PolyPublie
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

| | |
|---|---|
| **Titre:** Title: | Mémoire optique à changement de phase pour calcul neuromorphique destiné à l'intelligence artificielle |
| **Auteur:** Author: | Thomas Lacasse |
| **Date:** | 2023 |
| **Type:** | Mémoire ou thèse / Dissertation or Thesis |
| **Référence:** Citation: | Lacasse, T. (2023). Mémoire optique à changement de phase pour calcul neuromorphique destiné à l'intelligence artificielle [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/55106/ |

**Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

| | |
|---|---|
| **URL de PolyPublie:** PolyPublie URL: | https://publications.polymtl.ca/55106/ |
| **Directeurs de recherche:** Advisors: | Yves-Alain Peter |
| **Programme:** Program: | Génie physique |

**Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal**
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

# Mémoire optique à changement de phase pour calcul neuromorphique destiné à l'intelligence artificielle

**THOMAS LACASSE**
Département de génie physique

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie physique

Août 2023

© Thomas Lacasse, 2023.

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

Ce mémoire intitulé :

**Mémoire optique à changement de phase pour calcul neuromorphique destiné à l'intelligence artificielle**

présenté par **Thomas LACASSE**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

**Ludvik MARTINU**, président
**Yves-Alain PETER**, membre et directeur de recherche
**Juan Nicolás QUESADA MEJIA**, membre

# DÉDICACE

*À ma famille et à Béa...*

# REMERCIEMENTS

Je tiens à remercier mon directeur de recherche, Yves-Alain Peter de m'avoir acceuilli dans son laboratoire d'abord pour mon projet de maîtrise professionnelle et d'avoir cru en moi pour m'attribuer un projet de recherche aussi enrichissant et passionnant. Son support et ses connaissances ont été essentiels lors des multiples embûches rencontrées lors de ce projet.

Je tiens ensuite à remercier Emanuel Peinke et Mohamed Fouda, mes deux mentors, fondateurs de 3E8, soit la *startup* à partir de laquelle ce projet est né. Je tiens également à remercier Najla Najeeb, Md Mahadi Masnad, Kaveh Hassan Rahbardar Mojaver, Di Zhang et Sara Radfar avec qui j'ai eu la chance de travailler en tant que stagiaire chez 3E8 au cours des deux dernières années.

Je remercie également mes collègues du laboratoire de microphotonique de Polytechnique Montréal : Cédric Lemieux-Leduc, Régis Guertin, Marc-Antoine Bianki, Ahmed Sami Bahgat Abdelaziz, Salvador Poveda-Hospital et Alberto Acevedo Carrera grâce auxquels j'ai eu la chance d'acquérir des connaissances et des expériences inestimables.

Je suis également très reconnaissant de toute l'aide qui m'a été apportée par les techniciens et associée de recherche au laboratoire de microfabrication (LMF) dont Christophe Clément, Marie-Hélène Bernier, André Hoang-Gaona et Alexandre Perron, qui m'ont aidé et guidé à maintes reprises dans l'exécution de mes procédés de microfabrication.

Finalement, je tiens à remercier le jury d'avoir accepté d'évaluer mon travail de maîtrise.

# RÉSUMÉ

L'objectif principal de ce projet de maîtrise est d'explorer l'usage de matériaux à changement de phase comme mémoire optique non-volatile dans le but d'effectuer des opérations matricielles rapidement et efficacement dans le domaine optique. Les multiplications sont effectuées par modulation d'amplitude grâce au contraste de pertes optiques entre la phase amorphe et cristalline du chalcogénure étudié, dont le changement de phase est rapide et réversible. L'état de ces mémoires, lorsque intégrées à un circuit photonique actif faisant également l'objet de ce mémoire, peut être individuellement altéré par l'emploi d'éléments chauffants intégrés de sorte à encoder les éléments de la matrice à multiplier.

La proposition implique la caractérisation du $Ge_2Sb_2Se_5$, un matériau émergent prometteur dans le domaine des mémoires non-volatiles optiques. Le contraste optique mesuré entre les états vitreux et cristallins du matériau implique un différence d'absorption approximative de 0.8 dB/µm dans la bande C et des pertes d'insertion négligeables dans l'état amorphe, justifiant son utilité potentielle comme mémoire. La méthode d'intégration proposée consiste au dépôt du matériau sur guides d'onde en silicium, de sorte à maximiser l'interaction entre le mode de propagation et le chalcogénure.

Cette intégration a été effectuée par un procédé de microfabrication maison développé au Laboratoire de Microfabrication de Polytechnique Montréal. Un contraste optique de 0.5 à 1 dB/µm a été observé lorsque la mémoire est intégrée au circuit photonique suite à une cristallisation induite par recuit lent.

Une nouvelle architecture de multiplication matricielle optique, mieux adaptée aux grandes matrices a été présentée et simulée, démontrant une amélioration de 3 ordres de grandeur en terme de pertes et d'imbalance de calcul pour des matrices de taile $N = 100$.

# ABSTRACT

The main objective of this Master's project is to explore the use of phase-change materials as non-volatile optical memory to perform matrix operations rapidly and efficiently in the optical domain. Multiplications are achieved through amplitude modulation, taking advantage of the contrast in optical losses generated between the amorphous and crystalline phases of the studied chalcogenide material, which undergoes rapid and reversible phase changes. The states of these memories, when integrated into an active photonic architecture, which is also the subject of this project, can be individually altered using integrated micro-heaters, thereby encoding the elements of the matrix to be multiplied.

The proposal involves the characterization of $Ge_2Sb_2Se_5$, a promising and emerging material in the field of non-volatile optical memories. The measured optical contrast between the glassy and crystalline states of the material implies an approximate absorption difference of 0.8 dB/m in the C band with negligible insertion losses in the amorphous state, justifying its potential usefulness as memory. The proposed integration method involves depositing the material onto silicon waveguides to maximize the interaction between the propagation mode and the chalcogenide.

This integration was carried out using a custom microfabrication process developed at the Microfabrication Laboratory of Polytechnique Montréal. An optical contrast of 0.5 to 1 dB/m was measured when the memory was integrated into the photonic circuit following crystallization induced by slow annealing.

A new architecture for optical matrix multiplication, better suited for large matrices, was presented and simulated, demonstrating a three-orders-of-magnitude improvement in terms of computing losses and imbalance for matrices of size $N = 100$.

# TABLE DES MATIÈRES

DÉDICACE . . . . . iii

REMERCIEMENTS . . . . . iv

RÉSUMÉ . . . . . v

ABSTRACT . . . . . vi

TABLE DES MATIÈRES . . . . . vii

LISTE DES TABLEAUX . . . . . x

LISTE DES FIGURES . . . . . xi

LISTE DES SIGLES ET ABRÉVIATIONS . . . . . xiv

LISTE DES ANNEXES . . . . . xvii

CHAPITRE 1 INTRODUCTION . . . . . 1
1.1 Éléments de la problématique . . . . . 1
1.1.1 Limitations de l'électronique moderne . . . . . 1
1.1.2 Croissance de l'intelligence artificielle . . . . . 3
1.1.3 Efficacité des circuits intégrés spécifique . . . . . 4
1.1.4 Maturité de la photonique intégrée . . . . . 4
1.2 Objectifs de recherche . . . . . 5
1.2.1 Caractérisation du $Ge_2Sb_2Se_5$ . . . . . 5
1.2.2 Intégration photonique sur silicium . . . . . 6
1.2.3 Exploration d'une architecture photonique calculatoire . . . . . 6
1.3 Plan du mémoire . . . . . 6

CHAPITRE 2 DÉFINITIONS ET CONCEPTS THÉORIQUES . . . . . 7
2.1 Photonique intégrée sur silicium . . . . . 7
2.2 Pertes optiques . . . . . 8
2.2.1 Absorption diélectrique . . . . . 8
2.2.2 Pertes par chevauchement des modes . . . . . 10
2.2.3 Pertes plasmoniques . . . . . 10

2.3 Changement de phase . . . . . 12
2.3.1 Phase amorphe et vitrification . . . . . 12
2.3.2 État cristallin et cristallisation . . . . . 15
2.3.3 Ségrégation moléculaire et *dewetting* . . . . . 16
2.4 Réseaux de neurones artificiels . . . . . 17
2.5 Planarité des multiplications matrice-vecteur . . . . . 19
2.6 Effet Joule . . . . . 22

CHAPITRE 3 REVUE DE LITTÉRATURE . . . . . 23
3.1 Intelligence artificielle . . . . . 23
3.1.1 Demande en intelligence artificielle . . . . . 23
3.1.2 Solutions actuelles : GPU, FPGA, ASIC . . . . . 25
3.2 Calcul à faible précision . . . . . 26
3.3 Mémoires à changement de phase . . . . . 29
3.3.1 Matériaux considérés . . . . . 29
3.3.2 Méthode de dépôt . . . . . 31
3.3.3 Méthodes de cristallisation . . . . . 31
3.4 Calcul neuromorphique . . . . . 32
3.4.1 Définition . . . . . 32
3.4.2 Convertisseur digital-analogique optique . . . . . 33
3.4.3 Grille de memristors . . . . . 35
3.4.4 Cascade d'interféromètres Mach-Zehnder . . . . . 38
3.4.5 Processeur matriciel avec mémoire à changement de phase . . . . . 41
3.5 Croisements optiques . . . . . 42

CHAPITRE 4 SIMULATIONS . . . . . 46
4.1 Simulations optiques . . . . . 46
4.1.1 Modes . . . . . 46
4.1.2 Dimensions du matériau à changement de phase . . . . . 47
4.1.3 Distance de l'élément chauffant . . . . . 48
4.2 Simulations électrothermiques . . . . . 49
4.2.1 Régime stationnaire . . . . . 49
4.2.2 Régime transitoire . . . . . 51

CHAPITRE 5 CARACTÉRISATION DU $Ge_2Sb_2Se_5$ . . . . . 53
5.1 Préparation des échantillons . . . . . 53
5.1.1 Dépôt du matériau à changement de phase . . . . . 53

5.1.2 Cristallisation . . . . . 55
5.2 Ellipsométrie spectroscopique . . . . . 56
5.2.1 Théorie . . . . . 56
5.2.2 Méthodologie . . . . . 58
5.2.3 Résultats . . . . . 58
5.3 Spectrophotométrie . . . . . 61
5.3.1 Théorie . . . . . 61
5.3.2 Méthodologie . . . . . 63
5.3.3 Résultats . . . . . 64
5.4 Spectroscopie photoélectronique à rayons X (XPS) . . . . . 64
5.4.1 Théorie . . . . . 64
5.4.2 Méthodologie . . . . . 65
5.4.3 Résultats . . . . . 65
5.5 Diffraction par rayons X . . . . . 67
5.5.1 Théorie . . . . . 67
5.5.2 Résultats . . . . . 67

CHAPITRE 6 FABRICATION ET MÉTHODOLOGIE EXPÉRIMENTALE . . . . 70
6.1 Procédé de fabrication . . . . . 70
6.2 Montage optique . . . . . 71

CHAPITRE 7 ANALYSE DES RÉSULTATS . . . . . 74
7.1 Changement de phase thermique . . . . . 74
7.1.1 Ségrégation et *dewetting* . . . . . 76
7.2 Changement de phase électrothermique . . . . . 78
7.3 Architecture du mutliplicateur matrice-vecteur . . . . . 82

CHAPITRE 8 CONCLUSION . . . . . 89
8.1 Rappel des travaux . . . . . 89
8.2 Améliorations futures . . . . . 89

RÉFÉRENCES . . . . . 91

ANNEXES . . . . . 103

# LISTE DES TABLEAUX


Tableau 1.1 Évolution des meilleures résolutions atteintes par les diverses sources d'exposition depuis 1994. . . . . 3

Tableau 1.2 Méthodes de caractérisation du $Ge_2Sb_2Se_5$ et propriétés mesurées. . . 5

Tableau 3.1 Comparaison qualitative de divers accélérateur électroniques utilisés pour l'intelligence artificielle. . . . . 26

Tableau 3.2 Exactitude d'inférence de modèles selon leur précision pour divers jeux de données. . . . . 28

Tableau 3.3 Caractéristiques d'un bon matériau à changement de phase pour mémoire non-volatile . . . . . 29

# LISTE DES FIGURES


| | | |
|---|---|---|
| Figure 1.1 | Précision de la loi de Moore depuis 1971. | 2 |
| Figure 2.1 | Différences entre l'état cristallin ordonné et amorphe (vitreux). | 12 |
| Figure 2.2 | Changement de phase caractérisé par une abrupte variation de la capacité thermique isobare. | 14 |
| Figure 2.3 | Représentation non planaire de la multiplication matrice-vecteur ($N = 2$) | 20 |
| Figure 2.4 | Représentation planaire de la multiplication matrice-vecteur ($N = 2$) | 20 |
| Figure 2.5 | Graphe représentant la multiplication matrice-vecteur ($N = 3$) | 21 |
| Figure 3.1 | Croissance de la taille des modèles GPT de OpenAI depuis GPT-1. | 24 |
| Figure 3.2 | Envergure de l'entraînement de modèles d'intelligence artificielle populaires. | 24 |
| Figure 3.3 | Représentation de la centralisation des bits, une étape fondamentale des algorithmes de ML basés sur une faible précision. | 27 |
| Figure 3.4 | Architecture générale des accélérateurs optiques en interface électronique comprenant le convertisseur digital-optique (D2O), le processeur photonique (PTC) et le convertisseur optique-digital comprenant les photodiodes (PD), l'amplificateur transimpédance (TIA). | 34 |
| Figure 3.5 | Diagramme reliant les quatre composants passifs fondamentaux en électronique. | 35 |
| Figure 3.6 | Hystérèse observée dans la courbe IV d'un memristor. | 36 |
| Figure 3.7 | Schéma du fonctionnement d'une architecture neuromorphique en grille à memristors (MVM de dimension 4x3) | 37 |
| Figure 3.8 | Unité normalisée $U_2$ du calcul par cascade de MZI. | 39 |
| Figure 3.9 | Exemple d'architecture $U_4$ normalisée pour effectuer une multiplication matrice-vecteur de taille 4x4. | 39 |
| Figure 3.10 | Avantages et inconvénients des technologies de déphaseurs sur silicium. | 41 |
| Figure 3.11 | Processeur à produit scalaire (DPE) de taille 4 utilisant la modulation passive en amplitude par matériaux à changement de phase. | 42 |
| Figure 3.12 | Illustration de 3 types de croisements en photonique intégrée : a) par entonnoir, b) par MMI et c) par pontage. | 44 |
| Figure 4.1 | Modes fondamentaux $\|\mathbf{E}\|$ pour le guide d'onde i) sans mémoire, ii) avec mémoire en phase amorphe, iii) avec mémoire en phase cristalline à $\lambda = 1550$ nm. | 47 |

Figure 4.2 Pertes engendrées par une cellule de mémoire en fonction de l'épaisseur du $Ge_2Sb_2Se_5$. . . . . 48
Figure 4.3 Impact de la distance verticale de l'élément chauffant sur les pertes plasmoniques générées $\alpha$. . . . . 49
Figure 4.4 Distribution de la température de la mémoire à l'état stationnaire . . 50
Figure 4.5 Impact de la tension aux bornes de l'élément chauffant sur la température de divers régions de la mémoire (régime constant). . . . . 51
Figure 4.6 Régime transitoire du chauffage de la mémoire optique pour une impulsion de voltage de 250 µs. . . . . 52
Figure 5.1 Effet de la puissance sur le taux de dépôt (5 mTorr). . . . . 54
Figure 5.2 Effet de la pression sur le taux de dépôt (50 W). . . . . 55
Figure 5.3 Cellule de GSSe de 100 µm. . . . . 56
Figure 5.4 Cellule de GSSe de 20 µm. . . . . 56
Figure 5.5 Comparaison des modèles ellipsométriques pour le $Ge_2Sb_2Se_5$ amorphe. 59
Figure 5.6 Impact de la pression de dépôt sur l'indice de réfraction du matériau. 60
Figure 5.7 Comparaison des modèles ellipsométriques pour le $Ge_2Sb_2Se_5$ cristallin. 61
Figure 5.8 Coefficient d'absorption déterminé par spectrophotométrie. . . . . 64
Figure 5.9 Composition atomique du $Ge_2Sb_2Se_5$ déposé par pulvérisation au LMF (Polytechnique Montréal) . . . . . 66
Figure 5.10 Résultats de diffraction à rayons-X pour les deux états de $Ge_2Sb_2Se_5$ en comparaison avec la littérature et avec la base de données cristallographique ouverte *COD*. . . . . 68
Figure 5.11 Zoom sur les pics significatifs à 32.5 °. . . . . 69
Figure 6.1 Procédé de fabrication d'une cellule de mémoire. . . . . 71
Figure 6.2 Schéma du montage. . . . . 72
Figure 7.1 Absorption optique mesurée suite au recuit de cellules de $Ge_2Sb_2Se_5$. 75
Figure 7.2 Phénomènes indésirables observés lors de la cristallisation par recuit lent d'une cellule de $Ge_2Sb_2Se_5$ de dimension 50 µm x 20 µm x 50 nm. 76
Figure 7.3 Représentation du *dewetting* suite à un recuit. . . . . 77
Figure 7.4 Représentation de la migration du PCM suite à un recuit. . . . . 77
Figure 7.5 Confirmation de la rétraction du PCM sur le guide suite à un recuit thermique . . . . . 77
Figure 7.6 Problème d'adhésion lors de la fabrication des éléments chauffants. . 78
Figure 7.7 Image de la cellule de mémoire test fabriquée pour caractériser le changement de phase électrothermique. . . . . 79

Figure 7.8 Programmes de cristallisation testés pour une cellule de mémoire (1 impulsion, 50% cycle de travail). . . . . 80
Figure 7.9 Programmes de cristallisation testés pour une cellule de mémoire (1 impulsion, 50% cycle de travail). . . . . 81
Figure 7.10 Cellule de mémoire détruite par le bris de l'élément chauffant. . . . . 82
Figure 7.11 Architecture typique d'un MVM photonique avec PCM. . . . . 83
Figure 7.12 Architecture d'un MVM photonique avec PCM exploitant le pontage Si-SiN. . . . . 84
Figure 7.13 Comparaison des pertes totales des deux architectures étudiées. . . . 85
Figure 7.14 Imbalance maximale théorique entre les deux architectures étudiées. . 86
Figure 7.15 Réponse spectrale d'un coupleur contra-directionnel suggéré pour effectuer la WDM dans une architecture de MVM photonique. . . . . 87
Figure 7.16 Réponse de quatres filtres spectraux à large bande en cascade. . . . . 88
Figure A.1 Exemple d'images contenues dans les jeux de données standardisés présentés. . . . . 103
Figure B.1 Masques photolithographiques pour i) le guide d'onde fabriqué en fonderie, ii) les cellules de mémoire, iii) le microradiateur, iv) les connexions métalliques, v) le tout superposé. . . . . 104

# LISTE DES SIGLES ET ABRÉVIATIONS

| | |
|---|---|
| AGI | Artificial General Intelligence |
| AMF | Advanced Micro Foundry |
| ANT | Applied Nanotools |
| ASIC | Application-specific integrated circuit |
| BOX | Buried Oxide |
| CCD | Charge coupled device |
| CD | Critical dimension |
| CD | Compact disc |
| CI | Circuit intégré |
| CMOS | Complementary metal-oxide semiconductor |
| COD | Cristallography open database |
| CPU | Central processing unit |
| CVD | Chemical Vapour Deposition |
| CW | Continuous wave |
| CWDM | Coarse wavelength division multiplexing |
| D2O | Digital to optical |
| DAC | Digital to analog converter |
| DFB | Distributed feedback laser |
| DL | Deep Learning |
| DNN | Deep neural network |
| DPE | Dot-product engine |
| FAU | Fiber array unit |
| FLOP | Opérations à virgule flottante |
| FLOPS | Floating operations per second |
| FLOPS | Opérations à virgule flottante par seconde |
| FOM | Figure of merit |
| FPGA | Field-programmable gate array |
| GPT | Generative Pre-Transformer |
| GPU | Graphics processing unit |
| HALP | High-Accuracy Low-Precision |
| IA | Intelligence artificielle |
| ICDD | International centre for diffraction data |
| IoT | Internet of things |

| | |
|---|---|
| IPS | Instructions per second |
| LMF | Laboratoire de microfabrication |
| LPCVD | Low-Pressure Chemical Vapour Deposition |
| MAC | Multiply accumulate |
| MCW | Multi-contact wedge |
| MEMS | Micro-electromechanical system |
| ML | Machine Learning |
| MMI | Multimode interferometer |
| MPW | Multi-Project Wafer |
| MSE | Mean-squared error |
| MVM | Matrix-Vector Multiplication |
| MZI | Mach-Zehnder interferometer |
| NN | Neural Network |
| O2D | Optical to Digital |
| PAM | Pulse amplitude modulation |
| PCB | Printed Circuit Board |
| PCM | Phase-Change Material |
| PD | Photodétecteur |
| PECVD | Plasma-Enhanced Chemical Vapour Deposition |
| PIC | Photonic Integrated Circuit |
| PLD | Pulsed laser deposition |
| PRAM | Phase-change random-access memory |
| RF | Radio fréquence |
| RPM | Revolutions per second |
| SiP | Silicon Photonics |
| SOI | Silicon-On-Insulator |
| SWIR | Short-Wavelength Infrared |
| TE | Transverse électrique |
| TM | Transverse magnétique |
| TPU | Tensor processing unit |
| TTL | Transistor-transistor Logic |
| VASE | Variable angle spectroscopic ellipsometer |
| VCSEL | Vertical-cavity surface-emitting laser |
| WDM | Wavelength division multiplexing |
| XPS | Spectroscopie photoélectronique à rayons X |

XRD Diffraction par rayons X

# LISTE DES ANNEXES

Annexe A Jeux de données standardisés . . . . . 103
Annexe B Masques photolithographiques . . . . . 104
Annexe C Procédé photolithographique . . . . . 105

# CHAPITRE 1 INTRODUCTION

La loi de Moore et la mise à l'échelle de Dennard en combinaison avec l'essor de l'intelligence artificielle de la dernière décennie soulèvent aujourd'hui un déséquilibre entre la demande et l'offre de puissance de calcul. La miniaturisation des processeurs et l'ingéniosité de leur design ont permis, grâce entre-autres aux efforts consacrés à l'amélioration des techniques de microfabrication et à la découverte de matériaux de pointe, à l'industrie des circuits intégrés de combler la demande depuis 1960. Le ralentissement de ce progrès observé depuis 2010 a été comblé par l'arrivée d'infrastructures spécifiques dédiées à certains types de calcul informatique : les ASICs (*Application Specific Integrated Circuits*). Aujourd'hui, ni les processeurs graphiques (GPU), les matrices FPGA ou les processeurs de tenseur (TPU) ne répondent aux exigences de l'industrie. Il semble évident, vu les limites auxquelles font face les circuits intégrés (CI) modernes, que la prochaine révolution permettant de maintenir la croissance désirée sera basée sur un paradigme de calcul fondamentalement différent de la microélectronique telle qu'on la connaît.

## 1.1 Éléments de la problématique

### 1.1.1 Limitations de l'électronique moderne

La loi de Moore, proposée par Gordon Moore en 1965 est une prédiction empirique sur la croissance de la performance des circuits intégrés. Plus précisément, elle stipule que le nombre de transistors compris dans un CI dense double approximativement à chaque deux ans. Moore, se basant sur très peu de données historiques à l'époque, fonda Intel Corporation à peine trois ans plus tard et dédia ainsi la majorité de sa vie à la réalisation de sa prédiction.

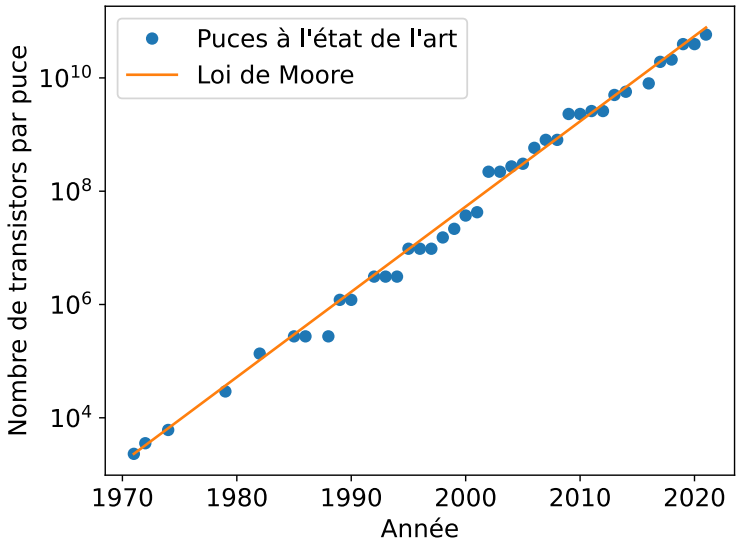



FIGURE 1.1 Précision de la loi de Moore depuis 1971.

Bien qu'elle prédise avec précision (voir figure 1.1 [1]) la croissance d'une industrie en entier sur plus de cinq décennies, la loi de Moore a rencontré son lot d'opposition de la part de la communauté scientifique au fil des années. Depuis les années 2000, on retrouve périodiquement des articles d'opinion offrant diverses prédictions et déclarations sur la fin de la loi de Moore [2–4]. Pourtant, 20 ans plus tard, son déclin suscite encore des débats dans la communauté scientifique avec des experts tels Jim Keller de *Intel*, *AMD* et *Apple* niant à ce jour son ralentissement [5].

La capacité de miniaturisation des transistors est toutefois fondamentalement limitée par les interactions atomiques indésirables prévalentes lorsque la grille du transistor n'est composée que d'une poignée d'atomes. Malgré que la poussée pour atteindre des résolutions de 20 nm est présente [6], la question sur les ressources nécessaires pour y parvenir perdure.

Depuis les années 1970, la majorité des avancées en terme de miniaturisation des CIs sont attribuables au progrès des systèmes de lithographie. Limitée approximativement par le critère de Rayleigh (équation 1.1), la dimension critique ($CD$) atteignable par un système de lithographie est proportionelle à la longueur d'onde d'exposition $\lambda$, inversement proportionnelle à l'indice de réfraction $n$ du milieu entourant le système de lentilles et $\alpha$, le demi-angle du cone de lumière maximal. Le tableau 1.1 présente les technologies charnières ayant contribué au progrès en photolithographie.

$$CD = \frac{k_1 \lambda}{NA} = \frac{k_1 \lambda}{n \sin \alpha} \tag{1.1}$$

Tableau 1.1 Évolution des meilleures résolutions atteintes par les diverses sources d'exposition depuis 1994.

| **Année** | **Source** | **Catégorie** | **Longueur d'onde** | **Noeud du procédé*** |
|---|---|---|---|---|
| 1994 | Hg (G-Line) | UV | 436 nm | 500 nm |
| 1998 | Hg (I-Line) | UV | 365 nm | 250 nm |
| 2001 | KrF | DUV | 248 nm | 130 nm |
| 2006 | ArF | DUV | 193 nm | 65 nm |
| 2017 | i-ArF** | DUV | 193 nm | 7 nm |
| 2022 | Sn | EUV | 13.5 nm | 5 nm |

* Traduction libre de *Process Node*, terme de marketing couramment utilisé par les fonderies et ne correspond pas aux dimensions critiques des transistors, mais plutôt à une métrique de performance relative à la génération précédente. En réalité, le procédé 3 nm permet une résolution de 24 nm.

** Émission à 193 nm utilisant l'immersion dans l'eau du système de lentilles pour améliorer la résolution globale.

### 1.1.2 Croissance de l'intelligence artificielle

Au cours de la dernière décénnie, cette loi empirique et auto-imposée par l'industrie est remplacée par une offre provenant de l'intelligence artificielle exigeant une demande proportionnelle. Depuis le début de l'ère moderne de l'intelligence artificielle (IA), celle-ci croit à un rythme cinq fois plus rapide que la loi de Moore. Bien que ces marchés soient similaires et inter reliés par le capitalisme de l'offre et de la demande, l'IA n'est pas confrontée à une limitation fondamentale au même titre que le progrès de CIs ou que la loi de Moore. Sa polyvalence et son accessibilité en fait un outil recherché pour une panoplie d'utilisations.

Aujourd'hui, avec GPT-4 et les milliards de dollars en investissements que ChatGPT a engendrés [7], ce taux de croissance est porté à augmenter, causant des enjeux en terme de consommation énergétique. Par exemple, entraîner un modèle de langues naturelles à l'état de l'art, requière approximativement 1 MWh, soit l'énergie consommée par un cerveau humain durant plus de six ans [8]. Extrapolée à partir de nos connaissances actuelles, l'électronique moderne est vouée à consommer $10^{27}$ Joules en 2040 uniquement pour des applications de logique binaires *Complementary Metal Oxide Semiconductor* (CMOS) [9]. Cette importante croissance nécessite davantage de ressources de calcul pour résoudre des problèmes de plus en plus ambitieux, imposant une réforme des solutions de calcul logicielles et *hardware*.

### 1.1.3 Efficacité des circuits intégrés spécifique

En vue des limitations de l'électronique intégrée moderne décrites à la section 1.1.1 et de la croissance exponentielle de la demande vue en 1.1.2, diverses alternatives sont actuellement étudiées. Aujourd'hui, il va sans dire qu'utiliser un processeur central (CPU), dédié au traitement d'instructions variées en série, pour rendre des modèles 3D ou pour entraîner un réseau de neurones profond (DNN) est inefficace. Il est donc nécessaire de concevoir des architectures adaptées à la figure de mérite (FOM) visée. Ainsi, une FOM appropriée pour mesurer l'efficacité d'un ASIC destiné à l'IA est basée sur la performance d'opérations en virgule flottante, mesurée en FLOPS (*FLoating point Operations Per Second*) et non selon par le nombre d'instructions par seconde (IPS), qui lui, est beaucoup plus approprié pour les CPUs.

Il s'avère que les limitations microscopiques intrinsèques à l'électronique moderne ne sont pas l'unique problème dans le développement d'accélérateurs de calcul. L'architecture de Von Neumann, basée sur la communication entre le processeur, l'unité arithmétique/logique et la mémoire vive, constitue un goulot d'étranglement fondamental en raison des nombreux appels et écritures en mémoire. Ainsi, la recherche dédiée aux processeurs neuromorphiques dits non-Von Neumann gagne en popularité pour intégrer le calcul en mémoire, sous forme d'accélérateur de calcul. Parmi les approches au calcul neuromorphique, l'optique se distingue par sa compatibilité avec les opérations MAC (Multiply ACcumulate), sa faible dissipation thermique et sa vitesse. Une alternative présentée pour remédier à ce problème constitue au développement d'accélérateur optiques.

### 1.1.4 Maturité de la photonique intégrée

L'industrie des télécommunications utilise l'optique et la photonique intégrée depuis des années sous forme de lasers, modulateurs, multiplexeurs et coupleurs. Ces derniers pouvant être utilisés à des fins calculatoires. Basée sur une technologie de fabrication quasi-identique à l'électronique intégrée, la photonique intégrée dispose de près de 8 décennies d'expertise en fabrication depuis l'invention du transistor. De plus, la plate-forme conventionnelle sur silicium est idéale pour des applications optiques en raison du contraste d'indice de réfraction entre celui-ci et son oxyde natif ainsi que sa large plage de transparence dans l'infrarouge.

Ceci mène au développement de procédés dédiés à la photonique intégrée de la part de géants des semiconducteurs comme *GlobalFoundries* et *Intel* en plus de la création de fonderies spécialisées comme *Advanced Micro Foundry* (AMF) à Singapour et *Applied Nanotools* (ANT) au Canada. Ces manufacturiers offrent des librairies de composants utiles à la fabrication

d'accélérateurs optiques. Le facteur manquant pour le développement d'ASICs photoniques demeure la mémoire, qui fait objet de recherche depuis quelques années, mais pour laquelle aucune solution ne semble faire l'unanimité. Ce besoin pour une mémoire optique, analogue à la mémoire flash, fait l'objet du présent mémoire. Les objectifs de recherche qui suivent décrieront les étapes parcourues afin de contribuer à remédier à cette problématique.

## 1.2 Objectifs de recherche

L'objectif global de ce mémoire est de contribuer au développement d'une architecture de calcul neuromorphique optique en fabricant, caractérisant et en intégrant une mémoire optique non-volatile. Les sous-objectifs permettant d'accomplir ce but principal font l'objet des sous-sections suivantes.

### 1.2.1 Caractérisation du $Ge_2Sb_2Se_5$

Le premier sous-objectif est la caractérisation du chalcogénure utilisé comme matériau pour la mémoire optique. Il s'agit d'un matériau à changement de phase (PCM) émergent ayant gagné en popularité au cours des dernières années pour ses propriétés optiques, électroniques ainsi que pour le contraste de celles-ci suite au changement de phase [10, 11]. Cette étape implique la mesure des propriétés électriques et optiques du matériau en phase amorphe et cristalline. La comparaison des résultats obtenus avec ceux répertoriés dans la littérature sera ensuite discutée. Les mesures et les propriétés dérivées de ces mesures sont présentées au tableau 1.2.

Tableau 1.2 Méthodes de caractérisation du $Ge_2Sb_2Se_5$ et propriétés mesurées.

| **Mesure/Test** | **Propriétés mesurées** | **Références** |
|---|---|---|
| Ellipsométrie | $n$, $\kappa$, $E_g$ | [12–14] |
| Spectrophotométrie | $\alpha$ | [13] |
| Spectroscopie photoélectronique à rayons X | Composition stoechiométrique | N/A |
| Diffraction de rayons X | Structure cristalline | [15–17] |

On définit ainsi $N = n + i\kappa$ comme l'indice de réfraction du matériau, $E_g$, sa largeur de bande et $\alpha$, l'absorption du matériau tel que

$$\alpha = \frac{4\pi\kappa}{\lambda}. \tag{1.2}$$

### 1.2.2 Intégration photonique sur silicium

Suite à la confirmation des propriétés optiques, il est essentiel d'intégrer le matériau à un circuit photonique. Le PCM n'étant pas communément utilisé dans l'industrie, sa compatibilité avec les procédés de fabrication *CMOS* ou logique transistor-transistor (TTL) doit être démontrée. Pour accomplir cet objectif, l'utilisation de *Multi-Project Wafer runs* (MPW) en photonique sur silicium (SiP) accessibles via l'abonnement à CMC Microsystems est privilégiée en raison du court délai de fabrication et du moindre coût. Divers méthodes d'intégration de la mémoire sont étudiés et comparés selon leur impact sur l'évolutivité du design.

### 1.2.3 Exploration d'une architecture photonique calculatoire

Une fois le fonctionnement de la mémoire établi, c'est-à-dire son changement de phase et les pertes optiques engendrées par celui-ci, l'exploration d'architectures de calcul permettant la mise à l'échelle des architectures de petite taille publiées sera explorée.

## 1.3 Plan du mémoire

Ce mémoire vise d'abord à établir les fondements théoriques des composants optiques et du matériaux à changement de phase utilisés dans le cadre de ce projet avant d'introduire brièvement la pertinence du développement d'accélérateurs de calcul optiques dans le domaine de l'intelligence artificielle. Ensuite, un recensement des écrits sur visant notamment à justifier le choix du $Ge_2Sb_2Se_5$ comme mémoire et à explorer l'état de l'art actuel du *hardware* dans le domaine du calcul matriciel est effectué. La présentation des simulations démontrant la viabilité de la solution choisie est subséquemment présentée, avant le recensement des résultats du processus de caractérisation du $Ge_2Sb_2Se_5$. Le procédé de fabrication de la mémoire ainsi que le montage développés au cours de ce projet feront l'objet finalement décrits, avant d'adresser la performance de la mémoire à changement de phase développée. Cette dernière étape inclut les nombreux défis rencontrés lors de l'intégration de la mémoire ainsi que la proposition d'une nouvelle architecture de calcul photonique intégré visant à maximiser la robustesse et minimiser les pertes de la multiplication matrice-vecteur.

# CHAPITRE 2 DÉFINITIONS ET CONCEPTS THÉORIQUES

## 2.1 Photonique intégrée sur silicium

La photonique ou l'optique intégrée réfère à l'utilisation de composantes optiques sur puce, contrairement à l'optique fibrée et à l'optique en espace libre. La plateforme d'intégration, soit la catégorie de matériaux utilisée pour un circuit photonique intégré (PIC), est variable et propre à l'application désirée. Parmi les technologies matures, on retrouve la photonique sur silicium (Si), basée sur une plateforme SOI (*Silicon-On-Insulator*), compatible avec les procédés CMOS et polyvalente pour l'intégration de composants passifs et actifs. La plateforme en nitrure de silicium (SiN) procure quant à elle de plus faibles pertes par propagation à 0.1 dB/m [18], ce qui en fait un matériau de choix pour des applications de détection et pour les ordinateurs quantiques [19]. Cependant, sa nature diélectrique du nitrure de silicium complexifie le développement de composants actifs. Pour l'émission laser sur puce, plusieurs plateformes peuvent être utilisées, dont le phosphure d'indium InP dans l'infrarouge rapproché (SWIR). Pour les raisons énoncées précédemment, la plateforme SOI pour photonique est offerte par plusieurs fonderies. Pour ces compagnies, la fabrication de composants photoniques est généralement triviale, puisque environ un ordre de grandeur distingue les dimensions critiques des dispositifs des limites de résolution de fabrication.

De manière générale, un procédé de photonique sur silicium actif destiné au SWIR débute à partir d'une gaufre SOI avec un oxyde enfoui (BOX) de 2 à 3 µm, et d'une couche de silicium monocristallin (100) de 200 à 400 nm. Ensuite, une nanostructuration par photolithographie UV ou par lithographie à faisceau d'électrons suivie d'une gravure plasma anisotrope du silicium définit le circuit photonique. Du germanium est ensuite ajouté par croissance épitaxiale et nanostructuré pour la photodétection/émission. Une variété d'implantations ioniques de dopants dans le silicium et le germanium sont ensuite effectuées pour ajuster les propriétés optoélectroniques des composantes actives.

Ensuite, des dépôts métalliques (Cr, W, Ti, Au, Cu...) connectant les composants actifs sont fabriqués afin de permettre le sondage électronique externe. Finalement, diverses couches de $SiO_2$ sont déposées par dépôt chimique en phase vapeur à basse pression (LPCVD) ou assisté par plasma (PECVD) au cours du procédé de fabrication agissant comme passivation, gaine des guide d'onde et isolation électrique. L'avantage d'utiliser un circuit de photonique intégrée en comparaison avec un CI électronique typique réside dans la vitesse d'opération, le parallélisme et la consommation énergétique. Les principaux désavantages sont toutefois au niveau de la difficulté des interconnexions et de la sensibilité thermique des dispositifs.

Les interconnexions englobent tout ce qui a trait à la gestion des entrées et sorties optiques de la puce. Ces interconnexions permettent donc le couplage optique planaire (e.g. *butt coupling*) ou hors plan (e.g. *grating coupling*). Le type de couplage priorisé dépend de l'état de la lumière externe (fibrée ou en espace libre), de la précision de couplage disponible, des pertes optiques envisageables et de la largeur de bande désirée. La propagation du signal, analogue aux routes métallique en électronique, est effectuée par guides d'ondes sous divers formes (*rib, strip, slot...*) générant des pertes de propagation d'environ $10^0 - 10^2$ dB/m [18,20]. Ensuite, le signal est manipulé pour produire le calcul désiré que ce soit par filtration, division, recombinaison et modulation. Pour effectuer ces manipulations, l'utilisation de coupleurs, d'interféromètres et de résonateurs est employée.

Finalement, le signal sortant peut être extrait dans le régime optique identiquement au couplage d'entrée ou dans le régime électronique via des photodétecteurs (PD) intégrés. La théorie concernant les principales structures photoniques utilisées dans ce mémoire fait l'objet des sous-sections suivantes.

## 2.2 Pertes optiques

La compréhension des interactions lumière-matière est fondamentale en optique. La nature de ces interactions peut être modélisée de différentes façons en fonction de la nature de l'onde électromagnétique et de la matière en question. En photonique intégrée, la dispersion, les pertes par propagation et les effets plasmoniques doivent être considérés lors du design d'un circuit pour en assurer son bon fonctionnement d'autant plus lorsque la technologie visée requière une bonne stabilité thermique, de grandes fréquences d'exécution et une précision de mesure considérable. L'essence théorique des divers phénomènes à considérer dans le cadre de ce projet fait l'objet des sous-sections suivantes.

### 2.2.1 Absorption diélectrique

L'absorption diélectrique est essentielle pour quantifier les pertes de propagation optique. On débute par introduire une composante imaginaire à la susceptibilité électrique $\chi$ :

$$\chi = \chi' + i\chi''. \tag{2.1}$$

Celle-ci est liée à la permittivité du matériau par l'équation 2.2 :

$$\varepsilon = \varepsilon_0(1 + \chi). \tag{2.2}$$

La composante imaginaire implique que la solution de l'équation d'Helmholtz produit un nombre d'onde imaginaire :

$$k = (1 + \chi' + i\chi'')^{1/2} k_0. \tag{2.3}$$

La composante réelle de $k$, $\beta$ représente la variation de la phase de l'onde avec $z$, alors que $\alpha$ correspond à l'absorption diélectrique :

$$\beta - \frac{i\alpha}{2} = k_0 \left(1 + \chi' + i\chi''\right)^{1/2}. \tag{2.4}$$

Jusqu'à maintenant, la susceptibilité électrique est considérée constante en fonction de la fréquence à laquelle le matériau est excité. Par contre, à part le vide, aucun médium est exactement non-dispersif (e.g. $\chi(\nu) \neq \chi_0$). Même l'air ambiant n'est pas transparent à de très courtes longueurs d'ondes et exhibe une dispersion mesurable dans le spectre visible. Or, la susceptibilité électrique, dont les composantes réelle et imaginaire sont liées à l'indice de réfraction et à l'absorption respectivement, varient en fonction de la fréquence selon les relations de Kramers-Kronig (équation 2.6 et 2.7) :

$$n + i\kappa = n + i\frac{\alpha}{2k_0} = \left(1 + \chi' + i\chi''\right)^{1/2}, \tag{2.5}$$

avec

$$\chi'(\nu) = \frac{2}{\pi} \int_0^\infty \frac{s\chi''(s)}{s^2 - \nu^2} \mathrm{d}s, \tag{2.6}$$

et

$$\chi''(\nu) = \frac{2}{\pi} \int_0^\infty \frac{\nu\chi'(s)}{\nu^2 - s^2} \mathrm{d}s. \tag{2.7}$$

En réalité, les matériaux diélectriques, via leur susceptibilité, exhibent des résonances lorsque la fréquence du champ électrique $E$ est en résonance avec les interactions électroniques modélisées par un oscillateur harmonique dont le coefficient d'amortissement est donné par $\zeta$. La susceptibilité électrique d'un milieu résonant est donc donnée par :

$$\chi(\nu) = \chi_0 \frac{\nu_0^2}{\nu_0^2 - \nu^2 + i\nu\zeta}. \tag{2.8}$$

Les matériaux diélectriques peuvent exhiber plusieurs types de résonances gouvernées par

différents mécanismes de polarisation, dont la polarisabilité électronique, ionique et dipolaire, en fonction de la nature du matériau.

### 2.2.2 Pertes par chevauchement des modes

Les pertes par chevauchement de modes sont omniprésentes en optique. Lorsque les modes de deux dispositifs sont différents, le transfert de puissance optique entre les deux est limité. Soit $\mathbf{E_0}$ et $\mathbf{H_0}$ les champs électriques et magnétiques décomposables en leur modes orthogonaux $\mathbf{e_i}$ et $\mathbf{h_i}$ tels que :

$$\mathbf{E_0} = \sum_i a_i \mathbf{e_i} \tag{2.9}$$

et

$$\mathbf{H_0} = \sum_i b_i \mathbf{h_i}, \tag{2.10}$$

où les coefficients de chevauchement $a_i$ et $b_i$ sont donnés par les intégrales de recouvrement

$$a_i = \frac{\int \mathbf{E_0} \times \mathbf{h_i^*} \cdot d\mathbf{S}}{\int \mathbf{e_i} \times \mathbf{h_i^*} \cdot d\mathbf{S}} \tag{2.11}$$

et

$$b_i^* = \frac{\int \mathbf{e_i} \times \mathbf{H_0^*} \cdot d\mathbf{S}}{\int \mathbf{e_i} \times \mathbf{h_i^*} \cdot d\mathbf{S}}. \tag{2.12}$$

La puissance électromagnétique est donnée par :

$$P_0 = \frac{1}{2} \operatorname{Re}\left\{ \int \mathbf{E_0} \times \mathbf{H_0^*} d\mathbf{S} \right\}. \tag{2.13}$$

La puissance transférée de l'onde $(\mathbf{E_0}, \mathbf{H_0})$ dans le ième mode propre est donc donnée par

$$\frac{P_i}{P_0} = \operatorname{Re}\left\{ \frac{\left(\int \mathbf{e_i} \times \mathbf{H_0^*} \cdot d\mathbf{S}\right)\left(\int \mathbf{E_0} \times \mathbf{h_i^*} \cdot d\mathbf{S}\right)}{\int \mathbf{e_i} \times \mathbf{h_i^*} \cdot d\mathbf{S}} \right\} \frac{1}{\operatorname{Re}\left\{\int \mathbf{E_0} \times \mathbf{H_0^*} \cdot d\mathbf{S}\right\}}. \tag{2.14}$$

### 2.2.3 Pertes plasmoniques

Une autre source de pertes à considérer lors du design d'un système photonique provient des métaux. Dans le cadre de ce projet, l'intégration d'éléments chauffants métallique est utilisée

pour induire un changement de phase. Celles-ci peuvent être modélisées par la résonance d'un gaz d'électrons libres. À partir de l'équation du mouvement d'un électron dans le gaz soumis à un champ électrique $\mathbf{E}$ et à un amortissement $\gamma$ [21] :

$$m\ddot{\mathbf{x}} + m\gamma\dot{\mathbf{x}} = -e\mathbf{E}, \tag{2.15}$$

on définit un champ électrique variable $\mathbf{E(t)} = \mathbf{E_0}e^{i\omega t}$, soit l'onde électromagnétique de fréquence angulaire $\omega$ :

$$\mathbf{x}(t) = \frac{e}{m\left(\omega^2 + i\gamma\omega\right)}\mathbf{E}(t). \tag{2.16}$$

Pour une densité électronique $n$ d'électrons, on obtient une polarisation macroscopique $\mathbf{P}$ :

$$\mathbf{P} = \frac{-ne^2}{m(\omega^2 + i\gamma\omega)}\mathbf{E}. \tag{2.17}$$

Avec cette expression de la polarisation et les relations de Maxwell entre les champs macroscopiques : $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$, on retrouve une expression pour la permittivité relative d'un gaz d'électron (équation 2.19) à partir de la fréquence plasma du métal $\omega_p$ :

$$\mathbf{D} = \varepsilon_0\left(1 - \frac{ne^2}{\varepsilon_0 m\left(\omega^2 + i\gamma\omega\right)}\right), \mathbf{E} \tag{2.18}$$

et

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2 + i\gamma\omega}, \text{avec } \omega_p = \sqrt{\frac{ne^2}{m\varepsilon_0}}. \tag{2.19}$$

Le coefficient d'absorption, définit précédemment pour un diélectrique, est ainsi donné en fonction de la fréquence de l'onde électromagnétique :

$$\alpha(\omega) = \sqrt{\frac{2\omega_p^2 w}{\gamma c^2}}, \tag{2.20}$$

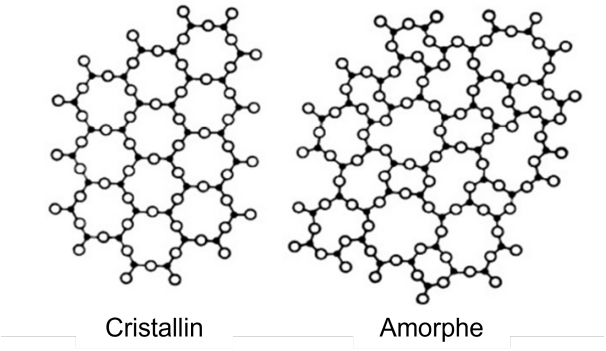
où $c$ est la vitesse de la lumière. Le coefficient $\alpha$ ($[\alpha] = \mathrm{m}^{-1}$) sera utilisé couramment pour quantifier les pertes d'absorption optique dans les divers matériaux utilisés.

## 2.3 Changement de phase

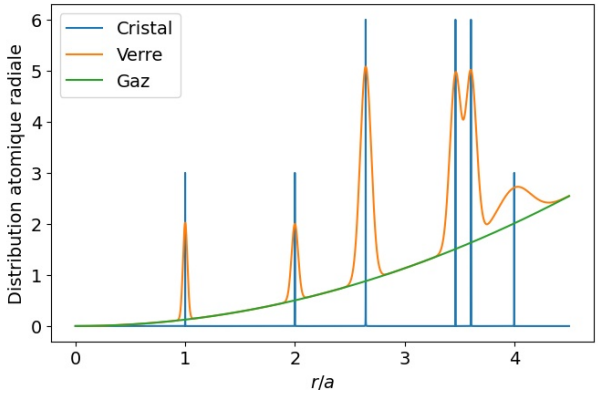
### 2.3.1 Phase amorphe et vitrification

Les matériaux à changement de phase ont la possibilité d'exhiber un changement entre deux phases ou états distincts. Il existe deux principales catégories de matériaux à changement de phase. La première s'intéresse à la différence entre l'énergie nécessaire pour atteindre un état classique métastable et sa chaleur spécifique dans cet état, à des fins de conservation d'énergie thermique. La seconde, s'intéresse au contraste des propriétés physiques entre deux phases, ou états non classiques pour manipuler les propriétés d'un système. Tout usage subséquent du terme «PCM» fera référence à cette seconde catégorie, particulièrement utile comme mémoire. Les termes «phase» et «état» seront dorénavant utilisés de manière interchangeable pour décrire l'organisation atomique d'un matériau X amorphe (noté $\alpha$-X) ou cristallin (noté c-X).

La phase amorphe ou vitreuse d'un chalcogénure à changement de phase constitue son état métastable, n'exhibant aucun ordre ou aucune périodicité de structure atomique sur longue distance (>1 nm). À cet effet, elle est identique à son état liquide ne serait-ce que pour la faible mobilité atomique caractéristique à l'état solide. La densité atomique radiale de l'état amorphe est une convolution entre l'état cristallin et gazeux tel qu'illustré à la figure 2.1 conjointement à une représentation 2D des deux états solides.




(a) Comparaison conceptuelle de l'organisation atomique d'un matériau amorphe et cristallin [22].




(b) Répartition des atomes dans divers états de la matière

FIGURE 2.1 Différences entre l'état cristallin ordonné et amorphe (vitreux).

La synthèse de ces verres n'étant pas thermodynamiquement préférentielle, exige un procédé particulier de vitrification ou d'amorphisation. Les verres sont formés en refroidissant le

matériau à partir de son état liquide assez rapidement pour ne générer aucun changement dans son volume spécifique. Ce taux de refroidissement est très variable et joue un rôle important dans le choix d'un matériau à changement de phase comme mémoire. Bien que les causes et la nature de la vitrification sont à ce jour méconnues, plusieurs modèles tentent d'expliquer le phénomène. Le consensus actuel pointe vers un hybride entre la théorie cinétique, impliquant que la vitesse d'un processus a un impact sur son comportement, et thermodynamique, pour laquelle les conditions d'un changement d'état ne sont pas satisfaites pour expliquer les observations phénoménologiques.

Le changement de phase classique est défini historiquement par Ehrenfest comme une discontinuité dans la dérivée première ou seconde de l'énergie libre $G$ d'un système par rapport à ses variables thermodynamiques. On définit l'énergie libre selon l'équation 2.21, où $H$ est l'enthalpie, $T$, la température, $S$, l'entropie, $U$, l'énergie interne, $P$, la pression et $V$, le volume du système.

$$G = H - TS = U + PV \tag{2.21}$$

Récemment, des définitions modernes plus robustes et adaptées aux particularités des matériaux émergents sont priorisées. Parmi celles-ci, on retrouve une discontinuité dans la variation de la chaleur latente du matériau. Ce changement se produit conjointement à une absorption ou émission fixe d'énergie et un mélange des deux états au point critique durant lequel, tout apport de chaleur n'augmente pas la température du matériau. Ces définitions classiques sont idéales pour décrire les transitions possibles entre les quatre états fondamentaux de la matière. Ainsi, tout phénomène thermodynamique exhibant ces conditions se doit de constituer un changement d'état. Depuis des décennies, la nature de la vitrification est questionnée et fait l'objet de paradoxes cinétiques et thermodynamiques [23, 24].

La température de transition vitreuse $T_g$ est la température ou la plage de température dans laquelle un matériau ayant été refroidit rapidement, se vitrifie. La nature de la solidification dépend donc du matériau et du taux de refroidissement. Lorsque la température du liquide initial devient inférieure à la température de fusion $T_m$, on qualifie ce dernier de liquide super-refroidit, indiscernable du liquide initial. À un certain point avoisinant $T_g$, celui-ci se solidifie subitement dans l'état vitreux. Cette transition se traduit donc, comme illustré à la figure 2.2 [25], par une chute de la capacité thermique isobare $C_p$ telle que

$$\left(\frac{\partial C_p}{\partial T}\right)_{T=T_g^{*-}} \neq \left(\frac{\partial C_p}{\partial T}\right)_{T=T_g^{*+}}. \tag{2.22}$$

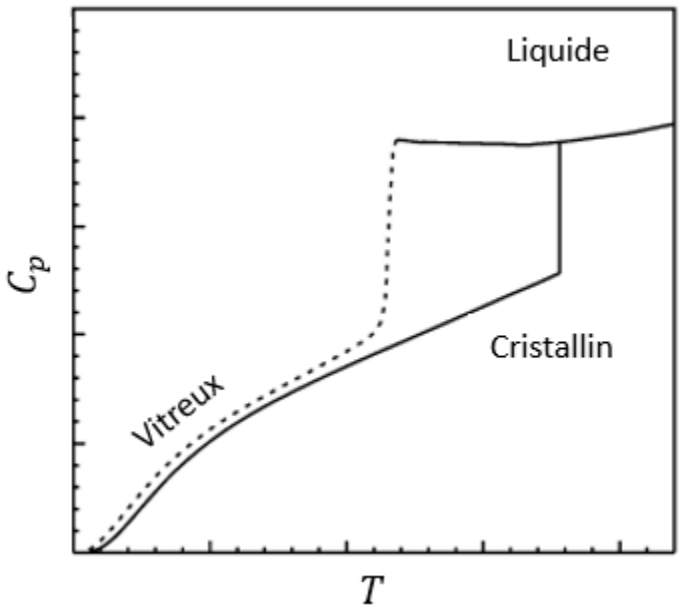



FIGURE 2.2 Changement de phase caractérisé par une abrupte variation de la capacité thermique isobare.

Dans ce contexte, la quantité $T_g^*$ est comprise dans l'intervalle de température $T_g$ (e.g. $T_g^* \in T_g$). Cette distinction est importante, car la température à laquelle la vitrification se produit varie selon les conditions expérimentales dont le taux de refroidissement et la méthode de synthèse du matériau. Puisque la discontinuité observée à $T_g^*$ n'est pas ponctuelle, des mesures calorimétriques ou dilatométriques sont généralement effectuées pour mesurer $T_g^*$ précisément.

La justification cinétique du phénomène de vitrification provient du fait que les atomes à l'état liquide, lorsque refroidis, perdent en mobilité et sont de plus en plus contraints aux sites de la maille cristalline. Si le taux de refroidissement est suffisamment élevé, le ralentissement des atomes est si précipité qu'ils ne disposent pas d'assez de «temps» pour s'ordonner. La composante cinétique de la réaction est inéquivoque, puisque $T_g^*$ augmente avec le taux de refroidissement.

Considérons deux expériences de vitrification effectuées sur deux matériaux identiques à l'état liquide refroidis relativement lentement et rapidement, notés L et R respectivement. Dans cette situation, le comportement cinétique du changement de phase est tel que $T_{g,\mathrm{R}}^* > T_{g,\mathrm{L}}^*$. Par définition de $T_g^*$, lorsque les deux matériaux atteignent une température $T_0$ $\left(T_{g,\mathrm{L}}^* < T_0 < T_{g,\mathrm{R}}^*\right)$, le matériau L est toujours à l'état liquide, alors que le matériaux R est vitreux. À partir de la définition thermodynamique de l'entropie (équation 2.23), on remarque que l'entropie du

solide vitreux à $T_0$ est supérieure à l'entropie du liquide super-refroidit, une contradiction directe de la troisième loi de la thermodynamique selon la formulation de Planck :

$$S(T_0) = S_0 + \int_0^{T_0} \frac{C_p}{T} \mathrm{d}T. \tag{2.23}$$

Cette contradiction est l'une de conclusions du paradoxe de Kauzmann, introduit par Walter Kauzmann en 1948 [26]. Ces implications font de l'amorphisation un phénomène complexe et impossible à modéliser adéquatement d'un point de vue purement thermodynamique ou cinétique en raison de son irréversibilité et de l'impact de l'histoire du matériau sur ses propriétés.

### 2.3.2 État cristallin et cristallisation

Un cristal est une structure solide avec une disposition régulière et ordonnée de ses composants atomiques ou moléculaires. Comme vu à la figure 2.1b cet ordre est périodique et permet ainsi la prédiction de l'emplacement des atomes à partir d'une maille fondamentale nommée maille élémentaire. Bien que le terme «cristal» est parfois utilisé pour décrire un matériau exhibant une périodicité atomique sur tout son volume, son utilisation dans ce mémoire sera analogue à la notion de polycristal décrivant plus adéquatement la structure atomique des PCMs en phase cristalline.

Les matériaux cristallins sont couramment étudiés et utilisés pour leur propriétés particulières générées par leur arrangement atomique. Plus précisément, leur réseau de Bravais et les symétries supportées par ce dernier ainsi que la composition du matériau lui permettent d'exhiber des propriétés telles que la biréfringence, piézoélectricité, pyroélectricité, ferroélectricité etc. On remarque également que pour un matériau donné, sa conductivité électrique peut varier de plusieurs ordres de grandeur selon sa structure atomique :

$$\sigma_\alpha \ll \sigma_\mathrm{c}. \tag{2.24}$$

La synthèse de couches minces monocristallines est effectuée par croissance épitaxiale, un procédé dispendieux et très lent en comparaison avec les procédés de dépots conventionnels. De plus, elle nécessite une concordance quasi parfaite entre la maille élémentaire du substrat et celle du matériau à croître. Les dépôts de matière à l'état amorphe ou polycristallin, en revanche, peuvent être effectués par toutes autre méthode, dont la pulvérisation, l'évaporation et le dépôt chimique en phase vapeur (CVD). La nature de l'arrangement atomique est dicté davantage par les conditions de dépôt que par la méthode. Par exemple, le dépôt de silicium

par LPCVD produit un arrangement polycristallin lorsque effectué à 630 °C et à amorphe 570 °C. Pour la synthèse en gros volumes (bulk) les matériaux vitreux sont fabriqués par *melt-quenching*, un procédé basé sur le mélange du composé en phase liquide suivi d'un refroidissement rapide pour causer la vitrification.

Comme vu précédemment, la cristallisation chez les PCMs est la transition thermodynamiquement favorable. L'état cristallin est à l'équilibre thermodynamique et possède donc une énergie interne $U$ inférieure à son analogue amorphe. On définit la température de cristallisation $T_c$ comme la température à laquelle un matériau amorphe se cristallise spontanément. Cette température de cristallisation est empiriquement liée à la température de transition vitreuse pour les chalcogénures selon $T_c \approx 1.4T_g$ [27]. Similairement à la transition vitreuse, la température de cristallisation dépend des conditions de dépôts du matériau et est plus adéquatement décrite par une plage de températures. De plus, certains composés peuvent se cristalliser selon divers réseaux de Bravais en fonction des conditions thermodynamiques.

La cristallisation se décompose en deux étapes fondamentales, soit la nucléation et la croissance. La nucléation est essentielle à la formation de la plupart de cristaux (sauf pour ceux fabriqués par épitaxie). Elle débute avec la formation de petites agrégations de particules appelées germes ou noyaux. Ces germes sont formés lorsque des particules individuelles, appelées molécules, atomes ou ions, s'assemblent de manière ordonnée dans une structure cristalline. La formation des germes est une étape cruciale, car elle représente une barrière énergétique à franchir pour passer d'un état désordonné à un état cristallin. Cette nucléation peut être homogène si les germes se forment spontanément lorsque des fluctuations aléatoires dans la distribution des particules conduisent à une surconcentration locale qui atteint un seuil critique, ou hétérogène à partir d'impuretés, au quel cas ces sites offrent des surfaces structurales où les particules peuvent s'arranger plus facilement dans un motif cristallin [28]. Ce germes deviennent ensuite les points de départs de la croissance cristalline qui se répartit sur l'ensemble du matériau. La nature et densité de ces sites influence la température et la vitesse de cristallisation qui sont intimement liés à la structure cristalline préférentielle.

### 2.3.3 Ségrégation moléculaire et *dewetting*

La ségrégation moléculaire est un effet indésirable du changement de phase. Elle se traduit, dans les matériaux hétéronucléaires comme les dichalcogénures et les trichalcogénures, par la migration de certains éléments ou molécules dans un alliage formant des agrégats suite à un traitement thermique. Ce phénomène affecte la durabilité des matériaux à changement de phase, puisque le déplacement des atomes généré lors de la ségrégation est quasi-irréversible. Ceci se traduit ensuite par une perte d'homogénéité graduelle du matériau.

Le phénomène de *dewetting* se produit dans une couche mince lorsque celle-ci se fragmente en gouttelettes ou en îlots discontinus. Ce processus se produit généralement sur une surface solide lorsque la tension superficielle de la couche mince devient suffisamment élevée pour surmonter les forces d'adhésion qui la maintiennent étalée.

Au début, une fine couche de matériau est déposée sur une surface solide, formant ainsi une couche continue. Au fur et à mesure que la couche mince s'étale sur la surface, des instabilités peuvent commencer à se former. Ces instabilités peuvent être causées par différentes forces, telles que les gradients de tension superficielle, les différences d'épaisseur ou de viscosité du matériau. Lorsque les instabilités deviennent suffisamment prononcées, des gouttelettes commencent à se former dans la couche mince. Ces gouttelettes sont généralement causées par la coalescence des instabilités locales en des zones plus denses. Ensuite, les gouttelettes nouvellement formées ont tendance à croître aux dépens de la couche mince environnante, car elles accumulent le matériau délaissé par le retrait de la couche. Si les gouttelettes continuent de croître, elles peuvent finir par se fusionner les unes avec les autres, formant ainsi des structures plus grandes.

Le phénomène de *dewetting* est courant dans de nombreux systèmes de revêtement mince, tels que les revêtements polymères, les films métalliques ou les revêtements de surface. Il peut être indésirable dans certaines applications, car il entraîne une distribution inhomogène du matériau et une mauvaise adhérence à la surface.

## 2.4 Réseaux de neurones artificiels

L'intelligence artificielle est omniprésente dans la société. Plusieurs tâches à priori bénignes sont le résultat d'algorithmes d'IA basés sur une grande quantité de calculs. L'IA couvre une multitude de domaines dont l'apprentissage machine (ML), la robotique et le traitement de langues naturelles. L'apprentissage machine est le terme général utilisé pour tout modèle effectuant des prédictions à partir de données qui lui sont procurées préalablement, sans instruction précise sur le traitement de celles-ci. Parmi ces modèles, on retrouve les réseaux de neurones artificiels (NN), fondements de l'apprentissage profond (DL). Tels des neurones transmettant des impulsions électriques dans le cerveau par l'entremise des connexions synaptiques, les réseaux de neurones artificiels, transmettent un vecteur d'entrées à travers plusieurs couches de neurones artificiels connectées pour en produire un vecteur de sortie, soit une prédiction, nommée inférence. On parle de DL lorsque le réseau contient une multitude de couches intermédiaires, augmentant significativement la quantité de connexions synaptiques et ainsi la complexité du modèle. L'objet de cette section est d'illustrer la nature des calculs nécessaires à l'entraînement et à l'inférence de réseaux de neurones conventionnels

afin de justifier le choix d'architecture pour l'accélérateur optique proposé.

Comme la majorité des modèles de ML, les NNs servent à produire une inférence sous la forme de classification (prédiction qualitative) ou de régression (prédiction quantitative). Les informations utilisées par le modèle sont nommées caractéristiques ($x_0$), alors que les prédictions sont nommées étiquettes ($y$). De manière générale, la propagation avant d'un réseau de neurones peut être représentée par une fonction $\mathcal{F}$ telle que

$$y = \mathcal{F}_{\gamma,p}(x_0) \tag{2.25}$$

où

- $\gamma$ représente les hyperparamètres du modèle, dictant sa taille et la manière dont il est entraîné ;
- $p = (w, b)$ représente les paramètres du modèle, sous-forme de poids $w$ et de biais $b$, dictant les prédictions de celui-ci ;
- $x_0$ représente le vecteur de *features* en entrée.

Le développement d'un modèle de ML suit généralement les étapes suivantes :

1. Prétraitement : Transformation des données brutes en caractéristiques et segmentation en trois jeux de données : d'entraînement, d'évaluation et de test
2. Entraînement : Établissement des paramètres du modèle en minimisant l'erreur sur un jeu de données spécifique à l'entraînement
3. Évaluation : Optimisation des hyperparamètres du modèle à l'aide du jeu de données spécifique à l'évaluation
4. Test : Caractérisation de la performance du modèle avec le jeu de données spécifique au test

Établissons d'abord le formalisme mathématique et le fonctionnement des réseaux de neurones artificiels. Par souci de simplicité, le NN illustré ci-dessous représente un modèle à propagation avant simple. Il ne considère pas de récurrence, de convolutions ou tout autre artefact de calcul permettant d'améliorer ses performances. Considérons un vecteur initial de caractéristiques noté $x_0$ se propageant à travers les $k$ couches de $n_{1,2,\ldots,k}$ neurones respectivement du réseau selon l'équation :

$$x_{i+1,j} = a_{i+1,j}\left(\left(\sum_j w_{i+1,j}x_{i,j}\right) - b_{i+1,j}\right), \tag{2.26}$$

où $x$ est la valeur du neurone, $w$ est le poids liant deux neurones, $b$ est la biais appliqué au neurone et $a$ est sa fonction d'activation. Avec cette notation, l'indice $i \in \{1, 2, ..., k\}$ réfère à la couche neuronale et $j \in \{1, 2, ..., n_k\}$ réfère au neurone dans la $k$ième couche. La fonction d'activation du neurone a pour but d'ajouter une composante non-linéaire au modèle afin de modéliser des phénomènes plus complexes. Un exemple de fonction d'activation est la fonction sigmoïde (voir équation 2.27), utilisée en classification pour favoriser la discrétisation binaire du neurone,

$$S(x) = \frac{1}{1 + e^{-x}}. \tag{2.27}$$

Cette représentation, bien que mathématiquement correcte, est complexe. Par souci de simplicité, on utilise généralement la notation vectorielle, où la valeur de tous les neurones de la $i$ème couche sont répertoriés dans le vecteur $\mathbf{x_i}$ et les poids liants tous les neurones de la $i$ème à la $i+1$ième couche dans la matrice $\mathbf{W_{i+1}}$. La fonction d'activation n'est pas nécessairement identique pour tous les neurones et toutes les couches, on utilise tout de même la notation simplifiée suivante,

$$\mathbf{x_{i+1}} = \mathbf{a_{i+1}} \left( \mathbf{W_{i+1}} \mathbf{x_i} - \mathbf{b_{i+1}} \right). \tag{2.28}$$

En somme, c'est un consensus que les opérations matricielles sont les plus coûteuses en terme de temps dans l'entraînement et l'inférence des modèles d'IA [29–31]. L'accélérateur optique à changement de phase proposé dans ce mémoire est idéal pour traiter de la linéarité et du volume d'opérations rapidement et efficacement.

## 2.5 Planarité des multiplications matrice-vecteur

La multiplication matrice-vecteur (MVM) peut être représentée par un graphe directionnel, où les éléments du vecteur $\mathbf{x}$ interagissent avec les éléments de la matrice $A$ pour produire le vecteur $\mathbf{y}$ en sortie. Cette représentation 2D est utile en design de CIs et de circuits photoniques, afin de minimiser le nombre de croisements. En électronique, le croisement direct de deux routes métalliques est inviable et doit être adressé par pontage entre deux couches conductrices par l'entremise de vias. Dans le cas d'une carte de circuits imprimés (PCBs), des cartes comprenant plusieurs couches conductrices sont employées, alors que pour les CIs, des étapes de microfabrication additionnelles sont nécessaires pour effectuer les pontage électronique. Ces deux solutions engendrent des coûts additionnels.

Le nombre de croisements minimums dans un circuit peut être déterminé en représentant le

circuit sous forme de graphe. Le graphe trivial de la MVM ($N = 2$) est illustré à la figure 2.3 de manière intuitive et non-planaire. Ce graphe est toutefois planaire par le théorème de Kuratowski [32] comme illustré à la figure 2.4.

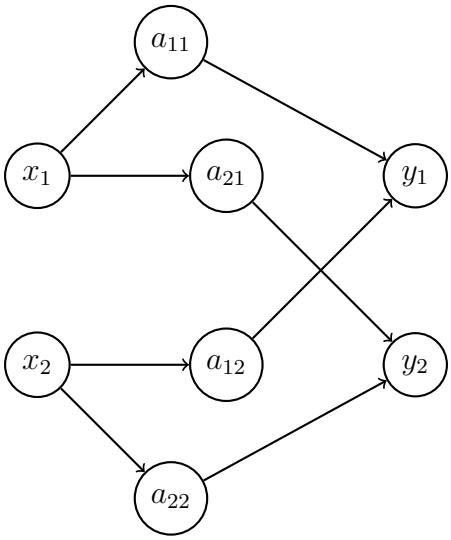



FIGURE 2.3 Représentation non planaire de la multiplication matrice-vecteur ($N = 2$)

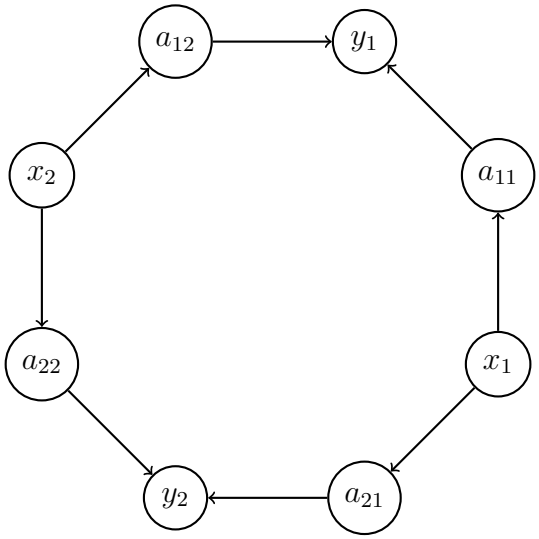



FIGURE 2.4 Représentation planaire de la multiplication matrice-vecteur ($N = 2$)

Il s'avère que tous les graphes de MVM pour $N \geq 3$ sont non-planaires et nécessitent l'emploi de croisements, tel que montré à la figure 2.5.

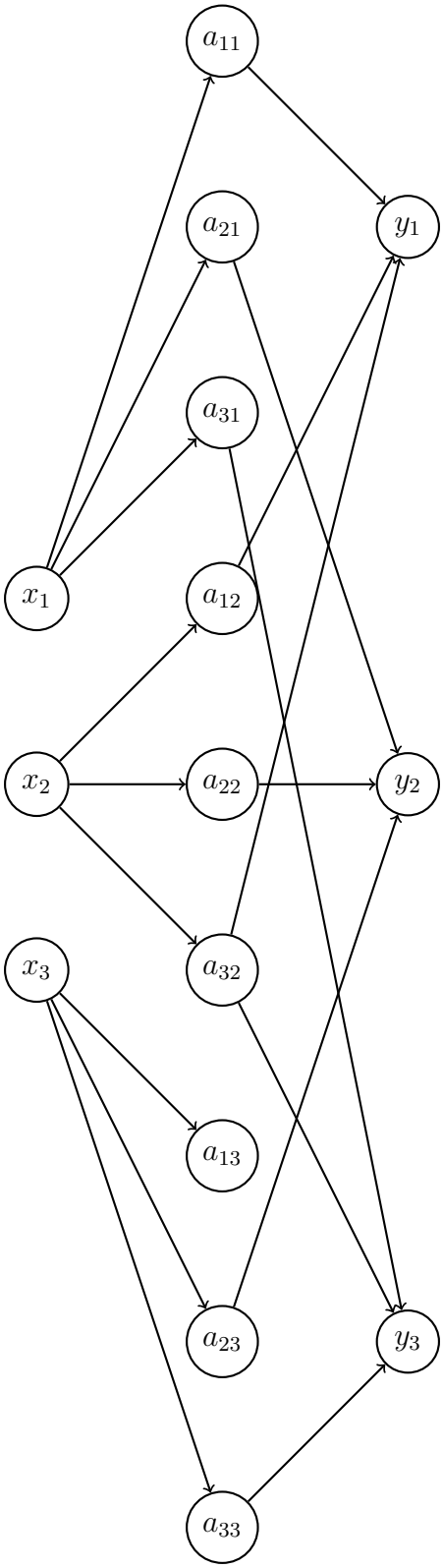



FIGURE 2.5 Graphe représentant la multiplication matrice-vecteur ($N = 3$)

Ceci est une propriété fondamentale des MVMs représentés de la sorte. Cette représentation sera utilisée à la section 7.3 pour déterminer le nombre minimal de croisements nécessaires sur une puce bidimensionnelle, mais aussi les défis rencontrés en terme de balancement et de pertes pour des matrices de taille supérieure.

## 2.6 Effet Joule

L'effet Joule désigne le phénomène de conversion de l'énergie électrique en chaleur lors du passage d'un courant électrique à travers un conducteur électrique. Ce phénomène est utilisé par les éléments chauffants permettant le changement de phase électrothermique des mémoires. D'un point de vue microscopique, ce phénomène est modélisé par l'accélération des porteurs de charge due au champ électrique causant des collisions entre les quasi-particules du conducteur, transférant de l'énergie cinétique à la maille qui est subséquemment relâchée sous forme radiative. L'équation 2.29 gouverne la puissance dissipée par effet Joule,

$$P = I^2 R. \tag{2.29}$$

L'effet Joule est indésirable dans la plupart des applications en raison de la dépendance en température de la résistivité des matériaux conducteurs et de la dissipation de chaleur nécessaire. Cette relation est généralement considérée linéaire pour des métaux :

$$\rho(T) = \rho_0 \left(1 + \alpha(T - T_0)\right), \tag{2.30}$$

où

$$\rho_0 = \rho(T = T_0). \tag{2.31}$$

Le taux d'augmentation de la résistivité ($\alpha$) pour les métaux communs comme l'aluminium, le cuivre et l'or varie entre $10^{-5} - 10^{-3}$ $\Omega$ m$/K$. Dans le cadre de ce projet, l'effet Joule est utilisé dans les éléments chauffants permettant le changement d'état de la mémoire optique. Il est donc nécessaire de maximiser le courant traversant l'élément chauffant ainsi que sa résistance pour obtenir l'effet désiré.

Maintenant les concepts fondamentaux établis, une revue de littérature est présentée pour justifier l'étude des mémoires optiques non-volatiles ainsi que de leur intégration potentielle dans un accélérateur optique.

# CHAPITRE 3 REVUE DE LITTÉRATURE

## 3.1 Intelligence artificielle

La motivation primaire appuyant le développement d'un accélérateur photonique provient des lacunes actuelles du paradigme électronique. Le progrès constant poussé par l'invention d'accélérateurs dédiés tel le GPU et le TPU, bien qu'ayant permis la création de modèles impressionnants comme ChatGPT de OpenAI et Autopilot de Tesla, demeure incrémental et limité par la durée d'entraînement et d'inférence.

### 3.1.1 Demande en intelligence artificielle

Le marché global de l'intelligence artificielle est évalué à plus de 500 milliards de dollars américains en 2023 et est estimé à connaître une croissance annuelle moyenne de plus de 20% jusqu'en 2030 [33]. De plus, le marché du calcul informatique via *Cloud*, la pierre angulaire de l'IA aujourd'hui, est évalué à près de 700 milliards de dollars US. La quête de l'intelligence artificielle généralisée (AGI) est illustrée par l'engouement généré par GPT-4. ChatGPT, le modèle de langues naturelles basé sur GPT-4, a atteint le plateau des 100 millions d'utilisateurs mensuels deux mois après son dévoilement, établissant un record de croissance web [34]. En comparaison, il a fallu plus de deux ans au réseau social Instagram avant d'atteindre un niveau d'engagement similaire [35]. Bien que la nature des modèles de la suite GPT (*Generative Pre-Transformer*) n'est pas publique, leur taille approximative en terme de nombre de paramètres est connue et présentée à la figure 3.1 [36]. Il est aussi établi que GPT-4 a coûté plus de 100 millions de dollars américains à entraîner [37]. Invariablement, la performance de ces modèles est fortement influencée par leur taille et la quantité de données avec laquelle ils sont entraînés [38, 39].

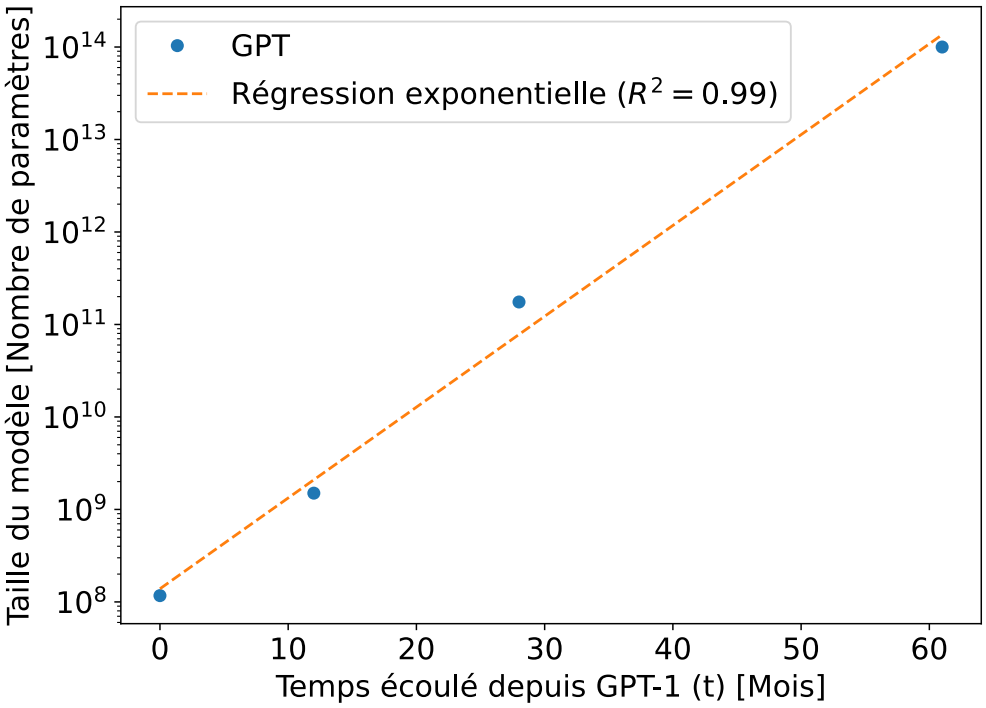



FIGURE 3.1 Croissance de la taille des modèles GPT de OpenAI depuis GPT-1.

Cette tendance est représentée à la figure 3.2 [40], où l'on observe la croissance exponentielle des besoins en entraînement dans les dernières années, et ce, particulièrement pour les modèles de type GPT, soient les modèles de langages et de voix.

FIGURE 3.2 Envergure de l'entraînement de modèles d'intelligence artificielle populaires.

La croissance de l'IA s'étend au-delà des modèles à l'état de l'art comme GPT-4. L'étendue d'applications de l'intelligence artificielle est également en expansion. Il est établi que les problèmes célèbres, comme la vision par ordinateur, la traduction instantanée, la conduite automatisée, le repliement des protéines et le problème du voyageur de commerce ont bénéficié significativement de l'émergence de l'IA. Il est également important de considérer la panoplie d'applications moins connues comme l'aide au diagnostic médical en médecine [41–43], les robots conversationnels pour le service à la clientèle [44,45], les publicités ciblées en marketing [46, 47] et la prédiction des marchés en finance [48, 49] qui bénéficient de cette technologie.

### 3.1.2 Solutions actuelles : GPU, FPGA, ASIC

Il a été établi que l'intelligence artificielle et la demande en calcul qu'elle génère est en croissance fulgurante. L'unité de comparaison appropriée pour comparer la performance des processeurs en terme d'apprentissage machine est le FLOPS. Elle correspond au nombre d'opérations à virgule flottante (FLOP) effectuées par seconde par un processeur. Le nombre de FLOPS d'un processeur est donc donné par [50] :

$$\#\ \text{FLOPS} = \#\ \text{coeurs} \times f_{\text{clock}} \times \frac{\#\ \text{FLOP}}{cycle}. \tag{3.1}$$

L'utilisation de GPUs pour l'apprentissage machine date de la fin des années 90 [51, 52]. Ces processeurs, originalement conçus pour accélérer le rendement graphique 3D, sont toujours considérés comme la meilleure solution locale en terme d'apprentissage profond, en raison de leur grand nombre de coeurs, permettant un parallélisme inégalé par les CPUs. Aujourd'hui, les processeurs centraux à l'état de l'art comme le Intel®Core™ i9-13900KS [53] disposent de dizaines de coeurs et opèrent à des fréquences d'horloge atteignant de 5 à 6 GHz. En revanche, un GPU comme le NVIDIA Titan RTX [54], dispose de plus de 5000 coeurs en maintenant une fréquence d'horloge de près de 3 GHz. De plus, les GPUs disposent d'une mémoire vive interne 100 à 1000 fois plus grande que la mémoire cache d'un processeur central, rendant les CPUs significativement plus vulnérables aux limitations de Von-Neumann.

Un des problèmes majeurs lié à l'utilisation de GPUs est leur consommation énergétique et leur latence. En effet, les GPUs, en moyenne, requièrent deux à six fois la puissance électrique pour fonctionner [55] en plus de générer une forte latence imprévisible, de par les nombreux échanges entre le CPU et le GPU lui-même [56]. Pour remédier à ce problème, Google a développé le processeur tensoriel, spécifiquement conçu pour effectuer efficacement plusieurs opérations tensorielles en parallèle. Les TPUs sont des infrastructures informatiques uniquement accessibles via *Cloud*. Ils sont constitués d'une composante dédiée aux multiplications matricielles, une composante vectorielle et une unité scalaire permettant d'effectuer jusqu'à 16 000 opérations par cycle d'horloge [57]. La dernière version (TPUv4) de Google produit 275 TFLOPS pour une consommation énergétique comparable à celle d'un CPU moderne haut de gamme.

La troisième alternative au CPUs pour le ML est le FPGA. Un FPGA est un circuit intégré spécifique conçu pour être configuré par un utilisateur après sa fabrication, offrant une flexibilité considérable. Les FPGAs sont des solutions de calcul qualifiées hétérogènes et séquentielles, car elles ne nécessitent pas d'appels d'instructions du processeur, mais traitent les données directement selon la configuration préalablement établie. Ainsi, les FPGAs ne sont

pas Von Neumann contraints et leur performance est liée à la vitesse de transport des données d'un bloc logique au prochain [58]. Ils sont considérés comme des solutions viables pour des applications de classification en temps réel dans des environnements limités en puissance de calcul [59]. Leur performance a été démontrée dans diverses conditions [60, 61].

Une comparaison qualitative des forces et des faiblesses des ASICs les plus populaires est présentée au tableau 3.1 [62].

Tableau 3.1 Comparaison qualitative de divers accélérateur électroniques utilisés pour l'intelligence artificielle.

| **Attributs** | **FPGA** | **GPU** | **TPU** |
|---|---|---|---|
| Performance | Faible | Modérée | Élevée |
| Latence | Modérée | Élevée | Faible |
| Efficacité énergétique | Modérée | Bonne | Élevée |
| Évolutivité | Faible | Modérée | Élevée |

Avec l'étendue d'applications de l'apprentissage machine [63] et la croissance de l'Internet des objets (IoT), le besoin de solutions dites «*Edge*» est considérable. Les applications *Edge* constituent l'exécution de modèles d'apprentissage automatique localement plutôt que sur des serveurs. Ces modèles sont donc conçus pour fonctionner à partir d'une puissance de calcul limitée, sur des téléphones intelligents, des drones, des voitures ou les machines d'une chaîne de production. Typiquement, ces applications ne disposent donc pas d'assez de ressources pour attribuer un ordinateur complet (CPU, GPU, RAM, stockage) à chaque maillon de la chaîne et doivent se fier à des ASICs préprogrammés pour traiter l'information virtuellement sans latence [64].

## 3.2 Calcul à faible précision

Une des avancées en apprentissage profond au cours de la dernière décennie est l'amélioration de l'exactitude d'un modèle observée au détriment de sa précision. À cet égard, on définit l'exactitude comme la qualité de l'inférence produite par un modèle, alors que sa précision correspond au nombre de bits attribués à ses variables. Par défaut, les deux librairies d'apprentissage majeures : TensorFlow et Pytorch, utilisent une précision de 32 bits par défaut [65, 66]. Cette section est dédiée aux récents travaux publiés dans le domaine du calcul à faible précision destiné à l'apprentissage machine. Cette solution potentielle peut sembler purement logicielle, mais le design d'un ASIC à faible précision comme le TPU de Google (8 bits) est un parfait exemple de l'impact du *hardware*, ayant permit de quadrupler le nombre

d'opérations par watt. Initialement, le calcul à faible précision était destiné uniquement à l'inférence, puisqu'il était assumé que la précision d'entraînement de 32 ou même 64 bits était un paramètre trop critique sur la précision globale du modèle. Or, il s'avère que l'entraînement, pour certaines applications, bénéficie considérablement d'une vitesse plus élevée et d'une consommation énergétique moindre. L'exactitude des modèles de classification d'image entraînés et inférés à 10 bits est rapportée similaire ($\Delta\text{Err} < 0.3\%$) à leur analogue en précision complète (32 bits) [67] pour les jeux de données MNIST et CIFAR-10. Ces travaux montrent que pour le jeu de données SVHN, la demie-précision est nécessaire pour atteindre une exactitude similaire. Actuellement, des algorithmes de propagation arrière à haute exactitude et faible précision optimisés comme le HALP font l'objet de recherche active. Cette méthode à quart de précision (8-bit) est inspirée d'une propagation arrière de gradient stochastique à variance réduite (SVRG), mais inclut une étape de centralisation de bit pour réduire le bruit induit par le manque de précision. La centralisation de bit est représentée à la figure 3.3 : elle constitue, une normalisation de l'étendue des solutions cherchant un µ satisfaisant l'équation 3.2 [68].

$$||w - w^*|| \leq \frac{1}{\mu}||\nabla f(w)|| \tag{3.2}$$

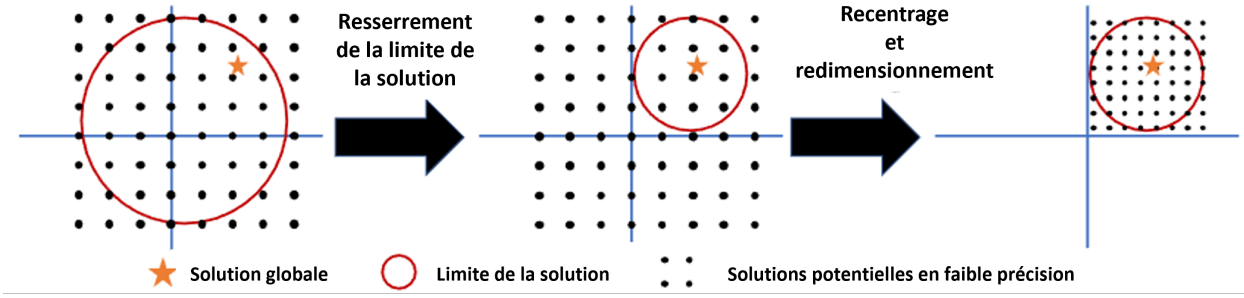



FIGURE 3.3 Représentation de la centralisation des bits, une étape fondamentale des algorithmes de ML basés sur une faible précision.

Dans cette équation, $w$ est la solution trouvée par la propagation arrière, $w^*$ la solution réelle du problème d'optimisation et $f$, la fonction à optimiser. Cette méthode a rapporté des résultats similaires sur CPU à des modèles entraînés et exécutés à double précision, soient un à mémoire à court-terme, à réseau de neurones convolutionnel et à régression logistique multi-classe. En utilisant une HALP et une SGD à quart de précision on réduit la durée d'entraînement d'un facteur 2 et 4 sur des jeux de données MNIST et synthétiques à 10000 caractéristiques respectivement par rapport à une descente de gradient stochastique à

précision complète [67]. Ceci indique la polyvalence et la perfomance des modèles à précision réduite lorsque les algorithmes d'entraînement sont adaptés à la précision. En apprentissage profond, vu la taille des modèles utilisés, la réduction de précision est encore plus avantageuse. Pour l'inférence en reconnaissance, la performance atteinte d'un modèle à 8 bits est illustrée au tableau 3.2.

Les travaux de recherche en intelligence artificielle sont généralement étalonnés selon des jeux de données standardisés. Dans les études citées, ces jeux de données sont appliqués à la reconnaissance d'image. Ils constituent des images de tumeurs au sein bénignes ou malignes [69], de champignons comestibles ou non [70], d'espèces d'iris [71], de chiffres écrits à la main [72], de types de vêtements [73], et d'images diverses [74]. Un exemple de ces jeux de données est illustré en annexe A.

Tableau 3.2 Exactitude d'inférence de modèles selon leur précision pour divers jeux de données.

| **Jeu de données** | **8 bits** | **32 bits** |
|---|---|---|
| WI Breast Cancer | 85.9 | 90.1 |
| Mushroom | 96.4 | 96.8 |
| MNIST | 98.5 | 98.5 |
| Iris | 96.0 | 98.0 |
| Fashion MNIST | 89.6 | 89.5 |

Cette réduction poussée à 4 bits a également rapporté une vitesse d'entraînement sept fois plus rapides qu'à 8 bits pour des modèles de classification multivariables, des réseaux de neurones convolutionnels et des réseaux de neurones à mémoire courte [75]. Le calcul à 2 bits a également été exploré en reconnaissance d'image à partir d'un accélérateur 2-bit dédié produisant 1 TFLOP/mm$^2$ [76]. Cette publication rapporte une $\Delta\text{Err} < 2\%$ pour l'inférence à 2 bits par rapport à l'inférence en précision complète. Bien-sûr, ce ne sont pas toutes les applications qui bénéficient du compromis en précision. Dans certains contextes de modélisation et simulations physiques comme la trajectoire d'orbites planétaires, la prédiction du climat et la diffraction des quarks, la précision double n'est pas suffisante [77].

Les arguments énoncés en faveur non seulement de l'utilisation de modèle à faible précision, mais aussi du développement d'ASICs à faible précision s'appliquent particulièrement à l'accélérateur optique proposé. Les mémoires optiques constituant un défi significatif dans le développement d'un tel accélérateur, l'encodage à faible précision favorise leur utilisation.

## 3.3 Mémoires à changement de phase

Comme mentionné précédemment à la section 2.3, il existe une variété de matériaux à changement de phase dont les chalcogénures, qui semblent idéaux pour agir en tant que mémoire non-volatile, et ce, en raison de leur stabilité et de leur durabilité. La section qui suit illustre les meilleurs matériaux à changement de phase dédiés à des applications de mémoire répertoriés dans la littérature. Les caractéristiques d'un bon matériau à changement de phase sont répertoriées dans le tableau 3.3.

TABLEAU 3.3 Caractéristiques d'un bon matériau à changement de phase pour mémoire non-volatile

| **Caractéristique** | **Symbole (Unités)** |
|---|---|
| Température de fusion basse | $T_f$ $(K)$ |
| Température de cristallisation assez élevée* | $T_c$ $(K)$ |
| Faible taux de refroidissement minimal | $\mathrm{d}T_{min}^{(a)}/\mathrm{d}t$ $(K/s)$ |
| Longue stabilité à l'état amorphe | $\tau_a$ $(s)$ |
| Longue durabilité | $N_{\text{cycles}}$ |
| Fort contraste des propriétés | $\Delta n$ ou $\Delta\sigma$ (S/m) |

* Température 3 à 5 fois plus élevée que la température d'opération de la mémoire pour maximiser la rétention de la phase amorphe, mais minimiser la quantité d'énergie nécessaire pour engendrer la cristallisation.

### 3.3.1 Matériaux considérés

Un des premiers matériaux à changement de phase est le AgInSbTe (AIST) en raison de son utilisation comme mémoire dans les disques compacts. Il s'agit d'un chalcogénure bien documenté, le SbTe, dopé avec de l'argent et de l'indium pour augmenter sa réflectivité dans le visible, une caractéristique essentielle au fonctionnement de lecteurs CD.

Lorsque riche en Sb, il maintient une température de cristallisation assez élevée (190 °C) ce qui permet de favoriser la rétention des données dans la phase amorphe et prévenir des *bitflips* involontaires, un effet indésirable des PCMs riches en Te. Puisque les CDs, DVDs et Blurays utilisent des lasers de 780, 650 et 405 nm respectivement, le contraste visé entre ses phases se produit majoritairement dans le visible. Le AIST riche en Sb présente une différence de réflectivité de 12 à 20 % dans le visible. Ceci se traduit par un faible contraste optique $\Delta n \approx 1$ mais une différence d'absorption $\Delta\kappa \approx 2$ à une longueur d'onde de 800 nm. La bande d'énergie interdite du Sb-AIST varie de 0.265 eV (4.67 µm) en phase amorphe à

1.127 eV (1.1 µm) lorsque cristallin [78].

Son point de fusion approximatif selon sa stochiométrie est de 540 °C [79] ce qui en fait une mémoire stable, mais dont la fusion est atteignable à faible coût énergétique. Cette stabilité s'étend sur 5 à 20 ans en fonction de la technologie du disque [80]. Bien qu'il soit traditionnellement utilisé exclusivement comme mémoire optique, le AgInSbTe présente un contraste électrique comparable aux PCMs électroniques comme le $Ge_2Sb_2Te_5$, soit de 4 à 6 ordres de grandeur [81]. Sa durabilité est assez faible, limitée commercialement à $10^3$ cycles de réécriture, mais un choix de couchage multi-couche judicieux montre une durabilité de $10^5$ - $10^6$ cycles lorsque le matériau est modifié électriquement [82, 83].

Le $Ge_2Sb_2Te_5$ (ou GST-225), est le chalcogénure le plus utilisé dans l'industrie électronique moderne. Il est considéré comme le PCM à la base des technologies PRAM (*Phase-change random-access memory*) modernes, développées par *STMicroelectronics*, *Samsung*, *IBM*, *Intel* et *Numonyx*. Le PRAM est considéré comme un candidat potentiel pour remplacer les mémoires Flash [84] que ce soit en intégration hétérogène ou hybride [85].

Il se cristallise d'abord dans une phase métastable cristalline cubique à face centrée (cfc) à 150 °C et se stabilise dans un arrangement hexagonal dense (hcp) au delà de 230 °C [86]. Cette faible température de cristallisation implique une mauvaise stabilité de la phase amorphe qui se traduit par un taux de refroidissement pour l'amorphisation complète de $10^{10}$ °C/s. Son indice de réfraction varie de $4 \pm 0.5$ à $7 \pm 0.5$ lors de la cristallisation, alors que son absorption augmente de $\kappa \approx 0.8$ à $\kappa \approx 2$ lors de la même transition [87, 88]. Ceci montre un contraste très élevé et des pertes significatives dans les deux états. Sa durabilité s'étend à plus de 1 millions de cycles de réécriture. Son point de fusion est de 600 °C, ce qui le rend accessible à amorphiser. Son utilisation comme mémoire a changement de phase électronique est justifiée par un chute de 5 à 6 ordres de grandeur en résistivité lors de la transition amorphe à cristalline hexagonale. Il possède une lanrgeur de bande $E_g^{opt}$ amorphe de 0.75 eV et cristallin de 0.5 eV pour les phases cfc et hcp [89].

Les pertes significatives du GST-225 dans l'état amorphe sont problématiques pour la mise à l'échelle de mémoires optiques. Pour remédier à ceci, le remplacement du Te par une chalcogène plus léger est suggéré, formant la catégorie des GSST (GeSbSeTe). Cet ajout de sélénium se traduit par une diminution du contraste optique $\Delta n$. Le GSST-2241 subit une double cristallisation similairement au GST-225 et exhibe un contraste de propriétés similaires, alors que la cristallisation du GSST-2223 est plus progressive et distinguée par davatange de liaisons Se-Ge/Sb, caractéristiques d'une structure cristalline orthorhombique [90].

### 3.3.2 Méthode de dépôt

Les propriétés des chalcogénures en couche mince varient en fonction de la méthode de dépôt utilisée. Leur durabilité, par exemple, dépend des traitement thermiques appliqués et imposent un certain budget thermique lors du dépôt. Les méthodes de dépôt des mémoires qui produisent une couche uniforme amorphe sont priorisées, car celles-ci préviennent un conditionnement non désiré, exhibant de la ségrégation moléculaire ou une cristallisation métastable. Les chalcogénures à changement de phase sont considérés comme faciles à déposer, en raison de leur faible température de fusion et d'évaporation. Ainsi, la plupart des méthodes de dépot conventionnelles, tel que l'évaporation, la pulvérisation et le dépôt par laser pulsé (PLD) peuvent être utilisées.

La composition et la conformité du dépôt sont les deux caractéristiques les plus importantes du dépôt. Pour assurer un dépôt amorphe uniforme, la synthèse de la cible est généralement effectuée par *melt-quenching*. Cette synthèse est caractérisée par le mélange de poudres à pureté très élevée (4 ou 5N) des éléments à composition désirée dans une ampoule en quartz. Ce mélange est maintenu à plus de 1000 °C durant plusieurs heures afin d'assurer l'homogénéité du mélange. Le refroidissement rapide propre à l'amorphisation est assuré en déposant l'ampoule dans un bain d'eau glacé afin de solidifier le mélange. Ensuite, l'ampoule de quartz est gravée chimiquement par un bain d'acide fluorhydrique. Finalement, la cible est pulvérisée mécaniquement en fonction de la taille des grains désirée. Cette taille de grain a également un impact sur la cristallisation du matériau [91].

Parmi les méthodes de dépôt, l'évaporation est la plus rapportée dans la littérature, cependant, le faible point de fusion du sélénium (221 °C) en comparaison au germanium (938 °C) et à l'antimoine (631 °C) le rend susceptible à des problèmes stoechiométriques. En revanche, la pulvérisation est reconnue comme une des meilleures pour permettre la conservation de la composition de la cible [92].

### 3.3.3 Méthodes de cristallisation

Comme vu à la section 2.3, le changement de phase est dicté par un traitement thermique du PCM. Il n'existe que trois méthodes intégrées pour engendrer cette génération de chaleur de manière locale.

La plus efficace, mais moins pratique à intégrer de manière compacte, est la méthode optique où le PCM est chauffé localement par un laser pulsé. Cette méthode est excellente pour l'amorphisation de matériaux exigeant un refroidissement très rapide, puisqu'elle permet de concentrer une forte densité de chaleur directement sur le PCM par la focalisation du

faisceau. Cette localisation thermique permet ensuite une dissipation énergétique très rapide et est donc idéale pour procurer une amorphisation reproductive. Cette méthode procure une aire decristallisation supérieure à 0.25 $\mu m^2$ [93] selon la taille du faisceau. D'un point de vue d'intégration informatique, l'utilisation d'un laser pulsé externe n'est pas justifiable logistiquement ni d'un point de vue de consommation énergétique.

La seconde méthode est celle utilisée dans les PRAMs. Elle implique l'application d'une différence de potentielle aux bornes du PCM amorphe et donc isolant, jusqu'à ce qu'un voltage seuil, nommé *Ovonic threshold*, soit atteint ($V_{threshold}$) et qu'un passage conductif cristallin soit établi. Cette méthode ne procure une cristallisation que sur un filament plus petit qu'un micromètre et dont la position et la structure est stochastique, la rendant inefficace pour des applications de mémoire optique, où une superficie de plusieurs microns est généralement nécessaire [94].

La troisième méthode est celle utilisée pour ce projet. Elle est basée sur la chaleur émise par un élément résistif par effet Joule. La fabrication d'un élément chauffant permet d'atteindre des températures de fusion tout en limitant la puissance électronique nécessaire. Des matériaux à haute température de fusion, comme le tungsten ou le titane sont généralement utilisés comme éléments chauffants [95]. Le principal inconvénient de cette méthode réside dans son empreinte et son intégration. Les métaux produisent énormément de pertes par résonance plasmonique (section 2.2.3) et doivent donc être situés de manière ingénieuse pour permettre un transfert thermique efficace sans générer trop de pertes optiques. À cet effet, des éléments chauffants en silicium dopé [96] ou en oxide d'indium-étain [97] peuvent être explorés, mais leur faible température de fusion les rend plus susceptibles d'être détruits lors du changement d'état de la mémoire.

## 3.4 Calcul neuromorphique

### 3.4.1 Définition

Le calcul neuromorphique est une facette du calcul informatique tirant inspiration sur la structure du cerveau humain pour accomplir des tâches sophistiquées qu'un processeur conventionnel n'est pas en mesure d'effectuer de manière rapide et efficace. Il s'agit de l'analogue *hardware* aux réseaus de neurones artificiels, soit une infrastructure de calcul en mémoire. Ce domaine s'étend de simples FPGAs à des infrastructures de calcul photoniques et biochimiques, comme l'indique le recensement de Schuman et al. comprenant plus de 2000 publications à ce sujet depuis les 30 dernières années [98]. Le cerveau humain effectue son travail à partir de quelques $10^{11}$ neurones interconnectés par environ $10^{15}$ synapses. Sa grande com-

plexité, combinant une performance, plasticité et efficacité énergétique inégalées par les superordinateurs modernes les plus puissants. Il est polyvalent, capable de mémoriser, identifier et interagir avec des stimuli auditifs, visuels et tactiles tout en consommant approximativement 20 W de puissance électrique. En comparaison, un modèle de langues naturelles à l'état de l'art nécessite plus de 1 MWh de puissance électrique pour son entraînement, équivalent à six ans de cerveau humain.

La compréhension de sa structure synaptique a mené au développement d'algorithmes comme les NNs qui, lorsque appliqués sur un processeur von Neumann, ont révolutionné le calcul informatique et le traitement de l'information tel qu'on le connaît depuis les années 2000. Le calcul neuromorphique cherche à étudier les infrastructures analogues au cerveau humain, dans lesquelles le traitement de l'information et son stockage sont fusionnées contrairement aux architectures de Von Neumann. En somme, le calcul neuromorphique vise à adresser la dissonance anthropologique entre le logiciel inspiré du cerveau humain et le *hardware* sur lequel il est exécuté. Un sommaire des technologies les plus prometteuses et de leur potentiel fait l'objet des sous-sections suivantes. À ce titre, les architectures présentées seront évaluées en comparaison à l'état de l'art actuel, selon des caractéristiques tel la consommation énergétique, la performance, l'évolutivité et la capacité d'intégration tenant compte des infrastructures actuelles.

### 3.4.2 Convertisseur digital-analogique optique

Pour comparer la performance d'accélérateurs neuromorphiques électroniquess et photoniques, il faut d'abord adresser le format dans lequel l'information est acheminée à l'accélérateur. Il est conservateur d'admettre que la grande majorité des données actuellement utilisées pour inférer ou entraîner les modèles d'apprentissage profond sont encodées électroniquement de manière digitale. À cet effet, pour assurer une juste comparaison entre les deux paradigmes de calcul, il faut inclure la conversion D2O (*Digital-to-optical*) et O2D (*Optical-to-digital*). Il existe plusieurs solutions matures et actuellement commercialisées en télécommunications pour effectuer ces conversions.

La conversion O2D peut être adressée directement par une série de photodétecteurs (PDs) suivi d'un amplificateur transimpédance (TIA). Ces PDs, dans la bande C, peuvent être fabriqués en germanium ou en InGaAs et opérer à des vitesses de transfert de plus de 43 Gb/s [99]. Le TIA, tant qu'à lui, sert à convertir le photocourant des PDs en tension. Plusieurs TIAs opérant à des débits de 25 à 40 Gb/s sont documentés [100, 101].

De son côté, la modulation du signal optique en entrée (D2O) peut être effectuée directement ou indirectement à très haut débit et ce, particulièrement pour des précisions faibles de 2 ou 4

bits. La modulation directe est définie par la modulation du signal électrique qui commande le laser, le rendant pulsé, alors que la modulation indirecte implique un laser continu (CW) qui est modulé par un composant photonique distinct à postériori. À partir d'une modulation directe en amplitude à 2 bit (PAM-4) de lasers intégrés à cavités verticales (VCSELs), un débit 64 Gb/s a été observé [102]. Pour des lasers à rétroaction répartie (DFB), un débit de transmission de 28 Gb/s a été rapporté [103]. Des modulateurs indirects en amplitude (PAM-4) à base d'interféromètres de Mach-Zehnder (MZI) et de résonateurs en anneau ont été rapportés opérant à des débits de 100 Gb/s [104] et 40 Gb/s [105] respectivement.

Les DACs optiques modernes présentent des performances 10 à 100 fois plus élevées en terme de vitesse de modulation que les meilleurs accélérateurs électronique modernes. Ceci suggère que même si les DACs s'avèrent être des facteurs limitants, ils sont à prioriser par rapport à l'état de l'art actuel. Un schéma d'architecture optique est illustré à la figure 3.4.

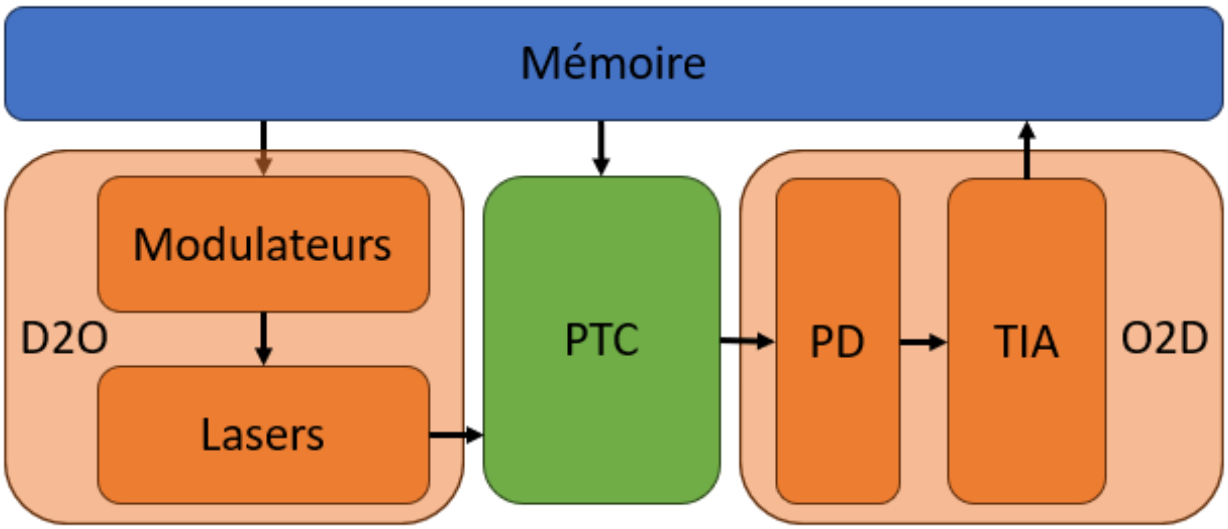



FIGURE 3.4 Architecture générale des accélérateurs optiques en interface électronique comprenant le convertisseur digital-optique (D2O), le processeur photonique (PTC) et le convertisseur optique-digital comprenant les photodiodes (PD), l'amplificateur transimpédance (TIA).

La conversion digitale D2O et O2D est actuellement essentielle, puisque le format digitial est priorisé pour virtuellement toutes les applications. Bien qu'il soit établi que ces conversions ne constituent pas un *bottleneck* de latence, ils constituent tout de même un défi considérable. Dans des applications de reconnaissance d'image, il est envisageable que la conversion D2O puisse éventuellement être omise, les données entrantes demeurant dans le domaine optique.

### 3.4.3 Grille de memristors

La première architecture est inspirée de la structure d'un FPGA conventionnel, mais en utilisant des mémoires non-volatiles électroniques pour effectuer le calcul directement en mémoire. Parmi les deux types de mémoires communément utilisées on retrouve les PCMs (section 3.3) et les memristors. Ces deux dispositifs permettent d'ajuster la résistance entre deux liens (neurones) afin de renforcer ou déprécier la connexion entre ceux-ci. Le memristor est le dernier des quatre composants passifs fondamentaux découverts en électronique. Avec la résistance, le condensateur et l'inducteur, il complète la dépendance entre les 4 quantités fondamentales en électronique, soient la tension ($V$), le courant ($I$), la charge électrique ($Q$) et le flux magnétique ($\Phi$) tel qu'illustré à la figure 3.5.

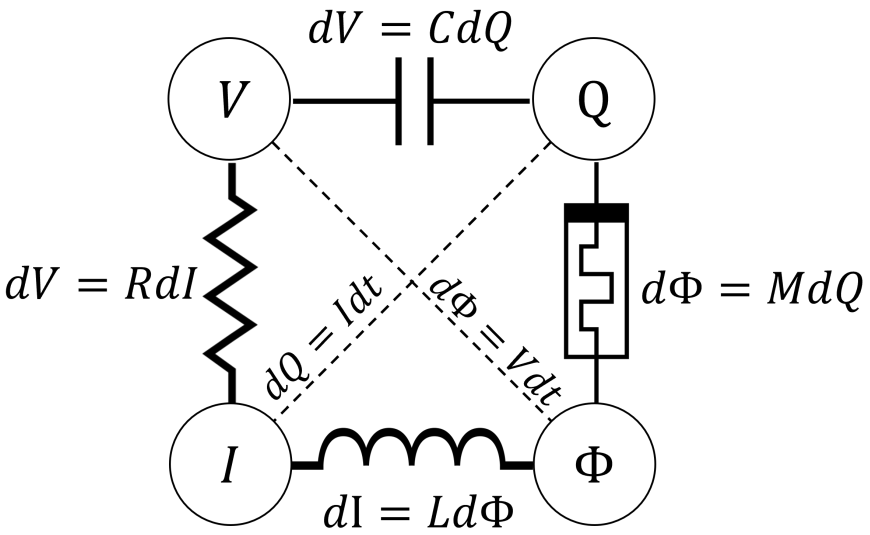



FIGURE 3.5 Diagramme reliant les quatre composants passifs fondamentaux en électronique.

Le nom «memristor» provient de la contraction de mémoire et *resistor*, la capacité de mémoire provient de l'impact de l'historique du flux de charges électriques sur la résistance du matériau, notée $M$ :

$$V(t) = M\left(Q(t)\right) I(t)\ ,\ [M] = \Omega = \mathrm{Wb/C}. \tag{3.3}$$

Cette mémoire du matériau se manifeste donc comme une hystérèse dans la courbe IV présentée à la figure 3.6 [106], dont la résistance du matériau est modulable et non-volatile, c'est-à-dire qu'elle est conservée sans apport énergétique constant. C'est en raison de cette capacité de rétention des tensions préalablement appliqués que les memristors sont considérés

pour des applications de calcul en mémoire analogues aux connexions synaptiques du cerveau humain.

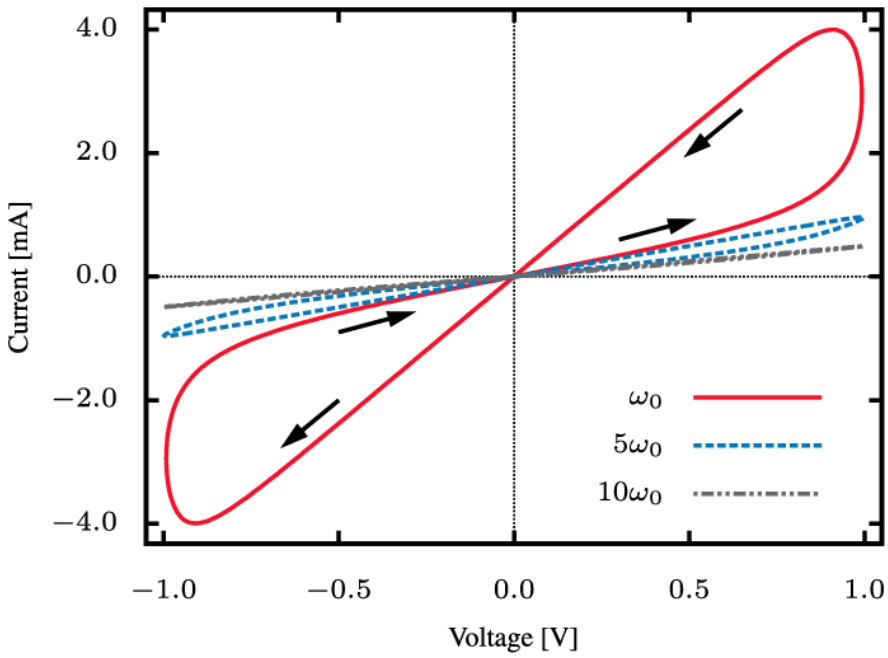



FIGURE 3.6 Hystérèse observée dans la courbe IV d'un memristor.

Une conjecture sur l'existence du memristor a été posée en 1971 pour compléter la symétrie des relations entre les quatres quantités électriques fondamentales [107], mais ce n'est que 37 ans plus tard que son invention a été publiée [108]. Le memristor conventionnel est constitué d'une région faite d'oxyde de titane ($TiO_2$) connectée via deux électrodes en platine (Pt). Une fraction de la région d'oxyde est dopée de manière à augmenter la conductivité du matériau. Lorsqu'une différence de potentiel significative est appliqué aux électrodes, les dopants diffusent dans la région non dopée du $TiO_2$, le rendant moins résistif.

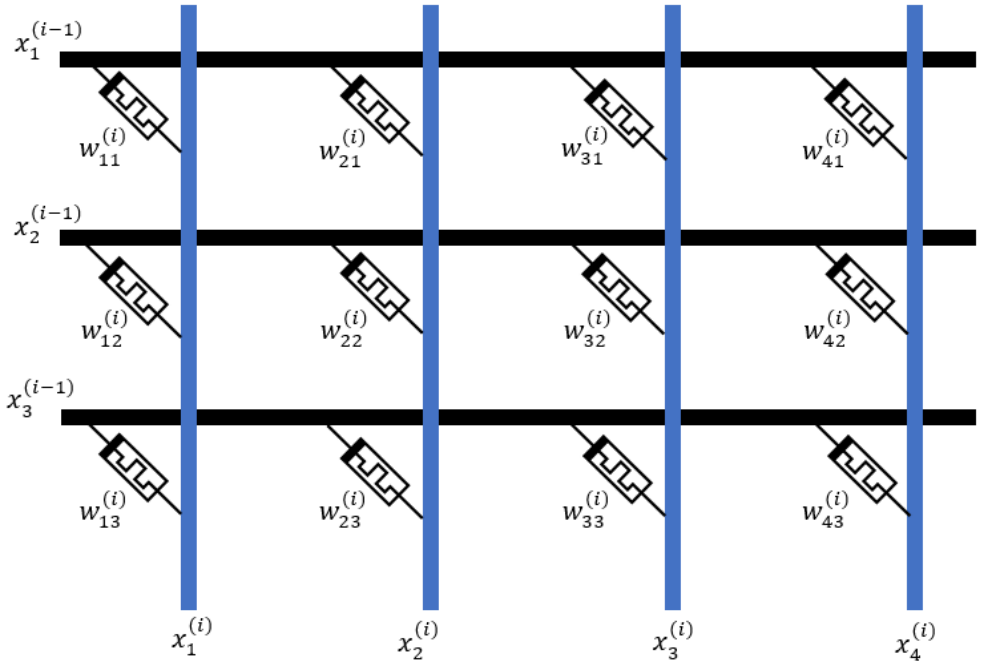



FIGURE 3.7 Schéma du fonctionnement d'une architecture neuromorphique en grille à memristors (MVM de dimension 4x3)

L'architecture proposée est une réseau de connexions électriques par nanofils métalliques où la valeur du neurone en entrée est caractérisée par un voltage dont le courant au neurone en sortie est donnée par la résistance qui représente les poids d'une matrice. Cette architecture purement électronique est très similaire au cerveau humain, capable de renforcer les connexions entre deux neurones virtuels par une diminution de la résistance du memristor, similairement à l'augmentation de densité d'ions dans les canaux dendritiques humains.

Les avantages d'une telle architecture, avec la modulation non-binaire par memristors, sont sa faible empreinte, sa grande capacité de mise à l'échelle et sa robustesse aux défectuosités. En effet il a été démontré qu'un réseau dense tel qu'illustré à la figure 3.7 possède une capacité de régénération, puisque les memristances défectueuses sont automatiquement ajustées de telle sorte à réduire la force de ses connexions synaptiques [109]. Des cellules de mémoire de 50 x 50 nm ont été fabriquées pour connecter les nanofils horizontaux et verticaux permettant une mise à l'échelle de 10 Gb/cm$^2$ [110]. Les principaux inconvénients de cette méthode sont intrinsèques à l'électronique soient la consommation énergétique, la vitesse de propagation du signal et la dissipation thermique. Dans le cas d'une architecture simple constituée de nanofils d'or, la performance du système est entièrement liée aux memristors utilisés. Sa composition et sa taille affecte les tensions nécessaires à l'opération, la durée de programmation de la

mémoire, sa durabilité et la nature binaire ou analogique de sa résistance [111].

La structure d'un memristor est généralement composées d'électrodes métalliques et d'un oxyde partiellement dopé. Plusieurs compositions ont été investiguées, dans le but de modifier l'hystérèse du memrisor, mais aussi pour faciliter la fabrication par ajout de couches d'adhésion. Des électrodes de Au, Pt, Pd et Ti sont généralement employées à cet égard conjointement avec des oxides de titane, de fer, de tantale, d'aluminium et de hafnium. Un couchage de Au(25 nm)/Pt(15 nm)/$TiO_{2-x}$(30 nm)/Pt(25 nm)/Ti(5 nm) a permis un contraste de 18 kΩ pour des voltages de -1.3 à 1.2 V par Prezioso et al. avec des temps d'écriture inférieurs à 1 µs [112].

Une lacune des memristors est leur calibration. Suite à la fabrication, la diffusion de dopants dans l'oxide memristif n'est pas ajustée à l'application désirée. Un potentiel nommé $V_{\text{forming}}$, 2 à 10 fois plus élevé que $|V_{\text{set,reset}}|$, doit être appliqué initialement, exigeant des composantes et circuits additionnels sur la puce. L'évolutivité n'est pas supportée par la réduction de cette quantité, puisque $V_{\text{forming}}$ est inversement proportionnel à l'empreinte du memristor [113].

### 3.4.4 Cascade d'interféromètres Mach-Zehnder

Les solutions électroniques, malgré qu'elles adressent le goulot d'étranglement de Von Neumann, demeurent fondamentalement énergivores. C'est pourquoi des infrastructures optiques sont également considérées. L'optique dispose de plusieurs avantages en terme de consommation énergétique, mais aussi de bande passante et de latence par rapport à l'électronique.

La première solution de calcul optique présentée est basée sur un réseau d'interféromètres de Mach-Zehnder (MZI) en cascade. Le principe fondamental est d'encoder les poids et le biais d'une matrice dans la différence de phase et l'amplitude entre les deux bras de l'interféromètre. En assignant un guide d'onde à chaque composante du vecteur en entrée, il est possible d'effectuer toute opération unitaire sur un vecteur 2x1 par une combinaison de coupleurs 2x2 et de déphaseurs. On représente cette opération unitaire par le formalisme matriciel selon l'équation 3.4.

Ce type d'accélérateur, pour faciliter son intégration en *datacenter* et minimiser son empreinte, est typiquement fabriqué sur puce à partir d'une plateforme SOI ou SiN (voir section 2.1). Un schéma de l'opération unitaire normalisée 2x2, notée $U_2$ est illustré à la figure 3.8.

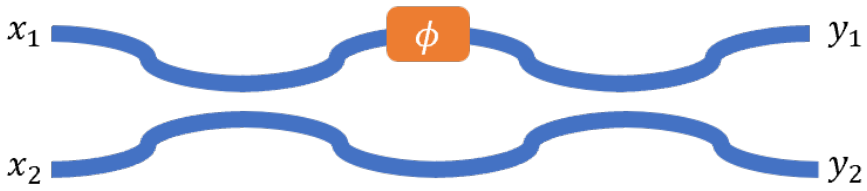



FIGURE 3.8 Unité normalisée $U_2$ du calcul par cascade de MZI.

La matrice de transfert de l'unité $U_2$ est présentée à l'équation 3.4.

$$U_2(\phi) = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} e^{i\phi} & 0 \\ 0 & e^{-i\phi} \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} i\sin\phi & -\cos\phi \\ \cos\phi & -i\sin\phi \end{bmatrix} \tag{3.4}$$

Le déphasage $\phi$ est variable de sorte a moduler le vecteur de sortie. En ajoutant des phases $\alpha$ et $\beta$ aux branches 1 et 2 respectivement, la restriction de normalisation est retirée, on obtient la matrice suivante :

$$\mathbf{y} = \begin{bmatrix} e^{i\alpha} & 0 \\ 0 & e^{i\beta} \end{bmatrix} U_2(\phi)\mathbf{x} = \begin{bmatrix} ix_1e^{i\alpha}\sin\phi - x_2e^{i\alpha}\cos\phi \\ x_1e^{i\beta}\cos\phi - ix_2e^{i\beta}\sin\phi \end{bmatrix} \tag{3.5}$$

On peut ensuite extrapoler pour une opération unitaire sur un vecteur 4x1. On remarque que 15 informations sont encodées par phase, correspondant à une matrice de poids 4x4 normalisée. Un potentiel design de $U_4$ est donné à la figure 3.9 [114].

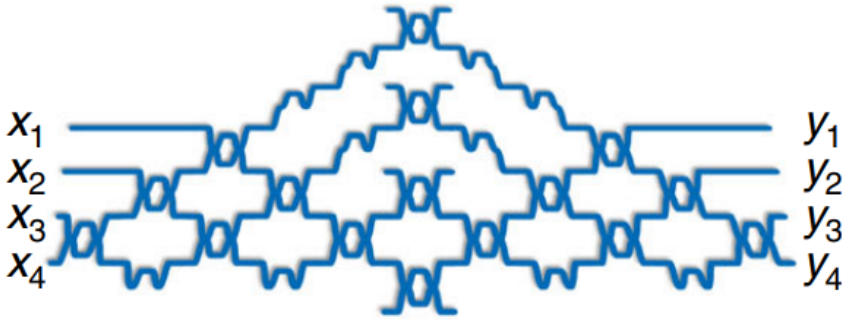



FIGURE 3.9 Exemple d'architecture $U_4$ normalisée pour effectuer une multiplication matrice-vecteur de taille 4x4.

Le principal inconvénient de ce type d'accélérateur provient de la nature de ses déphaseurs. Parmi ceux-ci le déphaseur thermique, soit l'intégration d'un simple élément chauffant à

proximité d'une des branches du MZI, est le plus communément utilisé en photonique intégrée. Étant basé sur la dilatation thermique $\left(\frac{\partial L}{\partial T}\right)$ des matériaux et leur coefficient thermo-optique $\left(\frac{\partial n}{\partial T}\right)$, il s'applique à toute platerforme. Ce type de déphaseur est extrêmement précis lorsque son isolation thermique par rapport aux autres déphaseurs à proximité est assurée [115]. Par contre, il nécessite un apport statique de puissance électrique et est lent par rapport aux autres solutions de modulation de phase. En considérant une potentielle mise à l'échelle, cette approche de déphasage est la moins prometteuse.

Le second déphaseur potentiel est basé sur la désertion des porteurs de charge. Il nécessite l'intégration de jonctions PN et donc de dopage, le limitant à toute fin pratique, à des applications en photonique sur silicium. Injecter ou retirer des porteurs de charges dans le guide d'onde a pour effet de varier l'indice de réfraction de manière similaire au déphaseur thermo-optique au détriment de l'ajout de pertes optiques. Il s'agit d'une solution à très grande vitesse de modulation, au détriment d'un empreinte importante et d'une consommation énergétique modérée [116]. Une solution similaire implique l'intégration hétérogène d'un matériau avec un grand coefficient électro-optique comme le niobiure de lithium ou le nitrure de gallium ce qui présente toutefois un défi de fabrication [117].

Finalement, un déphaseur par MEMS optique est la troisième solution la plus commune. Elle est basée sur une déformation électrostatique d'un des deux guides d'onde pour induire la différence de phase. Typiquement, le guide d'onde en question est sous-gravé pour permettre son déplacement vertical lorsqu'une différence de potentiel est appliquée entre le substrat et le guide d'onde lui-même. Il s'agit du même principe que les interrupteurs optiques [118]. Cette méthode est moins énergivore, assez dense et rapide, ce qui en fait un excellent candidat comme déphaseur [119]. Elle exige cependant des tensions de 2 à 10 fois les conventions usuelles (3.3 et 5 V) pour l'actionnement [120]. Une des limitations provient de la fabrication avec la sous gravure du guide d'onde qui n'est pas conforme à un procédé CMOS standard. Cette solution est tout de même priorisée par *Lightmatter*, possiblement la compagnie de calcul optique la plus avancée à ce jour [121]. Un sommaire des déphaseurs potentiels est illustré à la figure 3.10 [122].

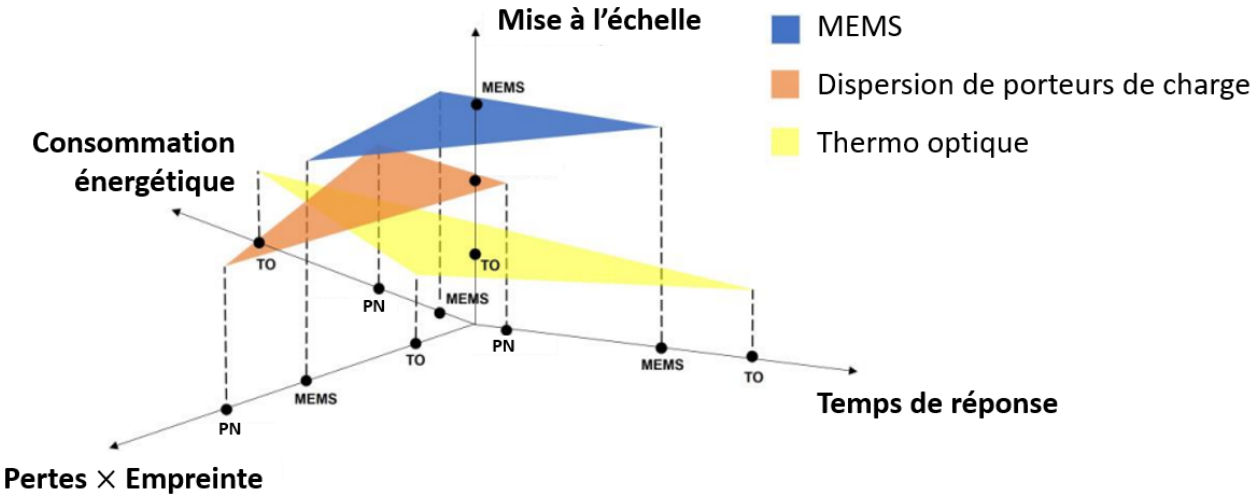



FIGURE 3.10 Avantages et inconvénients des technologies de déphaseurs sur silicium.

Peu importe le choix de déphaseur utilisé dans une telle structure, la mémoire qu'ils supportent est volatile et implique donc une consommation énergétique statique ou une tension non standard.

### 3.4.5 Processeur matriciel avec mémoire à changement de phase

La troisième solution de calcul neuromorphique présentée est celle qui fait l'objet de ce mémoire. Elle profite des avantages des deux designs présentés précédemment : soient la vitesse, la consommation énergétique et le parallélisme optique tout en utilisant une mémoire non volatile modulant directement l'amplitude du signal comme multiplication. Comme vu à la section 2.5, la multiplication matrice-vecteur est généralement décomposable en 5 blocs : 1) Division du vecteur d'entrée, 2) Décomposition des vecteurs 3) multiplication scalaires, 4) addition des scalaires, 5) recombinaison du vecteur en sortie. Les étapes 2, 3 et 4 constituent un produit scalaire simple, soit l'unité de base des processeurs en tenseur. Comme pour le design par cascade d'interféromètres, les connexions constituent des guides d'onde dans lesquels l'information encodée dans l'amplitude du champ électrique est modulée durant son parcours dans l'accélérateur photonique.

La première proposition d'accélérateur photonique basé sur des PCMs vient de l'Université George-Washington [123]. La structure suggérée est composée de coeurs à produits scalaires (*Dot Product Engine* (DPE)) dans lesquels les éléments individuels du vecteur en entrée sont spectralement filtrés par une série de résonateurs en anneaux. Chacune des composantes est ensuite individuellement multipliée par son interaction avec une cellule de mémoire située

directement au-dessus du guide d'onde. Cette mémoire est constituée d'une mince couche de PCM, qui, en phase amorphe est transparente aux longueurs d'ondes d'opération et qui, en phase cristalline absorbe une partie de la lumière parcourant le guide d'onde. Une fois la multiplication effectuée, les signaux sont spectralement recombinés par une seconde série de résonateurs en anneau identiques. Une architecture de multiplication matrice-vecteur (MVM) de dimension $n \times n$ est constituée de $n$ DPEs disposés en parallèle. Un exemple de DPE de dimension quatre est illustré à la figure 3.11.

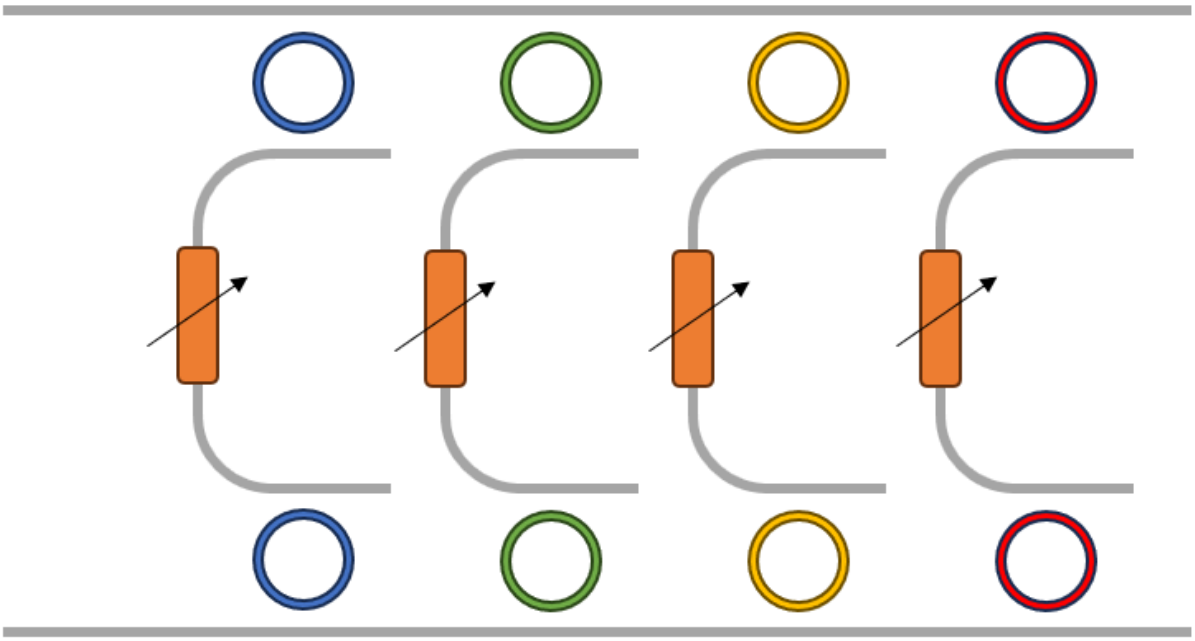


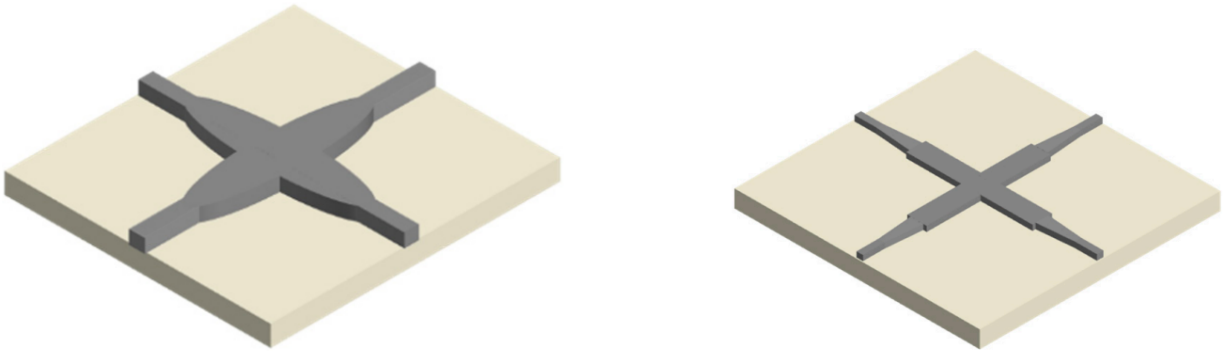
FIGURE 3.11 Processeur à produit scalaire (DPE) de taille 4 utilisant la modulation passive en amplitude par matériaux à changement de phase.

## 3.5 Croisements optiques

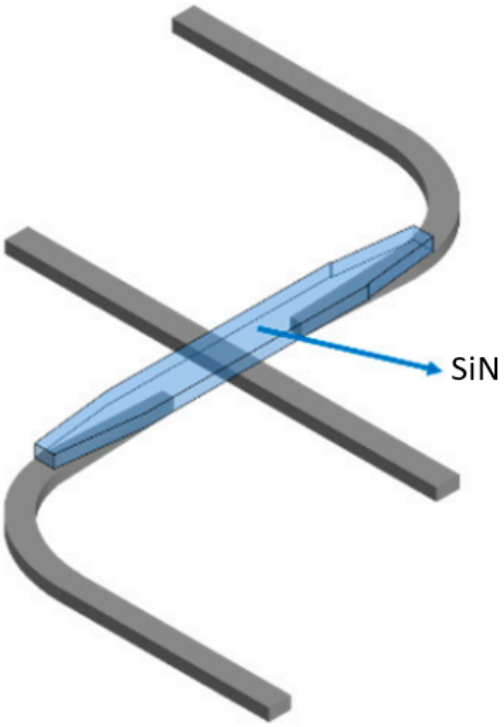
Comme vu à la section 2.5, la MVM est un algorithme fondamentalement non-planaire. Ceci pose problème lors du design de circuits photoniques sur puce. Pour minimiser les erreurs de calcul, il est essentiel d'établir quelle architecture et quels types de croisements doivent être utilisés pour minimiser les pertes, la diaphonie (*crosstalk*) et l'empreinte nécessaire. Cette section présentera les solutions potentielles en termes de croisements optiques comme les croisements directs, en entonnoirs et par pontage.

Le croisement direct est le plus fondamental, constitué de 2 guides d'ondes simples qui se croisent selon un certain angle. Bien que les angles de 60 et 120° sont répertoriés comme

optimaux, ceux-ci sont généralement employés à 90° pour faciliter le design [124]. Ceux fabriqués en fonderie répertorient des pertes de 0.1-1 dB et une diaphonie de -40 dB, mais sont fortement affectés par les défauts de fabrication. Les croisements en entonnoir a), emploient une augmentation progressive de la largeur du guide d'onde, typiquement gaussienne ou conique, pour minimiser les pertes par radiation dues au changement soudain du mode à l'intersection des deux guides d'onde. Les croisements par pontage c), quant à eux, sont fondamentalement différents. Analogues au vias en électronique, ils exploitent l'emploi d'un couplage vertical à une autre couche comprenant des guides d'ondes. Ce couplage utilise un guide d'onde en entonnoir pour augmenter le diamètre du mode et maximiser le couplage avec la couche supérieure ou antérieure. Avec la cointégration de la plateforme SiN et SiP, le croisement peut être adressé en dirigeant un des deux signaux dans la couche de nitrure au-dessus du guide d'onde en silicium. Cette technologie procure la diaphonie la plus faible et des pertes raisonnables ($\sim$0.1 dB), tout en minimisant l'empreinte et en permettant plusieurs croisements consécutifs sans pertes additionnelles. Une illustration des principaux types de croisements présentés fait l'objet de la figure 3.12.



(a) (b)




(c)

FIGURE 3.12 Illustration de 3 types de croisements en photonique intégrée : a) par entonnoir, b) par MMI et c) par pontage.

Les croisements par pontage sont ainsi très prometteurs et ce particulièrement lorsque l'architecture exploite les croisements consécutifs. L'architecture novatrice présentée à la section 7.3 quantifie la performance d'un tel DPE.

Suite à ce recensement de la littérature établissant les fondements du projet, des simulations

optiques et électriques présentant sa viabilité et justifiant le design des mémoires fait l'objet de la section suivante.

# CHAPITRE 4 SIMULATIONS

La caractérisation du $Ge_2Sb_2Se_5$ est essentielle pour simuler adéquatement les dispositifs et architectures photoniques.

## 4.1 Simulations optiques

Les simulations optiques présentées ci-dessous on été effectuées à l'aide du logiciel Lumerical (Ansys Optics).

### 4.1.1 Modes

Les modes de propagation fondamentaux d'une coupe transversale d'une cellule de mémoire à $\lambda = 1550$ nm sont illustrés à la figure 4.1.

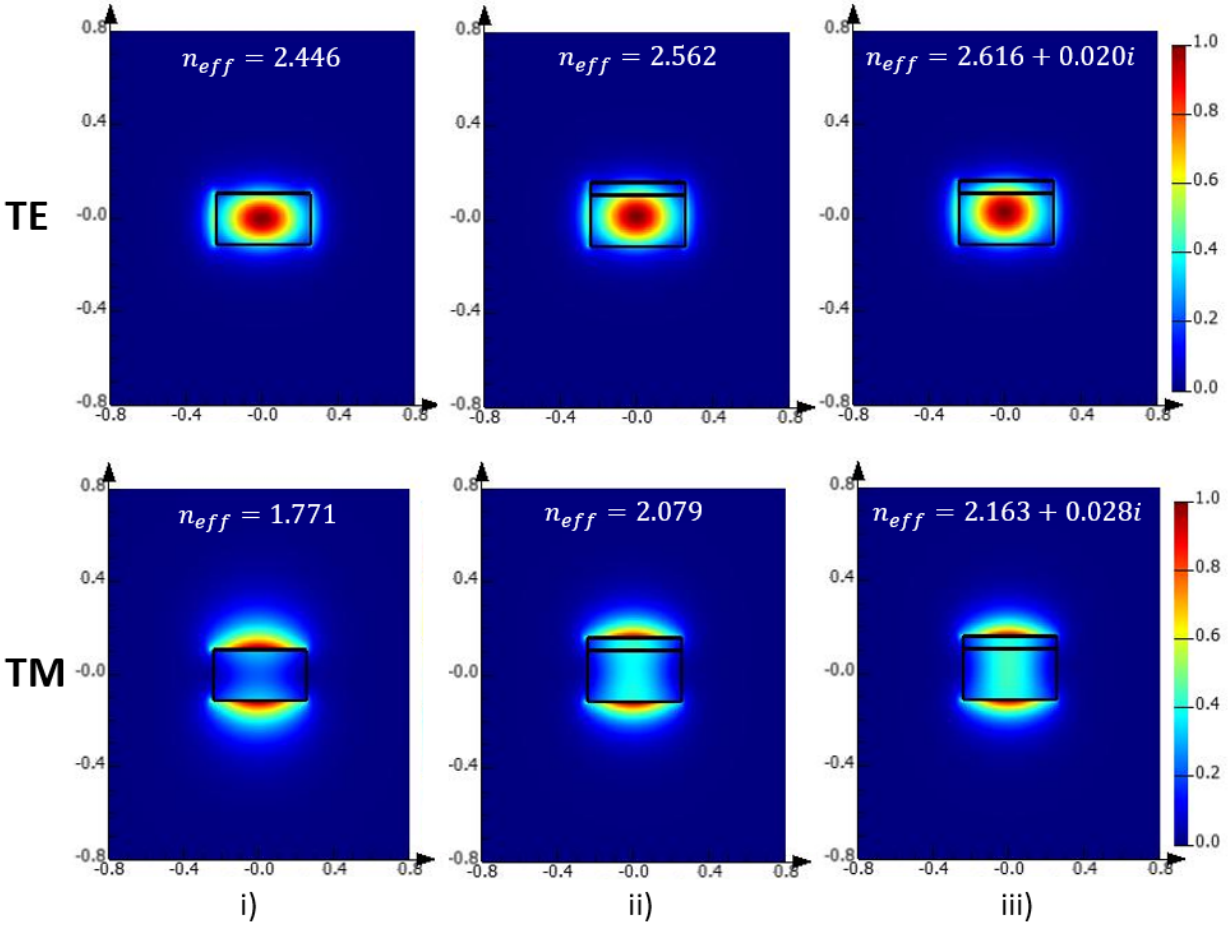


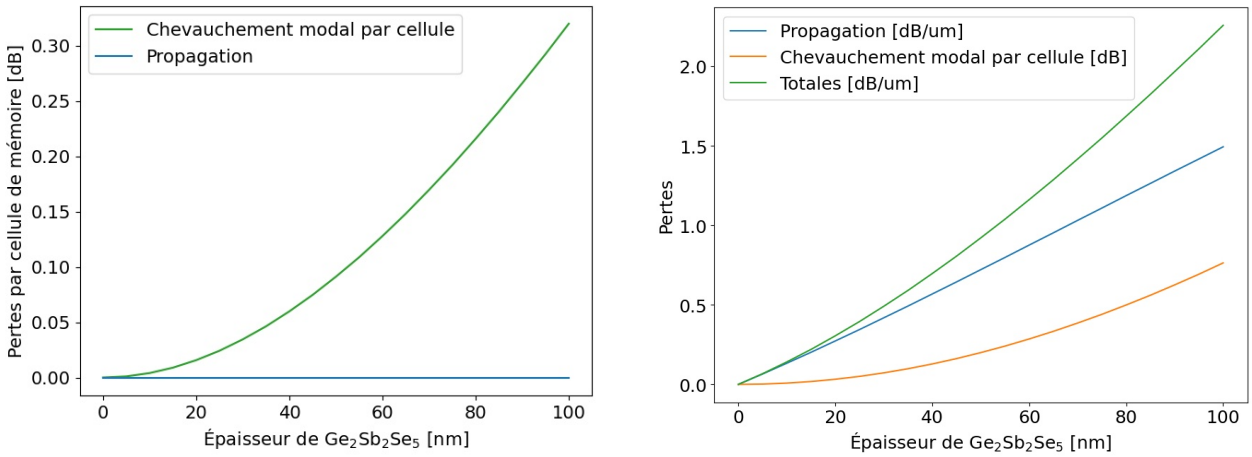

FIGURE 4.1 Modes fondamentaux $|\mathbf{E}|$ pour le guide d'onde i) sans mémoire, ii) avec mémoire en phase amorphe, iii) avec mémoire en phase cristalline à $\lambda = 1550$ nm.

on observe que les modes TM possèdent un contraste d'indice de réfraction très élevé entre le mode de propagation et les modes de mémoire. De plus, le contraste $\Delta\kappa$ entre les états amorphes et cristallins est similaire pour les deux modes. Pour ces raisons ainsi que les pertes de courbures plus importantes, le choix du mode TE est justifié.

### 4.1.2 Dimensions du matériau à changement de phase

Le dépôt du $Ge_2Sb_2Se_5$ sur le guide d'onde est une étape critique du procédé de fabrication de la mémoire optique. L'épaisseur de la couche de PCM dicte donc l'allure du mode de propagation dans le guide d'onde. On cherche à minimiser les pertes dues à la différence de diamètres de mode tout en assurant des pertes par propagation mesurables. À cet effet, des pertes approximatives de 1 dB/µm sont visées. Les pertes théoriques sont dictées par l'absorption du matériau à changement de phase déposé à proximité du guide d'onde présumé

idéal et les pertes par chevauchement de mode entre le mode propagé dans le guide d'onde et celui propagé dans le guide d'onde et le PCM. Cette dernière contribution se produit 2 fois par cellule de mémoire par réciprocité de l'optique passive. La figure 4.2 illustre la contribution de ces deux sources de pertes optiques pour l'état amorphe et cristallin à 1550 nm.




(a) Phase amorphe.

(b) Phase cristalline.

FIGURE 4.2 Pertes engendrées par une cellule de mémoire en fonction de l'épaisseur du $Ge_2Sb_2Se_5$.

On observe que le contraste de 1 dB désiré est obtenu pour une épaisseur approximative de 50 nm.

### 4.1.3 Distance de l'élément chauffant

La présence d'éléments chauffants dans le design de la mémoire implique la considération de routes métalliques indésirables dans le design du circuit photonique intégré (section 2.2.3). La distance minimale entre le guide d'onde et les couches métalliques doit être choisie telle que peu de pertes soient générées à l'échelle d'une puce conventionnelle (1-10 $cm^2$). En revanche, une faible distance entre le PCM et l'élément chauffant doit être telle que la cristallisation et l'amorphisation soient possibles sans accumulation latente de chaleur entre le métal et le PCM. À cet effet, les pertes plasmoniques à $\lambda = 1550$ nm ont été simulées et répertoriées à la figure 4.3.

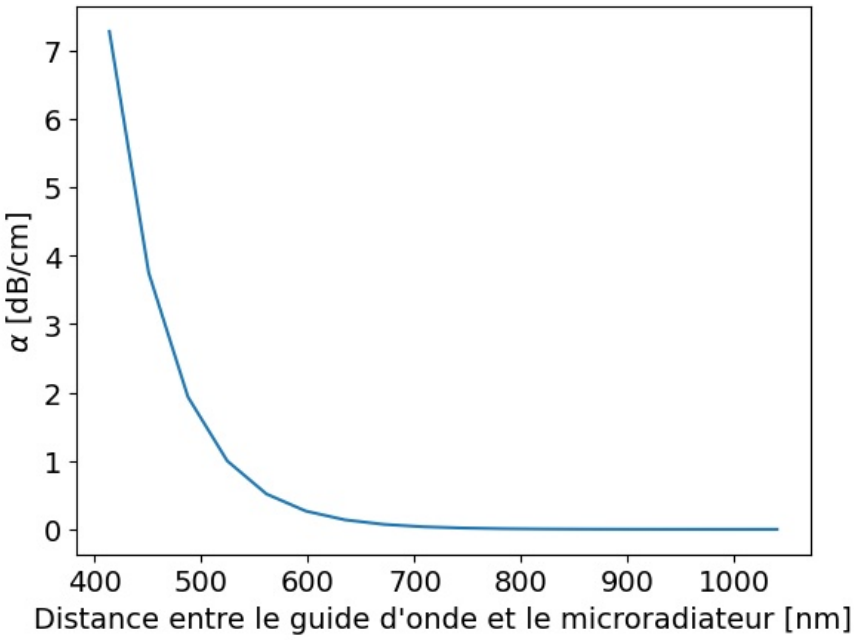



FIGURE 4.3 Impact de la distance verticale de l'élément chauffant sur les pertes plasmoniques générées $\alpha$.

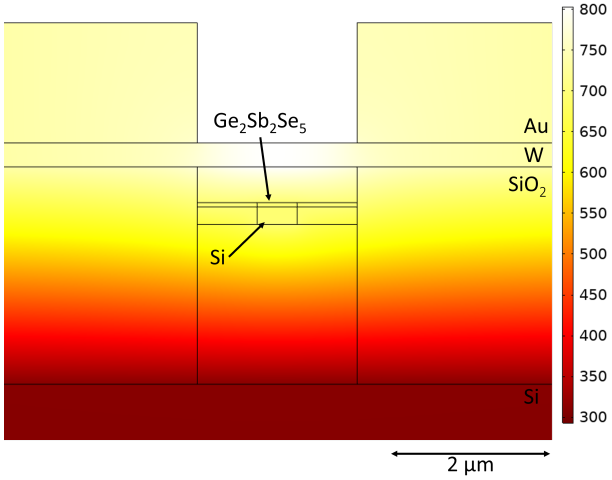
On remarque une diminution rapide des pertes à partir de 500 nm et des pertes négligeables à 800 nm. Le minimum d'épaisseur d'oxyde entre le guide d'onde et le W-Ti est donc établi à 500 nm et sera si ajusté la performance électrothermique le permet.

## 4.2 Simulations électrothermiques

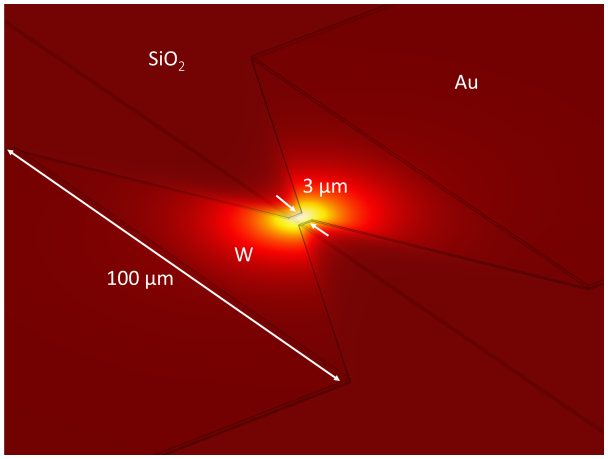
Les simulations électrothermiques ont été effectuées avec le logiciel *COMSOL Multiphysics* avec le module «Effet Joule», composé des unités «AC/DC» et «Transfert de chaleur».

### 4.2.1 Régime stationnaire

Les simulations stationnaires de l'élément chauffant à l'état stationnaire ont été effectuées pour diverses tensions afin de quantifier la dissipation de chaleur de la puce suite au changement d'une mémoire. Les conditions aux frontières utilisées impliquent une convection standard avec l'air ambiant sur le dessus de la puce et un dissipateur de chaleur idéal sous la puce, représentant le porte échantillon. La distribution de la température à l'état stationnaire illustrée en 2D pour un radiateur droit (figure 4.4a) et en 3D pour un radiateur en entonnoir (figure 4.4b) montre que la chaleur est très locale au dessus de la cellule de mémoire lorsque l'entonnoir est utilisé.




(a) Représentation 2D avec l'élément chauffant droit.




(b) Représentation 3D avec l'élément chauffant en entonnoir.

FIGURE 4.4 Distribution de la température de la mémoire à l'état stationnaire

Ces simulations sont utiles pour observer un la cristallisation DC et donc sans l'usage d'impulsions. Il est important de monitorer la température des matériaux voisinants l'élément chauffant afin de vérifier que les deux changements de phase sont possibles sans détruire le dispositif. Ainsi, la figure 4.5 confirme que $T_c$ est atteint aisément sans compromettre les métaux, mais que la température maximale de routes en or approche $T_{f,Au}$ lorsque la cellule de $Ge_2Sb_2Se_5$ atteint la température de fusion nécessaire à l'amorphisation.

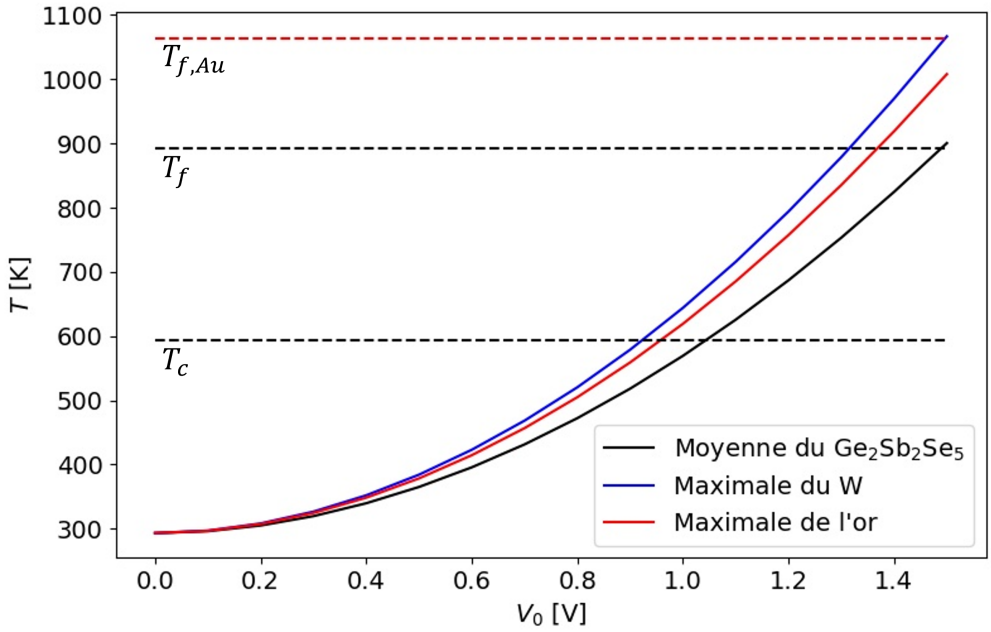



FIGURE 4.5 Impact de la tension aux bornes de l'élément chauffant sur la température de divers régions de la mémoire (régime constant).

On peut donc s'attendre à cristalliser la cellule de PCM en régime stationnaire aisément pour vérifier le bon fonctionnement de l'élément chauffant. Pour confirmer la viabilité du design pour la transition *RESET*, des simulations en régime transitoire sont nécessaires.

### 4.2.2 Régime transitoire

En réalité, la vitesse du changement de phase doit être optimisée. En effectuant des simulations en régime transitoire, on détermine la durée de l'impulsion nécessaire pour atteindre la température désirée. Pour l'amorphisation, il est impératif que le taux de refroidissement seuil soit atteint. La figure 4.6 illustre le comportement d'une cellule de mémoire telle que représentée à la figure 4.4.

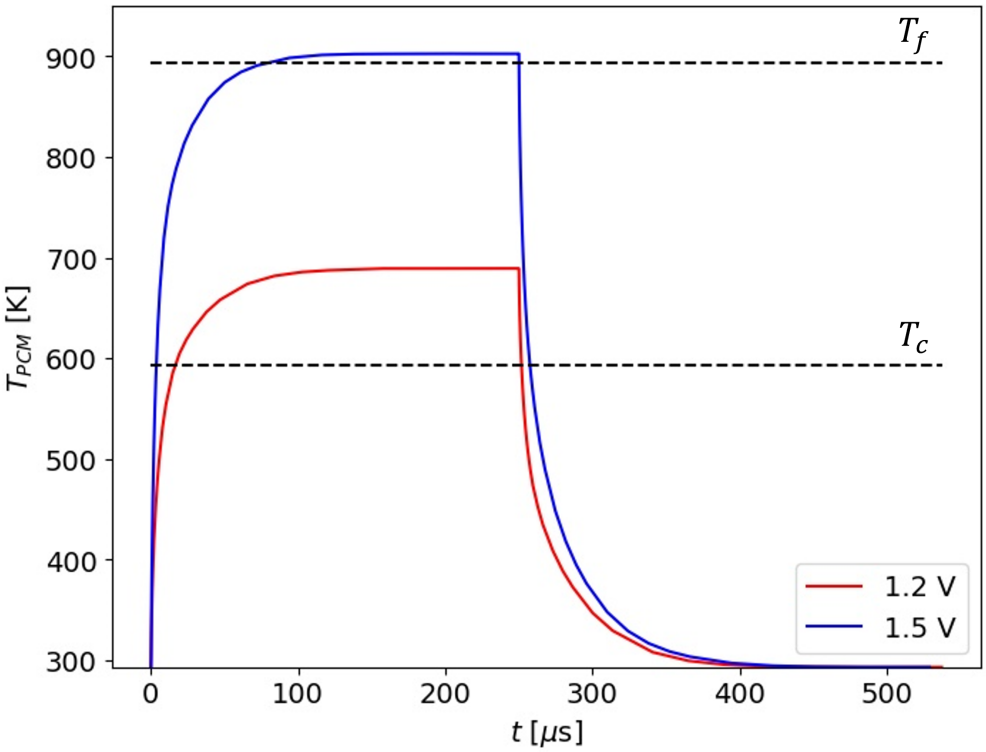



FIGURE 4.6 Régime transitoire du chauffage de la mémoire optique pour une impulsion de voltage de 250 µs.

Les simulations en régime transitoire montrent que l'état stationnaire est atteint rapidement ($< 100$ µs). On remarque que pour l'amorphisation, la température chute sous la température de cristallisation $T_f$ en $\approx 10$ µs, résultant en un taux de refroidissement $\frac{\mathrm{d}T}{\mathrm{d}t} \sim 10^9 - 10^6$ °C/s. On observe finalement que le dispositif simulé est très sensible en tension puisqu'une variation de quelques dixièmes de volt sépare les deux transitions.

Cette section a permis d'établir le point de départ de la méthode, des matériaux et des paramètres de fabrication de la mémoire optique. La section suivante présente la caractérisation du $Ge_2Sb_2Se_5$ pour confirmer les propriétés rapportées dans la littérature et ajuster le procédé de fabrication.

# CHAPITRE 5 CARACTÉRISATION DU $Ge_2Sb_2Se_5$

Cette section est dédiée à la caractérisation du matériau à changement de phase choisi. L'établissement des propriétés du $Ge_2Sb_2Se_5$ est essentielle afin de vérifier la concordance avec la littérature, confirmer son choix comme mémoire optique et caractériser son comportement lorsque intégré sur puce. Pour se faire, une caractérisation du dépôt et de la cristallisation suivie de mesures ellipsométriques, spectrophotométriques, spectroscopiques, et de diffraction à rayons X on été effectuées. Une brève explication du fonctionnement de ces instruments et la méthodologie employée précède l'analyse des résultats obtenus pour chacun des tests.

## 5.1 Préparation des échantillons

La préparation des échantillons pour la caractérisation englobe l'impact des paramètres de dépôt sur la qualité de la couche mince ainsi que les méthodes de cristallisation macroscopiques explorées. Finalement, l'encapsulation et l'épaisseur de PCM varient en fonction de la nature des tests effectués, par exemple, la spectroscopie photoélectronique est une mesure de surface pour laquelle l'encapsulation n'est pas viable.

### 5.1.1 Dépôt du matériau à changement de phase

Les échantillons dédiés à la caractérisation sont préparés au laboratoire de microfabrication (LMF) de l'École Polytechnique de Montréal. Le $Ge_2Sb_2Se_5$ est d'abord déposé sur un morceau de gaufre de silicium monocristallin (100) par pulvérisation à magnétron (13.56 MHz). Les 2 cibles de chalcogénure utilisées sont fabriquées par Angstrom Sciences par procédé conventionnel de compression à chaud et de fusion par induction sous vide suivi de compression à chaud respectivement. Le dépôt a été caractérisé selon la puissance RF et la pression de dépôt (figure 5.1 et 5.2).

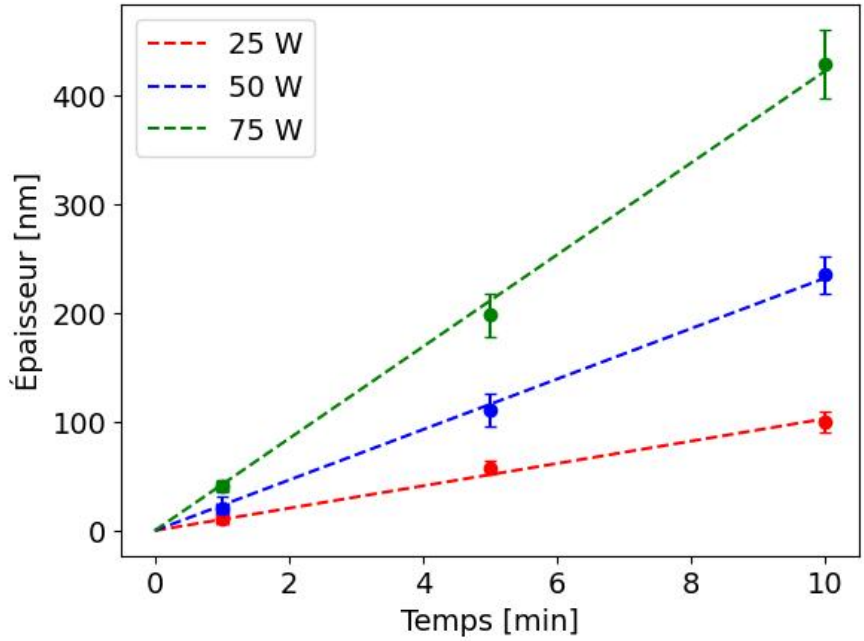



FIGURE 5.1 Effet de la puissance sur le taux de dépôt (5 mTorr).

Ce résultats indiquent que le $Ge_2Sb_2Se_5$ requière un faible puissance pour le dépôt. À 20 mTorr, le plasma s'allume entre 12 et 17 W, or la stabilité de celui-ci à 25 W est questionnable. À l'opposé, il a été observé que le plasma à 100 W génère un arc électrique endommageant la cible. Ainsi, puisque l'épaisseur visée pour les cellules de mémoire est de 50 nm, le taux de dépôt de 23 nm/min obtenu à 50 W est priorisé.

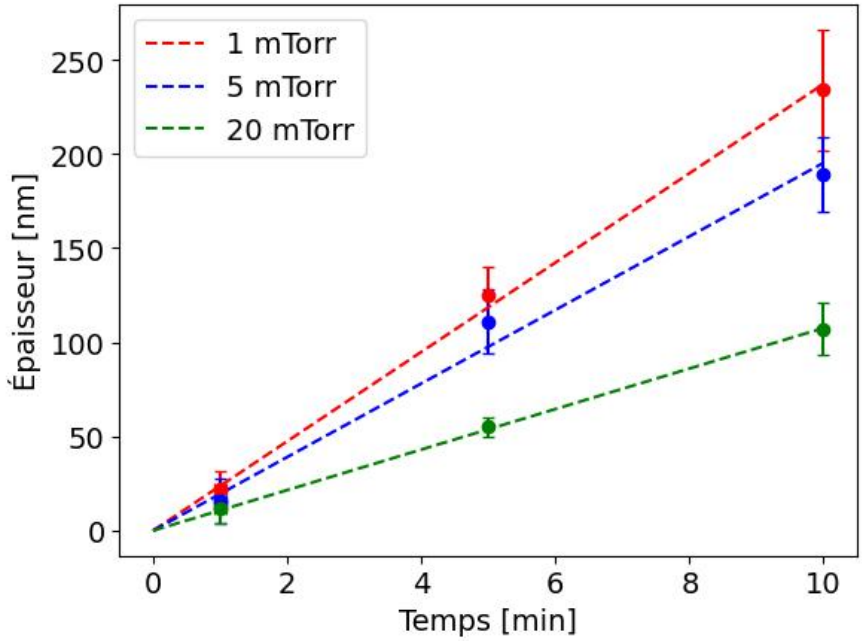



FIGURE 5.2 Effet de la pression sur le taux de dépôt (50 W).

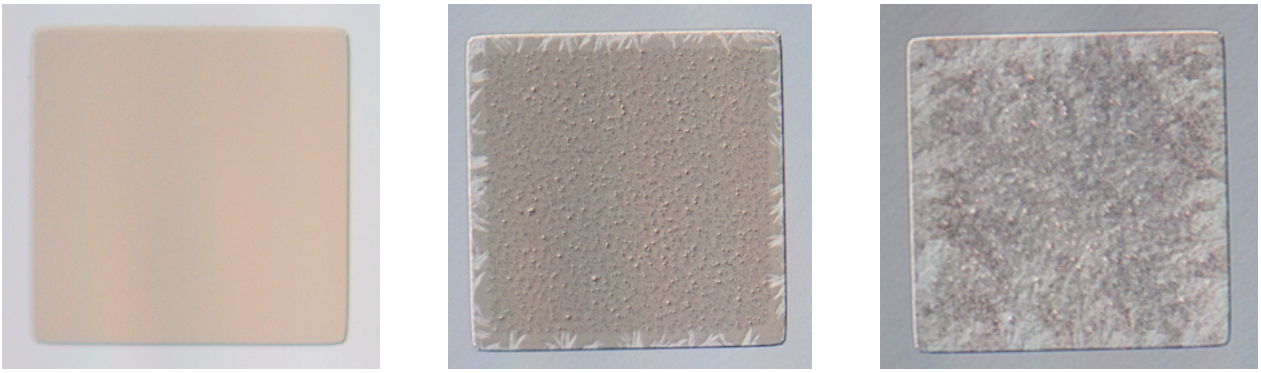
L'effet de la pression est inverse à ce que l'on observe pour des matériaux métalliques comme l'Al, le Pt et le Cu [125]. Il s'agit cependant d'un phénomène répertorié dans la littérature [126], expliqué par la faible énergie cinétique des particules heurtant le substrat en raison d'un plus grand nombre de collisions avec l'argon ionisé se traduisant par une réduction du libre parcours moyen des particules de la cible lorsque la pression dans la chambre est plus élevée.

Finalement, la pulvérisation choisie est effectuée à 50 W, à un débit de 10 sccm d'argon sous une pression de dépôt de 5 mTorr et produit un taux de dépôt d'environ 23 nm/min. Le pulvérisateur possède 3 magnétrons et permet la pulvérisation de trois matériaux différents sans interrompre le vide dans la chambre. Puisque les chalcogénures sont fortement susceptibles à l'oxydation [127], certains échantillons sont encapsulés par 15 nm de $SiO_2$ pulvérisé à 200 W, 10 sccm d'argon, 20 mTorr à un taux de dépôt approximatif de 5 nm/min.

### 5.1.2 Cristallisation

Les échantillons étant déposés dans l'état amorphe, il est essentiel d'établir une procédure de cristallisation en gros volume (bulk) de la couche mince de $Ge_2Sb_2Se_5$. Pour minimiser les contraintes de recuit dans les couches minces, un recuit lent sous environnement fermé d'azote a été effectué dans le four à oxydation thermique VBF-1200F, pour lequel un taux

de 1-5 K/min a été utilisé. Lorsque la température visée est atteinte, les échantillons sont maintenus durant 30 minutes pour maximiser l'homogénéité du processur de cristallisation. Finalement, l'échantillon est refroidit lentement par convection et atteint la température ambiante en quelques heures. Les figures 5.3 et 5.4 montrent les résultats du recuit lent sur le $Ge_2Sb_2Se_5$ tel que déposé.



(a) 290 °C (b) 310 °C (c) 330 °C

FIGURE 5.3 Cellule de GSSe de 100 µm.

(a) 290 °C (b) 310 °C (c) 330 °C

FIGURE 5.4 Cellule de GSSe de 20 µm.

On observe que malgré le recuit lent, la cristallisation émerge à 310 °C au périmètre des cellules. On observe également ce qui ressemble à des sites de nucléation émerger au centre des cellules. Pour l'échantillon à 330 °C on observe que la cristallisation est uniforme sur l'entièreté de la cellule. Ainsi, on établit $T_c \approx 330$ °C.

## 5.2 Ellipsométrie spectroscopique

### 5.2.1 Théorie

L'ellipsométrie spectroscopique ou l'ellipsométrie est une méthode de caractérisation optique de matériaux non-destructive. Elle est particulièrement utile pour les couches minces, puisque

leur indice de réfraction varie en fonction de l'épaisseur ($d$) de la couche mince et est généralement très différent de celui du matériau brut $\left(\frac{\mathrm{d}\varepsilon}{\mathrm{d}d} \neq 0\right)$. La méthode de dépôt de la couche mince influence également les propriétés optiques de celle-ci de par la conservation stoechiométrique, la porosité et la vitesse du dépôt [128]. Les ellipsomètres spectroscopiques sont plus puissants que les ellipsomètres conventionnels puisqu'ils permettent la mesure spectrale $(\Psi(\lambda), \Delta(\lambda))$ et donc sur plusieurs longueurs d'ondes. Recueillir un spectre pour divers angles d'incidences ($\theta_i$) permet également de modéliser plus précisément l'échantillon. Bien que le terme «ellipsométrie» ne date que des années 40 [129], l'idée d'utiliser la polarimétrie pour caractériser l'épaisseur de couches minces revient à Paul Drude à la fin du 19ème siècle [130].

Plus spécifiquement, son fonctionnement exploite la variation de polarisation d'un faisceau suite à sa réflection sur un ou plusieurs matériaux, typiquement sous forme de couche mince. Le ratio $\rho$ d'amplitude $\Psi$ et le déphasage $\Delta$ (voir équation 5.1) entre la composante du champ électrique parallèle ($r_p$) et normale ($r_s$) au plan d'incidence sont mesurés et comparés à un modèle pour déduire diverses propriétés du matériau, dont sont épaisseur, sa permittivité, sa rugosité et sa composition. Le développement de modèles semi-empiriques est fondamental à l'étude des matériaux, puisqu'il n'existe pas de modèle théorique universel expliquant la dispersion chromatique pour tous les types de matériaux [131].

$$\rho = \frac{r_p}{r_s} = \tan\Psi \; e^{i\Delta} \tag{5.1}$$

À partir de ces paramètres ellipsométriques, la constante pseudo diélectrique ($\langle\tilde{\varepsilon}\rangle$) est calculée à partir de l'équation 5.2. Cette constante représente la constante diélectrique moyenne de l'échantillon comme étant un seul matériel homogène. En réalité, l'ellipsométrie est généralement effectuée sur un échantillon multi-couches comprenant un substrat connu et le matériau étudié. La modélisation de l'échantillon est donc nécessaire pour tenir compte de l'impact du substrat sur $\langle\tilde{\varepsilon}\rangle$ et ainsi retirer l'épaisseur et/ou l'indice de réfraction du matériau à l'étude.

$$\langle\tilde{\varepsilon}\rangle = \sin^2\theta_i \left[1 + \tan^2\theta_i \left(\frac{1-\rho}{1+\rho}\right)^2\right] \tag{5.2}$$

La modélisation est une étape cruciale de l'analyse ellipsométrique. Par exemple, le modèle de Sellmeier, représenté à l'équation 5.3, est idéal pour modéliser la dispersion de certains matériaux, particulièrement les matériaux transparents dans le visible. Une fois le modèle adéquat choisi pour chaque matériau de l'échantillon, l'optimisation des paramètres est nécessaire afin de corréler la pseudo constante diélectrique aux permittivités de chaque matériau de la couche.

$$\varepsilon = 1 + \sum_i \frac{B_i \lambda^2}{\lambda^2 - C_i} \tag{5.3}$$

Le modèle d'oscillateur de Lorentz (équation 5.4) est, quant à lui, utilisé pour modéliser la réponse optique de diélectriques ou semiconducteurs exhibant une résonance de charges liées. Le modèle de Tauc-Lorentz (équations 5.5, 5.6) est dédié à la modélisation de matériaux semiconducteurs amorphes à des fréquences plus grandes que leur bande interdite.

$$\tilde{\varepsilon} = \varepsilon_1^{(\text{offset})} + \frac{AE_c}{E_c^2 - E^2 - iBE} \tag{5.4}$$

$$\varepsilon(E) = \varepsilon_\infty + \chi^{TL}(E) \tag{5.5}$$

$$\text{Im}\left\{\chi^{TL}(E)\right\} = \begin{cases} \frac{1}{E}\frac{AE_0C(E-E_g)^2)}{\left(E^2-E_0^2\right)^2+C^2E^2} & \text{si } E > E_g \\ 0 & \text{si } E \leq E_g \end{cases} \tag{5.6}$$

Par Kramers-Kronig la partie réelle de la susceptibilité électrique des matériaux est liée à sa composante réelle, une contrainte additionnelle durant la modélisation ellipsométrique :

$$\text{Re}\left\{\chi^{TL}(E)\right\} = \frac{2}{\pi}\int_{E_g}^{\infty} \frac{\xi \text{Im}\left\{\chi^{TL}(\xi)\right\}}{\xi^2 - E^2}\, d\xi. \tag{5.7}$$

### 5.2.2 Méthodologie

L'outil utilisé pour mesurer l'indice de réfraction du $Ge_2Sb_2Se_5$ est l'ellipsomètre spectroscopique à angle variable (VASE) RC2-XI J.A. Woollam Co., Inc. Les échantillons amorphes et cristallins de 50 nm d'épaisseur, encapsulés et non-encapsulés et préparés selon la procédure énoncée en 5.1 sont ensuite mesurés à l'ellipsomètre à des angles incidents de 45°, 55°, 65° et 75° sur une plage de longueurs d'onde de 500 nm à 2.5 µm.

### 5.2.3 Résultats

Les modèles communément employés pour caractériser les couches minces de di et trichalcogénures sont les modèles Tauc-Lorentz et Cody Lorentz [132], ce dernier comprenant un terme de queue d'Urbach, soit l'absorption au-delà de l'énergie du bandgap. Pour le $Ge_2Sb_2Se_5$

amorphe (figure 5.5 [90]), une excellente concordance avec un modèle de Tauc-Lorentz simple a été obtenue ($MSE = 1.9$) à partir des paramètres suivants : $\varepsilon_{inf} = 1.3$, $A = 126$, $C = 4.9$, $E_0 = 3.6$ et $E_g = 1.47$. En comparaison aux mesures obtenues par [90], l'indice de réfraction est moindre alors que $\kappa$ disparaît à 850 nm au lieu de 1.1 µm. Les deux matériaux exhibent tout de même une transparence dans la bande C, une condition essentielle au fonctionnement de la mémoire.

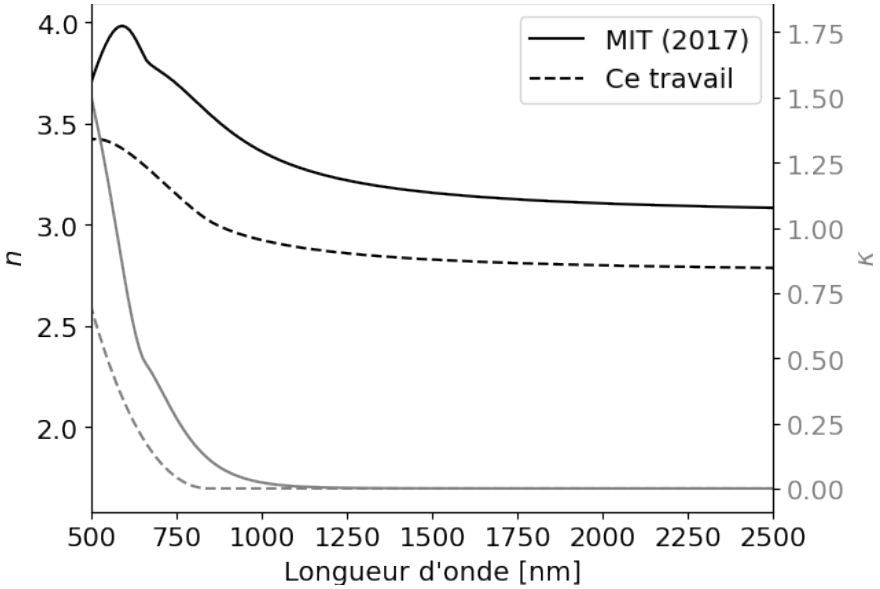



FIGURE 5.5 Comparaison des modèles ellipsométriques pour le $Ge_2Sb_2Se_5$ amorphe.

La différence d'indice de réfraction ($\Delta n \approx 0.5$) est considérable. Ceci peut être expliqué par la porosité du dépôt. En pulvérisation, la pression de dépôt est généralement liée à la porosité d'une couche mince [133–135]. Les résultats ellipsométriques en fonction de la pression de dépôt sont présentés à la figure 5.6.

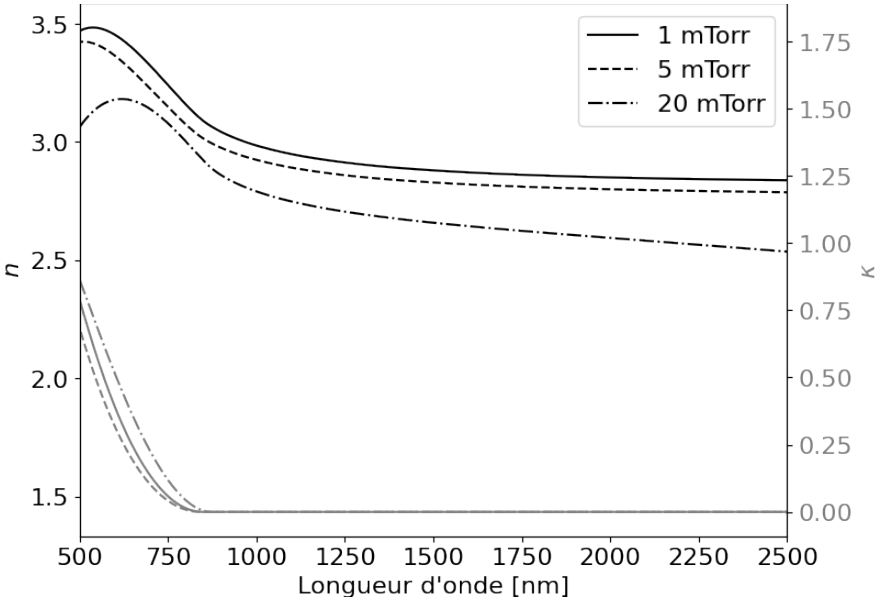



FIGURE 5.6 Impact de la pression de dépôt sur l'indice de réfraction du matériau.

On remarque que l'indice de réfraction augmente lors que la pression de dépôt diminue, indiquant que la porosité est moindre pour les couches minces déposées à 5 et 1 mTorr. Puisque le pulvérisateur utilisé ne permet pas les dépôts à pression $< 1$ mTorr, il est impossible de confirmer que l'écart d'indice de réfraction entre ce travail et ceux de Zhang et al. est entièrement dû à ce paramètre.

Pour l'échantillon cristallisé par recuit lent, des modèles similaires sont utilisés. Selon Zhang et al., un oscillateur additionnel est nécessaire pour obtenir une meilleure concordance aux résultats ellipsométriques. Les résultats pour le PCM cristallin sont présentés à la figure 5.7 [90].

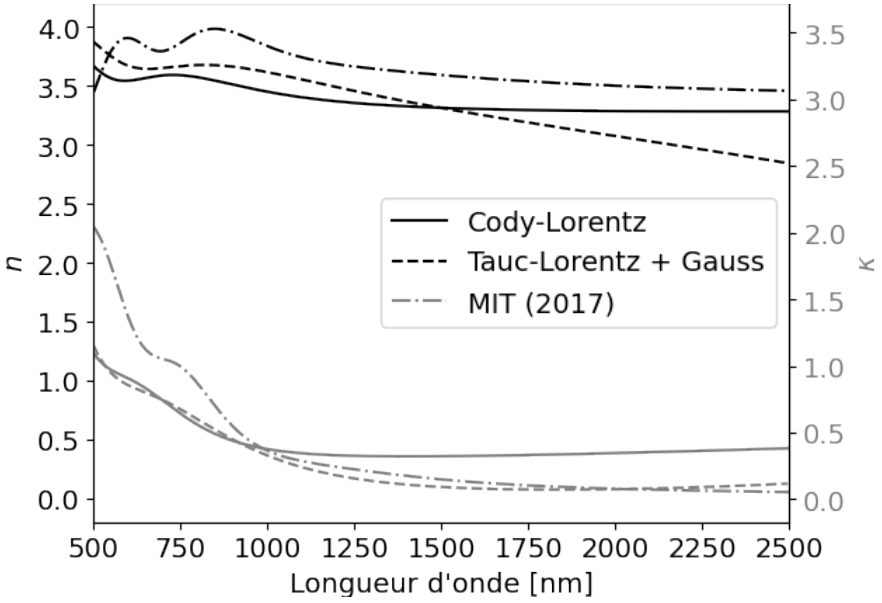



FIGURE 5.7 Comparaison des modèles ellipsométriques pour le $Ge_2Sb_2Se_5$ cristallin.

L'indice de réfraction obtenu est similaire pour les deux modèles étudiés et les travaux du MIT. L'absorption, liée au paramètre $\kappa$, quant à elle, est 5 à 8 fois inférieure à celle du MIT selon le modèle choisi à $\lambda = 1550$ nm.

## 5.3 Spectrophotométrie

### 5.3.1 Théorie

La spectrophotométrie est une technique dont la théorie est connue depuis le 19ème siècle, fortement inspirée des écrits de Newton sur la nature de la lumière. En 1929, Arthur C. Hardy présente les premiers résultats spectrophotométriques en réflexion recueillis «rapidement» de manière «automatique», sous le nom de *Color Analyzer* [136].

La spectrophotométrie est une méthode de mesure quantitative des propriétés optiques d'un matériau. Elle consiste en la mesure de 6 quantités spectrophotométriques soient la réflectance, la transmittance, l'absorbance, l'émittance, la diffraction et la fluorescence. À l'aide de ses 6 quantités, il est possible de calculer diverses propriétés optiques du matériau étudié comme son indice de réfraction ($n$), son coefficient d'absorption ($\alpha$), sa rugosité et sa densité de contaminants. Il s'agit d'une méthode polyvalente utilisée autant pour la caractérisation de simples CCDs que pour la métrologie de procédés nanotechnologiques de pointe en raison de son faible coût, de sa grande précision et de sa nature non-destructive. Les spectrophotomètres sondent généralement une grande plage de longueurs d'onde incidentes permettant la

caractérisation rapide d'un échantillon dans l'ultraviolet, le visible et l'infrarouge.

Les principales quantités d'intérêt en spectrophotométrie sont la réflectance et la transmittance. Elles consistent donc dans les cas simples, homogènes et isotropes, de scalaires normalisées, représentant le ratio de la puissance incidente réfléchie ou transmise par l'échantillon respectivement. En considérant la puissance optique $\Phi$ (W) et la radiance $L$ comme la densité de puissance optique émise par ou collectée sur une surface d'aire $\mathrm{d}A$, avec un angle solide $\mathrm{d}\Omega$. On calcule la réflectance et la transmittance comme les ratios suivants :

$$\rho\left(\Omega_i, A_i, \Omega_r, A_r; L_i\right) = \frac{\Phi_r\left(\Omega_i, A_i, \Omega_r, A_r; L_i\right)}{\Phi_i\left(\Omega_i, A_i; L_i\right)} \tag{5.8}$$

et

$$\tau\left(\Omega_i, A_i, \Omega_t, A_t; L_i\right) = \frac{\Phi_t\left(\Omega_i, A_i, \Omega_t, A_t; L_i\right)}{\Phi_i\left(\Omega_i, A_i; L_i\right)}. \tag{5.9}$$

Cette expression est peu utile et se doit d'être ajustée ou simplifiée selon les conditions de mesure et la nature de l'échantillon à analyser. En considérant une mince couche de matériau déposée uniformément, l'utilisation d'une source de radiance uniforme et isotrope ainsi qu'une méthode de détection hémisphérique, ces expressions se simplifient aux équations suivantes :

$$\rho\left(\theta_i, \phi_i; \Omega_r\right) = \int_{\Omega_r} \cos\theta_r f_r\left(\theta_i, \phi_i, \theta_r \phi_r\right) \mathrm{d}\Omega_r \tag{5.10}$$

et

$$\tau\left(\theta_i, \phi_i; \Omega_t\right) = \int_{\Omega_t} \cos\theta_t f_t\left(\theta_i, \phi_i, \theta_t \phi_t\right) \mathrm{d}\Omega_t. \tag{5.11}$$

Ici, pour l'instrument utilisé, le coefficient d'absorption est généralisé pour inclure les pertes par absorption et par diffusion non recuillies par le détecteur du spectrophotomètre. Ces pertes correspondent ainsi simplement à la puissance radiative manquante n'ayant ni été transmise, ni réfléchie :

$$\alpha\left(\theta_i, \phi_i\right) = 1 - \rho\left(\theta_i, \phi_i; 2\pi\right) - \tau\left(\theta_i, \phi_i; 2\pi\right). \tag{5.12}$$

Cette quantité est ensuite directement liée à la composante imaginaire de l'indice de réfraction :

$$\alpha = \frac{4\pi\kappa}{\lambda_0} \tag{5.13}$$

La fluorescence diffère des quantités présentées précédemment en raison de la différence entre

la longueur d'onde incidente et la longueur d'onde transmise par l'échantillon au photodétecteur. Sa caractérisation est donc basée sur le ratio du spectre émit sur le spectre mesuré. Finalement, l'émittance est la mesure spectrophotométrique analogue à l'énergie radiative émise par l'échantillon selon sa température et la loi de Planck :

$$B(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{k_B T}} - 1}. \tag{5.14}$$

L'émittance est définie comme le produit entre la radiance spectrale du corps noir et l'émissivité du matériau en question. Dans le cas d'un système parfaitement isolé, on retrouve que l'émissivité d'un matériau est identique à son absorbance [137].

### 5.3.2 Méthodologie

Les mesures spectrophotométriques on été effectuées avec le Cary 7000 de *Agilent Technologies* couvrant une plage de 175 nm à 3.3 µm. Deux échantillons de PCM ont été préparés avec un dépôt de 500 nm de $Ge_2Sb_2Se_5$ par pulvérisation RF sur un substrat de Si (100) *Double-side polished* (DSP). Ensuite, un mince couche de $SiO_2$ (15 nm) a été déposée pour prévnir l'oxidation du PCM. Le coefficient d'absorption $\alpha$ extrait de ces mesures en comparaison avec les celui mesuré par ellipsométrie est présenté à la figure 5.8.

### 5.3.3 Résultats

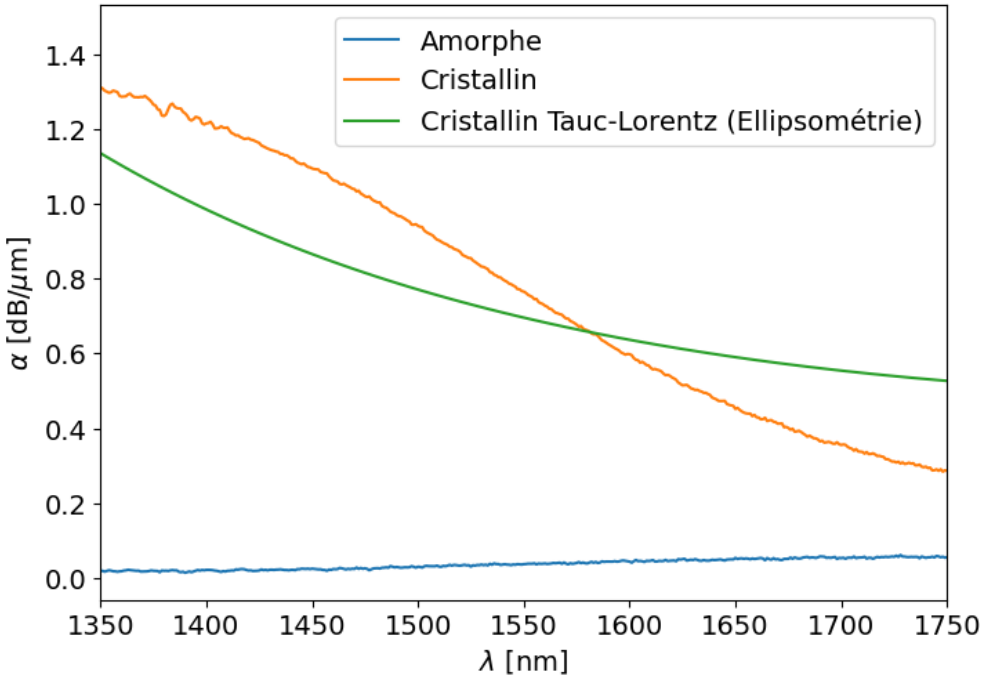



FIGURE 5.8 Coefficient d'absorption déterminé par spectrophotométrie.

On remarque que l'absorption dans l'état cristallin et amorphe est similaire pour les deux tests.

## 5.4 Spectroscopie photoélectronique à rayons X (XPS)

### 5.4.1 Théorie

Une analyse XPS est destinée à l'identification et à l'analyse en surface. Elle consiste en l'irradiance d'un échantillon par des rayons X de faible énergie sous un excellent vide, causant la photoionisation des atomes en surface de l'échantillon. La mesure de l'énergie cinétique des atomes émits suivi de l'analyse de la distribution de cette énergie permet de déterminer la nature des composés, leur état chimique ainsi que leur stochiométrie. Dans le cadre d'un dépot par couche mince, la XPS permet de mesurer la qualité du dépot par la mesure de ses contaminants et l'uniformité élémentaire du dépot. Lors de la mesure, les photoélectrons d'intérêt émis sont majoritairement ceux des orbitales en périphérie de l'atome, limitant cette méthode à une analyse de surface. L'énergie de liaison des photoélectrons $E_B$ est déterminée à partir de l'énergie de cinétique mesurée $E_K$ et par la fonction de travail $W$ du spectromètre

par la relation d'Einstein [138] :

$$E_B = h\nu - E_K - W. \tag{5.15}$$

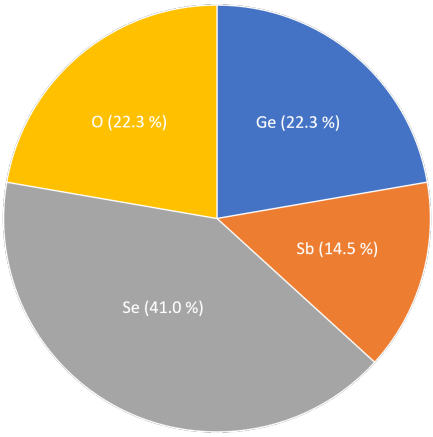
Cette méthode possède une faible résolution spatiale employant un faisceau de diamètre de quelques centaines de micromètres et mesurant une moyenne spatiale en fonction du niveau de focalisation du faisceau. Les pics du spectre obtenu incluent la contribution de photoémission des électrons profonds, de valence et Auger [139].

### 5.4.2 Méthodologie

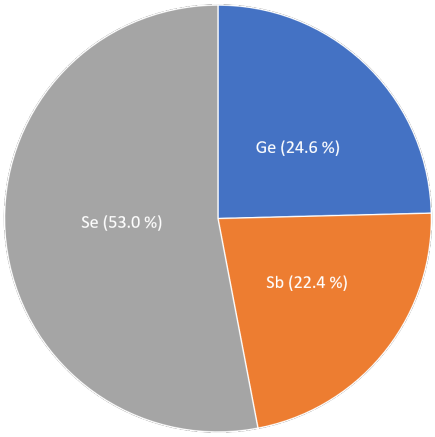
L'analyse XPS complète a été effectuée par Josianne Lefebvre, associée de recherche à Polytechnique Montréal avec le spectromètre photoélectronique à rayons X *ESCALAB 250Xi*. Un échantillon, composé d'une couche mince de 100 nm de $Ge_2Sb_2Se_5$ sur Si (100) a été analysé dans l'état «tel que déposé» (amorphe).

### 5.4.3 Résultats

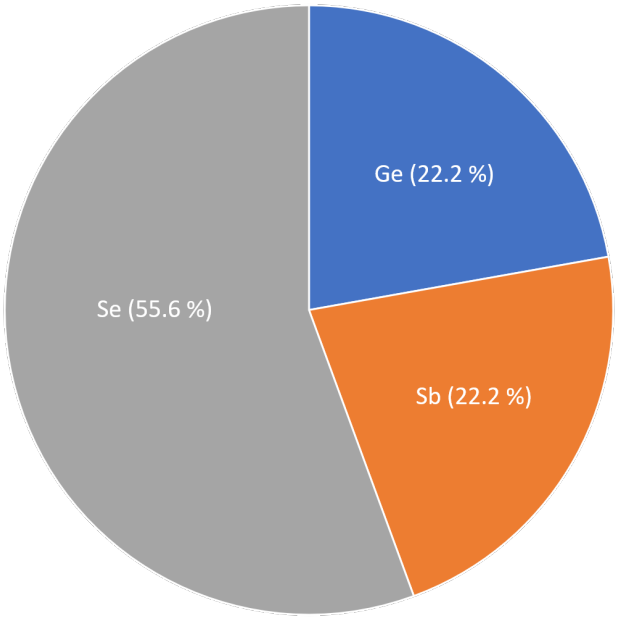
Le survol initial de l'échantillon a soulevé des traces de carbone adventice ainsi que de l'oxygène. Lors de l'analyse à haute résolution, l'oxydation en surface du chalcogénure répertoriée dans la littérature est confirmée par la présence de 22.3 % d'oxygène, sous forme d'oxyde de germanium et d'antimoine. L'importante contamination en surface est adressée par le décapage par faisceau d'argon polyatomique à 8 keV durant 60 secondes. Suite à ce décapage de l'oxyde natif, une seconde analyse de surface a été effectuée pour établir la réelle composition stoechiométrique produite par le dépot par pulvérisation. Les résultats sont représentés à la figure 5.9.




(a) Tel que déposé.




(b) Oxyde natif décapé.




(c) Stoechiométrie désirée.

FIGURE 5.9 Composition atomique du $Ge_2Sb_2Se_5$ déposé par pulvérisation au LMF (Polytechnique Montréal)

On note que l'échantillon a) «tel que déposé», sans encapsulation, s'oxyde de manière significative tel que près du quart de la surface sondée est constituée d'oxygène, soit 7 % sous forme d'oxyde de germanium et 2.2 % sous forme d'oxyde d'antimoine. Suite au décapage de l'oxyde b), la stoechiométrie est de la couche mince légèrement riche en germanium ($\Delta$Err = 2.2 %) et pauvre en sélénium ($\Delta$Err = 2.6 %) par rapport à la référence de la cible c). Ceci confirme à la fois la prévalence de l'oxydation du chalcogénure ainsi que la capacité du dépôt par

pulvérisation à conserver la stoechiométrie de la cible utilisée.

## 5.5 Diffraction par rayons X

### 5.5.1 Théorie

La diffraction par rayons X est une méthode non-destructive communément utilisée en cristallographie pour étudier et caractériser l'organisation des atomes d'un matériau cristallin. Comme son nom le suggère, elle est basée sur la diffraction de photons à très haute énergie (≈8000 eV) engendrée par la maille atomique. La très petite longueur d'onde de ces photons (source CuK$\alpha$ = 1.5418 Å) est similaire au paramètre de maille des cristaux, cette dernière agissant comme réseau de diffraction 3D pour les rayons X. Elle permet la mesure du paramètre de maille, du réseau de Bravais, de la taille des grains ainsi que la présence de défauts. Le patron d'interférence constructive obtenu obéit ainsi à la loi de Bragg :

$$n\lambda = 2d\sin\theta, \tag{5.16}$$

où $n$ est l'ordre de diffraction, $d$ le paramètre de maille cristalline et $\theta$, l'angle du faisceau de rayons X incident.

La signature spectrale permet, dans la plupart des cas, l'identification du matériau et nécessite généralement moins de 30 minutes à obtenir. Il s'agit également d'une méthode polyvalente, s'appliquant aux couches minces, tout comme aux fluides et aux matériaux sous forme de poudre. Un source de rayons X, généralement un tube à rayon cathodique, expose la surface de l'échantillon à caractériser, incliné à un angle theta. Les rayons S diffractés sont ensuite mesurés par un détecteur placé à angle $2\theta$ et l'intensité du signal diffracté est collectée sous forme de compte ou d'intensité relative. Un fort compte à un angle $\theta_0$ donné implique que la condition d'interférence constructive de Bragg vue précédemment est respectée. La distance d entre les atomes de la maille peut par la suite être calculée pour toutes les conditions d'interférence obtenues en fonction de l'angle ($2\theta$).

### 5.5.2 Résultats

La structure cristalline du $Ge_2Sb_2Se_5$ est peu répertoriée dans la littérature et aucune analyse XRD complète est répertoriée dans les bases de données *COD* (Base de données cristallographiques ouverte) et *ICDD* (Centre international de données de diffraction). Bien que les épaisseurs visées dans le cadre de ce projet sont inférieures à 100 nm, un échantillon de 400 nm de $Ge_2Sb_2Se_5$ a été employé afin de maximiser le signal XRD et réduire la durée de la

mesure.

Les résultats obtenus par diffraction par rayons-X sont présentés à la figure 5.10 et 5.11 [15].

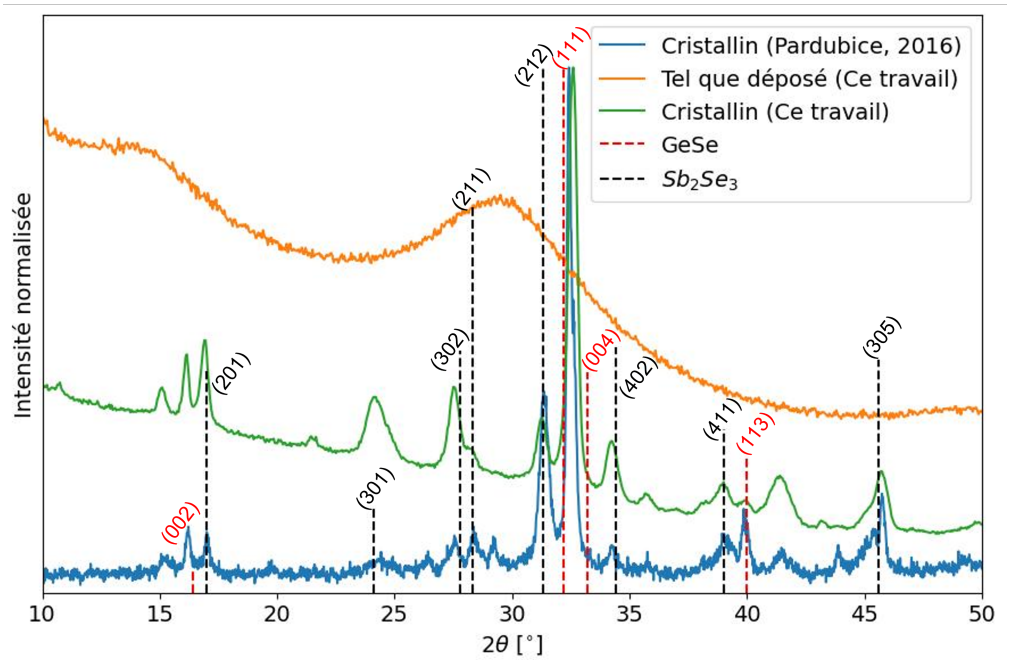



FIGURE 5.10 Résultats de diffraction à rayons-X pour les deux états de $Ge_2Sb_2Se_5$ en comparaison avec la littérature et avec la base de données cristallographique ouverte *COD*.

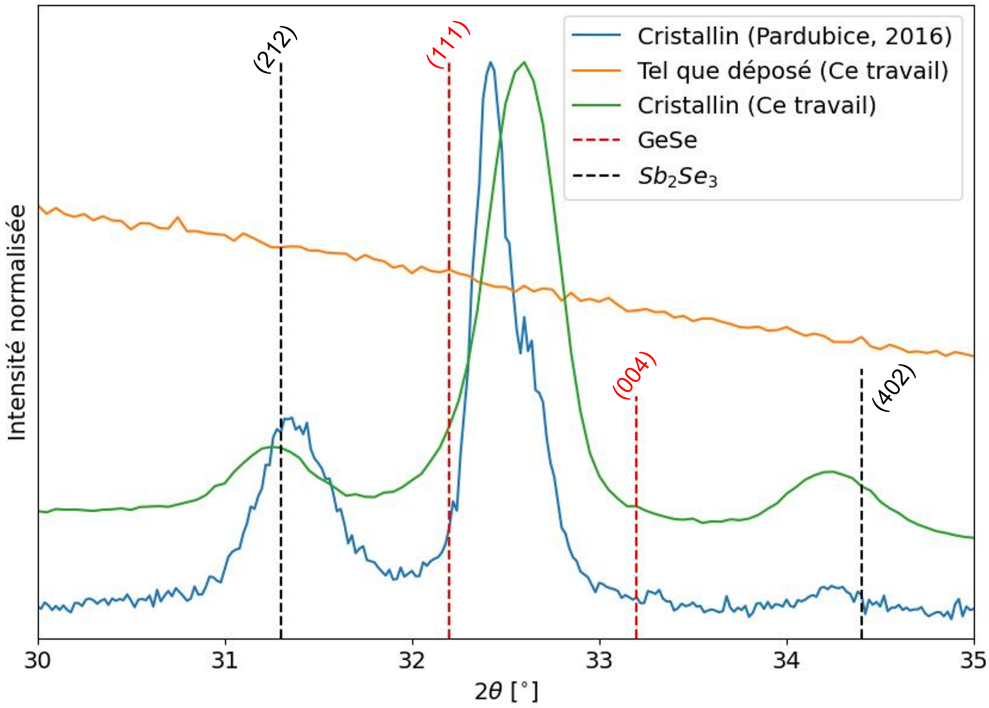



FIGURE 5.11 Zoom sur les pics significatifs à 32.5 °.

On observe d'abord une bonne concordance entre les mesures effectuées dans le cadre de ce travail avec les résultats de Barták et al., et ce particulièrement pour l'intensité relative des pics à $2\theta$ ≈16.2°, 17°, 31.3°, 32.5°, 34.2°, 39° et 45.7°. Ensuite, une analyse théorique effectuée à partir du logiciel *Match !* et de données de diffraction publiques du GeSe et du $Sb_2Se_3$ sont superposés pour illustrer les pics importants de ces deux composés. L'intensité relative de ces pics est déterminée à partir de la composition $2GeSe + 1Sb_2Se_3$ pour simuler le $Ge_2Sb_2Se_5$. Il est important de mentionner que tous ces composés cristallisent sous forme orthorhombique ce qui justifie une telle analyse.

La caractérisation du matériau à changement de phase est maintenant bien établie, permettant de procédé au développement d'un procédé de fabrication et d'intégration de la mémoire optique dans un circuit photonique. Les détails du procédé et de la méthodologie fait objet de la section suivante.

# CHAPITRE 6 FABRICATION ET MÉTHODOLOGIE EXPÉRIMENTALE

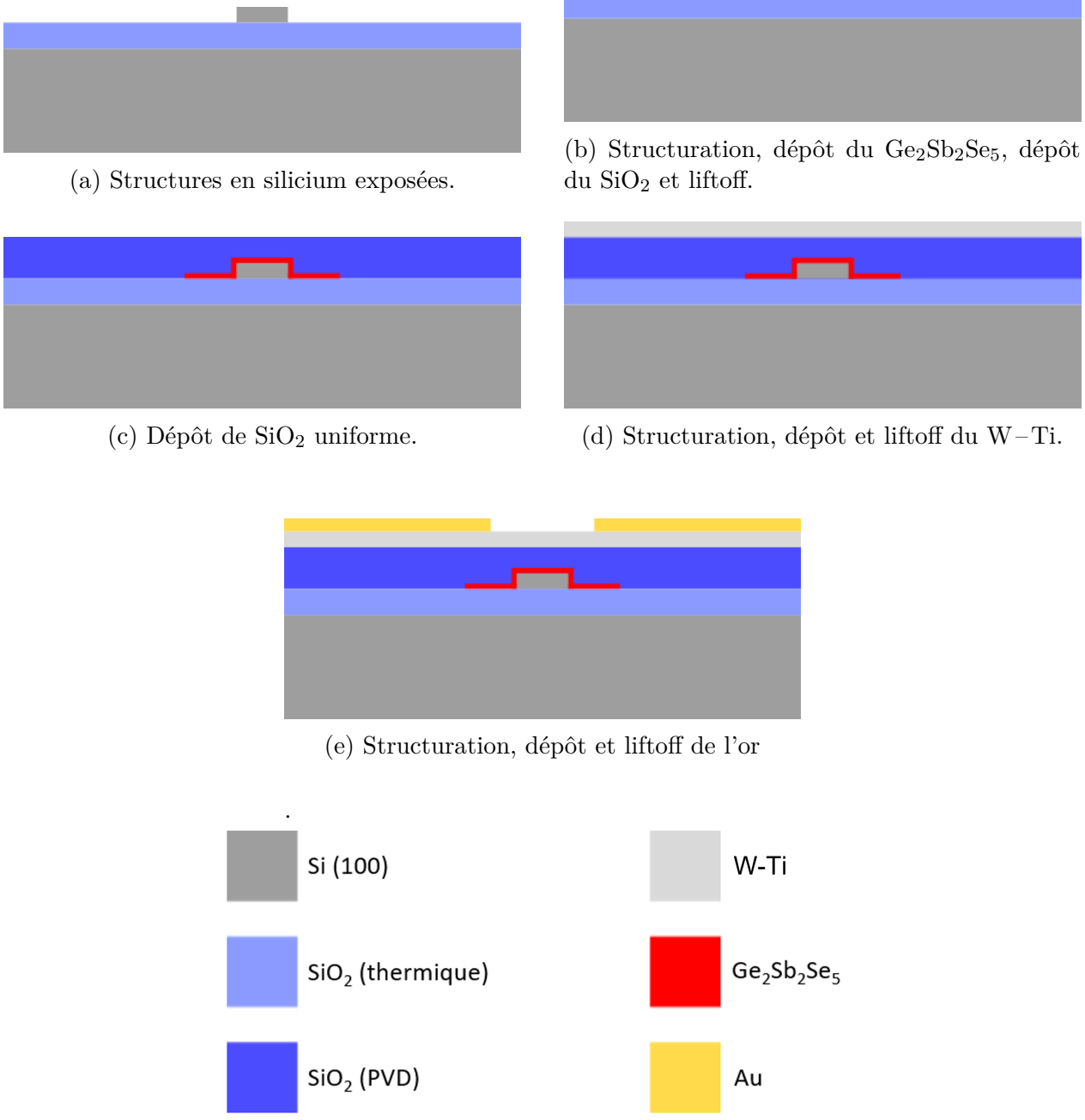
## 6.1 Procédé de fabrication

Le procédé de fabrication débute par une puce photonique sur silicium, composée d'un oxyde thermique enfoui d'une épaisseur de 2 µm agissant comme gaine pour les structures photoniques. La couche utile (*device*) de silicium possède une épaisseur de 220 nm et est préalablement gravée de manière anisotrope. Pour minimiser les pertes de propagation, la puce photonique est fabriquée par un procédé dédié en fonderie par *Advanced Micro-Foundry* ou *Applied Nanotools*.

Les étapes principales du procédé de fabrication sont présentées à la figure 6.1. Les cellules de mémoire sont d'abord structurées par liftoff, composé d'un procédé de photolithographie UV et d'un dépôt de 20 nm de $Ge_2Sb_2Se_5$ par pulvérisation à une puissance RF de 50 W, une pression de dépôt de 5 mTorr et un débit d'argon de 10 sccm. Ce procédé produit un taux de dépôt d'approximatif de 10 nm/min. Le GSSe est immédiatement encapsulé par 15 nm de $SiO_2$ à 200 W, 5 mTorr et 10 sccm d'argon (≈5 nm/min). Le liftoff est effectué dans un bain de Microposit™ Remover 1165 pour 12 h à température pièce b).

Un couchage de $SiO_2$ sur l'entièreté de la puce est ensuite effectué par le même procédé de pulvérisation jusqu'à obtention de 400 nm d'épaisseur c). Les éléments chauffants en W-Ti (90%-10%) sont ensuite structurés par un procédé lithographique identique à celui utilisé pour les cellules de mémoire. Le dépôt de 300 nm de tungstène-titane est effectué par pulvérisation RF à 200 W, 5 mTorr, et 10 sccm d'argon pour 30 minutes. La procédure de liftoff subséquente est identique à celle utilisée pour le GSSe d).

Un procédé identique de liftoff est effectué pour établir les connexions en or. 50 nm d'or sont déposés par évaporation à faisceau d'électrons à un taux de 1 à 2 Å avant de procéder au liftoff e). Les masques utilisés pour les diverses étapes photolithographiques sont présentés en annexe B. Les détails du procédés photolithographique font l'objet de l'annexe C.




(a) Structures en silicium exposées.

(b) Structuration, dépôt du $Ge_2Sb_2Se_5$, dépôt du $SiO_2$ et liftoff.

(c) Dépôt de $SiO_2$ uniforme.

(d) Structuration, dépôt et liftoff du W – Ti.

(e) Structuration, dépôt et liftoff de l'or

FIGURE 6.1 Procédé de fabrication d'une cellule de mémoire.

## 6.2 Montage optique

Le montage conçu (voir figure 6.2) est un montage typique de photonique intégrée active par couplage hors-plan, utilisant une unité de 8 fibres optiques, nommée FAU (*Fiber-Array*

*Unit*) provenant de Oz Optics (VGA-8-127-8-A-14.4-4.0-1.2-8-P-1550-8/125-3S-0.25-0.5-0). Ceci permet le couplage hors-plan avec 8 signaux d'entrée ou de sortie simultanément. Pour évaluer la qualité du couplage, les pertes par insertion et les résonances du circuit photonique étudié, le laser accordable Keysight N7776C effectue un balayage en longueur d'onde autour de la plage d'intérêt de 1540 à 1560 nm et mesure la puissance en sortie par synchronisation avec un puissance-mètre à 8 ports Keysight N7745C. Le contrôle électrique destiné à l'écriture des cellules de mémoire est contrôlé par un générateur de fonction 33220A 20 MHz de Agilent connecté à la puce par un réseau de 10 aiguilles de type MCW (*Multi-Contact Wedge*) de GGB Industries.

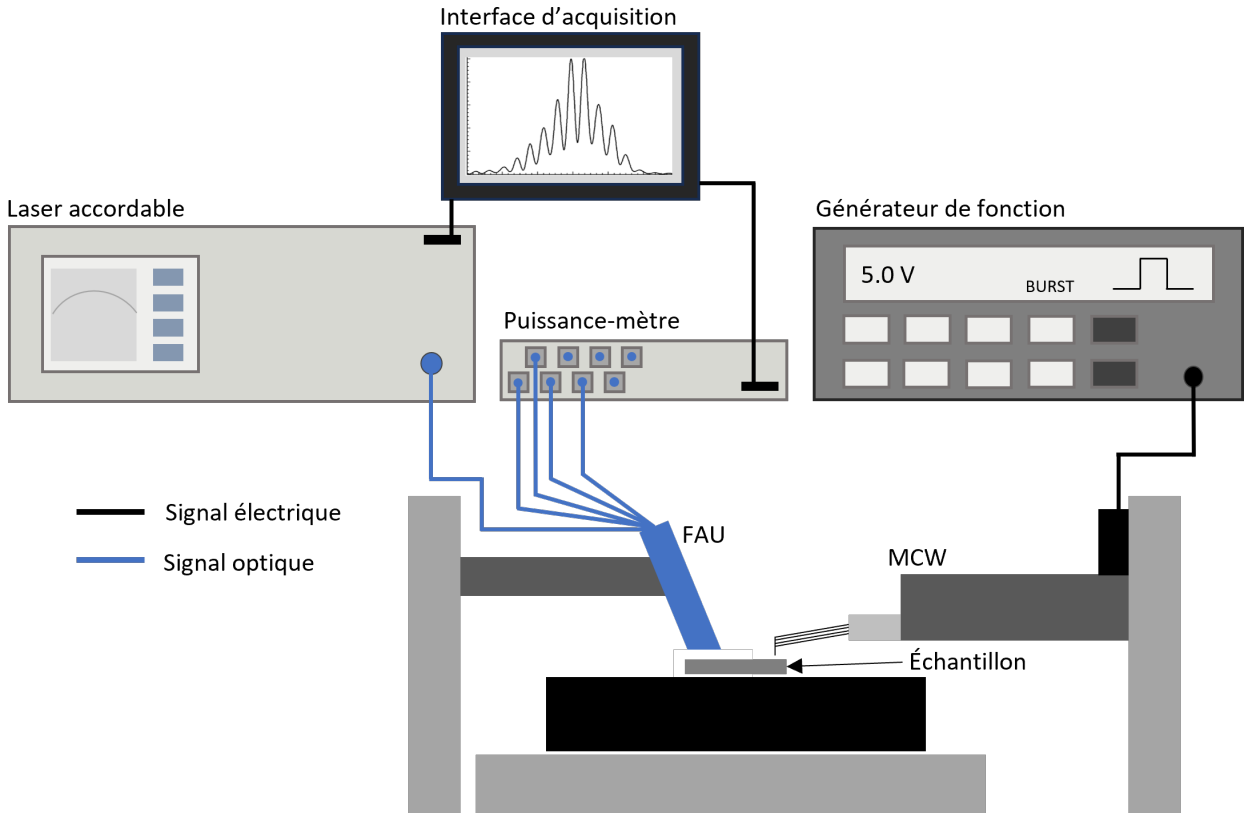



FIGURE 6.2 Schéma du montage.

L'échantillon est placé sur un porte-échantillon contrôlé thermiquement par une cellule Peltier qui posséde une sortie à vide pour éviter tout mouvement de la puce lors des tests. Le positionnement de l'échantillon est facilité par deux plateformes de translation motorisées VT-50L-12210 de *MICRONIX USA* disposant d'encodeurs magnétiques, permettant un déplacement de 10 cm et une précision de 50 nm sur le positionnement en $x, y$ de la puce. Les angles azimutaux et polaires de l'échantillon sont contrôlés par une plaque rotative 124-0055

et une plateforme d'inclinaison 123-2205 de Optosigma. Le FAU dispose également d'une plateforme motorisée similaire positionnée verticalement ($z$). Sur celle-ci, on retrouve une platforme rotative additionnelle pour l'inclinaison, ainsi qu'un goniomètre double GN2 de Thorlabs pour orienter les angle de roulement et de lacet des fibres. Les plateformes motorisées sont connectées à un controlleur Corvus eco de Micos. Pour le positionnement des aiguilles sur les contacts électriques, un positionneur 3-axes Newport 562 Series sur lequel le MCW est fixé est utilisé.

La prise de mesure débute par l'alignement du FAU au-dessus d'un réseau de 8 coupleurs de sorte à maximiser la puissance mesurée en sortie. Une caméra CMOS DCC1645C de Thorlabs munie d'un systèmes de lentilles à zoom variable x24 de Navitar est positionnée à un angle approximatif de 45° à l'opposé du FAU et utilisée pour faciliter l'alignement optique. L'angle de la caméra permet d'éviter l'usage de deux caméras simultanément ce qui prévient le FAU de percuter l'échantillon. Pour l'alignement électronique, une faible tension DC est appliquée aux bornes de deux aiguilles du MCW, ce circuit est complété par un ampèremètre indiquant la qualité du contact électrique entre le MCW et la puce. Pour cette étape, une caméra VCT-072-A-NTSC de Polytec GmbH avec vue de dessus est utilisée.

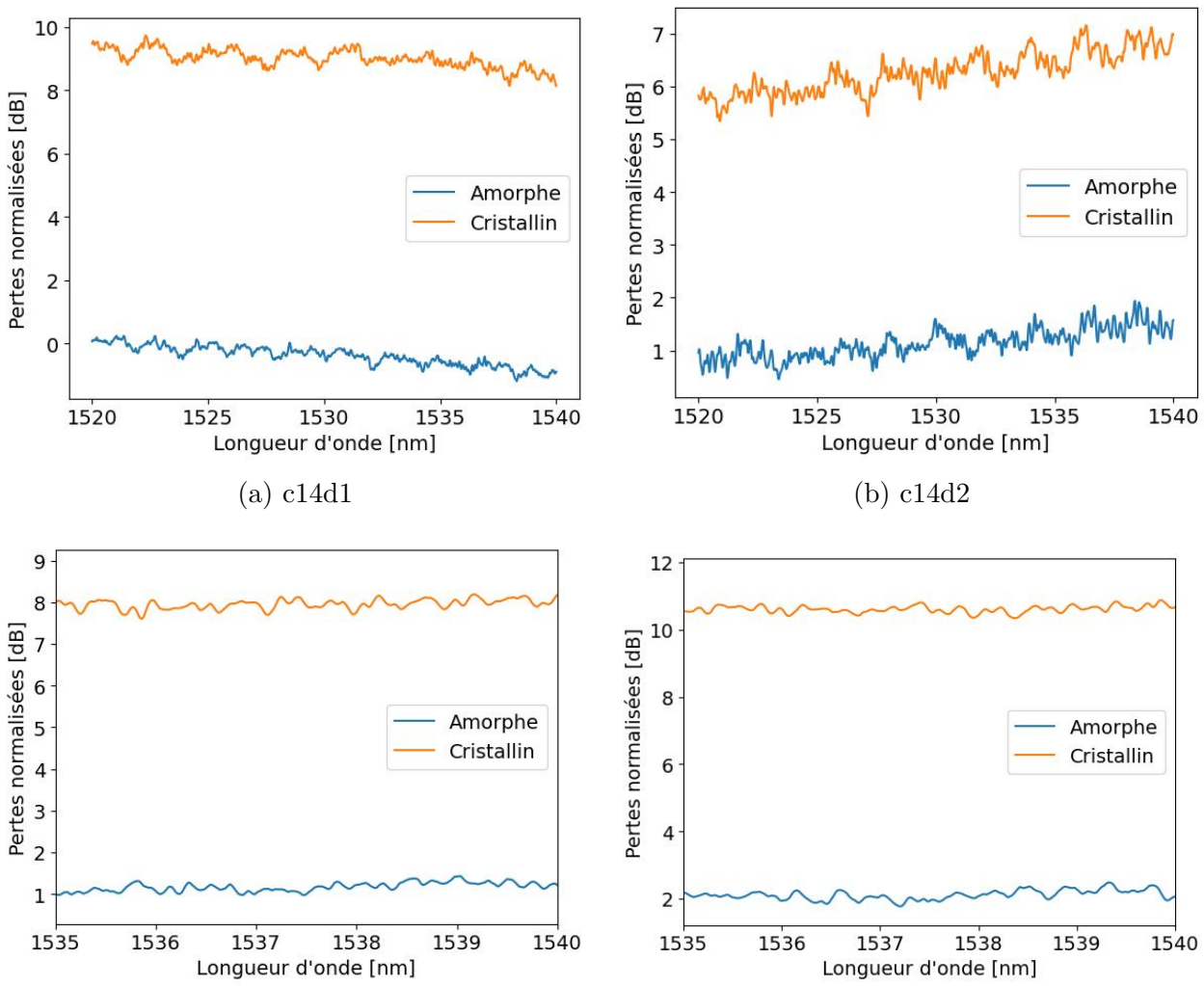
Une fois le couplage optique et les connexions électriques établies, un programme de tensions est envoyé par le générateur de fonction aux divers éléments chauffants associés aux cellules de mémoire monitorées optiquement. La mémoire étant initialement dans l'état amorphe, le programme d'impulsions vise à balayer l'espace des paramètres d'impulsion, soient son amplitude ($V_p$), sa durée ($t_p$), le nombre d'impulsions ($n_p$) et le régime d'utilisation (*duty-cycle*) ($D$), dans le but de déterminer les conditions de cristallisation. Une fois une absorption confirmée significative mesurée ($\sim 1 - 10$ dB), une faible tension est appliquée à l'élément chauffant pour vérifier la continuité du courant et assurer que les pertes optiques ne sont pas dues au bris de l'élément chauffant. Si l'absorption et la continuité du courant persiste, la cristallisation est confirmée. Un processus identique est ensuite effectué pour l'amorphisation durant lequel une diminution des pertes optiques est anticipée.

Maintenant que la méthodologie et la fabrication des mémoires est établie et optimisée, la présentation de la performance de la mémoire optique et de l'architecture MAC est présentée à la section suivante.

# CHAPITRE 7 ANALYSE DES RÉSULTATS

## 7.1 Changement de phase thermique

Avant d'étudier le changement de phase électrothermique, celui-ci est étudié sur une puce SiP pour tester le processus d'intégration de la mémoire sur guide d'onde. Des cellules de 50 nm de $Ge_2Sb_2Se_5$ (10 x 3 µm) ont été déposées sur divers guides d'onde. Des spectres de référence dans l'état amorphe ou «tel que déposé» sont mesurés dans la bande C avant d'effectuer un recuit de la puce dans un four à 350 C durant 30 minutes. Les spectres de la puce recuite et assumée cristallisée sont ensuite comparés avec la référence amorphe. Pour minimiser l'impact du couplage sur le résultat, les spectres sont normalisés selon une boucle d'alignement sans cellule de mémoire. La résultats pour 4 cellules de mémoire sont présentés à la figure 7.1




(a) c14d1

(b) c14d2

(c) c14d3

(d) c14d4

FIGURE 7.1 Absorption optique mesurée suite au recuit de cellules de $Ge_2Sb_2Se_5$.

On observe que l'échantillon recuit engendre significativement plus de pertes que l'échantillon tel que déposé. Selon les simulations (chapitre 4) effectuées et les propriétés du $Ge_2Sb_2Se_5$ (chapitre 5) mesurées, la différence d'absorption entre l'état amorphe et cristallin pour une cellule de 10 µm est estimée à 7-9 dB. Les pertes mesurées confirment une cristallisation, mais varient significativement entre 5 et 10.5 dB. Cette écart s'explique par les phénomènes de ségrégation et de *dewetting* confirmés à la section suivante.

### 7.1.1 Ségrégation et *dewetting*

Les phénomènes de ségrégation et de *dewetting* tels qu'expliqués à la section 2.3.3 sont suspectés être à l'origine de la grande différence d'absorption mesurée entre deux cellules de mémoire de même taille, situées sur la même puce et soumises au même recuit. Des images de ces deux phénomènes sont illustrées à la figure 7.2 pour une même cellule de mémoire suite à divers traitement thermiques.

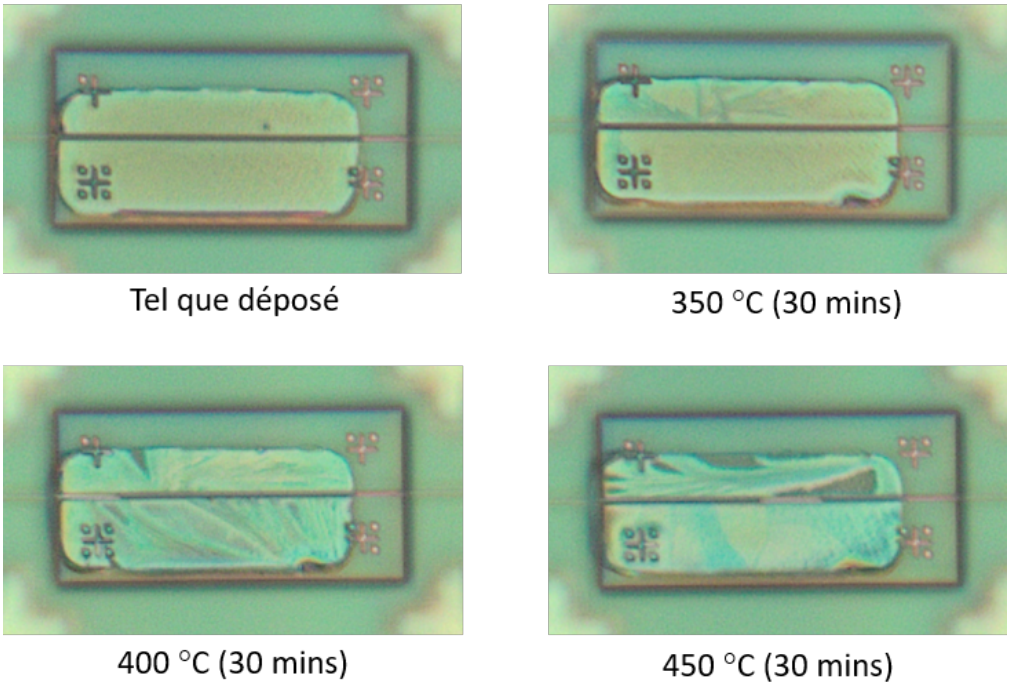



FIGURE 7.2 Phénomènes indésirables observés lors de la cristallisation par recuit lent d'une cellule de $Ge_2Sb_2Se_5$ de dimension 50 µm x 20 µm x 50 nm.

On peut confirmer que la cellule telle que déposée recouvre uniformément le guide d'onde par sa couleur foncée au microscope. Lorsque la cristallisation est atteinte, soit à 350 °C, le guide d'onde semble toujours recouvert, exhibant théoriquement l'absorption anticipée ($\approx$ 50 µm $\times$ 1 dB/ µm). Lorsque la cellule est chauffée à 400 °C, on observe la formation d'îlots et un retrait partiel du PCM situé sur le guide d'onde. En augmentant la température davantage, ce phénomène s'amplifie et peut justifier les grands écarts d'absorption entre les cellules identiques.

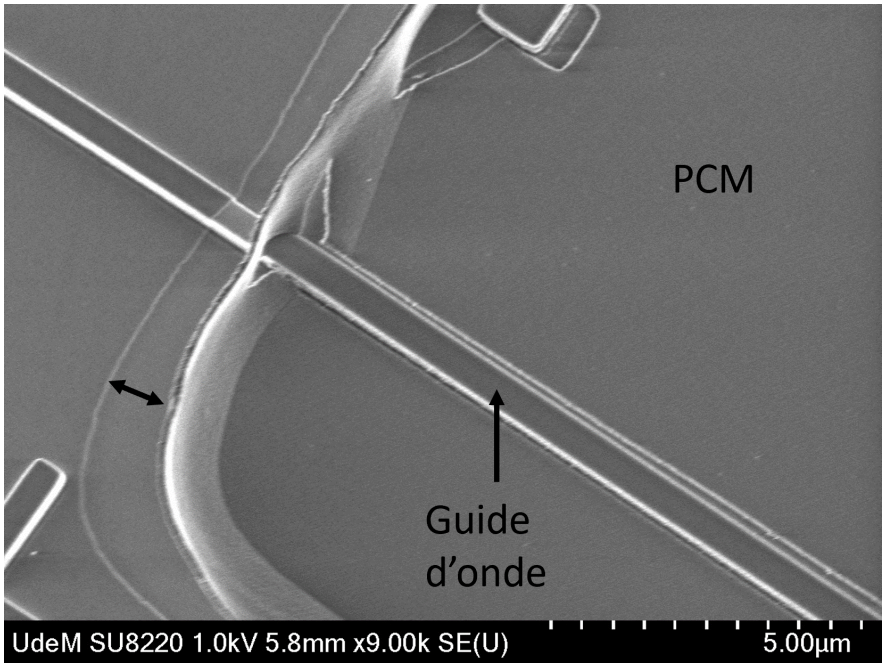



FIGURE 7.3 Représentation du *dewetting* suite à un recuit.

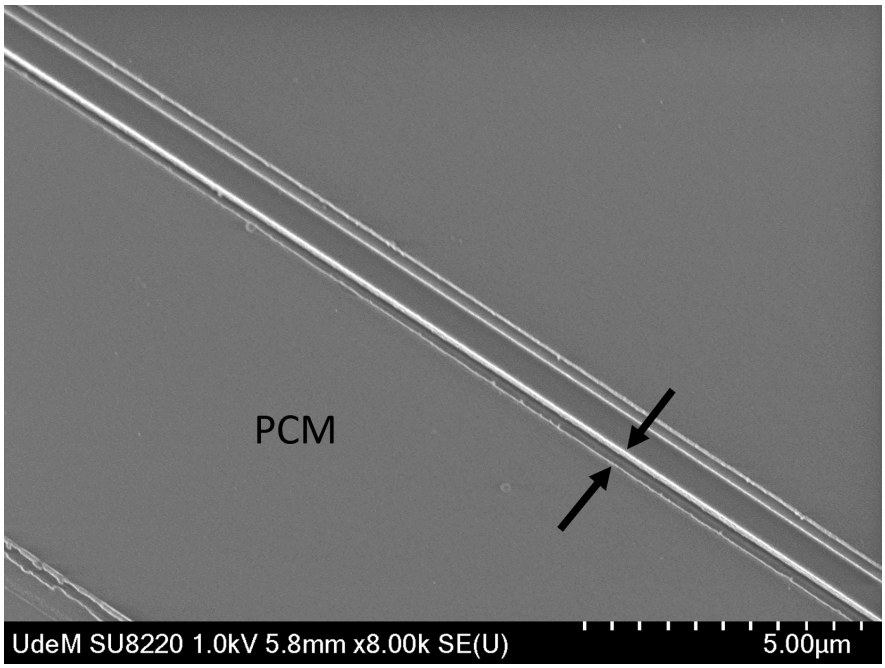



FIGURE 7.4 Représentation de la migration du PCM suite à un recuit.

FIGURE 7.5 Confirmation de la rétraction du PCM sur le guide suite à un recuit thermique

## 7.2 Changement de phase électrothermique

L'étape cruciale à la fabrication de la mémoire est la fabrication de l'élément chauffant. La nature du design et du procédé de fabrication en fonderie oblige la fabrication de ses composants sur Si et $SiO_2$, ceci entrainant des problèmes d'adhésion et un phénomène de *peel-off* lors du *lift-off* comme représenté à la figure 7.6.

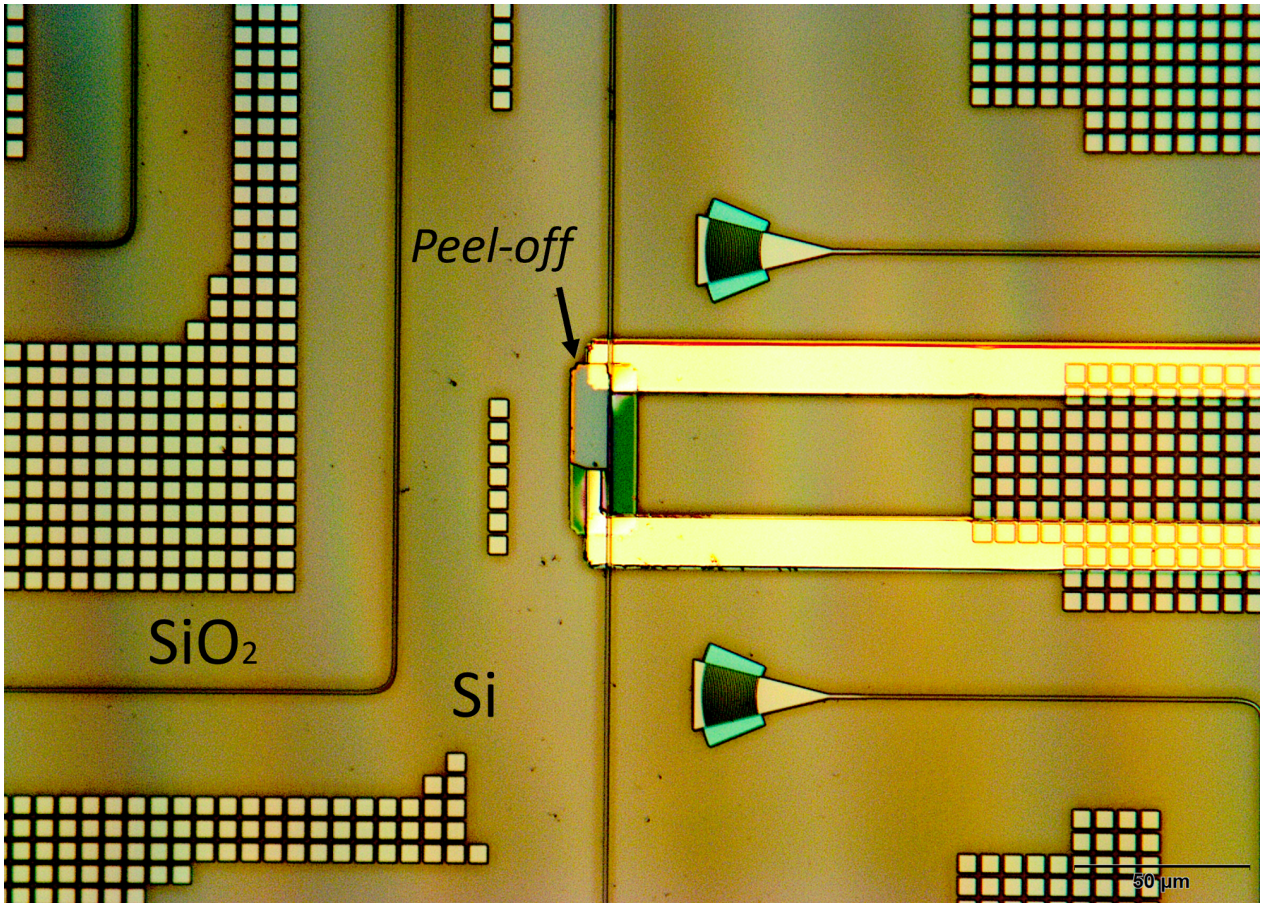



FIGURE 7.6 Problème d'adhésion lors de la fabrication des éléments chauffants.

De plus, la petite taille de la puce rend plus difficile l'étape de photolithographie des éléments chauffants, puisque l'application de résine à la tournette est très inhomogène, réduisant davantage le rendement de fabrication.

Ce problème s'est avéré atténué par l'ajout d'un trempage de 10 secondes de la puce dans l'acide fluorhydrique avant la structuration des éléments chauffants. Cette étape vise à homogénéiser la surface l'oxyde de silicium et ainsi maximiser l'adhésion du W-Ti. Une fois le dispositif fabriqué, la caractérisation du changement de phase s'effectue à partir d'un circuit test constitué d'une boucle contenant la mémoire et une boucle de référence pour minimiser

l'impact de la qualité du couplage sur la mesure. L'objectif de ce test est de déterminer la tension nécessaire pour cristalliser et amorphiser le PCM. Un exemple de dispositif est illustré à la figure 7.7.

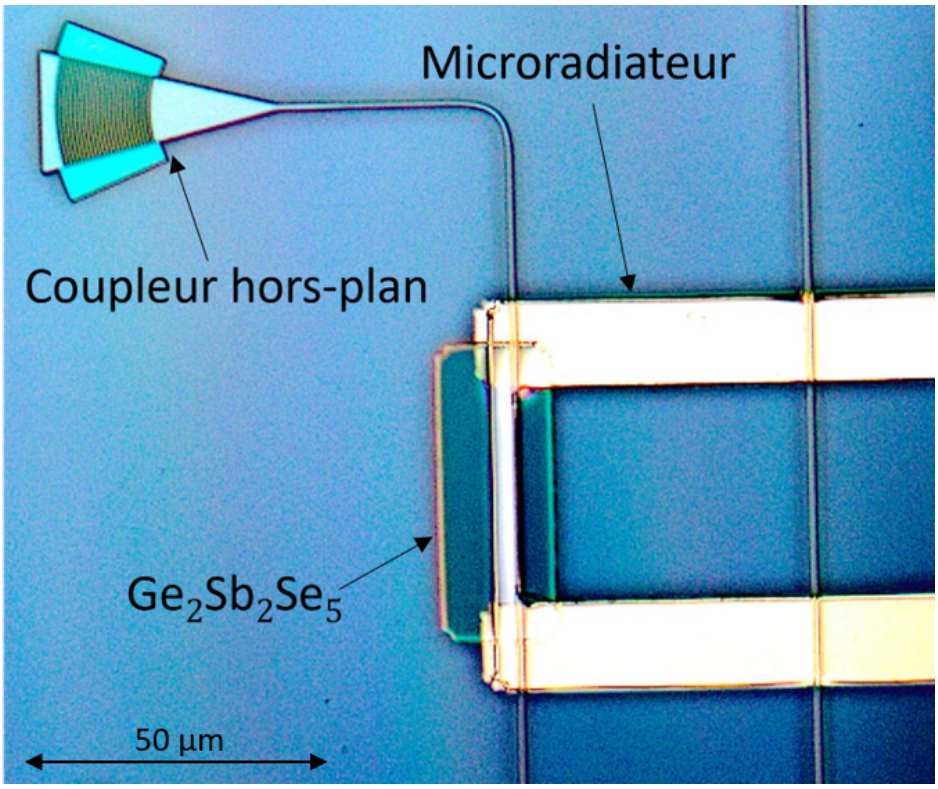
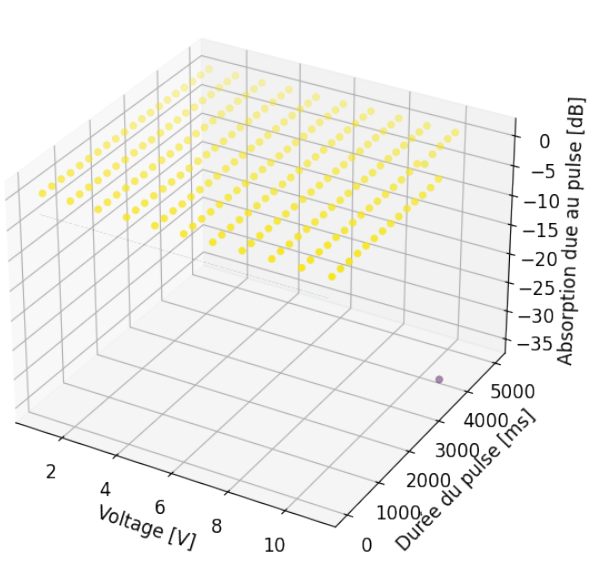



FIGURE 7.7 Image de la cellule de mémoire test fabriquée pour caractériser le changement de phase électrothermique.

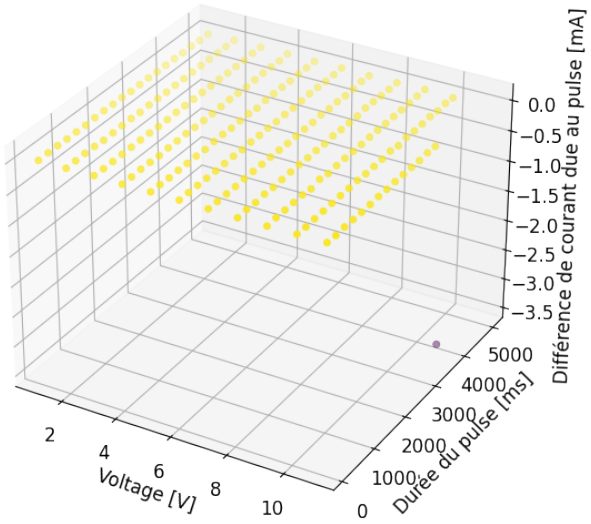
On observe que l'élément chauffant est adéquatement positionné au-dessus du guide d'onde et du PCM pour maximiser son efficacité thermique. Pour cette cellule de 50 x 20 µm, on s'attend, selon les résultats obtenus au chapitre 5, à observer des pertes de plus de 30 dB lors de la cristallisation.

Comme mentionné dans les travaux précédents similaires [123], la tension nécessaire pour la cristallisation et l'amorphisation est appliquée sous forme d'un ou d'une série d'impulsions, dans le but d'optimiser la vitesse du changement de phase et minimiser la consommation énergétique. Parmi les caractéristiques du programme d'impulsions, on retrouve son nombre, leur forme, leur tension, leur durée ainsi que leur cycle de travail. À partir des travaux de Miscuglio et Sorger, la cristallisation est atteinte suite à 20 impulsions carrées d'une durée de 1 µs et ayant un amplitude de 5 V. L'amorphisation est quant à elle obtenue suite à une seule impulsion carrée de 15 V d'une durée de 500 ns.

Au cours du projet, plus de 50 cellules de mémoire ont été testées afin de déterminer les conditions de première cristallisation. Sur ces cellules, une variété de programmes ont été testés, variant majoritairement la tension appliquée entre 0 et 15 $V_{pp}$ et la durée des impulsions entre 1 µs et 1 s. Comme mentionné dans la section 6.2, la cristallisation est confirmée uniquement si la continuité de l'élément chauffant est vérifiée suite à l'application du programme, impliquant que les pertes proviennent bien de la cristallisation et non du bris de l'élément chauffant. Deux exemples de résultats obtenus sont présentés aux figures 7.8 et 7.9.

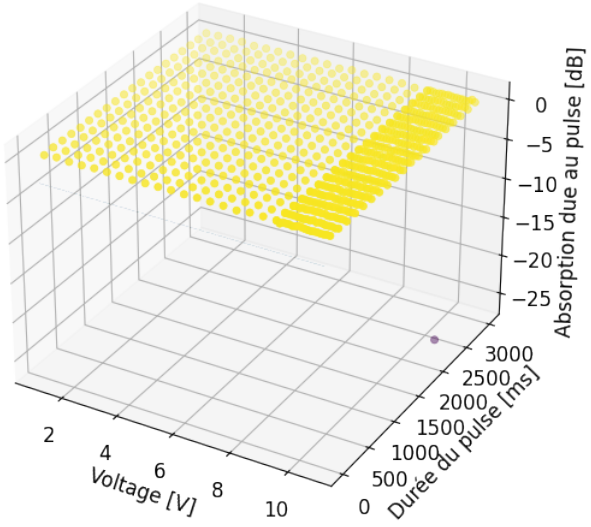



(a) Absorption mesurée suite aux impulsions.

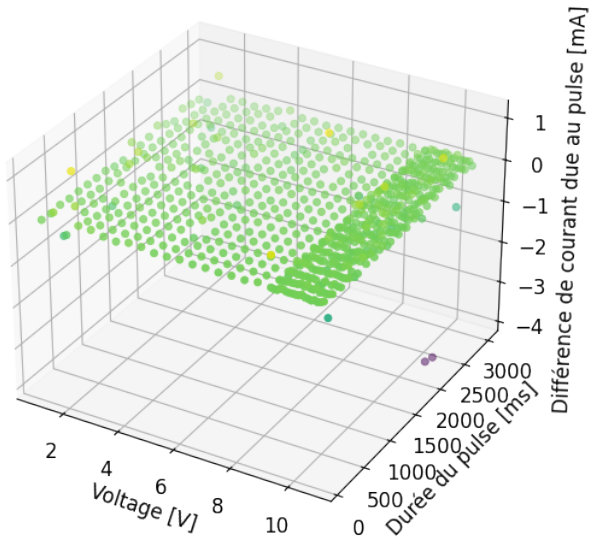



(b) Vérification du fonctionnement de l'élément chauffant suite aux impulsions.

FIGURE 7.8 Programmes de cristallisation testés pour une cellule de mémoire (1 impulsion, 50% cycle de travail).




(a) Absorption mesurée suite aux impulsions.




(b) Vérification du fonctionnement de l'élément chauffant suite aux impulsions.

FIGURE 7.9 Programmes de cristallisation testés pour une cellule de mémoire (1 impulsion, 50% cycle de travail).

On observe qu'une absorption significative de 35 et 25 dB a été rapportée suite à une impulsion de quelques secondes à 10.5 V pour les deux dispositifs a). Cependant, le bris de l'élément chauffant a été simultanément confirmé en b) par une diminution du courant. Ce phénomène a été observé sur des dizaines de dispositifs, indiquant que l'élément ne transmet pas efficacement la chaleur au PCM ou que ce dernier est trop fragile pour atteindre la $T_c$. Un exemple d'élément chauffant brisé est illustré à la figure 7.10.

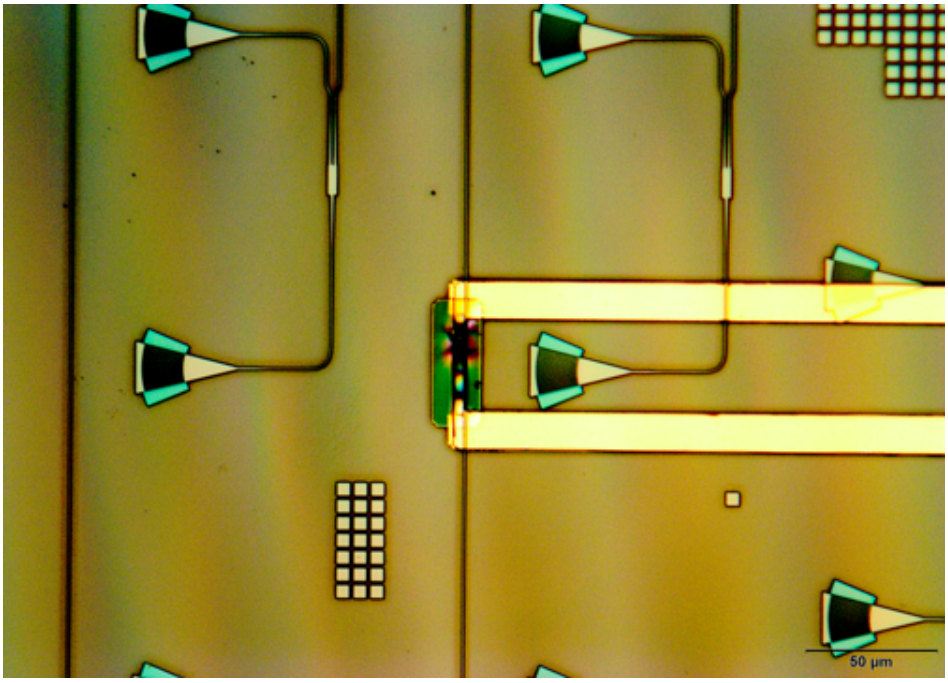



FIGURE 7.10 Cellule de mémoire détruite par le bris de l'élément chauffant.

Ce problème n'a pu être réglé au cours du projet, rendant impossible la vérification de l'amorphisation.

### 7.3 Architecture du mutliplicateur matrice-vecteur

Une fois la caractérisation de la mémoire effectuée, le choix d'architecture pour la multiplication matrice-vecteur doit être effectué. Comme vu théoriquement à la section 2.5 et plus spécifiquement à la section 3.4.5, la multiplication matricielle suggérée est fondamentalement contrainte par une division/recombinaison en amplitude du signal, une division/recombinaison spectrale du signal ainsi qu'une interaction avec les éléments de matrice.

Un exemple d'architecture MVM conventionnelle de dimension $N = 4$, constituée de 16 unités DPE ainsi que d'un accumulateur par photodiodes est présentée à la figure 7.11. Cette architecture présente intrinsèquement un nombre de croisements substantiel qui croît rapidement en fonction de la taille de la matrice selon :

$$N_c(N) = \frac{1}{4}(n-1)^2 n^2, \tag{7.1}$$

soit 36 croisements pour $N = 4$ et 24 502 500 croisements pour $N = 100$.

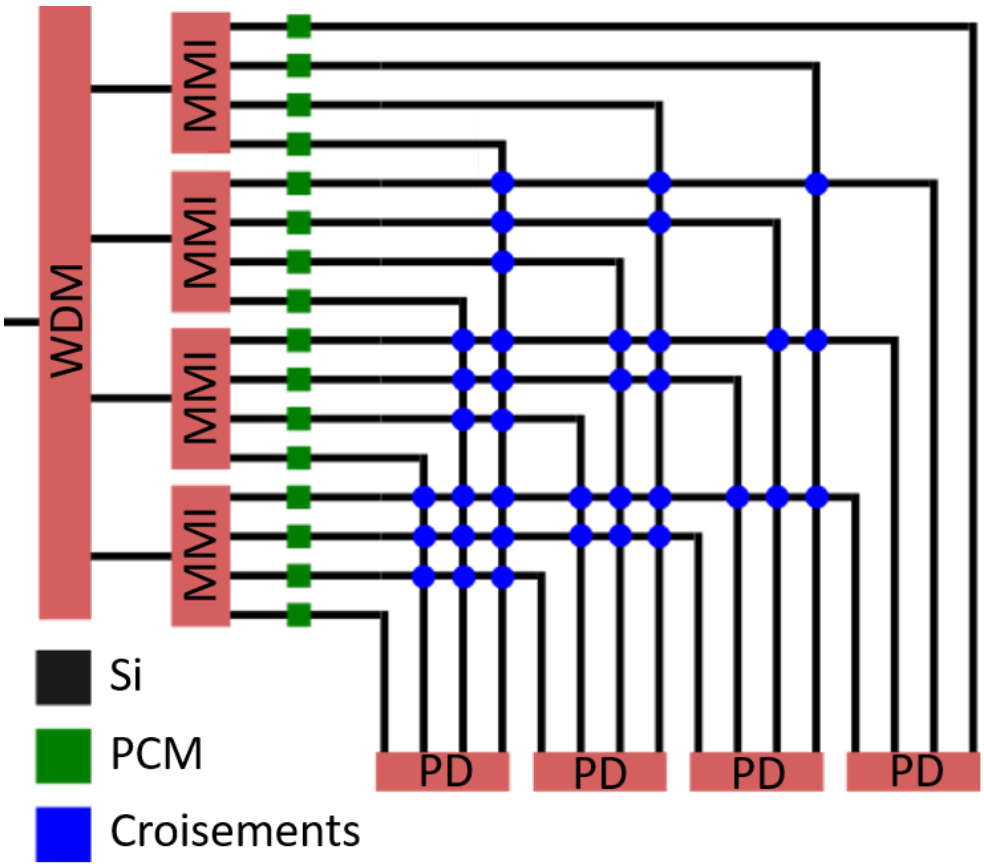



FIGURE 7.11 Architecture typique d'un MVM photonique avec PCM.

Ceci implique que la ligne la plus affectée écope de $(N-1)^2$ croisements. En revanche, en utilisant des croisements par pontage (e.g. Si-SiN) comme illustré à la figure 7.12, on peut remédier à ces croisements.

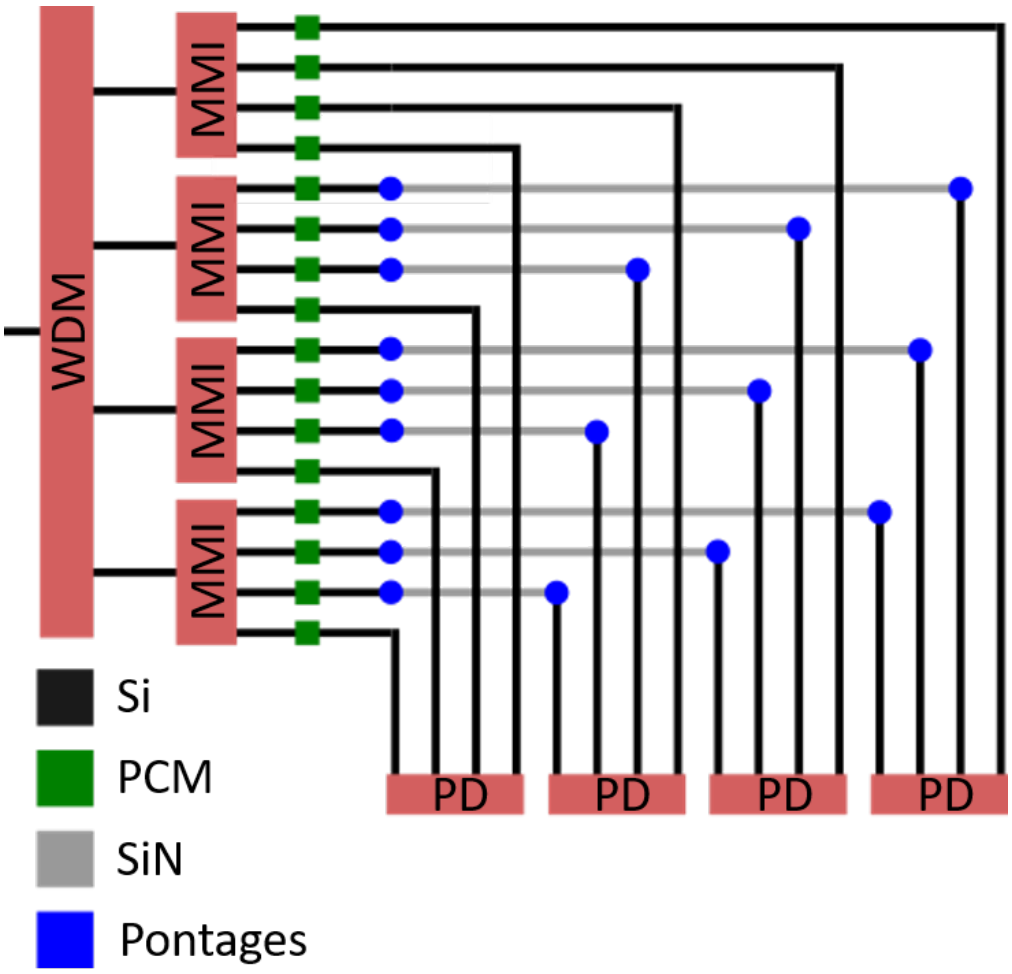



FIGURE 7.12 Architecture d'un MVM photonique avec PCM exploitant le pontage Si-SiN.

On remarque que le nombre de pontages ($N_p$) nécessaires est significativement inférieur à son analogue $N_c$. En effet, on a

$$N_p(N) = (N-1)^2 \tag{7.2}$$

soit 9 pontages pour $N = 4$ et 9801 pontages pour $N = 100$.

Une métrique importante pour la qualité du design est l'imbalance maximale, soit la différence entre la branche la plus affectée et celle la moins affectée par les croisements. En utilisant des valeurs conservatrices basées sur l'étude des croisements de la section 3.5, soit que les croisements en silicium engendrent des pertes de 0.1 dB et que les pontages en nitrure de silicium engendrent des pertes de 0.2 dB, on observe que le second design est beaucoup plus approprié pour des grandes matrices. À la figure 7.13 on compare les deux pertes totales

engendrées par un MVM selon sa taille pour les deux architectures, alors que la figure 7.14 compare l'imbalance maximale théorique, soit la différence maximale de signal entre deux lignes du MVM. Ces simulations n'incluent pas les pertes par propagation, celles-ci variant significativement selon la plateforme, le mode, et la configuration choisie.

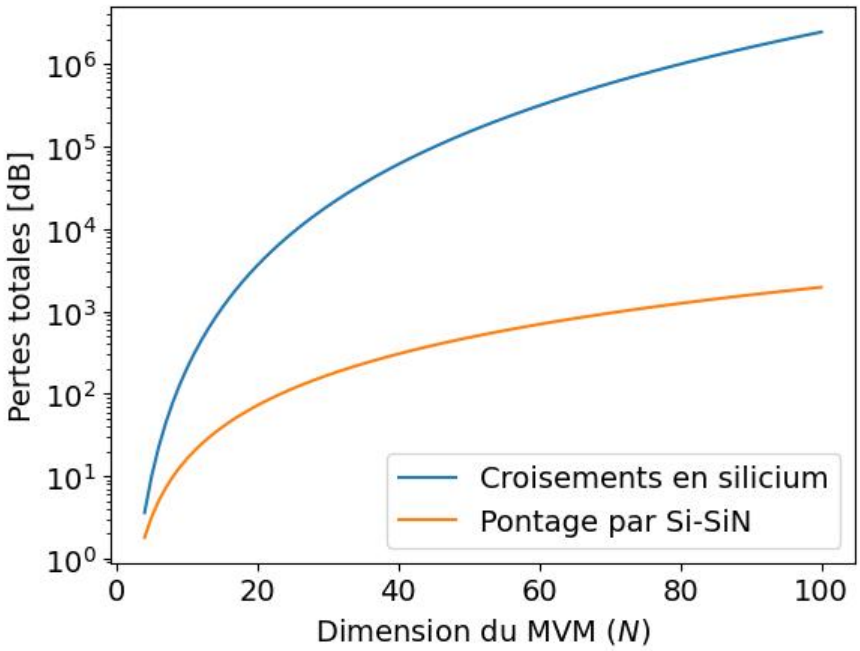



FIGURE 7.13 Comparaison des pertes totales des deux architectures étudiées.

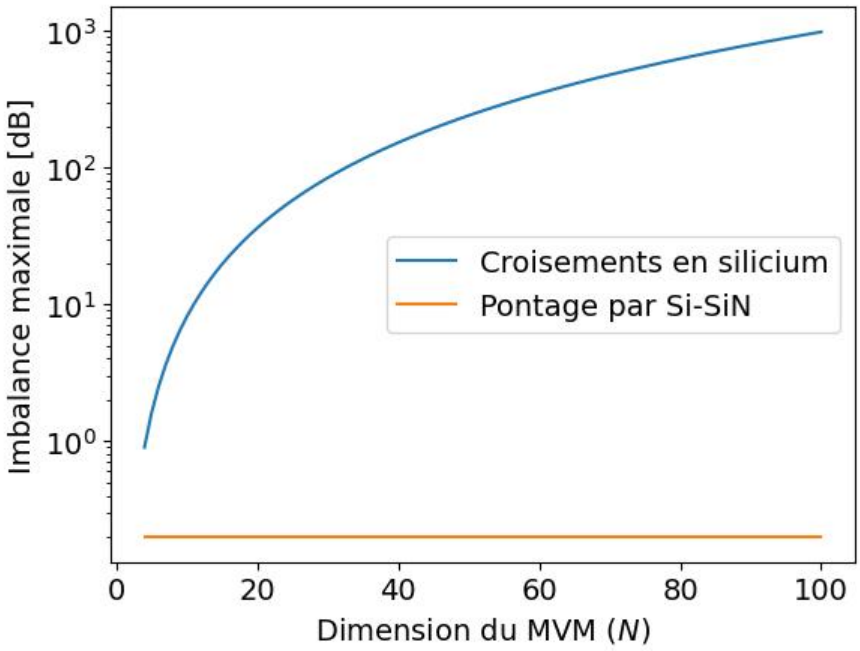



FIGURE 7.14 Imbalance maximale théorique entre les deux architectures étudiées.

Pour la division du signal, tout diviseur de type *Y-branch* ou interféromètre multimode (MMI) balancé dont la largeur de bande excède significativement l'encodage spectral utilisé est idéal. Ceci n'est toutefois pas le cas pour la recombinaison en amplitude, puisque les MMI 2x1 engendrent fondamentalement des pertes excédant 3 dB. L'accumulation électronique par l'entremise de photodétecteurs est privilégiée dans ce cas.

La division spectrale, telle que suggérée par [123], est effectuée par un duo de résonateurs en anneau. Ces résonateurs sont priorisés, en contraste au interféromètres de Mach-Zehnder, en raison de leur faible empreinte. Cependant, leur largeur de bande est très faible (10 pm), nécessitant un excellent contrôle en température pour aligner les longueurs d'ondes modulées à la résonance du résonateur en anneau qui se traduit par un apport énergétique supplémentaire. Pour éviter un tel ajustement, l'utilisation de coupleurs à réseau Bragg en cascade, nommés coupleurs contra-directionnels, possédant une largeur de bande de 10 nm permet de diviser le signal spectralement sans contrôle thermique. La réponse d'un tel coupleur possédant une largeur de bande 15 nm, fabriqué comme preuve de concept dans le cadre de ce projet est présenté à la figure 7.15.

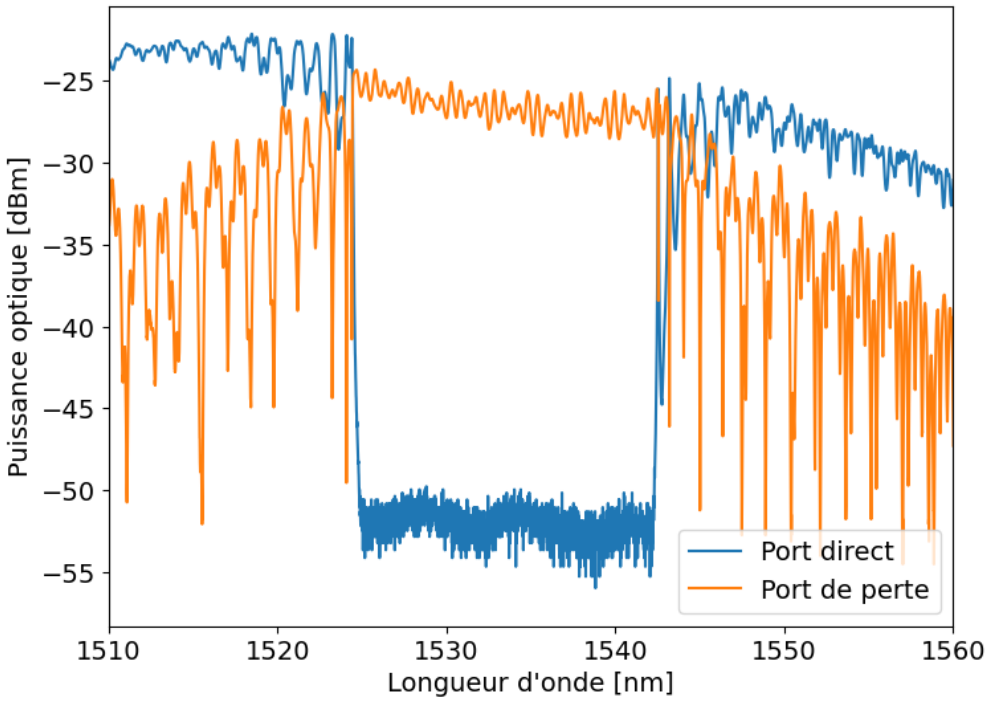



FIGURE 7.15 Réponse spectrale d'un coupleur contra-directionnel suggéré pour effectuer la WDM dans une architecture de MVM photonique.

Un exemple de 4 coupleurs contra-directionnels ainsi que sa signature spectrale pour une modulation CWDM est présentée à la figure 7.16.

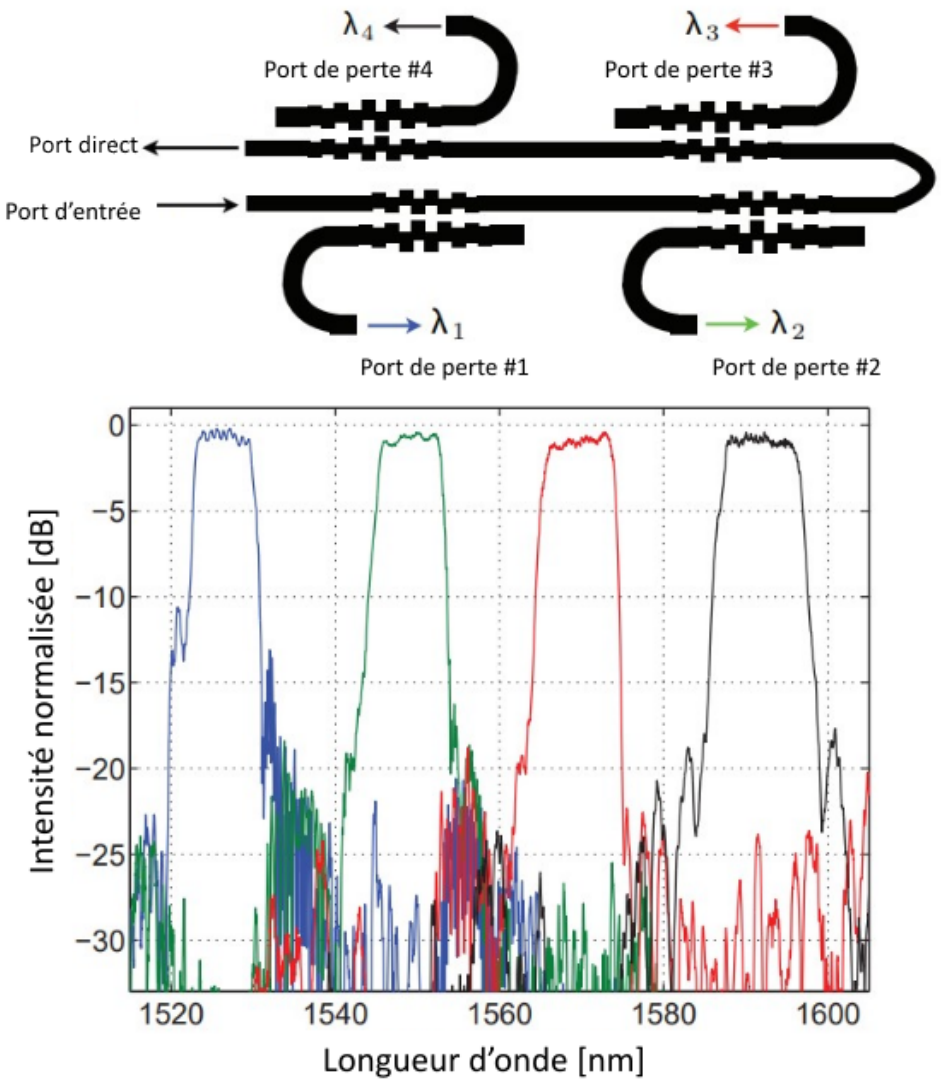



FIGURE 7.16 Réponse de quatres filtres spectraux à large bande en cascade.

Cette architecture est très robuste si la modulation du vecteur d'entrée $\mathbf{x}$ est telle que $x_0$, $x_1$, $x_2$ et $x_3$ sont encodés par des lasers à $\lambda = 1530, 1550, 1570$ et 1590 nm respectivement.

En somme, l'architecture et les améliorations proposées permettent de favoriser significativement la mise à l'échelle et augmenter la robustesse du multiplicateur matrice-vecteur.

# CHAPITRE 8 CONCLUSION

Pour conclure, rappelons que l'étude de processeurs optiques dédiés à l'exécution d'opération matricielles est poussée par une demande croissante de puissance de calcul due majoritairement à l'essor de l'intelligence artificielle. L'usage de la photonique intégrée sur silicium est proposée pour adresser ce problème en raison de sa maturité, sa polyvalence et sa taille. Il s'avère tout de même que le développement d'une mémoire optique non-volatile intégrée représente un défi substantiel dans le développement d'un accélérateur optique. Les travaux présentés dans ce mémoire visent à faire progresser le domaine du calcul optique, considéré par plusieurs comme une ajout non seulement viable, mais bénéfique au paradigme électronique actuel.

## 8.1 Rappel des travaux

Deux des quatres objectifs principaux de ce projet on été atteints. La caractérisation optique, stoechiométrique crisallographique du matériau émergent ($Ge_2Sb_2Se_5$) a été établie et comparée aux travaux précédents. L'indice de réfraction mesuré confirme le potentiel d'utilisation du chalcogénure comme mémoire optique à changement de phase. Les températures de fusion et de cristallisation ont également renforcé cette hypothèse.

L'intégration de cette mémoire sur une puce de photonique sur silicium a ensuite été effectuée. Ceci comprend le design de la puce, l'établissement du procédé de fabrication de la mémoire et l'étude de divers méthodes d'intégration. Le changement de phase a ensuite été partiellement confirmé par la cristallisation de la mémoire sur guide d'onde par recuit lent, générant une modulation de l'absorption mesurable.

Finalement, la proposition d'un design alternatif et plus propice aux grandes matrices a été présentée en raison de son imbalance maximale indépendante de la taille de la matrice à multiplier.

## 8.2 Améliorations futures

La principale embûche rencontrée au cours de ce projet est la fabrication d'éléments chauffants robustes sur topographie. Ceci s'est manifesté à maintes reprises lors des tentatives de caractérisation du changement de phase électrothermique par le bris précoce des éléments chauffants. Le développement d'un élément chauffant en polysilicium, par exemple, permettrait la fabrication monolithique de celui-ci par gravure et ainsi améliorer sa robustesse. Ceci

nécessiterait cependant plus de puissance électrique pour obtenir la température désirée et donc une efficacité thermique moindre. Aussi, la fragilité des éléments chauffants semble reliée à la topographie du guide d'onde, créant des points chauds faibles dans la couche métallique contribuant à leur défectuosité. Aussi, l'usage de guides d'onde planarisés serait à explorer pour adresser ce problème.

L'efficacité de l'élément chauffant, fortement influencée par sa largeur, serait également à investiguer. Les limitations en résolution du système photolithographique du LMF ne permettent pas la fabrication d'un élément de largeur $< 1$ µm. Cette largeur, lorsque réduite davantage, localise la dissipation de chaleur par effet Joule plus directement au matériau à changement de phase, nécessitant moins de puissance électrique pour atteindre les mêmes températures.

Une fois la cristallisation électrothermique atteinte, prouvant une génération de chaleur substantielle, l'étape d'amorphisation serait à vérifier et optimiser avant d'intégrer les mémoires dans une architecture de multiplication matrice-vecteur. Finalement, en utilisant un modulateur externe en entrée et un réseau de photodiodes en sortie, la multiplication matricielle pourrait être démontrée.

# RÉFÉRENCES


[1] K. Rupp *et al.*, "42 years of microprocessor trend data," *Microprocessor Trend Data Repository, https ://www. karlrupp. net/2018/02/42-years-of-microprocessor-trend-data*, 2018.

[2] C. C. Mann, "The end of moore's law ?" *Technology Review*, vol. 103, n°. 3, p. 42–42, 2000.

[3] K. Palem et A. Lingamneni, "What to do about the end of moore's law, probably !" dans *Proceedings of the 49th Annual Design Automation Conference*, 2012, p. 924–929.

[4] T. N. Theis et H.-S. P. Wong, "The end of moore's law : A new beginning for information technology," *Computing in Science & Engineering*, vol. 19, n°. 2, p. 41–50, 2017.

[5] L. Fridman. Jim keller : Moore's law, microprocessors, and first principles | lex fridman podcast 70. Youtube. [En ligne]. Disponible : https://www.youtube.com/watch?v=Nb2tebYAaOA

[6] T. Hiramoto, "Five nanometre cmos technology," *Nature Electronics*, vol. 2, n°. 12, p. 557–558, 2019.

[7] R. Metz, "Openai raises $175 million for startup investment fund," *Bloomberg*, 2023. [En ligne]. Disponible : https://www.bloomberg.com/news/articles/2023-05-25/openai-raises-175-million-for-startup-investment-fund

[8] D. Marković, A. Mizrahi, D. Querlioz et J. Grollier, "Physics for neuromorphic computing," *Nature Reviews Physics*, vol. 2, n°. 9, p. 499–510, 2020.

[9] V. K. Sangwan et M. C. Hersam, "Neuromorphic nanoelectronic materials," *Nature nanotechnology*, vol. 15, n°. 7, p. 517–528, 2020.

[10] A. Zakery et S. Elliott, "Optical properties and applications of chalcogenide glasses : a review," *Journal of Non-Crystalline Solids*, vol. 330, n°. 1, p. 1–12, 2003. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0022309303006379

[11] A. Zakery et S. R. Elliott, *Optical nonlinearities in chalcogenide glasses and their applications*. Springer, 2007, vol. 135.

[12] Y. S. Obeng, N. V. Nguyen, P. K. Amoah, J. Ahn, M. Y. Shalaginov, J. Hu et K. A. Richardson, "Dielectric spectroscopic investigation of reversible photo-induced changes in amorphous ge2sb2se5 thin films," *Journal of Applied Physics*, vol. 131, n°. 7, 2022.

[13] E. Mammadov, Y.-G. Shim, J. Sakamoto, K. Wakita, N. Mamedov et H. Uchiki, "Comparison of optical parameters of ge–as (sb)–se (te) glassy films," *Japanese Journal of Applied Physics*, vol. 50, n°. 5S2, p. 05FC12, 2011.

[14] Y. S. Obeng, N. V. Nguyen, P. K. Amoah *et al.*, "Dielectric spectroscopic investigation of reversible photo-induced changes in amorphous ge."

[15] J. Bartak, P. Kostal, V. Podzemna, J. Shanelova et J. Malek, "Crystal growth kinetics and viscous behavior in ge2sb2se5 undercooled melt," *The Journal of Physical Chemistry B*, vol. 120, n°. 32, p. 7998–8006, 2016.

[16] J. R. Rodriguez, Z. Qi, H. Wang, M. Y. Shalaginov, C. Goncalves, M. Kang, K. A. Richardson, J. Guerrero-Sanchez, M. G. Moreno-Armenta et V. G. Pol, "Ge2sb2se5 glass as high-capacity promising lithium-ion battery anode," *Nano Energy*, vol. 68, p. 104326, 2020.

[17] T. Matsunaga, R. Kojima, N. Yamada, K. Kifune, Y. Kubota, S. Kohara et M. Takata, "Structural analysis of gesbtese phase-change materials," *Libr E\PCOS*, vol. 2009, p. 99–104, 2009.

[18] M. J. Heck, J. F. Bauters, M. L. Davenport, D. T. Spencer et J. E. Bowers, "Ultra-low loss waveguide platform and its integration with silicon photonics," *Laser & Photonics Reviews*, vol. 8, n°. 5, p. 667–686, 2014.

[19] T. Sommer, N. Mange, P. Wegmann et M. Poot, "Efficient adiabatic-coupler-based silicon nitride waveguide crossings for photonic quantum computing," *Optics Letters*, vol. 48, n°. 11, p. 2981–2984, 2023.

[20] A. N. inc., "Nanosoi fabrication service," https ://www.appliednt.com/nanosoi-fabrication-service/, 2022, repéré le 21 juin 2023. [En ligne]. Disponible : https://www.appliednt.com/nanosoi-fabrication-service/

[21] S. A. Maier *et al.*, *Plasmonics : fundamentals and applications*. Springer, 2007, vol. 1.

[22] M. Çelikbilek, A. E. Ersundu et S. Aydın, "Crystallization kinetics of amorphous materials," *Advances in crystallization processes*, vol. 35347, 2012.

[23] T. M. Nieuwenhuizen, "Ehrenfest relations at the glass transition : solution to an old paradox," *Physical review letters*, vol. 79, n°. 7, p. 1317, 1997.

[24] F. H. Stillinger, "Supercooled liquids, glass transitions, and the kauzmann paradox," *The Journal of chemical physics*, vol. 88, n°. 12, p. 7818–7825, 1988.

[25] O. Yamamuro, I. Tsukushi, A. Lindqvist, S. Takahara, M. Ishikawa et T. Matsuo, "Calorimetric study of glassy and liquid toluene and ethylbenzene : Thermodynamic approach to spatial heterogeneity in glass-forming molecular liquids," *The Journal of Physical Chemistry B*, vol. 102, n°. 9, p. 1605–1609, 1998.

[26] W. Kauzmann, “The nature of the glassy state and the behavior of liquids at low temperatures.” *Chemical reviews*, vol. 43, n°. 2, p. 219–256, 1948.

[27] J. Pries, S. Wei, M. Wuttig et P. Lucas, “Switching between crystallization from the glassy and the undercooled liquid phase in phase change material ge2sb2te5,” *Advanced Materials*, vol. 31, n°. 39, p. 1900784, 2019. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.201900784

[28] J. Schmelzer, G. Röpke et V. B. Priezzhev, *Nucleation theory and applications*. Wiley-VCH Weinheim, 2005, vol. 76.

[29] A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatain, A. Novikov, F. J. R Ruiz, J. Schrittwieser, G. Swirszcz *et al.*, “Discovering faster matrix multiplication algorithms with reinforcement learning,” *Nature*, vol. 610, n°. 7930, p. 47–53, 2022.

[30] B. Brubaker, “Ai reveals new possibilities in matrix multiplication,” *Quanta Magazine*, 2022. [En ligne]. Disponible : https://www.quantamagazine.org/ai-reveals-new-possibilities-in-matrix-multiplication-20221123/

[31] R. Bulat, “Ai essentials : Working with matrices,” *Medium*, 2020. [En ligne]. Disponible : https://rossbulat.medium.com/ai-essentials-working-with-matrices-2ceb9ca3bd1b

[32] J. W. Kennedy, L. V. Quintas et M. M. Sysło, “The theorem on planar graphs,” *Historia Mathematica*, vol. 12, n°. 4, p. 356–368, 1985. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/031508608590045X

[33] F. B. Insights. Artificial intelligence market size, share  covid-19 impact analysis, by component (hardware, software/platform, and services), by function (human resources, marketing  sales, product/service deployment, service operation, risk, supply-chain management, and others (strategy and corporate finance)), by deployment (cloud and on-premises), by industry (healthcare, retail, it  telecom, bfsi, automotive, advertising  media, manudacturing, and others), and regional forecast, 2023-2030. [En ligne]. Disponible : https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114

[34] K. Hu, “Chatgpt sets record for fastest-growing user base - analyst note,” *Reuters*, Février 2023. [En ligne]. Disponible : https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[35] D. Ruby, “71+ instagram statistics for marketers in 2023 (data trends),” *DemandSage*, Mars 2023. [En ligne]. Disponible : https://www.demandsage.com/instagram-statistics/#:~:text=Instagram%20Monthly%20Active%20Users%20(MAUs,by%20the%20end%20of%202023.

[36] F. Ali, “Gpt-1 to gpt-4 : Each of openai’s gpt models explained and compared,” *Make Use Of*, April 2023. [En ligne]. Disponible : https://www.makeuseof.com/gpt-models-explained-and-compared/

[37] J. Stern, “Gpt-4 might just be a bloated, pointless mess,” *The Atlantic*, March 2023. [En ligne]. Disponible : https://www.theatlantic.com/technology/archive/2023/03/openai-gpt-4-parameters-power-debate/673290/

[38] J. C. Lin, D. N. Younessi, S. S. Kurapati, O. Y. Tang et I. U. Scott, “Comparison of gpt-3.5, gpt-4, and human user performance on a practice ophthalmology written examination,” *Eye*, p. 1–2, 2023.

[39] D. M. Katz, M. J. Bommarito, S. Gao et P. Arredondo, “Gpt-4 passes the bar exam,” *Available at SSRN 4389233*, 2023.

[40] J. Sevilla, P. Villalobos, J. F. Cerón, M. Burtell, L. Heim, A. B. Nanjajjar, A. Ho, T. Besiroglu, M. Hobbhahn et J.-S. Denain, “Parameter, compute and data trends in machine learning,” 2021.

[41] H. Fujita, “Ai-based computer-aided diagnosis (ai-cad) : the latest review to read first,” *Radiological physics and technology*, vol. 13, n°. 1, p. 6–19, 2020.

[42] S. Huang, J. Yang, S. Fong et Q. Zhao, “Artificial intelligence in the diagnosis of covid-19 : challenges and perspectives,” *International Journal of Biological Sciences*, vol. 17, n°. 6, p. 1581, 2021.

[43] B. Shabestri, M. A. Anastasio, B. Fei et F. Leblond, “Special series guest editorial : Artificial intelligence and machine learning in biomedical optics,” *Journal of biomedical optics*, vol. 26, n°. 5, 2021.

[44] T. Okuda et S. Shoda, “Ai-based chatbot service for financial industry,” *Fujitsu Scientific and Technical Journal*, vol. 54, n°. 2, p. 4–8, 2018.

[45] M. Adam, M. Wessel et A. Benlian, “Ai-based chatbots in customer service and their effects on user compliance,” *Electronic Markets*, vol. 31, n°. 2, p. 427–445, 2021.

[46] C. P. Lee, K. M. Lim, T. J. Yu et S. F. Abdul Razak, “Ai-based targeted advertising system,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, n°. 2, p. 787–793, 2019.

[47] C. Campbell, S. Sands, C. Ferraro, H.-Y. J. Tsao et A. Mavrommatis, “From data to action : How marketers can leverage ai,” *Business Horizons*, vol. 63, n°. 2, p. 227–243, 2020, artificial intelligence and machine learning. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0007681319301624

[48] L. Cao, “Ai in finance : challenges, techniques, and opportunities,” *ACM Computing Surveys (CSUR)*, vol. 55, n°. 3, p. 1–38, 2022.

[49] A. Zheng et J. Jin, “Using ai to make predictions on stock market,” *cs229. stanford. edu*, 2017.

[50] M. Fernandez, “Nodes, sockets, cores and flops, oh, my!” Novembre 2011. [En ligne]. Disponible : https://archive.wikiwix.com/cache/index2.php?url=http%3A%2F%2Fen.community.dell.com%2Ftechcenter%2Fhigh-performance-computing%2Fw%2Fwiki%2F2329.aspx#federation=archive.wikiwix.com&tab=url

[51] B. Zheng et G. Pekhimenko, “Ecornn : Efficient computing of lstm rnn on gpus,” *Memory*, vol. 9, p. 1735–1780, 1997.

[52] D. Steinkraus, I. Buck et P. Simard, “Using gpus for machine learning algorithms,” dans *Eighth International Conference on Document Analysis and Recognition (ICDAR’05)*, 2005, p. 1115–1120 Vol. 2.

[53] “Intel® core™ i9-13900ks processor,” 2023. [En ligne]. Disponible : https://ark.intel.com/content/www/us/en/ark/products/232167/intel-core-i913900ks-processor-36m-cache-up-to-6-00-ghz.html

[54] run.ai, “Best gpu for deep learning,” repéré le 5 juin 2023. [En ligne]. Disponible : https://www.run.ai/guides/gpu-deep-learning/best-gpu-for-deep-learning

[55] C. Collange, D. Defour et A. Tisserand, “Power consumption of gpus from a software perspective,” dans *Computational Science – ICCS 2009*, G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra et P. M. A. Sloot, édit. Berlin, Heidelberg : Springer Berlin Heidelberg, 2009, p. 914–923.

[56] D. Lustig et M. Martonosi, “Reducing gpu offload latency via fine-grained cpu-gpu synchronization,” dans *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2013, p. 354–365.

[57] “System architecture,” relevé le 4 juin 2023. [En ligne]. Disponible : https://cloud.google.com/tpu/docs/system-architecture-tpu-vm

[58] H. Yoshida, “Solving the vonneumann bottleneck with fpgas,” Janvier 2019, relevé le 4 juin 2023. [En ligne]. Disponible : https://community.hitachivantara.com/blogs/hubert-yoshida/2019/02/01/solving-the-von-neumann-bottleneck-with-fpgas

[59] S. Branco, A. G. Ferreira et J. Cabral, “Machine learning in resource-scarce embedded systems, fpgas, and end-devices : A survey,” *Electronics*, vol. 8, n°. 11, p. 1289, 2019.

[60] K. Kara, D. Alistarh, G. Alonso, O. Mutlu et C. Zhang, “Fpga-accelerated dense linear machine learning : A precision-convergence trade-off,” dans *2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2017, p. 160–167.

[61] R. Zhao, H.-C. Ng, W. Luk et X. Niu, “Towards efficient convolutional neural network for domain-specific applications on fpga,” dans *2018 28th International Conference on Field Programmable Logic and Applications (FPL)*, 2018, p. 147–1477.

[62] W. Dai et D. Berleant, “Benchmarking contemporary deep learning hardware and frameworks : A survey of qualitative metrics,” dans *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, 2019, p. 148–155.

[63] P. P. Shinde et S. Shah, “A review of machine learning and deep learning applications,” dans *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, p. 1–6.

[64] T. Sipola, J. Alatalo, T. Kokkonen et M. Rantonen, “Artificial intelligence in the iot era : A review of edge ai hardware and software,” dans *2022 31st Conference of Open Innovations Association (FRUCT)*, 2022, p. 320–331.

[65] “Numerical accuracy,” 2023, repéré le 5 juin 2023. [En ligne]. Disponible : https://pytorch.org/docs/stable/notes/numerical_accuracy.html?highlight=precision

[66] “tf.keras.backend.set_floatx,” 2023, repéré le 5 juin 2023. [En ligne]. Disponible : https://www.tensorflow.org/api_docs/python/tf/keras/backend/set_floatx

[67] M. Courbariaux, Y. Bengio et J.-P. David, “Training deep neural networks with low precision multiplications,” *arXiv preprint arXiv :1412.7024*, 2014.

[68] C. De Sa, M. Leszczynski, J. Zhang, A. Marzoev, C. R. Aberger, K. Olukotun et C. Ré, “Halp : High-accuracy low-precision training,” *Stanford Dawn*, 2018, repéré le 5 juin 2023. [En ligne]. Disponible : https://dawn.cs.stanford.edu/2018/03/09/low-precision/

[69] “Breast cancer wisconsin (diagnostic) data set,” 2023, repéré le 6 juin 2023. [En ligne]. Disponible : https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[70] “Mushroom classification,” 2023, repéré le 6 juin 2023. [En ligne]. Disponible : https://www.kaggle.com/datasets/uciml/mushroom-classification?select=mushrooms.csv

[71] “Iris flower dataser,” 2023, repéré le 6 juin 2023. [En ligne]. Disponible : https://www.kaggle.com/datasets/arshid/iris-flower-dataset

[72] “Digit recognizer,” 2023, repéré le 6 juin 2023. [En ligne]. Disponible : https://www.kaggle.com/competitions/digit-recognizer

[73] “Fashion mnist,” 2023, repéré le 6 juin 2023. [En ligne]. Disponible : https://www.kaggle.com/datasets/zalando-research/fashionmnist

[74] “Cifar-10 - object recognition in images,” 2023, repéré le 6 juin 2023. [En ligne]. Disponible : https://www.kaggle.com/c/cifar-10/

[75] X. Sun, N. Wang, C.-Y. Chen, J. Ni, A. Agrawal, X. Cui, S. Venkataramani, K. El Maghraoui, V. V. Srinivasan et K. Gopalakrishnan, "Ultra-low precision 4-bit training of deep neural networks," dans *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan et H. Lin, édit., vol. 33. Curran Associates, Inc., 2020, p. 1796–1807. [En ligne]. Disponible : https://proceedings.neurips.cc/paper_files/paper/2020/file/13b919438259814cd5be8cb45877d577-Paper.pdf

[76] G. Venkatesh, E. Nurvitadhi et D. Marr, "Accelerating deep convolutional networks using low-precision and sparsity," dans *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, p. 2861–2865.

[77] D. H. Bailey, "High-precision computation and mathematical physics," 2009.

[78] J. Pries, J. C. Sehringer, S. Wei, P. Lucas et M. Wuttig, "Glass transition of the phase change material aist and its impact on crystallization," *Materials Science in Semiconductor Processing*, vol. 134, p. 105990, 2021. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1369800121003371

[79] J. Kalb, F. Spaepen et M. Wuttig, "Calorimetric measurements of phase transformations in thin films of amorphous te alloys used for optical data storage," *Journal of applied physics*, vol. 93, n°. 5, p. 2389–2393, 2003.

[80] G. du Canada, "Longevity of recordable cds, dvds and blu-rays — canadian conservation institute (cci) notes 19/1," 2020, repéré le 19 juin 2023. [En ligne]. Disponible : https://www.canada.ca/en/conservation-institute/services/conservation-preservation-publications/canadian-conservation-institute-notes/longevity-recordable-cds-dvds.html

[81] F. Zhai, F. Zuo, H. Huang, Y. Wang, T. Lai, Y. Wu et F. Gan, "Optical–electrical properties of aginsbte phase change thin films under single picosecond laser pulse irradiation," *Journal of Non-Crystalline Solids*, vol. 356, n°. 18, p. 889–892, 2010. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0022309309007923

[82] M. Kim, H. Seo, J. Park, W. Choi et T. Yoo, "Cd-rewritable optical disks with enhanced media cyclability," dans *1997 Optical Data Storage Topical Meeting ODS Conference Digest*, 1997, p. 100–101.

[83] Y. Tao, X. Li, H. Xu, Z. Wang, W. Ding, W. Liu, J. Ma et Y. Liu, "Improved uniformity and endurance through suppression of filament overgrowth in electrochemical metallization memory with aginsbte buffer layer," *IEEE Journal of the Electron Devices Society*, vol. 6, p. 714–720, 2018.

[84] G. S. Choi, B.-W. On, K. Choi et S. Yi, "Ptl : Pram translation layer," *Microprocessors and Microsystems*, vol. 37, n°. 1, p. 24–32, 2013. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0141933112001172

[85] J. Guo, J. Yang, Y. Zhang et Y. Chen, "Low cost power failure protection for mlc nand flash storage systems with pram/dram hybrid buffer," dans *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2013, p. 859–864.

[86] I.-D. Simandan, F. Sava, A.-T. Buruiana, A.-C. Galca, N. Becherescu, I. Burducea, C. Mihai et A. Velea, "Influence of deposition method on the structural and optical properties of ge2sb2te5," *Materials*, vol. 14, n°. 13, 2021. [En ligne]. Disponible : https://www.mdpi.com/1996-1944/14/13/3663

[87] M. Wuttig, H. Bhaskaran et T. Taubner, "Phase-change materials for non-volatile photonic applications," *Nature photonics*, vol. 11, n°. 8, p. 465–476, 2017.

[88] W. H. Pernice et H. Bhaskaran, "Photonic non-volatile memories using phase change materials," *Applied Physics Letters*, vol. 101, n°. 17, p. 171101, 2012.

[89] T. Kato et K. Tanaka, "Electronic properties of amorphous and crystalline ge2sb2te5 films," *Japanese journal of applied physics*, vol. 44, n°. 10R, p. 7340, 2005.

[90] Y. Zhang, J. B. Chou, J. Li, H. Li, Q. Du, A. Yadav, S. Zhou, M. Y. Shalaginov, Z. Fang, H. Zhong *et al.*, "Broadband transparent optical phase change materials for high-performance nonvolatile photonics," *Nature communications*, vol. 10, n°. 1, p. 4279, 2019.

[91] R. Svoboda et J. Málek, "Particle size influence on crystallization behavior of ge2sb2se5 glass," *Journal of non-crystalline solids*, vol. 358, n°. 2, p. 276–284, 2012.

[92] S. Franssila, *Introduction to microfabrication*. John Wiley & Sons, 2010.

[93] Q. Wang, E. T. Rogers, B. Gholipour, C.-M. Wang, G. Yuan, J. Teng et N. I. Zheludev, "Optically reconfigurable metasurfaces and photonic devices based on phase change materials," *Nature photonics*, vol. 10, n°. 1, p. 60–65, 2016.

[94] A. Pohm, C. Sie, R. Uttecht, V. Kao et O. Agrawal, "Chalcogenide glass bistable resistivity (ovonic) memories," *IEEE Transactions on Magnetics*, vol. 6, n°. 3, p. 592–592, 1970.

[95] A. Masood, M. Pantouvaki, G. Lepage, P. Verheyen, J. Van Campenhout, P. Absil, D. Van Thourhout et W. Bogaerts, "Comparison of heater architectures for thermal control of silicon photonic circuits," dans *10th International Conference on Group IV Photonics*, 2013, p. 83–84.

[96] M. Kimura et K. Komatsuzaki, "Microheater made of heavily boron doped single crystal silicon beam," *MRS Online Proceedings Library (OPL)*, vol. 276, p. 271, 1992.

[97] H. Taghinejad, S. Abdollahramezani, A. A. Eftekhar, T. Fan, A. H. Hosseinnia, O. Hemmatyar, A. E. Dorche, A. Gallmon et A. Adibi, "Ito-based microheaters for reversible multi-stage switching of phase-change materials : towards miniaturized beyond-binary reconfigurable integrated photonics," *Opt. Express*, vol. 29, n°. 13, p. 20 449–20 462, Jun 2021. [En ligne]. Disponible : https://opg.optica.org/oe/abstract.cfm?URI=oe-29-13-20449

[98] C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose et J. S. Plank, "A survey of neuromorphic computing and neural networks in hardware," 2017.

[99] P. C. Eng, S. Song et B. Ping, "State-of-the-art photodetectors for optoelectronic integration at telecommunication wavelength," *Nanophotonics*, vol. 4, n°. 3, p. 277–302, 2015. [En ligne]. Disponible : https://doi.org/10.1515/nanoph-2015-0012

[100] G. Kalogerakis, T. Moran, T. Nguyen et G. Denoyer, "A quad 25gb/s 270mw tia in 0.13 $\mu$m bicmos with< 0.15 db crosstalk penalty," dans *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*. IEEE, 2013, p. 116–117.

[101] C. Caillaud, P. Chanclou, F. Blache, P. Angelini, B. Duval, P. Charbonnier, D. Lanteri et M. Achouche, "High sensitivity 40 gbit/s preamplified soa-pin/tia receiver module for high speed pon," dans *2014 The European Conference on Optical Communication (ECOC)*. IEEE, 2014, p. 1–3.

[102] H.-Y. Kao, C.-T. Tsai, S.-F. Leong, C.-Y. Peng, Y.-C. Chi, H.-Y. Wang, H.-C. Kuo, C.-H. Wu, W.-H. Cheng et G.-R. Lin, "Single-mode vcsel for pre-emphasis pam-4 transmission up to 64  gbit/s over 100–300  m in om4 mmf," *Photon. Res.*, vol. 6, n°. 7, p. 666–673, Jul 2018. [En ligne]. Disponible : https://opg.optica.org/prj/abstract.cfm?URI=prj-6-7-666

[103] J. M. Tang, P. M. Lane et K. A. Shore, "High-speed transmission of adaptively modulated optical ofdm signals over multimode fibers using directly modulated dfbs," *J. Lightwave Technol.*, vol. 24, n°. 1, p. 429, Jan 2006. [En ligne]. Disponible : https://opg.optica.org/jlt/abstract.cfm?URI=jlt-24-1-429

[104] D. Patel, A. Samani, V. Veerasubramanian, S. Ghosh et D. V. Plant, "Silicon photonic segmented modulator-based electro-optic dac for 100 gb/s pam-4 generation," *IEEE Photonics Technology Letters*, vol. 27, n°. 23, p. 2433–2436, 2015.

[105] S. Moazeni, S. Lin, M. Wade, L. Alloatti, R. J. Ram, M. Popović et V. Stojanović, "A 40-gb/s pam-4 transmitter based on a ring-resonator optical dac in 45-nm soi cmos," *IEEE Journal of Solid-State Circuits*, vol. 52, n°. 12, p. 3503–3516, 2017.

[106] T. Singh, "Hybrid memristor-cmos (memos) based logic gates and adder circuits," *arXiv [cs.ET]*, vol. 1506, p. 1–11, 06 2015.

[107] L. Chua, "Memristor-the missing circuit element," *IEEE Transactions on Circuit Theory*, vol. 18, n°. 5, p. 507–519, 1971.

[108] D. B. Strukov, G. S. Snider, D. R. Stewart et R. S. Williams, "The missing memristor found," *nature*, vol. 453, n°. 7191, p. 80–83, 2008.

[109] G. S. Snider, "Self-organized computation with unreliable, memristive nanodevices," *Nanotechnology*, vol. 18, n°. 36, p. 365202, aug 2007. [En ligne]. Disponible : https://dx.doi.org/10.1088/0957-4484/18/36/365202

[110] S. H. Jo, K.-H. Kim et W. Lu, "High-density crossbar arrays based on a si memristive system," *Nano letters*, vol. 9, n°. 2, p. 870–874, 2009.

[111] Y. Li, Z. Wang, R. Midya, Q. Xia et J. J. Yang, "Review of memristor devices in neuromorphic computing : materials sciences and device challenges," *Journal of Physics D : Applied Physics*, vol. 51, n°. 50, p. 503002, 2018.

[112] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev et D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, n°. 7550, p. 61–64, 2015.

[113] A. Chen, "Forming voltage scaling of resistive switching memories," dans *71st Device Research Conference*. IEEE, 2013, p. 181–182.

[114] H. Zhou, J. Dong, J. Cheng, W. Dong, C. Huang, Y. Shen, Q. Zhang, M. Gu, C. Qian, H. Chen *et al.*, "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light : Science & Applications*, vol. 11, n°. 1, p. 30, 2022.

[115] S. De, R. Das, R. K. Varshney et T. Schneider, "Design and simulation of thermo-optic phase shifters with low thermal crosstalk for dense photonic integration," *IEEE Access*, vol. 8, p. 141 632–141 640, 2020.

[116] Z. Zhang, H. Yu, Q. Huang, Z. Zhou, B. Chen, T. Dai, H. Qiu, Y. Wang et J. Yang, "High-speed and low-power silicon optical phased array based on the carrier-depletion mechanism," *IEEE Photonics Technology Letters*, vol. 34, n°. 5, p. 271–274, 2022.

[117] S. Sun, M. He, M. Xu, S. Gao, S. Yu et X. Cai, "Hybrid silicon and lithium niobate modulator," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 27, n°. 3, p. 1–12, 2021.

[118] S. Han, T. J. Seok, K. Yu, N. Quack, R. S. Muller et M. C. Wu, "Large-scale polarization-insensitive silicon photonic mems switches," *Journal of Lightwave Technology*, vol. 36, n°. 10, p. 1824–1830, 2018.

[119] S. Han, J. Beguelin, L. Ochikubo, J. Jacobs, T. J. Seok, K. Yu, N. Quack, C.-K. Kim, R. S. Muller et M. C. Wu, "32 × 32 silicon photonic MEMS switch with gap-adjustable directional couplers fabricated in commercial CMOS foundry," *Journal of Optical Microsystems*, vol. 1, n°. 2, p. 024003, 2021. [En ligne]. Disponible : https://doi.org/10.1117/1.JOM.1.2.024003

[120] T. J. Seok, K. Kwon, J. Henriksson, J. Luo et M. C. Wu, "240×240 wafer-scale silicon photonic switches," dans *2019 Optical Fiber Communications Conference and Exhibition (OFC)*, 2019, p. 1–3.

[121] C. Ramey, "Silicon photonics for artificial intelligence acceleration : Hotchips 32," dans *2020 IEEE hot chips 32 symposium (HCS)*. IEEE Computer Society, 2020, p. 1–26.

[122] H. Sun, Q. Qiao, Q. Guan et G. Zhou, "Silicon photonic phase shifters and their applications : A review," *Micromachines*, vol. 13, n°. 9, 2022. [En ligne]. Disponible : https://www.mdpi.com/2072-666X/13/9/1509

[123] M. Miscuglio et V. J. Sorger, "Photonic tensor cores for machine learning," *Applied Physics Reviews*, vol. 7, n°. 3, p. 031404, 2020.

[124] P. Sanchis, J. Galan, A. Griol, J. Marti, M. Piqueras et J. Perdigues, "Low-crosstalk in silicon-on-insulator waveguide crossings with optimized-angle," *IEEE Photonics Technology Letters*, vol. 19, n°. 20, p. 1583–1585, 2007.

[125] C. Priestland et S. Hersee, "The effects of pressure on the deposition rate in rf sputtering processes," *Vacuum*, vol. 22, n°. 3, p. 103–106, 1972. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/0042207X7290468X

[126] E. Baudet, M. Sergent, P. Němec, C. Cardinaud, E. Rinnert, K. Michel, L. Jouany, B. Bureau et V. Nazabal, "Experimental design approach for deposition optimization of rf sputtered chalcogenide thin films devoted to environmental optical sensors," *Scientific Reports*, vol. 7, n°. 1, p. 3500, 2017.

[127] M. Abd-Elrahman, R. M. Khafagy, S. A. Zaki et M. Hafiz, "Development and study of chalcogenide se60te30s10 thin film coated by carbon film : Structural and optical properties," *Materials Science in Semiconductor Processing*, vol. 27, p. 827–832, 2014. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1369800114004685

[128] G. Jellison, "Data analysis for spectroscopic ellipsometry," *Thin Solid Films*, vol. 234, n°. 1, p. 416–422, 1993. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/0040609093902984

[129] A. Rothen, “The Ellipsometer, an Apparatus to Measure Thicknesses of Thin Surface Films,” *Review of Scientific Instruments*, vol. 16, n°. 2, p. 26–30, 12 2004. [En ligne]. Disponible : https://doi.org/10.1063/1.1770315

[130] P. Drude, “Ueber die gesetze der reflexion und brechung des lichtes an der grenze absorbirender krystalle,” *Annalen der Physik*, vol. 268, n°. 12, p. 584–625, 1887. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.18872681205

[131] E. Stoumbou, I. Stavrakas, G. Hloupis, A. Alexandridis, D. Triantis et K. Moutzouris, “A comparative study on the use of the extended-cauchy dispersion equation for fitting refractive index data in crystals,” *Optical and Quantum Electronics*, vol. 45, 08 2013.

[132] D. Franta, I. Ohlıdal, M. Frumar et J. Jedelský, “Optical characterization of chalcogenide thin films,” *Applied Surface Science*, vol. 175-176, p. 555–561, 2001, 10th International Conference on Solid Films and Surfaces. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0169433201001489

[133] I. Chang, S. Woo, M. H. Lee, J. H. Shim, Y. Piao et S. W. Cha, “Characterization of porous pt films deposited via sputtering,” *Applied surface science*, vol. 282, p. 463–466, 2013.

[134] L.-J. Meng et M. P. dos Santos, “Investigations of titanium oxide films deposited by dc reactive magnetron sputtering in different sputtering pressures,” *Thin Solid Films*, vol. 226, n°. 1, p. 22–29, 1993.

[135] A. van der Rest, H. Idrissi, F. Henry, A. Favache, D. Schryvers, J. Proost, J.-P. Raskin, Q. Van Overmeere et T. Pardoen, “Mechanical behavior of ultrathin sputter deposited porous amorphous al2o3 films,” *Acta Materialia*, vol. 125, p. 27–37, 2017.

[136] A. C. Hardy, “A recording photoelectric color analyser,” *JOSA*, vol. 18, n°. 2, p. 96–117, 1929.

[137] T. A. Germer, J. C. Zwinkels et B. K. Tsai, “Theoretical concepts in spectrophotometric measurements,” dans *Experimental Methods in the Physical Sciences*. Elsevier, 2014, vol. 46, p. 11–66.

[138] *Electron Spectroscopy : Some Basic Concepts*. John Wiley Sons, Ltd, 2003, ch. 1, p. 1–15. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/0470867930.ch1

[139] *ESCALAB 250Xi*, Thermo Scientific, 12 2010, 2ème issue.

# ANNEXE A JEUX DE DONNÉES STANDARDISÉS

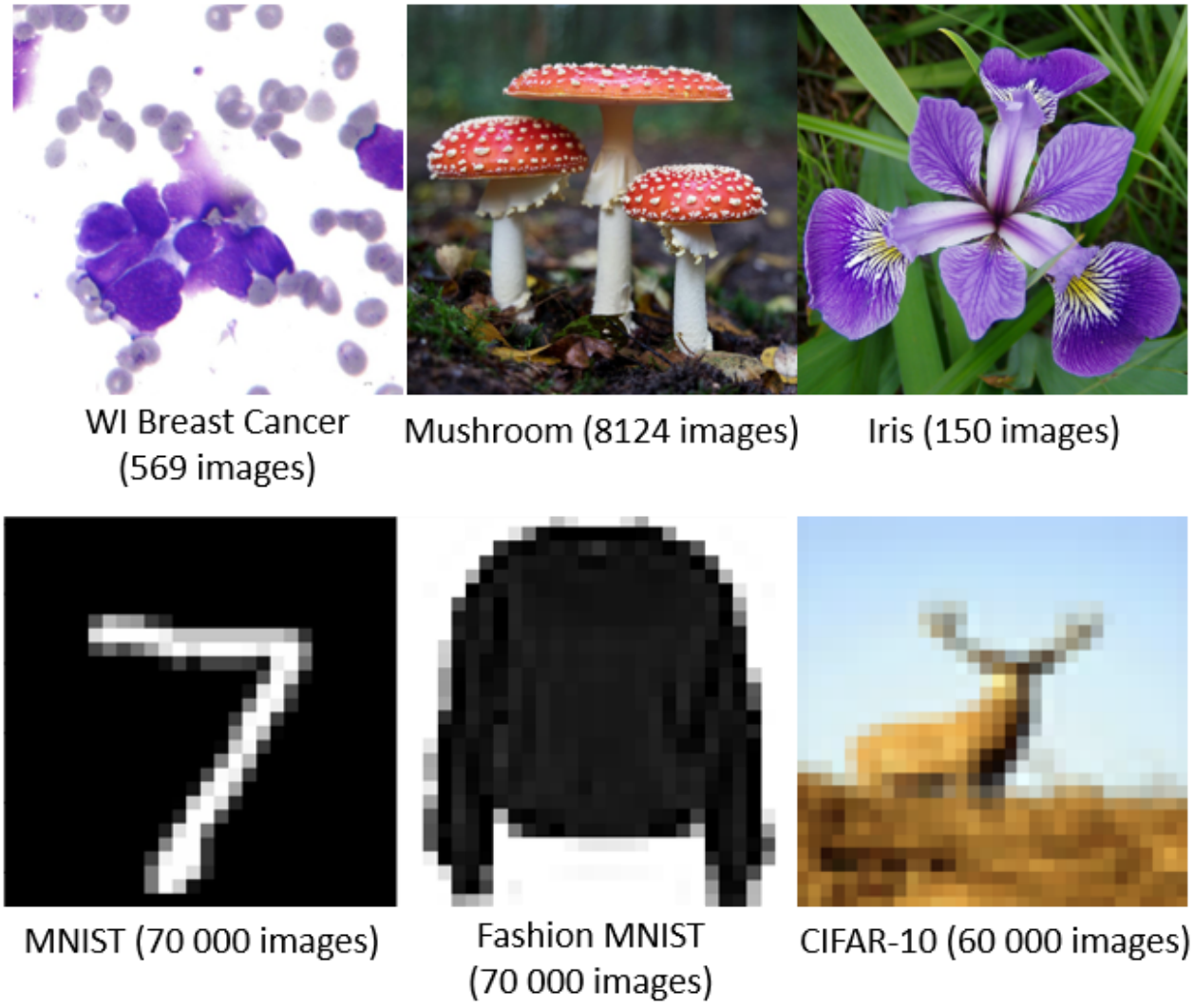



FIGURE A.1 Exemple d'images contenues dans les jeux de données standardisés présentés.

# ANNEXE B MASQUES PHOTOLITHOGRAPHIQUES

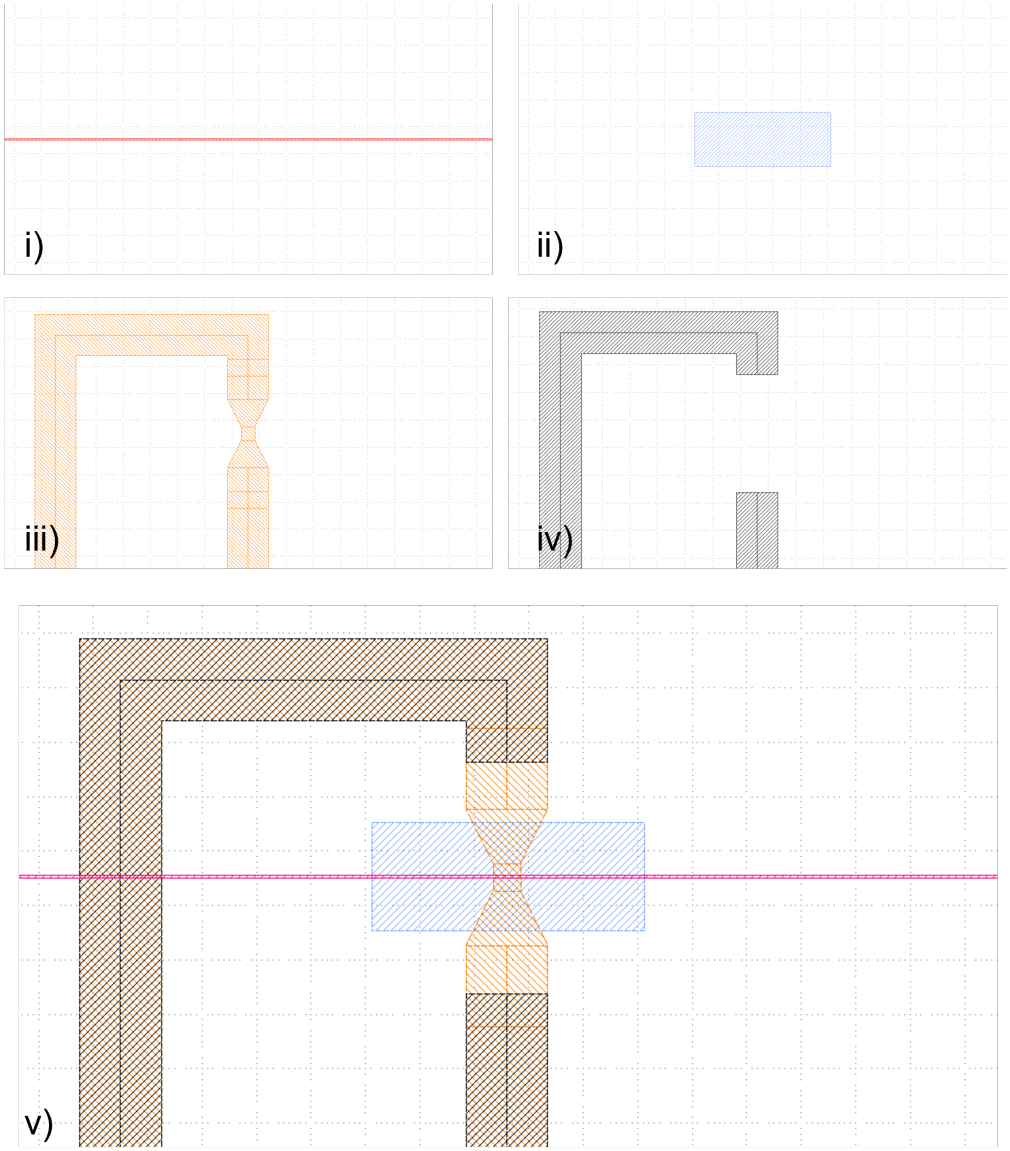



FIGURE B.1 Masques photolithographiques pour i) le guide d'onde fabriqué en fonderie, ii) les cellules de mémoire, iii) le microradiateur, iv) les connexions métalliques, v) le tout superposé.

# ANNEXE C PROCÉDÉ PHOTOLITHOGRAPHIQUE

1. Recuit de déshydratation dans le four YES-TA durant 15 minutes avec apprêt de HDMS
2. Application de résine photosensible à polarité réversible AZ5214-EIR par centrifugation à 4000 rpm pour obtenir une épaisseur de 1.4 µm
3. Recuit à 110 °C durant 50 secondes sur une plaque chauffante
4. Exposition UV de 15 mJ/cm$^2$ avec l'aligneuse MA6 de Karl Suss
5. Recuit d'inversion de polarité à 104 °C durant 120 secondes sur une plaque chauffante
6. Exposition UV complète (*flood exposure*) durant 30 secondes
7. Développement de la photorésine dans le développeur AZ 726 MIF durant 60 secondes