif ou d'un dispositif à l'autre. Ceci est dû à leur comportement intrinsèquement stochastique, plus particulièrement pour les mémoires ReRAM [83]. Des méthodes itératives basées sur le feedback, qui alternent entre des impulsions d'écriture et de lecture, ont été reportées pour atteindre des niveaux de précision plus élevés [75][84]. De plus, il a été montré que la stochasticité des mémoires RS peut être exploitée au niveau du logiciel pour certains types d'applications basées sur les ANNs [85].

2.3 Mémoires à commutation résistive

Comme évoqué précédemment, les dispositifs à commutation résistive (RS) forment une classe de mémoire ayant suscité un vif intérêt pour l'exécution d'opérations MAC dans des accélérateurs d'IA matériels. Cette section présentera les principales catégories de mémoires émergentes. Elle se concentrera ensuite de manière plus détaillée sur une type spécifique de mémoire résistive, à savoir les OxRAMs, sélectionnées pour ce projet.

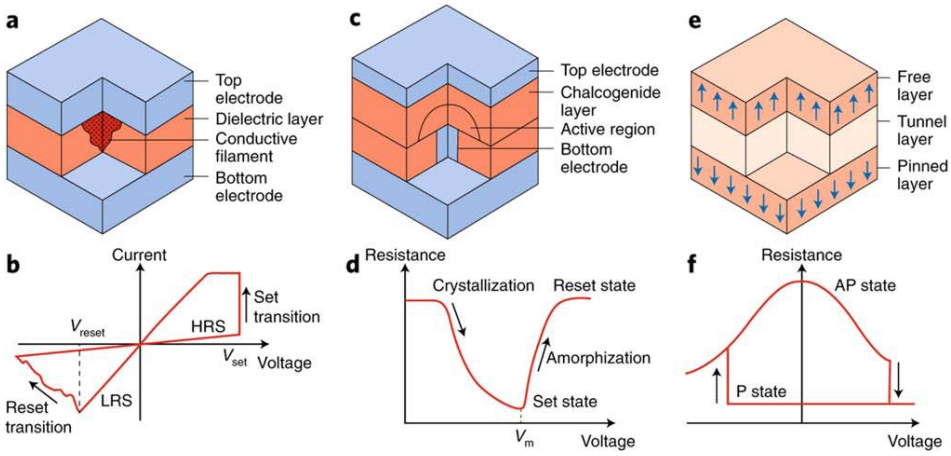


FIGURE 2.12 Illustration des principaux types de mémoire non volatiles émergentes avec leurs courbes I-V de fonctionnement respectives. (a-b) Mémoires résistives ReRAM basée sur la création et la résorption d'un filament métallique dans une couche d'oxide. (c-d) Mémoire à changement de phase (PCM) basée sur la différence de résistivité entre une phase cristalline et amorphe. (e-f) Mémoire magnétorésisitive à effet de transfère de spin (STT-RAM) basée sur l'effet de magnétorésistance à effet tunnel [86].

2.3.1 Mémoire résistive : ReRAM

Les mémoires résistives se caractérisent par une commutation basée sur la création et la résorption d'un filament conducteur au sein de la couche isolante d'une jonction MIM (*metal-insulator-metal*), comme illustré sur la figure 2.12(a). L'application d'un potentiel supérieur à une valeur seuil (V_{SET}) entraîne la migration d'éléments chargés et la création d'un chemin conducteur localisé, faisant passer la jonction MIM d'un état de résistance élevée à un état de résistance faible (transition SET). Inversement, l'opération de RESET se réalise par l'application d'un potentiel négatif inférieur à V_{RESET}, résorbant le filament et retournant la jonction à un état de résistance élevée (voir figure 2.12(b)). Les ReRAMs se divisent en deux catégories selon la charge des éléments migrants : les OxRAM (Oxide-based RAM), basées sur la migration de lacunes d'oxygène, et les CBRAM (conductive-bridging RAM), basées sur la migration de cations métalliques. Ces mémoires ont suscité un grand intérêt en raison de leur potentielle densité d'intégration élevée ($4F^2$),

grâce à leurs dimensions réduites pouvant atteindre moins de 10 nm [87]. Les matériaux de la couche active sont généralement compatibles avec les procédés CMOS et peuvent être intégrés dans le BEOL [88]. En outre, la possibilité de moduler la conductance de manière analogique rend cette technologie très pertinente pour les applications neuromorphiques. Néanmoins, les mécanismes de transport, concentrés dans des régions nanoscopiques de la couche active, sont difficiles à étudier, rendant les dynamiques de commutation partiellement comprises [89]. Le principal défi de cette technologie reste sa variabilité intrinsèque, due au mécanisme aléatoire de création et résorption du filament, qui génère des variations non seulement entre les dispositifs mais également entre différents cycles de commutation sur un même dispositif.

2.3.2 Mémoire à changement de phase : PCM

Les mémoires à changement de phase (PCM) sont un type de mémoire non volatile qui exploite la différence de résistivité entre deux phases d'un matériau chalcogénure : la phase cristalline (résistivité faible) et la phase amorphe (résistivité élevée). Comme illustré sur la figure 2.12(d), le passage de la phase amorphe à la phase cristalline est induit par l'application d'un courant qui, par effet Joule, augmente localement la température jusqu'à atteindre la température de cristallisation. Inversement, le retour à la phase amorphe se réalise par une élévation de la température jusqu'à la température de fusion, suivie d'un refroidissement rapide. Actuellement, les PCMs sont parmi les mémoires RS les plus matures. Les mécanismes de commutation sont bien compris, et leur utilisation en tant que synapses a été démontrée à de nombreuses reprises [90, 91, 92]. Elles offrent un rapport de résistance élevé et une endurance supérieure à celle des mémoires flash. De plus, le choix du matériau ($Ge_2Sb_2Te_5$) est relativement bien établi, contrairement aux mémoires résistives. Ces matériaux sont compatibles avec une intégration BEOL, et avec une gestion adéquate de la température, les PCMs peuvent offrir plusieurs niveaux de résistance, de façon similaire aux RRAMs. Cependant, la commutation étant dominée par des effets thermiques, le contrôle de la température reste un défi, surtout à petite échelle en raison des perturbations thermiques entre deux dispositifs voisins. Le processus de RESET nécessite un courant élevé pour atteindre la température de fusion, typiquement supérieur à $100\mu A$ [85], ce qui implique une consommation de puissance élevée. De plus, comme la transition amorphe/cristalline est un processus lent, le temps d'écriture est limité (entre 50 et 500 ns) [83]. Enfin, le principal facteur limitant réside dans la difficulté à maîtriser la stabilité de la phase métastable, ce qui entraîne une dérive de la résistance de l'état amorphe au fil du temps, et la rétention d'information est compromise à haute température en raison du risque de recristallisation [93].

2.3.3 Mémoire magnétorésistive : STT-RAM

Les mémoires magnétorésistives exploitent l'effet de magnétorésistance à effet tunnel. Elles se composent de deux couches ferromagnétiques : une avec une polarisation magnétique fixe (*pinned layer*) et l'autre avec une polarisation magnétique modifiable (*free layer*). Ces couches sont séparées par une couche d'oxyde, formant ainsi une jonction tunnel (*tunneling junction*). L'ensemble de ces trois couches constitue une MTJ (*jonction tunnel magnétique*). En fonction de la polarisation de la couche libre, une différence de résistance est observée entre la configuration parallèle (résistivité faible) et antiparallèle (résistivité élevée), comme indiqué sur les figures 2.12(e-f). La polarisation de la couche libre peut être contrôlée par plusieurs mécanismes, comme les STT-RAM (*Spin-transfer torque magnetic random-access memory*). Le mécanisme physique est bien compris et offre de bonnes performances. Le temps d'écriture est comparable à celui de la DRAM ($\sim 10ns$), la consommation d'énergie est inférieure à celle des PCMs ($100fJ$) et l'endurance est nettement supérieure à celle des mémoires Flash et PCM [83, 94]. Cependant, le rapport de résistance ON/OFF reste faible (environ 5 à température ambiante) et la réduction du courant d'écriture demeure un défi. Le budget thermique complique l'intégration BEOL et les procédés de fabrication influencent significativement la variabilité des MTJ, surtout à faible dimension, ce qui limite grandement la densité ($\sim 50F^2$). Plusieurs niveaux de résistance peuvent être atteints techniquement en ajoutant des domaines magnétiques dans la couche libre, permettant de moduler la proportion parallèle/antiparallèle entre plusieurs états distincts et de créer des mémoires analogiques [95]. Cette technologie est cependant encore peu développée.

2.3.4 Comparaison entre les principales mémoires

Le tableau 2.1 résume les performances des différentes mémoires RS et les compare avec les mémoires à effet de champ. En termes de compromis entre vitesse et consommation d'énergie, la DRAM et la SRAM offrent des performances qui peuvent rivaliser avec celles des mémoires émergentes. Cependant, leur volatilité, la faible densité et le coût élevé des SRAM, ainsi que la faible rétention de la DRAM qui nécessite une actualisation constante, rendent ces technologies peu adaptées pour les circuits de calcul MAC. Quant aux mémoires flash, elles souffrent d'une vitesse de lecture et d'écriture insuffisante.

Les PCM sont parmi les mémoires RS les plus matures actuellement. Avec les STT-RAM, elles représentent les technologies où les mécanismes de commutation sont bien compris. Cependant, il est difficile d'atteindre plusieurs niveaux de résistance avec ces technologies, ce qui les rend peu adaptées à des applications analogiques [82], tandis que les ReRAM ont déjà démontré la capacité de stocker plus de 6 bits d'information par dispositif [75]. Les

	ReRAM	PCM	STT-RAM	Flash(NAND)	SRAM	DRAM
	Mémoire RS			Mémoire à effet de champ		
Densité (F^2)	4	4-16	20-60	1-4	140	4-12
Énergie (pJ/bit)	0.1-3	2-25	0.1-2.5	0.00002	0.0005	0.005
t_{Read} (ns)	<10	10-70	10-35	$25*10^6$	0.1-0.3	10-25
t_{Write} (ns)	10	50-500	10-90	$200*10^6$	0.1-0.3	10-25
Rétention	Années	Années	Années	Années	Volatile	Secondes
Endurance	10^{6-12}	10^9	10^{15}	10^4	10^{16}	10^{16}

TABLEAU 2.1 Comparaison des performances entre les différentes technologies de mémoire RS et à effet de champ [83, 96, 97, 98, 85].

STT-RAM présentent des défis d'intégration, et les PCM nécessitent une consommation élevée d'énergie pendant l'écriture. Bien que les ReRAM nécessitent encore un développement significatif, elles offrent de bonnes performances techniques. En 2019, un circuit RS à base de ReRAM a affiché un temps de lecture de $t_{Read} = 6.7ns$ et une consommation énergétique de $E = 6.3nJ$ pour une opération MAC de 4 bits [97]. La ReRAM est donc une technologie prometteuse pour les systèmes VMM analogiques intégrés, grâce à ses performances en termes de densité, de vitesse et de consommation d'énergie. La section suivante examinera plus en détail la sous-classe de ReRAM basée sur la migration de lacune d'oxygène (OxRAM), qui sera utilisée dans ce projet.

2.3.5 OxRAM

Les OxRAMs sont une sous-classe de ReRAM basée sur le mouvement de lacunes d'oxygène qui modifie la résistance de l'oxyde en raison d'un changement de valence (aussi appelé *valence change memory*). Le comportement des OxRAMs peut être décrit par trois principaux paramètres :

- **La région de commutation**, localisée autour de dix à une centaine de nanomètres, qui influence fortement la stochasticité des dispositifs ;
- **Les forces motrices**, dues au potentiel électrique appliqué et à un effet joule ;
- **Les mécanismes de mouvement**, qui dirigent les espèces mobiles verticalement en raison du gradient électrique et horizontalement à cause du gradient thermique.

Selon la géométrie, il existe deux types de commutation. Les mémoires bipolaires nécessitent un voltage opposé pour former le chemin de conduction et pour le résorber, tandis que dans les mémoires unipolaires, les opérations de SET et de RESET s'effectuent avec des potentiels de même signe. Typiquement, les mémoires bipolaires sont influencées principalement par les effets du champ électrique, tandis que les unipolaires sont dominées par les effets thermiques.

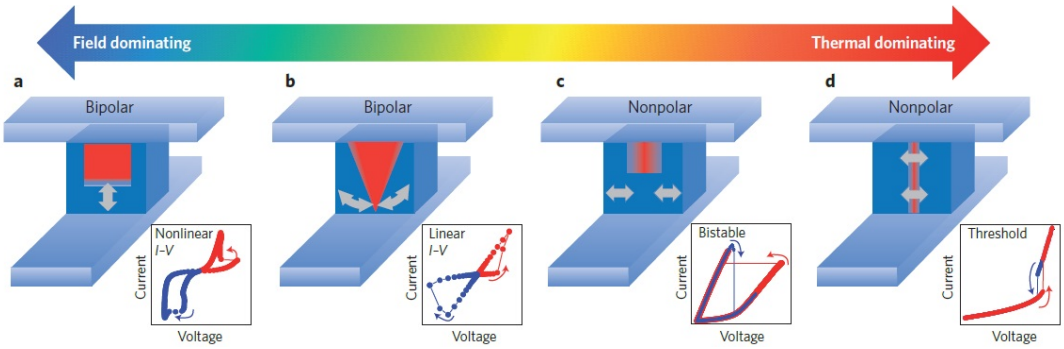


FIGURE 2.13 Illustration des différents types de OxRAM. Les mémoires bipolaires (a, b) sont majoritairement dominées par des effets de champ électrique qui induisent une formation du filament verticale, tandis que les mémoires non-polaires (c, d) sont dominées par des effets thermiques qui génèrent un élargissement du filament latéral [83].

On peut distinguer plusieurs types de mémoires. La figure 2.13(a) illustre une mémoire bipolaire non-linéaire dominée par les effets de dérive et d'électromigration, où la création et la résorption du chemin de conduction se font verticalement, dans le sens du champ électrique. La figure 2.13(b) montre le comportement d'une mémoire bipolaire linéaire, où la commutation est due aux effets combinés du champ électrique et du gradient thermique. Dans ce cas, le changement de résistance résulte de la modification de la composition et de la géométrie du chemin de conduction. Les figures 2.13(c,d) présentent deux types de mémoires non-polaires, principalement dominées par les effets thermiques.

Les OxRAM sont généralement basées sur des oxydes de transition et de nombreux types de matériaux ont été étudiés pour la couche de commutation, tels que TiO_2 [99], $SrTiO_3$ [100], NiO [101], CuO [102], ZnO [103], MnO_x [104], HfO_x [105], $TaO_{2.5}$ [106], Ti_2O_{5-x}/TiO_y [107], TaO_x/TiO_{2-x} [108], et TiO_x/Ti [40]. Le choix des électrodes est également crucial ; elles doivent empêcher la diffusion des lacunes d'oxygène depuis les électrodes vers la couche de commutation. La combinaison des électrodes métalliques avec la couche d'oxyde constitue une jonction MIM.

Le choix et la combinaison des matériaux, ainsi que leur épaisseur, ont un impact significatif sur les propriétés de commutation. Avec de nombreux oxydes de transition capables de commuter, aucun consensus clair n'a émergé sur le choix optimal. Le choix des matériaux est notamment influencé par l'application ciblée. Pour une utilisation analogique, les ReRAM en TiO_2 affichent de meilleures performances, permettant jusqu'à 7 bits de précision de programmation [75], tandis que d'autres matériaux tels que le HfO_x offrent une précision de 2 [109] à 3 bits [110].

2.4 Système VMM intégré

Un système VMM intégré se compose d'un circuit crossbar pour effectuer l'opération VMM, interfacé avec une circuitrie pour gérer les I/O du circuit ainsi que les conversions numérique-analogique et inversement. Le développement de ces systèmes a reçu une attention considérable [31, 32, 33], notamment pour des applications nécessitant des unités de calcul fonctionnant de façon autonome (*Edge computing*), c'est-à-dire sans connexion avec des serveurs massifs très coûteux en énergie (*The cloud*). Comme indiqué en vert sur la figure 2.14, ces systèmes VMM montrent des performances énergétiques supérieures de plusieurs ordres de grandeur comparées aux accélérateurs d'IA basés sur des ASIC.

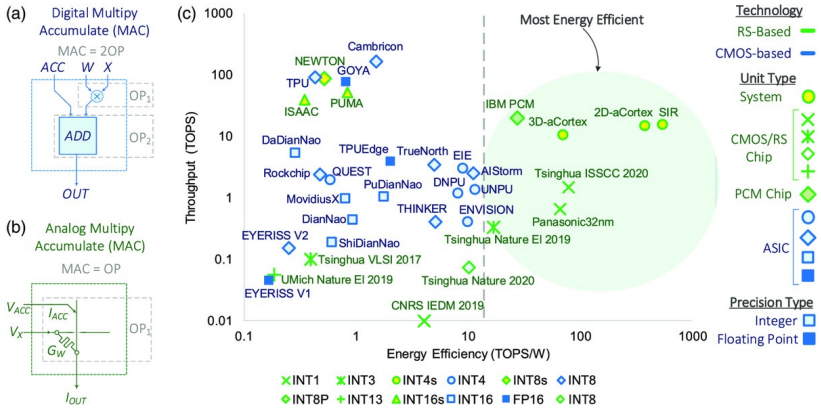


FIGURE 2.14 Comparaison des performances de calcul en fonction des performances énergétiques [16].

2.4.1 Architecture hybride numérique/analogique

Le comportement analogique des mémoires RS joue un rôle clé dans le choix de l'architecture des systèmes VMM. Ce choix n'influence pas uniquement la manière dont les poids sont encodés dans les unités de mémoire, mais affecte également la conception des circuits d'entrée et de sortie. Plusieurs configurations ont été expérimentées, allant de systèmes où les circuits d'entrée et les mémoires RS sont numériques, avec un signal de sortie analogique résultant de l'accumulation de multiplications [111], à des configurations où tant les circuits d'entrée que les mémoires et les sorties opèrent de manière numérique, normalisant le résultat analogique via un portail de majorité (*majority gate*) [112]. Pour ces approches, les mémoires RS n'ont besoin de commuter qu'entre deux états de résistance. Toutefois, certaines études explorent des configurations entièrement analogiques pour les entrées, sorties et mémoires, nécessitant des mémoires RS capables de commuter entre plusieurs états de résistance intermédiaires [97]. Plus le nombre de ces états est élevé, plus la pression sur les convertisseurs analogique-numérique (ADC) de sortie s'accroît, augmentant leur empreinte spatiale. En général, une approche plus analogique se traduit par une

empreinte spatiale plus importante pour les circuits CMOS par rapport aux mémoires RS, tout en offrant une plus grande flexibilité. Ces choix dépendent fortement de l'application spécifique visée et des contraintes de performance associées. Pour les applications nécessitant une haute précision, les architectures tendant vers des opérations analogiques avec des mémoires RS capables de multiples états de résistance offrent une précision améliorée au coût d'une complexité accrue et d'une consommation énergétique élevée due aux ADCs. À l'inverse, pour les applications tolérant une certaine imprécision, comme certaines tâches de traitement d'images où la perte de quelques bits de précision peut être négligeable, une approche plus numérique avec des états binaires simplifie l'architecture et réduit la consommation énergétique et l'espace de silicium nécessaire.

2.4.2 Compatibilité CMOS

Pour une industrialisation à grande échelle, un système VMM intégré dans le BEOL doit être compatible avec les lignes de fabrication industrielle CMOS. Pour cela, plusieurs éléments doivent être pris en considération.

Matériaux : Pour une intégration CMOS, les matériaux utilisés pour la couche de commutation doivent être compatibles avec les lignes CMOS. Des systèmes intégrés [113] et complètement intégrés dans le BEOL [114], utilisant tous deux des ReRAM à base d'Ag, ont été rapportés. Néanmoins, leur intégration CMOS semble difficile. Comme indiqué dans le Tableau 2.2, la majorité des démonstrations intégrées sont basées sur WO_x, TaO_x, TiO_x, SiO_x ou HfO_x, tous compatibles avec l'intégration CMOS. Des matériaux inertes comme le Pt ou le Pd, fréquemment utilisés comme électrodes pour les ReRAM, ou l'or pour les contacts ne peuvent être utilisés [85], et des matériaux comme le TiN et le TaN, courants dans les processus BEOL, pourront être envisagés.

Techniques de fabrication : De nombreux systèmes intégrés décrits dans la littérature emploient des procédés tels que le soulèvement (*lift-off*), inadaptés à la production de masse. Pour une compatibilité CMOS, des techniques de fabrication telles que le RIE et la CMP devraient être privilégiées.

Le tableau 2.2 répertorie plusieurs démonstrations de systèmes VMM et identifie leur compatibilité CMOS à partir des critères définis ci-dessus. La colonne "méthode de Fab." indique en vert les démonstrations qui utilisent des techniques de fabrication compatibles CMOS. De même pour la colonne "TE/BE" qui répertorie les matériaux utilisés comme électrode. La colonne "Comp. CMOS" indique la compatibilité CMOS en vert si les deux critères précédents sont respectés.

2.4.3 Niveau de maturité technologique

Le tableau 2.2 présente les différentes démonstrations de calcul en mémoire qui utilisent des ReRAM. On considère ici uniquement les démonstrations qui ont un potentiel d'industrialisation, c'est-à-dire avec des matériaux de commutation compatibles avec une fabrication CMOS et dont les matériaux des électrodes pourraient être adaptés pour être également compatibles. Ce tableau reporte les méthodes de fabrication et les matériaux utilisés pour chaque démonstration ainsi que leur intégration ou non dans le BEOL afin de définir un niveau de maturité. On définira ici de 0 à 4 le niveau de maturité technologique en fonction du nombre de critères suivants remplis :

- Intégration BEOL
- Intégration FI-BEOL (*fully integrated BEOL*)
- Compatibilité CMOS
- Procédé commercialisé

Le tableau 2.2 révèle que la majorité des démonstrations de systèmes ReRAM intégrés et complètement intégrés dans le BEOL optent pour des configurations 1T1R. Ceci s'applique également aux démonstrations utilisant des procédés de fabrication compatibles avec l'industrie CMOS. La technologie ReRAM la plus avancée actuellement (niveau 4) repose sur des mémoires numériques à base de HfOx, qui semblent issues d'un processus de fabrication développé au CEA-Leti [115]. À noter qu'un procédé d'intégration dans le FEOL, directement sur le drain d'un transistor, a également été rapporté [116].

Une seule démonstration de crossbar passif complètement intégré a été rapportée [78], avec cependant un procédé de fabrication utilisant des techniques de lift-off et des matériaux non compatibles avec l'industrie CMOS. Une seule démonstration non intégrée dans le BEOL avec un procédé compatible CMOS [40] a été rapportée. Ce procédé utilise une approche soustractive pour planariser les électrodes et est techniquement compatible avec une intégration BEOL industrielle.

À titre comparatif, le tableau 2.3 présente quelques exemples de calcul en mémoire VMM effectué sur des systèmes plus matures tels que les PCM et entièrement industriels utilisant des SRAM. Cependant, il est à noter que la taille des cellules augmente pour ces systèmes et qu'ils affichent une consommation énergétique plus élevée [117].

	Maté. de commut.	ReReRAM précision[a]	Taille - # cell.[b]	Type d'intégr.[c]	Noeud CMOS	Méthode de Fab.[d]	TE/BE[e]	Comp. CMOS	Niv./ Matu.[f]	Année/ ref.
0T1R	WOx	1-bit	25x20	SA	NA	Lift-off	Pd/W	Non	0	2017 [118]
		ND	11x3	SA	NA	Lift-off	ND	Non	0	2018 [119]
		6-bit	26x10	FI-BEOL	180 nm	Lift-off	Au/Pd	Non	2	2019 [78]
	Ta/TaOx	ND	18x2	SA	NA	Lift-off	Pd/Pd	Non	0	2017 [120]
		ND	4x3	SA	NA	Lift-off	Pd/Pd	Non	0	2018 [121]
	TiOx/Al2O3	ND	2x10x10	SA	NA	Ion-milling	TiN/Pt	Non	0	2017 [122]
	Ti/TiOx/ Al2O3	ND	10x6	SA	NA	Lift-off	Pt/Pt	Non	0	2015 [123]
		3-bit	17x20	SA	NA	Lift-off	Pt/Pt	Non	0	2018 [124]
		4-bit	64x64	SA	NA	RIE/CMP	Al/TiN	Oui	1	2021 [40]
	HfOx	1-bit	8x8	SA	NA	Lift-off	Pt/Pt	Non	0	2020 [125]
1T1R	HfOx/Al2O3	1-bit	1k	BEOL	1,2 µm	Lift-off	TiN/TiN	Non	1	2017 [126]
	HfOx	6-bit	8k	BEOL	2 µm	Lift-off	Ta/Pd	Non	1	2017 [127]
		5-bit	8k	BEOL	2 µm	Lift-off	Ta/Pd	Non	1	2018 [128]
		2-bit	448	BEOL	ND	ND	TiN/Pt	Non	1	2018 [129]
		1-bit	4k	FI-BEOL	150 nm	ND	ND	ND	2-3	2017 [111]
	Ti/HfOx	1-bit	1k	FI-BEOL	150 nm	ND	TiN/TiN	ND	2	2017 [130]
		1-bit	1k	FI-BEOL	130 nm	RIE/CMP	TiN/TiN	Oui	4	2018 [131]
		1-bit	18k	FI-BEOL	130 nm	RIE/CMP	TiN/TiN	Oui	4	2019 [132]
		1-bit	1k	FI-BEOL	130 nm	RIE/CMP	TiN/TiN	Oui	4	2020 [133]
		1-bit	1k	FI-BEOL	130 nm	RIE/CMP	TiN/TiN	Oui	4	2023 [134]
	TaOx/HfOx	4-bit	16k	FI-BEOL	130 nm	lift-off/ CMP	TiN/TiN	Non	2	2020 [77]
		4-bit	65k	FI-BEOL	130 nm	ND	TiN/TiN	ND	2-3	2020 [135]
		3-bit	158k	FI-BEOL	130 nm	ND	ND	ND	2-3	2020 [109]
	Ta/TaOx	ND	8k	BEOL	2 µm	lift-off/ CMP	Pt/Pt	Non	1	2021 [136]
		5-bit	4k	FI-BEOL	180 nm	ND	Pt/P	Non	2	2020 [137]
		4-bit	4k	FI-BEOL	180 nm	ND	Pt/P	Non	2	2021 [138]
	TaOx	1-bit	2M	FI-BEOL	180 nm	ND	ND	ND	2-3	2018 [139]
	SiOx	1-bit	1M	FI-FEOL	65 nm	ND	n-Si/TiON	ND	2-3	2019 [116]
		1-bit	1M	ND	55 nm	FI	ND	ND	2-3	2019 [140]

TABLEAU 2.2 Résumé des démonstrations de calcul en mémoire réalisées avec des ReRAM pour des configurations de crossbars 1T1R et passives (0T1R) : (a) Précision des poids. (b) Taille des réseaux de crossbars passifs et nombre de cellules mémoire utilisées en 1T1R. (c) Type d'intégration : SA (*Stand-alone*) pour une configuration sans intégration BEOL, BEOL pour une intégration dans le BEOL, FI-BEOL (*Fully-Integrated-BEOL*) lorsque les circuits périphériques d'entrée et de sortie sont intégrés, et FI-FEOL pour une démonstration entièrement intégrée avec les mémoires réalisées dans le FEOL directement sur le drain du transistor. (d) Type de procédé utilisé pendant la fabrication. (e) Matériaux des électrodes. (f) Niveau de maturité tel que défini dans la section 2.4.3.

	Type de Cellule	Précision des poids	# cellule	Integration	Noeud CMOS	Maturité	Année/ ref.
PCM	3T1C + 2PCM	>8-bit	524k	BEOL	90 nm	4	2018 [141]
	4T4R	4-bit	65k	FI-BEOL	14 nm	4	2022 [142]
	4T4R	5-bit	262k	FI-BEOL	14 nm	4	2021 [143]
SRAM	6T	1bit	ND	FI-FEOL	ND	4	2013 [144]
	10T1C	4-bit	4.5M	FI-FEOL	16 nm	4	2021 [145]
	8T	4-bit	4k	FI-FEOL	7 nm	4	2021 [146]

TABLEAU 2.3 Démonstration de calcul en mémoire VMM avec des systèmes basés sur des PCM et des SRAM.

Comme l'illustre la figure 2.15, il y a un compromis entre la maturité des technologies et la densité d'intégration ainsi que de la consommation énergétique. Les approches numériques utilisant des SRAMs ou des PCMs possèdent la maturité la plus aboutie mais avec des densités d'intégration largement inférieurs aux systèmes utilisant des ReRAMs. C'est également le cas pour l'utilisation des ReRAMs numériques comparée au ReRAM analogique. Finalement, l'approche passive montre le niveau de développement le plus bas dans l'optique d'une intégration industrielle, bien qu'elle soit celle avec le plus haut potentiel en termes de densité d'intégration et de consommation énergétique.

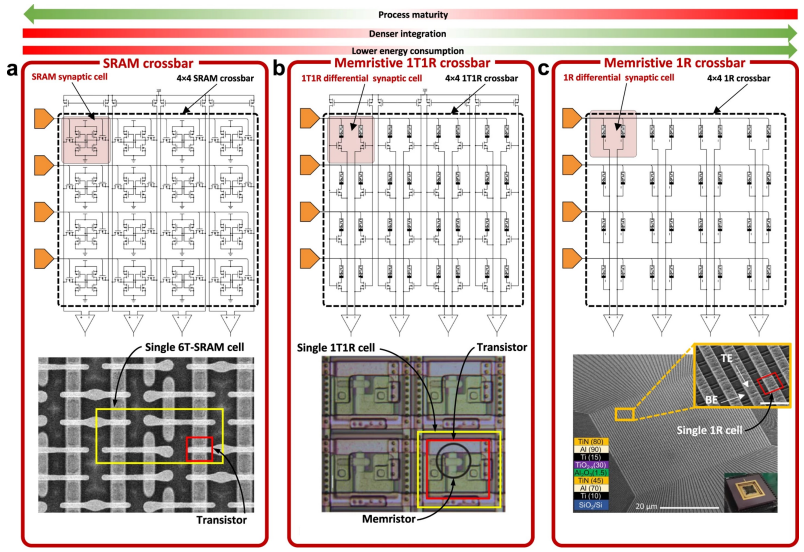


FIGURE 2.15 Illustration du compromis entre performance et maturité pour quelques exemples de technologies. (a) SRAM numérique possédant un fort niveau de maturité mais avec une empreinte spatiale élevée et une haute consommation énergétique [144]. (b) Structure 1T1R de ReRAM numérique [128], (c) ReRAM numérique passive [40]. Image tirée de la revue : [117].

2.4.4 Intégration BEOL

Comme montré dans le tableau 2.2, les architectures passives intégrées dans le BEOL sont beaucoup moins rapportées dans la littérature que l'architecture 1T1R. Les approches passives proposent des procédés de fabrication et des matériaux non compatibles avec une intégration CMOS, montrant le caractère encore peu abouti de ces approches. L'intégration dans le BEOL impose des contraintes distinctes pour les architectures 1T1R et passives des crossbars. Comme montré sur la figure 2.16(a), pour le 1T1R, les interconnexions du crossbar sont gérées par la circuiterie BEOL. Cette architecture favorise donc une intégration plus aisée dans le BEOL mais une empreinte accrue de chaque cellule due au transistor associé. Comme montré sur la figure 2.16(b), pour les architectures passives, en plus de connecter le crossbar aux interconnexions BEOL, les mémoires doivent être connectées entre elles, ce qui représente un défi d'intégration supplémentaire. Sans transistor pour gérer chaque mémoire, ces systèmes doivent optimiser les interconnexions pour minimiser les courants parasite, tout en maximisant la densité de mémoire.

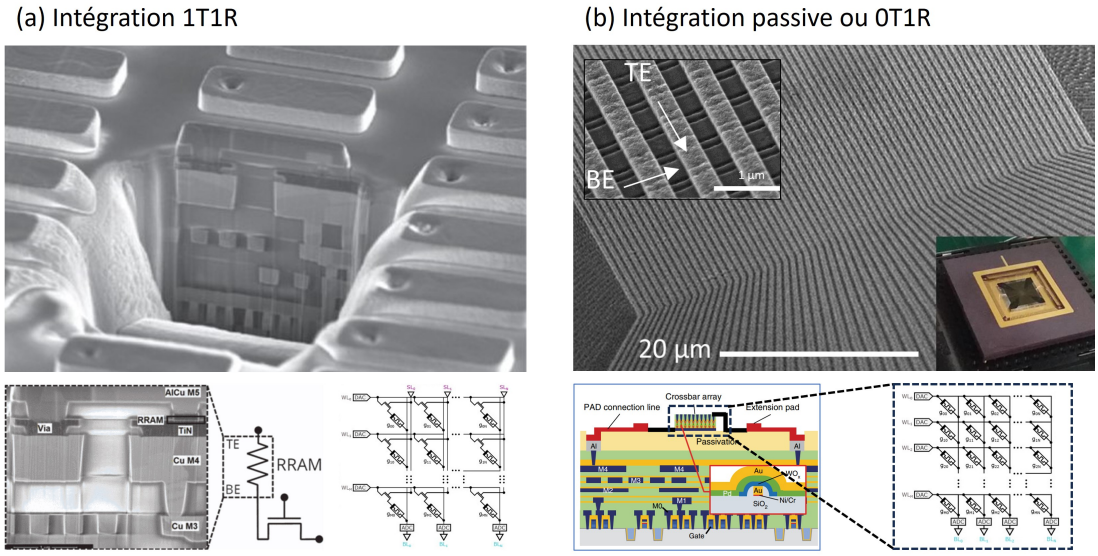


FIGURE 2.16 (a) Arrangement 1T1R avec un transistor par mémoire RS. (b) Arrangement passif (1R). Figures tirées de [78, 134, 147].

Le but des prochains chapitres sera donc de développer des procédés de fabrication de circuits passifs ReRAM et de les intégrer dans le BEOL de puces CMOS. Les procédés de fabrication adapteront les schémas de fabrication introduits dans la section 2.1. Ces choix seront discutés en fonction de leur potentiel à résoudre les défis inhérents au crossbar détaillés dans la section 2.2. Des mémoires ReRAMs utilisant un matériau de commutation de TiO_x pour ses propriétés analogiques, comme détaillé dans la section 2.3, et des électrodes de TiN pour garantir la compatibilité CMOS.

PARTIE I

Développement et fabrication

CHAPITRE 3

Procédé CMP de damascène vs soustractif

Avant-propos de l'article

Titre Complet : Damascene Versus Subtractive Line CMP Process for Resistive Memory Crossbars BEOL Integration.

Auteurs et affiliations : R. Dawant[1,2], M. Gaudreau[1,2], M.-A. Roy[1,2], P.-A. Mouny[1,2], M. Valdenaire[1,2], P. Gliech[1,2], J. Arias Zapata[1,2], Malek Zegaoui[3], F. Alibart[1,2], D. Drouin[1,2] et S. Ecoffey[1,2],

[1]Institut Interdisciplinaire d'Innovation Technologique (3IT), Université de Sherbrooke, Sherbrooke, Québec J1K 0A5, Canada
[2]Laboratoire Nanotechnologies Nanosystèmes (LN2) – CNRS UMI-3463 – 3IT, Sherbrooke, Québec J1K 0A5, Canada
[3]Institute of Electronics, Microelectronics and Nanotechnology (IEMN), Université de Lille, Villeneuve d'Ascq 59650, France

Date de publication : Avril 2024
Journal : Micro and Nano Engineering
Volume : 23, pp. 100251 (2024)
Référence : 10.1016/j.mne.2024.100251 [42]

Contribution du document :

Les configurations 1T1R ont permis d'importantes avancées dans l'intégration de systèmes VMM. Comme le montre la section de l'état de l'art 2.4.3, cette approche présente une maturité bien plus avancée que l'approche passive. Cependant, comme expliqué précédemment, les circuits ReRAM passifs offrent un avantage significatif en termes de densité lorsqu'ils sont intégrés de manière monolithique dans le BEOL. Cela suscite un intérêt prononcé pour le développement de structures pouvant être fabriquées au niveau BEOL afin de maximiser les bénéfices des circuits passifs.

Dans ces configurations passives, les principaux défis proviennent de la densité des éléments de mémoire et des interconnexions du circuit crossbar, tout en utilisant des techniques compatibles avec une intégration BEOL. S'inspirant des techniques de fabrication BEOL, plusieurs schémas d'intégration pour la jonction MIM dans un réseau crossbar à haute densité sont envisageables. Les méthodes traditionnelles telles que les interconnexions en aluminium par procédé soustractif, bien établies, ou le processus damascène, qui reste l'approche standard, ainsi que d'autres propositions innovantes inspirées des nœuds avancés, comme les techniques soustractives, sont envisagées.

Pour ces raisons, une étude a été réalisée afin de développer et comparer les performances de trois procédés de fabrication. Les stratégies d'intégration se distinguent principalement par la méthode employée pour planariser la surface : la méthode Damascène, caractérisée par le polissage du métal, et la méthode soustractive, où c'est le diélectrique qui est poli. L'objectif de cette étude est de développer ces trois procédés en SA (non intégrés dans le BEOL) et d'évaluer l'impact sur les performances des OxRAM dans le but d'une intégration sur CMOS. Le transfert et la démonstration de ces approches sur des puces CMOS de TSMC sont abordés dans le chapitre 6.

Résumé en français

Ces dernières années, les mémoires résistives se sont imposées comme une avancée clé dans le domaine de l'électronique, offrant de nombreux avantages en termes d'efficacité énergétique, de mise à l'échelle et de non-volatilité [148]. Caractérisées par leur comportement unique de commutation résistive, ces mémoires conviennent à une variété d'applications, allant du stockage de données haute densité à l'informatique neuromorphique [16]. Leur potentiel est encore renforcé par leur compatibilité avec les processus semi-conducteurs avancés, permettant une intégration sans heurts dans les circuits électroniques modernes [149]. Une voie particulièrement prometteuse pour la mémoire résistive réside dans son intégration au stade Back-End-of-Line (BEOL) de la fabrication de semi-conducteurs [79]. L'intégration BEOL implique des étapes se déroulant après la fabrication des transistors, se concentrant principalement sur la création d'interconnexions qui lient électriquement ces transistors. Intégrer des mémoires résistives à ce stade peut conduire à des architectures compactes, efficaces et performantes, cruciales pour les applications de calcul en mémoire où stockage et traitement des données sont co-localisés [31]. Cet article étudie trois manières d'intégrer la mémoire résistive à base de TiO_x dans des structures de circuit passif, en utilisant des processus de polissage mécanochimique (CMP), en se concentrant sur l'identification des techniques d'intégration optimales.

Damascene Versus Subtractive Line CMP Process for Resistive Memory Crossbars BEOL Integration

Abstract : In recent years, resistive memories have emerged as a pivotal advancement in the realm of electronics, offering numerous advantages in terms of energy efficiency, scalability, and non-volatility [148]. Characterized by their unique resistive switching behavior, these memories are well-suited for a variety of applications, ranging from high-density data storage to neuromorphic computing [16]. Their potential is further enhanced by their compatibility with advanced semiconductor processes, enabling seamless integration into modern electronic circuits [149]. A particularly promising avenue for resistive memory lies in its integration at the Back-End-of-Line (BEOL) stage of semiconductor manufacturing [79]. BEOL integration involves processes that occur after the fabrication of the transistors, primarily focusing on creating interconnections that electrically link these transistors. Integrating resistive memories at this stage can lead to compact, efficient, and high-performance architectures, pivotal for in-memory computing applications where data storage and processing are co-located [31]. This paper studies three ways to integrate TiO_x-based resistive memory into passive crossbar array structures, using chemical mechanical polishing (CMP) processes, focusing on identifying the optimal integration techniques.

3.1 Introduction

In recent years, resistive memories (ReRAM) have emerged as indispensable components for mixed-signal circuits, pivotal for executing vector-by-matrix multiplication (VMM) in artificial neural networks [148]. 1T1R memory arrays, combining a resistive switching element (1R) with a transistor (1T), have been widely studied and demonstrated breakthroughs in broad implementation of neuromorphic computing [76, 77, 78]. Yet, the technology is constrained by its cell size and the high conductance of devices, necessitating the use of large, energy-demanding peripheral circuits and challenging the drive towards miniaturization and improved energy efficiency [40]. In contrast, passive ReRAM circuits offer a substantial density advantage when integrated monolithically in the Back-End-of-Line (BEOL) [79]. In a $n \times n$ passive ReRAM array, due to the smaller footprint of the ReRAM in the BEOL compared to transistors located in the Front-End-of-Line (FEOL), the footprint of the $2n$ transistors, necessary for addressing n^2 ReRAM, often surpasses the footprint of the memory elements themselves in passive configurations [16]. Even for larger arrays where the total area of $2n$ transistors becomes lower than that of the n^2 memory array, ReRAM arrays can utilize the inherent potential for 3D integration provided

by the BEOL process. In 1T1R configurations, the footprint is primarily constrained by the transistor size. However, in passive arrays, challenges arise from the density of memory elements and the interconnections that form the circuit, referred to as a crossbar array. Consequently, there is significant interest in developing dense crossbar memory array structures that can be fabricated at the BEOL level to enhance the massive advantage of passive circuits. Inspired by the technique used to fabricate the CMOS interconnections, this paper aims to study different approaches to integrate crossbar arrays in the BEOL and enhance the 3D integration.

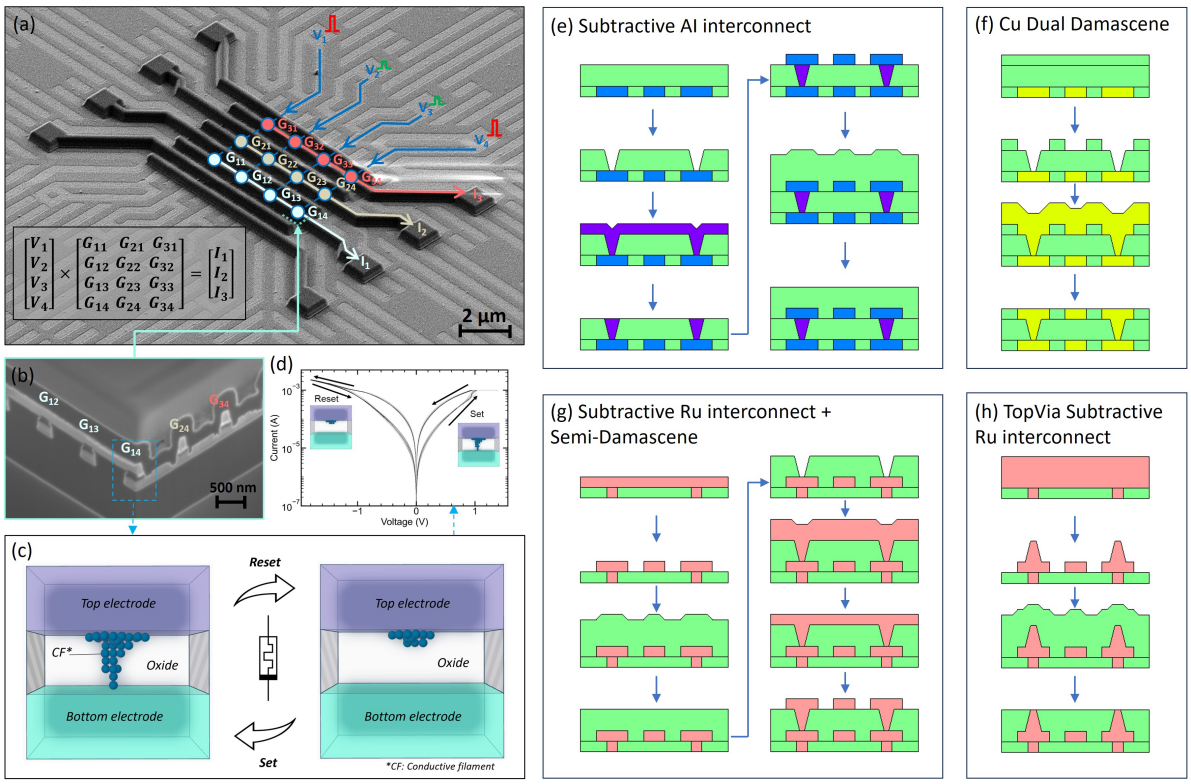


FIGURE 3.1 (a) SEM image of a crossbar array illustrating the use of resistive memory for dense in-memory computing, with (b) a FIB cross-section showing the MIM capacitance of the memory. (c) Diagram of the MIM junction depicting ion migration that enables switching between 'Set' and 'Reset' states in resistive memories, and (d) the current-voltage graph displaying the switching behavior. Main process flow used for BEOL interconnects : (e) Subtractive Al interconnect fabrication with IDL planarization [150]. (f) Cu Dual Damascene interconnect based on metal polishing [151]. (g) Ru interconnect combined with Semi-Damascene process [152] and (h) TopVia Subtractive Ru interconnect [153].

As shown in Fig.3.1(a), passive ReRAM circuits consist of metallic crossbar electrodes arrayed with a switching memory located at the intersection of each electrode (see Fig.3.1(b)). This circuit can perform direct VMM by taking advantage of Ohm's and Kirchhoff's laws.

The ReRAM used as the switching element are devices composed of a metal-insulator-metal (MIM) junction, as illustrated in Fig.3.1(c). As shown in Fig.3.1(d), applying a positive or negative voltage to the MIM junction can lead to the migration of ions in the insulating layer, creating a change in the junction's resistance. Different integration schemes for the MIM junction are compatible with the CMOS production lines and support a high interconnection density. Traditional methods, such as subtractive aluminum (Al) interconnects, shown in Fig. 3.1(e), involving Al line etching followed by interlayer dielectric (ILD) deposition and planarization, have established precedents [150]. However, the copper (Cu) dual Damascene process, which entails etching vias and trenches in ILD before the metal deposition and CMP (see Fig. 3.1(f)), remains the standard approach for advanced nodes [151]. Other innovative proposals for advanced CMOS nodes include subtractive metal line techniques complemented by semi-Damascene processes, as shown in Fig. 3.1(g) [152], and the TopVia subtractive method (see Fig. 3.1(h)) [153]. All of these proposals enable the creation of a planar Via and metal line block that can be repeated multiple times. Based on these integration strategies, two primary categories can be distinguished by the method employed to level the surface : the Damascene approach, characterized by the polishing of the metal, and the subtractive method, where the dielectric is polished. The next section presents different integration schemes that use these techniques to fabricate crossbar arrays and assesses their practicality for BEOL integration.

3.2 Fabrication process

We developed three integration schemes to fabricate passive resistive memory crossbar arrays : a Damascene approach (a), a single-subtractive approach (b), and a double-subtractive approach (c). Fig. 3.2 shows the process flow of the three fabrication approaches. The Damascene (a) and subtractive approaches (b–c) are distinguished by the planarization method used. The differences between the single (b) and double-subtractive approaches (c) come from the pattering of the TiO_x/Ti switching layers (SL). In the single-subtractive approach, the SL is patterned and etched a single time with the top electrode (TE). In the double-subtractive process, the SL is patterned and etched a first time with the bottom electrode (BE) and a second time during the pattering of the TE, which results in the pattering of the SL only at the intersection of each electrode without adding another lithography mask.

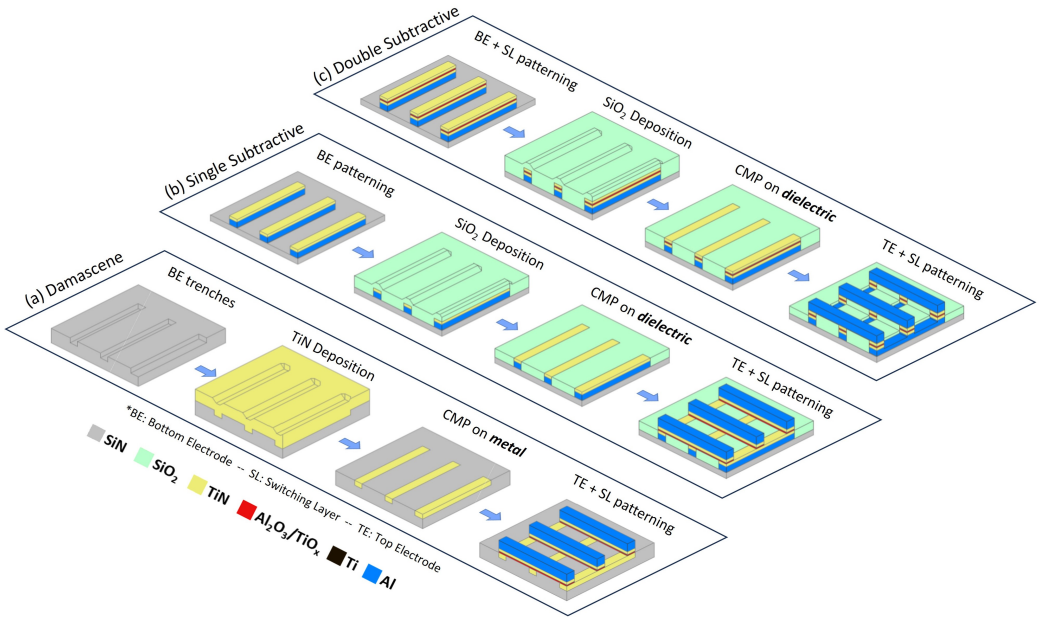


FIGURE 3.2 Process flow for the three approaches : (a) **Damascene** : BEs are etched in the SiN substrate. A 600-nm TiN layer is sputtered and polished by CMP. The switching stack ($Al_2O_3/TiO_{2-x}/Ti/TiN$) is deposited followed by the deposition of the Al TE. The SL and the TE are patterned. (b) **Single Subtractive** : Pattering of Al/TiN BEs, A 600-nm SiO_2 layer is deposited and planarized via CMP. Subsequent steps are as in the Damascene approach. (c) **Double Subtractive** : Follows the initial steps of the single-subtractive method, but after Al/TiN BEs deposition, the SL is added immediately. BEs and SL are defined. SiO_2 deposition and CMP follow, with TE deposition and patterning. During this step, a second etching of the SL at electrode intersections concludes the process.

3.2.1 Process Flow

For all the integration schemes tested, the MIM junction is composed of :

$$\mathrm{TiN(M)} - \mathrm{Al_2O_3/TiO_{2-x}/TiO_{2-x}(I)} - \mathrm{TiN(M)}. \tag{3.1}$$

The selected switching stack was chosen for its analog properties, demonstrating multi-bit precision in prior studies [154, 155, 156], and its proven applicability for VMM operations in passive ReRAM crossbar [40, 124].

The TiN metallic layers are at least 30 nm thick to prevent oxygen diffusion from the insulator stack inside the potential Al metallic contact electrodes. The insulator part of the MIM junction is a switching stack of Al_2O_3 (1.4 nm) deposited via atomic layer de-

position, and TiO_{2-x}(30 nm) deposited by reactive sputtering, along with Ti (10 nm)/TiN (30 −70 nm). The vacuum was not broken between the deposition of the TiO_{2-x}, Ti, and TiN layers. The stoichiometry of TiO_{2-x} was tuned by controlling the reactive sputtering regime and the oxidation state of the target as described in a previous study- [157]. With this process, we did demonstrate stable state retention at room temperature [156].

3.2.1.1 Damascene

For the Damascene approach, illustrated in Fig. 3.2(a), BEs were defined by electron beam lithography and patterned using inductively coupled plasma (ICP) etching within SiN to a depth of 150 nm. A 600 nm thick TiN metallic layer was deposited by sputtering and polished by CMP until the excess metal between electrodes was removed. The switching stack was deposited with 30 nm of TiN followed by the TEs of Al, 200 nm thick, by evaporation. The switching stack and the TEs were patterned using electron beam lithography (EBPG5200) and ICP etching with BCl_3/Cl_2/Ar chemistry.

3.2.1.2 Single Subtractive

For the single-subtractive approach, shown in Fig. 3.2(b), a metallic stack of Al/TiN BEs with a corresponding thickness of 150/70 nm was deposited by sputtering. The BEs were defined by electron-beam lithography and patterned by ICP etching using BCl_3/Cl_2/Ar chemistry. A 600 nm thick SiO_2 layer was deposited by PECVD and planarized via CMP. The overburden was removed until the metal electrodes were revealed. For the remainder of the process, the fabrication steps were the same as the Damascene approach. The switching stack was deposited with the 200 nm Al TE in the same way as explained in the Damascene approaches, then patterned using electron beam lithography and ICP etching with BCl_3/Cl_2/Ar chemistry.

3.2.1.3 Double Subtractive

For the double-subtractive approach, the metallic stack of Al/TiN BEs deposition is directly followed by the deposition of the SL using the same deposition techniques as presented for the previous approach but with a thicker top TiN layer of 70 nm. The BEs and the SL were defined by electron-beam lithography and patterned by ICP etching using BCl_3/Cl_2/Ar chemistry. A 600 nm thick SiO_2 layer was deposited by PECVD and planarized via CMP until the metal electrodes were visible. The same CMP recipe as the single-subtractive approach was used. TEs of Al, 200 nm thick, were then deposited by evaporation. The TEs were patterned using electron beam lithography and ICP etching with BCl_3/Cl_2/Ar chemistry as in the previous approaches but this was directly followed by a second etching of the SL using the mask of the TE. This results in the self-aligned patterning of the SL only at the intersection of each electrode, as shown in Fig. 3.2(c)

3.2.2 CMP Development

CMP is a critical step to obtain good planarity and enable the integration inside the planar levels of interconnections in the BEOL, as well as a 3D integration. It is a process that employs both chemical reactions and mechanical abrasion to remove materials to achieve a leveled surface. It requires the use of specific slurry solutions containing abrasive particles to control the rate of material removal and surface finish. Several factors influence the choice of the slurry : the material to be polished, the removal rates, and the selectivity between the different materials, among other parameters. In the case of the Damascene approach, high material removal rate (MRR) selectivity is desired between the TiN BE metal and the SiN substrate. On the other side, single and double-subtractive approaches require a good selectivity between the SiO_2 used to encapsulate the device and the TiN electrode. To address these requirements, we considered two types of slurries :

Silica- oxide-based slurries are usually used for the polishing of both SiO_2 and SiN but can also be used for the polishing of TiN with the addition of an oxidizing agent such as H_2O_2 or NaClO. The detachment of the polished material is facilitated by mechanical contact between the particles and the sample. These slurries are generally basic to promote hydration reactions [158, 159, 160].

Cerium-dioxide-based slurries are typically used to polish oxides and stop on nitride materials. The particles are softer than silicon dioxide slurries. Abrasion occurs in two steps [161] : First, the cerium dioxide molecule attaches to the silicon oxide surface, and then, the mechanical contact between the pad and the sample removes it along with the oxide [160, 162]. No literature has been found regarding the abrasion mode of CeO_2 on TiN. Nevertheless, preliminary tests have shown potential for TiN polishing and will be studied in the next section. More details on the mechanism for both of these slurries can be found in Supplementary Data, Sections 1–2.

CMP processes are also highly dependent on the layout density, and a non-uniform layout can result in unwanted effects, such as erosion and dishing, which can negatively affect the planarization process. For these reasons, we used filling dummy structures to obtain a density of structures between 40%–60%, and large structures were stripped to limit the maximum dimension/spacing in the three approaches studied.

3.2.2.1 Metal polishing

The highest selectivity between TiN and SiN was achieved by the CeO-based slurry with 1% H_2O_2 (see Supplementary Data, Section 3). However, it has been demonstrated that CeO_2 slurry is ineffective in polishing topographic structures [163]. CeO_2 particles tend to agglomerate at the bottom of trenches [164], preventing mechanical contact with the pad

required to remove the metal. In our tests, we observed that planarization was ineffective when the topography of the trenches was greater than 50 nm. For this reason, we used dummy structures below 500 nm, 150 nm depth trenches, and a 600 nm thick TiN layer. This results in a smoother surface that decreases the topography below 50 nm. Due to the good CeO_2 selectivity and the fact the filling structure matches the dimension of the electrode, erosion was reduced and planarity below 10 nm was achieved. This approach imposes strict design rules with a maximum dimension of 500 nm. Note that we were able to use SiO_2-based slurry with micrometer dummy structures to planarize the surface and use CeO_2 to end the process, resulting in higher topography (see Supplementary Data, Section 4).

3.2.2.2 Dielectric polishing

The CMP is performed on SiO_2 for both the subtractive and double-subtractive approaches. Most processes for resistive memory passive crossbar array fabrication used similar approaches and revealed the bottom electrode by plasma etching [40]. Note that we tested planarizing the structure and revealing the structure by plasma etching but this approach yielded poor results (see Supplementary Data, Section 5). Nevertheless, similar to the top via subtractive interconnect [153], CMP can be used to reveal the bottom electrode. In this case, the TiN layer acted as the stop layer. The slurry with the highest SiO_2/TiN selectivity considered in this study is a SiO_2-based particle (50 nm) slurry without any additive. The selectivity for this process is close to 1 (0.95). Because both materials have almost the same MRR, the challenge is end-point detection (EPD). For this reason, we use a relatively thick TiN layer of 70 nm as a buffer. 30 nm thick TiN is required to be efficient as an oxygen barrier, meaning that 40 nm of TiN can be polished. The EPD could be enhanced by adding a SiN layer on top of the TiN electrode to function as a stop layer, similar to the STI process [165], with a CeO_2-based slurry. The SiN on top of TiN could be selectively removed etching processes.

3.3 Results and discussion

3.3.1 Morphological Characterization

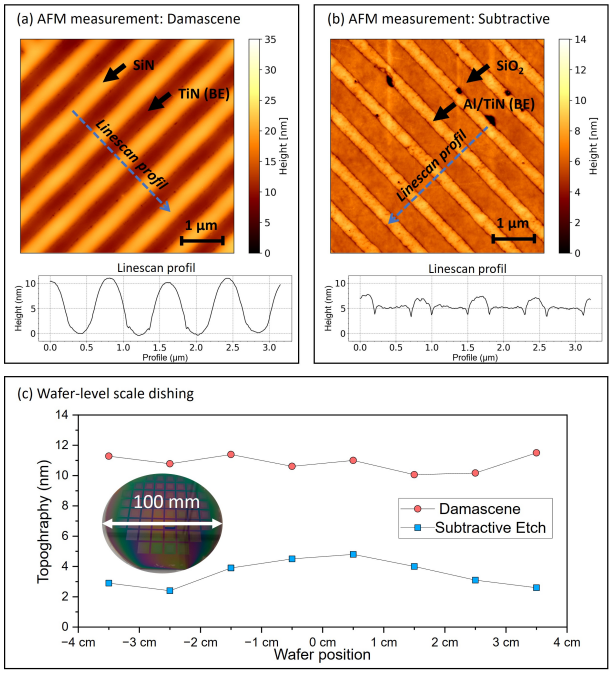


FIGURE 3.3 Topographical comparison : (a) AFM measurement after CMP of the Damascene approach (b) and the subtractive approaches. (c) Dishing comparison across a 4-inch wafer, highlighting Damascene's higher dishing versus the subtractive approaches.

Fig. 3.3(a,b) depicts the atomic force microscopy (AFM) measurements of the devices following CMP for the Damascene and subtractive approaches, and Fig. 3.3(c) compares the dishing measured by AFM on different devices over a 100-mm wafer. The Damascene approach exhibits a higher dishing, of 11±1 nm, while the subtractive process has a dishing of 3±1.5 nm. Although the subtractive technique has better planarity, it is essential to keep in mind that the EPD is also more difficult to find due to the low selectivity. Furthermore, the low dishing observed correlates with the low selectivity, as both materials are polished at a similar rate. For the Damascene approach, uniform dishing is observed over the wafer. In contrast, for the subtractive approach, a higher polishing rate is observed at the center of the wafer compared to the edges. Note that we didn't observe significant differences between the single and double-subtractive etching since both are SiO_2 polishing stopping on TiN.

Fig.3.4 shows Focused ion beam (FIB) microscopic images of the fabricated devices for the tree approaches. As shown in Fig. 3.4(a), the Damascene approach exhibits a trenching effect on the BEs, which is caused by the different orientations of the grain during the

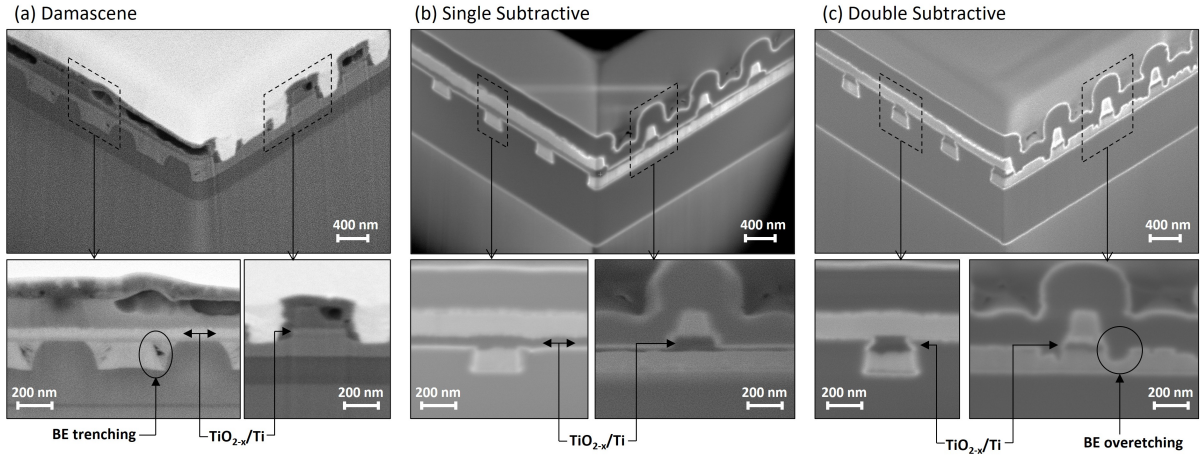


FIGURE 3.4 Cross-sectional comparison : (a) FIB cross-section of the Damascene process highlighting trenching effect. (b) FIB for the single-subtractive showing a cleaner etch profile, and (c) FIB for the double-subtractive demonstrating the results of the second etching with distinct memory layer patterning at the electrode intersections.

deposition of TiN. This creates a defective surface and air gap, which hinders the scaling down of interconnects. A similar trenching effect is also observed in the single and double-subtractive approach but occurs within the dielectric layer. As will be shown in the next section, this effect drastically degrades the performance of Damascene devices when scaling down the electrode. In contrast, in the single and double-subtractive approaches, this is limited by the resolution of the lithography.

Fig. 3.4(c) shows the result of the overetching after the second memory patterning in the double-subtractive approach. It results in the memory layer's being patterned only at the intersection of the BE and TE, unlike in the single-subtractive and Damascene approaches, where the memory layer is patterned with the TE. Non-uniform overetching in the BE can be observed close to the intersection. This originates from the chlorine-based gases used to etch the memory stack, which also etch the Al/TiN of the BE. This could be improved by using an etching stop layer between the Al and the TiN with a metallic layer that sustained chlorine gases like W.

3.3.2 Electrical Characterization

The devices performances for the three approaches were studied through current–voltage (I–V) characteristics. The impact of varying the widths of the electrodes was also investigated. For each approach and device dimension, 20 devices were tested by applying 5 bidirectional voltage sweeps. The high resistance state (HRS) and low resistance state (LRS) are defined at 0.2 V. A switching cycle occurs when the device transitions from HRS to LRS with a 2 :1 ratio. If the transitions do not meet this ratio, the device is not considered to have switched during that cycle. A device is only considered to be cycling if it undergoes two or more successful transitions. Otherwise, the device is considered faulty.

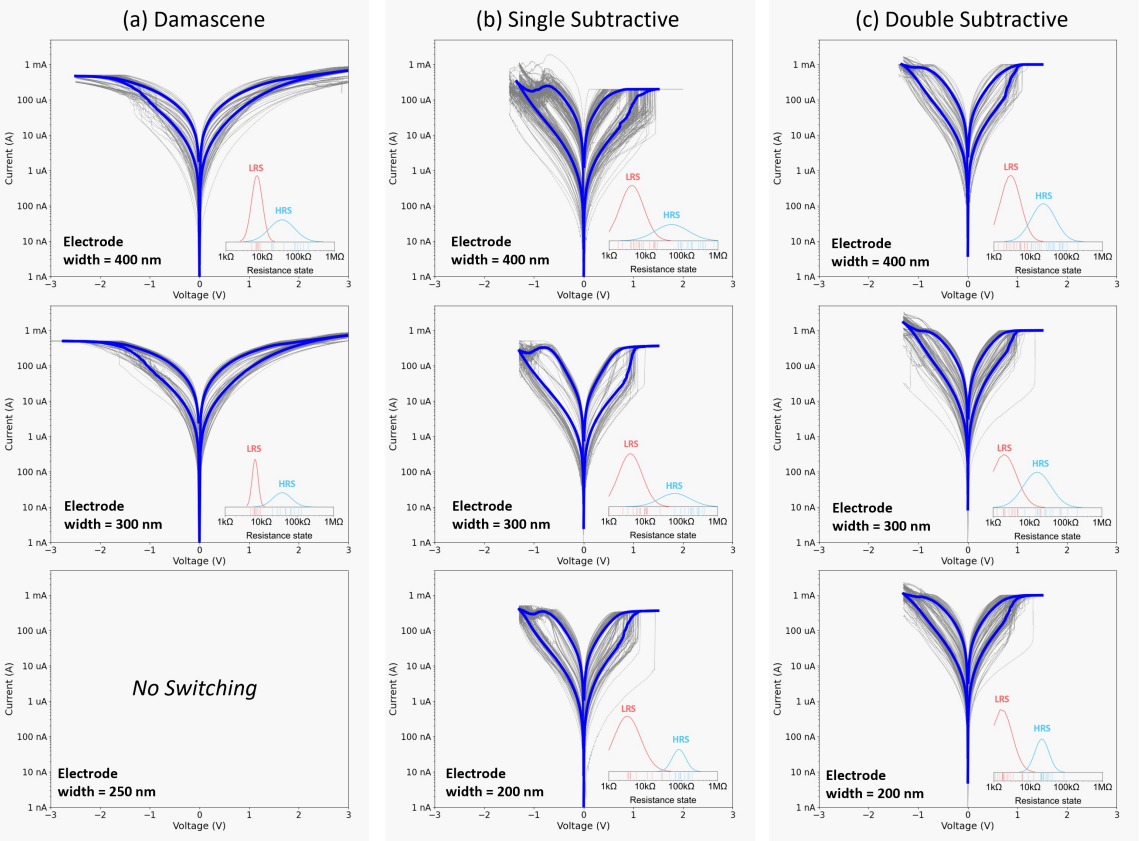


FIGURE 3.5 **I–V Characteristics** for (a) Damascene, (b) single, and (c) double-subtractive approaches, showing mean cycling curves for devices with different width in blue and corresponding HRS/LRS histogram inset. Damascene devices exhibit higher SET/RESET voltages (2.5/3 V) compared to the subtractive approaches (1/1.3 V), attributed to higher TiN electrode resistance for Damascene. Damascene shows lower ratios, with significant degradation below a width of 300 nm due to the trenching effect. The best performance is seen in the single-subtractive approach.

Fig. 3.5(a–c) shows the I–V switching cycle for devices with electrodes from 400 to 200 nm wide. The mean characteristic of cycling devices is shown in blue. The single and double-

subtractive approaches have similar SET/RESET voltages of approximately 1/1.3 V. For Damascene, the SET/RESET voltage was approximately 2.5/3 V. The higher switching voltage of Damascene is explained by the higher voltage drop along the TiN electrodes owing to their higher resistance. Fig. 3.5(a–c) inset shows the HRS and LRS histogram for each cycling device for the different approaches and dimensions. The Damascene approach exhibits a lower HRS/LRS ratio with an LRS and HRS of approximately 7 kΩ and 50 kΩ, respectively. For electrodes with widths less than 300 nm, none of the tested devices exhibits a cycling behavior. This can be explained by the trenching effect caused by the filling of the TiN inside the SiN trenches, showing that the trenching effect is the limiting factor for the scaling down, whereas, for the single and double-subtractive approaches, the scaling down is limited by the lithography resolution. The best ratio observed in this study was for the single-subtractive approach with an electrode width of 300 nm and a ratio of 35. The second-best ratio is with the same approach with an electrode width of 200 nm with a ratio of 20, but also with a lower device-to-device variability.

Since the active layer is the same for all three approaches, the lower performance in the Damascene can be explained by the trenching effect. For the double-subtractive approach, the differences can arise from the process flow. In the double-subtractive process, Al/TiN BEs and $Al_2O_3/TiO_{2-x}/Ti/TiN$ memory stacks were subsequently deposited without any other fabrication steps. In the single-subtractive process, before the deposition of the memory stack, the SiO_2 surface was polished by CMP and stopped on the TiN layer. The overpolishing of the CMP on the TiN reduces its roughness. This could indicate that the improvement of the roughness of the interface between the BE and memory stack improves the performance of the device.

3.4 Conclusion and Perspectives

This study demonstrates that Damascene and subtractive approaches are viable for integrating resistive memory. The low planarity below 10 nm, obtained after the CMP process, shows that the devices can be integrated inside the BEOL level of interconnects. Moreover, it enhances the stacking of the devices on top of each other, enabling 3D crossbar arrays.

The electrical measurements obtained are consistent with the current state-of-the-art performance for TiO_x-based resistive memory [40]. The Damascene approach suffers from worse electrical performance in terms of the SET/RESET voltage and the HRS/LRS ratio. The trenching effect and the formation of void cavities degrade the interface between the bottom electrode and the memory, limiting the scalability of the technology, as shown by the electrical measurements. In contrast, the subtractive approaches have been shown

to have certain advantages since they allow a higher aspect ratio, critical for reducing access resistance and increasing the number of memories in a single passive crossbar [16]. In addition, the subtractive approaches allow better metal stack optimization. Unlike the Damascene process, where the polishing of the metal limits the choices of the material for the BE, the subtractive etch process is compatible with the use of various materials. For example, the use of an Al layer beneath the TiN in the subtractive approaches decreases the access resistance and improves the conductance. These approaches present complexities in controlling the EPD. However, there is potential for further optimization in this regard, indicating that these challenges can be addressed in future iterations of the process. The single-subtractive approach achieves the best performance : for an electrode width of 300 nm, we obtain an HRS/LRS around 250/7 kΩ.

Although the double-subtractive approach demonstrated worse performance in this study, its process flow allowed the deposition of MIM memory without breaking the vacuum. This could lead to better interfaces between the memory stack and electrodes. Moreover, this approach has shown promising results at cryogenic temperatures for quantum applications [166]. Finally, drawing inspiration from the TopVia subtractive approach [153], the electrode line could be defined with a first lithography mask, followed by a second mask to define pillar-shaped memory. The same CMP process used in both subtractive approaches can be used, as shown in this study [43]. This allows excellent control of the sizes of the memory and electrodes, although with an additional lithography mask compared to the double-subtractive approach.

3.5 Supplementary Data

SiO_2 slurry mechanism

Typically, slurries based on colloidal silica are used to polish both SiO_2 and SiN but they can also be used for polishing TiN with the addition of an oxidizing agent such as H_2O_2 or NaClO. The polishing of SiO_2 was achieved through a hydration/abrasion cycle, which involves the following reaction [158] :

$$(SiO_2)_x + 2\,H_2O \longrightarrow (SiO_2)_{x-1} + Si(OH)_4 \tag{3.2}$$

The detachment of the $Si(OH)_4$ molecules was facilitated by mechanical contact between the particles and the sample. These slurries are generally basic, so as to promote hydration reactions.

For polishing TiN, the surface first undergoes oxidation via the following reaction [159] :

$$TiN + H_2O_2 \longrightarrow TiO_x + H_2O + N_xO \tag{3.3}$$

Mechanical contact between the titanium oxide and colloidal silica particles enables the removal of the oxide layer, and the process is repeated until the desired thickness is achieved. It is important to note that no chemical reaction occurrs between the abrasive particles and titanium oxide.

Finally, the CMP of SiN involves a complex surface reaction, which can be summarized as follows :

$$Si_3N_4 + H_2O \longrightarrow Si_2N_2O + NH_3 \tag{3.4}$$
$$Si_2N_2O + H_2O \longrightarrow SiN_2O_2 + NH_3 \tag{3.5}$$
$$SiN_2O_2 + H_2O \longrightarrow SiO_2 + 2\,NH_3 \tag{3.6}$$
$$SiN_2O_2 + SiO_2 \longrightarrow 2\,Si_2O_3 + 2\,NH_3 \tag{3.7}$$
$$Si_2N_2O + H_2O \longrightarrow 2\,Si_2O_3 + NH_3 \tag{3.8}$$
$$Si_2O_3 + H_2O \longrightarrow Si(OH)_4 \tag{3.9}$$

The reaction rate typically determines the polishing rate [160].

CeO_2 slurry mechanism

Slurries based on cerium dioxide are typically used to polish oxides on nitride substrates (as stop barriers). The particles are softer than silicon dioxide slurries.

Abrasion occurs in two steps [161]. First, the cerium dioxide molecule attaches to the silicon oxide surface, and then, the mechanical contact between the pad and the sample removes it along with some of the oxide. The reaction between the particles and the oxide is as follows :

$$-\mathrm{Ce}-\mathrm{OH} + -\mathrm{SiO}- \longrightarrow -\mathrm{Si}-\mathrm{O}-\mathrm{Ce} + \mathrm{OH}^- \tag{3.10}$$

Several links of this type form a single abrasive particle and a surface. No literature has been found regarding the abrasion mode of CeO_2 on TiN. Nevertheless preliminary tests have shown the potential for TiN polishing and will be studied in the next section.

Cerium-dioxide-based slurries typically contain acrylic copolymers [160]. Acrylic copolymers are chains of monomers that include acrylic groups, whose primary function is to attach to surfaces. Therefore, they act as surfactants, that is, they prevent particle aggregation. Moreover, they attach to silicon nitride and reduce polishing by inhibiting the reaction described above. Thus, their presence increases the oxide–nitride selectivity. The literature also reports an inhibiting effect of H_2O_2 on the polishing of SiN using CeO_2 particles [162].

Material Removal Rate

Table 3.1 lists the measured material removal rates (MRR) of TiN and SiN for different slurry solutions. Distinct effects on the material polishing rates were observed. When using the Allied slurry, the addition of H_2O_2 led to the expected increase in the TiN polishing rate. Conversely, the SiN polishing rate decreased, suggesting an inhibitory effect of H_2O_2 on SiN, which is consistent with observations in CeO_2 studies [162].

Upon diluting the slurry with water, both TiN and SiN experienced a reduction in polishing rates ; however, this effect was more pronounced for SiN. This disparity in rate reduction suggests a predominantly mechanical polishing mechanism for SiN, attributable to a lower particle concentration, as opposed to the oxidation-driven mechanism of TiN.

The results indicated that the CeO_2 slurry resulted in better selectivity than the silicate-based slurry when H_2O_2 was added. The inhibitory effect of H_2O_2 on SiN reported in the literature [162] was not observed in this study. This could be due to the surfactant present in the solution, inducing saturation in the inhibitory sites, leaving the SiN polishing rate constant. However, the TiN polishing rate increased in the presence of H_2O_2, reaching a plateau at a 1% concentration, indicating a potential maximum efficacy threshold for oxidation similar to that observed with SiO_2 particles. The result shows that an increase in pressure decreases the selectivity.

Slurry	Allied 50 nm (SiO_2)				Ultra-Sol STI (CeO_2)				
			H_2O						
			1 :1	1 :2					
Percent H_2O_2 (vol.)	-	2.5	1	1	-	-	0.2	1	2.5
Pressure (gr)	350	350	350	350	900	350	350	350	350
SiN rate (nm/min)	530	360	111	21	33	10	10	9	13
TiN rate (nm/min)	290	940	614	528	70	56	306	480	490
Selectivity : TiN/SiN	0.54	2.6	5.5	25.1	2.12	5.6	30.6	53.3	37.7

TABLEAU 3.1 Material removal rates (MRR) of TiN and SiN for silicon- and cerium-based slurries with various dilution levels and H_2O_2 concentrations.

Damascene development

As shown in Table I, the highest selectivities are achieved with the STI slurry (CeO_2) with H_2O_2 and Allied (SiO_2) with H_2O_2 and diluted on the wafer. We tested a damascene process on structured samples for CeO_2 slurry with 1% H_2O_2 and 50-nm SiO slurry with 1% H_2O_2 and 1 :2 water dilution.

We observed significant variability in the dishing for silica-based slurries at the end of the process from one sample to another due to the lower selectivity compared to CeO_2 slurry. Nevertheless, CeO_2 slurry is ineffective in polishing topographic structures [163]. CeO_2 particles tend to agglomerate at the bottom of trenches [164], preventing mechanical contact with the pad required to remove the metal. In our tests, we observed that planarization was ineffective when the topography of the trenches was greater than 50 nm.

To avoid this effect, the silica-based slurry can be used to planarize the surface and the CeO_2-based slurry to remove overburdened TiN and stop the process, as shown in Fig. 3.6(a). For better results, the maximum size of the structure (device and filling dummies structures) can be reduced below 500 nm to get a smoother surface and use CeO_2

slurry only. As shown in Fig. 3.6(b–c), this approach provided better planarity but with the drawback of imposing a constraining design rule.

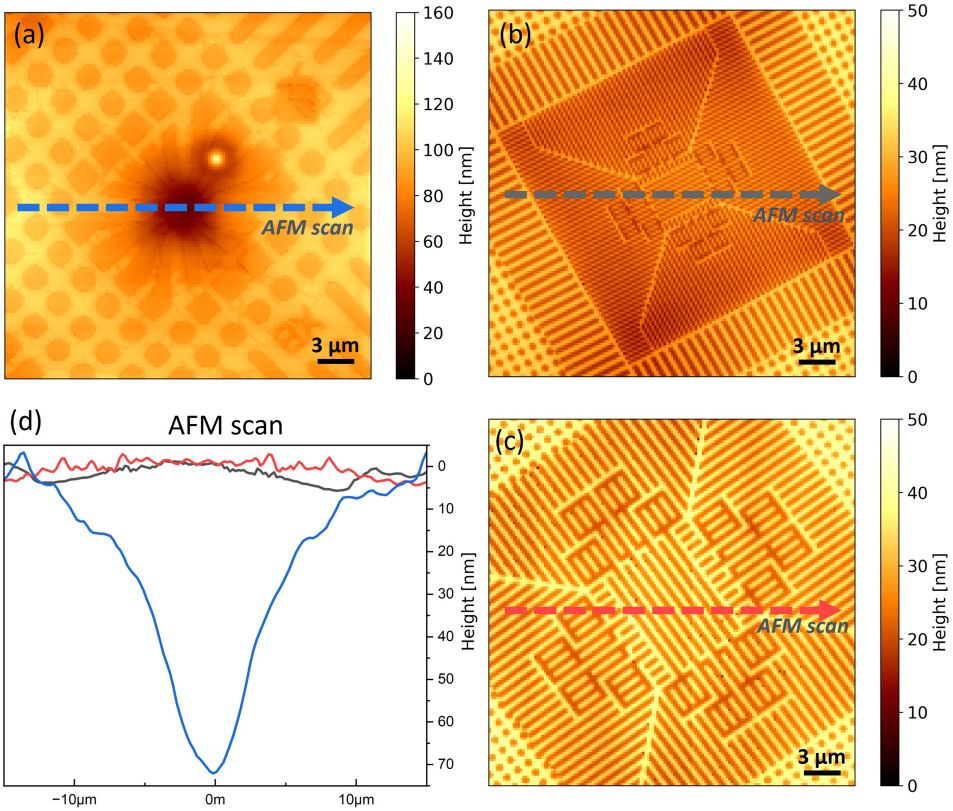


FIGURE 3.6 (a) AFM measurement showing a topography with over 60 nm of erosion after using a SiO_2 slurry on layout structures with micro-sized filling structures. Optimization of filling structures at the nanoscale for devices with electrode widths of 200 nm (b) and 400 nm (c), enabling the use of CeO_2 slurry, resulted in enhanced surface planarity.

Single and double subtractive development

For subtractive approaches, CMP is used to polish and planarize the SiO_2 layer as in IDL processes. To remove the overburden SiO_2 above the electrode, plasma-etching can be used as is done for most advanced approaches to the fabrication of TiO_x-based resistive memory passive crossbar array [40] or using CMP itself as in the top via subtractive interconnect process [153]. We tested both approaches using plasma-etching with SF_6/CF_4 chemistry and CMP with SiO_2-based particle (50 nm) slurry without any additive. As shown in Fig. 3.7, using CMP improves drastically the dishing and uniformity over the wafer compared to plasma etching.

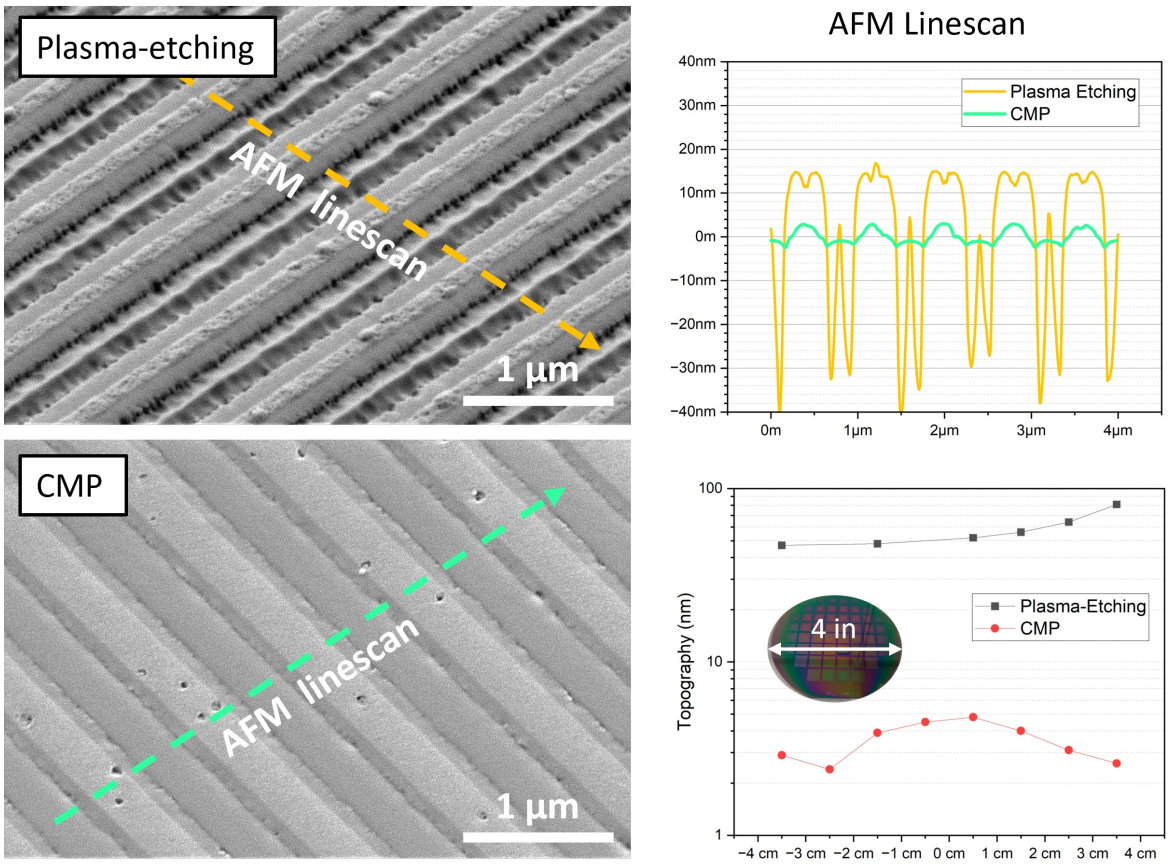


FIGURE 3.7 Comparative topography of device revelation after planarization : CMP yields a smoother topography and enhanced uniformity across a 4-inch wafer compared to plasma etching in the subtractive approaches

Discussions supplémentaires à l'article

Cette étude présente le développement de trois procédés de fabrication pour des circuits crossbar passifs. Une planéité inférieure à 10 nm a été atteinte après l'étape de CMP pour chaque procédé, respectant ainsi les exigences de planéité nécessaires à l'empilement des niveaux de BEOL.

Les mesures électriques obtenues sont conformes aux performances à l'état de l'art pour les mémoires résistives à base de TiO_x utilisant des procédés de fabrication similaires [40]. Les mesures de HRS et LRS présentent des performances comparables, bien que l'approche soustractive simple (*Single subtractive*) montre des performances légèrement supérieures en termes de rapport de résistance HRS/LRS.

Bien que toutes les approches démontrent un potentiel de transférabilité dans le BEOL, cette étude souligne les limitations de l'approche Damascène. Les tensions de SET/RESET sont plus élevées dans l'approche Damascène en raison de la résistance d'accès accrue due

à l'électrode en TiN. De plus, comme la CMP est effectuée sur le métal, un seul matériau peut être utilisé pour les électrodes métalliques, ce qui entraîne une résistance d'électrode élevée. Les cavités créées dans les électrodes métalliques dues aux problèmes de remplissage (*trenching*) sont également limitantes pour la réduction des dimensions. Comme expliqué dans la section 2.2.2, les résistances d'accès sont un paramètre limitant pour la mise à l'échelle des crossbars. L'utilisation d'interconnexions en Cu est envisageable, mais elle complique le procédé de fabrication, comme il sera étudié dans le chapitre 6.

Dans l'approche soustractive, un empilement de matériaux Al/TiN permet de réduire la résistance d'accès tout en conservant la couche de TiN nécessaire pour éviter la diffusion des ions d'oxygène vers les électrodes. Cette méthode permet également un procédé d'intégration plus avancé, où la couche de commutation est gravée une première fois avec la BE puis une deuxième fois avec la TE, sans rajouter de niveaux de masque, comme dans l'approche soustractive double.

L'approche soustractive permet non seulement une optimisation en sélectionnant des empilements de matériaux spécifiques mais ouvre également la voie à la fabrication de structures plus complexes. Au lieu d'utiliser uniquement des électrodes planes avec une couche de commutation entre chaque intersection, des structures de formes plus complexes peuvent être envisagées. La fabrication de ce type de structures fera l'objet du chapitre suivant.

CHAPITRE 4

Lithographie en niveaux de gris pour gravure multi-matériaux

Avant-propos de l'article

Titre Complet : Multiple material stack grayscale patterning using electron-beam lithography and a single plasma etching step.

Auteurs et affiliations : R. Dawant[1,2], S. Ecoffey[1,2] et D. Drouin[1,2]

[1]Institut Interdisciplinaire d'Innovation Technologique (3IT), Université de Sherbrooke, Sherbrooke, Québec J1K 0A5, Canada
[2]Laboratoire Nanotechnologies Nanosystèmes (LN2) – CNRS UMI-3463 – 3IT, Sherbrooke, Québec J1K 0A5, Canada

Date de publication : Novembre 2022
Journal : Journal of Vacuum Science & Technology B
Volume : 40, no. 6, pp. 062603 (2022)
Référence : 10.1038/s41467-021-25455-0 [43]

Contribution du document :

Le chapitre précédent a montré l'avantage d'une approche soustractive qui permet une optimisation des résistances d'accès grâce à une combinaison de matériaux pour augmenter la conductance des électrodes (approche soustractive simple) et une meilleure isolation de la couche mémoire (approche soustractive double). Ces deux procédés ont été rendus possibles par la capacité d'effectuer la CMP sur le diélectrique et d'encastrer des électrodes composées d'empilements de plusieurs matériaux conducteurs. L'approche soustractive ouvre également la voie à l'encapsulation de structures morphologiquement plus complexes. La figure 4.1 montre des exemples de structures avec une morphologie plus avancée à partir du même empilement de matériaux utilisé pour les approches soustractive simple (a) et double (b).

Comme illustré sur la figure 4.1(a), avec l'empilement de l'approche soustractive simple, on peut obtenir des points de contact aux intersections des électrodes où la commutation de la SL (*switching layer*) s'opère, formant une pointe dans la couche de TiN. Ainsi, le contact entre l'électrode et la couche de commutation se fait seulement sur la très petite surface de la pointe. Cette approche vise à améliorer la stabilité du mécanisme de commutation en permettant un meilleur contrôle de la localisation du filament. De plus, cette pointe crée un confinement et une augmentation du champ électrique localement, ce qui pourrait faciliter la création et la commutation du filament en réduisant la tension nécessaire à la commutation. Il a été montré que l'ajout de protubérances inférieures à 40 nm dans la zone de commutation améliorait grandement les performances des mémoires ReRAM à base de TiO_x [167].

La figure 4.1(b) montre comment, à partir de l'empilement utilisé dans l'approche soustractive double, on peut créer une structure composée d'une électrode métallique plane avec des mémoires ReRAM fabriquées sous forme de pilier à chaque intersection du crossbar. En plus de limiter potentiellement les courants de fuite entre les différentes mémoires, comme dans l'approche soustractive double, ce qui permet de réduire encore plus les dimensions, ce qui a pour effet de diminuer la variabilité, comme observé dans les mesures électriques du dernier chapitre (voir figure 3.5). Cela permet de réduire la taille des mémoires pour améliorer leur stabilité tout en conservant des électrodes de dimensions plus élevées, ce qui réduit les résistances d'accès.

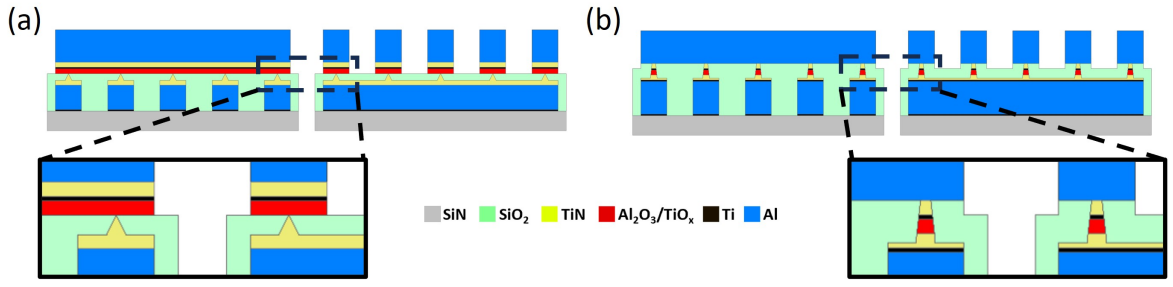


FIGURE 4.1 Optimisation structurelle des crossbars à partir des empilements de matériaux utilisés dans (a) l'approche soustractive simple pour créer un contact entre la BE et la mémoire extrêmement localisé et (b) dans l'approche soustractive double pour localiser la mémoire sous forme de pilier.

Dans ces deux exemples, une structure composée d'un empilement de plusieurs matériaux est gravée selon une morphologie spécifique. Ensuite, le reste du procédé de fabrication suit les étapes précédemment proposées : un diélectrique est déposé puis planarisé par CMP jusqu'à révéler certaines parties de la structure, avant que la TE, avec ou sans la SL,

ne soit gravée. La différence réside donc dans la manière de fabriquer la structure de la BE, avec ou sans la SL. Afin d'étudier l'effet de ces différentes morphologies, une méthode de fabrication a été développée pour réaliser de telles structures. La méthode proposée permet de contrôler la création de structures 2.5D à l'aide d'une technique de lithographie en niveaux de gris et suivi de la gravure d'un empilement de plusieurs matériaux en une seule étape.

Résumé en français

Dans cet article, nous présentons une méthode innovante pour réaliser la lithographie par faisceau d'électrons en niveaux de gris sur des empilements multicouches, où le transfert de motifs est effectué en une seule étape de gravure plasma. En raison des différences de taux de gravure des matériaux dans l'empilement, la forme de la résine après développement diffère significativement de celle de l'empilement multicouche après gravure. Pour obtenir la forme désirée dans l'empilement multicouche, la dose finale de résine est définie par une courbe de calibration de gravure qui établit la relation entre la dose du faisceau d'électrons et l'épaisseur restante des matériaux après la gravure plasma. Grâce à cette méthode, un crossbar de mémoire résistive est fabriqué, atteignant une résolution de gradient de hauteur de 10 nm et des dimensions de dispositifs à l'échelle nanométrique.

Multiple material stack grayscale patterning using electron-beam lithography and a single plasma etching step

Abstract : In this paper, we present a novel method to perform grayscale electron-beam lithography on multi-layer stacks where the pattern transfer is done in a single plasma etching step. Due to the differences of material etch rates in the stack, the shape of the resist after development versus the shape of the multi-layer stack after etching is significantly different. To be able to reach the desired shape in the multi-layer stack the final resist dose is defined by an etching calibration curve that describes the relation between the electron-beam dose and the remaining materials thickness after plasma etching. With this method, a resistive memory crossbar array is fabricated with height resolution of 10 nm and nanoscale dimensions devices.

4.1 Introduction

In this last decade, grayscale lithography has been studied [168, 169, 170] to pattern 3D structures for micro-optical [170], MEMS [171], and other applications. This technique can generate arbitrary shapes with height gradients in the resist profile. The resist profiles are often required to be transferred into an underlying material and it is usually done by deep reactive ion etching (DRIE). For greyscale lithography, material-to-resist etches selectivity, the in-process etch rate variations and the etch isotropy required careful calibration to get the proper profile transferred material [172]. These aspects have been widely studied to transfer the 3D resist profile using photolithography [173, 174, 175] as well as electron beam lithography (EBL) [176, 177]. The transfer is usually performed in a single bulk material meaning that only one material-to-resist etching selectivity is to be considered. The DRIE recipe is rather tuned to reach a 1 :1 selectivity to obtain the same profile in the resist as in the material after etching [178, 174] or applied a height shrink factor in the resist profile to obtain the desired profile after pattern transferring [173, 177]. In a case where the underlying materials is made of multiple material with different etch rates and selectivity, it will be extremely difficult to calibrate the required resist profile with this approach. Other approaches have been developed to create 3D structure by DRIE in a multiple material stack for nanoscale devices like hybrid photolithography and EBL for carbon nanotubes field effect transistors [179] or nanoimprint lithography to create thin-film transistors [180, 181] but different etching steps was used with selective etching rate to transfer the 3D resist profile in the different material etch. In this paper, we present an empirical method to calibrate the etching transfer after grayscale EBL in a multi-layer stack using a single plasma etching step to fabricate 3D structure for nanoscale devices.

4.1.1 Nanoscale devices fabrication

The proposed method is used to fabricate an array of TiOx resistive memory cylinder junctions on top of an Al bottom electrode for the fabrication of resistive memory crossbars. Resistive random-access memories (ReRAMs) can be integrated into crossbar array architecture to process vector-by-matrix multiplication (VMM) in the analog domain. These passive crossbars are made of two sets of electrode arrays, bottom electrodes (BEs) and top electrodes (TEs), perpendicular to each other with a resistive memory at the intersection of each BE and TE as shown in Fig.4.2. Passive crossbar circuit fabrication process reported in literature pattern the active memory stack in the same step than the TEs fabrication [182, 156, 183, 78, 184] as illustrated in Fig.4.2(a). Meaning that the memory stack is standing below the TE even when it is not needed. In Fig.4.2(b), we propose an alternative architecture where the memory stack is patterned as a cylinder shape at the intersection of the BEs and TEs. Studies show that decreasing the memory area can increase the resistance ratio and decrease the power consumption [185, 186].

The process flow is illustrated in Fig.4.3. The BE and the memory pillars are etched using grayscale lithography and a single plasma etching step. A dielectric layer is then deposited and planarized by chemical-mechanical polishing (CMP) to reveal the memory pillars. Finally, the TE is patterned by plasma etching.

This study focuses on the fabrication of the BE and the memory pillars shown in Fig.4.3. We will show that it can be made in a single etching step. It has the advantage to decrease the numbers of fabrication steps, moreover the memory pillar and the BE will be self-aligned. The stack used is composed of 7 different layers made of 4 different materials with thickness between 200 nm and 10 nm but the proposed method can be applied on any multi-material stack. The height resolution achieved in the transferred profile is in the order of the tenth nanometers.

4.1.2 Etching transfer using grayscale lithography

Grayscale lithography is used to create 3D shapes with height gradients, enabling the fabrication of textured surfaces with nanotopographies [168]. As shown in Fig.4.4(a), the resist can be exposed at different doses resulting in a specific resist height after development. The relation between the dose and the remaining height is called the contrast curve (CC), illustrated in Fig.4.4(d) for a negative resist. If the targeted shape in the resist is a grayscale image noted $T(x, y)$, where T is the target height depending on the lateral coordinate x, y, the applied ebeam doses D is :

$$D = CC^{-1}(T(x, y)) \quad [\mu C/cm^2] \tag{4.1}$$

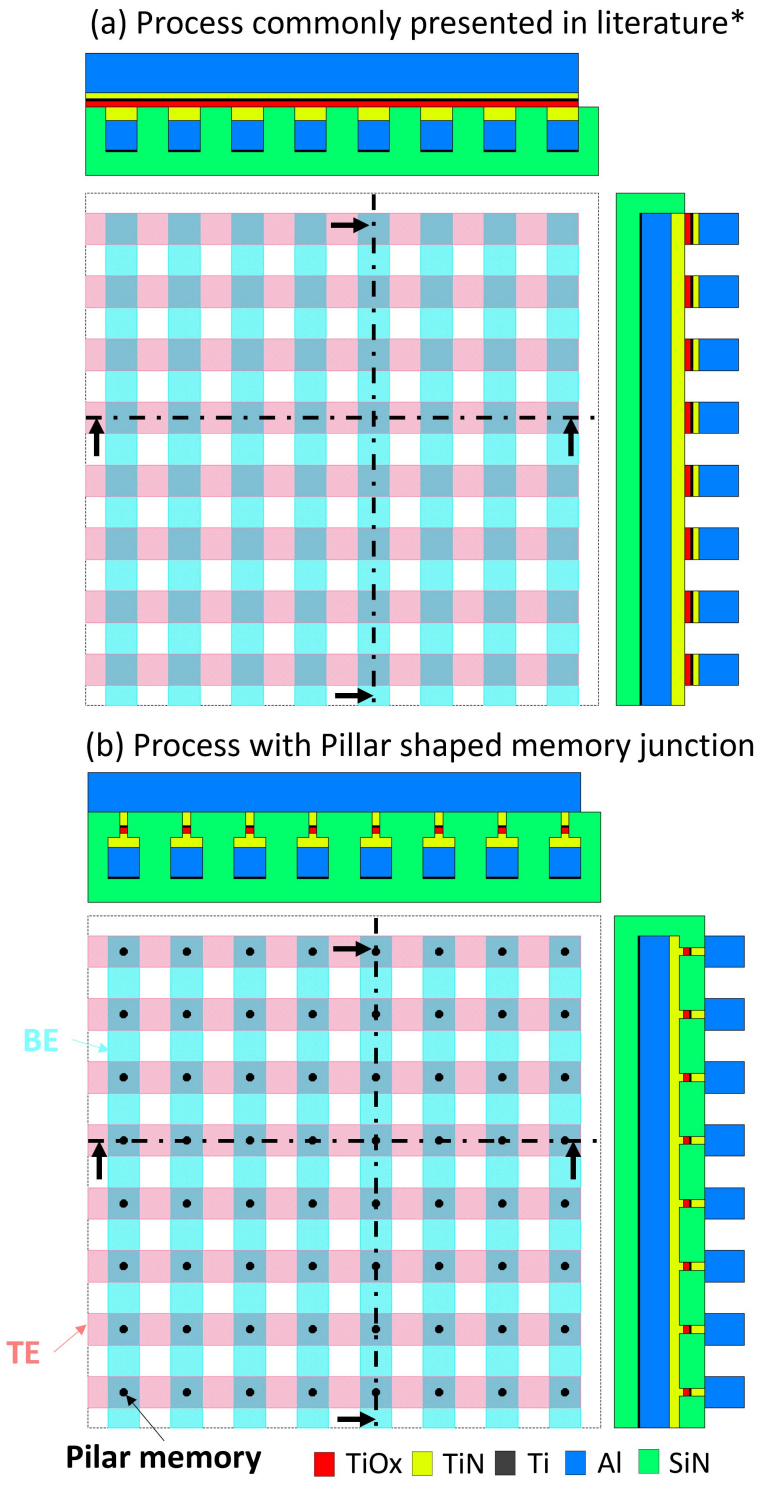


FIGURE 4.2 Comparison between the approach commonly used in literature where the active layer (in red) is patterned all along the TE as described in Ref [182, 156, 156, 183, 78, 184](a) and the proposed approach where the active layer is patterned in a pillar shape (b)

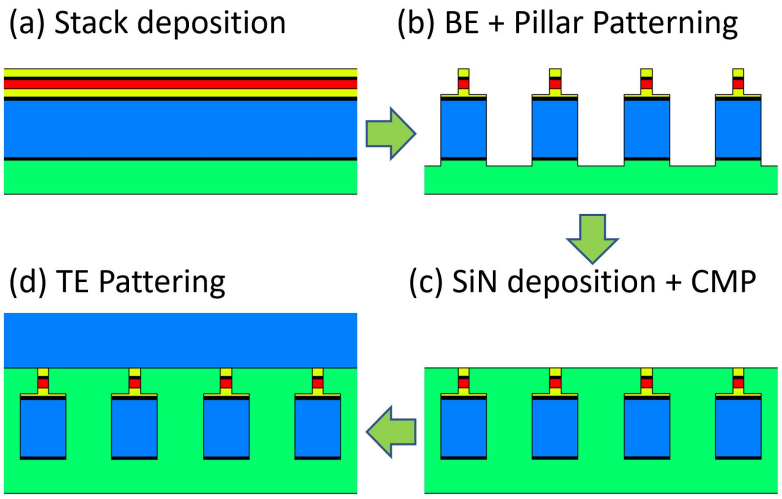


FIGURE 4.3 Process Flow : (a) The Al (200 nm) for the BE is deposited followed by the different layers used for the memory : Ti/TiN/TiOx/Ti/TiN (10/30/30/10/30 nm). (b) BE and memory pillars are patterned by grayscale EBL. (c) A dielectric layer is deposited and planarized by CMP. (d) TE is patterned by EBL and plasma etching.

The resist structure needs to be transferred into the underlying materials for the targeted application by DRIE. The etching recipe can be adjusted to get a selectivity between the resist and the material equal to 1 :1. In this case oxygen is added to the plasma as it is known to increase the resist etch rate until obtaining the same etch rate in the material [176, 178, 174] resulting in same resist and material profile. If the selectivity is not to 1 :1 the topography will be affected and the height shrink factor is applied to get the desired shape after transfer [173, 177]. As an example, for a pyramid shape with a specific resist slope, the slope transfer by plasma etching in a bulk material will have an angle increase or decrease depending on the selectivity. To create the desired topography, the effect of the angle needs to be considered. If S is the selectivity between the resist and the material, the applied dose for the target shape will be :

$$D = S \times CC^{-1}(T(x,y)) \quad [\mu C/cm^2]. \tag{4.2}$$

When the transfer is done on multiple materials stacked on each other with different etch rates, the resulting shape can become complex and difficult to control as shown in Fig.4.4(b,c). For a specific etching recipe, the final shape will depend on the etching rate of the resist and all the different materials etched but also on the thickness of each layer and the etching time. The model to determine the applied dose presented above could be

adapted if all the parameters are known individually but will be strongly dependent on their accuracy. Moreover, chemical diffusion at the interface of each material can increase the complexity of the etching rate evolution over the multi-layer stack. The applied dose required to get the targeted shape after the transfer can be challenging to obtain. A complex sequence of different etching steps with specific selectivity could be used as in [180, 181]. In the next section, a method to calibrate the transfer using a unique etching step for the full process will be presented.

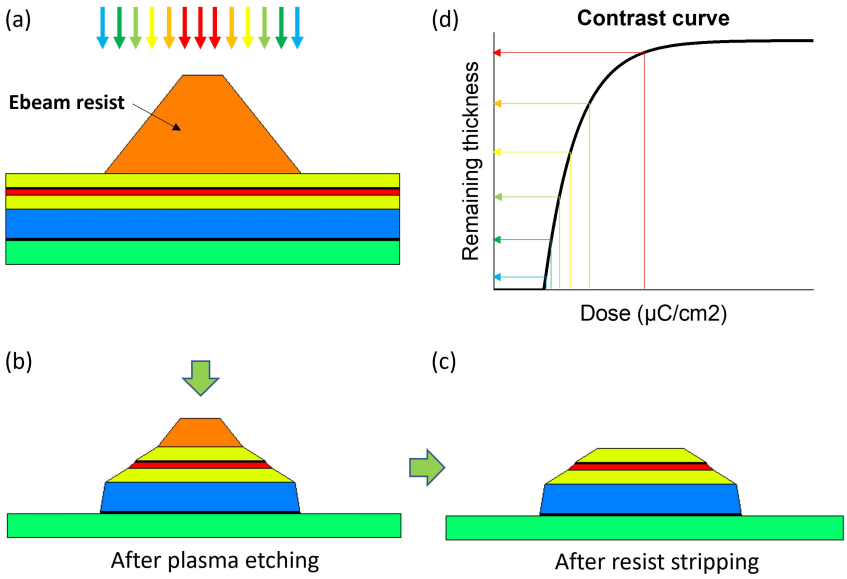


FIGURE 4.4 Grayscale lithography principle. (a) The EBL resist is exposed at different doses to create textured surfaces with height gradients. The resist structures are transferred by plasma etching (b) and the resist is stripped (c). The applied dose used for the exposition is determined with the contrast curve to get a specific shape in the resist (d), the contrast curve represents the thickness resist after development in function of the dose for a negative resist. The color gradient of the arrows in (a) and (d) illustrate the variation of the exposure dose from a lower dose (in blue) to a higher dose (in red). Since the etching rate of the resist and the material are different, the shape in the stack after the transfer is far from the initial shape in the resist.

4.2 Proposed method

We propose a method to calibrate the dose required to get a specific shape after plasma etching in a multi-layer stack without knowing precisely the thickness and the etch rates of the different materials in the stack. To measure a CC, large squares, larger than the 3β value, are patterned at different doses. The remaining thickness after the development is then measured for each square, resulting in a curve describing the relationship with the remaining height versus the applied dose (see Fig4.4(d)). The dose range can be determined

with the CC and is usually correspond to the remaining thickness between 25% and 75% of the starting resist thickness.

The test is repeated in this dose range with a higher number of points for better accuracy and a more precise CC and the resist patterned structures are etched by plasma etching. The structures are measured after resist striping and an etching calibration curve (ECC) is determined as shown in Fig.4.5. The ECC will therefore describe the relationship between the remaining height in the multi-layers stack after the plasma etching and the ebeam doses. The ECC can be used to etch a targeted shape in the stack. The applied dose will be :

$$D = ECC^{-1}(T(x,y)) \quad [\mu C/cm^2], \tag{4.3}$$

for a target structure defined by $T(x,y)$. This ECC has the advantage to be simple and fast to measure and requires only one sample will the full stack. The only information needed is an estimation of the times required to etch the full stack.

This study will use the method to fabricate restive memory on BEs with a stack made of Al, Ti, TiOx, and TiN material but the procedure is valid for any arbitrary stack as long as an etching recipe, can etch all the material, is used. For example, it could be used to fabricate thin-film transistors as published in [180, 181] but with grayscale lithography and a unique etching step. For this application, a negative resist is used for exposure time consideration but the method remains valid with a positive resist, the only requirement is a low contrast resist. In that case, the ECC will decrease with the dose rather than increase.

4.3 Experiments and results

This section presents the fabrication of BEs and memory pillars of a resistive memory crossbar array in a single step with the architecture presented Fig.4.2(b). As shown in Fig.4.3(a), the stack deposited for the BEs and pillars fabrication is the following : Ti/Al/Ti/TiN/TiOx/Ti/TiN (10/200/10/30/30/10/30 nm).

To measure the corresponding ECC of this process, large squares of 200x200 μm are exposed at 100 keV in the EBL negative resist ma-N 2405 (390 nm thick after spin coating) in a range of doses between 0 and 400 $\mu C/cm^2$ and with a higher number of doses point between 80 and 140 $\mu C/cm^2$ on the full stack. The remaining resist after the development is measured for each dose with a Dektak 150 profilometer to get the associated CC as shown Fig.4.6(a). The sample is then etched using an STS - Multiplex Inductively Coupled Plasma (ICP), with a pressure of 5 mTorr, gas flux of Cl_2 : 10 sccm, BCl_3 : 10 sccm, Ar :

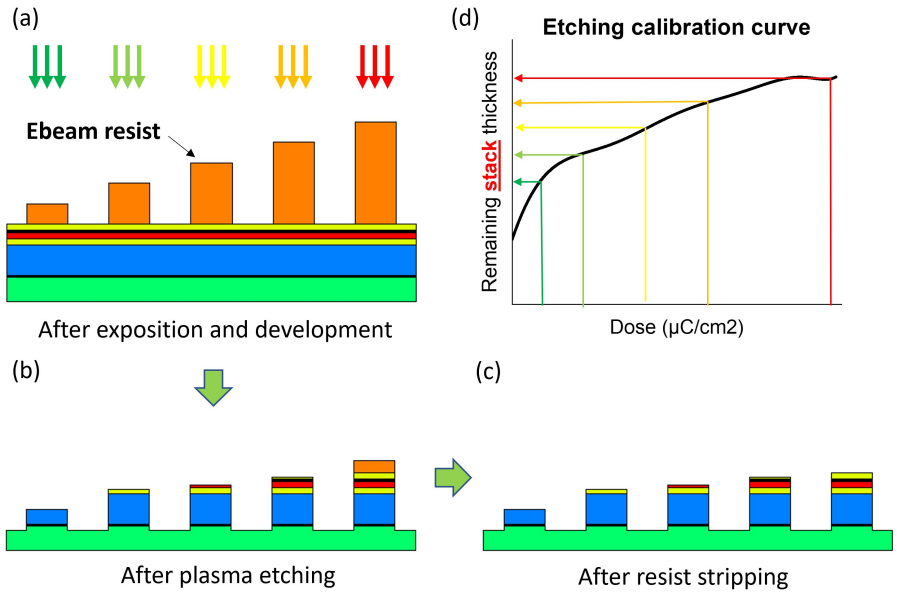


FIGURE 4.5 Etching calibration curve for grayscale lithography on multi-layer stack material. (a) Resist after development, squares are exposed at different doses. After the etching (b) and the resist striping (c), the remaining thickness for each square after the etching is measured (d). The calibration curve described the thickness remaining in the stack after the etching in function of the dose. The color gradient of the arrows in (a) and (d) illustrate the variation of the exposure dose from a lower dose (in blue) to a higher dose (in red).

10 sccm, a coil power of 500 W, a platen power of 50 W and a temperature of 20°C during 4 min 30 sec to etch the full stack, plus a slight overetching of 20nm when no resist remains after the development. After the etching of the sample test, the ECC is measured with a profilometer.

Fig.4.6(b) show the ECC for this process. Since on the dose range studied the relation between the dose and the remaining thickness of the resist is almost linear, the slope of the ECC can be considered proportional to the selectivity between the resist and the material (supposing that the etching rate of the resist is constant for every dose). It shows that the resist has a low selectivity with the TiN and TiOx layer (highlighted in yellow and red respectively) and the selectivity with the Al layer is higher (highlighted in blue). It also shows that for some layers like the top TiN layer and the aluminum layer, the etching rate varies over the junction for the same materials. It could come from diffusion between layers or oxidation after the deposition of these materials that change the etching rate behaviors.

As shown in Fig.4.6(c), the targeted structure is formed of rectangular BEs composed only of metallic layers (Ti, Al, and TiN) and circular pillars on top, composed of the full stack with the metallic and oxide (TiOx) layers. The ECC shown in Fig.4.6(b) indicates that the stack starts to be etched when the resist is exposed below 130 $\mu C/cm^2$. Since the pillar should not be etched at all, a dose above 130 $\mu C/cm^2$ is required but for process stability consideration, the resist is exposed at 400 $\mu C/cm^2$. It corresponds to the dose where the resist has no solubility with the developer to get the full resist thickness and protect the pillar from the etching during the full process. The BEs area needs to be partially exposed to stop the etching in the TiN layers between the memory (TiOx layer) and the BE (Al layers). This means that for the BE area, the remaining thickness after the etching needs to be between 210 and 250 nm. The targeted thickness is 230nm which corresponds, based on the ECC to an applied dose of 98 $\mu C/cm^2$ for this process.

The target for the BE is to stop in the layers between the Al and the TiOx. Since the process window is $\pm$ 20 nm, which corresponds to a variation of $\pm$ 6 $\mu C/cm^2$, the accuracy of the ECC measurement needs to be at least 20 nm. From the CC, the resist height after the development can be determined (here 280 nm) for the BE area. This can be useful to monitor the process quality after the development, an AFM measurement can be performed to check before the etching if the BE has the right thickness. It also gives the tolerance of thickness variation. For this process, a variation of $\pm$ 30 nm in the resist after the development can be accepted.

Fig.4.7 shows SEM images of the BE with memory pillars on top, patterned in the resist after the development (a) and after etching and resist striping (b). Fig.4.7(c) shows a FIB cross-section of the structure. The zoomed SEM images show the aluminum BE with a width of 1.5 µm and the TiO resistive memory patterned in a pillar shape with a width of 250 nm. The SEM images show that the BE was partially etched with a remaining thickness of 220 nm. Fig.4.7(c) was taken after the full process fabrication. A dielectric layer was deposited on the 3D structure and planarized by CMP. Only the memory pillars were revealed and the BE is encapsulated in the dielectric layer. The TE was then patterned.

This result shows that grayscale lithography can be used to shape nanoscale devices by plasma etching with gradient height shape in a single ebeam exposition followed by a single plasma etching step. With the proposed method, a height control of around 10 nm is achieved with a multi-stack layer composed of 7 layers with thicknesses in the tenth of nanometers.

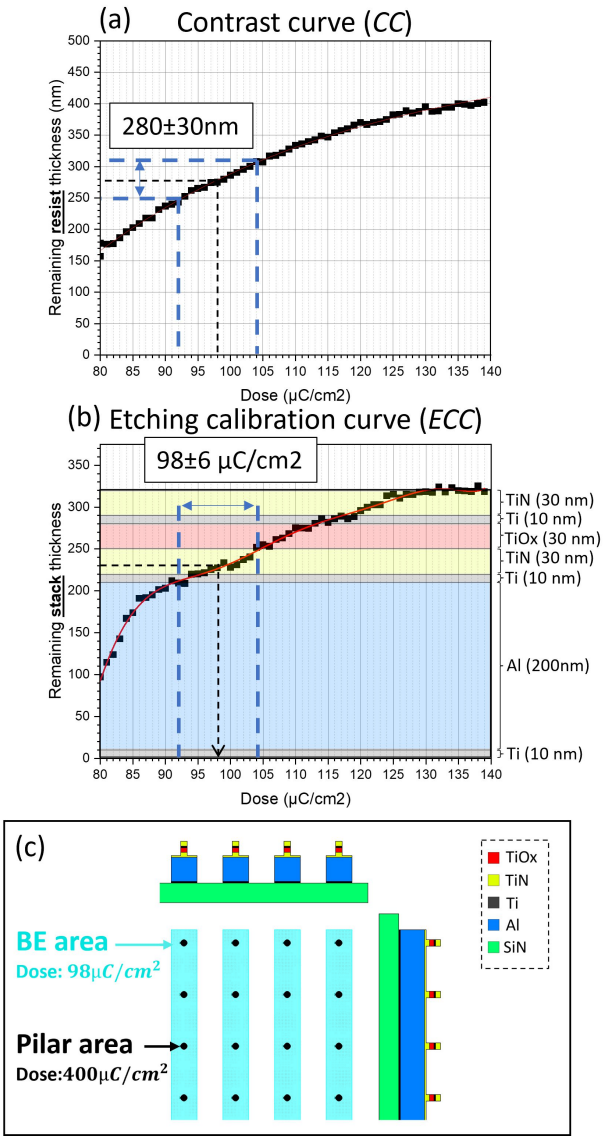


FIGURE 4.6 Contrast curve (a) and etching calibration curve (b) measured on the multi-layers stack composed of Ti/Al/Ti/TiN/TiOx/Ti/TiN (10/200/10/30/30/10/30 nm). (c) Layout and cross-section of the BEs (in blue) and the memory pillars (in black) area. Based on the ECC, the BEs area is exposed a 98 $\mu C/cm^2$ to stop the etching in the TiN layer between the TiOx and the Al, corresponding to a remaining thickness of 230 nm after etching.

4.4 Process window and limitations

The ECC depends on the resist, resist thickness, the composition of the stack, etching rates, and thickness of the different materials used. As shown in Fig.4.6, the process window is small. The ECC can drift due to shelf life, a drift of the CC, a variation in the etching plasma process, and development time variation. In that case, the applied dose will not result in the expected height after the etching. For those reasons, the calibration

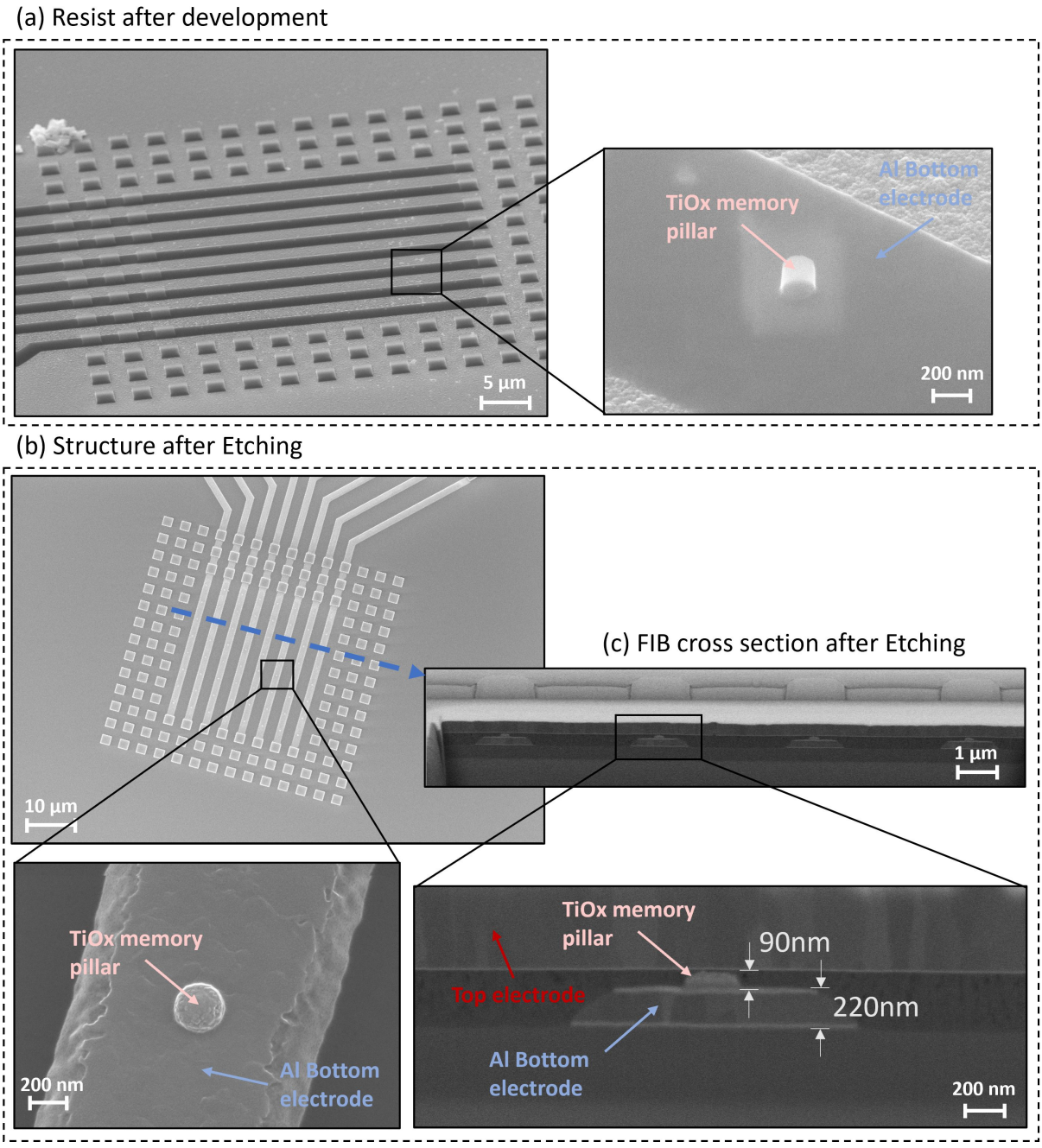


FIGURE 4.7 SEM images of the BEs with memory pillars patterned on top of it in the resist after the development (a) and the etching (b). The Fib cross-section of the electrode and the pillar (c) shows that the partial etching of the BE was stopped in the TiN layer with a thickness of 220 nm (target 230 nm) corresponding to a height resolution of 10 nm.

curve should be monitored from time to time to correct process drift and control process stability.

The stability of the process can be improved by increasing the process windows. It is well established [187] that using a low-contrast developer improves process control in grayscale lithography. Since the slope of the CC will be lower, the remaining thickness will be less sensitive to dose variation. In the same way, decreasing the etching selectivity between the resist and the material used will decrease the slope of the ECC and makes the process less sensitive to dose variation. The etching recipe can be therefore adapted to decrease the etching rate of the material to decrease selectivity.

The quality of the process will be strongly affected by the proximity effect correction used. For this process, a module of the BEAMER software from GenISys GmbH specially made for grayscale lithography is used. The module corrects the applied dose to compensate the backscatter signal. Nevertheless, the short- and mid-range effects were not corrected and can sometimes lead to failures, especially when the height gradient is steep.

Negative resists are crosslinked during exposure, the more the resist is exposed, the more it is crosslinked, and therefore the development rate decrease. A partially exposed resist will have lower mechanical stability and will be more sensitive to adhesion issues. ma-N 2400 has low adhesion and this limits the minimum dimension of the BE that is partially exposed. It was observed that below 1 µm width BE, the structures take off. A solution is to increase the development time to shift the contrast curve on the right and increase the applied dose required. Therefore, the effective dose absorbed for the same remaining thickness in the resist will be higher and get better resistance to adhesion but will increase the exposure time. An alternative solution could be to change the resist for another EBL negative resist Medusa 82 that has better adhesion than ma-N 2400 and has shown grayscale capability [188].

4.5 Conclusion and perspective

In this study, a method to calibrate the transfer by plasma-etching of grayscale ebeam lithography on multi-stack layers was presented. The method needs the same test structures used to measure contrast curves. Large squares exposed at different doses are etched and the remaining thickness is measured to get the ECC that describes the remaining materials thickness after the etching process in function of the dose. The applied dose can therefore be determined with this curve to etch the targeted profile in a full stack. It was shown that this method can be used for nanoscale devices.

With this method, a novel ReRAM crossbar architecture with metallic BEs and pillar shape oxide-based ReRAM on top was fabricated. The dimension achieved were 1.5 µm width for the BE and 250 nm diameters for the memory pillars. The partial etching of the BE could be stopped in the 30 nm TiN layer between the electrode and the memory with a height resolution of 10 nm. The proposed method can be extended to any application requiring grayscale topography on multi-layers material at the nanoscale level.

The limitation and challenges of this process was discussed. The use of ma-N 2400 negative resist limits the critical dimension of the BE. The process is mainly sensitive to the resist shelf-life and the etching recipe reproducibility. Careful monitoring of the ECC should be considered.

Discussions supplémentaires à l'article

Cette étude a démontré que la méthode proposée permet de fabriquer des structures avec des gradients de hauteur dans un empilement de plusieurs matériaux. En utilisant cette méthode, il a été possible de réaliser des structures constituées de lignes métalliques avec des piliers mémoire avec un contrôle de la hauteur d'environ 10 nm, ce qui a permis de réaliser la structure TopPilar proposée sur la figure 4.1(b). Cependant, la stabilité du procédé proposé s'est révélée insuffisante pour réaliser des électrodes avec des pointes pour effectuer le contact avec la mémoire, comme illustré sur la figure 4.1(a).

Procédé de fabrication : TopPilar

La fabrication d'électrodes avec des piliers mémoire sur le dessus ayant été réalisée, la fabrication du crossbar complet a pu être démontrée. Comme illustré sur la figure 4.8, après la fabrication des lignes BE avec le pilier mémoire (a), du SiO_2 a été déposé puis planarisé par CMP, en utilisant le même procédé que pour les approches soustractive simple et double développées au chapitre précédent (voir section 3.2.2.2). Après la planarisation de la surface, seuls les sommets des piliers sont révélés pour être en contact avec la TE, tandis que la BE est encastrée dans l'oxyde (b). La figure 4.8(c) montre une coupe FIB du crossbar après gravure de la TE. On observe la structure BE avec les piliers mémoire encastrés dans le SiO_2. Aucune topographie n'est observée sur la TE, ce qui indique une bonne planéité de la surface après l'étape de CMP.

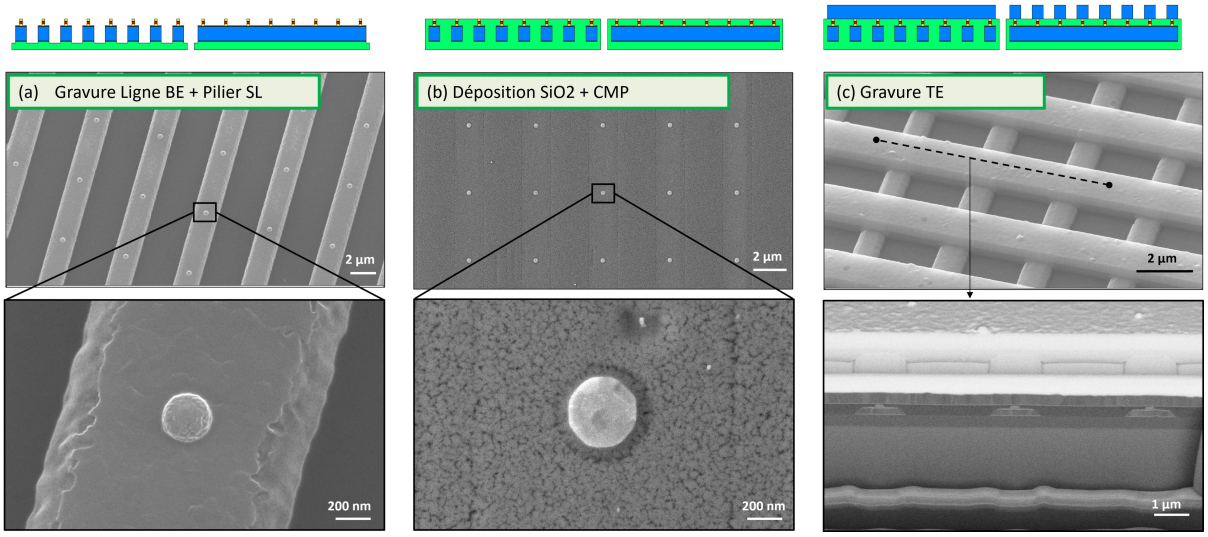


FIGURE 4.8 Procédé de fabrication : TopPilar. (a) Gravure de la BE avec les piliers mémoire. (b) Dépôt d'oxyde et planarisation par CMP jusqu'à révéler les piliers. (c) Gravure de la TE.

Commentaire sur le transfert industriel

La technique de lithographie en niveaux de gris utilisée dans cette recherche n'est pas directement transférable dans un procédé industriel. Ce type de lithographie se réalise par écriture directe, soit par lithographie à faisceaux d'électrons, soit par lithographie laser, qui ne sont pas couramment utilisées dans l'industrie en raison de leur faible rendement de production. Il est important de noter que les techniques de lithographie *nanoimprint* en niveaux de gris permettent un meilleur rendement de production, et que l'utilisation de transistors à base de film mince a déjà été rapportée avec cette méthode de fabrication [180].

La lithographie en niveaux de gris a été employée ici dans le but d'explorer différentes morphologies de structures pour la fabrication de crossbar. Pour une application industrielle des structures TopPilar présentées à la figure 4.8, l'usage de deux niveaux de masques, un pour l'électrode et un pour le pilier, similaire au procédé de fabrication TopVia soustractive introduit dans la section 2.1.4, semble plus approprié. Une approche similaire pourrait être envisagée pour le procédé proposé à la figure 4.1(a) en ajustant la recette de gravure DRIE pour créer une pente et former un cône dans le TiN.

PARTIE II

Transfert CMOS

CHAPITRE 5

Stratégie d'alignement sur puces CMOS

Avant-propos de l'article

Titre Complet : Hybrid cross correlation and line-scan alignment strategy for CMOS chips electron-beam lithography processing

Auteurs et affiliations : R. Dawant[1,2], R. Seils[3], S. Ecoffey[1,2], R. Schmid[3], D. Drouin[1,2],

[1]Institut Interdisciplinaire d'Innovation Technologique (3IT), Université de Sherbrooke, Sherbrooke, Québec J1K 0A5, Canada
[2]Laboratoire Nanotechnologies Nanosystèmes (LN2) – CNRS UMI-3463 – 3IT, Sherbrooke, Québec J1K 0A5, Canada
[3]Raith America, Inc., International Applications Center, 300 Jordan Road, Troy, New York 12180

Date de publication : Décembre 2021
Journal : Journal of Vacuum Science & Technology B
Volume : 40, no. 1, pp. 012601 (2021)
Référence : 10.1116/6.0001278 [44]

Contribution du document :

L'objectif de cette seconde partie est de transposer les structures élaborées dans la première section de ce manuscrit, après le dernier niveau d'interconnexion BEOL de puces CMOS conçues par une équipe de l'Université de Toronto et fabriquées par TSMC. Le premier défi réside dans la capacité de s'aligner sur ces puces CMOS en utilisant l'équipement de lithographie à faisceaux d'électrons du 3IT. Pour atteindre un alignement adéquat nécessaire à la réalisation des structures développées précédemment, il a été indispensable d'explorer une nouvelle technique d'alignement. Cette nécessité découle des contraintes spécifiques imposées par les règles de dessin de la technologie 130 nm de TSMC utilisée dans ce projet. Les marques d'alignement sont intégrées au dernier niveau de métal BEOL (M8) et doivent donc respecter des règles de densité de motifs qui empêchent l'utilisation de carrés isolés, généralement utilisés pour un alignement standard en utilisant un détecteur

d'électrons rétrodiffusés.. Cette étude a été réalisée en collaboration avec la compagnie Raith, le fabricant de l'équipement de lithographie par faisceau d'électrons utilisé.

Résumé en français

Dans cet article, nous présentons une stratégie d'alignement basée sur une approche hybride qui combine la corrélation croisée et l'alignement par balayage linéaire pour relever le défi du post-traitement de circuits intégrés CMOS utilisant la lithographie par faisceau d'électrons. En raison des règles de conception imposées par les fonderies pour le nœud technologique de 130 nm et en dessous, l'alignement classique par balayage linéaire n'est pas possible, et les formes des marqueurs sont limitées. La forme du marqueur est cruciale pour l'alignement par corrélation croisée. En mesurant avec précision le décalage d'alignement entre deux étapes de lithographie utilisant différentes formes de marqueurs compatibles avec les règles de conception, nous avons évalué l'influence de la forme des marqueurs sur les performances de l'alignement par corrélation croisée. Nous introduisons une méthode basée sur un tableau généré par bruit blanc pour concevoir des marqueurs haute performance pour la corrélation croisée, compatibles avec la technologie CMOS, en augmentant la netteté de leur pic d'autocorrélation. Nous démontrons que les performances d'alignement peuvent être améliorées en utilisant une stratégie hybride combinant corrélation croisée et alignement par balayage linéaire, atteignant un décalage moyen de 5,2 nm sur substrat CMOS.

Hybrid cross correlation and line-scan alignment strategy for CMOS chips electron-beam lithography processing

Abstract : In this paper, we show an alignment strategy based on a hybrid strategy using cross-correlation and line-scan alignment to address the challenge for CMOS integrated circuit post-processing using electron-beam lithography. Due to design rules imposed by the foundries at the 130nm node and below, classical line-scan alignment is not possible, and markers shapes are limited. The shape of the marker is essential for cross-correlation alignment. By measuring accurately the alignment offset between two lithography steps with different marker shapes compatible with the design rules, we tested the influence of the markers shape in the performances of the cross-correlation alignment. We present a method based on a white noise generated array to design high-performance markers for cross-correlation, compatible with CMOS technology, by increasing the sharpness of their autocorrelation peak. We show that the alignment performances can even be improved, using a hybrid strategy with cross-correlation and line-scan alignment and reach a mean offset of 5.2 nm on CMOS substrate.

5.1 Introduction

Back-end-of-line (BEOL) integration of microelectronic devices above CMOS integrated circuits (ICs) has attracted a lot of attention in the last decade either to increase device density or add functionality. Demonstrations have already been shown for in-memory computing applications [78], quantum computing [189] and integrated nano-photonics [190]. Electron-beam lithography (EBL) is a powerful platform for R&D and prototyping that provides design and layout flexibility together with the high-resolution capability for emerging technologies development. In this paper, we will discuss the challenges in terms of EBL alignment on top of ICs at the node 130 nm and below, and present an alignment procedure based on cross-correlation.

Different methods exist for EBL alignment. The classical method, that we will refer to as line-scan alignment, consists of scanning square marker or cross. The electrons beam scans the area where the marker is supposed to be along vertical and horizontal lines. The backscattered electrons detected signal is used to locate the edges of the shape and calculate the coordinate of the marker center [191]. Another known method for EBL alignment [192, 193] is based on a cross-correlation algorithm of two images. The full area of the marker location is scanned by the electrons beam and the shift between the mar-

ker in the backscattered electrons images and the reference image from the pattern file is extracted. The shift between the scanned image and the reference image is measured by a phase correlation algorithm. A Fourier transform is applied to both images and the cross-power spectrum between the Fourier transform is computed. By applying the inverse Fourier transform of the cross-power spectrum, the shift between the two images can be determined.

The line-scan method is commonly used for its simple implementation, speed, and precision below 5 nm [194]. It is limited by the quality of the marker edges and requires large clear space around the marker to avoid signal disturbance during marker scanning. Different factors can impact the quality of alignments like the material contrast between the substrate and the marker or the marker depth [195]. The alignment performance can be improved by scanning the mark repeatedly and averaging the result of each lines-scan [196]. For the cross-correlation alignment, the marker design is a critical factor that affects the accuracy of the overlay alignment [197, 198]. It is more time-consuming compared to the line-scans approach due to the requirement to scan the whole marker area. The beam step size will also affect the accuracy even though it has been shown that sub-pixel accuracy is achievable [199]. This technique is useful when the marker is partially damaged, sub-5 nm alignment can be reached even with up to 80 % of the marker pattern missing [192].

BEOL materials, more specifically thick dielectrics, and design rules are fixed for each CMOS technology node and represent challenges for both the EBL of nanostructures and the alignment with the last metal level patterns for further post-processing. Cu damascene, that first appeared at the 130 nm CMOS node, forced the use of dummy structures that do lead to design rules specific to each CMOS node that were introduced afterwards [200]. These rules impede minimal and maximal dimensions as well as minimal and maximal spacing, usually around few micrometers and about a tenth of micrometers for minimal and maximal dimensions. As shown in Fig. 5.1, the maximal spacing does not allow the drawing of large clear space around the marks to perform an alignment by line-scans.

To address this issue, cross-correlation is a strategy that can be exploited for the EBL alignment on top of CMOS ICs. Multiple studies have reported [192, 193, 201] that Penrose tiles are good candidates for correlation-based alignment due to their aperiodic properties. Penrose tilings cannot be implemented on CMOS ICs due to the design rule that only allows specific angled edges (orthogonal and 45°). For this reason, different types of markers, compatible with CMOS ICs, are proposed and their performances in terms of processing and overlay accuracy will be detailed. We will use a method to quantify the accuracy of

the alignment process with an overlay of two lithography levels and present an alignment strategy based on a hybrid method using cross-correlation and line-scan alignment to address the challenges of CMOS ICs alignment.

5.2 Experiments

To measure the alignment performances, the overlay test structures shown Fig. 5.2(a) are used. The design Layers 1 and 2 (EBL 1 and EBL 2) are exposed independently with an alignment before each exposure in a single resist layer. The procedure is the following : a negative resist (MaN 2410), 100nm thick, is spun on the sample and loaded in the EBPG5200 EBL system. The first alignment by cross-correlation is performed and EBL 1 is exposed in the resist. The sample is unloaded from the tool and reloaded to simulate a real microfabrication process where the wafer position will be shift and misaligned in comparison to the previous lithography step. A second alignment by cross-correlation is performed and EBL 2 is exposed. Both exposures were performed at 100 keV, with a dose of 400 $\mu C/cm^2$, a current of 5nA, and a beam step size of 5 nm. The resist is then developed 2 min with an MF319 solution and observed with a Raith150 Two SEM, after the deposition of a thin Au/Pt layer of a few nanometers, to avoid charging effects.

The offsets between the EBL exposure are extracted by ProSEM software from GenIsys. The relative position of every feature is obtained in each quadrant and the X and Y offset of the two EBL exposures are calculated by the software automatically from the SEM images (see Fig. 5.2(b)). The exposure is performed using a 1 mm write field.

The figure of merit used to qualify the alignment is the mean distance-to-center defined as :

$$\overline{DTC} = \frac{\sum^{N} \sqrt{\Delta X^2 + \Delta Y^2}}{N}, \tag{5.1}$$

where ΔX and ΔY are the offset in X and Y and N the number of tests.

Two sets of experiments were performed. The first set on Si substrates, to test the alignment performance of different types of markers. The second set was done on CMOS die to evaluate the best alignment conditions on CMOS. Every test was repeated 25 times in the same conditions. The Fig.5.3(a) shows the layout of the full sample used for the experiment on Si, 25 chips with the different markers on each was used and the overlay test exposed inside the chip. As show on Fig.5.3(b), a single chip of 1x1 cm^2 was used for the CMOS test.

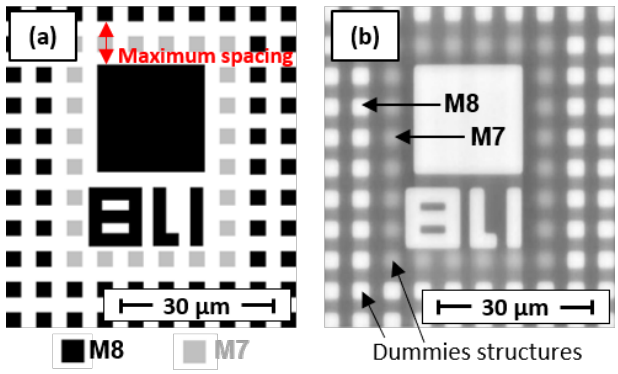


FIGURE 5.1 Marker on CMOS. (a) Layout of the last metal level (M8 in black) and the penultimate metal level (M7 in grey) with squared dummies. (b) Backscattered electron image captured by the EBPG5200 at 100 keV. This marker was designed to perform a line-scan alignment on the central square. These images highlight (i) that the design rule does not allow large clear space around the markers to perform line-scans since the maximum spacing is usually around a tenth of micrometers ; (ii) M7 Cu structures underneath M8 are also detected disturbing the alignment. For these reasons, line-scan alignment was impossible on this marker. Backscattered electron image parameters : EHT :100keV, WD :40mm, beam step size : 50nm

Three different types of markers were tested on a Si wafer by cross-correlation and compared with the classical line-scan alignment method, the same test was performed with this technique. The markers are made of 80 nm thick W directly sputtered on Si substrate. They were exposed by EBL and transferred to the W layer by a Plasma Etcher STS Multiplex Inductively Coupled Plasma (STS ICP) with the following recipe ; O_2 : 5 sccm, SF_6 : 65 sccm, temperature : 20°C, pressure : 10 mTorr, Coil power : 500W and Platen power : 30W. The three types of markers shown on top of Fig. 5.4 are a Cross, a QRCode, and a periodic marker known as a Sierpinski carpet. Each marker is 65 µm × 65 µm large. The pixel size used by the algorithm for the alignment is 50 nm.

On CMOS TSMC 130 nm ICs samples, a marker similar to the Sierpinski carpet shown in Fig. 5.4 was tested in different configurations. The markers are part of the last BEOL metal level (M8). There are made of 2 µm thick copper and capped by a 900 nm thick layer of SiN. The first test was performed in the same conditions as the markers on Si, i.e. 65 µm × 65 µm large and with a pixel size of 50 nm.

5.3 Results and discussion

The alignment results for each marker tested on Si are shown in Fig. 5.4. The QRcode shows the best $\overline{DTC}$ of all markers tested by cross-correlation with 6.50 nm (5.4(c)). The

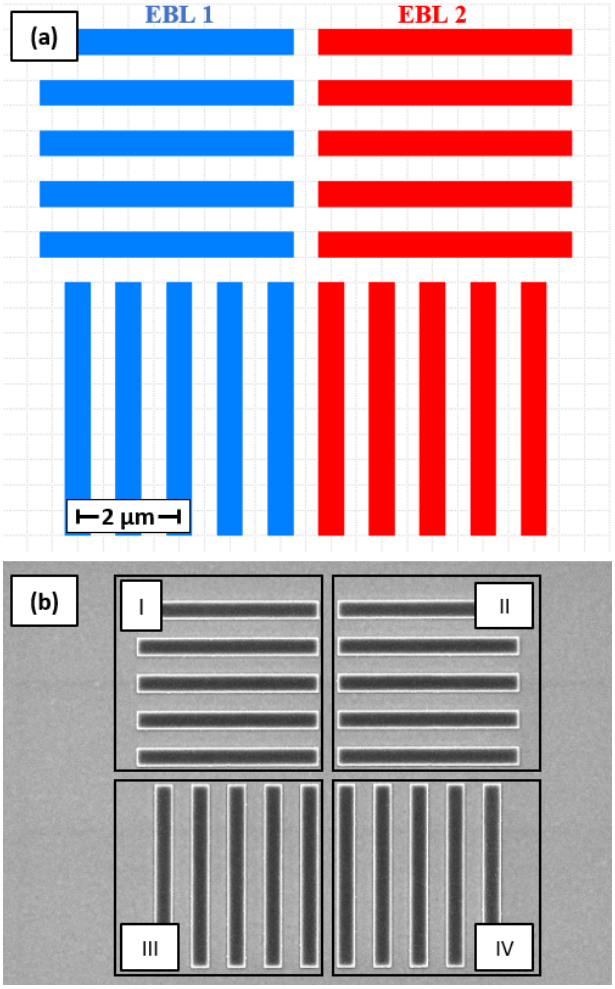


FIGURE 5.2 Overlay Test. (a) Layout of the overlay test performed in two steps in a single layer resist. The first lithography (EBL 1 / red) is aligned on M8 ; the sample is unloaded and reloaded into the system ; the second lithography (EBL 2 / blue) is carried out with a similar procedure to EBL 1. The sample is developed and imaged by SEM. (b) The overlay accuracy is measured by GenISys ProSEM software. From the SEM images, the software extracts automatically the relative position of every feature in each quadrant. The offset in X is measure by extracting the mean position of the feature in quadrant I (exposed during EBL 1) and quadrant II (exposed during EBL 2). The same procedure is used the extract the Y offset with quadrants III and IV. Backscattered electron image parameters : EHT :3keV, WD :7.4mm, beam step size : 4nm.

cross marker with a $\overline{DTC}$ of 8.23 nm also shows a good accuracy but with lower precision than the QRcode markers(5.4(b)), i.e. there is a shift from the target position. Finally, the Sierpinski carpet shows the highest mean offset with a $\overline{DTC} = 19.57$ nm(5.4(d)). As

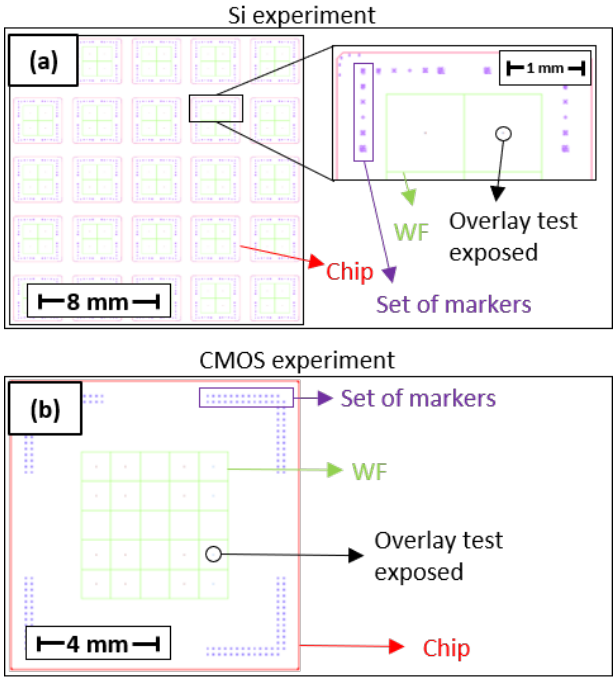


FIGURE 5.3 Samples layouts of the experiment. (a) Layout of the experiment on Si and (b) CMOS chip layout. The chips are in red, the markers location in purple, and the writefield (WF) in green. Every overlay test was exposed in the center of the WF.

shown in Fig.5.4(a), the line-scan alignment presents a $\overline{DTC} = 1.85$ nm showing that the classical alignment technique should be favored when possible.

The result of the QRCode and the Sierpinski carpet seems to show that the periodicity decreases the precision of the alignment. For correlation-based alignment, the important feature of a marker pattern is the autocorrelation peak of the marker pattern [193]. The sharpness of the peak will determine the sensitivity to a small positional offset and thus a sharper peak will increase the alignment performance. Fig.5.5 shows the autocorrelation peak of the reference images tested in this study. The sharpness of the peak confirms our experimental results. A sharper peak results in a lower offset.

The best marker design results in an autocorrelation that could be approximated by a 2D δ-function. This implies a uniform sampling of frequency space, meaning that the marker should be perfectly aperiodic. For those reasons, it has been shown [192, 193] that Penrose tiling [202] is a good candidate for correlation-based alignment due to their aperiodic tiling properties. However, Penrose tiles are not compatible with the design rules imposed by CMOS foundries since a minimum spacing between features is required and only specific angled edges are allowed.

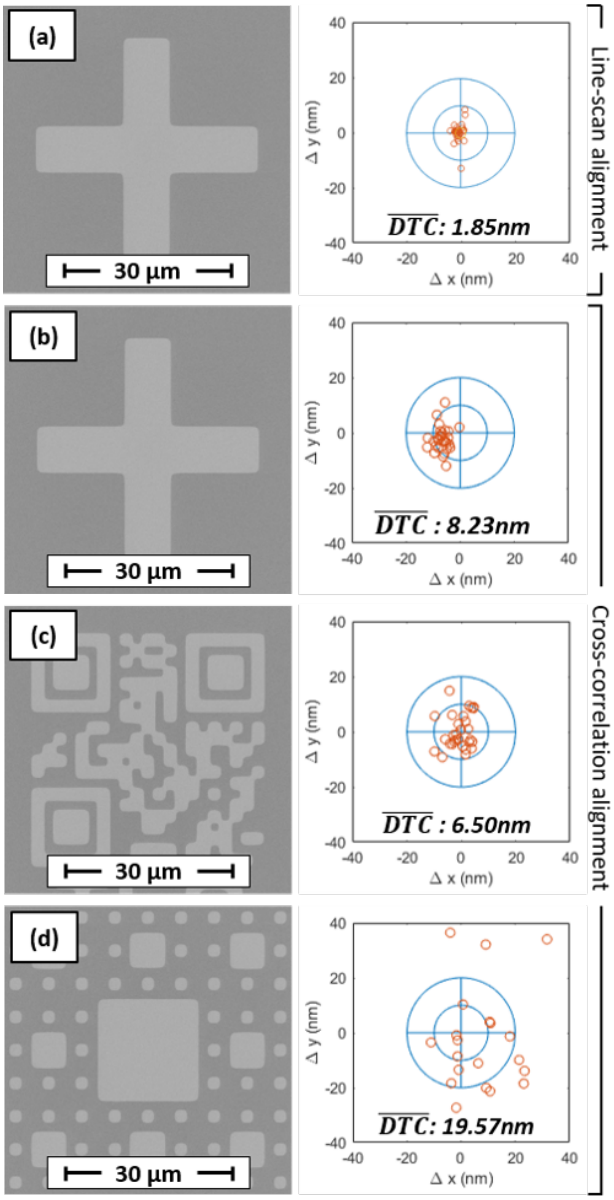


FIGURE 5.4 (a) Result of the experiment by line-scan alignment. (b-d) Result of the experiment by cross-correlation alignment. On the left ; backscattered electron image markers with the corresponding offset of the correlation-based alignment for each type of marker on Si. (b) Cross, (c) QRCode, and (d) the Sierpinski carpet. The plot show the offset in X (ΔX) and Y (ΔY) for the corresponding marker with their mean distance-to-center define as $\overline{DTC} = \frac{\sum^{N} \sqrt{\Delta X^2 + \Delta Y^2}}{N}$. Backscattered electron image parameters : EHT :100keV, WD :40mm, beam step size : 50nm

To decrease the peak sharpness of the autocorrelation while remaining compatible with specific design rules, a marker made of an array of structures generated by white noise could be investigated. White noise is a random signal having equal intensity at different

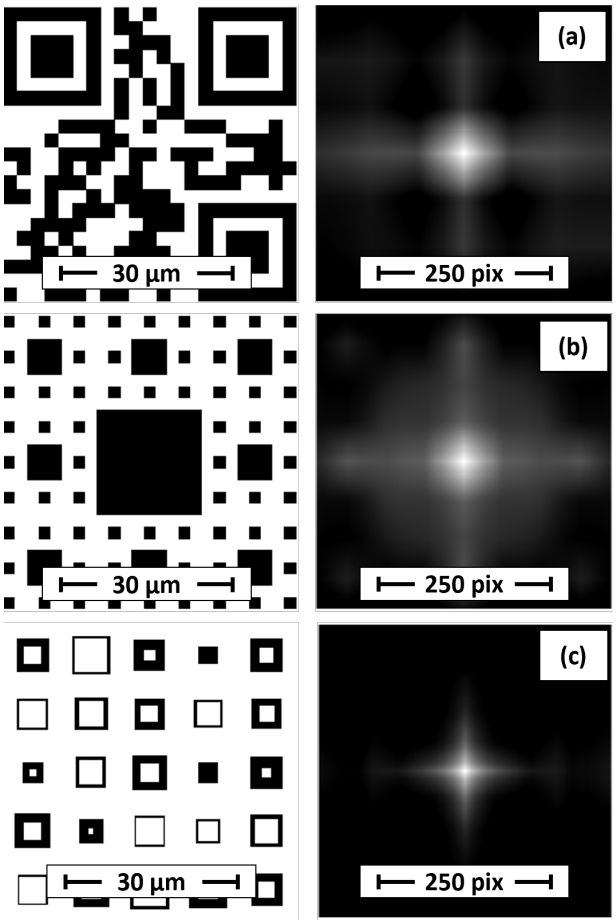


FIGURE 5.5 Reference image used for the cross-correlation with the corresponding autocorrelation function of the (a) QRcode, (b) Sierpinski carpet, and (c) white noise generated array. Choosing a marker shape with a function that tends towards a 2D-Dirac function is a way to improve the alignment accuracy. The autocorrelations are computed with an undersampling of 100 nm by pixel and zoomed around the central peak. All are shown with the same normalized brightness scale to enable a better comparison of the central peaks.

frequencies, giving it, therefore, a δ-function autocorrelation. Starting from an $n \times n$ array of inverted squares, the inner and outer squares can be adjusted following a white noise with a maximum and minimum amplitude set to comply with the dimension and spacing imposed by the design rules. Fig.5.5 shows that such patterns can drastically increase the autocorrelation peak sharpness and removing the subsidiary peaks around the central peak that are usually present, even in Penrose tilling pattern [193]. If the design rule allows it, the pattern can be rotated 45 degrees to exhibit a lack of coherence with the orthogonal pixel array. Small variations in the sampling position will provide significant changes in the detected pattern and therefore increase the sensitivity to misalignment.

As shown in Fig. 5.6(b), the result on CMOS is twice as large as on Si substrate for a similar marker shape (Sierpinski Carpet). This is due to the difference in contrast. The sample on Si has a higher contrast (W/Si) than the CMOS (Cu/SiOx). Moreover, the markers on CMOS are capped by a thick layer of SiN that decreases the contrast of the scanned markers. The results on CMOS could be improved with a marker shape that has a sharper autocorrelation peak but will always be worse than on a Si substrate. The alignment can be improved with a hybrid alignment strategy based on first coarse cross-correlation alignment and fine line-scan alignment. As we show in Fig. 5.1, line-scan alignment is impossible due to the maximum spacing/dimensions. Yet, with a first coarse alignment of 40.79 nm by correlation, it is now possible to target a specific area to perform a line-scan alignment on a small marker square. With this hybrid procedure, we reach a mean offset of 5.2 nm as shown in Fig. 5.6(c).

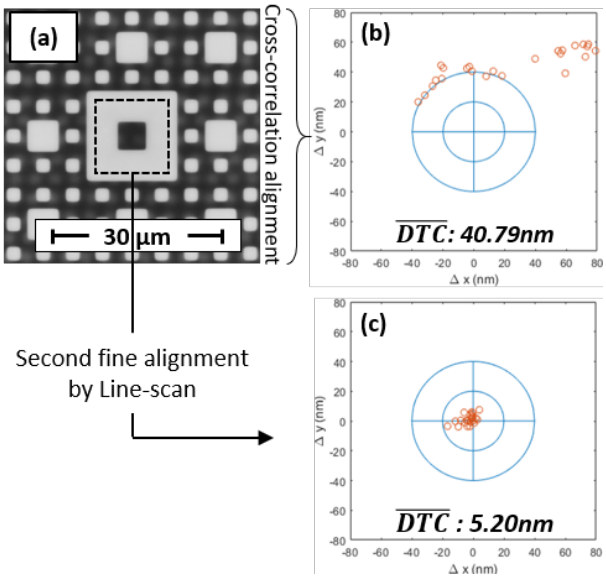


FIGURE 5.6 Backscattered electron image of marker (a) and offset of the correlation-based alignment on CMOS (b). To improve the alignment, a hybrid alignment strategy is proposed. A first coarse alignment by cross-correlation is performed with the marker shown in (a). The alignment is sufficiently accurate (b) to target a specific area (dash square shows in (a)) and perform a second fine alignment by line-scan that comply with the maximum dimensions constraint imposed by the design rules. (c) Results of the hybrid strategy with a second fine alignment by line-scan. With this strategy, a mean offset of 5.2 nm is reached. Backscattered electron image parameters : EHT :100keV, WD :40mm, beam step size : 50nm

Table 5.1 summarizes the result performed in this study compared with the literature and their applicability to CMOS IC. The result on Si substrate shows that for cross-correlation

alignment the marker design impacts the alignment accuracy. Large numbers of edges should be maximized, and periodicity of the marker shape should be avoided. Literature shows that Penrose tilling is the best candidate but cannot be applied to CMOS IC due to design rule limitations. The tests show that line-scan alignment provided a better accuracy compared to cross-correlation of marker tested in this study, but also reported with Penrose marker [192, 201] Although, sub-1nm accuracy could technically be achieved with Penrose marker [193]. The hybrid procedure proposed in this works allows benefiting from line-scan accuracy by taking advantage of the cross-correlation method to target a specific area of the marker and perform a successful line-scan even with the design restriction imposed by CMOS technology. This hybrid method is still limited by the quality of the marker inherent to a line-scan procedure. Designing markers with ultra-low periodicity and with a sharp autocorrelation peak, compatible with the design rules as shown in Fig.5.5 could improve the overlay accuracy.

Marker	Method	CMOS Compatibility	Mean offset (nm)	Ref
Test on Si substrate				
Cross	Line-scan	No	1.85	-
Cross	Cross-correlation	Yes	8.23	-
QRCode	Cross-correlation	Yes	6.50	-
Sierpinski carpet	Cross-correlation	Yes	19.57	-
Test on CMOS substrate				
Sierpinski carpet	Cross-correlation	Yes	40.79	-
Sierpinski carpet	Hybrid	Yes	5.20	-
Reported test				
Penrose tiling	Cross-correlation	No	sub-10nm	[201]
Penrose tiling	Cross-correlation	No	sub-5nm	[192]
Penrose tiling	Cross-correlation	No	sub-1nm	[193]

TABLEAU 5.1 Comparison of the different alignment accuracy tested on Si substrate and CMOS substrate, and with the literature. The applicability of the methods on CMOS substrate, due to the design constrain is shown in the CMOS compatibility. The hybrid procedure is the better alignment method that can be performed on CMOS.

5.4 Conclusion

In this study, we showed a method to test the alignment accuracy of different marker shapes, based on a cross-correlation algorithm. The shapes of the markers play an important role in the performance of cross-correlation-based alignment. The alignment accuracy is strongly related to the periodicity of the markers and the number of edges. For good alignment performances, high numbers of edges and aperiodic structures are required

In this context, the sharpness of the autocorrelation peak could be used as a metric to establish the quality of the markers. We proposed a method based on white noise signals to generate markers constrained by CMOS design rules with ultra-sharp autocorrelation peaks.

The result we obtained on CMOS samples led to a mean distance-to-center of 40.79 nm due to the periodicity of the structure but also due to the low contrast of the substrate. This result was improved by using the correlation-based alignment as a coarse alignment, followed by a fine alignment using line-scan. With this hybrid cross-correlation and line-scan alignment approach, we decreased the $\overline{DTC}$ from 40.79 nm down to 5.2 nm.

We showed that the cross-correlation alignment can be used on CMOS ICs to workaround the challenges of alignment imposed by the layout design rules. With a hybrid cross-correlation and line-scan alignment, a sub-10nm overlay accuracy can be achieved.

Discussion supplémentaire à l'article

Cette étude a développé une méthode permettant de s'aligner avec une précision ≤ 10 nm sur les puces CMOS utilisées dans ce projet. Suite à sa publication, l'étude a également conduit à la création de marques d'alignement optimisées pour un alignement par corrélation croisée pour des projets futurs. La marque présentée sur la figure 5.7 (a) est générée à partir de bruit, tandis que la structure affichée sur la figure 5.7 (b) est construite à partir d'une grille de Penrose, tout en respectant les règles de conception de TSMC. Cette dernière marque a été utilisée dans un autre projet dirigé par le Professeur Dominique Drouin, nommé UNICO, où un alignement précis sur des puces CMOS TSMC était également requis.

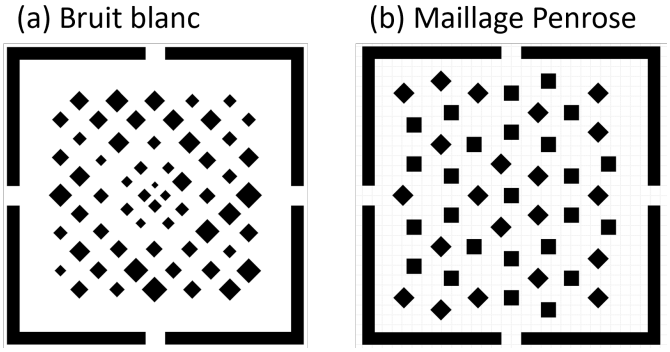


FIGURE 5.7 Marque d'alignement par corrélation croisée : (a) Marque générée à partir d'une matrice de carrés dont les dimensions ont été modifiées à l'aide de bruit blanc. (b) Marque créée à partir d'un maillage de Penrose où des carrés ont été positionnés à chaque nœud du maillage.

En plus des contraintes de conception CMOS, ces marques se sont révélées utiles pour les procédés de fabrication non intégrés au CMOS, car elles facilitent l'alignement même en présence d'un contraste de matériaux très faible. C'est notamment le cas dans les procédés de fabrication damascène où les marques d'alignement sont composées de TiN et de SiN, qui ont des numéros atomiques similaires.

CHAPITRE 6

Intégration CMOS

Le chapitre 3 compare le développement et la fabrication de trois schémas d'intégration de crossbars passifs, en termes de caractéristiques morphologiques et de performances électriques. Chaque schéma utilise uniquement deux niveaux de masque pour fabriquer chaque série d'électrodes, la couche de commutation ReRAM étant structurée pendant ces étapes. Le chapitre 4 décrit la fabrication d'un schéma d'intégration de crossbar, où la couche de commutation est spécifiquement gravée sous forme de pilier. L'étude présentée dans le chapitre 5 démontre la capacité de s'aligner avec le huitième niveau de métallisation des puces CMOS de TSMC, ouvrant ainsi la possibilité de fabriquer monolithiquement les structures crossbar développées précédemment sur ces puces.

Ce chapitre débutera par une présentation de la puce CMOS utilisée pour l'intégration et une brève description des circuits fabriqués sur la puce. Les procédés de fabrication développés dans ce projet seront présentés, suivis d'une discussion sur les adaptations nécessaires ou souhaitables pour le transfert industriel de ces schémas.

6.1 Matériels

Pour montrer l'intégration complète de crossbars passifs dans le BEOL, nous utiliserons une puce CMOS fabriquée par TSMC et conçue par Roman Genov, Amirali Amirsoleimani et leurs équipe de l'Université de Toronto.

6.1.1 Description de la puce CMOS HiData

La puce CMOS, fabriquée par TSMC avec la technologie de 130 nm, est montrée dans la figure 6.1 où une coupe FIB de la section de la puce est présentée. Le transistor dans le FEOL et les huit niveaux d'interconnexion métallique constituant le BEOL sont visibles. Ces niveaux sont composés d'interconnexions en cuivre, fabriquées selon un schéma de dual-damascène. Les niveaux 2 à 7 présentent un pas minimum de 450 nm et des dimensions critiques de 200 nm. Le dernier niveau métallique (M8), également appelé UTM (*ultra thick metal*), possède des CDs de 2 µm et une épaisseur nettement supérieure aux autres niveaux, car il est généralement utilisé pour l'interconnexion avec les PCBs et est conçu pour minimiser la résistance, étant donné que les courants d'alimentation sont généralement distribués à travers ce niveau. Le niveau M8 est encapsulé par trois couches de diélectrique dont l'épaisseur totale est de 870 nm. Bien que la composition chimique exacte ne soit

pas divulguée par TSMC pour des raisons de confidentialité, l'analyse des permittivités suggère que l'empilement pourrait se rapprocher d'une composition $SiN/SiO_2/SiN$, avec des épaisseurs respectives de 70/400/400 nm. Comme détaillé dans la section suivante, les crossbars seront fabriqués sur la dernière couche de SiN et seront connectés au M8 à travers cet empilement.

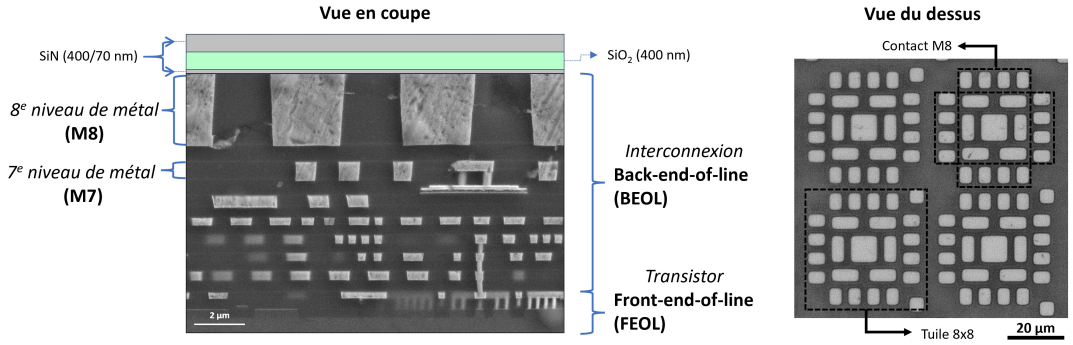


FIGURE 6.1 Description de la puce CMOS TSMC : (a) Vue en coupe du FEOL, du BEOL et des couches d'encapsulation. (b) Vue de dessus des plots de contact M8 pour l'interconnexion des crossbars avec le BEOL.

6.1.2 Agencement des circuits

Le concept de ce système intégré vise à subdiviser une large matrice de mémoire de 128 × 128 en tuiles de crossbars passifs de 8 × 8 pour surmonter les non-idéalités des crossbars passifs dans les configurations de mémoire de grande taille ($n \times n$). Cette modularité permet un contrôle indépendant de chaque tuile, réduisant ainsi les défis liés aux courants parasites et limitant les niveaux de courant qui peuvent devenir significatifs, comme expliqué dans le chapitre 2.2.2.

L'image 6.2(a) présente une tuile crossbar 8 × 8 et la figure 6.2(b) illustre le schéma électrique du découpage de la matrice 128 × 128 en tuiles de 8 × 8. La figure 6.2(c) montre le réseau de crossbars passifs 8 × 8 après la fabrication de la couche de terminaison électrique et les étapes d'encapsulation et d'interconnexion. Comme montré sur la figure 6.3, plusieurs cellules de tailles variées – 8 × 8 (une seule tuile crossbar), 16 × 16 (2 × 2 tuiles) et 128 × 128 (16 × 16 tuiles) – sont présentes pour tester les différents blocs indépendamment dans le cadre de cette démonstration de la matrice complète avec circuits périphériques. On note que, dans les cellules équipées de DAC/ADC, ce sont les circuits périphériques qui occupent le plus d'espace.

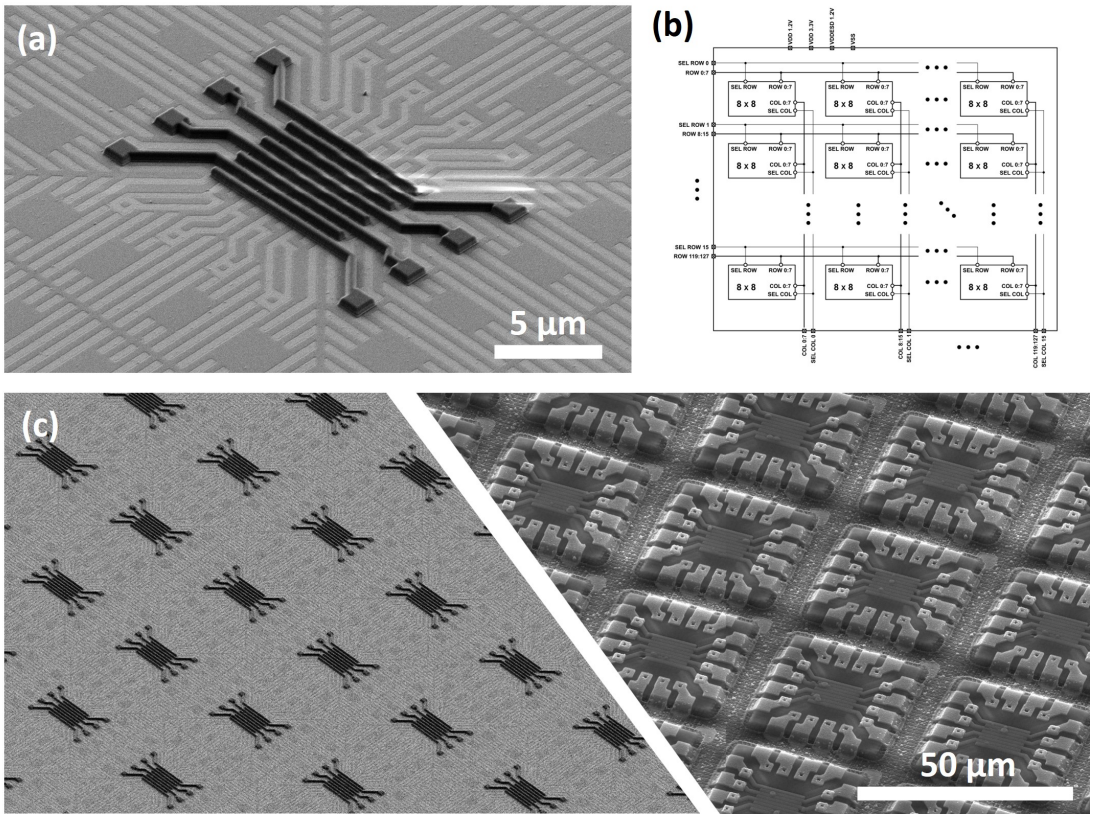


FIGURE 6.2 (a) Tuile crossbar 8 × 8 après la fabrication de la TE. (b) Schéma électrique montrant la segmentation de la matrice 128 × 128 en tuiles 8 × 8. (c) Maillage de crossbars 8 × 8 après la fabrication de la TE et après la fabrication des interconnexions.

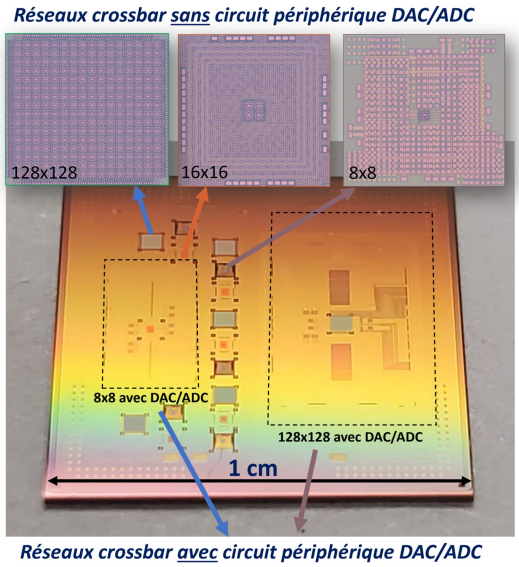


FIGURE 6.3 Puce de 1 cm^2 montrant les différentes cellules de tailles 8 × 8, 16 × 16 et 128 × 128, avec et sans les circuits DAC/ADC périphériques.

6.2 Procédé de fabrication

Le procédé de fabrication est divisé en deux étapes principales. La première, décrite dans la première section du manuscrit, concerne la fabrication des crossbars. On commencera par comparer les schémas d'intégration étudiés pour sélectionner ceux qui seront intégrés au BEOL. La seconde étape, détaillée ci-après, implique l'encapsulation et l'interconnexion des crossbars avec les contacts M8.

6.2.1 Choix du schéma de fabrication

Le tableau 6.1 répertorie les quatre schémas développés dans les chapitres précédents, en énumérant respectivement leurs avantages et inconvénients.

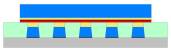
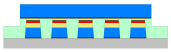
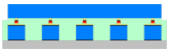
	Damascène	**Single subtractive**	**Double subtractive**	**TopPilar**
				
Résistivité BE	-	+	+	++
AR ↗	-	+	+	++
CD ↘	-	+	+	+
SL isolé	-	-	+	+
Standard BEOL	++	+	+	-
# de masques				-
Potentiel	- -	+++	+++++	++++
Stabilité CMP	+	-	-	- -
Perf. électrique	+	+ +	+ +	n.d.
Maturité	++	+	+	- -

TABLEAU 6.1 Comparaison des schémas d'intégration crossbar. AR ↗ : Potentiel d'amélioration du rapport d'aspect. SL : *Switching layer*. CD ↘ : potentiel de réduction des dimension critique. n.d. : non-démontré.

L'approche Damascène, bien que limitée par la résistance d'accès qui restreint l'augmentation de la taille du réseau, bénéficie de la stratégie consistant à subdiviser de grandes matrices en petits crossbars pour atténuer cette contrainte. Néanmoins, la réduction des dimensions demeure entravée par les difficultés associées au remplissage des tranchées. En revanche, les approches soustractives, en intégrant un matériau moins résistif pour le BE, améliorent les résistances et offrent de meilleurs rapports d'aspect. Ceci réduit davantage les résistances et facilite une mise à l'échelle plus efficace des crossbars. Par ailleurs, dans ces approches, la limitation des dimensions est essentiellement due à la résolution de la lithographie. L'approche soustractive double et TopPilar permettent d'isoler la SL à chaque intersection. Spécifiquement, l'approche TopPilar se distingue en découplant les dimensions des électrodes de celles de la mémoire. Cela permet de réduire la taille de

la mémoire pour améliorer sa stabilité électrique tout en maintenant des électrodes plus larges, évitant ainsi une augmentation excessive des résistances. À noter également que, bien qu'un procédé de lithographie en niveau de gris ait été utilisé pour réduire l'ajout de niveaux de masque, un transfert industriel nécessiterait trois niveaux de masque (deux pour les électrodes et un pour le pilier), ce qui impliquerait un niveau supplémentaire dans cette approche.

Le procédé damascène est le plus avancé au sein du groupe de recherche INPAQT, ayant permis la réalisation de plusieurs fabrications de crossbars. Les approches soustractives, simples et doubles, pourraient offrir de meilleures performances électriques que l'approche damascène, comme montré dans le chapitre 3. L'approche double soustractive s'est avérée prometteuse pour des applications cryogéniques [166], bien qu'une programmation complète des crossbars n'ait pas encore été démontrée. Concernant l'approche TopPilar, elle a été validée uniquement sur le plan morphologique et ses performances électriques nécessitent des études plus approfondies. La stabilité limitée des approches soustractives est attribuée à la faible sélectivité entre le SiO_2 et le TiN, compliquant le contrôle du processus. Cette problématique est particulièrement prononcée dans l'approche TopPilar, où la présence de piliers de petites dimensions isolés peut intensifier les effets d'érosion.

Les approches soustractives, bien que plus prometteuses, sont encore en cours de développement. Pour cette raison, il a été choisi d'effectuer une première démonstration en utilisant l'approche damascène, qui sera suivie par l'approche soustractive simple. La figure 6.4 illustre les deux schémas d'intégration choisis pour être intégrés après le dernier niveau de métal de la puce CMOS. La fabrication des crossbars suit le même procédé que celui décrit dans le Chapitre 3.

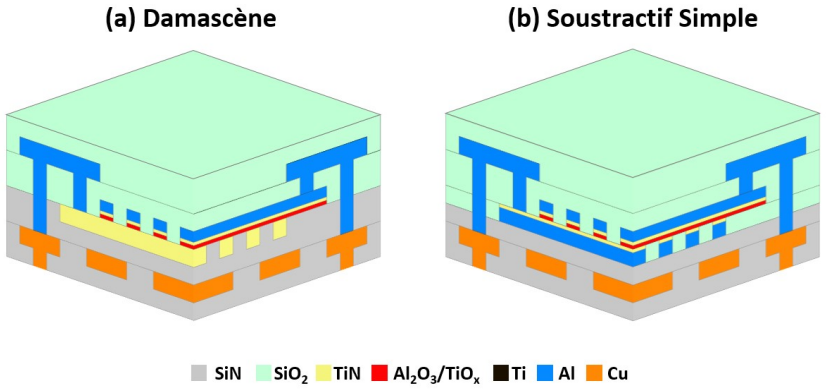


FIGURE 6.4 Illustration de l'intégration au-dessus du dernier niveau d'interconnexion de la puce CMOS : (a) l'approche damascène ; (b) l'approche soustractive simple.

Pour l'approche damascène, les BEs sont gravées directement dans le SiN qui encapsule le niveau M8 de la puce. Ces BEs sont définies par lithographie de faisceaux d'électrons et gravées par DRIE avec une profondeur de 150 nm et une largeur de 400 nm pour limiter le rapport d'aspect et minimiser les problèmes liés au remplissage. La TE est dimensionnée avec une largeur de 200 nm et est définie comme détaillé dans la section 3.2.1.
Pour l'approche Soustractive simple, le même empilement de métal est déposé sur le SiN d'encapsulation de la puce, et les mêmes conditions de fabrication sont appliquées, également détaillé dans la section 3.2.1.

6.2.2 Procédé d'encapsulation et d'interconnexion

Une fois les crossbars fabriqués, le procédé de fabrication utilisé pour connecter les crossbars au M8 est identique pour les deux approches. Comme illustré sur la figure 6.6(a), une couche d'oxyde de 500 nm est déposée pour encapsuler les TEs et la SL. Des vias sont ensuite ouverts au-dessus des BEs et des TEs par une même exposition aux faisceaux d'électrons et gravés à l'aide d'une gravure plasma en utilisant un mélange de gaz $C_4F_8/SF_6/H_2$ (50/20/25 sccm) (b). Des ouvertures au-dessus des contacts M8 sont créées avec la même recette de gravure par DRIE (c). Finalement, une couche d'aluminium de 500 nm est déposée par pulvérisation, puis exposée et définie à l'aide d'une gravure plasma utilisant un mélange de gaz $BCl_3/Cl_2/Ar$ (17.5/2.5/10 sccm) (d).

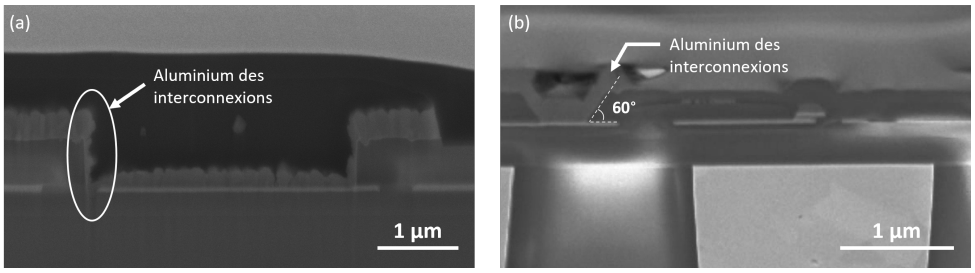


FIGURE 6.5 Couverture d'aluminium sur les flancs des vias sur les électrodes : (a) Couverture à 90°. (b) Couverture à 60°.

Étant donné que l'épaisseur du dépôt d'aluminium est inférieure à la profondeur des vias, la couverture de l'aluminium sur les flancs devient un paramètre critique. Comme le montre l'image MEB de la figure 6.5(a), des flancs de gravure à 90° résultent en une couverture d'aluminium très faible, ce qui peut causer des problèmes de contact électrique ou augmenter la résistance due à la minceur de l'aluminium sur ces flancs. Pour résoudre ce problème, la recette de gravure a été modifiée pour obtenir des flancs à 60°. Cette modification a été réalisée en ajoutant du C_4F_8 au mélange de gaz pour favoriser la polymérisation de la résine, entraînant un redépôt de fluorocarbone. L'angle ainsi créé assure une épaisseur d'aluminium plus uniforme autour des vias, comme illustré dans la figure 6.5(b).

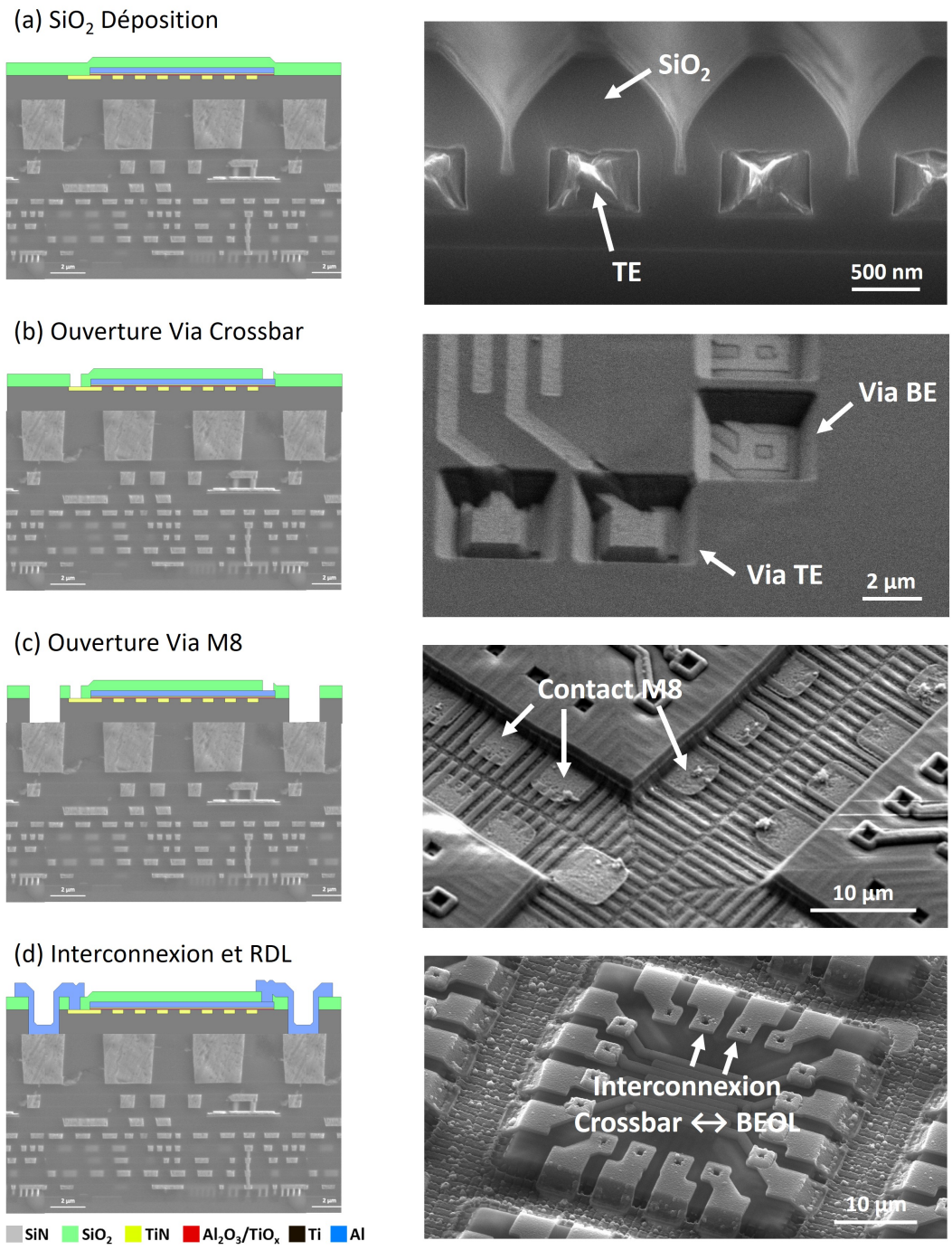


FIGURE 6.6 Procédé de fabrication de l'interconnexion des crossbars avec le niveau de métal M8. À gauche : Schéma en coupe de la puce. À droite : Vue au MEB. (a) Encapsulation par une déposition de SiO_2 de 500 nm. (b) Ouverture des vias sur les électrodes du crossbar par gravure plasma fluoré. (c) Ouverture des vias sur les contacts M8 par gravure plasma fluoré. (d) Connexion entre les vias du crossbar et M8 par un dépôt suivi d'une gravure chlorée de lignes d'Al.

6.3 Résultats morphologiques

Les approches présentées dans la première partie du manuscrit ont été développées et fabriquées sur des tranches de 100 mm. Compte tenu du coût de fabrication, les puces CMOS ont des dimensions de 1 x 1 cm, ce qui est suffisamment grand pour permettre les étapes de microfabrication. Cependant, le polissage de petits échantillons carrés accentue les effets de non-uniformité qui apparaissent sur les bords de l'échantillon. La figure 6.7 montre une image microscopique de l'échantillon complet après CMP pour chacune des deux approches.

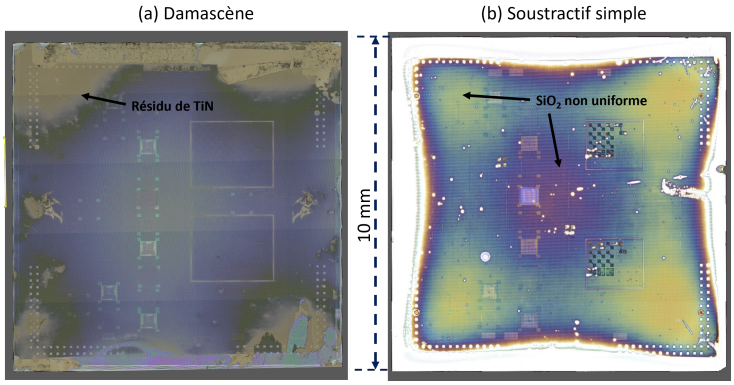


FIGURE 6.7 Image microscopique après CMP de la puce complète pour les approches (a) damascène et (b) soustractive. On observe des différences de taux de polissage entre les bords et le centre de la puce, entraînant un polissage non uniforme pour les deux approches

Dans les deux cas, nous observons un polissage non uniforme dû à des différences de pression entre le centre et le bord de la puce, ce qui entraîne des taux de polissage différents. Dans le cas de l'approche damascène, le centre se polit plus rapidement que les bords, ce qui peut provoquer un court-circuit des cellules proches des bords à cause des résidus de TiN. De même, le SiO_2 est poli plus rapidement au centre, laissant un excédent de SiO_2 sur les bords de la puce et rendant les BEs inaccessibles pour le contact avec les TEs. Pour ces raisons, le procédé est calibré principalement sur les cellules situées au centre, et les crossbars sont de préférence fabriqués sur ces cellules centrales.

La figure 6.8 présente des images MEB après CMP des cellules situées au centre de la puce. Des structures de remplissage sacrificielles (*CMP dummies*) sont ajoutées autour des crossbars, et les zones de contact des électrodes sont striées pour maintenir une densité de structures uniforme, afin d'éviter les problèmes d'érosion et d'affaissement (*dishing*). Nous observons encore ici des défauts sur les bords des électrodes dus aux problèmes de remplissage des tranchées dans l'approche damascène. Il est à noter que le même phénomène

existe dans l'approche soustractive mais dans le diélectrique, ce qui n'affecte pas dans ce cas la qualité de l'interface entre la BE et la SL.

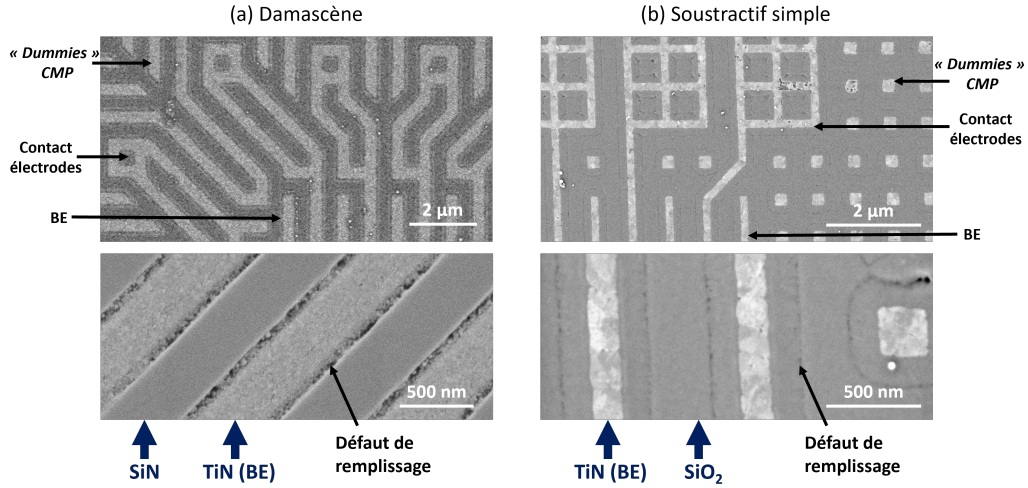


FIGURE 6.8 Images MEB après CMP des cellules au centre de la puce pour (a) l'approche damascène et (b) l'approche soustractive.

La figure 6.9 montre des images AFM des mêmes zones sur la puce. La topographie y est plus marquée que celle présentée dans les résultats du chapitre 3, due aux défis d'uniformité engendrés par le polissage de petits échantillons. On observe une topographie de ~ 25 nm pour l'approche damascène et de ~ 10 nm pour l'approche soustractive, cette dernière présentant à nouveau la topographie la plus faible.

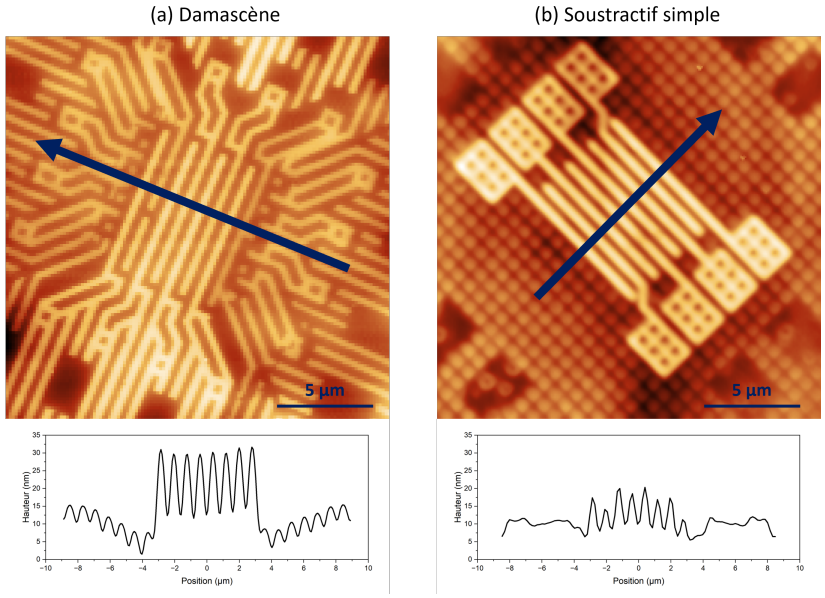


FIGURE 6.9 Images AFM après CMP de la BE, dans des cellules au centre de la puce pour (a) l'approche damascène et (b) l'approche soustractive.

L'image FIB de la figure 6.10 montre une section de la puce après fabrication, encapsulation, et interconnexion des crossbars avec l'approche damascène. Cette réalisation technique illustre le succès du processus, malgré les défis inhérents à une intégration CMOS. Sur l'image du bas, nous observons la section des BEs fabriquées avec le procédé de damascène. Contrairement aux tranches de Si utilisées dans le chapitre précédent, qui étaient parfaitement planaires, la puce ici présente une topographie non planaire au niveau du contact de Cu du M8 qui affecte directement l'épaisseur des BEs. Après la gravure des tranches de BE, la profondeur est initialement identique, mais la CMP réalisée sur le TiN planarise la topographie résiduelle du M8. Cela entraîne une réduction de l'épaisseur finale des BEs de dimensions de 150 nm à 50 nm dans les zones où le BEOL de la puce est le plus surélevé.

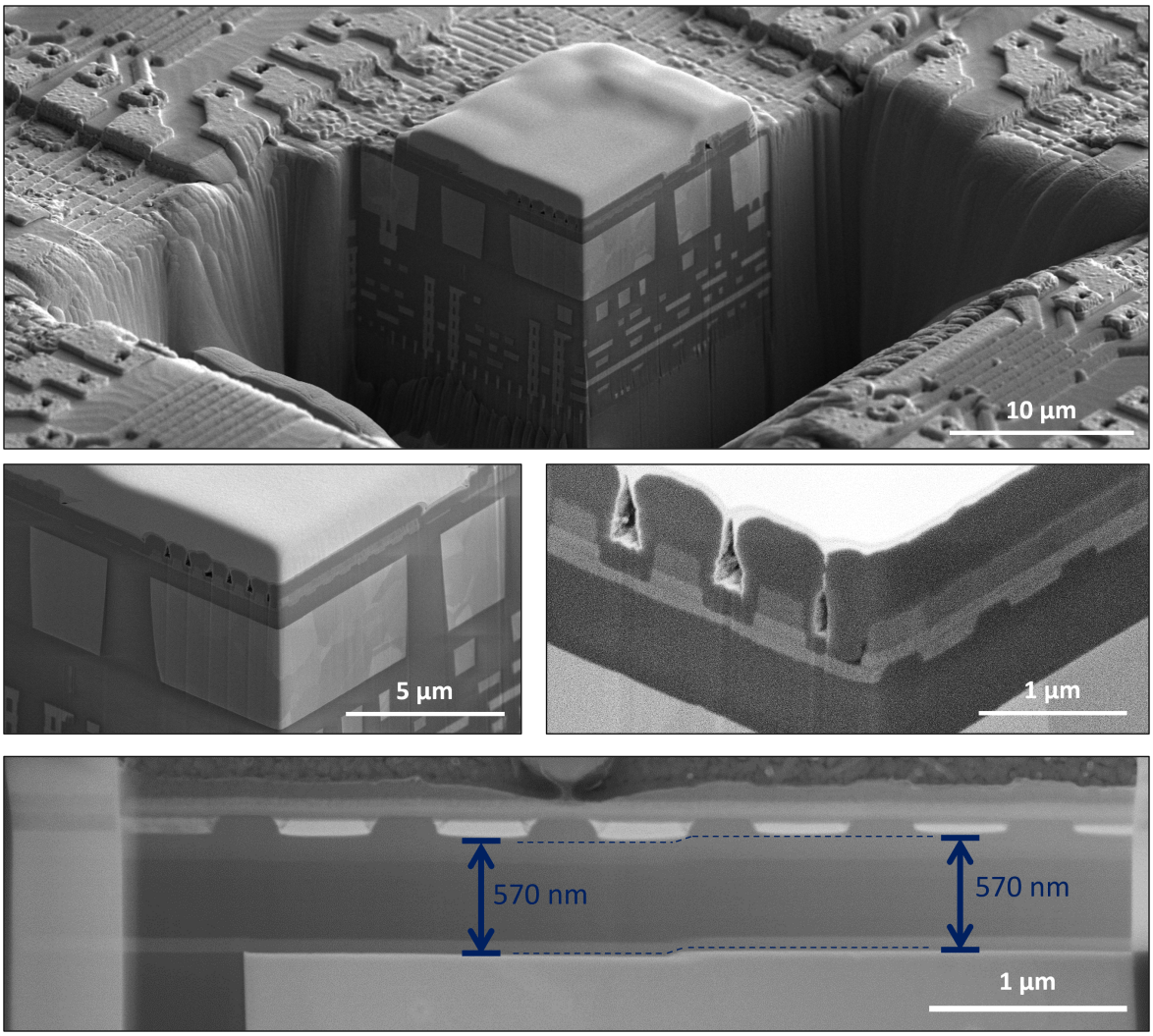


FIGURE 6.10 Coupe FIB d'un crossbar intégré au-dessus du BEOL et connecté à la puce CMOS. L'effet de la topographie résiduelle du M8, transférée sur les crossbars, peut être observé.

Cependant, ces résultats montrent que les étapes de fabrication, notamment l'étape critique de polissage CMP, ont pu être réalisées avec succès pour les approches damascène et soustractive simple, malgré les défis d'uniformité dus à la petite taille des puces et à la topographie occasionnelle causée par les lignes métalliques M8. Ils illustrent la capacité à fabriquer des circuits crossbar avec une planarité inférieure à 30 nm et confirment que les étapes de connexion des crossbars avec la puce ont pu être réalisées.

6.4 Optimisation et Perspectives

Bien que l'intégration ait été réalisée avec succès, plusieurs pistes peuvent être explorées pour améliorer la stabilité du procédé. Les optimisations seront illustrées avec l'approche damascène mais restent applicables aux autres approches soustractives.

6.4.1 Ouverture M8

Étant donné que la profondeur de gravure est plus importante à l'étape d'ouverture des vias M8, une résine plus épaisse avec une résolution moindre est utilisée pendant l'exposition. Pour cette raison, un anneau autour du crossbar est ouvert, ce qui permet d'obtenir des dimensions critiques plus larges, comme montré sur l'image MEB de la figure 6.6(c). Cependant, cela impacte la dernière étape où les interconnexions sont définies. Comme indiqué sur la figure 6.11(a), la topographie ainsi créée réduit la résolution de la lithographie en raison d'une différence d'épaisseur de la résine entre le haut et le bas du via. De plus, une résine plus épaisse est nécessaire pour assurer qu'il reste suffisamment de résine au-dessus des structures.

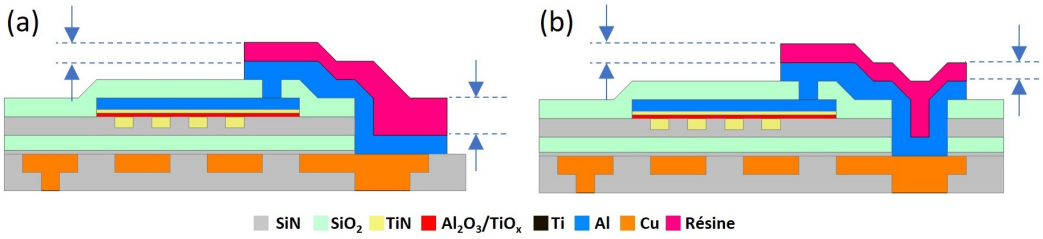


FIGURE 6.11 Illustration des problèmes de lithographie engendrés par la topographie : (a) Large ouverture M8 : La gravure de l'aluminium se fait sur le haut et le bas de la résine, où la différence d'épaisseur limite la résolution à cette étape. (b) Ouverture M8 juste au-dessus des contacts M8 : Malgré la topographie, les zones développées présentent la même épaisseur de résine.

L'utilisation de masques durs à l'étape d'ouverture des contacts M8 ou l'amincissement de l'épaisseur du diélectrique avant le début du procédé de fabrication pour réduire l'épaisseur à graver lors de l'étape d'ouverture M8, permettrait d'utiliser une résine moins épaisse pour n'ouvrir les vias que juste au-dessus des contacts. Dans ce cas, la gravure de l'aluminium

se ferait uniquement au-dessus des structures avec une épaisseur de résine identique, ce qui améliorerait la résolution de la lithographie comme illustré sur la figure 6.11(b).

6.4.2 Réduction de la topographie

Pour atténuer les défis liés à la réalisation d'étapes de lithographie sur une topographie significative, plusieurs solutions peuvent être envisagées. Tout d'abord, réduire l'épaisseur de la couche diélectrique au-dessus du M8 avant de commencer la fabrication des crossbars permet de diminuer la profondeur des vias. Après le dépôt de SiO_2 suivant la fabrication des crossbars, la surface peut être planarisée par CMP pour réduire la topographie avant l'étape de fabrication des interconnexions métalliques, comme le montre la figure 6.12(b). Une autre stratégie consiste à augmenter l'épaisseur de l'aluminium jusqu'à combler le via, ce qui présente également l'avantage de prévenir les problèmes de couverture sur les flancs des vias, illustrés dans la figure 6.12(c). Il est à noter que la sélectivité actuelle pourrait ne pas être suffisante pour graver une couche plus épaisse avec la recette de gravure actuelle, et l'utilisation d'un masque dur pourrait s'avérer nécessaire. Pour éliminer complètement la topographie, une approche dual-damascène peut être envisagée en remplaçant les interconnexions en cuivre et en ajoutant une étape de CMP supplémentaire mais sans l'ajout de niveaux de masques supplémentaires, comme illustré dans la figure 6.12(d).

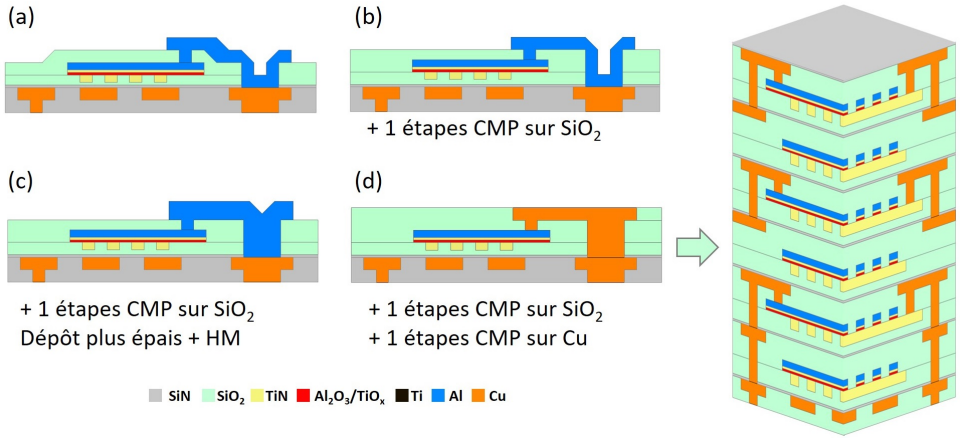


FIGURE 6.12 Proposition pour réduire la topographie du procédé de fabrication. (a) Schémas présentés précédemment. (b) Planarisation de la couche d'encapsulation. (c) Augmentation de l'épaisseur du dépôt d'aluminium. (d) Schémas dual-damascène avec du Cu pour éliminer complètement la topographie. Cette dernière proposition ouvre la possibilité d'un empilement des crossbars.

Dans cette dernière approche, l'interconnexion métallique est similaire à un niveau BEOL classique et permet une intégration non plus au-dessus du BEOL mais au sein même du BEOL. Cette configuration facilite l'empilement de plusieurs niveaux d'interconnexion avec des crossbars, ouvrant ainsi la voie à l'intégration 3D.

6.4.3 Couche d'arrêt pour les étapes de gravure

Durant l'étape de gravure de la TE et de la SL avec une chimie à base de Cl_2/BCl_3, le TiN et l'Al sont partiellement attaqués lors de la surgravure. Pour améliorer la stabilité de cette étape, il pourrait être envisagé d'ajouter une couche d'arrêt sur la BE. Pour des raisons précédemment évoquées, il n'est pas possible d'utiliser un empilement de plusieurs matériaux pour l'approche damascène. Cependant, pour les approches soustractives, une fine couche de tungstène (W) pourrait être utilisée. À noter que pour l'approche TopVia, la BE n'est pas affectée par la gravure de la TE, car elle est encapsulée dans le SiO_2 après la CMP, mais l'ajout d'une couche de W est tout de même à envisager pour servir de couche d'arrêt lors de la gravure des piliers mémoire. Pendant l'ouverture des contacts dans le diélectrique au-dessus des électrodes avec une gravure fluorée, le TiN est également gravé par ce type de chimie. Ce n'est pas problématique pour les approches soustractives car, même si la fine couche de TiN est gravée, la majorité de l'électrode en Al possède une très bonne sélectivité au fluor. Cependant, cela pose plus de problèmes pour l'approche damascène où la BE est uniquement en TiN et l'ajout d'une couche d'arrêt n'est pas possible. À moyen terme, il y a donc plus de possibilités pour stabiliser le procédé de fabrication avec les approches soustractives qu'avec l'approche damascène.

6.4.4 Couche d'arrêt pour l'étape de CMP

L'approche soustractive développée dans ce projet est limitée par une absence de sélectivité de polissage entre le SiO_2 et le TiN, proche de 1 avec une slurry à base de silice, ce qui rend l'arrêt du processus difficilement contrôlable. Cependant, l'ajout d'une couche d'arrêt de SiN pour arrêter le polissage de SiO2 peut permettre d'augmenter la stabilité de cette étape. Des tests effectués avec notre équipement de CMP ont montré qu'une sélectivité de 3.7 entre le SiO2 et le SiN pouvait être atteinte avec des paramètres de pression de 900 g, des rotations plateau/tête de 70/60 rpm avec une slurry à base de silice (Allied 0.05 µm). De plus, la slurry à base de cérium utilisée pour polir le TiN dans le procédé damascène devrait également fournir une bonne sélectivité puisque cette slurry a été développée pour des procédés STI où le SiO2 est poli avec du SiN comme couche d'arrêt [162]. Cependant, la couche d'arrêt de SiN devra être enlevée avec un procédé sélectif au SiO_2 avant le dépôt des étapes suivantes.

6.5 Perspective d'intégration BEOL

Les schémas d'intégration CMOS réalisés dans ce chapitre fabriques les crossbars sur le dernier niveau de BEOL (Intégration Top-BEOL illustrée sur la figure 6.13(a)). Comme expliqué dans la section précédente, l'utilisation de schémas d'interconnexion planaire, comme le procédé dual Damascène, permet un empilement de plusieurs niveaux. Dans cette configuration, l'intégration des crossbars se fait non plus sur, mais dans les niveaux de BEOL (Intégration 3D-BEOL illustrée sur la figure 6.13(b)). Comme il sera expliqué dans cette section, l'optimisation du procédé de fabrication, où les électrodes fusionnent avec les interconnexions du BEOL, permet, dans le cas des approches soustractives, une intégration où chaque empilement de réseaux de mémoire est connecté avec ceux du dessus et du dessous (Intégration 3D-FC - *fully connected* - illustrée sur la figure 6.13(c)).

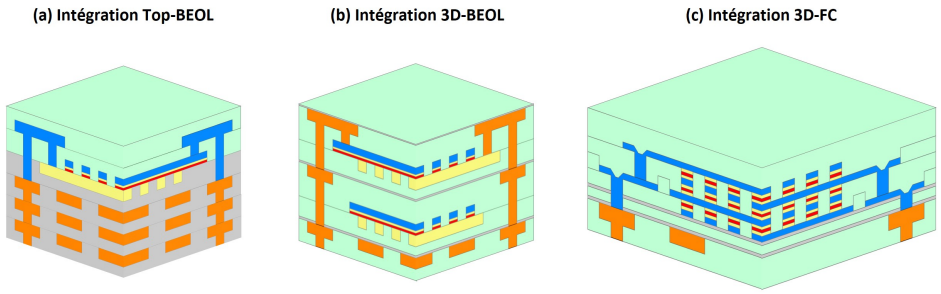


FIGURE 6.13 Types de schémas d'intégration BEOL : (a) Intégration sur le dessus du dernier niveau de BEOL. (b) Intégration dans les niveaux d'interconnexion BEOL qui permet un empilement 3D. (c) Intégration 3D où chaque réseau crossbar est complètement connecté (*fully connected*) avec le suivant.

L'avantage de l'approches 3D-BEOL réside dans la possibilité de sélectionner les meilleurs schémas d'intégration pour les crossbars, notamment soustractif tout en conservant le standard des interconnexions en cuivre damascène, qui est bien établi dans une perspective de transfert industriel. Cependant en fusionnant les interconnexions métalliques qui connectent les mémoires (BE et TE) avec celles qui se connectent au BEOL, le nombre d'étapes pour un réseaux crossbar peux être optimiser. Dans ce cas, le schéma d'interconnexion doit être étudié spécifiquement pour chaque approche.

6.5.1 Dual Damascène

L'approche damascène est intéressante pour intégrer la technologie dans un système 3D-BEOL, en conservant les méthodes traditionnelles d'interconnexion damascène. Comme illustré sur la figure 6.14, cette méthode peut être adaptée pour utiliser uniquement deux niveaux d'interconnexion. La BE est fabriquée selon un schéma dual-damascène, permettant ainsi un contact direct avec le BEOL. Toutefois, la situation devient plus complexe

pour la TE et la SL, déposées simultanément, rendant difficile un contact par le bas. Cependant, cela est réalisable si le via est ouvert après le dépôt du TiN pour la SL. Ensuite, l'aluminium de la TE est déposé puis gravé, et la SL est également gravée lors de cette étape. Ce schéma d'intégration nécessite l'utilisation de quatre masques : Via BE, ligne BE, Via TE, ligne TE et une étape de CMP.

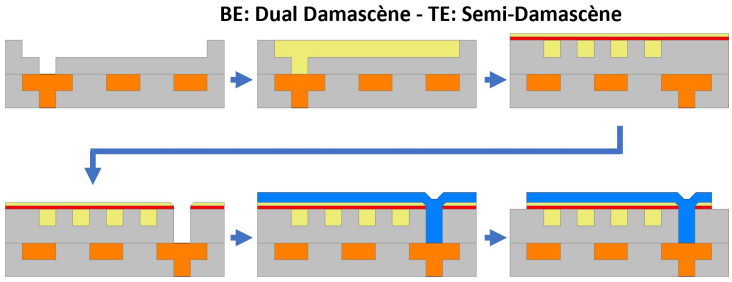


FIGURE 6.14 Procédé de fabrication pour fusionner les interconnexions BEOL et les électrodes du crossbar. La BE est fabriquée avec un procédé de dual Damascène de TiN. La SL est déposée puis, comme dans le procédé semi-Damascène présenté à la section 2.1.4, le via est ouvert dans le diélectrique, puis l'Al de la TE est déposé et gravé.

À long terme, l'utilisation du TiN pour la BE est impossible en raison de sa résistivité élevée. Le TiN est nécessaire entre la SL et la BE, et aucun autre métal ne peut être ajouté sous cette couche, car le polissage est limité à un seul matériau. Si la TE est réalisée en damascène, comme illustré dans la figure 6.15(b), elle peut être remplacée par du cuivre, puisque la SL et le TiN sont alors au fond de la tranchée. Cependant, comme le damascène n'est pas adapté pour la BE, un schéma où la BE est réalisée par soustraction et la TE par damascène pourrait être envisagé comme le montre la figure 6.15(c). Cette configu-

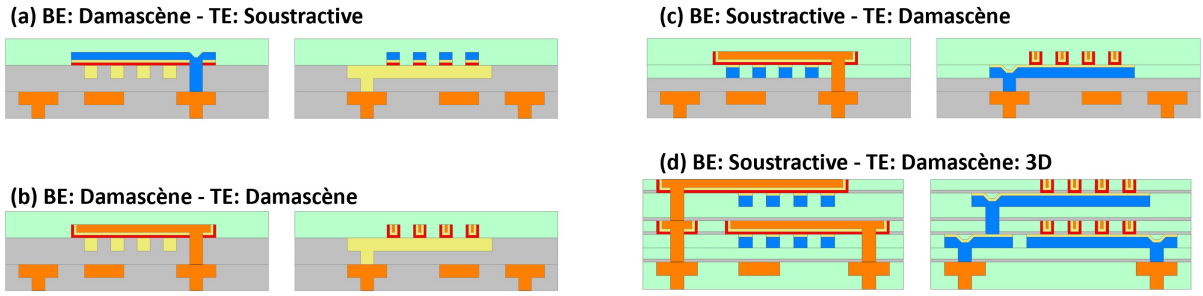


FIGURE 6.15 (a) Dual Damascène en TiN pour la BE et TE soustractive en Al. (b) Dual Damascène en TiN pour la BE et dual Damascène en Cu pour la TE, comme la SL est déposée dans le fond de la tranchée de la TE, l'utilisation d'un matériau moins résistif est possible pour la TE. (c) TE soustractive en Al et dual Damascène en Cu pour la TE, dans cette configuration il n'y a plus d'électrode de TiN résistive et l'intégration 3D-BEOL est possible (d)

ration permet une intégration 3D-BEOL sans présenter d'interconnexion trop résistive. Il est cependant nécessaire d'évaluer les impacts potentiels d'une couche active déposée dans une tranchée. L'intégration 3D-FC reste irréalisable avec ce schéma en raison de la présence nécessaire du TiN entre le Cu et la SLComme il sera montré par la suite, les seules perspectives viables pour les approches damascène impliquent l'intégration de piliers soustractifs.

6.5.2 Soustraction simple et double

Des changements similaires peuvent être mis en place pour les approches de soustraction simple et double, visant une intégration efficace dans un système 3D-BEOL, y compris avec une connectivité complète. Dans l'approche de soustraction simple, la fabrication de la BE commence par la gravure d'un via, suivie par le dépôt et la gravure de l'aluminium pour former la BE. L'oxyde est ensuite déposé puis planarisé par CMP. Pour la TE, le processus reste similaire à celui pratiqué précédemment : le via est ouvert après le dépôt de la SL. Ensuite, l'aluminium de la TE est déposé et gravé en même temps que la SL. La soustraction double simplifie encore plus le processus pour la TE. La SL est déposée en même temps que la BE. Dans ce cas, la SL n'a pas besoin d'être traitée séparément pour la TE, ce qui allège le processus général et réduit potentiellement les risques d'erreurs ou de contamination entre les couches.

Comme montré sur la figure 6.16, ces approches permettent non seulement une intégration 3D-BEOL mais également 3D-FC. L'intégration complètement connectée 3D-FC permet de réduire encore plus le nombre de masques utilisés. Pour n niveaux de réseau crossbar empilés, l'approche 3D-BEOL avec des interconnexions dual damascene nécessite $4n$ niveaux de masque et $3n$ étapes de CMP. L'approche 3D-BEOL avec des interconnexions soustractives nécessite également $4n$ niveaux de masque mais réduit le nombre d'étapes de

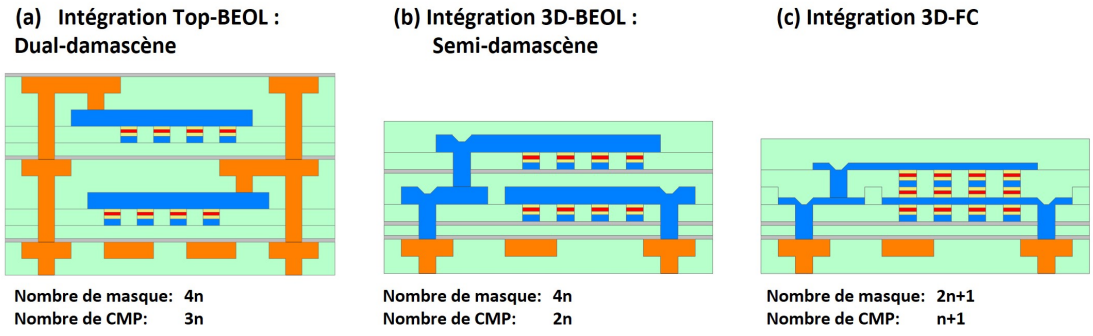


FIGURE 6.16 Schémas d'intégration 3D pour l'approche soustractive double. L'intégration complètement connectée 3D-FC est rendue possible en fusionnant les interconnexions BEOL et les électrodes pour les approches soustractives.

CMP à $2n$. Cependant, l'approche 3D-FC ne demande que $2n+1$ niveaux de masque et $n+1$ étapes de CMP. L'approche 3D-FC permet donc un gain significatif dans le nombre d'étapes de fabrication.

6.5.3 Pilier Soustractif

La dernière approche examinée est celle utilisant des piliers. Dans une intégration soustractive, le schéma d'intégration peut être adapté de façon similaire en ayant un niveau de masque et de gravure spécifique pour les piliers. Cela nécessite cinq niveaux de masques — Via BE, ligne BE, Pilier, Via TE, et ligne TE — et une étape de CMP pour polir le SiO_2 et planariser le haut des piliers. Tout comme les approches soustractives simple et double, ce schéma d'intégration est compatible avec les approches 3D-BEOL et 3D-FC, comme illustré sur la figure 6.17(a).

Bien que l'approche avec piliers requière un processus plus complexe, elle permet de combiner l'utilisation d'interconnexions en cuivre suivant un schéma damascène avec les piliers dans une approche soustractive. Comme montré sur la figure 6.17(b), la BE est fabriquée

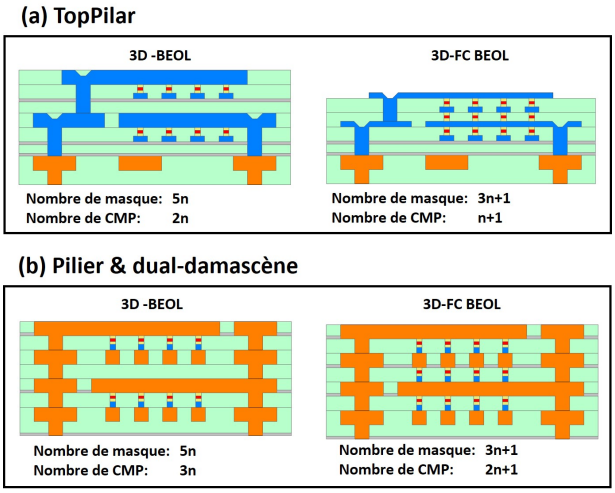


FIGURE 6.17 Intégration 3D : (a) TopPilar et (b) des piliers soustractifs avec des interconnexions dual damascene.

selon une approche dual damascène classique. Un empilement de matériaux formant le pilier est déposé puis gravé, suivi du dépôt de SiO_2 pour la planarisation. Ce schéma est également compatible avec une intégration 3D-BEOL et 3D-FC. Pour l'approche 3D-FC, cinq niveaux de masques et deux étapes de CMP sont requit, une sur le métal pour les électrodes et une sur le diélectrique. Ce dernier schéma est le plus complexe en termes d'étapes (1 étape de CMP supplémentaire), il tire parti de la maturité avancée des interconnexions en cuivre dual damascène du BEOL, présentant un fort potentiel pour un transfert industriel à long terme.

6.6 Conclusion du chapitre

Ce chapitre expose le succès de l'intégration de crossbars passifs dans le BEOL de puces CMOS au nœud de 130 nm, réalisée grâce aux procédés damascène et soustractif simple. Il a été possible de maintenir une topographie inférieure à 30 nm après l'étape de CMP pour les deux méthodes. Cette performance, utilisant des matériaux et des techniques de fabrication compatibles avec les lignes CMOS, souligne le potentiel de transfert industriel. Des améliorations visant à stabiliser le procédé de fabrication ont été proposées, notamment par la réduction de la topographie pour favoriser une intégration non seulement au-dessus, mais également entre les interconnexions du BEOL, exploitant ainsi l'empilement 3D pour accroître la densité de la mémoire intégrée.

Les schémas d'intégration élaborés dans cette thèse ont permis de discuter des modifications nécessaires pour fusionner les interconnexions métalliques reliant les mémoires (TE et BE) avec celles connectant les circuits ReRAM au BEOL ce qui permettrait de réduire le nombre d'étapes et éventuellement améliorer le rendement. Les méthodes damascène et soustractive simple requièrent la gravure du Via TE entre le dépôt de la couche de commutation et l'aluminium de la TE, tandis que l'approche soustractive double nécessite peu de modifications par rapport au procédé initial.

L'analyse met en avant les contraintes de l'approche damascène pour le BE, en particulier l'impossibilité de remplacer le TiN par un autre matériau ayant une plus faible résistivité telle que le Cu. Bien qu'une combinaison des approches BE soustractive et TE damascène soit envisageable, une intégration 3D totalement connectée est impossible. Seule l'approche avec des piliers soustractifs permet l'intégration standard d'interconnexions Cu dual damascène.

Dans l'optique d'une intégration avancée, deux méthodes apparaissent prometteuses. Premièrement, l'approche soustractive double qui isole efficacement la couche mémoire à chaque intersection et permet une intégration FC-3D, limitant le nombre d'étapes de CMP (1 par niveau) et de lithographie (2 par niveau), mais exigeant un changement du standard damascène actuellement utilisé pour les interconnexions.

Deuxièmement, l'intégration de piliers associée aux interconnexions dual damascène se distingue également. Bien que cette méthode nécessite davantage d'étapes de CMP (2 par niveau) et de lithographie (3 par niveau) que l'approche soustractive double, elle offre l'avantage de ne pas limiter la taille de la mémoire aux dimensions des interconnexions et de profiter du standard bien établi en industrie du dual damascène.

CHAPITRE 7

Conclusion

Ce projet de doctorat a pour objectif de développer un procédé de microfabrication dans le but d'intégrer des réseaux de mémoires ReRAM passifs sur le BEOL de puces CMOS fabriquées par TSMC. Des mémoires ReRAM développées par d'autres membres du groupe INPAQT ont été utilisées pour le développement et la caractérisation du procédé de fabrication. Des mémoires à base de TiOx ont été utilisées pour leur caractère analogique. L'objectif étant que ce procédé de fabrication soit au plus proche des standards industriels CMOS dans la limite des ressources disponibles. Des matériaux compatibles avec l'industrie CMOS et des techniques de fabrication typiques des technologies BEOL ont été utilisés.

La problématique introduite dans le premier chapitre et la revue détaillée de la littérature ont mis en évidence le besoin d'intégrer des systèmes intégrés utilisant des crossbars passifs pour augmenter massivement la densité d'intégration de ReRAM pour le calcul en mémoire afin de répondre aux besoins énergétiques et de performance des algorithmes d'ANN. La revue de la littérature a montré que l'approche passive est la moins mature du point de vue de la compatibilité industrielle ; de plus, peu de démonstrations intégrées au BEOL sont rapportées. Contrairement à l'approche 1T1R qui ne nécessite qu'une connexion de la mémoire avec le BEOL, l'approche passive nécessite un niveau d'interconnectivité supplémentaire entre les mémoires, référé sous le terme TE et BE dans ce manuscrit. Pour cette raison, une revue des principaux procédés de fabrication BEOL a été effectuée dans le but d'adapter ces techniques au besoin de fabrication des crossbars passifs.

La revue de la littérature des technologies BEOL met en lumière deux classes de fabrication d'interconnexion : l'approche soustractive où le métal est gravé et le diélectrique poli, et l'approche damascène où le diélectrique est gravé et le métal est poli. La première étude rapportée dans le **chapitre 3** a étudié le développement et la fabrication de 3 schémas d'intégration, un procédé de damascène avec des électrodes de TiN et deux procédés soustractifs avec des électrodes d'Al/TiN. Les résultats morphologiques montrent une topographie inférieure à 10 nm après CMP pour toutes les approches. Les analyses électriques montrent des performances relativement similaires et proches de l'état de l'art pour des mémoires de TiOx. Cependant, l'approche damascène montre des résistances élevées dues à l'électrode de TiN qui a pour conséquence d'augmenter les tensions pour former et cycler

les mémoires. Le procédé damascène est plus stable du fait d'une meilleure sélectivité de polissage entre le TiN et le SiN, contrairement à l'approche soustractive. Cependant, la possibilité de n'utiliser que du TiN comme électrode limite fortement cette approche par rapport à l'approche soustractive où l'empilement de plusieurs matériaux peut être utilisé. L'approche soustractive simple utilise une structure similaire à l'approche damascène mais avec un empilement de Al/TiN pour minimiser la résistance des électrodes. L'approche soustractive double intègre l'empilement de matériaux de la mémoire ReRAM et permet de créer un procédé de fabrication où la mémoire est isolée à chaque interconnexion.

En plus de permettre l'utilisation d'un empilement de plusieurs matériaux pour fabriquer l'électrode avec ou sans la couche de commutation, l'approche soustractive permet également d'intégrer des structures avec des formes plus complexes qu'une ligne métallique. Plusieurs structures peuvent être envisagées comme la fabrication de pointes au-dessus de la BE pour effectuer une connexion électrique sur une zone très faible ou la fabrication des piliers avec la couche mémoire sur une ligne métallique qui forme la BE. Ces structures permettent un potentiel d'optimisation supplémentaire pour le fonctionnement des mémoires. Afin d'étudier ces différentes structures de crossbar avec une méthode unique, un procédé de fabrication utilisant une exposition par faisceaux d'électrons en niveau de gris a été développé pour graver ces structures, présentées au **chapitre 4**. Afin de contrôler le transfert par gravure plasma dans un empilement de matériaux avec des taux de gravure différents, une méthode de calibration a été proposée pour effectuer le transfert en une seule étape de gravure. L'étude publiée montre que la méthode de calibration a permis la réalisation d'un crossbar avec des piliers mémoires fabriqués sur une ligne métallique et réalisée en une étape avec une précision de 10 nm dans les profondeurs de gravure.

La deuxième partie de ce manuscrit vise à utiliser les procédés de fabrication de crossbar développés dans la première partie et à les intégrer au-dessus du BEOL de puces CMOS TSMC 130 nm, conçues spécialement pour la démonstration intégrée d'un système matériel capable de réaliser les opérations VMM. La fabrication monolithique sur des puces CMOS, commandées auprès d'un fournisseur industriel, impose des défis spécifiques, notamment la capacité de s'aligner pour réaliser les étapes de lithographie. Les règles de conception ne sont pas compatibles avec les exigences des marques d'alignement standard du système d'écriture par faisceau d'électrons utilisé au 3IT. À cet égard, une étude a été menée pour développer une méthode d'alignement alternative, compatible avec les règles de conception CMOS, en collaboration avec le fabricant de la machine de lithographie. Cette étude, présentée au **Chapitre 5**, montre une précision d'alignement inférieure à 10 nm avec la méthode d'alignement développée.

Finalement, le **chapitre 6**, après une comparaison des différents procédés de fabrication développés, basée sur leur potentiel futur et leur stabilité, a démontré l'intégration dans le BEOL des puces CMOS TSMC des approches damascène et soustractive. Malgré les défis de polissage liés à la petite taille des échantillons, une topographie inférieure à 30 nm a été obtenue, montrant ainsi la faisabilité d'une intégration de crossbars passifs dans le BEOL avec des procédés de fabrication compatibles avec les procédés CMOS. Ce chapitre met en lumière le potentiel de transfert industriel grâce à l'utilisation de matériaux et techniques compatibles avec les lignes de production CMOS et propose des améliorations pour stabiliser la fabrication, incluant la réduction de la topographie pour favoriser une intégration plus avancée et l'utilisation de l'empilement 3D.

Une discussion sur des procédés de fabrication plus avancés, visant à réduire le nombre d'étapes et à permettre une intégration 3D, a également été présentée. L'analyse souligne les contraintes de l'approche damascène pour le BE, notamment l'impossibilité de remplacer le TiN résistif par un autre matériau, sauf dans le cas d'une approche utilisant des piliers mémoire. Deux méthodes principales ont été identifiées : la méthode soustractive double, qui réduit les étapes de CMP et de lithographie tout en isolant la couche mémoire, et l'intégration de piliers avec des interconnexions dual damascène, offrant une plus grande flexibilité dans le dimensionnement de la mémoire et conforme aux standards BEOL.

LISTE DES RÉFÉRENCES

[1] Chrislb. Français : Voix schématique d'un neurone artificiel avec un index j.

[2] Figure 12 : Artificial neural network-deep learning model. Library Catalog : www.researchgate.net.

[3] David Held, Sebastian Thrun, and Silvio Savarese. Deep learning for single-view instance recognition.

[4] Nandini Sengupta, Md Sahidullah, and Goutam Saha. Lung sound classification using cepstral-based statistical features. 75 :118–129.

[5] Leonardo Bottaci, Philip J Drew, John E Hartley, Matthew B Hadfield, Ridzuan Farouk, Peter WR Lee, Iain MC Macintyre, Graeme S Duthie, and John RT Monson. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. 350(9076) :469–472.

[6] Dr. N. Ganesan, Dr.K. Venkatesh, Dr. M. A. Rama, and A. Malathi Palani. Application of neural networks in diagnosing cancer disease using demographic data. 1(26) :81–97.

[7] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. 33(8) :831–838.

[8] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation : Volume 2, Shared Task Papers*, pages 131–198. Association for Computational Linguistics.

[9] Jordan French. The time traveller's CAPM. 46(2) :81–96.

[10] Wei Ren Tan, Chee Seng Chan, Hernan E. Aguirre, and Kiyoshi Tanaka. Ceci n'est pas une pipe : A deep convolutional network for fine-art paintings classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3703–3707. IEEE.

[11] Modeling the global economic impact of AI | McKinsey. Library Catalog : www.mckinsey.com.

[12] Future of artificial intelligence economic growth | accenture. Library Catalog : www.accenture.com.

[13] Alex Woodie. Deep learning has hit a wall, intel's rao says, Nov 2019.

[14] Xiaowei Xu, Yukun Ding, Sharon Xiaobo Hu, Michael Niemier, Jason Cong, Yu Hu, and Yiyu Shi. Scaling for edge inference of deep neural networks. 1(4) :216–222. Number : 4 Publisher : Nature Publishing Group.

[15] Z. Waszczyszyn. Fundamentals of artificial neural networks. In Zenon Waszczyszyn, editor, *Neural Networks in the Analysis and Design of Structures*, CISM International Centre for Mechanical Sciences, pages 1–51. Springer.

[16] Amirali Amirsoleimani, Fabien Alibart, Victor Yon, Jianxiong Xu, M. Reza Pazhouhandeh, Serge Ecoffey, Yann Beilliard, Roman Genov, and Dominique Drouin. In-memory vector-matrix multiplication in monolithic complementary metal–oxide–semiconductor-memristor integrated circuits : Design choices, challenges, and perspectives. *Advanced Intelligent Systems*, 2(11) :2000115, 2020.

[17] Ragav Venkatesan and Baoxin Li. *Convolutional Neural Networks in Visual Computing : A Concise Guide*. CRC Press, October 2017.

[18] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1) :1929–1958, 2014.

[19] M. A. Talib, S. Majzoub, Q. Nasir, and Dina Jamal. A systematic literature review on hardware implementation of artificial intelligence algorithms. *The Journal of Supercomputing*, 77 :1897 – 1938, 2020.

[20] Kumar Chellapilla, Sidd Puri, and Patrice Simard. High Performance Convolutional Neural Networks for Document Processing. In Guy Lorette, editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France), October 2006. Université de Rennes 1, Suvisoft. http ://www.suvisoft.com.

[21] Michael K. Gschwind, Valentina Salapura, and O. Maischberger. A generic building block for hopfield neural networks with on-chip learning. *1996 IEEE International Symposium on Circuits and Systems. Circuits and Systems Connecting the World. ISCAS 96*, Supplement :49–52, 1996.

[22] N. Jouppi, C. Young, Nishant Patil, and David A. Patterson. Motivation for and evaluation of the first tensor processing unit. *IEEE Micro*, 38 :10–19, 2018.

[23] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net : Imagenet classification using binary convolutional neural networks, 2016.

[24] Jun-Seok Park, Changsoo Park, S. Kwon, Taeho Jeon, Yesung Kang, Heonsoo Lee, Dongwoo Lee, James Kim, Hyeong-Seok Kim, YoungJong Lee, Sangkyu Park, Min-Seong Kim, Sanghyuck Ha, Jihoon Bang, Jinpyo Park, Sukhwan Lim, and Inyup Kang. A multi-mode 8k-mac hw-utilization-aware neural processing unit with a unified multi-precision datapath in 4-nm flagship mobile soc. *IEEE Journal of Solid-State Circuits*, 58 :189–202, 2023.

[25] Chung-Bin Wu, Yu-Cheng Hsueh, Ching-Shun Wang, and Yen-Chih Lai. High throughput hardware implementation for deep learning ai accelerator. *2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, pages 1–2, 2019.

[26] Gagandeep Singh, Lorenzo Chelini, Stefano Corda, Ahsan Javed Awan, S. Stuijk, Roel Jordans, H. Corporaal, and A. Boonstra. Near-memory computing : Past, present, and future. *Microprocess. Microsystems*, 71, 2019.

[27] Sanghoon Kang, Gwangtae Park, Sangjin Kim, Soyeon Kim, Donghyeon Han, and H. Yoo. An overview of sparsity exploitation in cnns for on-device intelligence with software-hardware cross-layer optimizations. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 11 :634–648, 2021.

[28] Swagath Venkataramani, V. Srinivasan, W. Wang, Sanchari Sen, Jintao Zhang, A. Agrawal, Monodeep Kar, Shubham Jain, Alberto Mannari, H. Tran, Yulong Li,

Eri Ogawa, K. Ishizaki, H. Inoue, M. Schaal, M. Serrano, Jungwook Choi, Xiao Sun, Naigang Wang, Chia-Yu Chen, Allison Allain, J. Bonanno, N. Cao, Robert Casatuta, Matthew Cohen, B. Fleischer, Michael Guillorn, Howard Haynie, Jinwook Jung, Mingu Kang, Kyu-Hyoun Kim, S. Koswatta, Sae Kyu Lee, Martin Lutz, S. Mueller, Jinwook Oh, Ashish Ranjan, Z. Ren, Scot Rider, Kerstin Schelm, M. Scheuermann, J. Silberman, Jie quan Yang, V. Zalani, Xin Zhang, Ching Zhou, M. Ziegler, Vinay Shah, Moriyoshi Ohara, P. Lu, B. Curran, Sunil Shukla, Leland Chang, and K. Gopalakrishnan. Rapid : Ai accelerator for ultra-low precision training and inference. *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pages 153–166, 2021.

[29] Abu Sebastian, Manuel Le Gallo, Riduan Khaddam-Aljameh, and Evangelos Eleftheriou. Memory devices and applications for in-memory computing. *Nature Nanotechnology*, 15(7) :529–544, March 2020.

[30] L. Ceze, J. Hasler, K. K. Likharev, J. . Seo, T. Sherwood, D. Strukov, Y. Xie, and S. Yu. Nanoelectronic neurocomputing : Status and prospects. In *2016 74th Annual Device Research Conference (DRC)*, pages 1–2, 2016.

[31] Mohammed A. Zidan, John Paul Strachan, and Wei D. Lu. The future of electronics based on memristive systems. 1(1) :22–29.

[32] Alex Pappachen James. A hybrid memristor–CMOS chip for AI. 2(7) :268–269.

[33] E. Chicca and G. Indiveri. A recipe for creating ideal hybrid memristive-CMOS neuromorphic processing systems. 116(12) :120501.

[34] Soo Gil Kim, Tae Jung Ha, Seonghyun Kim, Jae Yeon Lee, Kyung Wan Kim, Jung Ho Shin, Yong Taek Park, Suk Pyo Song, Beom Yong Kim, Wan Gee Kim, Jong Chul Lee, Hyun Sun Lee, Jong Ho Song, Eung Rim Hwang, Sang Hoon Cho, Ja Chun Ku, Jong Il Kim, Kyu Sung Kim, Jong Hee Yoo, Hyo Jin Kim, Hoe Gwon Jung, Kee Jeung Lee, Suock Chung, Jong Ho Kang, Jung Hoon Lee, Hyeong Soo Kim, Sung Joo Hong, Gary Gibson, and Yoocharn Jeon. Improvement of characteristics of nbo2 selector and full integration of 4f2 2x-nm tech 1s1r reram. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 10.3.1–10.3.4, 2015.

[35] M. Barlas, A. Grossi, L. Grenouillet, E. Vianello, E. Nolot, N. Vaxelaire, P. Blaise, B. Traoré, J. Coignus, F. Perrin, R. Crochemore, F. Mazen, L. Lachal, S. Pauliac, C. Pellissier, S. Bernasconi, S. Chevalliez, J. F. Nodin, L. Perniola, and E. Nowak. Improvement of hfo2 based rram array performances by local si implantation. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 14.6.1–14.6.4, 2017.

[36] Giuk Kim, Hunbeom Shin, Taehyong Eom, Minhyun Jung, Taeho Kim, Sangho Lee, Minki Kim, Yeongseok Jeong, Ji-Sung Kim, Kab-Jin Nam, Bong Jin Kuh, and Sanghun Jeon. Design guidelines of thermally stable hafnia ferroelectrics for the fabrication of 3d memory devices. In *2022 International Electron Devices Meeting (IEDM)*, pages 5.4.1–5.4.4, 2022.

[37] J.-H. Park, J. H. Kim, J. M. Kim, J. Kim, D. Apalkov, A. Okada, H. Sato, J. H. Jeong, Y. J. Cho, U. Pi, Y. Kim, Y. S. Park, K. M. Song, K. Kim, D.-E. Jeong, D. S. Kim, C. Kim, I. Kim, S. H. Han, K. Lee, J. H. Lee, Y. J. Song, G. H. Koh, B. J. Kuh, J. M. Lee, and J. H. Song. Highly reliable stt-mram adopting advanced

mtjs with controlled domain wall pinning. In *2022 International Electron Devices Meeting (IEDM)*, pages 10.6.1–10.6.4, 2022.

[38] N. Gong, M.J. Rasch, S.-C. Seo, A. Gasasira, P. Solomon, V. Bragaglia, S. Consiglio, H. Higuchi, C. Park, K. Brew, P. Jamison, C. Catano, I. Saraf, F.F. Athena, C. Silvestre, X. Liu, B. Khan, N. Jain, S. Mcdermott, R. Johnson, I. Estrada-Raygoza, J. Li, T. Gokmen, N. Li, R. Pujari, F. Carta, H. Miyazoe, M.M. Frank, D. Koty, Q. Yang, R. Clark, K. Tapily, C. Wajda, A. Mosden, J. Shearer, A. Metz, S. Teehan, N. Saulnier, B. J. Offrein, T. Tsunomura, G. Leusink, V. Narayanan, and T. Ando. Deep learning acceleration in 14nm cmos compatible reram array : device, material and algorithm co-optimization. In *2022 International Electron Devices Meeting (IEDM)*, pages 33.7.1–33.7.4, 2022.

[39] M. Stanisavljevic, H. Pozidis, A. Athmanathan, N. Papandreou, T. Mittelholzer, and E. Eleftheriou. Demonstration of reliable triple-level-cell (tlc) phase-change memory. In *2016 IEEE 8th International Memory Workshop (IMW)*, pages 1–4, 2016.

[40] H. Kim, M. R. Mahmoodi, H. Nili, and D. B. Strukov. 4K-memristor analog-grade passive crossbar circuit. *Nature Communications*, 12(1) :5198, 2021.

[41] Masoud Zabihi, Salonik Resch, Hüsrev Cilasun, Zamshed I. Chowdhury, Zhengyang Zhao, Ulya R. Karpuzcu, Jian-Ping Wang, and Sachin S. Sapatnekar. Exploring the feasibility of using 3-d xpoint as an in-memory computing accelerator. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, 7(2) :88–96, 2021.

[42] R. Dawant, Matthieu Gaudreau, Marc-Antoine Roy, Pierre-Antoine Mouny, Matthieu Valdenaire, Pierre Gliech, Javier Arias Zapata, Fabien Alibart, Dominique Drouin, and Serge Ecoffey. Damascene versus subtractive line cmp process for resistive memory crossbars beol integration. *Micro and Nano Engineering*, page 100251, 2024.

[43] R. Dawant, S. Ecoffey, and D. Drouin. Multiple material stack grayscale patterning using electron-beam lithography and a single plasma etching step. *Journal of Vacuum Science & Technology B*, 40(6) :062603, 11 2022.

[44] Raphaël Dawant, Robyn Seils, Serge Ecoffey, Rainer. Schmid, and Dominique Drouin. Hybrid cross correlation and line-scan alignment strategy for cmos chips electron-beam lithography processing. *Journal of Vacuum Science amp ; Technology B*, 40(1), December 2021.

[45] R.H. Havemann and J.A. Hutchby. High-performance interconnects : an integration overview. *Proceedings of the IEEE*, 89(5) :586–601, 2001.

[46] Yi-Lung Cheng and Chih-Yen Lee. *Porous Low-Dielectric-Constant Material for Semiconductor Microelectronics*. IntechOpen, August 2020.

[47] T. Saito, T. Imai, J. Noguchi, M. Kubo, Y. Ito, S. Omori, N. Ohashi, T. Tamaru, and H. Yamaguchi. A novel copper interconnection technology using self aligned metal capping method. In *Proceedings of the IEEE 2001 International Interconnect Technology Conference (Cat. No.01EX461)*. IEEE, 2001.

[48] J.R. Lloyd, C.E. Murray, S. Ponoth, S. Cohen, and E. Liniger. The effect of cu diffusion on the tddb behavior in a low-k interlevel dielectrics. *Microelectronics Reliability*, 46(9–11) :1643–1647, September 2006.

[49] Sung-Kwon Lee Sung-Kwon Lee, Sung-Soon Chun Sung-Soon Chun, ChanYong Hwang ChanYong Hwang, and Won-Jong Lee Won-Jong Lee. Reactive ion etching mechanism of copper film in chlorine-based electron cyclotron resonance plasma. *Japanese Journal of Applied Physics*, 36(1R) :50, jan 1997.

[50] J. Kriz, C. Angelkort, M. Czekalla, S. Huth, D. Meinhold, A. Pohl, S. Schulte, A. Thamm, and S. Wallace. Overview of dual damascene integration schemes in cu beol integration. *Microelectronic Engineering*, 85(10) :2128–2132, October 2008.

[51] Hyung-Woo Kim. Recent trends in copper metallization. *Electronics*, 11(18) :2914, September 2022.

[52] J. Gambino, A. Stamper, T. McDevitt, V. McGahay, S. Luce, T. Pricer, et al. Integration of copper with low-k dielectrics for 0.13m technology. In *Proceedings of the IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits*, pages 111–117, 2002.

[53] S.W. Lim, Y. Shimogaki, Y. Nakano, K. Tada, and H. Komiyama. Changes in the orientational polarization and structure of silicon dioxide film by fluorine addition. *J. Electrochem. Soc.*, 146 :4196–4202, 1999.

[54] A. Grill. Low and ultralow dielectric constant films prepared by plasma-enhanced chemical vapor deposition. In M. Baklonov, M. Green, and K. Maex, editors, *Dielectric Films for Advanced Microelectronics*, pages 1–32. Wiley, New York, n.d.

[55] S. Gates, A. Grill, C. Dimitrakopoulos, D. Restaino, M. Lane, V. Patel, et al. A porous sicoh dielectric with k = 2.4 for high performance beol interconnects. In *AMC Proceedings*, pages 351–357, 2006.

[56] Y. Travaly, J. van Aelst, V. Truffert, P. Verdonck, T. Dupont, E. Camerotto, O. Richard, H. Bender, C. Kroes, D. de Roest, G. Vereecke, M. Claes, Q. T. Le, E. Kesters, M. van Cauwenberghe, J. Beynet, S. Kaneko, H. Struyf, M. Baklanov, K. Matsushita, N. Kobayashi, H. Sprey, and G. Beyer. Key factors to sustain the extension of a mhm-based integration scheme to medium and high porosity pecvd low-k materials. In *2008 International Interconnect Technology Conference*, pages 52–54, 2008.

[57] Jeff Gambino, Fen Chen, and John He. Copper interconnect technology for the 32 nm node and beyond. In *2009 IEEE Custom Integrated Circuits Conference*, pages 141–148, 2009.

[58] J. Gambino, F. Chen, and J. He. Copper interconnect technology for the 32nm node and beyond. In *IEEE Custom Integrated Circuits Conference Proceedings*, pages 141–148, Warrendale, PA, 2009.

[59] S. Y. Chang, C. C. Wan, and Y. Y. Wang. Selectivity enhancement of electroless co deposition for cu capping process via spontaneous diazonium ion reduction. *Electrochemical and Solid-State Letters*, 10(5) :D43, feb 2007.

[60] D. Edelstein, J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, et al. Full copper wiring in a sub-0.25m cmos ulsi technology. In *IEEE International Electron Device Meeting Proceedings*, pages 773–776, 1997.

[61] C. Witt, K.B. Yeap, A. Leśniewska, D. Wan, N. Jordan, I. Ciofi, C. Wu, and Z. Tokei. Testing the limits of tan barrier scaling. In *2018 IEEE International Interconnect Technology Conference (IITC)*, pages 54–56, 2018.

[62] Davide Tierno, M. Hosseini, M. van der Veen, A. Dangol, K. Croes, S. Demuynck, Zs. Tőkei, E.D. Litta, and N. Horiguchi. Reliability of barrierless pvd mo. In *2021 IEEE International Interconnect Technology Conference (IITC)*, pages 1–3, 2021.

[63] Dooho Choi and Katayun Barmak. On the potential of tungsten as next-generation semiconductor interconnects. *Electronic Materials Letters*, 13(5) :449–456, September 2017.

[64] T. Nogami, R. Patlolla, J. Kelly, B. Briggs, H. Huang, J. Demarest, J. Li, R. Hengstebeck, X. Zhang, G. Lian, B. Peethala, P. Bhosale, J. Maniscalco, H. Shobha, S. Nguyen, P. McLaughlin, T. Standaert, D. Canaperi, D. Edelstein, and V. Paruchuri. Cobalt/copper composite interconnects for line resistance reduction in both fine and wide lines. In *2017 IEEE International Interconnect Technology Conference (IITC)*, pages 1–3, 2017.

[65] Liang Gong Wen, Philippe Roussel, Olalla Varela Pedreira, Basoene Briggs, Benjamin Groven, Shibesh Dutta, Mihaela I. Popovici, Nancy Heylen, Ivan Ciofi, Kris Vanstreels, Frederik W. Østerberg, Ole Hansen, Dirch H. Petersen, Karl Opsomer, Christophe Detavernie, Christopher J. Wilson, Sven Van Elshocht, Kristof Croes, Jürgen Bömmels, Zsolt Tőkei, and Christoph Adelmann. Atomic layer deposition of ruthenium with tin interface for sub-10 nm advanced interconnects beyond copper. *ACS Applied Materials amp ; Interfaces*, 8(39) :26119–26125, September 2016.

[66] Xunyuan Zhang, Huai Huang, Raghuveer Patlolla, Wei Wang, Frank W. Mont, Juntao Li, Chao-Kun Hu, Eric G. Liniger, Paul S. McLaughlin, Cathy Labelle, E. Todd Ryan, Donald Canaperi, Terry Spooner, Griselda Bonilla, and Daniel Edelstein. Ruthenium interconnect resistivity and reliability at 48 nm pitch. In *2016 IEEE International Interconnect Technology Conference / Advanced Metallization Conference (IITC/AMC)*, pages 31–33, 2016.

[67] Shibesh Dutta, Shreya Kundu, Anshul Gupta, Geraldine Jamieson, Juan Fernando Gomez Granados, Jürgen Bömmels, Christopher J. Wilson, Zsolt Tőkei, and Christoph Adelmann. Highly scaled ruthenium interconnects. *IEEE Electron Device Letters*, 38(7) :949–951, 2017.

[68] Marleen H. van der Veen, N. Heyler, O. Varela Pedreira, I. Ciofi, S. Decoster, V. Vega Gonzalez, N. Jourdan, H. Struyf, K. Croes, C. J. Wilson, and Zs. Tőkei. Damascene benchmark of ru, co and cu in scaled dimensions. In *2018 IEEE International Interconnect Technology Conference (IITC)*, pages 172–174, 2018.

[69] K. Motoyama, N. Lanzillo, S. Mukesh, B. Peethala, T. Spooner, D. Edelstein, and K. Choi. Metal-induced line width variability challenge and mitigation strategy in advanced post-cu interconnects. In *2022 IEEE International Interconnect Technology Conference (IITC)*, pages 55–57, 2022.

[70] O. Varela Pedreira, M. Stucchi, A. Gupta, V. Vega Gonzalez, M. van der Veen, S. Lariviere, C.J. Wilson, Zs Tőkei, and K. Croes imec. Metal reliability mechanisms in ruthenium interconnects. In *2020 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–7, 2020.

[71] Chris Penny. Challenges and innovations for advanced beol scaling at the 1nm node and beyond. Short course presented at the VLSI 2023 Conference, 2023.

[72] Danny Wan, Sara Paolillo, Nouredine Rassoul, Bogumila Kutrzeba Kotowska, Victor Blanco, Christoph Adelmann, Frederic Lazzarino, Monique Ercken, Gayle Murdoch, Jürgen Bömmels, Christopher J. Wilson, and Zsolt Tökei. Subtractive etch of ruthenium for sub-5nm interconnect. In *2018 IEEE International Interconnect Technology Conference (IITC)*, pages 10–12, 2018.

[73] C. Penny, K. Motoyama, S. Ghosh, T. Bae, N. Lanzillo, S. Sieg, C. Park, L. Zou, H. Lee, D. Metzler, J. Lee, S. Cho, M. Shoudy, S. Nguyen, A. Simon, K. Park, L. Clevenger, B. Anderson, C. Child, T. Yamashita, J. Arnold, T. Wu, T. Spooner, K. Choi, K-I. Seo, and D. Guo. Subtractive ru interconnect enabled by novel patterning solution for euv double patterning and topvia with embedded airgap integration for post cu interconnect scaling. In *2022 International Electron Devices Meeting (IEDM)*, pages 12.1.1–12.1.4, 2022.

[74] Réseau de neurones artificiels. Page Version ID : 169602792.

[75] Fabien Alibart, Ligang Gao, Brian D Hoskins, and Dmitri B Strukov. High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. *Nanotechnology*, 23(7) :075201, jan 2012.

[76] Cheng-Xin Xue, Tsung-Yuan Huang, Je-Syu Liu, Chang Ting-Wei, Hui-Yao Kao, Jing-Hong Wang, Ta-Wei Liu, Shih-Ying Wei, Sheng-Po Huang, Wei-Chen Wei, Yi-Ren Chen, Tzu-Hsiang Hsu, Yen-Kai Chen, Yun-Chen Lo, Tai-Hsing Wen, Chung-Chuan Lo, Ren-Shuo Liu, Chih-Cheng Hsieh, Kea-Tiong Tang, and Meng-Fan Chang. 15.4 a 22nm 2mb reram compute-in-memory macro with 121-28tops/w for multibit mac computing for tiny ai edge devices. pages 244–246, 02 2020.

[77] Peng Yao, Huaqiang Wu, Bin Gao, Jianshi Tang, Qingtian Zhang, Wenqiang Zhang, J. Joshua Yang, and He Qian. Fully hardware-implemented memristor convolutional neural network. *Nature*, 577(7792) :641–646, January 2020.

[78] Fuxi Cai, Justin M. Correll, Seung Hwan Lee, Yong Lim, Vishishtha Bothra, Zhengya Zhang, Michael P. Flynn, and Wei D. Lu. A fully integrated reprogrammable memristor–cmos system for efficient multiply–accumulate operations. *Nature Electronics*, 2(7) :290–299, July 2019.

[79] Tommaso Zanotti, Cristian Zambelli, Francesco Maria Puglisi, Valerio Milo, Eduardo Pérez, Mamathamba K. Mahadevaiah, Oscar G. Ossorio, Christian Wenger, Paolo Pavan, Piero Olivo, and Daniele Ielmini. Reliability of logic-in-memory circuits in resistive memory arrays. *IEEE Transactions on Electron Devices*, 67(11) :4611–4615, 2020.

[80] M. A. Zidan, H. Omran, R. Naous, A. Sultan, H. a. H. Fahmy, W. D. Lu, and K. N. Salama. Single-readout high-density memristor crossbar. 6(1) :1–9. Number : 1 Publisher : Nature Publishing Group.

[81] Mohammed Affan Zidan, Hossam Aly Hassan Fahmy, Muhammad Mustafa Hussain, and Khaled Nabil Salama. Memristor-based memory : The sneak paths problem and solutions. 44(2) :176–183.

[82] Stefan Slesazeck and Thomas Mikolajick. Nanoscale resistive switching memory devices : a review. 30(35) :352003. Publisher : IOP Publishing.

[83] J. Joshua Yang, Dmitri B. Strukov, and Duncan R. Stewart. Memristive devices for computing. 8(1) :13–24.

[84] C. Yakopcic, M. McLean, T.M. Taha, R. Hasan, and D. Palmer. Memristor-based neuron circuit and method for applying learning algorithm in SPICE. 50(7) :492–494.

[85] Yibo Li, Zhongrui Wang, Rivu Midya, Qiangfei Xia, and J Joshua Yang. Review of memristor devices in neuromorphic computing : materials sciences and device challenges. 51(50) :503002.

[86] Daniele Ielmini and Stefano Ambrogio. Emerging neuromorphic devices. 31(9) :092001.

[87] Shuang Pi, Peng Lin, and Qiangfei Xia. Cross point arrays of 8nm×8nm memristive devices fabricated with nanoimprint lithography. 31(6) :06FA02. Publisher : American Vacuum Society.

[88] I.G. Baek, M.S. Lee, S. Seo, M.J. Lee, D.H. Seo, D.-S. Suh, J.C. Park, S.O. Park, H.S. Kim, I.K. Yoo, U.-In. Chung, and J.T. Moon. Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses. In *IEDM Technical Digest. IEEE International Electron Devices Meeting, 2004.*, pages 587–590. ISSN : null.

[89] Cheol Seong Hwang. Prospective of semiconductor memory devices : from memory system to materials. 1(6) :1400056. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/aelm.201400056.

[90] Duygu Kuzum, Rakesh G. D. Jeyasingh, Byoungil Lee, and H.-S. Philip Wong. Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. 12(5) :2179–2186.

[91] Bryan L. Jackson, Bipin Rajendran, Gregory S. Corrado, Matthew Breitwisch, Geoffrey W. Burr, Roger Cheek, Kailash Gopalakrishnan, Simone Raoux, Charles T. Rettner, Alvaro Padilla, Alex G. Schrott, Rohit S. Shenoy, Bülent N. Kurdi, Chung H. Lam, and Dharmendra S. Modha. Nanoscale electronic synapses using phase change devices. 9(2) :12 :1–12 :20.

[92] Geoffrey W. Burr, Robert M. Shelby, Severin Sidler, Carmelo di Nolfo, Junwoo Jang, Irem Boybat, Rohit S. Shenoy, Pritish Narayanan, Kumar Virwani, Emanuele U. Giacometti, Bülent N. Kurdi, and Hyunsang Hwang. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. 62(11) :3498–3507.

[93] A. Pirovano, A.L. Lacaita, F. Pellizzer, S.A. Kostylev, A. Benvenuti, and R. Bez. Low-field amorphous state resistance and threshold voltage drift in chalcogenide materials. 51(5) :714–719.

[94] Navnidhi K. Upadhyay, Hao Jiang, Zhongrui Wang, Shiva Asapu, Qiangfei Xia, and J. Joshua Yang. Emerging memory devices for neuromorphic computing. 4(4) :1800589.

[95] Steven Lequeux, Joao Sampaio, Vincent Cros, Kay Yakushiji, Akio Fukushima, Rie Matsumoto, Hitoshi Kubota, Shinji Yuasa, and Julie Grollier. A magnetic synapse : multilevel spin-torque memristor with perpendicular anisotropy. 6(1) :1–7. Number : 1 Publisher : Nature Publishing Group.

[96] Syed Ghazi Sarwat. Materials science and engineering of phase change random access memory. 33(16) :1890–1906. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/02670836.2017.1341723.

[97] Fuxi Cai, Justin M. Correll, Seung Hwan Lee, Yong Lim, Vishishtha Bothra, Zhengya Zhang, Michael P. Flynn, and Wei D. Lu. A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations. 2(7) :290–299.

[98] Adrian M. Caulfield, Joel Coburn, Todor Mollov, Arup De, Ameen Akel, Jiahua He, Arun Jagatheesan, Rajesh K. Gupta, Allan Snavely, and Steven Swanson. Understanding the impact of emerging non-volatile memories on high-performance, IO-intensive computing. In *2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11. IEEE.

[99] M H Tang, Z P Wang, J C Li, Z Q Zeng, X L Xu, G Y Wang, L B Zhang, Y G Xiao, S B Yang, B Jiang, and J He. Bipolar and unipolar resistive switching behaviors of sol–gel-derived srtio3 thin films with different compliance currents. *Semiconductor Science and Technology*, 26(7) :075019, apr 2011.

[100] Kyung Ho Sun and Yoon Young Kim. Design of magnetoelectric multiferroic heterostructures by topology optimization. *Journal of Physics D : Applied Physics*, 44(18) :185003, apr 2011.

[101] L. Goux, J. G. Lisoni, M. Jurczak, D. J. Wouters, L. Courtade, and Ch. Muller. Coexistence of the bipolar and unipolar resistive-switching modes in NiO cells made by thermal oxidation of Ni layers. *Journal of Applied Physics*, 107(2) :024512, 01 2010.

[102] R. Yasuhara, K. Fujiwara, K. Horiba, H. Kumigashira, M. Kotsugi, M. Oshima, and H. Takagi. Inhomogeneous chemical states in resistance-switching devices with a planar-type Pt/CuO/Pt structure. *Applied Physics Letters*, 95(1) :012110, 07 2009.

[103] Pai-Chun Chang and Jia Grace Lu. Temperature dependent conduction and UV induced metal-to-insulator transition in ZnO nanowires. *Applied Physics Letters*, 92(21) :212113, 05 2008.

[104] Xiao-Yu Zhang, Peng Wang, Su Sheng, Yungui Ma, Feng Xu, and C K Ong. A novel structure for dc bias on varactors in composite right/left-handed transmission lines phase shifter using ba0.25sr0.75tio3 thin film. *Journal of Physics D : Applied Physics*, 42(17) :175103, aug 2009.

[105] C. H. Lien, Y. S. Chen, H. Y. Lee, P. S. Chen, F. T. Chen, and M.-J. Tsai. The highly scalable and reliable hafnium oxide reram and its future challenges. In *2010 10th IEEE International Conference on Solid-State and Integrated Circuit Technology*. IEEE, November 2010.

[106] Feng Miao, John Paul Strachan, J. Joshua Yang, Min-Xian Zhang, Ilan Goldfarb, Antonio C. Torrezan, Peter Eschbach, Ronald D. Kelley, Gilberto Medeiros-Ribeiro, and R. Stanley Williams. Anatomy of a nanoscale conduction channel reveals the mechanism of a high-performance memristor. *Advanced Materials*, 23(47) :5633–5640, November 2011.

[107] Yuanjun Yang, Meng Meng Yang, Z. L. Luo, Haoliang Huang, Haibo Wang, J. Bao, Chuansheng Hu, Guoqiang Pan, Yiping Yao, Yukuai Liu, X. G. Li, Sen Zhang, Y. G.

Zhao, and C. Gao. Large anisotropic remnant magnetization tunability in (011)-La2/3Sr1/3MnO3/0.7Pb(Mg2/3Nb1/3)O3-0.3PbTiO3 multiferroic epitaxial heterostructures. *Applied Physics Letters*, 100(4) :043506, 01 2012.

[108] Yuanjun Yang, Meng Meng Yang, Z. L. Luo, Haoliang Huang, Haibo Wang, J. Bao, Chuansheng Hu, Guoqiang Pan, Yiping Yao, Yukuai Liu, X. G. Li, Sen Zhang, Y. G. Zhao, and C. Gao. Large anisotropic remnant magnetization tunability in (011)-La2/3Sr1/3MnO3/0.7Pb(Mg2/3Nb1/3)O3-0.3PbTiO3 multiferroic epitaxial heterostructures. *Applied Physics Letters*, 100(4) :043506, 01 2012.

[109] Qi Liu, Bin Gao, Peng Yao, Dong Wu, Junren Chen, Yachuan Pang, Wenqiang Zhang, Yan Liao, Cheng-Xin Xue, Wei-Hao Chen, Jianshi Tang, Yu Wang, Meng-Fan Chang, He Qian, and Huaqiang Wu. 33.2 a fully integrated analog reram based 78.4tops/w compute-in-memory chip with fully parallel mac computing. In *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, pages 500–502, 2020.

[110] B. Q. Le, A. Grossi, E. Vianello, T. Wu, G. Lama, E. Beigne, H. . P. Wong, and S. Mitra. Resistive ram with multiple bits per cell : Array-level demonstration of 3 bits per cell. *IEEE Transactions on Electron Devices*, 66(1) :641–646, 2019.

[111] Fang Su, Wei-Hao Chen, Lixue Xia, Chieh-Pu Lo, Tianqi Tang, Zhibo Wang, Kuo-Hsiang Hsu, Ming Cheng, Jun-Yi Li, Yuan Xie, Yu Wang, Meng-Fan Chang, Huazhong Yang, and Yongpan Liu. A 462gops/j rram-based nonvolatile intelligent processor for energy harvesting ioe system featuring nonvolatile logics and processing-in-memory. In *2017 Symposium on VLSI Technology*, pages T260–T261, 2017.

[112] M. Bocquet, T. Hirztlin, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 20–26, Piscataway, NJ, 2018. IEEE.

[113] Hanwool Yeon, Peng Lin, Chanyeol Choi, Scott H. Tan, Yongmo Park, Doyoon Lee, Jaeyong Lee, Feng Xu, Bin Gao, Huaqiang Wu, He Qian, Yifan Nie, Seyoung Kim, and Jeehwan Kim. Alloying conducting channels for reliable neuromorphic computing. *Nature Nanotechnology*, 15(7) :574–579, June 2020.

[114] Kuk-Hwan Kim, Siddharth Gaba, Dana Wheeler, Jose M. Cruz-Albrecht, Tahir Hussain, Narayan Srinivasa, and Wei Lu. A functional hybrid memristor crossbar-array/cmos system for data storage and neuromorphic applications. *Nano Letters*, 12(1) :389–395, December 2011.

[115] A. Grossi, E. Nowak, C. Zambelli, C. Pellissier, S. Bernasconi, G. Cibrario, K. El Hajjam, R. Crochemore, J.F. Nodin, P. Olivo, and L. Perniola. Fundamental variability limits of filament-based rram. In *2016 IEEE International Electron Devices Meeting (IEDM)*, pages 4.7.1–4.7.4, 2016.

[116] Wei-Hao Chen, Chunmeng Dou, Kai-Xiang Li, Wei-Yu Lin, Pin-Yi Li, Jian-Hao Huang, Jing-Hong Wang, Wei-Chen Wei, Cheng-Xin Xue, Yen-Cheng Chiu, Ya-Chin King, Chorng-Jung Lin, Ren-Shuo Liu, Chih-Cheng Hsieh, Kea-Tiong Tang, J. Joshua Yang, Mon-Shu Ho, and Meng-Fan Chang. Cmos-integrated memristive non-volatile computing-in-memory for ai edge processors. *Nature Electronics*, 2(9) :420–428, August 2019.

[117] Fernando Aguirre, Abu Sebastian, Manuel Le Gallo, Wenhao Song, Tong Wang, J. Joshua Yang, Wei Lu, Meng-Fan Chang, Daniele Ielmini, Yuchao Yang, Adnan

Mehonic, Anthony Kenyon, Marco A. Villena, Juan B. Roldán, Yuting Wu, Hung-Hsi Hsu, Nagarajan Raghavan, Jordi Suñé, Enrique Miranda, Ahmed Eltawil, Gianluca Setti, Kamilya Smagulova, Khaled N. Salama, Olga Krestinskaya, Xiaobing Yan, Kah-Wee Ang, Samarth Jain, Sifan Li, Osamah Alharbi, Sebastian Pazos, and Mario Lanza. Hardware implementation of memristor-based artificial neural networks. *Nature Communications*, 15(1), March 2024.

[118] Patrick M. Sheridan, Fuxi Cai, Chao Du, Wen Ma, Zhengya Zhang, and Wei D. Lu. Sparse coding with memristor networks. *Nature Nanotechnology*, 12(8) :784–789, May 2017.

[119] Jong Hoon Shin, Yeon Joo Jeong, Mohammed A. Zidan, Qiwen Wang, and Wei D. Lu. Hardware acceleration of simulated annealing of spin glass by rram crossbar array. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 3.3.1–3.3.4, 2018.

[120] Shinhyun Choi, Jong Hoon Shin, Jihang Lee, Patrick Sheridan, and Wei D. Lu. Experimental demonstration of feature extraction and dimensionality reduction using memristor networks. *Nano Letters*, 17(5) :3113–3118, May 2017.

[121] YeonJoo Jeong, Jihang Lee, John Moon, Jong Hoon Shin, and Wei D. Lu. K-means data clustering with memristor networks. *Nano Letters*, 18(7) :4447–4453, June 2018.

[122] G. C. Adam, B. D. Hoskins, M. Prezioso, F. Merrikh-Bayat, B. Chakrabarti, and D. B. Strukov. 3-d memristor crossbars for analog and neuromorphic computing applications. *IEEE Transactions on Electron Devices*, 64(1) :312–318, 2017.

[123] M. Prezioso, I. Kataeva, F. Merrikh-Bayat, B. Hoskins, G. Adam, T. Sota, K. Likharev, and D. Strukov. Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer pt/al2o3/tio2-x/pt memristors. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 17.4.1–17.4.4, 2015.

[124] F. Merrikh Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, and D. Strukov. Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits. *Nature Communications*, 9(1), June 2018.

[125] Peng Lin, Can Li, Zhongrui Wang, Yunning Li, Hao Jiang, Wenhao Song, Mingyi Rao, Ye Zhuo, Navnidhi K. Upadhyay, Mark Barnell, Qing Wu, J. Joshua Yang, and Qiangfei Xia. Three-dimensional memristor circuits as complex neural networks. *Nature Electronics*, 3(4) :225–232, April 2020.

[126] Peng Yao, Huaqiang Wu, Bin Gao, Sukru Burc Eryilmaz, Xueyao Huang, Wenqiang Zhang, Qingtian Zhang, Ning Deng, Luping Shi, H.-S. Philip Wong, and He Qian. Face classification using electronic synapses. *Nature Communications*, 8(1), May 2017.

[127] Can Li, Miao Hu, Yunning Li, Hao Jiang, Ning Ge, Eric Montgomery, Jiaming Zhang, Wenhao Song, Noraica Dávila, Catherine E. Graves, Zhiyong Li, John Paul Strachan, Peng Lin, Zhongrui Wang, Mark Barnell, Qing Wu, R. Stanley Williams, J. Joshua Yang, and Qiangfei Xia. Analogue signal and image processing with large memristor crossbars. *Nature Electronics*, 1(1) :52–59, December 2017.

[128] Can Li, Daniel Belkin, Yunning Li, Peng Yan, Miao Hu, Ning Ge, Hao Jiang, Eric Montgomery, Peng Lin, Zhongrui Wang, Wenhao Song, John Paul Strachan, Mark

Barnell, Qing Wu, R. Stanley Williams, J. Joshua Yang, and Qiangfei Xia. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. *Nature Communications*, 9(1), June 2018.

[129] Xin Zheng, Ryan Zarcone, Dylan Paiton, Joon Sohn, Weier Wan, Bruno Olshausen, and H. S. Philip Wong. Error-resilient analog image storage and compression with analog-valued rram arrays : An adaptive joint source-channel coding approach. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 3.5.1–3.5.4, 2018.

[130] Fang Su, Wei-Hao Chen, Lixue Xia, Chieh-Pu Lo, Tianqi Tang, Zhibo Wang, Kuo-Hsiang Hsu, Ming Cheng, Jun-Yi Li, Yuan Xie, Yu Wang, Meng-Fan Chang, Huazhong Yang, and Yongpan Liu. A 462gops/j rram-based nonvolatile intelligent processor for energy harvesting ioe system featuring nonvolatile logics and processing-in-memory. In *2017 Symposium on VLSI Technology*, pages T260–T261, 2017.

[131] M. Bocquet, T. Hirztlin, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz. In-memory and error-immune differential rram implementation of binarized deep neural networks. In *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE, December 2018.

[132] Tony F. Wu, Binh Q. Le, Robert Radway, Andrew Bartolo, William Hwang, Seungbin Jeong, Haitong Li, Pulkit Tandon, Elisa Vianello, Pascal Vivet, Etienne Nowak, Mary K. Wootters, H. S. Philip Wong, Mohamed M. Sabry Aly, Edith Beigne, and Subhasish Mitra. 14.3 a 43pj/cycle non-volatile microcontroller with 4.7s shutdown/wake-up integrating 2.3-bit/cell resistive ram and resilience techniques. In *2019 IEEE International Solid-State Circuits Conference - (ISSCC)*, pages 226–228, 2019.

[133] Tifenn Hirtzlin, Marc Bocquet, Bogdan Penkovsky, Jacques-Olivier Klein, Etienne Nowak, Elisa Vianello, Jean-Michel Portal, and Damien Querlioz. Digital biologically plausible implementation of binarized neural networks with differential hafnium oxide resistive memory arrays. *Frontiers in Neuroscience*, 13, January 2020.

[134] Djohan Bonnet, Tifenn Hirtzlin, Atreya Majumdar, Thomas Dalgaty, Eduardo Esmanhotto, Valentina Meli, Niccolo Castellani, Simon Martin, Jean-François Nodin, Guillaume Bourgeois, Jean-Michel Portal, Damien Querlioz, and Elisa Vianello. Bringing uncertainty quantification to the extreme-edge with memristor-based bayesian neural networks. *Nature Communications*, 14(1), November 2023.

[135] Weier Wan, Rajkumar Kubendran, Clemens Schaefer, Sukru Burc Eryilmaz, Wenqiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, Siddharth Joshi, Huaqiang Wu, H.-S. Philip Wong, and Gert Cauwenberghs. A compute-in-memory chip based on resistive random-access memory. *Nature*, 608(7923) :504–512, August 2022.

[136] Fatemeh Kiani, Jun Yin, Zhongrui Wang, J. Joshua Yang, and Qiangfei Xia. A fully hardware-based memristive multilayer neural network. *Science Advances*, 7(48), November 2021.

[137] Can Li, Jim Ignowski, Xia Sheng, Rob Wessel, Bill Jaffe, Jacqui Ingemi, Cat Graves, and John Paul Strachan. Cmos-integrated nanoscale memristive crossbars for cnn and optimization acceleration. In *2020 IEEE International Memory Workshop (IMW)*. IEEE, May 2020.

[138] Giacomo Pedretti, Piergiulio Mannocci, Can Li, Zhong Sun, John Paul Strachan, and Daniele Ielmini. Redundancy and analog slicing for precise in-memory machine learning—part i : Programming techniques. *IEEE Transactions on Electron Devices*, 68(9) :4373–4378, 2021.

[139] Reiji Mochida, Kazuyuki Kouno, Yuriko Hayata, Masayoshi Nakayama, Takashi Ono, Hitoshi Suwa, Ryutaro Yasuhara, Koji Katayama, Takumi Mikawa, and Yasushi Gohou. A 4m synapses integrated analog reram based 66.5 tops/w neural-network processor with cell current controlled writing and flexible network architecture. In *2018 IEEE Symposium on VLSI Technology*, pages 175–176, 2018.

[140] Cheng-Xin Xue, Wei-Hao Chen, Je-Syu Liu, Jia-Fang Li, Wei-Yu Lin, Wei-En Lin, Jing-Hong Wang, Wei-Chen Wei, Ting-Wei Chang, Tung-Cheng Chang, Tsung-Yuan Huang, Hui-Yao Kao, Shih-Ying Wei, Yen-Cheng Chiu, Chun-Ying Lee, Chung-Chuan Lo, Ya-Chin King, Chorng-Jung Lin, Ren-Shuo Liu, Chih-Cheng Hsieh, Kea-Tiong Tang, and Meng-Fan Chang. 24.1 a 1mb multibit reram computing-in-memory macro with 14.6ns parallel mac computing time for cnn based ai edge processors. In *2019 IEEE International Solid-State Circuits Conference - (ISSCC)*, pages 388–390, 2019.

[141] Stefano Ambrogio, Pritish Narayanan, Hsinyu Tsai, Robert M. Shelby, Irem Boybat, Carmelo di Nolfo, Severin Sidler, Massimo Giordano, Martina Bodini, Nathan C. P. Farinha, Benjamin Killeen, Christina Cheng, Yassine Jaoudi, and Geoffrey W. Burr. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature*, 558(7708) :60–67, June 2018.

[142] Riduan Khaddam-Aljameh, Milos Stanisavljevic, Jordi Fornt Mas, Geethan Karunaratne, Matthias Brändli, Feng Liu, Abhairaj Singh, Silvia M. Müller, Urs Egger, Anastasios Petropoulos, Theodore Antonakopoulos, Kevin Brew, Samuel Choi, Injo Ok, Fee Li Lie, Nicole Saulnier, Victor Chan, Ishtiaq Ahsan, Vijay Narayanan, S. R. Nandakumar, Manuel Le Gallo, Pier Andrea Francese, Abu Sebastian, and Evangelos Eleftheriou. Hermes-core—a 1.59-tops/mm2 pcm on 14-nm cmos in-memory compute core using 300-ps/lsb linearized cco-based adcs. *IEEE Journal of Solid-State Circuits*, 57(4) :1027–1038, 2022.

[143] P. Narayanan, S. Ambrogio, A. Okazaki, K. Hosokawa, H. Tsai, A. Nomura, T. Yasuda, C. Mackin, S. C. Lewis, A. Friz, M. Ishii, Y. Kohda, H. Mori, K. Spoon, R. Khaddam-Aljameh, N. Saulnier, M. Bergendahl, J. Demarest, K. W. Brew, V. Chan, S. Choi, I. Ok, I. Ahsan, F. L. Lie, W. Haensch, V. Narayanan, and G. W. Burr. Fully on-chip mac at 14 nm enabled by accurate row-wise programming of pcm-based weights and parallel vector-transport in duration-format. *IEEE Transactions on Electron Devices*, 68(12) :6629–6636, 2021.

[144] *Green Computing with Emerging Memory : Low-Power Computation for Social Innovation*. Springer New York, 2013.

[145] Hongyang Jia, Murat Ozatay, Yinqi Tang, Hossein Valavi, Rakshit Pathak, Jinseok Lee, and Naveen Verma. 15.1 a programmable neural-network inference accelerator based on scalable in-memory computing. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 64, pages 236–238, 2021.

[146] Mahmut E. Sinangil, Burak Erbagci, Rawan Naous, Kerem Akarvardar, Dar Sun, Win-San Khwa, Hung-Jen Liao, Yih Wang, and Jonathan Chang. A 7-nm compute-in-memory sram macro supporting multi-bit input, weight and output and achieving 351 tops/w and 372.4 gops. *IEEE Journal of Solid-State Circuits*, 56(1) :188–198, 2021.

[147] Corey Lammie, Wei Xiang, Bernabé Linares-Barranco, and Mostafa Rahimi Azghadi. Memtorch : An open-source simulation framework for memristive deep learning systems. *Neurocomputing*, 485 :124–133, May 2022.

[148] Sneh Saurabh Varshita Gupta, Shagun Kapur and Anuj Grover. Resistive random access memory : A review of device challenges. *IETE Technical Review*, 37(4) :377–390, 2020.

[149] Sung Hyun Jo and Wei Lu. CMOS compatible nanoscale nonvolatile resistance switching memory. *Nano Letters*, 8(2) :392–397, 2008.

[150] Gautam Banerjee and Robert Rhoades. Chemical mechanical planarization historical review and future direction. *ECS Transactions*, 13, 10 2008.

[151] Microelectronic applications of chemical mechanical planarization, no date. https ://www.wiley.com/en-us/Microelectronic+Applications+of+Chemical+Mechanical+Planarization-p-9780470180891, Accessed on 2023-11-24.

[152] G. Murdoch, M. O'Toole, G. Marti, A. Pokhrel, D. Tsvetanova, S. Decoster, S. Kundu, Y. Oniki, A. Thiam, Q.T. Le, O. Varela Pedreira, A. Lesniewska, G. Martinez-Alanis, S. Park, and Zs. Tokei. First demonstration of two metal level semi-damascene interconnects with fully self-aligned vias at 18mp. In *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pages 1–2, 2022.

[153] K. Motoyama, D. Metzler, C. Park, N. Lanzillo, L. Zou, S. Ghosh, and K. Choi. A novel integration scheme for self-aligend Ru topvia as post-Cu alternative metal interconnects. In *2023 IEEE International Interconnect Technology Conference (IITC) and IEEE Materials for Advanced Metallization Conference (MAM)(IITC/MAM)*, pages 1–3, 2023.

[154] Fabien Alibart, Ligang Gao, Brian D Hoskins, and Dmitri B Strukov. High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. *Nanotechnology*, 23(7) :075201, January 2012.

[155] Fabien Alibart, Elham Zamanidoost, and Dmitri B. Strukov. Pattern classification by memristive crossbar circuits using ex situ and in situ training. *Nature Communications*, 4(1), June 2013.

[156] Abdelouadoud El Mesoudy, Gwénaëlle Lamri, Raphaël Dawant, Javier Arias-Zapata, Pierre Gliech, Yann Beilliard, Serge Ecoffey, Andreas Ruediger, Fabien Alibart, and Dominique Drouin. Fully cmos-compatible passive tio2-based memristor crossbars for in-memory computing. *Microelectronic Engineering*, 255 :111706, 2022.

[157] Abdelouadoud El Mesoudy, Denis Machon, Andreas Ruediger, Abdelatif Jaouad, Fabien Alibart, Serge Ecoffey, and Dominique Drouin. Band gap narrowing induced by oxygen vacancies in reactively sputtered TiO2 thin films. *Thin Solid Films*, 769 :139737, 2023.

[158] Hong Jin Kim. Abrasive for chemical mechanical polishing. In *Abrasive Technology-Characteristics and Applications*, pages 183–201. IntechOpen, 2018.

[159] Dao-Huan Feng, Ruo-Bing Wang, Ao-Xue Xu, Fan Xu, Wei-Lei Wang, Wei-Li Liu, and Zhi-Tang Song. Mechanism of titanium–nitride chemical mechanical polishing. *Chinese Physics B*, 30(2) :028301, 2021.

[160] Mahadevaiyer Krishnan, Jakub W Nalaskowski, and Lee M Cook. Chemical mechanical planarization : Slurry chemistry, materials, and mechanisms. *Chemical Reviews*, 110(1) :178–204, 2010.

[161] Tetsuya Hoshino, Yasushi Kurata, Yuuki Terasaki, and Kenzo Susa. Mechanism of polishing of SiO2 films by CeO2 particles. *Journal of Non-Crystalline Solids*, 283(1-3) :129–136, 2001.

[162] R Manivannan and S Ramanathan. The effect of hydrogen peroxide on polishing removal rate in CMP with various abrasives. *Applied Surface Science*, 255(6) :3764–3768, 2009.

[163] Ji Chul Yang, Dong Oh, Gae Lee, Chang Song, and Taesung Kim. Step height removal mechanism of chemical mechanical planarization (CMP) for sub-nano-surface finish. *Wear*, 268 :505–510, 02 2010.

[164] D. Lim, J. Ahn, H. Park, and J. Shin. The effect of CeO 2 abrasive size on dishing and step height reduction of silicon oxide film in STI–CMP. *Surface & Coatings Technology*, 200 :1751–1754, 11 2005.

[165] K. M. Robinson, K. DeVriendt, and D. R. Evans. Integration issues of CMP. In Michael R. Oliver, editor, *Chemical-Mechanical Planarization of Semiconductor Materials*, pages 351–417. Springer-Verlag, Berlin, 2004.

[166] Pierre-Antoine Mouny, Raphaël Dawant, Bastien Galaup, Serge Ecoffey, Michel Pioro-Ladrière, Yann Beilliard, and Dominique Drouin. Analog programming of CMOS-compatible Al2O3/TiO2-x memristor at 4.2K after metal–insulator transition suppression by cryogenic reforming. *Applied Physics Letters*, 123(16) :163505, 10 2023.

[167] Fabien Alibart, Elham Zamanidoost, and Dmitri B. Strukov. Pattern classification by memristive crossbar circuits using ex situ and in situ training. 4(1) :2072.

[168] Anya Grushina. Direct-write grayscale lithography. *Advanced Optical Technologies*, 8(3-4) :163–169, 2019.

[169] Amritha Rammohan, Prabhat K. Dwivedi, Rodrigo Martinez-Duarte, Hari Katepalli, Marc J. Madou, and Ashutosh Sharma. One-step maskless grayscale lithography for the fabrication of 3-dimensional structures in su-8. *Sensors and Actuators B : Chemical*, 153(1) :125–134, 2011.

[170] Curt McKenna, Kevin Walsh, Mark Crain, and Joseph Lake. Maskless direct write grayscale lithography for mems applications. In *2010 18th Biennial University/Government/Industry Micro/Nano Symposium*, pages 1–4, 2010.

[171] Inês S. Garcia, Carlos Ferreira, Joana D. Santos, Marco Martins, Rosana A. Dias, Diogo E. Aguiam, Jorge Cabral, and João Gaspar. Fabrication of a mems micromirror based on bulk silicon micromachining combined with grayscale lithography. *Journal of Microelectromechanical Systems*, 29(5) :734–740, 2020.

[172] Melissa A. Smith, Shaun Berry, Lalitha Parameswaran, Christopher Holtsberg, Noah Siegel, Ronald Lockwood, Michael P. Chrisp, Daniel Freeman, and Mordechai Rothschild. Design, simulation, and fabrication of three-dimensional microsystem components using grayscale photolithography. *Journal of Micro/Nanolithography, MEMS, and MOEMS*, 18(4) :043507, 2019.

[173] B. Morgan, C.M. Waits, J. Krizmanic, and R. Ghodssi. Development of a deep silicon phase fresnel lens using gray-scale lithography and deep reactive ion etching. *Journal of Microelectromechanical Systems*, 13(1) :113–120, 2004.

[174] Marcel Heller, Dieter Kaiser, Maik Stegemann, Georg Holfeld, Nicoló Morgana, Jens Schneider, and Daniel Sarlette. Grayscale lithography : 3D structuring and thickness control. In Will Conley, editor, *Optical Microlithography XXVI*, volume 8683, page 868310. International Society for Optics and Photonics, SPIE, 2013.

[175] Robert R. Benoit, Delaney M. Jordan, Gabriel L. Smith, Ronald G. Polcawich, Sarah S. Bedair, and Daniel M. Potrepka. Direct-write laser grayscale lithography for multilayer lead zirconate titanate thin films. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 65(5) :889–894, 2018.

[176] Jihoon Kim, David Joy, and Soo-Young Lee. Controlling resist thickness and etch depth for fabrication of 3d structures in electron-beam grayscale lithography. *Microelectronic Engineering*, 84 :2859–2864, 12 2007.

[177] Liya Yu, Richard J. Kasica, Robert N. Newby, Lei Chen, and Vincent K. Luciani. The evaluation of photo/e-beam complementary grayscale lithography for high topography 3D structure. In Mark H. Somervell, editor, *Advances in Resist Materials and Processing Technology XXX*, volume 8682, page 868212. International Society for Optics and Photonics, SPIE, 2013.

[178] Jianming Zhang, Chuanfei Guo, Yongsheng Wang, Junjie Miao, Ye Tian, and Qian Liu. Micro-optical elements fabricated by metal-transparent-metallic-oxides grayscale photomasks. *Appl. Opt.*, 51(27) :6606–6611, Sep 2012.

[179] Laura Vera Jenni, Lalit Kumar, and Christofer Hierold. Hybrid lithography based fabrication of 3d patterns by deep reactive ion etching. *Microelectronic Engineering*, 209 :10–15, 2019.

[180] Han-Jun Kim, Marcia Almanza-Workman, Bob Garcia, Ohseung Kwon, Frank Jeffrey, Steve Braymen, Jason Hauschildt, Kelly Junge, Don Larson, Dan Stieler, Alison Chaiken, Bob Cobene, Richard Elder, Warren Jackson, Mehrban Jam, Albert Jeans, Hao Luo, Ping Mei, Craig Perlov, and Carl Taussig. Roll-to-roll manufacturing of electronics on flexible substrates using self-aligned imprint lithography (sail). *Journal of the Society for Information Display*, 17(11) :963–970, 2009.

[181] Shunpu Li and Daping Chu. A review of thin-film transistors/circuits fabrication with 3d self-aligned imprint lithography. *Flexible and Printed Electronics*, 2(1) :013002, mar 2017.

[182] H. Kim, M. R. Mahmoodi, H. Nili, and D. B. Strukov. 4k-memristor analog-grade passive crossbar circuit. *Nature Communications*, 12(1) :5198, Aug 2021.

[183] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, 521(7550) :61–64, May 2015.

[184] Gina C. Adam, Brian D. Hoskins, Mirko Prezioso, Farnood Merrikh-Bayat, Bhaswar Chakrabarti, and Dmitri B. Strukov. 3-d memristor crossbars for analog and neuromorphic computing applications. *IEEE Transactions on Electron Devices*, 64(1) :312–318, 2017.

[185] Yanfei Qi, Zongjie Shen, Chun Zhao, and Ce Zhou Zhao. Effect of electrode area on resistive switching behavior in translucent solution-processed alox based memory device. *Journal of Alloys and Compounds*, 822 :153603, 2020.

[186] Takeshi Yanagida, Kazuki Nagashima, Keisuke Oka, Masaki Kanai, Annop Klamchuen, Bae Ho Park, and Tomoji Kawai. Scaling effect on unipolar and bipolar resistive switching of metal oxides. *Scientific Reports*, 3(1) :1657, Apr 2013.

[187] R. Kirchner, V.A. Guzenko, I. Vartiainen, N. Chidambaram, and H. Schift. Zep520a — a resist for electron-beam grayscale lithography and thermal reflow. *Microelectronic Engineering*, 153 :71–76, 2016. Micro- and Nanofabrication 2015.

[188] Mandy Grube, Benjamin Schille, Matthias Schirmer, Maik Gerngroß, Uwe Hübner, Paul Voigt, and Sascha Brose. Medusa 82—hydrogen silsesquioxane based high sensitivity negative-tone resist with long shelf-life and grayscale lithography capability. *Journal of Vacuum Science & Technology B*, 39(1) :012602, 2021.

[189] Donggyu Kim, Mohamed I. Ibrahim, Christopher Foy, Matthew E. Trusheim, Ruonan Han, and Dirk R. Englund. A cmos-integrated quantum sensor based on nitrogen–vacancy centres. *Nature Electronics*, 2(7) :284–289, Jul 2019.

[190] Yurii A. Vlasov. Silicon cmos-integrated nano-photonics for computer and data communications beyond 100g. *IEEE Communications Magazine*, 50(2) :s67–s72, 2012.

[191] Min Zhao, Tang Xu, Baoqin Chen, and Jiebin Niu. Technology of alignment mark in electron beam lithography. In Xiangang Luo and Harald Giessen, editors, *7th International Symposium on Advanced Optical Manufacturing and Testing Technologies : Smart Structures and Materials for Manufacturing and Testing*, volume 9285, pages 66 – 71. International Society for Optics and Photonics, SPIE, 2014.

[192] K.E. Docherty, K.A. Lister, J. Romijn, and J.M.R. Weaver. High robustness of correlation-based alignment with penrose patterns to marker damage in electron beam lithography. *Microelectronic Engineering*, 86(4) :532–534, 2009. MNE '08.

[193] K.E. Docherty, S. Thoms, P. Dobson, and J.M.R. Weaver. Improvements to the alignment process in a commercial vector scan electron beam lithography tool. *Microelectronic Engineering*, 85(5) :761–763, 2008. Proceedings of the Micro- and Nano-Engineering 2007 Conference.

[194] High resolution lithography : EBL tool EBPG5200 | raith group.

[195] Min Zhao, Tang Xu, Baoqin Chen, and Jiebin Niu. Technology of alignment mark in electron beam lithography. In *7th International Symposium on Advanced Optical Manufacturing and Testing Technologies : Smart Structures and Materials for Manufacturing and Testing*, volume 9285, pages 66–71. SPIE.

[196] A. D. Wilson, T. H. P. Chang, and A. Kern. Experimental scanning electron-beam automatic registration system. 12(6) :1240–1245. Publisher : American Vacuum Society.

[197] A.M. Bruckstein, L. O'Gorman, and A. Orlitsky. Design of shapes for precise image registration. *IEEE Transactions on Information Theory*, 44(7) :3156–3162, 1998. cited By 25.

[198] Y. Chen, W. Huang, and X. Dang. Design and analysis of two-dimensional zero-reference marks for alignment systems. *Review of Scientific Instruments*, 74(7) :3549–3553, 2003. cited By 15.

[199] E. H. Anderson, D. Ha, and J. A. Liddle. Sub-pixel alignment for direct-write electron beam lithography. 73-74 :74–79.

[200] Eitan N. Shauly and S. Rosenthal. Coverage layout design rules and insertion utilities for cmp-related processes. *Journal of Low Power Electronics and Applications*, 2020.

[201] Stephen Thoms, Douglas S. Macintyre, Kevin E. Docherty, and John M.R. Weaver. Alignment verification for electron beam lithography. *Microelectronic Engineering*, 123 :9–12, 2014. Nano Lithography 2013.

[202] Roger Penrose and N David Mermin. The emperor's new mind : Concerning computers, minds, and the laws of physics. *American Journal of Physics*, 58(12) :1214–1216, 1990.